

POLS 205
Political Science as a Social Science

Analyzing Multivariate Relationships
Part 2

Christopher Adolph

University of Washington, Seattle

May 26, 2010

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

Today we tackle these issues, and try to make linear regression into a flexible modeling tool

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

Regression with binary and categorical covariates

Regression with interaction terms

Regression with transformed variables

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

This is complicated by two factors:

- 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

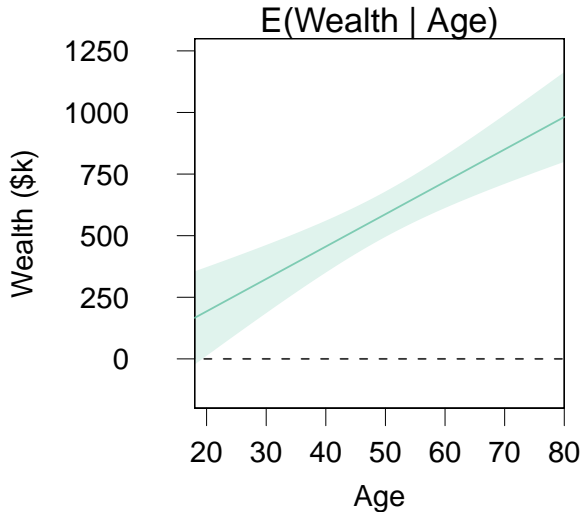
Before addressing these issues, let's consider a bivariate regression of Wealth on our only continuous covariate, Age

Regression of Wealth on Age

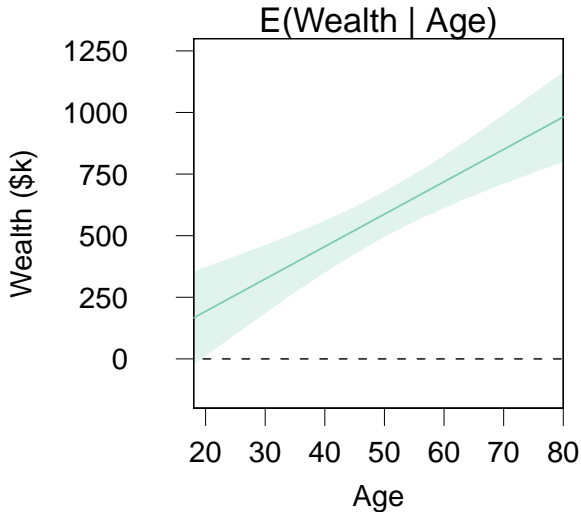
Variable	Estimate	se	<i>t</i> -stat	<i>p</i> -value
Age	13.2	2.6	5.04	< 0.001
Intercept	-70.1	137.4	-0.51	0.61
<i>N</i>	10000			
<i>R</i> ²	0.002			
RMSE	4484			

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 Education = 1 implies less than high school, Education = 2 implies high school, and so on?

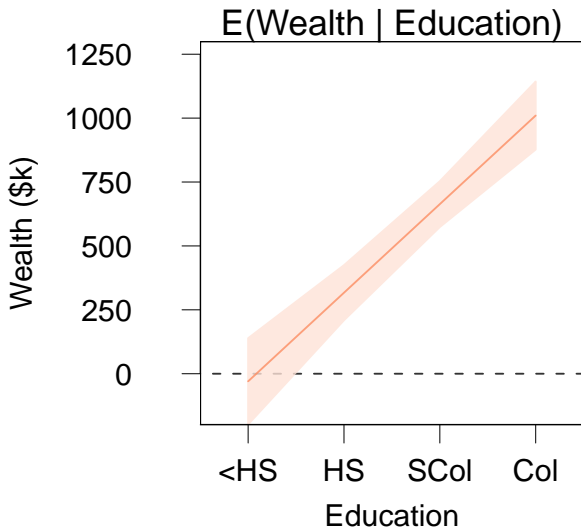
Including it in our regression would assume each step (from less than high school to high school, and from high school to some college, etc.) has the same effect on wealth

Is this reasonable?

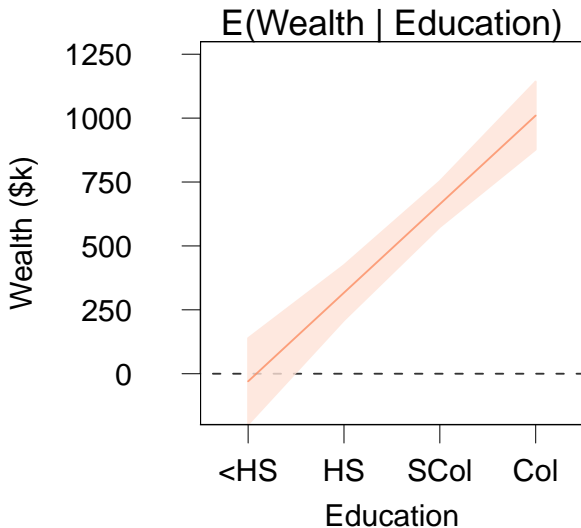
Regression of Wealth on Education (as an additive variable)

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

How 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

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

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

In words, a 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	<i>t</i> -stat	<i>p</i> -value
College	867.2	87.7	9.89	< 0.001
Intercept	271.0	51.9	5.22	< 0.001
<i>N</i>	10000			
<i>R</i> 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

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

The above table shows $E(\text{Wealth}|\text{College})$, 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:

E(Wealth Less than HS)	=	β_0		
E(Wealth HS)	=	β_0	$+\beta_1$	
E(Wealth Some College)	=	β_0		$+\beta_2$
E(Wealth College)	=	β_0		$+\beta_3$

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	<i>t</i> -stat	<i>p</i> -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
<i>N</i>	10000			
<i>R</i> ²	0.01			
RMSE	4185			

How do we interpret the above?

Expected wealth under different levels of education

Variable	Estimate	95% CI	
		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 and CIs for each education level

Does this remind you of anything?

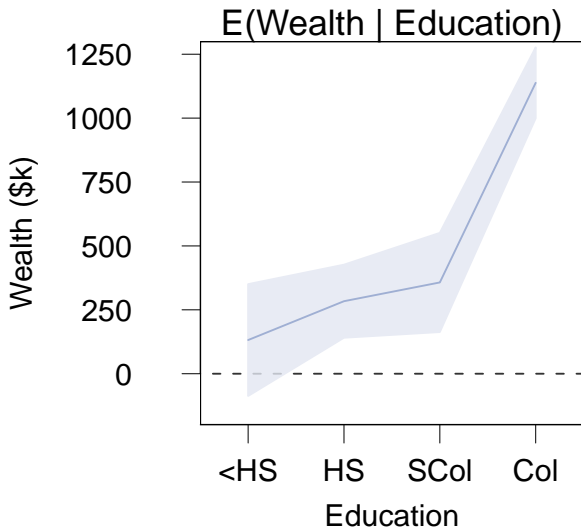
Expected wealth under different levels of education

Variable	Estimate	95% CI	
		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

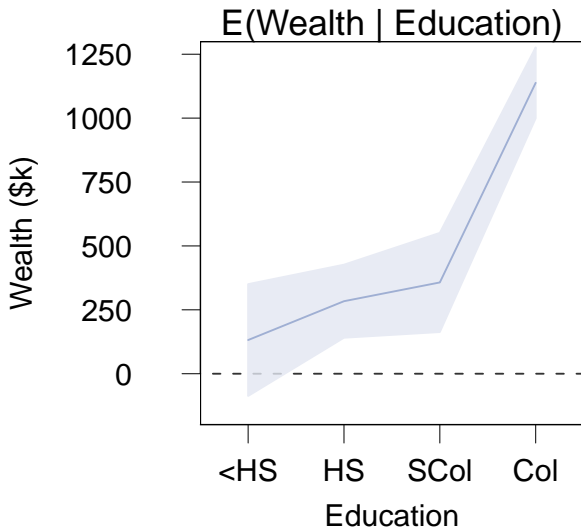
The above shows the fitted values and CIs for each education level

Does this remind you of anything? Comparison of means!

