

# PBAF 528

More Interpreting Results

Prediction

(Teaching Notes: finish p. 16-20 from last week;  
Sections 5 and 6, Pages 21-30 new for today)

# Recap

- What about other causal factors?
- Framing a research question, unit of analysis
- Experiments and Quasi-Experiments (Salk, Homelessness, MTO)
- Multivariate Regression

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \dots \hat{\beta}_k X_k + e$$

- Goodness of fit-  $R^2$ , Adjusted  $R^2$
- Hypothesis testing for 1 coefficient
- Interpretation

# More Recap

- Theory and Hypotheses
- Sampling distribution of parameter estimates

$$E(\hat{\beta}_1) = \beta_1 \quad SD(\hat{\beta}_1) = SE_{\hat{\beta}_1}$$

- Hypothesis testing for one coefficient
  - Is one coefficient really zero?
  - t-tests, confidence intervals, p-values.
  - Interpretation
    - A one-unit change in x is associated with a beta-hat unit change in y.
    - Interpretation of confidence intervals

# Today

- More Multivariate Regression
  - Finish discussion of homelessness example
  - What about other causal factors?
    - Bias and the impact of omitted variables (bias)
  - Classical Regression Assumptions
  - Reporting Requirements
- Hypothesis testing for multiple coefficients
  - F-tests
    - Is more than one coefficient really zero?
- Prediction
- The Hunt for Data

# Do cities with rent control laws have higher levels of homelessness?

(Was: does rent control cause homelessness)

- Data from 50 US Cities
- Outcome is log of rate of homelessness
- Rent control is “dummy variable” (0=no law, 1=rent control law)
- Table 1 and Table 2

# Does Rent Control Cause Homelessness?

**Table 2.** Regression of homelessness upon selected characteristics of cities.<sup>a</sup>

| Variable                                 | Tucker's model   | Some simple alternatives |                  |                  |
|--|------------------|--------------------------|------------------|------------------|
|  |                  | I                        | II               | III              |
| Rent control                             | 0.420<br>(3.99)  | 0.234<br>(1.69)          | 0.103<br>(0.60)  | —                |
| Temperature                              | 0.017<br>(2.50)  | 0.013<br>(1.83)          | 0.015<br>(2.09)  | 0.016<br>(2.14)  |
| Percent growth                           | -0.005<br>(1.89) | -0.006<br>(1.99)         | -0.006<br>(1.91) | -0.006<br>(2.03) |
| Poverty rate                             | —                | 0.014<br>(1.22)          | 0.017<br>(1.44)  | 0.020<br>(1.92)  |
| Average rent                             | —                | 0.003<br>(2.04)          | 0.004<br>(2.26)  | 0.005<br>(3.45)  |
| Vacancy rate                             | —                | —                        | -0.018<br>(1.31) | -0.023<br>(2.08) |
| Constant                                 | -0.484<br>(1.28) | -1.173<br>(2.34)         | -1.264<br>(2.52) | -1.402<br>(3.16) |
| Explained variation<br>(adjusted $R^2$ ) | 0.311            | 0.342                    | 0.353            | 0.362            |

<sup>a</sup>  $t$ -ratios are reported in parentheses. The rate of homelessness is measured in common logarithms in a manner consistent with the results reported by Tucker.

# Discussion Questions

1. What do the adjusted R-squares for the models in Table 2 tell you?
2. How did the coefficient for rent control change from Tucker's Model to Model I? How about the change from Model I to Model II? What does the change suggest to you about rent control's relationship with homelessness?
3. What does the change in the t-statistic for rent control for Tucker's Model to Models I and II tell you?
4. How does Model III differ from the other two models? In your opinion, given what you have learned about hypothesis testing and goodness of fit, does this model explain homelessness better than the other models?

# Impact of Omitted Variables

- Included coefficient is too high (toward positive) if
  - the omitted factor is positively correlated with the outcome and included factoror
  - the omitted factor is negatively correlated with the outcome and included factor
- Included coefficient is too low (toward negative) if
  - the omitted factor is positively correlated with the outcome and negatively correlated with the included factoror
  - the omitted factor is negatively correlated with the outcome and positively correlated with the included factor

# Classical Assumptions

1. The regression model is linear in the coefficients, is correctly specified, and has an additive error term.
2. The error term has zero population mean.
3. All explanatory variables are uncorrelated with the error term.
4. Observations of the error term are uncorrelated with each other.
5. The error term has a constant variance.
6. No explanatory variable is a perfect linear function of any other explanatory variable.
7. The error term is normally distributed.

# Classical Assumptions

- So, OLS coefficient estimators have the following properties:
  - Unbiased
  - Minimum Variance
  - Consistent
  - Normally distributed
- OR, OLS is BLUE (Best Linear Unbiased Estimator)

# Reporting Requirements

- Required elements
  - ⇒ Data sources
  - ⇒  $n$
  - ⇒ Descriptives
  - ⇒ Coefficients
  - ⇒ Standard errors
  - ⇒ Goodness of fit
  - ⇒ Issues of concern or debate
- Possible Additions
  - ⇒ Predicted values
  - ⇒ CIs for coefficients
  - ⇒ Standardized coefficients
  - ⇒ Alternative models
  - ⇒ Regressions on selected subsets

Is the coefficient really zero?  
Is more than one coefficient really  
zero?

