CSSS/SOC/STAT 321 · Case-Based Statistics I

MULTIPLE REGRESSION

part 2

Christopher Adolph

Department of Political Science

and

Center for Statistics and the Social Sciences

University of Washington, Seattle

Motivation: Making Linear Regression Useful

So far linear regression is a limited tool for us:

It *can* control for confounders

But we don't yet know:

1. How to control for categorical variables,
2. What to do if we think our variables are related by a curve instead of a line

This week we tackle these issues,
making linear regression a flexible modeling tool

As consumers of social science, we'll learn how to understand linear regression tricks used in almost every application of the technique

Outline

Regression with binary and categorical covariates

Regression with interaction terms

Regression with transformed variables

Goodness of Fit using Cross-validation

Regression when the Dependent Variable is Binary

Dealing with Outliers

Correlation and Causation Revisited

What makes some American households wealthier than others?

We take the following data from the 2007 Survey of Consumer Finances:

Net Household Wealth The sum of financial and non-financial assets (e.g., vehicle and home equity), minus debt, in thousands of dollars

Education The education of the head of household, coded as less than high school, high school, some college, and college

Age The age of the head of household, in years

Race The self-identified race of the head of household: non-Hispanic white, black, Hispanic, Asian, or other

Specifying a regression model for Wealth

Our goal is to select and fit an appropriate *specification*, or set of explanatory variables, for Wealth. Two complications:

1. Two of our covariates are categorical: Education (which is ordered) and Race (which is nominal).
   We need additive or ratio level variables for our covariates in linear regression

Specifying a regression model for Wealth

Our goal is to select and fit an appropriate *specification*, or set of explanatory variables, for Wealth. Two complications:

1. Two of our covariates are categorical: Education (which is ordered) and Race (which is nominal).
   We need additive or ratio level variables for our covariates in linear regression

2. Wealth was strongly skewed to the right, and so the residuals in our regression are unlikely to be Normally distributed

   But linear regression assumes the errors are Normal

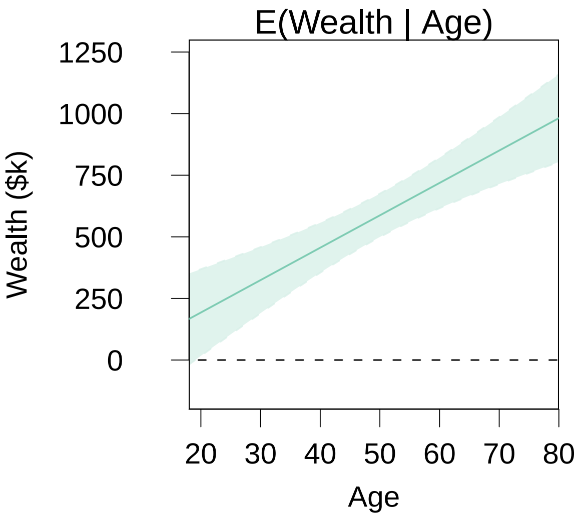
Before addressing these issues, let's consider a bivariate regression of Wealth on our only continuous covariate, Age

Regression of Wealth on Age

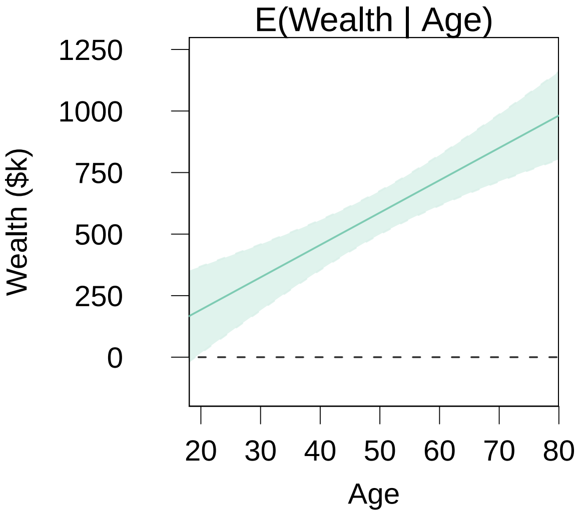
Variable	Estimate	se	t-stat	p-value
Age	13.2	2.6	5.04	$<$0.001
Intercept	-70.1	137.4	-0.51	0.61
N	10000			
R^2	0.002			
RMSE	4484			

How do we interpret the above?

Is this a good model?



How do we interpret this graph?



How do we interpret this graph?

Sometimes graphs like these will be much easier to understand than regression tables

Controlling for a categorical variable in linear regression

Few variables are likely to affect wealth as much as education

But we've measured Education as a categorical variable

What happens if we treat Education as an additive variable, where:
$\text{Education} = 1$ implies less than high school,
$\text{Education} = 2$ implies high school, and so on?

Controlling for a categorical variable in linear regression

Few variables are likely to affect wealth as much as education

But we've measured Education as a categorical variable

What happens if we treat Education as an additive variable, where:
$\text{Education} = 1$ implies less than high school,
$\text{Education} = 2$ implies high school, and so on?

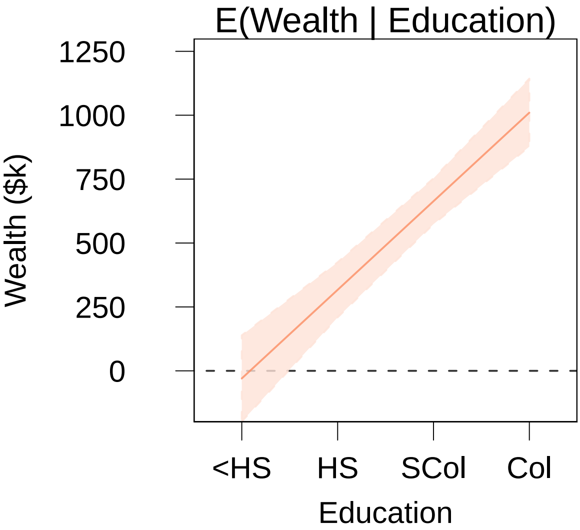
Including it in our regression would assume each step:
(from less than high school to high school,
and from high school to some college, etc.)
has the same effect on wealth

Is this reasonable?

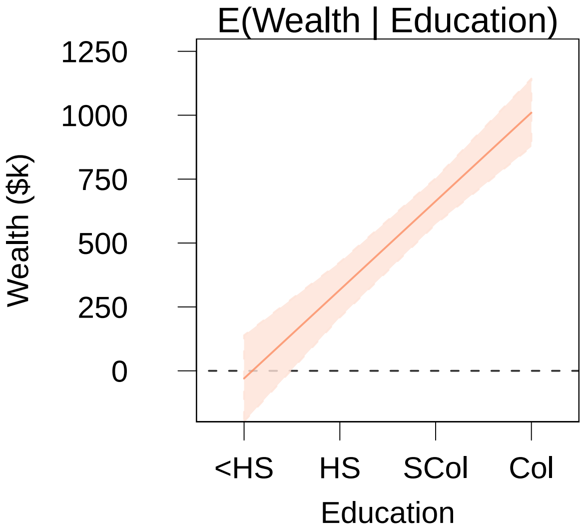
Regression of Wealth on Education (as an additive variable)

Variable	Estimate	se	t-stat	p-value
Education	355.9	113.8	9.21	$<$0.001
Intercept	-400	113.8	-3.51	$<$0.001
N	10000			
R^2	0.01			
RMSE	4187			

Ho do we interpret the above table?



Under this model, each stepwise increase in Education has the same predicted increase in wealth



Under this model, each stepwise increase in Education has the same predicted increase in wealth

Very strong assumption – probably unwarranted

Binary and Categorical Covariates in Linear Regression

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

Linear regression can easily accommodate binary covariates

Suppose that x_{2i} is binary

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

Linear regression can easily accommodate binary covariates

Suppose that x_{2i} is binary

Then if $x_{2i} = 1$,

Binary and Categorical Covariates in Linear Regression

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

Linear regression can easily accommodate binary covariates

Suppose that x_{2i} is binary

Then if $x_{2i} = 1$, our fitted y is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2$

Binary and Categorical Covariates in Linear Regression

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

Linear regression can easily accommodate binary covariates

Suppose that x_{2i} is binary

Then if $x_{2i} = 1$, our fitted y is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2$

But if $x_{2i} = 0$,

Binary and Categorical Covariates in Linear Regression

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

Linear regression can easily accommodate binary covariates

Suppose that x_{2i} is binary

Then if $x_{2i} = 1$, our fitted y is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2$

But if $x_{2i} = 0$, our fitted y is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i}$

In words, each binary variable increases y by its β coefficient when present

Let's use binary covariates to solve our problem with Education

Regression of Wealth on College (as a binary variable)

Variable	Estimate	se	t-stat	p-value
College	867.2	87.7	9.89	$<$0.001
Intercept	271.0	51.9	5.22	$<$0.001
N	10000			
R^2	0.01			
RMSE	4185			

Suppose instead we just control for College as a binary variable

How do we interpret the above table?

Expected wealth by college and non-college status

		95% CI	
Variable	Estimate	Lower	Upper
Less than College	271	169.2	372.8
College	1138.1	999.6	1276.7

The above table shows $\mathrm{E}(\mathrm{Wealth}|\mathrm{College})$, or the expected level of wealth given a college education, according to the model

What about other levels of education besides College?

Categorical covariates as sets of dummy variables

Let's create a set of dummy variables (another name for binary variables) for Education

That is, let's create High School, Some College, and College such that:

	High School	Some College	College
Someone with $<$ high school	0	0	0
Someone with high school	1	0	0
Someone with some college	0	1	0
Someone with college	0	0	1

Why don't we need a dummy for "less than high school"?

Categorical covariates as sets of dummy variables

Let's create a set of dummy variables (another name for binary variables) for Education

That is, let's create High School, Some College, and College such that:

	High School	Some College	College
Someone with $<$ high school	0	0	0
Someone with high school	1	0	0
Someone with some college	0	1	0
Someone with college	0	0	1

Why don't we need a dummy for "less than high school"?

