

# PBAF 528

What About Other Causal Factors?

and

Interpreting Results

(Sections 3 and 4, Page 10-20 in Teaching Notes)

# Recap

- Correlation
- Simple Linear Regression
  - Graphed the line  $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
  - A one-unit change in  $X$  is associated with a Beta-hat 1 change in  $Y$ .
  - The line explains some variation in the outcome, but not all.
- The Research Process

# Today

- What about other causal factors?
  - 2 Examples
    - Rent Control and Homelessness
    - Salk Polio Vaccine
- Experiments and Quasi-Experiments
- Multivariate Regression
  - Goodness of fit
  - hypothesis testing
  - Interpretation
  - Impact of omitted variables (bias)
  - Classical Regression Assumptions
  - Reporting Requirements

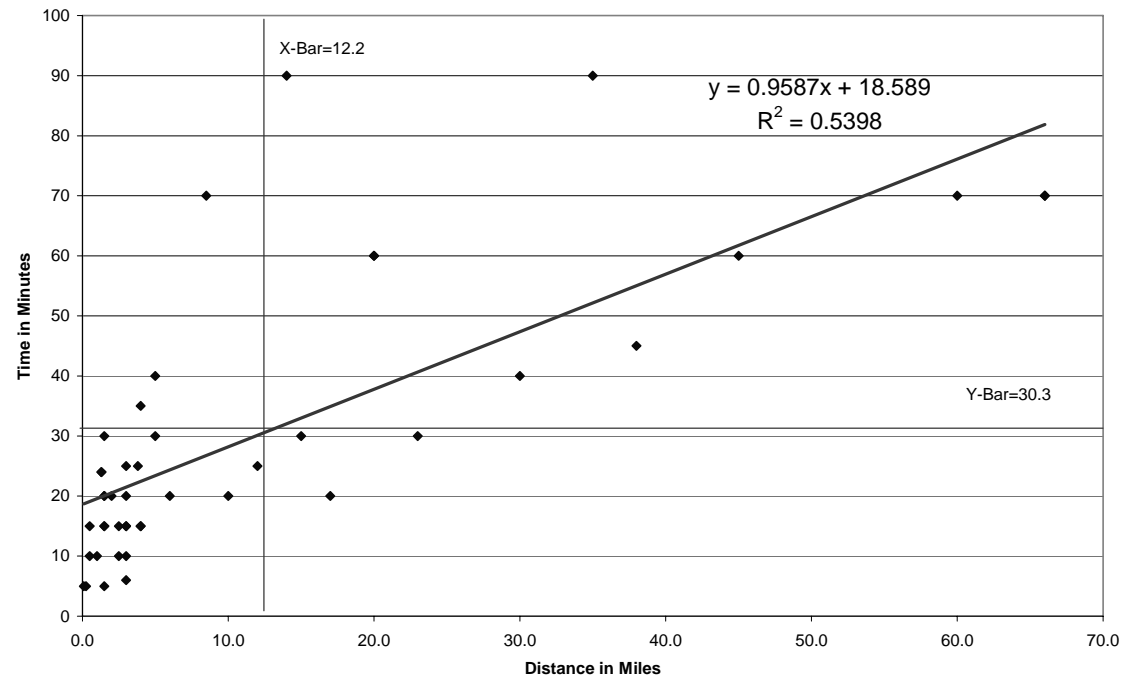
# Recap: Time & Distance From Campus

$$\hat{Y}_i = 18.589 + .9587X_i$$

$$\bar{Y} = 30.3$$

$$(Y_i - \bar{Y}) = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)$$

$$= (\hat{Y}_i - \bar{Y}) + e_i$$



# What about other causal factors?

- Framing a research question
- Designing the research
- What if there are multiple causes??

Why might rent control cause homelessness?

# Experiments & Quasi-Experiments

- Independent Var → change in Dependent Variable
- One Shot Case Study
- Pretest-posttest
- Experiment and Control group
- Validity
  - Internal
  - External

# Example: Salk Polio Vaccine

- Until 1980's most expensive medical experiment in history (\$5 million in 1954)
- Is the killed virus vaccine an effective protection against poliomyelitis?

