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
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

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>se</th>
<th>t-stat</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>13.2</td>
<td>2.6</td>
<td>5.04</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Intercept</td>
<td>-70.1</td>
<td>137.4</td>
<td>-0.51</td>
<td>0.61</td>
</tr>
<tr>
<td>$N$</td>
<td>10000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.002</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMSE</td>
<td>4484</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

How do we interpret the above?

Is this a good model?
How do we interpret this graph?

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

\[ \text{Education} = 1 \text{ implies less than high school,} \]
\[ \text{Education} = 2 \text{ implies high school, and so on?} \]
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:
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)**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>se</th>
<th>t-stat</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Education</td>
<td>355.9</td>
<td>113.8</td>
<td>9.21</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Intercept</td>
<td>-400</td>
<td>113.8</td>
<td>-3.51</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

- $N = 10000$
- $R^2 = 0.01$
- 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
$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$

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 \( \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 \).
\[ y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \]

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 \( \beta \) coefficient when present

Let’s use binary covariates to solve our problem with Education
### Regression of Wealth on College (as a binary variable)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>se</th>
<th>t-stat</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>College</td>
<td>867.2</td>
<td>87.7</td>
<td>9.89</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Intercept</td>
<td>271.0</td>
<td>51.9</td>
<td>5.22</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

| N          | 10000    |
| R²         | 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

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than College</td>
<td>271</td>
<td>169.2</td>
</tr>
<tr>
<td>College</td>
<td>1138.1</td>
<td>999.6</td>
</tr>
</tbody>
</table>

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:

<table>
<thead>
<tr>
<th></th>
<th>High School</th>
<th>Some College</th>
<th>College</th>
</tr>
</thead>
<tbody>
<tr>
<td>Someone with &lt; high school</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Someone with high school</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Someone with some college</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Someone with college</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

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:

<table>
<thead>
<tr>
<th></th>
<th>High School</th>
<th>Some College</th>
<th>College</th>
</tr>
</thead>
<tbody>
<tr>
<td>Someone with &lt; high school</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Someone with high school</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Someone with some college</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Someone with college</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

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:

\[ \text{E(Wealth|Less than HS)} = \]
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:

\[ \begin{align*}
\text{E}(\text{Wealth} | \text{Less than HS}) &= \beta_0 \\
\text{E}(\text{Wealth} | \text{HS}) &= \\
\end{align*} \]
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:

\[
\begin{align*}
E(\text{Wealth} | \text{Less than HS}) &= \beta_0 \\
E(\text{Wealth} | \text{HS}) &= \beta_0 + \beta_1 \\
E(\text{Wealth} | \text{Some College}) &= \\
\end{align*}
\]
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:

\[
\begin{align*}
\text{E(Wealth|Less than HS)} &= \beta_0 \\
\text{E(Wealth|HS)} &= \beta_0 + \beta_1 \\
\text{E(Wealth|Some College)} &= \beta_0 + \beta_2 \\
\text{E(Wealth|College)} &= \\
\end{align*}
\]
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:

\[
\begin{align*}
E(\text{Wealth} | \text{Less than HS}) & = \beta_0 \\
E(\text{Wealth} | \text{HS}) & = \beta_0 + \beta_1 \\
E(\text{Wealth} | \text{Some College}) & = \beta_0 + \beta_2 \\
E(\text{Wealth} | \text{College}) & = \beta_0 + \beta_3
\end{align*}
\]

Each possible case is *uniquely* defined and allowed to have its own effect on wealth.
Regression of Wealth on Education (as dummy variables)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>se</th>
<th>t-stat</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>High School</td>
<td>151.9</td>
<td>133.3</td>
<td>1.14</td>
<td>0.25</td>
</tr>
<tr>
<td>Some College</td>
<td>225.3</td>
<td>149.1</td>
<td>1.51</td>
<td>0.13</td>
</tr>
<tr>
<td>College</td>
<td>1006.4</td>
<td>132.1</td>
<td>7.62</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Intercept</td>
<td>131.7</td>
<td>111.6</td>
<td>1.18</td>
<td>0.24</td>
</tr>
</tbody>
</table>

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td>10000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMSE</td>
<td>4185</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