Because that case is uniquely defined by the absence of our three dummy variables

What happens if we include our dummies in a regression?

$$\text{Wealth}_i = \beta_0 + \beta_1 \text{HS}_i + \beta_2 \text{SomeCol}_i + \beta_3 \text{College}_i + \varepsilon_i$$

Consider the fitted values for individuals at each level of education:

$$\text{E}(\text{Wealth}|\text{Less than HS}) \quad =$$

What happens if we include our dummies in a regression?

$$\text{Wealth}_i = \beta_0 + \beta_1 \text{HS}_i + \beta_2 \text{SomeCol}_i + \beta_3 \text{College}_i + \varepsilon_i$$

Consider the fitted values for individuals at each level of education:

$$\begin{aligned} \text{E}(\text{Wealth}|\text{Less than HS}) &= \beta_0 \\ \text{E}(\text{Wealth}|\text{HS}) &= \end{aligned}$$

What happens if we include our dummies in a regression?

$$\text{Wealth}_i = \beta_0 + \beta_1 \text{HS}_i + \beta_2 \text{SomeCol}_i + \beta_3 \text{College}_i + \varepsilon_i$$

Consider the fitted values for individuals at each level of education:

$$\begin{aligned}
\text{E}(\text{Wealth}|\text{Less than HS}) &= \beta_0 \\
\text{E}(\text{Wealth}|\text{HS}) &= \beta_0 \quad +\beta_1 \\
\text{E}(\text{Wealth}|\text{Some College}) &=
\end{aligned}$$

What happens if we include our dummies in a regression?

$$\text{Wealth}_i = \beta_0 + \beta_1 \text{HS}_i + \beta_2 \text{SomeCol}_i + \beta_3 \text{College}_i + \varepsilon_i$$

Consider the fitted values for individuals at each level of education:

$$\begin{array}{lclll}
\text{E}(\text{Wealth}|\text{Less than HS}) & = & \beta_0 & & \\
\text{E}(\text{Wealth}|\text{HS}) & = & \beta_0 & +\beta_1 & \\
\text{E}(\text{Wealth}|\text{Some College}) & = & \beta_0 & & +\beta_2 \\
\text{E}(\text{Wealth}|\text{College}) & = & & &
\end{array}$$

What happens if we include our dummies in a regression?

$$\text{Wealth}_i = \beta_0 + \beta_1 \text{HS}_i + \beta_2 \text{SomeCol}_i + \beta_3 \text{College}_i + \varepsilon_i$$

Consider the fitted values for individuals at each level of education:

$$\begin{array}{lclllll}
\text{E(Wealth|Less than HS)} & = & \beta_0 & & & \\
\text{E(Wealth|HS)} & = & \beta_0 & +\beta_1 & & \\
\text{E(Wealth|Some College)} & = & \beta_0 & & +\beta_2 & \\
\text{E(Wealth|College)} & = & \beta_0 & & & +\beta_3
\end{array}$$

Each possible case is *uniquely* defined and allowed to have its own effect on wealth

Regression of Wealth on Education (as dummy variables)

Variable	Estimate	se	t-stat	p-value
High School	151.9	133.3	1.14	0.25
Some College	225.3	149.1	1.51	0.13
College	1006.4	132.1	7.62	$<$0.001
Intercept	131.7	111.6	1.18	0.24
N	10000			
R^2	0.01			
RMSE	4185			

How do we interpret the above?

Expected wealth under different levels of education

		95% CI	
Scenario	Estimate	Lower	Upper
Less than HS	131.7	-86.9	350.4
High School	283.7	140.8	551.1
Some College	357.1	163.1	551.1
College	1138.1	999.6	1276.7

The above shows the fitted values of y and CIs for each possible education level

Does this remind you of anything?

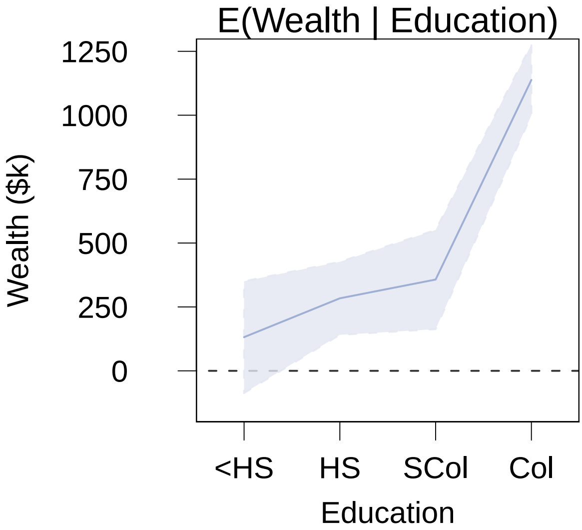
Expected wealth under different levels of education

		95% CI	
Scenario	Estimate	Lower	Upper
Less than HS	131.7	-86.9	350.4
High School	283.7	140.8	551.1
Some College	357.1	163.1	551.1
College	1138.1	999.6	1276.7

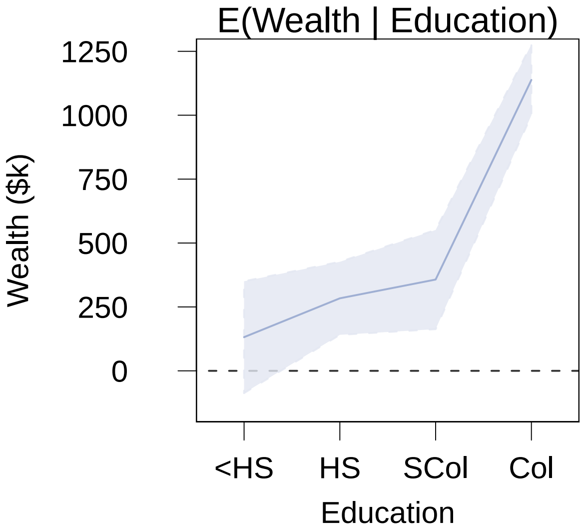
The above shows the fitted values of y and CIs for each possible education level

Does this remind you of anything? Comparison of means!

Linear regression encompasses the comparison of means test. Can do it by specifying the appropriate regression on dummy variables

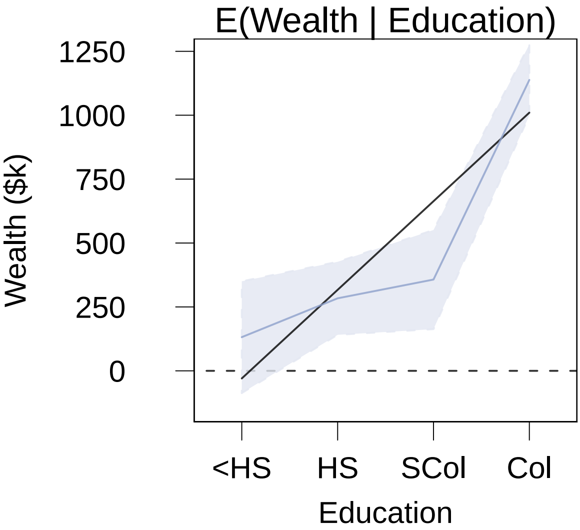


We can also present our results as a graphic

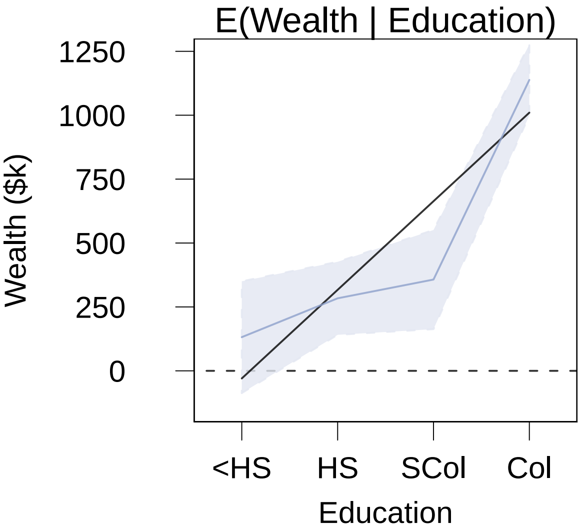


We can also present our results as a graphic

Note that the dummy variable specification is *flexible*. Doesn't have to follow a straight line



Compare the assumption that categories of Education have additive linear effects on Wealth



Converting to dummies for all but one category avoids over-simplication

Regression with continuous and dummy covariates

$$\text{Wealth}_i = \beta_0 + \beta_1 \text{HS}_i + \beta_2 \text{SomeCol}_i + \beta_3 \text{College}_i + \beta_4 \text{Age} + \varepsilon_i$$

When a regression includes continuous and categorical covariates,
think of the categories as *shifting* the sloped lines defined by the continuous covariates

$$\text{Wealth}_i = \beta_0 + \beta_1\text{HS}_i + \beta_2\text{SomeCol}_i + \beta_3\text{College}_i + \beta_4\text{Age} + \varepsilon_i$$

When a regression includes continuous and categorical covariates,
think of the categories as *shifting* the sloped lines defined by the continuous covariates

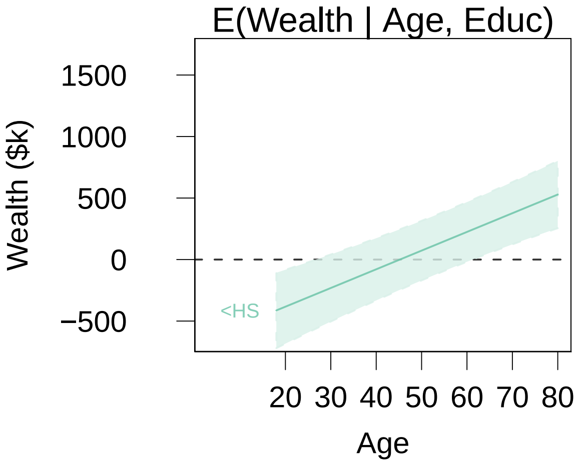
$$\begin{array}{lclllll}
\text{E(Wealth}|\text{Less than HS)} & = & \beta_0 & & & & +\beta_4\text{Age} \\
\text{E(Wealth}|\text{HS)} & = & \beta_0 & +\beta_1 & & & +\beta_4\text{Age} \\
\text{E(Wealth}|\text{Some College)} & = & \beta_0 & & +\beta_2 & & +\beta_4\text{Age} \\
\text{E(Wealth}|\text{College)} & = & \beta_0 & & & +\beta_3 & +\beta_4\text{Age}
\end{array}$$

The terms β_0 through β_3 only shift the *intercept* of the regression line whose slope is β_4

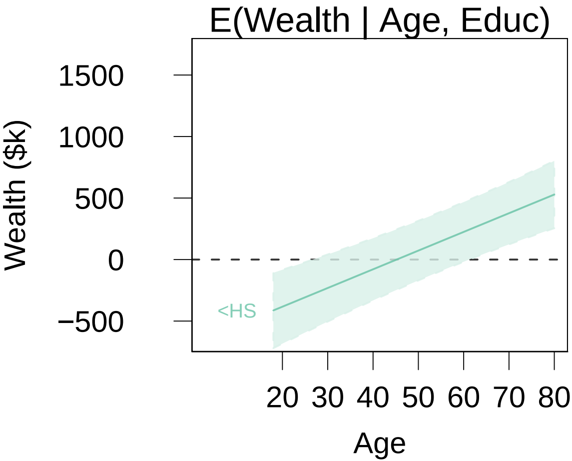
Regression of Wealth on Age and Education

Variable	Estimate	se	t-stat	p-value
Age	15.2	2.63	5.79	$<$0.001
High School	182.5	146.1	1.25	0.212
Some College	446.4	163.0	2.74	0.006
College	1038.8	144.6	7.18	$<$0.001
Intercept	-687.1	189.3	-3.63	$<$0.001
N	10000			
R^2	0.01			
RMSE	4466			

How do we interepret the above table?