# A little about Polio

- Polio not common, but frightening
  - 1950s only 6% of death ages 5-9
  - Paralysis, iron lungs
- Occurred more often in more hygienic environments
- Caused by a virus
- Vaccine creates antibodies
  - live virus (smallpox) (risk, bad experiences)
  - killed virus (like the flu vaccine) (Salk)

# How to Evaluate?

- Sample size?
  - Say occurrence is 50 in 100,000 and we think the vaccine is effective 50% of the time.
    - If select 40,000 controls (no vaccine)--20 cases
    - And 40,000 who get vaccine--10 cases
    - Not a significant difference!
  - So, a much larger study required because of the relatively low incidence of the disease!
  - Over a million people needed

# How to Evaluate?

- Vital Statistics
  - Distribute widely to schools to see if rate is lower than in the previous season
  - What's the problem?
- Observed Control
  - Volunteer Control group and Treatment group-- known to all.
  - What's the problem?
- But no room for doubt about this. What to do?

# Randomization and Placebo Control

- Placebo controls (given salt solution)
- Double blind evaluation
- Assigned at Random to treatment or control
- Why?

# Dealing with Multiple Causation

- Experiments (Salk)
- Quasi-Experiments
  - Moving to Opportunity
  - Gateway
- Multiple Regression

# Multiple Regression

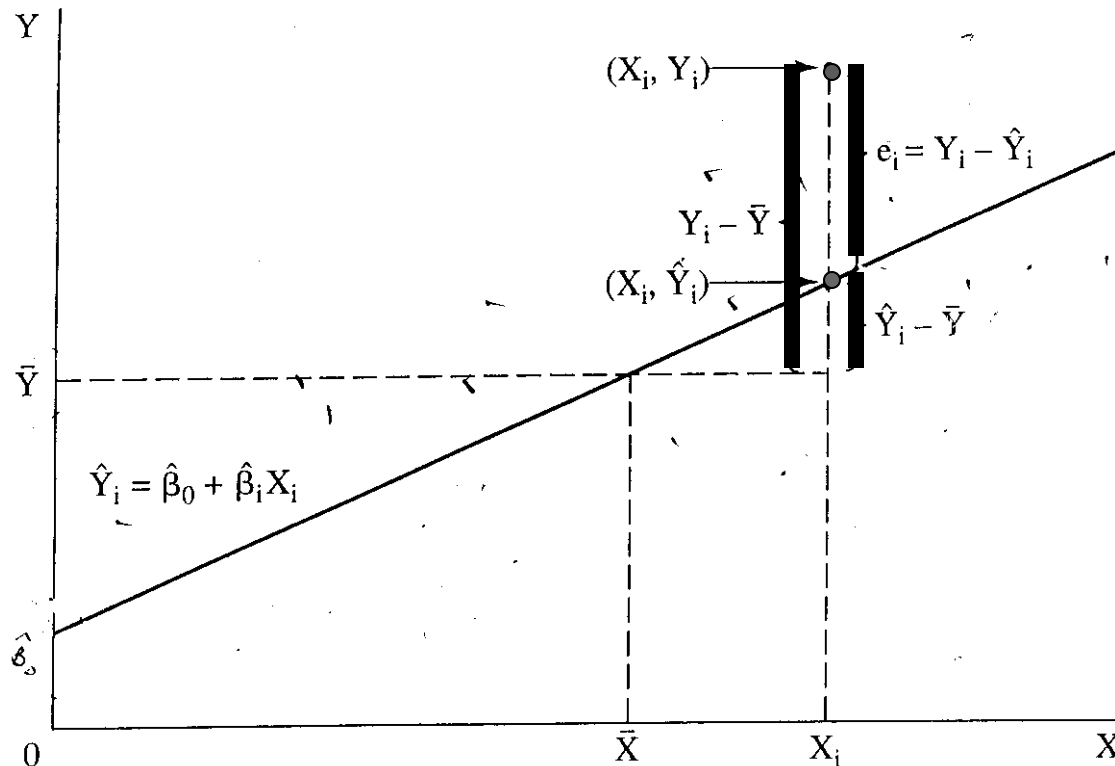
- Population

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \varepsilon$$

- Sample

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

# What is the regression line doing? (simple example)



**Figure 2.1 DECOMPOSITION OF THE VARIANCE IN Y**

The variation of  $Y$  around its mean ( $Y_i - \bar{Y}$ ) can be decomposed into two parts: 1)  $(\hat{Y}_i - \bar{Y})$ , the difference between the estimated value of  $Y$  ( $\hat{Y}$ ) and the mean value of  $Y$  ( $\bar{Y}$ ); and 2)  $(Y_i - \hat{Y}_i)$ , the difference between the actual value of  $Y$  and the estimated value of  $Y$ .

# Goodness of Fit

$$\text{TSS} = \text{ESS} + \text{RSS}$$

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n e_i^2$$

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

# Goodness of Fit

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

$$\bar{R}^2 = 1 - \frac{\left(\sum_{i=1}^n e_i^2\right) / (n - (k + 1))}{\left(\sum_{i=1}^n (Y - \bar{Y})^2\right) / (n - 1)} = R^2 - (1 - R^2) \left(\frac{k}{n - k - 1}\right)$$

# What causes homelessness?

# SPSS Regression Results

Variables Entered/Removed<sup>b</sup>

Model	Variables Entered	Variables Removed	Method
1	Distance <sup>a</sup>	.	Enter

a. All requested variables entered.

b. Dependent Variable: Time

$R^2$  Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.735 <sup>a</sup>	.540	.529	15.840

a. Predictors: (Constant), Distance

ESS

ANOVA<sup>b</sup>

RSS

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	12655.155	1	12655.155	50.440	.000 <sup>a</sup>
	Residual	10788.490	43	250.895		
	Total	23443.644	44			

TSS

a. Predictors: (Constant), Distance

b. Dependent Variable: Time

Coefficients<sup>a</sup>

$\hat{\beta}_0$

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	18.589	2.881		6.452	.000
	Distance	.959	.135	.735	7.102	.000

