Conclusion

December 7, 2022

Today

- Course Wrap-up and focus on evaluating research
- Brief discussion of logistic regression
  - Not expected to know logit, but should know basic interpretation (and how it differs from OLS) if you see it
- Evaluating research
  - 8 questions to ask to evaluate hypothesis testing
- Discussion of final exam

Regression with non-interval DVs

- OLS Regression is generally advisable only with an interval DV
  - Although, the linear probability model (LPM) can be used
- Can use regression with non-interval DV, but it’s not linear
- Dichotomous DV: Logit or Probit
  - Ex: civil war vs. no civil war
  - Ex: win before Supreme Court vs. loss before Court
- Nominal DV: Multinomial Logit or Multinomial Probit
  - Ex: violence, non-violent direct action, conventional politics, nothing
  - Ex: liberal decision, conservative decision, neither, holdover to next term
- Ordinal DV: Ordered Logit or Ordered Probit
  - Ex: no conflict, minor internal armed conflict, civil war, large-scale civil war
  - Ex: liberal decision, mixed decision, conservative decision

Logistic Regression

- Basic logic of OLS regression: examine how changes in *X* affect the (expected) mean value of *Y*
- Basic logic of logit: examine how changes in *X* affect probability of observing a “1” on a dichotomous variable
  - Ex—impact of an individual’s income on probability he/she will vote
  - Ex—impact of a country’s GDP (per capita) on the probability it will experience a civil war
- Key difference between OLS regression and logistic regression:
  - OLS generates predicted values of the DV
    - Predicted vote share for President Biden in a state based on the number of college graduates
  - Logistic regression generates predicted probabilities:
    - Probability an individual will vote (or not) based on her/his income, etc.

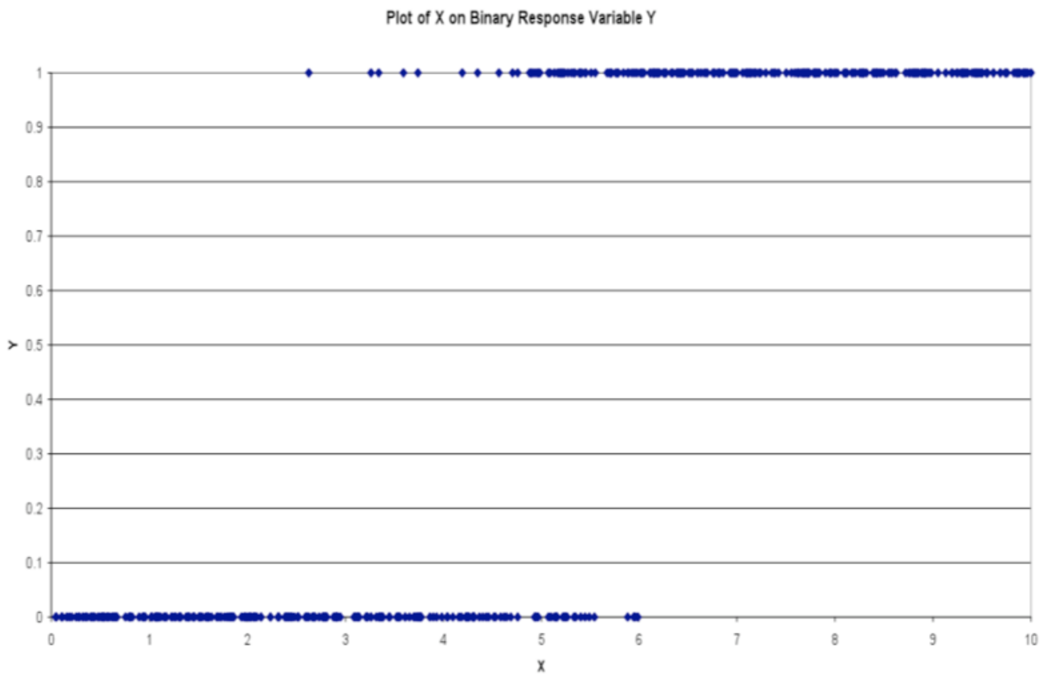
Maximum Likelihood Estimation Intuition

- Steps & Intuition
  - Start by thinking about the process that generated the data—that is, how was our dependent variable generated?
  - Then, we specify the probability distribution that describes the dependent variable
  - Our regression estimates are the values that maximize the probability (i.e., likelihood) that we observe the dependent variable, *Y*, that we actually observed
- Full details in GVPT 729a

