POLS 205
Political Science as a Social Science

Analyzing Multivariate Relationships
Part 3

Christopher Adolph

University of Washington, Seattle

May 26, 2010

Outline

More on Transformed Variables in Linear Regression

Goodness of Fit using Cross-validation

Regression when the Dependent Variable is Binary

Dealing with Outliers

Correlation and Causation Revisited

Regression with transformed variables

Last time we saw an example in which logging the response variable was appropriate

We expected percentage changes in wealth to depend on the unit changes in our covariates

This suggested exponential growth, or a model in which the log of wealth is a function of linear covariates

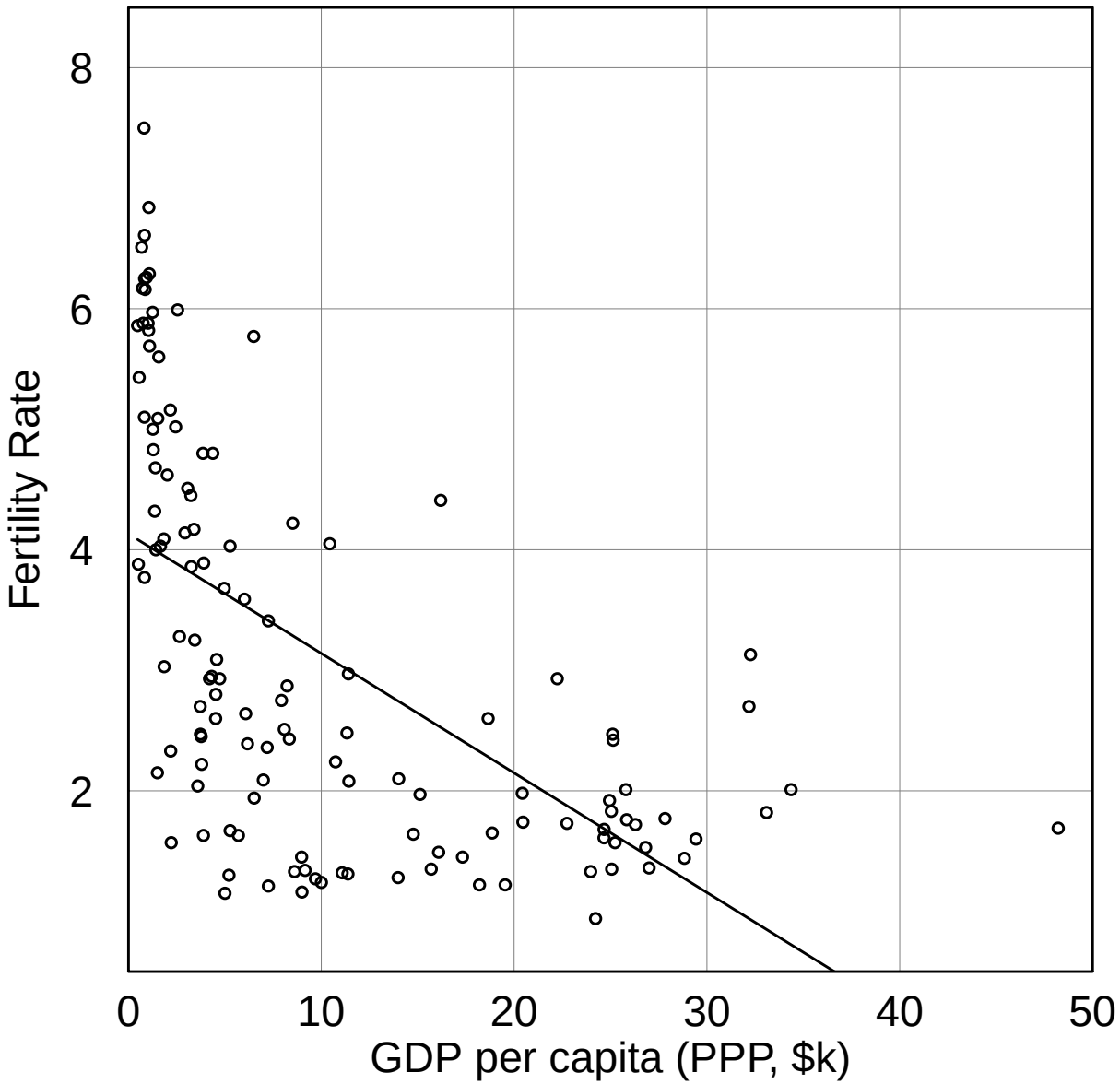
Sometimes, we will instead expect percentage changes in a *covariate* to lead to level changes in our response variable

Regression of Fertility on Education Ratio & GDP

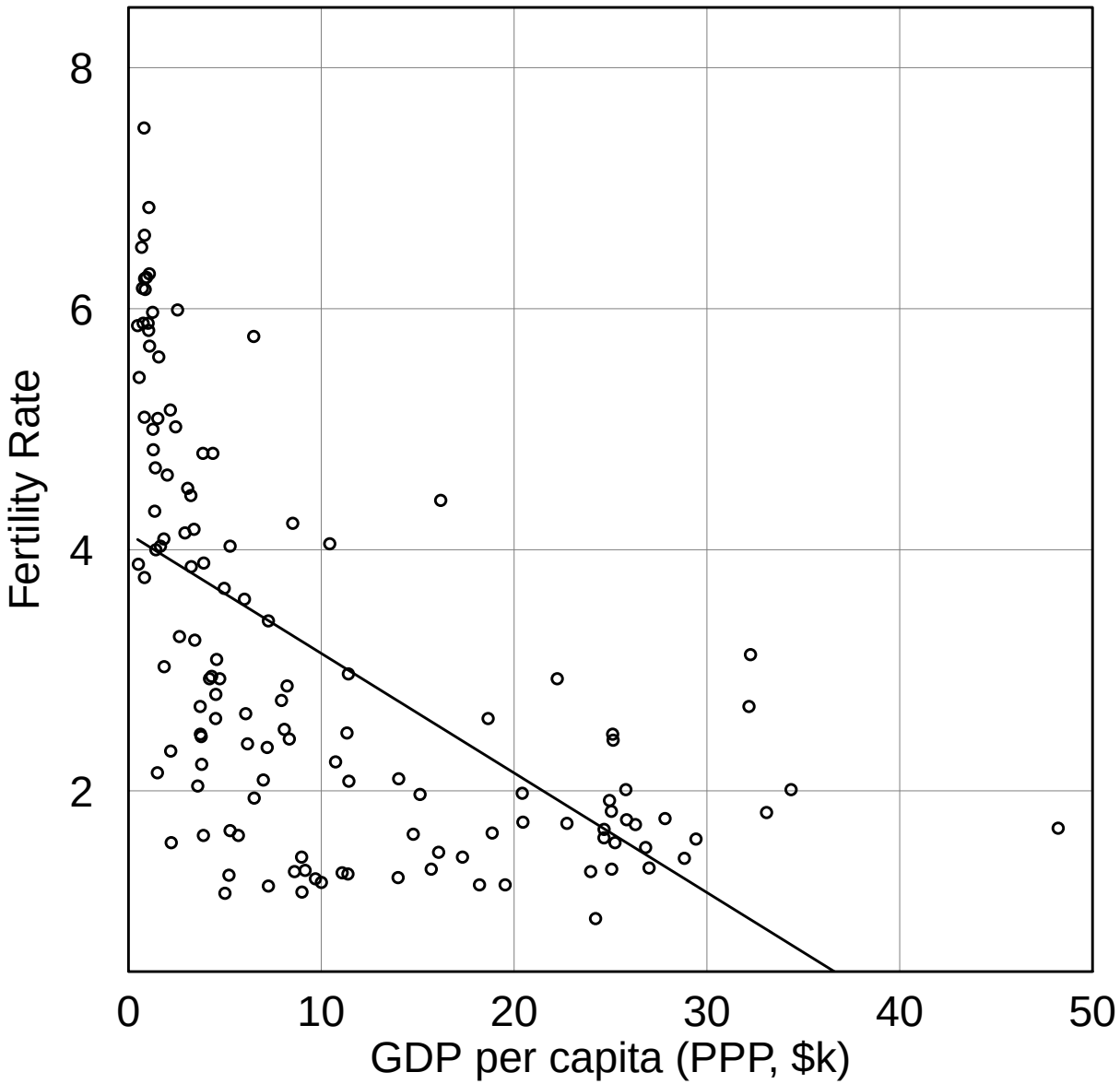
Variable	Estimates	se	t-stat	p-value
Intercept	11.25	(0.73)	15.46	< 0.001
Education Ratio	-0.08	(0.01)	-9.93	< 0.001
GDP per capita ($k)	-0.05	0.01	-5.32	< 0.001
N	130			
R^2	0.64			
RMSE	1.01			

Recall the fertility example.

In this example, we found both female education and GDP per capita reduced fertility



But we worried that GDP and Fertility might not have a linear relationship



But we worried that GDP and Fertility might not have a linear relationship

Perhaps the log of GDP affects fertility

Regression of Fertility on Education Ratio & log(GDP)

Variable	Estimate	se	*t*-stat	*p*-value
Education Ratio	−0.05	0.01	−6.26	< 0.001
log(GDP per capita)	−0.72	0.09	−8.02	< 0.001
Intercept	9.48	0.73	13.03	< 0.001
N	130			
R^2	0.71			
RMSE	0.91			

If we think percentage changes in GDP induce level changes in fertility, we should log GDP before including it in our model

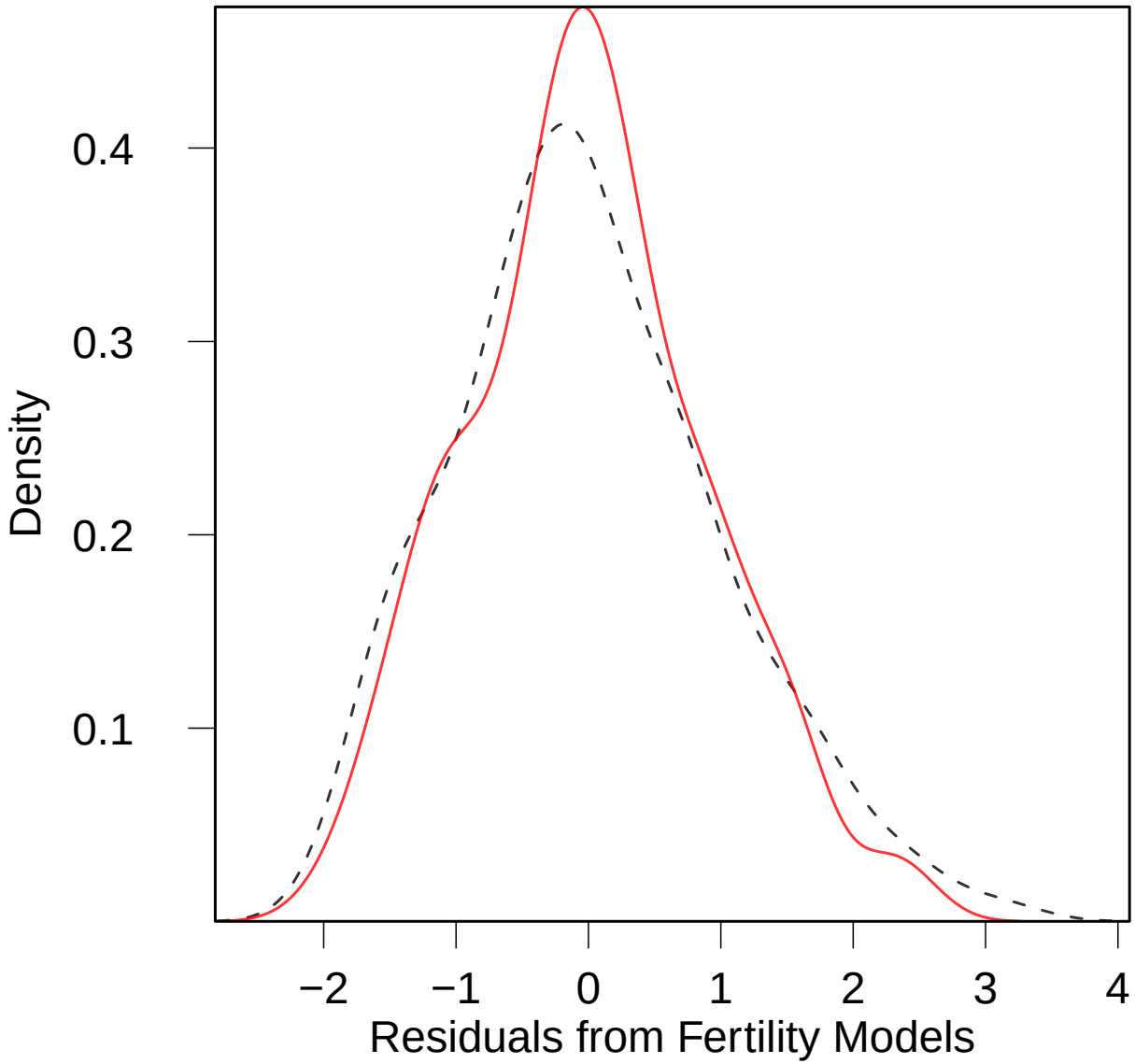
This model now assumes diminishing returns to GDP increases

