STATISTICS FOR ECN 132 HEALTH ECONOMICS

Colin Cameron

Economics, UCD

Spring 2021

Overview

- This handout presents the key statistical methods used in this course.
- Descriptive statistics for a sample $x_1, ..., x_n$
  - e.g. sample mean $\bar{x}$.
- Statistical inference on the population mean μ
  - confidence intervals and hypothesis tests for μ.
- Linear (OLS) regression for a sample $(y_1, x_1), ..., (y_n, x_n)$
  - fitting a linear relationship between y and x.
- Inference on the slope β of the regression line
  - confidence intervals and hypothesis tests for β.
- The special case of regression with a binary regressor (or regressors)
  - difference in means and difference in several means.
- Differences in differences.
- Elasticities and regression with logs.

Outline

1. Statistical software
2. Sample descriptive statistics
3. Statistical inference on μ
    1. What is statistical inference?
    2. Estimation, confidence interval and tests on μ
4. The general principle for statistical inference
5. Difference in two means
6. Linear (OLS) regression of y on x
7. Statistical inference following regression
8. Regression on only an intercept (inference on mean)
9. Regression on a binary regressor (difference in two means)
10. Regression on several binary regressors (difference in several means)
11. Differences in differences (for a natural experiment)
12. Elasticities and regression with logs
13. Stata code used for these notes

1. STATISTICAL SOFTWARE

- In this course we use **Stata**
  - see http://cameron.econ.ucdavis.edu/stata/stataintro.html
  - and for more see http://cameron.econ.ucdavis.edu/stata/stata.html
  - and the code given at the end of these slides.
- Stata is available in some university computer labs
  - see http://www.stata.com/order/new/edu/gradplans/student-pricing/ to purchase your own copy
  - the cheapest is Stata/IC which is enough for all of this course and other Economics courses.
- Other software that we might have used includes
  - R (free), GretI (free), Eviews, SPSS, SAS.

2. SAMPLE: Rand Health Insurance Experiment year 5

```
. * The complete dataset
. use tr132statistics.dta, clear

. describe

Contains data from tr132statistics.dta
  obs:         1,035
 vars:             9                          15 Oct 2016 12:45
 size:        38,295
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
outspend        double  %9.0g                 Outpatient spending in year 5 (2012$)
age             byte    %10.0g                Age in years
badhealth       float   %9.0g                 =1 if health poor and =0 otherwise
coins0          float   %9.0g                 =1 if Free Care and =0 otherwise
coins25         float   %9.0g                 =25% Coins and =0 otherwise
coins50         float   %9.0g                 =25% Coins and =0 otherwise
coins95         float   %9.0g                 =95% Coins and =0 otherwise
coinsindiv      float   %9.0g                 =Indiv Deduct and =0 otherwise
plan            float   %9.0g                 Coinsurance type =1,2,4,5, or 6
-------------------------------------------------------------------------------
```

Summary Statistics

- Summary statistics for outpatient spending (excludes inpatient hospital)
  - Mean is the **sample mean:** $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 = x_2 + \cdots x_n}{n}$.
  - Std. dev. is the **sample standard deviation:** $s_x = \sqrt{s_x^2}$
  - where s_X^2 is the **sample variance:** $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.
  - Min and Max are the minimum and maximum values.

```
. * Summary statistics for outpatient spending
. summarize outspend
```

Variable	Obs	Mean	Std. Dev.	Min	Max
outspend	1,035	1200.503	2085.393	0	28519.19

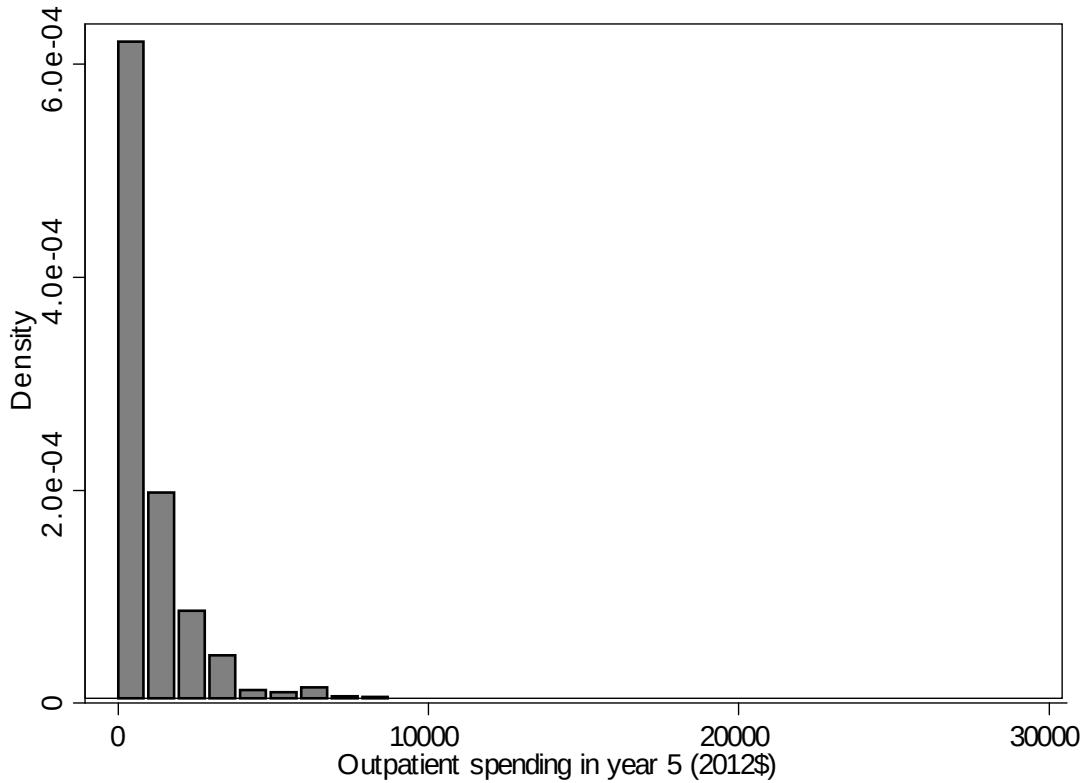
Histogram

- For a histogram
  - `set scheme s1mono` is best for black-and-white printing
  - here export as Windows metafile (`.wmf`) but other formats possible.
- Here do histogram only for positive spending (13% had zero).

```
. * Histogram if outspend > 0
. set scheme s1mono

. hist outspend if outspend > 0
(bin=29, start=8.858e-07, width=983.42051)

. graph export histogram.wmf, replace
```

