Univariate Regression

(Part 3 - Sampling Distribution of the OLS Estimator)

Dragos Ailoae

Advanced Economics and Business Statistics
ECON-4400w

Brooklyn College

Today

1. Probability framework for linear regression
2. The ordinary least squares (OLS) estimator and the sample regression line
3. Measures of fit of the sample regression
4. **The least squares model assumptions**
5. **The sampling distribution of the OLS estimator**

Motivation

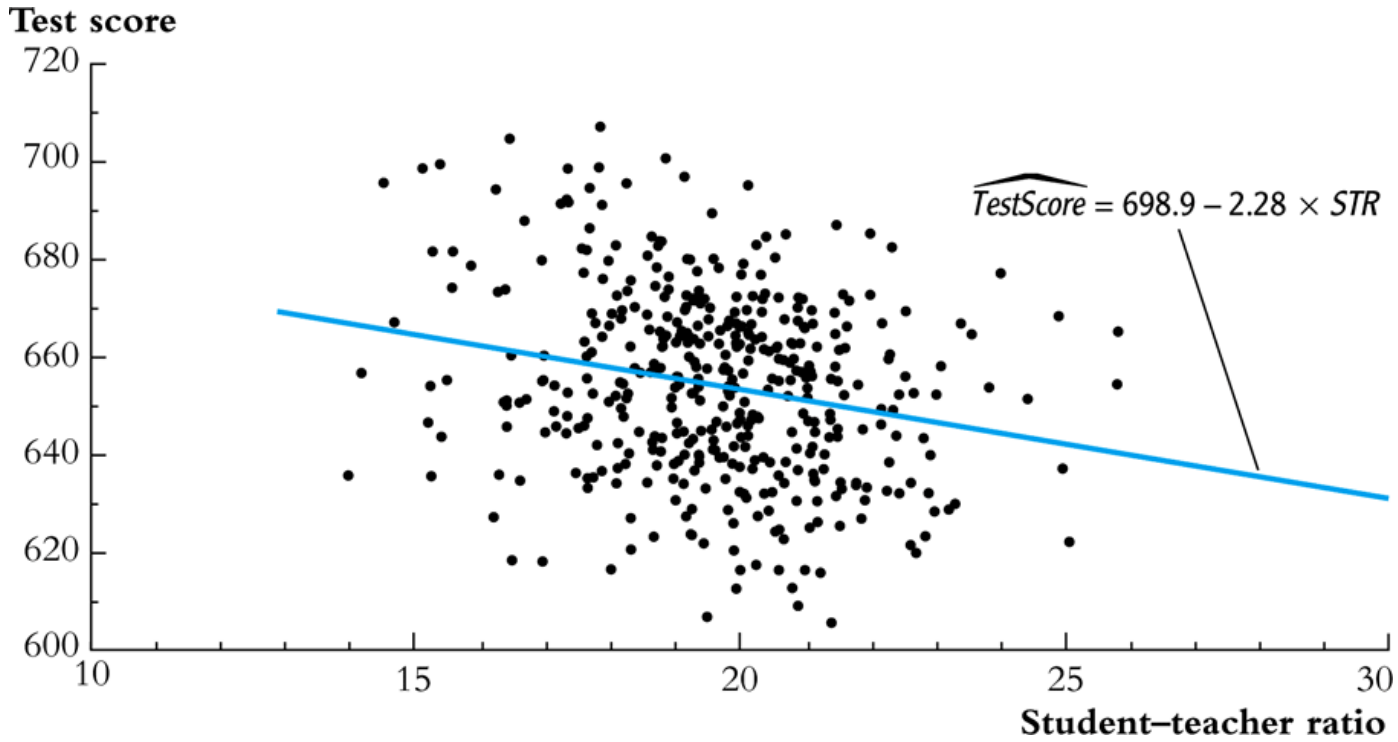
- So far we have treated OLS as a way to draw a straight line through the data on Y and X. Under what conditions does the slope of this line represent the true parameter β_1? That is, when will the OLS estimator be unbiased for the marginal effect on Y of X?
- What is the variance of the OLS estimator over repeated samples?
- To answer these questions, we need to make some assumptions about how Y and X are related to each other, and about how they are collected (the sampling scheme)
- These assumptions – there are three – are known as the Least Squares Assumptions

Sample mean: Flashback to lecture 5

In a sample of $25, \bar{x} = 1.63$ and $s = 0.51$. Construct a 95 percent confidence interval for μ.

