Lecture 13: Regression Interpretations

Pierre Biscaye

Fall 2022

Agenda

1. Midterm 1 recap
2. Hypothesis testing recap
3. Units and regression interpretation
4. Functional form and regression interpretation
5. (Time permitting) interaction terms

Midterm 1 recap

- Mean 37/50, median 39.25 - good work!
- Approximate curve posted on bcourses; similar to how I will curve overall final grade
- Solutions posted; regrade requests close on Sunday
- Challenging questions:
  - Interpreting $\hat{\beta}$s with logs: $gradrate_i = \beta_0 + \beta_1 lsalary_i + \beta_2 lnchprg_i + v_i$
  - Using formula for MLR SE (typo in solutions) to think about changes in SE: $SE(\hat{\beta}_2) = \frac{SSR}{(n-k-1)SST_{lnchprg}(1-R^2_{lnchprg})}$
  - Hypothesis testing: setting up H0 and H1, calculating test statistic, finding critical value, interpreting test and concluding about null
- Midterm 2 is Tuesday November 1

Hypothesis testing recap

- Simple hypothesis test: $H_0 : \beta_j = \beta_{j0}$; $t = \frac{\hat{\beta}_j - \beta_{j0}}{SE(\hat{\beta}_j)} \sim t_{n-k-1}$
  - Same for other parameter estimates: just replace β_j with the parameter
  - Exception is binary variables/proportions: they are computed the same but are $z \sim N(0, 1)$ because the mean under the null tells us the SD so we don't need to estimate the SE
- Confidence intervals: don't specify β_{j0}, estimate $\left[\hat{\beta}_j - c_{\frac{\alpha}{2}} * SE(\hat{\beta}_j), \hat{\beta}_j + c_{\frac{\alpha}{2}} * SE(\hat{\beta}_j)\right]$
- Linear combinations: $H_0 : \beta_1 - \beta_2 = b$; $t = \frac{(\hat{\beta_1} - \hat{\beta_2}) - b}{SE(\hat{\beta_1} - \hat{\beta_2})} \sim t_{n-k-1}$
  - Tricky part is the SE of the linear combination: solve by defining a parameter $\hat{\theta} = \hat{\beta_1} - \hat{\beta_2}$ and doing simple hypothesis test
  - For regression parameter estimates, use substitution to rewrite model to estimate $\hat{\theta}$

Joint hypothesis tests

- Want to test multiple restrictions, e.g., $H_0 : \beta_1 = 0 \textit{ and } \beta_2 = 0$
  - Are these variables *jointly* significant?
- Consider two models:
  - Unrestricted $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + u$
  - Restricted $y = \beta_0 + \beta_3 x_3 + \cdots + u$
- Construct test statistic based on model fit:
  - $F = \frac{(SSR_r - SSR_u)/q}{SSR_u/(n-k-1)} = \frac{(R_u^2 - R_r^2)/q}{(1-R_u^2)/(n-k-1)}$
  - $F \sim F(q, n-k-1)$
  - Reject H_0 if F is larger than critical value
- Extreme case: overall F-statistic: do *any* of the variables have explanatory power?
  - $H_0 : \beta_1 = 0 \textit{ and } \beta_2 = 0 \textit{ and } \ldots \textit{ and } \beta_k = 0$
  - $R_r^2 = 0$, compare to R_u^2 via F stat

Practice: match questions to approach and write H_0

Suppose we model

$$bwght_i = \beta_0 + \beta_1 cigs_i + \beta_2 parity_i + \beta_3 faminc_i + \beta_4 motheduc_i + \beta_5 fatheduc_i + u_i$$

Questions

1. What range of values is β_1 likely to take with 95% probability?
2. Do socioeconomic characteristics matter for birthweight, holding cigarette smoking and birth order constant?
3. Which is worse for birth weight holding other variables constant: smoking an additional cigarette per day or decreasing annual family income by \$10,000?
4. Does being born after siblings (*parity*) significantly increase birth weight?

Approach

A. Simple hypothesis test
B. Confidence interval
C. Linear combination hypothesis test
D. Joint hypothesis test

Units and regression interpretation

- Units matter: every interpretation of a $\hat{\beta}_j$ must specify the units of both the dependent and independent variables
- Why do they matter? Suppose we estimate

$$CO_2 / pop = 0.75 + 0.24 GDP / pop$$

- Economic significance changes a lot if GDP / pop is in dollars vs. thousands of dollars (it is in 1000s of 2005 PPP USD)
- Economic significance similarly changes a lot if CO_2 / pop is in kilograms vs. tons (it is in tons)
- What happens to coefficient estimates when we manipulate units in a regression model?