Histogram (continued)

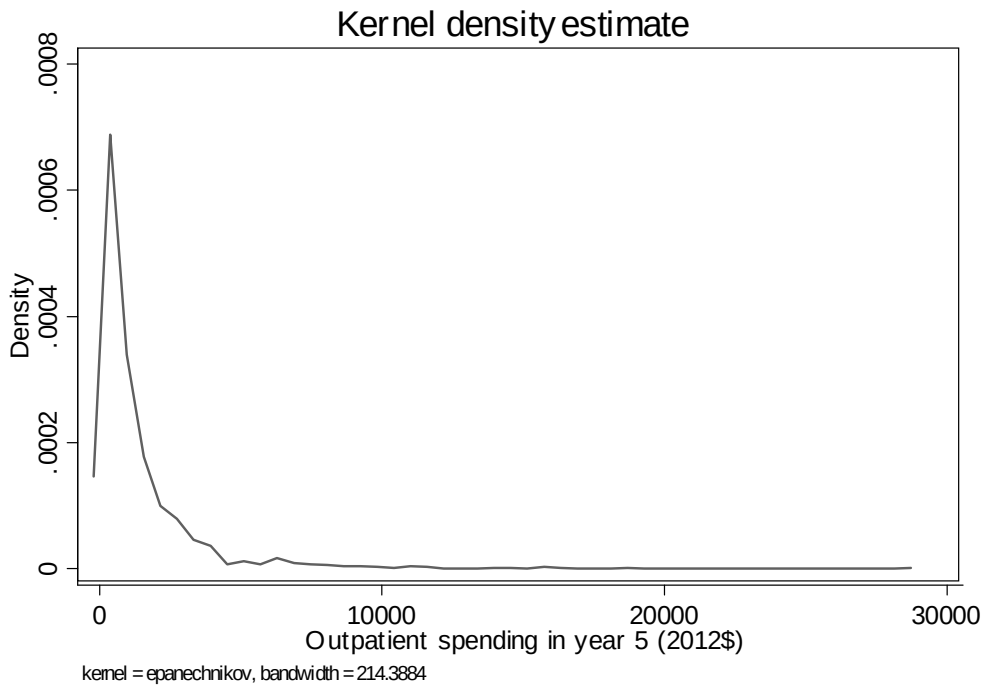
- Histogram counts the number of observations in each bin
  - this default version then sets the vertical axis so that the shaded area equals one.
- Outpatient spending is very right skewed with some very large values.



Smoothed Histogram

- For continuous data it is better to use a smoothed histogram (a kernel density estimate).

```
. * Smoothed histogram (kernel density estimate) if > 0
. kdensity outspend if outspend > 0
```



3. STATISTICAL INFERENCE

- Statistical inference **extrapolates** from a **sample** to the **population**.
- We let X denote the **random variable** of interest
  - e.g. X = outpatient spending of an individual in the population.
- Three common measures (called parameters) of the distribution of values that X takes are
  - **mean** $\mu = E[X]$: the probability-weighted population average of X
  - **variance** $\sigma^2 = E[(X - \mu)^2]$: the probability-weighted population average of $(\mathrm{X}-\mu)^2$
  - **standard deviation** $\sigma = \sqrt{\sigma^2}$: variance rescaled to the units of X.
- If X is normally distributed then we expect
  - 68% of observations are in the range $(\mu - \sigma,\ \mu + \sigma)$
  - 95% of observations are in the range $(\mu - 2\sigma,\ \mu + 2\sigma)$
  - 99.7% of observations are in the range $(\mu - 3\sigma,\ \mu + 3\sigma)$.

Example of population mean and standard deviation

- Let X = number of doctor visits and suppose

x	0	1	2
$\Pr[X = x]$	0.5	0.3	0.2

  - aside: note that probabilities sum to one (0.5+0.3+0.2=1).

- **Mean**: $\mu = E[X] = \sum_x \Pr[X = x] \times x$
  $= 0.5 \times 0 + 0.3 \times 1 + 0.2 \times 2 = 0.7.$
- **Variance**: $\sigma^2 = E[(X - \mu)^2] = \sum_x \Pr[X = x] \times (x - \mu)^2$
  $= 0.5 \times (0 - 0.7)^2 + 0.3 \times (1 - 0.7)^2 + 0.2 \times (2 - 0.7)^2$
  $= 0.5 \times 0.49 + 0.3 \times 0.09 + 0.2 \times 1.69 = 0.61.$
- **Standard deviation**: $\sigma = \sqrt{0.61} = 0.78102.$

Statistical Inference on the population mean

- We focus on determining the likely range of values for the population mean μ given the sample at hand.
- The sample is assumed to be a random sample
  - sample values are independently drawn
  - from common distribution with mean μ and standard deviation σ.
- 1. The standard single estimate of μ is the sample mean $\bar{x}$.
- 2. A confidence interval gives a range of likely values of μ.
- 3. A hypothesis test accepts or rejects whether μ lies in a hypothesized range.

Estimation of the population mean

- The standard single **estimate** of μ is the sample mean $\bar{x}$.
- Different samples have different sample means and hence different estimates of μ
  - the estimated standard deviation of $\bar{x}$ measures this variability
  - this is called the standard error of $\bar{x}$.
- The **standard error of** $\bar{x}$:
  - $se(\bar{x}) = s_x/\sqrt{n}$ where $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.
- Earlier Stata output for `outspend` gave $s_x = 2085.393$ and $n = 1035$ so

$$se(\bar{x}) = \frac{s_x}{\sqrt{n}} = \frac{2085.393}{\sqrt{1035}} = 64.8231.$$

Confidence interval for the population mean

- A 95% confidence interval is one that 95% of the time (in repeated samples) will include the unknown μ and 5% of the time will not.
- A **95% confidence interval** for μ is approximately

$$\bar{x} \pm 1.96 \times se(\bar{x}) \text{ equals } \bar{x} \pm 1.96 \times (s_x/\sqrt{n}).$$

