Johan Harlan

ANALISIS DATA SURVEI

Rancangan Sampling Kompleks

Penerbit Gunadarma

ANALISIS DATA SURVEI

Rancangan Sampling Kompleks

Johan Harlan

Penerbit Gunadarma

Analisis Data Survei: Rancangan Sampling Kompleks

Penulis : Johan Harlan

Cetakan Pertama, Januari 2018

Disain cover : Joko Slameto

Diterbitkan pertama kali oleh Gunadarma

Jl. Margonda Raya No. 100, Pondokcina, Depok 16424

Telp. +62-21-78881112, 7863819 Faks. +62-21-7872829

e-mail : sektor@gunadarma.ac.id

KATA PENGANTAR

Survei bukan merupakan sebuah metode penelitian, melainkan merupakan salah satu strategi penelitian. Umumnya survei dilakukan dengan sampel berukuran besar yang ditarik dari populasi yang terdefinisi secara jelas, untuk mengumpulkan data mengenai relatif sedikit variabel, dan diolah terutama secara deskriptif untuk menggambarkan karakteristik populasi. Data survei juga dapat diolah secara analitik, namun uji statistik umumnya bersifat eksploratorik, yaitu lebih ke arah pembentukan hipotesis (*hypothesis generating*) dan tidak bersifat konfirmatorik.

Pelaksanaan survei sesungguhnya mencakup pentahapan yang sangat panjang, mulai dari perencanaan dan penyusunan rancangan penelitian, pengumpulan data, pengolahan dan analisis data, serta pelaporan dan diseminasi hasil penelitian. Dalam buku ini hanya akan dibahas mengenai aspek pengolahan dan analisis data survei. Pengolahan dan analisis data dilakukan dengan menggunakan program komputer statistik Stata 15. Secara umum Stata dikenal sebagai program komputer statistik yang terutama memiliki keunggulan antara lain dalam pengolahan dan analisis data survei.

Pembaca diharapkan telah memiliki pengetahuan dasar tentang Statistika, terutama mengenai analisis regresi dan *Generalized Linear Models*, serta regresi data survival. Setiap saran dan kritik dari pembaca akan dihargai dan diterima demi untuk perbaikan isi buku selanjutnya.

Januari 2018

Johan Harlan

DAFTAR ISI

Kata Pengantar **v**

Daftar Isi **vii**

Bab 1 Pendahuluan **1**

Beberapa Konsep 1

Data Survei 5

Bab 2 Deklarasi Rancangan Survei **7**

Deklarasi untuk Rancangan Multi-Tahap 7

Spesifikasi Deklarasi 7

Contoh 2.1 9

Contoh 2.2 10

Bab 3 Deskripsi Data Survei **13**

Deskripsi Data Survei dengan svydescribe 13

Contoh 3.1 13

Deskripsi Data Survei dengan svy estimation 16

Contoh 3.2 17

Contoh 3.3 20

Bab 4 Tabel dan Grafik untuk Data Survei Tertimbang **23**

Tabel untuk Data Survei Tertimbang 23

Contoh 4.1 23

Grafik untuk Data Survei Tertimbang 27

Contoh 4.2 27

Bab 5 Analisis Regresi Linear 31
Regresi Linear dengan Perintah svy 31
Contoh 5.1 31
Regresi Linear: Postestimasi 34
Contoh 5.2 35

Bab 6 Analisis Regresi Logistik 43
Estimasi Koefisien Regresi 43
Contoh 6.1 43
Estimasi Rasio Odds 45
Contoh 6.2 46

Bab 7 Regresi Logistik Politomi 49
Regresi Logistik Multinomial 49
Contoh 7.1 49
Regresi Logistik Ordinal 55
Contoh 7.2 56

Bab 8 Regresi dengan Respons Data Cacah 61
Regresi Poisson 61
Contoh 8.1 61
Regresi Binomial Negatif 64
Contoh 8.2 64

Bab 9 Regresi Data Survival 69
Model Hazard Proporsional Cox 69
Contoh 9.1 69

Model Survival Parametrik 72
Contoh 9.2 73

Bab 10 Model SEM dan GSEM untuk Data Survei 79
Model SEM untuk Data Survei 79
Contoh 10.1 79
Model GSEM untuk Data Survei 83
Contoh 10.2 84

Kepustakaan 89

Lampiran 1 Bobot Sampling 90

Lampiran 2 Uji *t* untuk Data Survei 94

BAB 1
PENDAHULUAN

❖ Beberapa Konsep

Survei adalah studi observasional, yang umumnya bersifat deskriptif dengan skala besar, untuk mengumpulkan data secara terencana dan sistematik, dengan maksud untuk mengestimasi karakteristik tertentu dalam populasi. Walaupun metodologi survei mencakup dari tahap perencanaan sampai dengan diseminasi hasil survei, di sini hanya akan dibahas mengenai metode analisis data survei yang telah terkumpul.

▪ Rancangan Sampling Kompleks

Dalam teori dasar tentang pengambilan sampel acak, dikenal 4 metode sampling dasar, yaitu sampling acak sederhana, sampling acak stratifikasi, sampling acak kluster, dan sampling acak sistematik. Sebagian besar teori statistika didasarkan atas dasar asumsi sampling acak sederhana, namun dalam pelaksanaan survei sesungguhnya sampling acak sederhana seringkali tidak layak untuk digunakan. Dalam survei sesungguhnya, yang lazim digunakan adalah rancangan sampling kompleks, yang memiliki satu atau lebih di antara fitur berikut:

- Stratifikasi
- Klustering
- Probabilitas seleksi yang tidak sama
- Seleksi multi-tahap

Adanya satu atau lebih fitur di atas menyebabkan prosedur estimasi dan pengujian standar Statistika menjadi tak relevan, dan diperlukan penyesuaian dengan prosedur tersendiri dalam metode survei sampling.

- **Kluster dan Stratum**

Pembahasan metode statistika dalam kepustakaan sebagian besar didasarkan atas asumsi bahwa data berasal dari sampling acak sederhana. Pada pengumpulan data sampel untuk mengestimasi karakteristik populasi penelitian yang besar, sampling acak sederhana praktis sukar dilakukan. Bagi populasi besar demikian, pengumpulan data sampel lazimnya dilakukan dengan metode sampling kompleks, yang pengumpulan datanya menggunakan kluster dan/atau stratifikasi.

Kluster adalah kumpulan individu yang disampel sebagai satu unit dalam tahapan sampling. Kluster yang terpilih dapat mengalami subsampling lagi ataupun diambil seluruh anggotanya menjadi anggota sampel. Pengelompokan individu dalam suatu kluster biasa didasarkan atas kesamaan dalam hal spatial (kota, desa, sekolah, kelas, dan sebagainya).

Stratum adalah subpopulasi yang menjalani sampling secara independen. Pengacakan dilakukan dalam tiap stratum, bukan di dalam populasi secara keseluruhan. Stratifikasi pada metode sampling kompleks dapat dikerjakan tersendiri ataupun secara bersama dengan klustering.

- **Unit Sampling**

Unit sampling dalam metode sampling kompleks adalah kluster atau individu yang dipilih sebagai satuan secara acak dalam suatu tahap sampling. Dalam sampling acak sederhana, tiap unit sampling adalah individu. Dalam metode sampling kompleks multi-tahap, unit sampling dalam tiap tahap sebelum tahap akhir adalah kluster, sedangkan untuk tahap akhir unit samplingnya dapat berupa kluster atau individu.

Unit sampling pada tahap pertama sampling dinamakan unit sampling primer (*primary sampling units*; PSUs). Jika ada tahap kedua, unit sampling pada tahap kedua dinamakan unit sampling sekunder (*secondary sampling units*; SSUs). Secara umum unit sampling pada tahap ke-# dituliskan dengan lambang **su#** (**su1**, **su2**, dan seterusnya).

- **Bobot Sampling dan Efek Desain**

Dalam sampling acak sederhana dengan pengembalian, tiap unit sampling (individu) selalu memiliki probabilitas yang sama untuk terpilih menjadi anggota sampel). Pada metode sampling kompleks, walaupun observasi dipilih secara acak, tiap observasi berbeda memiliki probabilitas berbeda untuk terpilih. Untuk menghindari bias dalam estimasi, harus dilakukan pembobotan terhadap nilai-nilai observasi.

Bobot sampling bernilai sama dengan atau proporsional terhadap kebalikan probabilitas terpilihnya observasi tersebut. Bobot di sini tidak sama dengan kebalikan variansi seperti pada metode Kuadrat Terkecil Tertimbang (*Weighted Least Squares*). Jika subsampel berukuran n_j dipilih dari subpopulasi berukuran N_j, maka probabilitas tiap unit sampling untuk terpilih adalah n_j / N_j. Bobot sampling akan memperkecil variansi dan meningkatkan efisiensi estimator tertimbang.

Jumlah seluruh bobot sampling sama dengan ukuran populasi. Ukuran populasi ini selalu dilaporkan dalam keluaran analisis data survei dengan Stata.

Efek desain (*design effect*; DEFT) adalah rasio antara *standard error* data sampling kluster dengan pembobotan terhadap *standard error* data yang sama yang dihitung dengan asumsi data diperoleh melalui sampling acak sederhana (tanpa pembobotan):

$$\text{DEFT} = \frac{\textit{Standard error} \text{ data tertimbang}}{\textit{Standard error} \text{ data tak terimbang}}$$

Kuadratnya yaitu DEFF, seringkali disebut sebagai efek desain juga:

$$\text{DEFF} = \frac{\text{Variansi data tertimbang}}{\text{Variansi data tak terimbang}}$$

Penggunaan efek desain (estimasinya) antara lain adalah pada perhitungan ukuran sampel minimum yang dibutuhkan sebelum pelaksanaan survei. Jika ukuran sampel yang dibutuhkan dengan sampling acak sederhana adalah *n*, maka ukuran sampel minimum yang dibutuhkan pada rancangan sampling kompleks dengan estimasi efek desain de adalah $\text{de} \times n$.

- **FPC**

Rumus-rumus perhitungan estimator parameter secara tipikal dikembangkan untuk data survei pada sampling dengan pengembalian. Pada sampling tanpa pengembalian, harus dilakukan koreksi dengan faktor yang dinamakan koreksi populasi berhingga (*finite population correction*; FPC). Walaupun demikian, jika rasio ukuran sampel dengan ukuran populasi lebih kecil daripada 5%, beda hasil analisis data survei yang berasal dari sampling tanpa pengembalian dengan sampling dengan pengembalian dapat diabaikan, sehingga FPC tidak perlu digunakan.

Sebagai contoh, diperlihatkan beberapa nilai FPC berikut:

Sample size (*n*)	FPC
1	1.000
10	0.9995
100	0.9950
500	0.9747
1000	0.9487
5000	0.7071
9000	0.3162

- **VCE**

VCE (*variance estimator*) adalah metode yang digunakan untuk estimasi variansi estimator pada rancangan sampling kompleks. Metode yang lazim digunakan, yang juga merupakan metode *default* untuk Stata adalah metode linearisasi Taylor.

Metode estimasi variansi lain yang dapat digunakan yaitu metode *bootstrapping* dan metode *jackknife* tidak dibahas di sini. Selanjutnya semua perhitungan estimasi parameter yang akan dibahas di sini akan menggunakan metode linearisasi Taylor, yang tidak akan dibahas dasar teoretisnya dalam buku ini.

❖ Data Survei

Dalam garis besarnya, perintah Stata untuk analisis data survei dapat dibagi menjadi 4 kelompok, yaitu:

1. Deklarasi dataset menjadi data survei dengan perintah **svyset**.
2. Deskripsi muatan data survei dengan perintah **svydescribe**.
3. Estimasi parameter pada data survei dengan perintah **svy estimation**.
4. Evaluasi kesesuaian model peneliti dengan data survei menggunakan perintah **svy postestimation**.

Secara umum, seluruh perintah Stata untuk menganalisis data survei harus didahului dengan deklarasi dataset menjadi data survei. Deklarasi ini dimaksud untuk menyelaraskan isi dataset dengan beberapa karakteristik data survei untuk memudahkan proses analisis data selanjutnya.

Perintah **svydescribe** untuk data proporsi menghasilkan estimasi proporsi berikut tampilan tabulasi distribusi frekuensi berbagai variabel data proporsi tersebut dalam dataset. Untuk data kontinu akan diperoleh estimasi nilai-nilai tengah, baik secara menyeluruh, per stratum, ataupun per kelompok yang didefinisikan menurut persyaratan tertentu.

Estimasi parameter dengan perintah **svy estimation** mencakup pengestimasian nilai-nilai deskriptif sampel maupun pengestimasian parameter berbagai pemodelan statistik. Estimasi nilai-nilai deskriptif dengan perintah **svy estimation** meliputi rentang ragam yang lebih luas daripada yang dihasilkan dengan perintah **svydescribe** di atas. Estimasi parameter pemodelan mencakup praktis seluruh tipe pemodelan yang ada dalam Statistika dengan asumsi data diperoleh dari proses survei.

Perintah **svy postestimation** diberikan langsung menyusul perintah **svy estimation**, antara lain untuk mengevaluasi kesesuaian model yang dispesifikasikan pada perintah **svy estimation**, berikut beberapa nilai estimasi lainnya.

Secara singkat, disimpulkan bahwa data survei memiliki beberapa karakteristik tertentu, yaitu:

- **Bobot sampling**

  Pada sampel survei, observasi dipilih secara acak, tetapi dengan probabilitas yang berbeda antar observasi. Bobot sama dengan atau proporsional terhadap kebalikan probabilitas terpilihnya observasi tersebut. Berbagai penyesuaian pasca-sampling juga seringkali dilakukan. Bobot w_j untuk rerata observasi ke-*j* berarti observasi ke-*j* tersebut merepresentasikan w_j unsur dalam populasi penarikan sampel.