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

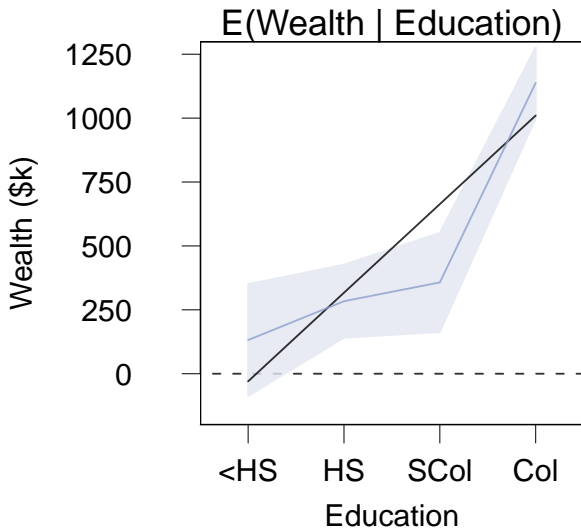


We can also present our results as a graphic

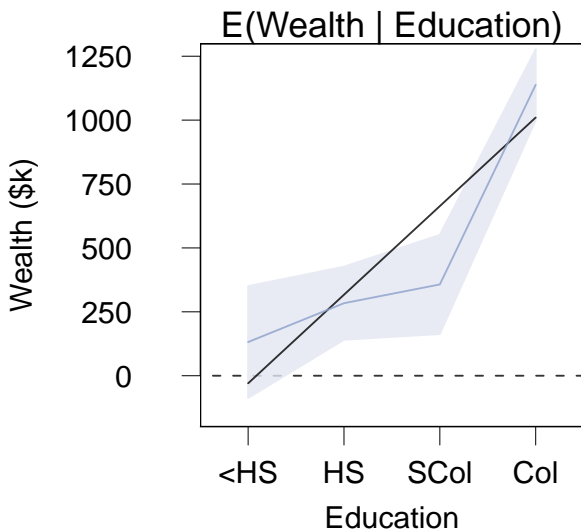


We can also present our results as a graphic

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
oversimpli-
cation

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

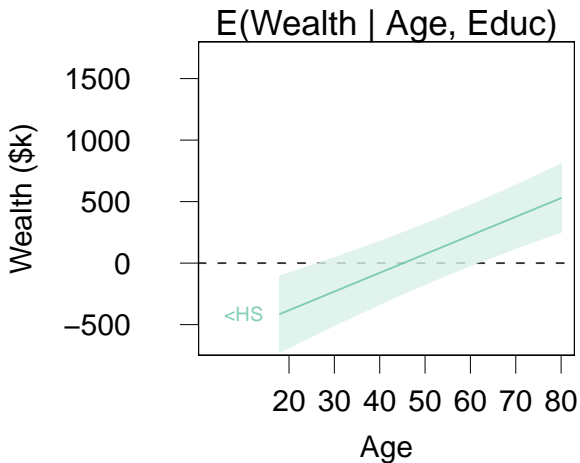
$E(\text{Wealth} \text{Less than HS})$	$=$	β_0			$+ \beta_4 \text{Age}$
$E(\text{Wealth} \text{HS})$	$=$	β_0	$+ \beta_1$		$+ \beta_4 \text{Age}$
$E(\text{Wealth} \text{Some College})$	$=$	β_0		$+ \beta_2$	$+ \beta_4 \text{Age}$
$E(\text{Wealth} \text{College})$	$=$	β_0			$+ \beta_3 + \beta_4 \text{Age}$

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

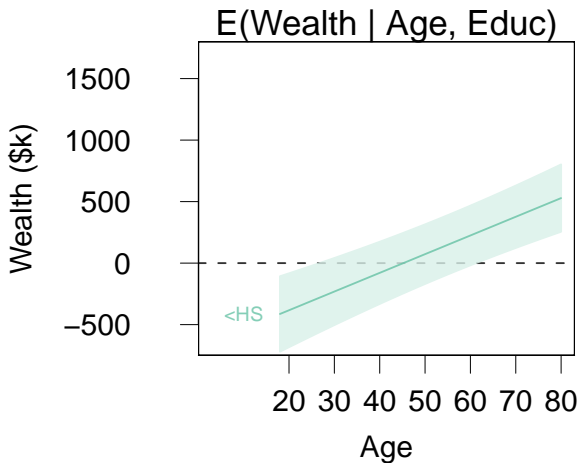
Regression of Wealth on Age and Education

Variable	Estimate	se	<i>t</i> -stat	<i>p</i> -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	2.74	0.006
College	1038.8	144.6	7.18	< 0.001
Intercept	-687.1	189.3	-3.63	< 0.001
<i>N</i>	10000			
<i>R</i> ²	0.01			
RMSE	4466			

How do we interpret the above table?

