**STA 6207 – Practice Problems – Multiple Regression**

**Part A: Estimating and Testing**

QA.1. You obtain the following partial output from a regression program. Fill in all missing parts.





p.1.a. R2 = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ p.1.b. n = \_\_\_\_\_\_\_\_\_\_\_\_ p.1.c. dfReg = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

p.1.d. MS(Regression) = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ p.1.e. Fobs = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

p.1.f. Critical F-value ( = 0.05) = \_\_\_\_\_\_\_\_\_\_\_\_ p.1.g. 

p.1.h.  p.1.i. t-stat for H0: 1 = 0: vs HA: 1 ≠ 0: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

p.1.j. t-stat for H0: 2 = 0: vs HA: 2 ≠ 0: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ p.1.k. Critical t-value ( = 0.05) \_\_\_\_\_\_\_\_\_\_\_\_

QA.2. A multiple linear regression model is fit, relating height (Y, mm) to hand length (X1, mm) and foot length (X2, mm), for a sample of n = 20 adult males. The following partial computer output is obtained, for model 1 with 2 predictors.



p.2.a Complete the table. Do you reject the null hypothesis H0: ? **Yes** or **No**

p.2.b. Give the predicted height of a man with a hand length of 210mm and a foot length of (260mm). Just give the point estimate, not confidence interval for the mean or a prediction interval.

p.2.c. Give an unbiased estimate of the error variance 2

p.2.d. The coefficient of determination represents the proportion of variation in heights “explained” by the model with hand and foot length as predictors. What is the proportion explained for this model?

QA.3. For the Analysis of Variance in model 2, with *n* observations and *p* predictors, complete the following parts.

p.3.a. Write the Regression and Residual sums of squares as quadratic forms.

p.3.b. Derive the distributions of SSRegression/ and SSResidual/

p.3.c. Show that SSRegression/ and SSResidual/ are independent

p.3.d. What is the sampling distribution of MSRegression/MSResidual when p = 0?

QA.4. A multiple regression model is fit, based on Model 2, with *p* predictors and an intercept. Define the projection matrix as: , where Define

p.4.a. Show that **P** and are symmetric and idempotent. (Hint: **(X’X)-1** is symmetric). **SHOW ALL WORK**.

p.4.b. Obtain  **SHOW ALL WORK**

p.4.c. Obtain the rank of **P**,, and  **SHOW ALL WORK**

p.4.d. What is the sampling distribution of ? **SHOW ALL WORK**

p.4.e. Show that and are independent. **SHOW ALL WORK**.

QA.5. Use the following output to obtain the quantities given below:

**   **

** **

Total Corrected: Sum Of Squares Degrees of Freedom

Regression: Sum Of Squares Degrees of Freedom

Residual: Sum Of Squares Degrees of Freedom

S2 

Testing H0: F-stat Num df Den df

Predicted Value for Y2

QA.6. Use the following output to obtain the quantities given below:





Complete the following elements of the regression model:

 

S2 = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 

Tests of (TS=Test statistic, RR=Rejection Region) each based on  = 0.05 significance level:



Predicted Value for Y4 based on each of these two forms and its residual (show work):



QA.7. A large electronics retailer is interested in the relationship between net revenue of plasma TV sales (Y, $1000s) , and the following 4 predictors: X1= shipping costs ($/unit), X2= print advertising ($1000s), X3= electronic media ads ($1000s), and X4= rebate rate (% of retail price). A sample of n=50 stores is selected and the resulting (partial) regression output is obtained:





p.7.a. Complete the ANOVA table.

p.7.b. Give the prediction for net revenue, when ShipCost=10, PrintAds=50, WebAds=40, Rebate%=15.

p.7.c. Controlling for all other factors, give a 95% confidence interval for the change in expected net revenue ($1000s) when Rebate% is increased by 1.

p.7.d. Test H0: PrintAds - WebAds = 0 vs HA: PrintAds - WebAds ≠ 0 at  = 0.05 significance level:

p.7.d.i. Test Statistic: p.7.d.ii. Rejection Region

p.7.e. What proportion of variation in revenues is “explained” by the regression model?