  - Simpler is to replace 1.96 with 2.
- The exact confidence interval (assuming data are normally distributed) replaces 1.96 with $t_{.025}(n-1)$
  - this is the value of Students $T(n-1)$ distribution that has area 0.95 in the middle (and hence area 0.05 in the two tails).
  - e.g. with $n = 100$ use Stata command `invttail(99,.025)`.
- For `outspend` we have $\bar{x} = 1200.503$ and $se(\bar{x}) = 64.8231$ so the 95% confidence interval is approximately

$$1200.503 \pm 1.96 \times 64.8231 = (1073.45,\ 1327.56).$$

Stata command mean

- Stata command `mean` computes the standard error and 95% confidence interval

```
. * Standard error and 95% confidence interval for mu
. mean outspend

Mean estimation                   Number of obs    =       1,035

------------------------------------------------------------------------------
             |       Mean   Std. Err.      [95% Conf. Interval]
-------------+----------------------------------------------------------------
    outspend |   1200.503   64.82131       1073.307    1327.699
------------------------------------------------------------------------------
```

- The standard error $se(\bar{x}) = s_x/\sqrt{n} = 64.821$ as already computed.
- The 95% confidence interval is slightly different from the simple one computed using 1.96
  - it uses invttail(1034,0.025) = 1.9622609 rather than 1.96.
- To get e.g. a 90% confidence interval
  - give command mean outspend, level(90)

Hypothesis tests on the population mean

- A hypothesis test tests whether or not μ lies in a hypothesized range.
- A two sided test is a test of $H_0 : \mu = \mu^*$ against $H_a : \mu \neq \mu^*$.
- The test is based on how far the estimate $\bar{x}$ is from μ^* after normalizing by the standard error of $\bar{x}$ which is our estimate of precision.
- The t-test statistic is
$$t = \frac{\bar{x} - \mu^*}{se(\bar{x})}.$$
- We reject H_0 if $|t|$ is large
  - if testing at level 0.05 then reject H_0 if $|t| > 1.96$.
- Alternatively compute the p-value $\Pr[|T_{n-1}| > |t|]$
  - where T_{n-1} is a random variable with $T(n-1)$ distribution
  - if testing at level 0.05 then reject H_0 if $p < 0.05$.

Stata command test

- For our example test $H_0 : \mu = 1000$ against $H_a : \mu \neq 1000$.

$$t = \frac{\bar{x} - \mu^*}{se(\bar{x})} = \frac{1200.503 - 1000}{64.8231} = 3.093.$$

  - Reject H_0 at level 0.05 as $|t| = 3.093 > 1.96$.
- Stata command `test` does this but uses an F-test which equals t^2

```
. * Hypothesis test that population mean outspend = 1000
. quietly summarize outspend

. test outspend = 1000

 ( 1)  outspend = 1000

       F(  1,  1034) =    9.57
            Prob > F =    0.0020
```

- Reject H_0 at level 0.05 as $p = .0020 < 0.05$
  - Note that $t^2 = 3.093^2 = 9.57$ which equals F.

4. THE GENERAL PRINCIPLE FOR STATISTICAL INFERENCE

- The following works for standard estimators such as in this course.
- Notation:
  - θ denotes the unknown population parameter
  - $\widehat{\theta}$ denotes the estimate from our sample
  - $\text{se}(\widehat{\theta})$ denotes the standard error of the estimate $\widehat{\theta}$
- Assumption:
  - the sample is large enough that we can view $\widehat{\theta}$ as approximately normally distributed.
- A **95% confidence interval** for θ is approximately

$$\widehat{\theta} \pm 1.96 \times se(\widehat{\theta}) \qquad \text{or} \qquad \text{estimate} \pm 1.96 \times \text{standard error}$$

- A **test statistic** for $H_0 : \theta = \theta^*$ against $H_a : \theta \neq \theta^*$ is

$$t = \frac{\widehat{\theta} - \theta^*}{se(\widehat{\theta})} \qquad \text{or} \qquad t = \frac{\text{estimate} - \text{hypothesized value}}{\text{standard error}}$$

  - we reject at level α if p-value $< \alpha$.

5. DIFFERENCE IN TWO MEANS

- We have two different populations and wish to summarize the differences in their means or test whether they have the same mean
  - e.g. compare mean spending for those with 0% coinsurance (free health care) to those with other insurance.
- We use subscript 1 to denote the first population (e.g. coins0=1) and subscript 2 to denote the second population (e.g. coins0=0).
- We define
  - X to be the random variable of interest (e.g. outpatient spending)
  - μ_1 and μ_2 to be the population mean of X in the first and second populations
  - $\bar{x}_1$ and $\bar{x}_2$ to be the sample averages in the first and second populations
  - s_1 and s_2 to be the standard deviations in the first and second populations
  - $s_{\bar{X}_1}$ and $s_{\bar{X}_2}$ to be the standard errors of $\bar{x}_1$ and $\bar{x}_2$ in the first and second populations.

Example: Difference in two means due to insurance type

- coins0==1 if free health care and coins0==0 otherwise.
- Following gives $\bar{x}_1 = 1692.908$, $s_1 = 2085.393$, $n_1 = 353$, $\bar{x}_2 = 945.637$, $s_1 = 1535.926$, $n_2 = 682$.

```
. * Obtain the two different means
. summarize outspend
```

Variable	Obs	Mean	Std. Dev.	Min	Max
outspend	1,035	1200.503	2085.393	0	28519.19

```
. summarize outspend if coins0==1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
outspend	353	1692.908	2800.32	0	28519.19