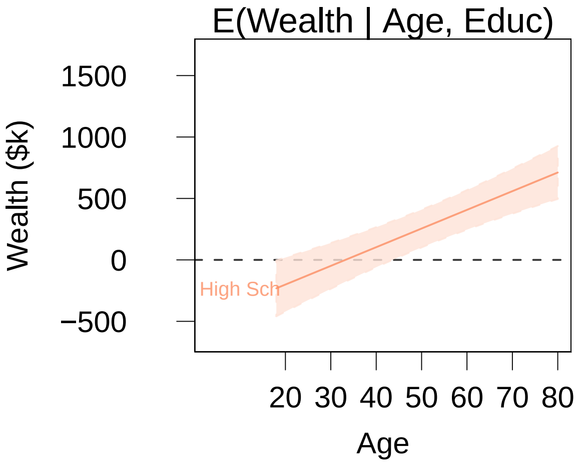


As models get more complicated, graphics become more helpful

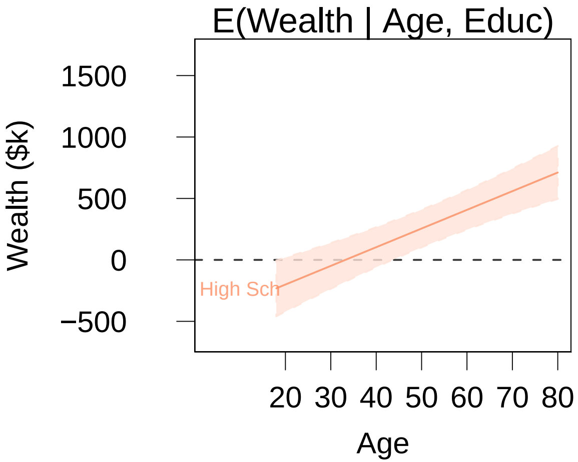


As models get more complicated, graphics become more helpful

Here we plot out the expected wealth for high school dropouts at different ages

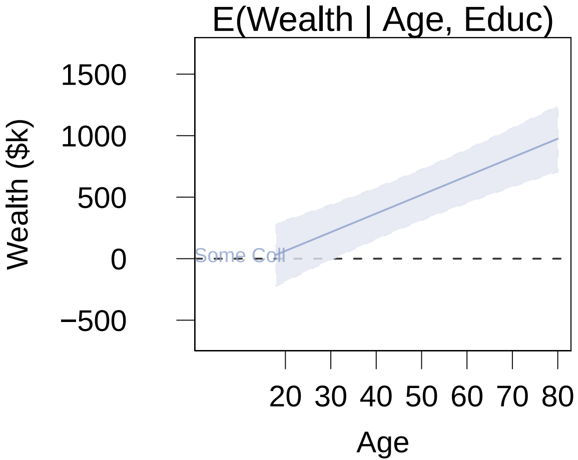


By changing the hypothetical level of education, we get a new picture

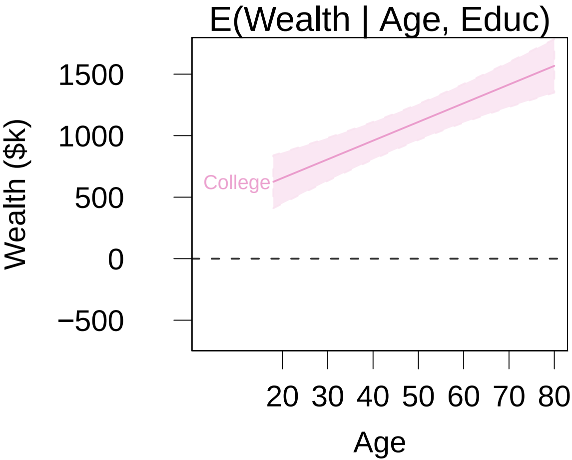


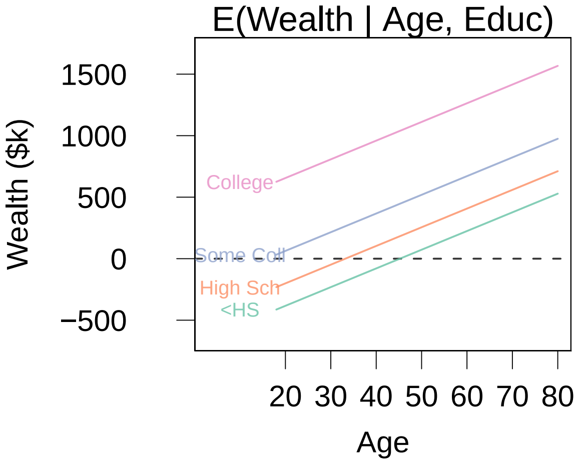
By changing the hypothetical level of education, we get a new picture

Here we plot out the expected wealth for high school graduates at different ages

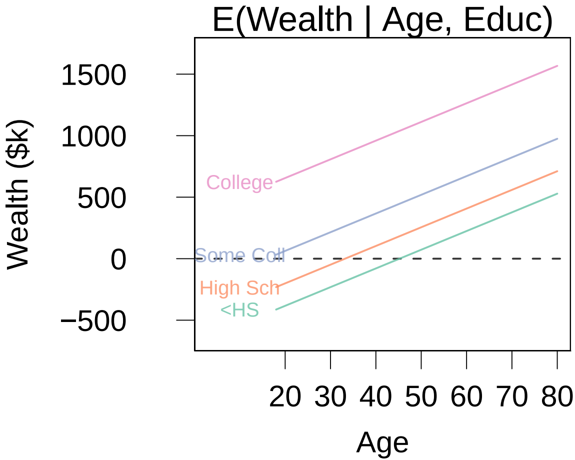


Comparing the pictures reveals that changing categories only changes the intercept, not the slope


E(Wealth | Age, Educ)
Wealth ($k)
1500
1000
500
0
−500
College
20 30 40 50 60 70 80
Age

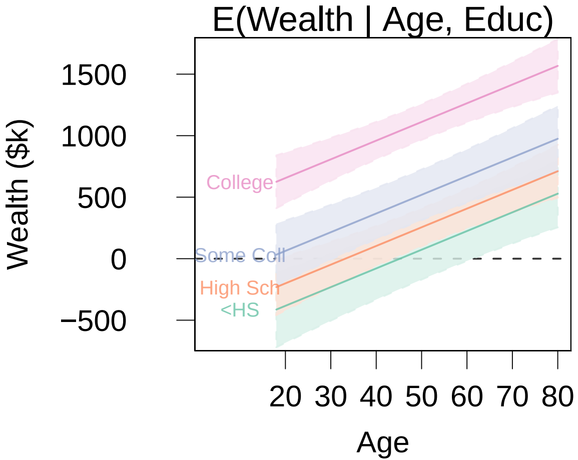


We can collect the whole model on a single slide

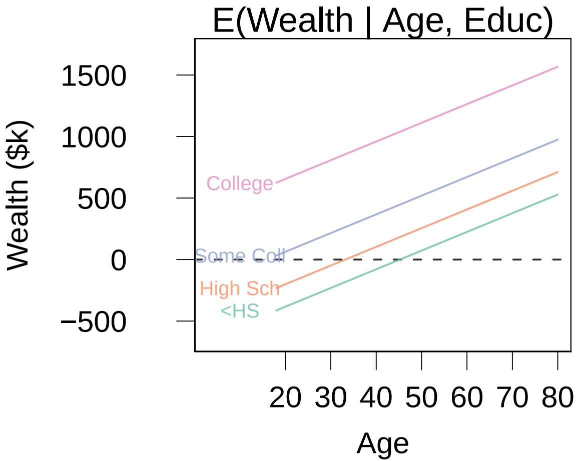


We can collect the whole model on a single slide

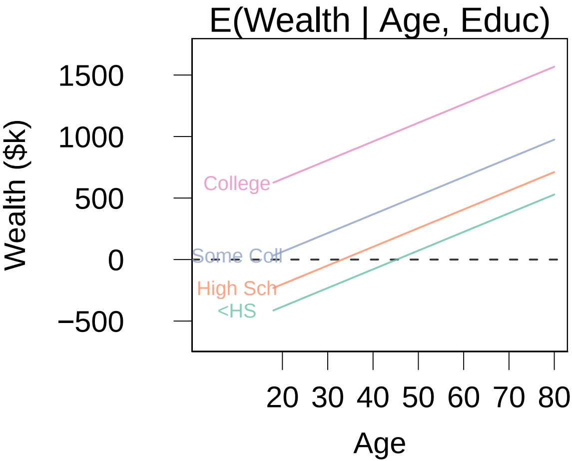
How do we interpret this picture?



Adding confidence intervals



Reasonable to assume these slopes are the same?



Reasonable to assume these slopes are the same?

What if college grads' wealth grows faster as they age than high school drop-outs'?

Conditional slopes

We've assumed the same slope applies to everyone in our sample

But what if this is too restrictive?

What if our theory implies different slopes for different groups?

In that case, we need *conditional slopes*

A different conditional slope applies for each group

Interaction terms

$$\text{Wealth}_i \quad = \quad \beta_0 + \beta_1 \text{HS}_i + \beta_2 \text{SomeCol}_i + \beta_3 \text{College}_i + \beta_4 \text{Age} +$$

Interaction terms

$$\begin{aligned}\text{Wealth}_i &= \beta_0 + \beta_1\text{HS}_i + \beta_2\text{SomeCol}_i + \beta_3\text{College}_i + \beta_4\text{Age} + \\ &\quad \beta_5\text{Age} \times \text{HS}_i\end{aligned}$$

Interaction terms

$$\begin{aligned}\text{Wealth}_i &= \beta_0 + \beta_1 \text{HS}_i + \beta_2 \text{SomeCol}_i + \beta_3 \text{College}_i + \beta_4 \text{Age} + \\ &\quad \beta_5 \text{Age} \times \text{HS}_i + \beta_6 \text{Age} \times \text{SomeCol}_i\end{aligned}$$

Interaction terms

$$\begin{aligned} \text{Wealth}_i &= \beta_0 + \beta_1 \text{HS}_i + \beta_2 \text{SomeCol}_i + \beta_3 \text{College}_i + \beta_4 \text{Age} + \\ &\quad \beta_5 \text{Age} \times \text{HS}_i + \beta_6 \text{Age} \times \text{SomeCol}_i + \beta_7 \text{Age} \times \text{College}_i \\ &\quad + \varepsilon_i \end{aligned}$$

What does the above imply for different levels of education?

$$\begin{aligned}\text{Wealth}_i &= \beta_0 + \beta_1 \text{HS}_i + \beta_2 \text{SomeCol}_i + \beta_3 \text{College}_i + \beta_4 \text{Age} + \\ &\quad \beta_5 \text{Age} \times \text{HS}_i + \beta_6 \text{Age} \times \text{SomeCol}_i + \beta_7 \text{Age} \times \text{College}_i \\ &\quad + \varepsilon_i\end{aligned}$$

What does the above imply for different levels of education?

$$\text{E}(\text{Wealth} \mid \text{Less than HS}) = \beta_0 \qquad + \beta_4 \text{Age}$$

$$\begin{aligned}\text{Wealth}_i &= \beta_0 + \beta_1\text{HS}_i + \beta_2\text{SomeCol}_i + \beta_3\text{College}_i + \beta_4\text{Age} + \\ &\quad \beta_5\text{Age}\times\text{HS}_i + \beta_6\text{Age}\times\text{SomeCol}_i + \beta_7\text{Age}\times\text{College}_i \\ &\quad +\varepsilon_i\end{aligned}$$

What does the above imply for different levels of education?

