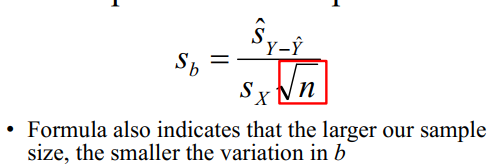
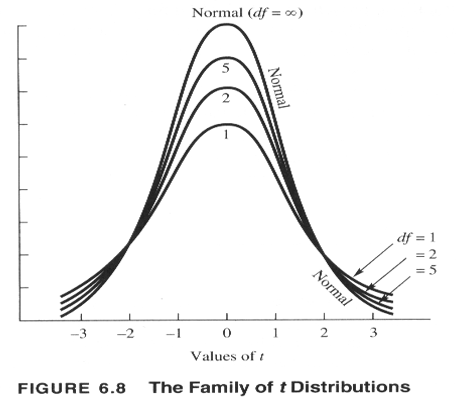
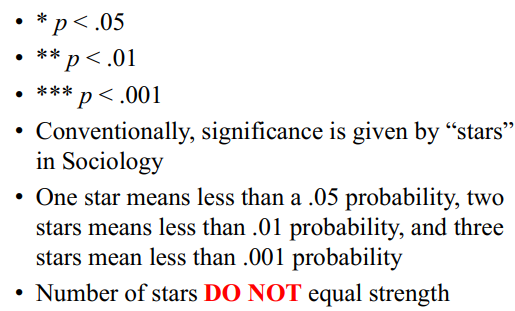
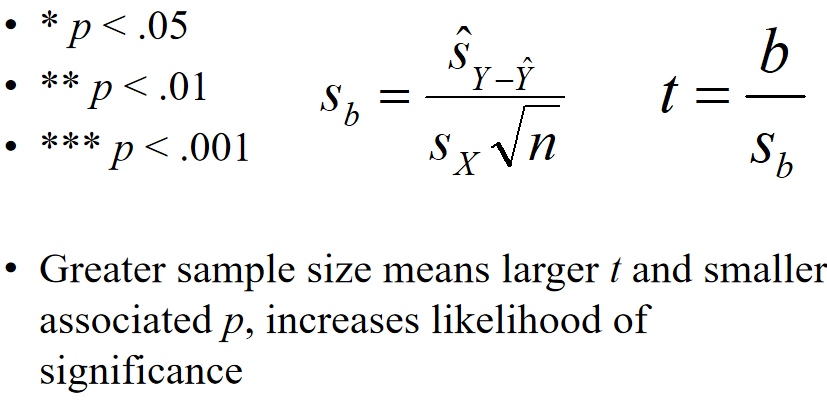
Assignment 4 – Barbara Cernosa 30279243

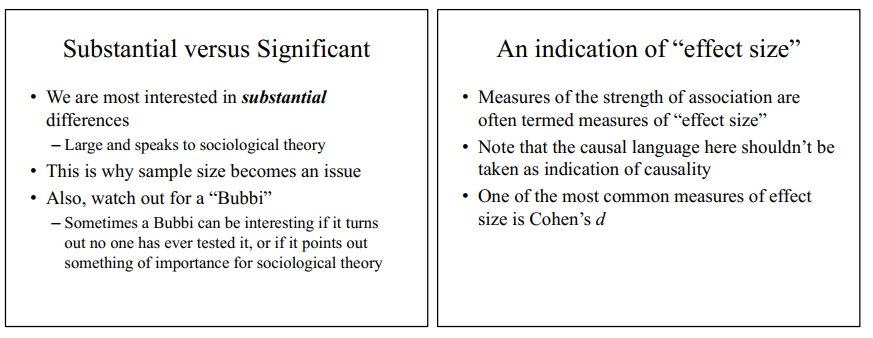
1. No, the student is not correct. Significance depends on sample size—larger samples (like 10,000) are more likely to detect smaller effects as significant, while smaller samples (like 1,000) require larger effects to reach significance. To compare predictor strength, we should examine effect sizes (standardized coefficients or R-squared values), not just p-values.



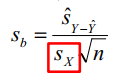
The higher the n, the lower the s (for example,1/sqrt1000 (0.03) vs 1/sqrt10000(0.01).