- **Klustering**

  Pada kebanyakan rancangan survei, individu tidak dipilih secara independen. Kumpulan individu yang dinamakan kluster, dipilih secara berkelompok.

  Dalam kluster dapat dilakukan subsampling. Misalnya pada survei di Kotamadya Jakarta Selatan dapat dipilih kecamatan, lalu kelurahan dalam kecamatan, lalu rumah tangga dalam kelurahan, dan akhirnya individu dalam rumah tangga. Kluster pada sampling tahap pertama dinamakan unit sampling primer (*primary sampling units*; PSUs), dalam contoh ini PSUs adalah kecamatan. Jika tidak ada klustering (sampling 1-tahap), PSUs adalah individu atau dikatakan sebagai kluster berukuran 1.

- **Stratifikasi**

  Berbagai kelompok kluster seringkali disampel secara terpisah. Pengelompokan kluster ini disebut strata. Misalnya dari sebuah propinsi terdapat 53 kota yang dibagi dalam 2 strata, yaitu kota besar (berpenduduk 1 juta atau lebih) dan kota kecil (berpenduduk kurang daripada 1 juta). Dari stratum kota besar dipilih 3 kota dan dari stratum kota kecil dipilih 5 kota. Subsampling dikerjakan secara terpisah pada tiap stratum, dan stratifikasi sudah harus ditetapkan sebelum proses sampling.

BAB 2
DEKLARASI RANCANGAN SURVEI

❖ Deklarasi untuk Rancangan Multi-Tahap

Secara umum, sebelum mengerjakan analisis data survei dengan perintah **svy**, dataset hatus dideklarasikan dulu sebagai data survei dengan perintah **svyset**. Berdasarkan tahapannya, bentuk perintah **svyset** dibagi menjadi:

- **Rancangan 1-tahap**:

  svyset **. . .**
  [spesifikasi tahap 1]

- **Rancangan 2-tahap**:

  svyset **. . .** || **. . .**
  [spesifikasi tahap 1] [spesifikasi tahap 2]

- **Rancangan 3-tahap:**

  svyset **. . .** || **. . .** || **. . .**
  [spesifikasi tahap 1] [spesifikasi tahap 2] [spesifikasi tahap 3]

- Dan seterusnya

Tampak bahwa spesifikasi deklarasi untuk tiap tahapan dipisahkan dengan operator "*or*" (atau) berupa dua palang tegak sejajar ("||").

❖ Spesifikasi Deklarasi

Secara umum, komponen spesifikasi deklarasi adalah:

svyset su# [**pweight=***pweight*], **fpc**(*fpc*) **strata**(*strata*)

su# menyatakan unit sampling tahap ke-**#**; **#** adalah nomor tahapan; **su1** = unit sampling tahap 1, **su2** = unit sampling tahap2, **su3** = unit sampling tahap3, dan seterusnya (**su** = *sampling unit*).

pweight : Variabel pengidentifikasi bobot sampling. *Default*-nya adalah bobot sampling sama dengan 1.
fpc : Variabel pengidentifikasi koreksi populasi berhingga (FPC). FPC hanya digunakan pada sampling tanpa pengembalian.
strata : Variabel pengidentifikasi strata.

Secara lebih rinci, sintaks deklarasi adalah:

- **Desain satu-tahap**:
  svyset [*psu*] [*weight*] [, *options*]

- **Desain multi-tahap**:
  svyset *psu* [*weight*] [, *options*] [| | *ssu*, *options*] . . . [*options*]

psu : Unit sampling primer (*primary sampling unit*), dapat berupa _n atau *varname*. Dalam sintaks satu-tahap, *psu* bersifat optional, *default*-nya adalah _n.
_n menyatakan bahwa individu disampel secara acak jika desain tidak menggunakan sampling kluster.
varname berisi pengidentifikasi untuk kluster pada rancangan sampling kluster.
ssu : *ssu* adalah _n atau *varname* yang mengidentifikasikan unit sampling (kluster) dalam tahap berikutnya pada rancangan survei.
_n menyatakan bahwa individu disampel secara acak pada tahap sampling terakhir.

Opsi:

strata(*varname*) : variabel pengidentifikasi strata
fpc(*varname*) : koreksi populasi berhingga
weight(*varname*) : bobot sampling tingkatan tahap

Setelah **svyset** dispesifikasikan, setiap saat dapat dispesifikasikan **svyset** baru yang akan menggantikan isi spesifikasi **svyset** lama dan membatalkan isi **svyset** lama tersebut.

Contoh 2.1:

. use "D:\Survey\Data\stage5a.dta", clear

Diasumsikan data diperoleh melalui sampling acak sederhana dengan pengembalian (tidak memerlukan FPC). Data tidak memiliki stratifikasi. Pada sampling acak sederhana tidak ada klustering, sehingga PSUs pada tahap pertama adalah individu.

. svyset _n

```
      pweight: <none>
          VCE: linearized
  Single unit: missing
     Strata 1: <one>
         SU 1: <observations>
        FPC 1: <zero>
```

Karena digunakannya sampling acak sederhana, hanya ada 1 tahap pengumpulan data dan sampling langsung dilakukan terhadap individu pada tahap pertama ini, dinyatakan dengan **_n** pada deklarasi data survei.

svyset menyatakan bahwa metode estimasi variansi di sini adalah vce(linearized) yang menggunakan linearisasi Taylor dan merupakan metode *default* untuk estimasi variansi. Metode lain yang kurang lazim misalnya adalah vce(bootstrap) yang menggunakan estimasi variansi *bootstrap* dan vce(jackknife) yang menggunakan estimasi variansi *jackknife*. **svyset** juga akan melaporkan nilai kosong (*missing values*) dalam dataset jika ada. Dengan sampling acak sederhana tidak ada stratifikasi, sehingga **strata** = 1 dan PSUs adalah langsung berupa observasi individual. FPC tidak ada karena digunakan sampling dengan pengembalian.

Contoh 2.2: Data survei multi-tahap

Dimiliki data fiktif tentang siswa SMA di AS (kelas 12) dengan pengumpulan data 2-tahap. Tahap pertama, *counties* dipilih secara independen dalam tiap negara bagian. Tahap kedua, sekolah dipilih di tiap *county* yang terpilih. Di tiap sekolah yang terpilih, dibagikan kuesioner untuk diisi oleh tiap siswanya. Data disimpan dalam dataset bernama **multistage.dta**. Variabel rancangan survei adalah:

- state : memuat identifikasi stratum.
- county : memuat unit sampling tahap-pertama.
- ncounties : memuat jumlah *counties* dalam tiap negara bagian.
- school : memuat unit sampling tahap-kedua.
- nschools : memuat jumlah sekolah dalam tiap *county*.
- sampwgt : memuat bobot sampling untuk tiap individu yang terpilih.

Selanjutnya dataset dimasukkan ke dalam memori dan dideklarasikan sebagai data survei dengan perintah **svyset**:

```
. use "D:\Survey\Data\multistage.dta", clear
. svyset county [pw=sampwgt], strata(state) fpc(ncounties) ||
  school, fpc(nschools)

      pweight: sampwgt
          VCE: linearized
  Single unit: missing
     Strata 1: state
         SU 1: county
        FPC 1: ncounties
     Strata 2: <one>
         SU 2: school
        FPC 2: nschools
```

Perintah berikut memperlihat ringkasan data (summarize) untuk pweight (= sampwgt), strata (= state), dan psu (= county).

. summ sampwgt state county

```
Variable |   Obs      Mean  Std. Dev.      Min       Max
---------+--------------------------------------------------
 sampwgt | 4,071  1965.119  1261.919  350.9584  10387.62
   state | 4,071  25.57799   14.2365         1        50
  county | 4,071  1.473348  .4993505         1         2
```

Tampak bahwa survei dilakukan di 50 negara bagian dengan *state* sebagai **strata** untuk starifikasi. Data survei dikumpulkan dalam 2-tahap, untuk tahap pertama PSUs adalah *counties* dan untuk tahap kedua SSUs (*secondary sampling units*) adalah *schools*. Selanjutnya di sekolah tidak ada proses sampling dan semua siswa di sekolah yang terpilih dijadikan anggota sampel. FPC untuk tahap pertama adalah ncounties dan FPC untuk tahap kedua adalah nschools.

Data disimpan (di-*saved*) dengan nama file **highschool.dta**.

. save highschool

```
file highschool.dta saved
```

BAB 3
DESKRIPSI DATA SURVEI

❖ Deskripsi Data Survei dengan svydescribe

Perintah **svydescribe** adalah untuk mendeskripsikan data survei. Perintah ini selalu harus didahului dengan deklarasi data survei **svyset.**

Sintaks:

svydescribe [*varlist*] [*if*] [*in*] [, *options*]

varlist : Daftar variabel yang diingini deskripsinya.

Opsi:

stage(#) : Tahapan sampling yang diinginkan deskripsinya, *default*-nya adalah *stage*(1).

finalstage : Menampilkan informasi per unit sampling pada tahap terakhir.

Contoh 3.1:

. use "D:\Survey\Data\nhanes2b.dta", clear

Tanpa menggunakan perintah **svy**, dengan perintah standar Stata **describe** diperoleh:

. describe psuid finalwgt stratid

```
variable  storage  display value
    name    type   format  label variable label
-------------------------------------------------------------
psuid       byte   %9.0g          primary sampling unit, 1 or 2
finalwgt    long   %9.0g          sampling weight (except lead)
stratid     byte   %9.0g          stratum identifier, 1-32
```

Terdapat 32 strata (variabel **stratid**). PSUs adalah **psuid**, dipilih secara acak di tiap stratum, lalu semua individu anggota **psuid** yang terpilih dijadikan anggota sampel.

. svyset psuid [pweight=finalwgt], strata(stratid)

```
      pweight: finalwgt
          VCE: linearized
  Single unit: missing
     Strata 1: stratid
         SU 1: psuid
        FPC 1: <zero>
```

. svydescribe

```
Survey: Describing stage 1 sampling units

      pweight: finalwgt
          VCE: linearized
  Single unit: missing
     Strata 1: stratid
         SU 1: psuid
        FPC 1: <zero>
```

			#Obs per Unit		
Stratum	#Units	#Obs	min	mean	max
1	2	380	165	190.0	215
2	2	185	67	92.5	118
3	2	348	149	174.0	199
4	2	460	229	230.0	231
5	2	252	105	126.0	147

..	.	...	...	.....	...
28	2	299	136	149.5	163
29	2	503	215	251.5	288
30	2	365	166	182.5	199
31	2	308	143	154.0	165
32	2	450	211	225.0	239
31	62	10,351	67	167.0	288

. svydescribe hdresult, finalstage

```
Survey: Describing final stage sampling units

      pweight: finalwgt
          VCE: linearized
  Single unit: missing
     Strata 1: stratid
         SU 1: psuid
        FPC 1: <zero>
```

Stratum	Unit	#Obs with complete data	#Obs with missing data
1	1	0	215
1	2	114	51
2	1	98	20
2	2	0	67
3	1	161	38
3	2	116	33
4	1	160	71
4	2	180	49

```
         5          1         92         55
         5          2         81         24
        ..          .        ...         ..
        30          1        147         19
        30          2        179         20
        31          1        121         22
        31          2        158          7
        32          1        180         59
        32          2        203          8
--------  --------  --------  --------
        31         62      8,720      1,631
                             ------------------
                                    10,351
```

❖ Deskripsi Data Survei dangan svy estimation

Setelah deklarasi dataset menjadi data survei dengan **svyset**, selanjutnya sejumlah besar perintah Stata dapat dilakukan terhadap dataset tersebut dengan *prefix* **svy**. Bentuk seperti ini selanjutnya disebut perintah **svy**. Beberapa permintaan tampilan deskripsi data lain yang dapat diperoleh dengan perintah **svy** sintaks-nya adalah sebagai berikut:

svy: mean *cont_varlist*

svy: proportion *cat_varlist*

svy: ratio *var1/var2*

svy: total *varlist*

cont_varlist : Variabel kontinu

cat_varlist : Variabel kategorik

Tabulasi 1-arah:

svy: tabulate *varname* [, *options*]

Tabulasi 2-arah:

svy: tabulate *varname1 varname2* [, *options*]

Opsi:

count	:	cacah tertimbang (*weighted count*)
se	:	*standard error*
ci	:	interval konfidensi
deff	:	efek desain
obs	:	observasi sel

Contoh 3.2:

. use "D:\Survey\Data\nmihs.dta", clear

. svyset [pweight=finwgt], strata(stratan)

```
      pweight: finwgt
          VCE: linearized
  Single unit: missing
     Strata 1: stratan
         SU 1: <observations>
        FPC 1: <zero>
```

. svy: mean age birthwgt

```
(running mean on estimation sample)

Survey: Mean estimation

Number of strata =     6        Number of obs   =      9,946
Number of PSUs   = 9,946        Population size =  3,895,562
                                Design df       =      9,940
```

```
------------------------------------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   26.28983   .0776105       26.1377    26.44196
    birthwgt |   3355.452   6.402741      3342.902    3368.003
------------------------------------------------------------------------------
```

Tanpa pembobotan diperoleh:

. mean age birthwgt

```
Mean estimation                 Number of obs    =      9,946

------------------------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    25.6106   .0580008       25.4969    25.72429
    birthwgt |   2845.094   9.861422      2825.764    2864.424
------------------------------------------------------------------------------
```

Tampak bahwa bobot dapat memperkecil ataupun memperbesar *standard error*.

