Simple Linear Regression I

Econ UA 266: Intro to Econometrics

Linear regression model

Policy questions in economics are about cause and effect:

> *If I change policy variable X (e.g. class size, tax rates, Fed funds rate) by some amount, how will outcome variable Y (e.g. test scores, GDP, inflation) change?*

Suppose the effect is constant:

$$\beta_{CS} = \frac{dTestScore}{dClassSize}$$

Then:

$$TestScore = \beta_0 + \beta_{CS} ClassSize$$

Linear regression model

Write as an econometric model:

$$TestScore_i = \beta_0 + \beta_{CS} ClassSize_i + u_i$$

where:

- subscript i denotes observations (school districts)
- error term u_i is a random variable that represents all other factors that determined test scores in school district i that aren't accounted for in this linear relationship, including:
  - everything that might explain test scores besides class size (e.g. parental resources, teacher quality, test conditions, etc)
  - misspecification error: true relationship between average test scores and class size may not be linear (more on this later...)

Linear regression model and causality

When we run the regression

$$TestScore_i = \beta_0 + \beta_{CS} ClassSize_i + u_i$$

we're merely making a statement about correlation

What $\beta_{CS} < 0$ tells us: class size is negatively correlated with test scores

- could be a confounding variable that's the reason for us observing this correlation
    - e.g. teacher quality or parental resources

What $\beta_{CS} < 0$ does not tell us: reducing class size by 1 increases average test scores by $|\beta_{CS}|$

Linear regression model and causality

Nevertheless, linear regression is used for prediction

- it gives us the best linear predictor of Y given X
  - draw another observation at random from the population
  - regression gives us a "best" predictor of Y_{n+1} given X_{n+1}
  - but, the prediction error may be large (compare $s_{\hat{u}}$ to s_Y)

So, knowing X may help us to predict Y, but this is different from saying that X causes Y

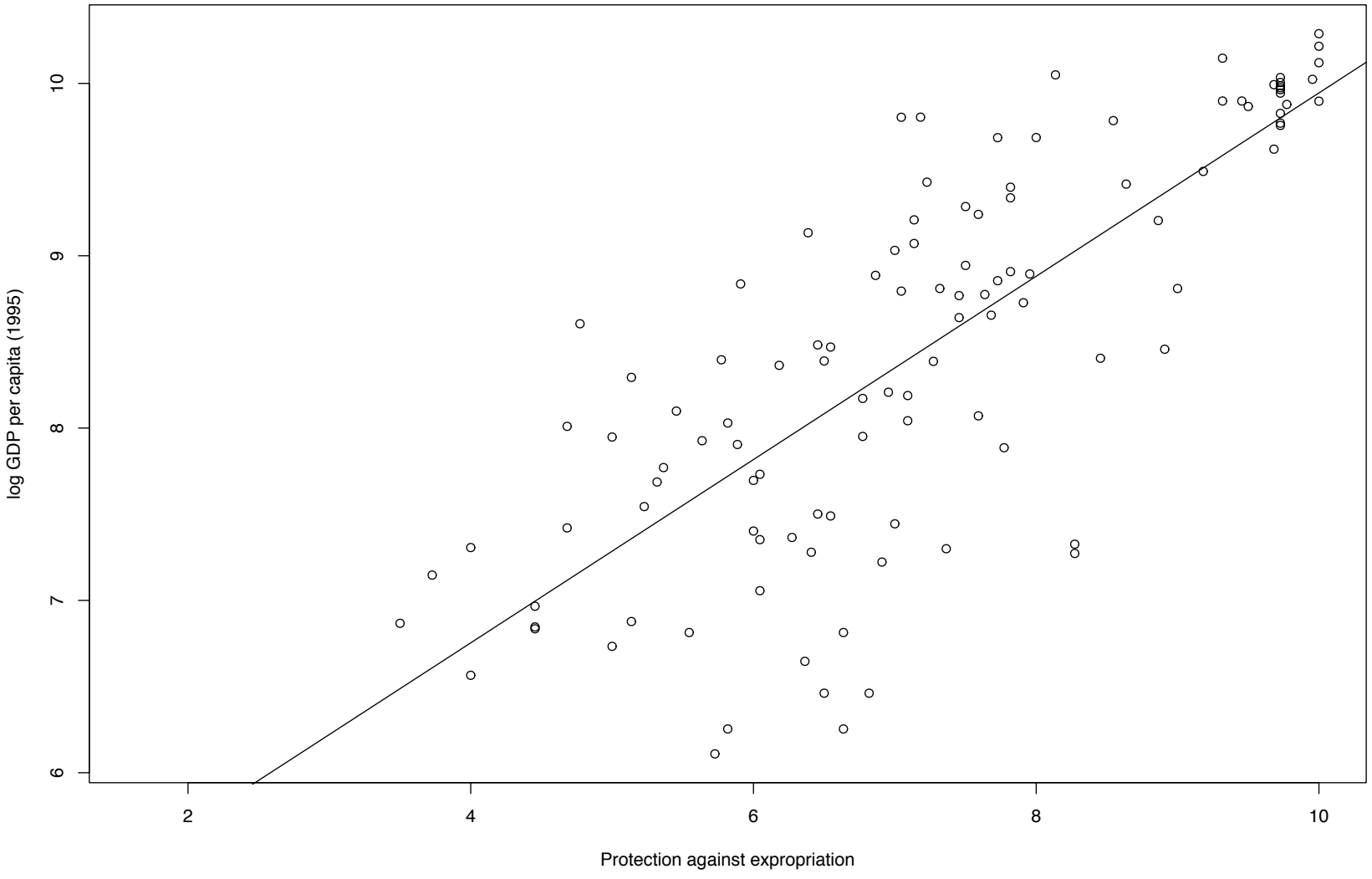
- to make a statement about causality, we need to know something about how identical individuals react under different policy settings
- we can do this with randomized experiments
- but we generally can't do this with observational data, unless:
  - we add all other variables to the regression that might explain Y aside from X (we need multiple regression for this)
  - and we assume that, once we condition on these variables, X is as good as randomly assigned

