COMPARING STUDY DESIGNS

TYPES OF STUDIES

- **Completely randomized design**: Each subject is randomly assigned to only one treatment* group; we do not have any other information about the subjects when we assign them
  - * Note: In this section, the word "treatment" is used in a generic sense to refer to any level of the explanatory variable; this includes the active treatment(s) and the control condition
- **Matched-pairs design**: Recall there are 2 possibilities…
  - Each subject is assigned to both treatment groups
    - Classic example: anything measured before and after on the same subject
  - Each subject is randomly assigned to only one treatment group; we know other information about the subjects that we use to pair specific subjects together *before* we assign them to a treatment group, for example:
    - Two people with the same age, weight, sex, and race are paired together; one is randomly assigned to treatment, the other to control
    - There is a natural or genetic link between subjects such as comparing a husband to his wife or a father to his son
  - Reduces variability in the response due to the matching variable(s), which makes it easier to determine if there is an effect of treatment
- **Block design**: Each subject is randomly assigned to only one treatment group; we know other information about the subjects that we use to group them together (in groups with more than 2 subjects) *before* we assign them to a treatment group
  - Example: Subjects are grouped based on sex; half of the males and half of the females are randomly assigned to treatment, the other half of each sex receive control
  - Note: the "blocking variable" (sex in the previous example) is a characteristic of the subjects; it cannot be randomly assigned!
  - Reduces variability in the response due to the blocking variable(s), which makes it easier to determine if there is an effect of treatment
- We have also talked about **observational studies**, which are not experiments since the research does not intervene to assign subjects to treatment groups (they simply observe which treatment group the subject selects for themselves)

SELECTED STUDY DESIGN CONSIDERATIONS

- **Random assignment**: Avoids bias due to lurking variables by evenly distributing them between the treatment groups, allowing us to say any difference between the groups at the end of the study is caused by the treatment
  - Can only exist in an experiment, not an observational study
  - As such, observational studies inherently give weaker evidence to support the study conclusion
- **Control group**: Acts as a baseline or reference against which to evaluate the effect of treatment, to determine if the treatment group would have an impact above and beyond what would have happened naturally (whether or not the control group was given a **placebo**)
  - Can exist in an experiment or an observational study; key is that in an observational study, subjects choose to be in the control group
- **Blinding**: Avoids bias due to researcher <u>or</u> subject (depending on who was blinded); **Double-blinding**: avoids bias to researcher <u>and</u> subject
  - Can only exist in an experiment, not an observational study

COMPARING STUDIES

- Questions about comparing studies come down to choosing the one with a stronger design, which will inherently be an experiment instead of an observational study
    - Because observational studies cannot incorporate every element of good design (e.g. randomization, blinding blocking)
    - If choosing between experiments, choose the one that makes the best use of the elements of good design, keeping in mind that there is no perfect study!

PRACTICE PROBLEMS

1. A pharmaceutical company has developed a new flu vaccine. They would like to see if the it protects against more strains of the flu than the current vaccine or a placebo.
    a. For each description below, state the type of experimental design that was used.
        i. They take a sample of 450 subjects and group them based on age (20-29, 30-39,etc). A third of the subjects in each age group is randomly assigned to receive the new vaccine, a third the current vaccine, and a third receive a placebo vaccine.
        ii. They randomly assign 150 subject to receive the new vaccine, 150 subjects to receive the current vaccine, and 150 subjects to receive a placebo vaccine.
        iii. They group subjects together in trios based on age and biological sex. One member of each trio is randomly assigned to each treatment (new vaccine, current vaccine, or placebo).
    b. Which study (i, ii, or iii) would give the strongest evidence of the effectiveness of the new vaccine? Explain.
2. A study examines the effect of using personal response devices (e.g. "clickers") on performance in an introductory physics class.
    a. For each description below, state the type of experimental design that was used.
        i. Before the first semester exam (half-way through the course), the students do not use clickers. Between the first and second exam (at the end of the course), the students do use clickers every class period to answer questions about the content. The difference in exam scores (2nd exam – 1st exam) is analyzed.
        ii. One section of the course is randomly assigned to use clickers while the other section of the course does not. The difference in final exam scores (section 1 – section 2) is analyzed.
        iii. Students are allowed to use clickers if they choose to. The difference in final exam scores (for those students that reported using clickers – those that did not) is analyzed.
    b. Which study (i, ii, or iii) would give the strongest evidence of the effectiveness of clickers on performance? Explain.