. svy: proportion miscar

```
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =     6        Number of obs    =      9,953
Number of PSUs   = 9,953        Population size =  3,898,922
                                 Design df        =      9,947
```

```
------------------------------------------------------------
             |            Linearized           Logit
             | Proportion   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------
miscar       |
    nomiscar |   .8388045   .0051787      .8283932    .8486996
      miscar |   .1611955   .0051787      .1513004    .1716068
------------------------------------------------------------
```

. svy: ratio age/race

```
(running ratio on estimation sample)

Survey: Ratio estimation

Number of strata =     6        Number of obs   =      9,953
Number of PSUs   = 9,953        Population size =  3,898,922
                                 Design df       =      9,947

     _ratio_1: age/race

------------------------------------------------------------
             |            Linearized
             |      Ratio   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------
    _ratio_1 |   153.9555   .3941266      153.1829     154.728
------------------------------------------------------------
```

. svy: total highbp

```
(running total on estimation sample)

Survey: Total estimation

Number of strata =     6        Number of obs    =      9,953
Number of PSUs   = 9,953        Population size =  3,898,922
                                 Design df       =      9,947

------------------------------------------------------------------
             |             Linearized
             |      Total   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------
      highbp |   186196.7    11286.4      164073.1    208320.3
------------------------------------------------------------------
```

Contoh 3.3:

. use "D:\Survey\Data\nhanes2b.dta", clear

. svyset psuid [pweight=finalwgt], strata(stratid)

. svy: tabulate race, count ci deff

```
(running tabulate on estimation sample)

Number of strata   = 31          Number of obs    =      10,351
Number of PSUs     = 62          Population size = 117,157,513
                                 Design df       =          31
```

```
------------------------------------------------------------
1=white,   |
2=black,   |
3=other    |      count          lb          ub        deff
-----------+------------------------------------------------
    White  |    1.0e+08     9.7e+07     1.1e+08       60.19
    Black  |    1.1e+07     8.2e+06     1.4e+07       18.58
    Other  |    3.0e+06     4.1e+05     5.5e+06       47.87
           |
    Total  |    1.2e+08
------------------------------------------------------------
  Key:  count =  weighted count
        lb    =  lower 95% confidence bound for weighted count
        ub    =  upper 95% confidence bound for weighted count
        deff  =  deff for variance of weighted count
```

BAB 4
TABEL DAN GRAFIK UNTUK DATA SURVEI TERTIMBANG

❖ Tabel untuk Data Survei Tertimbang

Tabel yang dihasilkan dengan perintah **svy** di sini berbeda dengan tabel yang dihasilkan oleh perintah standar Stata **tab**, karena frekuensi atau persentase sel yang didapatkan dengan perintah **svy** memperhitungkan pembobotan **pweight** pada deklarasi data survei.

Dengan Stata prosedurnya adalah sebagai berikut.

Sintaks:

svy: tab *cat_var* [, *options*]

Opsi:

percent	:	Persentase sel
count	:	Frekuensi cacah sel
se	:	*Linearized standard error*
ci	:	Interval konfidensi

Contoh 4.1:

```
. use "D:\Survey\Data\nmihs.dta", clear
. svyset [pweight=finwgt], strata(stratan)

      pweight: finwgt
          VCE: linearized
  Single unit: missing
     Strata 1: stratan
         SU 1: <observations>
        FPC 1: <zero>
```

. svy: tab agegrp, percent

```
(running tabulate on estimation sample)

Number of strata =     6          Number of obs   =     9,953
Number of PSUs   = 9,953          Population size = 3,898,922
                                  Design df       =     9,947

     ----------------------
     Age         |
     groups      |
     1-5         | percentage
     ------------+-----------
      age15-19   |      12.26
      age20-24   |      27.38
      age25-29   |      31.74
      age30-34   |      20.56
        age35+   |      8.045
                 |
           Total |        100
     ----------------------
  Key:  percentage  =  cell percentage
```

. svy: tab agegrp, count

```
(running tabulate on estimation sample)

Number of strata =     6          Number of obs   =     9,953
Number of PSUs   = 9,953          Population size = 3,898,922
                                  Design df       =     9,947
```

```
     ----------------------
     Age         |
     groups      |
     1-5         |      count
     ----------+-----------
      age15-19 |    4.8e+05
      age20-24 |    1.1e+06
      age25-29 |    1.2e+06
      age30-34 |    8.0e+05
        age35+ |    3.1e+05
               |
         Total |    3.9e+06
     ----------------------
  Key:  count     =  weighted count
```

. svy: tab agegrp, obs percent se ci

```
(running tabulate on estimation sample)

Number of strata =     6         Number of obs   =     9,953
Number of PSUs   = 9,953         Population size = 3,898,922
                                 Design df       =     9,947

-------------------------------------------------------------
Age         |
groups      |
1-5         | percentage       se       lb       ub      obs
----------+--------------------------------------------------
 age15-19 |      12.26    .4583    11.39    13.19     1653
 age20-24 |      27.38    .6227    26.18    28.62     2866
 age25-29 |      31.74    .6541    30.47    33.04     2848
```

```
 age30-34 |      20.56     .561    19.49    21.69     1851
   age35+ |      8.045     .383    7.326    8.829      735
          |
    Total |        100                              9953
-------------------------------------------------------------

Key:  percentage  =  cell percentage
      se          =  linearized standard error of cell percentage
      lb          =  lower 95% confidence bound for cell percentage
      ub          =  upper 95% confidence bound for cell percentage
      obs         =  number of observations
```

Tanpa pembobotan akan diperoleh hasil yang berbeda sebagai berikut.

. tab agegrp

```
 Age groups |
        1-5 |      Freq.     Percent        Cum.
------------+-----------------------------------
   age15-19 |      1,653       16.61       16.61
   age20-24 |      2,866       28.80       45.40
   age25-29 |      2,848       28.61       74.02
   age30-34 |      1,851       18.60       92.62
     age35+ |        735        7.38      100.00
------------+-----------------------------------
      Total |      9,953      100.00
```

❖ Grafik untuk Data Survei Tertimbang

Seperti halnya dengan perintah **svy** untuk tabulasi data survei tertimbang, dengan grafik juga diperlukan penyesuaian untuk memperoleh hasil grafik data tertimbang. Perintah Stata adalah sebagai berikut.

Sintaks:

graph bar [aweight = *weight*]**, over**(*cat_var*)

graph hbar [aweight = *weight*]**, over**(*cat_var*)

graph bar [aweight = *weight*]**, over**(*cat_var1*) **by**(*cat_var2*)

cat_var : Variabel kategorik

Perintah **graph bar** tidak dapat menerima pembobotan **pweight**. (*probability weight*), yaitu pembobotan yang proporsional terhadap kebalikan probabilitas terpilihnya suatu observasi, yang merupakan pembobotan yang biasa digunakan pada hampir semua perintah **svy**.

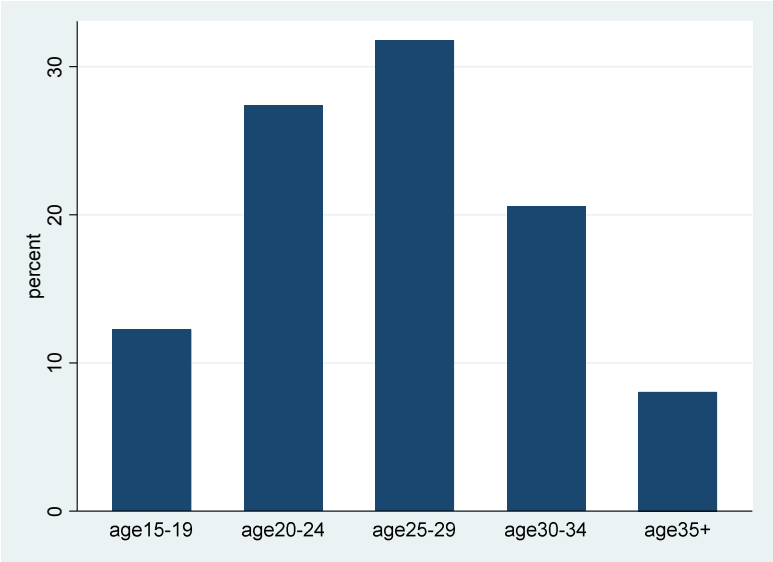
aweight (*analytical weight*) adalah pembobotan pada regresi *weighted least squares*, tetapi untuk **graph bar** tertimbang dapat digunakan karena efek visualnya sesuai dengan hasil persentase **svy: tab**.

Contoh 4.2:

```
. use "D:\Survey\Data\nmihs.dta", clear
. svyset [pweight=finwgt], strata(stratan)
```

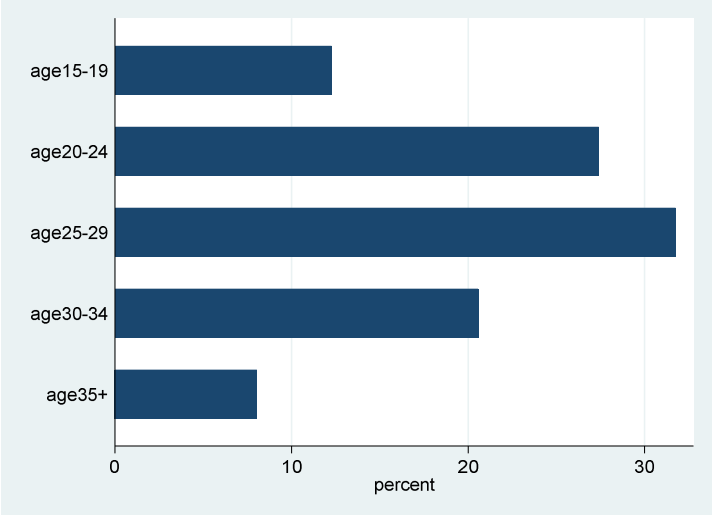
Akan dibuat grafik batang tertimbang untuk kelompok usia.

```
. graph bar [aweight = finwgt], over(agegrp)
```



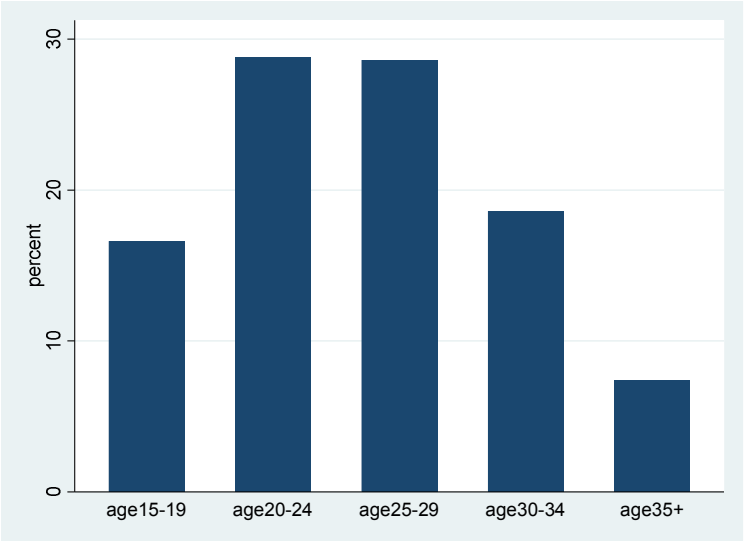
Grafik batang dengan batang horizontal adalah sebagai berikut.

. **graph hbar [aweight = finwgt], over(agegrp)**



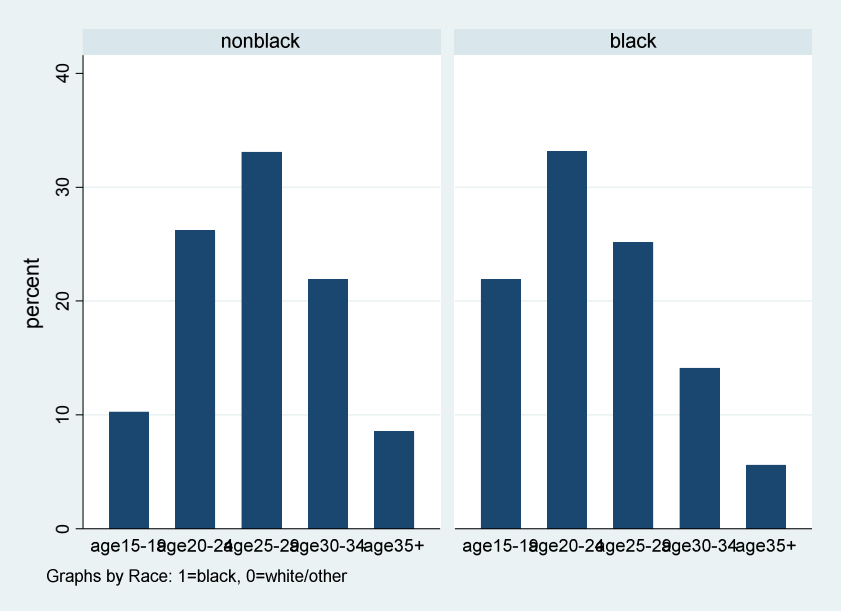
Seandainya tidak dilakukan penyesuaian untuk pembobotan data survei, akan diperoleh grafik yang berbeda:

. **graph bar, over(agegrp)**



Grafik tertimbang kelompok usia menurut ras adalah sebagai berikut:

. graph bar [aweight = finwgt], over(agegrp) by(race)



BAB 5
ANALISIS REGRESI LINEAR

❖ Regresi Linear dengan Perintah svy

Mulai bab ini akan dibahas beberapa uji statistik untuk data survei tertimbang. Perintah **svy** tidak dapat menerima perintah untuk uji statistik sederhana seperti uji *t* dan uji khi-kuadrat. Di sini akan dibahas prosedur analisis regresi untuk data survei tertimbang dengan Stata.