Units in dependent variable: example

$$bwght_i = \beta_0 + \beta_1 cigs_i + \beta_2 parity_i + u_i \tag{1}$$

- The mother consuming one more cigarette per day during pregnancy is associated with a change of β_1 ounces at birth, holding birth order constant.
- What if our dependent variable was pounds at birth?
- $bwghtlbs_i = bwght_i/16$

To Jupyter!

Units in dependent variable: mathematically

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u \quad (2)$$

$$\alpha y = \alpha\beta_0 + \alpha\beta_1 x_1 + \alpha\beta_2 x_2 + ... + \alpha\beta_k x_k + \alpha u \quad (3)$$

- All $\hat{\beta}_j$ scale by the same α is the dependent variable.
- Our test statistics do not change. Why not?
- We also still draw exactly the same conclusions about statistical significance.

Units in independent variables: example

$$bwght_i = \beta_0 + \beta_1 cigs_i + \beta_2 parity_i + u_i$$

- What if we were interested in the effect of *packs* of cigarettes per day?
- $packs_i = cigs_i / 20$

To Jupyter!

Units in independent variable: mathematically

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 ... + \beta_k x_k + u \tag{4}$$

$$y = \beta_0 + (\frac{\beta_1}{\alpha})\alpha x_1 + \beta_2 x_2 + ... + \beta_k x_k + u \tag{5}$$

- Once again, changing the units only scales $\hat{\beta}$ estimates.
- Changing units for one x_j only affects $\hat{\beta}_j$, no other $\hat{\beta}$s
- No changes to test statistics, or conclusions.
- Sometimes changing units may lead to easier interpretations: e.x. CO_2 and GDP
  - $CO_2 / pop = 0.75 + 0.24 GDP / pop$
  - Useful to rescale units if $\hat{\beta}$ is a small decimal, for example, or if range of a variable is very wide so small marginal changes are not as interesting

Standardizing variables

- Units will often be difficult to compare across variables, and sometimes may not have a clear interpretation.
  - E.g., test scores when grade inflation is different across contexts
- When we want to compare effects of variables with different units or have variables with units that are hard to interpret, can be helpful to *standardize* variables.
  - Standardizing: $\tilde{x} = (x - \bar{x})/\sigma_x$
  - $\tilde{x}$ now measured in units of standard deviations, with magnitude indicating distance away from the mean.
  - This is how we construct t and z statistics.

Example: Pollution and hedonic pricing

- Suppose we want to know how badly pollution reduces welfare. How to estimate this?
- Could ask people how much they would pay to reduce pollution: *contingent valuation*.
  - But how reliable are these stated preferences?
- An alternative approach uses revealed preferences: *hedonic pricing*.
  - Common in environmental economics to assume that housing prices reflect how much people are willing to pay for a bundle of amenities.
  - Differences in house prices with different levels of an amenity reveal willingness to pay for that amenity
- An issue with hedonic pricing: amenities that can affect house prices have very different units (e.g., rooms in a house, distance from elementary school in miles, etc.): how to compare relative importance of these amenities?
  - Use standardized variables.

Example: Pollution and hedonic pricing

- We have data on house prices and amenities for communities in the Boston area (decades ago):
  - Median house price in $
  - Nitrogen oxide concentration in parts per 100m
  - Crimes committed per capita in a year
  - Average number of rooms
  - Weighted distance to 5 nearest employment centers, miles
- We first estimate

$$price = \beta_0 + \beta_1 nox + \beta_2 crime + \beta_3 rooms + \beta_4 dist + u \qquad (6)$$

To Jupyter!

Comparing *crime* and *nox*

- How to compare an increase of 1 crime per capita in a year to an increase of 1 part per 100m of nox?
  - $\beta_{nox} = -2381.2$, $\beta_{crime} = -213.5$
  - $|\beta_{nox}| > |\beta_{crime}|$: is nox more important for housing prices? Can't tell. What to do?
- Standardize the data.

$$
\begin{aligned}
y &= \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u \\
\bar{y} &= \beta_0 + \beta_1 \bar{x_1} + ... + \beta_k \bar{x_k} \\
(y - \bar{y}) &= \beta_1 (x_1 - \bar{x_1}) + ... + \beta_k (x_k - \bar{x_k}) + u \\
\frac{y - \bar{y}}{\sigma_y} &= \frac{\beta_1}{\sigma_y}(x_1 - \bar{x_1}) + ... + \frac{\beta_k}{\sigma_y}(x_k - \bar{x_k}) + \frac{u}{\sigma_y} \\
\frac{y - \bar{y}}{\sigma_y} &= \beta_1 \frac{\sigma_{x_1}}{\sigma_y} \frac{x_1 - \bar{x_1}}{\sigma_x} + ... + \beta_k \frac{\sigma_{x_k}}{\sigma_y} \frac{x_k - \bar{x_k}}{\sigma_{x_k}} + \frac{u}{\sigma_y}
\end{aligned}
$$