Solution:
2.064 is the 95% critical value from a t distribution with 24 degrees of freedom. Thus, the confidence interval is $1.63 \pm [2.064(0.51)/5]$ or $[1.4195, 1.8405]$.

OLS estimator: California *Test Score vs Class Size* data



- Estimated slope $= \hat{\beta}_1 = -2.28$
- Estimated intercept $= \hat{\beta}_0 = 698.9$
- Estimated regression line: $\widehat{TestScore} = 698.9 - 2.28 \times STR$

Can we similarly set a confidence interval for β_1 ? What if it includes 0?

Task at hand: to characterize the sampling distribution of the OLS estimator. To do so, we make three assumptions:

The Least Squares Assumptions

Let β_1 be the effect on Y of a change in X:

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \quad i = 1, \ldots, n$$

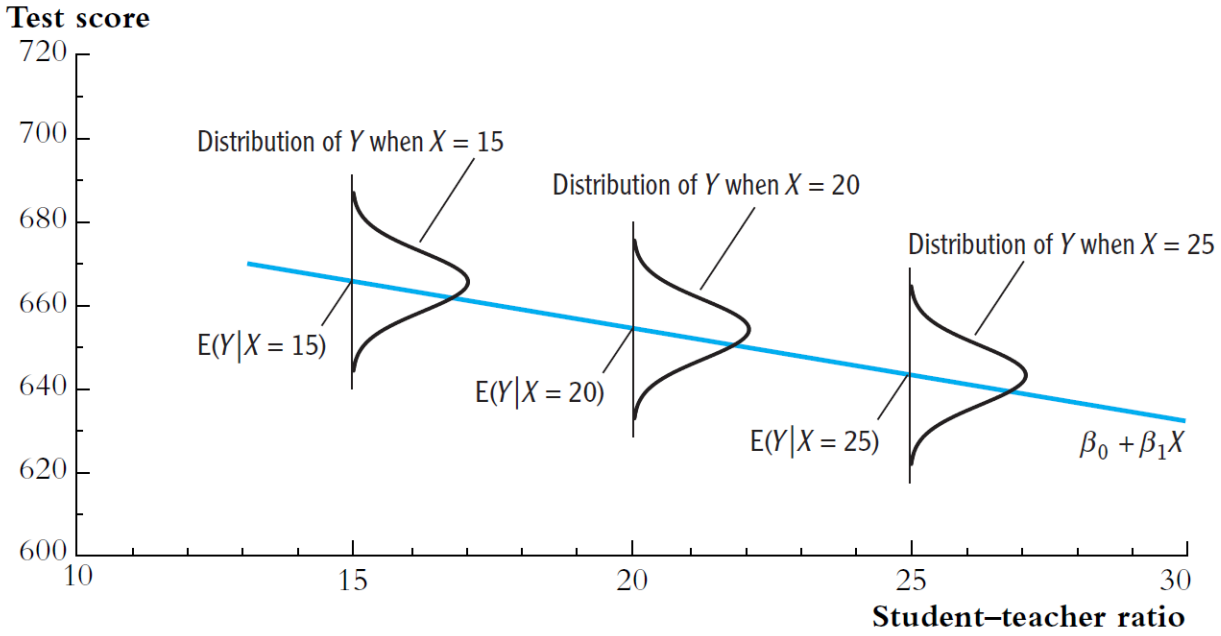
1. The conditional distribution of u given X has mean zero, that is, $E(u \mid X = x) = 0$.

   - This implies that $\hat{\beta}_1$ is unbiased, that is $E[\hat{\beta}_1] = \beta_1$

2. (X_i, Y_i), $i = 1, \ldots, n$, are i.i.d.
   - *This is true if (X, Y) are collected by simple random sampling*
   - *This delivers the sampling distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$*

3. Large outliers in X and/or Y are rare.
   - *Technically, X and Y have finite fourth moments*
   - *Outliers can result in meaningless values of $\hat{\beta}_1$*

Least squares assumption #1: $E(u|X = x) = 0$. (1 of 3)



The figure shows the conditional probability of test scores for districts with class sizes of 15, 20, and 25 students. The mean of the conditional distribution of test scores, given the student–teacher ratio, $E(Y|X)$, is the population regression line. At a given value of X, Y is distributed around the regression line and the error, $u = Y - (\beta_0 + \beta_1 X)$, has a conditional mean of zero for all values of X.