```
. summarize outspend if coins0==0
```

Variable	Obs	Mean	Std. Dev.	Min	Max
outspend	682	945.6365	1535.926	0	16002.1

- So $\bar{x}_1 - \bar{x}_2 = 1692.908 - 945.637 = 747.271$.

Formula for difference in two means

- A 95% confidence interval for the difference $\mu_1 - \mu_2$ is
  - estimate $\pm$ critical value $\times$ standard error or

$$(\bar{x}_1 - \bar{x}_2) \pm t_{.025;df} \times se(\bar{x}_1 - \bar{x}_2).$$

- Test of equal means: test $H_0 : \mu_1 - \mu_2 = 0$ against $H_a : \mu_1 - \mu_2 \neq 0$.
  - in general use we use $t = \frac{\text{estimate} - \text{hypothesized values}}{\text{standard error}}$ so

$$t = \frac{\bar{x}_1 - \bar{x}_2}{se(\bar{x}_1 - \bar{x}_2)}$$

- For critical values and p-values we use $t(df)$ distribution.
- Usual case is $\sigma_1 \neq \sigma_2$. Then

$$se(\bar{x}_1 - \bar{x}_2) = \sqrt{s^2_{\bar{x}_1} + s^2_{\bar{x}_2}} = \sqrt{s_1^2/n_1 + s_2^2/n_2}$$

where n_1 and n_1 and n_2 are sample sizes and $df = n_1 + n_2 - 2$ or sometimes Satterthwaite's degrees of freedom.

Example: Differences in two means (continued)

- IMPORTANT: There are different ways to calculate $se(\bar{x}_1 - \bar{x}_2)$.
- We use $se(\bar{x}_1 - \bar{x}_2) = \sqrt{s^2_{\bar{x}_1} + s^2_{\bar{x}_2}}$ and get

$$\begin{aligned} se(\bar{x}_1 - \bar{x}_2) &= \sqrt{(2800.32/\sqrt{353})^2 + (1535.926/\sqrt{682})^2} \\ &= \sqrt{149.046^2 + 58.814^2} = 160.230. \end{aligned}$$

- A 95% confidence interval is then (using $df = n_1 + n_2 - 2 = 1033$)

$$\begin{aligned} & (\bar{x}_1 - \bar{x}_2) \pm t_{.025;1033} \times se(\bar{x}_1 - \bar{x}_2) \\ = \; & (1692.908 - 945.637) \pm 1.9622631 \times 160.230 \\ = \; & 747.271 \pm 314.413 \\ = \; & (433,\ 1062). \end{aligned}$$

- And $t = \dfrac{\bar{x}_1 - \bar{x}_2}{se(\bar{x}_1 - \bar{x}_2)} = \dfrac{1692.908 - 945.637}{160.230} = 4.644 >$ $t_{.025;1033} = 1.962$.
  Reject $H_0 : \mu_1 - \mu_2 = 0$ against $H_a : \mu_1 - \mu_2 \neq 0$ at level .

Command ttest for difference in two means

- Command `ttest` with unequal variances (`ttest, unequal`) yields

```
. ttest outspend, by(coins0) unequal

Two-sample t test with unequal variances
```

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
0	682	945.6365	58.81365	1535.926	830.1587	1061.114
1	353	1692.908	149.046	2800.32	1399.776	1986.041
combined	1,035	1200.503	64.82131	2085.393	1073.307	1327.699
diff		-747.2718	160.2303		-1062.138	-432.4055

```
    diff = mean(0) - mean(1)                                      t =  -4.6637
Ho: diff = 0                     Satterthwaite's degrees of freedom =  464.335

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

- Compared to the earlier manual calculations
  - same estimate, standard errors and t-test statistic
  - slightly different confidence interval
    - ★ as uses Satterthwaite's $df = 464.335$ with invttail(464.3, .025) = 1.9651

6. LINEAR (OLS) REGRESSION

- Best fitting line from scatterplot of y against x
  - example is outpatient spending and age

```
. * Plot of linear regression example
. graph twoway (scatter outspend age) (lfit outspend age, lwidth(thick) lcol(black))
```

Ordinary least squares

- Basic regression is also called ordinary least squares (OLS).
- We are fitting a line, say $\widehat{y} = a + bx$.
- We choose a and b so that y and $\widehat{y}$ are as close to each other in the following sense.
  - The residual is the difference $y - \widehat{y}$
  - We minimize the sum over observations of the squared residuals.
- So $\widehat{y} = a + bx$ where a and b minimize

$$\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2.$$

- It can be shown that

$$\begin{aligned} a &= \bar{y} - b\bar{x} \\ b &= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}. \end{aligned}$$

Stata command regress

- Stata command `regress` gives the OLS estimates and more
  - NOTE: use option vce(robust) throughout.

```
. * OLS regression estimates, standard error and confidence interval
. regress outspend age, vce(robust)

Linear regression                               Number of obs     =      1,035
                                                F(1, 1033)        =      60.67
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0699
                                                Root MSE          =     2012.2
```

outspend	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	34.49203	4.428282	7.79	0.000	25.80257	43.18148
_cons	343.4013	84.1483	4.08	0.000	178.2802	508.5224

- Results include $y = 343.40 + 34.49 \times age$
  - so outpatient spending increases by \$34.49 with each year of aging.

7. STATISTICAL INFERENCE FOLLOWING REGRESSION

- We need to specify a model generating the sample data.
- We assume the true model is linear with

$$E[y|x] = \alpha + \beta x.$$

- This means that on average, in the population, y equals $\alpha + \beta x$.
- We are then interested in inference on the slope parameter β
  - it tells us how much y on average increases as x increases by one unit.
- This has the complication that there are different ways to compute the standard error of the OLS estimates a and b
  - these lead to different confidence intervals and hypothesis tests.
  - we most often use `regress` with option `vce(robust)`.

Computing the standard errors

- The basic model is

$$y_i = \alpha + \beta x_i + u_i$$

  where the error u_i has mean 0.

- 1. Default standard errors: `regress y x`
  - assume that the errors u_i are independent and have the same variance
  - rarely used.
- 2. Heteroskedastic-robust standard errors: `regress y x, vce(robust)`
  - assume errors are independent and have variance that may vary with x_i
  - standard to use this with independent cross-section data.
- 3. Cluster-robust standard errors: `regress y x, vce(cluster id_cluster)`
  - assume that the errors u_i can be grouped (e.g. family, village) where independent across groups but correlated within group.

Confidence interval for beta

- An approximate 95% confidence interval is in general

$$\text{estimate} \pm 1.96 \times \text{standard error}.$$

- An approximate 95% confidence interval for the slope parameter β is

$$b \pm 1.96 \times se(b) = 34.492 \pm 1.96 \times 4.428 = (25.81,\ 43.17).$$

- This is directly given in the regression output as (25.80, 43.18)
  - which uses invttail(1033,.025)= 1.9622631 rather than 1.96.

Test of statistical significance

- A test of statistical significance is a t-test of $H_0 : \beta_2 = 0$ against $H_0 : \beta_2 \neq 0$
  - if $\beta_2 = 0$ then $E[y|x] = \beta_1 + 0 \times x = \beta_2$ does not vary with x.
- In general we use $t = \frac{\text{estimate} - \text{hypothesized values}}{\text{standard error}}$.
- Here
$$t = \frac{b - 0}{se(b)} = \frac{34.492}{4.428} = 7.79.$$
- Stata output gives this and the p-value $= 0.000$
  - we reject H_0 at level 0.05 since $p < 0.05$
  - we conclude that age is statistically significant.