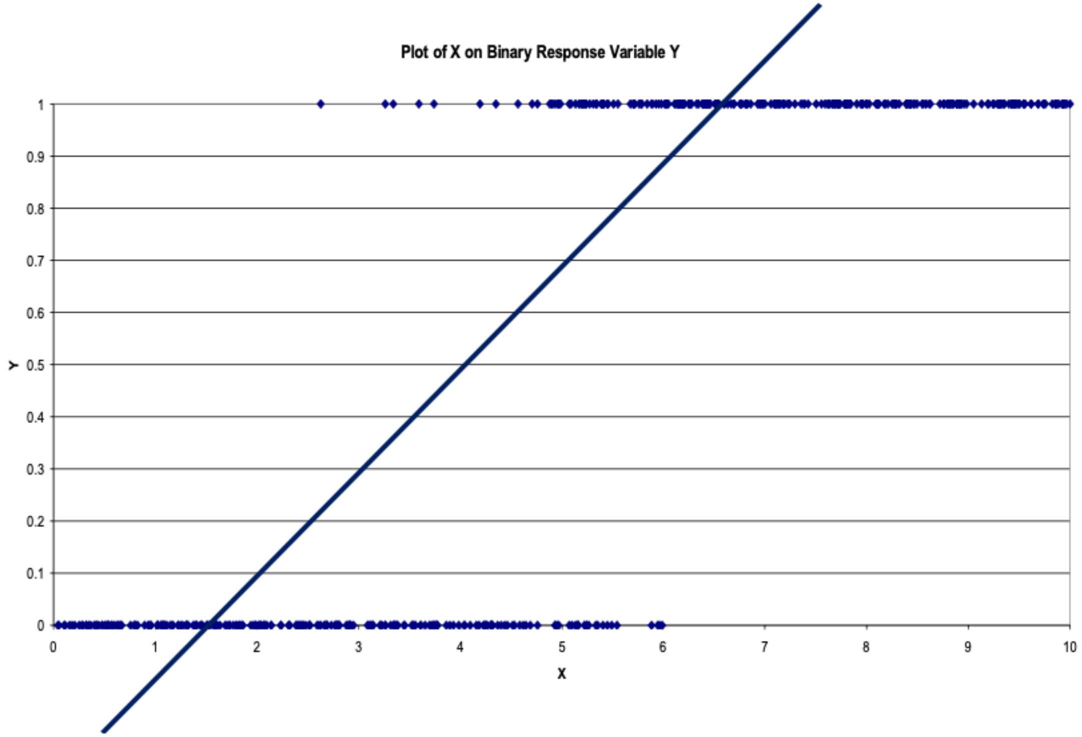
Empirical Set-up

- For convenience, code the dependent variable to “1” (success) and “0” (failure)
  - This is generally arbitrary & we set the outcome of greatest interest to “1”
- Example—individual turnout (i.e., did a survey respondent vote?)
  - 1=reported voting
  - 0=reported not voting
- What might the data look like?