- This condition requires **independence;** stronger than "**no correlation**"
- https://wwwedu.github.io/BC4400/Lecture3/Lecture3.pdf#page=19

Least squares assumption #1: $E(u|X = x) = 0$. (2 of 3)

- Example: Assumption #1 and the class size example
  - $Test\ Score_i = \beta_0 + \beta_1 STR_i + u_i$, u_i = other factors

"Other factors:"

- parental involvement
- outside learning opportunities (extra math class,..)
- home environment conducive to reading
- family income is a useful proxy for many such factors

So $E(u|X=x) = 0$ means $E(Family\ Income \mid STR)$ = constant (which implies that family income and *STR* are uncorrelated). *This assumption is not innocuous! We will return to it often.*

Least squares assumption #1: $E(u|X = x) = 0$. (3 of 3)

- The benchmark for understanding this assumption is to consider an ideal randomized controlled experiment:
- X is randomly assigned to people (students randomly assigned to different size classes; patients randomly assigned to medical treatments). Randomization is done by computer – using no information about the individual.
- Because X is assigned randomly, all other individual characteristics – the things that make up u – are distributed independently of X, so u and X are independent
- Thus, in an ideal randomized controlled experiment, $\boldsymbol{E(u|X = x) = 0}$ (that is, LSA #1 holds)
- In actual experiments, or with observational data, we will need to think hard about whether $E(u|X = x) = 0$ holds.

Least squares assumption #2: (X_i, Y_i), $i = 1,\ldots,n$ are i.i.d.

This arises automatically if the entity (individual, district) is sampled by simple random sampling:

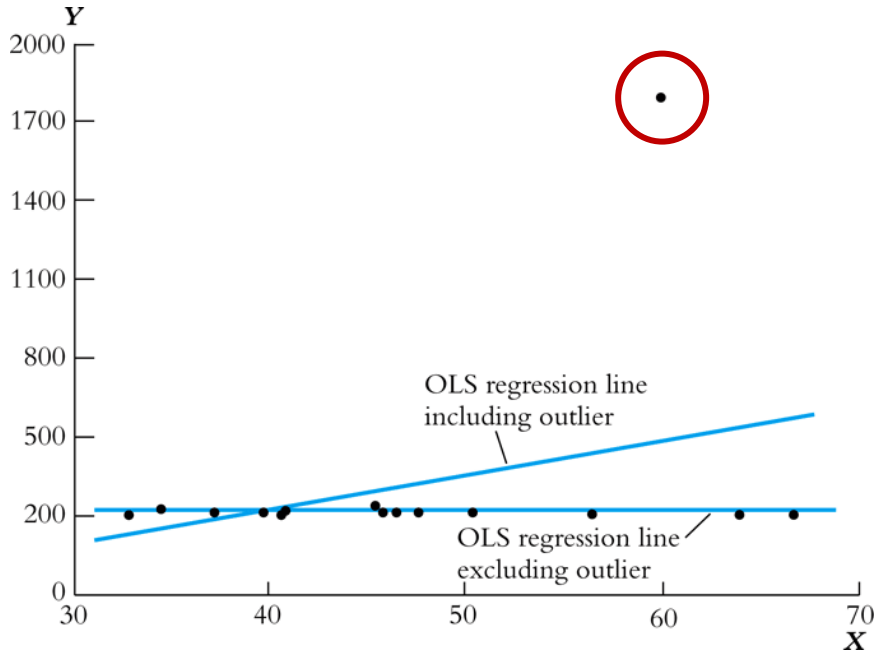
- The entities are selected from the same population, so (X_i, Y_i) are *identically distributed* for all $i = 1,\ldots, n$.
- The entities are selected at random, so the values of (X, Y) for different entities are *independently distributed*.

The main place we will encounter non-i.i.d. sampling is when data are recorded over time for the same entity (panel data and time series data) – we will deal with that complication when we cover panel data.

Least squares assumption #3: *Large outliers are rare*
Technical statement: $E(X^4) < \infty$ and $E(Y^4) < \infty$

- A large outlier is an extreme value of X or Y
- On a technical level, if X and Y are bounded, then they have finite fourth moments. (Standardized test scores automatically satisfy this; *STR*, family income, etc. satisfy this too.)
- The substance of this assumption is that a large outlier can strongly influence the results – so we need to rule out large outliers.
- Look at your data! If you have a large outlier, is it a typo? Does it belong in your data set? Why is it an outlier?

OLS can be sensitive to an outlier:



- *Is the lone point an outlier in X or Y?*
- In practice, outliers are often data glitches (coding or recording problems). Sometimes they are observations that really shouldn't be in your data set. Plot your data!