HYPOTHESIS TESTING: ALPHA, BETA, POWER

- **Alpha** (α) is the *probability* of making a Type I Error
  - **Type I Error** = rejecting the null when it is true = incorrectly rejecting the null
    = saying/concluding there is an effect when in fact one does not exist
- **Beta** (β) is the *probability* of making a Type II Error
  - **Type II Error** = failing to reject the null when it is false = incorrectly failing to reject the null
    = saying/concluding there is not an effect when in fact one does exist
- **Power** = probability of detecting an effect when one exists **=** probability of correctly rejecting null
  = probability of saying/concluding there is an effect when in fact one does exist = $1 - \beta$
- Notice that α and β are probabilities, whereas the errors themselves are definitions/things that can occur. *Power* by definition is a probability.
- Type I and Type II Errors occur when the true state of the world does not match the conclusion we reached at the end of the hypothesis test.
  - They are not due to mistakes in applying the hypothesis test.
  - Instead they are due to sampling variability / random chance; in some sense, these errors are due to "bad luck"…we just happened to get a sample where the calculated value of the statistic was in the extremes of the sampling distribution.
- Hypothesis tests are designed so to have certain properties, meaning that they have fixed values for α, β, and power; Lecture 6.4 discusses this.
  - Calculating the probabilities depends on specifics of a particular test (see practice problems).

PRACTICE PROBLEMS

3. A fabric manufacturer is investigating the proportion of orders for their fabric that are shipped out late. They plan to test $H_0: p = 0.3$ vs. $H_a: p < 0.3$ using a random sample of 50 orders. If 20% or less (i.e. 10 or fewer) of the sampled orders are shipped late, they will reject the null hypothesis.
   a. Define a Type I Error in the context of this problem.
   b. What is one example of a real-world consequence of a Type I Error in this context?
   c. For this test, $\alpha = 0.06$; write a sentence interpreting α in the context of this problem.
   d. Define a Type II Error in the context of this problem.
   e. What is one example of a real-world consequence of a Type II Error in this context?
   f. If $p = 0.25$ under the alternative, $\beta = 0.79$; write a sentence interpreting β and one interpreting power in the context of this problem.
   g. If $p = 0.15$ under the alternative, $\beta = 0.16$; write a sentence interpreting β and one interpreting power in the context of this problem.
   h. Why is the value for β lower (and power higher) for part (f) than part (g)?
4. Consider testing $H_0: \mu = 50$ vs. $H_a: \mu > 50$ using a 5% significance level. Considering all possible combinations from parts (a) to (c), use PROC POWER in SAS to calculate n needed to:
   a. Achieve 80% or 90% power…
   b. When we believe the true mean under the alternative is 55 or 60…
   c. When we believe the true standard deviation is 20 or 30.
   d. What happens to n as: We require higher power? The effect size $(\mu_{alternative} - \mu_{null})$ increases? The standard deviation increases?
   e. What is the predicted power is budget constraints limit sample size to 75 (considering all possible combinations from parts (b) and (c))?

CHI-SQUARE TESTS

- **Test of Goodness of Fit**
    - Goal: test fit of a discrete distribution
    - Features of the data*: One categorical response variable, one population (1-dimensional)
- **Test of Homogeneity**
    - Goal: compare distributions
    - Features of the data*: One categorical response variable, two or more populations
- **Test of Independence**
    - Goal: determine if there is a relationship
    - Features of the data*: Two categorical variables (response and explanatory), one population

* Caution: Determining what is a variable and what is a population can be difficult sometimes.

- If we targeted specific groups when we sampled (e.g. stratified sampling), then these groups represent populations.
    - Example: If we took separate samples based on year in school (e.g. freshman, sophomore, junior, senior), then year in school represents populations.
- Sometimes, however, we take a variable and treat the categories as populations (e.g. we take a single random sample but divide it by year in school for the analysis).
- When in doubt about whether something is "just" a variable or if it represents populations, use the goal of the analysis to help you determine the type of test to use.