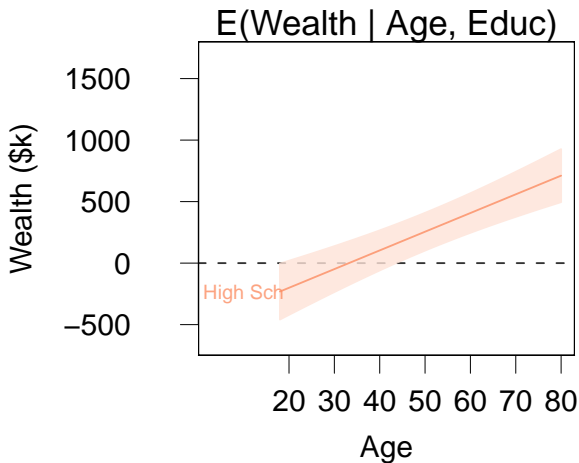


As models get more complicated, graphics become more helpful

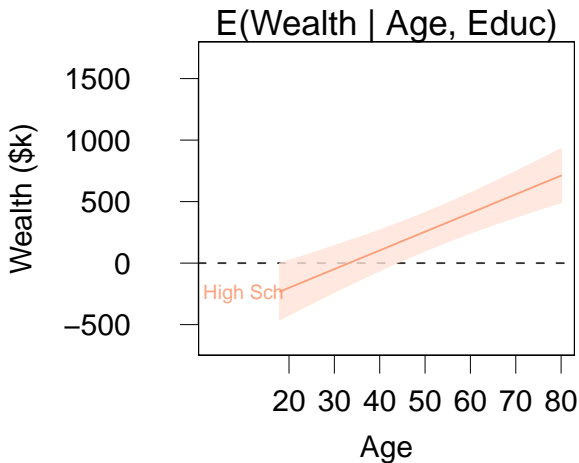


As models get more complicated, graphics become more helpful

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

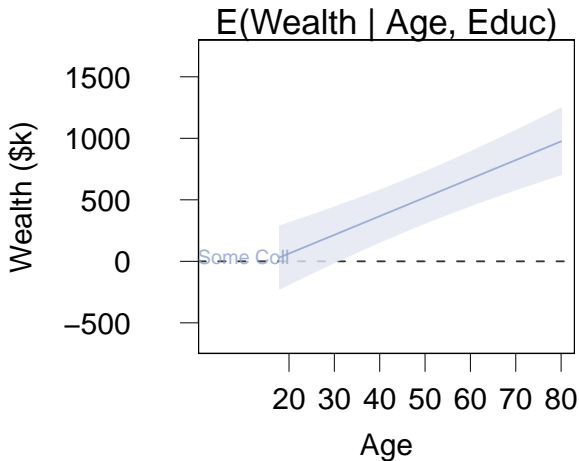


By changing the hypothetical level of education, we get a new picture

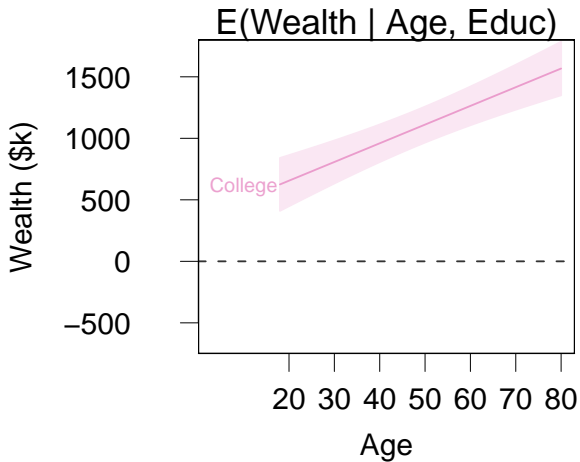


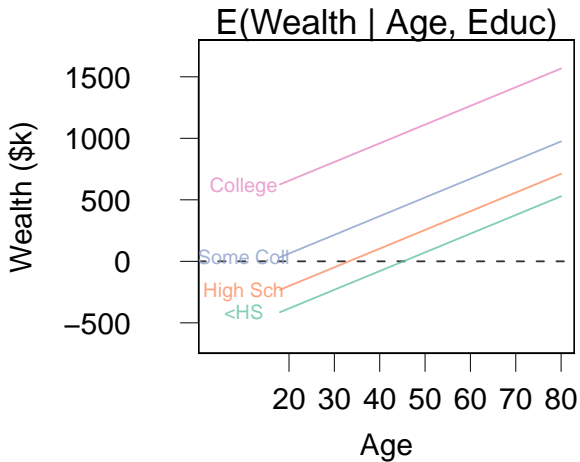
By changing the hypothetical level of education, we get a new picture

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

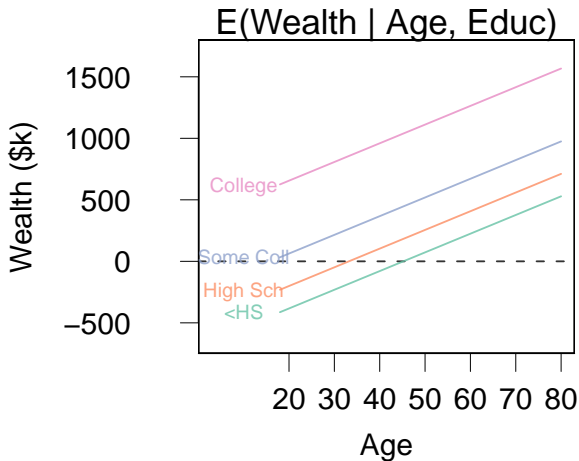


Comparing the pictures reveals that changing categories only changes the intercept, not the slope



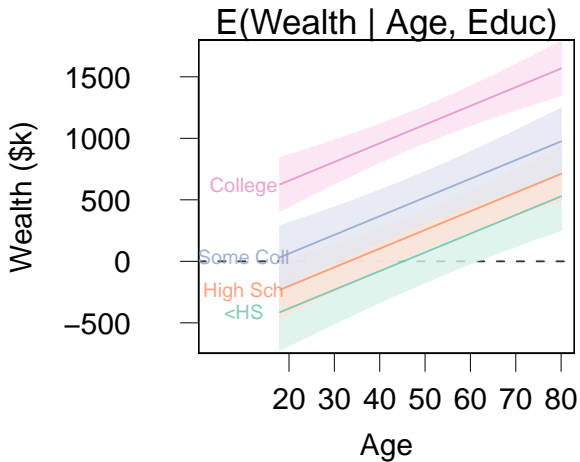


We can collect the whole model on a single slide

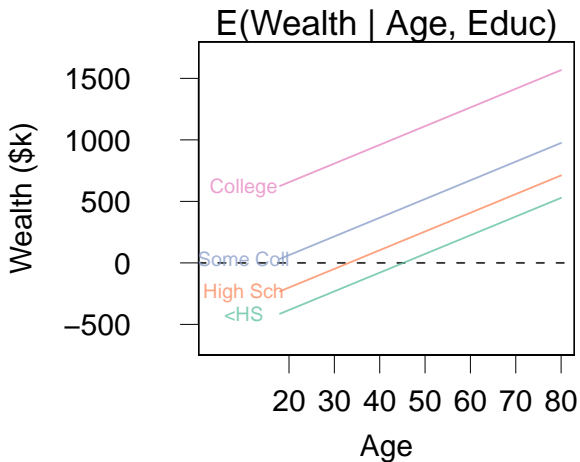


We can collect the whole model on a single slide

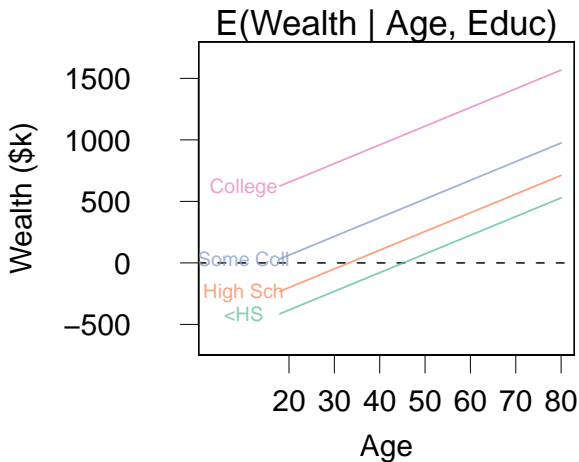
How do we interpret this picture?



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

$$\begin{aligned} \text{Wealth}_i = & \beta_0 + \beta_1 \text{HS}_i + \beta_2 \text{SomeCol}_i + \beta_3 \text{College}_i + \beta_5 \text{Age} \\ & \beta_6 \text{Age} \times \text{HS}_i + \beta_7 \text{Age} \times \text{SomeCol}_i + \beta_8 \text{Age} \times \text{College}_i \\ & + \varepsilon_i \end{aligned}$$

What does the above imply for different levels of education?

$E(\text{Wealth} \text{Less than HS})$	$=$	β_0		$+ \beta_4 \text{Age}$
$E(\text{Wealth} \text{HS})$	$=$	β_0	$+ \beta_1$	$+ (\beta_4 + \beta_6) \text{Age}$
$E(\text{Wealth} \text{Some College})$	$=$	β_0	$+ \beta_2$	$+ (\beta_4 + \beta_7) \text{Age}$
$E(\text{Wealth} \text{College})$	$=$	β_0	$+ \beta_3$	$+ (\beta_4 + \beta_8) \text{Age}$

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	39.8	396.7	0.1	0.920
Some College	-150.7	436.9	-0.35	0.730
College	-854.6	408.4	-2.09	0.036
Age × High School	2.53	6.96	0.36	0.717
Age × Some College	9	8.12	1.11	0.268
Age × College	38.76	7.36	5.27	< 0.001
Intercept	-103.4	329.9	-0.31	0.754
<i>N</i>	10000			
<i>R</i> ²	0.02			
RMSE	4167			

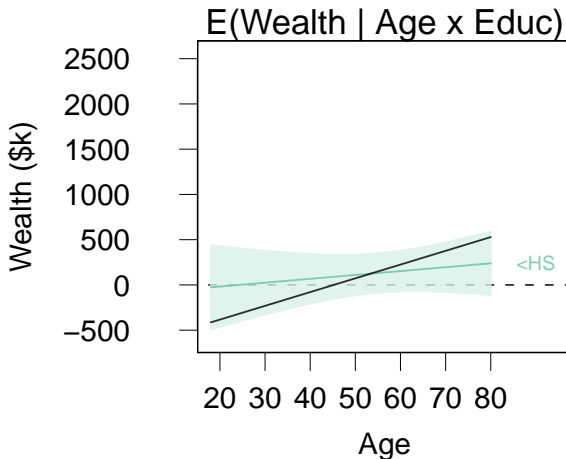
How do we interpret the above?