Model regresi linear adalah:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Perintah standar Stata untuk mengestimasi parameter model ini adalah:

regress *depvar indepvars* [*if*] [*in*] [, *options*]

Dengan data survei, diperlukan penyesuaian sebagai berikut.

Sintaks:

svy: regress *depvar indepvars* [, *options*]

depvar : Dependen variabel

indepvars : Variabel independen

Hasil analisis regresi dengan perintah **svy** ini tidak sama dengan hasil analisis regresi dengan perintah standar **regress**.

Contoh 5.1:

```
. use "D:\Survey\Data_complete\nhanes2d.dta"
. svyset
. label list
```

```
highlead:
           0 lead<25
           1 lead>=25
agegrp:
           1 age20-29
           2 age30-39
           3 age40-49
           4 age50-59
           5 age60-69
           6 age 70+
race:
           1 White
           2 Black
           3 Other
sex:
           1 Male
           2 Female
region:
           1 NE
           2 MW
           3 S
           4 W
```

. svyset

```
      pweight: finalwgt
          VCE: linearized
  Single unit: missing
     Strata 1: strata
         SU 1: psu
        FPC 1: <zero>
```

Regresi dengan data tertimbang adalah:

. svy: regress bpdiast age i.race diabetes

```
(running regress on estimation sample)
```

```
Survey: Linear regression

Number of strata = 31                  Number of obs   =      10,349
Number of PSUs   = 62                  Population size = 117,131,111
                                       Design df       =          31
                                       F(   4,     28) =      340.62
                                       Prob > F        =      0.0000
                                       R-squared       =      0.0852

------------------------------------------------------------------------------
             |             Linearized
     bpdiast |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .2334752   .0060689    38.47   0.000     .2210975     .2458528
             |
        race |
       Black |    2.90551   .6030235     4.82   0.000     1.675635     4.135384
       Other |   .8202038   .6666638     1.23   0.228    -.5394659     2.179873
             |
    diabetes |   1.422486   .8085378     1.76   0.088    -.2265377      3.07151
       _cons |   70.80624    .561676   126.06   0.000     69.6607     71.95179
------------------------------------------------------------------------------
```

Tanpa menggunakan pembobotan, dengan perintah standar **regress** akan diperoleh hasil yang berbeda:

. regress bpdiast age i.race diabetes

```
      Source |       SS       df       MS              Number of obs = 10,349
-------------+------------------------------           F(4, 10344)   = 213.75
       Model |  132038.73      4  33009.6824           Prob > F      = 0.0000
    Residual | 1597440.33 10,344  154.431586           R-squared     = 0.0763
```

```
---------+------------------------------   Adj R-squared = 0.0760
   Total | 1729479.06 10,348 167.131722   Root MSE      = 12.427

------------------------------------------------------------------------------
 bpdiast |    Coef. Std. Err.    t   P>|t|  [95% Conf. Interval]
---------+--------------------------------------------------------------------
     age | .1991704 .0072122  27.62  0.000   .1850331   .2133077
         |
    race |
  Black  | 3.335804 .3999102   8.34  0.000   2.551903   4.119706
  Other  | .6896976  .888784   0.78  0.438  -1.052491   2.431886
         |
diabetes | .6701793 .5797047   1.16  0.248  -.4661539   1.806513
   _cons | 71.84372 .3656954 196.46  0.000   71.12689   72.56056
------------------------------------------------------------------------------
```

❖ Regresi Linear: Postestimasi

Seperti pada perintah standar Stata, dengan perintah **svy** pun analisis regresi dapat dilanjutkan dengan *postestimation statistics*. Di sini hanya akan dibahas beberapa perintah *postestimation statistics*, sebagian di antaranya tidak untuk analisis regresi.

Sintaks:

Menampilkan efek desain untuk estimasi titik:

estat effects

Menghitung ukuran subpopulasi:

estat size

Mengestimasi standar deviasi subpopulasi:

estat sd

Menghitung koefisien variasi data survei:

estat cv

Uji kebaikan-suai model regresi logistik untuk data survei:

estat gof

Contoh 5.2:

. use "D:\Survey\Data\nhanes2.dta", clear

. svyset

```
      pweight: finalwgt
          VCE: linearized
  Single unit: missing
     Strata 1: strata
         SU 1: psu
        FPC 1: <zero>
```

Regresi **bpdiast** secara bersama terhadap **age**, **sex**, **tcresult**, dan **tgresult** untuk subpopulasi **female** = 1 tanpa pembobotan adalah:

. regress bpdiast age sex tcresult tgresult if female==1

```
note: sex omitted because of collinearity

  Source |       SS       df       MS      Number of obs =   2,609
---------+------------------------------   F(3, 2605)    =  122.96
   Model | 56156.2492     3 18718.7497   Prob > F      =  0.0000
Residual | 396577.626 2,605 152.237092   R-squared     =  0.1240
---------+------------------------------   Adj R-squared =  0.1230
   Total | 452733.875 2,608 173.594277   Root MSE      =  12.338
```

```
------------------------------------------------------------------------------
 bpdiast |    Coef.  Std. Err.   t   P>|t| [95% Conf. Interval]
---------+--------------------------------------------------------------------
     age | .2041154  .0161123 12.67  0.000   .1725213  .2357095
     sex |        0 (omitted)
tcresult | .0054392  .0059577  0.91  0.361  -.0062432  .0171215
tgresult | .0253385  .0035005  7.24  0.000   .0184746  .0322025
   _cons | 64.97445  1.111928 58.43  0.000    62.7941   67.1548
------------------------------------------------------------------------------
```

Dengan perintah **svy** diperoleh hasil berbeda:

. svy, subpop(female): regress bpdiast age female tcresult tgresult

```
(running regress on estimation sample)

Survey: Linear regression

Number of strata = 31                     Number of obs   =      7,524
Number of PSUs   = 62                     Population size = 85,220,634
                                          Subpop. no. obs =      2,609
                                          Subpop. size    = 29,061,154
                                          Design df       =         31
                                          F(   3,     29) =     117.73
                                          Prob > F        =     0.0000
                                          R-squared       =     0.1374

------------------------------------------------------------------------------
         |             Linearized
 bpdiast |     Coef.  Std. Err.    t   P>|t| [95% Conf. Interval]
---------+--------------------------------------------------------------------
     age |  .2218233  .0162074  13.69  0.000   .1887681  .2548784
  female |         0 (omitted)
```

```
tcresult |   .0204581   .0082171     2.49   0.018    .0036992    .037217
tgresult |   .0210723   .0055222     3.82   0.001    .0098097   .0323349
   _cons |   61.81568   1.461041    42.31   0.000    58.83586   64.79549
------------------------------------------------------------------------------
```

Efek desain adalah sebagai berikut:

. estat effects

```
--------------------------------------------------------------
             |             Linearized
     bpdiast |      Coef.   Std. Err.       DEFF       DEFT
-------------+------------------------------------------------
         age |   .2218233   .0162074     .816386   .903541
      female |  (omitted)
    tcresult |   .0204581   .0082171     1.64125   1.28111
    tgresult |   .0210723   .0055222     1.80948   1.34517
       _cons |   61.81568   1.461041     1.76207   1.32743
--------------------------------------------------------------
```

Efek desain dinyatakan dengan DEFF dan DEFT. DEFF adalah rasio antara estimasi variansi suatu parameter hasil perintah **svy** dengan estimasi hasil perintah standar dengan sampling acak sederhana. DEFT adalah akar DEFF, yaitu rasio antara kedua *standard error*. Definisi lain untuk DEFF yaitu:

$$\text{DEFF} = 1 + \rho\,(n - 1)$$

Secara *default*, DEFF dan DEFT dihitung dengan asumsi sampling acak sederhana dikerjakan terhadap seluruh populasi. Jika diasumsikan bahwa sampling acak sederhana hanya dikerjakan terhadap subpopulasi, perintahnya adalah:

. estat effects, srssubpop

```
------------------------------------------------------------
             |             Linearized
     bpdiast |      Coef.   Std. Err.       DEFF      DEFT
-------------+----------------------------------------------
         age |   .2218233   .0162074     .829935   .911008
      female |  (omitted)
    tcresult |   .0204581   .0082171     1.66849    1.2917
    tgresult |   .0210723   .0055222     1.83951   1.35628
       _cons |   61.81568   1.461041     1.79131    1.3384
------------------------------------------------------------
```

Berikut diperlihatkan prosedur *postestimation* terhadap perintah **svy: mean**.

. svy: mean tcresult tgresult

```
(running mean on estimation sample)

Survey: Mean estimation

Number of strata =      31       Number of obs   =       5,050
Number of PSUs   =      62       Population size =  56,820,832
                                 Design df       =          31

--------------------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
    tcresult |   211.3975   1.252274      208.8435    213.9515
    tgresult |    138.576   2.071934      134.3503    142.8018
--------------------------------------------------------------
```

. estat effects, deff deft

```
------------------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.       DEFF       DEFT
-------------+----------------------------------------------
    tcresult |   211.3975   1.252274     3.57141   1.88982
    tgresult |    138.576   2.071934     2.35697   1.53524
------------------------------------------------------------
```

Sebagai kelanjutan dari perintah **svy: mean** dapat diestimasi ukuran subpopulasi.

. estat size

```
------------------------------------------------------------------
         |             Linearized
         |       Mean   Std. Err.         Obs            Size
---------+--------------------------------------------------------
tcresult |   211.3975   1.252274       5,050      56,820,832
tgresult |    138.576   2.071934       5,050      56,820,832
------------------------------------------------------------------
```

Estimasi rerata **tcresult** untuk masing-masing kelompok jenis kelamin adalah:

. svy: mean tcresult, over(sex)

```
(running mean on estimation sample)

Survey: Mean estimation
```

```
Number of strata =      31      Number of obs    =       10,351
Number of PSUs   =      62      Population size  =  117,157,513
                                Design df        =           31

         Male: sex = Male
       Female: sex = Female

------------------------------------------------------------------------------
             |             Linearized
        Over |       Mean   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------------------
tcresult     |
        Male |   210.7937   1.312967      208.1159    213.4715
      Female |   215.2188   1.193853       212.784    217.6537
------------------------------------------------------------------------------
```

Estimasi ukuran subpopulasinya masing-masing adalah:

. estat size

```
         Male: sex = Male
       Female: sex = Female

------------------------------------------------------------------------------
          |             Linearized
     Over |       Mean   Std. Err.           Obs          Size
----------+-------------------------------------------------------------------
tcresult  |
     Male |   210.7937   1.312967         4,915    56,159,480
   Female |   215.2188   1.193853         5,436    60,998,033
------------------------------------------------------------------------------
```

Estimasi standar deviasi kolesterol serum masing-masing subpopulasi adalah:

. estat sd

```
         Male: sex = Male
       Female: sex = Female

------------------------------------
        Over |       Mean   Std. Dev.
-------------+----------------------
tcresult     |
        Male |   210.7937    45.79065
      Female |   215.2188    50.72563
------------------------------------
```

Tanpa pembobotan, estimasi standar deviasi kolesterol serum masing-masing subpopulasi yaitu:

. tab sex, sum(tcresult)

```
            |    Summary of serum cholesterol
    1=male, |                (mg/dL)
   2=female |        Mean   Std. Dev.       Freq.
------------+------------------------------------
       Male |   213.17314   45.979654       4,915
     Female |   221.73528    51.94704       5,436
------------+------------------------------------
      Total |   217.66969   49.386943      10,351
```

Estimasi koefisien variasi untuk data survei adalah:

```
. estat cv

          Male: sex = Male
        Female: sex = Female

------------------------------------------------
             |             Linearized
        Over |       Mean   Std. Err.     CV (%)
-------------+----------------------------------
tcresult     |
        Male |   210.7937   1.312967     .622868
      Female |   215.2188   1.193853     .554716
------------------------------------------------
```

BAB 6
ANALISIS REGRESI LOGISTIK

❖ Estimasi Koefisien Regresi

Regresi logistik adalah regresi dengan respons biner. Model regresinya adalah:

$$\text{logit } y_i = \ln \frac{y_i}{1-y_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots$$

Perintah standar Stata untuk mengestimasi parameternya adalah:

logit *depvar indepvars* [*if*] [*in*] [, *options*]

Untuk data survei diperlukan penyesuaian sebagai berikut.

Sintaks:

svy: logit *depvar indepvars* [, *options*]

Dengan perintah **svy** ini akan diperoleh hasil yang berbeda.

