Analysis of Economics Data
Chapter 6: The Least Squares Estimator

© A. Colin Cameron
Univ. of Calif. Davis

November 2022

CHAPTER 6: The Least Squares Estimator

- The **sample** leads to a fitted regression line $\widehat{y} = b_1 + b_2 x$.
  - But different samples will lead to different fitted regression lines.
  - Example: in a random sample individual earnings increase by 7% with an extra year of schooling
    - what can we say about the increase in the entire population?
- We suppose that there is an unknown **population line** $\beta_1 + \beta_2 x$
  - then the regression slope b_2 is an estimate of β_2
- This chapter
  - **distribution of the regression estimates** b_1 and b_2.
- The subsequent chapter
  - confidence intervals and hypothesis tests for the slope parameter β_2.

- Key regression output for statistical inference:

Variable	Coefficient	Standard Error	t statistic	p value	95% conf. interval	
Size	73.77	11.17	6.60	0.000	50.84	96.70
Intercept	115017.30	21489.36	5.35	0.000	70924.76	159109.8

- The standard error of `Size` is an estimate of the precision of b_2 as an estimate of β_2
  - we need to explain how this is obtained
  - different assumptions lead to different standard errors
  - so important to go into details.
- The remaining statistics are studied in Chapter 7
  - the confidence interval for `Size` is one for β_2.
  - the t statistic for `Size` is a test of $H_0 : \beta_2 = 0$ against $H_a : \beta_2 \neq 0$
    - ★ i.e. is there any relationship between `Size` and `Price`?

Outline

1. Population and Sample
2. Examples of Sampling from a Population
3. Properties of the Least Squares Estimator
4. Estimators of Model Parameters

Datasets: GENERATEDDATA, GENERATEDREGRESSIONS, CENSUSREGRESSIONS

6.1 Population Model: Conditional Mean of y given x

- The **sample model** is a line $b_1 + b_2 x$.
- So we assume that the **population model** is also a line, denoted $\beta_1 + \beta_2 x$
  - where β is "beta" and we use Greek letters for (unknown) parameters.
- More formally the **conditional mean** of y is assumed to be linear in x

$$\mathrm{E}[Y|X = x] = \beta_1 + \beta_2 x.$$

- The **population conditional mean** of Y given $X = x$
  - is the probability-weighted average of all possible values of Y for a given value of x; e.g. earnings conditional on years of schooling
  - is denoted $\mathrm{E}[Y|X = x]$
  - generalizes $\mathrm{E}[Y]$ in chapter 3 that is the probability-weighted average of all possible values of Y.

Population Conditional Mean (continued)

- We assume that the conditional mean is linear in x

$$\mathrm{E}[Y|X=x]=\beta_1+\beta_2 x.$$

- Commonly-used simpler notation is

$$\mathrm{E}[y|x]=\beta_1+\beta_2 x.$$

- Note: In general the conditional mean need not be linear.
  - Case 1: $\mathrm{E}[Y|X=1]=5$, $\mathrm{E}[Y|X=2]=7$, $\mathrm{E}[Y|X=3]=9$
    - ★ linear since this implies $\mathrm{E}[Y|X=x]=3+2x$.
  - Case 2: $\mathrm{E}[Y|X=1]=5$, $\mathrm{E}[Y|X=2]=7$, $\mathrm{E}[Y|X=3]=12$
    - ★ nonlinear as increase by 2 from $X=1$ to $X=2$ but increases by 5 from $X=2$ to $X=3$.
  - In Chapter 9 we consider nonlinear conditional means.

Error Term

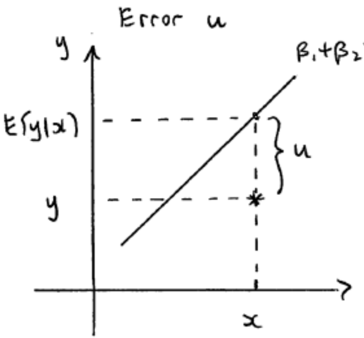
- y does not exactly equal $\beta_1 + \beta_2 x$
  - instead $\mathrm{E}[y|x] = \beta_1 + \beta_2 x$.
- The difference between y and $\mathrm{E}[y|x]$ is called the **error term** u

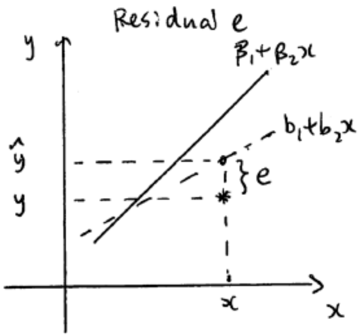
$$\begin{aligned} u &= y - \mathrm{E}[y|x] \\ &= y - (\beta_1 + \beta_2 x). \end{aligned}$$

- The error term u is **not observed as** β_1 and β_2 are unknown.