Regression of Wealth on Age and Education (with interactions)

Variable	Estimate	se	<i>t</i> -stat	<i>p</i> -value
Age	4.27	5.64	0.76	0.449
High School	39.8	396.7	0.1	0.920
Some College	-150.7	436.9	-0.35	0.730
College	-854.6	408.4	-2.09	0.036
Age × High School	2.53	6.96	0.36	0.717
Age × Some College	9	8.12	1.11	0.268
Age × College	38.76	7.36	5.27	< 0.001
Intercept	-103.4	329.9	-0.31	0.754
<i>N</i>	10000			
<i>R</i> ²	0.02			
RMSE	4167			

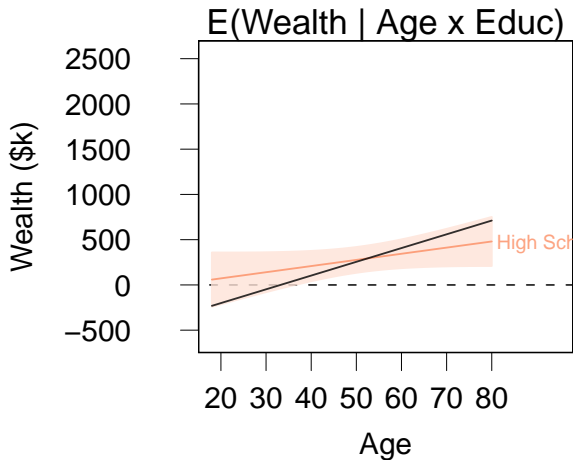
How do we interpret the above? Hard to do!

Interaction terms best interpreted graphically!

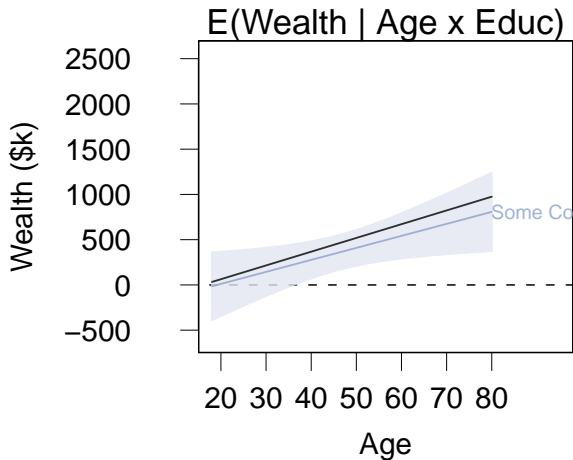


The education-conditional slope is in color

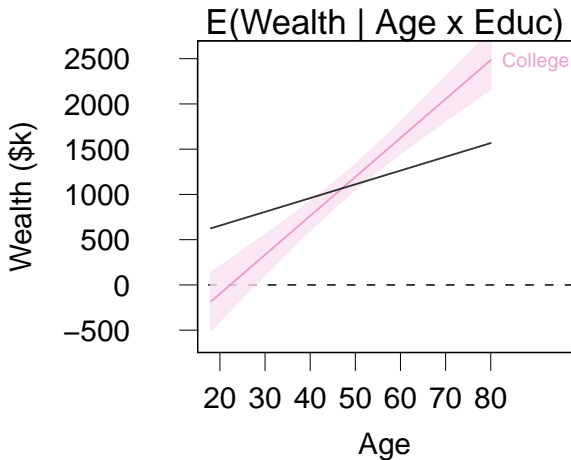
The uninteracted slope is in black for comparison



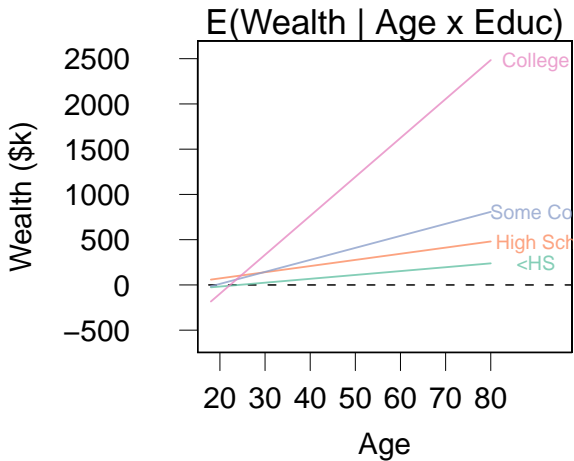
What happens to the Age slope as we increase Education?



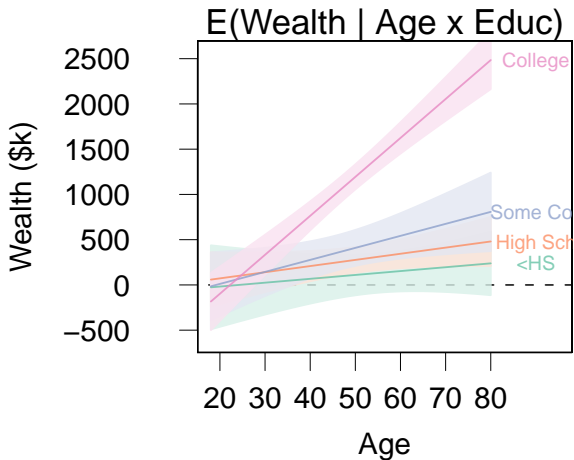
Can we summarize the substance of this effect in words?



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



All
conditional
slopes

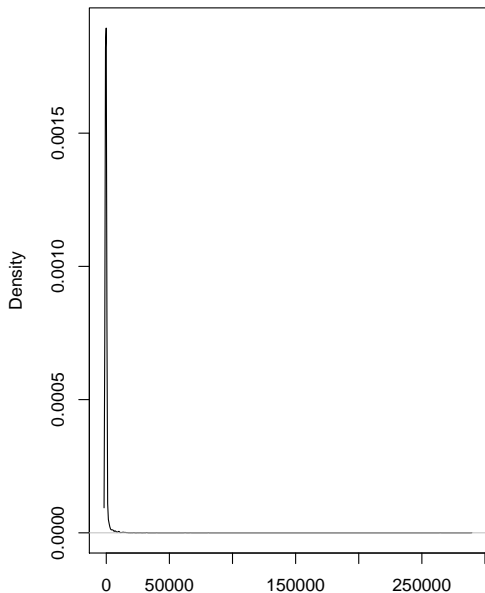


With 95%
CIs

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

Residuals from Wealth = f(Age, Education)

Tremendous
right skew!