Binary Response

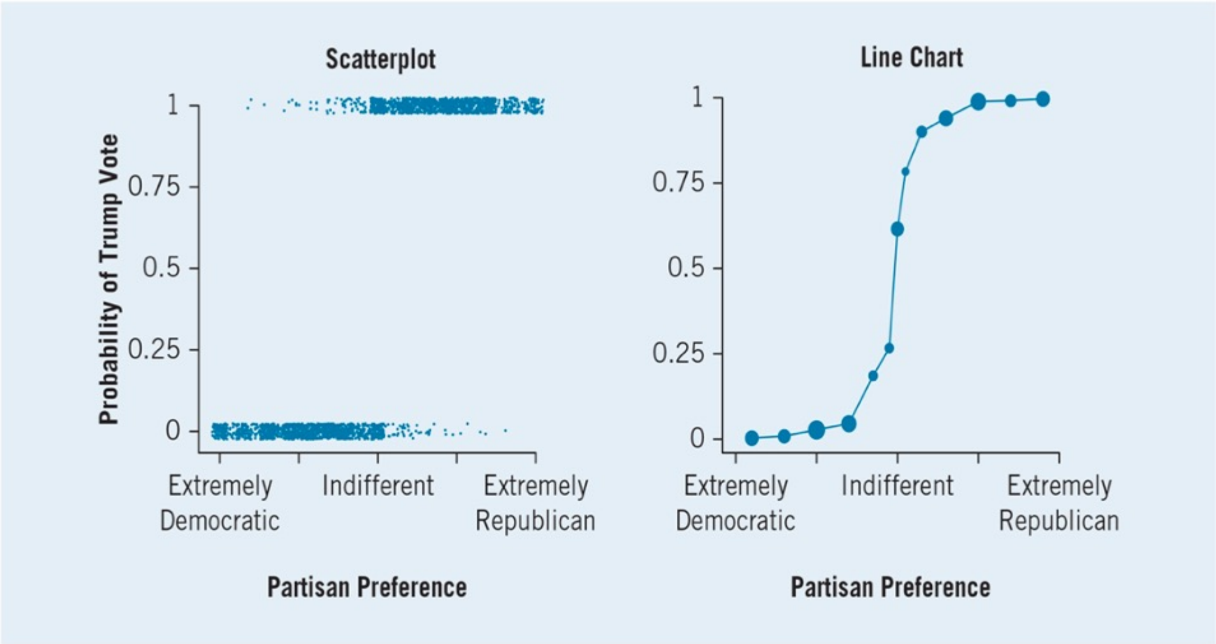


Binary Response



Binary Response

Figure 9-1 Observed Probabilities of Trump Vote by Partisan Preference



Note: The size of the line chart points is proportional to the number of observations in each group.

Challenges with the Basic Linear Probability Model

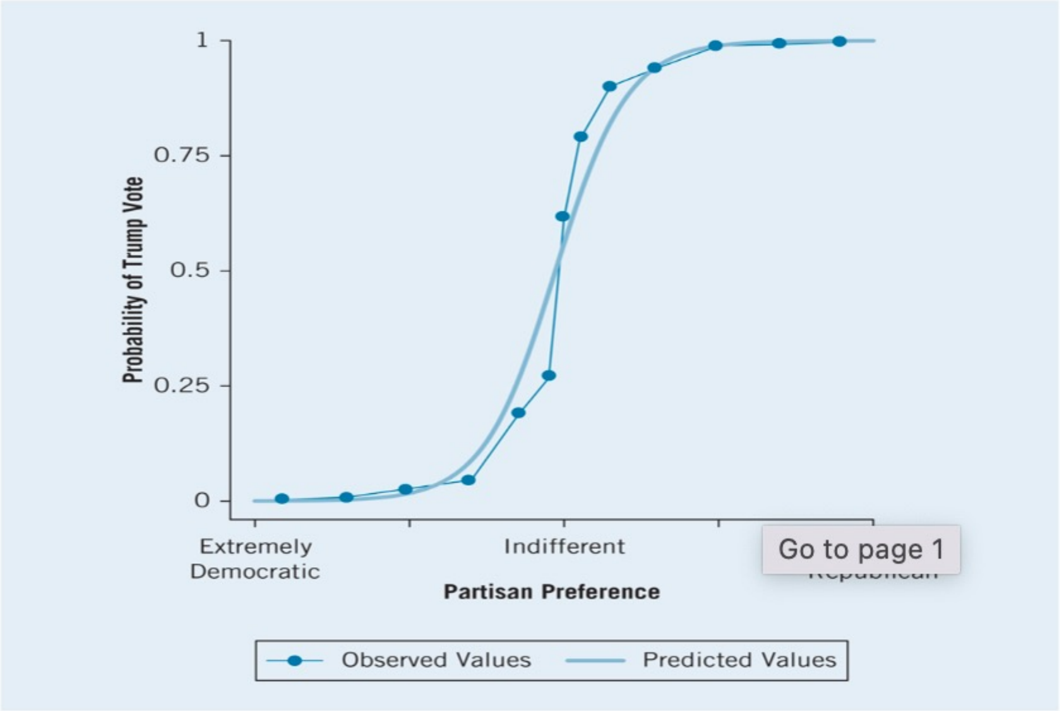
- OLS may produce nonsensical predictions—recall that OLS assumes continuous interval-level data
  - Predictions that exceed the bounds of 0 and 1
- Disturbances are not normally distributed—they follow the Bernoulli distribution
  - More problematic with smaller samples—i.e., hypothesis testing & inference
- Y is distributed with a mean of P_i and variances of P_i(1- P_i) and P_i is a function of the *X*'s—thus we have heteroscedasticity with variance related to the *X*'s
  - Although, we can adjust the standard errors accordingly, or use WLS
- Linearity assumption of OLS is suspect
  - A potentially serious problem—i.e. functional form issues
- Goodness of fit measures are even less helpful