The p-value reflects statistical significance, which is heavily influenced by sample size—larger samples (10,000) increase the t-value and reduce the p-value, making significance more likely even for a smaller effect. The difference likely reflects sample size power rather than true predictive strength. Remembering from lecture 1:

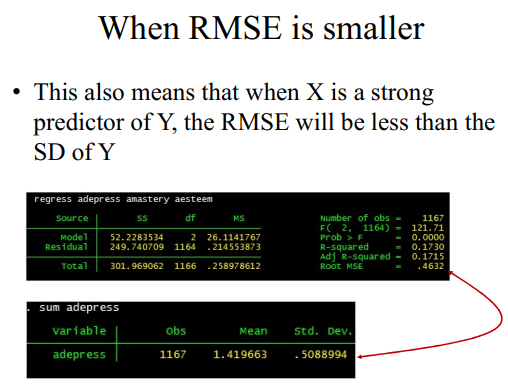
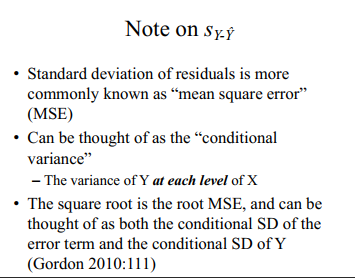
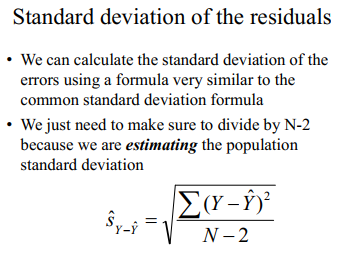


2. Student A will likely have a smaller standard error for b. Greater diversity in X values increases the denominator in the formula for the standard error of the slope.



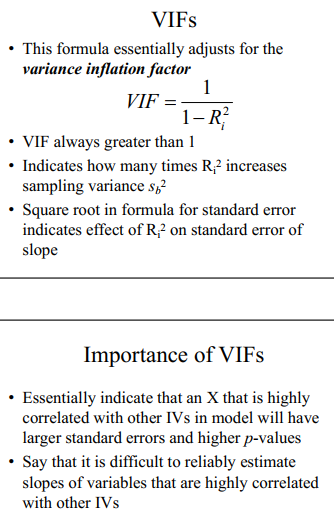
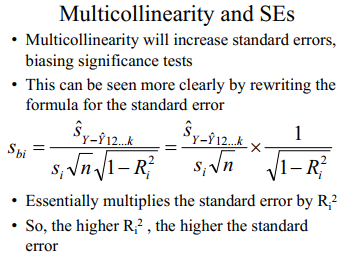
Since both students have the same sample size and similar error variance, the difference lies in Sx , the standard deviation of X. Student A, with a diverse set of values for X, will have a larger Sx , while Student B, with only a few similar values, will have a smaller Sx . A larger Sx in the denominator reduces Sb , meaning a smaller standard error for Student A.

3. The variance of Y will be larger than the MSE of the OLS regression model. The variance of Y is captured in the SST, where SST is. When using a very accurate predictor, X being a strong predictor of Y reduces the RMSE (Root Mean Squared Error, which is the square root of MSE). We should also remember SST = SSR + SSE, and when SSR is not 0, SST > SSR even with a small sample size.

.