General tests of beta

- We may want to test other hypothesized values of β.
- Example: Test $H_0 : \beta_2 = 40$ against $H_0 : \beta_2 \neq 40$

$$t = \frac{b - 40}{se(b)} = \frac{34.492 - 40}{4.428} = -1.244.$$

  - We do not reject H_0 at level 0.05 as $|t| = 1.407 < 1.96$.
- We can use command `test`
  - it gives $F = 1.55$ (which $= (-1.244)^2$)
  - and $p = 0.2138 > 0.05$ so do not reject H_0.

```
. * Test H0: beta = 40 against Ha: beta not = 40
. quietly regress outspend age, vce(robust)

. test age = 40

 ( 1)  age = 40

       F(  1,  1033) =    1.55
            Prob > F =    0.2138
```

Confidence intervals and tests on means

- There are two identical ways to implement confidence intervals and tests on the mean μ
  - using a statistical command for inference on the mean
    - ★ in Stata this is command `mean`.
  - using a linear regression on only an intercept
    - ★ in Stata this is command `regress`.
- There are two ways to implement confidence intervals and tests on the difference in means $\mu_2 - \mu_1$
  - using a statistical command for difference in two means
    - ★ in Stata this is command `ttest` with option `unequal` (so $\sigma_1 \neq \sigma_2$).
  - using a linear regression on an intercept and binary regressor
    - ★ in Stata this is command `regress` with option `vce(robust)`.
- It is better to use regression (command `regress`) as
  - it allows including additional variables for controls
  - it extends to difference in more than two means (whereas `ttest` does not).

8. SINGLE MEAN (regress on just an intercept)

- Regression of y on an intercept (command `regress y`) is identical to inference on the mean of y (command `mean y`)

```
. * regress y same as mean y
. regress outspend, noheader
```

outspend	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
_cons	1200.503	64.82131	18.52	0.000	1073.307	1327.699

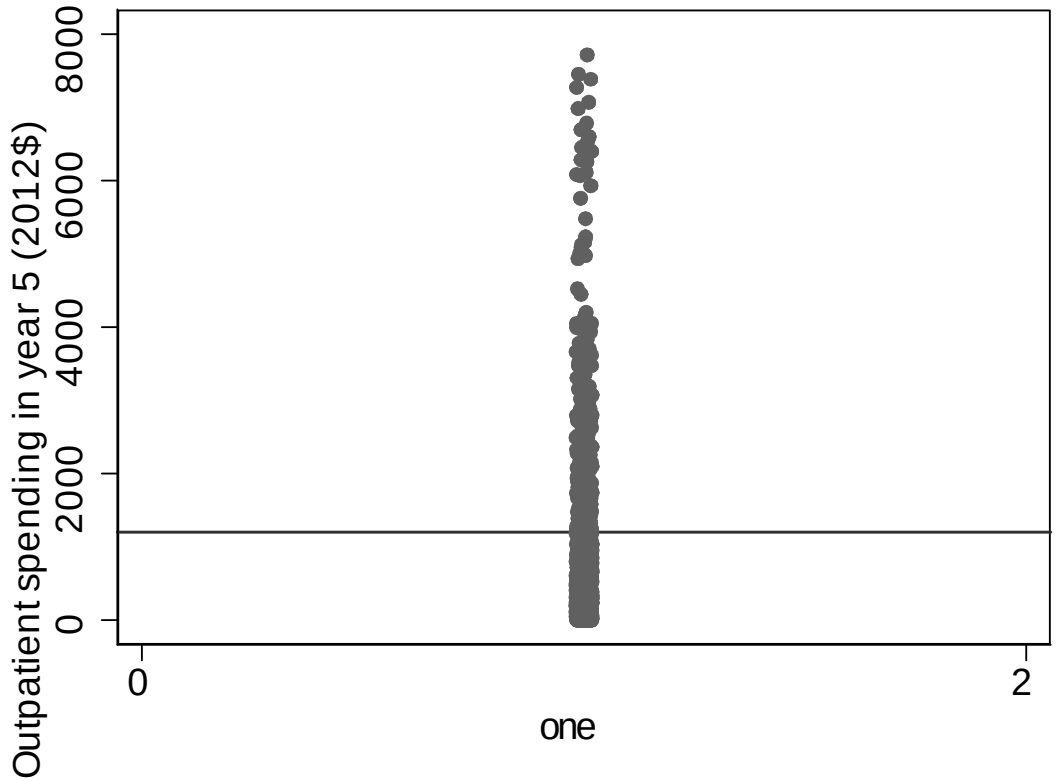
```
. mean outspend

Mean estimation                   Number of obs    =      1,035
```

	Mean	Std. Err.	[95% Conf.	Interval]
outspend	1200.503	64.82131	1073.307	1327.699

- Aside: In this special case only `regress y` and `regress y, vce(robust)` give the same standard error.

Explanation of why regress on just intercept gives the mean

- Intercept only model: $y_i = \alpha + \varepsilon_i$.
- OLS minimizes sum of squared residuals: $\sum_{i=1}^{n}(y_i - \alpha)^2$.
  - $\frac{d}{d\alpha}\sum_{i=1}^{n}(y_i - \alpha)^2 = \sum_{i=1}^{n}\frac{d}{d\alpha}(y_i - \alpha)^2 = \sum_{i=1}^{n} -2(y_i - \alpha)$.
  - Set $\frac{d}{d\alpha} = 0$ gives $\sum_{i=1}^{n} -2(y_i - \alpha) = 0 \Rightarrow \sum_{i=1}^{n}(y_i - \alpha) = 0$ $\Rightarrow \sum_{i=1}^{n} y_i = n\alpha \Rightarrow \widehat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$.



9. Regression with binary regressor

- Define the binary regressor

$$d_i = 0 \text{ or } 1$$

- Then OLS regression on a binary regressor gives fitted value

$$\begin{aligned} \widehat{y}_i &= a + bd_i \\ &= \begin{cases} a+b & \text{if } d_i = 1 \\ a & \text{if } d_i = 0 \end{cases} \end{aligned}$$

- The intercept a is the predicted value when $d = 0$
  - it turns out it equals $\bar{y}_{d=0}$, the sample mean of y when $d = 0$
- The intercept plus slope $a + b$ is the predicted value when $d = 1$
  - it turns out it equals $\bar{y}_{d=1}$, the sample mean of y when $d = 1$
- The slope b is the difference in the predicted value of y
  - it turns out it equals $\bar{y}_{d=1} - \bar{y}_{d=0}$, the difference in the sample means of y.