Contoh 6.1:

```
. use "D:\Survey\Data_complete\nhanes2d.dta"
. svyset
```

Estimasi koefisien regresinya adalah:

```
. svy: logit highbp height weight age age2 female
(running logit on estimation sample)

Survey: Logistic regression
```

```
Number of strata = 31                  Number of obs   =       10,351
Number of PSUs   = 62                  Population size =  117,157,513
                                       Design df       =           31
                                       F(   5,     27) =       284.33
                                       Prob > F        =       0.0000

------------------------------------------------------------------------------
             |             Linearized
      highbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |   -.034962   .0053467    -6.54   0.000    -.0458667   -.0240573
      weight |    .051559   .0025105    20.54   0.000     .0464389    .0566791
         age |    .054326   .0126512     4.29   0.000     .0285238    .0801282
        age2 |  -.0000601   .0001379    -0.44   0.666    -.0003413    .0002211
      female |  -.4682471   .0582887    -8.03   0.000    -.5871278   -.3493665
       _cons |  -.4251878   .8788481    -0.48   0.632     -2.21761    1.367235
------------------------------------------------------------------------------
```

Tanpa penyesuaian untuk data survei, dengan perintah standar Stata akan diperoleh hasil yang berbeda:

. logit highbp height weight age age2 female

```
Iteration 0:   log likelihood = -7050.7655
Iteration 1:   log likelihood = -5855.0582
Iteration 2:   log likelihood = -5839.5092
Iteration 3:   log likelihood = -5839.4541
Iteration 4:   log likelihood = -5839.4541
```

```
Logistic regression                             Number of obs =  10,351
                                                LR chi2(5)    = 2422.62
                                                Prob > chi2   =  0.0000
Log likelihood = -5839.4541                     Pseudo R2     =  0.1718

------------------------------------------------------------------------------
      highbp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |  -.0356301   .0036617    -9.73   0.000    -.0428068   -.0284534
      weight |   .0499014   .0018444    27.06   0.000     .0462864    .0535163
         age |   .0520305   .0103084     5.05   0.000     .0318265    .0722346
        age2 |  -.0000539   .0001076    -0.50   0.617    -.0002648     .000157
      female |  -.3770696   .0642985    -5.86   0.000    -.5030924   -.2510469
       _cons |    -.16185   .6471727    -0.25   0.803    -1.430285    1.106585
------------------------------------------------------------------------------
```

❖ Estimasi Rasio Odds

Estimasi rasio odds untuk model regresi logistik dengan perintah standar Stata adalah:

logistic *depvar indepvars* [*if*] [*in*] [, *options*]

Untuk data survei, diperlukan penyesuaian sebagai berikut.

Sintaks:

svy: logistic *depvar indepvars* [, *options*]

Estimasi rasio odds dengan perintah **svy** ini akan memberi hasil berbeda.

Contoh 6.2:

. use "D:\Survey\Data_complete\nhanes2d.dta"

. svyset

. svy: logistic highbp height weight age age2 female

```
(running logistic on estimation sample)

Survey: Logistic regression

Number of strata = 31          Number of obs   =       10,351
Number of PSUs   = 62          Population size =  117,157,513
                               Design df       =           31
                               F(   5,     27) =       284.33
                               Prob > F        =       0.0000

------------------------------------------------------------------------------
             |             Linearized
      highbp | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |   .9656421    .005163    -6.54   0.000     .9551693    .9762298
      weight |   1.052911   .0026433    20.54   0.000     1.047534    1.058316
         age |   1.055829   .0133575     4.29   0.000     1.028935    1.083426
        age2 |   .9999399   .0001379    -0.44   0.666     .9996588    1.000221
      female |   .6260988   .0364945    -8.03   0.000     .5559217    .7051347
       _cons |    .653647   .5744564    -0.48   0.632      .108869    3.924483
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Hasil dengan perintah standar Stata adalah:

. logistic highbp height weight age age2 female

```
Logistic regression                     Number of obs =  10,351
                                        LR chi2(5)    = 2422.62
                                        Prob > chi2   =  0.0000
Log likelihood = -5839.4541             Pseudo R2     =  0.1718

------------------------------------------------------------------------------
highbp | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------+----------------------------------------------------------------
height |   .9649972   .0035335    -9.73   0.000     .9580965    .9719476
weight |   1.051167   .0019388    27.06   0.000     1.047374    1.054974
   age |   1.053408   .0108589     5.05   0.000     1.032338    1.074907
  age2 |   .9999461   .0001076    -0.50   0.617     .9997352    1.000157
female |   .6858683   .0441003    -5.86   0.000     .6046579    .7779859
 _cons |   .8505688   .5504649    -0.25   0.803     .2392407    3.024014
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

BAB 7
REGRESI LOGISTIK POLITOMI

❖ Regresi Logistik Multinomial

Regresi logistik multinomial adalah model regresi dengan respons berskala nominal. Tiap kategori respons diperbandingkan dengan 1 kategori *basedline*, yang secara *default* adalah kategori pertama.

Tiap kategori respons diperbandingkan dengan kategori *basedline* dalam model regresi logistik:

$$\text{logit } y_i = \ln \frac{y_i}{1-y_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots$$

Perintah standar Stata untuk mengestimasi parameternya adalah:

mlogit *depvar indepvars* [*if*] [*in*] [, *options*]

Untuk data survei diperlukan penyesuaian sebagai berikut.

Sintaks:

svy: mlogit *depvar indepvars* [, *options*]

Dengan perintah **svy** ini akan diperoleh hasil yang berlainan. Untuk mengestimasi rasio odds, ditambahkan opsi "**rrr**".

Contoh 7.1:

```
. use "D:\Survey\Data\nhanes2f.dta", clear
. svyset psuid [pweight=finalwgt], strata(stratid)

      pweight: finalwgt
          VCE: linearized
  Single unit: missing
```

```
      Strata 1: stratid
          SU 1: psuid
         FPC 1: <zero>
```

. tab region

```
1=NE, 2=MW, |
   3=S, 4=W |      Freq.     Percent        Cum.
------------+-----------------------------------
         NE |      2,086       20.18       20.18
         MW |      2,773       26.83       47.01
          S |      2,853       27.60       74.61
          W |      2,625       25.39      100.00
------------+-----------------------------------
      Total |     10,337      100.00
```

Akan dilakukan regresi logistik multinomial dengan respons keempat kategori **region** dan yang dipilih sebagai kategori respons *basedline* adalah kategori ke-4.

. svy: mlogit region i.sex i.race c.age##c.age sizplace, baseoutcome(4)

```
(running mlogit on estimation sample)

Survey: Multinomial logistic regression

Number of strata = 31                  Number of obs   =      10,337
Number of PSUs   = 62                  Population size = 117,023,659
                                       Design df       =          31
                                       F(  18,     14) =        3.10
                                       Prob > F        =      0.0183
```

```
-------------------------------------------------------------------------------
             |             Linearized
      region |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+-----------------------------------------------------------------
NE           |
         sex |
      Female |   -.052379   .0588247    -0.89   0.380    -.1723529    .0675948
             |
        race |
       Black |   -.316601   .5906444    -0.54   0.596    -1.521228    .8880261
       Other |  -2.559074     .69193    -3.70   0.001    -3.970275   -1.147874
             |
         age |   .0165421   .0135663     1.22   0.232    -.0111266    .0442108
             |
   c.age#    |
      c.age |  -.0001134    .000151    -0.75   0.458    -.0004213    .0001945
             |
    sizplace |  -.0678906   .0616589    -1.10   0.279    -.1936448    .0578636
       _cons |  -.3541028   .3091841    -1.15   0.261    -.9846879    .2764823
-------------+-----------------------------------------------------------------
MW           |
         sex |
      Female |   .0017105   .0665411     0.03   0.980    -.1340011     .137422
             |
        race |
       Black |   .5412428   .5554036     0.97   0.337    -.5915103    1.673996
       Other |  -2.185329   .6539244    -3.34   0.002    -3.519016   -.8516411
             |
         age |   .0091949    .016302     0.56   0.577    -.0240533    .0424431
             |
```

```
 c.age#    |
    c.age | -.0000993   .0001928  -0.51  0.610  -.0004926    .000294
          |
 sizplace | -.0297954   .0479628  -0.62  0.539  -.1276162   .0680254
    _cons | -.1389536   .3225095  -0.43  0.670  -.7967161   .5188088
----------+------------------------------------------------------------
S         |
      sex |
  Female  |  .0461749   .0651536   0.71  0.484  -.0867068   .1790565
          |
     race |
   Black  |  1.682673   .6038539   2.79  0.009   .4511048   2.914241
   Other  | -2.052808   .7123293  -2.88  0.007  -3.505613  -.6000024
          |
      age |  .0289217   .0148612   1.95  0.061   -.001388   .0592314
          |
 c.age#   |
    c.age | -.0002509   .0001671  -1.50  0.143  -.0005918   .0000899
          |
 sizplace |  .1960466   .0604354   3.24  0.003   .0727878   .3193053
    _cons | -1.936347   .4913482  -3.94  0.000  -2.938458  -.9342361
----------+------------------------------------------------------------
W         |  (base outcome)
------------------------------------------------------------------------
```

Untuk mendapatkan estimasi rasio odds:

. svy: mlogit region i.sex i.race c.age##c.age sizplace, baseoutcome(4) rrr

```
(running mlogit on estimation sample)
```

Survey: Multinomial logistic regression

Number of strata = 31	Number of obs = 10,337
Number of PSUs = 62	Population size = 117,023,659
	Design df = 31
	F(18, 14) = 3.10
	Prob > F = 0.0183

region	RRR	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
NE						
sex						
Female	.9489691	.0558229	-0.89	0.380	.8416821	1.069932
race						
Black	.7286214	.4303561	-0.54	0.596	.2184435	2.430328
Other	.0773763	.053539	-3.70	0.001	.0188682	.3173107
age	1.01668	.0137926	1.22	0.232	.988935	1.045203
c.age#						
c.age	.9998866	.0001509	-0.75	0.458	.9995788	1.000195
sizplace	.9343627	.0576118	-1.10	0.279	.8239505	1.05957
_cons	.7018028	.2169863	-1.15	0.261	.3735558	1.318484
MW						
sex						

```
  Female  | 1.001712   .0666551     0.03   0.980    .8745891    1.147312
          |
     race |
    Black | 1.718141   .9542616     0.97   0.337    .5534907    5.333437
    Other | .1124408   .0735278    -3.34   0.002    .0296286    .4267141
          |
      age | 1.009237   .0164526     0.56   0.577    .9762337    1.043357
          |
 c.age#   |
    c.age | .9999007   .0001928    -0.51   0.610    .9995076    1.000294
          |
 sizplace | .9706441   .0465548    -0.62   0.539    .8801911    1.070393
    _cons | .8702684   .2806698    -0.43   0.670    .4508069    1.680025
----------+-----------------------------------------------------------
S         |
      sex |
  Female  | 1.047258   .0682326     0.71   0.484    .9169459    1.196088
          |
     race |
    Black | 5.379917   3.248684     2.79   0.009    1.570046    18.43482
    Other |  .128374   .0914445    -2.88   0.007    .0300284    .5488103
          |
      age | 1.029344   .0152973     1.95   0.061     .998613    1.061021
          |
 c.age#   |
    c.age | .9997491   .0001671    -1.50   0.143    .9994084     1.00009
          |
 sizplace | 1.216584   .0735247     3.24   0.003    1.075502    1.376171
    _cons | .1442298   .0708671    -3.94   0.000    .0529473    .3928859
----------+-----------------------------------------------------------
```

```
W          | (base outcome)
-------------------------------------------------------------------------
Note: _cons estimates baseline relative risk for each outcome.
```

Dengan menggunakan regio W (*West*) sebagai kategori *baseline*, dan distribusi karakteristiknya sebagai referensi, maka tampak:

- Untuk regio NE (*North-East*) dan MW (*Mid-West*), proporsi ras *Other* (*Non-White* dan *Non-Black*) berbeda bermakna dibandingkan dengan proporsinya pada regio W (*West*).
- Untuk regio S (*South*), proporsi ras *Black* dan *Other* berbeda bermakna dibandingkan dengan proporsinya pada regio W (*West*).

Dengan melihat tanda pada koefisien regresinya, disimpulkan bahwa proporsi ras *Other* pada regio NE, MW, dan S lebih sedikit daripada di regio W, sedangkan proporsi ras *Black* pada regio S lebih banyak daripada di regio W.

❖ Regresi Logistik Ordinal

Regresi logistik ordinal adalah model regresi dengan respons berskala ordinal. Untuk regresi logistik ordinal dengan k kategori respons, dilakukan $(k - 1)$ perbandingan model logistik, masing-masing dengan titik potong berbeda pada kategori responsnya. $(k - 1)$ perbandingan ini akan memiliki koefisien regresi yang sama, tetapi intersepnya berbeda. Model logistiknya yaitu:

$$\text{logit } y_i = \ln \frac{y_i}{1-y_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots$$

Perintah standar Stata untuk mengestimasi parameternya adalah:

ologit *depvar indepvars* [*if*] [*in*] [, *options*]

Untuk data survei dilakukan penyesuaian sebagai berikut.

Sintaks:

svy: ologit *depvar indepvars* [, *options*]

Jika diperlukan estimasi rasio odds, ditambahkan opsi "**or**". Dengan perintah **svy** ini akan diperoleh hasil yang berlainan.

