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

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 (20 total points)

Using 14 years of monthly data on Acme stock prices you estimate a capital asset pricing model (CAPM). The response variable is the percentage change in Acme stock value. The explanatory variable *Market* is the percentage change in the value of the entire stock market.

SUMMARY OUTPUT

Regression Statistics	
R Square	0.2793
Adjusted R Square	0.2749
Standard Error	9.22
Observations	168

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-0.279	0.713	-0.39	0.696
Market	1.250	0.156	8.02	0.000

A. (5 points) Interpret in words the coefficient corresponding to *Market*. Be sure to use the correct units.

B. (5 points) Test whether the slope coefficient corresponding to *Market* exceeds 1. Can you reject the null $\beta_1 \leq 1$ at a 5% significance level? Approximately what is the *p*-value?

You next estimate an augmented model that includes an additional explanatory variable *High-Low* which is the monthly percentage *difference* in returns between value stocks and growth stocks. Here "high" and "low" corresponds to the companies' book to market ratio. The results from the augmented regression are described below.

SUMMARY OUTPUT

Regression Statistics	
R Square	0.3072
Adjusted R Square	0.2988
Standard Error	9.07
F-Statistic	36.58
F-Statistic (*p*-value)	0.0000
Observations	168

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*VIF*
Intercept	-0.003	0.710	-0.00	0.996	
Market	1.057	0.171	6.20	0.000	1.24
High-Low	-0.572	0.222	-2.58	0.011	1.24

C. (5 points) Do you have a collinearity problem? How do you know?

D. (5 points) Is the correlation between *Market* and *High-Low* positive or negative? How do you know?

2. U.S. Open (20 points)

This table summarizes the performance of golfers on the first and last rounds in the 2010 U.S. Open golf tournament. The table shows the number of holes in each round completed in fewer than the allowed number of strokes "birdies", at the allowed number of strokes "par", or with more than the allowed number of strokes "bogey". Only players who score among the lower half after two rounds participate in the final two rounds. The PGA would like to know whether the distribution of scores is independent of round.

	Thursday	Sunday
Birdie (or better)	341	194
Par	1,603	836
Bogey (or worse)	846	450

A. (3 points) What does it mean for two variables to be "independent"? What would it mean for the distribution of scores to be independent of round?

B. (3 points) Under the assumption of independence, what is the expected number of bogeys on Sunday?

C. (5 points) Calculate the appropriate test statistic for this test of independence.

D. (2 points) Which cell contributes the most to the final test statistic?

E. (2 points) What are the degrees of freedom for this test?

F. (5 points) Calculate the approximate *p*-value. What are the conclusions for this test?

3. Mayor Tom Bates (10 points)

A random sample of Berkeley residents are interviewed to determine voter satisfaction with mayor Tom Bates. Let p be the true proportion of Berkeley residents who are satisfied with the mayor. After the survey is completed it is announced that (.705269, .794731) is a 95% confidence interval for p.

Using this confidence interval determine how many residents were interviewed, n, and what proportion of those who were interviewed were satisfied with the mayor.

4. Diamonds (25 points)

You have data on a sample of 144 emerald-cut diamonds. For each diamond you know the price (in dollars), the weight (in carats), and the clarity grade. The diamonds have clarity grade either VVS1 or VS1. The VVS1 category is nearly flawless whereas VS1 diamonds have more visible (but still small) flaws. You create a dummy variable "VVS1" for diamonds in the "nearly flawless" category (i.e. VVS1 = 1 for nearly flawless diamonds). You also create an interaction term that is the product of this dummy variable and weight (in carats). Using price as the response variable you estimate the following regression.

SUMMARY OUTPUT

Regression Statistics	
R Square	0.4980
Adjusted R Square	0.4872
Standard Error	162.33
F-Statistic	46.29
F-Statistic (*p*-value)	0.0000
Observations	144

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*VIF*
Intercept	-52.54	131.9	-0.40	0.691	
Weight (carats)	2863.50	316.3	9.05	0.000	1.58
VVS1	214.13	216.0	0.99	0.323	59.75
VVS1 * Weight (carats)	-211.54	522.2	-0.41	0.686	59.76

A. (5 points) What is the equation for the predicted price of a .4 carat diamond with clarity grade VS1?

B. (5 points) Calculate a 95% prediction interval for a single diamond with these characteristics. What conditions would you check before presenting this prediction interval in a board meeting?

C. (5 points) What does the model predict for the *difference* in price between a .3 carat and .5 carat diamond, both with clarity grade VVS1?

