Gov 2000 - 7. Simple Linear Regression

Matthew Blackwell
Harvard University
mblackwell@gov.harvard.edu

Where are we? Where are we going?

- Last week: motivating the idea of regression and deriving an estimator for the parameters of a linear regression model.
- This week: investigating the properties of the least squares estimator and the assumptions of the linear regression model.

REVIEW

The population linear regression function

- The (population) simple linear regression model can be stated as the following:

$$r(x) = E[Y|X = x] = \beta_0 + \beta_1 x$$

- This describes the data generating process in the population
- Y = dependent variable
- X = independent variable
- β_0, β_1 = population intercept and population slope (what we want to estimate)

The sample linear regression function

- The estimated or sample regression function is:

$$\widehat{r}(X_i) = \widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$$

- $\widehat{\beta}_0, \widehat{\beta}_1$ are the estimated intercept and slope
- $\widehat{Y}_i$ is the fitted/predicted value
- We also have the residuals, $\widehat{u}_i$ which are the differences between the true values of Y and the predicted value:

$$\widehat{u}_i = Y_i - \widehat{Y}_i$$

- You can think of the residuals as the prediction errors of our estimates.

GOALS

Overall goals

- Learn how to run and read regression
- Mechanics: how to estimate the intercept and slope?
- Properties: when are these good estimates?
- Uncertainty: how will the OLS estimator behave in repeated samples?
- Testing: can we assess the plausibility of no relationship ($\beta_1 = 0$)?
- Interpretation: how do we interpret our estimates?

More narrow goal

- A more narrow goal is to understand everything from an R regression output:

```r
ajr <- foreign::read.dta(”ajr.dta”)
out <- lm(logpgp95 ~ logem4, data = ajr)
summary(out)
```

```
##
## Call:
## lm(formula = logpgp95 ~ logem4, data = ajr)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.71304 -0.53326  0.01954  0.47188  1.44673
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.66025    0.30528   34.92  < 2e-16 ***
## logem4      -0.56412    0.06389   -8.83 2.09e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7563 on 79 degrees of freedom
##   (82 observations deleted due to missingness)
## Multiple R-squared:  0.4967, Adjusted R-squared:  0.4903
## F-statistic: 77.96 on 1 and 79 DF,  p-value: 2.094e-13
```

MECHANICS OF OLS

What is OLS?

- An estimator for the slope and the intercept of the regression line
- We talked last week about ways to derive this estimator and we settled on deriving it by minimizing the squared prediction errors of the regression, or in other words, minimizing the sum of the squared residuals:
- **Ordinary Least squares** (OLS):

$$(\widehat{\beta}_0, \widehat{\beta}_1) = \underset{b_0, b_1}{\arg\min} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

- In words, the OLS estimates are the intercept and slope that minimize the sum of the squared residuals.

Intuition of the OLS estimator

- The intercept equation tells us that the regression line goes through the point $(\overline{Y}, \overline{X})$:

$$\overline{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 \overline{X}$$

- The slope for the regression line can be written as the following:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2} = \frac{\text{Sample Covariance between } X \text{ and } Y}{\text{Sample Variance of } X}$$

- The higher the covariance between X and Y, the higher the slope will be.

- Negative covariances = negative slopes; positive covariances = positive slopes

- What happens when X_i doesn't vary?
- What happens when Y_i doesn't vary?

Mechanical properties of OLS

- Later we'll see that under certain assumptions, OLS will have nice statistical properties. But some properties of OLS are mechanical in the sense that they are just a function of how we estimated the slope and intercept.
- Each of these can be derived from the first order conditions of OLS.
- The residuals will be 0 on average:

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{u}_i = 0$$

- The residuals will be uncorrelated with the predictor ($\widehat{\text{cov}}$ is the sample covariance):

$$\widehat{\text{cov}}(X_i, \widehat{u}_i) = 0$$

- The residuals will be uncorrelated with the fitted values:

$$\widehat{\text{cov}}(\widehat{Y}_i, \widehat{u}_i) = 0$$

- Note that these are properties of the estimated residuals, $\widehat{u}_i$, not the true errors, u_i!

OLS slope as a weighted sum of the outcomes

- One useful derivation that we'll do moving forward is to write the OLS estimator for the slope as a weighted sum of the outcomes.

$$\begin{aligned}
\widehat{\beta}_1 &= \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \\
&= \frac{\sum_{i=1}^{n}(X_i - \overline{X})Y_i}{\sum_{i=1}^{n}(X_i - \overline{X})^2} - \frac{\sum_{i=1}^{n}(X_i - \overline{X})\overline{Y}}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \\
&= \frac{\sum_{i=1}^{n}(X_i - \overline{X})Y_i}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \\
&= \sum_{i=1}^{n} W_i Y_i
\end{aligned}$$

- Where here we have the weights, W_i as:

$$W_i = \frac{(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

- This is important for two reasons. First, it'll make derivations later much easier. And second, it shows that is just the sum of a random variable. Therefore it is also a random variable.

PROPERTIES OF THE OLS ESTIMATOR

Sampling distribution of the OLS estimator

- Remember: OLS is an estimator—it's a machine that we plug data into and we get out estimates. Just like the sample mean, sample difference in means, or the sample variance. It's a more complicated estimator, to be sure, but it still has the same basic structure as the others.