How do we interpret the above?
## Expected wealth under different levels of education

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Estimate</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than HS</td>
<td>131.7</td>
<td>-86.9</td>
<td>350.4</td>
</tr>
<tr>
<td>High School</td>
<td>283.7</td>
<td>140.8</td>
<td>551.1</td>
</tr>
<tr>
<td>Some College</td>
<td>357.1</td>
<td>163.1</td>
<td>551.1</td>
</tr>
<tr>
<td>College</td>
<td>1138.1</td>
<td>999.6</td>
<td>1276.7</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Estimate</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than HS</td>
<td>131.7</td>
<td>-86.9</td>
<td>350.4</td>
</tr>
<tr>
<td>High School</td>
<td>283.7</td>
<td>140.8</td>
<td>551.1</td>
</tr>
<tr>
<td>Some College</td>
<td>357.1</td>
<td>163.1</td>
<td>551.1</td>
</tr>
<tr>
<td>College</td>
<td>1138.1</td>
<td>999.6</td>
<td>1276.7</td>
</tr>
</tbody>
</table>

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 oversimplication.
Regression with continuous and dummy covariates

\[ \text{Wealth}_i = \beta_0 + \beta_1 \text{HS}_i + \beta_2 \text{SomeCol}_i + \beta_3 \text{College}_i + \beta_4 \text{Age} + \epsilon_i \]

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

\[
\begin{align*}
E(\text{Wealth} | \text{Less than HS}) &= \beta_0 + \beta_4 \text{Age} \\
E(\text{Wealth} | \text{HS}) &= \beta_0 + \beta_1 + \beta_4 \text{Age} \\
E(\text{Wealth} | \text{Some College}) &= \beta_0 + \beta_2 + \beta_4 \text{Age} \\
E(\text{Wealth} | \text{College}) &= \beta_0 + \beta_3 + \beta_4 \text{Age}
\end{align*}
\]

The terms \( \beta_0 \) through \( \beta_3 \) only shift the intercept of the regression line whose slope is \( \beta_4 \).
# Regression of Wealth on Age and Education

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>se</th>
<th>t-stat</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>15.2</td>
<td>2.63</td>
<td>5.79</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>High School</td>
<td>182.5</td>
<td>146.1</td>
<td>1.25</td>
<td>0.212</td>
</tr>
<tr>
<td>Some College</td>
<td>446.4</td>
<td>163.0</td>
<td>2.74</td>
<td>0.006</td>
</tr>
<tr>
<td>College</td>
<td>1038.8</td>
<td>144.6</td>
<td>7.18</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Intercept</td>
<td>-687.1</td>
<td>189.3</td>
<td>-3.63</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

- \( N = 10000 \)
- \( R^2 = 0.01 \)
- \( \text{RMSE} = 4466 \)

How do we interpret the above table?
As models get more complicated, graphics become more helpful.

\[ E(\text{Wealth} \mid \text{Age, Educ}) \]

Wealth ($\text{k}$)

\[ -500 \quad 0 \quad 500 \quad 1000 \quad 1500 \]

Age

\[ 20 \quad 30 \quad 40 \quad 50 \quad 60 \quad 70 \quad 80 \]

<HS
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)
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?

Chris Adolph (University of Washington)
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

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

What does the above imply for different levels of education?
Interaction terms

\[ \text{Wealth}_i = \beta_0 + \beta_1 \text{HS}_i + \beta_2 \text{SomeCol}_i + \beta_3 \text{College}_i + \beta_4 \text{Age} + \beta_5 \text{Age} \times \text{HS}_i \]
Interaction terms

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

What does the above imply for different levels of education?
Interaction terms

\[ \text{Wealth}_i = \beta_0 + \beta_1 \text{HS}_i + \beta_2 \text{SomeCol}_i + \beta_3 \text{College}_i + \beta_4 \text{Age} + \beta_5 \text{Age} \times \text{HS}_i + \beta_6 \text{Age} \times \text{SomeCol}_i + \beta_7 \text{Age} \times \text{College}_i + \varepsilon_i \]

What does the above imply for different levels of education?
Interaction terms

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

What does the above imply for different levels of education?