Regression for difference in two means

- Regression with unequal variances (`regress, vce(robust)`) yields

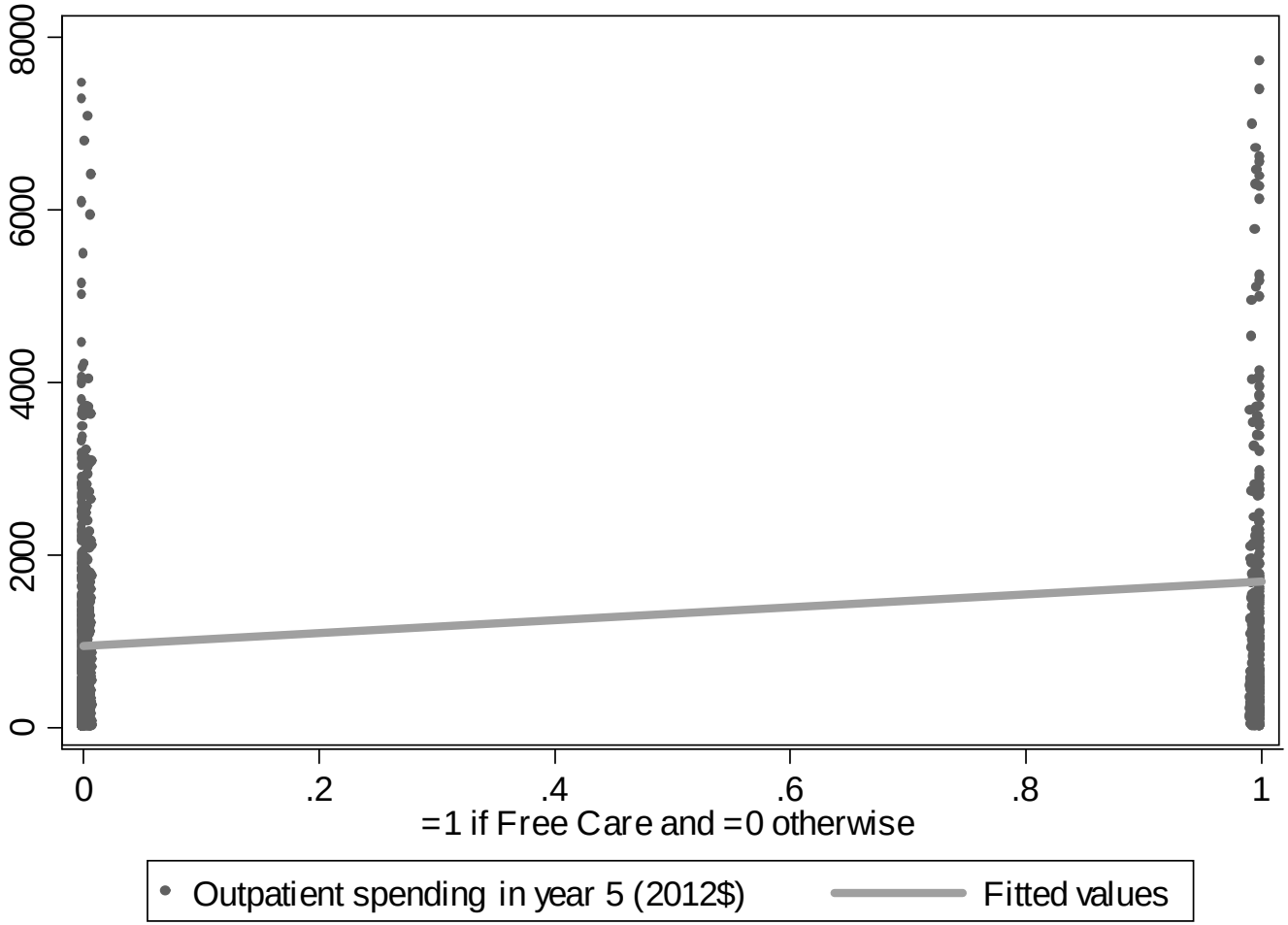
```
. * regress y on binary regressor gives same difference in means
. regress outspend coins0, vce(robust) noheader
```

outspend	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
coins0	747.2718	160.1728	4.67	0.000	432.9707	1061.573
_cons	945.6365	58.82738	16.07	0.000	830.2017	1061.071

- Compared to the earlier manual calculations
  - same difference in means of 747.272
  - slightly different standard error (160.173 versus 160.230)
    - ★ leading to slightly different confidence interval and test statistic.

Regression for difference in two means: graph

- `twoway (scatter outspend coins0) (lfit outspend coins0)`
- Intercept is $\bar{y}_{d=0}$. value at $d = 1$ is $\bar{y}_{d=1}$, slope is $\bar{y}_{d=1} - \bar{y}_{d=0}$.



10. DIFFERENCE IN SEVERAL MEANS (binary regressors)

- There are five mutually exclusive health insurance types.

```
. * Verify that mutually exclusive
. sum coins*
```

Variable	Obs	Mean	Std. Dev.	Min	Max
coins0	1,035	.3410628	.4742956	0	1
coins25	1,035	.2270531	.4191297	0	1
coins50	1,035	.0859903	.2804853	0	1
coins95	1,035	.1768116	.3816936	0	1
coinsindiv	1,035	.1690821	.3750056	0	1

```
. gen total = coins0 + coins25 + coins50+ coins95 + coinsindiv

. sum total
```

Variable	Obs	Mean	Std. Dev.	Min	Max
total	1,035	1	0	1	1

Regression without intercept gives sample means

```
. * regress y on mutually exclusive variables without intercept gives means
. regress outspend coins0 coins25 coins50 coins95 coinsindiv, noconstant vce(robust)

Linear regression                               Number of obs     =      1,035
                                                F(5, 1030)        =      81.57
                                                Prob > F          =     0.0000
                                                R-squared         =     0.2718
                                                Root MSE          =     2057.6

------------------------------------------------------------------------------
             |               Robust
   outspend |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------+----------------------------------------------------------------
      coins0 |   1692.908   149.1955    11.35   0.000     1400.146    1985.67
     coins25 |   1072.902   115.2773     9.31   0.000     846.6965    1299.107
     coins50 |   874.5381   117.5911     7.44   0.000     643.7926    1105.284
     coins95 |   852.5867   118.2486     7.21   0.000      620.551    1084.622
   coinsindiv |   908.1999   98.40113     9.23   0.000     715.1103    1101.289
------------------------------------------------------------------------------

. sum outspend if coins50==1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
   outspend |         89    874.5381     1112.94         0    7399.413
```