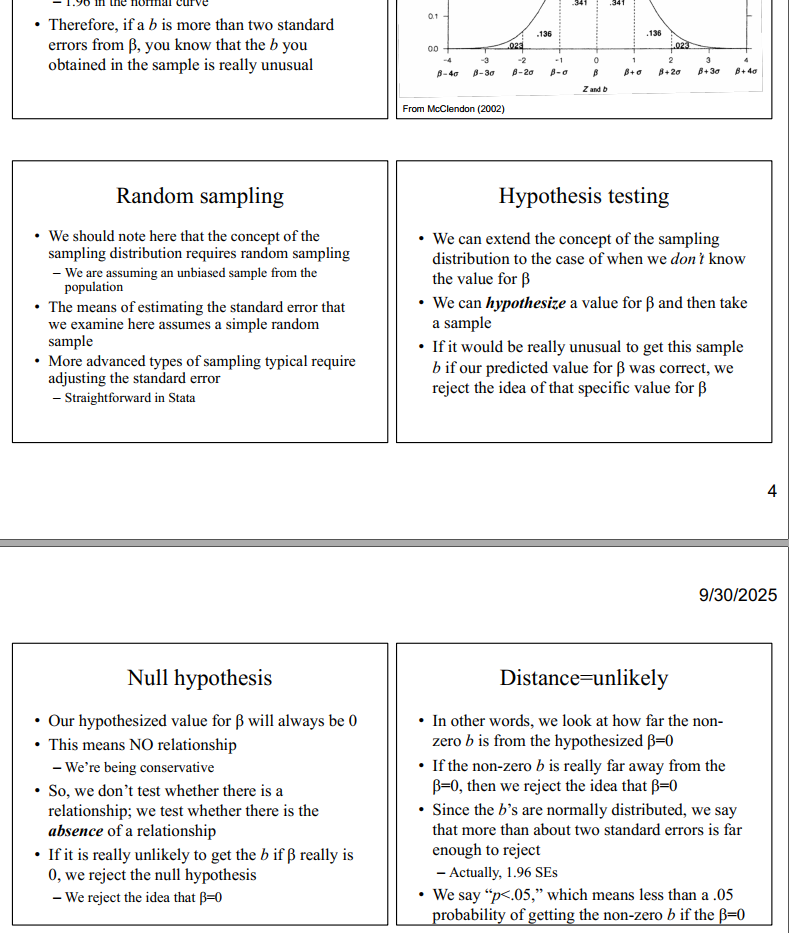
4. One advantage is that multiple regression provides the total explained variance (R-squared) for both predictors combined, and it allows for a better assessment of individual contributions to predicting Y while controlling for the other variable, they each may still have a unique effect on Y that could be masked or confounded if analyzed separately. With multiple regression, we can detect spuriousness, redundancy and suppression.

5. This occurs due to multicollinearity: the high correlation between the X variables increases their standard errors in multiple regression. In a multiple regression, the model attempts to isolate the unique contribution of each X to Y, but when X1 and X2 are highly correlated, their individual effects become difficult to distinguish. This increases the standard errors of the regression coefficients, making the predictors appear non-significant, even though together they may explain a substantial portion of Y's variance (as indicated by a potentially highR2). One way to check is to calculate the variance inflation factor (VIF) for each predictor and a high VIF (>5 or 10) indicates multicollinearity.

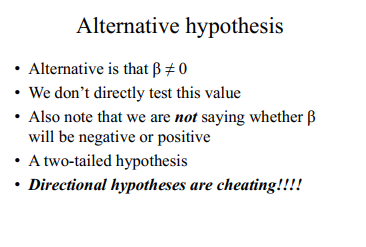


6. (skipped)

7. No, Dr. Vishnik is not correct. The 68% figure appears to represent a p-value of 0.68, which is the probability of observing the sample b (or more distant) assuming the null hypothesis (β = 0) is true. Whether Dr. Vishnik’s conclusion holds depends on the chosen significance level (α). The 68% chance of obtaining the sample b if β=0 reflects the probability under the null hypothesis, which relates to the p-value or confidence interval in hypothesis testing. The decision to reject or fail to reject the null hypothesis (H0 :β=0) is based on comparing this p-value toα.



8. No, Dr. Bierman would not agree, he has mentioned us only using two-tailed test on several occasions, especially in scholarly presentations. Also, difference between statistical and substantive significance. With a large sample size like 1,000, even trivial effects can achieve p < .05 due to increased statistical power, which makes it easier to detect small differences as "significant" without them speaking to broader sociological theory or real-world importance. Dr. Bierman would argue that true scholarly value lies in substantial differences (large effect sizes like Cohen's d > 0.5 or Hedge’s g or meaningful R² values), not just p-values.