$$E(\text{Wealth}|\text{Less than HS}) = \beta_0 + \beta_4 \text{Age}$$
Wealth\(_i\) = \(\beta_0 + \beta_1 HS_i + \beta_2 SomeCol_i + \beta_3 College_i + \beta_4 Age + \beta_5 Age \times HS_i + \beta_6 Age \times SomeCol_i + \beta_7 Age \times College_i + \varepsilon_i\)

What does the above imply for different levels of education?

\[
\begin{align*}
E(\text{Wealth}|\text{Less than HS}) &= \beta_0 + \beta_4 \text{Age} \\
E(\text{Wealth}|\text{HS}) &= \beta_0 + \beta_1 + (\beta_4 + \beta_5) \text{Age}
\end{align*}
\]
Interaction terms

\[ \text{Wealth}_i = \beta_0 + \beta_1 \text{HS}_i + \beta_2 \text{SomeCol}_i + \beta_3 \text{College}_i + \beta_4 \text{Age} + \beta_5 \text{Age} \times \text{HS}_i + \beta_6 \text{Age} \times \text{SomeCol}_i + \beta_7 \text{Age} \times \text{College}_i + \epsilon_i \]

What does the above imply for different levels of education?

\[
\begin{align*}
E(\text{Wealth} | \text{Less than HS}) &= \beta_0 + \beta_4 \text{Age} \\
E(\text{Wealth} | \text{HS}) &= \beta_0 + \beta_1 + (\beta_4 + \beta_5) \text{Age} \\
E(\text{Wealth} | \text{Some College}) &= \beta_0 + \beta_2 + (\beta_4 + \beta_6) \text{Age}
\end{align*}
\]
Interaction terms

\[
\text{Wealth}_i = \beta_0 + \beta_1 HS_i + \beta_2 \text{SomeCol}_i + \beta_3 \text{College}_i + \beta_4 \text{Age} + \beta_5 \text{Age} \times HS_i + \beta_6 \text{Age} \times \text{SomeCol}_i + \beta_7 \text{Age} \times \text{College}_i + \varepsilon_i
\]

What does the above imply for different levels of education?

\[
\begin{align*}
E(\text{Wealth} \mid \text{Less than HS}) &= \beta_0 + \beta_4 \text{Age} \\
E(\text{Wealth} \mid \text{HS}) &= \beta_0 + \beta_1 + (\beta_4 + \beta_5) \text{Age} \\
E(\text{Wealth} \mid \text{Some College}) &= \beta_0 + \beta_2 + (\beta_4 + \beta_6) \text{Age} \\
E(\text{Wealth} \mid \text{College}) &= \beta_0 + \beta_3 + (\beta_4 + \beta_7) \text{Age}
\end{align*}
\]

A different slope and intercept for each educational category
### Regression of Wealth on Age and Education (with interactions)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>se</th>
<th>t-stat</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>4.27</td>
<td>5.64</td>
<td>0.76</td>
<td>0.449</td>
</tr>
<tr>
<td>High School</td>
<td>39.80</td>
<td>396.70</td>
<td>0.10</td>
<td>0.920</td>
</tr>
<tr>
<td>Some College</td>
<td>-150.70</td>
<td>436.90</td>
<td>-0.35</td>
<td>0.730</td>
</tr>
<tr>
<td>College</td>
<td>-854.60</td>
<td>408.40</td>
<td>-2.09</td>
<td>0.036</td>
</tr>
<tr>
<td>Age × High School</td>
<td>2.53</td>
<td>6.96</td>
<td>0.36</td>
<td>0.717</td>
</tr>
<tr>
<td>Age × Some College</td>
<td>9.00</td>
<td>8.12</td>
<td>1.11</td>
<td>0.268</td>
</tr>
<tr>
<td>Age × College</td>
<td>38.76</td>
<td>7.36</td>
<td>5.27</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Intercept</td>
<td>-103.40</td>
<td>329.90</td>
<td>-0.31</td>
<td>0.754</td>
</tr>
</tbody>
</table>

| N                         | 10000    |
| R²                        | 0.02     |
| RMSE                      | 4167     |

How do we interpret the above?
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
Multiple Regression, part 2

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!