The $\hat{\beta}$ for GDP is now harder to interpret, but the $\hat{\beta}$ for Education Ratio has the same interpretation as before

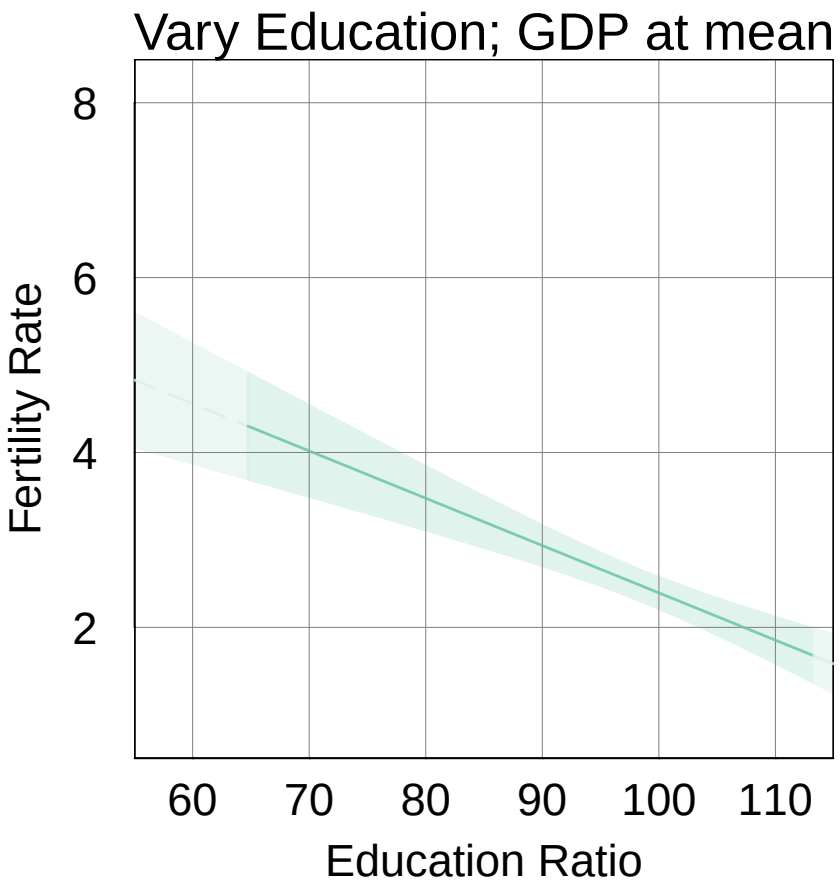
Four regression models of fertility

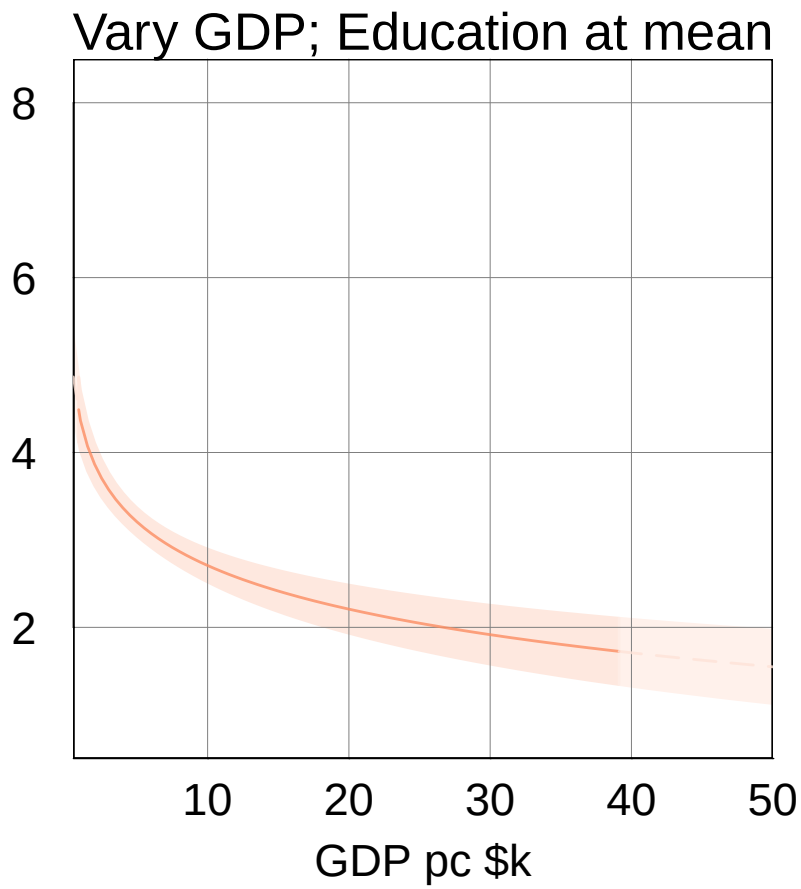
	Model			
Variable	1	2	3	4
Intercept	12.59	4.13	11.25	9.48
	(0.75)	(0.17)	(0.73)	(0.73)
Education Ratio	−0.10		−0.08	−0.05
	(0.01)		(0.01)	(0.01)
GDP per capita		−0.10	−0.05	
		(0.01)	(0.01)	
log(GDP per capita)				−0.72
				(0.09)
N	130	130	130	130
R^2	0.55	0.35	0.64	0.71
RMSE	1.12	1.35	1.01	0.91

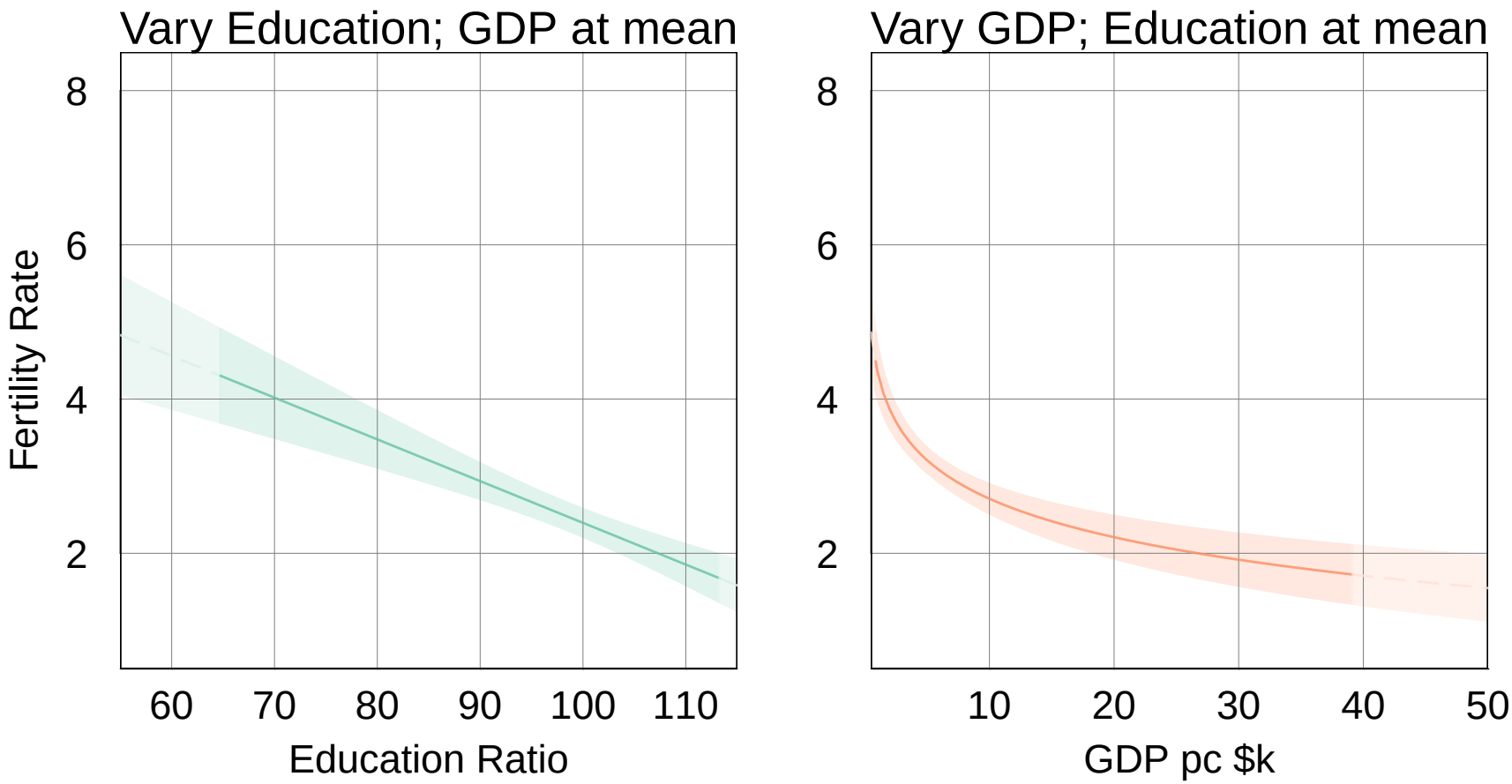
Standard errors in parentheses



Logging GDP has made the residuals a bit smaller, and a bit more symmetrical

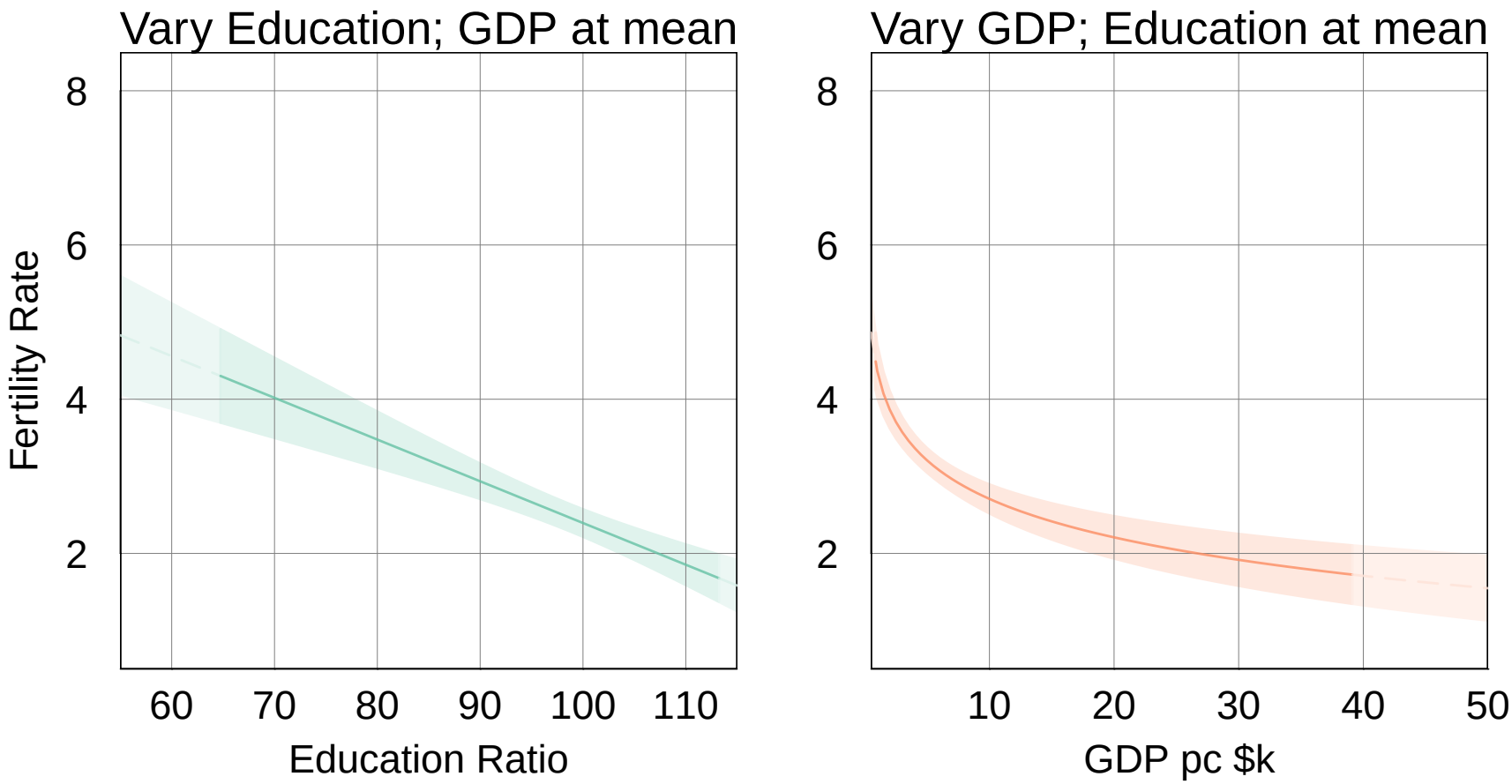

Vary Education; GDP at mean
Fertility Rate
8
6
4
2
60
70
80
90
100
110
Education Ratio