PRACTICE PROBLEMS

5. For each of the following, determine which type of Chi-square test would be appropriate.
    a. We would like to learn if opinions related to gun control (stronger laws needed, laws good as is, weaker laws needed) is independent of political affiliation (Republican, Democrat, Other).
    b. We would like to learn if machines in a particular manufacturing plant produce defective parts at the same rate or not. Specifically, we are concerned that Machine A may produce twice as many defective parts than Machines B or C.
    c. We would like to learn if the distribution of drug use (never, rarely, occasional, frequently, daily) is the same across levels of educational attainment (no or some high school [HS], HS graduate, some college, college graduate, graduate school).
    d. Data on the starting position of the winning horses in 144 races is available; there are 8 starting positions (gates) to consider. We would like to determine if starting position has an effect on the chance of winning a race, or if horses starting in each are equally likely to win.
    e. A genetics experiment on the characteristics of tomato plants counts the number of plants that are one of the four phenotypes: (1) tall, cut leaf; (2) dwarf, cut leaf; (3) tall, potato leaf; (4) dwarf, potato leaf. We would like to determine if these four phenotypes will appear in following proportions: $p_1 = 0.56, p_2 = 0.19, p_3 = 0.19, p_4 = 0.06$.
    f. The recruitment director for a large firm categorizes universities as: most desirable, desirable, adequate, or undesirable. They review the performance of a random sample of employees as: outstanding, average, or poor. The director would like to determine if there is a relationship between university type and performance rating for employees.
    g. Integrated Pest Management (IPM) adopters apply less insecticide to crops than non-adopters. We have randomly selected farmers from six different states and would to determine if the distribution of IPM adopters (compared to non-adopters) is the same for all six states.

MEASURES OF VARIABILITY FOR ANOVA OR REGRESSION

- I think the confusion arises because all measures of variability relate back to the original definition of variance (and standard deviation) that was presented in Note Outline 2.
    - Thus the measures can sound similar when you are describing them.
    - Some are in fact directly related to each other, while there are (sometimes subtle) distinctions between others based on what and how we are measuring.
    - I do my best to describe them below, including pointing out connections between measures as appropriate, but I do not repeat the detailed formulas from the notes.
- For this discussion, the measures are defined in the context of ANOVA and SLR, but the ideas generalize to multiple regression.
- $\boldsymbol{s} = \sqrt{MSE}$
    - In SLR:
        - Called the *standard deviation of the residuals* or the *residual standard error*
        - Is the average distance of the points around the line; said another way: it is the average distance of the actual values of y from what we predicted
    - We didn't talk about this much in ANVOVA, but it would represent an estimate of the common population variance (recall assumption that population variances are all equal)
- Sums of Squares: Overall measures of variability
    - Total variability in the response variable y
        - Measured by *Total Sums of Squares*: $\boldsymbol{SST} = \boldsymbol{SS_{yy}}$ (occasionally denoted $\boldsymbol{S_{yy}}$)
        - *SST* is the same for ANOVA and regression
    - Variability in y that <u>can</u> be explained (by the treatment or by the model)
        - In ANOVA: Measured by *Sums of Squares for Groups* ($\boldsymbol{SSG}$)
            - Sometimes also called *Sums of Squares for Treatment* ($\boldsymbol{SS_{treatment}}$)
        - In SLR: Measured by *Regression Sums of Squares* ($\boldsymbol{SSR}$)
            - Sometimes also called *Model Sums of Squares* ($\boldsymbol{SSM}$)
        - *SSG* and *SSR* are really the same thing, the difference in notation reflects the different goals of the analysis (comparing groups, estimating a model)
    - Variability in y that <u>cannot</u> be explained (by the treatment or by the model)
        - Measured by *Sums of Squares for Error* ($\boldsymbol{SSE}$)
        - Sometimes also called *Error Sums of Squares* ($\boldsymbol{SSE}$)
        - *SSE* is the same for ANOVA and regression
    - Variability in the explanatory variable x
        - Measured by the sums of squares for the *X*s ($\boldsymbol{SS_{xx}}$; occasionally denoted $\boldsymbol{S_{xx}}$)
        - SS_{xx} is the same for ANOVA and SLR, though we only focused on it for regression
    - Covariance between y and x (e.g. how x and y vary together)
        - Measured by $\boldsymbol{SS_{xy}}$ (occasionally denoted $\boldsymbol{S_{xy}}$)
        - SS_{xy} is the same for ANOVA and SLR, though we only focused on it for regression
- Mean Squares: Average measures of variability
    - $\boldsymbol{MSG}$ & $\boldsymbol{MSR}$ measure the same thing as *SSG* & *SSR*; they just now represents averages
    - $\boldsymbol{MSE}$ measures the same thing as *SSE*; it just now represents an average
- Most of these measure variability pertaining to the response variable
    - In SLR, *s* specifically compares the actual values of y with the predicted values
        - It is used to evaluate the fit of the model and is important in inference
    - *SSG* & *SSR* and *SSE* (and the corresponding mean squares) divide the overall variability in the response (*SST*) into what can be explained (*SSG* & *SSR*) and what cannot (*SSE*)

