

Working with epidemiologic data in Stata

Gabriela Ortiz

StataCorp LLC

April 12, 2022

Goals

- Learn how to
 - Work with cross-sectional, incidence-rate, cumulative incidence, and survival-time data
 - Create contingency tables and tables of summary statistics
 - Fit logistic, Poisson, and Cox proportional hazards models
 - Obtain goodness-of-fit statistics

Terminology

- Epidemiology is the study of the distribution and determinants of disease in human populations (Woodward, 2014)
 - For example, we might look at the occurrence of a disease across regions or social class
- We'll use the term **risk factor** to refer to attributes or factors that could potentially cause the disease

Epidemiologic studies

- Studies may be classified as prospective or retrospective
- These terms have been defined in different ways. Rothman et al. (2008) recommend defining them in terms of whether the disease could influence our information on exposure. So we will define them as follows:
 - a **prospective study** is one in which the recording of exposure occurs before the occurrence of disease
 - a **retrospective study** is one in which the occurrence of the disease precedes recording of the exposure
 - A study in which we ask lung cancer patients about their smoking habits (a risk factor) over the last few years would be an example of a retrospective study

Epidemiologic studies

- Cohort studies
 - We follow people over time and monitor the outcome (e.g., disease and death). We define two or more cohorts, or groups, based on their exposure to the risk factor. We can then study the occurrence of the disease and compare across the cohorts.
- Case-control studies
 - We select a number of individuals who have the disease of interest (cases) and a number who don't (controls). We can then study which risk factors differ across the two groups.
- Cross-sectional studies
 - We select a representative sample of the population, regardless of exposure or disease status

NHANES II data

```
. use nhanes2l, clear
(Second National Health and Nutrition Examination Survey)
. describe sex bmi race agegrp diabetes highbp heartatk
```

| Variable name | Storage type | Display format | Value label | Variable label |
|------------------|-----------------|-------------------|----------------|-----------------------|
| sex | byte | %9.0g | sex | Sex |
| bmi | float | %9.0g | | Body mass index (BMI) |
| race | byte | %9.0g | race | Race |
| agegrp | byte | %8.0g | agegrp | Age group |
| diabetes | byte | %12.0g | yesno | Diabetes |
| highbp | byte | %8.0g | yesno | * High BP |
| heartatk | byte | %16.0g | yesno | Prior heart attack |

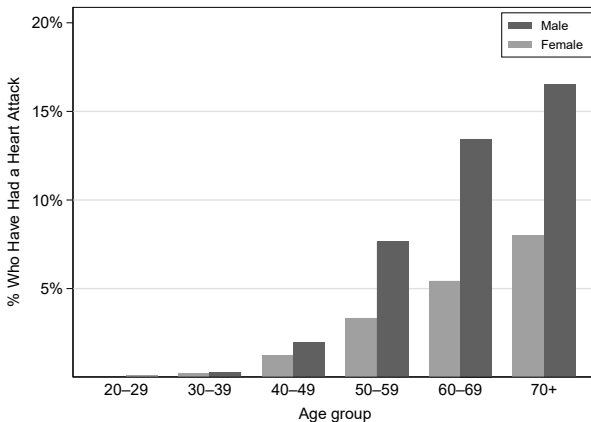
Characteristics of our sample

```
. table (var) (region), ///
> statistic(fvpercent sex race agegrp diabetes highbp heartatk)
```

| | Region | | | | |
|------------------------|--------|-------|-------|-------|-------|
| | NE | MW | S | W | Total |
| Sex=Male | 48.57 | 47.22 | 46.69 | 47.75 | 47.48 |
| Sex=Female | 51.43 | 52.78 | 53.31 | 52.25 | 52.52 |
| Race=White | 94.42 | 88.75 | 80.06 | 89.04 | 87.58 |
| Race=Black | 5.06 | 10.63 | 19.21 | 5.21 | 10.49 |
| Race=Other | 0.52 | 0.61 | 0.74 | 5.75 | 1.93 |
| Age group=20-29 | 21.18 | 24.66 | 20.64 | 22.95 | 22.41 |
| Age group=30-39 | 15.94 | 15.61 | 15.53 | 15.68 | 15.67 |
| Age group=40-49 | 11.74 | 13.55 | 12.79 | 10.84 | 12.29 |
| Age group=50-59 | 13.93 | 11.07 | 13.49 | 11.68 | 12.47 |
| Age group=60-69 | 28.01 | 25.99 | 28.43 | 28.20 | 27.63 |
| Age group=70+ | 9.21 | 9.12 | 9.11 | 10.65 | 9.53 |
| Diabetes=No | 95.32 | 95.49 | 94.36 | 95.62 | 95.18 |
| Diabetes=Yes | 4.68 | 4.51 | 5.64 | 4.38 | 4.82 |
| High BP=No | 55.92 | 59.99 | 57.34 | 57.19 | 57.72 |
| High BP=Yes | 44.08 | 40.01 | 42.66 | 42.81 | 42.28 |
| Prior heart attack=No | 96.32 | 95.64 | 95.41 | 94.41 | 95.40 |
| Prior heart attack=Yes | 3.68 | 4.36 | 4.59 | 5.59 | 4.60 |