Error Term versus Residual - a crucial distinction

- u is **not observed** - it is the difference between y and the unknown population line $\beta_1 + \beta_2 x$ (the solid line)
- e is **observed** - it is the difference between y and the known fitted line $b_2 + b_2 x$ (the dashed line)





Error Term is assumed to have mean zero

- Since $u = y - (\beta_1 + \beta_2 x)$ we have

$$y = \beta_1 + \beta_2 x + u.$$

- The **error term is assumed to be zero on average for each** x **value**
  - sometimes $u_i > 0$ and so y_i is above the population line
  - sometimes $u_i < 0$ and so y_i is below the population line
  - but the long-run average of u_i (at each value of x) is zero.
- More precisely the **error term has conditional mean zero**

$$\mathrm{E}[u|x] = 0.$$

- This ensures that the population line is indeed $\beta_1 + \beta_2 x$.

$$\begin{aligned} \mathrm{E}[y|x] &= \mathrm{E}[\beta_1 + \beta_2 x + u|x] \\ &= \beta_1 + \beta_2 x + \mathrm{E}[u|x] \\ &= \beta_1 + \beta_2 x \qquad \text{if } \mathrm{E}[u|x] = 0. \end{aligned}$$

Population Conditional Variance of y given x

- The variability of the error term around the line will determine in part the precision of our estimates
  - greater variability is greater noise so less precision.
- We initially assume that the **error variance is constant** and does not vary with x

$$\text{Var}[u|x] = \sigma_u^2.$$

- This is called the assumption of **homoskedastic** errors
  - "skedastic" based on the Greek word for scattering
  - "homos" is the Greek word for same
  - this assumption can be relaxed (and is often relaxed - later).
- The error term provides the only variation in y around the population line so then

$$\text{Var}[y|x] = \text{Var}[u|x] = \sigma_u^2.$$

Summary

- The bottom line:
  - Univariate analysis: $y_1, ..., y_n$ is a simple random sample with
$$Y_i \sim (\mu, \sigma^2).$$
  - Regression analysis: $(x_1, y_1), ..., (x_n, y_n)$ is a simple random sample that allows the mean to vary with x, so
$$y_i | x_i \sim (\beta_1 + \beta_2 x, \sigma_u^2).$$

6.2 Examples of Sampling from a Population

- We consider two examples of sampling from a population
  - regression generalizations of the two examples in chapter 4.
- 1. Generate by computer 400 samples from an explicit model $y = \beta_1 + \beta_2 x + u$.
- 2. Select 400 samples from a finite population - the U.S. 1880 Census for males aged 60-69 years.
- In both cases we run 400 regressions giving 400 estimates b_1 and b_2 and find
  - the average of the 400 slopes b_2 is close to β_2
  - the distribution of the 400 slopes b_2 is approximately normal
  - similar results hold for the intercept b_1.

Single Sample Generated from an Experiment

- Example with $n = 5$ is generate data from

$$\begin{aligned} y &= \beta_1 + \beta_2 x + u = 1 + 2x + u \\ u &\sim N(0, \sigma_u^2 = 4) \\ x &= 1, 2, 3, 4, 5. \end{aligned}$$

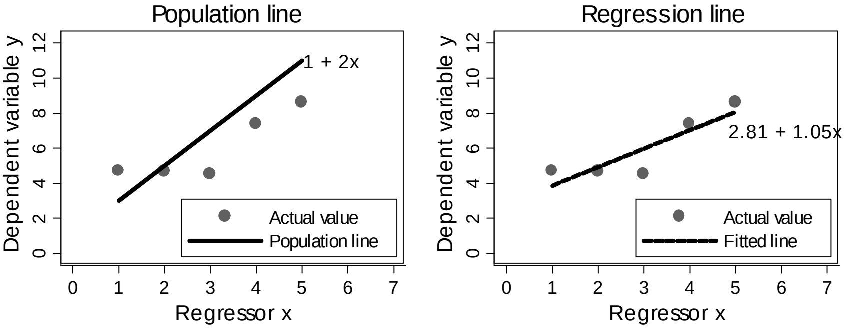
  - note: added the assumption that errors are normally distributed