Transforming variables

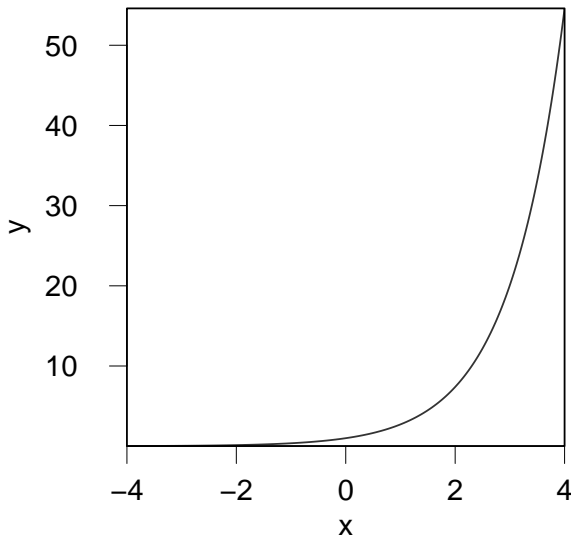
Wealth is clearly not a linear function of Education and Age

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 *transformed* version of Wealth is

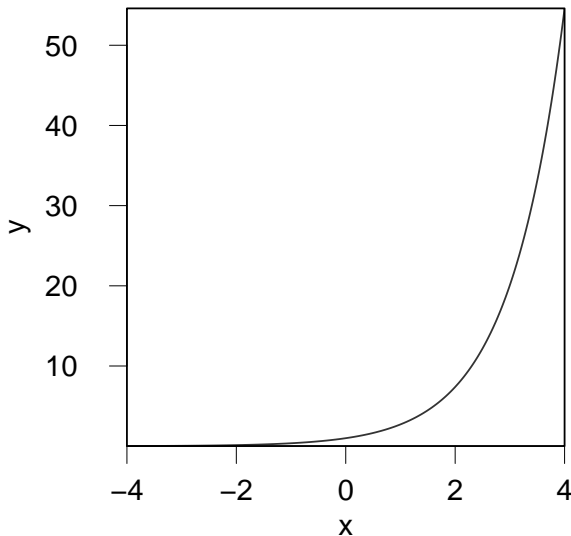
For example, perhaps $\log(\text{Wealth})$ is a linear function of Age and Education

$$\log(y) = x$$



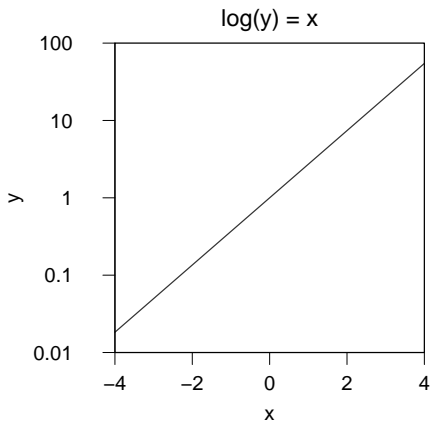
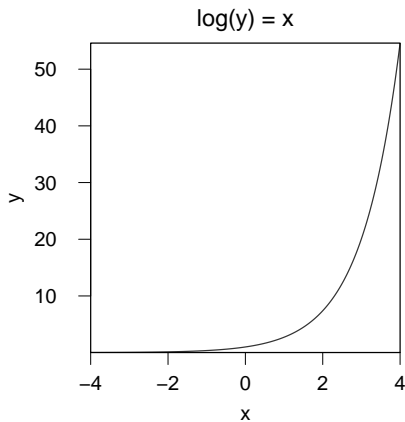
At left is a
log-linear
relationship

$$\log(y) = x$$



At left is a
log-linear
relationship

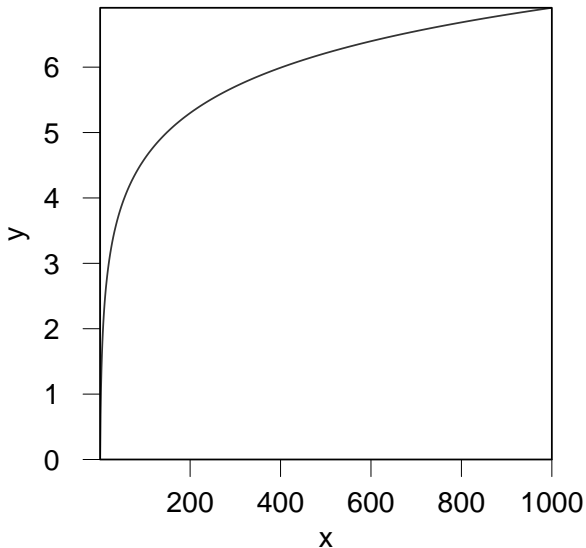
$\log(y)$ is a
function of
 x



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

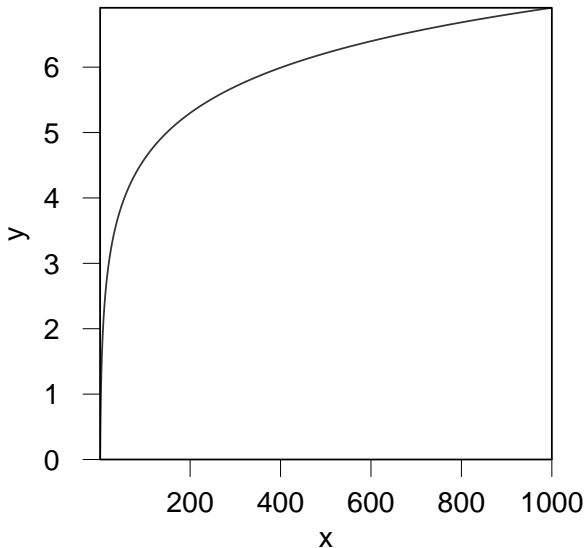
We can think of this as a case where *a level change in x causes a percentage change in y*

$$y = \log(x)$$



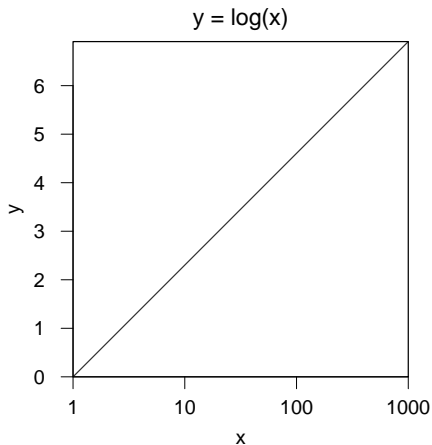
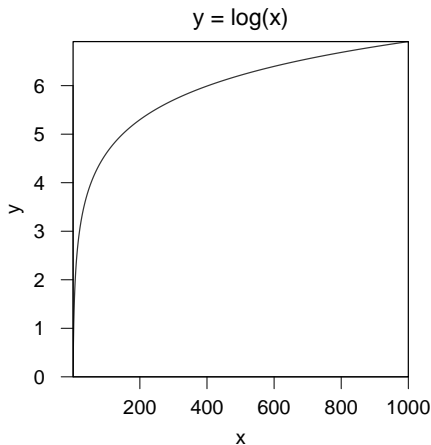
Another possibility is that the level of y is a function of the $\log(x)$

$$y = \log(x)$$



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

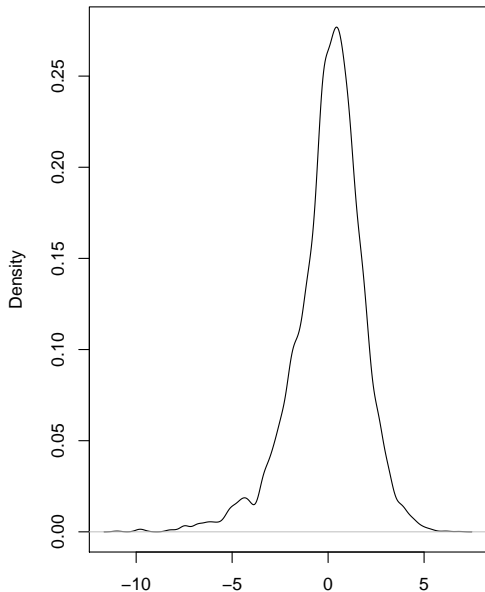
We can think of this as a case where *a percentage change in x causes a level change in y*