Prevalence of heart attacks

```
. graph bar heartatk, over(sex) over(agegrp) asyvars
```



Two-way table

```
. tabulate highbp heartatk, row chi2
```

| Key |
|-----------------------|
| <i>frequency</i> |
| <i>row percentage</i> |

| High BP | Prior heart attack | | Total |
|--------------------------------------|--------------------|-------------|------------------|
| | No | Yes | |
| No | 5,776 96.70 | 197 3.30 | 5,973 100.00 |
| Yes | 4,097 93.62 | 279 6.38 | 4,376 100.00 |
| Total | 9,873 95.40 | 476 4.60 | 10,349 100.00 |
| Pearson chi2(1) = 54.5144 Pr = 0.000 | | | |

The cc command

- `cc` is used to obtain point estimates and confidence intervals for the odds ratio
- It is designed to work with case-control (CC) and cross-sectional data
- The syntax is as follows: `cc var_case var_exposed`
 - `var_case` indicates whether an individual is a case or control
 - `var_exposed` indicates whether an individual was exposed to the risk factor

Odds ratios with cc

```
. cc heartatk highbp
```

| | Exposed | Unexposed | Total | Proportion exposed |
|----------------------------------|----------------|-----------|----------------------|-----------------------|
| Cases | 279 | 197 | 476 | 0.5861 |
| Controls | 4097 | 5776 | 9873 | 0.4150 |
| Total | 4376 | 5973 | 10349 | 0.4228 |
| | Point estimate | | [95% conf. interval] | |
| Odds ratio | 1.996637 | | 1.649529 | 2.417111 (exact) |
| Attr. frac. ex. | .4991579 | | .3937664 | .586283 (exact) |
| Attr. frac. pop | .2925737 | | | |
| chi2(1) = 54.51 Pr>chi2 = 0.0000 | | | | |

Confounding variables

```
. mlogit heartatk highbp, by(race)
```

Maximum likelihood estimate of the odds ratio comparing highbp==1 vs. highbp==0 by race

| race | Odds ratio | chi2(1) | P>chi2 | [95% conf. interval] | |
|-------|------------|---------|--------|----------------------|----------|
| White | 2.006772 | 49.49 | 0.0000 | 1.64630 | 2.44617 |
| Black | 1.823414 | 3.86 | 0.0496 | 0.99208 | 3.35137 |
| Other | 5.397590 | 2.77 | 0.0963 | 0.57998 | 50.23292 |

Mantel-Haenszel estimate controlling for race

| Odds ratio | chi2(1) | P>chi2 | [95% conf. interval] | |
|------------|---------|--------|----------------------|----------|
| 2.006005 | 55.16 | 0.0000 | 1.663129 | 2.419570 |

Approximate test of homogeneity of odds ratios: chi2(2) = 0.91
 Pr>chi2 = 0.6330

Logistic regression

```
. logistic heartatk bmi i.diabetes i.highbp i.agegrp i.race
```

Logistic regression

Number of obs = 10,349

LR chi2(10) = 576.83

Prob > chi2 = 0.0000

Pseudo R2 = 0.1494

Log likelihood = -1642.1778

| heartatk | Odds ratio | Std. err. | z | P> z | [95% conf. interval] | |
|----------|------------|-----------|-------|-------|----------------------|----------|
| bmi | 1.005872 | .0101807 | 0.58 | 0.563 | .986115 | 1.026025 |
| diabetes | | | | | | |
| No | 1 (base) | | | | | |
| Yes | 1.828345 | .2696365 | 4.09 | 0.000 | 1.369388 | 2.441123 |
| highbp | | | | | | |
| No | 1 (base) | | | | | |
| Yes | .9669678 | .0984323 | -0.33 | 0.741 | .792071 | 1.180483 |
| agegrp | | | | | | |
| 20-29 | 1 (base) | | | | | |
| 30-39 | 5.662283 | 6.333842 | 1.55 | 0.121 | .6321728 | 50.71628 |
| 40-49 | 36.04339 | 36.96869 | 3.50 | 0.000 | 4.828057 | 269.0785 |
| 50-59 | 125.1551 | 126.2456 | 4.79 | 0.000 | 17.33127 | 903.7883 |
| 60-69 | 222.8061 | 223.5441 | 5.39 | 0.000 | 31.18213 | 1592.019 |
| 70+ | 288.5501 | 290.4329 | 5.63 | 0.000 | 40.12969 | 2074.803 |
| race | | | | | | |
| White | 1 (base) | | | | | |
| Black | .9832512 | .1600566 | -0.10 | 0.917 | .7146664 | 1.352775 |
| Other | .6411202 | .2983426 | -0.96 | 0.339 | .2575351 | 1.596035 |
| _cons | .0003793 | .00039 | -7.66 | 0.000 | .0000505 | .0028464 |