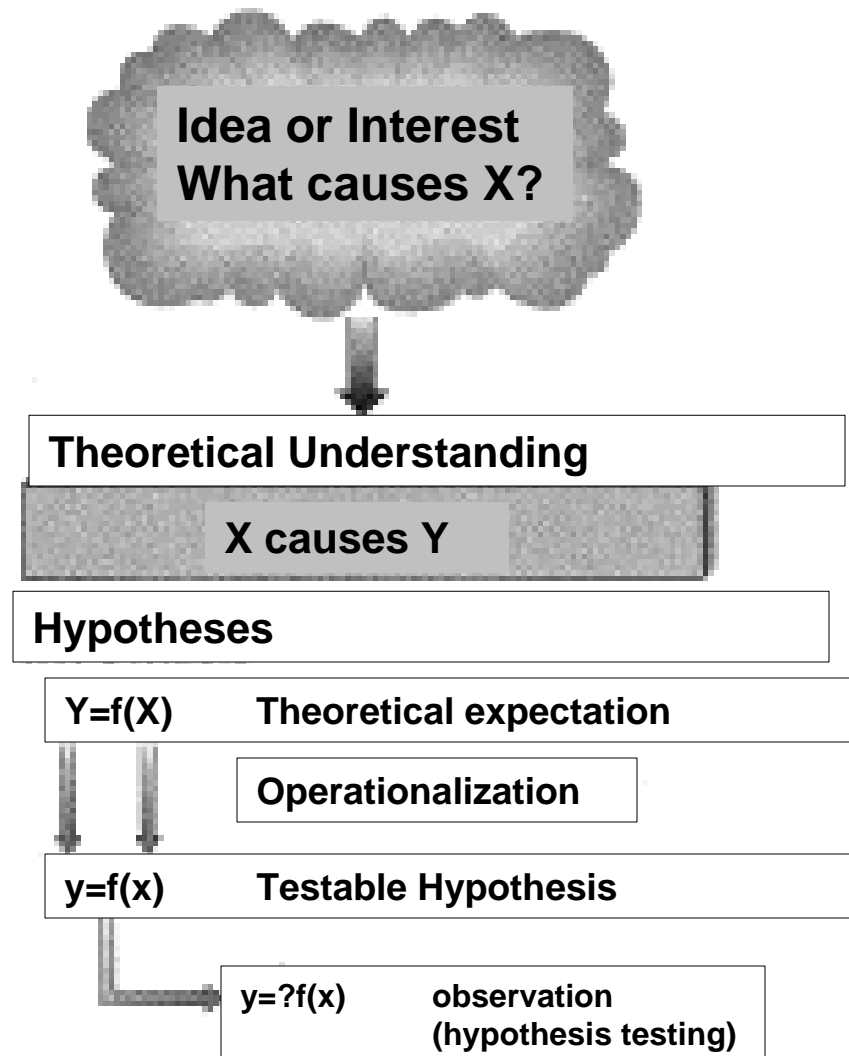
P-value

$\hat{\beta}_1$

a. Dependent Variable: Time

# Theory and Hypotheses

(teaching notes p. 15)