Regression of $\log(\text{Wealth})$ on Age and Education

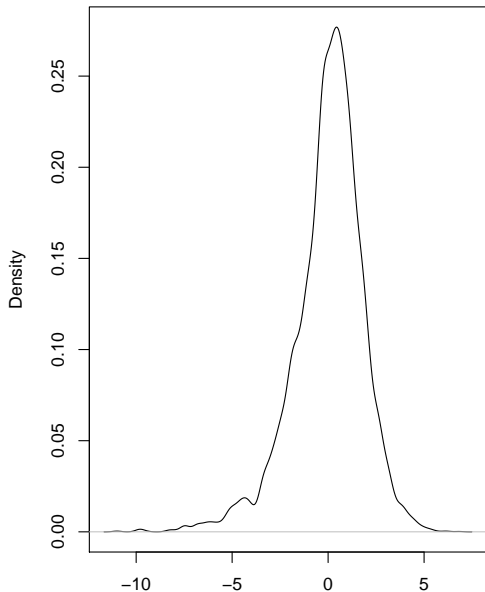
Variable	Estimate	se	<i>t</i> -stat	<i>p</i> -value
Age	0.045	0.001	38.49	< 0.001
High School	1.016	0.063	16.05	< 0.001
Some College	1.424	0.072	19.84	< 0.001
College	2.552	0.063	40.55	< 0.001
Intercept	0.890	0.086	10.53	< 0.001
<i>N</i>	9024			
<i>R</i> ²	0.26			
RMSE	1.84			

Logging the dependent variable has drastically improved the fit

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

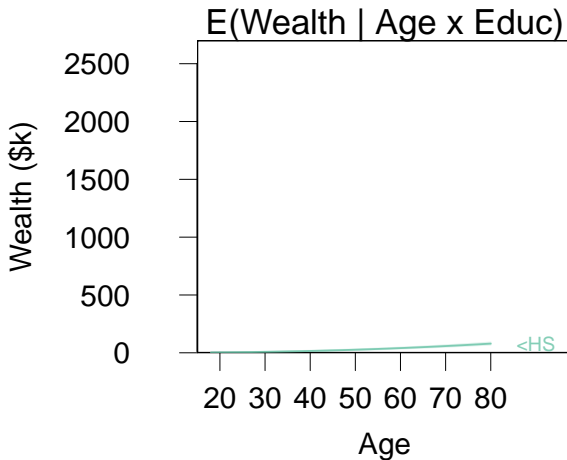
Residuals from $\log(\text{Wealth}) = f(\text{Age}, \text{Education})$ 

Now the
residuals
look approx
Normal

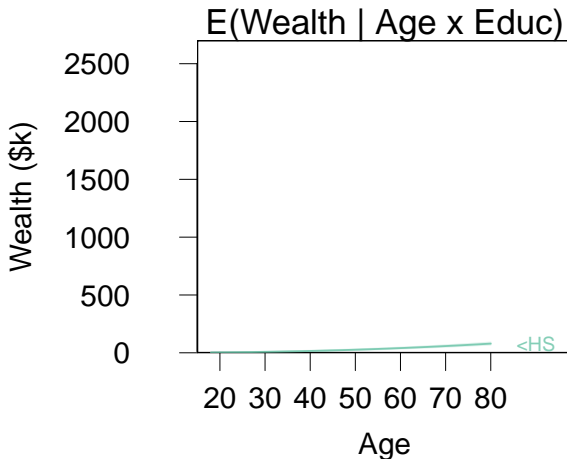
Residuals from $\log(\text{Wealth}) = f(\text{Age}, \text{Education})$ 