The Sampling Distribution of the OLS Estimator (SW Section 4.5)

The OLS estimator is computed from a sample of data. A different sample yields a different value of $\hat{\beta}_1$. This is the source of the "sampling uncertainty" of $\hat{\beta}_1$. We want to:

- quantify the sampling uncertainty associated with $\hat{\beta}_1$
- use $\hat{\beta}_1$ to test hypotheses such as $\beta_1 = 0$
- construct a confidence interval for β_1
- All these require figuring out the sampling distribution of the OLS estimator. Two steps to get there…
  - Probability framework for linear regression
  - Distribution of the OLS estimator

Probability Framework for Linear Regression

The probability framework for linear regression is summarized by the three least squares assumptions.

Population

- The group of interest (ex: all possible school districts)

Random variables*: *Y*, *X

- Ex: (*Test Score*, *STR*)

***Joint distribution of* (*Y*, *X*). We assume:**

- The population regression function is linear
- $E(u|X) = 0$ (1st Least Squares Assumption)
- X, Y have nonzero finite fourth moments (3rd L.S.A.)

Data Collection by simple random sampling implies:

- $\{(X_i, Y_i)\}$, $i = 1,\ldots, n$, are i.i.d. (2nd L.S.A.)

The Sampling Distribution of $\hat{\beta}_1$

- Like $\bar{Y}$, $\hat{\beta}_1$ has a sampling distribution.
- What is $E(\hat{\beta}_1)$?
  - If $E(\hat{\beta}_1) = \beta_1$, then OLS is unbiased – a good thing!
- What is var($\hat{\beta}_1$)? (measure of sampling uncertainty)
  - We need to derive a formula so we can compute the standard error of β_1.
- What is the distribution of $\hat{\beta}_1$ in small samples?
  - It is very complicated in general
- What is the distribution of $\hat{\beta}_1$ in large samples?
  - In large samples, $\hat{\beta}_1$ is normally distributed.

The mean and variance of the sampling distribution of $\hat{\beta}_1$ (1 of 3)

Some preliminary algebra:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

$$\bar{Y} = \beta_0 + \beta_1 \bar{X} + \bar{u}$$

SO $$Y_i - \bar{Y} = \beta_1 (X_i - \bar{X}) + (u_i - \bar{u})$$

Thus,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})[\beta_1 (X_i - \bar{X}) + (u_i - \bar{u})]}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

The mean and variance of the sampling distribution of $\hat{\beta}_1$ (2 of 3)

$$\hat{\beta}_1 = \beta_1 \frac{\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} + \frac{\sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

SO $$\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

Now $$\sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u}) = \sum_{i=1}^{n}(X_i - \bar{X})u_i - \left[\sum_{i=1}^{n}(X_i - \bar{X})\right]\bar{u}$$

$$= \sum_{i=1}^{n}(X_i - \bar{X})u_i - \left[\left(\sum_{i=1}^{n}X_i\right) - n\bar{X}\right]\bar{u}$$

$$= \sum_{i=1}^{n}(X_i - \bar{X})u_i$$

The mean and variance of the sampling distribution of $\hat{\beta}_1$ (3 of 3)

Substitute $\sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u}) = \sum_{i=1}^{n}(X_i - \bar{X})u_i$ into the expression for $\hat{\beta}_1 - \beta_1$:

$$\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

SO

$$\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Now we can calculate $E(\hat{\beta}_1)$ and $var(\hat{\beta}_1)$:

$$E(\hat{\beta}_1) - \beta_1 = E\left[\frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]$$

Law of iterated expectations (SW 2.20)

$$= E\left\{E\left[\frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right] \middle| X_1, \ldots, X_n\right\}$$

$= 0$ **because $E(u_i \mid X_i = x) = 0$ by LSA #1**

- Thus LSA #1 implies that $E(\hat{\beta}_1) = \beta_1$
- That is, $\hat{\beta}_1$ **is an unbiased estimator of β_1.**
- For details see App. 4.3