Example: Acemoglu, Johnson, Robinson (2001)

- On relationship between "institutions" and economic development
- GDP per capita in each country in 1995 (*logpgp95*)
- Regress on a measure of protection against expropriation (*avexpr*)

```
# load package to read stata data files
require(foreign)
# import ajr data
ajr <- read.dta("maketable2.dta")
# regress log gdp on protection against expropriation
z <- lm(logpgp95~avexpr,data=ajr)
# print summary
summary(z)
# plot the data and line of best fit
plot(logpgp95~avexpr,data=ajr,
    xlab="Protection against expropriation",
    ylab="log GDP per capita (1995)")
abline(z)
```

Example: Acemoglu, Johnson, Robinson (2001)



Example: Acemoglu, Johnson, Robinson (2001)

```
lm(formula = logpgp95 ~ avexpr, data = ajr)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9020 -0.3160  0.1380  0.4225  1.4406

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.62609    0.30058   15.39   <2e-16 ***
avexpr       0.53187    0.04062   13.09   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7179 on 109 degrees of freedom
  (52 observations deleted due to missingness)
Multiple R-squared:  0.6113,Adjusted R-squared:  0.6078
F-statistic: 171.4 on 1 and 109 DF,  p-value: < 2.2e-16
```

Example: Acemoglu, Johnson, Robinson (2001)

We estimate:

$$\widehat{logpgp95}_i = 4.62 + 0.53 avexpr_i$$

What this tells us: log gdp per capita and property rights are positively correlated

What this does not tell us: property rights *cause* gdp per capita to be higher

- could be a confounding variable
  - maybe democracy is the real cause of higher gdp, and that's also correlated with strong property rights
- could be reverse causality

Linear regression model

In general terms:

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

- Y_i = dependent variable
- X_i = independent variable/regressor
- u_i = error term

Population parameters:

- β_1 = slope coefficient (policy-relevant parameter)
- β_0 = intercept

Regression line

Regression line: relationship between Y and X that holds on average across the population (according to the model):

$$\mathrm{E}[Y|X = x] = \beta_0 + \beta_1 x$$

- model for the expected value of Y when we know the value of X
- β_1: change in the expected value of Y for a unit increase in X

$$\beta_1 = \frac{d\mathrm{E}[Y|X = x]}{dx} = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}$$

  - measured in units of Y per unit of X
  - not causal: just a statement about what we observe in the data
- β_0: expected value of Y when $X = 0$
  - may not make sense in some applications