Now the residuals look approx Normal

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

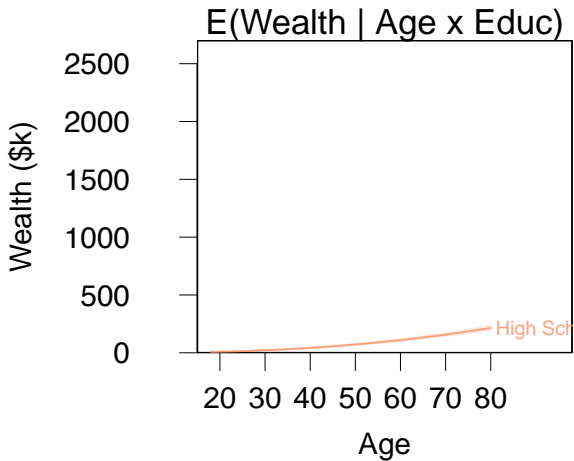


We'll plot
the fitted
wealth
given Age
and
Education
as usual

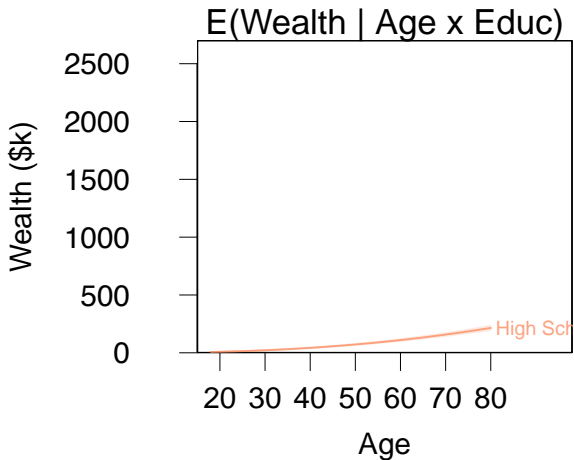


We'll plot the fitted wealth given Age and Education as usual

Note the linear scales

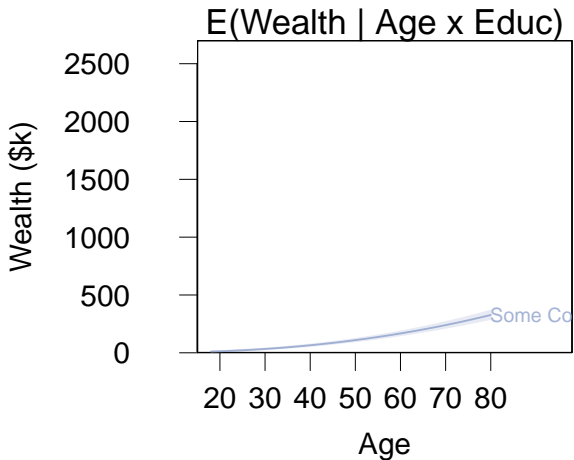


The log specification has added curvature

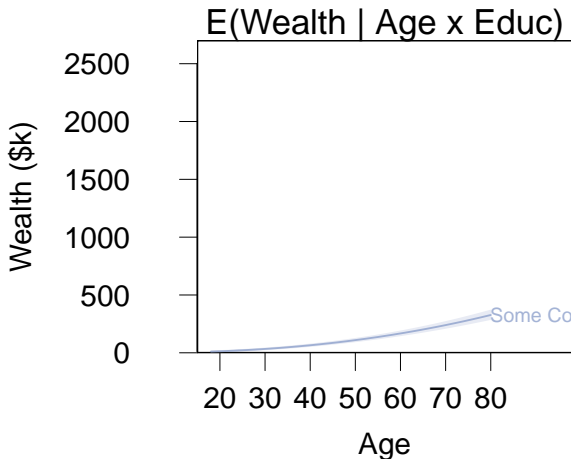


The log specification has added curvature

The slope is no longer constant

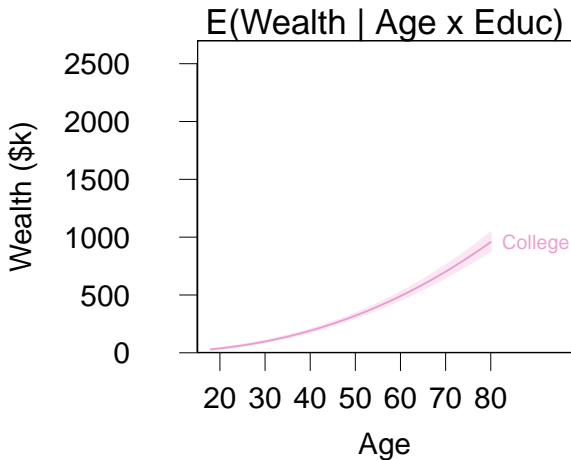


In fact, the slope gets bigger as Age rises

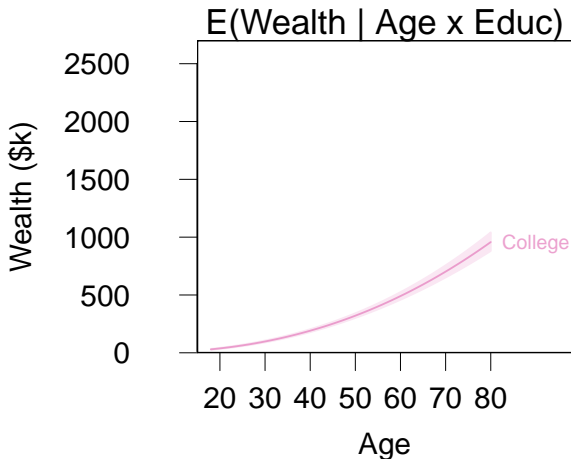


In fact, the slope gets bigger as Age rises

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

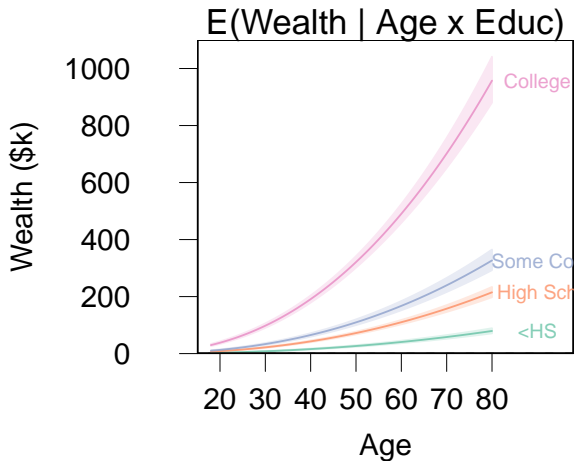


Note that the log transformation is capturing something similar to our interactions

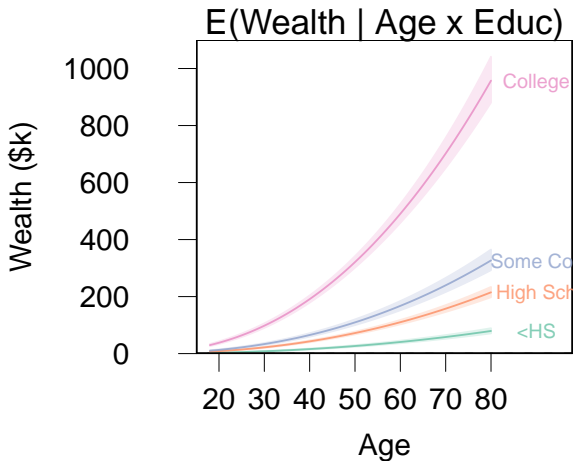


Note that the log transformation is capturing something similar to our interactions

Steeper slopes for higher education

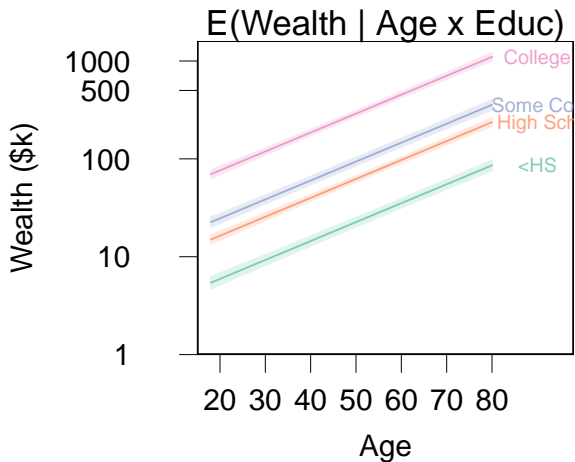


We now see the reason for the differing slopes

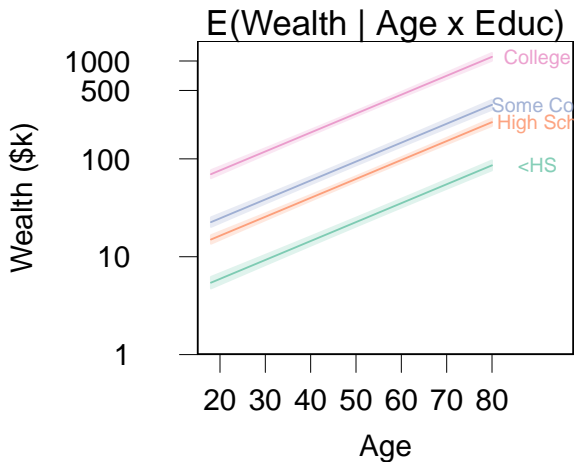


We now see the reason for the differing slopes

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

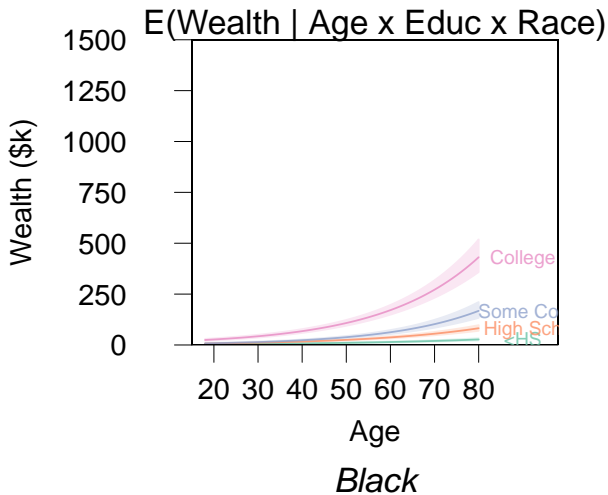
We “found” that Blacks and Hispanics were at a major, and essentially equal, disadvantage in wealth

Will that finding hold up controlling for Age and Education?

