Data, Environment and Society: Lecture 12: Linear Regression Wrapup

Instructor: Duncan Callaway
GSI: Seigi Karasaki

October 2, 2018

Announcements

Today

- Some regression loose ends
- Discuss Novotny *et al*

Reading

- Today: Novotny *et al*
- Thursday: DS100 Ch 11 (Gradient descent)
- Next Tuesday: Clark, Millet and Marshall, *PloS One* 2014.

Prediction models in the news

The Upshot

Detailed New National Maps Show How Neighborhoods Shape Children for Life

Some places lift children out of poverty. Others trap them there. Now cities are trying to do something about the difference.

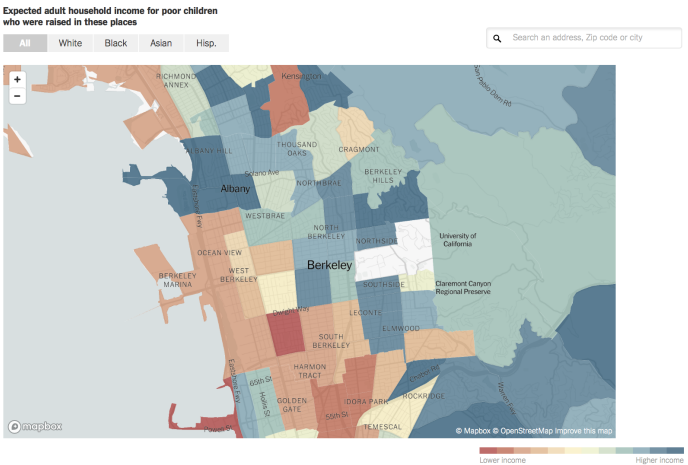
By Emily Badger and Quoctrung Bui

Oct. 1, 2018

225

"The Opportunity Atlas is built using anonymized data on 20 million Americans who are in their mid-thirties today. We map these individuals back to the Census tract (geographic units consisting of about 4,200 people) in which they grew up. Then, for each of the 70,000 tracts in America, we estimate children's average earnings, incarceration rates, and other outcomes by their parental income level, race, and gender."

Prediction in the news, continued



Poor indicates families making about $27,000 a year (in 2015 dollars), at the 25th percentile of the national income distribution. Adult incomes were measured when children were in their mid-30s.

Source: The Opportunity Atlas

More exploring **here**

Is this a qualitative predictor?

Pop-up shops are an annoying way to get me to think shopping is fun.

1. Strongly disagree
2. Disagree
3. Neither agree nor disagree
4. Agree
5. Strongly agree

Answer:

Is this a qualitative predictor?

Pop-up shops are an annoying way to get me to think shopping is fun.

1. Strongly disagree
2. Disagree
3. Neither agree nor disagree
4. Agree
5. Strongly agree

Answer: This is not a qualitative variable: it categorizes someone's opinion on a numeric scale. The responses can be *ordered* from one extreme to another. (The scale is called the *Likert scale*.)

Qualitative or not?

What is your favorite type of soup?

1. Split pea
2. Minestrone
3. Other
4. I don't like soup

Answer:

Qualitative or not?

What is your favorite type of soup?

1. Split pea
2. Minestrone
3. Other
4. I don't like soup

Answer: This is a qualitative predictor. There is no context in which these answers can be sorted or summed together.

Qualitative or not?

1. Red
2. Orange
3. Yellow
4. Green
5. Blue
6. Indigo
7. Violet

Answer:

Qualitative or not?

1. Red
2. Orange
3. Yellow
4. Green
5. Blue
6. Indigo
7. Violet

Answer: This could be a qualitative variable *or* a quantitative one. Quantitative might make sense if we're doing research that relates to color frequency spectra (these are ordered in the sequence of the rainbow). Qualitative might make sense if we're trying to understand what color clothing people like to buy.

Qualitative predictor?

What type of roofing material?

1. Thatched
2. Corrugated Metal
3. Composition shingle
4. Clay tile

Answer:

Qualitative predictor?

What type of roofing material?

1. Thatched
2. Corrugated Metal
3. Composition shingle
4. Clay tile

Answer: Again, depends on context. Often roof material is used as a measure of wealth for household surveys in low income countries. But one might also be asking for people's aesthetic preferences, in which case the connection to wealth may not be relevant.

Qualitative predictors, defined

Quantitative predictors

- have a natural order, or
- values can be summed, and
- often have units of measurement.

Qualitative predictors do not have these characteristics.

Any qualitative predictors in the Novotny data set?

parameter	units
impervious surface	%
tree canopy	%
population	no.
major road length[35]	km
minor road length[35]	km
total road length[35]	km
elevation[36]	km
distance to coast	km
OMI NO_2[25,26]	ppb

	Monitor_ID	State	Latitude	Longitude	Observed_NO2_ppb
0	04-013-0019-42602-1	AZ	33.48385	-112.14257	23.884706
1	04-013-3002-42602-6	AZ	33.45793	-112.04601	25.089886
2	04-013-3003-42602-1	AZ	33.47968	-111.91721	19.281969
3	04-013-3010-42602-1	AZ	33.46093	-112.11748	30.645138
4	04-013-4011-42602-1	AZ	33.37005	-112.62070	11.070412

The **State** variable could be treated as a qualitative predictor.

What about **Monitor_ID**?

Any qualitative predictors in the Novotny data set?

parameter	units
impervious surface	%
tree canopy	%
population	no.
major road length[35]	km
minor road length[35]	km
total road length[35]	km
elevation[36]	km
distance to coast	km
OMI NO_2[25,26]	ppb

	Monitor_ID	State	Latitude	Longitude	Observed_NO2_ppb
0	04-013-0019-42602-1	AZ	33.48385	-112.14257	23.884706
1	04-013-3002-42602-6	AZ	33.45793	-112.04601	25.089886
2	04-013-3003-42602-1	AZ	33.47968	-111.91721	19.281969
3	04-013-3010-42602-1	AZ	33.46093	-112.11748	30.645138
4	04-013-4011-42602-1	AZ	33.37005	-112.62070	11.070412

The **State** variable could be treated as a qualitative predictor.

What about **Monitor_ID**? Then each data point would have a unique intercept. The model would overfit the data in the extreme.

Can we do this?

$$
\begin{aligned}
x \equiv \text{soup preference.} & \\
\text{Split pea} \rightarrow x &= 1 \\
\text{Minestrone} \rightarrow x &= 2 \\
\text{Other} \rightarrow x &= 3 \\
\text{Don't like soup} \rightarrow x &= 4
\end{aligned}
$$

Then fit some data, for example age of respondent, to x.

Can we do this?

$$
\begin{aligned}
x \equiv \text{soup preference.} & \\
\text{Split pea} \rightarrow x &= 1 \\
\text{Minestrone} \rightarrow x &= 2 \\
\text{Other} \rightarrow x &= 3 \\
\text{Don't like soup} \rightarrow x &= 4
\end{aligned}
$$

Then fit some data, for example age of respondent, to x.

The problem with this mapping is that it forces the answers to be quantitative when they are not.

Can we do this?

$$
\begin{aligned}
x \equiv \text{soup preference.} & \\
\text{Split pea} \rightarrow x &= 1 \\
\text{Minestrone} \rightarrow x &= 2 \\
\text{Other} \rightarrow x &= 3 \\
\text{Don't like soup} \rightarrow x &= 4
\end{aligned}
$$

Then fit some data, for example age of respondent, to x.

The problem with this mapping is that it forces the answers to be quantitative when they are not.

We actually need need $n - 1$ variables for n mutually exclusive possibilities in a qualitative predictor.

This is how you do it

$$x_1 = \begin{cases} 1, & \text{Likes split pea.} \\ 0, & \text{otherwise.} \end{cases}$$

$$x_2 = \begin{cases} 1, & \text{Likes minestrone.} \\ 0, & \text{otherwise.} \end{cases}$$

$$x_3 = \begin{cases} 1, & \text{Doesn't like soup.} \\ 0, & \text{otherwise.} \end{cases}$$

Question: What about the "other" category?

This is how you do it

$$x_1 = \begin{cases} 1, & \text{Likes split pea.} \\ 0, & \text{otherwise.} \end{cases}$$

$$x_2 = \begin{cases} 1, & \text{Likes minestrone.} \\ 0, & \text{otherwise.} \end{cases}$$

$$x_3 = \begin{cases} 1, & \text{Doesn't like soup.} \\ 0, & \text{otherwise.} \end{cases}$$

Question: What about the "other" category?
Answer: The answers are mutually exclusive, so if x_1, x_2, x_3 are all zero, then the answer must be "other".

Cooked-up example: Predicting age

$$\text{AGE}_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \beta_4 x_{4,i} + \epsilon_i$$

where x_1, x_2, x_3 are defined on previous slide and x_4 is how spicy the respondent likes their food.

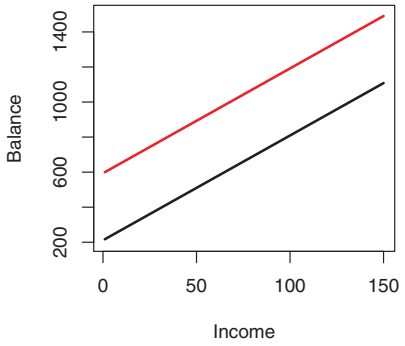
With these variables the qualitative predictors are just modifying the intercept.

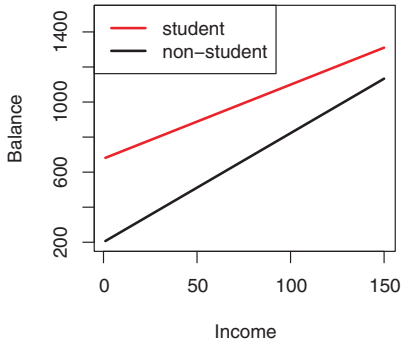
- When $x_1 = x_2 = x_3 = 0$ (i.e. the answer is "other") then the intercept is β_0.
- Otherwise the intercept is $\beta_0 + \beta_i$ where i is the x variable that is nonzero.

Which equation belongs to which picture? (From ISLR)

$$\texttt{balance}_i \approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 & \text{if } i\text{th person is a student} \\ 0 & \text{if } i\text{th person is not a student} \end{cases}$$

$$\texttt{balance}_i \approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 + \beta_3 \times \texttt{income}_i & \text{if student} \\ 0 & \text{if not student} \end{cases}$$





Which equation belongs to which picture?

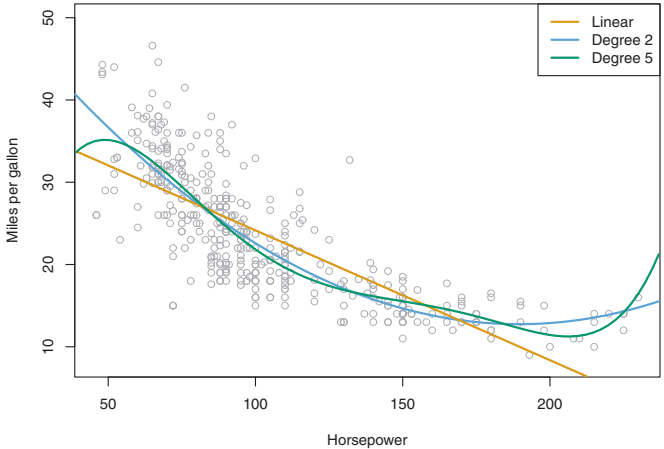
$$\texttt{balance}_i \approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 & \text{if } i\text{th person is a student} \\ 0 & \text{if } i\text{th person is not a student} \end{cases}$$

$$\texttt{balance}_i \approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 + \beta_3 \times \texttt{income}_i & \text{if student} \\ 0 & \text{if not student} \end{cases}$$

The first equation produces two lines with different intercepts $\rightarrow$ left figure.

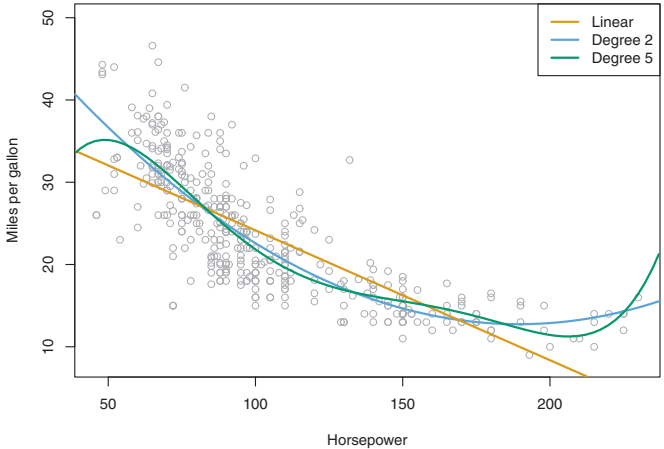
The second equation produces two lines with different intercepts *and* slopes $\rightarrow$ right figure.

What if the relationship seems nonlinear?



What's wrong with this statement: We're doing *linear regression*, so we can only capture *linear relationships* between our data?

What if the relationship seems nonlinear?



What's wrong with this statement: We're doing *linear regression*, so we can only capture *linear relationships* between our data?

Answer: We can *make* new nonlinear predictors to capture the relationship of interest.

Nonlinear predictors

We can specify virtually any nonlinear model you can think of. For example:

$$\hat{y}_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^{\frac{1}{3}} + \beta_4 f(x_i)$$

$f(x_i)$ can be any function you want!

Let's see how this might play out in the Novotny data. Check out the Lecture 12 Jupyter notebook.

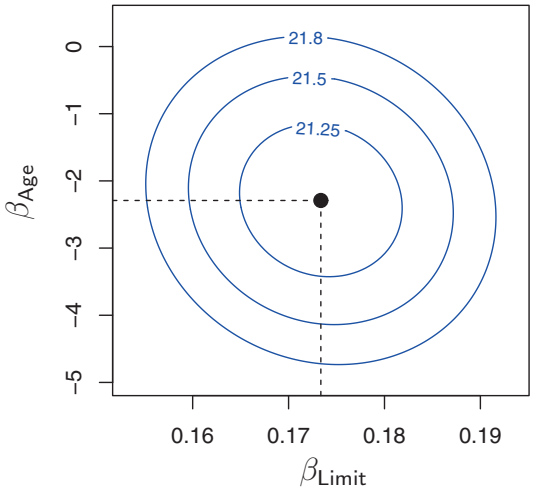
Potential problems

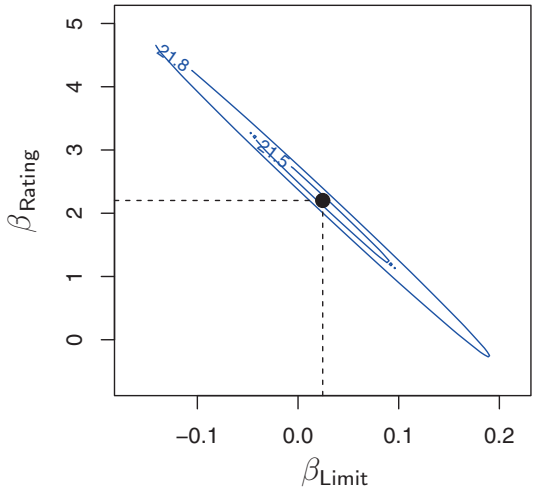
- Non-linearity of the response-predictor relationships.
- Correlation of error terms.
- Non-constant variance of error terms.
- Outliers.
- High-leverage points.
- Collinearity.

Collinearity

- Collineartity is the condition in which independent variables are strongly correlated with each other.
- It doesn't need to be that variable x_j is correlated with variable x_i.
- For example, it could be that $2x_j + 1.3x_k$ is correlated with x_i.
- In other words, *linear combinations* of variables could be correlated with each other.
- **Key problem:** results in inflated standard errors for coefficient estimates.

Colinearity – example from ISLR





Variance inflation

We can measure the extent to which collinearity seems to be impacting results by the VIF, or *variance inflation factor*.

$$\begin{aligned} \text{VIF} &= \frac{\text{variance of } \hat{\beta}_j \text{ when fit with all other variables}}{\text{variance of } \hat{\beta}_j \text{ when it is the only independent variable}} \\ &\geq 1 \end{aligned}$$

Question: If we are focusing on prediction, should we care about evaluating VIF?

Reading questions: Alstone et al

- Review Figure 1 in detail. From a visualization perspective, what features of the figure do you appreciate? Do you think it could be improved upon?
- In the section **Electricity and human development**, the authors state that their figure is 'consistent with an aggregate view of household-level diminishing returns on energy consumption....' Discuss what the authors mean by this statement. Do you agree?
- In the section **The electricity continuum**, the authors argue that 'By overcoming access barriers, often through market-based structures, these systems provide incremental and often substantial increases in access to services, compared with the status quo.' Contrast this statement to the premise of Lee *et al* (which we read last week). Is Alstone *et al*'s view consistent or in conflict with Lee *et al*?

Reading questions: Novotny et al

- Describe the basic model selection process and how it differs from AIC (which we learned briefly in class on Thursday). This question can be answered if you read Section 2.2 and 2.3 carefully and do a little background research to understand unfamiliar terms.
- The authors compare R2 values for the test data versus R2 for the training data. What do they observe? What does this imply about their data and model? Read Section 3.3 to answer this question.
- Also in section 3.3 the authors state "We also investigated the extent to which monitor locations span the (independent) variable space, an important issue for any LUR". What does this mean? Why is spanning the variable space important?
- The authors discuss several limitations in the Discussion section. Review these. How important are their limitations? Are there others you think are worth considering?
- Why use the VIF metric if they are doing prediction?