**STA 403** **Name**: Gary Feng

**Final Exam** (150 pts) **Class Number:** 8

**Instructions**: When answering the following questions, be sure to follow all guidelines for homework submission. Your answers may be placed within this document. Submit a PDF version of your answers to the Assignments section of UBLearns by 3:30 PM EST on Friday, 12/17. Unless otherwise stated, carry out inferences using . Hypotheses should be written in terms of parameters when possible, and conclusions of hypothesis tests should be given using the context of the problem.

1. As the semester draws to a close, some students worry about their performance on the upcoming final exams. Data from 73 students in a college statistics course from a previous semester are recorded in the file grade\_pred.txt (UBLearns). We wish to model students’ final exam performance (200 points maximum) as a function of their homework average and their scores on three tests given during the semester. (34 pts)

1. Ensure that you have read the complete data set into Minitab by presenting the sample mean of the response variable as . (1)

Mean of Final

Graphical user interface

Description automatically generated with medium confidence

1. Suppose a first-order model in four predictor variables will be fit. State the model using mathematical symbols. Also write any assumptions to be made about the error term in the model. (4)

State a first order model.

Yi = B0 + B1Xi1 + B2Xi2 + B3Xi3 + B4Xi4 + Ei, Ei ~ Mu(0, sigma^2)

1. Produce a table of sample correlations between and the four predictor variables. Identify the predictor least likely to be highly explanatory of final exam score. Comment on your level of concern about multicollinearity; quote at least one value from the output. (4)

Give a table of sample correlation between Y vs X1, X2, X3, X4. What’s the predictor that is least likely to affect final exam score.

Chart, scatter chart

Description automatically generated

Test 1 will be the least likely to highly explanatory for the final exam. There isn’t a predictor that is high multicollinearity. One output is 0.469

1. Fit the first-order model and provide the coefficients table. (2)

Table

Description automatically generatedFit the first-order model. Give coefficient table.

1. Two of the students in this data set took a zero on one of their mid-semester tests. This represents unusual behavior, and there is a fairly strong argument for excluding these students. Refit the full first-order model after removing the two rows in question, and provide the updated coefficients table. (2)

Remove 2 rows with 0 in it. Fit another first-order model and give an updated coefficient table.

Table

Description automatically generated

1. Test whether a regression relationship exists between and this set of four predictor variables. Provide the hypotheses, test statistic, p-value, and conclusion. (4)

Test regression relationship between Y vs the 4 predictors. What’s the hypotheses, test statistic, p-value and conclusion.

H0: There is a regression relationship between Y and 4 predictor

Ha: There is not a regression relationship between Y and 4 predictor

p-value: > 0.100

Fail to reject H0. Yes there is a relationship between Y and the 4 predictors.

1. State the value of the estimated regression coefficient on test2, and give an interpretation; keep in mind that this is a multiple regression model. (3)

What is the value in the estimated regression coefficient on test2? Explain.

0.581

Test 2 grade has count as 0.581 times towards the final grade, however, this only applies to when all the other variables are calculated for final grade estimate.

1. Use the output from the fitted model to comment further on the degree of multicollinearity present among this set of predictor variables. (2)

Using the output from the fitted model, comment on the multicollinearity.

There are correlation but very little from the set of predictors.

1. Given that all three test scores are included as predictors, test whether homework score may be dropped from the model. Give the hypotheses, test statistic, p-value, and conclusion.(4)

H0: B1 = 0

Ha: B1 != 0

Test statistic: (10235.8-4246.4)/(4-3) / (4246.4)/(71-4-1) = 93.0907

P-value: 0.000

P value < 0.05. We reject null hypotheses. Homework cannot be dropped.

1. After taking any action implied by the results of the test above, obtain the model residuals. Apply the Ryan-Joiner test for normality, stating the hypotheses, p-value, and conclusion. (3)

Get the residuals and do a ryan joiner test. What’s the hypotheses, p-value, and conclusion?

H0: the data come from a normal population

Ha: the data come from not a normal population

P-value: 0.029

P value < 0.05, so we reject null hypotheses. The data is not normally distributed.

1. Provide a plot of residuals vs. fitted values, and comment on the validity of the model assumptions. (3)

Give a residual vs fitted plot. Explain the validity.

Chart, scatter chart

Description automatically generated

The graph is a huge mess. The validity of the model is very bad.

1. Use the fitted model to estimate the mean final exam grade corresponding to students who score 75, 80, and 95 on the three tests given throughout the semester. Also provide a 90% confidence interval for this quantity. (2)

Estimate with a 90% CI for test 1 = 75, test 2 = 80, test 3 = 95

164.110

(157.39, 170.83)

2. Continue to use the data set from problem 1, after removing the two unrepresentative student records. For notational convenience, define . (15 pts)

1. Obtain numeric values for the following quantities. (6)

Solve the following:

20608.7

4377

16231

5024

971.6

1196

1. Compute the following coefficients of partial determination. (3)

Do partial R^2 for the following:

5024/16231 = .3095

971.6/10235.8 = .09492

1196/10236 = 0.1168

1. Explain what your value of signifies. (2)

What does Ry^2 3|12 mean?

When adding X3 to a model containing X1 and X2; X3 explains 9.492% of the remaining unexplained variability in Y.

1. Using the following model,

conduct a partial F-test of whether and can be dropped from the model simultaneously. Give the hypotheses, test statistic, p-value, and conclusion. (4)

Do a partial F test to see if we can drop X1 and X3. What’s the hypotheses, test statistic, p-value, and conclusion?

H0: B2 = B3 = 0

Ha: Not all B’s = 0

Test statistic: (13532-10235.8)/(3-1) / (10235.8)/(71-3-1)

= 1648.1/152.77 = 10.79

P-value: .000087

10.79 > 3.1337, we reject null hypotheses. We cannot drop X1 and X3 from the model.

3. Recall the cake preference data set discussed in class. The response variable is a preference score, where a high number represents high preference. Cake moisture content is recorded as the continuous variable , while can be treated as a binary categorical variable recorded as “sweet” or “not sweet.” The 16 observations are housed in the file cake\_ratings.txt (UBLearns). (29 pts)

1. To ensure all the data have been read in, reproduce the mean of the response variable as

. (1)

Graphical user interface, text, application

Description automatically generatedFind the mean of rating in cake\_ratings.txt

1. Suppose we wish to fit an ANCOVA model using these two predictors. State the ANCOVA model using mathematical symbols. (2)

What is the ANCOVA model using mathematical symbols?

Yi = B0 + B1Xi12 + B2Xi12 + Ei

1. We would like to visualize whether there is a linear relationship between moisture content and cake rating *at each level of sweetness*. Produce one or more separate scatter plots that include two regression lines, one for each sweetness category. (2)

Produce one or more scatter plot with regression lines based on the sweetness category.

Chart, line chart

Description automatically generated

1. Use the appearance of the plots above to explain whether the ANCOVA model seems appropriate. (2)

Does the plot for c look appropriate as an ANCOVA model?

Yes while the data is limited, the plots are very close to the line.

1. Use Minitab to fit the ANCOVA model. Provide a table of output that will permit you to execute the overall F-test. (2)

Fit the model, and give the output that allows us to do an F test.

Table

Description automatically generated

1. Give numeric evidence of the degree to which this ANCOVA model fits the data. Is the fit particularly good/bad? (2)

Is this ANCOVA model a good fit?

Yes, the adjusted R squared is 94.47%, so it’s an extremely good fit.

1. Perform the overall F-test, stating the hypotheses, test statistic, p-value, and conclusion. (4)

Do an overall F-test, what’s the hypotheses, test statistic, p-value, and conclusion.

H0: B1 = B2 = 0

Ha: Not all B’s = 0

Test statistic: 129.08

P-value: 0.000

P value is < 0.05. We reject null hypotheses. There is a linear association between y and the predictors.

1. Give interpretations of the quantities and as obtained from this model. (4)

How do you interpret b1 and b2 in this model?

Based on the moisture and sweetness of the cake, it will receive a different rating.

1. Use the fitted model to estimate the mean cake rating for batches with a moisture value of 5 and a sweetness category of “not sweet.” (2)

Estimate the rating of a cake with the moisture = 5 and not sweet.

68.525

1. A new batch of cake batter is mixed and is about to be placed in the oven. The ingredients are known to produce cake with moisture value 5 of the “not sweet” variety. Produce a 95% prediction interval, and give a detailed explanation about the meaning of this interval, in terms of the liquid batter. (4)

What’s the 95% PI rating for moisture of 5 and not sweet? Explain the PI.

(62.2179, 74.8321)

Just because we know the ingredients doesn’t mean we can make a good cake. With the liquid batter before putting it into the oven, we also don’t know the time it’ll take to cook as well as the temperature and with all the missing factors, we can predict it’ll at least have a rating of 62.2179 to 74.8321.

1. A colleague suggests fitting a larger model that includes the interaction between moisture content and sweetness, for the sole purpose of testing whether that interaction term was required. Fit the appropriate model, and test whether the interaction term can be eliminated. (4)

Fit a model between moisture content and sweetness and see if the interaction term can be removed.