Logistic Regression

- Linear regression assumes that a one-unit change in *X* has the same effect on *Y* across all values of *X*
  - But, this is not generally true for probabilities
  - Ex—a country going from $1,000 to $2,000 GDPpc reduces its risk of civil war more than a country going from $20,000 to $21,000 GDPpc
    - Logic—the probability of civil war at $20,000 (per capita) is already low, hard to decrease much more

Binary Response

Figure 9-2 Predicted Probabilities of Trump Vote by Partisan Preference



An Important Point

- The model is inherently interactive in all of the variables
  - This is critical-–the effect of a one-unit change in X is **NOT CONSTANT**—it depends on where we are on the curve
  - Compare this to OLS—the effect of β_k is constant across the entire range of the predictor

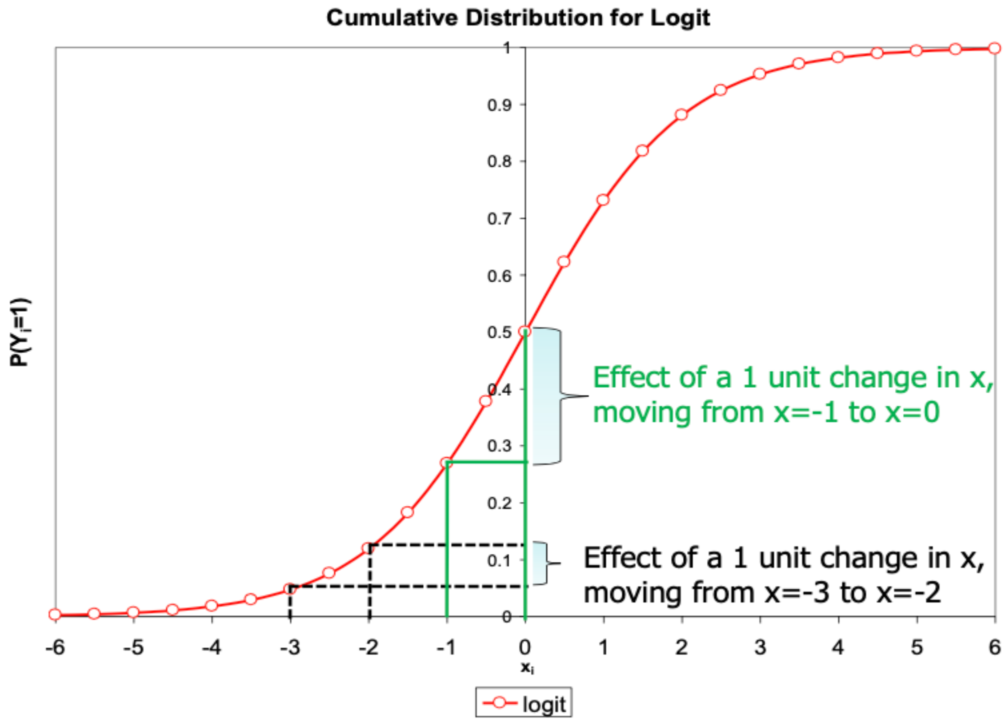
- The output tells us (assuming no interaction terms)
  - The sign of the effect of X_k on the probability of success
  - Overall statistical significance of the effect

- But, we do NOT directly interpret the coefficients as we did with OLS
  - We need to transform them in order to determine substantive significance

Intuition Behind These Nonlinear Effects

- Note in the following CDF for logit that:
  - A one-unit shift in X when P_i is close to 0 results in a small change
  - The effect gets larger as we move closer to $P_i = 0.5$
  - The effect then tapers as we get closer to $P_i = 1$
- Thus, the initial probability of success plays an important role in the determination of the effect of changes in the independent variables
- Back to turnout example:
  - For those almost sure not to vote, mobilization will do little
  - For those close to 50-50, mobilization will have a large effect
  - For those almost sure to vote, mobilization will do little

