Review of Probability and Statistics

Juergen Meinecke

Roadmap

Bivariate Probability

- Covariance and Correlation

So far, we've only looked at a single random variable Y

Now we turn to studying relationships between two random variables X and Y

We are moving from univariate analysis to bivariate analysis

(Soon enough we will also do multivariate analysis)

We assume that we have available a random sample of observations on ordered pairs (X_i, Y_i) from an unknown population

We don't know the *marginal* population distributions of X_i and Y_i

We also don't know the *joint* population distributions of X_i and Y_i

Our focus in this section is to study parameters of the joint distribution

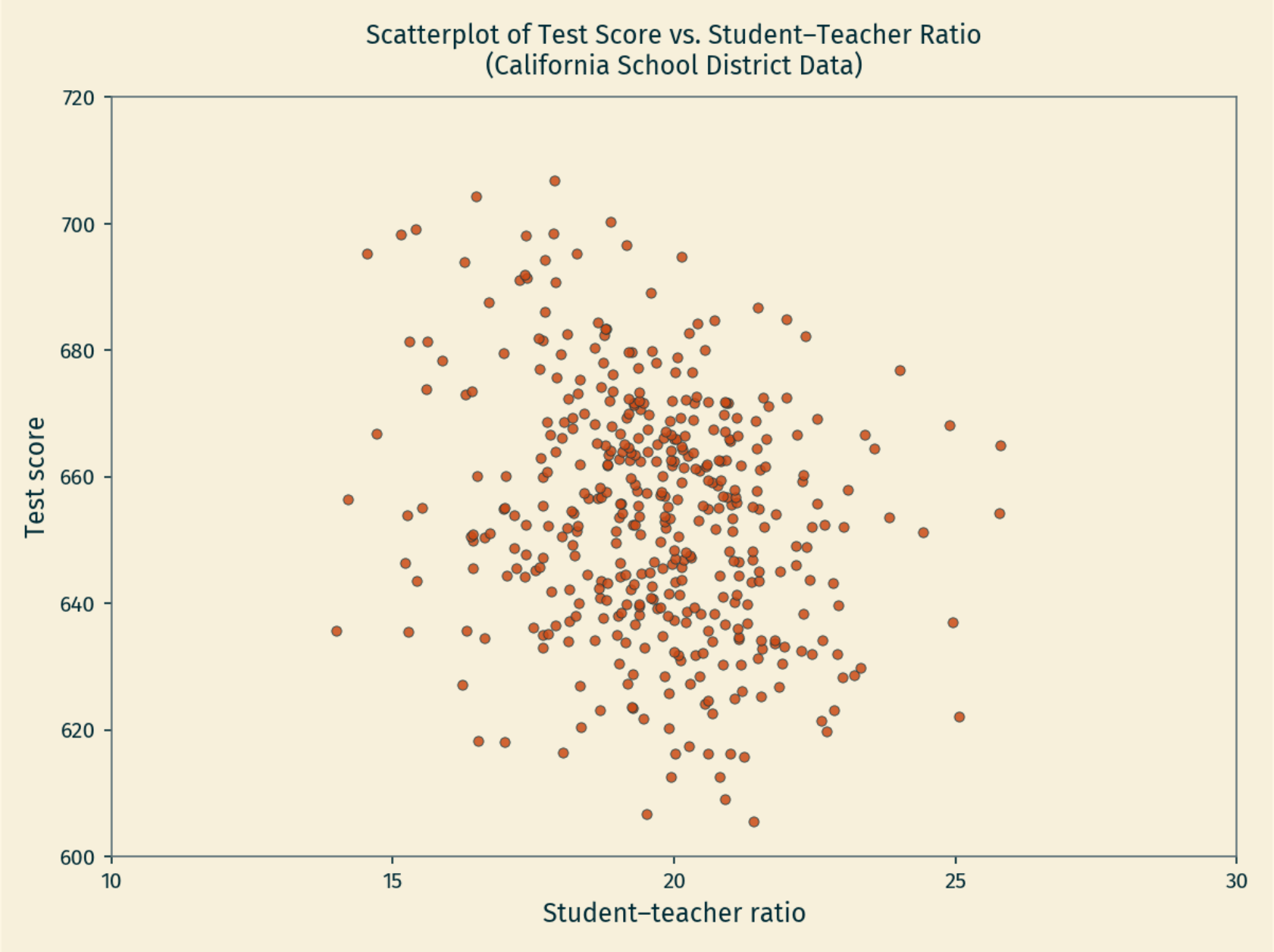
(We could, of course, use our knowledge from the previous two weeks to study X_i and Y_i separately; for example, we could calculate sample averages of both of them)

The easiest thing to do when studying the joint distribution is to take a look at the scatterplot

Let’s say we are interested in looking at the joint distribution of test scores and student-teacher ratios

The primitive idea is that schools with lower student-teacher ratios achieve better results in standardized test scores

Using data from 420 Californian school districts, the scatterplot looks like this...


Scatterplot of Test Score vs. Student–Teacher Ratio
(California School District Data)
Test score
Student–teacher ratio
720
700
680
660
640
620
600
10
15
20
25
30

In the univariate world, the main parameters of interest were the population mean and the population variance

We have leared that the sample average is the blue-stimator of the population mean

In the bivariate world, the main parameters of interest are the population covariance and the population correlation

We will now define the population parameters and their sample analogs

Definition

The **population covariance** between X and Y is defined by

$$\mathrm{Cov}(X, Y) := E\left[(X - E[X])(Y - E[Y])\right]$$

The population covariance is sometimes also denoted by σ_{XY}.

Definition

The **sample covariance** between X and Y is defined by

$$s_{XY} := \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

Definition

The **population correlation** between X and Y is defined by

$$\rho_{XY} := \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}$$

Definition

The **sample correlation** between X_i and Y_i is defined by

$$r_{XY} := \frac{s_{XY}}{s_X \cdot s_Y}$$

Do you remember what s_X and s_Y are?

Covariance and correlation measure joint variation in X and Y

If above average values of X tend to occur together with above average values of Y, then both covariance and correlation will be positive

If below average values of X tend to occur together with below average values of Y, then both covariance and correlation will be positive (not a typo)

If above average values of X tend to occur together with below average values of Y, then both covariance and correlation will be negative

If below average values of X tend to occur together with above average values of Y, then both covariance and correlation will be negative

What's the difference between covariance and correlation?

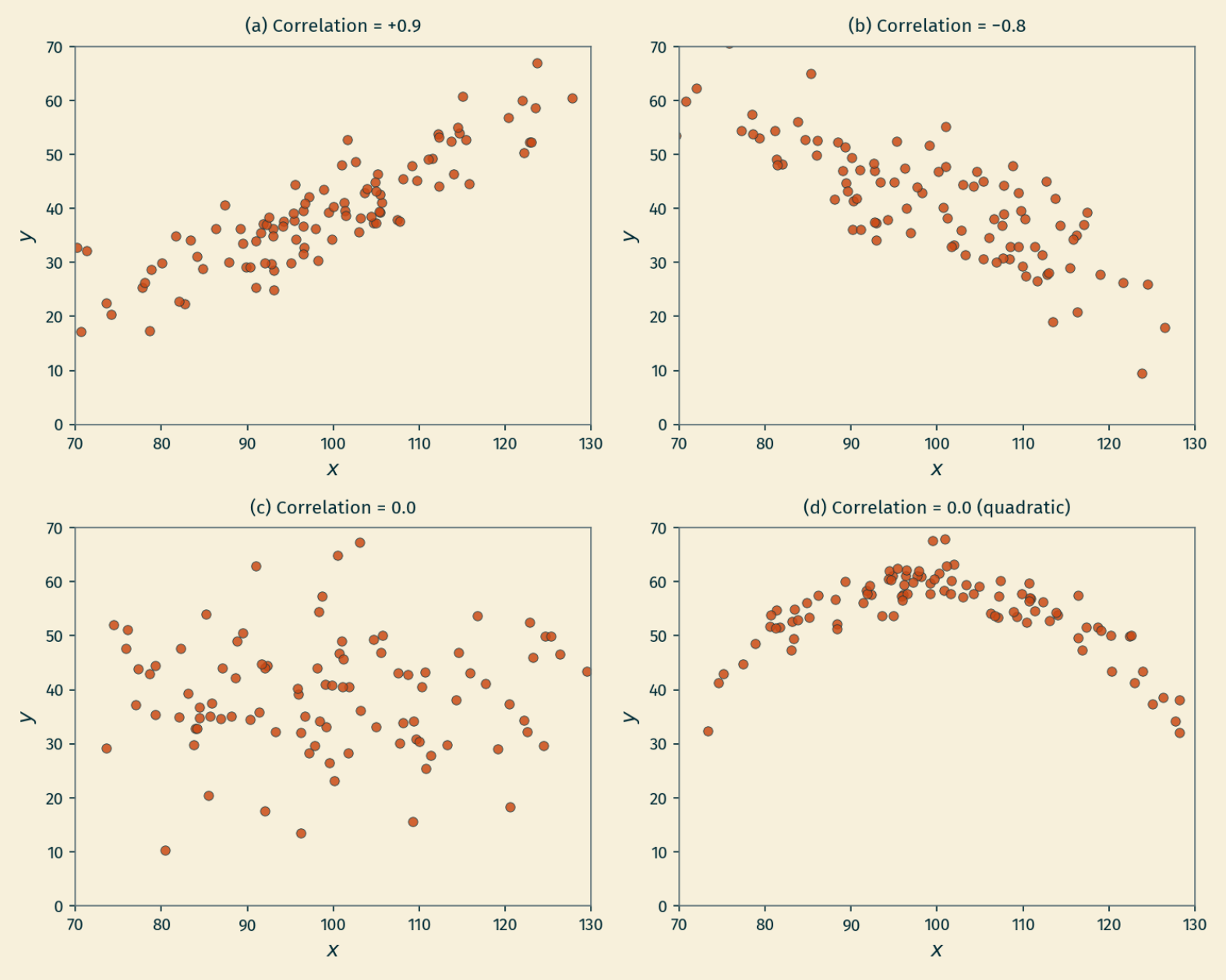
Covariance inherits the units of measurement of both X and Y, in fact it multiplies them

Hard to interpret what the unit of measurement of the product of X and Y are

Correlation, by dividing the covariance by the standard deviations, removes the units of measurement and instead results in a unit-free number that has to lie between -1 and 1

In that sense, the correlation is easier to interpret

Limitation: covariance and correlation only capture linear relationship between X and Y



Review of Probability and Statistics

Juergen Meinecke

Roadmap

Ordinary Least Squares Estimation

- Population Regression Function

We currently consider the statistical relationship between two variables X and Y

To demonstrate important concepts, let's focus on the example given by the textbook:

> *How do student-teacher ratios affect student performance?*

Before we get technical, let's first think about the content of that question; i.e., what is the hypothesized relationship b/w the two variables?

(It's always a good idea to first think about your econometric model intuitively)

The primitive argument goes like this:

The fewer students there are per teacher,
the more individualized instruction can be

Individualized instruction helps students

We would hypothesize that a lower student-teacher ratio, all else equal, should have a positive effect on student performance

It’s a simple argument, really

Now, turning technical, we believe there is a functional relationship between student-teacher ratio and student performance:

$$TestScore = f(STR, u),$$

where

- $TestScore$ captures student performance
- STR is the student-teacher ratio
- u captures all other things that explain $TestScore$ (over and above STR)
  Examples: intelligence, luck

In this example, the three different variables can be given generic names

- $TestScore$ is called the dependent variable
- STR is called the independent variable or explanatory variable or regressor
- u is called the error term

We hypothesize a negative relationship b/w $TestScore$ and STR, formally: $\partial TestScore / \partial STR < 0$

The equation

$$TestScore = f(STR, u),$$

together with the hypothesis

$$\frac{\partial TestScore}{\partial STR} < 0$$

summarize an *econometric model*
(a way an economist would think of a relationship between two variables statistically)

Digression: why do we need to include u?

If we did not include u as part of the function $f(\cdot)$ then we would presume that the relationship b/w $TestScore$ and STR was *deterministic*

It would mean that once we know STR we also know $TestScore$

This is almost like saying that they are one and the same thing

Deterministic relationships often make sense in the natural sciences, example: relationship b/w Celsius and Fahrenheit

In economics, relationships b/w variables are never deterministic but are subject to some degree of randomness and the presence of the *error term* u allows for that

So we think there is a functional relationship between $TestScore$ and STR

Problem: which function should $f(\cdot)$ be?

There are infinitely many possibilities!

There is no way of knowing

We will now make three important simplifying assumptions regarding the functional form of $f(\cdot)$

1. $f(X, u)$ is additively separable in X and u:

$$f(X, u) = g(X) + h(u)$$

2. $g(X)$ is a linear function:

$$g(X) = \beta_0 + \beta_1 X,$$

where β_0 and β_1 are called *coefficients* of the model

3. $h(u)$ is the simple identity function:

$$h(u) = u$$

Combining these three results in

$$f(X, u) = \beta_0 + \beta_1 X + u$$

The function $g(X) = \beta_0 + \beta_1 X$ plays an important role

It has its own name:

Definition

$g(X)$ is called the **population regression function (PRF)**.

The coefficients β_0 and β_1 are unknown

In a sense, we still don't really know the PRF because we do not know β_0 and β_1

But at least we know the principal class of the PRF
(it's linear in the coefficients)

But we will see soon that this is not so bad,
there is a good way of estimating the coefficients

Applying $f(X, u)$ to our example results in

$$TestScore = \beta_0 + \beta_1 STR + u,$$

Our hypothesis $\frac{\partial TestScore}{\partial STR} < 0$ from earlier simply translates to $\beta_1 < 0$

Our econometric model therefore is summarized in one line:

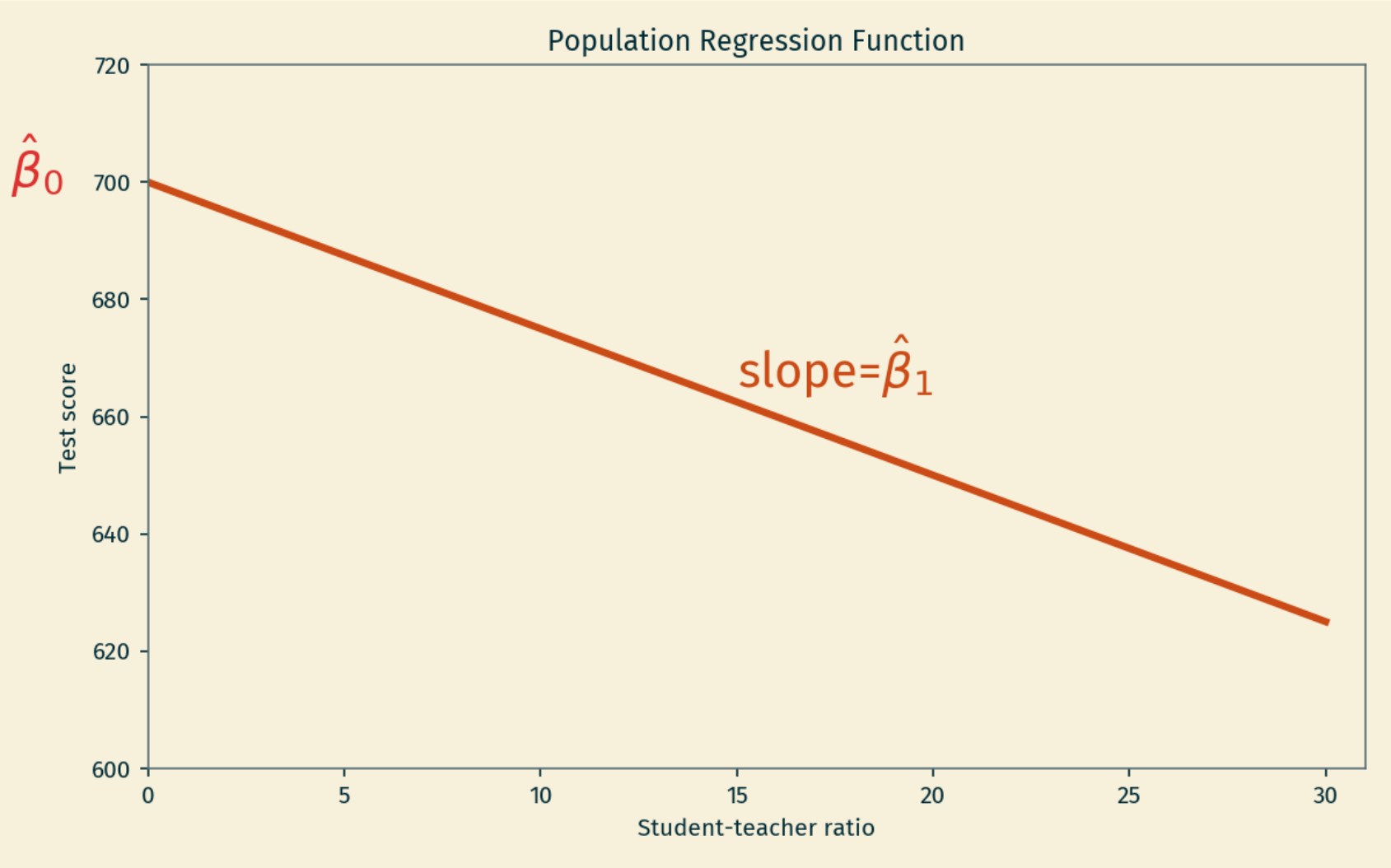
$TestScore = \beta_0 + \beta_1 STR + u$, with $\beta_1 < 0$

Our main obsession this semester will be to learn how to estimate the coefficients β_0 and β_1 and study the characteristics of these estimates

But let's suppose, for the sake of illustration, that an oracle was so kind to tell us that $\beta_0 = 700$ and $\beta_1 = -2.50$

What would we make of this information?

Let's graph the PRF



β_0 is the intercept, β_1 is the slope

A closer look at the PRF

Claim:
The PRF tells us the *expected* TestScore for a person with a given value of STR

More technically, this is called the *conditional expectation of TestScore given STR*

Let's prove that claim

Recall: $TestScore = \beta_0 + \beta_1 STR + u$

The expected $TestScore$ given STR is given by the conditional expectation $E[TestScore|STR]$

Plugging in the top equation and then solving results in

$$\begin{aligned} \mathsf{E}[TestScore|STR] &= \mathsf{E}[\beta_0 + \beta_1 STR + u|STR] \\ &= \beta_0 + \beta_1 \mathsf{E}[STR|STR] + \mathsf{E}[u|STR] \\ &= \beta_0 + \beta_1 STR + 0 \\ &= g(STR) \end{aligned}$$

The lhs is the expected $TestScore$ for a person with a given value of STR; the rhs is the PRF; therefore we have just proved the claim

In the previous calculations we have made one important assumption, namely

$$E[u|STR] = 0$$

This is the so-called **conditional mean independence assumption**

We now need to look at conditional mean independence from two angles:

1. mathematical
2. intuitive

The mathematical definition is straightforward

Definition

Two random variables X and Y are **conditionally mean independent (CMI)** if

$$E[Y|X] = E[Y].$$

Corollary

If X and Y are CMI then $E[X|Y] = E[X]$.

Corollary

If X and Y are CMI then $\sigma_{XY} = \rho_{XY} = 0$.

On the previous slide, X and Y played the role of generic random variables

Applying the CMI definition to u and STR would mean, strictly speaking, that we require

$$E[u|STR] = E[u]$$

in order for u and STR to be CMI

But we wrote earlier that $E[u|STR] = 0$

Reconciliation: without loss of generality we set $E[u] = 0$

Offering intuition for CMI

Example: average daytime maximum temperature

- in Canberra throughout the year is 20 degrees Celsius
- in Canberra in summmer is 27 degrees Celsius
- in Canberra in winter is 13 degrees Celsius

Mathematically:

$$\mathsf{E}[Temp] = 20 \quad \mathsf{E}[Temp|summer] = 27 \quad \mathsf{E}[Temp|winter] = 13$$

and therefore

$$\mathsf{E}[Temp] \neq \mathsf{E}[Temp|summer] \quad \text{and} \quad \mathsf{E}[Temp] \neq \mathsf{E}[Temp|winter]$$

Temperature is not mean independent of season

More examples:

- E[Intelligence|Education]
- E[Intelligence|identify as male]
- E[Height|identify as female]

Which ones are conditionally mean independent?

Intuitively, when are u and STR CMI?

In analogy to the previous examples, they would be CMI if STR is not predictive of u

It seems obvious that if u were a true random error, then STR is not predictive of u and they would therefore be CMI

Remember, earlier we said that u may be capturing things such as intelligence or luck

Do you believe that STR is predictive of these?

Let's say I told you that when I went to high-school (many many years ago) I had 28 class mates (including myself)

Then the econometric model would assign the following *expected* $TestScore$ for me:

$$Testscore = 700 - 2.5 \cdot 28 = 630$$

This is merely the *expected* $TestScore$

Actual $TestScores$ for individual students most likely differ

In any class of 28 students, some students will do better and some will do worse for various reasons unrelated to the student-teacher ratio

In other words, it is not deterministic that every student in a class of 28 will receive a test score of 630 (obvious, isn't it!?)

The role of the error term u is to bring in some randomness across students

Actual $TestScore$ (as opposed to expected) is as follows:

$$TestScore = 700 - 2.5 \cdot 28 + u = 630 + u$$

The way to think about it is this:

- Each student in my class of 28 students starts out on a basis of 630 (that is explained through the effect of student-teacher ratio on $TestScore$)
- On top of that each student has her/his own error term u that shifts her/him further up or down

Consider three of my class mates

Student	STR	u
A	28	20
B	28	-15
C	28	0

How would we compare students A and B?

- Student A has a test score of 650, student B of 615
- Reasons: student A is cleverer and/or is a better test taker and/or was more lucky on test taking day
- These reasons are unrelated to STR and influence $TestScore$ as well

What important role does student C play here?

Student C is the average student, with a score of 630

We see that the econometric model can make predictions for the *average student*

In the absence of any other influences on $TestScore$, the model predicts a score of 630

This offers us a new view of the function $f(\cdot)$

When we write the model as $TestScore = \beta_0 + \beta_1 STR + u$
we make an assumption about how we think about the average student

Put differently we have an econometric model about the *expected* test score:

$$\mathrm{E}[TestScore|STR] = \beta_0 + \beta_1 STR$$

So we are not actually studying

- how STR determines $TestScore$
  but instead
- how STR determines *expected* $TestScore$

This is an important difference!