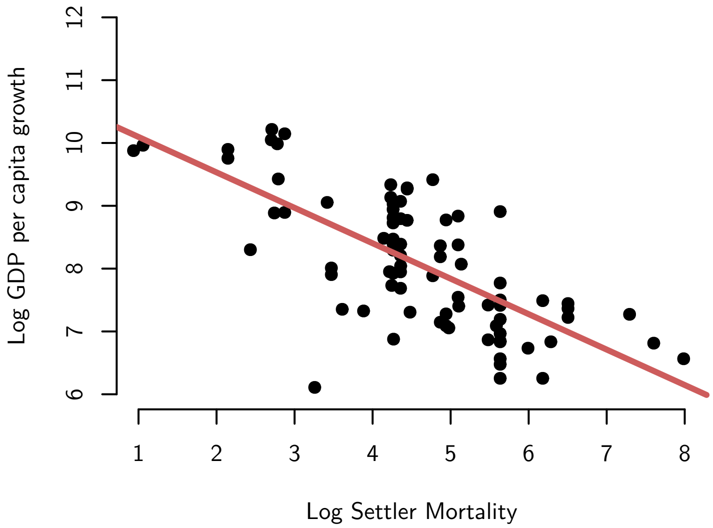
- It has a sampling distribution, with a sampling variance/standard error, etc.

- Let's simulate some data to get a sense for how the sampling distribution of the OLS estimators works.

- To do this, we're going to pretend that the AJR data represents the population of interest and we are going to take samples from it to see how the regression line varies from sample to sample. (Note that this is just for demonstration since we'll never actually have the whole population)

- First, let's plot the population regression line:

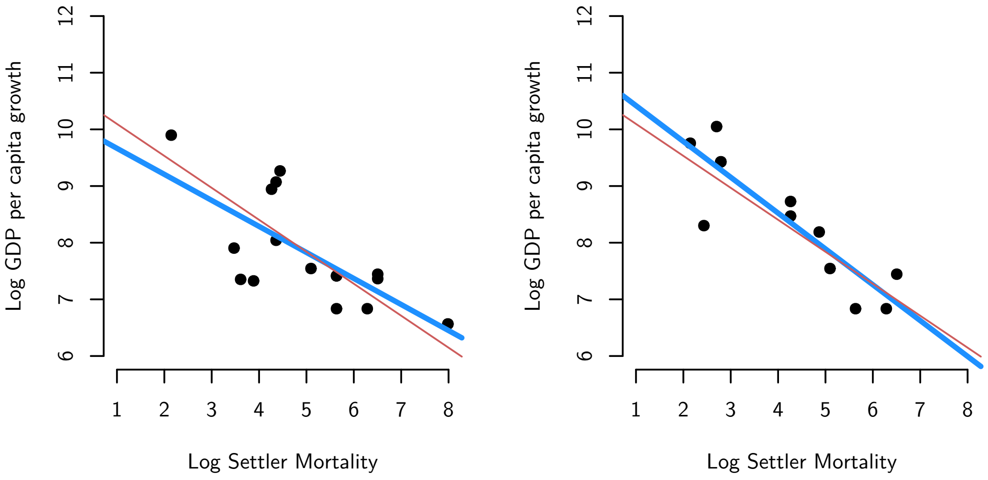
```
ajr <- foreign::read.dta(”ajr.dta”)
plot(ajr$logem4, ajr$logpgp95, xlab = ”Log Settler Mortality”, ylab = ”Log GDP per capita growth”,
    pch = 19, bty = ”n”, xlim = c(1, 8), ylim = c(6, 12))
abline(lm(logpgp95 ~ logem4, data = ajr), col = ”indianred”, lwd = 3)
```



- Now, let's take two random samples of size $n = 30$ from this "population" and the plot the results, with the true population line overlaid:

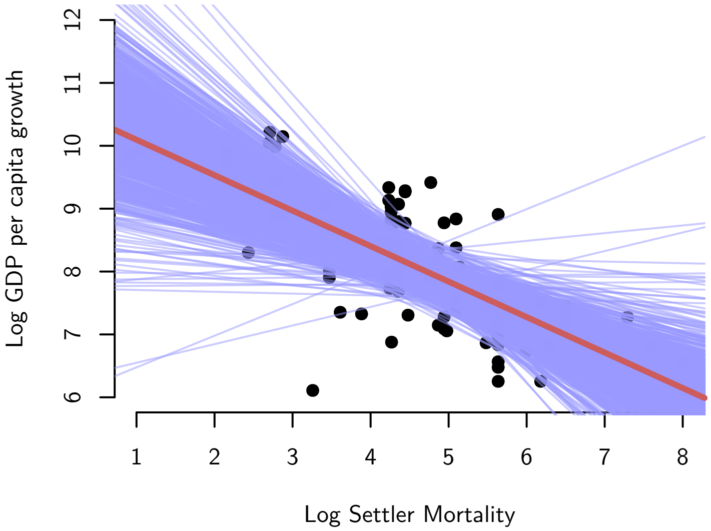
```
set.seed(2143)
par(mfrow = c(1, 2))
ajr.samp <- ajr[sample(1:nrow(ajr), size = 30, replace = TRUE), ]
plot(ajr.samp$logem4, ajr.samp$logpgp95, xlab = "Log Settler Mortality", ylab = "Log GDP per capita growth",
    pch = 19, bty = "n", xlim = c(1, 8), ylim = c(6, 12))
abline(lm(logpgp95 ~ logem4, data = ajr.samp), col = "dodgerblue", lwd = 3)
abline(lm(logpgp95 ~ logem4, data = ajr), col = "indianred", lwd = 1)

ajr.samp2 <- ajr[sample(1:nrow(ajr), size = 30, replace = TRUE), ]
plot(ajr.samp2$logem4, ajr.samp2$logpgp95, xlab = "Log Settler Mortality", ylab = "Log GDP per capita growth",
    pch = 19, bty = "n", xlim = c(1, 8), ylim = c(6, 12))
abline(lm(logpgp95 ~ logem4, data = ajr.samp2), col = "dodgerblue", lwd = 3)
abline(lm(logpgp95 ~ logem4, data = ajr), col = "indianred", lwd = 1)
```



