Lecture 14: Modeling Choices and Predicted Values

Pierre Biscaye

Fall 2022

Agenda

1. More on functional form
2. Interaction terms
3. Choosing X variables to include
4. Predicted values

Functional form choices

- We have seen how changing units can affect regression interpretation but not inference.
- Changing functional form *can* affect inference.
- Common functional forms include:

$$y = \beta_0 + \beta_1 x + u \qquad \textit{linear} \tag{1}$$
$$y = \beta_0 + \beta_1 log(x) + u \qquad \textit{log} \tag{2}$$
$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u \qquad \textit{quadratic} \tag{3}$$

- Many variations of these combining different logs and polynomials, some less common transformations, and interaction terms.

Selecting functional form

1. Which interpretations are *a priori* resonable?
   - Think about how you would theoretically model the relationship.
   - Is it likely to be simple linear? If no reason to think not, linear may make sense. If not, think about likely shape.
     - A bit more complicated in MLR setting, but general intuition may still hold.
   - Do you think what matters is unit changes or percent changes in x?
   - With logs, effects will be monotonic and decreasing.
   - Quadratic forms estimate effects with a u or inverted-u shape. Higher order polynomials are more complicated to interpret.
   - Quadratic turning point: $x = -\frac{\beta_1}{2\beta_2}$

Selecting functional form

2. Which form fits the data best?
   - Linear model finds the best line to fit the data.
   - Log model finds the best logarithm shape to fit the data.
   - Quadratic model finds the best parabola to fit the data.
   - Adjusted-R^2 is one indicator of which fit is best. Accounts for different number of variables included.
   - Provides a way to test across non-nested models (can't do an F test).

$$\bar{R^2} = 1 - \frac{\frac{SSR}{n-k-1}}{\frac{SST}{n-1}} = 1 - (\frac{SSR}{SST})(\frac{n-1}{n-k-1}) = 1 - \frac{n-1}{n-k-1}(1-R^2) \tag{4}$$

   - Can think of Adjusted-R^2 as a kind of tiebreaker, but don't lean too heavily on it: only tells you about the fit for your particular sample.

Selecting functional form

3. What does X look like in terms of density and support?
   - Levels often used for variables with a (relatively) narrow range of values.
   - Logs often used for variables with a long range (e.g., things measured in \$s).
   - Logs place low weight on differences between large values and high weight on changes at low values: may be desirable if there are large positive outliers.
   - $log(0)$ is undefined. If X has a lot of zeros sometimes $log(1 + x)$ is used.

Functional form practice

Choose the functional form you might use for the following relationships:

1. Y = probability of being employed and X = age
2. Y = probability of completing 4-year college and X = parents' combined annual income
3. Y = grade in this class and X = lecture attendance
4. Y = grade in this class and X = time spent studying for exams

Justify your decision!

What about functional forms of Y?

- Consider two models

$$wage_i = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + \beta_3 exper_i^2 + u_i \quad (5)$$

$$log(wage_i) = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + \beta_3 exper_i^2 + u_i \quad (6)$$

- How can we select between functional forms of y?

Consider the Adjusted-R^2

$$\bar{R}^2 = 1 - \frac{\frac{SSR}{n-k-1}}{\frac{SST}{n-1}} \tag{7}$$

- $n-k-1$ and $n-1$ are the same for both equations.
- But both SSR and SST change when we move from levels to logs.

$$\bar{R^2}_{levels} > \bar{R^2}_{logs} \quad if$$
$$\frac{SSR_{logs}}{SST_{logs}} > \frac{SSR_{levels}}{SST_{levels}}$$

- Can't make a clear comparison based on $\bar{R}^2$
  - If $Y > 1$, $log(Y)$ will always have smaller SST than Y because it is compressed.
  - Unlike comparing $\bar{R}^2$ across functional forms of X where the denominator SST is not changing.
- Will often need to rely on the other rules for choosing functional form.

What if the relationship between Y and X depends on Z?

- There are many situations where we might expect the relationship between dependent variable Y and independent variable X will vary by some other variable Z.
- Let's go back to the idea of hedonic pricing: assume that housing prices reflect how much people are willing to pay for a bundle of amenities.
- Have data on house price ($1000s), square footage, and number of bedrooms, and model:

$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + u \tag{8}$$

- What if relationship between bedrooms and price is different depending on square footage?
  - Why might we think this?

Another functional form: interaction terms

- What if relationship between bedrooms and price is different depending on square footage?
- To estimate this, we use an *interaction term*:

$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + \beta_3 sqrft * bdrms + u \tag{9}$$

- What then is the effect of an increase in the number of bedrooms? Partial derivative:

$$\Delta price = (\beta_2 + \beta_3 sqrft)\Delta bdrms \tag{10}$$

- Similar to total effect with quadratic functional form: the relationship between Y and X varies with the level of X itself.

To Jupyter!

Interprations with interactions

$$\widehat{price} = 181.69 + 0.033 sqrft - 35.96 bdrms + 0.023 sqrft * bdrms \quad (11)$$

- $\hat{\beta_2} < 0$, but $\hat{\beta_2}$ is the relationship between a bedroom and price in a 0 square foot house.
  - Why?
  - Total effect of an additional bedroom is $-35.96 + 0.023 sqrft$: depends on square footage.

Interprations with interactions

$$\widehat{price} = 181.69 + 0.033 sqrft - 35.96 bdrms + 0.023 sqrft * bdrms \quad (11)$$

- $\hat{\beta_2} < 0$, but $\hat{\beta_2}$ is the relationship between a bedroom and price in a 0 square foot house.
  - Why?
  - Total effect of an additional bedroom is $-35.96 + 0.023 sqrft$: depends on square footage.
- Interpretation: adding a bedroom adds value when you have space for it.
  - Becomes positive at around 1560 square feet

Interprations with interactions

$$\widehat{price} = 181.69 + 0.033 sqrft - 35.96 bdrms + 0.023 sqrft * bdrms \quad (11)$$

- $\hat{\beta_2} < 0$, but $\hat{\beta_2}$ is the relationship between a bedroom and price in a 0 square foot house.
  - Why?
  - Total effect of an additional bedroom is $-35.96 + 0.023 sqrft$: depends on square footage.
- Interpretation: adding a bedroom adds value when you have space for it.
  - Becomes positive at around 1560 square feet
- How to summarize effect of bedrooms? One approach is to interpret at the mean for square feet.
  - Mean house is about 2000 square feet (in these data).
  - Average effect of a bedroom is $\hat{\beta_2} + 2000 * \hat{\beta_3} = -35.96 + 0.023 * 2000 = 10.04$
  - Close to what we find in the simple linear regression.

Another interaction example

- Consider the relationship between weekly wages and years of experience.
- We might think wages change differently with experience in different sectors.
  - For example, could think that raises and possibility of promotion may be greater in "professional" (i.e., office) jobs relative to other jobs
- Use an interaction term to test this.

$$lwage = \beta_0 + \beta_1 exper + \beta_2 profocc + \beta_3 exper * profocc + u \quad (12)$$

$$\Delta lwage = (\beta_1 + \beta_3 profocc)\Delta exper \quad (13)$$

- How do I interpret these coefficients?
- How do I test whether wages increase with experience?
- How do wages increase with experience for professional occupations?

To Jupyter!

How to determine which X variables to include?

Need to balance a few objectives:

1. Want to include variables that make sense to include theoretically.
2. Want to diminish potential for omitted variables bias.
3. Want to minimize sum of squared residuals to reduce standard errors in β_j.
   - But be careful about adding highly collinear variables which increase R_j^2.

$$SE(\beta_j) = \frac{\hat{\sigma^2}}{SST_j * (1 - R_j^2)} = \frac{SSR}{(n - k - 1)SST_j * (1 - R_j^2)} \tag{14}$$

4. Want to make sure *ceteris paribus* interpretations make sense.
   - Again, worry about collinearity.

Example 1: housing prices

- Consider two hedonic models for housing prices:

$$price_i = \beta_0 + \beta_1 sqrft_i + \beta_2 bdrms_i + u_i \tag{15}$$

$$price_i = \beta_0 + \beta_1 sqrft_i + \beta_2 bdrms_i + \beta_3 assess_i + u_i \tag{16}$$

- Have data on house price ($1000s), square footage, number of bedrooms, and assessed price ($1000s)
- Any concerns about the second model?

To Jupyter!

Example 2: alcohol taxes

- A policy maker wants to analyze the effectivness of a tax on alcohol in reducing drunk driving fatalities.
- Have data from counties on annual drunk driving fatalities, the level of alcohol tax, average annual miles driven per household, and average annual beer consumption per household in liters.
- Consider two models:

$$fatalities_i = \beta_0 + \beta_1 tax_i + \beta_2 miles_i + u_i \tag{17}$$

$$fatalities_i = \beta_0 + \beta_1 tax_i + \beta_2 miles_i + \beta_3 beercons_i + u_i \tag{18}$$

- How does the interpretation of β_1 change?
- Which model should we use?

Another use of regression models: predicted values

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + ... + \hat{\beta}_k x_{ki} \tag{19}$$

- So for values $x_1 = c_1, x_2 = c_2, ..., x_k = c_k$
- If we want to estimate $\theta_0 = \beta_0 + \beta_1 c_1 + \beta_2 c_2 + ... + \beta_k c_k = E[y_i | x_1 = c_1, x_2 = c_2, ..., x_k = c_k]$
- We would estimate

$$\hat{\theta_0} = \hat{\beta}_0 + \hat{\beta}_1 c_1 + \hat{\beta}_2 c_2 + ... + \hat{\beta}_k c_k \tag{20}$$

- If MLR assumptions hold, then $\hat{\beta}_j$s are consistent estimators for β_j and so $\hat{\theta_0}$ is also a consistent estimator.
- This is useful!
  - Think back to our early model where we wanted to predict how CO_2/cap would change across countries as GDP/cap increases.

Variance of predictive values

- We care not just about the predicted value but also its precision.
  - A very imprecise/variable predicted value is not very useful.
- Calculating the variance of the predicted value is not straightforward:

$$\begin{aligned} var(\hat{\theta_0}) &= var(\hat{\beta_1}c_1 + \hat{\beta_2}c_2 + ... + \hat{\beta_k}c_k) \\ &\neq var(\hat{\beta_1}c_1) + var(\hat{\beta_2}c_2) + ... + var(\hat{\beta_k}c_k) \end{aligned} \tag{21}$$

- We can use a variable substitution trick:

$$\theta_0 = \beta_0 + \beta_1 c_1 + \beta_2 c_2 + ... + \beta_k c_k \tag{22}$$

$$\beta_0 = \theta_0 - \beta_1 c_1 - \beta_2 c_2 - ... - \beta_k c_k \tag{23}$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \tag{24}$$

$$y = \theta_0 + \beta_1(x_1 - c_1) + \beta_2(x_2 - c_2) + ... + \beta_k(x_k - c_k) \tag{25}$$

- So we can regress y on $x_1 - c_1$, $x_2 - c_2$, ..., $x_k - c_k$ and use the constant to estimate θ_0 and its standard error.

Example: predicting college GPA

- Suppose we were running college admissions.
- We want to predict success in college (proxied by GPA) using SAT scores, high school class sizes (in 100s), and high school class percentile (lower percentile indicates higher relative rank).
- Using data on our current students, we regress

$$colgpa_i = \beta_0 + \beta_1 SAT_i + \beta_2 hsperc_i + \beta_3 hsize_i + \beta_4 hsize_i^2 + u_i \quad (26)$$

- What is the predicted college gpa for someone with a SAT of 1200, in the 30th percentile of their graduating class, with a high school graduating class of 500?

To Jupyter!

Predicting college GPA

- $E[colgpa|SAT = 1200, hsperc = 30, hsize = 5] = 2.7$
- And $SE(E[colgpa|SAT = 1200, hsperc = 30, hsize = 5]) = 0.02$
- So, our 95% confidence interval for $E[colgpa|SAT = 1200, hsperc = 30, hsize = 5]$ is $[2.66, 2.74]$
- Note, this does *not* mean we expect 95% of people with these characteristics to have a college GPA in this range.
- Instead, we expect people with these characteristics to have a college GPA in this range *on average* (with 95% probability). There will still be variation!