Regression line and misspecification

True conditional mean of Y given $X = x$ need not be linear

In general:

$$\mathrm{E}[Y|X = x] = m(x)$$

for some function m

- If the linear regression model is correctly specified (i.e. $m(x)$ is a linear function of x), then $m(x) = \beta_0 + \beta_1 x$
- If the linear regression model is misspecified (i.e. $m(x)$ is *not* a linear function of x), then

  $$\beta_0 + \beta_1 x$$

  is the best approximation to $m(x)$ in the space of linear functions
- this is the best linear predictor property

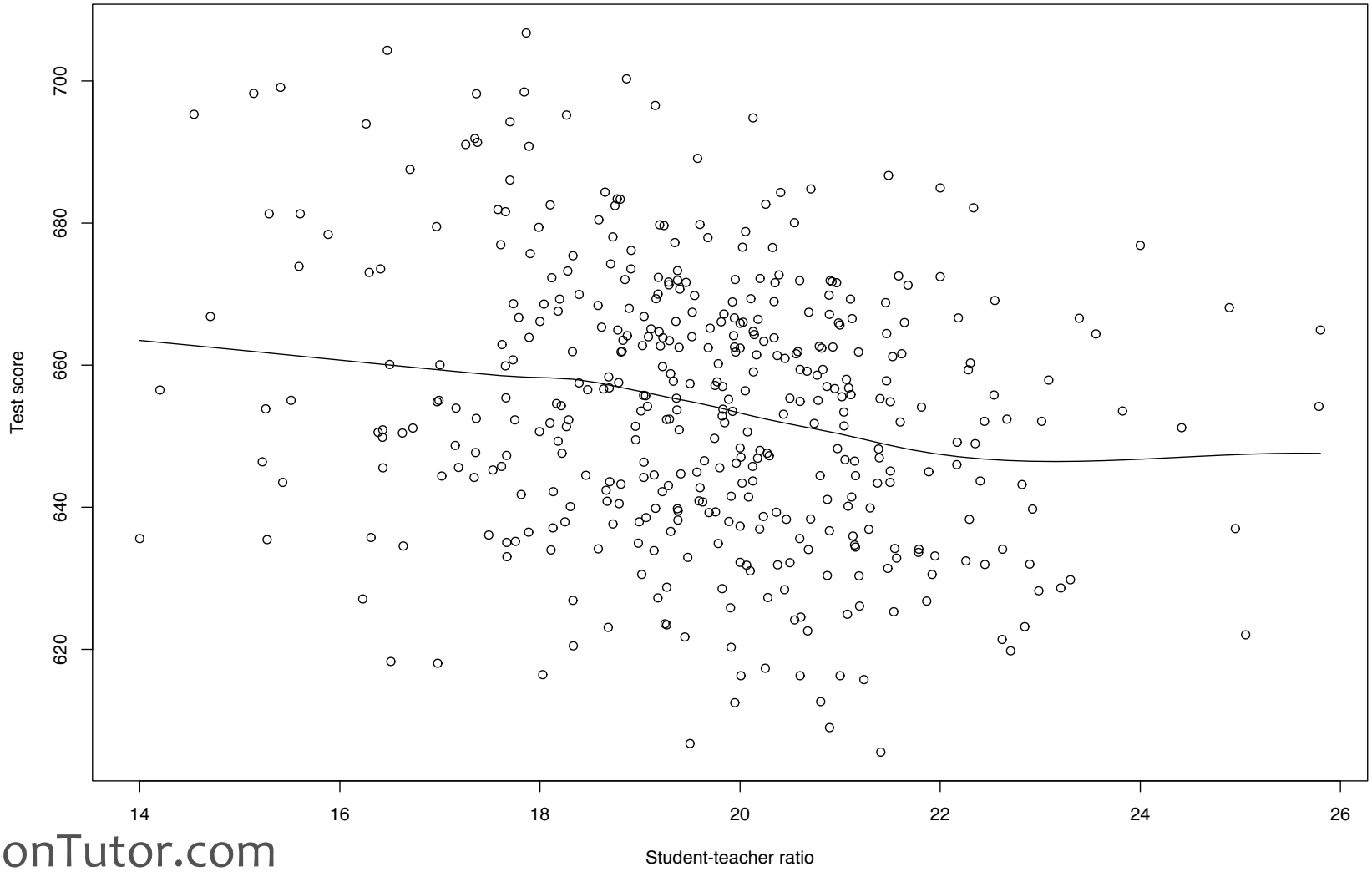
California schools data

- Data on 420 schools in California from 1999
- Avg reading and maths scores for 5th graders
- School characteristics:
  - enrollment
  - number of teachers
  - number of computers per classroom
  - expenditures per student
  - income
- Demographic variables:
  - % of students in CalWorks (welfare program)
  - % of students that qualify for reduced price lunch
  - % of students that are English learners

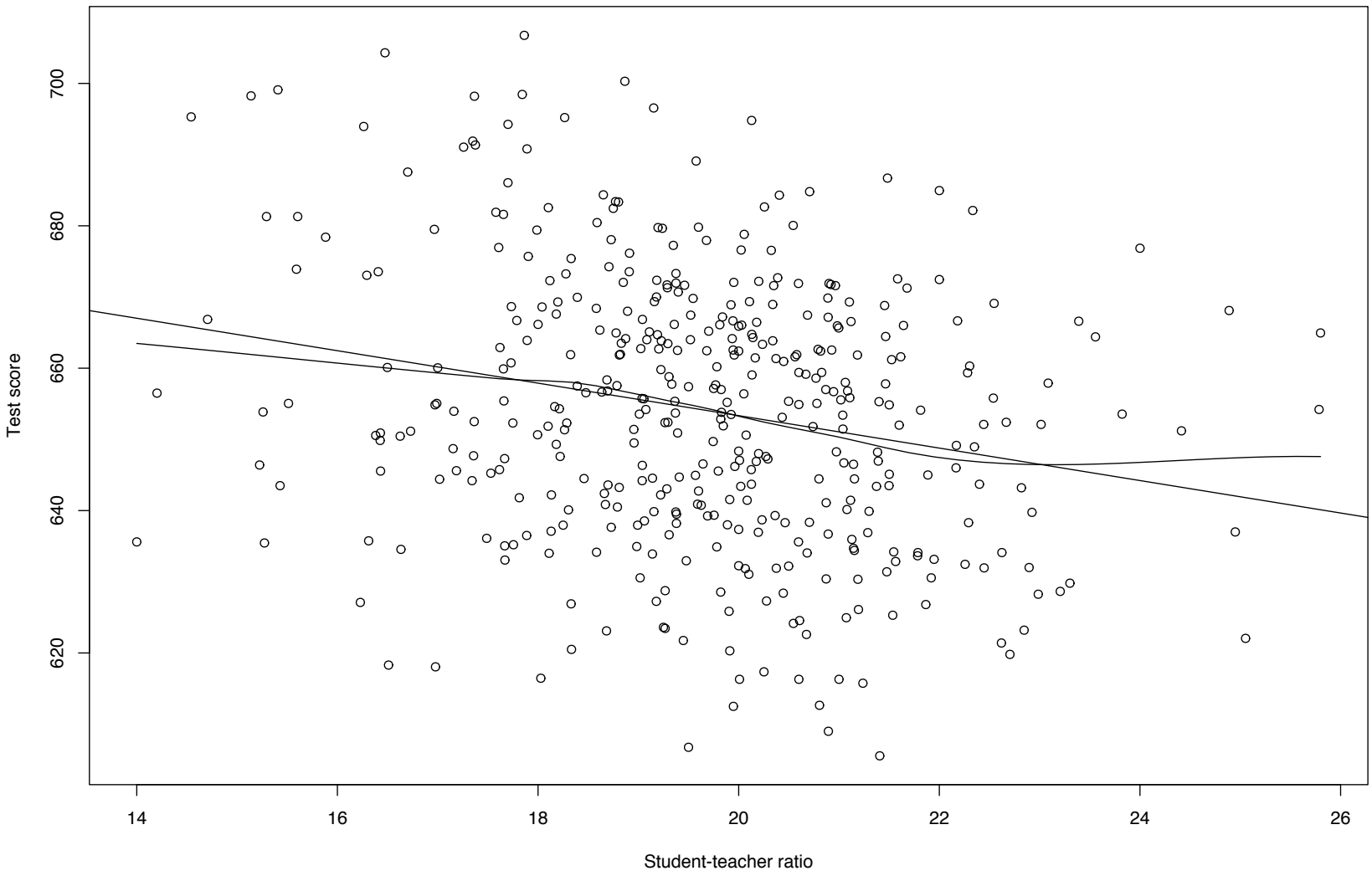
California schools data

```
# import californian school data
ca <- read.csv("Caschool.csv")

# plot the data and a nonparametric estimate of
    the true regression function
with(ca,scatter.smooth(str,testscr,
    xlab="Student-teacher ratio",ylab="Test score"))

# estimate linear regression model
z <- lm(testscr~str,data=ca)

# add the regression line to the plot
abline(z)

# print coefficients
summary(z)
```

California schools data



California schools data



California schools data

```
Call:
lm(formula = testscr ~ str, data = ca)

Residuals:
    Min      1Q  Median      3Q     Max
-47.727 -14.251   0.483  12.822  48.540

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 698.9330     9.4675  73.825  < 2e-16 ***
str          -2.2798     0.4798  -4.751 2.78e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.58 on 418 degrees of freedom
Multiple R-squared:  0.05124,Adjusted R-squared:  0.04897
F-statistic: 22.58 on 1 and 418 DF,  p-value: 2.783e-06
```

California schools data: interpreting the output

The OLS estimates are:

$$\hat{\beta}_0 = 698.9 \qquad \hat{\beta}_1 = -2.28$$

The OLS regression line is:

$$\widehat{TestScore} = 698.9 - 2.28STR$$

So for school i the model predicts:

$$\widehat{TestScore}_i = 698.9 - 2.28STR_i$$

which are called the fitted values

The residual is the difference:

$$\hat{u}_i = TestScore_i - \widehat{TestScore}_i$$

California schools data: interpreting the output

```
Residuals:
    Min      1Q  Median      3Q     Max
-47.727 -14.251   0.483  12.822  48.540
```

- Smallest, largest, 25th, 50th, and 75th percentiles of the residuals

```
Residual standard error: 18.58
```

- called standard error of the regression (SER)
- estimate of the standard deviation of u_i
- formula:

$$SER = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2}$$

- measure of fit: tells us how wide the data points are dispersed around the regression line

California schools data: interpreting the output

```
Multiple R-squared:  0.05124,Adjusted R-squared:  0.04897
```

- regression R^2 is the proportion of sample variance in Y_i explained by the model
- formula:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = \hat{\rho}^2_{XY}$$

where

$$TSS = \sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2 \qquad ESS = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y}_n)^2$$

$$RSS = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} \hat{u}_i^2$$

- $R^2 \approx 1$ indicates X is a good predictor ($R^2 = 1$ if there is a perfect linear relationship between X_i and Y_i)
- $R^2 \approx 0$ indicates X is a poor predictor

California schools data: interpreting the output

```
# verify residuals have mean zero
mean(z$residuals)

# compute residual standard error
RSS <- sum(z$residuals^2)
sqrt(RSS/418)

# compute R^2
TSS <- sum((ca$testscr-mean(ca$testscr))^2)
ESS <- sum((z$fitted.values-mean(ca$testscr))^2)
ESS/TSS
1-RSS/TSS
```

California schools data: interpreting the output

To summarize:

- $R^2 = 0.05$ indicates the model only explains about 5% of the variation in test scores
- to be expected: there are plenty of other factors (e.g. teacher quality, variation in student body) that may also explain variation in test scores
- $SER = 18.58$ tells us the standard deviation of the residuals is about 18.6 (measured in test score points)
- the standard deviation in test scores is 19.05
- so, any prediction we make about performance in a specific school district may be off by a large amount

The Big Picture

- We start from a model:

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

- we need to estimate it, i.e. to estimate β_0 and β_1
- an estimator is needed $\rightarrow$ Ordinary Least Squares (OLS)
- under certain assumptions, the OLS is a good estimator
- armed with the OLS estimator, we use the data to compute our estimates $\hat{\beta}_0$ and $\hat{\beta}_1$
- after estimation, we ask how well the estimated model fits the data (R^2 and standard error of the regression $s_{\hat{u}}$)
- we also try to quantify the uncertainty around our estimates

Least squares estimators: Idea

- suppose we predicted Y_i using $\hat{Y}_i = b_0 + b_1 X_i$
- linear prediction error is $Y_i - b_0 - b_1 X_i$
- a "good" estimator should minimize the average prediction errors
- so, we choose estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the sum of squared linear prediction errors:

$$S(b_0, b_1) = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

- solve for the estimators by taking FOCs

$$\frac{\partial S(\hat{\beta}_0, \hat{\beta}_1)}{\partial b_0} = 0 \quad \text{and} \quad \frac{\partial S(\hat{\beta}_0, \hat{\beta}_1)}{\partial b_0} = 0$$

Least squares estimators

- after some algebra, we see that the OLS estimators are:[1]

$$\hat{\beta}_1 = \frac{s_{XY}}{s_X^2}$$

$$\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{X}_n$$

- $\hat{\beta}_1$: estimate of the change in the expected value of Y for a unit increase in X
  - nothing causal: merely a statement about how X and Y covary in the data
  - whether we can interpret causally will depend on relationship between X_i and u_i
- $\hat{\beta}_0$: estimate of the expected value of Y when $X = 0$

[1] Detailed derivation posted on NYU Classes.

The linear regression model is:

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

The population regression line is:

$$E[Y|X] = \beta_0 + \beta_1 X$$

The sample or estimated regression line is:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

The predicted or forecasted or fitted value of Y_i given X_i is:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

The residuals are:

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

Properties of the residuals

Mean zero. FOC for $\hat{\beta}_0$ is:

$$0 = \frac{\partial S(\hat{\beta}_0, \hat{\beta}_1)}{\partial b_0} = -2 \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = -2 \sum_{i=1}^{n} u_i$$

Uncorrelated with X. FOC for $\hat{\beta}_1$ is:

$$0 = \frac{\partial S(\hat{\beta}_0, \hat{\beta}_1)}{\partial b_1} = -2 \sum_{i=1}^{n} X_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = -2 \sum_{i=1}^{n} X_i u_i$$

Measures of fit

R^2: fraction of sample variance of Y_i explained by the model

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

- always between 0 (explains nothing) and 1 (perfect fit)

Standard error of the regression: estimator of the standard deviation of the regression error u_i

$$SER = s_{\hat{u}}, \qquad s_{\hat{u}}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2 = \frac{RSS}{n-2}$$

- ideally as small as possible
- should compare it with standard deviation of Y_i

OLS Assumptions

1. Conditional distribution of u_i given X_i has mean zero: $\mathrm{E}[u_i|X_i] = 0$
   - this is the most controversial assumption—it may not hold!
   - in particular, this implies $\mathrm{Cov}(X_i, u_i) = 0$
   - so, we're effectively assuming that all of the "other factors" packed into u_i that explain Y_i are all uncorrelated with X_i
   - also assumes $m(x)$ is linear in x

2. $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ are a random sample (i.i.d.)
   - we will relax this when dealing with time-series data
   - e.g. macro variables, stock returns, etc

3. Large outliers are unlikely
   - formally, $0 < \mathrm{E}[X_i^4] < \infty$ and $0 < \mathrm{E}[Y_i^4] < \infty$
   - use this to justify applying LLN/CLT

Properties of OLS estimators

Under these assumptions:

1. $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased
2. $\hat{\beta}_0$ and $\hat{\beta}_1$ are consistent
3. The sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ are approximately $N(\beta_0, s.e.(\hat{\beta}_0)^2)$ and $N(\beta_1, s.e.(\hat{\beta}_1)^2)$; equivalently:

$$\frac{\hat{\beta}_0 - \beta_0}{s.e.(\hat{\beta}_0)} \qquad \frac{\hat{\beta}_1 - \beta_1}{s.e.(\hat{\beta}_1)}$$

are approximately distributed as $N(0,1)$

Properties of OLS estimators

Therefore, under assumptions $1-3$, OLS is a good estimator!

1. Unbiasedness $\rightarrow$ on average across samples, we estimate $\hat{\beta}_0 = \beta_0$ and $\hat{\beta}_1 = \beta_1$
2. Consistency $\rightarrow$ as sample grows larger, our estimates become closer and closer to β_0 and β_1
3. Sampling Distributions of $\hat{\beta}_0$ and $\hat{\beta}_1 \rightarrow$ we can quantify estimation uncertainty and test hypotheses