MBA 200S Data and Decisions, Fall 2014

Lucas Davis and Noam Yuchtman

The University of California at Berkeley

Haas School of Business

Practice Waiver Exam, Solutions

COURSE OVERVIEW: Data and Decisions is not a typical statistics course. We move quickly, covering during seven weeks everything from statistical tests and confidence intervals all the way through interpretation and inference in multiple regression models. Along the way, we emphasize deep ideas rather than memorizing formulas, discussing, for example, how smart companies use experiments to increase profits and the business and ethical implications of "big data".

WAIVER EXAM: Statistics majors and other individuals with unusually strong backgrounds in statistics should consider taking the waiver exam.

TOPICS COVERED: The waiver exam will draw questions from the all seven weeks of Data and Decisions (MBA 200S). The practice exam provides a sense of the type of questions that we are likely to ask. But of course, no practice exam can cover all topics that may show up on the actual waiver exam.

TIME ALLOWED: You will have two hours for the waiver exam. Each problem will indicate how many points it is worth, so allocate your time accordingly.

SHOW YOUR WORK: Please show all relevant calculations and/or reasoning for each problem. Write your answer in the space provided using clear, legible, normal-size writing. You will lose points for excessively long or unfocused answers.

MATERIALS ALLOWED: The waiver exam is closed book and no laptops, tablets, or cell phones are allowed. You are allowed to use a single 8½ x 11 sheet of paper (double-sided) with notes and any kind of calculator. You will be provided with a Z-table, a t-Table, and a chi-squared table.

1. **Acme Stock.**

A. Acme stock increases by 1.25 percent for every 1.0 percent increase in the market.

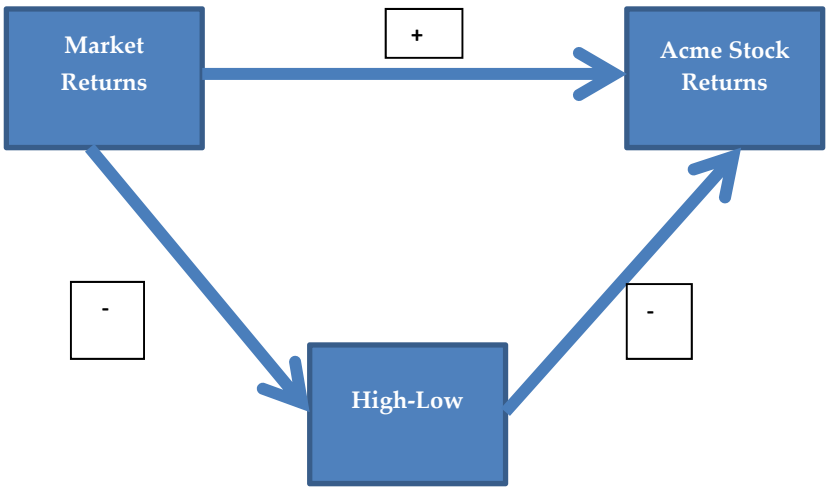
B. The t-statistic is,

$$\frac{b_1 - 1}{se(b_1)} = \frac{1.25 - 1.0}{0.156} = 1.603.$$

With 166 degrees of freedom, this t-statistic is slightly smaller than the t-value corresponding to p-value of .05. With software it corresponds to p-value = .0554. We cannot reject the null at a 5% significance level.

C. There is no collinearity problem. The variance inflation factors are very low here and both of the explanatory variables are statistically significant.

D. Negative. The easiest way to answer this question is to construct a path diagram. The partial slopes for the two explanatory variables are revealed in the MRM. The other piece of information comes from comparing the partial slope for "Market" in the two regressions. Notice that the slope becomes <u>smaller</u> after controlling for "High-Low". This implies that the two explanatory variables are <u>negatively</u> correlated. Consequently, in the SRM the coefficient on "Market" reflects both the positive direct effect and the positive indirect effect (negative*negative=positive).