$$\begin{aligned}\text{E(Wealth}|\text{Less than HS)} &= \beta_0 && +\beta_4\text{Age} \\ \text{E(Wealth}|\text{HS)} &= \beta_0 \quad +\beta_1 && +(\beta_4+\beta_5)\text{Age}\end{aligned}$$

Interaction terms

$$\begin{aligned}\text{Wealth}_i \quad = \quad & \beta_0 + \beta_1 \text{HS}_i + \beta_2 \text{SomeCol}_i + \beta_3 \text{College}_i + \beta_4 \text{Age} + \\ & \beta_5 \text{Age} \times \text{HS}_i + \beta_6 \text{Age} \times \text{SomeCol}_i + \beta_7 \text{Age} \times \text{College}_i \\ & + \varepsilon_i\end{aligned}$$

What does the above imply for different levels of education?

$$\begin{array}{lcllll}
\text{E(Wealth | Less than HS)} & = & \beta_0 & & & +\beta_4 \text{Age} \\
\text{E(Wealth | HS)} & = & \beta_0 & +\beta_1 & & +(\beta_4 + \beta_5)\text{Age} \\
\text{E(Wealth | Some College)} & = & \beta_0 & & +\beta_2 & +(\beta_4 + \beta_6)\text{Age}
\end{array}$$

Interaction terms

$$\begin{aligned}
\text{Wealth}_i = \; & \beta_0 + \beta_1 \text{HS}_i + \beta_2 \text{SomeCol}_i + \beta_3 \text{College}_i + \beta_4 \text{Age} + \\
& \beta_5 \text{Age} \times \text{HS}_i + \beta_6 \text{Age} \times \text{SomeCol}_i + \beta_7 \text{Age} \times \text{College}_i \\
& + \varepsilon_i
\end{aligned}$$

What does the above imply for different levels of education?

$$\begin{array}{lclllll}
\text{E(Wealth|Less than HS)} & = & \beta_0 & & & & +\beta_4 \text{Age} \\
\text{E(Wealth|HS)} & = & \beta_0 & +\beta_1 & & & +(\beta_4 + \beta_5)\text{Age} \\
\text{E(Wealth|Some College)} & = & \beta_0 & & +\beta_2 & & +(\beta_4 + \beta_6)\text{Age} \\
\text{E(Wealth|College)} & = & \beta_0 & & & +\beta_3 & +(\beta_4 + \beta_7)\text{Age}
\end{array}$$

A different slope and intercept for each educational category

Regression of Wealth on Age and Education (with interactions)

Variable	Estimate	se	t-stat	p-value
Age	4.27	5.64	0.76	0.449
High School	39.80	396.70	0.10	0.920
Some College	-150.70	436.90	-0.35	0.730
College	-854.60	408.40	-2.09	0.036
Age $\times$ High School	2.53	6.96	0.36	0.717
Age $\times$ Some College	9.00	8.12	1.11	0.268
Age $\times$ College	38.76	7.36	5.27	$<$0.001
Intercept	-103.40	329.90	-0.31	0.754
N	10000			
R^2	0.02			
RMSE	4167			

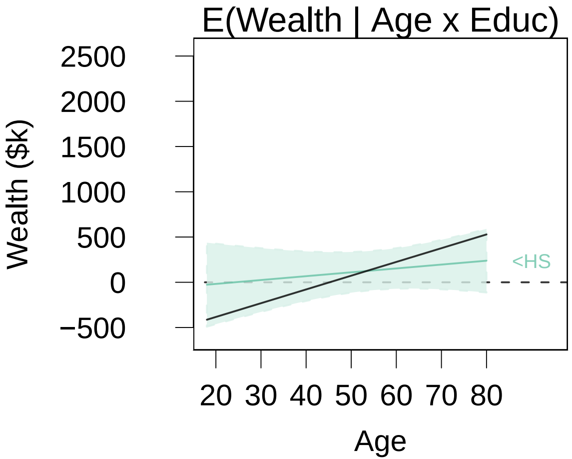
How do we interpret the above?

Regression of Wealth on Age and Education (with interactions)

Variable	Estimate	se	t-stat	p-value
Age	4.27	5.64	0.76	0.449
High School	39.80	396.70	0.10	0.920
Some College	-150.70	436.90	-0.35	0.730
College	-854.60	408.40	-2.09	0.036
Age $\times$ High School	2.53	6.96	0.36	0.717
Age $\times$ Some College	9.00	8.12	1.11	0.268
Age $\times$ College	38.76	7.36	5.27	$<$0.001
Intercept	-103.40	329.90	-0.31	0.754
N	10000			
R^2	0.02			
RMSE	4167			

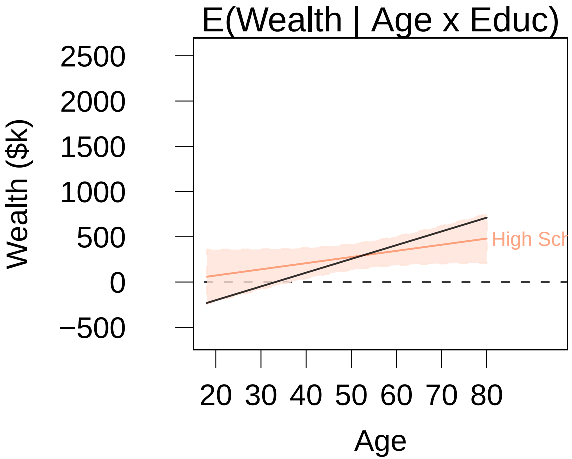
How do we interpret the above? Hard to do!

Interaction terms best interpreted graphically!

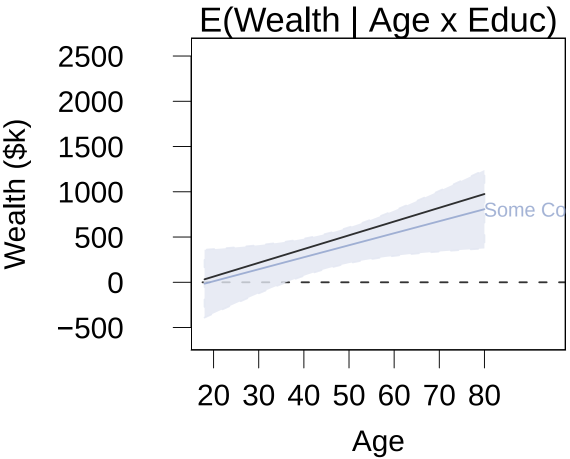


The education-conditional slope is in color

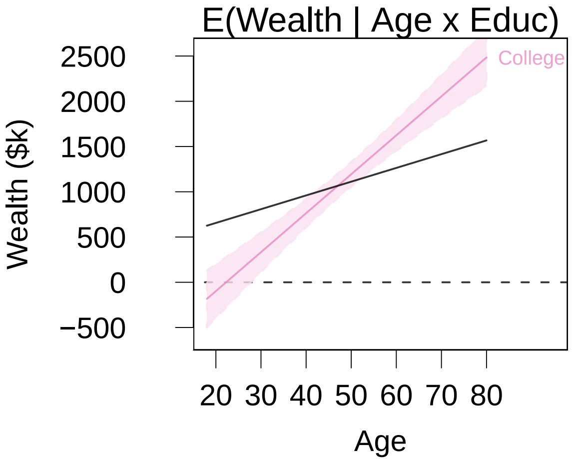
The uninteracted slope is in black for comparison



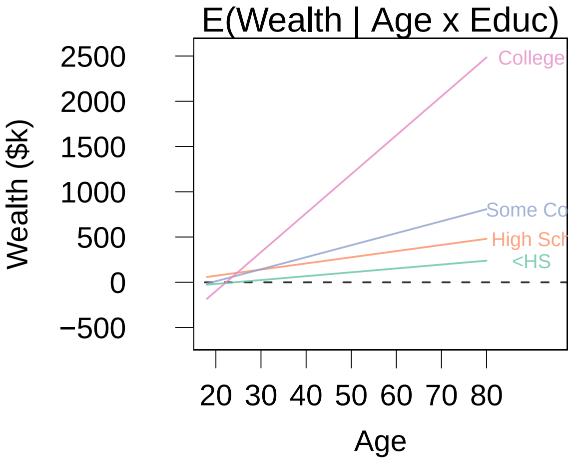
What happens to the Age slope as we increase Education?



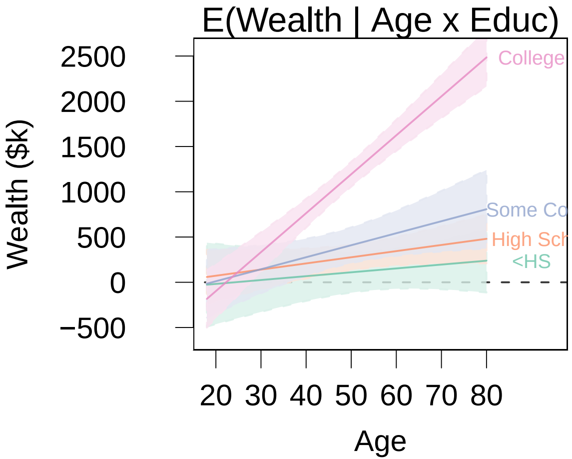
Can we summarize the substance of this effect in words?



Does the interactive model make high education more or less attractive?



All conditional slopes

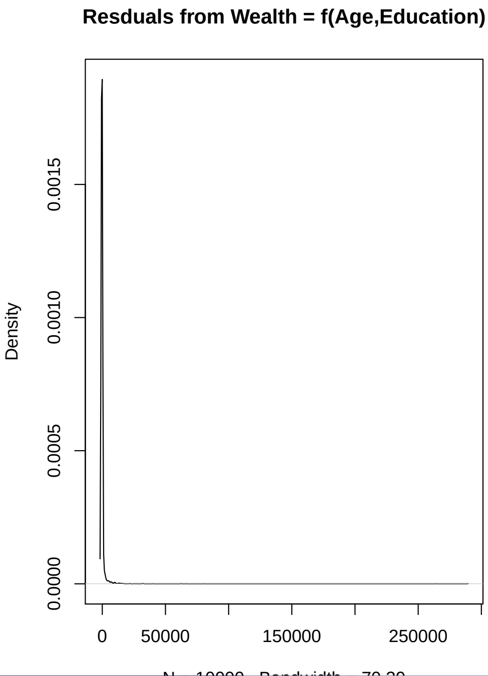

E(Wealth | Age x Educ)
Wealth ($k)
2500
2000
1500
1000
500
0
−500
20
30
40
50
60
70
80
Age
College
Some Co
High Sch
<HS

With 95% CIs

The fit for our models so far seems quite poor

We mispredict on average by over 4 million in wealth!

A look at our residuals may clarify matters



Tremendous right skew!