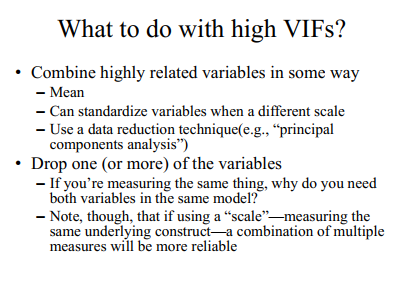


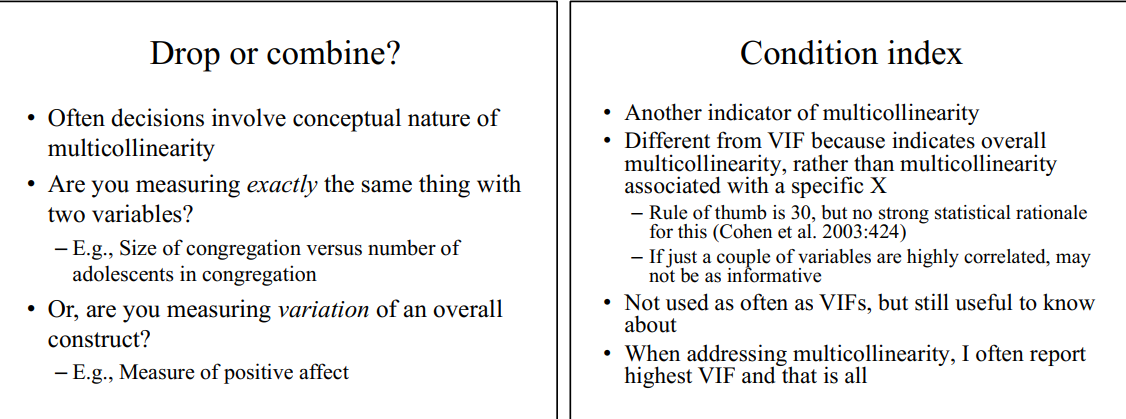
9. Two strategies:

- Combine highly related variables in some way (mean, standardization using a different scale, using a data reduction technique (PCA),

- Drop one (or more) of the variables.

The choice is based on several factors, such as the conceptual overlap between the predictors, the research goal (what do we want to communicate or show), and diagnostic tools like the VIF and the condition index.



10. Syntax

\*create and recode variables\*

\*tabulation of demands, question Q15D in the HRS\*

tab PLB015D

\*tabulation without labels\*

tab PLB015D, nol

\*creating new variable\*

gen deman=PLB015D

\*recode\*

recode deman (1=4) (2=3) (3=2) (4=1) (else=.)

\*compare\*

tab1 PLB015D deman

\*tabulation of critique, question Q15E in the HRS\*

tab PLB015E

\*tabulation without labels\*

tab PLB015E, nol

\*creating new variable\*

gen crit=PLB015E

\*recode\*

recode crit (1=4) (2=3) (3=2) (4=1) (else=.)

\*compare\*

tab1 PLB015E crit

\*tabulation of down, question Q15F in the HRS\*

tab PLB015F

\*tabulation without labels\*

tab PLB015F, nol

\*creating new variable\*

gen down=PLB015F

\*recode\*

recode down (1=4) (2=3) (3=2) (4=1) (else=.)

\*compare\*

tab1 PLB015F down

\*tabulation of nerves, question Q15G in the HRS\*

tab PLB015G

\*tabulation without labels\*

tab PLB015G, nol

\*creating new variable\*

gen nerve=PLB015G

\*recode\*

recode nerve (1=4) (2=3) (3=2) (4=1) (else=.)

\*compare\*

tab1 PLB015G nerve

\*calculates numbers of questions missing for negints\*

egen missing=rowmiss(deman crit down nerve)

\*labels variable overall\*

label variable missing "Number of missing for negints"

\*tabulation\*

tab missing

\*calculates mean of the negints if missing less than 1 of the items\*

egen negints=rmean(deman crit down nerve) if missing<=1

\*labels variable overall\*

label variable negints "Negative Interactions"

\*tabulation\*

tab negints

\*mean and sd\*

sum negints

A screenshot of a data sheet

AI-generated content may be incorrect.

11. Syntax

\*create and recode variables\*

\*tabulation of anyfriends, question Q14 in the HRS\*

tab PLB014

\*tabulation without labels\*

tab PLB014, nol

\*creating new variable\*

gen anyfriends=PLB014

\*recode\*

recode anyfriends (1=1) (5=0) (else=.)

\*labels variable overall\*

label variable anyfriends "Do you have any friends"

\*creates labels\*

label define anyfriendslab 0 "No friends" 1 "Friends"

\*labels categories\*

label values anyfriends anyfriendslab

\*compare\*

tab1 PLB014 anyfriends

A screenshot of a computer

AI-generated content may be incorrect.