Pearson goodness-of-fit test

```
. estat gof
```

Goodness-of-fit test after logistic model
Variable: heartatk

```
      Number of observations =    10,349  
Number of covariate patterns =    10,309  
      Pearson chi2(10298) = 10192.58  
              Prob > chi2 =     0.7680
```

Logistic regression

```
. logistic highbp bmi i.sex i.agegrp i.race
```

Logistic regression

Number of obs = 10,351
 LR chi2(9) = 2412.47
 Prob > chi2 = 0.0000
 Pseudo R2 = 0.1711

Log likelihood = -5844.528

| highbp | Odds ratio | Std. err. | z | P> z | [95% conf. interval] | |
|--------|------------|-----------|--------|-------|----------------------|----------|
| bmi | 1.145456 | .0058679 | 26.51 | 0.000 | 1.134013 | 1.157015 |
| sex | | | | | | |
| Male | 1 (base) | | | | | |
| Female | .6154242 | .0277887 | -10.75 | 0.000 | .5632999 | .6723718 |
| agegrp | | | | | | |
| 20-29 | 1 (base) | | | | | |
| 30-39 | 1.651423 | .1362054 | 6.08 | 0.000 | 1.404925 | 1.941169 |
| 40-49 | 2.695465 | .226855 | 11.78 | 0.000 | 2.285574 | 3.178867 |
| 50-59 | 4.951629 | .41051 | 19.30 | 0.000 | 4.209011 | 5.825271 |
| 60-69 | 6.135048 | .4329176 | 25.71 | 0.000 | 5.342607 | 7.045028 |
| 70+ | 9.487106 | .8681766 | 24.59 | 0.000 | 7.929381 | 11.35085 |
| race | | | | | | |
| White | 1 (base) | | | | | |
| Black | 1.447319 | .106917 | 5.00 | 0.000 | 1.252229 | 1.672802 |
| Other | 1.606787 | .2580201 | 2.95 | 0.003 | 1.172927 | 2.201129 |
| _cons | .0083758 | .0011958 | -33.50 | 0.000 | .0063314 | .0110804 |

Note: _cons estimates baseline odds.

Pearson goodness-of-fit test

```
. estat gof
```

Goodness-of-fit test after logistic model
Variable: highbp

```
      Number of observations =    10,351  
Number of covariate patterns =    10,295  
      Pearson chi2(10285) = 10185.79  
              Prob > chi2 =     0.7547
```

Hosmer-Lemeshow goodness-of-fit test

```
. estat gof, group(10) table
```

note: obs collapsed on 10 quantiles of estimated probabilities.

Goodness-of-fit test after logistic model
 Variable: highbp

Table collapsed on quantiles of estimated probabilities

| Group | Prob | Obs_1 | Exp_1 | Obs_0 | Exp_0 | Total |
|-------|--------|-------|-------|-------|-------|-------|
| 1 | 0.1197 | 56 | 91.8 | 980 | 944.2 | 1036 |
| 2 | 0.1761 | 153 | 152.5 | 882 | 882.5 | 1035 |
| 3 | 0.2531 | 209 | 220.9 | 826 | 814.1 | 1035 |
| 4 | 0.3395 | 334 | 306.4 | 701 | 728.6 | 1035 |
| 5 | 0.4219 | 417 | 394.6 | 618 | 640.4 | 1035 |
| 6 | 0.5010 | 507 | 477.9 | 528 | 557.1 | 1035 |
| 7 | 0.5752 | 569 | 557.4 | 466 | 477.6 | 1035 |
| 8 | 0.6441 | 629 | 631.1 | 406 | 403.9 | 1035 |
| 9 | 0.7280 | 691 | 708.3 | 344 | 326.7 | 1035 |
| 10 | 0.9869 | 811 | 835.2 | 224 | 199.8 | 1035 |

```
Number of observations = 10,351
Number of groups      = 10
Hosmer-Lemeshow chi2(8) = 30.52
Prob > chi2 = 0.0002
```

Classification table

```
. estat classification
```

Logistic model for highbp

| Classified | True | | Total |
|------------|------|------|-------|
| | D | ~D | |
| + | 2708 | 1446 | 4154 |
| - | 1668 | 4529 | 6197 |
| Total | 4376 | 5975 | 10351 |