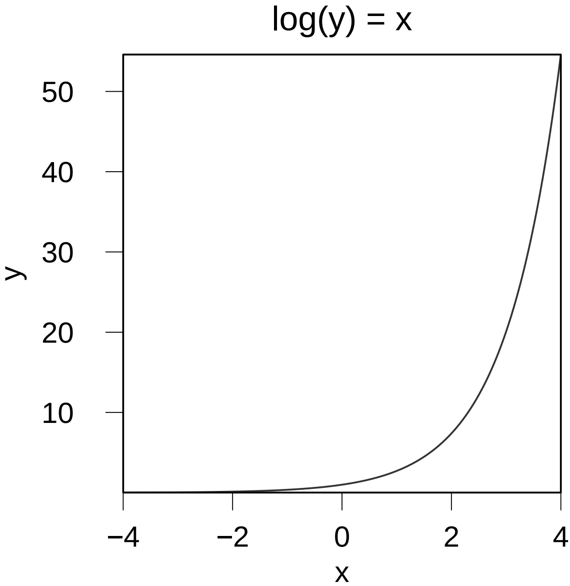
Transforming variables

Wealth is clearly not a linear function of Education and Age

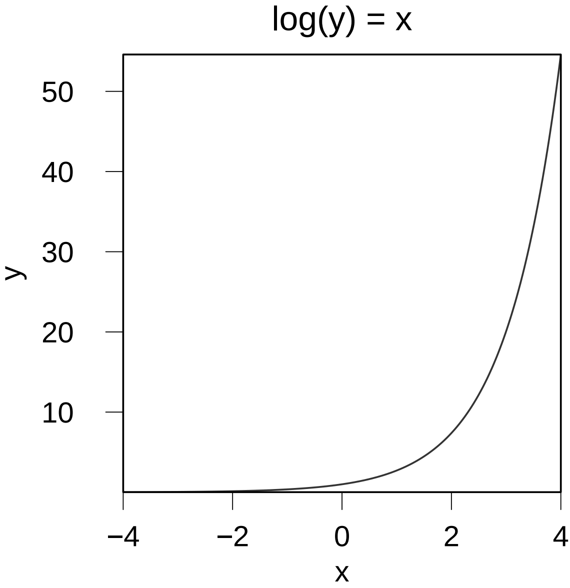
Some individuals amass many millions in wealth;
our model predicts everyone will have somewhere under 1.5 million

But even if Wealth isn't a linear function of our covariates, perhaps some *tranformed* version of Wealth is

For example, perhaps $\log(\text{Wealth})$ is a linear function of Age and Education

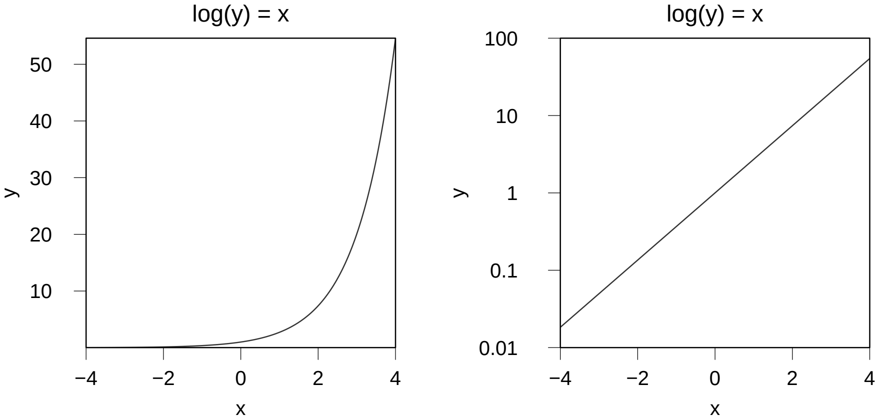


At left is a log-linear relationship



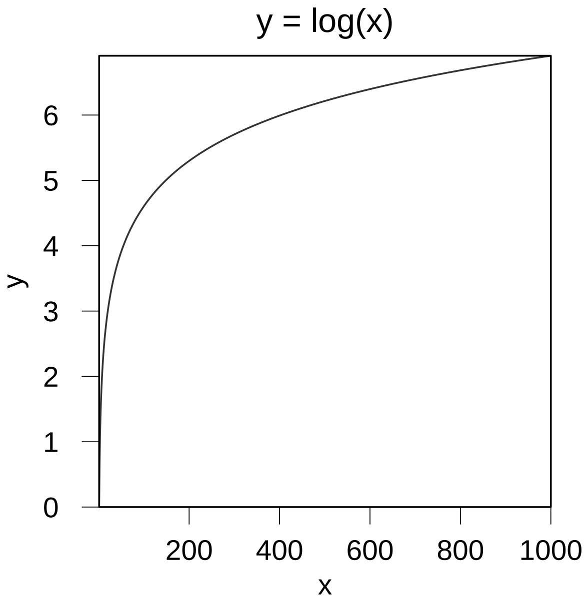
At left is a log-linear relationship

$\log(y)$ is a function of x

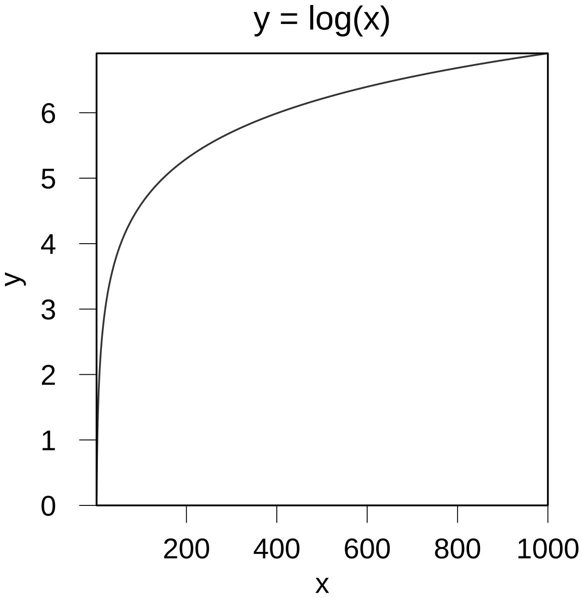


With the right "squeezing" of the y-axis, this relationship can appear linear

We can think of this as a case where
a level change in x causes a percentage change in y

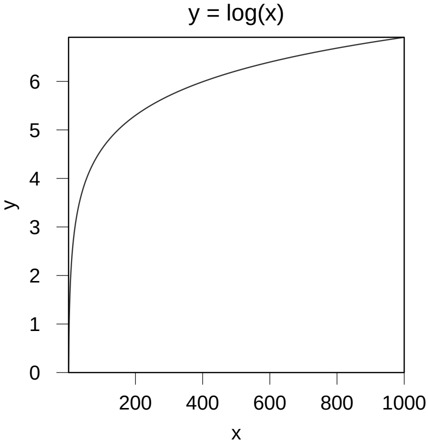


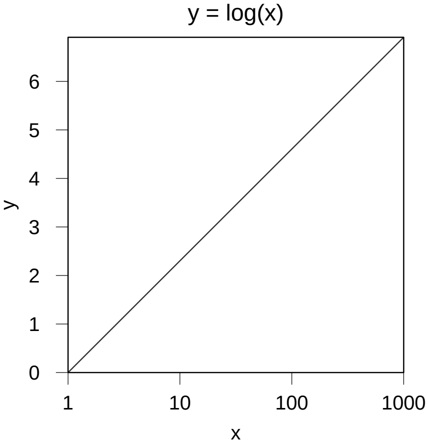
Another possibility is that the level of y is a function of the $\log(x)$



Another possibility is that the level of y is a function of the $\log(x)$

This implies diminishing returns for x on y





With the right "squeezing" of the x-axis, this relationship can appear linear

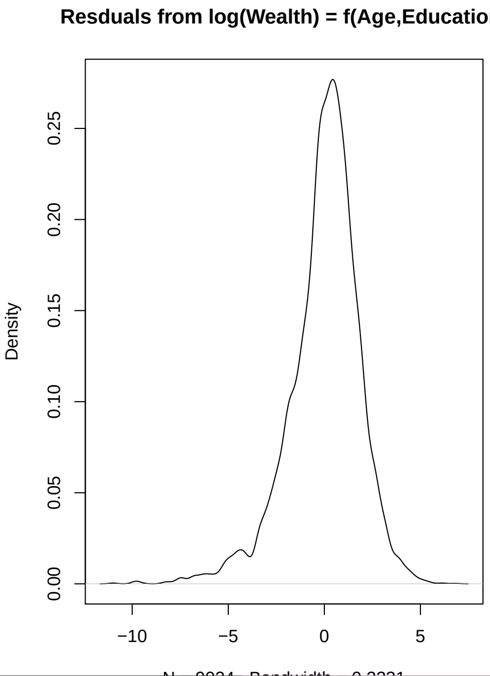
We can think of this as a case where
a percentage change in x causes a level change in y

Regression of log(Wealth) on Age and Education

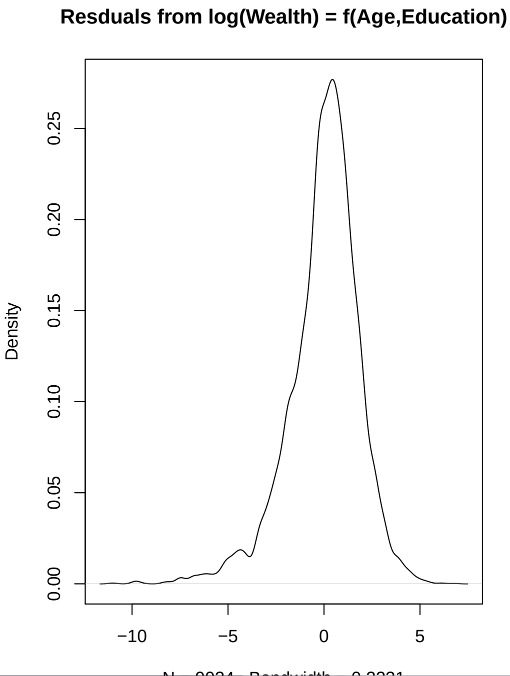
Variable	Estimate	se	t-stat	p-value
Age	0.045	0.001	38.49	$<$0.001
High School	1.016	0.063	16.05	$<$0.001
Some College	1.424	0.072	19.84	$<$0.001
College	2.552	0.063	40.55	$<$0.001
Intercept	0.890	0.086	10.53	$<$0.001
N	9024			
R^2	0.26			
RMSE	1.84			

Logging the dependent variable has drastically improved the fit

But the coefficients are harder to interpret – so we'll use graphs!