- Then a random normal generator for u yielded

Observation	x	E[y|x]=1+2x	u	y=1+2x+u
1	1	1+2×1=3	1.689889	4.689889
2	2	1+2×2=5	-.3187171	4.681283
3	3	1+2×3=7	-2.506667	4.493333
4	4	1+2×4=9	-1.63328	7.366720
5	5	1+2×5=11	-2.390764	8.609236

- Five generated observations
  - left panel: population regression line $y = \beta_1 + \beta_2 x = 1 + 2x$
  - right panel: sample regression line $\widehat{y} = b_1 + b_2 x = 2.81 + 1.05x$

  - note that $b_1 \neq \beta_1$ and $b_2 \neq \beta_2$.



Many Samples Generated from an Experiment

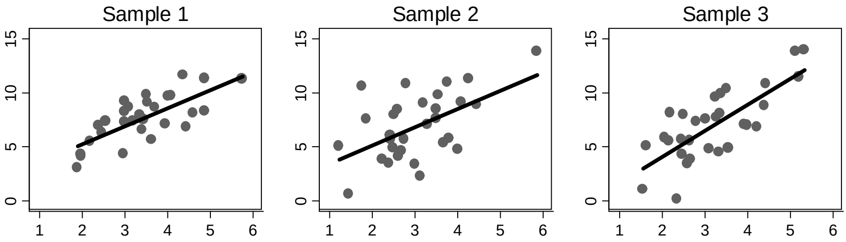
- Samples of size 30 from

$$\begin{aligned} y &= \beta_1 + \beta_2 x + u = 1 + 2x + u \\ u &\sim N(0, \sigma_u^2 = 4) \\ x &\sim N(0, 1). \end{aligned}$$

- This is the same model for y as above
  - except now regressors are draws from a standard normal distribution
  - and $n = 30$.
- Next slide gives results from three samples.

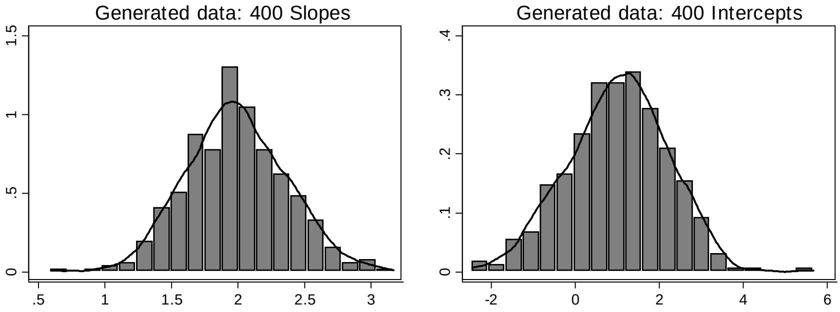
Three Generated Samples yield three different lines

- Scatterplots and regression lines from three samples of size 30
  - intercepts and slopes vary across samples.



400 Generated Samples of Size 30

- 400 such samples were generated and fitted
  - left panel: $\beta_2 = 2$ and average of 400 slopes equals 1.979.
  - right panel: $\beta_1 = 1$ and average of 400 intercepts equals 1.039.
  - both histograms are approximately normal.



Many Samples Generated from a Finite Population

- Data from the 1880 Census
  - complete enumeration of the U.S. population in 1880.
- Relationship between
  - y = labforce = labor force participation
    - ★ 1 if in the labor force; 0 if not in the labor force
  - and x = age = 60 to 70 years.
- Population is of size 1,058,475 (men aged 60-70 years)
- Population mean of labforce is 0.8945
  - so 89.45% were in the labor force.