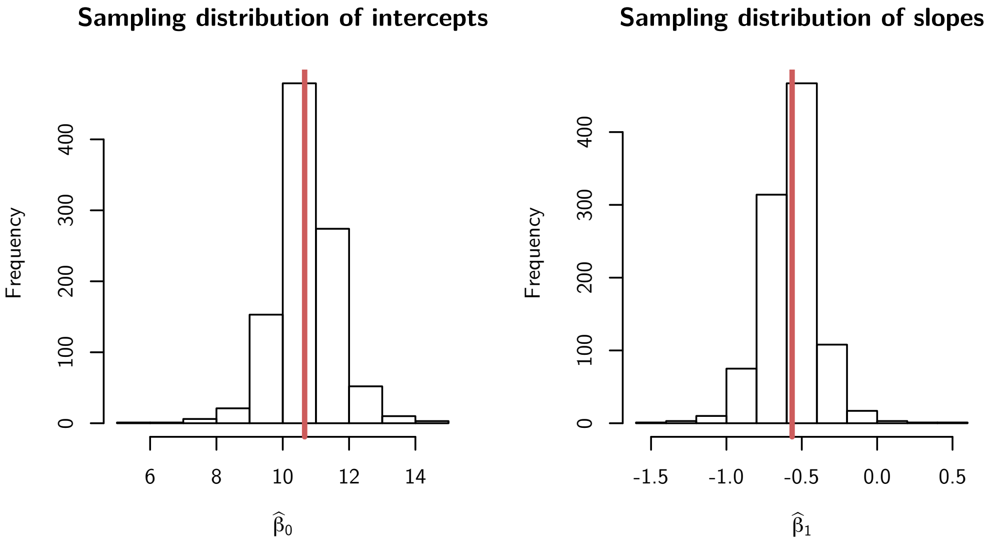
- Note how in our two samples, the slope and the intercept are not exactly the same as in the population. In the first, the estimated intercept is *lower* than the population intercept, while in the second the estimated intercept is higher.
- In the first sample, the estimated slope is closer to 0 than the true slope. In the second sample, it's more negative than the true slope.
- This is just due to random sampling!
- Now let's repeat this process 1000 times to see how the slopes and intercepts vary in lots of repeated samples:

```
set.seed(2143)
true.reg <- lm(logpgp95 ~ logem4, data = ajr)
sims <- 1000
inters <- rep(NA, times = sims)
slopes <- rep(NA, times = sims)
plot(ajr$logem4, ajr$logpgp95, xlab = ”Log Settler Mortality”, ylab = ”Log GDP per capita growth”,
    pch = 19, bty = ”n”, xlim = c(1, 8), ylim = c(6, 12))
for (i in 1:sims) {
    ajr.samp <- ajr[sample(1:nrow(ajr), size = 30, replace = TRUE), ]
    this.reg <- lm(logpgp95 ~ logem4, data = ajr.samp)
    abline(this.reg, col = rgb(0.6, 0.6, 1, alpha = 0.5), lwd = 1)
    inters[i] <- coef(this.reg)[1]
    slopes[i] <- coef(this.reg)[2]
}
abline(true.reg, col = ”indianred”, lwd = 3)
```



- You can see that the estimated slopes and intercepts vary from sample to sample, but that the “average” of the lines looks about right. We can look at the sampling distribution of the parameters separately to see that this is about right:

```
par(mfrow = c(1, 2))
hist(inters, xlab = expression(hat(beta)[0]), main = ”Sampling distribution of intercepts”)
abline(v = coef(true.reg)[1], col = ”indianred”, lwd = 3)
hist(slopes, xlab = expression(hat(beta)[1]), main = ”Sampling distribution of slopes”)
abline(v = coef(true.reg)[2], col = ”indianred”, lwd = 3)
```



- The sampling distribution of the OLS estimators are centered roughly around their true value. Remember that we call this property unbiasedness of the estimators.
- Here's the question: will OLS always be unbiased? Under what assumptions will it be unbiased or consistent?

Assumptions needed for the unbiasedness of the sample mean

- What assumptions did we make to prove that the sample mean was unbiased?
- Just one: that we had a random or iid sample from the population.
- We'll need more than this for the regression case

Assumptions for unbiasedness and consistency of OLS

- Generally we'll need different assumptions to derive different properties of the OLS estimator.
- For unbiasedness and consistency, we'll need the following assumptions.

1. Linearity
2. Random (iid) sample
3. Variation in X_i
4. Zero conditional mean of the errors

Assumption 1: Linearity

Assumption 1 - The population regression function is linear in the parameters:

$$Y = \beta_0 + \beta_1 X_i + u$$

- u is the unobserved error or disturbance term that represents all factors influencing Y other than X.
- Violation of the linearity assumption:

$$Y = \frac{1}{\beta_0 + \beta_1 X} + u$$

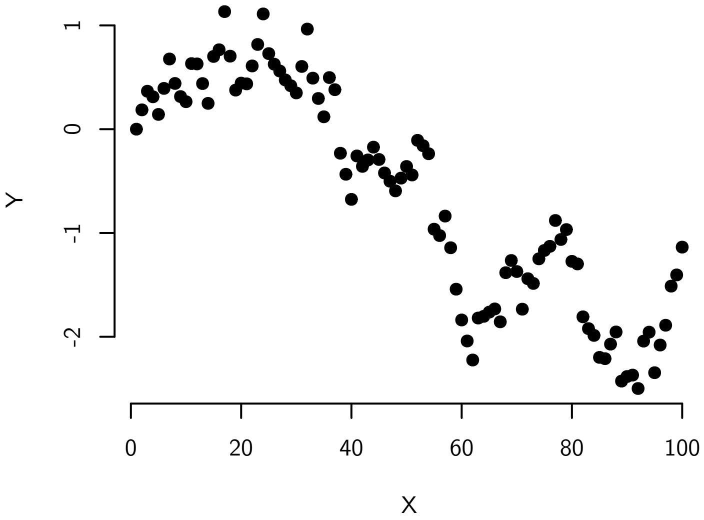
- **Not** a violation of the linearity assumption:

$$Y = \beta_0 + \beta_1 X^2 + u$$

Assumption 2: Random Sample

Assumption 2 - We have a iid random sample of size n, $\{(Y_i, X_i) : i = 1, 2, \ldots, n\}$ from the population regression model above.

- Violation of the linearity assumption: time-series, selected samples.



- Think about the weight example from last week, where Y_i was my weight on a given day and X_i was my number of active minutes the day before:

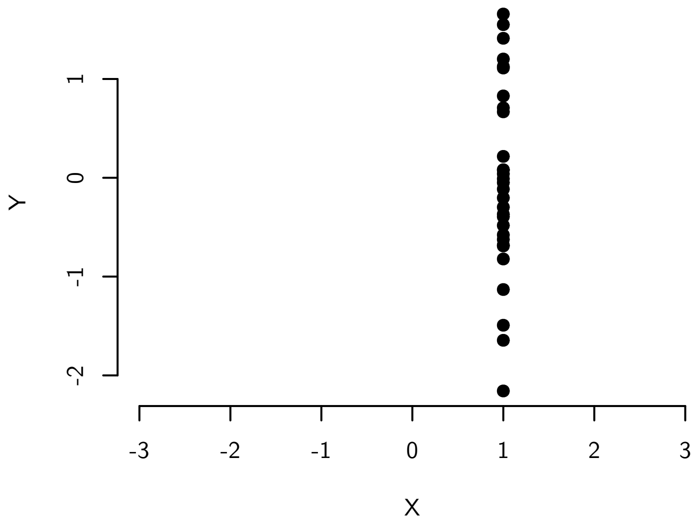
$$\text{weight}_i = \beta_0 + \beta_1 \text{activity}_i + u_i$$

- What if I only weighed myself on the weekdays?

Assumption 3: Variation in X

Assumption 3 - The in-sample independent variables, $\{X_i : i = 1, \ldots, n\}$, are not all the same value.

- Why does this matter? How would you draw the line of best fit through this scatterplot, which is a violation of this assumption?



- Also remember the formula for the OLS slope estimator:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

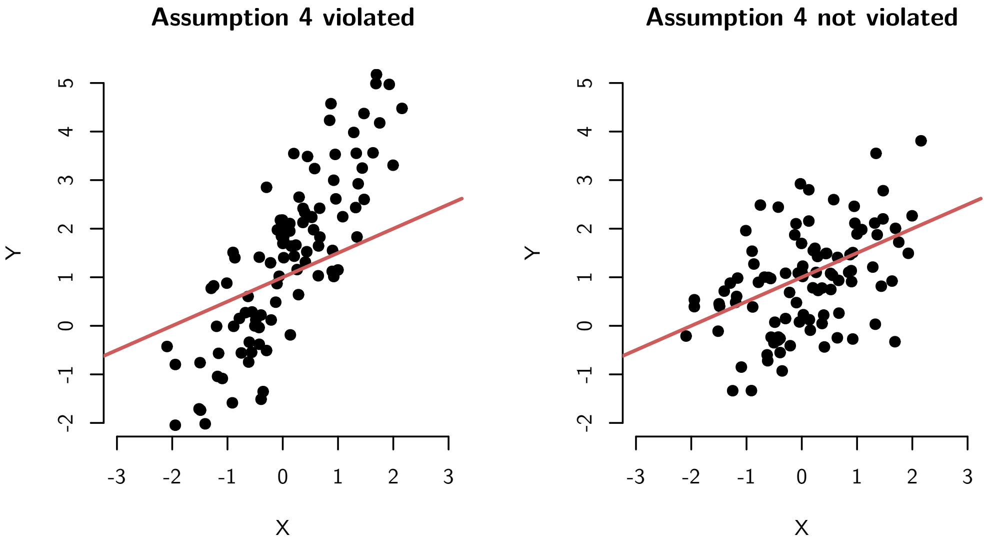
- What happens here when X_i doesn't vary?

Assumption 4: Zero conditional mean of the errors

Assumption 4 - The error, u_i, has expected value of 0 given any value of the independent variable:

$$\mathbb{E}[u_i|X_i = x] = 0 \quad \forall x.$$

- How does this assumption get violated? Let's generate data from the following model:
$$Y_i = 1 + 0.5X_i + u_i$$
- But let's compare two situations. One where X_i and u_i are correlated so that the mean of u_i depends on X_i (a violation of Assumption 4) and one where there is no correlation (not a violation). Let's plot this data along with the true regression line ($\beta_0 = 1$ and $\beta_1 = 0.5$):



- For the violation, you can see that for low values of X_i most of the errors are negative and for high values of X_i, most of the errors are positive. You can also see that the sample of data points doesn't really fit the regression line at all.

- Compare this to a situation with no correlation between X_i and u_i, where the errors are roughly 0 on average, no matter the value of X_i.