Classified + if predicted $\Pr(D) \geq .5$

True D defined as highbp != 0

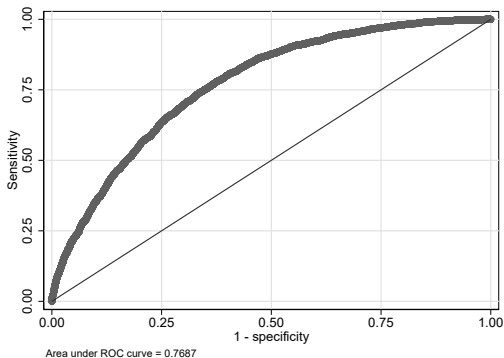
| | | |
|---------------------------|-----------------|--------|
| Sensitivity | $\Pr(+ D)$ | 61.88% |
| Specificity | $\Pr(- \sim D)$ | 75.80% |
| Positive predictive value | $\Pr(D +)$ | 65.19% |
| Negative predictive value | $\Pr(\sim D -)$ | 73.08% |

| | | |
|-------------------------------|-----------------|--------|
| False + rate for true ~D | $\Pr(+ \sim D)$ | 24.20% |
| False - rate for true D | $\Pr(- D)$ | 38.12% |
| False + rate for classified + | $\Pr(\sim D +)$ | 34.81% |
| False - rate for classified - | $\Pr(D -)$ | 26.92% |

| | |
|----------------------|--------|
| Correctly classified | 69.92% |
|----------------------|--------|

ROC curve

```
. lroc
```



Case-control study from Doll and Hill

- Sir Richard Doll and Sir Austin Bradford Hill conducted a case-control study in 1948 and 1949
- They asked lung cancer patients (cases) about their smoking habits
- For each case they recruited a non-cancer patient from the same hospital, of a similar age group and same sex
- They found that cases were more likely to be heavy smokers
- But, this study was susceptible to bias

Seminal cohort study from Doll and Hill

- To establish a non-negligible link between smoking and death, Doll and Hill began conducting a cohort study in 1951
- They sent a questionnaire to everyone on the British medical register and followed the subjects for many years
- In cohort studies, it is possible that the risk behavior (smoking in this case) could change with time
- To address this concern, Doll and Hill would send follow-up questionnaires to update their records of the patients' smoking habits
- This cohort study confirmed a relationship between smoking and death
 - All deaths were recorded, so this dataset was also used to study mortality for other diseases

Doll and Hill (1966)

```
. webuse dollhill3, clear
(Doll and Hill (1966))
. list
```

| | agecat | smokes | deaths | pyears |
|-----|--------|--------|--------|--------|
| 1. | 35-44 | 1 | 32 | 52,407 |
| 2. | 45-54 | 1 | 104 | 43,248 |
| 3. | 55-64 | 1 | 206 | 28,612 |
| 4. | 65-74 | 1 | 186 | 12,663 |
| 5. | 75-84 | 1 | 102 | 5,317 |
| 6. | 35-44 | 0 | 2 | 18,790 |
| 7. | 45-54 | 0 | 12 | 10,673 |
| 8. | 55-64 | 0 | 28 | 5,710 |
| 9. | 65-74 | 0 | 28 | 2,585 |
| 10. | 75-84 | 0 | 31 | 1,462 |

Incidence rate

- We have the number of deaths (cases) for each age group and category of smoke
- Smoke is our exposure variable
- Person-years is a measure of how long individuals were at risk of death
 - If patient 1 lived to 40 and patient 2 lived to 60, together they contribute 100 person-years
- The incidence rate tells us how many individuals died out of all the person-years contributed. More generally:

$$IR = \frac{\# \text{ of disease onsets}}{\text{total person-time}}$$

Incidence-rate ratios

```
. ir deaths smokes pyears
```

Incidence-rate comparison

| | Whether person smokes | | Total |
|------------------|-----------------------|-----------|----------------------|
| | Exposed | Unexposed | |
| Number of deaths | 630 | 101 | 731 |
| Person-years | 142247 | 39220 | 181467 |
| Incidence rate | .0044289 | .0025752 | .0040283 |
| | Point estimate | | [95% conf. interval] |
| Inc. rate diff. | .0018537 | .0012439 | .0024635 |
| Inc. rate ratio | 1.719823 | 1.391992 | 2.14353 (exact) |
| Attr. frac. ex. | .4185447 | .281605 | .5334797 (exact) |
| Attr. frac. pop | .3607157 | | |

Mid p-values for tests of incidence-rate difference:

Adj Pr(Exposed Number of deaths <= 630) = 1.0000 (lower one-sided)

Adj Pr(Exposed Number of deaths >= 630) = 0.0000 (upper one-sided)

