STATS 210P

Homework 3

Sevan Krikor Gulesserian

Round to the nearest 2nd decimal place (nearest hundredth) when possible.
Use $\alpha = 0.05$ significance level where needed.

1. Say we have two quantitive explanatory variables, X_1 and X_2, and a quantitative continuous response Y. Say we have two population models.

Model 1 is: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$.

Model 2 is: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \varepsilon_i$.

a. Is Model 1 nested under Model 2? Yes or no.

b. In Model 1, what would a null hypothesis of $\beta_1 = 0$ imply with how X_1 affects the response?

c. In Model 2, if β_3 is not equal to 0, what does that imply with how X_1 is affecting Y. Explain in a sentence or two.

d. In Model 2, given that X_1 is in the model already, say you want to test if X_2 affects the response directly or indirectly (through X_1). What should be the null and alternative be?

e. In Model 2, if you want to test if X_1 and X_2 have any association with Y, what would the null and alternative be?

f. In Model 2, now say X_2 is a categorical yes/no covariate (0 for no and 1 for yes). Interpret the effect of X_1 on the response variable Y.

2. Say you have two models with respect to the same response variable Y. Model 1 has all the p-many explanatory variables that Model 2 has but also an additional new explanatory variable. Thus Model 1 is Model 2 (X_1 to X_p in it) but with an extra explanatory variable (X_{p+1}). Model 2 is nested under Model 1.

For each part, state whether the answer is yes (or very likely to be yes), no (or very likely to be no), or need more information/maybe. Explain in a sentence.

a. The **adjusted** R-squared (not the regular R squared, but the adjusted one) of Model 1 is more than Model 2.

b. Model 1 has a higher sum of square total (SSTO) than Model 2.

c. Model 1 has a higher sum of square error (SSE) than Model 2.

d. The slope β_1 on the first explanatory variable, X_1, is the same in both models.

3. This question will use the pulse.txt dataset we have been using. Let the response Y be Rest (resting heart rate), and covariates X_1 be Height in inches (Hgt), X_2 be Weight in pounds (Wgt) and X_3 smoking status (Smoke, 1 for smokers and 0 for non-smokers).

a. Say we want a multiple linear regression model with all the covariates listed in it, along with an interaction between height and weight. Write out this population model.

b. Why does it make sense to have an interaction between weight and height in the model? Explain.

c. Fit the model from part a. in R and write out the estimated model. What is the adjusted R-squared value?

d. What is the estimated SSE (sum of squared errors) of your model? Use the R output from the summary(model), along with some formulas, to compute this.

e. Test if your model has any significance. Write out the null and alternative hypothesis, the test statistic (and what distribution it follows), p-value, and make a conclusion.

f. Test the interaction term between height and weight. State the null and alternative hypothesis and p-value. What can you conclude with respect to the effect of height and weight on the response of resting heart rate ?

g. Now a researchers states that you do not need weight in the model in any way or form. Write out the null and alternative hypothesis for this (write it out in terms of slope coefficients).

h. Conduct the test from part g. Make a conclusion in context of the study.

i. Show the output from R for the sequential sum of regressions table (keep the order of X_1, X_2,

X_3, and X_1X_2 in your model). Does adding weight when height is already in the model add to the explanatory strength of the model?

j. Using the sequential sum of regression table from part i, what is the SSTO (sum of squared total) for this model?

k. Now say you have a model with only height in it. From your table in part i, what is the SSE, SSR, and SSTO of this model with only height as covariate? Note: Do not run a new model, use the table from part i.

l. Explain in a few sentences how daily exercise amount could be a potential confounder in the model from part a.

m. Using the model from part a., create and present the scatterplot of the residual versus fitted values (fitted values go on the X-axis). Is there any evidence that our assumption of constant variance (a single σ^2 for the entire model) is invalid?

n. Using the residual vs. fitted plot from part m., comment on the linearity assumption of the model that was fit.

o. Using the model from part a., now create a QQ plot of the residuals. Do we seem to have any issues with our normality assumption for the errors?

p. Using the model from part a., do a summary(data) to see the summary statistics of the weight (Wgt) explanatory variable. Does it make sense to use your model from part a. to predict the resting heart rate for someone who weight 350 pounds? Why or why not.

4. An environmental expert is interested in modeling the concentration of various chemicals in well water over time. Identify the regression population models that would be used to (just write out the population models):

a. Predict the amount of arsenic (Arsenic) in a well based on Year, the distance (Miles) from a mining site, and the interaction of these two variables.

b. Predict the amount of lead (Lead) in a well based on Year with two different lines depending on whether or not the well has been cleaned (Iclean).

c. Predict the amount of titanium (Titanium) in a well based on a possible quadratic relationship with the distance (Miles) from a mining site.

d. Predict the amount of sulfide (Sulfide) in a well based on Year, distance (Miles) from a mining site, depth (Depth) of the well, and any interactions of pairs of explanatory variables.

5. This question will use the ThreeCars.txt dataset on our class Canvas page. The response variable is Price (the price of the car) and the explanatory variable is Mileage (miles of the car in 1000's of miles).

a. Create a scatterplot of Mileage on the X-axis and Price on the Y-axis. Show the plot and comment on what you see with respect to the association between Mileage and Price.

b. Fit a simple linear regression model with the response being Price and the covariate being Mileage. Show the residuals plot versus fitted values plot.

c. What do you see in the plot from part b. with respect to the linear relationship between Mileage and Price?

d. What do you see in the plot from part b. with respect to the constant variance assumption of the linear regression model?

e. Create a qq-plot of the residuals, and comment on the normality of the errors assumption.