Vary GDP; Education at mean
8
6
4
2
10
20
30
40
50
GDP pc $k

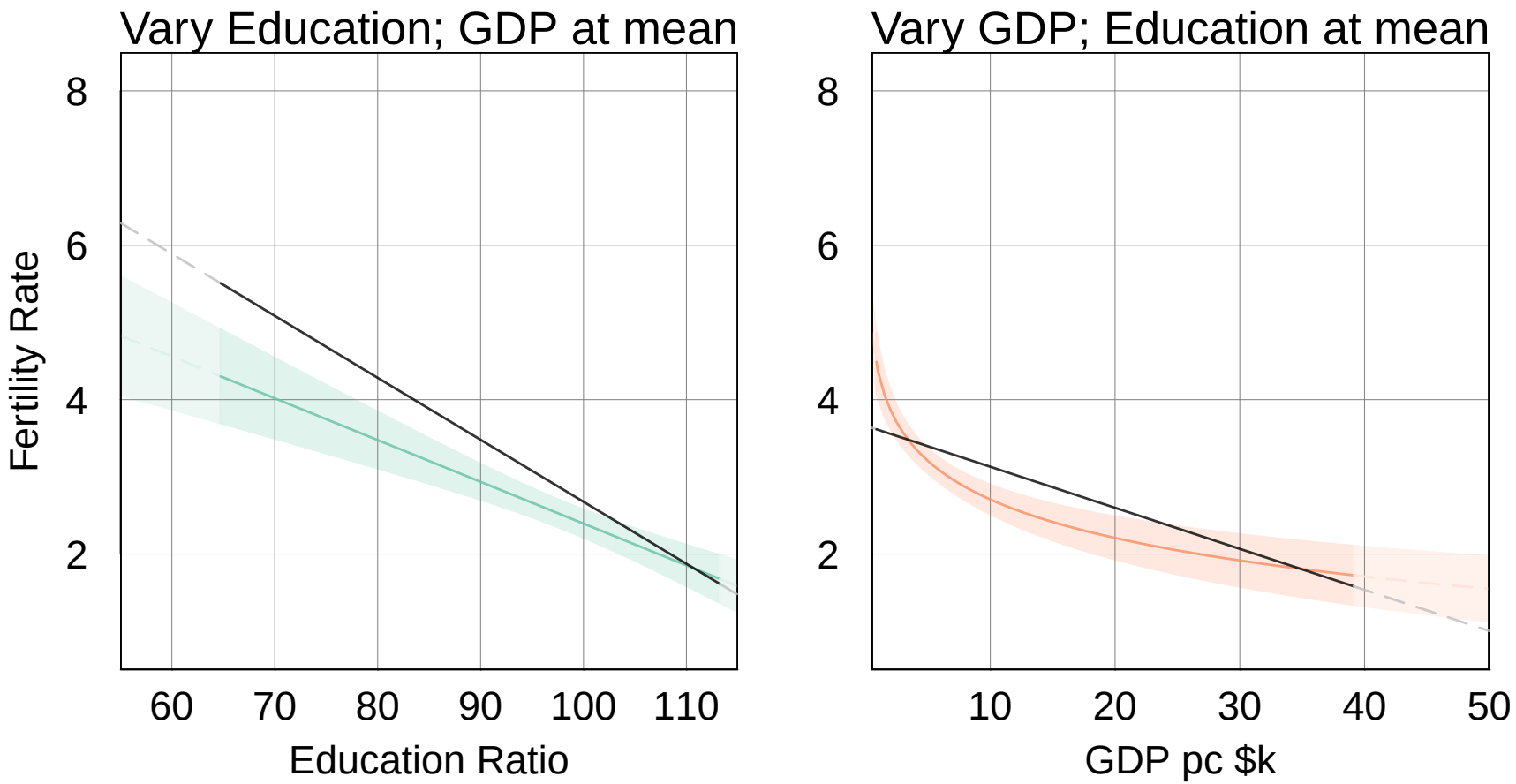


Logging GDP now allows small increases in GDP per capita is poor countries to dramatically lower fetility

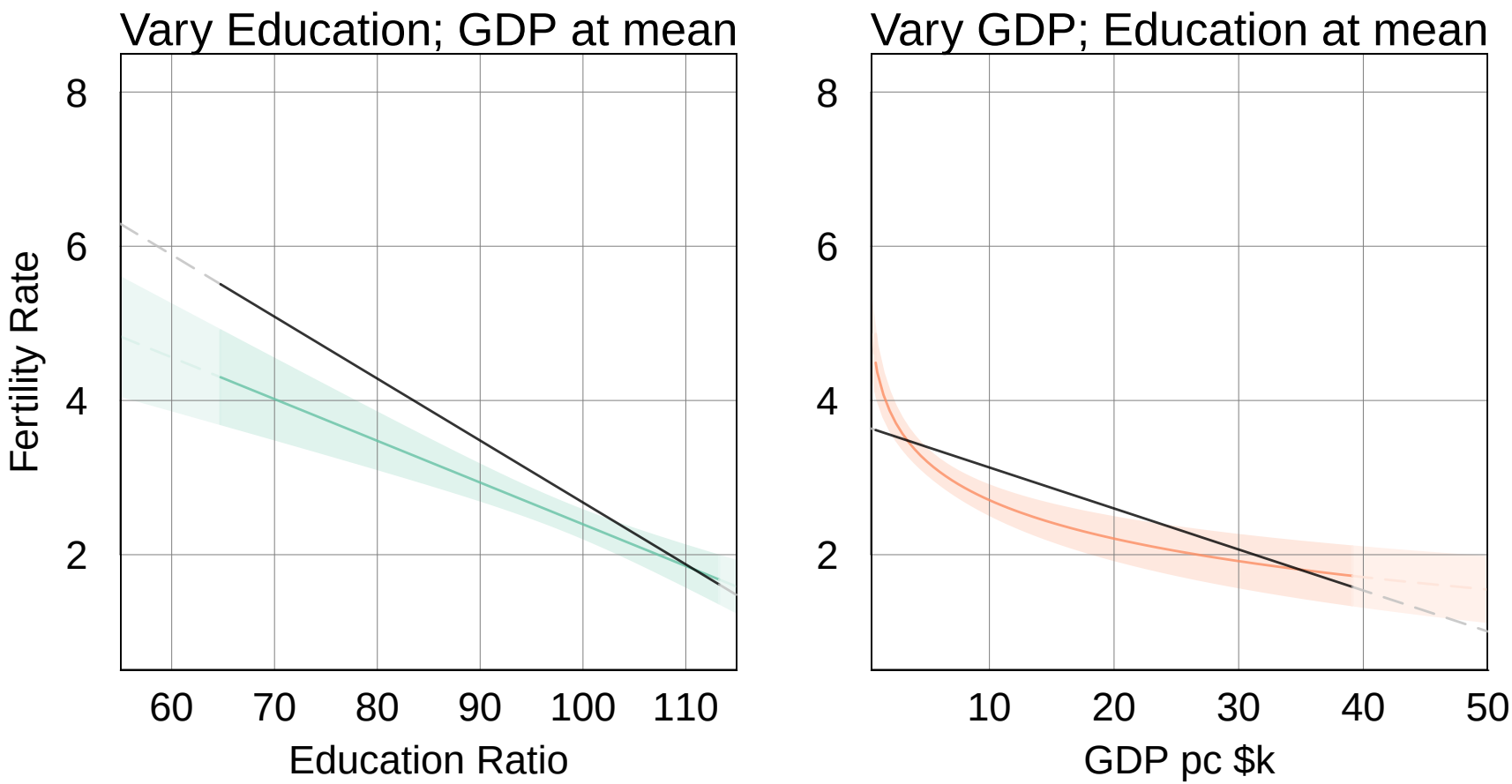
But small changes in GDP have very little effect in rich countries



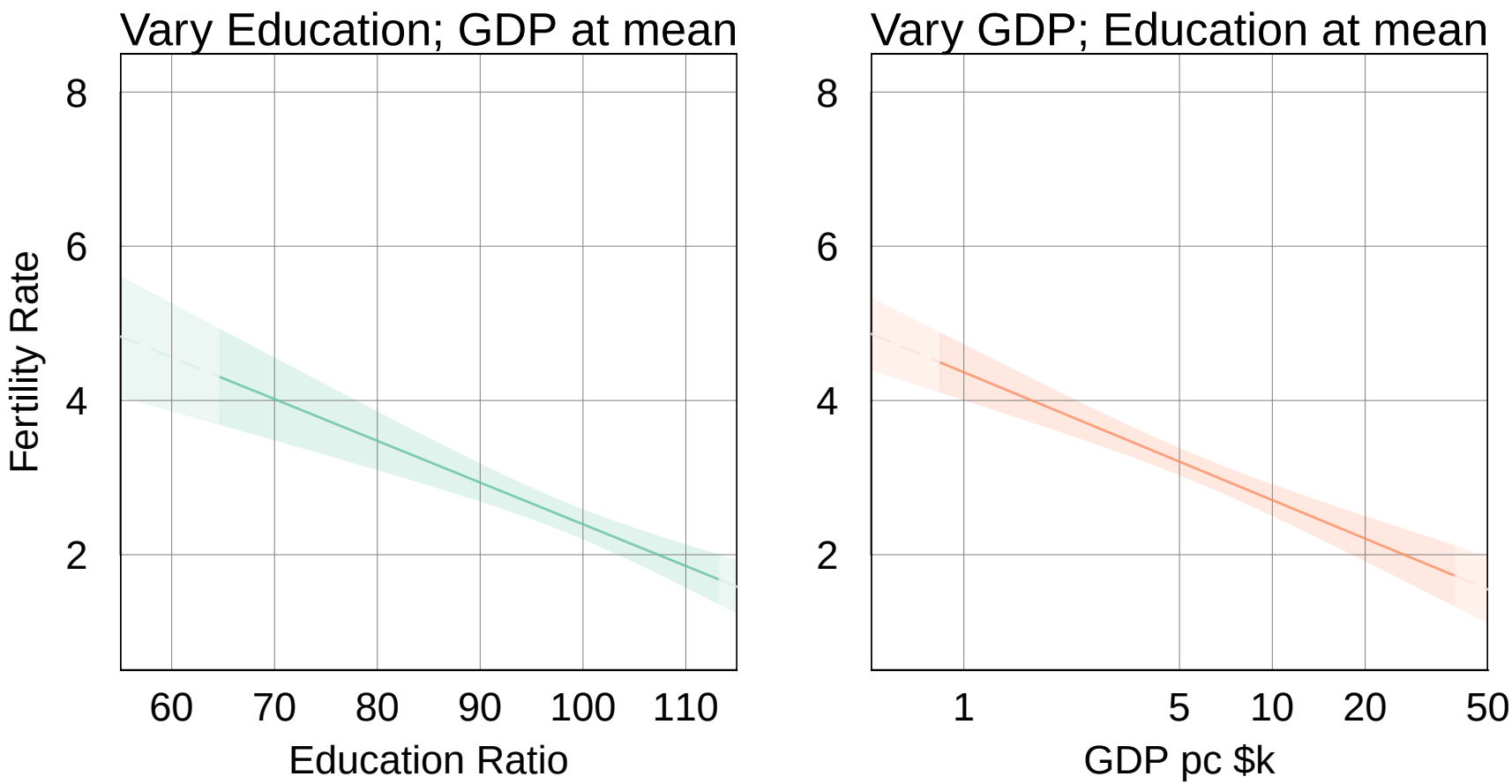
This is a pattern of diminishing marginal effects of economic development on fertility



The black lines show the fits from the model with a linear GDP control (last week's model)



Notice that the effect of education has shrunk a bit:
before, with an incorrect specification of GDP per capita (which needed to be logged)
we obtained a potentially biased estimate of the effect of female education



Finally, remember that the log of GDP per capita still has a linear effect in this model

If we squeeze the horizontal axis in a certain way, a linear relationship will reappear

Out of sample tests

The RMSE, or standard error of the regression, measures how much the model missed the sample data on average

But the model has an unfair advantage: it was *estimated* using the sample: of course it should fit!

The real question is usually whether the model would fit *all* samples drawn from the population

If we have a second sample of data, we can leave it out of our estimation, and then use the model to predict it

The standard error from this prediction is a measure of *Out of Sample Prediction Error*

You class project model will be judged on this metric, using the 100 surveyed individual's reserved by your TA

Cross-validation

Testing our model's predictions on out of sample data is a tough and valuable test

But *expensive*: we have to collect more data, and can't use it to improve our model

Cross-validation is a cheaper way to the same end

Step 1 Leave out one observation from our sample. Call the 1 left out case the *test set* and the $n - 1$ retained cases the *training set*

Cross-validation

Testing our model's predictions on out of sample data is a tough and valuable test

But *expensive*: we have to collect more data, and can't use it to improve our model

Cross-validation is a cheaper way to the same end

Step 1 Leave out one observation from our sample. Call the 1 left out case the *test set* and the $n - 1$ retained cases the *training set*

Step 2 Estimate your model using the training set

Cross-validation

Testing our model's predictions on out of sample data is a tough and valuable test

But *expensive*: we have to collect more data, and can't use it to improve our model

Cross-validation is a cheaper way to the same end

Step 1 Leave out one observation from our sample. Call the 1 left out case the *test set* and the $n - 1$ retained cases the *training set*

Step 2 Estimate your model using the training set

Step 3 Use the model estimated in Step 2 to predict the test set; record the error

Cross-validation

Testing our model's predictions on out of sample data is a tough and valuable test

But *expensive*: we have to collect more data, and can't use it to improve our model

Cross-validation is a cheaper way to the same end

Step 1 Leave out one observation from our sample. Call the 1 left out case the *test set* and the $n - 1$ retained cases the *training set*

Step 2 Estimate your model using the training set

Step 3 Use the model estimated in Step 2 to predict the test set; record the error

Step 4 Repeat Steps 1 through 3 n times, leaving out each observation in turn.