Cumulative Distribution Function for Logit



Probability and Odds

- Logistic regression uses odds, which are a function of probabilities
  - Odds=probability of something happening/probability of it not happening
    - Probability of someone voting: .8
    - Probability of someone not voting: .2
    - Odds: 4 (i.e., 4 to 1 odds that the person will vote)
- Logistic regression makes predictions about the odds of observing the DV at different values of the IV
- Interested in the odds ratio—that is, the odds of observing the DV at one value of the IV compared to the next value of the IV
  - Ex. odds of a country with a GDPpc of $5,000 having a civil war compared to a country with a GDPpc of $4,000

Odds Ratios

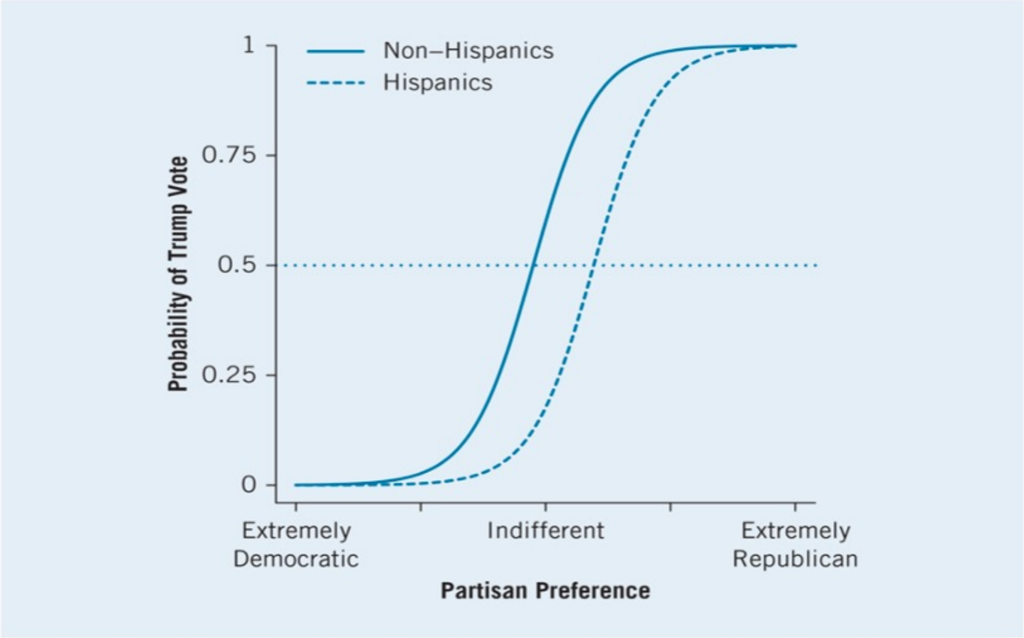
- The difference between OLS regression and logistic regression:
  - OLS regression—changes in the value of the IV have a consistent effect on the values of the DV
  - Logistic regression—changes in the value of the IV have a consistent effect on the odds ratio of the DV
- The logistic regression equation is:
  - Log odds(Y)= β_0 + β_1(X_1)
- We can use the logistic regression equation to make predictions about the probability of observing the DV based on specific values of the IV
  - Find the odds (by taking the exponent of the log-odds) to determine the probability
    - $e^{logOdds} = \text{Odds}$
    - Probability=$\frac{Odds}{1+Odds}$

Logistic Regression with Multiple I.V.s

- Just like OLS, Logistic regression is easily extended to use multiple IVs
  - Log odds(Y)= $\beta_0 + \beta_1(X_1) + \beta_2(X_2) + \beta_3(X_3) +\dots.\ \beta_k(X_\mathrm{k})$
- In multiple logistic regression, the predicted log odds are a function of the values of all of the IVs
- The coefficient on each IV tells us the impact of a change in that variable on the log odds of the DV, controlling for the other variables
- Like OLS, separate IVs are assumed to be additive (even though the functional form is inherently non-linear)
- If effect is interactive, need to include a multiplicative term

Binary Response

Figure 9-4 Predicted Probabilities at Representative Values of an Independent Variable