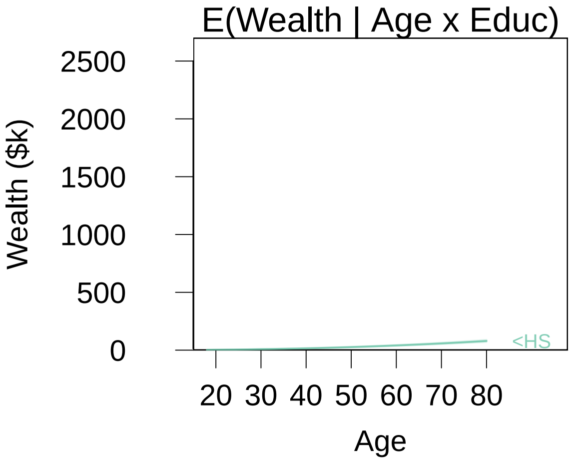


Now the residuals look approx Normal

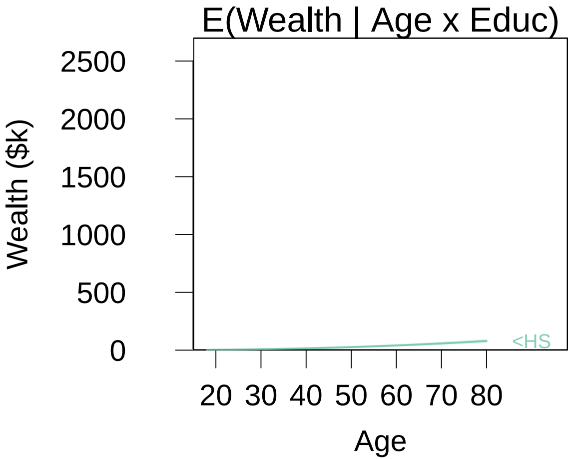


Now the residuals look approx Normal

Suggests $\log(\text{Wealth})$ was an appropriate transformation

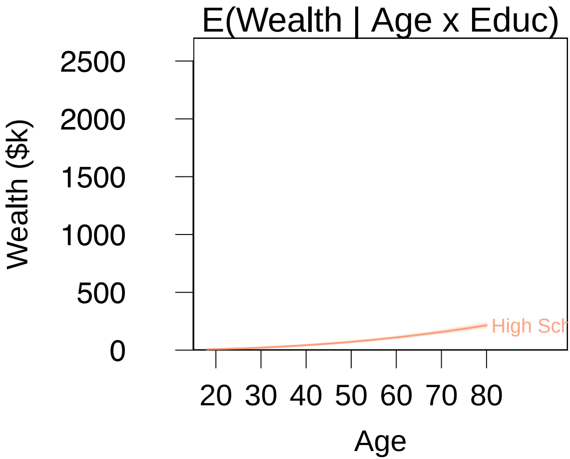


We'll plot the fitted wealth given Age and Education as usual

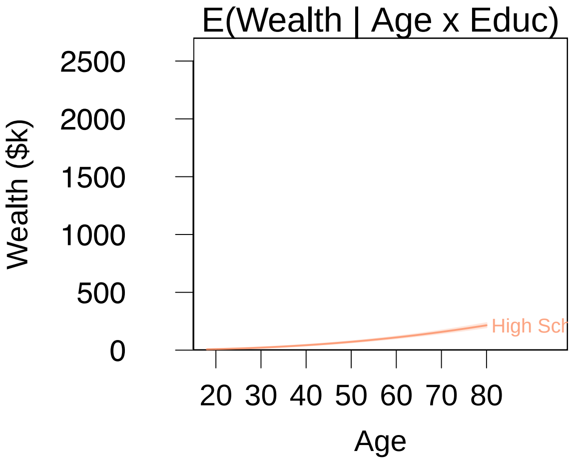


We'll plot the fitted wealth given Age and Education as usual

Note the linear scales

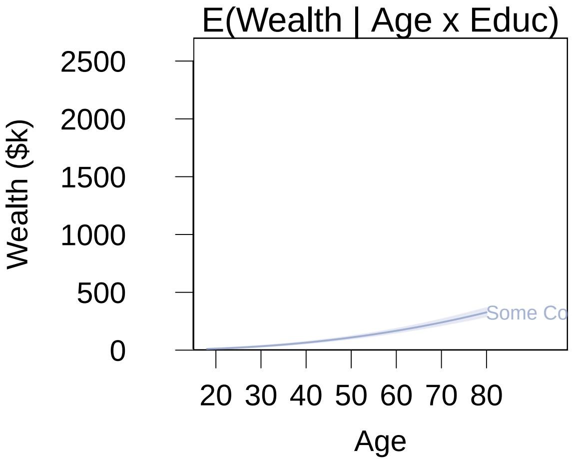


The log specification has added curvature

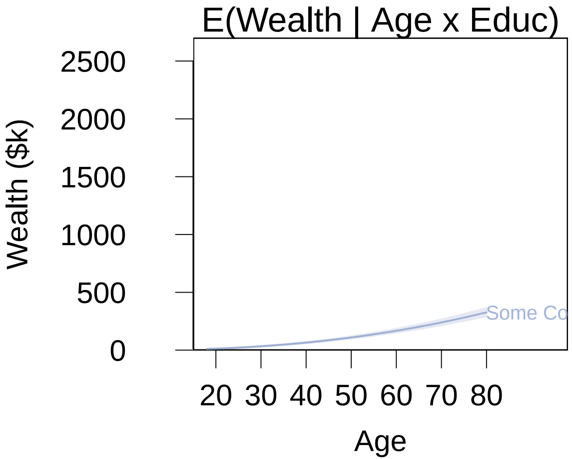


The log specification has added curvature

The slope is no longer constant

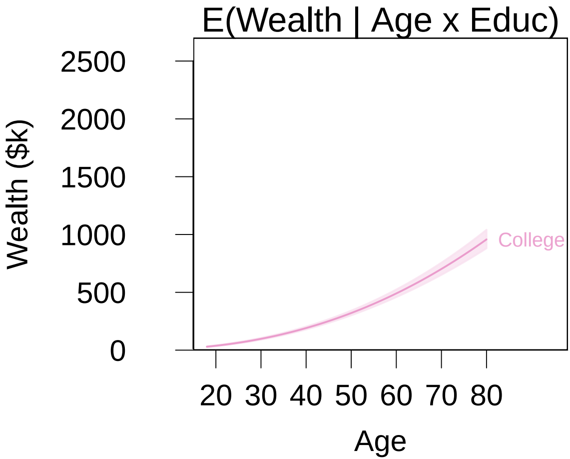


In fact, the slope gets bigger as Age rises

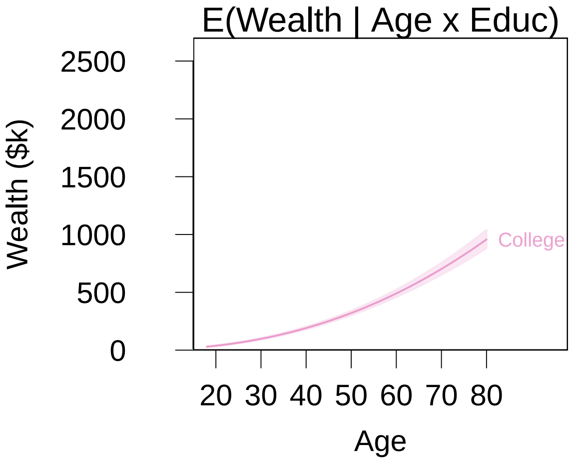


In fact, the slope gets bigger as Age rises

Fits the model of compound interest, and our expectations about earnings

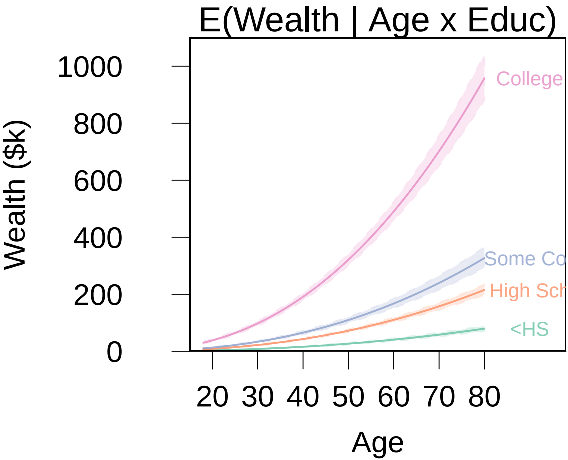


Note that the log transformation is capturing something similar to our interactions

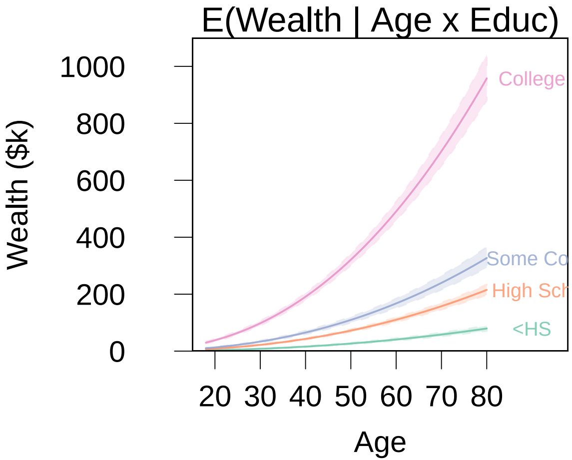


Note that the log transformation is capturing something similar to our interactions

Steeper slopes for higher education

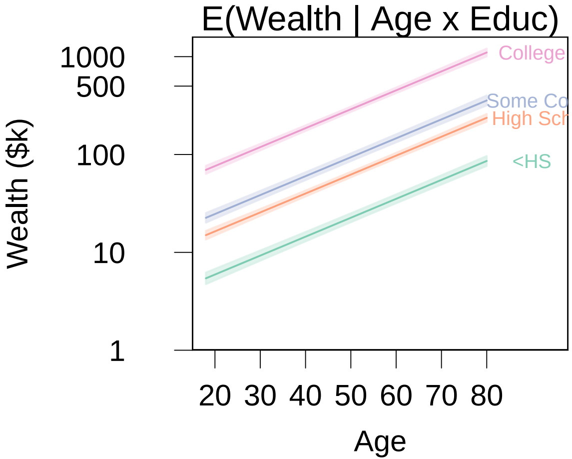


We now see the reason for the differing slopes

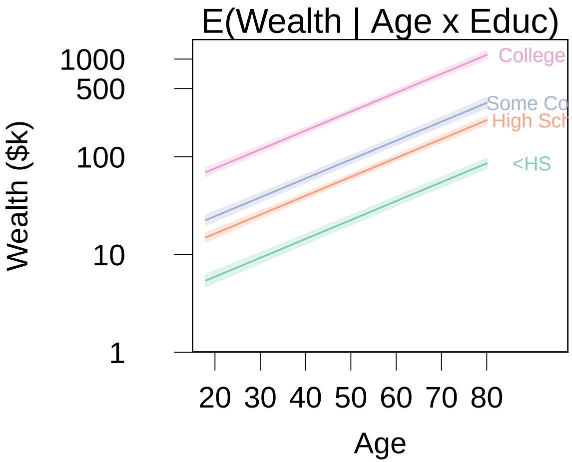


We now see the reason for the differing slopes

The more money you have, the easier it is to make more



If we squeezed the Wealth axis just right, a linear relationship will appear



If we squeezed the Wealth axis just right, a linear relationship will appear

This is still *linear* regression, just on a log scale

A final model: Adding Race

Several weeks ago, we considered the relationship of Race and Wealth

We used primitive methods: histograms of the data for each group; no controls

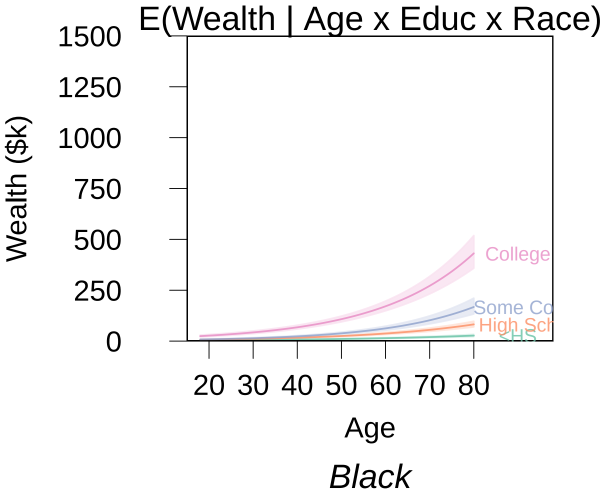
We “found” that Blacks and Hispanics were at a major, and essentially equal, disadvantage in wealth

Will that finding hold up controlling for Age and Education?