The square root of the average of the squared error across these iterations is the *Cross-Validation standard error*

Goodness of fit, fertility models

Model	RMSE	CV Error
Education	1.12	1.25
GDP	1.35	1.84
Education, GDP	1.01	1.04
Education, log GDP	0.91	0.85

The above shows the in sample and cross-validation standard errors for each model

Cross-validation performance is usually worse than in sample

Leave-one-out cross-validation is the best estimate of out of sample performance, and thus one of the best goodness of fit measures

Example: The decision to vote

Suppose we want to understand the determinants of voter turnout

We use data from the 2000 National Election Study, which surveys thousands of respondents:

Vote Whether the respondent voted in 2000

Age The age of the respondent in 2000

Education The education level of the respondent, measured as less than high school, high school, or college

Vote06 Whether the respondent voted in 1996

Regression with binary variables

Linear regression assumes the error term is Normally distributed

We can often transform our y or x to make this happen

But if the *response* variable is binary or categorical, transformation isn't enough

Instead we need to model directly the probability that the observation falls in each possible category

We do this using advanced methods called *logit* (for binary dependent variables) or *multinomial logit* (for categorical variables)

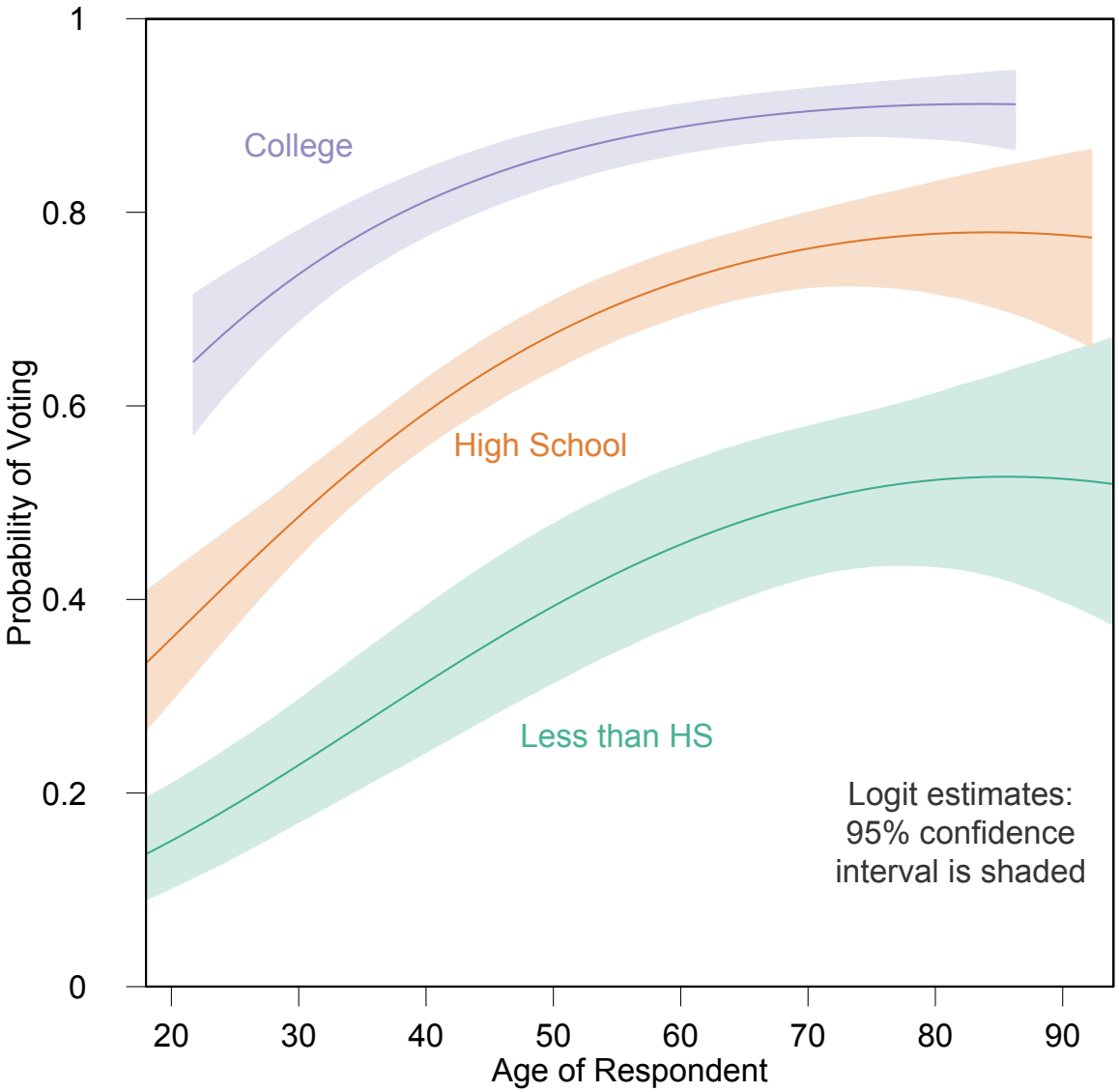
Logit

Multinomial logit is the model your TA will use to estimate your models of music preference

Multinomial logit is a complex model well beyond POLS 205, but we can get an idea of what it is doing by using binary logit to estimate the probability that people vote

The best way to understand logit models is with a graph showing the relationship between the levels of our covariates x and the probability of our outcome, y

Here, we will look at the changing probability a person turns out to vote given their age and education



What does this graphic tell us?

Goodness of fit with binary response

One way to judge fit in a binary or multinomial response model is to calculate the *Percent Correctly Predicted*

For any y_i, we say the model has correctly predicted if the event the model considered most likely is also the event that actually happened

For binary logit, correct predictions occur if and only if

1. $y_i = 1$ and the predicted $\Pr(y_i) > 0.5$
2. $y_i = 0$ and the predicted $\Pr(y_i) < 0.5$

Goodness of fit with binary response

The best we can do is 100% correctly predicted

What's the worst we can do?

The most naïve model of a binary variable is one that always predicts the most likely outcome will occur

Such a model will be right $100 \times p\%$ of the time, where p is the probability of the most likely outcome

65.4% of people in our sample voted, so we will judge our models by their ability to predict better than a model which just says "everyone votes"

	Percent correctly predicted	
	Full data	Cross-validated
"Everyone votes"	65.4	65.4

	Percent correctly predicted	
	Full data	Cross-validated
"Everyone votes"	65.4	65.4
Age	65.7	65.5

	Percent correctly predicted	
	Full data	Cross-validated
"Everyone votes"	65.4	65.4
Age	65.7	65.5
Education	67.4	67.4

	Percent correctly predicted	
	Full data	Cross-validated
"Everyone votes"	65.4	65.4
Age	65.7	65.5
Education	67.4	67.4
Age, Education	69.8	69.7

Cross-validation is very similar to in sample prediction here

Our model is better than a naïve model, but only a little

Prediction vs. Explanation

One reason our models make only slight improvements is that our goal is testing a causal model

So we've limited ourselves to variables which might have *caused* voting

But what if we could include any covariate, whether it caused the vote or not?

What would you include?

We could surely predict individual voters' behavior much better with a variable, even if we don't understand *why* they voted

	Percent correctly predicted	
	Full data	Cross-validated
"Everyone votes"	65.4	65.4
Age	65.7	65.5
Education	67.4	67.4
Age, Education	69.8	69.7
Age, Education, Vote06	77.5	77.4

Adding Vote06 makes a better predictive model

It doesn't help us explain anything, though: to understand *why* people vote, we should leave Vote96 out

Dealing with Outliers

Let's return to linear regression and our Wealth example

One problem we had was the inclusion of a handful of very rich households

Logging wealth helped a bit with the extreme cases, but still some remained outliers, even in a logged model

What if the very rich are just different: what if a different causal model explains their wealth?

Dealing with Outliers

Let's return to linear regression and our Wealth example

One problem we had was the inclusion of a handful of very rich households

Logging wealth helped a bit with the extreme cases, but still some remained outliers, even in a logged model

What if the very rich are just different: what if a different causal model explains their wealth?

That model might include inheritance (not a big factor for most people), luck in being in the right position when a corporation has a windfall profit, or entrepreneurial skills

Dealing with Outliers

Including observations from a "different model" or different world can bias our results for explaining the wealth of ordinary people

One solution is to remove extreme outliers from the model altogether

Then we'd have a cleaner result focusing on the relationships between wealth, education, age, and race in the heart of the dataset: the poor and middle class

Very dangerous to choose which observations to exclude, though:
we'd likely just create selection bias

We need a model which figures out which cases are really outliers, and excludes them for us

This type of model is known as *robust regression*

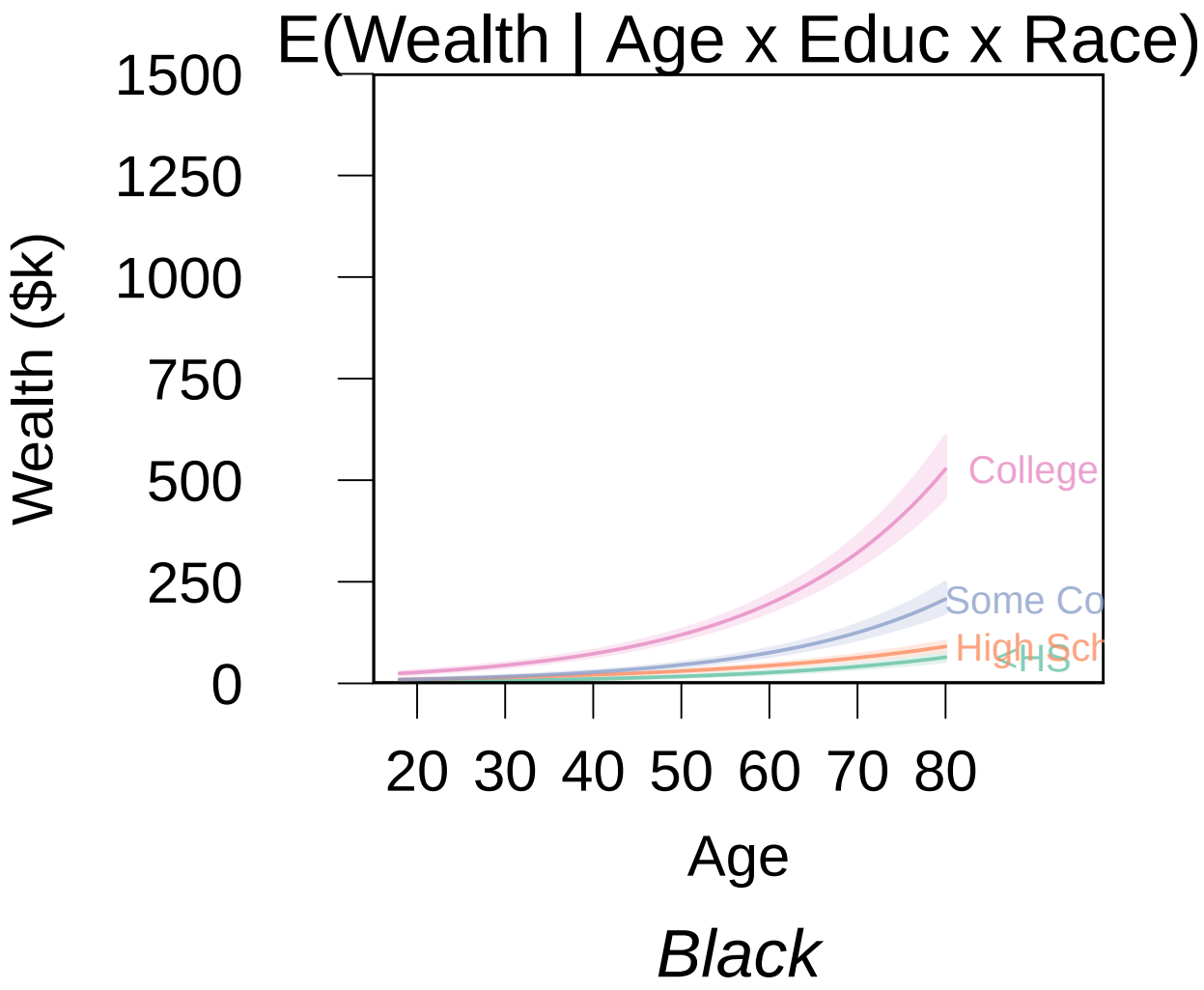
Robust regression

Robust regression techniques are a powerful tool for mitigating the effect of outliers on your results