Contoh 7.2:

. use "D:\Survey\Data\nhanes2f.dta", clear

. label list

```
highlead:
           0 lead<25
           1 lead>=25
agegrp:
           1 age20-29
           2 age30-39
           3 age40-49
           4 age50-59
           5 age60-69
           6 age 70+
race:
           1 White
           2 Black
           3 Other



sex:
```

```
           1 Male
           2 Female
region:
           1 NE
           2 MW
           3 S
           4 W
hlthgrp:
           1 poor
           2 fair
           3 average
           4 good
           5 excellent
```

. svyset psuid [pweight=finalwgt], strata(stratid)

```
      pweight: finalwgt
          VCE: linearized
  Single unit: missing
     Strata 1: stratid
         SU 1: psuid
        FPC 1: <zero>
```

. svy: ologit health female black age age2

```
(running ologit on estimation sample)
```

```
Survey: Ordered logistic regression

Number of strata = 31                Number of obs   =      10,335
Number of PSUs   = 62                Population size = 116,997,257
                                     Design df       =          31
                                     F(   4,     28) =      223.27
                                     Prob > F        =      0.0000

------------------------------------------------------------------------------
             |             Linearized
      health |      Coef.  Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.1615219   .0523678    -3.08   0.004    -.2683267    -.054717
       black |   -.986568   .0790277   -12.48   0.000    -1.147746   -.8253899
         age |  -.0119491   .0082974    -1.44   0.160    -.0288717    .0049736
        age2 |  -.0003234    .000091    -3.55   0.001     -.000509   -.0001377
-------------+----------------------------------------------------------------
       /cut1 |  -4.566229   .1632561                     -4.899192   -4.233266
       /cut2 |  -3.057415   .1699944                     -3.404121   -2.710709
       /cut3 |  -1.520596   .1714342                     -1.870239   -1.170954
       /cut4 |   -.242785   .1703965                      -.590311    .104741
------------------------------------------------------------------------------
```

Tampak bahwa kelompok wanita melaporkan memiliki status kesehatan lebih buruk daripada kelompok pria, warga kulit berwarna melaporkan status kesehatannya lebih buruk daripada warga kulit putih, dan kelompok usia lanjut melaporkan status kesehatannya lebih buruk daripada kelompok usia muda.

Untuk mendapatkan estimasi rasio odds:

. svy: ologit health female black age age2, or

```
(running ologit on estimation sample)

Survey: Ordered logistic regression

Number of strata = 31                Number of obs   =      10,335
Number of PSUs   = 62                Population size = 116,997,257
                                     Design df       =          31
                                     F(   4,     28) =      223.27
                                     Prob > F        =      0.0000

------------------------------------------------------------------------------
             |             Linearized
      health | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   .8508479   .044557    -3.08   0.004     .7646579    .946753
       black |   .3728541   .0294658   -12.48   0.000     .3173512    .4380642
         age |    .988122   .0081988    -1.44   0.160     .9715411    1.004986
        age2 |   .9996767    .000091    -3.55   0.001     .9994911    .9998623
-------------+----------------------------------------------------------------
       /cut1 |  -4.566229   .1632561                     -4.899192   -4.233266
       /cut2 |  -3.057415   .1699944                     -3.404121   -2.710709
       /cut3 |  -1.520596   .1714342                     -1.870239   -1.170954
       /cut4 |   -.242785   .1703965                      -.590311    .104741
------------------------------------------------------------------------------
Note: Estimates are transformed only in the first equation.
```

BAB 8
REGRESI DENGAN RESPONS DATA CACAH

❖ Regresi Poisson

Regresi Poisson adalah model regresi dengan respons berupa data cacah, yang berdistribusi Poisson. Model regresi Poisson adalah:

$$\ln Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots$$

dengan Y_i data cacah; $Y_i = 0, 1, 2, \ldots$

Asumsi terpenting untuk Y_i dengan distribusi Poisson adalah asumsi ekidispersi (*equidispersion*), yaitu variansinya sama dengan (atau dapat dianggap sama dengan) reratanya.

Perintah standar Stata untuk mengestimasi parameternya adalah:

.**poisson** *depvar indepvars* [*if*] [*in*] [, *options*]

Untuk data survei, perintah **svy**-nya menjadi:

Sintaks:

svy: poisson *depvar indepvars* [, *options*]

Untuk memperoleh estimasi rasio *rate*, ditambahkan opsi "**irr**".

Contoh 8.1:

. use "D:\Survey\Data\poisson_3.dta", clear

. sum num_awards

```
    Variable |   Obs   Mean   Std. Dev.    Min     Max
-------------+--------------------------------------------
  num_awards |   200    .63   1.052921       0       6
```

. svyset psuid [pw=sampwgt], strata(stratid)

```
      pweight: sampwgt
          VCE: linearized
  Single unit: missing
     Strata 1: stratid
         SU 1: psuid
        FPC 1: <zero>
```

. svy: poisson num_awards math i.prog

```
(running poisson on estimation sample)

Survey: Poisson regression

Number of strata = 4                  Number of obs   =    200
Number of PSUs   = 8                  Population size = 67,761
                                      Design df       =      4
                                      F(   3,      2) =  10.97
                                      Prob > F        = 0.0847

------------------------------------------------------------------------------
             |             Linearized
num_awards |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        math |   .0646063   .0120638     5.36   0.006     .0311118    .0981008
             |
        prog |
    academic |   .7075928   .2304017     3.07   0.037     .0678952     1.34729
    vocation |  -.3200741   .3690818    -0.87   0.435    -1.344809    .7046613
             |
       _cons |  -4.550341   .8846441    -5.14   0.007    -7.006507   -2.094175
------------------------------------------------------------------------------
```

```
. svy: poisson num_awards math i.prog, irr
(running poisson on estimation sample)

Survey: Poisson regression

Number of strata = 4                  Number of obs   =    200
Number of PSUs   = 8                  Population size = 67,761
                                      Design df       =      4
                                      F(   3,      2) =  10.97
                                      Prob > F        = 0.0847

------------------------------------------------------------------------------
             |             Linearized
num_awards |       IRR   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        math |   1.066739   .0128689     5.36   0.006     1.031601    1.103074
             |
        prog |
   academic  |   2.029101   .4675083     3.07   0.037     1.070253    3.846988
   vocation  |   .7260952   .2679886    -0.87   0.435     .2605893    2.023161
             |
       _cons |   .0105636    .009345    -5.14   0.007     .000906    .1231718
------------------------------------------------------------------------------
Note: _cons estimates baseline incidence rate.
```

Tampak bahwa siswa dengan nilai math yang lebih tinggi memiliki *rate* yang lebih besar untuk mendapatkan *awards*, juga yang berasal dari program academic agak lebih besar *rate* perolehan *awards*-nya dibandingkan dengan siswa program general.

❖ Regresi Binomial Negatif

Seperti halnya regresi Poisson, regresi binomial negatif adalah model regresi dengan respons juga berupa data cacah. Model regresi binomial negatif juga adalah:

$$\ln Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots$$

dengan Y_i data cacah; $Y_i = 0, 1, 2, \ldots$

Berbeda dengan distribusi Poisson, asumsi untuk Y_i pada distribusi binomial negatif adalah asumsi overdispersi (*overdispersion*), yaitu variansinya lebih besar daripada reratanya.

Perintah standar Stata untuk mengestimasi parameternya adalah:

.nbreg *depvar indepvars* [*if*] [*in*] [, *options*]

Untuk data survei, perintah **svy**-nya menjadi:

Sintaks:

svy: nbreg *depvar indepvars* [, *options*]

Untuk memperoleh estimasi rasio *rate*, ditambahkan opsi "**irr**".

Contoh 8.2:

```
. use "D:\Survey\Data\lahigh_2.dta", clear
. svyset psuid [pw=sampwgt], strata(strata)

      pweight: sampwgt
          VCE: linearized
  Single unit: missing
     Strata 1: strata
         SU 1: psuid
        FPC 1: <zero>
```

Estimasi parameter untuk regresi dengan respons data cacah dianjurkan selalu dimulai dengan regresi Poisson.

. svy: poisson daysabs mathnce langnce gender

```
(running poisson on estimation sample)

Survey: Poisson regression
Number of strata =  5                     Number of obs   =     316
Number of PSUs   = 10                     Population size = 499,904
                                          Design df       =       5
                                          F(   3,      3) =   16.68
                                          Prob > F        =  0.0224

------------------------------------------------------------------------------
             |             Linearized
     daysabs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    mathnce |   .0005623   .0063969     0.09   0.933    -.0158815    .0170061
    langnce |  -.0146343   .0052641    -2.78   0.039    -.0281661   -.0011025
     gender |   .4752252   .1548074     3.07   0.028     .0772802    .8731702
      _cons |   2.156165    .290054     7.43   0.001     1.410557    2.901772
------------------------------------------------------------------------------
```

Tampak bahwa model cukup baik (`Prob > F = 0.0224`), tetapi nilai *p* tidak terlalu kecil sebaiknya dilakukan pengujian asumsi lebih lanjut.

. sum daysabs

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     daysabs |       316    5.810127    7.449003         0         45
```

Estimasi variansinya adalah 55.488 (= 7.449^2), hampir sepuluh kali lebih besar daripada estimasi reratanya 5.810, sehingga asumsi distribusi Poisson tidak terpenuhi.

Perintah untuk melihat efek desainnya adalah:

. svy: mean daysabs

```
(running mean on estimation sample)

Survey: Mean estimation
Number of strata =       5          Number of obs   =     316
Number of PSUs   =      10          Population size = 499,905
                                    Design df       =       5

--------------------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
daysabs |   5.599699   .9364053      3.192593    8.006806
--------------------------------------------------------------
```

. estat effects

```
-------------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.       DEFF      DEFT
-------------+-----------------------------------------
daysabs |   5.599699   .9364053     4.98923   2.23366
-------------------------------------------------------
```

Selanjutnya **daysabs** diregresikan terhadap **mathnce**, **langnce**, dan **gender** dalam model regresi binomial negatif:

. svy: nbreg daysabs mathnce langnce gender

```
(running nbreg on estimation sample)

Survey: Negative binomial regression

Number of strata =  5                    Number of obs    =     316
Number of PSUs   = 10                    Population size = 499,904
                                         Design df        =       5
                                         F(   3,      3) =   13.84
Dispersion       = mean                  Prob > F         =  0.0291

------------------------------------------------------------------------------
             |             Linearized
     daysabs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     mathnce |   .0003722   .0047593     0.08   0.941    -.011862    .0126064
     langnce |  -.0163684   .0045831    -3.57   0.016    -.0281495   -.0045873
      gender |  -.5001316   .1673956    -2.99   0.031     .0698275    .9304358
       _cons |   2.236401   .2761484     8.10   0.000     1.526539    2.946263
-------------+----------------------------------------------------------------
    /lnalpha |   .2822735   .1060883                      .0095648    .5549822
-------------+----------------------------------------------------------------
       alpha |   1.326141   .1406881                      1.009611     1.74191
------------------------------------------------------------------------------
```

Tampak bahwa siswa dengan nilai kemampuan bahasa (**langnce**) lebih tinggi cenderung memiliki *rate* absensi (**dayabs**) yang lebih rendah, juga gender pria cenderung memiliki *rate* absensi (**dayabs**) yang lebih rendah daripada gender wanita.

. svy: nbreg daysabs mathnce langnce gender, irr

```
(running nbreg on estimation sample)

Survey: Negative binomial regression

Number of strata =  5                  Number of obs   =     316
Number of PSUs   = 10                  Population size = 499,904
                                       Design df       =       5
                                       F(   3,      3) =   13.84
Dispersion       = mean                Prob > F        =  0.0291

------------------------------------------------------------------------------
             |             Linearized
     daysabs |       IRR   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     mathnce |   1.000372   .0047611     0.08   0.941     .9882081    1.012686
     langnce |   .9837648   .0045086    -3.57   0.016      .972243    .9954232
      gender |   1.648938   .2760251     2.99   0.031     1.072323    2.535614
       _cons |   9.359588   2.584635     8.10   0.000     4.602222     19.0347
-------------+----------------------------------------------------------------
    /lnalpha |   .2822735   .1060883                      .0095648    .5549822
-------------+----------------------------------------------------------------
       alpha |   1.326141   .1406881                      1.009611     1.74191
------------------------------------------------------------------------------
Note: Estimates are transformed only in the first equation.
Note: _cons estimates baseline incidence rate.
```

BAB 9
REGRESI DATA SURVIVAL

❖ Model Hazard Proporsional Cox

Model hazard proporsional Cox adalah model regresi yang mempergunakan waktu sampai dengan terjadinya kegagalan sebagai variabel respons. Perintah standar Stata untuk mengestimasi parameter model hazard proporsional Cox adalah:

stcox *indepvars* [*if*] [*in*] [, *options*]

yang dikerjakan setelah melakukan deklarasi data survival dengan "**stset**".

Untuk data survei perintahnya disesuaikan sebagai berikut.

Sintaks:

svy: stcox *indepvars* [, *options*]

Perintah di atas akan menghasilkan estimasi rasio hazard. Untuk memperoleh estimasi koefisien regresi, ditambahkan opsi "**nohr**".

Contoh 9.1:

```
. use "D:\Survey\Data\nhefs.dta", clear
. svyset psu2 [pw=swgt2], strata(strata2)

      pweight: swgt2
          VCE: linearized
  Single unit: missing
     Strata 1: strata2
         SU 1: psu2
        FPC 1: <zero>
```

. stset age_lung_cancer [pw=swgt2], fail(lung_cancer)

```
     failure event:  lung_cancer != 0 & lung_cancer < .
obs. time interval:  (0, age_lung_cancer]
 exit on or before:  failure
            weight:  [pweight=swgt2]

--------------------------------------------------------------------------
 14,407  total observations
  5,126  event time missing (age_lung_cancer>=.)   PROBABLE ERROR
--------------------------------------------------------------------------
  9,281  observations remaining, representing
     83  failures in single-record/single-failure data
599,691  total analysis time at risk and under observation
                                        at risk from t = 0
                              earliest observed entry t = 0
                                   last observed exit t = 97
```

. svy: stcox former_smoker smoker male urban1 rural

```
(running stcox on estimation sample)

Survey: Cox regression

Number of strata =  35                Number of obs    =       9,149
Number of PSUs   = 105                Population size = 151,327,827
                                      Design df       =          70
                                      F(   5,     66) =       14.07
                                      Prob > F        =      0.0000
```