Next calculate var($\hat{\beta}_1$) (1 of 2)

write

$$\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\frac{1}{n}\sum_{i=1}^{n} v_i}{\left(\frac{n-1}{n}\right)s_X^2}$$

where $v_i = (X_i - \bar{X})u_i$. If n is large, $s_X^2 \approx \sigma_X^2$ and $\frac{n-1}{n} \approx 1$, so

$$\hat{\beta}_1 - \beta_1 \approx \frac{\frac{1}{n}\sum_{i=1}^{n} v_i}{\sigma_X^2},$$

where $v_i = (X_i - \bar{X})u_i$ (see App. 4.3). Thus,

Next calculate var($\hat{\beta}_1$) (2 of 2)

$$\hat{\beta}_1 - \beta_1 \approx \frac{\frac{1}{n}\sum_{i=1}^{n} v_i}{\sigma_X^2}$$

so $$\text{var}(\hat{\beta}_1 - \beta_1) = \text{var}(\hat{\beta}_1) = \text{var}\left(\frac{1}{n}\sum_{i=1}^{n} v_i\right) \Big/ (\sigma_X^2)^2 = \frac{\text{var}(v_i)/n}{(\sigma_X^2)^2}$$

where the final equality uses assumption 2. Thus,

$$\text{var}(\hat{\beta}_1) = \frac{1}{n} \times \frac{\text{var}[(X_i - \mu_x)u_i]}{(\sigma_X^2)^2}.$$

Summary so far

1. $\hat{\beta}_1$ is unbiased: under LSA#1, $E(\hat{\beta}_1) = \beta_1$ – just like $\bar{Y}$!
2. var($\hat{\beta}_1$) is inversely proportional to n – just like $\bar{Y}$!

What is the sampling distribution of $\hat{\beta}_1$*?*

The exact sampling distribution is complicated – it depends on the population distribution of (Y, X) – but when n is large we get some simple (and good) approximations:

1) Because $\text{var}(\hat{\beta}_1) \propto 1/n$ and $E(\hat{\beta}_1) = \beta_1$, $\hat{\beta}_1 \xrightarrow{p} \beta_1$
2) When n is large, the sampling distribution of $\hat{\beta}_1$ is well approximated by a normal distribution (CLT)

Recall the ***CLT***: suppose $\{v_i\}$, $i = 1, \ldots, n$ is i.i.d. with $E(v) = 0$ and $\text{var}(v) = \sigma^2$. Then, when n is large, $\frac{1}{n}\sum_{i=1}^{n} v_i$ is approximately distributed $N(0, \sigma_v^2/n)$.

Large-n approximation to the distribution of $\hat{\beta}_1$:

$$\hat{\beta}_1 - \beta_1 = \frac{\frac{1}{n}\sum_{i=1}^{n} v_i}{\left(\frac{n-1}{n}\right) S_X^2} \approx \frac{\frac{1}{n}\sum_{i=1}^{n} v_i}{\sigma_X^2}, \text{ where } v_i = (X_i - \bar{X})u_i$$

- When n is large, $v_i = (X_i - \bar{X})u_i \approx (X_i - \mu_X)u_i$, which is i.i.d. (*why*?) and $\text{var}(v_i) < \infty$ (*why*?). So, by the CLT, $\frac{1}{n}\sum_{i=1}^{n} v_i$ is approximately distributed $N(0,\ \sigma_v^2/n)$.
- Thus, for n large, $\hat{\beta}_1$s approximately distributed

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_v^2}{n(\sigma_X^2)^2}\right), \text{ where } v_i = (X_i - \mu_X)u_i$$