Zero conditional mean in the error example

- Think about the weight example from last week, where Y_i was my weight on a given day and X_i was my number of active minutes the day before:

$$\text{weight}_i = \beta_0 + \beta_1 \text{activity}_i + u_i$$

- What might in u_i here? Amount of food eaten, workload, etc etc.
- We have to assume that all of these factors have the same mean, no matter what my level of activity was. Plausible?
- When is this assumption most plausible? When X_i is randomly assigned.

Unbiasedness

- With Assumptions 1-4, we can show that the OLS estimator for the slope is unbiased, that is $\mathbb{E}[\widehat{\beta}_1] = \beta_1$.

- There are two ways that we use the above assumptions. First, we can establish that the conditional expectation function (CEF)

$$\begin{aligned}\mathbb{E}[Y_i|X_1,\ldots,X_n] &= \mathbb{E}[Y_i|X_i] \qquad \text{(A2: iid)}\\ &= \mathbb{E}[\beta_0+\beta_1X_i+u|X_i] \qquad \text{(A1: linearity)}\\ &= \beta_0+\beta_1X_i+\mathbb{E}[u_i|X_i]\\ &= \beta_0+\beta_1X_i \qquad \text{(A4: zero mean error)}\end{aligned}$$

- Second, note that we can only calculate $\widehat{\beta}_1$ when Assumption 3 (variation in X) holds.
- With these two facts, we can show show that $\mathbb{E}[\widehat{\beta}_1|X_1,\ldots,X_n]=\beta_1$.
- Remember that we showed that $\widehat{\beta}_1=\sum_{i=1}^n W_iY_i$. We're going to use this fact. Also remember that W_i is a function of all observations of the independent variable since it contains the mean, so conditional on $\mathbf{X}=(X_1,\ldots,X_n)$, it is constant.

$$\begin{aligned}\mathbb{E}[\widehat{\beta}_1|X_1,\ldots,X_n] &= E\left[\sum_{i=1}^n W_iY_i\middle|X_1,\ldots,X_n\right]\\ &= \sum_{i=1}^n E\left[W_iY_i|X_1,\ldots,X_n\right]\\ &= \sum_{i=1}^n W_iE\left[Y_i|X_1,\ldots,X_n\right]\\ &= \sum_{i=1}^n W_i(\beta_0+\beta_1X_i) \qquad \text{(result above)}\\ &= \beta_0\sum_{i=1}^n W_i+\beta_1\sum_{i=1}^n W_iX_i\end{aligned}$$

- Are we stuck? No! Because we can show that $\sum_{i=1}^n W_i=0$ and $\sum_{i=1}^n W_iX_i=1$:

$$\begin{aligned}\sum_{i=1}^{n} W_i &= \sum_{i=1}^{n} \frac{(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \\ &= \frac{1}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \sum_{i=1}^{n}(X_i - \overline{X}) \\ &= \frac{1}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \cdot 0 = 0\end{aligned}$$

- This works because the sum of deviations from the mean are 0! Now, the second fact:

$$\begin{aligned}\sum_{i=1}^{n} W_i X_i &= \sum_{i=1}^{n} \frac{X_i(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \\ &= \frac{1}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \sum_{i=1}^{n} X_i(X_i - \overline{X}) \\ &= \frac{1}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \left[\sum_{i=1}^{n} X_i(X_i - \overline{X}) - \sum_{i=1}^{n} \overline{X}(X_i - \overline{X})\right] \\ &= \frac{1}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X}) \\ &= 1\end{aligned}$$

- Plugging this back into our original derivation, we get the following:

$$\mathbb{E}[\widehat{\beta}_1 | X_1, \ldots, X_n] = \beta_0 \cdot 0 + \beta_1 \cdot 1 \quad = \beta_1$$

- Now, noticed that we conditioned on $X_1, \ldots, X_n$. But we need to show that $\mathbb{E}[\widehat{\beta}_1] = \beta_1$. Let's use the law of iterated expectations!

$$\begin{aligned}\mathbb{E}[\widehat{\beta}_1] &= \mathbb{E}[\mathbb{E}[\widehat{\beta}_1 | X_1, \ldots, X_n]] \\ &= \mathbb{E}[\beta_1] \\ &= \beta_1\end{aligned}$$

- The basic intuition here that the condition mean given the independent variable is the same, no matter the value of the independent variables. Therefore, the overall mean must just be equal to that constant.

- Recap: linearity, random sampling, variation in X, and zero conditional mean for the error will get us unbiasedness.

Consistency

- Under the same set of assumptions, we can show that the OLS estimator is consistent, so that $\widehat{\beta}_1 \xrightarrow{p} \beta_1$.

- It's not very hard to prove, but we'll skip the proof because it involves properties of convergence in probability. You can find the proof in the appendix of these notes.

Where are we?

- Now we know that, under Assumptions 1-4, we know that $\widehat{\beta}_1 \sim ?(\beta_1, ?)$

- That is we know that the sampling distribution is centered on the true population slope, but we don't know the population variance.

Sampling variance of estimated slope

- In order to derive the sampling variance of the OLS estimator,

1. Linearity
2. Random (iid) sample
3. Variation in X_i
4. Zero conditional mean of the errors
5. Homoskedasticity

Assumption 5: Homoskedasticity

Assumption 5 - The conditional variance of Y_i given X_i is constant:

$$\mathbb{V}(Y_i|X_i = x) = \mathbb{V}(u_i|X_i = x) = \sigma_u^2.$$

- The conditional variance of Y given X is sometimes called the **skedastic function**, thus the name homoskedasticity.