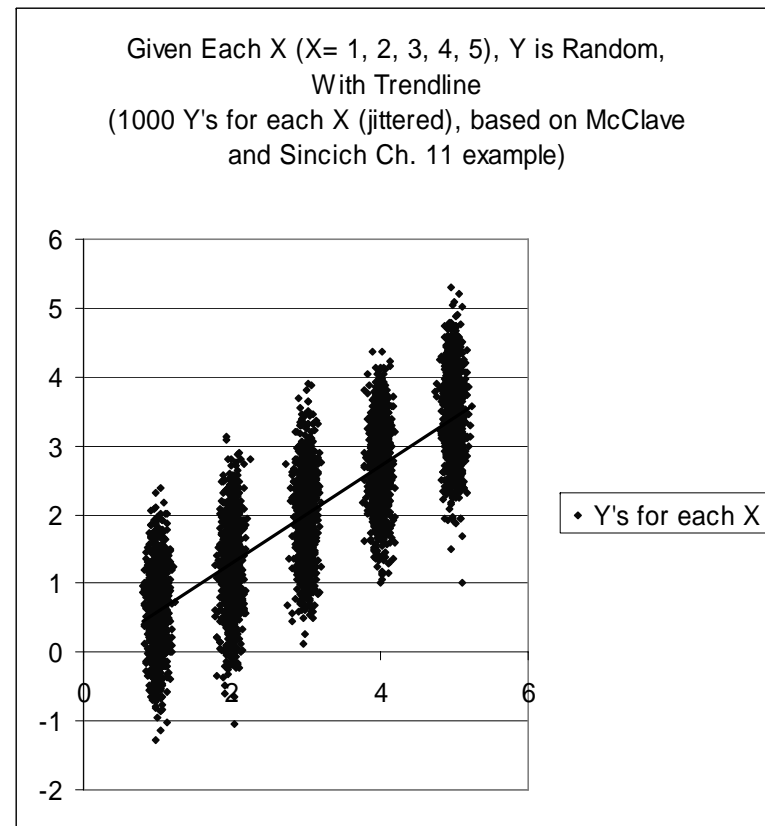
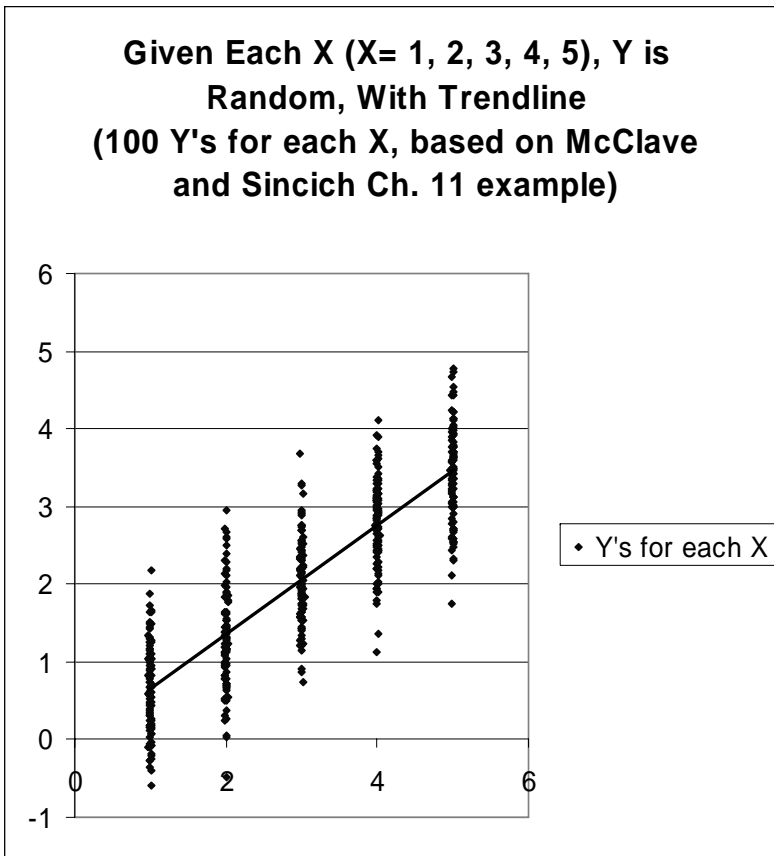
# Sampling distribution of parameter estimates