```
------------------------------------------------------------------------------
             |               Linearized
          _t | Haz. Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
former_smoker |   2.788113   .6205102     4.61   0.000     1.788705    4.345923
       smoker |   7.849483   2.593249     6.24   0.000     4.061457    15.17051
         male |   1.187611   .3445315     0.59   0.555     .6658757    2.118142
       urban1 |   .8035074   .3285144    -0.54   0.594     .3555123    1.816039
        rural |   1.581674   .5281859     1.37   0.174     .8125799    3.078702
------------------------------------------------------------------------------
```

Tampak bahwa status **former_smoker** dan terutama **smoker** meningkatkan *hazard rate* kejadian kanker paru dibandingkan dengan status **non-smoker**.

. stcox former_smoker smoker male urban1 rural, nohr

```
         failure _d:  lung_cancer
   analysis time _t:  age_lung_cancer
             weight:  [pweight=swgt2]

(sum of wgt is 151,327,827)
Iteration 0: log pseudolikelihood = -607.69761
Iteration 1: log pseudolikelihood =  -563.0561
Iteration 2: log pseudolikelihood = -559.97974
Iteration 3: log pseudolikelihood = -559.91054
Iteration 4: log pseudolikelihood = -559.91053
Refining estimates:
Iteration 0: log pseudolikelihood = -559.91053
```

```
Cox regression -- Breslow method for ties

No. of subjects      = 151,327,827        Number of obs =  9,149
No. of failures      =   1,261,494
Time at risk         =  9573934271
                                          Wald chi2(5)  =  71.17
Log pseudolikelihood =  -559.91053        Prob > chi2   = 0.0000

------------------------------------------------------------------------------
             |               Robust
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
former_smoker |   1.025365   .3118372     3.29   0.001     .4141753    1.636555
       smoker |   2.060448    .322937     6.38   0.000     1.427503    2.693393
         male |   .1719435   .3090661     0.56   0.578    -.433815     .7777019
       urban1 |  -.2187688   .3840601    -0.57   0.569    -.9715128    .5339751
        rural |   .4584836   .3529357     1.30   0.194    -.2332576    1.150225
------------------------------------------------------------------------------
```

❖ Model Survival Parametrik

Model survival parametrik adalah model survival dengan asumsi fungsi survivalnya memiliki distribusi parametrik tertentu. Perintah standar Stata untuk mengestmasi parameternya adalah:

streg *indepvars* [*if*] [*in*], **distribution**(*dist*) [*options*]

yang dilakukan setelah mendeklarasikan dataset sebagai data survival dengan perintah "**stset**".

Untuk data survei perintahnya disesuaikan menjadi:

Sintaks:

svy: streg *indepvars*, **distribution**(*dist*) [*options*]

Opsi yang tersedia untuk **distribution** antara lain adalah:

exponential : Distribusi eksponensial

weibull : Distribusi Weibull

loglogistic : Distribusi loglogistik

Dengan distribusi eksponensial dan Weibull, perintah di atas akan menghasilkan rasio hazard. Untuk memperoleh estimasi koefisien regresi, ditambahkan opsi "**nohr**".

Untuk opsi distribusi loglogistik, estimasi yang diperoleh selalu berupa koefisien regresi. Pada model loglogistik, tidak ada rasio hazard, karena model loglogistik termasuk model odds proporsional (model PO), bukan model hazard proporsional (model PH) seperti model PH Cox.

Contoh 9.2:

```
. use "D:\Survey\Data\nhefs.dta", clear
. svyset psu2 [pw=swgt2], strata(strata2)

      pweight: swgt2
          VCE: linearized
  Single unit: missing
     Strata 1: strata2
         SU 1: psu2
        FPC 1: <zero>
. stset age_lung_cancer [pw=swgt2], fail(lung_cancer)

     failure event:  lung_cancer != 0 & lung_cancer < .
obs. time interval:  (0, age_lung_cancer]
 exit on or before:  failure
           weight:  [pweight=swgt2]
```

```
------------------------------------------------------------------------------
 14,407  total observations
  5,126  event time missing (age_lung_cancer>=.)   PROBABLE ERROR
------------------------------------------------------------------------------
  9,281  observations remaining, representing
     83  failures in single-record/single-failure data
599,691  total analysis time at risk and under observation
                                              at risk from t =  0
                                   earliest observed entry t =  0
                                        last observed exit t = 97
```

. svy: streg former_smoker smoker, dist(exp)

```
(running streg on estimation sample)

Survey: Exponential PH regression

Number of strata =  35                    Number of obs    =       9,149
Number of PSUs   = 105                    Population size = 151,327,827
                                          Design df       =          70
                                          F(   2,     69) =       34.57
                                          Prob > F        =      0.0000

------------------------------------------------------------------------------
             |             Linearized
          _t | Haz. Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
former_smoker |   5.472978   1.280159     7.27   0.000     3.432584    8.726221
       smoker |   7.410505   2.545939     5.83   0.000     3.734798    14.70376
        _cons |   .0000286   7.90e-06   -37.83   0.000     .0000165    .0000496
```

```
------------------------------------------------------------------------------
Note: _cons estimates baseline hazard.
```

Dengan data yang sama seperti pada Contoh 9.1, secara kualitatif hasil di sini sama dengan model hazard proporsional Cox, tetapi secara kuantitatif peningkatan *hazard rate* **former smoker** dibandingkan dengan **non-smoker** pada model parametrik eksponensial ini (HR = 5.473) hampir dua kali lebih besar dibandingkan dengan pada model hazard proporsional Cox (HR = 2.788).

. svy: streg former_smoker smoker, dist(exp) nohr

```
(running streg on estimation sample)

Survey: Exponential PH regression

Number of strata =  35                Number of obs    =       9,149
Number of PSUs   = 105                Population size = 151,327,827
                                      Design df       =          70
                                      F(   2,     69) =       34.57
                                      Prob > F        =      0.0000

------------------------------------------------------------------------------
             |             Linearized
          _t |     Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
former_smoker |   1.699823   .2339053     7.27   0.000     1.233313    2.166332
       smoker |   2.002899    .343558     5.83   0.000     1.317694    2.688103
        _cons | -10.46334   .2765963   -37.83   0.000    -11.01499   -9.911684
------------------------------------------------------------------------------
```

. svy: streg former_smoker smoker, dist(wei)

```
(running streg on estimation sample)

Survey: Weibull PH regression

Number of strata =  35                    Number of obs    =       9,149
Number of PSUs   = 105                    Population size  = 151,327,827
                                          Design df        =          70
                                          F(   2,     69)  =       30.12
                                          Prob > F         =      0.0000

------------------------------------------------------------------------------
             |             Linearized
          _t | Haz. Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
former_smoker |   2.792858   .6644426     4.32   0.000     1.737719    4.488674
       smoker |   8.445133   2.823761     6.38   0.000      4.33499    16.45223
        _cons |   5.76e-18   1.32e-17   -17.30   0.000     5.93e-20    5.59e-16
-------------+----------------------------------------------------------------
        /ln_p |   2.065687   .0663027    31.16   0.000      1.93345    2.197923
-------------+----------------------------------------------------------------
            p |   7.890714   .5231759                      6.91332     9.00629
          1/p |   .1267312   .0084026                     .1110335    .1446483
------------------------------------------------------------------------------
Note: Estimates are transformed only in the first equation.
Note: _cons estimates baseline hazard.
```

. svy: streg former_smoker smoker, dist(wei) nohr

```
(running streg on estimation sample)
```

```
Survey: Weibull PH regression

Number of strata =  35                    Number of obs    =       9,149
Number of PSUs   = 105                    Population size = 151,327,827
                                          Design df       =          70
                                          F(   2,     69) =       30.12
                                          Prob > F        =      0.0000

------------------------------------------------------------------------------
             |             Linearized
          _t |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
former_smoker |   1.027065   .2379078     4.32   0.000     .5525732    1.501557
       smoker |    2.13359   .3343655     6.38   0.000     1.466719    2.800461
        _cons | -39.69603    2.29395   -17.30   0.000    -44.27117   -35.12089
-------------+----------------------------------------------------------------
       /ln_p |   2.065687   .0663027    31.16   0.000      1.93345    2.197923
-------------+----------------------------------------------------------------
           p |   7.890714   .5231759                      6.91332     9.00629
         1/p |   .1267312   .0084026                     .1110335    .1446483
------------------------------------------------------------------------------
```

Perbandingan estimasi koefisien regresi dan *hazard ratio* untuk *smoker* vs. *non-smoker* pada ketiga model survival diperlihatkan sebagai berikut:

Model	Estimasi	
	Koefisien regresi	**Hazard Ratio**
Model PH Cox	2.060	7.849
Parametrik eksponensial	2.003	7.411
Parametrik Weibull	2.134	8.445

.

BAB 10
MODEL SEM DAN GSEM UNTUK DATA SURVEI

❖ Model SEM untuk Data Survei

Model SEM dan GSEM Stata untuk data survei praktis masih dalam tahap pengembangan. Belum semua perintah standar Stata dapat dikerjakan modifikasinya dalam perintah **svy**. Dalam garis besarnya, model SEM dan GSEM dibedakan menjadi model struktural (Analisis Jalur) dan model pengukuran (Analisis Faktor Konfirmatorik), serta kombinasinya yaitu model regresi struktural. Di sini hanya akan dibahas contoh untuk model struktural.

Perintah standar Stata untuk model SEM adalah:

sem [*path 1*] [*path 2*] . . . [, *options*]

Modifikasi penyesuaiannya untuk perintah **svy** adalah:

svy: sem [*path 1*] [*path 2*] . . . [, *options*]

Contoh 10.1:

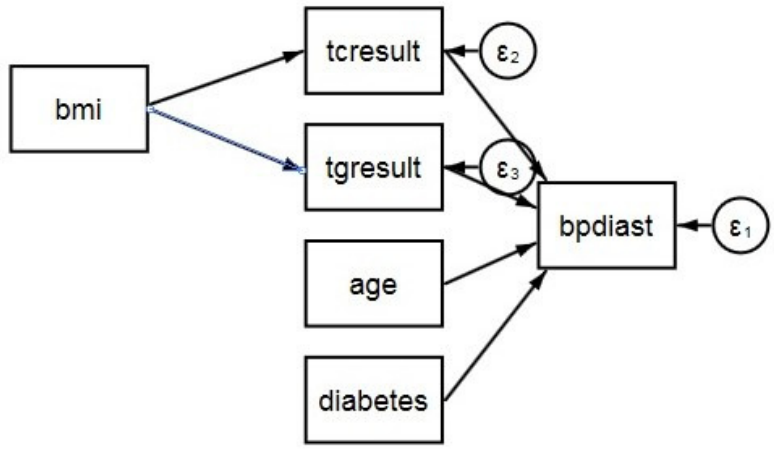
```
. use "D:\Survey\Data\nhanes2d.dta"
. svyset

      pweight: finalwgt
          VCE: linearized
  Single unit: missing
     Strata 1: strata
         SU 1: psu
        FPC 1: <zero>
```

. gen bmi = weight/((height/100)^2)

. save "D:\Survey\Data\nhanes2d1.dta"

```
file D:\Survey\Data\nhanes2d1.dta saved
```

Akan dilakukan estimasi terhadap model berikut:



. svy: sem (tcresult <– bmi) (tgresult <– bmi) (bpdiast <– tcresult tgresult age diabetes)

```
(running sem on estimation sample)

Survey: Structural equation model          Number of obs    =       5,049
Number of strata  =    31                  Population size  =  56,806,965
Number of PSUs    =    62                  Design df        =          31
```

	Linearized					
	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Structural						
tcresult						
bmi	1.946319	.1306044	14.90	0.000	1.679949	2.212688
_cons	162.2527	3.624923	44.76	0.000	154.8596	169.6458

```
-------------+----------------------------------------------------------------
tgresult     |
         bmi |  5.523548   .4526919    12.20   0.000     4.600277    6.44682
       _cons | -.9206394   10.65858    -0.09   0.932    -22.65895   20.81767
-------------+----------------------------------------------------------------
bpdiast      |
    tcresult |  .0252644   .0062454     4.05   0.000     .0125267    .038002
    tgresult |  .0205274   .0030245     6.79   0.000     .0143588   .0266959
         age |  .1751973   .0094619    18.52   0.000     .1558997    .194495
    diabetes |  1.305233    1.12309     1.16   0.254    -.985325    3.595791
       _cons |  64.69383   1.085145    59.62   0.000     62.48066     66.907
-------------+----------------------------------------------------------------
var(e.tcresult)|  2127.779   57.37395                      2013.924   2248.072
var(e.tgresult)|  8484.064    945.706                      6758.818   10649.69
 var(e.bpdiast)|  147.6975   5.763342                      136.3987   159.9323
------------------------------------------------------------------------------
```

Untuk perintah **svy postestimation estat eqtest** tidak diperlukan prefix **svy**:

. estat eqtest

```
Adjusted Wald tests for equations
--------------------------------------
             |         F   df         p
-------------+------------------------
observed     |
    tcresult |   222.08    1    0.0000
    tgresult |   148.88    1    0.0000
     bpdiast |   185.97    4    0.0000
--------------------------------------
      Design                 31
```

Tanpa menggunakan bobot sampling, hasil yang diperoleh yaitu:

. sem (tcresult <- bmi) (tgresult <- bmi) (bpdiast <- tcresult tgresult age diabetes)

```
(5302 observations with missing values excluded)

Endogenous variables

Observed:  tcresult tgresult bpdiast

Exogenous variables

Observed:  bmi age diabetes

Fitting target model:

Iteration 0: log likelihood = -112491.3
Iteration 1: log likelihood = -112491.3

Structural equation model            Number of obs = 5,049
Estimation method  = ml
Log likelihood     =  -112491.3