Regression with intercept gives difference in means

- We need to omit a category
  - The difference is with respect to the omitted category, here `coins0`.

```
. * regress y on several mutually exclusive variables
. regress outspend coins25 coins50 coins95 coinsindiv, vce(robust)

Linear regression                               Number of obs     =      1,035
                                                F(4, 1030)        =       6.34
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0302
                                                Root MSE          =     2057.6
```

outspend	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
coins25	-620.0067	188.5422	-3.29	0.001	-989.9774	-250.036
coins50	-818.3703	189.9657	-4.31	0.000	-1191.134	-445.6063
coins95	-840.3216	190.3734	-4.41	0.000	-1213.886	-466.7576
coinsindiv	-784.7085	178.7235	-4.39	0.000	-1135.412	-434.0048
_cons	1692.908	149.1955	11.35	0.000	1400.146	1985.67

Test joint statistical significance of insurance type

- They are jointly statistically significant at level 0.05 since $p = 0.0000 < 0.05$.
- We get exactly the same test statistic value regardless of which category was omitted
  - e.g. `regress outspend coins0 coins25 coins50 coins95, vce(robust)`
  - `test coins0 coins25 coins50 coins95`

```
. * test joint significance
. quietly regress outspend coins25 coins50 coins95 coinsindiv, vce(robust)

. test coins25 coins50 coins95 coinsindiv

 ( 1)  coins25 = 0
 ( 2)  coins50 = 0
 ( 3)  coins95 = 0
 ( 4)  coinsindiv = 0

       F(  4,  1030) =    6.34
            Prob > F =    0.0000
```

Can add additional regressors

- A big advantage of regression is that can add regressors
  - - here add age and badhealth as regressors and test for joint

```
. * add regressors regress y on several mutually exclusive variables
. regress outspend coins25 coins50 coins95 coinsindiv ///
>    age badhealth, vce(robust) noheader
```

outspend	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
coins25	-593.9946	178.3616	-3.33	0.001	-943.989	-244.0001
coins50	-816.2792	185.3261	-4.40	0.000	-1179.94	-452.6185
coins95	-826.4517	183.4846	-4.50	0.000	-1186.499	-466.4046
coinsindiv	-833.9904	175.8659	-4.74	0.000	-1179.088	-488.8932
age	33.45397	4.217604	7.93	0.000	25.17788	41.73007
badhealth	378.1198	327.8335	1.15	0.249	-265.1795	1021.419
_cons	830.7078	128.6758	6.46	0.000	578.2105	1083.205

```
. test coins25 coins50 coins95 coinsindiv

 ( 1)  coins25 = 0
 ( 2)  coins50 = 0
 ( 3)  coins95 = 0
 ( 4)  coinsindiv = 0

       F(  4,  1028) =    7.01
            Prob > F =    0.0000
```

11. DIFFERENCES IN DIFFERENCES: Motivation

- Regression of y on x gives only correlation.
- Unless x is experimentally assigned (as in the Rand experiment) we cannot say that an association between y and x is not due to the reverse (y actually causes x) or due to some third variable (such as health status in the case of health care service use and health insurance type).
- Economists cannot usually run experiments.
- Instead use various "quasi-experimental" methods, some of which are covered in ECN 140 (econometrics)
  - instrumental variables
  - propensity score matching
  - regression discontinuity design
  - differences in differences for a natural experiment.
- Each method gives a causal estimate, though a causal interpretation relies on assumption(s) that vary with the specific method.

Treatment effect of a natural experiment

- Consider a "natural" experiment where a policy (called a treatment) comes into being that effects one group more than another.
- Let y denote the outcome and S denote the treatment
  - with $S = 1$ if treated and $S = 0$ if not treated.
- **1. Method 1: Treatment-control comparison** (at a point in time)
  - Treatment effect = ($\bar{y}$ for treated) - ($\bar{y}$ for not treated)= $\bar{y}_{S=1} - \bar{y}_{S=0}$.
  - Problem: This is misleading if the treated and untreated groups differ in their characteristics
    - ★ e.g. if the policy was targeted towards poor people.
- **2. Method 2: Before-after comparison** over time for treated only
  - Treatment effect = ($\bar{y}$ for treated after treatment) - ($\bar{y}$ for treated before treatment)
  - Problem: Misleading if other things also effect the treated over time.
- **3. Differences-in-differences** combines methods 1. and 2.
  - it uses change over time for the untreated to control for nontreatment changes over time.

Differences in differences example

- Example where have data on treated and nontreated before and after the time of treatment
  - **pre: before treatment**
    - ★ the treated had average 10 and the nontreated had 15
    - ★ treatment was targeted to those with less on average
  - **post: after treatment**
    - ★ the treated had average 14 and the nontreated had 16
- Differences-in-difference estimate
  - before-after comparison for treated $= 14 - 10 = 4$
  - before-after comparison for untreated $= 16 - 15 = 1$
  - the treated improved over time by 4 while our estimate of background changes is the untreated change of 1
  - differences-in-difference estimate of the treatment effect is $4 - 1 = 3$.

Differences in differences formula

- Introduce time before (pre) and after (post) the policy comes into effect
  - $T = 0$ is a time period before and $T = 1$ is a time period after.
- Then the difference in difference estimate of the effect of treatment is
  - DinD= $\Delta\bar{y}$ for those treated $-$ $\Delta\bar{y}$ for those not treated
  - or DinD= $(\bar{y}_{S=1,\text{post}} - \bar{y}_{S=1,\text{pre}}) - (\bar{y}_{S=0,\text{post}} - \bar{y}_{S=0,\text{pre}})$.
- Equivalently we can use
  - DinD= $(\bar{y}_{S=1,\text{post}} - \bar{y}_{S=0,\text{post}}) - (\bar{y}_{S=1,\text{pre}} - \bar{y}_{S=0,\text{pre}})$
  - the post-period difference in the two groups less that in the pre-period.
- DinD can be estimated by computing the four separate means and then computing the differences.
- DID gives a causal estimate of treatment
  - under the assumption that in the absence of treatment the change over time would be the same for the treated and untreated groups
  - called the parallel trends assumption.

Differences in differences manual computation

- The following table illustrates manual computation.