Source: 2016 American National Election Study

Inference and Logistic Regression

- Statistical inference in logit is very similar to OLS
  - Null hypothesis: $\beta_1 = 0$
  - Each coefficient has a standard error
  - Can use standard error to find the p-value
- In logit, we can use the normal distribution, so not affected by degrees of freedom
  - So, we find z-statistics (and not t-statistics)
- Can also use the standard error to estimate a confidence interval for β_1

Logistic Regression Output from R

```
Call:
glm(formula = own_dog ~ prochoice_scale + dem_age6 + obama_therm,
    family = "binomial", data = nes)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5525  -1.0352  -0.8837   1.2476   1.7858

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       0.730192   0.181920   4.014 5.97e-05 ***
prochoice_scale  -0.005875   0.003192  -1.841  0.06568 .
dem_age630-39     0.125267   0.154726   0.810  0.41816
dem_age640-49     0.224402   0.160981   1.394  0.16333
dem_age650-59     0.035993   0.156161   0.230  0.81771
dem_age660-69    -0.202141   0.176782  -1.143  0.25285
dem_age670-older -0.613725   0.212812  -2.884  0.00393 **
obama_therm      -0.011200   0.001719  -6.516 7.21e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Evaluating Research

- We have focused a lot on hypothesis testing
- Much of political science designed to test hypotheses
- Let’s discuss several big-picture questions to ask about hypothesis testing when evaluating research
- First steps:
    - Is the theory logically consistent?
    - Do hypotheses flow logically from the theory?
- Several other questions to ask…

Hypothesis Testing

(1) Do the variables properly measure the concepts?

- Remember—a hypothesis is prediction about the relationship between concepts
- Hypothesis testing will evaluate the relationship between measures
  - If measures and concepts are not highly correlated, results of analysis don't tell you much
- Want to ask—is a measure reliable? Is it valid?
- Also important to think about level of measurement—is it appropriate?

Hypothesis Testing

(2) Is the sample unbiased?

- Hypothesis predicts relationship between variables in population
- Generally will only observe a sample, not the population
- Are some members of the population systematically more likely to end up in sample?
- Related question—what can you generalize to?
    - Ex—relationship between democracy and war?
        - Should it apply during 19th century?
        - In the future?

Hypothesis Testing

(3) Does study control for other factors that could affect relationship between IV and DV?

- Remember, test of hypothesis will examine correlation between IV and DV
- In non-experimental design consider what determines the distribution of values of the IV
- If factors that affect distribution of the IV also affect distribution of the DV, can lead to omitted variable bias
  - Ex—Ice cream sales and drownings

Hypothesis Testing

(4) Are the statistical techniques appropriate?

- Remember—appropriate techniques a function of measurement of variables
  - Ex—OLS is only appropriate if DV is interval (IV can be interval or dichotomous)
  - Ex—Difference of proportions only appropriate if DV is dichotomous, IV is dichotomous

- Also want to consider distribution of variable
  - Ex—mean is less useful if variable is skewed
  - Regression focused on mean—if mean is not useful, regression may not be appropriate

Hypothesis Testing

(5) Are the assumptions underlying the statistical models met?

- All statistical tests have assumptions, important to consider whether they are violated and the potential consequences
- Ex—linear regression assumes a linear relationship between the IV and DV
  - If not, linear regression is not appropriate
- Ex—regression assumes errors are uncorrelated
  - May not be true, if not, need to do something about it

Hypothesis Testing

(6) Is the interpretation of the analysis correct?

- Are there basic mistakes?
  - This can happen surprisingly frequently
- Interpretation can be complicated
  - Ex-Interaction terms in non-linear models (such as logit)

Hypothesis Testing

(7) How big (and substantively meaningful) is the effect?

- Many studies focus excessively on statistical significance
- Important—mostly interested in substantive effect
  - Ex—getting a Ph.D. increases average income by $1000 (p=0.001)
  - Ex—for every additional year of schooling, average incoming increases by $50000 (p=0.001)
- Some factors that are statistically significant have little substantive effect/importance

Hypothesis Testing

(8) Would the results hold up to replication?

- Remember, 95% significance level means 5% chance we will reject the null hypothesis when it is actually correct
  - True even if sample is representative of population, concepts perfectly measured, etc.
- Important to replicate studies—chances of committing Type I error twice are dramatically lower

Final Exam

- Final exam posted online by Wednesday December 14
- Due by Monday, December 19th, 12:00 noon
- Same format as first two exams
- Not cumulative, focused only on material since 2nd midterm
- Questions?