D. (5 points) What does the model predict for the *difference* in price between a .5 carat diamond with clarity grade VS1, and a .5 carat diamond with clarity grade VVS1?

E. (5 points) Given that neither *VVS1* nor *VVS1 * Weight (carats)* are statistically significant can you conclude that clarity is not a statistically significant factor in determining prices for this type of diamond? Why or why not?

5. Curved Patterns (15 points)

Mark the following statements True or False and then provide a brief explanation.

A. (3 points) If the correlation between y and x is larger than .7, then a linear equation is appropriate to describe the association.

(True / False)

B. (3 points) If the response variable of a regression is in logs, the interpretation of the intercept should be avoided.

(True / False)

C. (3 points) The slope of a regression model which uses log x as an explanatory variable is known as an elasticity.

(True / False)

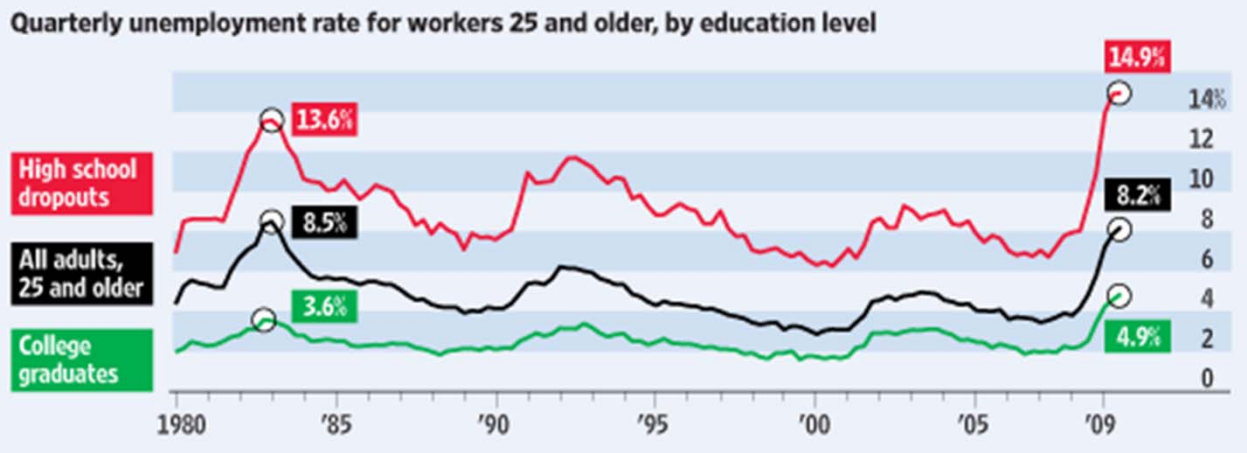
D. (3 points) When fitting the model $1/y = b_0 + b_1 x$ the standard deviation of the residuals (s_e) has the same units as $1/y$.

(True / False)

E. (3 points) When returned to the original scale, the fitted values of a model with a transformed response variable produce a curved set of predictions.

(True / False)

6. Unemployment in the United States (10 points)



Source: Wall Street Journal, December 2009.

As of the end of 2009 there was still debate in the United States as to whether the current economic slump was worse than the recession of the early 1980s. The *overall* unemployment rate for workers 25 and older was 8.2%, compared to 8.5% in 1982.

However, the picture was different when viewed separately by education level. The unemployment rate among both high school dropouts and college graduates was *higher* in 2009 than in 1982. Although not included in the figure above, the same also holds for high school graduates. In fact for all three education levels, the unemployment rate was higher in 2009 than it was in 1982.

How can you reconcile the *overall* pattern with this pattern by education level? What is the name of this statistical phenomenon? What specifically must have happened between 1982 and 2009 for this to be possible?

7. Knee Surgery (20 total points)

You are working for a medical company that has developed a new knee surgery technique. For a particular class of knee surgeries, the average recovery time using the old technique is 50 days. Using the new surgical technique on a random sample of 40 patients, you find that the average recovery time is 43 days with a sample standard deviation of 22. You may assume in this problem that the sample size condition is satisfied.

A. (4 points) Define an appropriate null and alternative hypothesis to test whether your results provide evidence that the new technique reduces average recovery time.

B. (4 points) In this problem what is a Type I error? What is a Type II error?

C. (8 points) Calculate the *p*-value. Interpret the *p*-value in words using as little statistical jargon as possible.

D. (4 points) Are your results significant at a 5% significance level? Circle One.

(Yes / No)