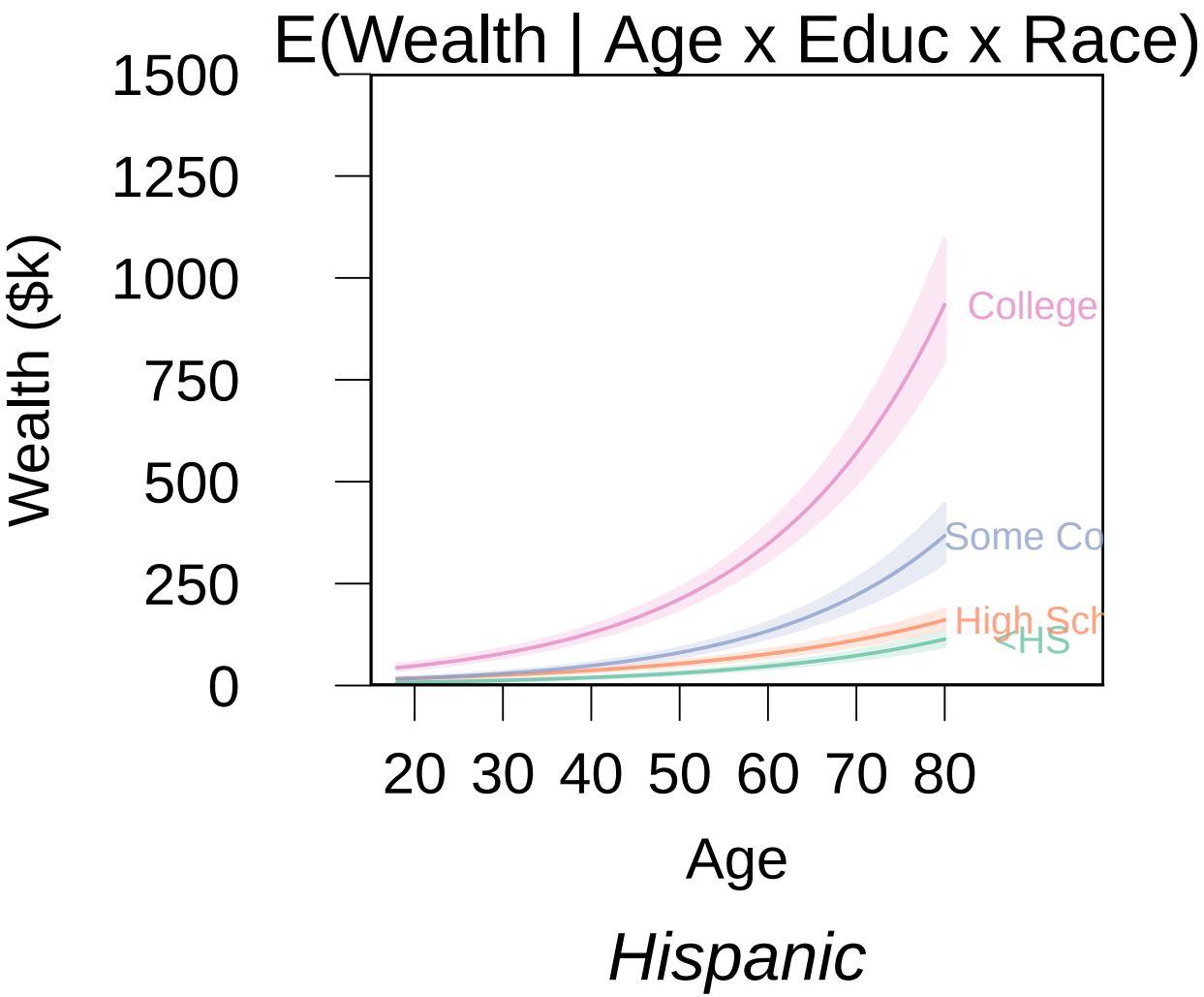
Can interpret robust regression coefficients just like linear regression coefficients

Here, we will just look at some graphics

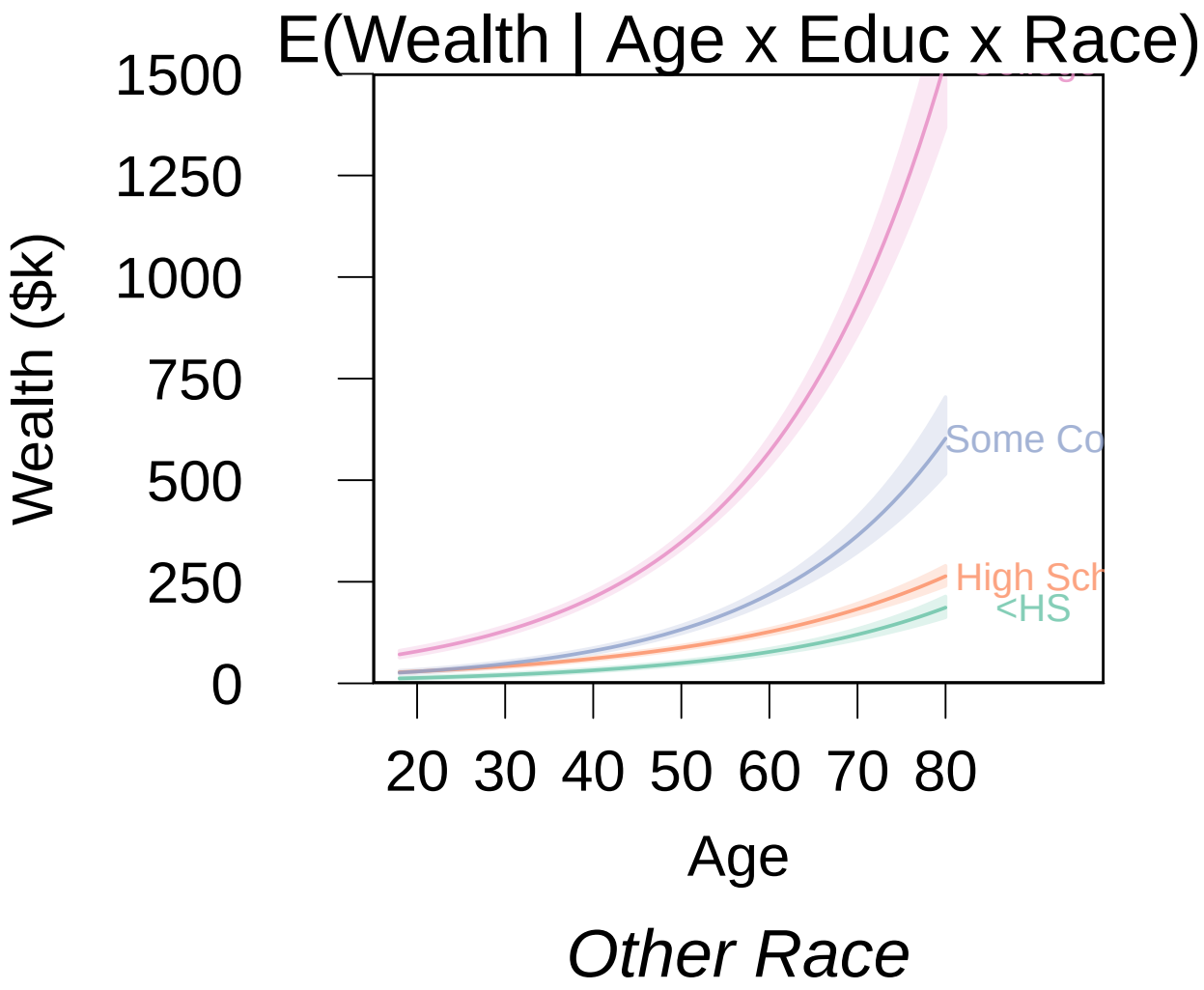
As we look at the model predictions, remember that our new model excludes the extremely poor and hyper-rich, to focus our findings on people with lower or middle class wealth



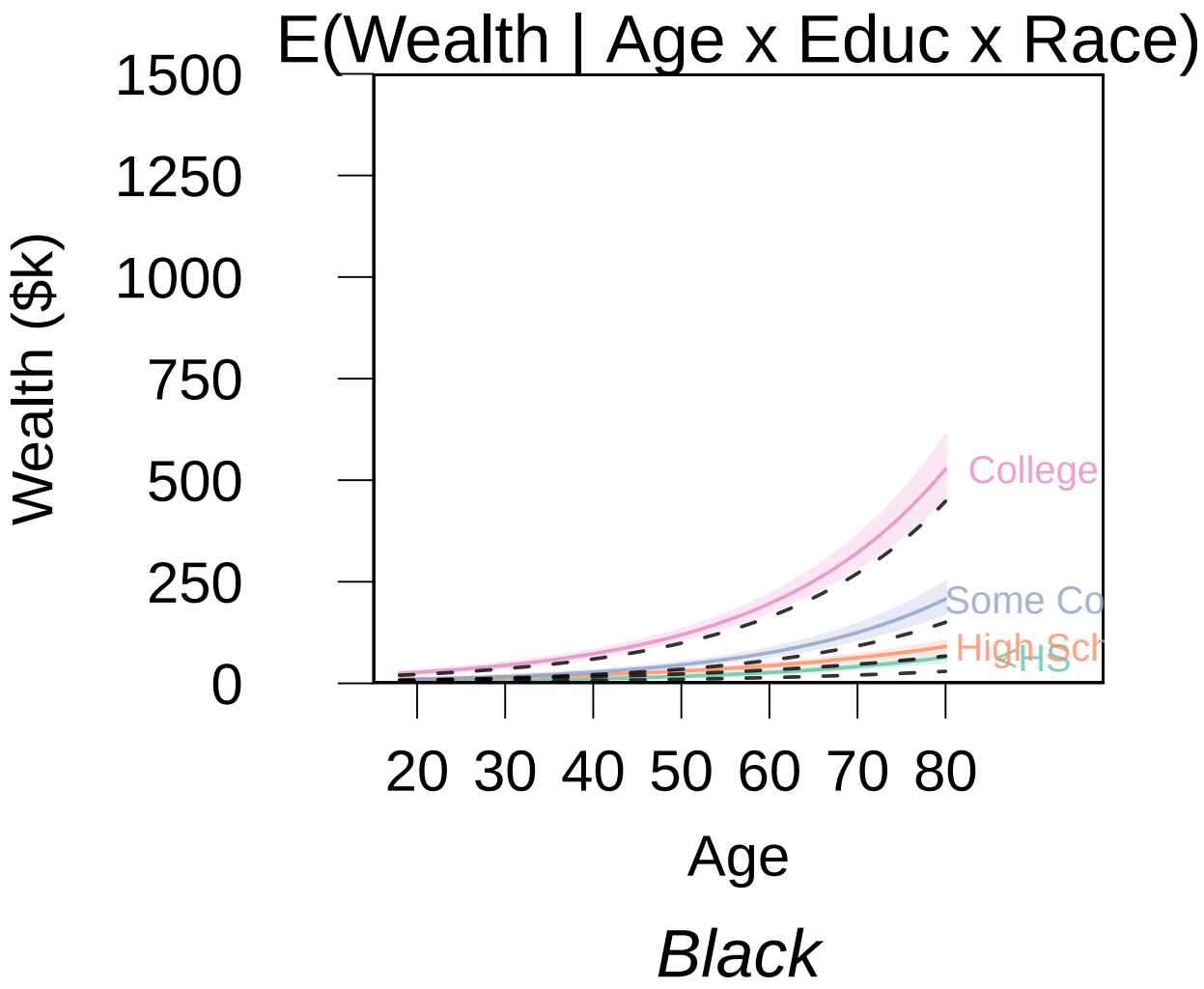
The *robust* expected levels of wealth for black households