Two-sided p-value = 0.0000

Stratified IRRs

```
. ir deaths smokes pyyears, by(age)
```

Stratified incidence-rate analysis

| Age category | IRR | [95% conf. interval] | | M-H weight |
|--------------|----------|----------------------|----------|------------------|
| 35-44 | 5.736638 | 1.463557 | 49.40468 | 1.472169 (exact) |
| 45-54 | 2.138812 | 1.173714 | 4.272545 | 9.624747 (exact) |
| 55-64 | 1.46824 | .9863624 | 2.264107 | 23.34176 (exact) |
| 65-74 | 1.35606 | .9081925 | 2.096412 | 23.25315 (exact) |
| 75-84 | .9047304 | .6000757 | 1.399687 | 24.31435 (exact) |
| Crude | 1.719823 | 1.391992 | 2.14353 | (exact) |
| M-H combined | 1.424682 | 1.154703 | 1.757784 | |

Test of homogeneity (M-H): $\chi^2(4) = 10.41$

$\text{Pr} > \chi^2 = 0.0340$

Poisson regression model

- Below is a representation of the Poisson regression model

$$\ln(C) = b_0 + b_1x_1 + b_2x_2 + \dots + b_kx_k + \ln(T)$$

where C represents a count, T represents person-time, and we have k explanatory variables

- The estimated coefficients are the natural logs of the rate ratios ($b = \ln(rr)$)
- We can exponentiate the coefficients to get the rate ratios ($\exp^b = \exp^{\ln(rr)} = rr$)
- Note that the coefficient for person-time is constrained to 1; this term is an offset

Poisson regression

```
. poisson deaths i.smokes i.agecat, exposure(pyears) irr
```

Poisson regression

Number of obs = 10
LR chi2(5) = 922.93
Prob > chi2 = 0.0000
Pseudo R2 = 0.9321

Log likelihood = -33.600153

| deaths | IRR | Std. err. | z | P> z | [95% conf. interval] | |
|------------|----------|------------|--------|-------|----------------------|----------|
| smokes | | | | | | |
| 0 | 1 | (base) | | | | |
| 1 | 1.425519 | .1530638 | 3.30 | 0.001 | 1.154984 | 1.759421 |
| agecat | | | | | | |
| 35-44 | 1 | (base) | | | | |
| 45-54 | 4.410584 | .8605197 | 7.61 | 0.000 | 3.009011 | 6.464997 |
| 55-64 | 13.8392 | 2.542638 | 14.30 | 0.000 | 9.654328 | 19.83809 |
| 65-74 | 28.51678 | 5.269878 | 18.13 | 0.000 | 19.85177 | 40.96395 |
| 75-84 | 40.45121 | 7.775511 | 19.25 | 0.000 | 27.75326 | 58.95885 |
| _cons | .0003636 | .0000697 | -41.30 | 0.000 | .0002497 | .0005296 |
| ln(pyears) | 1 | (exposure) | | | | |

Note: _cons estimates baseline incidence rate.

Predicting the number of events

```
. margins agecat
```

```
Predictive margins
```

Number of obs = 10

```
Model VCE: OIM
```

```
Expression: Predicted number of events, predict()
```

| | Delta-method | | z | P> z | [95% conf. interval] | |
|--------|--------------|-----------|-------|-------|----------------------|----------|
| | Margin | std. err. | | | | |
| agecat | | | | | | |
| 35-44 | 8.800113 | 1.509654 | 5.83 | 0.000 | 5.841245 | 11.75898 |
| 45-54 | 38.81363 | 3.604262 | 10.77 | 0.000 | 31.74941 | 45.87786 |
| 55-64 | 121.7865 | 7.977974 | 15.27 | 0.000 | 106.15 | 137.4231 |
| 65-74 | 250.9509 | 17.18329 | 14.60 | 0.000 | 217.2723 | 284.6296 |
| 75-84 | 355.9752 | 30.86697 | 11.53 | 0.000 | 295.477 | 416.4733 |

```
. display 8.800113*4.4105
```

```
38.812898
```

Predicting incidence rates

```
. margins agecat, predict(ir)
```

Predictive margins

Number of obs = 10

Model VCE: OIM

Expression: Predicted incidence rate, predict(ir)

| | Delta-method | | | | | |
|--------|--------------|-----------|-------|-------|----------------------|----------|
| | Margin | std. err. | z | P> z | [95% conf. interval] | |
| agecat | | | | | | |
| 35-44 | .000441 | .0000763 | 5.78 | 0.000 | .0002915 | .0005905 |
| 45-54 | .0019451 | .0001889 | 10.30 | 0.000 | .001575 | .0023153 |
| 55-64 | .0061033 | .0004418 | 13.81 | 0.000 | .0052374 | .0069692 |
| 65-74 | .0125764 | .000943 | 13.34 | 0.000 | .0107281 | .0144247 |
| 75-84 | .0178397 | .0016197 | 11.01 | 0.000 | .0146652 | .0210141 |

```
. display 0.000441*40.45
.01783845
```