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

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

- t-tests, confidence intervals, and p-values work for coefficient estimates.

# Variation in Multiple Samples



# Hypothesis test of one coefficient

- Set up hypothesis  $H_0: \beta_1=0$   $H_a: \beta_1 \neq 0$
- Set decision rule  
If  $|t| > t_{\alpha}$  then reject  $H_0$ ;  $t_{\alpha}$  with  $n-k-1$  df
- Find test (or t) statistic

$$t = \frac{\hat{\beta}_1 - \beta_{H_0}}{SE_{\hat{\beta}_1}}$$

- Compare critical t and test statistic

# Problems with tests

- Type I Error
  - Reject the null when the null is true.
  - The p-value is the chance of making this type of error
- Type II Error
  - We fail to reject the null when the null is false.
  - Chances decrease w/
    - larger sample sizes
    - larger standard errors of the parameter estimate
    - the size of the true parameter.

# Confidence Interval for parameter estimate

- A  $(1-\alpha)\%$  confidence interval for the true coefficient

$$\hat{\beta}_1 \pm t_{1-\alpha/2, df} SE_{\hat{\beta}_1}$$

use t at  $n-k-1$  degrees of freedom

# P-values

- How likely is it that we'd get a coefficient of that far away from 0?
  - Probability that you would get an estimate so far (in SEs) from  $\beta_{H_0}$  if  $H_0$  were true.
  - p-values give the level of support for  $H_0$
  - you can look up t-statistic in t or z table (or use Excel) to find the probability in the tail(s).

# Example: Hypothesis Testing

- (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

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

(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  $|t| > t_c$

# Confidence Interval for parameter estimate

- A  $(1-\alpha)\%$  confidence interval for the true coefficient

$$\hat{\beta}_1 \pm t_{\alpha, df} SE_{\hat{\beta}_1}$$

- Example from Quigley/Tucker

- 95% CI for rent control.  $.42 \pm (2.02).105$   
 $= .42 \pm 0.2121$   
 $= [.21, .63]$

# P-values

- How likely is it that we'd get a coefficient of .42?

$P[|t| > 3.99 \mid \beta_1 = 0] < .01$  (for  $df=46$ , 2-sided from table)

- Therefore, we would expect a t statistic this big (an estimate this far from our hypothesized one) less than 1 percent of the time, if the null hypothesis were true.

# 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

# Recap Today

- What about other causal factors?
  - 2 Examples
    - Rent Control and Homelessness
    - Salk Polio Vaccine
- Experiments and Quasi-Experiments
- Multivariate Regression
  - Goodness of fit
  - hypothesis testing
  - Interpretation
  - Impact of omitted variables (bias)
  - Classical Regression Assumptions
  - Reporting Requirements

# Next Time

- More Interpreting Results
- More Hypothesis Testing
- Prediction
  
- The Hunt for Data