------------------------------------------------------------------------------
             |                 OIM
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Structural   |
  tcresult   |
         bmi |   1.565127   .1341434    11.67   0.000     1.302211    1.828043
       _cons |   176.0087   3.487575    50.47   0.000     169.1731    182.8442
  ---------- +----------------------------------------------------------------
  tgresult   |
         bmi |   4.958535   .2652444    18.69   0.000     4.438666    5.478404
       _cons |   17.36096   6.896053     2.52   0.012     3.844946    30.87698
  ---------- +----------------------------------------------------------------
```

```
  bpdiast      |
      tcresult |  .0189016  .0042523   4.45  0.000   .0105672   .027236
      tgresult |  .0189612  .0019704   9.62  0.000   .0150993  .0228232
           age |  .1580245  .0112583  14.04  0.000   .1359586  .1800903
      diabetes | -.2373632  .8424418  -0.28  0.778  -1.888519  1.413792
         _cons |  66.69479  .8319971  80.16  0.000   65.06411  68.32548
---------------+------------------------------------------------------------
var(e.tcresult)|  2226.803  44.31942                2141.611  2315.384
var(e.tgresult)|  8706.343  173.2799                8373.259  9052.676
 var(e.bpdiast)|  154.0241  3.065498                148.1315  160.1511
------------------------------------------------------------------------------
LR test of model vs. saturated: chi2(6) = 2159.24,
Prob > chi2 = 0.0000
```

. Tampak bahwa tanpa pembobotan diperoleh hasil estimasi koefisien regresi dan *standard error* yang agak berbeda.

❖ Model GSEM untuk Data Survei

Perintah standar Stata untuk model GSEM adalah:

gsem [*path 1*] [*path 2*] . . . , [*link*] [*options*]

Modifikasinya untuk perintah svy adalah:

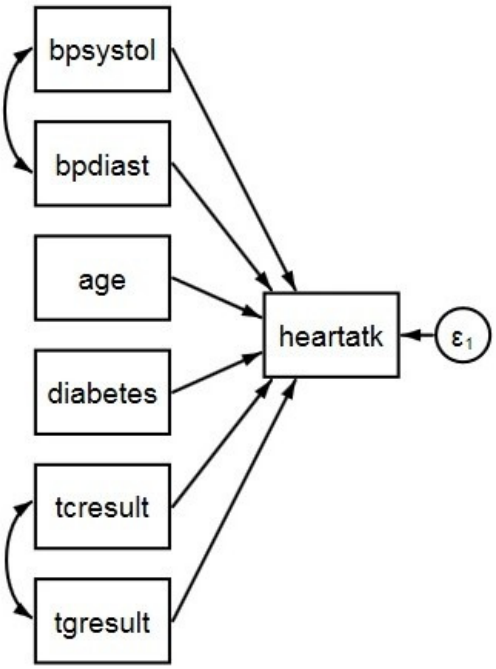
svy: gsem [*path 1*] [*path 2*] . . . , [*link*] [*options*]

Contoh 10.2:

```
. use "D:\Survey\Data\nhanes2d1.dta"
. svyset
```

```
      pweight: finalwgt
          VCE: linearized
  Single unit: missing
     Strata 1: strata
         SU 1: psu
        FPC 1: <zero>
```

Model yang akan diuji adalah:



Dengan perintah gsem, variansi dan kovariansi variabel eksogen teramati tak diestimasi, sehingga perintah **svy gsem** adalah sebagai berikut

. svy: gsem (heartatk <- bpsystol bpdiast age diabetes tcresult tgresult), logit

(running gsem on estimation sample)

```
Survey: Generalized structural equation model

Number of strata = 31                  Number of obs   =      5,049
Number of PSUs   = 62                  Population size = 56,806,965

Design df        = 31
Response     : heartatk
Family       : Bernoulli
Link         : logit

------------------------------------------------------------------------------
             |             Linearized
             |      Coef.  Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
heartatk     |
    bpsystol |   -.00406   .0037884    -1.07   0.292    -.0117865    .0036664
     bpdiast |  -.0042786   .0064826    -0.66   0.514    -.0174999    .0089427
         age |    .089641   .0058926    15.21   0.000     .0776229    .1016591
    diabetes |    .496719   .2639028     1.88   0.069    -.0415144    1.034952
    tcresult |  -.0014159    .002146    -0.66   0.514    -.0057926    .0029609
    tgresult |   .0028082   .0007379     3.81   0.001     .0013032    .0043133
       _cons |  -7.392649   .4949259   -14.94   0.000    -8.402058   -6.383241
------------------------------------------------------------------------------
```

Dengan perintah standar Stata **gsem** diperoleh hasil:

. gsem (heartatk <- bpsystol bpdiast age diabetes tcresult tgresult), logit

```
Iteration 0: log likelihood = -976.30422
Iteration 1: log likelihood = -792.64941
Iteration 2: log likelihood = -780.62976
Iteration 3: log likelihood = -780.09596
Iteration 4: log likelihood = -780.09474
Iteration 5: log likelihood = -780.09474

Generalized structural equation model      Number of obs = 5,049
Response       : heartatk
Family         : Bernoulli
Link           : logit
Log likelihood = -780.09474

------------------------------------------------------------------------------
             |      Coef.  Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
heartatk     |
    bpsystol |  -.0064624   .0041118    -1.57   0.116    -.0145215    .0015966
     bpdiast |  -.0008168   .0072727    -0.11   0.911     -.015071    .0134374
         age |   .0865489   .0075776    11.42   0.000      .071697    .1014008
    diabetes |    .521423   .2227335     2.34   0.019     .0848734    .9579726
    tcresult |  -.0026478   .0016194    -1.64   0.102    -.0058217    .0005262
    tgresult |   .0024396   .0005653     4.32   0.000     .0013316    .0035475
       _cons |  -6.913039   .6788565   -10.18   0.000    -8.243574   -5.582505
------------------------------------------------------------------------------
```

Tampak hasil yang diperoleh agak berbeda, selain itu uji Wald pada perintah standar Stata **gsem** menggunakan statistik *Z*, bukan statistik *t*.

KEPUSTAKAAN

Acock AC. **A Gentle Introduction to Stata**, 4th Ed. College Station, Texas: Stata Press, 2014.

de Leeuw ED, Hox JJ, Dillman DA. **International Handbook of Survey Methodology**. New York, NY: Lawrence Erlbaum Associates, 2008.

Hamilton LC. **Statistics with Stata: Version 12**. Boston: Brooks/Cole, Cenage Learning, 2013.

Heeringa SG, West BT, Berglund PA. **Applied Survey Data Analysis**. Boca Raton, FL: Chapman & Hall/CRC, 2010.

Johnson DR. **Using Weights in the Analysis of Survey Data**. Population Research Institute, The Pennsylvania State University, November 2008. Available from http://www.nyu.edu/classes/jackson/design.of. social.research, viewed on December 20, 2017.

Kohler U, Kreuter F. **Data Analysis Using Stata**, 3rd Ed. College Station, Texas: Stata Press, 2012.

Levy PS, Lemeshow S. **Sampling of Populations: Methods and Applications**, 4th Ed. Hoboken, NJ: John Wiley & Sons, Inc, 2008.

Scheaffer RL, Mendelhall III W, Ott RL, Gerow K. **Elementary Survey Sampling**, 7th Ed. Boston: Brooks/Cole, Cenage Learning, 2012.

StataCorp LLC. **Stata Survey Data Reference Manual: Release 15.** College Station, Texas: Stata Press, 2017.

Thompson SK. **Sampling**, 3rd Ed. Hoboken, New Jersey: John Wiley & Sons, Inc, 2012.

Lampiran 1

BOBOT SAMPLING

Dalam contoh-contoh analisis data survei terdahulu, dataset yang digunakan telah memuat nilai-nilai pembobotan tanpa pembahasan rinci mengenai cara perolehannya. Dalam penelitian survei sesungguhnya, peneliti harus menghitung sendiri bobot sampling untuk tiap anggota sampel dalam dataset-nya.

Pembobotan dapat dilakukan dengan 2 metode, yaitu *self-weighting* (pembobotan-sendiri) dan *probability weighting* (pembobotan probabilitas). *Self-weighting* dikenal juga sebagai sampling acak stratifikasi dengan alokasi proporsional, yaitu jika sampel berukuran n akan dipilih dari populasi berukuran N, maka ukuran tiap subsampel n_i yang akan dipilih dari subpopulasi berukuran N_i adalah sedemikian hingga $n_i/N_i = n/N$. Diperoleh nilai bobot yang sama untuk tiap anggota sampel. Walaupun relatif mudah, kelemahan metode ini yaitu umumnya hanya dapat digunakan pada sampling 1-tahap dan menjadi tidak valid jika ada *non-response rate* yang tidak homogen antar-stratum. Metode ini tidak akan dibahas lebih lanjut di sini.

Metode kedua yang lazim digunakan dalam analisis data survei adalah pembobotan probabilitas. Pembobotan probabilitas terdiri atas 2 bagian, yaitu pembobotan desain (*design weighting*) yang selalu diperlukan dan pembobotan post-stratifikasi (*post-stratification*), yaitu penyesuaian bobot yang dilakukan hanya jika dibutuhkan.

▪ Bobot desain

Bobot desain ditujukan untuk mengatasi bias yang timbul akibat fitur tertentu pada desain sampling ataupun hal-hal terjadi secara aksidental dalam proses pengumpulan data. Kedua faktor ini menyebabkan data sampling tidak merepresentasikan keadaan sebenarnya untuk populasi yang diteliti.

Untuk mengkompensasikan bias tersebut, digunakan bobot probabilitas yaitu inversi fraksi sampling.

Untuk rancangan 1-tahap, jika subsampel berukuran n_i akan dipilih dari subpopulasi berukuran N_i, maka probabilitas tiap unit sampling dalam subpopulasi tersebut untuk terpilih sama dengan **fraksi sampling**-nya yairu:

$$\lambda_i = \frac{n_i}{N_i}$$

Bobot untuk tiap unit sampling yang terpilih disebut **bobot desain** yang besarnya sama dengan inversi fraksi sampling:

$$w_i = \frac{1}{\lambda_i} = \frac{N_i}{n_i}$$

Untuk rancangan multi-tahap, fraksi sampling dan probabilitas sampling harus dihitung sendiri-sendiri untuk tiap tahap sampling. Bobot desain multi-tahap adalah:

$$w_D = w_{i1} \times w_{i2} \times w_{i3} \times \ldots$$

$$= \frac{1}{\lambda_{i1}} \times \frac{1}{\lambda_{i2}} \times \frac{1}{\lambda_{i3}} \times \ldots$$

$$= \frac{N_{i1}}{n_{i1}} \times \frac{N_{i2}}{n_{i2}} \times \frac{N_{i3}}{n_{i3}} \times \ldots$$

▪ Bobot post-stratifikasi

Bobot post-stratifikasi perlu diperhitungkan jika setelah data sampel terkumpul didapatkan disparitas antara distribusi karakteristik tertentu dalam data sampel dengan distribusi karakteristik tersebut dalam populasi.

Misalkan dimiliki data distribusi gender dalam suatu penelitian sebagai berikut (Johnson, 2008):

Gender	Proporsi populasi	Proporsi sampel	Bobot post-stratifikasi
Pria	0.5	0.4	$0.5/0.4 = 1.2500$
Wanita	0.5	0.6	$0.5/0.6 = 0.8333$
Total	1	1	

Jika dalam suatu penelitian diperlukan pembobotan post-stratifikasi dan w_{Ps} menyatakan bobot post-stratifikasi, maka bobot probabilitas total adalah:

$$w_T = w_D \times w_{Ps}$$

Selain bobot post-stratifikasi, setelah data sampel terkumpul ada kalanya dibutuhkan pula bobot non-respons untuk mengkompensasikan bias akibat *non-response rate* yang tidak sama antar-stratum. Walaupun demikian, karena dalam setiap penelitian dibutuhkan *non-response rate* yang rendah agar data penelitian valid, bobot non-respons umumnya dapat diabaikan dan tidak akan dibahas lebih lanjut di sini.

Lampiran 2

Uji *t* untuk Data Survei

Walaupun uji *t* tidak dapat dikerjakan secara langsung dengan perintah **svy**, jika diperlukan uji *t* tersebut dapat juga dikerjakan seperti pada contoh berikut ini.

```
. use "D:\Survey\Data\hsb2.dta", clear
(highschool and beyond (200 cases))

. svyset [pw=socst], strata(ses)

      pweight: socst
          VCE: linearized
  Single unit: missing
     Strata 1: ses
         SU 1: <observations>
        FPC 1: <zero>

. svy: mean write, over(female)
(running mean on estimation sample)

Survey: Mean estimation

Number of strata =   3        Number of obs   =    200
Number of PSUs   = 200        Population size = 10,481
                              Design df       =    197

          male: female = male
        female: female = female
```

```
------------------------------------------------------------
             |             Linearized
        Over |       Mean   Std. Err.  [95% Conf. Interval]
-------------+----------------------------------------------
write        |
        male |   51.65351   1.041066   49.60045    53.70658
      female |   55.81467    .721354    54.3921    57.23723
------------------------------------------------------------
```

Uji *t* berikut memperbandingkan kemampuan menulis siswa pria dan siswa wanita:

. test [write]male = [write]female

```
Adjusted Wald test

 ( 1)  [write]male - [write]female = 0

       F(  1,   197) =   10.45
            Prob > F =    0.0014
```

Tampak adanya perbedaan bermakna antara kemampuan menulis kedua kelompok siswa.