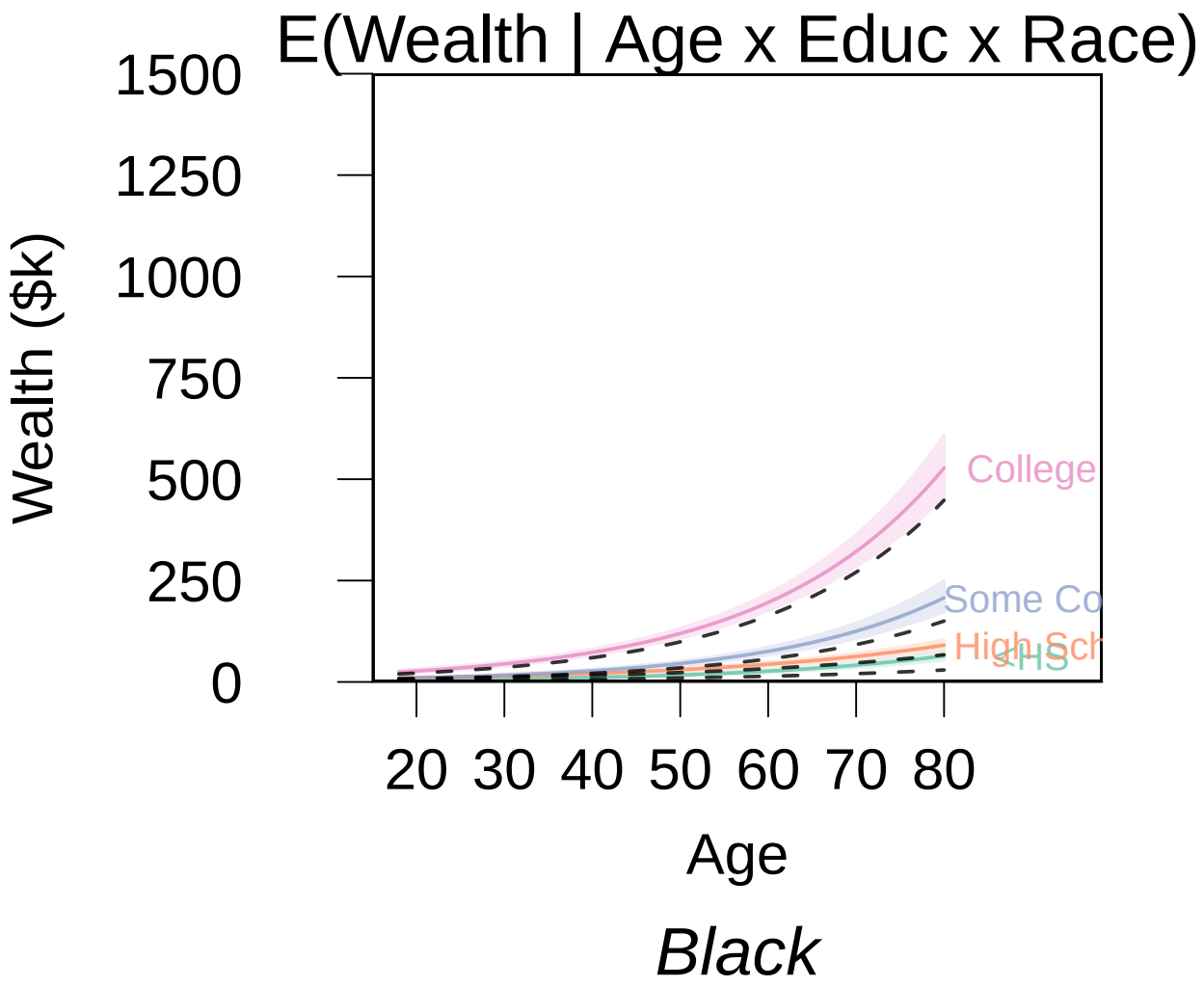
The *robust* expected levels of wealth for Hispanic households



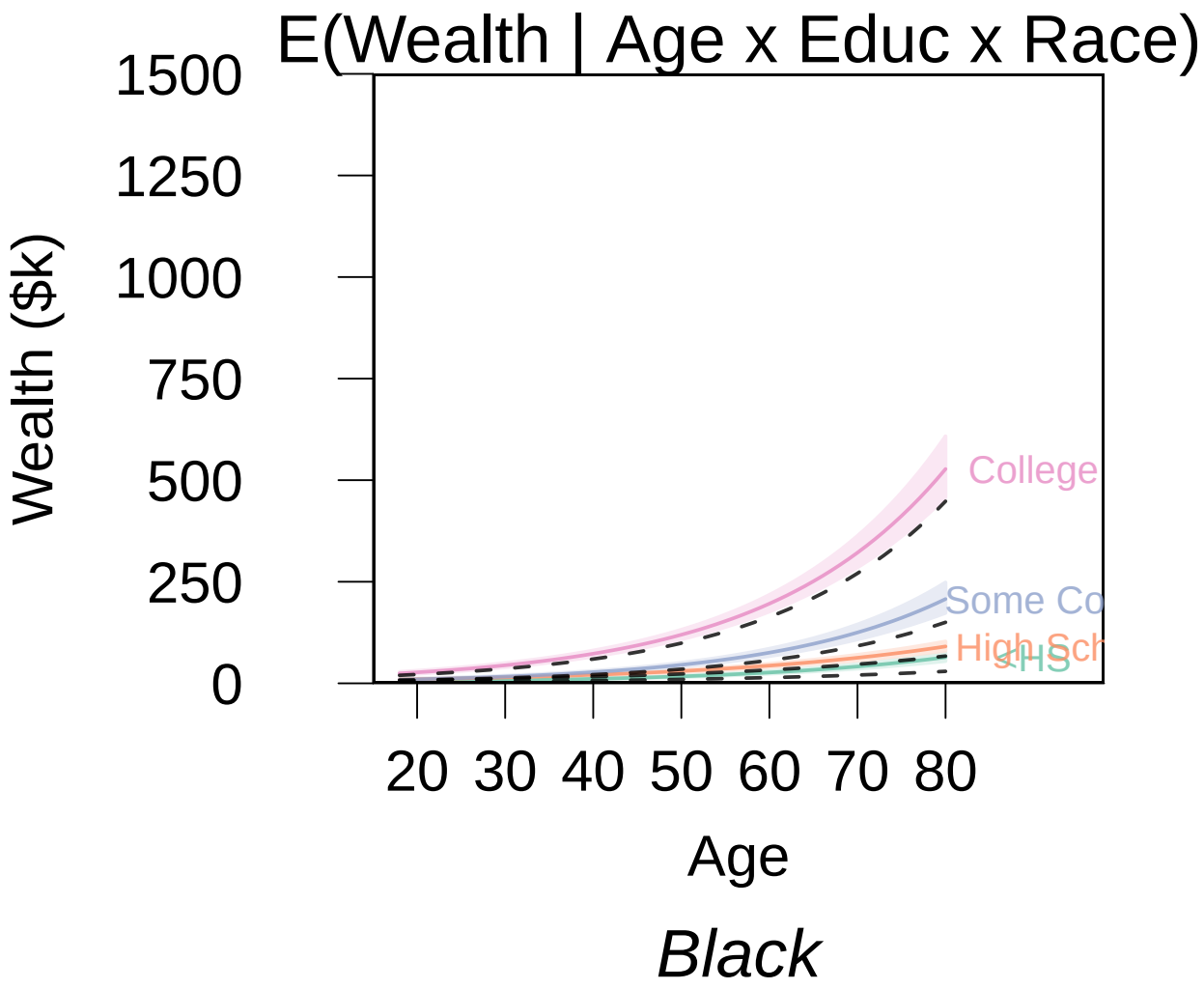
The *robust* expected levels of wealth for other households



Comparing to the non-robust fit (in black)