Goodness-of-fit tests

```
. estat gof
```

```
Deviance goodness-of-fit = 12.13237  
Prob > chi2(4)           = 0.0164  
  
Pearson goodness-of-fit  = 11.15533  
Prob > chi2(4)           = 0.0249
```


Fitted/expected counts

```
. predict counts_hat
(option n assumed; predicted number of events)
. summ counts_hat
```

| Variable | Obs | Mean | Std. dev. | Min | Max |
|------------|-----|------|-----------|----------|----------|
| counts_hat | 10 | 73.1 | 73.81032 | 6.832935 | 205.2639 |

Cumulative incidence

- Another way to express the occurrence of a disease is the proportion of those at risk who become diseased
- This is referred to as the cumulative incidence or the incidence proportion:

$$\frac{\text{\# of disease onsets}}{\text{population at risk}}$$

University Group Diabetes Program

- We have data from the University Group Diabetes Program, which was designed to assess the effectiveness of certain drugs on complications of diabetes
- In this clinical trial, some individuals were given a placebo and others were given tolbutamide, a hypoglycemic drug

Cumulative incidence data

```
. webuse ugdg, clear
(University Group Diabetes Program 1970)
. list, sepby(case)
```

| | age | case | exposed | pop |
|----|-----|-----------|-------------|-----|
| 1. | <55 | Surviving | Placebo | 115 |
| 2. | <55 | Surviving | Tolbutamide | 98 |
| 3. | <55 | Dead | Placebo | 5 |
| 4. | <55 | Dead | Tolbutamide | 8 |
| 5. | 55+ | Surviving | Placebo | 69 |
| 6. | 55+ | Surviving | Tolbutamide | 76 |
| 7. | 55+ | Dead | Placebo | 16 |
| 8. | 55+ | Dead | Tolbutamide | 22 |

- pop reflects the number of individuals in each group

Risk ratios

```
. cs case exposed [fweight = pop], by(age)
```

| Age category | Risk ratio | [95% conf. interval] | | M-H weight |
|--------------|------------|----------------------|----------|------------|
| <55 | 1.811321 | .6112044 | 5.367898 | 2.345133 |
| 55+ | 1.192602 | .6712664 | 2.11883 | 8.568306 |
| Crude | 1.435574 | .8510221 | 2.421645 | |
| M-H combined | 1.325555 | .797907 | 2.202132 | |

Test of homogeneity (M-H) $\chi^2(1) = 0.447$ $\text{Pr}>\chi^2 = 0.5037$

Survival-time data

- We measure time to an event of interest
- The occurrence of the event is typically called a failure
- An observation is censored if we don't know the exact time of failure
- Stata's `st` suite of commands is designed for analyzing survival-time data

A look at survival data

- One record per patient

| Patient ID | Sex | Days | Died |
|------------|--------|------|------|
| 1 | Male | 89 | Yes |
| 2 | Female | 91 | No |
| 3 | Male | 90 | Yes |

A look at survival data

- Multiple-record data

| Patient ID | Sex | Days | Died |
|------------|--------|------|------|
| 1 | Male | 33 | No |
| 1 | Male | 89 | Yes |
| 2 | Female | 33 | No |
| 2 | Female | 91 | No |
| 3 | Male | 32 | No |
| 3 | Male | 90 | Yes |

Other notes on survival data

- There are other varieties
 - A subject might be diagnosed before the study starts, meaning they are at risk before we observe them (delayed entry).
 - There might be a gap between the time the subject entered the study and the time the study ended. Suppose the patient was traveling and unable to be reached for a month in the middle of the study but returned before the study ended.
 - You might have multiple-failure data.
- We won't be focusing on these types of complications, but Stata's commands for analyzing survival-time data accommodate data with these features.

Single-observation survival-time data

```
. webuse drugtr, clear
(Patient survival in drug trial)
. describe studytime-age
```

| Variable name | Storage type | Display format | Value label | Variable label |
|------------------|-----------------|-------------------|----------------|--------------------------------|
| studytime | byte | %8.0g | | Months to death or end of exp. |
| died | byte | %8.0g | | 1 if patient died |
| drug | byte | %8.0g | | Drug type (0=placebo) |
| age | byte | %8.0g | | Patient's age at start of exp. |

Display survival-time settings

```
. stset
-> stset studytime, failure(died)

Survival-time data settings

      Failure event: died!=0 & died<.
Observed time interval: (0, studytime]
      Exit on or before: failure
```