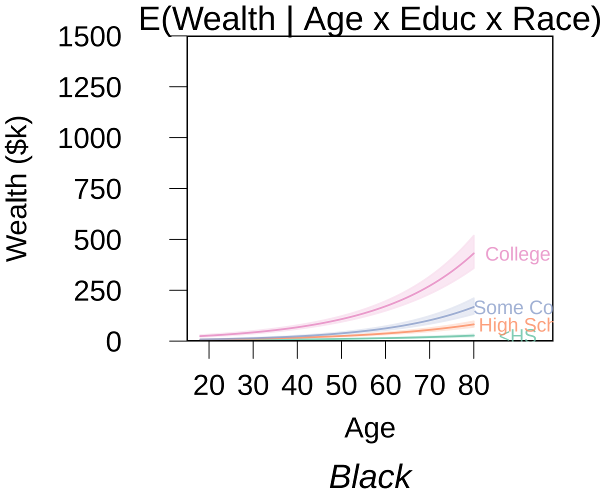
Let’s add dummies for Race into a model incorporating everything else we’ve explored

Regression of log(Wealth) on Age, Education, and Race

Variable	Estimate	se	t-stat	p-value
Age	0.031	0.003	11.58	$<$0.001
High School	0.464	0.192	2.42	0.016
Some College	0.396	0.215	1.84	0.066
College	1.588	0.201	7.89	$<$0.001
Age $\times$ High School	0.008	0.003	2.49	0.013
Age $\times$ Some College	0.018	0.004	4.61	$<$0.001
Age $\times$ College	0.015	0.003	4.23	$<$0.001
Black	-1.082	0.063	-16.97	$<$0.001
Hispanic	-0.428	0.072	-5.95	$<$0.001
Intercept	1.866	0.168	11.11	$<$0.001
N	9024			
R^2	0.29			
RMSE	1.81			

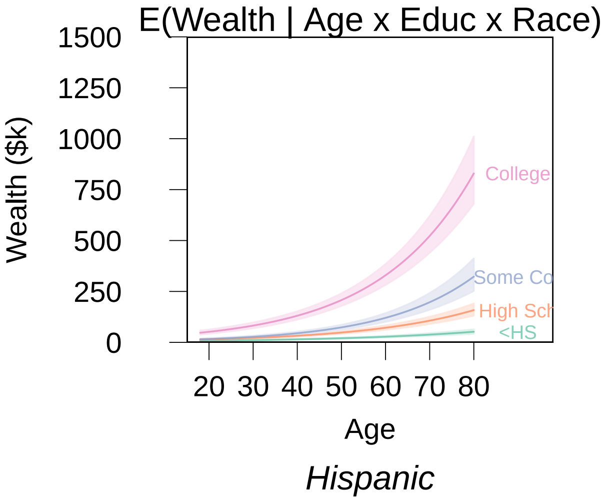


Graphics again help clarify a complex model

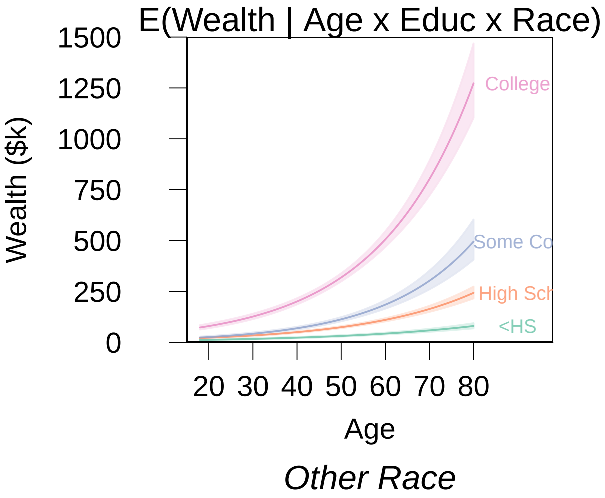


Graphics again help clarify a complex model

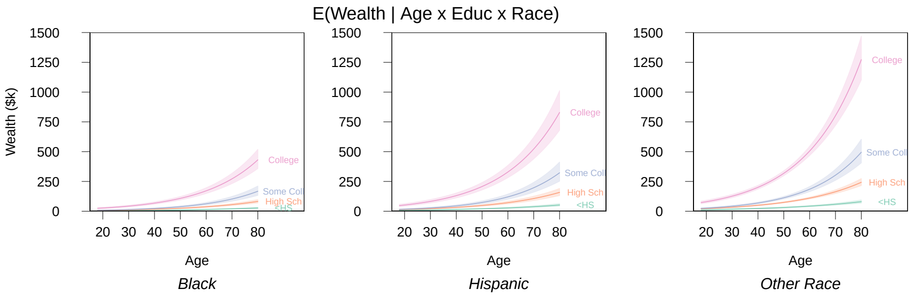
At left is the expected wealth for Black households, at different levels of Education and Age



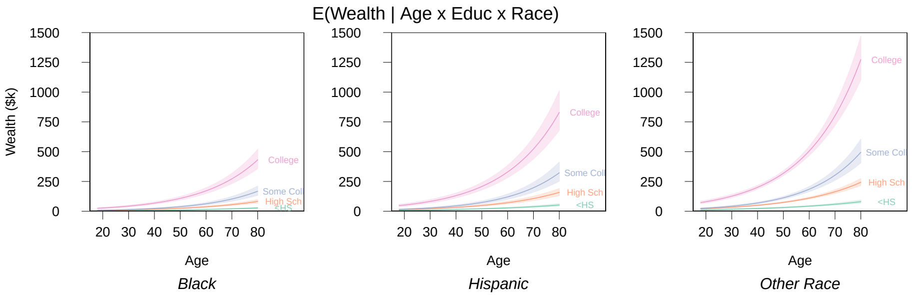
We can compare to Hispanics



And all other households (mostly Whites)

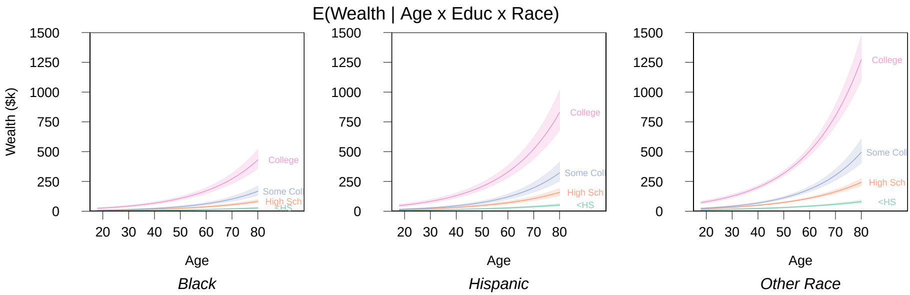


What do we make of the above results?



What do we make of the above results?

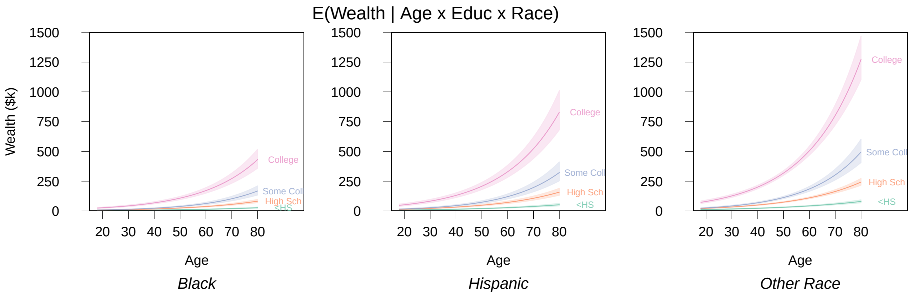
How much would we expect an individual household to deviate from this prediction?



Potentially a lot: RMSE = 1.81 on the log scale.

For example, if we predict that a household will have \$500k, given their characteristics, we would be unsurprised if they really had

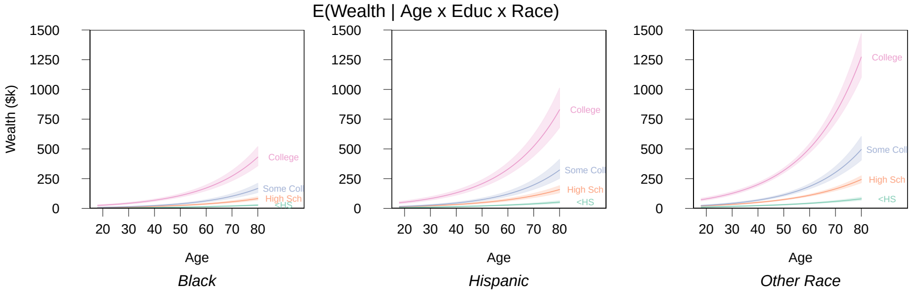
$$\log(500) \pm 1.8 \quad = \quad 6.2 \pm 1.8 \text{ logged dollars}$$



Potentially a lot: RMSE = 1.81 on the log scale.

For example, if we predict that a household will have $500k, given their characteristics, we would be unsurprised if they really had

$$\begin{aligned}\log(500) \pm 1.8 &= 6.2 \pm 1.8 \text{ logged dollars} \\ &= \text{between } \exp(6.2 - 1.8) \text{ and } \exp(6.2 + 1.8)\end{aligned}$$



Potentially a lot: RMSE = 1.81 on the log scale.