        - They are useful for inference and also calculating R-square
    - *SST* is like the numerator of the variance formula that was presented on page 17 of Note Outline 2, so it measures the variability in *y* ignoring *x*
        - $\boldsymbol{s_y}$ would represent the <u>standard deviation</u> of the response variable
    - *SSG, SSR, SSE,* and SS_{xy} each then measure the variability in *y* when we take *x* in account
- SS_{xx} is the only measure that does not consider *y*. It is like the numerator of the variance formula when considering only the explanatory variable, so it measures the variability in *x* ignoring *y*
    - It is useful for inference and also for calculating the sample slope
    - $\boldsymbol{s_x}$ would represent the <u>standard deviation</u> of the explanatory variable

SUMMARY

- $\boldsymbol{SST = SS_{yy} = S_{yy}}$ = overall variability in response variable
- $\boldsymbol{s_y}$ = standard deviation of response variable
- $\boldsymbol{SSG = SSR = SSM}$ = variability in response that can be explained
    - $\boldsymbol{MSG = MSR = MSM}$ = average variability in response that can be explained
- $\boldsymbol{SSE}$ = variability in response that cannot be explained
    - $\boldsymbol{MSE}$ = average variability in response that cannot be explained
- $\boldsymbol{SS_{xx} = S_{xx}}$ = overall variability in explanatory variable
- $\boldsymbol{s_x}$ = standard deviation of explanatory variable
- $\boldsymbol{SS_{xy} = S_{xy}}$ = covariance between response and explanatory variable

CONNECTION TO R-SQUARE

- R^2 = proportion of variability in y that is explained by the model
    - Think about this in light of the definitions above: SSR measures variability in *y* that can be explained by the model; SST measures overall variability in *y*
    - Thus: $R^2 = \frac{SSR}{SST}$
        - And since $SSR = SST - SSE$: $R^2 = \frac{SST-SSE}{SST} = \frac{SST}{SST} - \frac{SSE}{SST} = 1 - \frac{SSE}{SST}$

CONNECTION TO FORMULAS IN ANOVA TABLE

- $SST = SSG + SSE$
    - So: $SSG = SST - SSE$ and $SSE = SST - SSG$
- $MSG = \frac{SSG}{df_{groups}}$
    - So: $SSG = MSG \times df_{groups}$
    - Recall: $df_{groups} = t - 1 =$ number of groups minus 1
- $MSE = \frac{SSE}{df_{error}}$
    - So: $SSE = MSG \times df_{error}$
    - Recall: $df_{error} = n - t =$ sample size minus number of groups
- $F = \frac{MSG}{MSE}$
    - So: $MSG = F \times MSE$ and $MSE = \frac{MSG}{F}$
- $df_{total} = n - 1 =$ sample size minus 1

CONFIDENCE VS. PREDICTION INTERVALS FOR A REGRESSION ANALYSIS

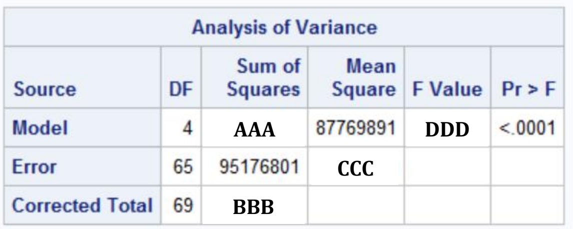
- **Confidence interval (CI) for the <u>average</u> value of the response** is used to estimate the true *average* value of the response for all people who have some specific value of the explanatory variable
- **Prediction interval (PI) for an <u>individual</u> value of the response** is used to estimate the *individual* value of the response for a future person who has some specific value of the explanatory variable
- Key distinction: CI is for *average value* while PI is for an *individual value*
  - This is reflected in the interpretation of each type of interval.
  - It is also reflected in the width of the intervals (see next bullet).
- Recall that there is less variability in the distribution of the averages (i.e. the sampling distribution of the sample mean) than there is in the distribution of the individuals (e.g. the population), thus the CI will always be narrower than the PI.
  - This is seen in the formulas for the *standard error* of each interval:
    - For the CI: $s\sqrt{\frac{1}{n}+\frac{(x_i-\bar{x})^2}{SS_{xx}}}$
    - For the PI: $s\sqrt{1+\frac{1}{n}+\frac{(x_i-\bar{x})^2}{SS_{xx}}}$
    - That extra "1" in the standard error for the PI accounts for additional variability in the distribution of the individuals because it is really an extra s^2 (the square of the residual standard error)! This can be seen with some basic algebra:

$$s\sqrt{1+\frac{1}{n}+\frac{(x_i-\bar{x})^2}{SS_{xx}}}=\sqrt{s^2\left(1+\frac{1}{n}+\frac{(x_i-\bar{x})^2}{SS_{xx}}\right)}=\sqrt{s^2+\frac{s^2}{n}+s^2\frac{(x_i-\bar{x})^2}{SS_{xx}}}$$

- The "fanning" seen in each interval is due to the $\frac{(x_i-\bar{x})^2}{SS_{xx}}$ term in each standard error.
  - Numerator of this term will be smaller when the specific value of the explanatory variable we are considering (x_i) is closer to the mean ($\bar{x}$), thus the intervals will be narrower.
  - As x_i gets further from $\bar{x}$, the intervals will get wider.
  - When you use lines to connect the end points for each interval for all possible values of *x*, you get the fanning pattern observed in the graphic on page 21 of Note Outline 9.
    - Both the CI (solid band) and PI (dashed band) "fan" but it is less obvious in the PI because there is already so much additional variability.

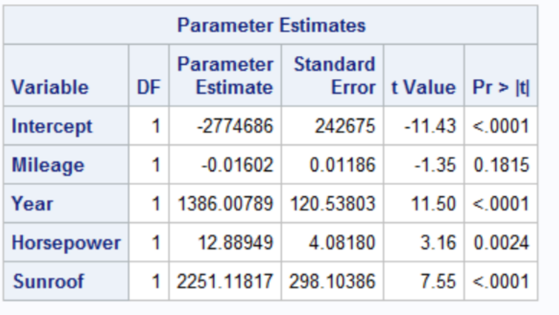
PRACTICE PROBLEMS

6. At the right are the results of a regression analysis where the price of a car (in dollars) is predicted based on its: mileage, model year, horsepower, and whether it has a sunroof (sunroof = 1 if yes and 0 if no).
   a. Use the formulas on page 6 of this document to fill in the missing values in the ANOVA table (denoted: AAA, BBB, CCC, and DDD).
   b. Write the formula, with values plugged in, that would be used to predict the price of a 2012 car with 58,000 miles, a 271 HP engine, and a sunroof.
   c. What percent of the variation in price is explained by the model?
   d. On average, how far are the actual values of price from what we predicted?
   e. How much would we expect price to change, on average, for each additional mile on the odometer (holding the other variables fixed)?
   f. Is there evidence that the model is significant overall?
   g. Is there evidence that horsepower is a significant predictor of price (holding the other variables fixed)?
   h. To check the necessary assumptions for validity of these tests, two plots were made. For each, state which assumption (normality, linearity, constant variance) the plot is checking and explain if the assumption(s) appear to be met or not.