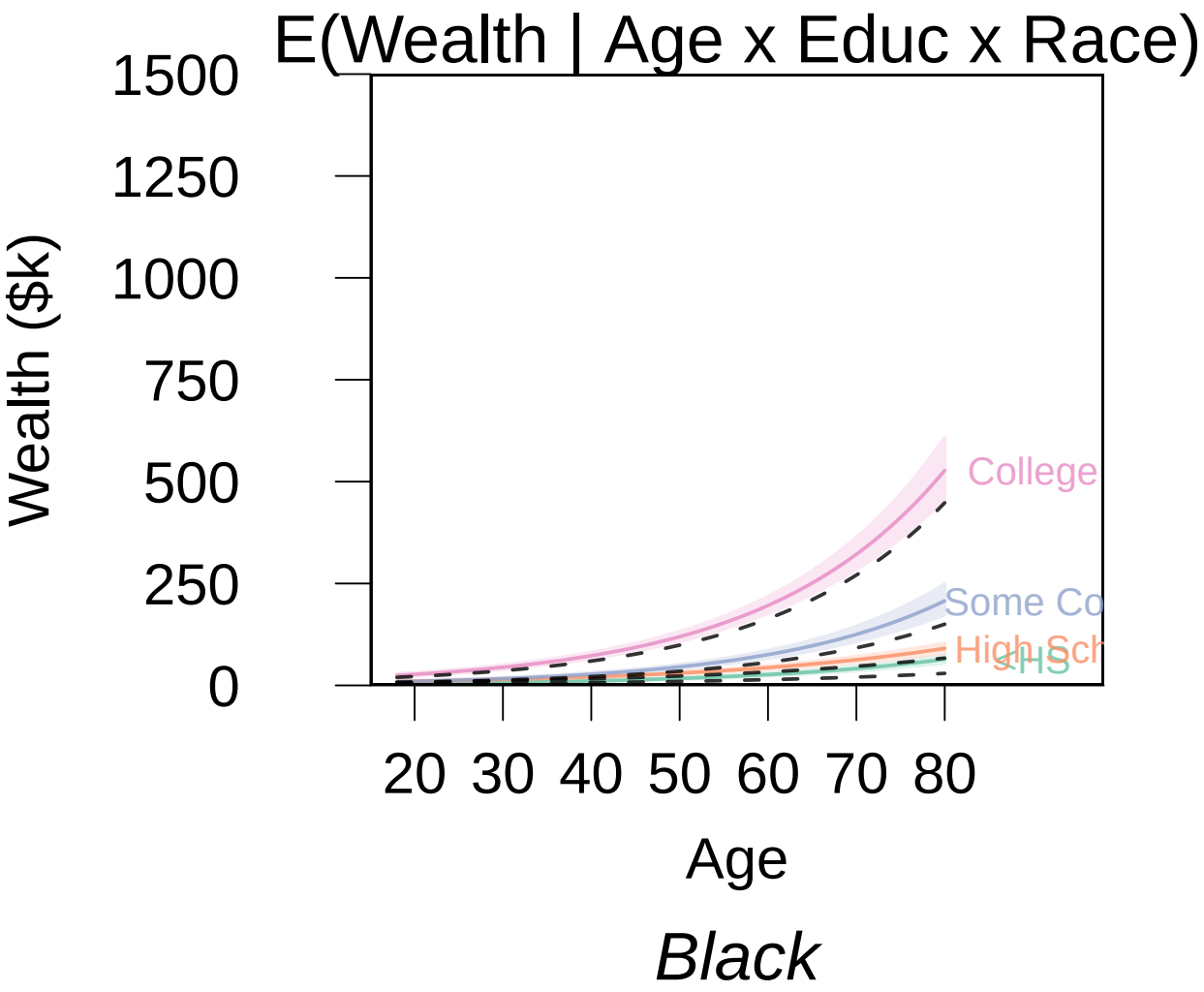


Education gap shrinks—extremely wealthy are nearly all well educated

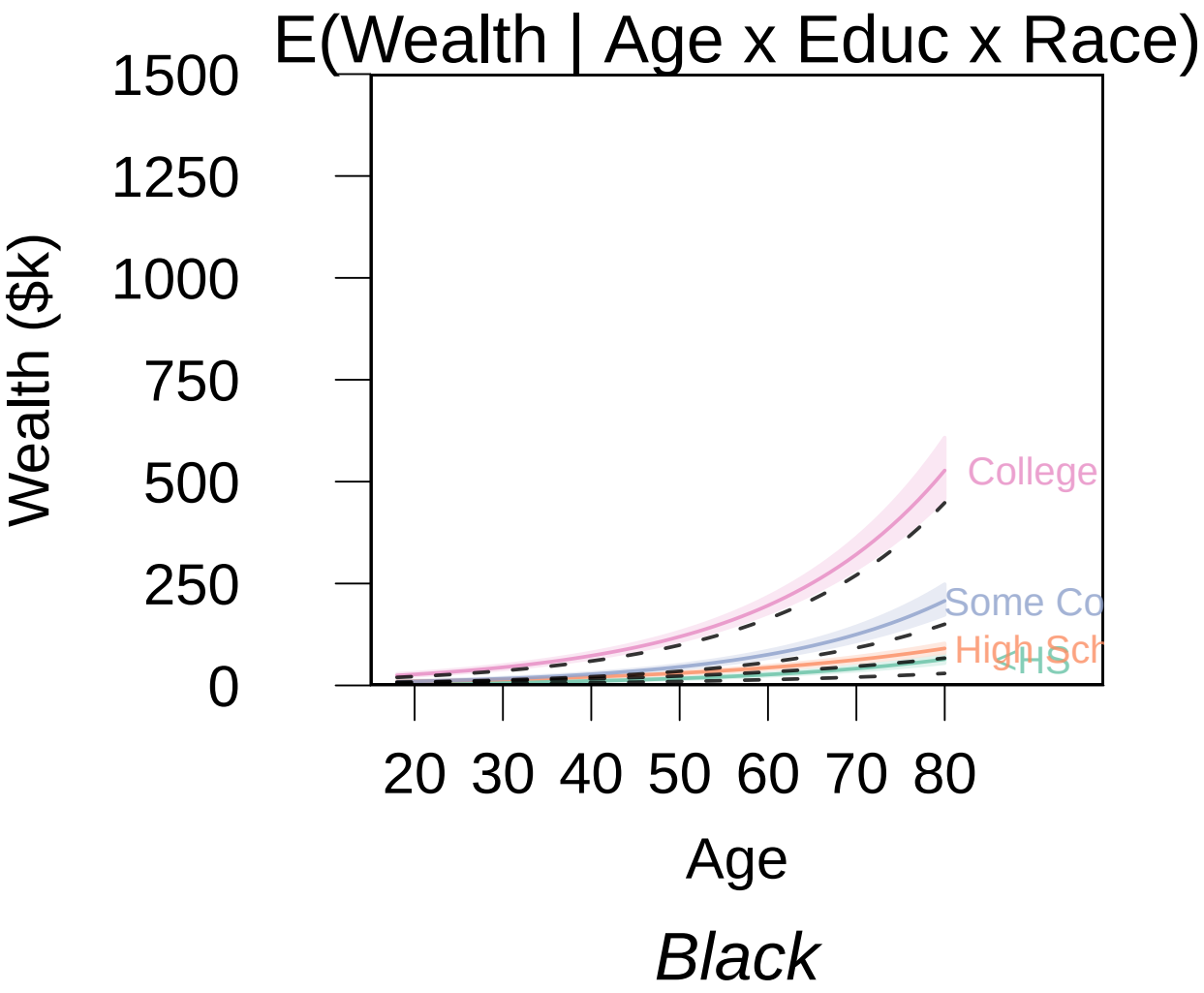


Education gap shrinks—extremely wealthy are nearly all well educated

But this is not a *likely* outcome from education for anyone

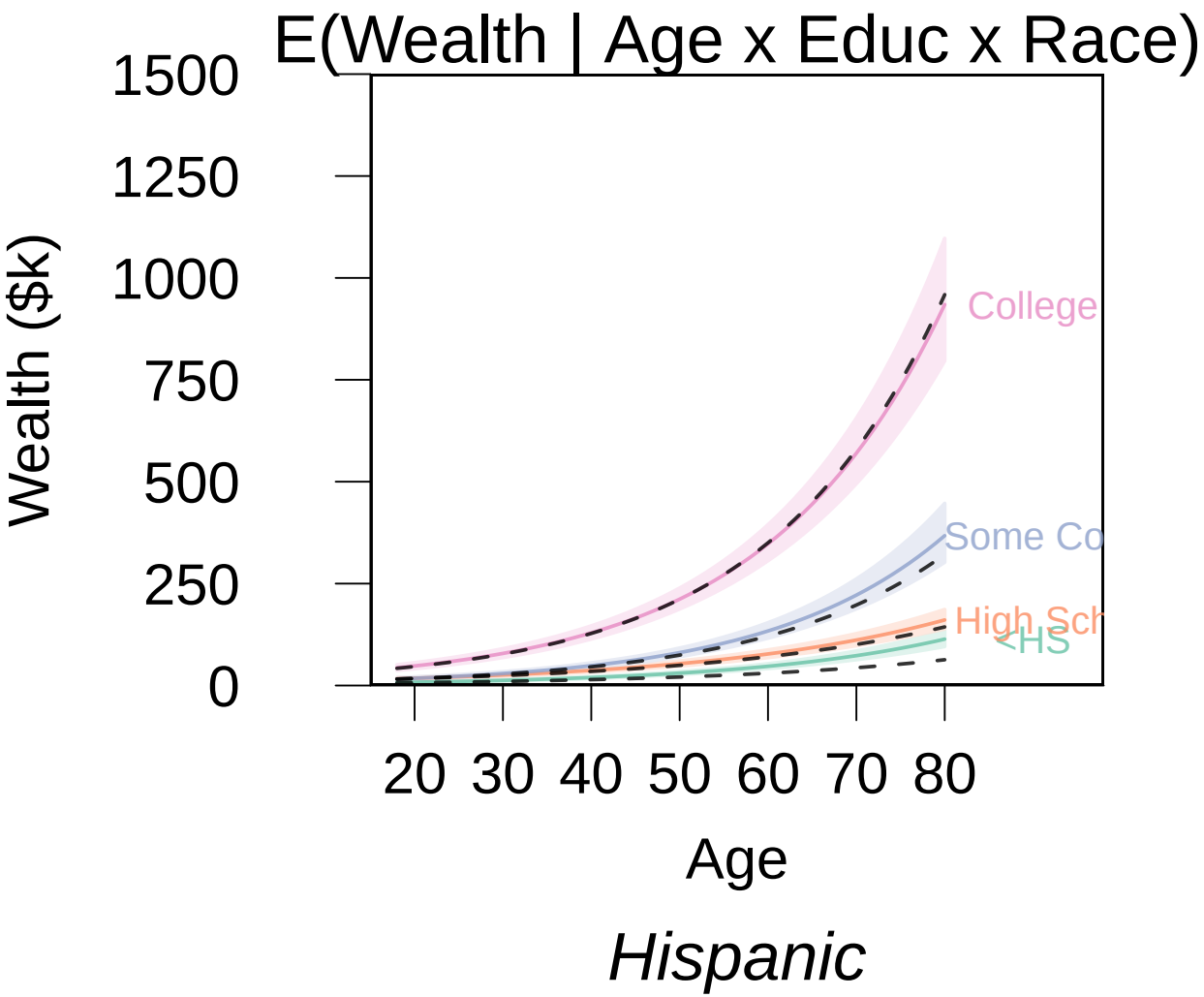


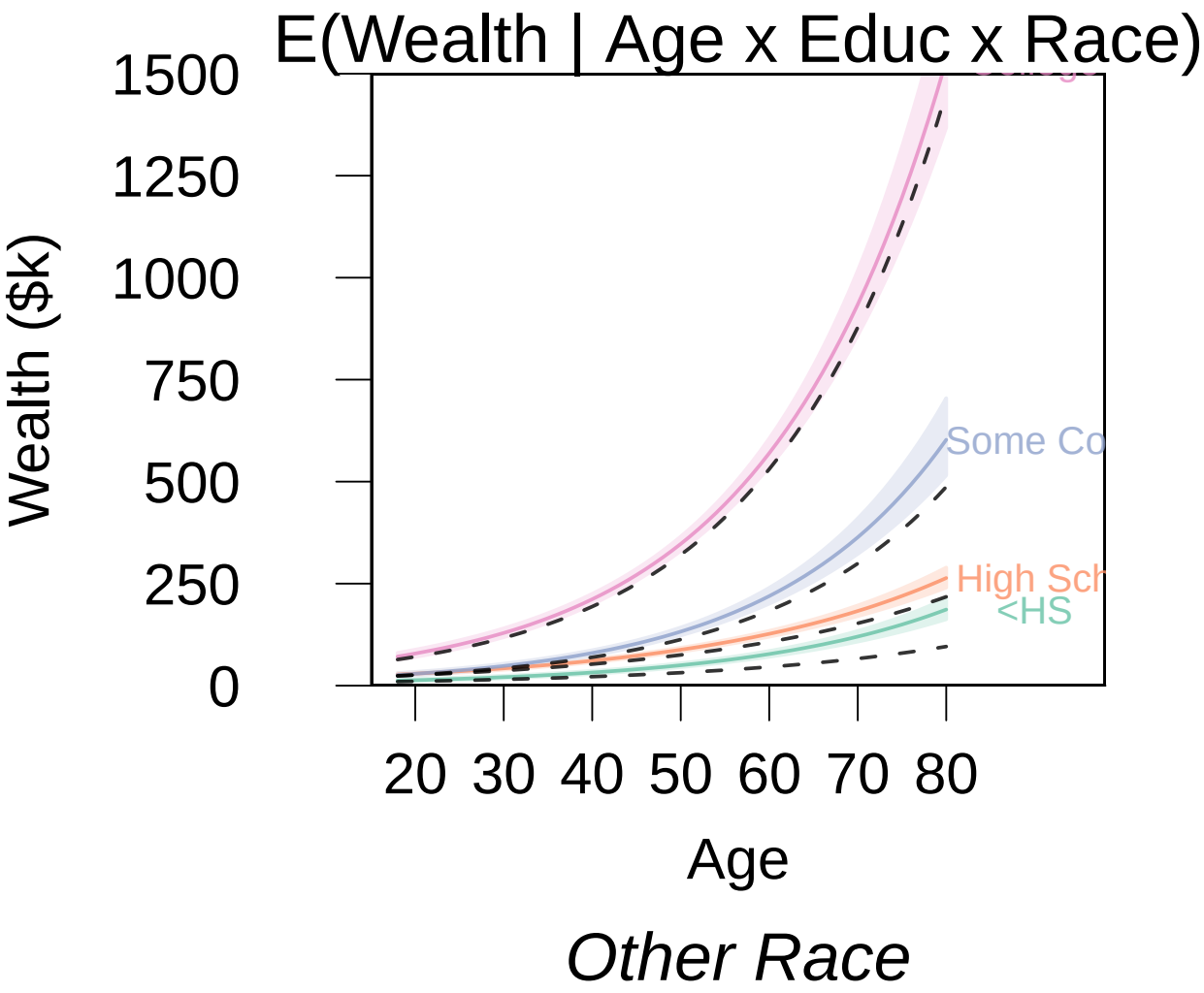
Race gap shrinks—extremely wealthy are mostly white



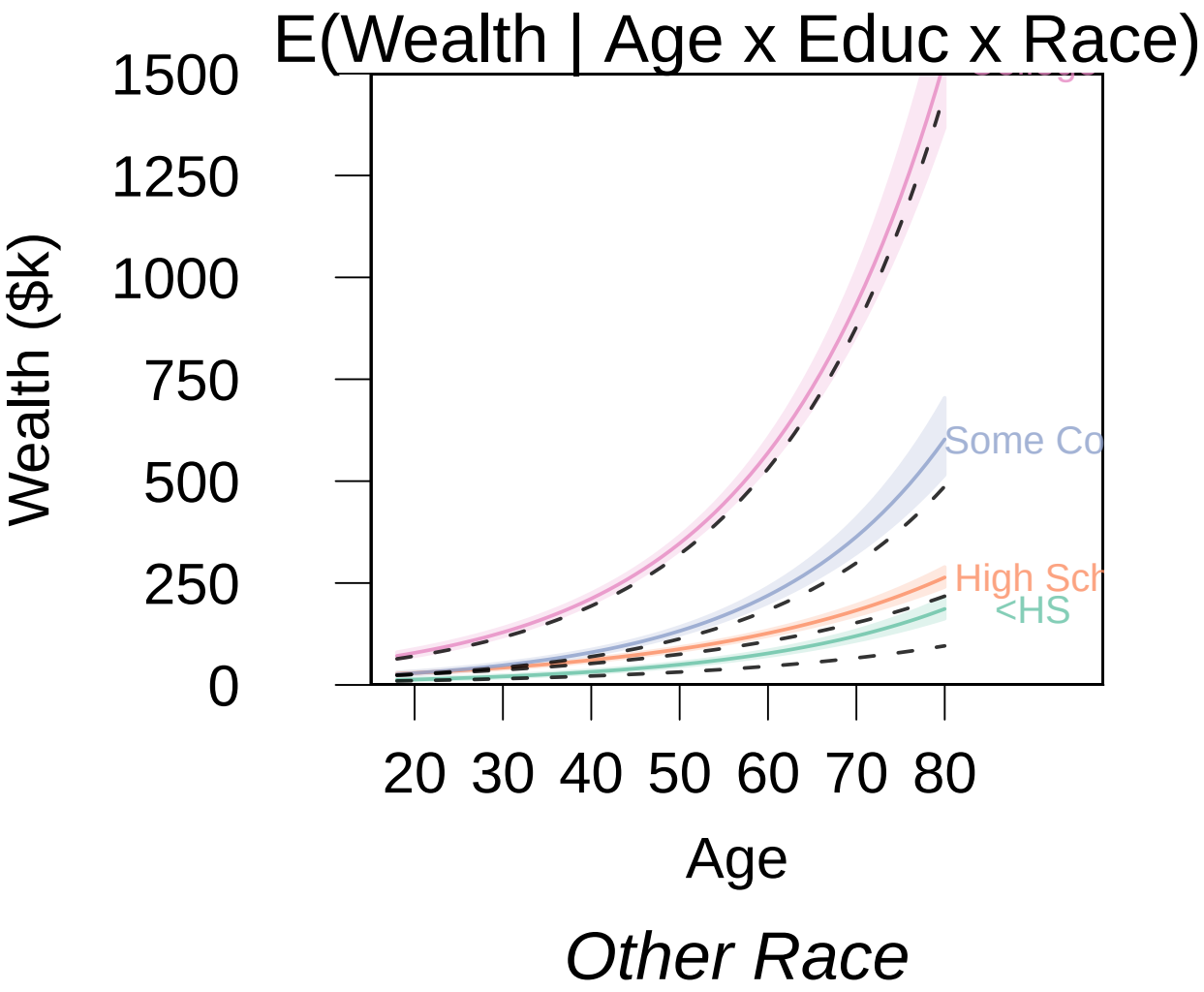
Race gap shrinks—extremely wealthy are mostly white

But this is not a *likely* outcome from education for any specific white household


E(Wealth | Age x Educ x Race)
Wealth ($k)
1500
1250
1000
750
500
250
0
College
Some Co
High Sch
<HS
20 30 40 50 60 70 80
Age
Hispanic



Once again, a shrinking education gap



Once again, a shrinking education gap

College is valuable, but attributing extreme wealth to college is misleading: probably the other way around!