QA.8. You obtain the following partial output from a regression program. Fill in all missing parts.





p.8.a. R2 = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ p.8.b. n = \_\_\_\_\_\_\_\_\_\_\_\_ p.8.c. dfReg = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

p.8.d. MS(Regression) = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ p.8.e. Fobs = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

p.8.f. Critical F-value ( = 0.05) = \_\_\_\_\_\_\_\_\_\_\_\_ p.8.g. 

p.8.h.  p.8.i. t-stat for H0: 1 = 0: vs HA: 1 ≠ 0: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

p.8.j. t-stat for H0: 2 = 0: vs HA: 2 ≠ 0: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ p.8.k. Critical t-value ( = 0.05) \_\_\_\_\_\_\_\_\_\_\_\_

QA.9. A linear regression model is fit, relating mean January temperatures (Y, in ᵒF) to Elevation (X1, in 100s of feet) and Latitude (X2, in degrees north latitude) for a random sample of n = 63 weather stations in Texas. The (partial) computer results are given below.



p.9.a. Complete the ANOVA table.

p.9.b. Dallas/Fort Worth International Airport (DFW) was not in the sampled locations, and is located at an elevation of X1 = 5.6 and a latitude of X2 = 32.9. Give the predicted value for DFW.

p.9.c. For DFW, we obtain the following values: .

Compute the 95% **Prediction Interval** for DFW’s mean January temperature.

QA.10. For the multiple regression model, with an intercept term, complete the following parts, **show all of your work.**

p.10.a. 

p.10.b. Derive the sampling distributions of 

QA.11. A linear regression model was fit, relating weekly number of passengers (Y, in 10000s) to number of street cars in operation (X1, in 100s) and number of miles street cars ran (X2, in 10000 miles) over a period of n=20 consecutive weeks. The following EXCEL spreadsheet summarizes the model.



p.11.a. Complete the sheet.

Note: VIF = Variance Inflation Factor and DW = Durbin-Watson Statistic for Autocorrelation



p.11.b. The critical values for the Durbin-Watson test with n=20 and p=2 are: dL = 1.10 and dU = 1.54. Does the assumption of uncorrelated errors seem reasonable? **Yes** or **No**

p.11.c. Based on the VIF, is there evidence of serious multicollinearity? **Yes** or **No**

**Part B: General Linear Hypothesis Tests**

QB.1. A forensic study related Hand (X1) and Foot (X2) lengths to Stature (Y) for a sample of n = 75 adult females (each variable in 100s of mms). Consider the following three models.





p.1.a. Compute  and the Total (Corrected) Sum of Squares.

p.1.b. Compute the Residual (Error) Sum of Squares for each model.

p.1.c. Compute 

p.1.d. Use the general linear test for Model 3 to test 

QB.2. A firm has 2 types of expenditures that can varied in their marketing plan: advertising and in-store promotion. A regression model is fit, relating Y=weekly sales to levels of these expense variables (X1=advertising, X2=in-store promotion). The model fit is: E(Y) = X1+2X2. Set up the **K’** matrix and **m** vector for testing: (a) whether mean sales are 500 when no advertising or in-store promotion is conducted, and (b) the effects of increasing X1 and X2 by 1 unit have the same effect on mean sales. That is, H0A: 0=500 H0B: .

QB.3. A marketing department is interested in the effects of changing advertising levels for television and internet on sales. They vary X1=TV ad $, and X2=internet ad $ and obtain the following regression results:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **X'X** |  |  |  | **X'Y** |
| **20** | **416.5343** | **406.487** |  | **3676.373** |
| **416.5343** | **9546.826** | **8733.245** |  | **78940.14** |
| **406.487** | **8733.245** | **9111.308** |  | **77022.41** |
|  |  |  |  |  |
|  |  |  |  |  |
| **(X'X)^(-1)** |  |  |  | **betahat** |
| **0.800494** | **-0.01832** | **-0.01815** |  | **98.54071** |
| **-0.01832** | **0.00127** | **-0.0004** |  | **2.093241** |
| **-0.01815** | **-0.0004** | **0.001303** |  | **2.050869** |
|  |  |  |  |  |
| **SS(Resid)** |  |  |  |  |
| **608.6247** |  |  |  |  |

Give the analysis of variance.

Set up and conduct the general linear test that the effects of changing each type of advertising are equal in terms of sales at the =0.05 significance level.