No the interaction term cannot be eliminated

4. A new video hosting platform has been trying to recruit content creators. The platform earns money every time a viewer subscribes to a given channel, but the platform has to pay web hosting fees for every video supported by the platform. The file web\_hosting.txt contains data on 72 channels. The Earnings column contains the average monthly channel earnings (which may be negative) in thousands of dollars. The Length column gives the average length of a video on the channel, in minutes. Finally, a column labeled Group classifies a particular channel with the label I, II, or III. (26 pts)

1. Ensure that you have read in all the data by reproducing the sample mean of the Earnings column as . (1)

Graphical user interface, application

Description automatically generatedWhat is the mean of ‘Earnings’

1. Suppose we wish to model Earnings as a linear function of Length, Group, and their interaction. State the regression model using mathematical symbols. (2 pts)

What is the regression model?

Yi = B0 + B1X1 + B2X2 + B3X3 Ei

1. Explain (using mathematical symbols) how Minitab will deal with the Group variable when instructed to fit a linear model and estimate parameters via least squares. (2)

Explain how minitab will deal with the group variable when you fit the linear model and estimate param via least square.

Minitab will split it up into 3 equations

Yi\_groupI = B0 – B1Xi1

Yi\_groupII = B0 – B1Xi1

Yi\_groupIII = B0 – B1Xi1

1. Write a set of expressions (using mathematical symbols) for the expected value of ; one for each possible value of . (3 pts)

What’s the expected value of Y for each of X2?

For group 1: E[Yi] = B0 + B2B2 + B3

For group 2: E[Yi] = B0 + B1 + B2X2

For group 3: E[Yi] = B0 + B1 + B2X2 + B3

1. Fit the full model in Minitab; provide the coefficients table and the ANOVA table. (3)

Fit the full model and give the coefficient table and ANOVA table.

Table

Description automatically generatedTable

Description automatically generated

1. Interpret the value of from the Coefficients table using the context of the problem. Also choose one of the estimates under the heading “Group” and provide an interpretation. (4)

What does b1 mean from? Also interpret a group coeff.

The longer the length of the video, the more it would cost to host it. Depending on the group, it’ll earn a lot more money, most likely due to the views.

1. One question of interest is whether the group-specific regression functions written in part (c) have the same slope. Produce three separate scatter plots of Earnings vs. Length; one for each Group classification. Include a regression line in each plot, and comment on what you see. (4)

Make 3 separate scatter plots of earning vs length based on the groups.

Chart, scatter chart

Description automatically generated

They do not have the same slope. Group I earning is actually decreasing as their video length increases.

1. Perform a single hypothesis test of whether the three group-specific regression lines have the same slope. Give the hypotheses, test statistic, p-value, and conclusion. (4)

Do a single hypothesis test for all group’s slope are the same. What’s the hypotheses, test statistic, p-value, and conclusion?

H0: B1 = B2 = B3

Ha: Not all B’s = 0

Test statistic: 163.24

P-value: 0.000

P value is < 0.05. We reject null hypotheses. The slopes are not the same.

1. State the p-value (from the full model) corresponding to a test of whether Length is a significant predictor of Earnings. (1)

What is the p value of length?

0.250

1. Given the results of the test in part (i), can we drop Length from the model? Explain. (2)

Can we drop length based on part i.

P value > 0.05. No we cannot drop Length as it has a huge impact on the model. Also if we dropped length from the model, the model would just be based on a categorical variable with no continuous variable.

5. Obesity researchers randomly selected 50 male patients to a local clinic, and obtained permission to access their medical records; the data appear in a file called bodyfat21.txt (UBLearns). The continuous variables in the data set are the subjects’ percent body fat, their age (years), waist size (inches), thigh circumference (cm), bicep circumference (cm), and wrist circumference (cm). (25 pts)

1. Ensure that you have correctly read in the full data set by reproducing the sample mean of the percent body fat variable as 17.37. (1)

Text

Description automatically generated with medium confidenceMean of Pct.BF

1. The researchers wish to build a model in which percent body fat is the response variable, and the available predictors are the five measurements listed earlier. Generate the matrix of sample correlations between the response variable and the five predictors. Comment on any variables that appear to be strongly related to the response. Also comment on any cause for concern that you see. (4)

Chart, diagram, scatter chart

Description automatically generatedGenerate a matrix of correlation between response and the 5 predictors. Any strong correlations? Any concerns?

There are no strong correlation. Age seems to be the least impactful, so it might be valid to remove it

1. Fit the full first-order model in five predictors. Provide a table of output displaying estimates of the regression parameters, and a separate table containing . Comment on the variance inflation factors for this model. (4)

Fit a first-order model with the 5 predictors. Give table with estimates of regression parameter. Give adjusted R squared table. Comment on VIF for the model.

Table

Description automatically generated with low confidenceTable

Description automatically generated

There are no VIF that are above 10, so there’s no multicollinearity.

1. Use Minitab to obtain a model chosen by forward selection, using . Provide the estimated regression function, the coefficients table, and a table of output detailing the actions taken at each step in the automated procedure. Write a summary statement describing what happened at each step of the process. (4)

Do a forward selection with alpha = 0.05. Give estimated regression function, the coefficient table, and table of steps. Write a summary describing the process.

Table

Description automatically generatedText

Description automatically generated

A screenshot of a computer

Description automatically generated with low confidence

Waist was added in first because it has a p-value close to 0 which explained 60% of the model followed by wrist with a p-value of 0.001 which explains 68% of the model and then it stops because the p values of the other variables are too large.

1. Use Minitab to obtain a model chosen by backward selection, using . Provide the estimated regression function, the coefficients table, and a table of output detailing the actions taken at each step in the automated procedure. Write a summary statement describing what happened at each step of the process. (4)

Do a backwards elimination with alpha = 0.05. What is the estimated regression function, coefficient table, and table of steps. Write a summary explaining the process.

Text

Description automatically generated with medium confidence Table

Description automatically generated

Table

Description automatically generated

All the variables were fit. Age has a huge p-value, so it was removed first. Then biceps and then finally thighs until we’re left with p-values that are smaller than 0.05.

1. Compare the final models chosen by the two automated selection procedures, and explain whether you are surprised by the results. (2)

Compare the 2 models. What surprised you.

It ended up being the same model, which I sort of expected. I wasn’t surprised that age got removed first. I am surprised wrist stayed. From what I understand about wrist width and strength, that it’s usually genetics and wrist tend to not have that much fat. I would have figured waist and thighs would be the final model because a lot of fat ends up going there.

1. Use the best subsets approach to consider a collection of first-order models. (4)
   1. Provide concise output representing the “best” models, restricted to .

Table of the best subset with p = 3

Table

Description automatically generated

* 1. Which is the best model () if is used as the selection criterion?

Using adjusted predict R squared what is the best mode?

Waist + Thigh + Wrist

* 1. Would the same model be chosen if were the criterion?

Instead of adjusted predict R square, use AICp. Is it the same?

No assuming p = 3, it would be Waste + Bicep + Wrist

* 1. Would the same model be chosen if were the criterion?

Instead of adjusted predict R square, use PRESSp. Is it the same?

No it would be the same as iii.

6. Dengue fever is a viral infection that is spread chiefly by mosquitoes. A set of 196 subjects in a Mexican city were enrolled in a study, and followed for a set period of time. The data set dengue.txt contains an the age of the subject, the sector of the city in which the subject lives (1=lower, 2=upper), and an indicator of dengue infection (1=Yes, 0=No). (23 pts)

1. Suppose we wish to model the probability of dengue infection as a function of age using logistic regression. Produce a graphic that displays the estimated response function. Describe what happens to the estimated probability of infection as age increases. (4)

Plot a logistic regression disease vs age. Explains what happens to the probability of infection as age increases.

Chart, line chart

Description automatically generated

When age increases, the probability of getting the dengue virus also increase.

1. Suppose we wish to use logistic regression to model the probability of dengue infection, using age and domicile sector as predictor variables. Provide a table of output showing the estimated regression coefficients and their standard errors. (2)

Give a coefficient table of a logistic regression disease vs age + sector.

Table

Description automatically generated

Sector seems to matter quite a bit on where people get the virus.

1. We would like to test whether the regression parameters corresponding to age and domicile sector are different from 0. Provide a single table of output that contains the p-value for the overall test (whether both parameters are 0 simultaneously), and for individual tests that a given parameter is equal to 0. Can either predictor be dropped from the current model ? (4)

Provide a table to test whether you can drop either predictor from the current model

Table

Description automatically generated

No we cannot drop any predictors from the current model

1. Provide Minitab output containing an estimate of the odds ratio corresponding to domicile sector. Describe this value using the context of the problem, and be clear about whether a particular sector is found to be more dangerous. (3)

Give minitab output of odds ratio for domicile sector. Explain using context of the problem.

Text

Description automatically generated with medium confidence

If you are in sector 2 your estimated odds of infection is multiplied by 3.2599.

1. Describe whether the 95% confidence interval for the sector odds ratio is consistent with the appropriate p-value in the output from part (c). (2)

Is the 95% CI for sector OR appropriate for p-value from output from c?

(1.0099, 1.0447)

1. Provide Minitab output containing an estimate of the odds ratio corresponding to a five-year increase in age. Interpret this value. (4)

What is the OR with increment of 5? Explain.

1.1435

When age increases by 5, the estimated odds of infection are multiplied by 1.1435

1. Use the fitted model to estimate the probability of dengue infection for a 40-year-old subject living in sector 1. (2)

What is the fitted probability of dengue infection of a 40 year old who lives in sector 1?

0.2522

1. Provide a 90% confidence interval for the probability estimated above. (2)

What’s the 90% CI for g?

(0.1797, 0.3417)