	Treated $(S=1)$	Not Treated $(S=0)$	Difference over treatment
Pre $(T=0)$	A	C	$(A-C)$
Post $(T=1)$	B	D	$(B-D)$
Change over time	$(B-A)$	$(D-C)$	

$$(B-D)-(A-C)$$

or equivalently

$$(B-A)-(D-C)$$

Regression computation

- The same difference-in-difference estimate can be obtained as the coefficient of $T \times S$ in the OLS regression

$$y_i = \beta_1 + \beta_2 T_i + \beta_3 S_i + \beta_4 T_i \times S_i + \text{error}.$$

  - where $T_i = 1$ in the post-period and $T_i = 0$ in the pre-period
  - and $S_i = 1$ if treated and $S_i = 0$ if not treated
  - $T_i \times S_i = 1$ if treated and in the post-period and $= 0$ otherwise.

- The model implies that y equals the following

	Treated ($S = 1$)	Not Treated ($S = 0$)	Difference over treatment
Pre ($T = 0$)	$\beta_1 + \beta_3$	β_1	β_3
Post ($T = 1$)	$\beta_1 + \beta_2 + \beta_3 + \beta_4$	$\beta_1 + \beta_2$	$\beta_3 + \beta_4$
Change over time	$\beta_2 + \beta_4$	β_2	

Diff in diff = β_4!

Differences in differences regression computation

- So suppose we have data on each individual, not just the means.
- The OLS regression is

$$y_i = \beta_1 + \beta_2 T_i + \beta_3 S_i + \beta_4 T_i \times S_i + \text{error}.$$

- This is often written as

$$y_i = \beta_1 + \beta_2 Post_i + \beta_3 Treat_i + \beta_4 Post_i \times Treat_i + \text{error}.$$

- The difference-in-differences estimate is β_4.
- The advantage of using an OLS regression are
  - 1. A t-test $H_0 : \beta_4 = 0$ is a test of statistical significance of the treatment
  - 2. We can add control variables as additional regressors.
  - 3. We can compute robust standard errors of $\widehat{\beta}_4$.
- A data example will be given in a class assignment.

12. REGRESSION WITH LOGS

- The elasticity of y given a change in x is
  - the proportionate change in y divided by the proportionate change in x
    - ★ Elasticity$= \frac{\Delta y / y}{\Delta x / x}$.
- Note that $d \ln y / dy = 1/y$ so $d \ln y = dy/y$
  - hence $\frac{d \ln y}{d \ln x} = \frac{dy/y}{dx/x}$ = elasticity.
- So elasticity can be estimated by OLS in the regression model in logs
  - $\ln y = \beta_1 + \beta_2 \ln x +$ error
  - since $d \ln y / d \ln x = \beta_2$
- A semi-elasticity of y given a change in x is
  - the proportionate change in y divided by the change in x
    - ★ Semi-elasticity$= \frac{\Delta y / y}{\Delta x}$.
- It can be estimated by OLS in the log-linear regression model
  - $\ln y = \beta_1 + \beta_2 x +$ error.

Semi-elasticity example: Log spending on age

- Because outspend=0 for 13% of sample use lny=ln(y+1)
  - semi-elasticity: a one year increase in age is associated with a 3.267% increase in outpatient spending

```
. * Generate a new variable allowing for outspend == 0
. gen lnout = ln(outspend + 1)

. * Log-linear gives semielasticity
. regress lnout age, vce(robust) noheader
```

lnout	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0326705	.0043515	7.51	0.000	.0241318	.0412093
_cons	4.829998	.1336079	36.15	0.000	4.567824	5.092172

Elasticity example: Log spending on log age

- Elasticity: A 1% increase in age is associated with a 0.503% increase in outpatient spending.

```
. * Log-log gives elasticity
. gen lnage = ln(age)
(10 missing values generated)

. regress lnout lnage, vce(robust) noheader
```

lnout	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnage	.5026852	.0800434	6.28	0.000	.3456171	.6597533
_cons	4.17108	.2442859	17.07	0.000	3.691722	4.650439

12. STATA CODE USED IN THESE NOTES

```
*** Summary statistics
* The complete dataset
use tr132statistics.dta, clear
describe
* Summary statistics for outpatient spending
summarize outspend
* Histogram if outspend > 0
set scheme s1mono
hist outspend if outspend > 0
graph export histogram.wmf, replace
* Smoothed histogram (kernel density estimate) if > 0
kdensity outspend if outspend > 0
graph export kdensity.wmf, replace
```

```
*** Statistical Inference on Population Mean
* Standard error and 95% confidence interval for mu
mean outspend
* Hypothesis test that population mean outspend = 1000
quietly summarize outspend
test outspend = 1000
```

```
*** Linear Regression
* Plot of linear regression example
graph twoway (scatter outspend age) ///
    (lfit outspend age, lwidth(thick) lcol(black))
graph export regression.wmf, replace
* OLS regression estimates, standard error & conf.  int.
regress outspend age, vce(robust)
*** Test on the slope parameter
* Test H0:  beta = 40 against Ha:  beta not = 40
quietly regress outspend age, vce(robust)
test age = 40
*** Regression on an intercept same as inference on mean
* regress y same as mean y
regress outspend, vce(robust) noheader
mean outspend
```

```
*** Regression on a binary variable is diff in two means
* regress y on binary regressor gives difference in means
summarize outspend if coins0==1
summarize outspend if coins0==0
* ttest gives difference in means (use option unequal)
ttest outspend, by(coins0) unequal
* regress gives difference in means (use vce(robust))
regress outspend coins0, vce(robust) noheader
```

```
*** Regression on several mutually exclusive variables
* Verify that mutually exclusive
sum coins*
gen total = coins0+coins25+coins50+coins95+coinsindiv
sum total
* regress y on mutually exclusive variables
* without intercept gives means
regress outspend coins0 coins25 coins50 coins95 ///
    coinsindiv, noconstant vce(robust)
sum outspend if coins50==1
* regress y on several mutually exclusive variables
* (omit one category)
regress outspend coins25 coins50 coins95 coinsindiv, ///
    vce(robust)
* test joint significance
quietly regress outspend coins25 coins50 coins95 ///
    coinsindiv, vce(robust)
test coins25 coins50 coins95 coinsindiv
```

```
*** Elasticity
* Generate a new variable allowing for outspend == 0
gen lnout = ln(outspend + 1)
* Log-linear gives semielasticity
regress lnout age, vce(robust) noheader
* Log-log gives elasticity
gen lnage = ln(age)
regress lnout lnage, vce(robust) noheader
```