Population Regression Line

- Population regression line is

$$labforce = \beta_1 + \beta_2 \times age$$

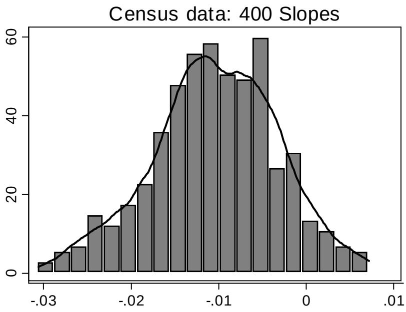
- Population regression line based on 1,058,475 observations is

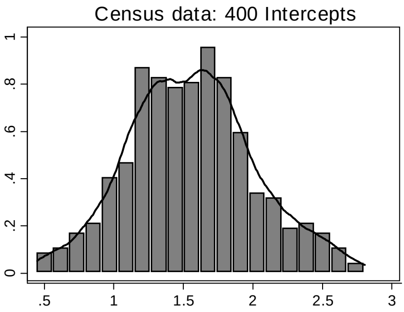
$$labforce = 1.593 - 0.0109 \times age$$

  - so $\beta_1 = 1.593$ and $\beta_2 = -0.0109$
  - with each extra year the probability of being in the labor force falls by 0.0109 or by 1.09 percentage points.

400 Samples of Size 200

- Draw 400 samples of size 200; regress labforce on age in each sample
  - large sample sizes as regression fit is poor: $R^2 \simeq 0.01$.
  - left panel: $\beta_2 = -0.0109$ and average of 400 slopes is -0.0115
  - right panel: $\beta_1 = 1.593$ and average of 400 intercepts is 1.636
  - both histograms are approximately normal.





6.3 Properties of the Least Squares Estimator

- **Slope estimate** is a random variable

$$b_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

  - different samples have different data and hence different $b_2's$.
- We want to find $E[b_2]$, $\text{Var}[b_2]$ and a distribution for inference.
- If we assume the model is $y_i = \beta_1 + \beta_2 x_i + u_i$ then some algebra leads to the re-expression of the formula for b_2 as

$$b_2 = \beta_2 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

- Conditioning on the regressors x_i, the only source of randomness is the errors u_i.
- It follows that $E[b_2]$ and $\text{Var}[b_2]$ **depend crucially on assumptions about the error** u_i.

Data Assumptions

- Always assume that **there is variation in the regressors**
  - we rule out the case $x_i = \bar{x}$ for all i
  - this ensures $\sum_{i=1}^{n}(x_i - \bar{x})^2 > 0$.
- Otherwise cannot compute b_1 and b_2.
- Also at least 3 observations.

Population Assumptions

- Standard assumptions are that:
  - **1.** The **population model** is $y_i = \beta_1 + \beta_2 x_i + u_i$ for all i.
  - **2.** The **error for the** i^{th} **observation has mean zero conditional on x**: $\mathrm{E}[u_i|x_i] = 0$ for all i.
  - **3.** The **error for the** i^{th} **observation has constant variance conditional on x**: $\mathrm{Var}[u_i|x_i] = \sigma_u^2$ for all i.
  - **4.** The **errors for different observations are statistically independent**: u_i is independent of u_j for all $i \neq j$.
- Assumptions 1-2 are the crucial assumptions that ensure

$$\mathrm{E}[y_i|x_i] = \beta_1 + \beta_2 x_i.$$

- Assumption 3 is called conditionally homoskedastic errors

Mean and Variance of the OLS Slope Coefficient

- Given assumptions 1-2 ($y = \beta_1 + \beta_2 x + u$ and $\mathrm{E}[u|x] = 0$)

$$\mathrm{E}[b_2] = \beta_2.$$

- Given assumptions 1-4 (add $\mathrm{V}[u|x] = \sigma_u^2$ and independent errors)

$$\sigma_{b_2}^2 = \mathrm{Var}[b_2] = \frac{\sigma_u^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

- These results are proved in Appendix C.1
    - in the simpler case of a model without intercept.