# Remember

## Hypothesis Testing for 1 Coefficient

- (From Table 2, Tucker's Model)
- Set hypotheses:  $H_0: \beta_1=0$       $H_a: \beta_1 \neq 0$  (two-sided)
- Set decision rule: if  $|t| > t_c$  then reject  $H_0$  at 5% sig level

$$df = n - (k + 1) = 50 - (3 + 1) = 46 \quad \implies t_{\alpha/2} = 2.02$$

(from the t table)

- Find t statistic:

$$t = \frac{\hat{\beta}_1 - \beta_{H_0}}{SE_{\hat{\beta}_1}} = \frac{.42 - 0}{.105} = 3.99$$

- Therefore, can reject the null hypothesis since  $|t| > t_c$

# Hypotheses for 1 Coefficient

- A hypothesis states the relationship between two variables.
- Hypothesis testing is about statistical significance, not theoretical validity.
- T-tests do not test importance of the coefficient.
- T-tests are intended for use on samples.

# Hypothesis Testing for Multiple Coefficients

- Are all the coefficients in the model really zero?
- How do we know if a particular subset of variables adds to the model?

# F-test #1: Does this model have significant explanatory power?

- Set up hypothesis
  - Are all the coefficients really equal to zero?
- Set decision rule: if  $F \geq F_c$  then reject  $H_0$

- Find test (or F) statistic:

$$F = \frac{\frac{ESS}{k}}{\frac{RSS}{n-k-1}} = \frac{\frac{\sum (\hat{Y}_i - \bar{Y})^2}{k}}{\frac{\sum e_i^2}{n-k-1}} = \frac{\frac{R^2}{1-R^2}}{\frac{k}{n-k-1}}$$

- Compare critical F and test statistic

# F-test #2: Do a subset of predictors add explanatory power?

- **Hypothesis**

- Are a subset of coefficients really equal to zero?

Restricted equation (null model)

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

Unrestricted equation

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_M X_M + \varepsilon$$

- **Decision rule**

- if  $F_{(M,n-k-1)} \geq F_c$  then reject  $H_0$

- **Find test (F) statistic**

$$F_{(M,n-k-1)} = \frac{\frac{R_{UR}^2 - R_R^2}{M}}{\frac{1 - R_{UR}^2}{n - k - 1}}$$

OR

$$F_{(M,n-k-1)} = \frac{\frac{RSS_R - RSS_{UR}}{M}}{\frac{RSS_{UR}}{n - k - 1}}$$

**Compare critical F and test statistic**

# Prediction

- What's the predicted level of homelessness in an average city with or without rent control?
- Tucker's Model:

$$\widehat{\text{Homelessness}} = -0.484 + 0.42(\text{Rent Control}) + 0.017(\text{Temperature}) - 0.005(\text{Percent Growth})$$

**With**

$$\begin{aligned}\widehat{\text{Homelessness}} &= -0.484 + 0.42(1) + 0.017(56.94) - 0.005(-1.49) \\ &= -0.484 + 0.42 + 0.968 - 0.00745 \\ &= .91 \quad e^{.91} = 2.484 \text{ homeless per thousand}\end{aligned}$$

**Without**

$$\begin{aligned}\widehat{\text{Homelessness}} &= -0.484 + 0.42(0) + 0.017(56.94) - 0.005(-1.49) \\ &= -0.484 + 0 + 0.968 - 0.00745 \\ &= .49 \quad e^{.49} = 1.631 \text{ homeless per thousand}\end{aligned}$$

We only exponentiate because the DV is logged!!

# In-Class Exercise: Valuing Property

- An local community development corporation (CDC) is trying to value 2 multifamily properties it redeveloped when it first started as a CDC, 15 years ago.
- As their development manager, your boss at the CDC asks you to look into this.
- Your boss took a sample of 25 multifamily buildings at random from all buildings sold during a recent year as a way to begin the analysis and then ran some regression models.

# Discussion Questions

1. Do each of the models presented have significant explanatory power? How do you know?
2. Does the number of parking areas in combination with the age of the structure and lot size add to the explanatory power of the model? (Compare model 1 with model 2)
3. Given the print out you've seen so far and the tests that you've conducted, what variables do you think should remain in the model? Is there other information you'd like before you settle on a model?
4. Based on Model 2, tell your boss the predicted sale values for the two buildings. Here are the characteristics:

# The Hunt for Data

- Where can we find data?
  - The Web
  - CSSCR (<http://julius.csscr.washington.edu/>)
  - Publications
  - Personal contacts (privacy issues?)
- In what form is data when we get it?
  - Text file (ASCII): delimited or not?
  - SPSS, SAS
- How do we need to manipulate data when we get it?
  - Continuous or Discrete?
  - Transformations
  - Cleaning
  - Missing cases, missing variables

# Quiz Preview

- Conditions for causation
- Correlation
- Simple and multiple linear regression
  - The relationship between correlation and regression
  - Coefficient interpretation
  - Measures of Goodness of Fit: R-Square, Adjusted R-Square
  - Confidence intervals for estimated coefficient values
  - Hypothesis testing for one coefficient (t-tests, p-values)
  - What tests would you use to test hypothesis for multiple coefficients?
    - why use Global F-test (#1) or Nested F-test (#2)?
  - Why is the estimated regression coefficient a good estimate of the true regression coefficient?
  - Prediction with regression equations.