For example, if we predict that a household will have $500k, given their characteristics, we would be unsurprised if they really had

$$\begin{aligned}\log(500) \pm 1.8 &= 6.2 \pm 1.8 \text{ logged dollars} \\ &= \text{between } \exp(6.2 - 1.8) \text{ and } \exp(6.2 + 1.8) \\ &= \text{between } \$81.5\text{k and } \$2981.0\text{k}\end{aligned}$$

Regression with transformed variables

In the last example,
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

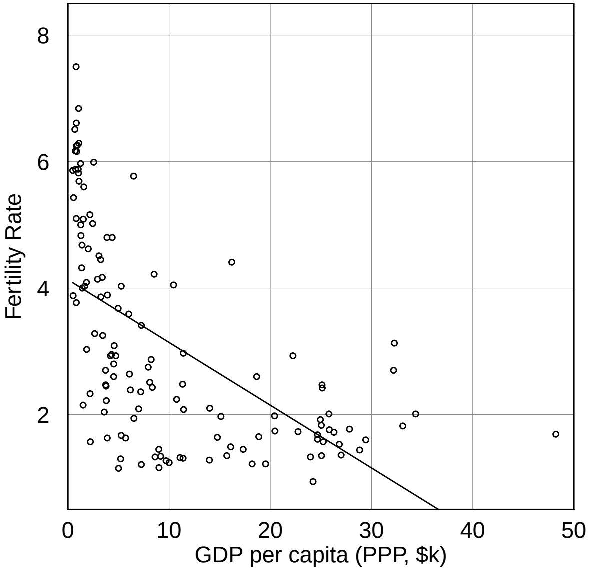
Sometimes, we will instead expect percentage changes in a *covariate* to lead to level changes in our response variable

Regression of Fertility on Education Ratio & GDP

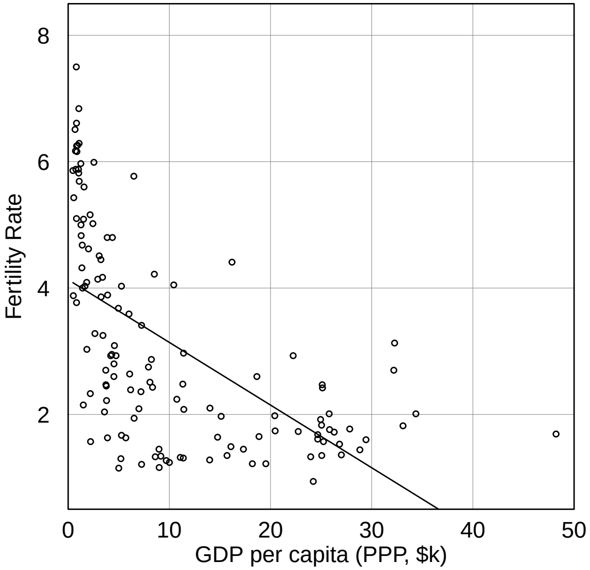
Variable	Estimates	se	t-stat	p-value
Intercept	11.25	(0.73)	15.46	$<$0.001
Education Ratio	-0.08	(0.01)	-9.93	$<$0.001
GDP per capita ($k)	-0.05	(0.01)	-5.32	$<$0.001
N	130			
R^2	0.64			
RMSE	1.01			

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

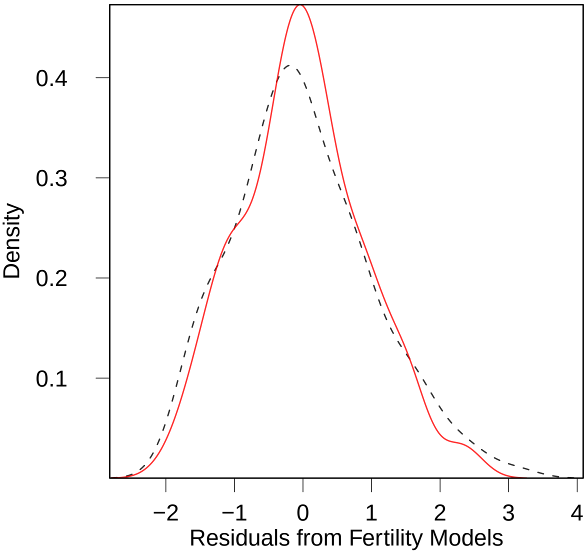
This model now assumes diminishing returns to GDP increases

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

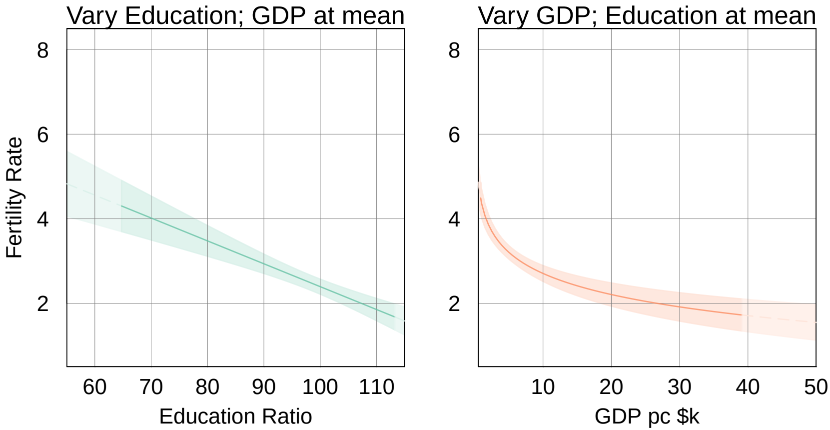
Four regression models of fertility

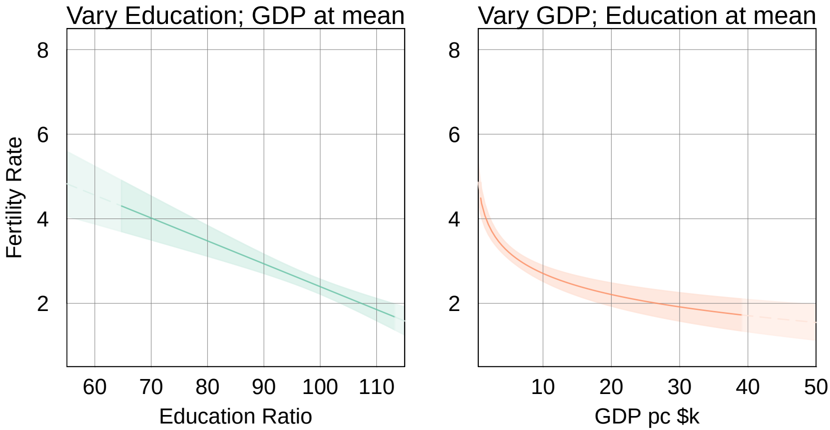
	Model			
Variable	1	2	3	4
Intercept	12.59	4.13	11.25	9.48
	(0.75)	(0.17)	(0.73)	(0.73)
Education Ratio	-0.10		-0.08	-0.05
	(0.01)		(0.01)	(0.01)
GDP per capita		-0.10	-0.05	
		(0.01)	(0.01)	
log(GDP per capita)				-0.72
				(0.09)
N	130	130	130	130
R^2	0.55	0.35	0.64	0.71
RMSE	1.12	1.35	1.01	0.91

Standard errors in parentheses



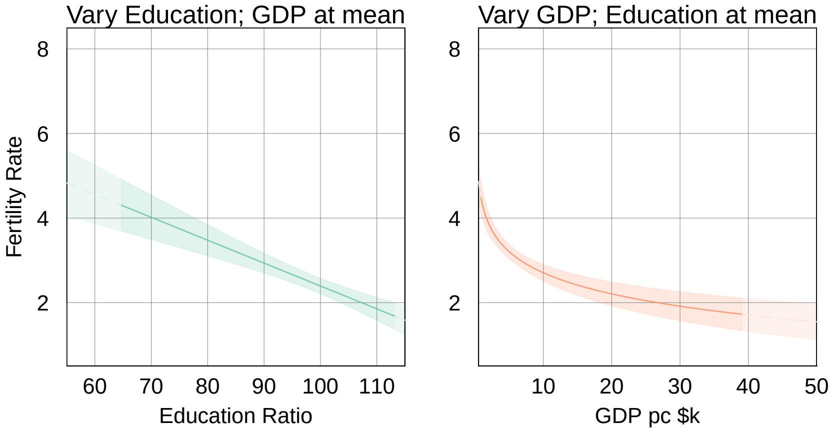
Logging GDP has made the residuals a bit smaller, and a bit more symmetrical


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

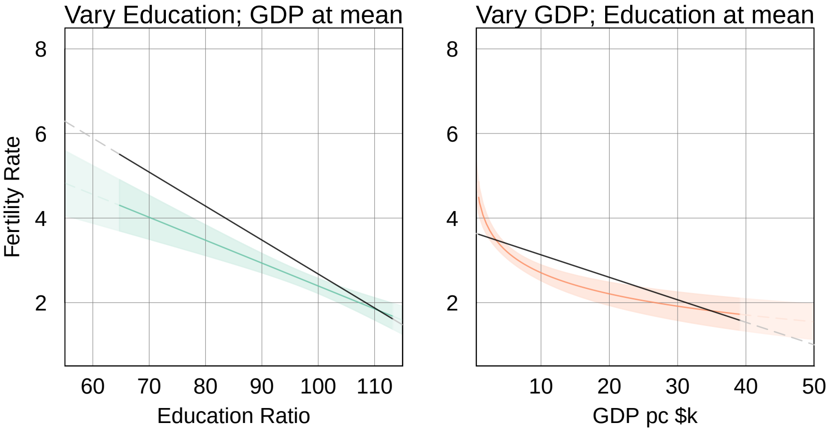


Logging GDP now allows small increases in GDP per capita is poor countries to dramatically lower fetility

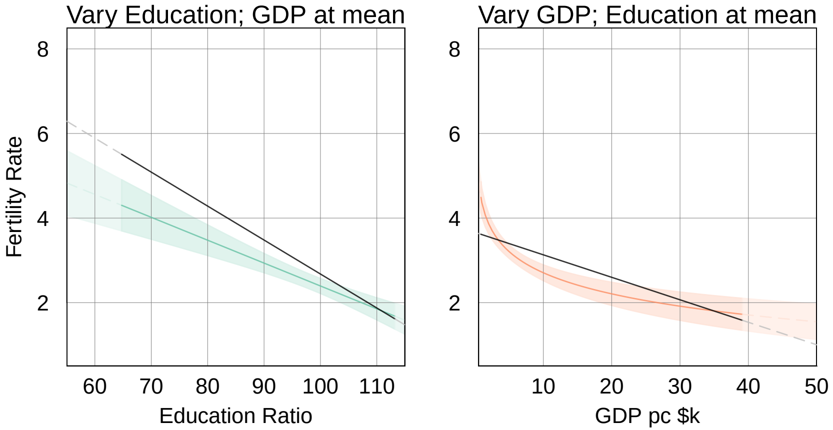
But small changes in GDP have very little effect in rich countries



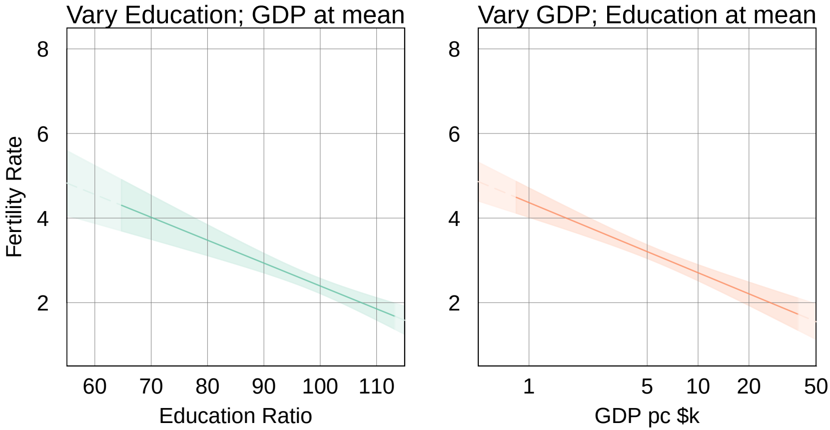
This is a pattern of diminishing marginal effects of economic development on fertility



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

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!

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

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

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*

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