Estimate of the Error Variance

- $\sigma^2_{b_2} = \text{Var}[b_2]$ depends in part on σ^2_u which is unknown.
- So estimation of $\text{Var}[b_2]$ requires an estimate of σ^2_u.
- Estimate variance of the error σ^2_u by the sample variance of the residuals

$$s^2_e = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$$

- We use $1/(n-2)$ as this guarantees s^2_e is unbiased for σ^2_u.
  - the "intuition" is that $\widehat{y} = b_1 + b_2 x$ is based on two estimated coefficients leaving $(n-2)$ degrees of freedom.
- The **standard error of the regression** or the **root mean squared error** takes the square root to give an estimate of σ_u

$$s_e = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}$$

Estimate of the Variance of the OLS Slope Coefficient

- Under assumptions 1-4

$$\mathrm{Var}[b_2] = \sigma^2_{b_2} = \frac{\sigma_u^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

- Replace σ_u^2 with estimate $s_e^2 = \frac{1}{n-2}\sum_{i=1}^n (y_i - \widehat{y}_i)^2$.
- The estimated variance of b_2 is then

$$\frac{s_e^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

- Taking the square root, **the standard error of** b_2 is

$$\begin{aligned} se(b_2) &= \sqrt{\frac{s_e^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \\ &= \frac{s_e}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}. \end{aligned}$$

Example: Computation of the Standard Error

- Artificial data on a sample of size five
  - (y, x) equals $(1,1)$, $(2,2)$, $(2,3)$, $(2,4)$ and $(3,5)$.
  - From chapter 5: $\widehat{y} = 0.8 + 0.4x$.
  - so $\widehat{y}_1 = 1.2$, $\widehat{y}_2 = 1.6$, $\widehat{y}_3 = 2.0$, $\widehat{y}_4 = 2.4$, $\widehat{y}_5 = 2.8$.
- Standard error of the regression $s_e = \sqrt{.1333333} = 0.365148$ since

$$\begin{aligned} s_e^2 &= \tfrac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 \\ &= \frac{1}{3}\{(1-1.2)^2+(2-1.6)^2+(2-2)^2+(2-2.4)^2+(3-2.8)^2\} \\ &= 0.13333. \end{aligned}$$

- $\sum_{i=1}^{n}(x_i - \bar{x})^2 = 10$ calculated earlier in computing b_2. So

$$se(b_2)^2 = \frac{s_e^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{0.133333}{10} = 0.0133333.$$

- Standard error of the slope b_2 is $se(b_2) = \sqrt{0.013333} = 0.115$.

When is the Slope Coefficient Precisely Estimated?

- The **standard error** of b_2 is $se(b_2) = \sqrt{\frac{s_e^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$.
- **Better precision** = smaller standard error occurs if
  - **1.** Model fits well (s_e^2 is smaller)
  - **2.** Many observations (then $\sum_{i=1}^n (x_i - \bar{x})^2$ is larger).
  - **3.** Regressors are widely scattered (then $\sum_{i=1}^n (x_i - \bar{x})^2$ is larger).

Normal Distribution and the Central Limit Theorem

- Under assumptions 1-4

$$b_2 \sim (\beta_2, \sigma^2_{b_2}).$$

- The standardized variable

$$\begin{aligned} Z &= \frac{b_2 - \beta_2}{\sigma_{b_2}} \\ &\sim (0,1) \text{ by construction} \\ &\sim N(0,1) \text{ as } n \rightarrow \infty \text{ if a central limit theorem holds.} \end{aligned}$$

- In practice, σ_{b_2} is unknown as error standard deviation σ_u is unknown
  - this will lead to use of the T distribution in chapter 7.

Aside: The OLS Intercept Coefficient

- Under assumptions 1-2

$$\mathrm{E}[b_1] = \beta_1.$$

- Given assumptions 1-4

$$\sigma^2_{b_1} = \mathrm{Var}[b_1] = \frac{\sigma^2_u \sum_{i=1}^n x_i{}^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}.$$

- The standard error of b_2 is

$$se(b_2) = \sqrt{\frac{s^2_e \sum_{i=1}^n x_i{}^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}}.$$

- And $Z = (b_2 - \beta_2)/\sigma_{b_2}$ is $N(0, 1)$ as $n \rightarrow \infty$.

Summary for the OLS Slope Coefficient

A summary given assumptions 1-4 is the following.