N = 10000   Bandwidth = 70.39

Density
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
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 $y$*
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

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>se</th>
<th>t-stat</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>0.045</td>
<td>0.001</td>
<td>38.49</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>High School</td>
<td>1.016</td>
<td>0.063</td>
<td>16.05</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Some College</td>
<td>1.424</td>
<td>0.072</td>
<td>19.84</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>College</td>
<td>2.552</td>
<td>0.063</td>
<td>40.55</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Intercept</td>
<td>0.890</td>
<td>0.086</td>
<td>10.53</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

| N             | 9024     |
| R²            | 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.
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(Wealth) on Age, Education, and Race

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>se</th>
<th>t-stat</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>0.031</td>
<td>0.003</td>
<td>11.58</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>High School</td>
<td>0.464</td>
<td>0.192</td>
<td>2.42</td>
<td>0.016</td>
</tr>
<tr>
<td>Some College</td>
<td>0.396</td>
<td>0.215</td>
<td>1.84</td>
<td>0.066</td>
</tr>
<tr>
<td>College</td>
<td>1.588</td>
<td>0.201</td>
<td>7.89</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Age × High School</td>
<td>0.008</td>
<td>0.003</td>
<td>2.49</td>
<td>0.013</td>
</tr>
<tr>
<td>Age × Some College</td>
<td>0.018</td>
<td>0.004</td>
<td>4.61</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Age × College</td>
<td>0.015</td>
<td>0.003</td>
<td>4.23</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Black</td>
<td>-1.082</td>
<td>0.063</td>
<td>-16.97</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Hispanic</td>
<td>-0.428</td>
<td>0.072</td>
<td>-5.95</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Intercept</td>
<td>1.866</td>
<td>0.168</td>
<td>11.11</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

$N$ 9024  
$R^2$ 0.29  
RMSE 1.81
At left is the expected wealth for Black households, at different levels of Education and Age. 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

$$\log(500) \pm 1.8 = 6.2 \pm 1.8$$
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

\[
\log(500) \pm 1.8 = 6.2 \pm 1.8 \text{ logged dollars} = \text{between } \exp(6.2 - 1.8) \text{ and } \exp(6.2 + 1.8)
\]
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

\[ \log(500) \pm 1.8 = 6.2 \pm 1.8 \text{ logged dollars} \]
\[ = \text{ between } \exp(6.2 - 1.8) \text{ and } \exp(6.2 + 1.8) \]
\[ = \text{ between } $81.5k \text{ and } $2981.0k \]
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

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimates</th>
<th>se</th>
<th>t-stat</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>11.25</td>
<td>(0.73)</td>
<td>15.46</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Education Ratio</td>
<td>-0.08</td>
<td>(0.01)</td>
<td>-9.93</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>GDP per capita ($k)</td>
<td>-0.05</td>
<td>(0.01)</td>
<td>-5.32</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

N = 130

R² = 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)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>se</th>
<th>t-stat</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Education Ratio</td>
<td>-0.05</td>
<td>0.01</td>
<td>-6.26</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>log(GDP per capita)</td>
<td>-0.72</td>
<td>0.09</td>
<td>-8.02</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Intercept</td>
<td>9.48</td>
<td>0.73</td>
<td>13.03</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

\[ N = 130 \]
\[ R^2 = 0.71 \]
\[ \text{RMSE} = 0.91 \]

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

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

Standard errors in parentheses
Logging GDP has made the residuals a bit smaller, and a bit more symmetrical.
Logging GDP now allows small increases in GDP per capita to dramatically lower fertility. But small changes in GDP have very little effect in rich countries.
Logging GDP now allows small increases in GDP per capita to dramatically lower fertility.

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 Prediction Error

Chris Adolph (University of Washington)
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.
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

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE</th>
<th>CV Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Education</td>
<td>1.12</td>
<td>1.25</td>
</tr>
<tr>
<td>GDP</td>
<td>1.35</td>
<td>1.84</td>
</tr>
<tr>
<td>Education, GDP</td>
<td>1.01</td>
<td>1.04</td>
</tr>
<tr>
<td>Education, log GDP</td>
<td>0.91</td>
<td>0.85</td>
</tr>
</tbody>
</table>

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.