2. **U.S. Open**

A. Two variables are independent if the joint probability is equal to the product of the marginal probabilities, or, equivalently, if the conditional probabilities are equal to the unconditional probabilities. In the context of this problem, independence would mean that the distribution of scores is the same in both rounds i.e. the proportion of birdies, pars, and bogeys is the same on Thursday and Sunday.

B. First calculate the column and row totals.

	Thursday	Sunday	TOTAL
Birdie (or better)	341	194	535
Par	1603	836	2439
Bogey (or worse)	846	450	1296
TOTAL	2790	1480	4270

The expected number of bogeys on Sunday is,

$$\frac{1296}{4270} * \frac{1480}{4270} * 4270 = 449.2.$$

The observed number of bogeys on Sunday is almost exactly equal to the expected number.

C. For testing whether two categorical variables are independent we use the chi-squared test of independence. The first step is to calculate the expected count for all six cells under independence. These are indicated below in parentheses.

	Thursday	Sunday	TOTAL
Birdie (or better)	341 (349.6)	194 (185.4)	535
Par	1603 (1593.6)	836 (845.4)	2439
Bogey (or worse)	846 (846.8)	450 (449.2)	1296
TOTAL	2790	1480	4270

On Thursday, there are more pars and fewer birdies than would be expected. On Sunday, there are more birdies and fewer pars. The observed numbers of bogeys are very close to the expected number on both days.

Now calculate the chi-squared statistic.

$$\chi^2 = \sum_{i=1,\,2,\,\ldots} \frac{(observed_i - expected_i)^2}{expected_i} =$$

$$\frac{(341-349.6)^2}{349.6}+\frac{(1603-1593.6)^2}{1593.6}+\frac{(846-846.8)^2}{846.8}+$$
$$\frac{(194-185.4)^2}{185.4}+\frac{(836-845.4)^2}{845.4}+\frac{(450-449.2)^2}{449.2}$$

$$= 0.212 + 0.055 + .001 + 0.399 + 0.105 + 0.001 = 0.773.$$

Overall, the actual counts are very similar to the expected counts under independence so the chi-squared statistic is a small number.

D. Birdies on Sunday. This is the biggest discrepancy from what would be predicted under independence. That single cell contributes more than half (.399) of the entire chi-squared statistic (.773).

E. The degrees of freedom test for the chi-squared test is (r-1)(c-1) = (3-1)(2-1) = 2.

F. We use the chi-squared table to calculate the *p*-value. Find on the table the row corresponding to 2 degrees of freedom. This chi-squared statistic is smaller than any of the cutoff values indicated on the table, so we conclude that the p-value is larger than 0.10. Using the chi-squared distribution with two degrees of freedom the cutoff for 10% significance level is 4.605 – compared to our test statistic of 0.773.

We conclude that the test statistic is not statistically significant at any conventional significance level and we do not reject the null hypothesis of independence. Despite the fact that on Sunday there is a different group of golfers, we cannot reject that the distribution of scores is the same as on Thursday.

3. **Mayor Tom Bates.**

At the middle of the confidence interval is 75% so that is the proportion of those who were interviewed who are satisfied with Tom Bates.

What is the sample size? The 95% confidence interval is +/- 1.96 times the estimated standard error which is given by the formula,

$$\text{se}(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n}.$$

The width of the confidence interval above is .044731 so the estimated standard error is .022822, and solving for n you get n = 360. We conclude that 270 out of 360 surveyed residents were satisfied with the mayor.

4. **Diamonds.**

A. The clarity category VS1 is the excluded category so the fitted equation excludes the slopes from both explanatory variables containing *VVS1*,

$$\hat{y} = -52.54 + 2863.50\,[\text{Weight}].$$

For a .4 carat diamond the predicted price is $1092.86. This is very similar to the SRM estimated in the book in Chapter 19 with this same dataset.