Correlation and Causation Revisited

For most people, going to college raises your expected lifetime wealth

But if your parents are hyper-rich, they will both send you to college (with very high probability) and give you the bulk of your wealth

In this case, causation is *reversed*

Reverse causation is probably very common

Yet linear regression completely fails when it is present!

Reverse causation and fertility

We theorized that as countries got wealthier, women would have fewer children

Lots of potential reasons:

1. Development leads to female education, which leads to contraception

Reverse causation and fertility

We theorized that as countries got wealthier, women would have fewer children

Lots of potential reasons:

1. Development leads to female education, which leads to contraception
2. Development leads to the welfare state, which replaces children as a source of security in old age

Reverse causation and fertility

We theorized that as countries got wealthier, women would have fewer children

Lots of potential reasons:

1. Development leads to female education, which leads to contraception
2. Development leads to the welfare state, which replaces children as a source of security in old age
3. Development lowers child mortality, so families don't need to have lots of kids to ensure a few survive

Reverse causation and fertility

We theorized that as countries got wealthier, women would have fewer children

Lots of potential reasons:

1. Development leads to female education, which leads to contraception
2. Development leads to the welfare state, which replaces children as a source of security in old age
3. Development lowers child mortality, so families don't need to have lots of kids to ensure a few survive
4. Development shrinks the agricultural sector, and lowers demand for child labor

Reverse causation and fertility

We theorized that as countries got wealthier, women would have fewer children

Lots of potential reasons:

1. Development leads to female education, which leads to contraception
2. Development leads to the welfare state, which replaces children as a source of security in old age
3. Development lowers child mortality, so families don't need to have lots of kids to ensure a few survive
4. Development shrinks the agricultural sector, and lowers demand for child labor
5. Development raises the premium for education, which means parents need to invest more in each child

Reverse causation and fertility

But what if high *fertility* also stops development from happening?

1. Many children leads to fewer years of education per child, and less education means fewer skilled workers, and less innovation

Reverse causation and fertility

But what if high *fertility* also stops development from happening?

1. Many children leads to fewer years of education per child, and less education means fewer skilled workers, and less innovation
2. Farmers with many children often divide their farms into smaller plots at death, to provide for each child. Smaller farms discourages industrial agriculture

Reverse causation and fertility

But what if high *fertility* also stops development from happening?

1. Many children leads to fewer years of education per child, and less education means fewer skilled workers, and less innovation
2. Farmers with many children often divide their farms into smaller plots at death, to provide for each child. Smaller farms discourages industrial agriculture
3. More children may discourage female education, as patriarchal societies focus limited educational resources on male offspring

Linear regression can't tell us which direction causality flows (if any!)

To see this, watch what happens when we flip the dependent and independent variables

Regression of log(GDP per capita $k) on Fertility

toprule Variable	Estimate	se	*t*-stat	*p*-value
Fertility	-0.56	0.0390	-14.33	< 0.001
Intercept	3.49	0.14	25.08	< 0.001
N	130			
R^2	0.61			
RMSE	0.746			

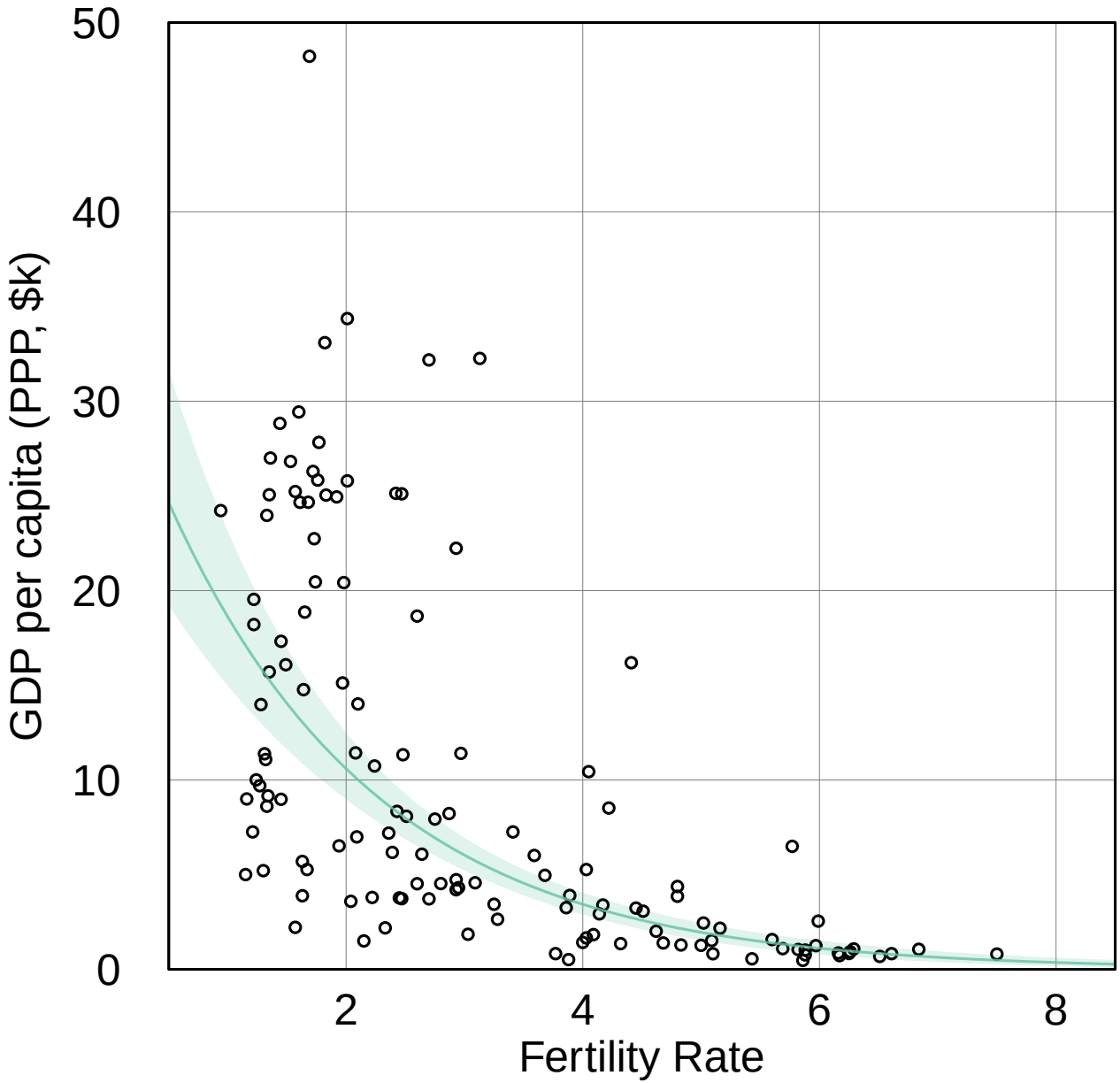
What does this table say?

Regression of log(GDP per capita $k) on Fertility

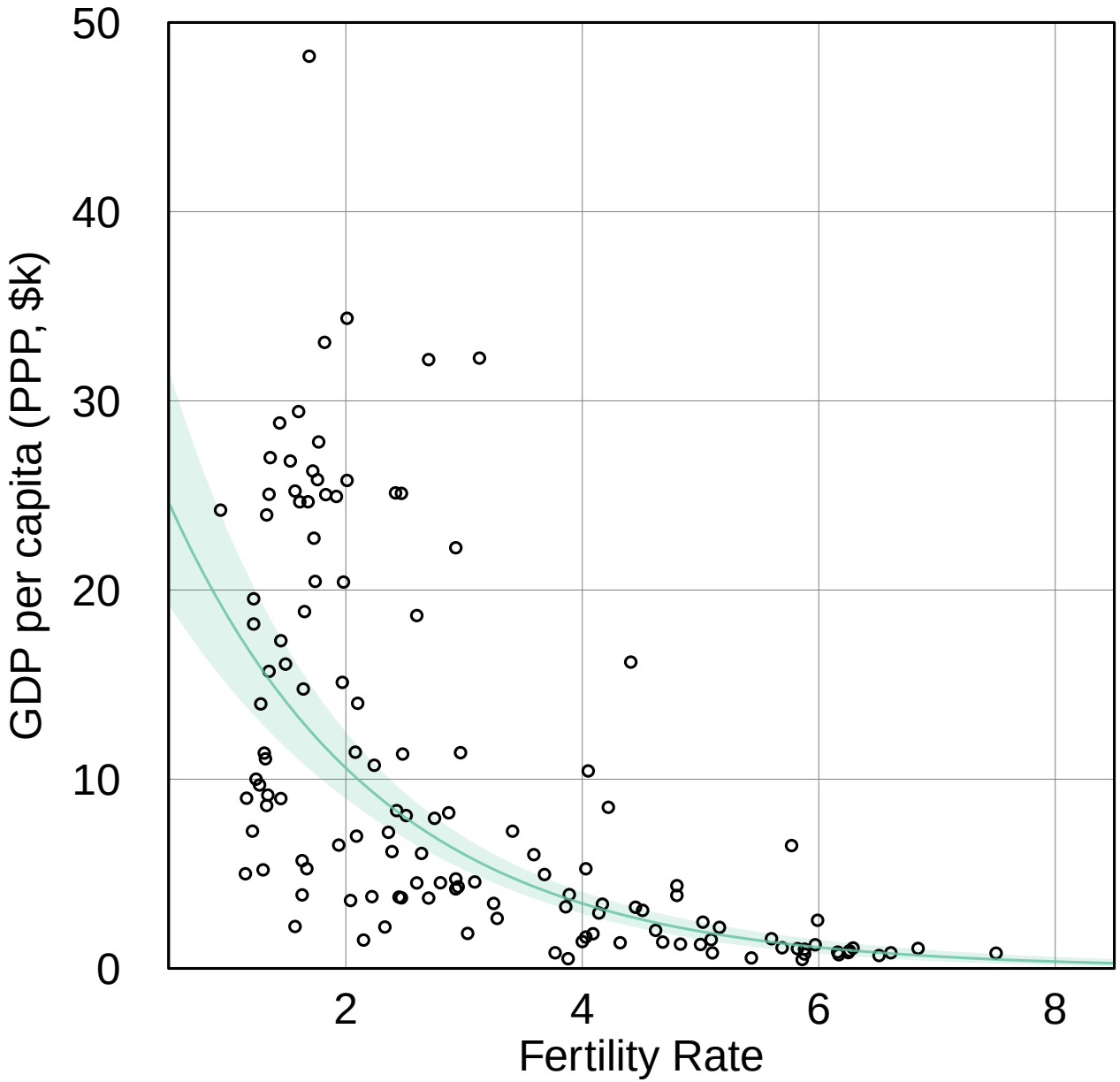
toprule Variable	Estimate	se	*t*-stat	*p*-value
Fertility	−0.56	0.0390	−14.33	< 0.001
Intercept	3.49	0.14	25.08	< 0.001
N	130			
R^2	0.61			
RMSE	0.746			

What does this table say?

If GDP explains fertility in a linear regression, then fertility will “explain” GDP!



Scatterplots can be rotated 90 deg, too



Scatterplots can be rotated 90 deg, too

Linear regression doesn't reveal anything about the direction of causation!

Correlation and causation

Correlation may be a big hint of causation, but it's not enough to demonstrate it

What would help?

How to establish causal relationships

An instrument Add to our linear regression a special variable which we are *certain* explains GDP but not Fertility, or vice versa. Very hard to find.

How to establish causal relationships

An instrument Add to our linear regression a special variable which we are *certain* explains GDP but not Fertility, or vice versa. Very hard to find.

An experiment Random assignment of high-fertility societies to high or low GDP would tell us if GDP affects fertility.

Random assignment of poor countries to high and low fertility would tell us if fertility affects GDP growth.

Neither of these experiments is remotely possible!

How to establish causal relationships

An instrument Add to our linear regression a special variable which we are *certain* explains GDP but not Fertility, or vice versa. Very hard to find.

An experiment Random assignment of high-fertility societies to high or low GDP would tell us if GDP affects fertility.

Random assignment of poor countries to high and low fertility would tell us if fertility affects GDP growth.

Neither of these experiments is remotely possible!

Careful qualitative research How about a little history?

Careful case studies of poor societies over time might help trace the process—the mechanisms—by which GDP growth discourages fertility (or vice versa)

Quantitative and qualitative methods can be *complementary*, especially for understanding the *why* of relationships uncovered in quantitative studies

Final thoughts on linear regression

Linear regression is a powerful tool for isolating conditional expectations of y given x after removing confounding variables

But vulnerable to many hazards:

- Outliers
- Reverse causation
- Selection bias
- Omitted variable bias

Advanced techniques can mitigate these problems, as well as deal with others

Topic of a sequence of required courses in most graduate social science programs

Also a topic of SOC/CSSS/STAT 321 Case-based Statistics, offered by your instructor in Fall 2010