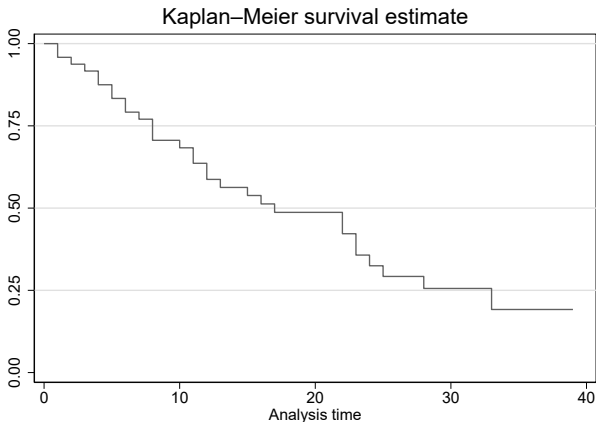
```
      48  total observations
      0  exclusions
```

```
      48  observations remaining, representing
      31  failures in single-record/single-failure data
      744 total analysis time at risk and under observation

                                At risk from t =           0
                                Earliest observed entry t =       0
                                Last observed exit t =          39
```

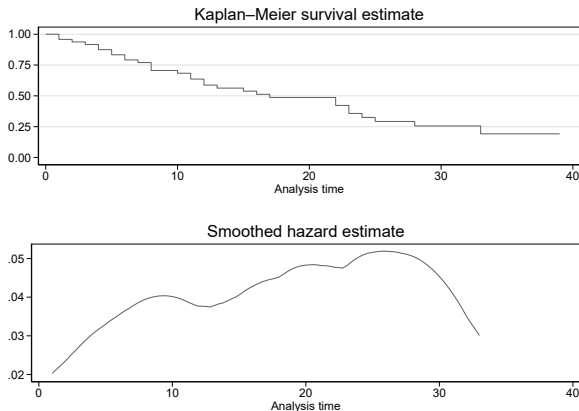
Kaplan-Meier survivor function

```
. sts graph
```



• $S(t) = \Pr(T > t)$

Survivor and hazard functions



Cox Proportional hazards model

$$h(t) = h_0(t) \exp(\beta_1 x_1 + \dots + \beta_k x_k)$$

where $h_0(t)$ is the baseline hazard

- The hazard depends on the covariates; we estimate their coefficients (β_k).
- We assume the hazard ratio ($\exp(\beta_k)$) is fixed over time.

Cox Proportional hazards model

```
. stcox drug age
      Failure _d: died
      Analysis time _t: studytime
Cox regression with Breslow method for ties

No. of subjects = 48                Number of obs =      48
No. of failures = 31
Time at risk    = 744

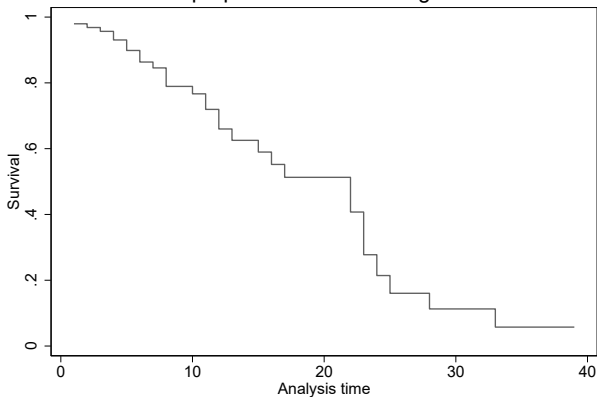
Log likelihood = -83.323546          LR chi2(2)      = 33.18
                                      Prob > chi2     = 0.0000
```

| _t | Haz. ratio | Std. err. | z | P> z | [95% conf. interval] | |
|------|------------|-----------|-------|-------|----------------------|----------|
| drug | .1048772 | .0477017 | -4.96 | 0.000 | .0430057 | .2557622 |
| age | 1.120325 | .0417711 | 3.05 | 0.002 | 1.041375 | 1.20526 |