QB.4. A researcher fits a **simple linear regression** model, relating yield of a chemical process to temperature when all inputs beside temperature are at a specific level. She wishes to test the following two hypotheses simultaneously (the temperature range the experiment was conducted was: 55ºF – 85ºF):

* The average yield increases by 2 units when temperature increases by 1ºF
* The average yield is 400 when the temperature is set to 70ºF

p.4.a. For model 2, fill in the following matrix and vectors that she is testing (this is her null hypothesis):



p.4.b. She obtains the following results from fitting the regression based on *n* = 18 measurements while conducting the experiment:



p.4.c. Conduct her test at the  = 0.05 significance level.

* Test Statistic:
* Reject H0 if the Test Statistic falls in the range: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

QB.5. A researcher fits a **multiple linear regression** model, relating yield (Y) of a chemical process to temperature (X1), and the amounts of 2 additives (X2 and X3, respectively). She fits the following model:



She wishes to test the following three hypotheses simultaneously:

* The mean response when X1=70, X2=10, X3=10 is 80
* The average yield increases by 4 units when temperature increases by 1ºF, controlling for X2 and X3
* The partial effect of increasing each additive is the same (controlling for all other factors)

p.5.a. Fill in the following matrix and vectors that she is testing (this is her null hypothesis):



p.5.b. She obtains the following results from fitting the regression based on *n* = 24 measurements while conducting the experiment:



p.5.c. Conduct her test at the  = 0.05 significance level.

* Test Statistic:

Reject H0 if the Test Statistic falls in the range: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

QB.6. A research firm is interested in the effects of 4 types of advertising (**T**elevision, **R**adio, **N**ewspaper, and **I**nternet) on a firm’s sales. They hold all other variables constant over the study period (such as price and store promotion). The sample is based on n=30 sales periods. They fit the following 2 regressions based on Model 1 (note that SS(Total Corrected)=5000):



p.6.a. Test H0: T = R = N = I = 0 at  = 0.05 significance level.

Test Statistic \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Rejection Region \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

p.6.b. Set up the test of H0: T = R = N = I in the form of a general linear test by giving K’, , and m, and the degrees of freedom. Note that there are several ways K’ can be formed.

p.6.c. Test H0: T = R = N = I at  = 0.05 significance level.

Test Statistic \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Rejection Region \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

QB.7. A study was conducted, relating female heights (*Y*, in 100s of mm) to hand length (*X*1, in 100s of mm) and foot length (*X*2 in 100s of mm), based on a sample of *n* = 15 adult females. The following model was fit, with matrix results given below.



p.7.a. Set this null hypothesis in the form

H0: **K’ - m = 0**



p.7.b. Obtain the estimate of **K’ - m:**

p.7.c. Obtain **K’(X’X)-1K**

p.7.d. Obtain the estimate of 2

p.7.e. Compute the test statistic, give the rejection region, and conclusion for the test:

Test Statistic: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Rejection Region: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Reject H0? **Yes** or **No**

QB.8. A regression model is fit, relating total team payroll (Y, in millions of £s) to offensive goals scored (X1) and defensive goals allowed (X2) for the n=20 teams during the 2013 English Premier League season. For this problem, we will treat this as a sample from a population of all possible league teams.



p.8.a. Complete the following ANOVA table.



p.8.b. Test whether the offensive goals scored and defensive goals allowed effects are of equal magnitude, but opposite direction: H0: 

 Test Statistic: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Rejection Region: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Reject H0? **Yes** / **No**

QB.9. A forensic study related Hand (X1) and Foot (X2) lengths to Stature (Y) for a sample of n = 75 adult females (each variable in 100s of mms). Consider the following three models.





p.9.a. Compute  and the Total (Corrected) Sum of Squares.

p.9.b. Compute the Residual (Error) Sum of Squares for each model.

p.9.c. Compute 

p.9.d. Use the general linear test for Model 3 to test 

QB.10. Consider a sequence of regression models to be fit, each based on n observations:



p.10.a. Set-up ***P***0, the projection matrix for model 0.

p.10.b. Obtain R(0) in terms of the data *Y*1,…,*Y*n.

p.10.c. Suppose . Complete the following table:



p.10.d. Suppose SS(Total Corrected) = 1000.

* P.10.d.i. Give the proportion of variation in Y that is explained by X1 alone
* P.10.d.i.. Give the proportion of variation in Y that is not explained by X1 that is explained by X2

QB.11. A study considered noise level of the Teheran-Karaj express train (Y, in dB) in terms of distance to the center of the track (X1, in meters) and speed of the train (X2, in km/h), with n = 50. