POLS 205
Political Science as a Social Science

Analyzing Bivariate Relationships

Christopher Adolph

University of Washington, Seattle

May 17, 2010

Motivating Example

We have cross-national data from several sources:

Fertility The average number of children born per adult female, in 2000 (United Nations)

Education Ratio The ratio of girls to boys in primary and secondary education, in 2000 (Word Bank Development Indicators)

GDP per capita Economic activity in thousands of dollars, purchasing power parity in 2000 (Penn World Tables)

What are the levels of measurement of these variables?

Our question: how are these variables related to each other?

Motivating Example: Fertility, Female Education, and Development

Specifically, we ask:

Motivating Example: Fertility, Female Education, and Development

Specifically, we ask:

- If the level of female education changed by a certain amount, how much would we expect Fertility to change?

Motivating Example: Fertility, Female Education, and Development

Specifically, we ask:

- If the level of female education changed by a certain amount, how much would we expect Fertility to change?
- If the level of GDP per capita changed by a certain amount, how much would we expect Fertility to change?

Motivating Example: Fertility, Female Education, and Development

Specifically, we ask:

- If the level of female education changed by a certain amount, how much would we expect Fertility to change?
- If the level of GDP per capita changed by a certain amount, how much would we expect Fertility to change?
- How much would we expect our predictions to be off because of other random factors (noise)?

Motivating Example: Fertility, Female Education, and Development

Specifically, we ask:

- If the level of female education changed by a certain amount, how much would we expect Fertility to change?
- If the level of GDP per capita changed by a certain amount, how much would we expect Fertility to change?
- How much would we expect our predictions to be off because of other random factors (noise)?
- How much would we expect our predictions to be off because of sampling variability (poor estimation)?

Answering these questions will go far towards towards answering hypotheses about relationships between variables

Outline

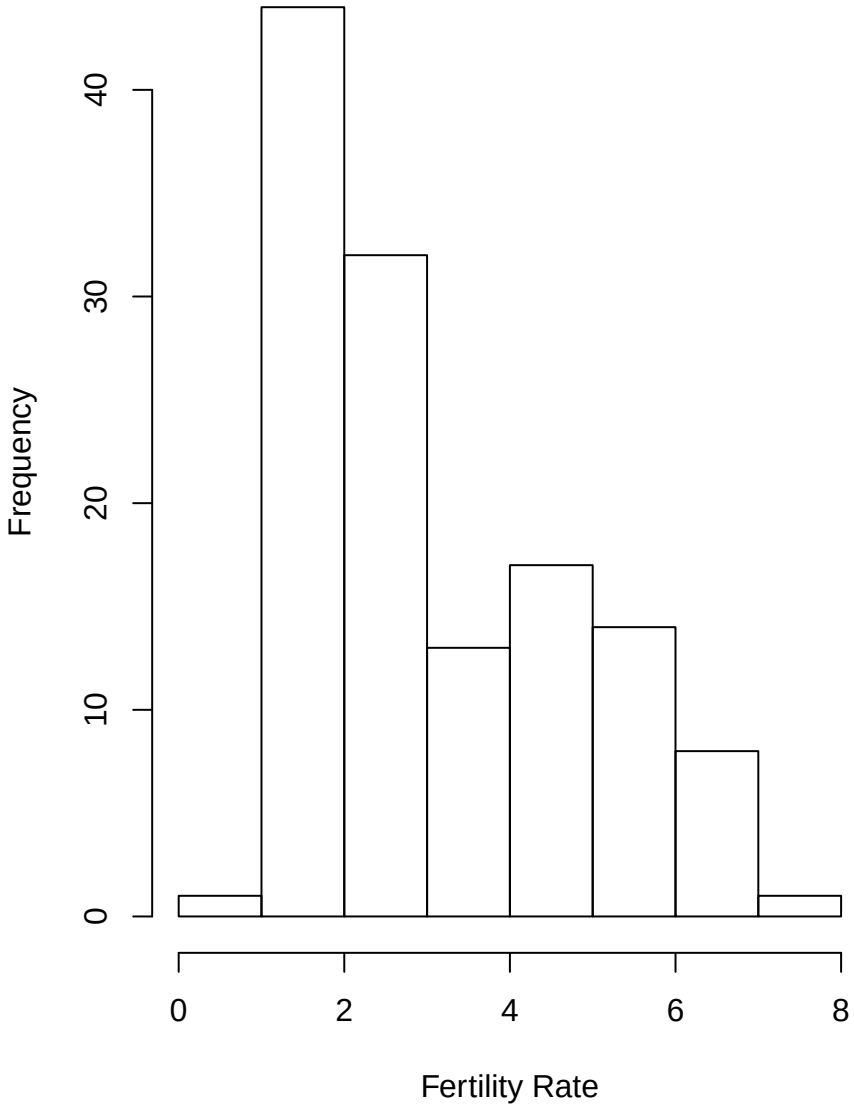
Review the Univariate Summary Statistics for our Example

Explore the Bivariate Relationship between Fertility & Education Ratio

Explore the Bivariate Relationship between Fertility & GDP per capita

Throughout, develop an understanding of *linear regression*

Summary of Univariate Distribution: Fertility

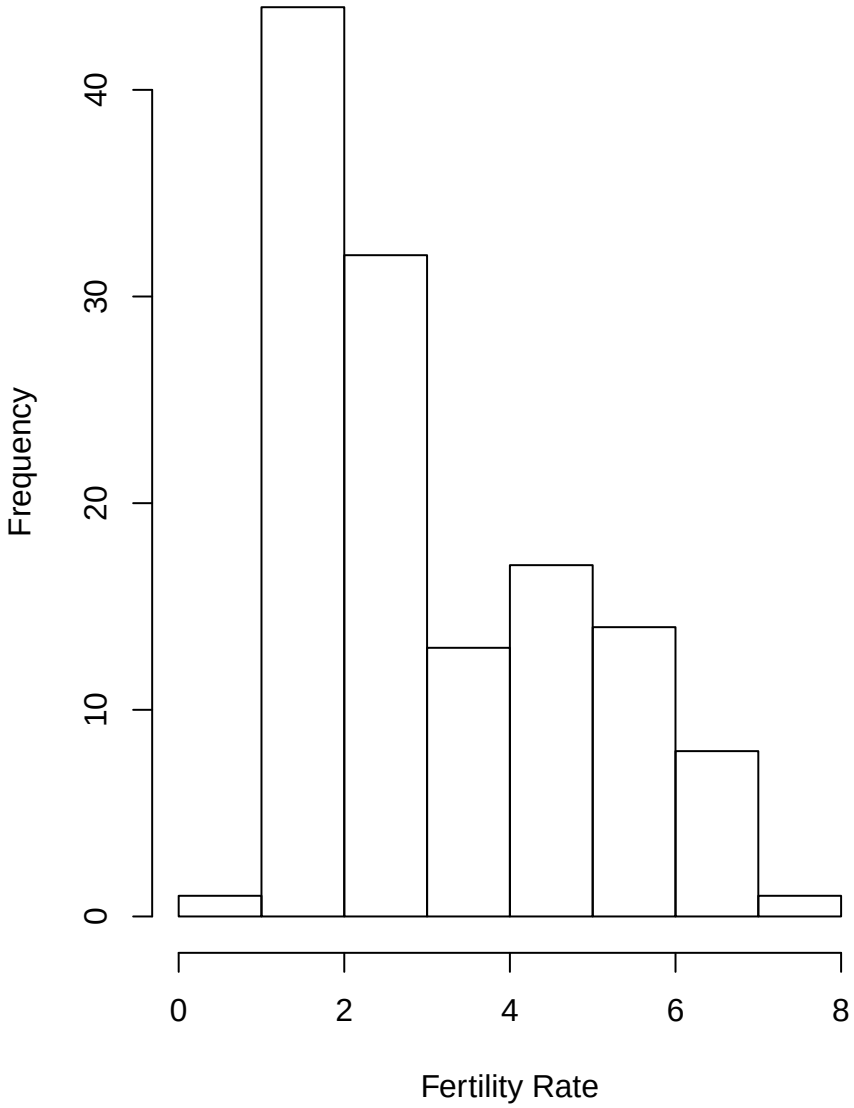


Median = 2.60

Mean = 3.12 children

std dev = 1.67 children

Summary of Univariate Distribution: Fertility



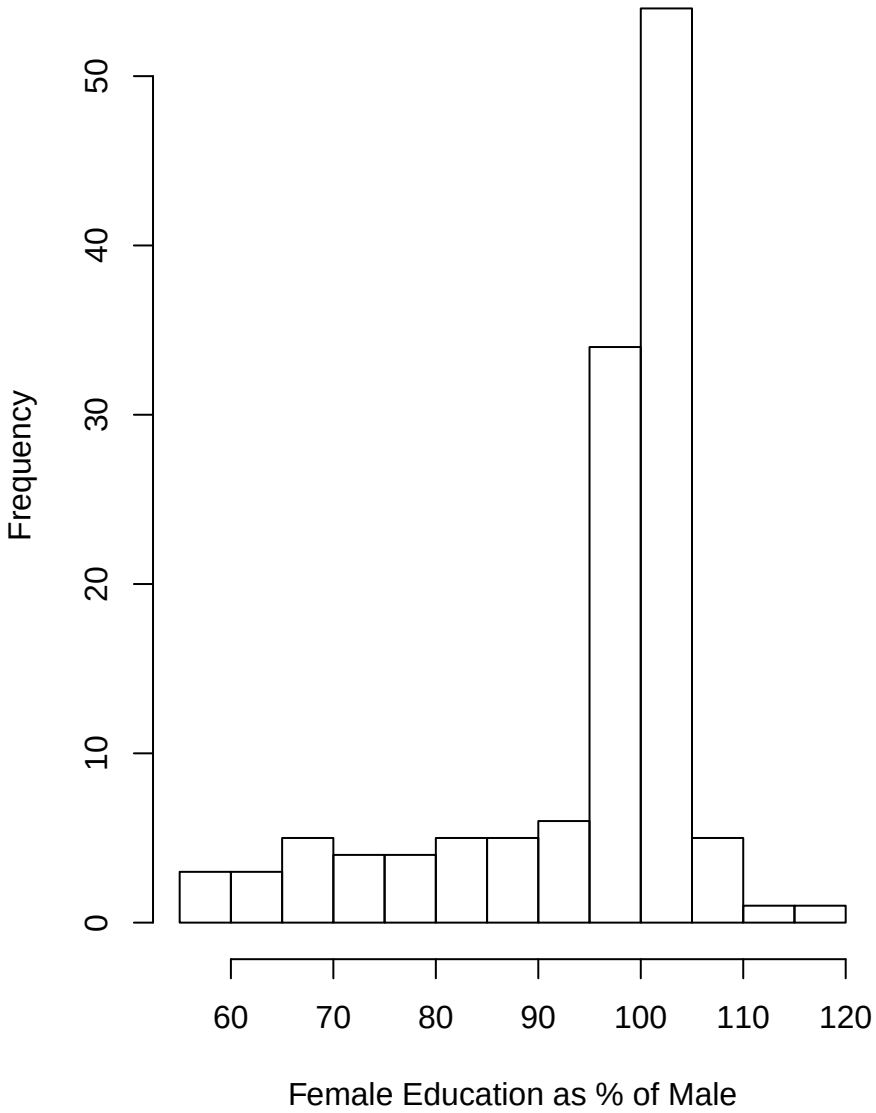
Median = 2.60

Mean = 3.12 children

std dev = 1.67 children

How would you describe this distribution?

Summary of Univariate Distribution: Education Ratio

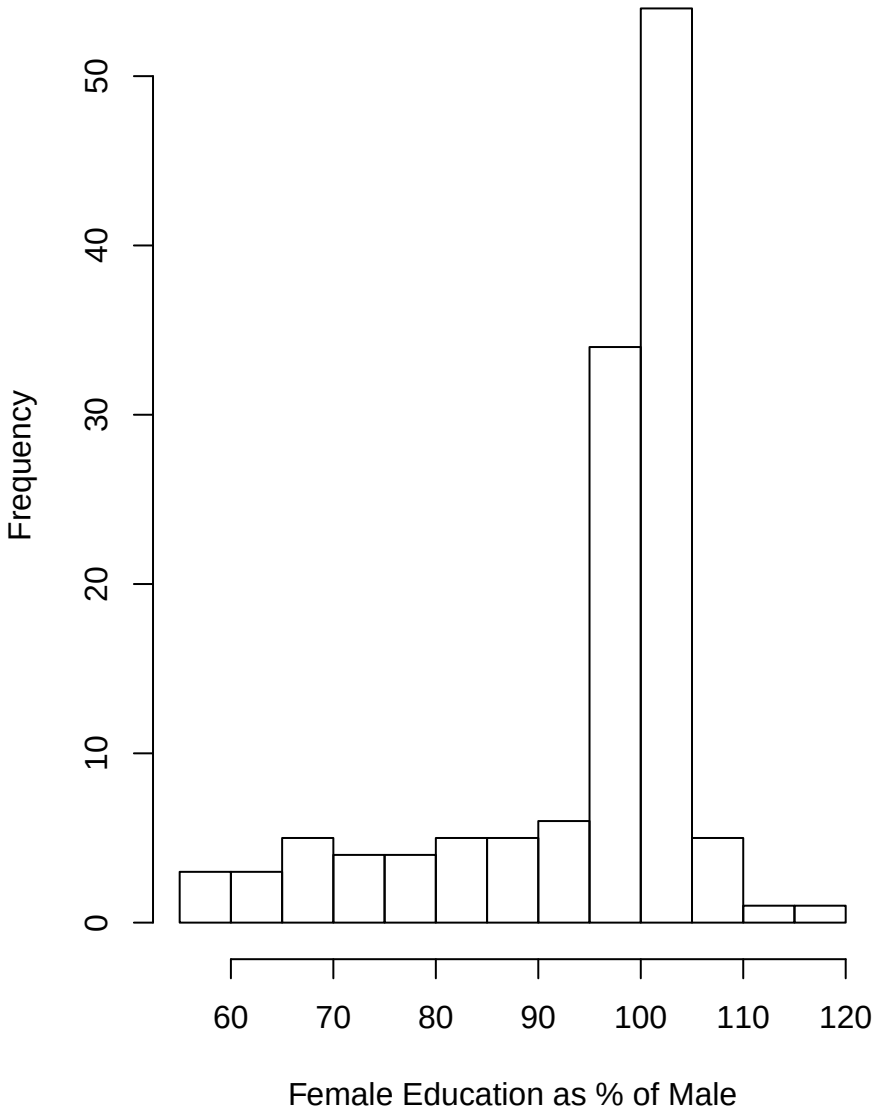


Median = 99.60%

Mean = 94.48%

std. dev. = 12.45%

Summary of Univariate Distribution: Education Ratio



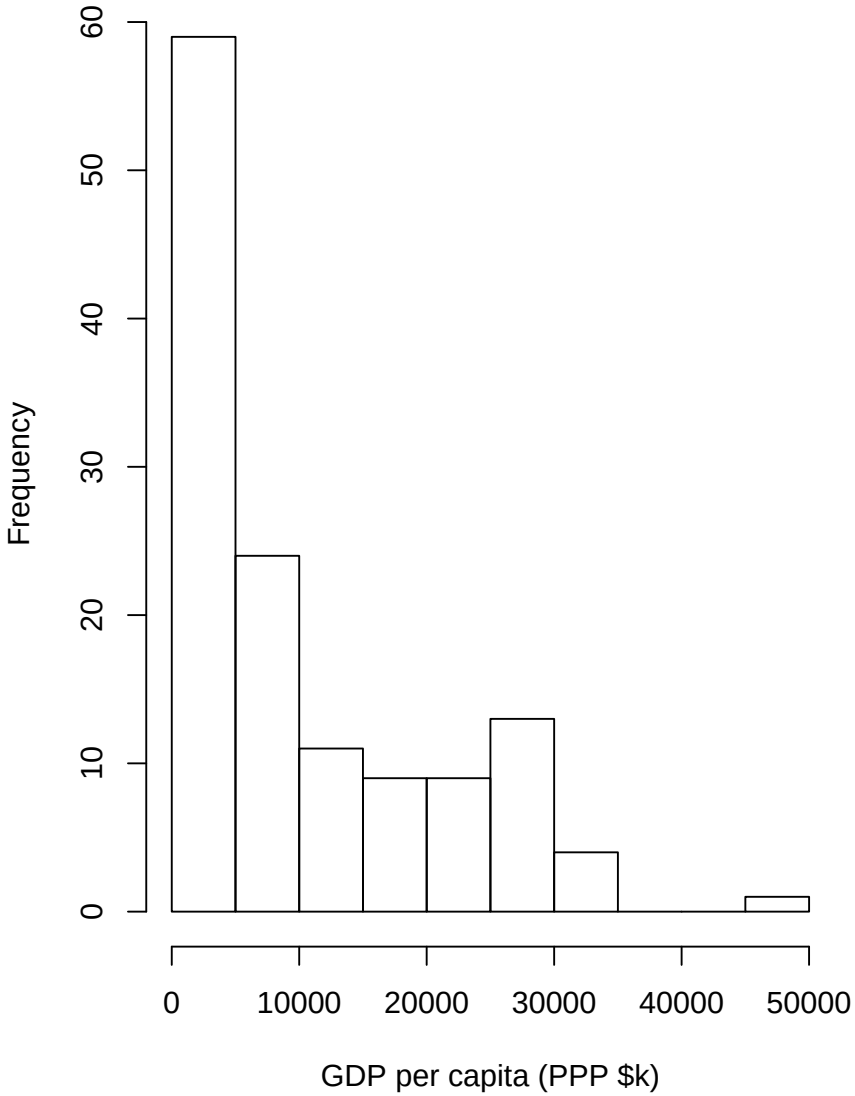
Median = 99.60%

Mean = 94.48%

std. dev. = 12.45%

How would you describe this distribution?

Summary of Univariate Distribution: GDP per capita

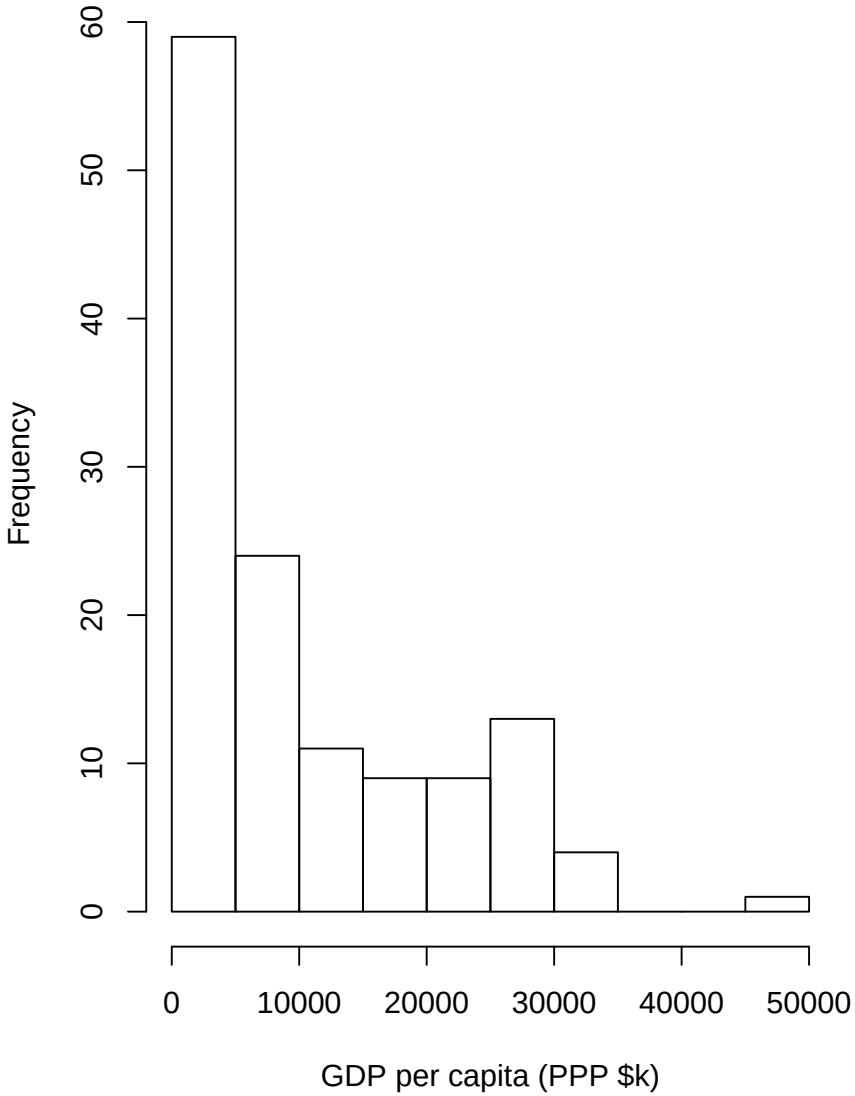


Median = $6047

Mean = $10,200

std. dev. = $10,078

Summary of Univariate Distribution: GDP per capita



Median = $6047

Mean = $10,200

std. dev. = $10,078

How would you describe this distribution?

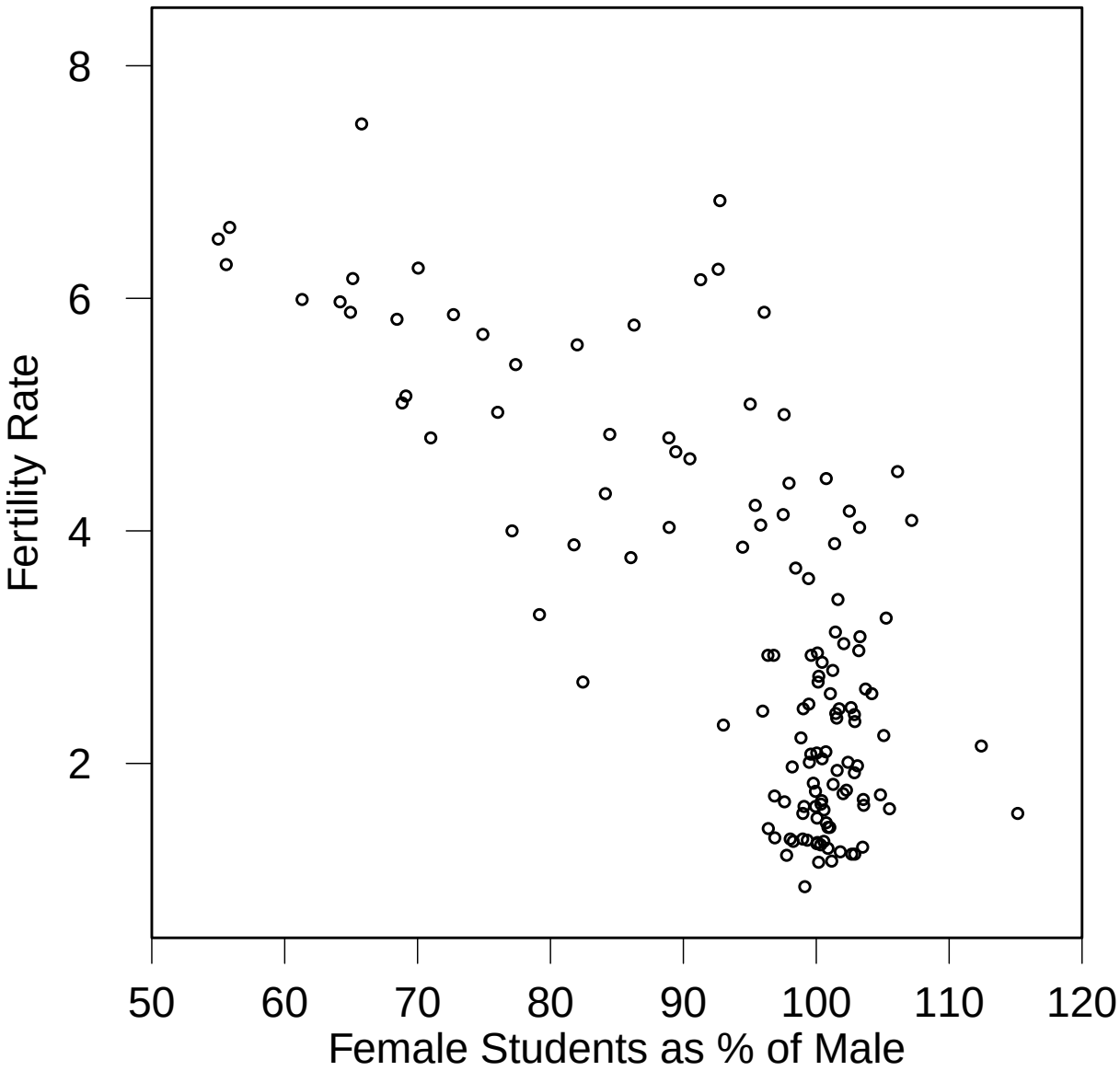
Scatterplots

Most powerful tool for bivariate data analysis

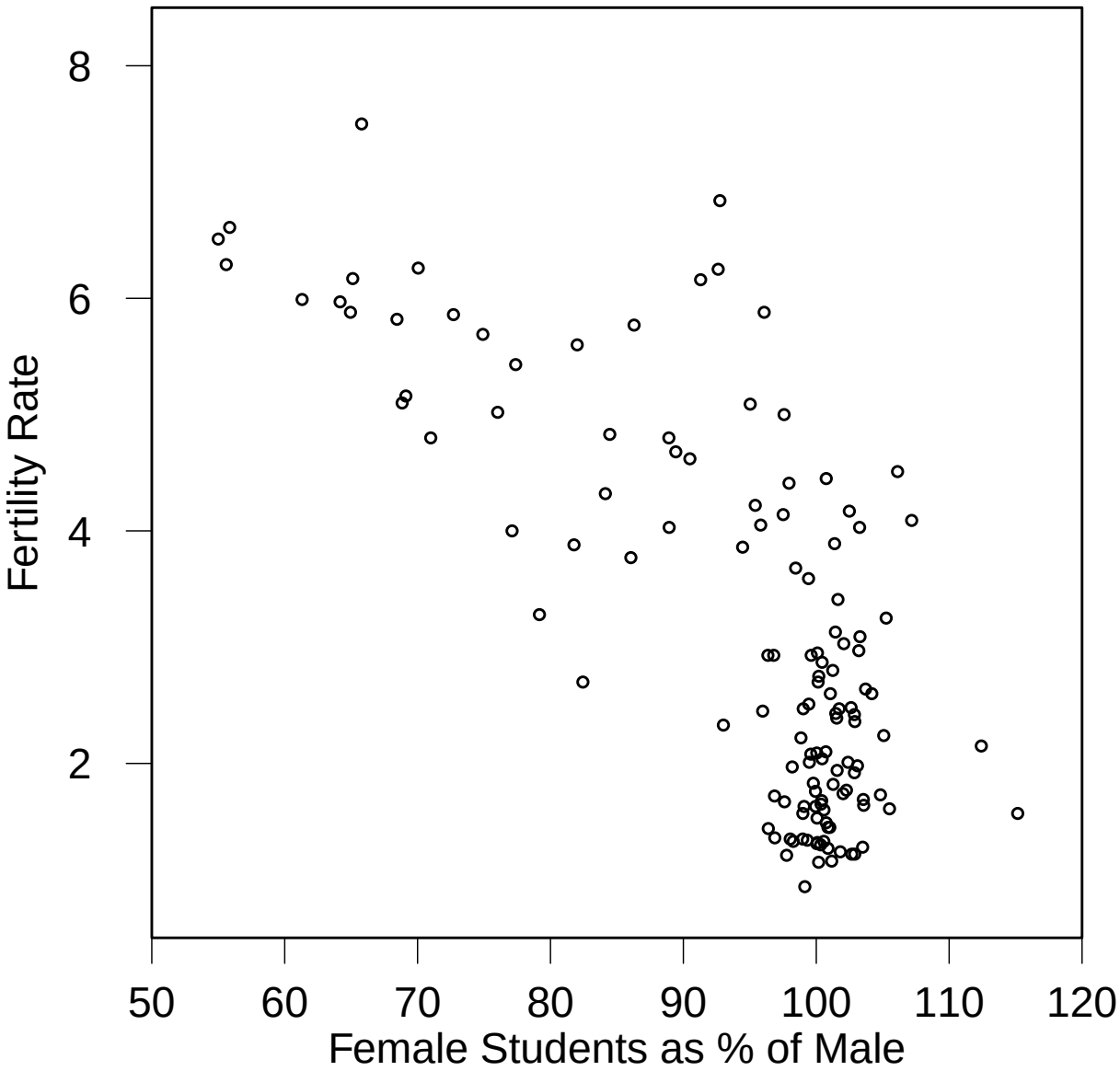
Will show full pattern of conditional expectation, including nonlinearity & outliers

Need additive level variables though!

No matter what advanced methods we learn, scatterplots will *always* be useful

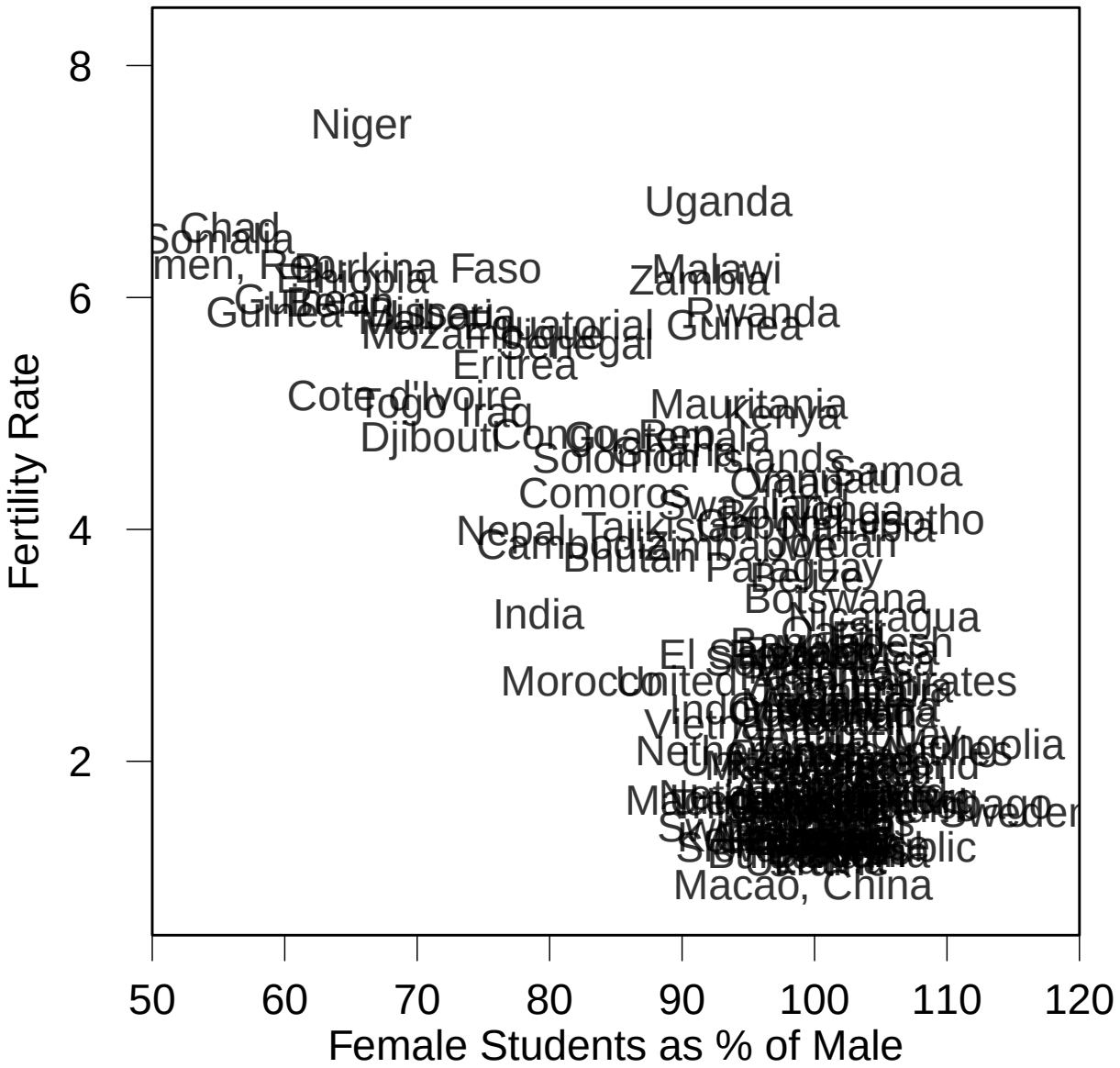


How would you describe the relationship between Fertility & Education Ratio?

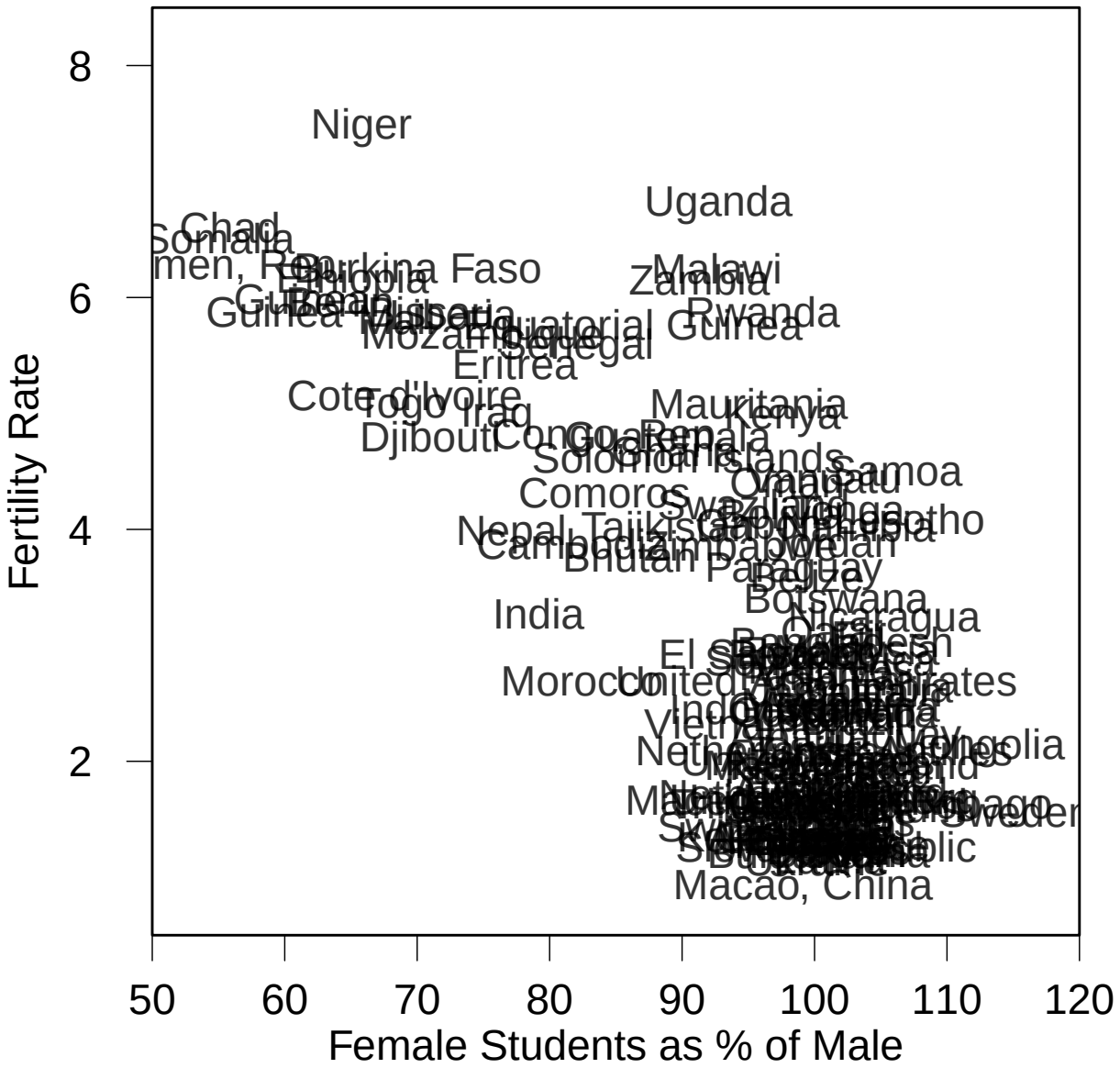


How would you describe the relationship between Fertility & Education Ratio?

If I asked you to predict Fertility for a country not sampled, how accurate do you expect your prediction to be?

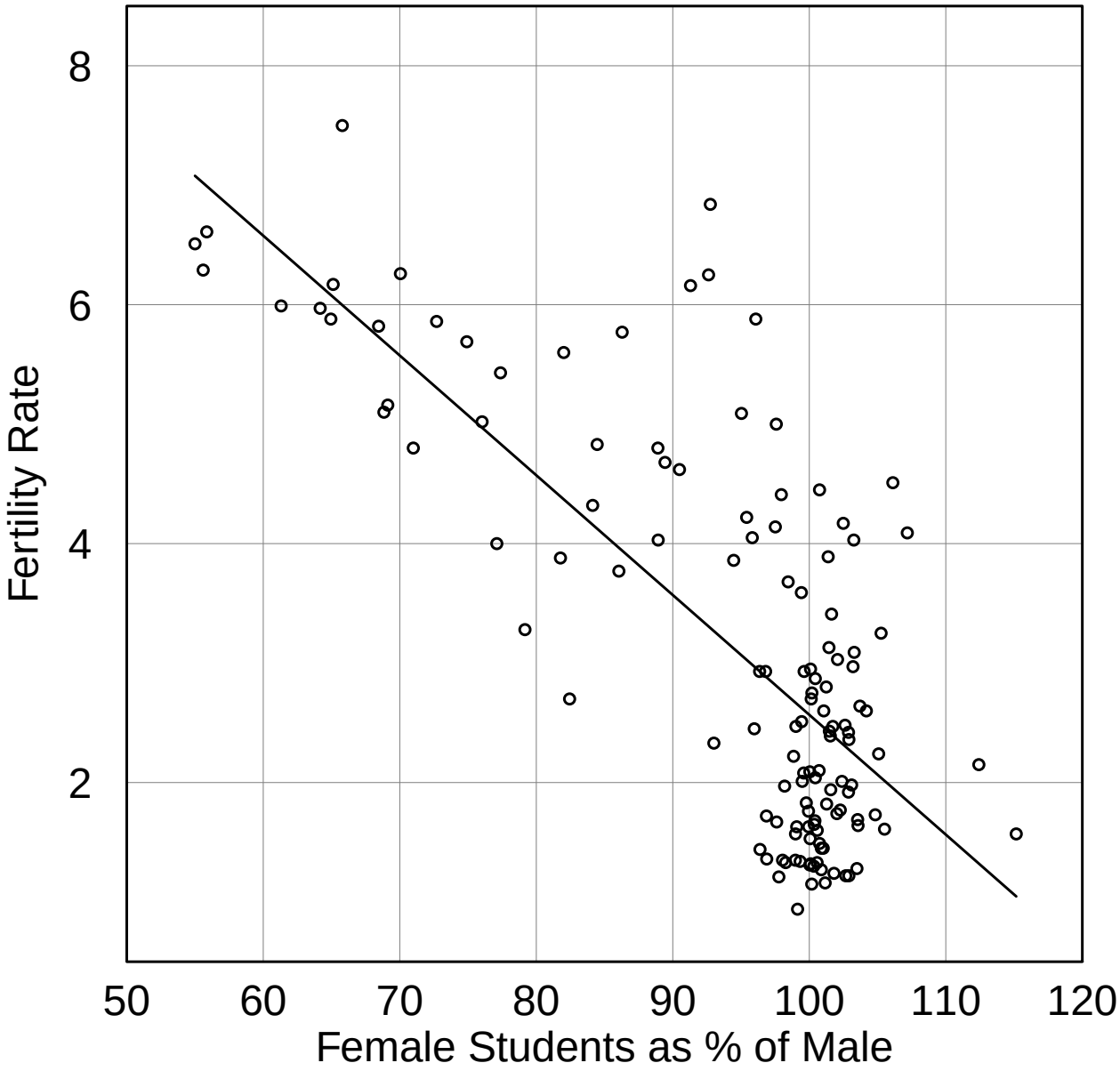


Labelling cases sometimes helps, especially for identifying *outliers*

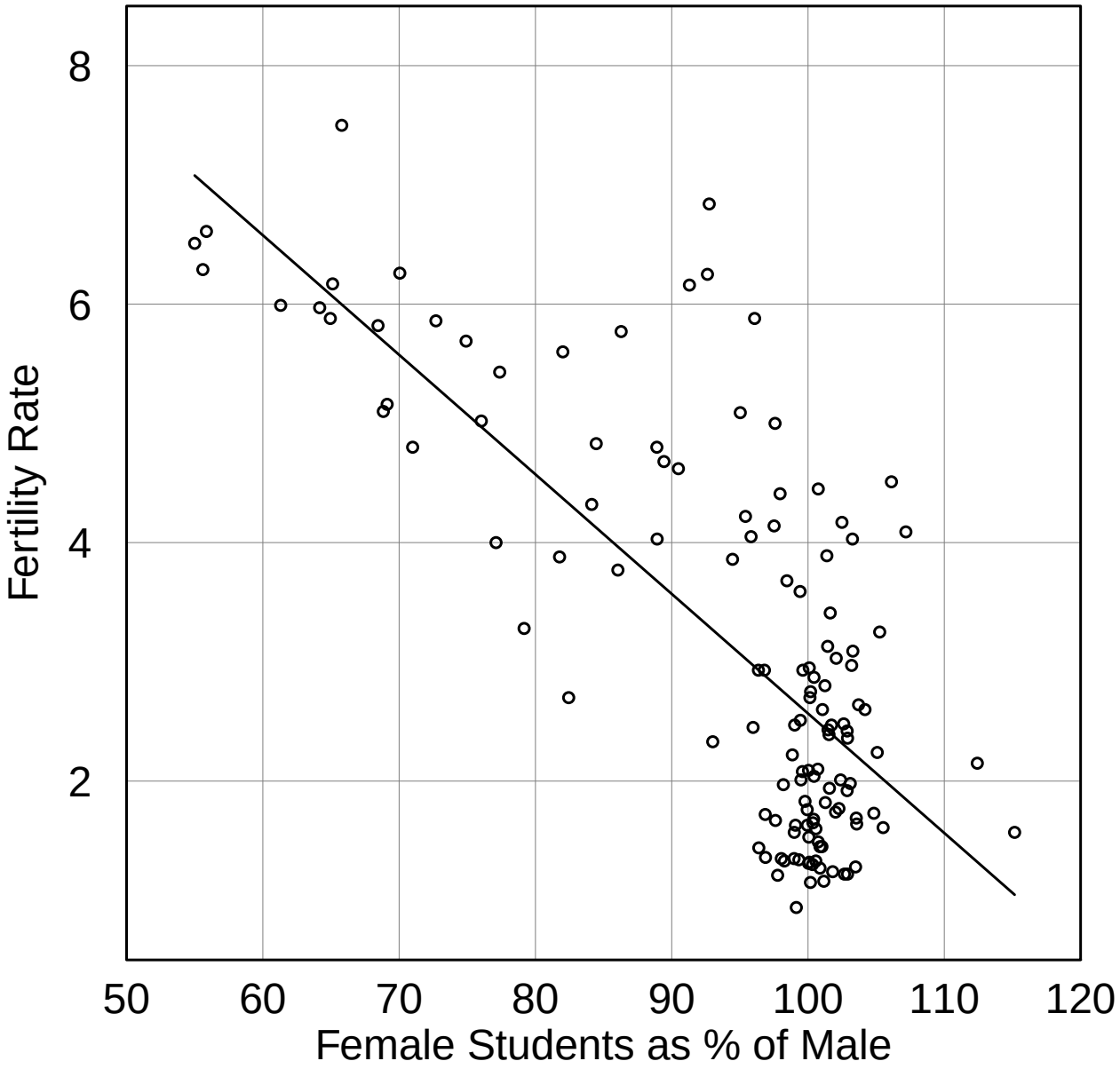


Labelling cases sometimes helps, especially for identifying *outliers*

What makes a point an outlier?



The best fit line is the line that passes closest to the majority of the points



The best fit line is the line that passes closest to the majority of the points

If we take this line to be our model of Fertility, how do we interpret it?

Best fit lines

From high school math, a line on a plane follows this equation:

$$y = b + mx$$

where:

- y is the dependent variable,
- x is the independent variable,
- m is the slope of the line,
  or the change in y for a 1 unit change in x,
- and b is the intercept,
  or value of y when $x = 0$

Best fit lines

Customarily, in statistics, we write the equation of a line as:

$$y = \beta_0 + \beta_1 x$$

where:

- y_i is the dependent variable
- x is the independent variable,
- β_1 is the slope of the line,
  or the change in y for a 1 unit change in x,
- and β_0 is the intercept,
  or value of y when $x = 0$

Best fit for fertility against education ratio

$$\begin{aligned} \widehat{\text{Fertility}} &= \hat{\beta}_0 + \hat{\beta}_1 \text{EduRatio} \\ \widehat{\text{Fertility}} &= 12.59 - 0.10 \times \text{EduRatio} \end{aligned}$$

The above equation is the best fit line given by *linear regression*

The $\hat{\beta}$'s are the estimated linear regression *coefficients*

$\widehat{\text{Fertility}}$ is the *fitted value*, or model prediction, of the level of Fertility givem the EduRatio

Intrepreting regression coefficients

$$\begin{aligned} \widehat{\text{Fertility}} &= \hat{\beta}_0 + \hat{\beta}_1 \text{EduRatio} \\ \widehat{\text{Fertility}} &= 12.59 - 0.10 \times \text{EduRatio} \end{aligned}$$

Interpreting $\hat{\beta}_1 = -0.10$:

Increasing EduRatio by 1 unit lowers Fertility by 0.10 units.

Because EduRatio is measured in percentage points, this means a 10% increase in female education (relative to males) will lower the number of children a woman has over her lifetime by 1 on average.

Intrepreting regression intercepts

$$\begin{aligned} \widehat{\text{Fertility}} &= \hat{\beta}_0 + \hat{\beta}_1 \text{EduRatio} \\ \widehat{\text{Fertility}} &= 12.59 - 0.10 \text{EduRatio} \end{aligned}$$

Interpreting $\hat{\beta}_0 = 12.59$:

If EduRatio is 0, Fertility will be 12.59.

If there are no girls in primary or secondary education, then women are expected to have 12.59 children on average over their lifetimes.

Can we trust this prediction?

Intrepreting regression intercepts

$$\begin{aligned} \widehat{\text{Fertility}} &= \hat{\beta}_0 + \hat{\beta}_1 \text{EduRatio} \\ \widehat{\text{Fertility}} &= 12.59 - 0.10 \text{EduRatio} \end{aligned}$$

Interpreting $\hat{\beta}_0 = 12.59$:

If EduRatio is 0, Fertility will be 12.59.

If there are no girls in primary or secondary education, then women are expected to have 12.59 children on average over their lifetimes.

Can we trust this prediction? No.
No country has 0 female education, so this is an *extrapolation* from the model.

Using regression coefficients to predict specific cases

$$\begin{aligned} \widehat{\text{Fertility}} &= \hat{\beta}_0 + \hat{\beta}_1 \text{EduRatio} \\ \widehat{\text{Fertility}} &= 12.59 - 0.10 \text{EduRatio} \end{aligned}$$

How many children do we expect women to get if girls get half the education boys do?

If EduRatio is 50, Fertility will be $12.59 - 0.10 \times 50 = 7.59$*.*

How many children do we expect women to have if girls get the same education boys do?

If EduRatio is 100, Fertility will be $12.59 - 0.10 \times 100 = 2.59$*.*

Using regression coefficients to predict specific cases

$$\begin{aligned} \widehat{\text{Fertility}} &= \hat{\beta}_0 + \hat{\beta}_1 \text{EduRatio} \\ \widehat{\text{Fertility}} &= 12.59 - 0.10 \text{EduRatio} \end{aligned}$$

If EduRatio is 100, Fertility will be $12.59 - 0.10 \times 100 = 2.59$.

Does this hold exactly for any country with education parity?

Using regression coefficients to predict specific cases

$$\begin{aligned} \widehat{\text{Fertility}} &= \hat{\beta}_0 + \hat{\beta}_1 \text{EduRatio} \\ \widehat{\text{Fertility}} &= 12.59 - 0.10 \text{EduRatio} \end{aligned}$$

If EduRatio is 100, Fertility will be $12.59 - 0.10 \times 100 = 2.59$*.*

Does this hold exactly for any country with education parity?

No. It holds on average. In any specific case *i*, there is some error between the expected and actual levels of Fertility

The linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

To account for the random deviation of each case from the underlying trend, we add an *error term*, ε_i.

We will assume our y_i's follow the above model

That is, we will assume there is some "true" β_0 and β_1 which generated the y_i we observe, and some "true" error from this trend

The linear regression model

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\varepsilon}_i$$

When we estimate this model, we designate the estimates by adding "hats"

The estimates $(\hat{\beta}_0, \hat{\beta}_1, \hat{\varepsilon}_i)$ probably differ from the (usually unknown) true values $(\beta_0, \beta_1, \varepsilon_i)$

To emphasize this, we will call $\hat{\varepsilon}_i$ the residual, since it is not the true error, but only an estimate

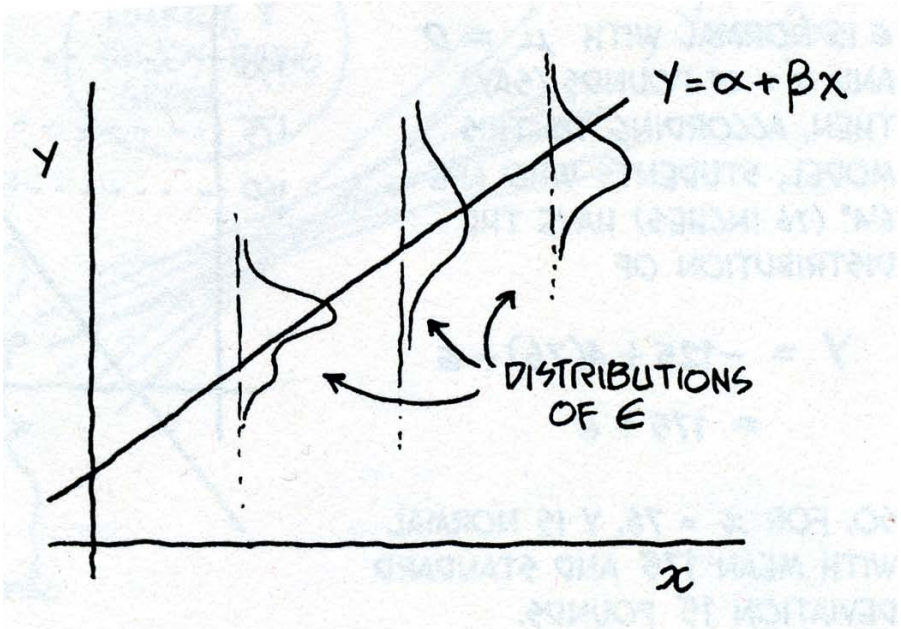
Estimating linear regression coefficients

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\varepsilon}_i$$

How do we obtain our estimates of the β's?

The full details are beyond the scope of 205

Key assumption is that ε_i is Normally distributed, $\mathrm{N}(0, \sigma^2)$



(Source: Larry Gonick & Wollcott Smith, *The Cartoon Guide to Statistics*)

The distribution of ε_i determines how closely or widely the y_i's are spaced around the best fit line

Our key simplifying assumption is that everywhere around the line, the y_i's are spread with the same Normal distribution

Estimating $\hat{\beta}$

With this assumption in mind, how do we find the best fit line?

Estimating $\hat{\beta}$

With this assumption in mind, how do we find the best fit line?

Perhaps the line that minimizes the total residuals?

Estimating $\hat{\beta}$

With this assumption in mind, how do we find the best fit line?

Perhaps the line that minimizes the total residuals?

But some residuals are positive, and others negative—their sum is always 0

Estimating $\hat{\beta}$

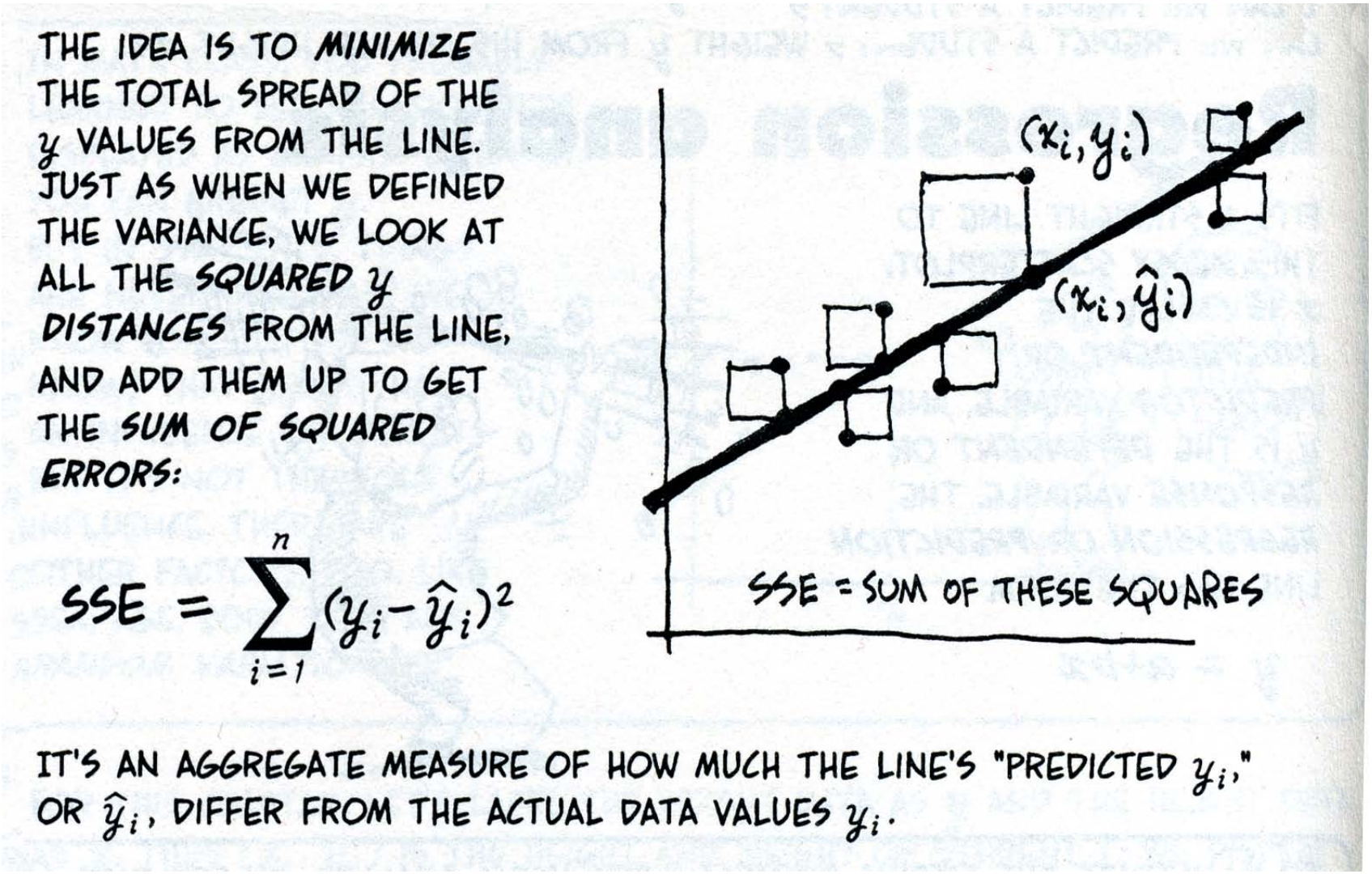
With this assumption in mind, how do we find the best fit line?

Perhaps the line that minimizes the total residuals?

But some residuals are positive, and others negative—their sum is always 0

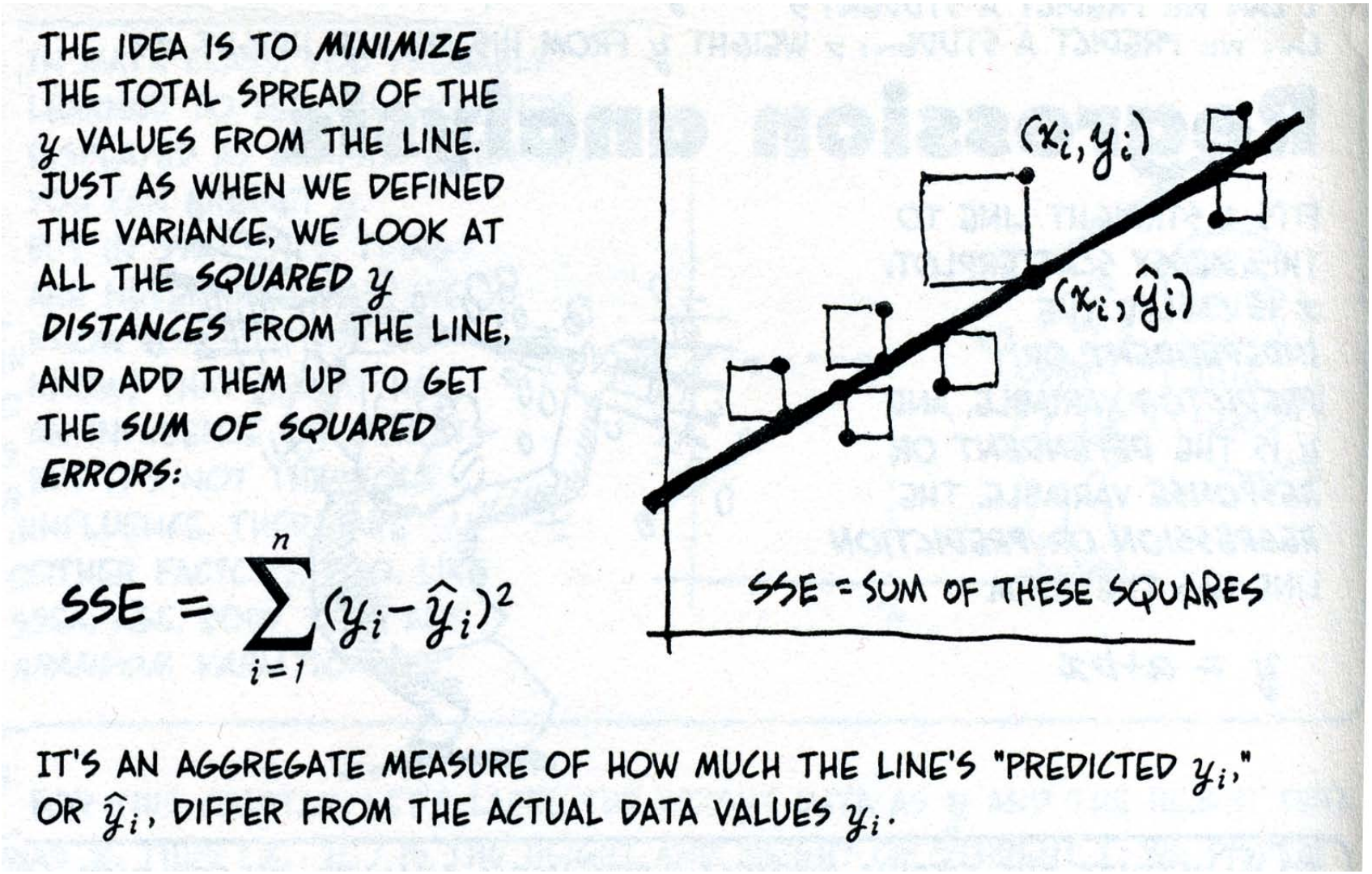
So lets minimize the sum of squared error!

Linear regression is fitted using the *least squares* procedure



(Source: Larry Gonick & Wollcott Smith, *The Cartoon Guide to Statistics*)

The *least squares estimates* are the $\hat{\beta}$'s that minimize the total area of the above squares

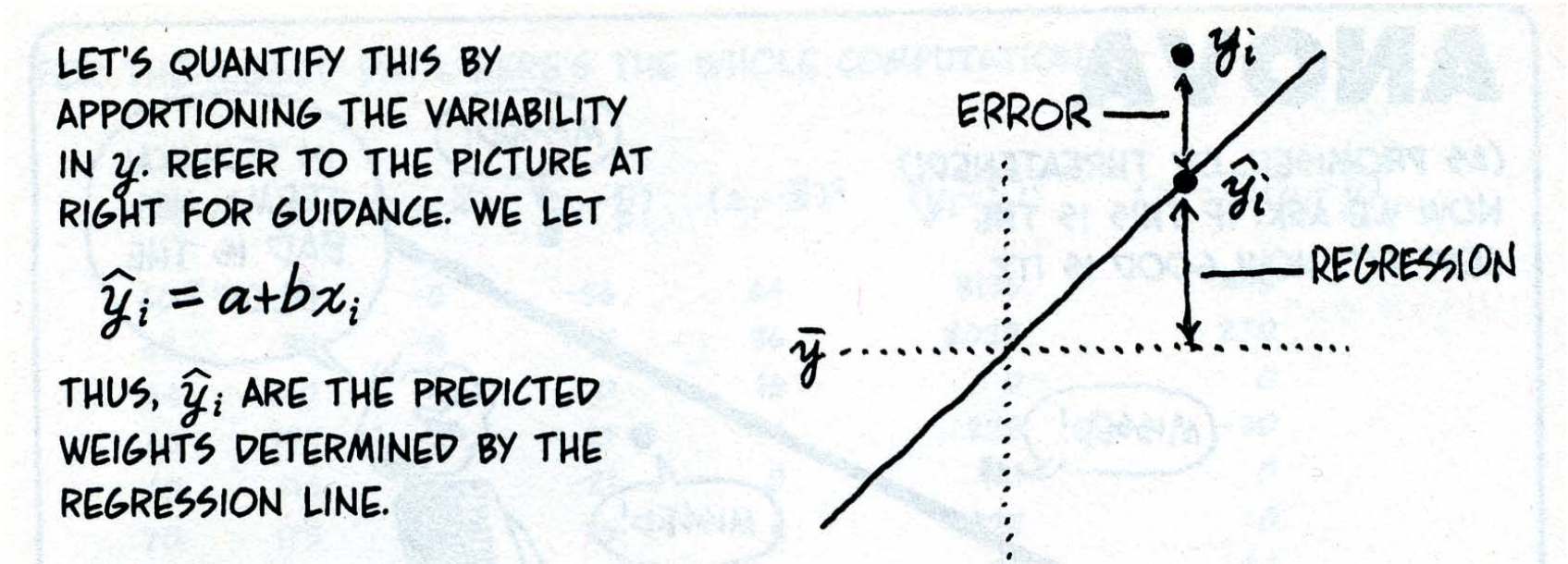


(Source: Larry Gonick & Wollcott Smith, *The Cartoon Guide to Statistics*)

STATA can find these $\hat{\beta}$'s easily

Residuals

Notice the distinction between what we *explain* and what is left *unexplained*



(Source: Larry Gonick & Wollcott Smith, *The Cartoon Guide to Statistics*)

Analysis of variance

The total variation in y_i is its total variance from the mean $\bar{y}$, or $\sum_{i=1}^{n}(y_i - \bar{y})^2$

Using least squares, we can break down the variance in y_i into two components:

Sum of square errors (SSE)	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
Regression sum of squares (RSS)	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
Total sum of squares (TSS)	$\sum_{i=1}^{n}(y_i - \bar{y})^2$

The Regression sum of squares (RSS) is what we have explained

The Sum of squared errors (SSE) is what is left unexplained

Analysis of variance

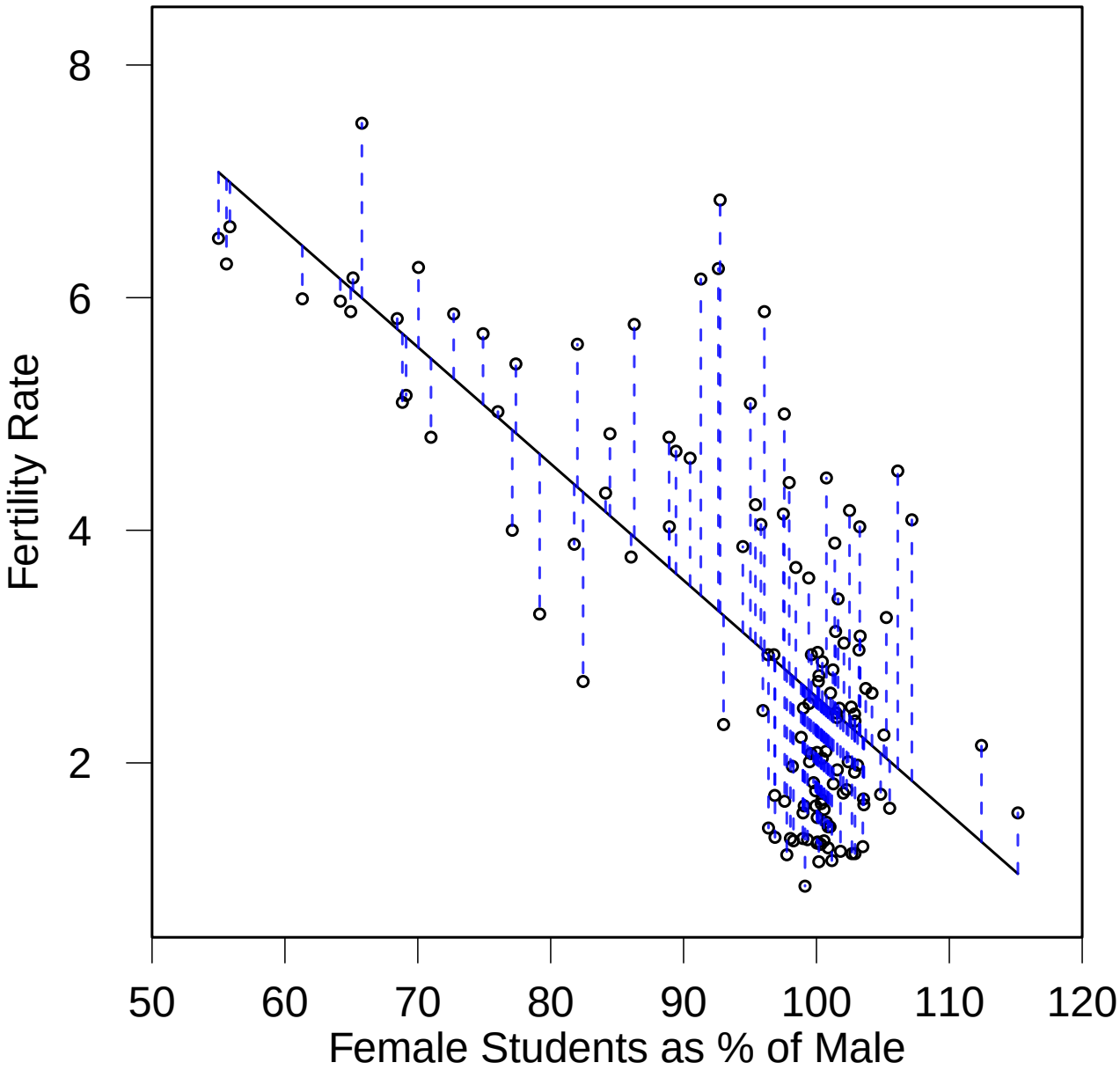
The Sum of squared errors ($\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^2$) is what is left unexplained

A very useful summary of this is the square root of the mean squared error:

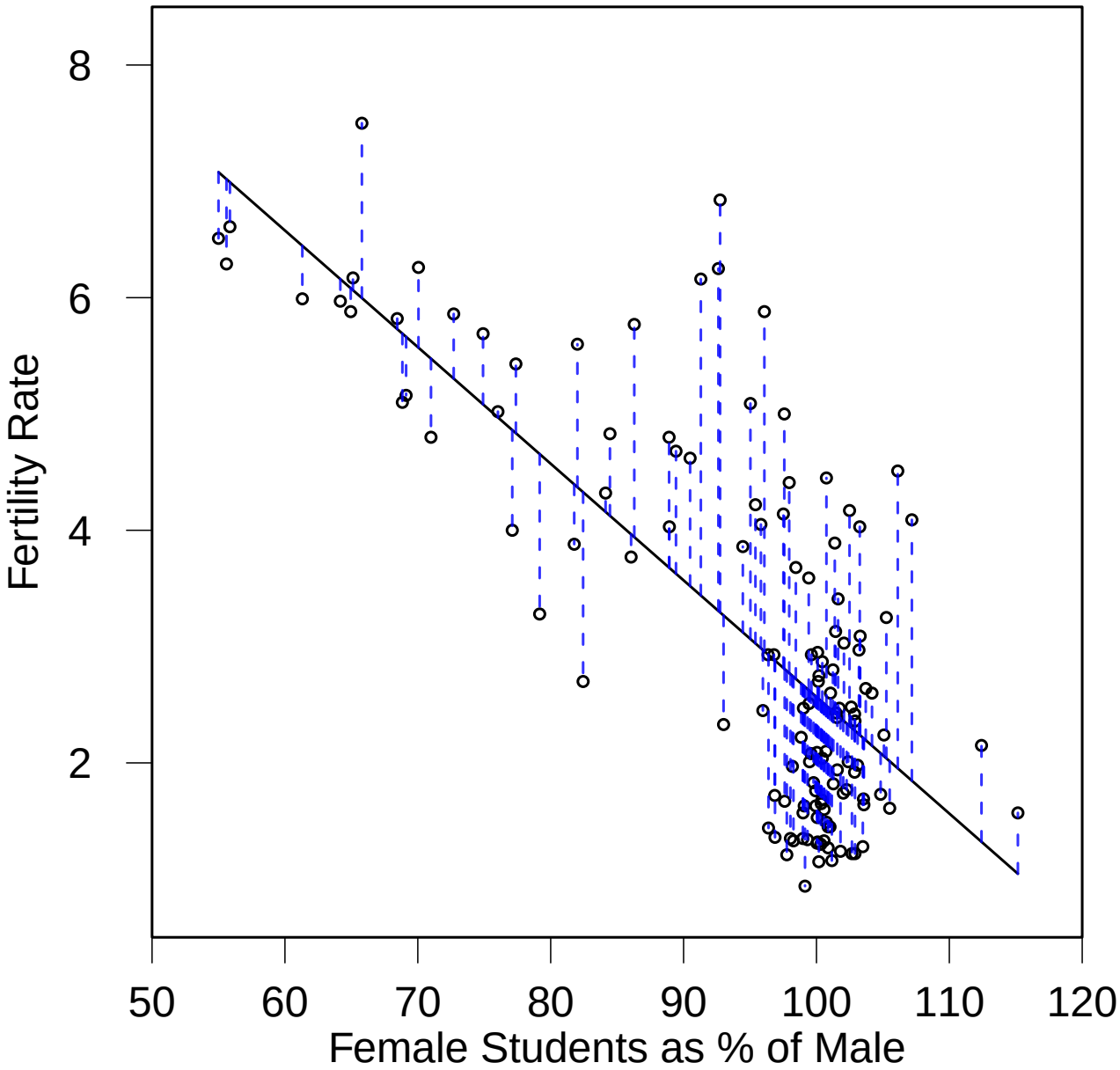
$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

This is how much a prediction from this linear regression will differ from the true y_i on average

Also known as the *standard error of the regression*

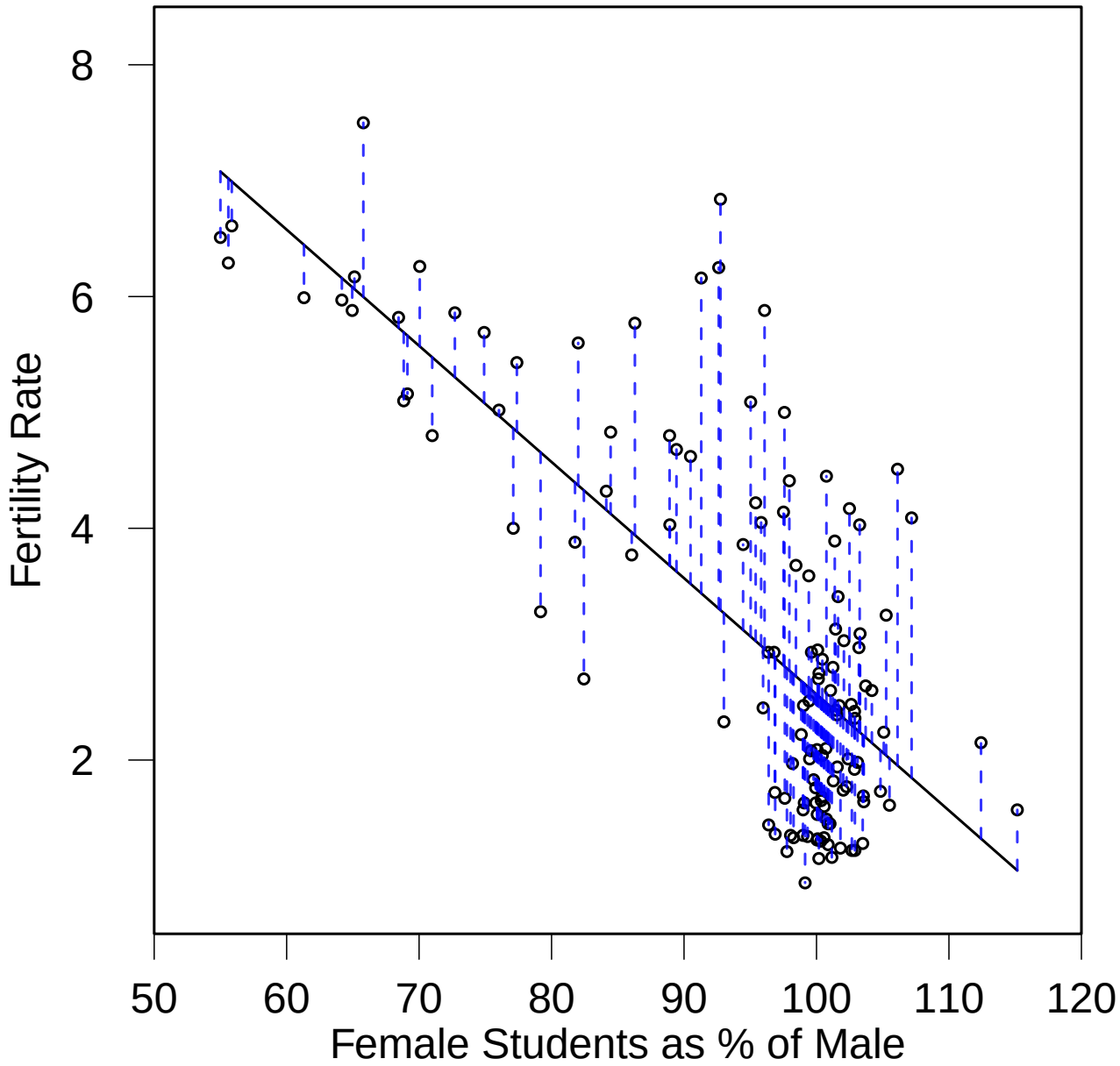


The residuals for the regression of Fertility on Education Ratio

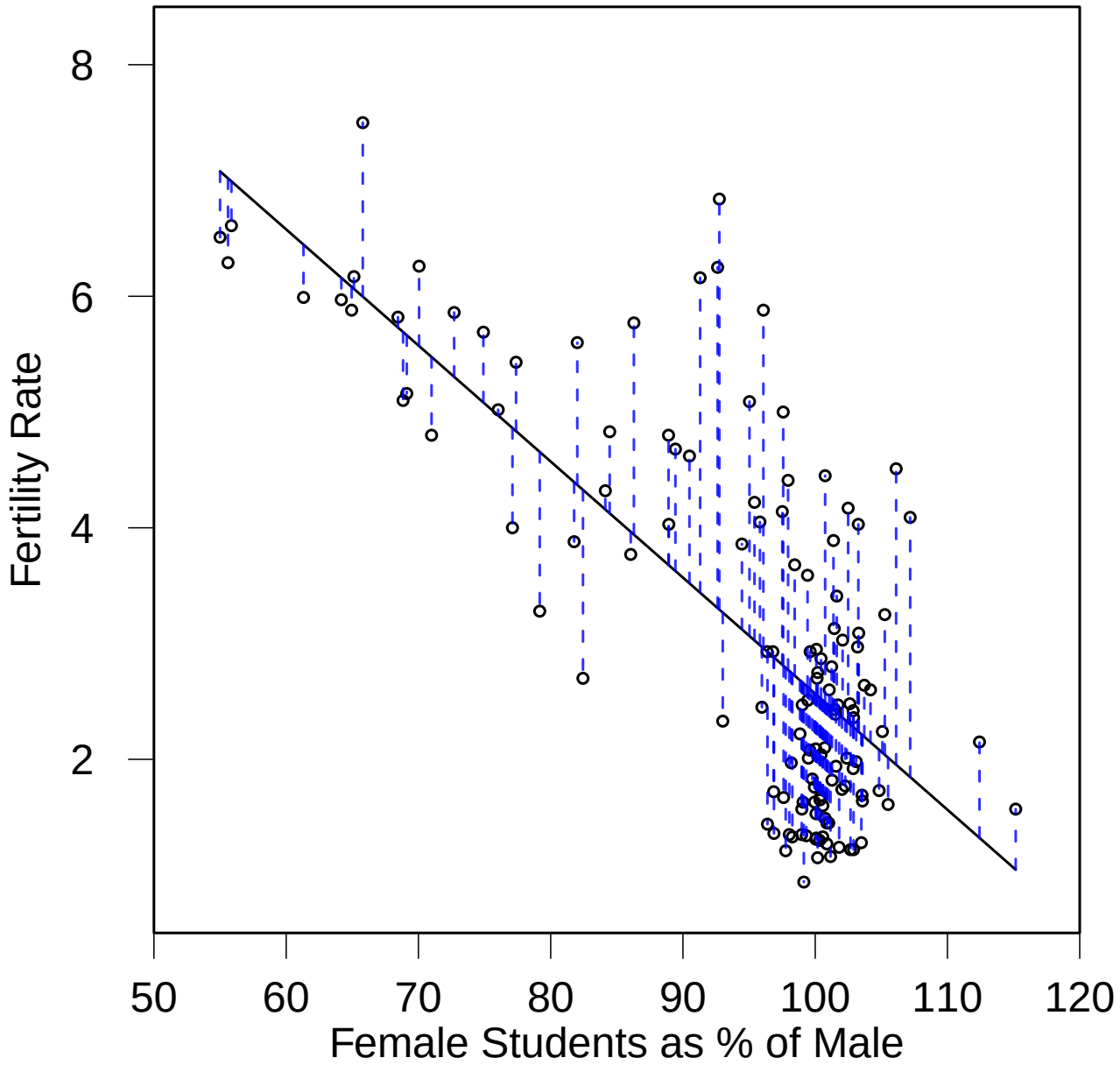


The residuals for the regression of Fertility on Education Ratio

This line minimizes the squared deviations on the dependent variable



The smaller the sum of squared residuals, the better the model *fits* the data.



The smaller the sum of squared residuals, the better the model *fits* the data.

The quality of model fit is a separate issue from the substantive strength of the relationship, which is given by β, or the change in *y* for a one unit change in *x*

Important distinction: regression coefficients (β) & correlation coefficients (r)

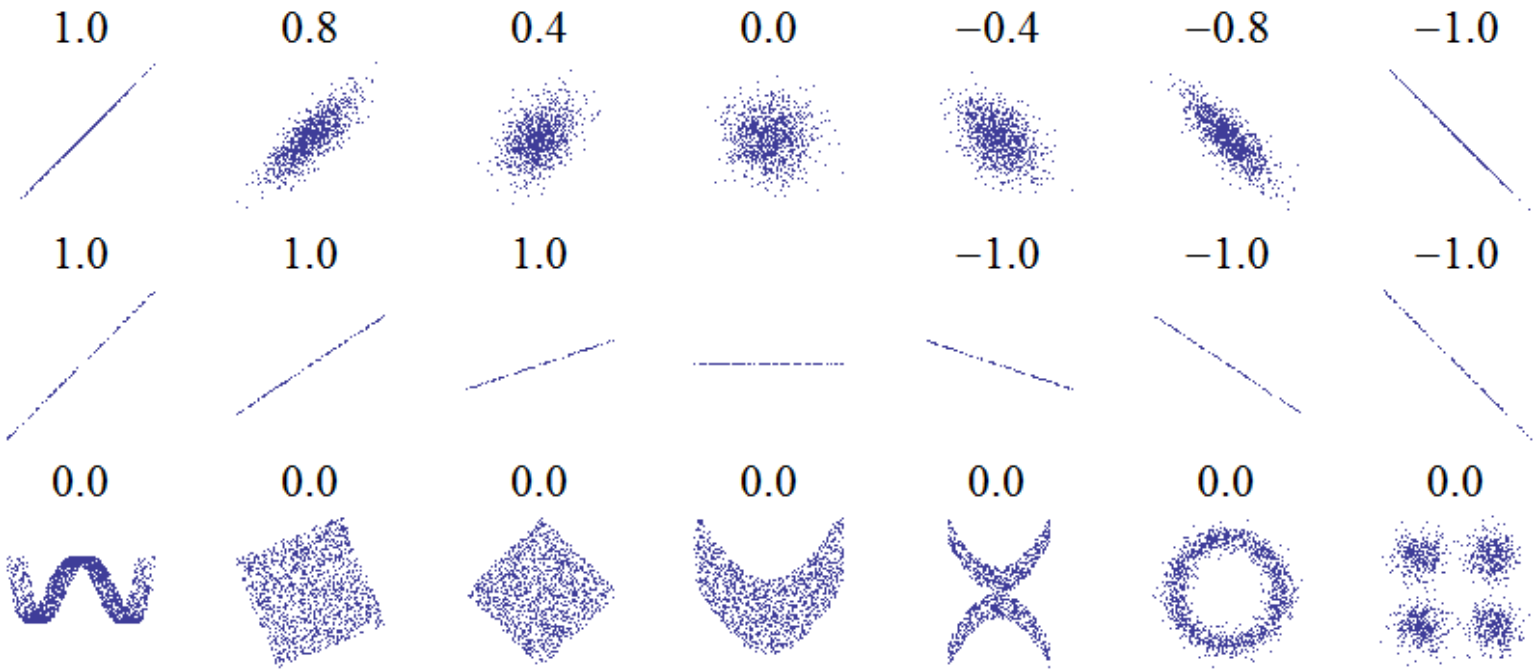
We've already discussed *correlation* between two variables, and noted that the *correlation coefficient* between two variables is:

$$\text{corr}(X, Y) = \frac{\text{E}\left((X - \text{E}(X))\left(Y - \text{E}(Y)\right)\right)}{\sqrt{\text{var}(X)}\sqrt{\text{var}(Y)}}$$

Note: the numerator of the above is also known as the *covariance*, or $\sigma_{X,Y}$, so the correlation $r_{X,Y}$ can also be stated as:

$$r_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$$

Correlation examples



Correlation coefficients measure strength of association

When two variables X and Y are highly *correlated*:

- They have $|r_{X,Y}| \approx 1$
- If we know X, we can narrow the expected range of Y down to a small interval
- If we know Y, we can narrow the expected range of X down to a small interval

Regression coefficients measure the substance of relationships

However, the correlation coefficient *r* does *not* tell us the substance of the relationship:

r does not tell us how much *Y* changes for a particular change in *X*:
that's $\hat{\beta}$

Nor does *r* tell us how much sampling variability might induce *estimation* uncertainty in $\hat{\beta}$: that's $\mathrm{se}(\hat{\beta})$

r tells us about a *different* kind of uncertainty:
the uncertainty caused by *real* random variation in the world that muddles the relationship of *X* and *Y*

Contrasting *r* and β

Low *r* between Fertility and Education Ratio, for example, would tell us that many other random factors besides female education intervene in causing Fertility in a particular case

High *r* would tell us that few stochastic factors intervene in any particular case

Low β would tell us that it takes a lot of female education to lower Fertility, on average

High β would tell us that a little bit of female education lowers Fertility a lot, on average

Goodness of fit

Our model is captured in the β's, or regression coefficients

r is an example of a goodness of fit measure;
larger values imply better fit of the model to the data

In our example, *r* between Fertility and Education Ratio is -0.75

Substantively, this number is hard to interpret

(What's a "big" *r*? A "small" *r*? Arbitrary)

The coefficient of determination, R^2

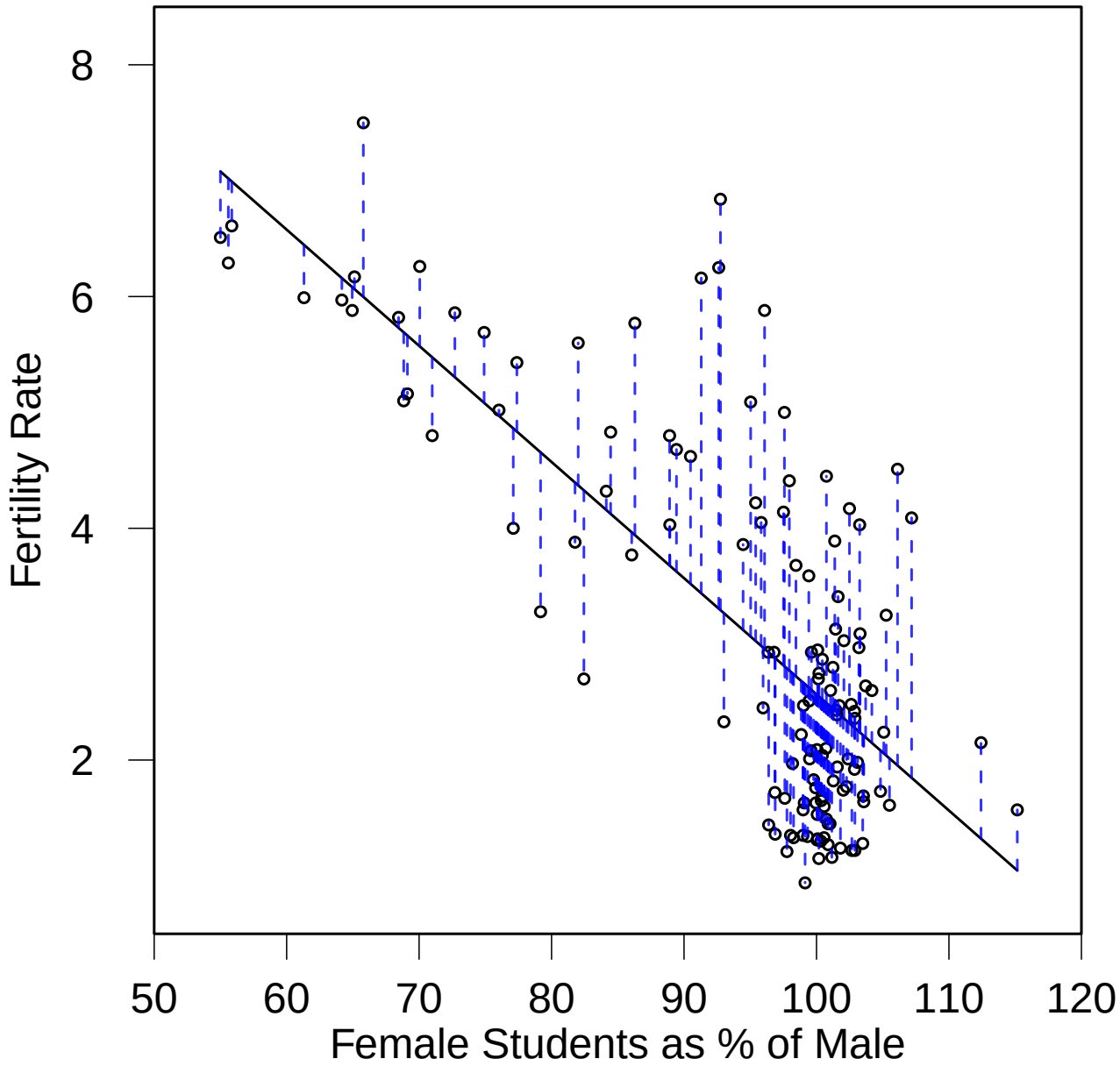
One easy to interpret goodness of fit measure is R^2,
known as the coefficient of determination

In general, R^2 is the ratio of the variance the model explains to the total variance:

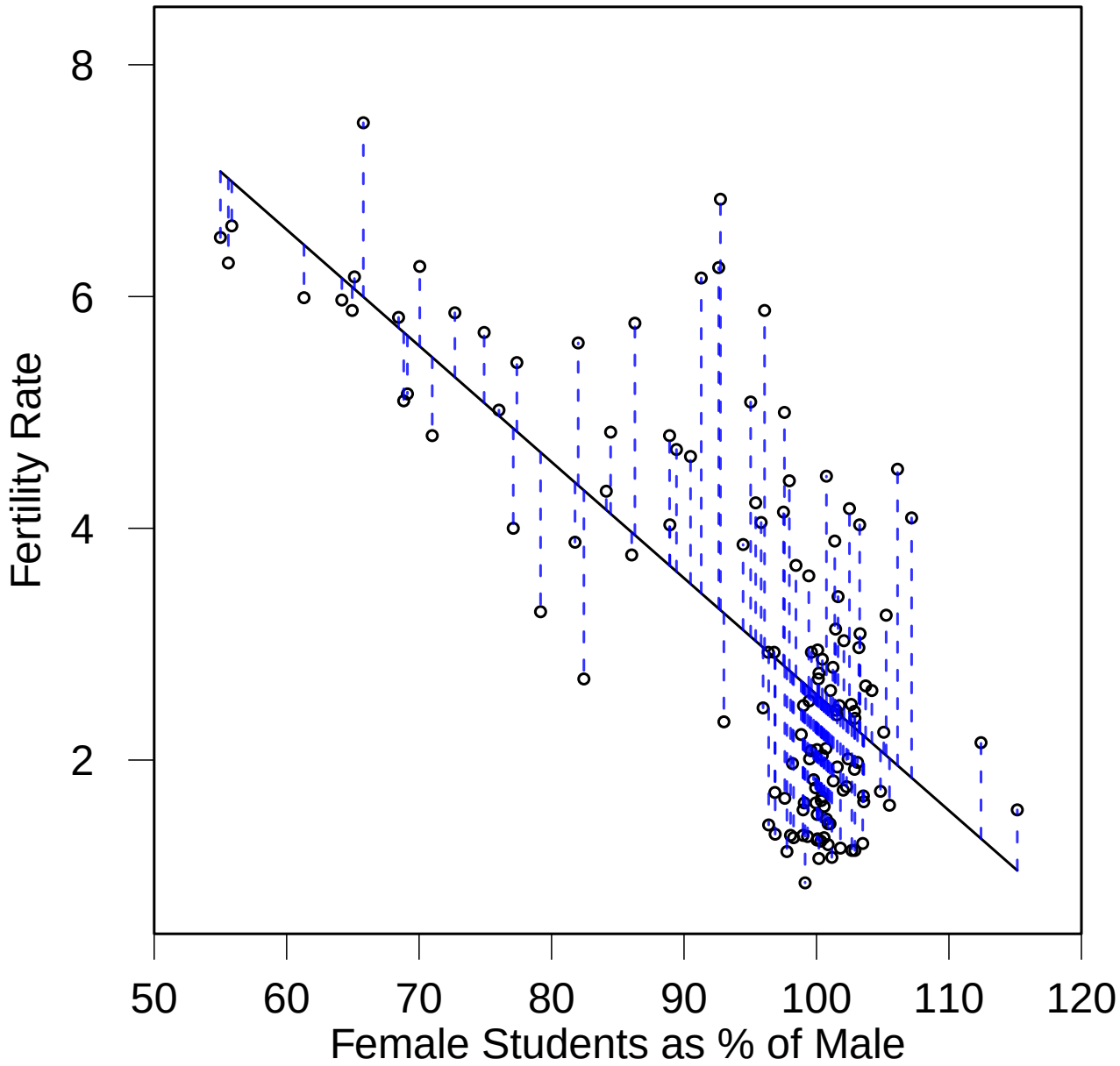
$$R^2 = \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\text{SSE}}{\text{TSS}}$$

In *bivariate* regression only, R^2 also the square of $r_{X,Y}$

In our example, $R^2 = 0.56$, which says that Education Ratio "explains" 56% of the variation in Fertility, *and vice versa*

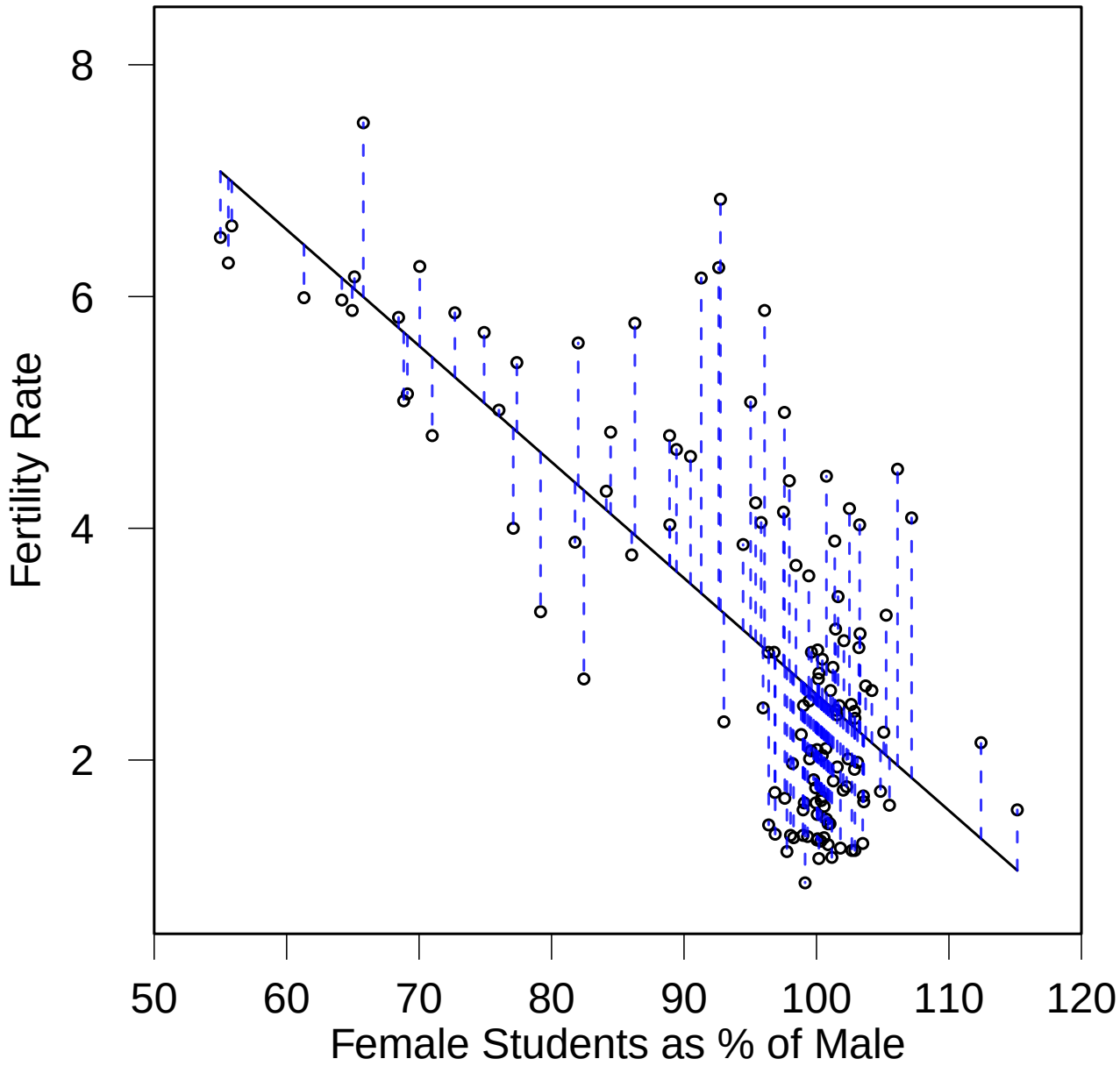


I prefer a more tangible measure of goodness of fit, the root mean squared error (RMSE).



I prefer a more tangible measure of goodness of fit, the root mean squared error (RMSE).

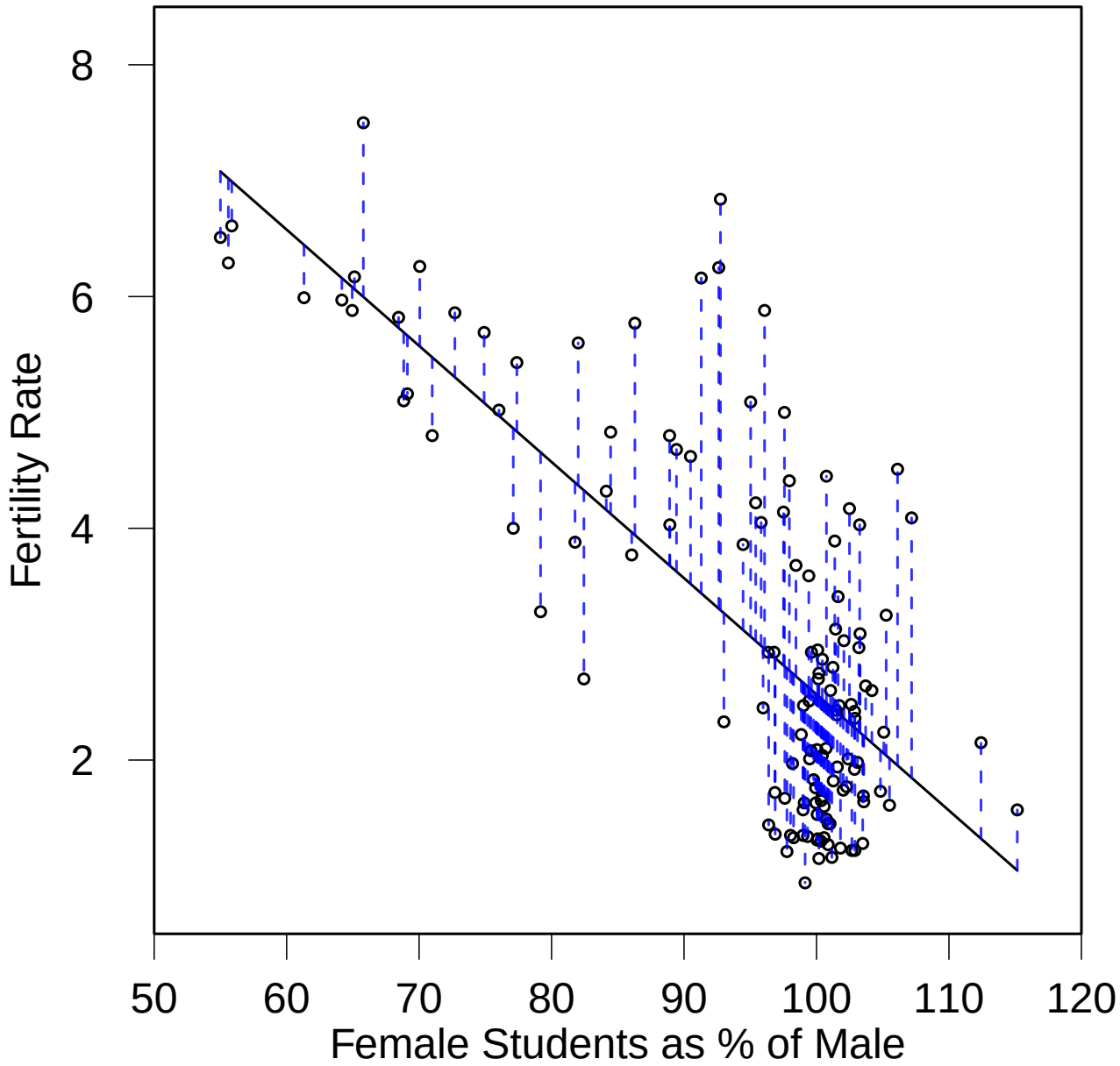
RMSE is "how much your model predictions miss by":



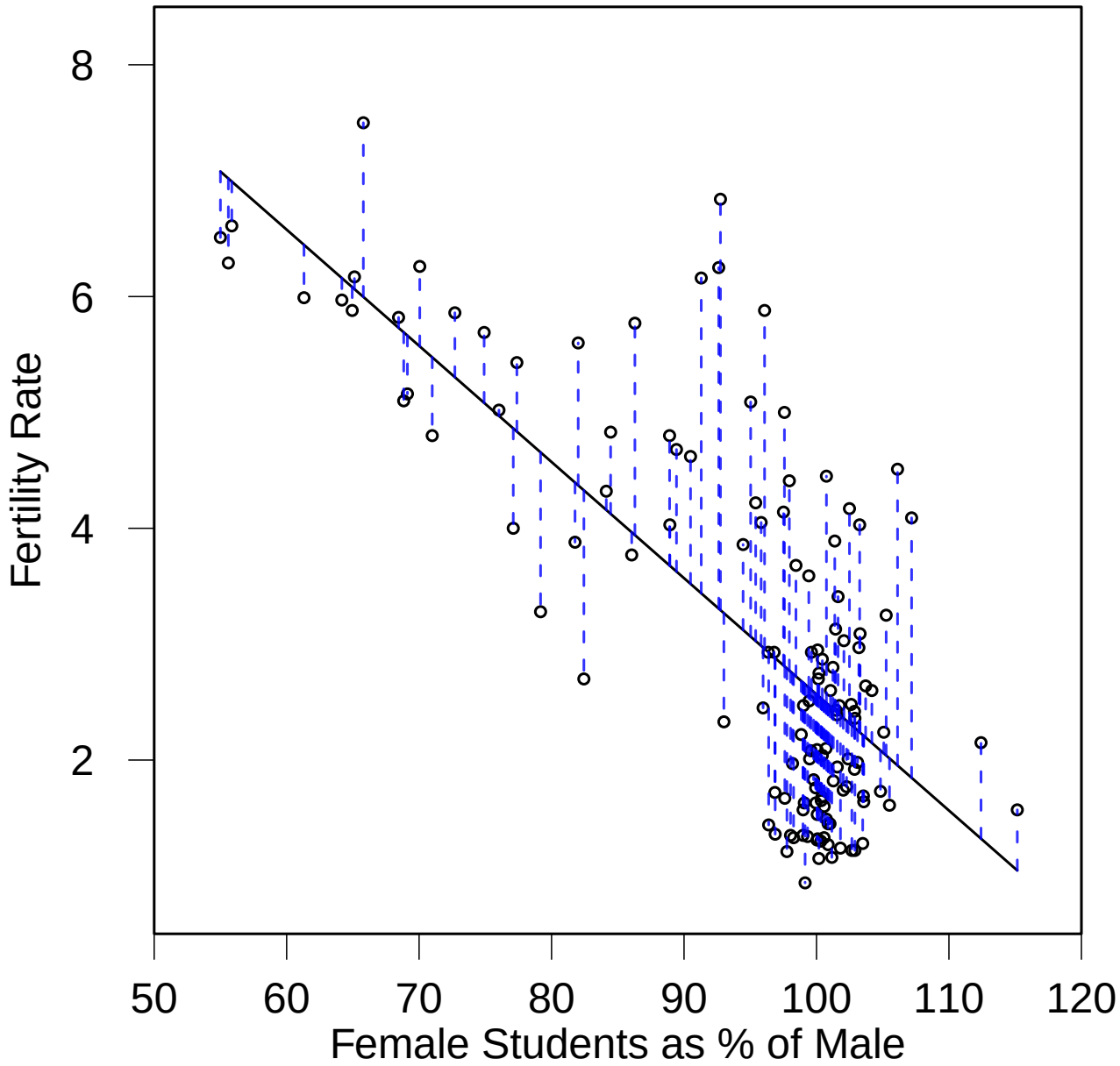
I prefer a more tangible measure of goodness of fit, the root mean squared error (RMSE).

RMSE is “how much your model predictions miss by”:

here, 1.12 children per female

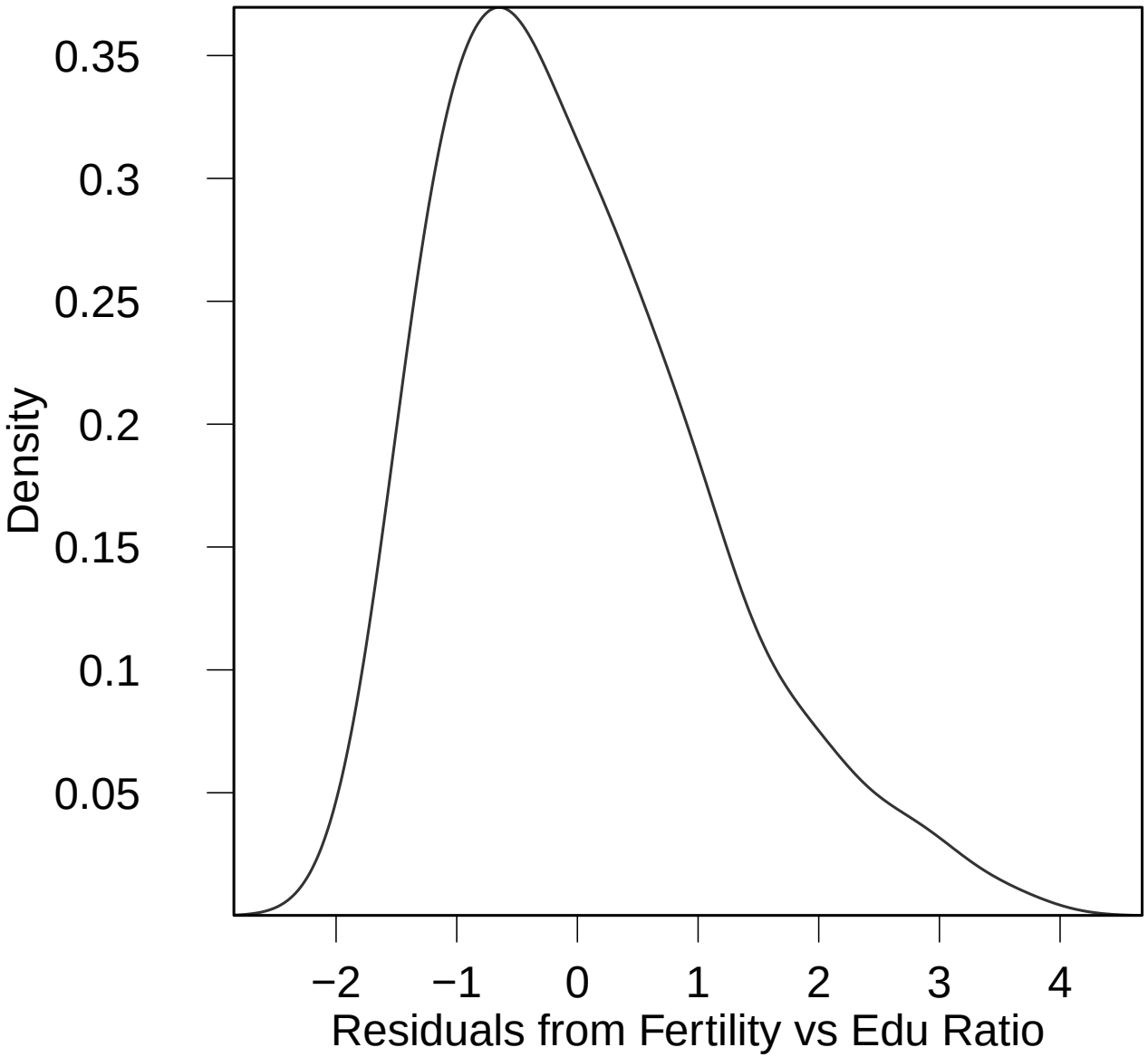


RMSE is better than R^2 because it can be compared across models and datasets—R^2 can't.

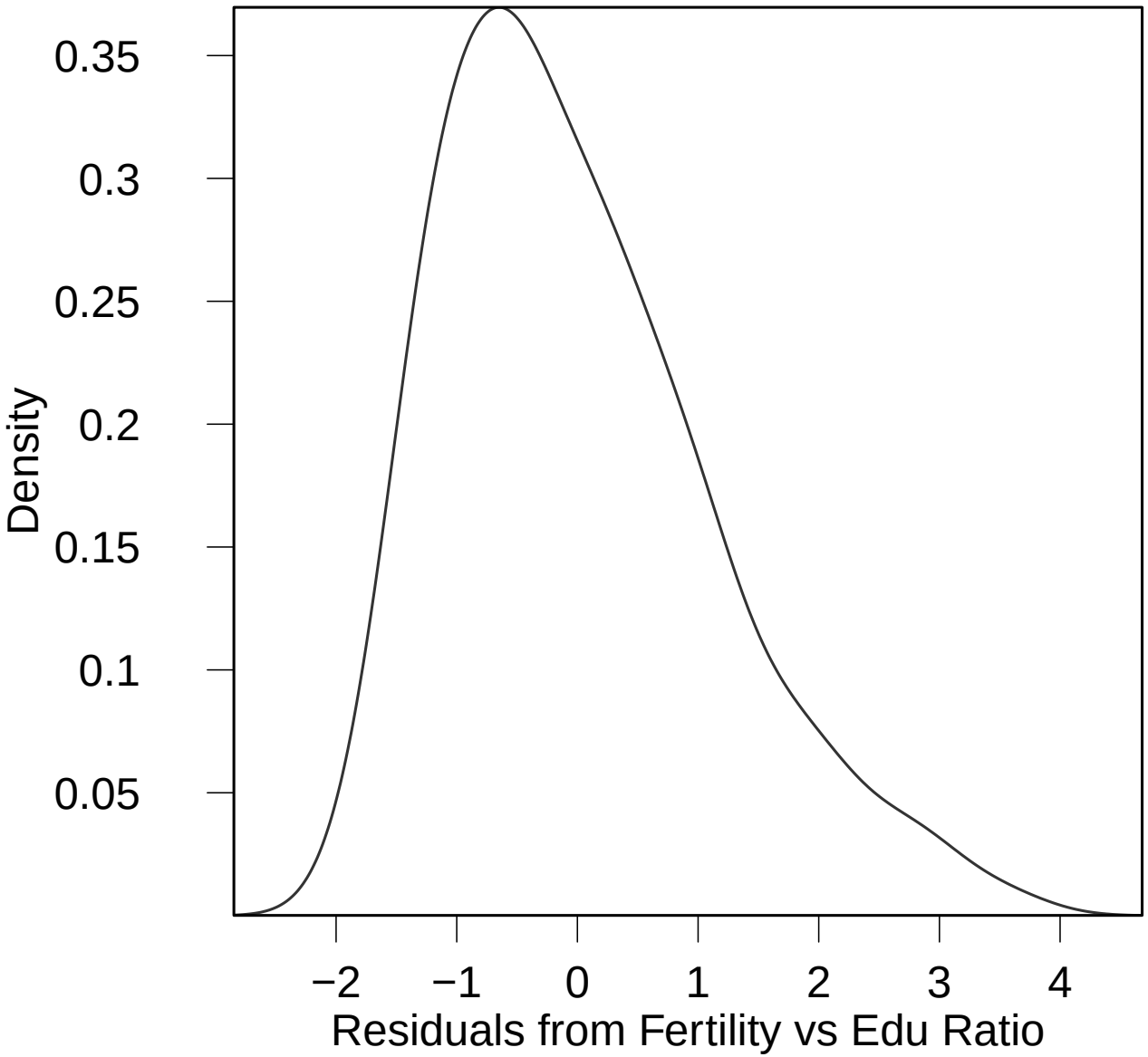


RMSE is better than R^2 because it can be compared across models and datasets—R^2 can't.

A question: we assumed the errors would be Normal—are they?

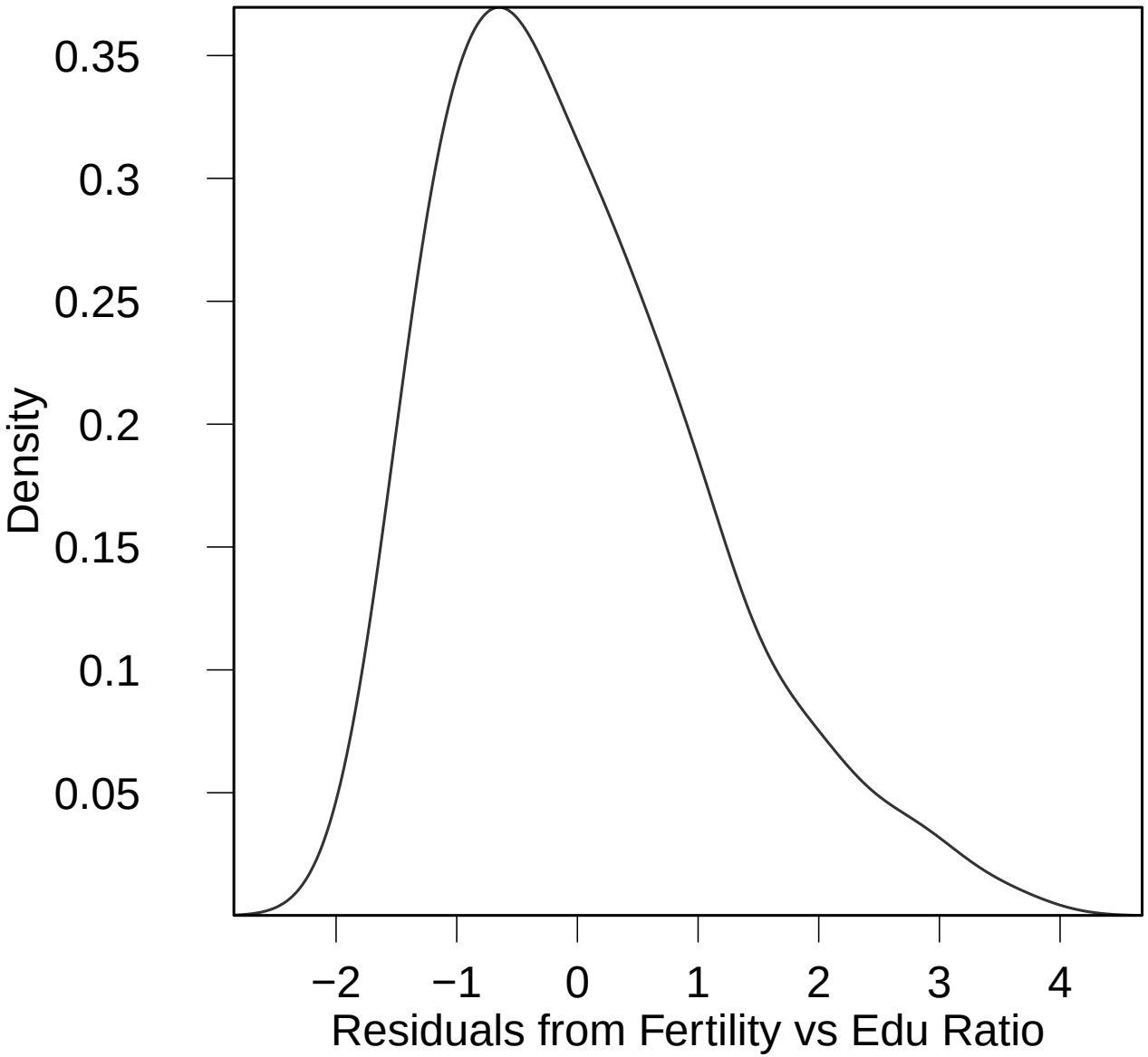


Recall that linear regression assumes the ε_i's are Normally distributed.



Recall that linear regression assumes the ε_i's are Normally distributed.

We do *not* assume that y_i follows a bell curve, except after controlling for x_i



Recall that linear regression assumes the ε_i's are Normally distributed.

We do *not* assume that y_i follows a bell curve, except after controlling for x_i

Do the residuals appear Normally distributed in this case?

Uncertainty of $\hat{\beta}$

When estimating a mean or difference of means, we worried that by chance, our sample might not reflect the population

That's a worry in linear regression as well

Does $\hat{\beta}$ estimated from our sample reflect the true population β?

Or did we get an unusual result due to sampling variability?

Uncertainty of $\hat{\beta}$

As with estimating a mean, we can calculate the *standard error* of $\hat{\beta}$

Uncertainty of $\hat{\beta}$

As with estimating a mean, we can calculate the *standard error* of $\hat{\beta}$

$\text{se}(\hat{\beta})$ is the amount we expect to miss the population β by on average over regression using repeated samples

Uncertainty of $\hat{\beta}$

As with estimating a mean, we can calculate the *standard error* of $\hat{\beta}$

$\text{se}(\hat{\beta})$ is the amount we expect to miss the population β by on average over regression using repeated samples

Remarkably, the $\hat{\beta}$'s themselves are Normally distributed, no matter what we are modeling

Uncertainty of $\hat{\beta}$

As with estimating a mean, we can calculate the *standard error* of $\hat{\beta}$

$\text{se}(\hat{\beta})$ is the amount we expect to miss the population β by on average over regression using repeated samples

Remarkably, the $\hat{\beta}$'s themselves are Normally distributed, no matter what we are modeling

So we can use a *t*-test to see if our $\hat{\beta}$'s would differ from the null hypothesis purely by chance

Uncertainty of $\hat{\beta}$

As with estimating a mean, we can calculate the *standard error* of $\hat{\beta}$

$\text{se}(\hat{\beta})$ is the amount we expect to miss the population β by on average over regression using repeated samples

Remarkably, the $\hat{\beta}$'s themselves are Normally distributed, no matter what we are modeling

So we can use a *t*-test to see if our $\hat{\beta}$'s would differ from the null hypothesis purely by chance

Usually, we will consider the null hypothesis to be $\beta^{\text{null}} = 0$, but sometimes we might want a different null

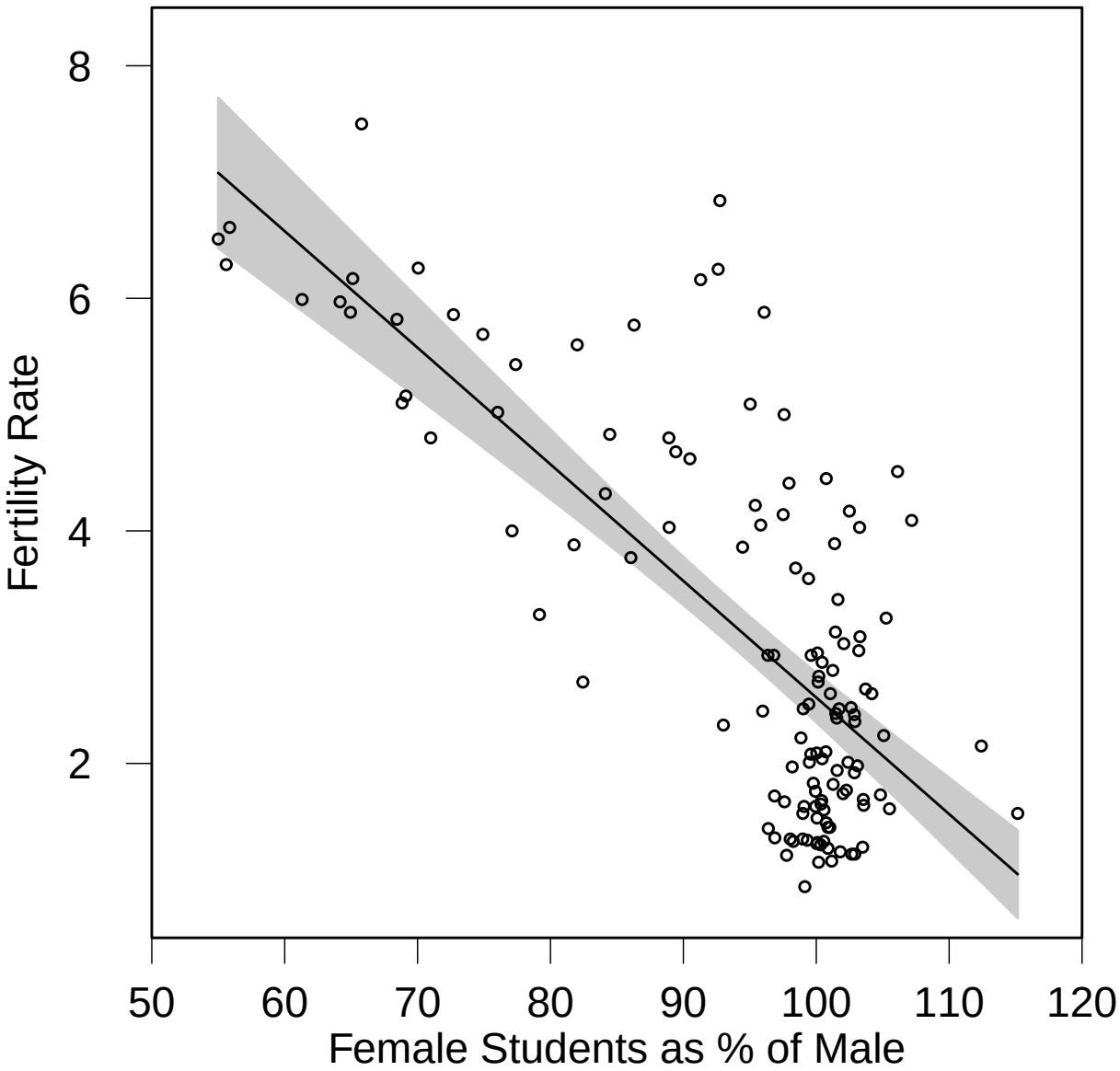
Uncertainty of $\hat{\beta}$

We can also construct confidence intervals around $\hat{\beta}_0$ and $\hat{\beta}_1$

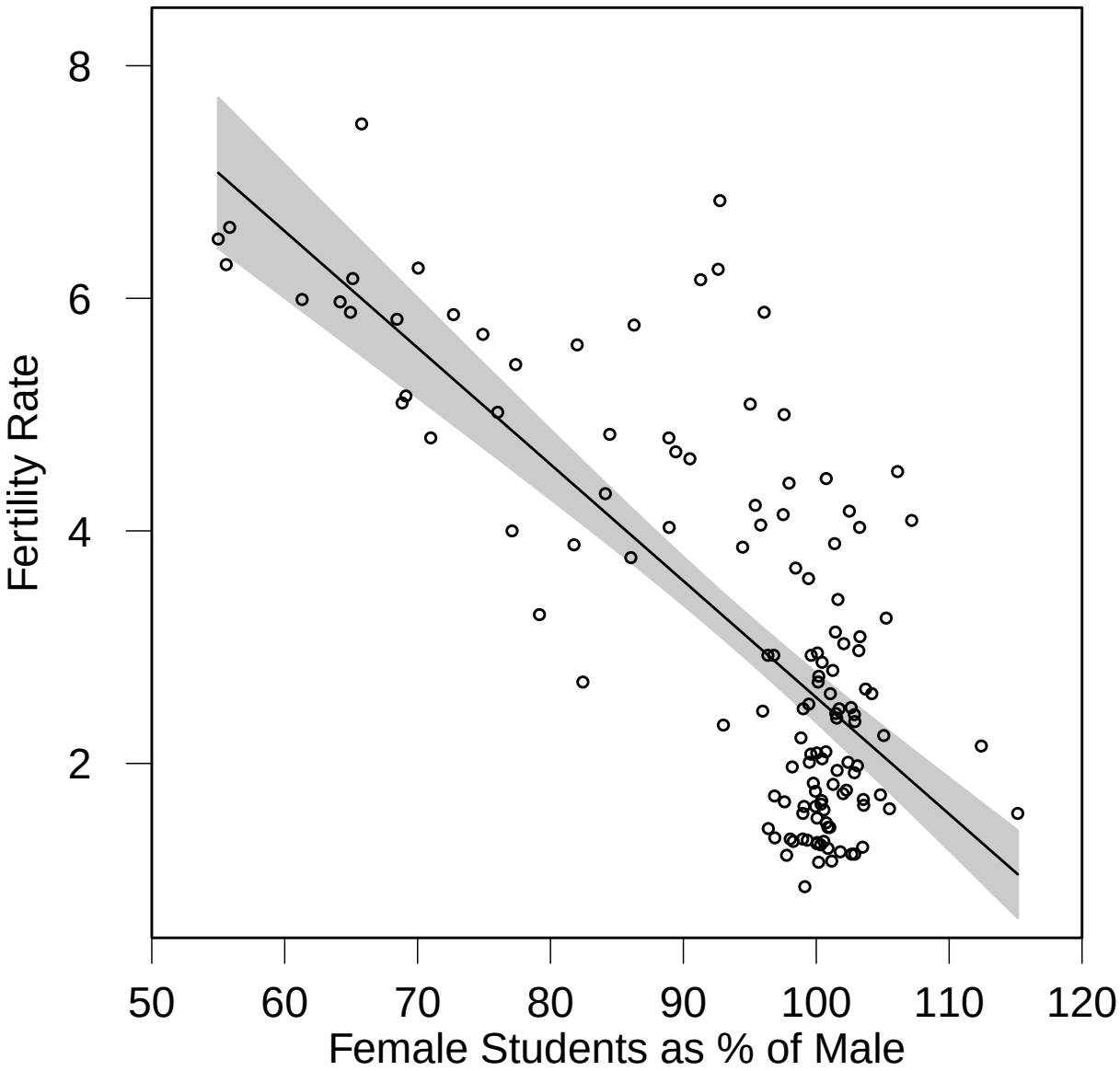
These CIs reflect the uncertainty created by randomly sampling our data from the population

95% of the time, the true population β's should lie in the 95% confidence intervals

Roughly, these intervals will be ± 2 standard errors, if we have a lot of data

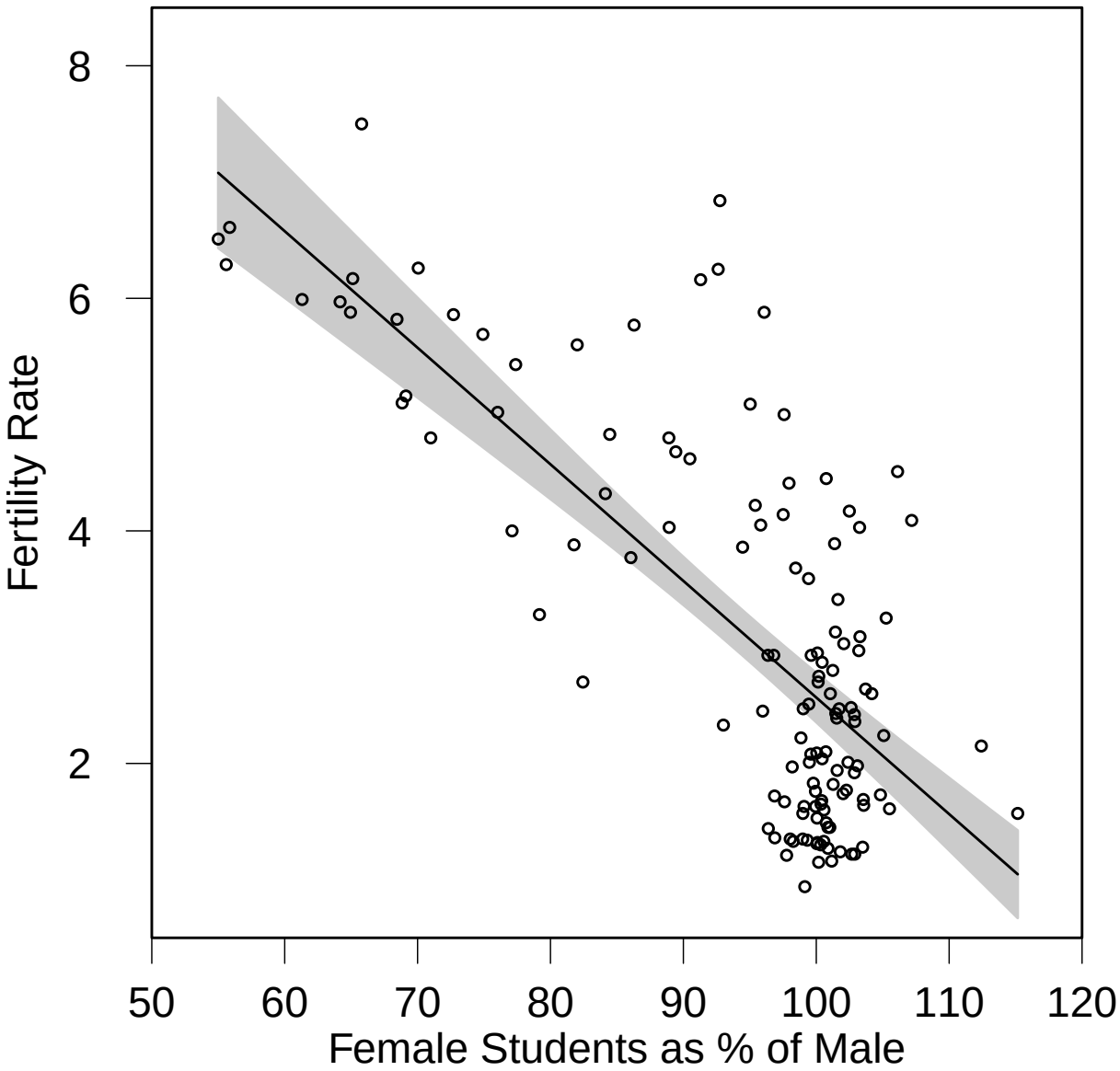


The standard errors of $\hat{\beta}$ reflect the fact that the true best fit line for the population lies within range of the estimated line

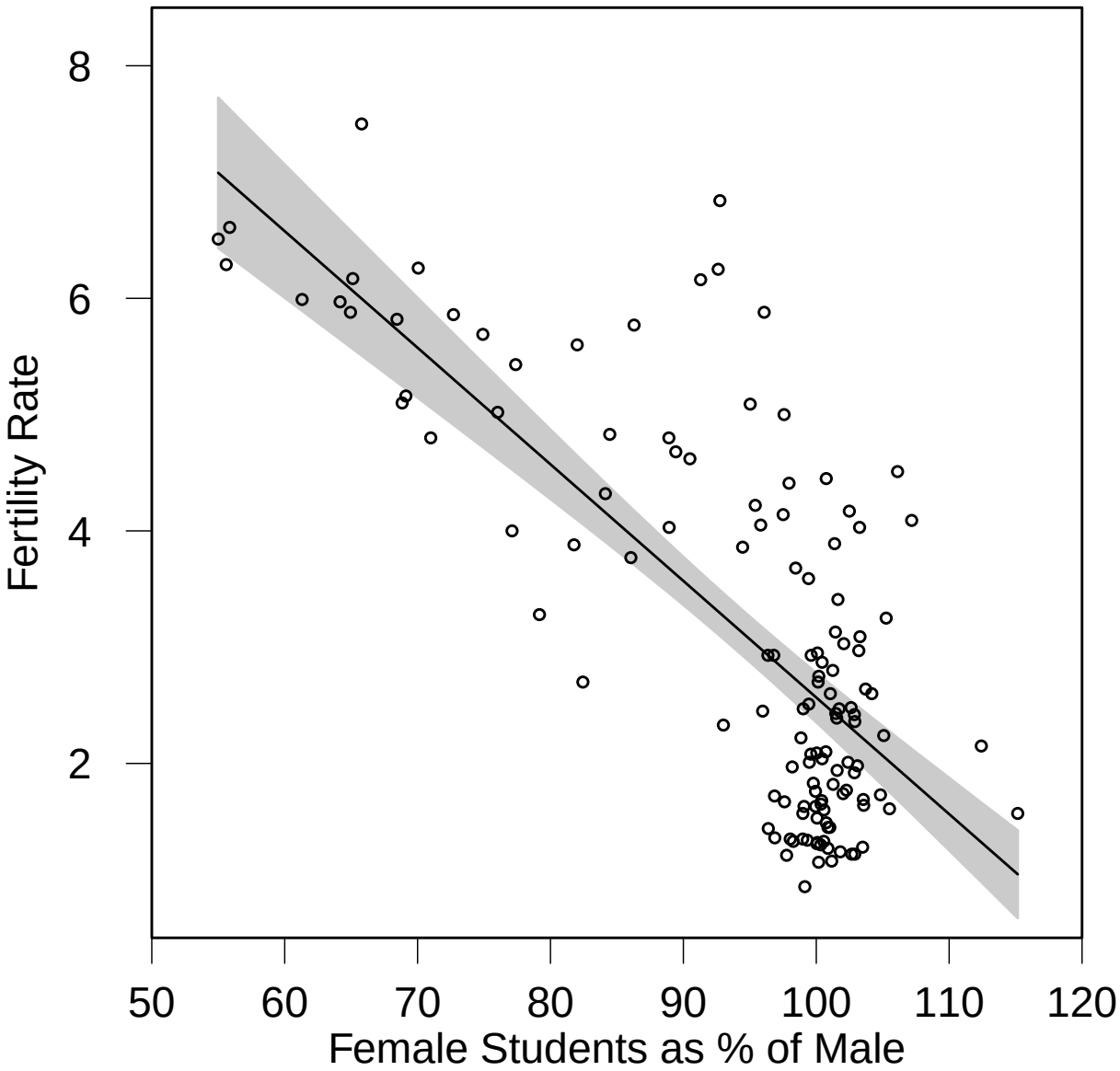


The standard errors of $\hat{\beta}$ reflect the fact that the true best fit line for the population lies within range of the estimated line

We can capture this "wiggle room" graphically

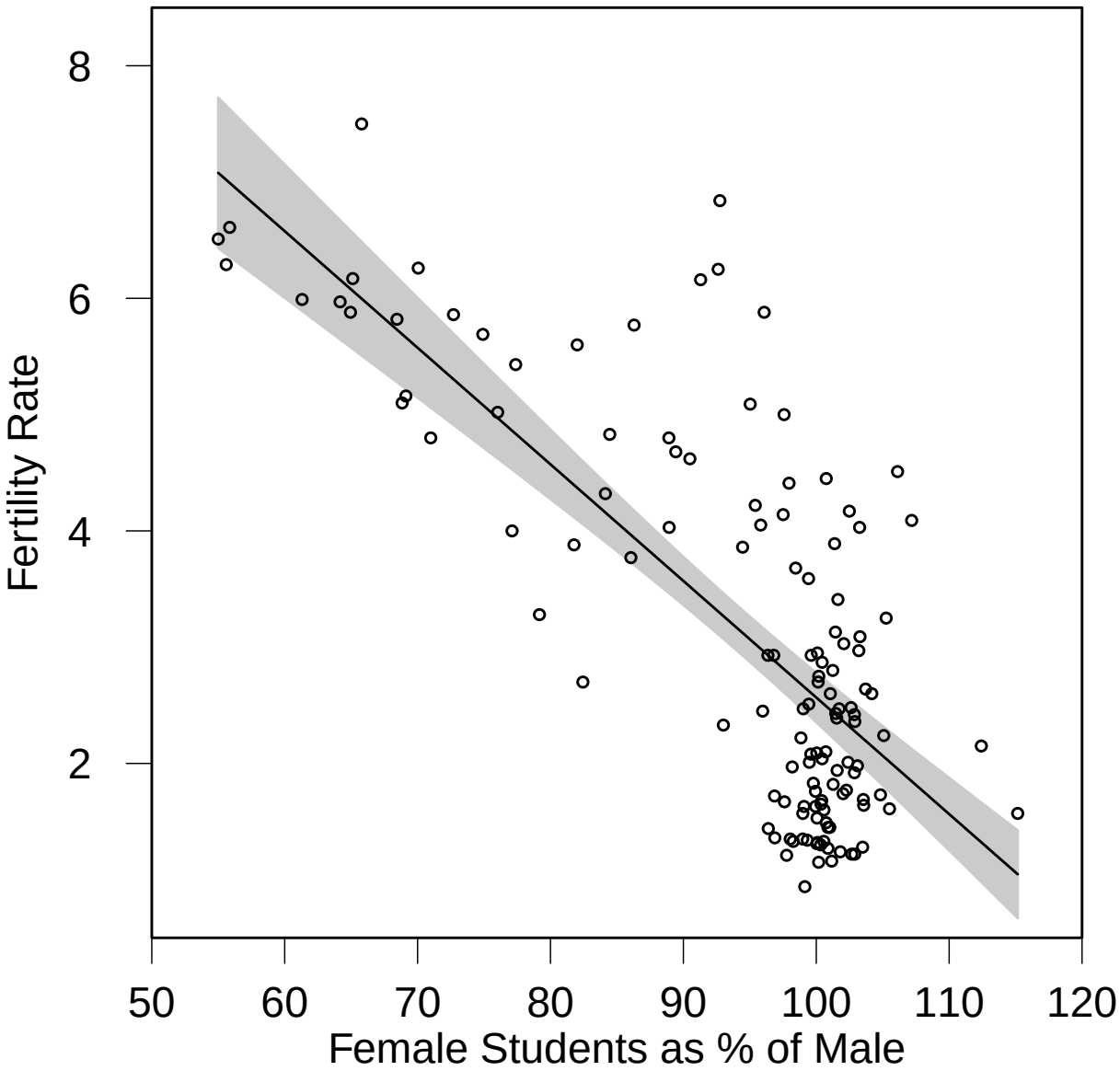


Why don't 95% of the points lie inside this interval?



Why don't 95% of the points lie inside this interval?

Because of fundametal uncertainty, or RMSE



Why don't 95% of the points lie inside this interval?

Because of fundametal uncertainty, or RMSE

The CIs just measure uncertainty in the best fit line, not in the data itself

A standard regression table

Regression of Fertility on Education Ratio

Variable	Estimates	se	t-stat	p-value
Intercept	12.59	(0.75)	16.75	<0.001
Education Ratio	-0.10	(0.01)	-12.71	<0.001
N	130			
R^2	0.56			
RMSE	1.12			

The most common presentation of a linear regression is the above table

Usually, graphics are more informative and easier to read, but older articles rely heavily on this tabular format

Understanding these tables will be important for the final exam. Let's take this one apart

A standard regression table

Regression of Fertility on Education Ratio

Variable	Estimates	se	t-stat	p-value
Intercept	12.59	(0.75)	16.75	<0.001
Education Ratio	-0.10	(0.01)	-12.71	<0.001
N	130			
R^2	0.56			
RMSE	1.12			

The middle of the table contains several important quantities regarding our independent variable(s):

- *Estimates*: the $\hat{\beta}$'s, or regression coefficients
- *se*: the standard errors of $\hat{\beta}$
- *t-stat*: the t-statistic for the regression coefficient, or $\hat{\beta}/\text{se}(\hat{\beta})$
- *p-value*: the probability of seeing such a large t-stat by chance

Regression of Fertility on GDP per capita

Variable	Estimates	95% Confidence Interval Lower	Upper
Intercept	12.59	[11.11,	14.08]
Education Ratio	-0.10	[-0.12,	-0.08]
N	130		
R^2	0.36		
RMSE	1.35		

Just as will our other estimates, we can construct *confidence intervals* around our $\hat{\beta}$'s

Our results show 95% confidence that a 1 unit (1%) increase in education of girls relative to boys lowers fertility by between 0.08 and 0.12 children per woman

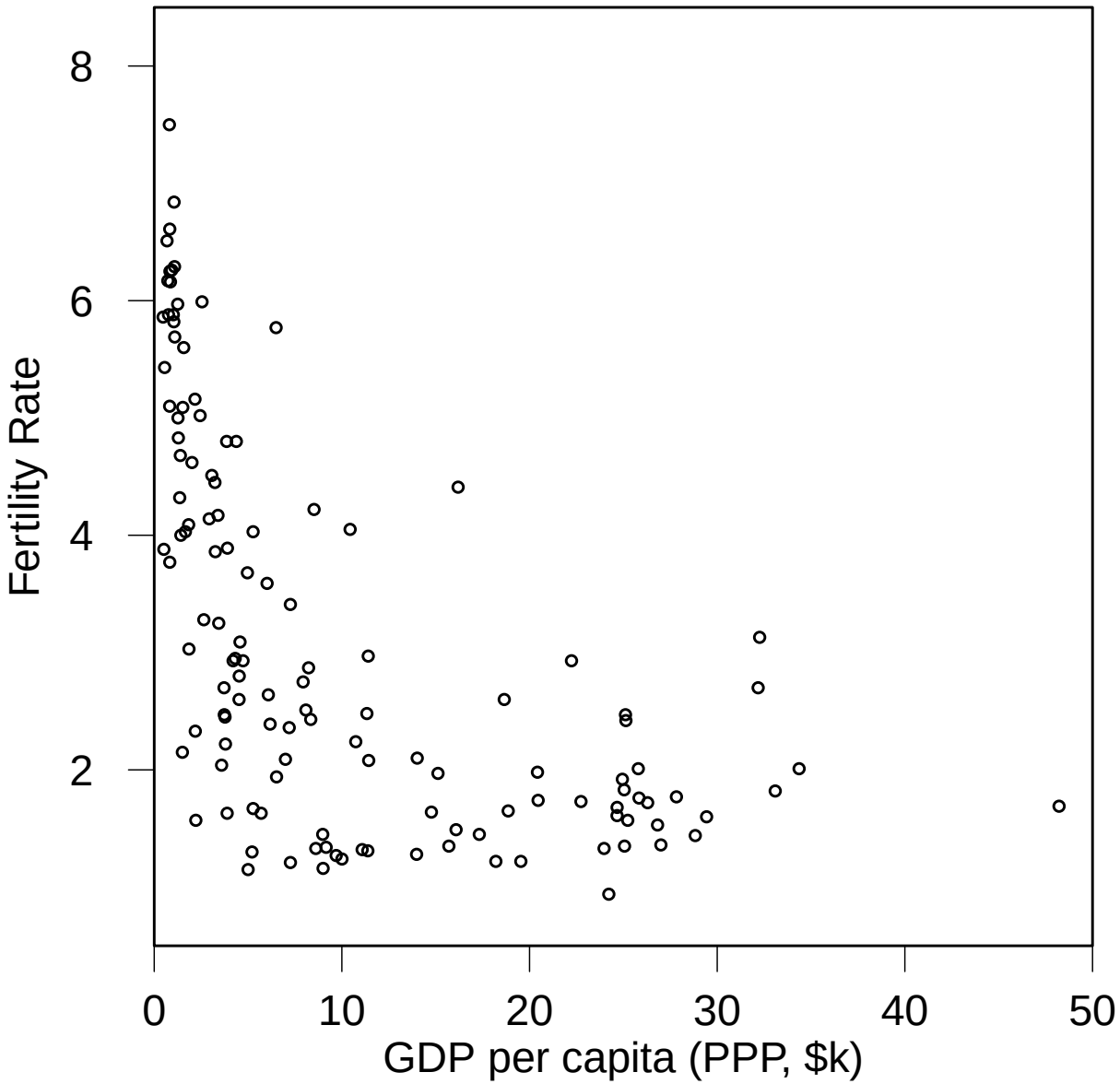
We would only expect the truth to lie outside this interval in 1 of 20 random samples

Wait a minute!

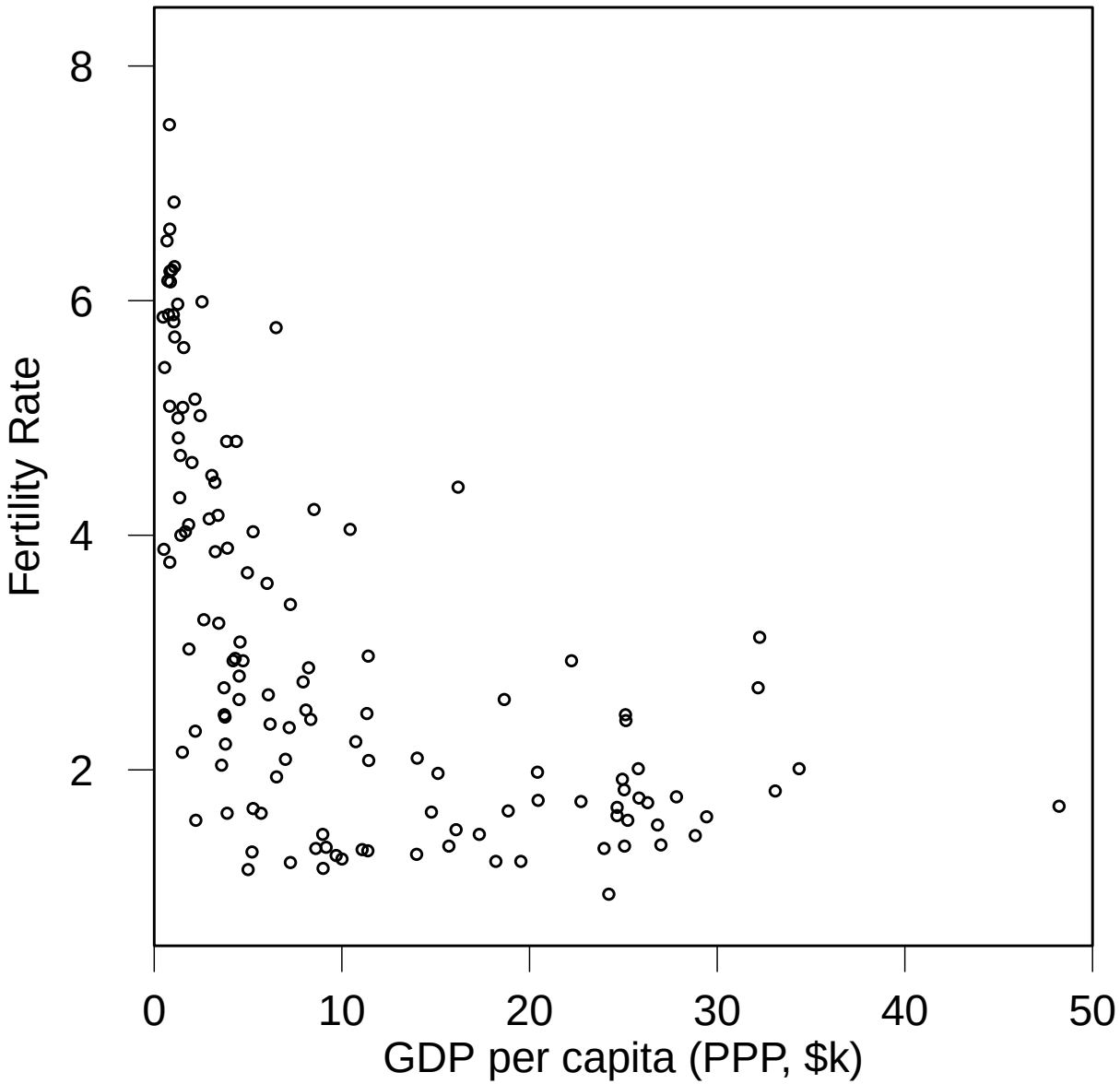
When we considered the relationship of female education and fertility, we also hypothesized an effect of GDP per capita

We suspected this might be an indirect effect, flowing through female education

Can we use regression to check for an effect of GDP?

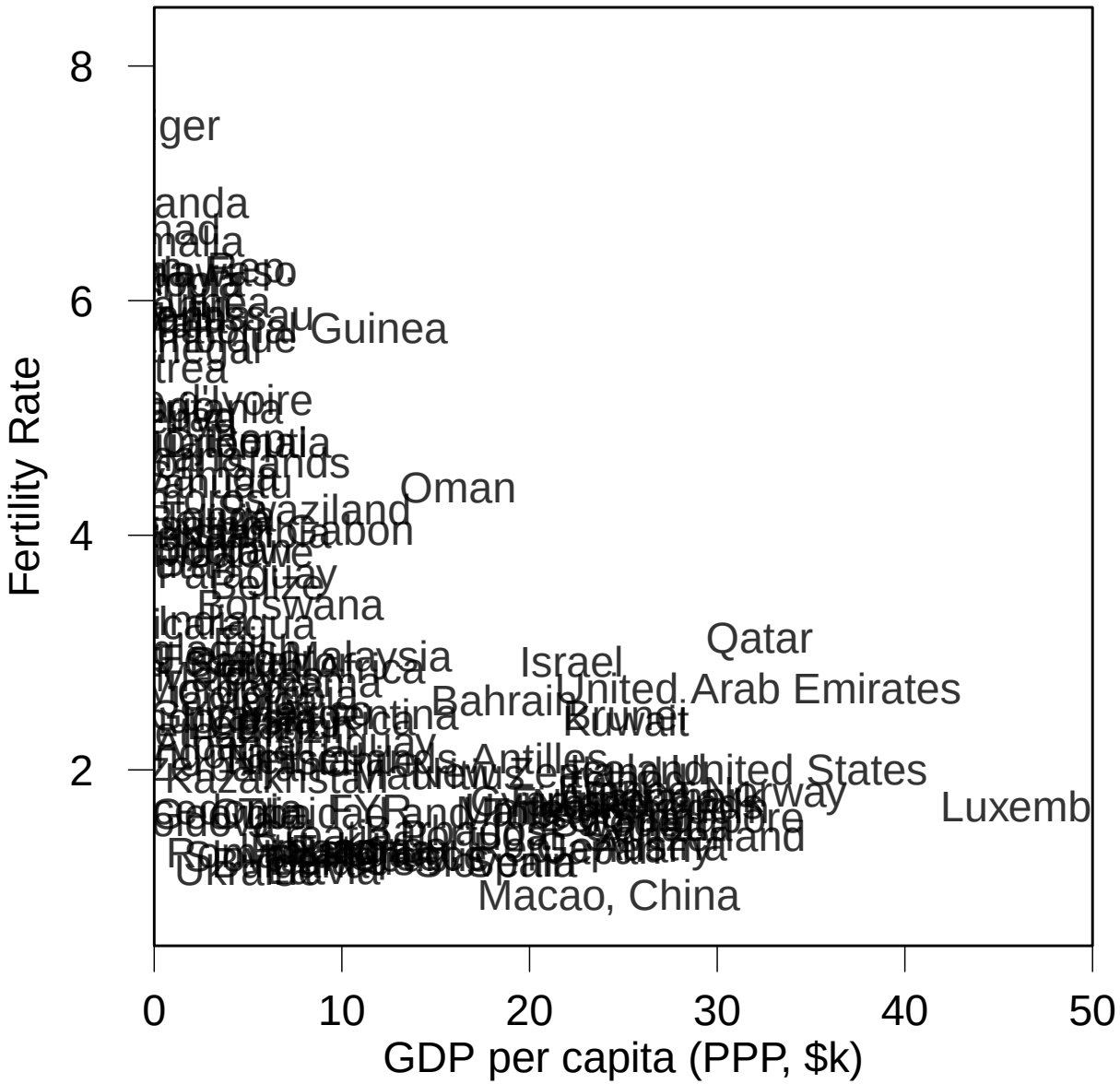


Let's regress Fertility on GDP per capita

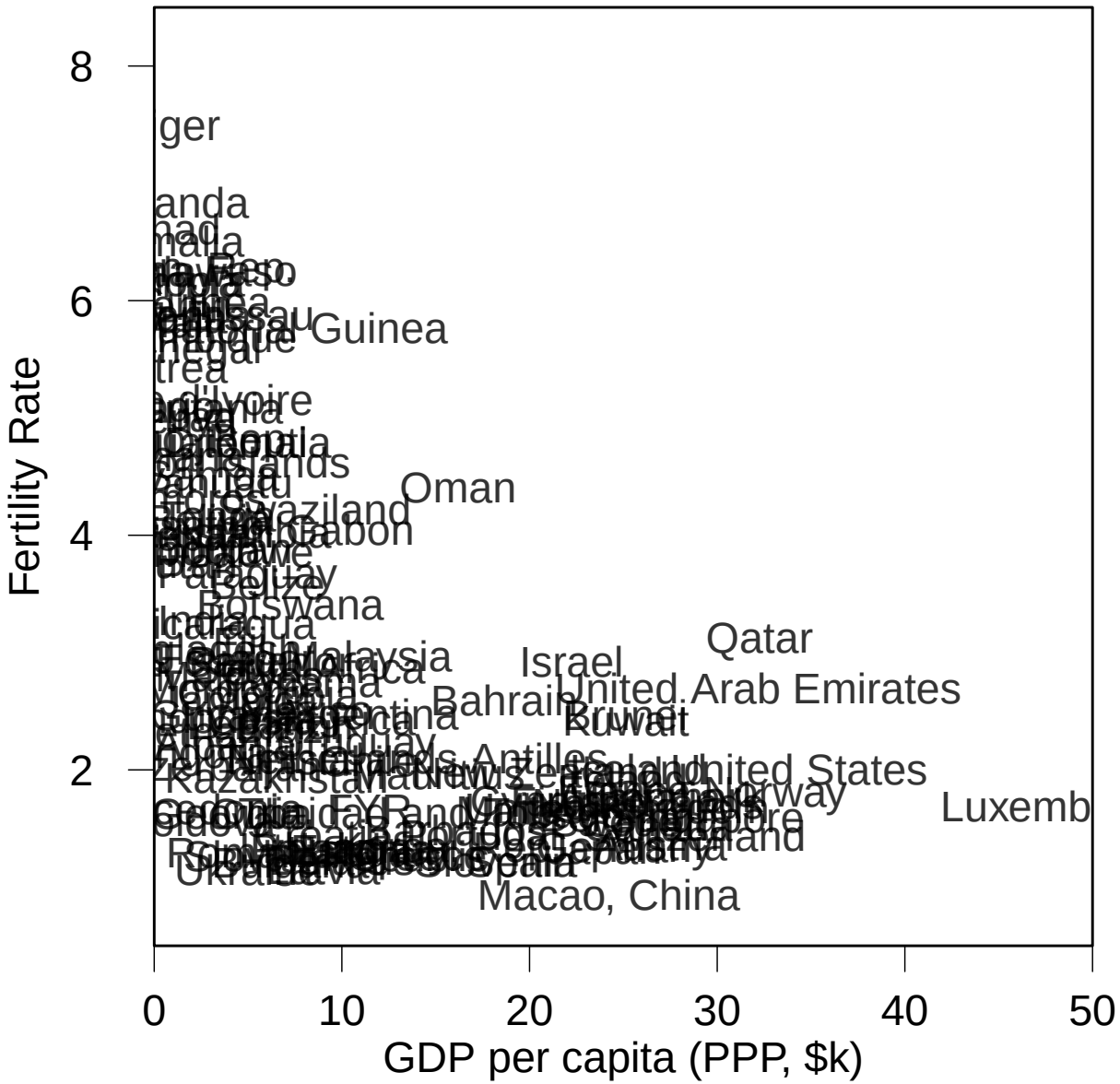


Let's regress Fertility on GDP per capita

Does this scatterplot suggest a linear relationship?

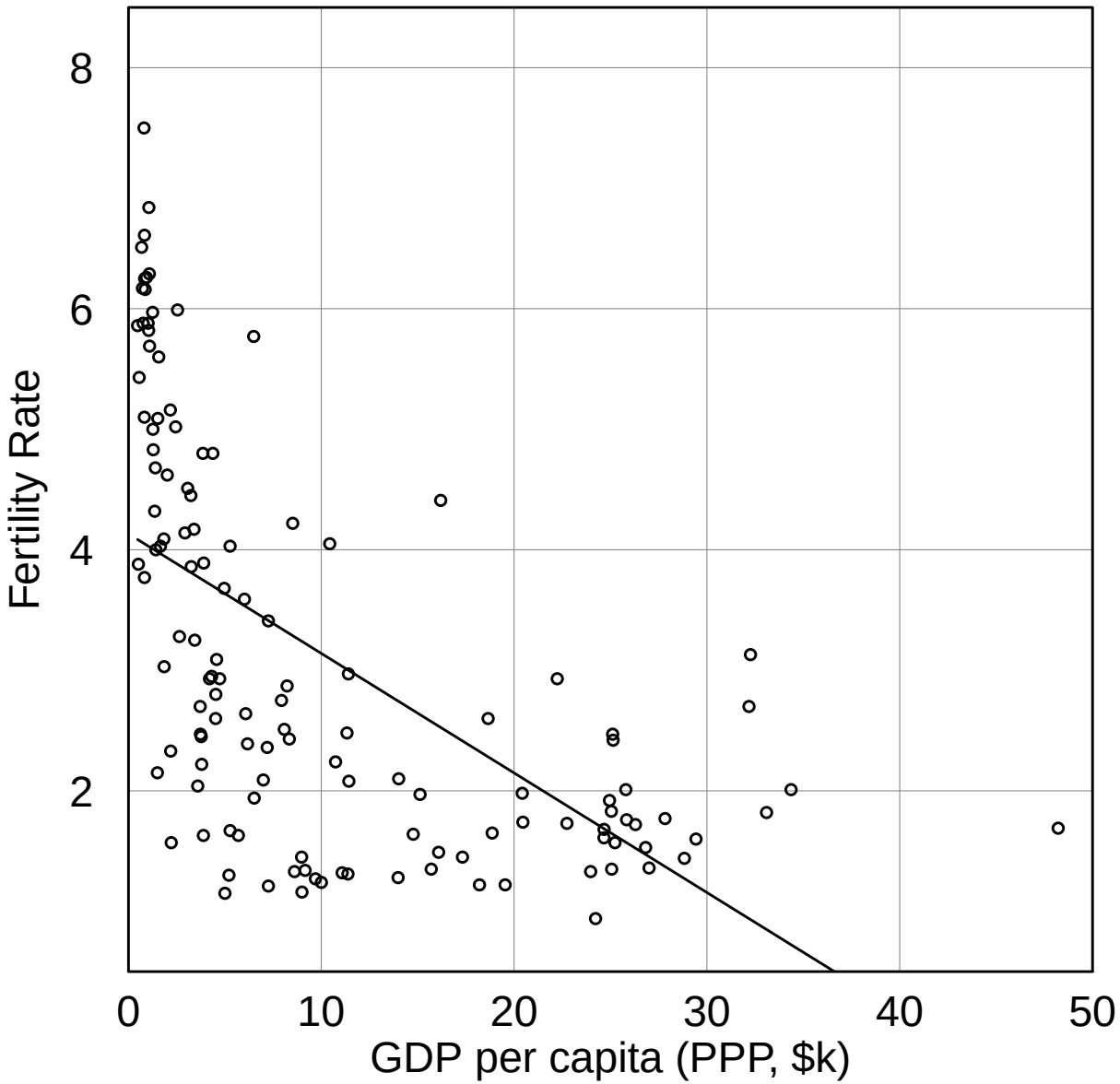


Not really. Next week we'll discuss solutions for curved relation-ships

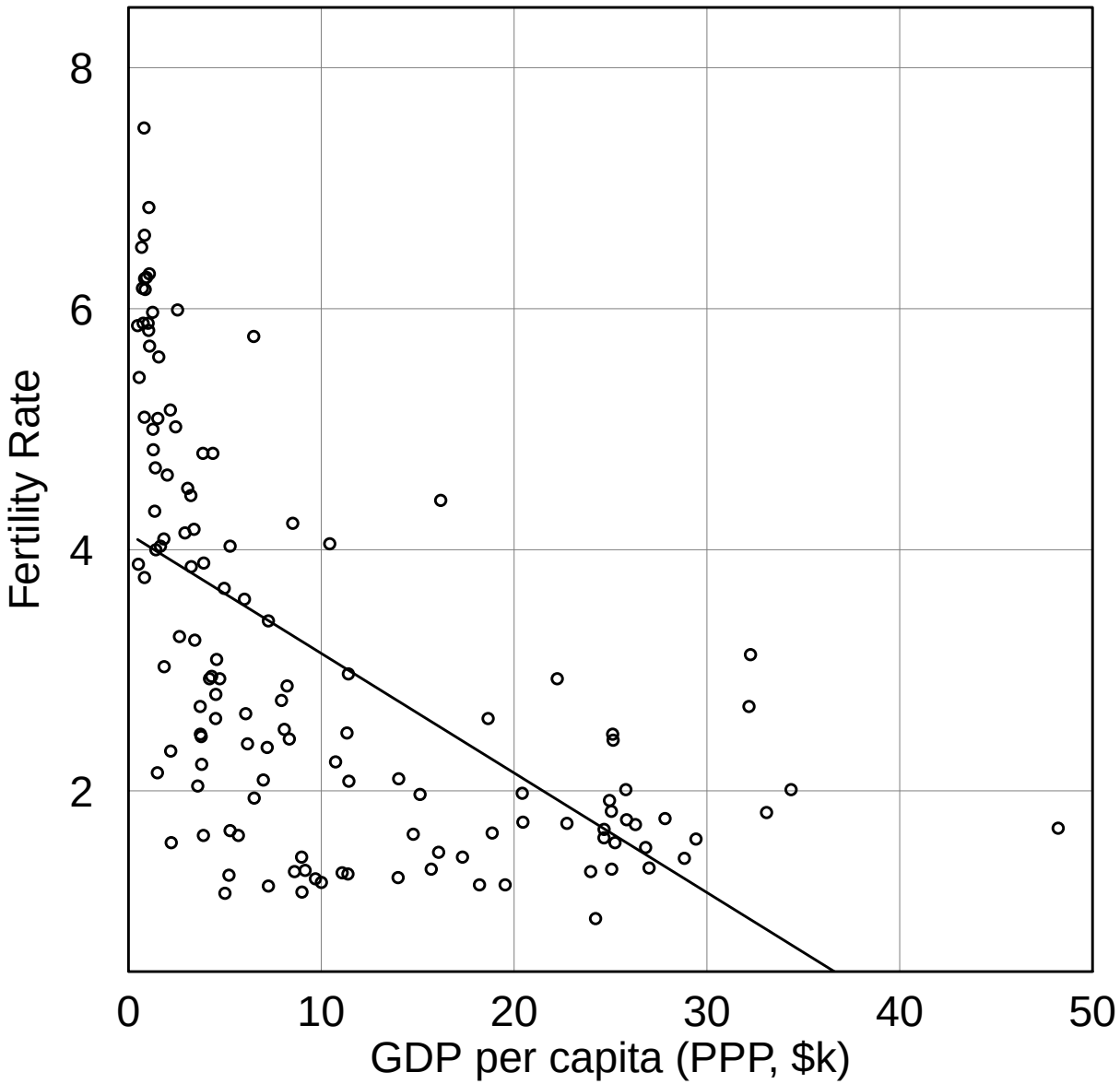


Not really. Next week we'll discuss solutions for curved relation-ships

For now, let's proceed with the best linear fit

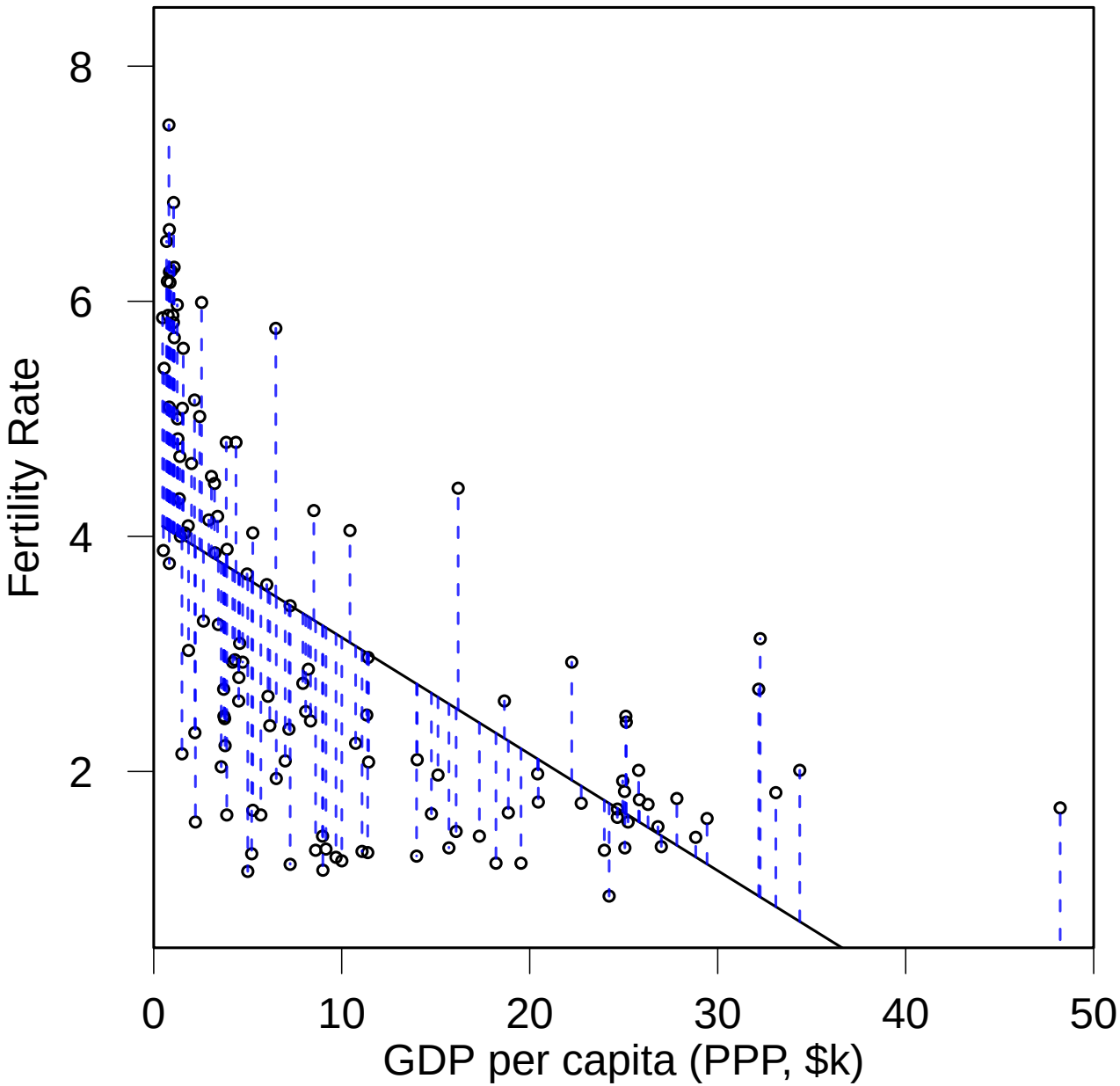


This is the least squares fit (What does that mean?)

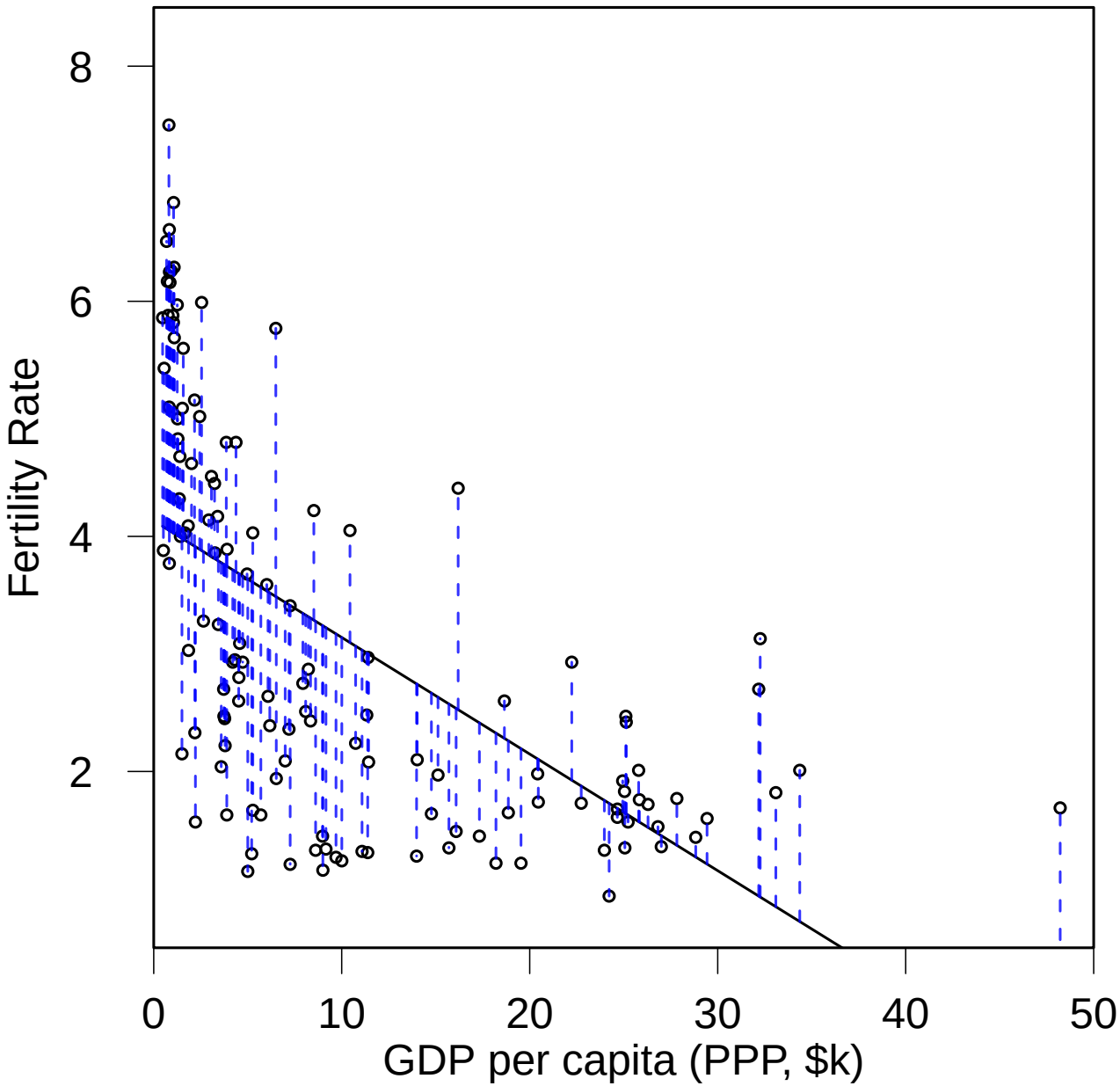


This is the least squares fit (What does that mean?)

How good does this fit look?

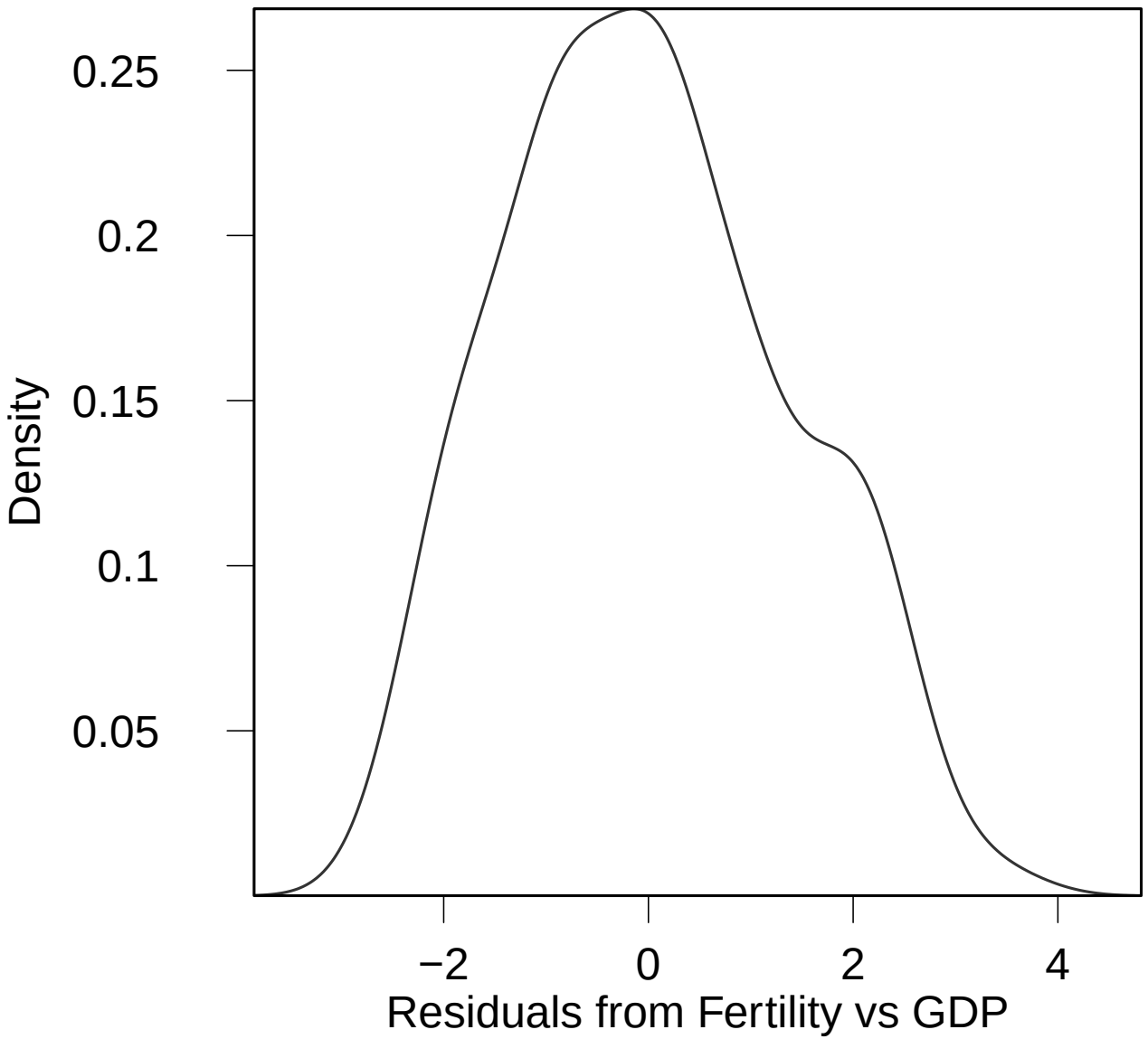


Can you imagine an alternative model that would reduce the sum of squared residuals further?

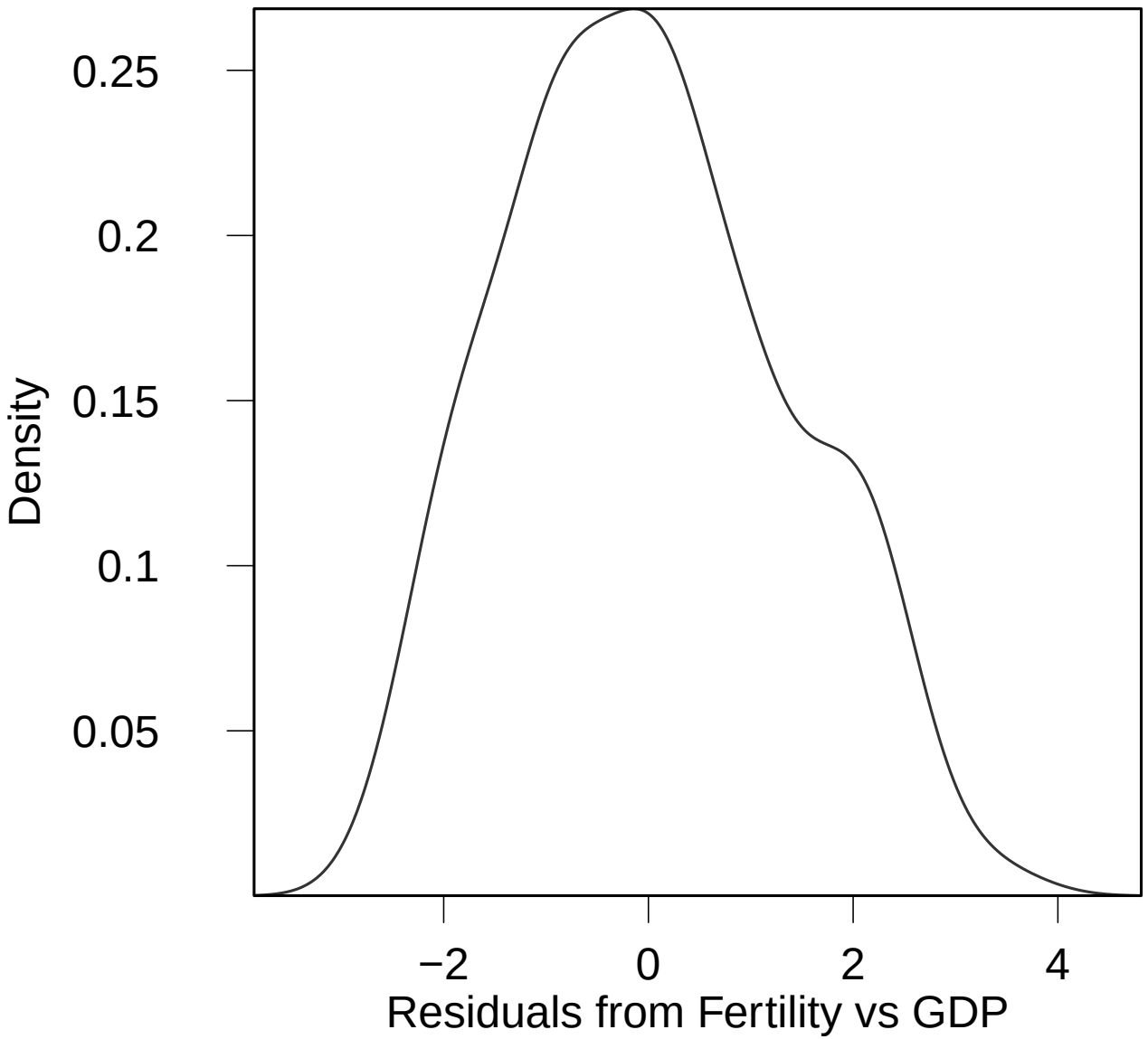


Can you imagine an alternative model that would reduce the sum of squared residuals further?

Perhaps a concave curve?

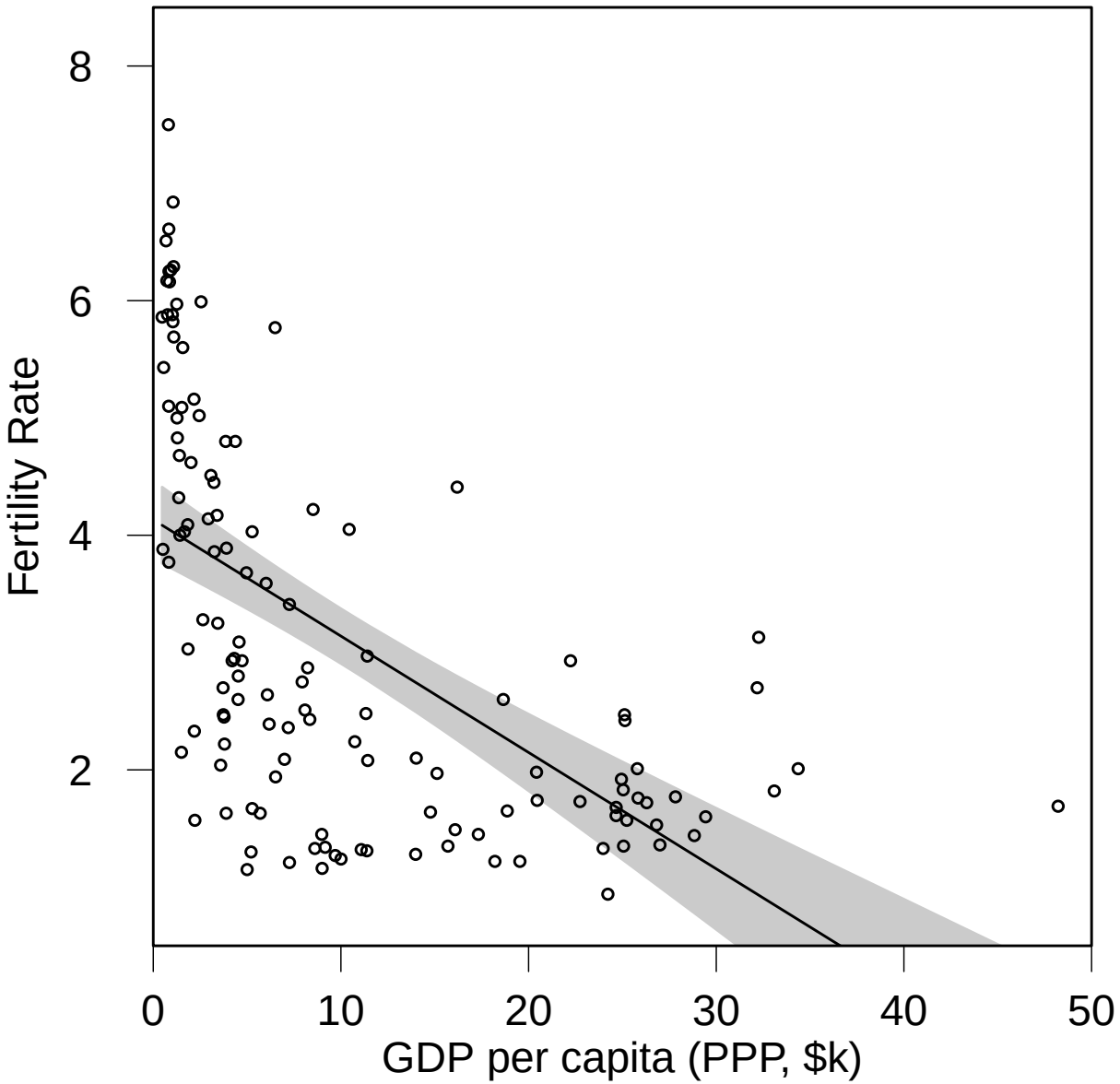


Do the residuals look Normally dis-tributed?

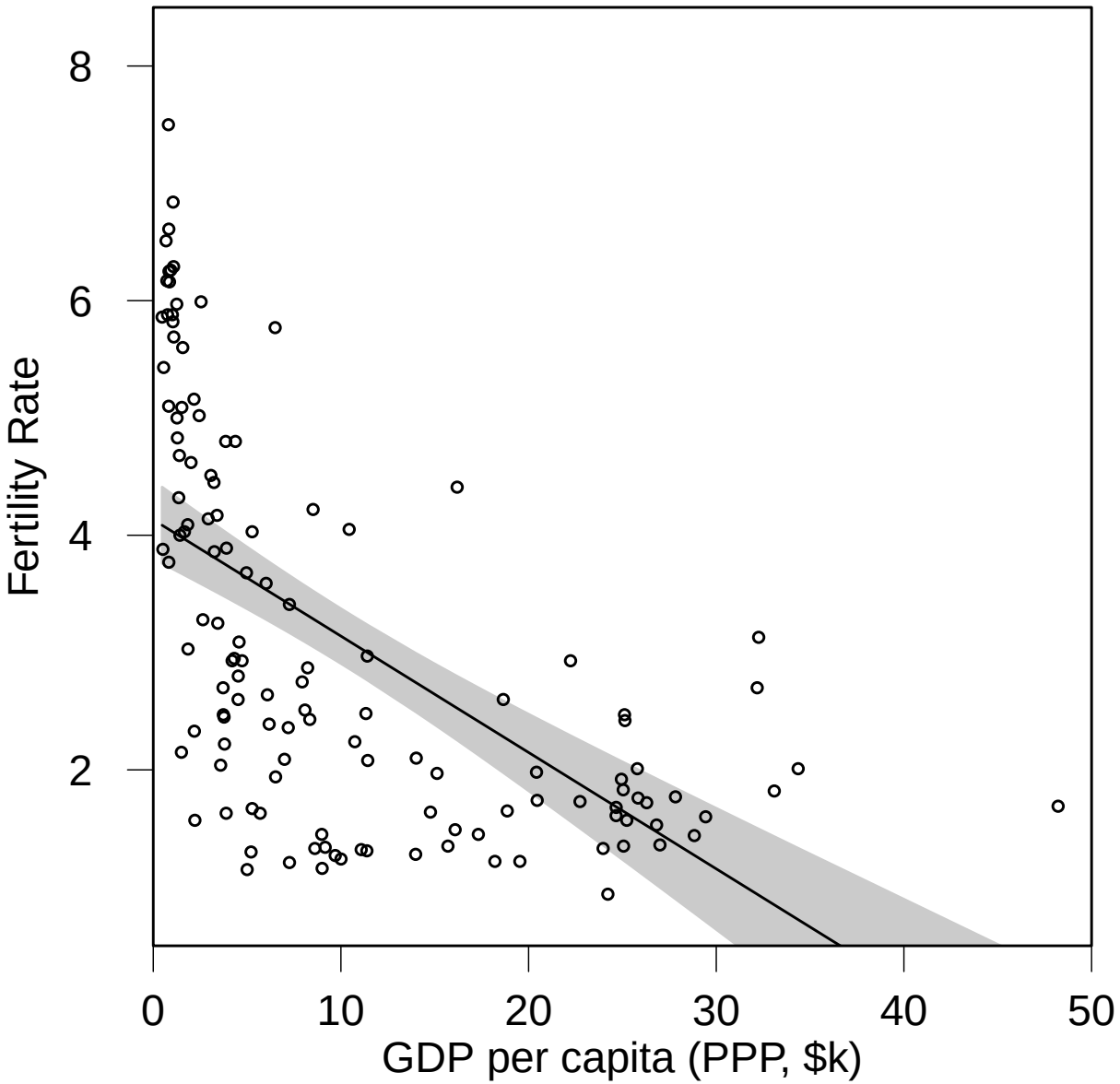


Do the residuals look Normally dis-tributed?

A strongly skewed distribution of errors is cause for concern. More next week



How do we interpret this 95% confidence interval?



How do we interpret this 95% confidence interval?

Why don't 95% of the points lie inside it?

Another regression table

Regression of Fertility on GDP per capita

Variable	Estimates	se	t-stat	p-value
Intercept	4.13	(0.17)	24.57	<0.001
GDP per capita ($k)	-0.10	(0.01)	-8.44	<0.001
N	130			
R^2	0.36			
RMSE	1.35			

How do we interpret this table?

Another regression table

Regression of Fertility on GDP per capita

Variable	Estimates	se	t-stat	p-value
Intercept	4.13	(0.17)	24.57	<0.001
GDP per capita ($k)	-0.10	(0.01)	-8.44	<0.001
N	130			
R^2	0.36			
RMSE	1.35			

1. How much do we expect Fertility to change when we increase GDP by $1000?

Another regression table

Regression of Fertility on GDP per capita

Variable	Estimates	se	t-stat	p-value
Intercept	4.13	(0.17)	24.57	<0.001
GDP per capita ($k)	-0.10	(0.01)	-8.44	<0.001
N	130			
R^2	0.36			
RMSE	1.35			

1. How much do we expect Fertility to change when we increase GDP by $1000? *decrease* by 0.1 children

2. What would Fertility be if GDP were $1000? $10,000? $30,000?

Another regression table

Regression of Fertility on GDP per capita

Variable	Estimates	se	t-stat	p-value
Intercept	4.13	(0.17)	24.57	<0.001
GDP per capita ($k)	-0.10	(0.01)	-8.44	<0.001
N	130			
R^2	0.36			
RMSE	1.35			

1. How much do we expect Fertility to change when we increase GDP by $1000? *decrease* by 0.1 children
2. What would Fertility be if GDP were $1000? $10,000? $30,000? 4.03, 3.13, and 1.13, respectively.
3. What would Fertility be if GDP were 0? Do you trust this estimate?

Another regression table

Regression of Fertility on GDP per capita

Variable	Estimates	se	t-stat	p-value
Intercept	4.13	(0.17)	24.57	<0.001
GDP per capita ($k)	-0.10	(0.01)	-8.44	<0.001
N	130			
R^2	0.36			
RMSE	1.35			

1. How much do we expect Fertility to change when we increase GDP by $1000? *decrease* by 0.1 children

2. What would Fertility be if GDP were $1000? $10,000? $30,000? 4.03, 3.13, and 1.13, respectively.

3. What would Fertility be if GDP were 0? Do you trust this estimate? 4.13. No—this is an extrapolation.

Another regression table

Regression of Fertility on GDP per capita

Variable	Estimates	se	t-stat	p-value
Intercept	4.13	(0.17)	24.57	<0.001
GDP per capita ($k)	-0.10	(0.01)	-8.44	<0.001
N	130			
R^2	0.36			
RMSE	1.35			

1. Suppose we drew another sample of countries. Would we expect to see a GDP different from zero in that case?

Another regression table

Regression of Fertility on GDP per capita

Variable	Estimates	se	t-stat	p-value
Intercept	4.13	(0.17)	24.57	<0.001
GDP per capita ($k)	-0.10	(0.01)	-8.44	<0.001
N	130			
R^2	0.36			
RMSE	1.35			

1. Suppose we drew another sample of countries. Would we expect to see a GDP different from zero in that case? Yes.

2. Why?

Another regression table

Regression of Fertility on GDP per capita

Variable	Estimates	se	t-stat	p-value
Intercept	4.13	(0.17)	24.57	<0.001
GDP per capita ($k)	-0.10	(0.01)	-8.44	<0.001
N	130			
R^2	0.36			
RMSE	1.35			

1. Suppose we drew another sample of countries. Would we expect to see a GDP different from zero in that case? Yes.
2. Why?
   The se is small relative to $\hat{\beta}$, so the true β is probably far from 0.
3. How likely is it that we would see a t statistic this large if $\beta = 0$?

Another regression table

Regression of Fertility on GDP per capita

Variable	Estimates	se	t-stat	p-value
Intercept	4.13	(0.17)	24.57	<0.001
GDP per capita ($k)	-0.10	(0.01)	-8.44	<0.001
N	130			
R^2	0.36			
RMSE	1.35			

1. Suppose we drew another sample of countries. Would we expect to see a GDP different from zero in that case? Yes.

2. Why?
   The se is small relative to $\hat{\beta}$, so the true β is probably far from 0.

3. How likely is it that we would see a t statistic this large if $\beta = 0$?
   Very unlikely—less than 1 in 1000 samples.

Another regression table

Regression of Fertility on GDP per capita

Variable	Estimates	95% Confidence Interval Lower	Upper
Intercept	4.13	[3.80,	4.46]
GDP per capita ($k)	-0.10	[-0.12,	-0.08]
N	130		
R^2	0.36		
RMSE	1.35		

1. What do these confidence intervals mean?

Another regression table

Regression of Fertility on GDP per capita

Variable	Estimates	95% Confidence Interval Lower	Upper
Intercept	4.13	[3.80,	4.46]
GDP per capita ($k)	-0.10	[-0.12,	-0.08]
N	130		
R^2	0.36		
RMSE	1.35		

1. What do these confidence intervals mean?
   In 95% of random samples, the true β's will lie inside these intervals

Another regression table

Regression of Fertility on GDP per capita

Variable	Estimates	se	t-stat	p-value
Intercept	4.13	(0.17)	24.57	<0.001
GDP per capita ($k)	-0.10	(0.01)	-8.44	<0.001
N	130			
R^2	0.36			
RMSE	1.35			

1. How much of the variance in Fertility does this model explain?

Another regression table

Regression of Fertility on GDP per capita

Variable	Estimates	se	t-stat	p-value
Intercept	4.13	(0.17)	24.57	<0.001
GDP per capita ($k)	-0.10	(0.01)	-8.44	<0.001
N	130			
R^2	0.36			
RMSE	1.35			

1. How much of the variance in Fertility does this model explain?
   36 percent

2. When using the model to predict fertility for a specific country, how much does it miss by on average?

Another regression table

Regression of Fertility on GDP per capita

Variable	Estimates	se	t-stat	p-value
Intercept	4.13	(0.17)	24.57	<0.001
GDP per capita ($k)	-0.10	(0.01)	-8.44	<0.001
N	130			
R^2	0.36			
RMSE	1.35			

1. How much of the variance in Fertility does this model explain?
   36 percent
2. When using the model to predict fertility for a specific country, how much does it miss by on average? 1.35
3. How many cases were used in this analysis?

Another regression table

Regression of Fertility on GDP per capita

Variable	Estimates	se	t-stat	p-value
Intercept	4.13	(0.17)	24.57	<0.001
GDP per capita ($k)	-0.10	(0.01)	-8.44	<0.001
N	130			
R^2	0.36			
RMSE	1.35			

1. How much of the variance in Fertility does this model explain?
   36 percent
2. When using the model to predict fertility for a specific country, how much does it miss by on average? 1.35
3. How many cases were used in this analysis? 130

Foreshadowing

How do we reconcile our two sets of results?

Which model, if any, is right?

To solve this conundrum, we need *multiple regression*:
A method for regressing a dependent variable on several independent variables at once

Then, at last, we can say something about confounders

Fortunately, all of today's concepts will carry over to multiple regression

Important linear regression concepts

Regression coefficient	β
Estimate of regression coefficient	$\hat{\beta}$
Standard error of est. of reg. coef.	$\text{se}(\hat{\beta})$
Fitted values	$\hat{y}_i$
Regression errors	ε_i
Residuals	ε_i
Coefficient of determination	R^2
Sum of squared errors (SSE)	$\sum_{i=1}^{n} \varepsilon_i$
Regression sum of squares (SSR)	$\sum_{i=1}^{n} \hat{y}_i - \bar{y}$