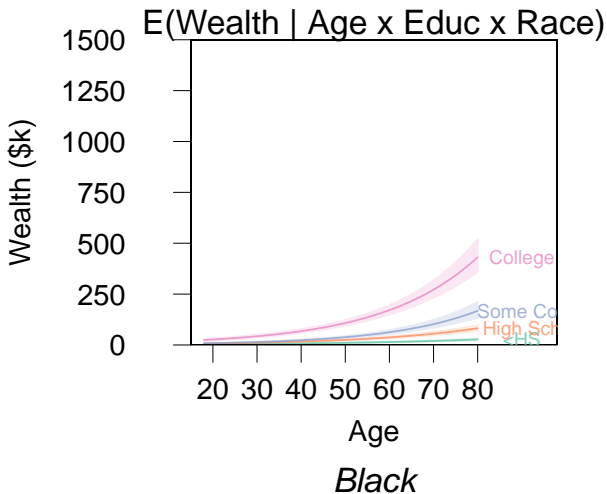
Let's add dummies for Race into a model incorporating everything else we've explored

Regression of $\log(\text{Wealth})$ on Age, Education, and Race

Variable	Estimate	se	<i>t</i> -stat	<i>p</i> -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
<i>N</i>	9024			
<i>R</i> ²	0.29			
RMSE	1.81			

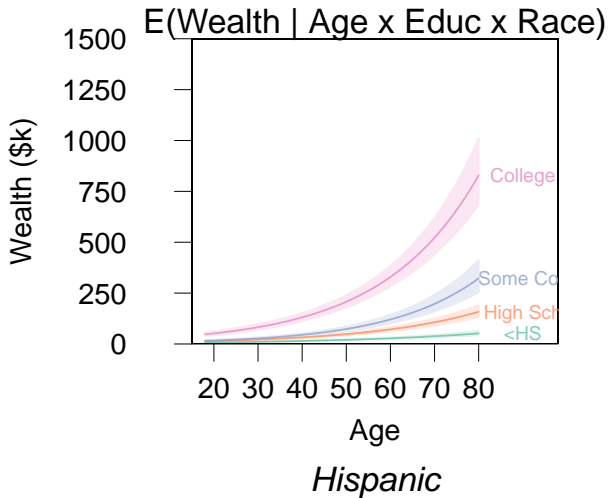


Graphics again help clarify a complex model

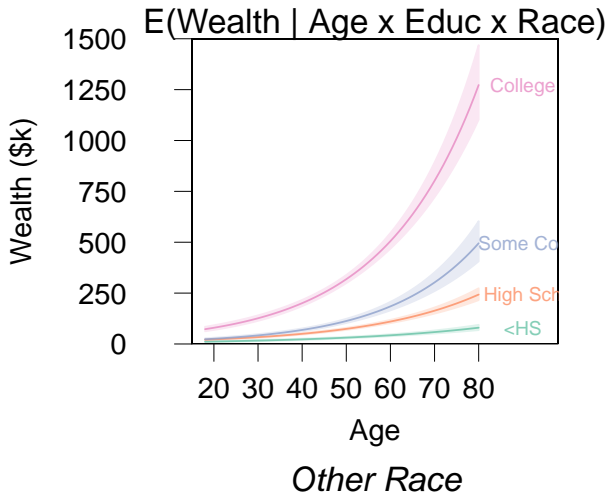


Graphics again help clarify a complex model

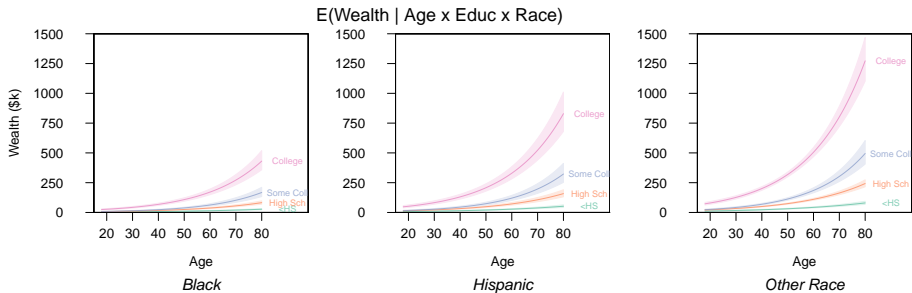
At left is the expected wealth for Black households, at different levels of Education and Age



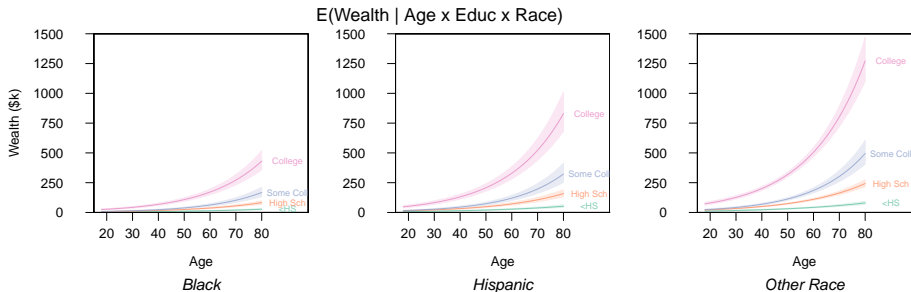
We can compare to Hispanics



And all
other
households
(mostly
Whites)

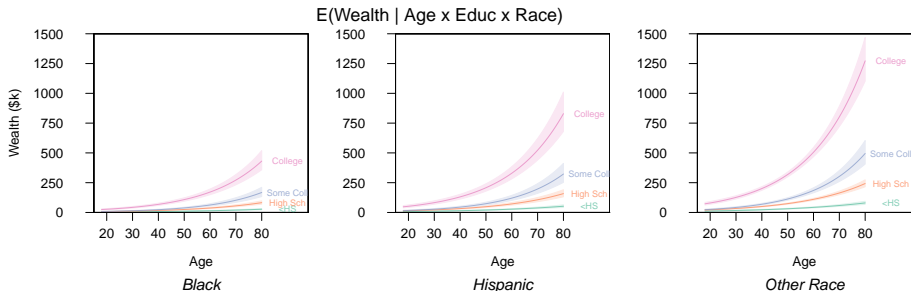


What do we make of the above results?



What do we make of the above results?

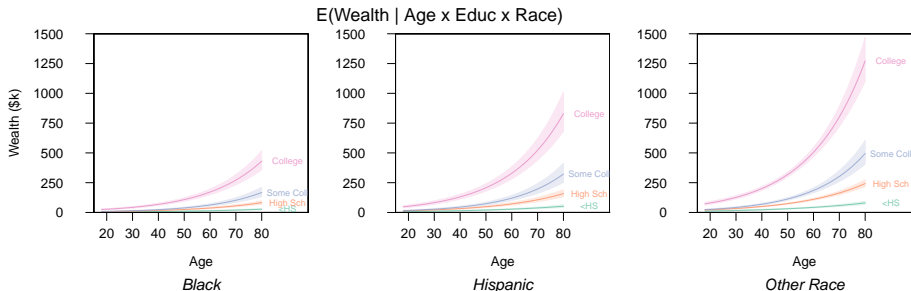
How much would we expect an individual household to deviate from this prediction?



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

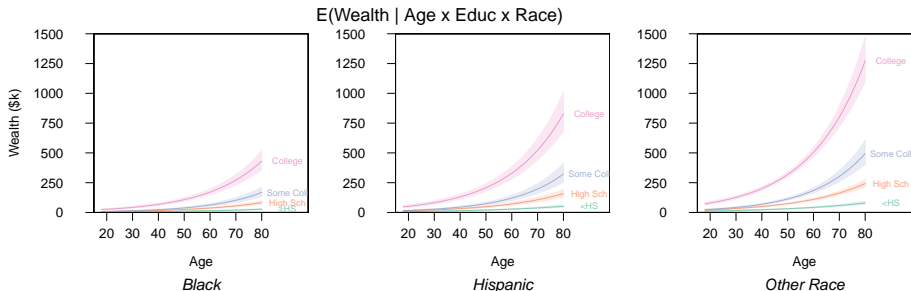
$$\log(500) \pm 1.8 = 6.2 \pm 1.8 \text{ logged dollars}$$



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

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