How many people in the sample indicated that they did not have any friends: 638

12. Syntax

\*tabulation of number of negative interactions, including those without friends\*

\*tabulation of number of negative interactions\*

tab negints

\*tabulation without labels\*

tab negints, nol

\*creates label for those without friends\*

gen negfriends=negints

\*replaces response with 1, if they have 0 friends\*

replace negfriends=1 if anyfriends==0

\*labels variable overall\*

label variable negfriends "Negative Interactions, 1 if no friends"

\*compare\*

tab1 negints negfriends

\*tabulate negfriends\*

tab negfriends

A screenshot of a graph

AI-generated content may be incorrect.

13. Syntax

\*crosstab of negfriends and anyfriends\*

tab negfriends anyfriends, m

A white sheet with numbers and black text

AI-generated content may be incorrect.



*Explain how you can tell that your imputation for respondents who did not have any friends was carried out correctly:* The number of respondents who had “No friends” (anyfriends) now only has values of 1 under “Negative interactions” (variable negint), 638 respondents.



*Indicate how many individuals missing on the measure of whether a respondent has any friends are included in the negative interactions variable.* 122

14. Conceptual diagram:

negative interactions with friends ---------+--------à shame

15. Syntax (weird paste)

\*create and recode shame variable tabulation of shame, question Q26O in the HRS\*

tab PLB026O

\*tabulation without labels\*

tab PLB026O, nol

\*creating new variable \*

gen shame=PLB026O

\*recode\*

recode shame (1=5) (2=4) (3=3) (4=2) (5=1) (else=.)

\*labels variable overall \*

label variable shame "Ashamed"

\*compare\*

tab1 PLB026O shame

\*create and recode education variable\*

\*tabulation of education, raedyrs in the HRS\*

tab raedyrs

\*tabulation without labels\*

tab raedyrs, nol

\*creating new variable\*

gen education=raedyrs if raedyrs>=0 & raedyrs<20

\*labels variable overall\*

label variable education "Education"

\*compare\*

tab1 raedyrs education

\*create sss var\*

\*tabulation, question Q36 in the HRS\*

tab PLB036

\*tabulation without labels\*

tab PLB036, nol

\*creating new variable\*

gen sss=PLB036 if PLB036>0 & PLB036<11

\*labels variable overall\*

label variable sss "SSS"

\*tabulation before and after recoding\*

tab1 PLB036 sss

\*create age var\*

\*tabulation, PAGE in the HRS\*

tab PAGE

\*tabulation without labels\*

tab PAGE, nol

\*creating new variable\*

gen age=PAGE if PAGE>49 & PAGE<99

\*labels variable overall\*

label variable age "Age"

\*tabulation before and after recoding\*

tab1 PAGE age

\*creates flag\*

gen flag=0

replace flag=1 if shame!=. & education!=. & negfriends!=. & sss!=. & age!=.

\*tabulate flag\*

tab flag

\*regression for the association between shame and negfriends\*

\*controls for sss age and education\*

regress shame negfriends age education if flag==1, beta

A table with numbers and text

AI-generated content may be incorrect.

5324 respondents will be retained.

16. Syntax

\*regression for the association between shame and negfriends\*

\*controls for age and education\*

regress shame negfriends age education if flag==1, beta

A screenshot of a computer

AI-generated content may be incorrect.

It is statistically significant P>|t| = 0.000. Negative interactions with friends (negfriends) are associated with more shame, as the coefficient is positive (beta = 0.1609). According Brydge’s criterion the association is between small and medium. A positive coefficient means that higher negfriends values (more negative interactions) predict higher shame scores.

“Brydges (2019) had conducted a survey of gerontological research. In this survey, he also indicates “The 25th (small effect), 50th (medium effect), and 75th (large effect) percentiles corresponded to Pearson’s r values of .12, .20, and .32, respectively” (Brydges 2019:3).”

17. Syntax

\*regression for the association between shame and negfriends\*

\*controls for sss age and education\*

regress shame negfriends age education sss if flag==1, beta

\*obtains vif\*

vif

The VIFs are almost 1, and a value of 1 indicates almost no inflation whatsoever, so we can see little evidence of multicollinearity. To observe VIFs at the commonly accepted threshold of 10 (or closer to 6, practically, according to the professor), the variables would have to be so highly related that they likely measure the same thing. 