1. y_i given x_i has conditional mean $\beta_1 + \beta_2 x_i$ and conditional variance σ_u^2.
2. Slope coefficient b_2 has mean β_2 and variance $\sigma_{b_2}^2 = \sigma_u^2 / \sum_{i=1}^n (x_i - \bar{x})^2$.
3. Standard error of b_2 is s_{b_2} where $se(b_2)^2 = s_e^2 / \sum_{i=1}^n (x_i - \bar{x})^2$ and $s_e^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$.
4. $Z = (b_2 - \beta_2)/\sigma_{b_2}$ has mean 0 and variance 1.
5. As sample size $n \rightarrow \infty$, Z is standard normal distributed by the central limit theorem.

Least Squares in Practice

- **Assumptions 1-2 are essential** for least squares to be unbiased and consistent.
  - in particular assumption 2 rules out any correlation between x and u
    - ★ e.g. rules out high x being associated with high u
  - we maintain these assumptions
  - chapter 16 discusses failures
  - chapter 17 has some possible solutions.
- Assumptions 3-4 can be relaxed
  - **a crucial practical part of regression is choosing the correct variation of assumptions 3 and 4**
  - this is necessary to get correct standard errors
    - ★ and hence correct confidence intervals and hypothesis tests
  - chapters 7.7 and 12.1 provide methods.

6.4 Estimators of Model Parameters

- Ideal properties of estimators were presented in Chapter 3.6 for estimation of the population mean.
- For centering
  - unbiasedness (on average)
  - consistency (almost perfect in infinitely large samples).
- For being best
  - minimum variance among all possible correctly-centered estimators.
- Bottom line: Under assumptions 1-4 OLS is essentially the best estimator of β_1 and β_2.

Unbiased Estimator

- Given assumptions 1-2

$$E[b_2|x_1, ..., x_n] = \beta_2.$$

  - b_2 is **unbiased** for β_2 (and b_1 is unbiased for β_1)
  - if we obtain many samples yielding many b_2 then on average $b_2 = \beta_2$.
- Essentially we need sampling such that $E[y_i|x_i] = \beta_1 + \beta_2 x_i$.

Consistent Estimator

- A sufficient condition for a consistent estimator is that as $n \longrightarrow \infty$
  - any bias disappears and the variance goes to zero.
- So b_2 is **consistent** for β_2 as
  - b_2 is unbiased for β_2 given assumptions 1-2
  - $\mathrm{Var}[b_2] \longrightarrow 0$ as $n \longrightarrow \infty$ given assumptions 1-4
    - note: assumptions 3-4 can be relaxed and still get consistency.

Minimum Variance Estimator

- We want as precise an estimator as possible.
- OLS is the **best linear unbiased estimator** (**BLUE**) of β_2 under assumptions 1-4
  - lowest variance of all unbiased estimators that are a linear combination of the $y's$
    - recall $b_2 = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) / \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} w_i y_i$
    - so linear in y_i.
- OLS is the **best unbiased estimator** (**BUE**) of β_2 if additionally u is normally distributed
  - so lowest variance of all unbiased estimators.
- OLS is the **best consistent estimator** (**BUE**) in standard settings under assumptions 1-4,
  - it has smallest variance among consistent estimators.

Key Stata Commands

```
* Generated data
clear
set obs 5
set rng kiss32 // uses old Stata random number generator
generate x = _n // set x to equal the observation number
generate Eygivenx = 1 + 2*x
set seed 123456
generate u = rnormal(0,2)
generate y = Eygivenx + u
list
regress y x
twoway (scatter y x) (lfit y x)
twoway (scatter y x) (lfit ytrue x)
```

Some in-class Exercises

1. Suppose we know that $y = 8 + 5x + u$ where $E[u|x] = 0$. Give the conditional mean of y given x and the error term for the observation $(x, y) = (5, 30)$.
2. OLS regression of y on x on a large sample leads to slope coefficient equal to 10 with standard error 4. Provide an approximate 95% confidence interval for β_2 in the model $y = \beta_1 + \beta_2 x + u$.
3. OLS regression of y on x on a large sample leads to slope coefficient equal to 20 with standard error 5. Test at level 0.05 the claim that the population slope coefficient equals 8.
4. You are given the following $\sum_{i=1}^{27}(x_i - \bar{x})^2 = 20$ and $\sum_{i=1}^{27}(y_i - \widehat{y}_i)^2 = 400$. Compute the standard error of the OLS slope coefficient under assumptions 1-4.
5. Which of assumptions 1-4 ensure that OLS estimates are unbiased?