The larger the variance of X, the smaller the variance of $\hat{\beta}_1$

The math

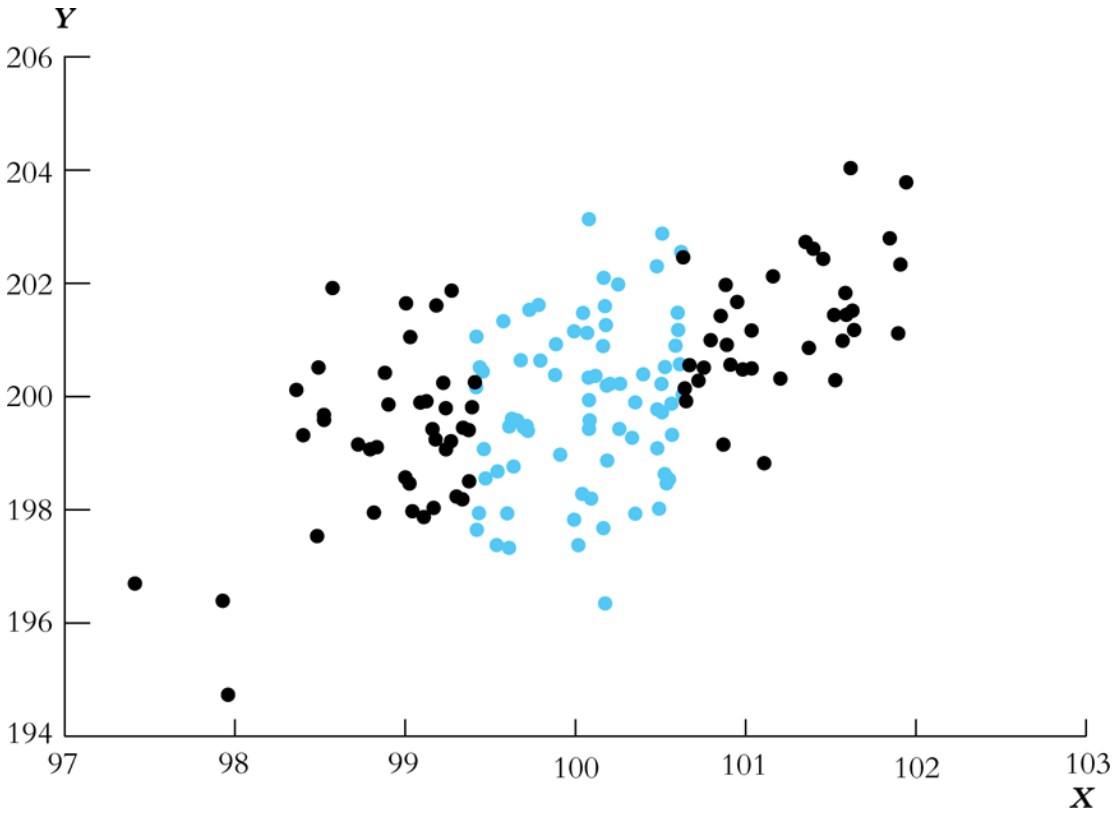
$$\text{var}(\hat{\beta}_1 - \beta_1) = \frac{1}{n} \times \frac{\text{var}[(X_i - \mu_x)u_i]}{(\sigma_X^2)^2}$$

Where $\sigma_X^2 = \text{var}(X_i)$. The variance of X appears (squared) in the denominator – so increasing the spread of X decreases the variance of β_1.

The intuition

If there is more variation in X, then there is more information in the data that you can use to fit the regression line. This is most easily seen in a figure…

The larger the variance of X, the smaller the variance of $\hat{\beta}_1$



The number of black and blue dots is the same. Using which would you get a more accurate regression line?

Summary of the sampling distribution of $\hat{\beta}_1$:

If the three Least Squares Assumptions hold, then

- The exact (finite sample) sampling distribution of $\hat{\beta}_1$ has:
  - $E(\hat{\beta}_1) = \beta_1$ (that is, $\hat{\beta}_1$ is unbiased)
  - $\text{var}(\hat{\beta}_1) = \frac{1}{n} \times \frac{\text{var}[(X_i - \mu_x)u_i]}{\sigma_X^4}$
- Other than its mean and variance, the exact distribution of $\hat{\beta}_1$ is complicated and depends on the distribution of (X, u)
- $\hat{\beta}_1 \xrightarrow{p} \beta_1$ (that is, $\hat{\beta}_1$ is consistent)
- When n is large, $\frac{\hat{\beta}_1 - E(\hat{\beta}_1)}{\sqrt{\text{var}(\hat{\beta}_1)}} \sim N(0,\ 1)$ (CLT)
- *This parallels the sampling distribution of* $\bar{Y}$.

Key Concept 4.4: Large-Sample Distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$

If the least squares assumptions in Key Concept 4.3 hold, then in large samples $\hat{\beta}_0$ and $\hat{\beta}_1$ have a jointly normal sampling distribution. The large-sample normal distribution of $\hat{\beta}_1$ is $N(\beta_1, \sigma^2_{\hat{\beta}_1})$, where the variance of this distribution, $\sigma^2_{\hat{\beta}_1}$, is

$$\sigma^2_{\hat{\beta}_1} = \frac{1}{n}\frac{\text{var}[(X_i - \mu_X)u_i]}{[\text{var}(X_i)]^2}. \quad (4.21)$$

The large-sample normal distribution of $\hat{\beta}_0$ is $N(\beta_0, \sigma^2_{\hat{\beta}_0})$, where

$$\sigma^2_{\hat{\beta}_0} = \frac{1}{n}\frac{\text{var}(H_i u_i)}{[E(H_i^2)]^2}, \text{ where } H_i = 1 - \left[\frac{\mu_X}{E(X_i^2)}\right]X_i. \quad (4.22)$$

We are now ready to turn to hypothesis tests & confidence intervals…Chapter 5