Stat 5870, section 2 – HW 10 – due Tuesday, 19 Nov, 11:59 pm to Canvas

1) Some of you know about the Challenger disaster of 1986. The space shuttle Challenger disintegrated shortly after liftoff. The cause was determined to be the o-rings that sealed joints in the booster rockets. At low temperatures, they became stiff and failed to seal. The temperature at Challenger liftoff was close to freezing, significantly colder than any previous shuttle liftoff. Further description, if you're interested, is in case study 4.1.

An aside, before the Challenger launch, engineers at the company that made the booster rockets analyzed the relationship between liftoff temperature and the number of failed o-rings (shown in case study 4.1). That analysis was completely wrong; the response variable was log(count), which removed all the 0 observations and found no evidence of a relationship with temperature.

We will use logistic regression. Fit the logistic regression that models the relationship between failure (0/1) and temperature. The data in oring.csv are whether one or more o-rings failed (failure = 1) or not (failure = 0) and temperature at liftoff (in degrees Fahrenheit) or the 24 liftoffs prior to the 1986 liftoff.

Software notes: SAS and R: you need to use the failure variable (0/1) as the response. SAS: remember to set the event as 1 (failure). JMP: you need to use the categorical variable (failureText) as the response.

Fit the logistic regression that predicts the log odds of failure as a linear function of launch temperature.

a) Report the estimated intercept and slope.

b) Calculate a 95% confidence interval for the slope.

c) Report whether that confidence interval is a Wald interval (based on a Z score) or a profile likelihood interval. Your answer will depend on your software.

d) Use the estimated slope and its standard error to construct a Wald test of slope = 0. Report the test statistic and associated p-value.

Note: You should be able to do this using software.

e) Use model comparison to construct a drop-in-deviance (aka Likelihood Ratio) test of slope = 0. Report the test statistic and associated p-value.

Note: You should be able to do this using software.

f) Consider two liftoffs. Liftoff A is at temperature T degrees; liftoff B is at temperature T-10 degrees. Use thinking about these two liftoffs to estimate the change in the log odds of failure when the temperature at liftoff is lower by 10 degrees F.

g) Estimate the odds ratio that completes this sentence:

The odds of failure is _____ times higher when the liftoff temperature is 10 degrees F lower.

h) Predict the probability of failure at a temperature of 32 F.

2) This problem demonstrates the interpretation of multiple linear regression slopes and why they usually differ from simple linear regression slopes.

The goal is to assess the "value" of education. Specifically, it is to estimate the relationship between Income and the number of years of education (6-20), where more than 16 includes time in graduate school. A potential confounding variable is student aptitude. The results given below come from analyses of data from a national (US) survey of youth over time. We have previously looked at these data using education categories. The results described below come from analyses treating years of education as a continuous variable. The response is income in 2005. This is log transformed to improve linearity. Educ is the number of years of education completed by 2006. Student aptitude is defined as the score on the Armed Forces Qualifying Test (AFQT), which ranges from 0 to 100. The results below are based on the subset of data for males.

Here are the coefficients for two fitted regression models. One is the simple linear regression with X = Educ. The second is the multiple linear regression with 2 X variables, Educ and AFQT.

Model	Intercept	Educ	AFQT
$Log\ Income\ =\ \beta_0 + \beta_1 Educ$	9.12	0.118	
$Log\ Income\ =\ \beta_0 + \beta_1 Educ + \beta_2 AFQT$	9.38	0.072	0.0067

a) The correlation between AFQT and Educ is 0.59. Does this help explain why the two estimates of the Educ slope are not the same? Briefly explain your answer.

Note: You should be able to answer question b) without predicting any values.

b) Consider two individuals: Individual A has 17 years of education, B has 18 years of education. That's all you know. Using the appropriate model, estimate how much larger individual B's log Income is (i.e. how much more than individual A)

c) Use the multiple linear regression to predict the log income for these two individuals:

Individual	Educ	AFQT
C	17	80
D	18	80

Then calculate how much larger individual D's log income is (compared to individual C). Your answer is the two predicted log income values and their difference.

d) You have fitted a regression to predict the average AFQT for a given education level. That gives the following information for individuals E and F

Individual	Educ	AFQT
E	17	77.370
F	18	84.252

AFQT in the table is the predicted average AFQT for an individual with that education. Use the multiple regression to predict the log income for these two individuals. Your answer is the two predicted log income values.

e) Use the results from f) to calculate how much larger individual F's log income is (compared to individual E). Your answer is the difference.

f) You have computed two estimated changes in log income (d and f). Which one matches the simple regression slope for Educ? Which one matches the multiple regression slope for Educ?

3) Problem 9:14 (pace of life), with my questions. This study examines the suggestion that individuals in cities with a faster "pace of life" are more likely to have heart attacks. Most of the data are explained in detail in the description of problem 9.14 in the text. A summary is that bank, walk, talk, and jog are 4 measures of the pace of life, with higher values meaning a faster pace of life. Note: I suspect that the description of bank is incorrect (scale is reversed) and jog is not in the data set provided by the book. The question is whether these are associated with the age-adjusted death rate from ischemic heart disease (heart attacks). The data are in pace2.csv. We will not use the avg variable in the data set.

The purpose of these questions is to illustrate some of the issues with multiple linear regression.

a) Plot a scatterplot matrix of the five variables ***excluding avg***. The most obvious visual pattern is the large variability in the plots of bank, walk, talk, and job against heart. Ignoring that, is it appropriate to assume a linear relationship between heart and the other 4 variables? Your answer is the plot, whether a linear relationship is appropriate and a briefly explaination.

b) Fit the linear regression predicting heart using 4 variables **bank, walk, talk, and jog**. Look at the T tests for the slope coefficients. Do any have p-values less than 0.05? If so, which variables are those?

c) Look at (or compute) the ANOVA table for the overall regression. This compares the model fit in 1b to the model with only the intercept. Report the p-value for this test and write a one-sentence conclusion.

Note: JMP and SAS users get this ANOVA table in the standard output. R users will need to compute the overall regression test by fitting the intercept-only model and explicitly comparing the two models. (Code in the early part of dietall.r, lab 7, shows you how to explicitly compare a model to an intercept-only model.)

d) The results in 1b suggest that three coefficients are "not significant". Fit a model predicting heart using only **bank**, then construct the second model comparison test (i.e., using an ANOVA table) that compares the bank model to the model in 1b. What is the null hypothesis for this model comparison?

e) Calculate the F statistic for the model comparison test in 1d.

Note: R users can compute this by explicitly comparing the two models using anova(). JMP and SAS users should compute the table by hand from the Error (aka Residual) df and SS for the two models.

f) Fit the linear regression predicting heart using 3 variables **bank, walk, and talk** Look at the T tests for the 3 slope coefficients. Do any of these have p-values less than 0.05? If so, which variables are those?

g) no answer needed. Think about why a variable may "become significant" (e.g., walk in 3b vs. 3f). We'll talk about this in lecture.