Deriving the sampling variance

$$
\begin{aligned}
\mathbb{V}[\widehat{\beta}_1|X_1,\ldots,X_n] &= \mathbb{V}\left[\sum_{i=1}^{n} W_i Y_i \middle| X_1,\ldots,X_n\right] \\
&= \sum_{i=1}^{n} W_i^2 \mathbb{V}\left[Y_i|X_1,\ldots,X_n\right] && \text{(A2: iid)} \\
&= \sum_{i=1}^{n} W_i^2 \mathbb{V}\left[Y_i|X_i\right] && \text{(A2: iid)} \\
&= \sigma_u^2 \sum_{i=1}^{n} W_i^2 && \text{(A5: homoskedastic)} \\
&= \frac{\sigma_u^2 \sum_{i=1}^{n}(X_i-\overline{X})^2}{\left(\sum_{i=1}^{n}(X_i-\overline{X})^2\right)^2} \\
&= \frac{\sigma_u^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}
\end{aligned}
$$

- What drives the sampling variability of the OLS estimator?
  - The higher the variance of Y_i, the higher the sampling variance
  - The lower the variance of X_i, the higher the sampling variance
  - As we increase n, the denominator gets large, while the numerator is fixed and so the sampling variance shrinks to 0.

Estimating the sampling variance/standard error

- We just saw that $\mathbb{V}(\widehat{\beta}_1|X) = \frac{\sigma_u^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}$
- But we don't observe σ_u^2—it is the variance of the errors, which we don't observe. What can we do? Estimate it using the residuals!

$$\widehat{\sigma}_u^2 = \frac{1}{n-2}\sum_{i=1}^{n}\widehat{u}_i^2$$

- Why $n-2$ instead of n or $n-1$? Remember that OLS is designed to minimize the sum of the squared residuals, so it tends to slightly underestimate the variance. The $n-2$ corrects this.
- With this, we can find the estimated standard error of our OLS estimator of the slope:

$$\widehat{SE}[\widehat{\beta}_1] = \frac{\sqrt{\widehat{\sigma}_u^2}}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2}} = \frac{\widehat{\sigma}_u}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2}}$$

Gauss-Markov Theorem

Theorem Under assumptions 1-5, the OLS estimator is BLUE, or the Best Linear Unbiased Estimator, where by "best" we mean it lowest sampling variance.

- The proof is very detailed, so we'll skip it. See Wooldridge, Appendix 3A.6 for details.
- Fails to hold when the assumptions are violated!

Large-sample inference distribution of OLS estimators

- Remember that we can write $\widehat{\beta}_1 = \sum_{i=1}^n W_i Y_i$, so that the OLS estimator is the sum of independent r.v.'s.
- Also remember the mantra of the central limit theorem: "the sums and means of r.v.'s tend to be Normally distributed in large samples."
- Also true here, by a more advanced version of the CLT (Liapounov CLT for fixed regressors, Maringale Difference CLT for general regressors), we know that in large samples:
$$\frac{\widehat{\beta}_1 - \beta_1}{SE[\widehat{\beta}_1]} \sim N(0, 1)$$
- Also, in large samples, remember that we can replace the true standard error with our estimate of the standard error, so that:
$$\frac{\widehat{\beta}_1 - \beta_1}{\widehat{SE}[\widehat{\beta}_1]} \sim N(0, 1)$$

Sampling distribution in small samples

- What if we have a small sample? What can we do then?
- We still know that $\widehat{\beta} \sim ?(\beta_1, SE[\widehat{\beta}]^2)$ since we know that unbiasedness holds and we know how to calculate the sampling variance. We just don't know the form of the sampling distribution.
- Can't get something for nothing, but we can make progress if we make another assumption:

1. Linearity
2. Random (iid) sample
3. Variation in X_i
4. Zero conditional mean of the errors
5. Homoskedasticity
6. Errors are conditionally Normal

Assumption 6: conditionally Normal errors

Assumption 6 - The conditional distribution of u given X is Normal with mean 0 and variance σ_u^2.

- This implies that the distribution of Y_i given X_i is: $N(\beta_0 + \beta_1 X_i, \sigma_u^2)$.

Sampling distribution of OLS slope

- If we have Y_i given X_i is distributed $N(\beta_0 + \beta_1 X_i, \sigma_u^2)$, then we have the following at any sample size:

$$\frac{\widehat{\beta}_1 - \beta_1}{SE[\widehat{\beta}_1]} \sim N(0,1)$$

- Furthermore, if we replace the true standard error with the estimated standard error, then we get the following:

$$\frac{\widehat{\beta}_1 - \beta_1}{\widehat{SE}[\widehat{\beta}_1]} \sim t_{n-2}$$

- The standardized coefficient follows a t distribution $n-2$ degrees of freedom. We take off an extra degree of freedom because we had to one more parameter than just the sample mean.
- All of this depends on Normal errors! We can check to see if the error do look Normal.

HYPOTHESIS TESTS FOR REGRESSION

Null and alternative hypotheses review

- Null: $H_0 : \beta_1 = 0$
  - The null is the straw man we want to knock down.
  - With regression, almost always null of no relationship
- Alternative: $H_a : \beta_1 \neq 0$
  - Claim we want to test
  - Almost always "some effect"
  - Could do one-sided test, but you shouldn't, for reasons we've already discussed

- Notice these are statements about the population parameters, not the OLS estimates. ## Test statistic
- Under the null of $H_0 : \beta_1 = c$, we can use the following familiar test statistic:

$$T = \frac{\widehat{\beta}_1 - c}{\widehat{SE}[\widehat{\beta}_1]}$$

- As we saw in the last section, if the errors are conditionally Normal, then under the null hypothesis we have:

$$T \sim t_{n-2}$$

- In large samples, we know that T is approximately (standard) Normal, but we also know that t_{n-2} is approximately (standard) Normal in large samples too, so this statement works there too, even if Normality of the errors fails.
- Thus, under the null, we know the distribution of T and can use that to formulate a rejection region and calculate p-values.