B. The prediction interval is given,

$$\hat{y} \pm 2s_e = \$1092.86 \pm 2\,(\$162.33) = [\$768.20, \$1417.52].$$

As usual, however, you would want to check a few conditions such as SRS and no obvious lurking variables. In addition, conditions for inference in the multiple regression model need to be satisfied: i. errors are independent, ii. errors are distributed normal, and iii. errors all have the same variance. You would want to make sure, in particular, that the residuals from both clarity types have similar variances. Finally, prediction intervals perform poorly under extrapolation, so you would want to make sure that you have many diamonds in your dataset of approximately .4 carats.

C. The fitted equation for VVS1 diamonds includes all four coefficients. The intercept is the sum of the intercept and the slope on *VVS1*, and the slope is the sum of the coefficient on weight and the interaction term,

$$\hat{y} = (-52.54 + 214.13) + (2863.50 - 211.54)[\text{Weight}] = 161.59 + 2651.96[\text{Weight}].$$

Somewhat surprisingly, the cost per carat is actually lower for the higher clarity diamonds. The difference in price between a .3 carat and a .5 carat diamond is just .2 times the slope for weight for a VVS1 diamond, equal to $530.39.

D. The first prediction equation implies that a .5 carat diamond with clarity VS1 has a predicted price of $1379.21. The second equation implies that a .5 carat diamond with clarity VVS1 has a predicted price of $1487.57. The difference is $108.36.

E. The two variables are highly collinear as indicated with the very large variance inflation factors. This collinearity could explain the lack of statistical significance. It would be interesting to estimate the model again but without the interaction term. It turns out that when you estimate the model without the interaction term the coefficient on the dummy for VVS1 is strongly statistically significant.

5. **Curved Patterns.**

A. False. Even if the correlation is high, you need to check the scatterplot to assess whether there is a non-linearity.

B. True. The intercept is measured in logs and thus cannot be easily interpreted. In contrast, the *slope* can be interpreted in percent (when the explanatory variable is in levels), or as an elasticity (when the explanatory variable is in logs).

C. False. In a regression the slope is an elasticity only if both x and y are measured in logs.

D. True. The standard deviation of the residuals always has the same units as the response variable.

E. True. Assuming the transformation worked and the relationship was approximately linear with a transformed response variable, then it will be curved when returned to the original scale. For example a model that is linear in *gallons per mile* will be nonlinear in *miles per gallon.*

6. Unemployment in the United States.

This is an example of Simpson's paradox. This is the idea that changes in the sizes of different groups can lead an aggregate pattern to differ from the pattern exhibited in groups. Education levels have increased dramatically since 1982. In 2009 there are many more college graduates, and many fewer high school graduates. Because unemployment rates in general tend to decrease with education, this compositional change explains how the overall employment rate can be lower than 1982 despite higher rates for each of the individual groups. As Princeton economist Hank Farber explains, "We have more skilled workers than we had before, and more-skilled workers are less susceptible to unemployment." See Wall Street Journal, "Good Data and Flawed Conclusions", December 2, 2009.

7. Knee Surgery.

A. The null and alternative hypotheses are H_0: $\mu \geq 50$ and H_a: $\mu < 50$. As with most business problems it makes sense to set this up as a one-sided test. The default is that the new technique is no better than the old technique. Only in the case in which we have evidence to reject this null, might the company take action, for example, marketing the new technique.

B. Type 1 error here is that the surgery isn't any better than conventional technique but you reject the null and, for example, market the new technique. Type II error is that the new technique is better, but that you fail to reject the null. Both types of errors are costly to the business.

C. We are told that n = 40, s = 22, and $\bar{x}$ = 43. Constructing the t-statistic we get,

$$\frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{43 - 50}{22/\sqrt{40}} = -2.012$$

On the t-table for 39 degrees of freedom this is slightly smaller than the cutoff corresponding to 95%. Because the table at the back of the book gives cutoffs for two-sided tests, we conclude that the p-value is slightly larger than .025. With software the exact p-value is .0256. Therefore, under the null hypothesis (i.e. that the true mean were indeed 50) the probability of observing a sample mean of 43 or smaller is 2.56%.

D. Yes. The p-value is less than 5%.