Standardized variables

$$\frac{y - \bar{y}}{\sigma_y} = \beta_1 \frac{\sigma_{x_1}}{\sigma_y} \frac{x_1 - \bar{x_1}}{\sigma_x} + ... + \beta_k \frac{\sigma_{x_k}}{\sigma_y} \frac{x_k - \bar{x_k}}{\sigma_{x_k}} + \frac{u}{\sigma_y} \tag{7}$$

- If we run this regression, we estimate $\frac{\widehat{\beta_j \sigma_{x_j}}}{\sigma_y}$
- Interpretation: effect of a one *standard deviation* increase in x_j on *standard deviations* of y, holding all else constant.
- Allows comparability between variables.
- Estimated coefficients when fully standardizing the model are called "Standardized effects" or (unfortunately) "Beta coefficients".
- Now what can we say about the relative effects of crime and nox?

To Jupyter!

Functional form choices and interpretation

- We have seen how changing units can affect regression interpretation.
- But changing units does not change statistical significance/inference.
- Changing functional form *can* affect inference.
- Common functional forms include:

$$y = \beta_0 + \beta_1 x + u \qquad \textit{linear} \tag{8}$$

$$y = \beta_0 + \beta_1 log(x) + u \qquad \textit{log} \tag{9}$$

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u \qquad \textit{quadratic} \tag{10}$$

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

Functional form differences

- Changes in functional form *affect* interpretations of β estimates.
  - Linear: one unit increase in x increases y by β_1 units.
  - Level-Log: one percent increase in x increases y by $\beta_1/100$ units.
  - Quadratic: one unit increase in x increases y by $\beta_1 + 2\beta_2 x$ units.
- In quadratic models, when $\hat{\beta_1}$ and $\hat{\beta_2}$ have different signs there wll be a turning point.
  - Turning point is $x = -\frac{\beta_1}{2\beta_2}$.
  - $\beta_1 > 0$ and β_2 <0: y increases with x until $-\frac{\beta_1}{2\beta_2}$ and decreases after.
- These are big changes.
- Each model will find the best fit for that shape.
  - 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.

Functional form differences: example

- Consider an example using our familiar data wages (per hour) and experience (years of work):
  - $log(wage) = \beta_0 + \beta_1 exper + u$
  - $log(wage) = \beta_0 + \beta_1 log(exper) + u$
  - $log(wage) = \beta_0 + \beta_1 exper + \beta_2 exper^2 + u$

To Jupyter!

How to test significance with quadratic functional form?

$$log(wage_i) = \beta_0 + \beta_1 exper_i + \beta_2 exper_i^2 + u_i \tag{11}$$

- How do you test if there is a relationship between wages and experience?

How to test significance with quadratic functional form?

$$log(wage_i) = \beta_0 + \beta_1 exper_i + \beta_2 exper_i^2 + u_i \tag{11}$$

- How do you test if there is a relationship between wages and experience?
- Does $H_0 : \beta_2 = 0$ or $H_0 : \beta_3 = 0$ deliver the right test?

How to test significance with quadratic functional form?

$$log(wage_i) = \beta_0 + \beta_1 exper_i + \beta_2 exper_i^2 + u_i \tag{11}$$

- How do you test if there is a relationship between wages and experience?
- Does $H_0 : \beta_2 = 0$ or $H_0 : \beta_3 = 0$ deliver the right test?
- Need an F test: $H_0 : \beta_2 = 0$ *and* $\beta_3 = 0$
  - This is a really common use for F tests.

Comparing model fit: R^2

- We've talked about using R^2 to compare between models. Here it is highest for the quadratic model.
- One concern: the R^2 will be mechanically higher when we control for more variables.
  - Can't help you determine whether you should include more variables in your specification.
  - The quadratic model has an extra variable!
  - How to compare models with different numbers of variables?

Comparing model fit: R^2

- We've talked about using R^2 to compare between models. Here it is highest for the quadratic model.
- One concern: the R^2 will be mechanically higher when we control for more variables.
  - Can't help you determine whether you should include more variables in your specification.
  - The quadratic model has an extra variable!
  - How to compare models with different numbers of variables?
- Another concern: R^2 is a biased estimator of ρ^2, what we're really trying to get at: share of the population variance of y that the model explains.

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{\frac{SSR}{n}}{\frac{SST}{n}} \tag{12}$$

$$\rho^2 = 1 - \frac{\sigma_u^2}{\sigma_y^2} \tag{13}$$

- What then should we look at to compare models?

Adjusted-R^2

- Unlike R^2, the Adjusted-R^2 ($\bar{R^2}$) is an unbiased estimator of ρ^2.

$$\bar{R^2} = 1 - \frac{\frac{SSR}{n-k-1}}{\frac{SST}{n-1}} = 1 - \frac{\hat{\sigma_u^2}}{\hat{\sigma_y^2}} \tag{14}$$

$$E[\bar{R^2}] = \rho^2 \tag{15}$$

- Adjusted-R^2 penalizes for additional regressors (since k goes up).
  - Allows you to compare models with different numbers of variables.
  - Does an additional variable increase your statistical power?
- In fact, Adjusted-R^2 can be <0:

$$\bar{R^2} = 1 - \frac{\frac{SSR}{n-k-1}}{\frac{SST}{n-1}} = 1 - (\frac{SSR}{SST})(\frac{n-1}{n-k-1}) = 1 - \frac{n-1}{n-k-1}(1 - R^2) \tag{16}$$

Selecting functional forms

How to choose which functional form to use?

1. Which interpretations are *a priori* resonable?
   - Think about how you would theoretically model the relationship. Is it likely to be simple linear?

Selecting functional forms

How to choose which functional form to use?

1. Which interpretations are *a priori* resonable?
   - Think about how you would theoretically model the relationship. Is it likely to be simple linear?
2. Which form fits the data best? Adjusted-R^2 is one indicator.
   - A way to test across non-nested models.
   - Lowest for simple linear; highest for quadratic.
   - But, does it make sense for effect of experience to turn negative after 25 years? Could be an omitted variable or a problem with the quadratic form requiring a sign change.

Selecting functional forms

How to choose which functional form to use?

1. Which interpretations are *a priori* resonable?
   - Think about how you would theoretically model the relationship. Is it likely to be simple linear?
2. Which form fits the data best? Adjusted-R^2 is one indicator.
   - A way to test across non-nested models.
   - Lowest for simple linear; highest for quadratic.
   - But, does it make sense for effect of experience to turn negative after 25 years? Could be an omitted variable or a problem with the quadratic form requiring a sign change.
3. What does X look like in terms of density and support?
   - Logs place low weight on differences between large values and high weight on changes at low values; this may or may not be desirable.
   - $log(0)$ is undefined. If X has a lot of zeros sometimes $log(1+x)$ is used.

Functional form summary

- Levels
  - Easy interpretation.
  - Often used for variables with a (relatively) narrow range of values.

Functional form summary

- Levels
  - Easy interpretation.
  - Often used for variables with a (relatively) narrow range of values.
- Logs
  - Percentage interpretation.
  - Undefined for $x \leq 0$.
  - Treats high values as smaller changes than small values: may be desirable if there are large positive outliers.
  - Effects will be monotonic and decreasing,
  - Often used for variables with a long range (e.g., things measured in $s)

Functional form summary

- Levels
  - Easy interpretation.
  - Often used for variables with a (relatively) narrow range of values.
- Logs
  - Percentage interpretation.
  - Undefined for $x \leq 0$.
  - Treats high values as smaller changes than small values: may be desirable if there are large positive outliers.
  - Effects will be monotonic and decreasing,
  - Often used for variables with a long range (e.g., things measured in $s)
- Quadratic (and higher polynomial) forms
  - Quadratic forms estimate effects with a u or inverted-u shape.
  - Sometimes desirable and sometimes not.
  - Interpretations can be challenging.

Interaction terms

- Let's go back to the idea of hedonic pricing.

$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + u \tag{17}$$

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

$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + \beta_3 sqrft * bdrms + u \tag{18}$$

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

$$\Delta price = (\beta_2 + \beta_3 sqrft)\Delta bdrms \tag{19}$$

To Jupyter!

Interprations with interactions

$$\widehat{price} = 181.69 + 0.033sqrft - 35.96bdrms + 0.023sqrft * bdrms \quad (20)$$

- $\hat{\beta_2} < 0$, but $\hat{\beta_2}$ is the relationship between a bedroom and price in a 0 square foot house.
  - $\hat{\beta_2}$ is no longer capturing the total effect of an additional bedroom.
  - Total effect is $-35.96 + 0.023sqrft$: depends on square footage.

Interprations with interactions

$$\widehat{price} = 181.69 + 0.033 sqrft - 35.96 bdrms + 0.023 sqrft * bdrms \quad (20)$$

- $\hat{\beta_2} < 0$, but $\hat{\beta_2}$ is the relationship between a bedroom and price in a 0 square foot house.
  - $\hat{\beta_2}$ is no longer capturing the total effect of an additional bedroom.
  - Total effect 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

$$\widehat{price} = 181.69 + 0.033sqrft - 35.96bdrms + 0.023sqrft * bdrms \quad (20)$$

- $\hat{\beta_2} < 0$, but $\hat{\beta_2}$ is the relationship between a bedroom and price in a 0 square foot house.
  - $\hat{\beta_2}$ is no longer capturing the total effect of an additional bedroom.
  - Total effect is $-35.96 + 0.023sqrft$: 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? 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.