Analysis of Variance

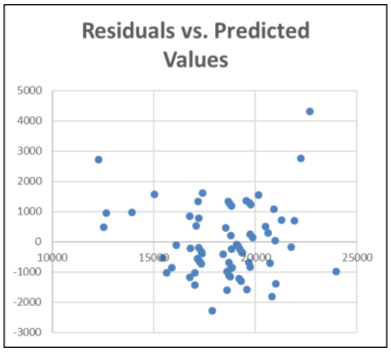
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	**AAA**	87769891	**DDD**	<.0001
Error	65	95176801	**CCC**		
Corrected Total	69	**BBB**			



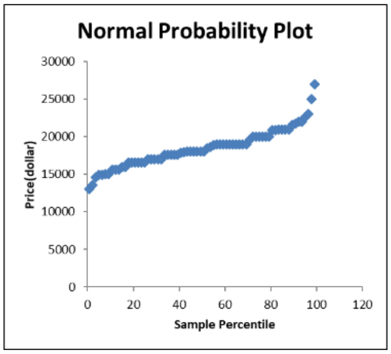
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-2774686	242675	-11.43	<.0001
Mileage	1	-0.01602	0.01186	-1.35	0.1815
Year	1	1386.00789	120.53803	11.50	<.0001
Horsepower	1	12.88949	4.08180	3.16	0.0024
Sunroof	1	2251.11817	298.10386	7.55	<.0001

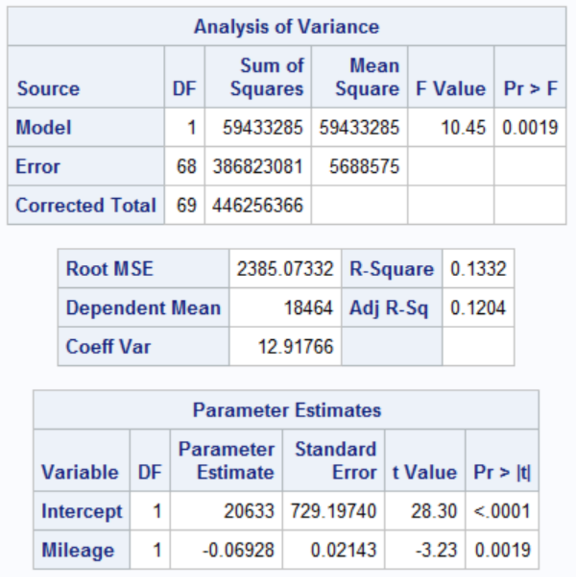
Plot #1



Plot #2



7. Below are results of the simple linear regression of price on mileage.



Analysis of Variance

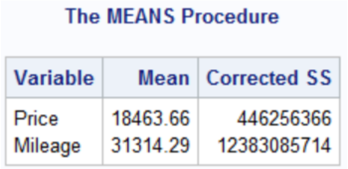
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	59433285	59433285	10.45	0.0019
Error	68	386823081	5688575		
Corrected Total	69	446256366			

Root MSE	2385.07332	R-Square	0.1332
Dependent Mean	18464	Adj R-Sq	0.1204
Coeff Var	12.91766		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	20633	729.19740	28.30	<.0001
Mileage	1	-0.06928	0.02143	-3.23	0.0019

The following output may also be helpful to answer the questions below:



The MEANS Procedure

Variable	Mean	Corrected SS
Price Mileage	18463.66 31314.29	446256366 12383085714

h. What price would we predict for a car that has 58,000 miles?

i. Write the formula, with values plugged in, that would be used to calculate the 95% confidence interval for the average price of cars that have 58,000 miles.

j. The confidence interval is (15340, 17890). Write a sentence interpreting this interval in context.

k. Write a sentence interpreting the confidence level in context.

l. Write the formula, with values plugged in, that would be used to calculate the 95% prediction interval the price of a car that has 58,000 miles.

m. The prediction interval is (11689, 21541). Write a sentence interpreting this interval in context.