Rejection region

- Choose a level of the test, α, and find rejection regions that correspond to that value under the null distribution:

$$\mathbb{P}(-t_{\alpha/2,n-2} < T < t_{\alpha/2,n-2}) = 1 - \alpha$$

- This is exactly the same as with sample means and sample differences in means, except that the degrees of freedom on the t distribution have changed.

p-value

- The interpretation of the p-value is the same: *the probability of seeing a test statistic at least this extreme if the null hypothesis were true*
- Mathematically:

$$\mathbb{P}\left(\left|\frac{\widehat{\beta}_1 - c}{\widehat{SE}[\widehat{\beta}_1]}\right| \geq |T_{obs}|\right)$$

- If the p-value is less than α we would reject the null at the α level.

R output

- By default, R shows you the T_{obs} for the test statistic with the null of $\beta_1 = 0$, which is just the estimate divided by the standard error:

$$T_{obs} = \frac{\widehat{\beta}_1 - 0}{\widehat{SE}[\widehat{\beta}_1]} = \frac{\widehat{\beta}_1}{\widehat{SE}[\widehat{\beta}_1]}$$

- R also calculates the p-values for you.
- In the AJR data:

```
out <- lm(logpgp95 ~ logem4, data = ajr)
coef(summary(out))
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 10.6602465 0.30528441 34.919066 8.758878e-50
## logem4      -0.5641215 0.06389003 -8.829569 2.093611e-13
```

- We could have calculated that directly:

```
r  -0.5641/0.06389
## [1] -8.829238
```

CONFIDENCE INTERVALS FOR REGRESSION

Confidence intervals

- Very similar to the approach with sample means. By the sampling distribution of the OLS estimator, we know that we can find t-values such that:

$$\mathbb{P}\left(- t_{\alpha/2,n-2} \leq \frac{\widehat{\beta}_1 - \beta_1}{\widehat{SE}[\widehat{\beta}_1]} \leq t_{\alpha/2,n-2} \right) = 1 - \alpha$$

- If we rearrange this as before, we can get an expression for confidence intervals:

$$\mathbb{P}\left(\widehat{\beta}_1 - t_{\alpha/2,n-2}\widehat{SE}[\widehat{\beta}_1] \leq \beta_1 \leq \widehat{\beta}_1 t_{\alpha/2,n-2}\widehat{SE}[\widehat{\beta}_1] \right) = 1 - \alpha$$

- Thus, we can write the confidence intervals as:

$$\widehat{\beta}_1 \pm t_{\alpha/2,n-2}\widehat{SE}[\widehat{\beta}_1]$$

- We can derive these for the intercept as well:

$$\widehat{\beta}_0 \pm t_{\alpha/2,n-2}\widehat{SE}[\widehat{\beta}_0]$$

Confidence intervals in R

- Confidence intervals are not outputted by default, but you grab them for any regression using the `confint()` function:

```r
confint(lm(logpgp95 ~ logem4, data = ajr))
```

```
##                  2.5 %     97.5 %
## (Intercept) 10.0525931 11.2678999
## logem4      -0.6912914 -0.4369515
```

REVIEW OF ASSUMPTIONS

- What assumptions do we need to make what claims with OLS?

  1. Data description: variation in X
  2. Unbiasedness/Consistency: linearity, iid, variation in X, zero conditional mean error.
  3. Large-sample inference: linearity, iid, variation in X, zero conditional mean error, homoskedasticity.
  4. Small-sample inference: linearity, iid, variation in X, zero conditional mean error, homoskedasticity, Normal errors.

- Can we weaken these? In some cases, yes. Bootstrap to weaken homoskedasticity, for instance.

GOODNESS OF FIT

Prediction error

- How do we judge how well a line fits the data? Is there some way to judge?
- One way is to find out how much better we do at predicting Y once we include X into the regression model.
- Prediction errors without X: best prediction is the mean, so our squared errors, or the **total sum of squares** (SS_{tot}) would be:

$$SS_{tot} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

- Once we have estimated our model, we have new prediction errors, which are just the sum of the squared residuals or SS_{res}:

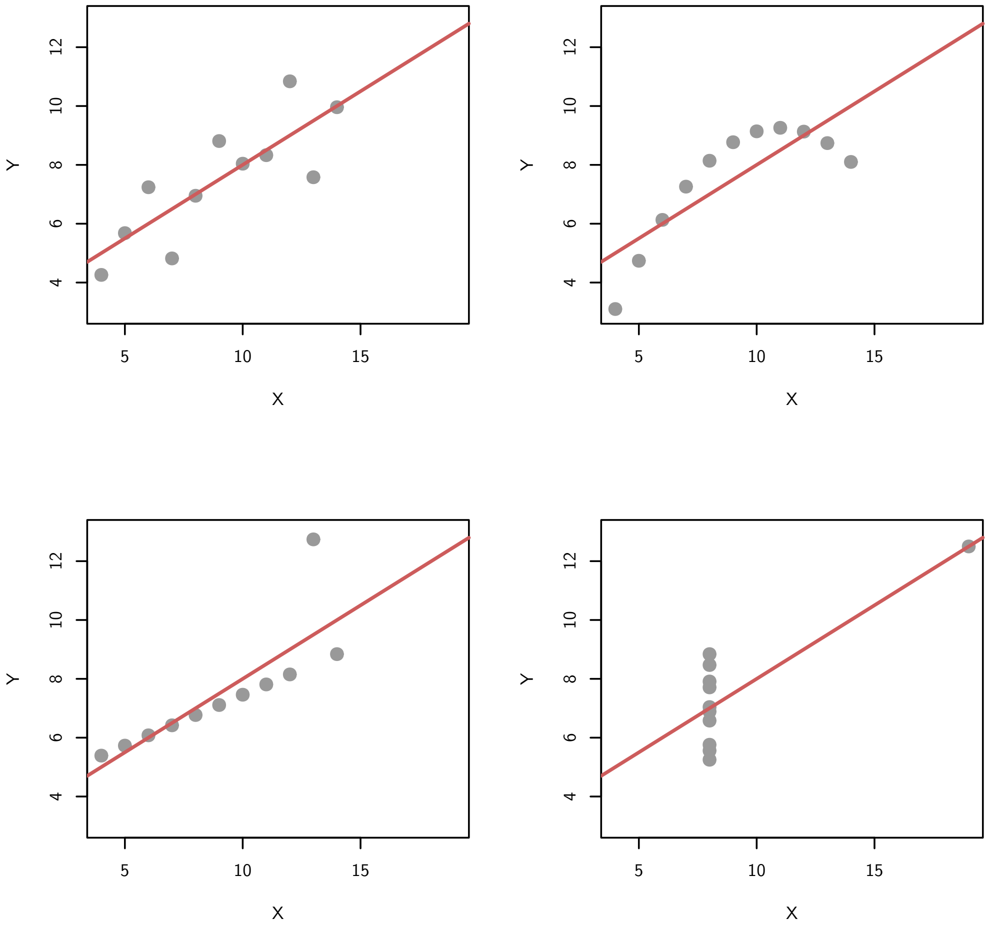
$$SS_{res} = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$$

R-square

- By definition, the residuals have to be smaller than the deviations from the mean, so we might ask the following: how much lower is the SS_{res} compared to the SS_{tot}?
- We quantify this question with the **coefficient of determination** or R^2. This is the following:
$$R^2 = \frac{SS_{tot} - SS_{res}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}}$$
- This is the fraction of the total prediction error eliminated by providing information on X.
- Alternatively, this is the fraction of the variation in Y is "explained by" X.
- $R^2 = 0$ means no relationship
- $R^2 = 1$ implies perfect linear fit

Is R-squared useful?

- Can be very misleading. Each of these samples have the same R^2 even though they are vastly different:



APPENDIX

Proof of sums and means trick

- In the derivation of the OLS estimator, we relied on a trick with the means and sums. Here is the proof:

$$\begin{aligned}
\sum_{i=1}^{n} X_i(Y_i - \overline{Y}) &= \sum_{i=1}^{n} X_i(Y_i - \overline{Y}) - n\overline{XY} + n\overline{XY} \\
&= \sum_{i=1}^{n} X_i(Y_i - \overline{Y}) - \overline{X}\left(\sum_{i=1}^{n} Y_i\right) + \overline{X}\left(\sum_{i=1}^{n} \overline{Y}\right) \\
&= \sum_{i=1}^{n} X_i(Y_i - \overline{Y}) - \overline{X}\left(\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \overline{Y}\right) \\
&= \sum_{i=1}^{n} X_i(Y_i - \overline{Y}) - \overline{X}\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right) \\
&= \sum_{i=1}^{n} X_i(Y_i - \overline{Y}) - \sum_{i=1}^{n} \overline{X}\left(Y_i - \overline{Y}\right) \\
&= \sum_{i=1}^{n} \left[X_i(Y_i - \overline{Y}) - \overline{X}\left(Y_i - \overline{Y}\right)\right] \\
&= \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})
\end{aligned}$$

- Replace $(Y_i - \overline{Y})$ with $(X_i - \overline{X})$ to prove that

$$\sum_{i=1}^{n} X_i(X_i - \overline{X}) = \sum_{i=1}^{n} (X_i - \overline{X})^2$$

Proof of OLS Consistency

- How can we prove this? Well, we can use the

$$\begin{aligned}
\widehat{\beta}_1 &= \sum_{i=1}^{n} W_i Y_i \\
&= \sum_{i=1}^{n} W_i(\beta_0 + \beta_1 X_i + u_i) \\
&= \beta_0 \sum_{i=1}^{n} W_i + \beta_1 \sum_{i=1}^{n} W_i X_i + \sum_{i=1}^{n} W_i u_i \\
&= \beta_1 + \sum_{i=1}^{n} W_i u_i \\
&= \beta_1 + \frac{\sum_{i=1}^{n}(X_i - \overline{X})u_i}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \\
&= \beta_1 + \frac{1/n \sum_{i=1}^{n}(X_i - \overline{X})(u_i - \overline{u})}{1/n \sum_{i=1}^{n}(X_i - \overline{X})^2}
\end{aligned}$$

- Note that, by the law of large numbers:

$$\begin{aligned}
1/n \sum_{i=1}^{n}(X_i - \overline{X})(u_i - \overline{u}) &\xrightarrow{p} \text{cov}(X, u) \\
1/n \sum_{i=1}^{n}(X_i - \overline{X})^2 &\xrightarrow{p} \mathbb{V}(X)
\end{aligned}$$

- With these facts in hand and relying on the properties of convergence in probability (see Wooldridge PLIM property 2 in Appendix C):

$$\begin{aligned}
\widehat{\beta}_1 &\xrightarrow{p} \beta_1 + \frac{\text{cov}(X, u)}{\mathbb{V}(X)} \\
&= \beta_1 + \frac{0}{\mathbb{V}(X)} \qquad \text{(A4: zero conditional mean of error)} \\
&= \beta_1 \qquad \text{(A3: variation in IV)}
\end{aligned}$$