Survivor function

```
. stcurve, survival
```

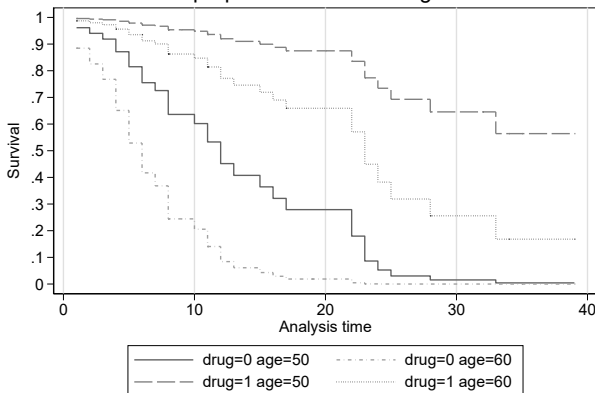
Cox proportional hazards regression



Survivor function

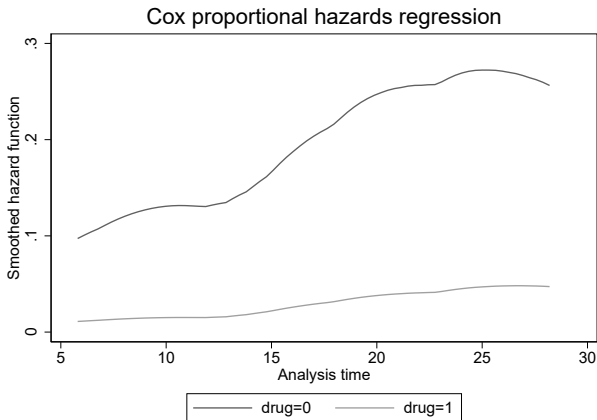
```
. stcurve, survival at(drug=(0 1) age=(50 60))
```

Cox proportional hazards regression



Hazard function

```
. stcurve, hazard at(drug=(0 1))
```



Assessing our model

- Statistics
 - How well do our predictions agree with the outcomes?
 - Does the proportional-hazards assumption hold?
- Diagnostic plots
 - Plot of residuals versus time
 - Log-log plots
 - Comparison of the observed survival curve and the Cox predicted curve

Concordance probability

```
. estat concordance, gheller
      Failure _d: died
      Analysis time _t: studytime
Gonen and Heller's K concordance statistic
      Number of subjects (N)      =      48
      Gonen and Heller's K =      0.7748
      Somers' D =      0.5496
```

Test the proportional hazards assumption

```
. estat phtest, detail
```

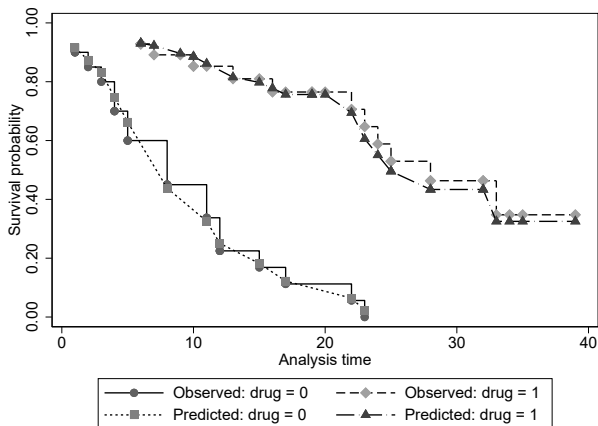
Test of proportional-hazards assumption

Time function: Analysis time

| | rho | chi2 | df | Prob>chi2 |
|-------------|----------|------|----|-----------|
| drug | 0.00949 | 0.00 | 1 | 0.9603 |
| age | -0.11758 | 0.42 | 1 | 0.5168 |
| Global test | | 0.43 | 2 | 0.8064 |

Kaplan–Meier and predicted survival plots

```
. stcoxkm, by(drug)
```



More analyses of interest

- For *matched case-control data*, obtain differences and ratios of the proportion of subjects with the factor with `mcc`
- Fit *generalized linear models*
- Fit *negative binomial regression models*
- Learn more about working with *survival-time data*
- Perform *meta analysis*

Other analyses of interest

- Additionally, Stata has elegant tools for
 - working with *complex survey data*
 - performing *multiple imputation* and *Bayesian analysis*
 - performing *precision and sample-size analysis*
 - fitting *multilevel models* and *structural equation models*
 - analyzing *treatment effects* in observational data

Resources

- The Stata *YouTube channel*
- Stata's Technical Support (tech-support@stata.com)
- The *Stata Blog*
 - Learn to import COVID-19 data from Johns Hopkins University and create choropleth maps with these data
 - Learn to fit Bayesian regression models, create animated graphics, and more

References

- Cummings, P. 2019. *Analysis of Incidence Rates*. Boca Raton, FL: CRC Press.
- McDowell, A., A. Engel, J. T. Massey, and K. Maurer. 1981. Plan and operation of the Second National Health and Nutrition Examination Survey, 1976–1980. *Vital and Health Statistics* 1(15): 1–144.
- Rothman, K. J., Greenland, S., & Lash, T. L. (2008). *Modern Epidemiology* 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins.
- University Group Diabetes Program. 1970. *A study of the effects of hypoglycemic agents on vascular complications in patients with adult-onset diabetes, II: Mortality results*. *Diabetes* 19, supplement 2: 789–830.
- Woodward, M. 2014. *Epidemiology: Study Design and Data Analysis*. 3rd ed. Boca Raton, FL: Chapman & Hall/CRC.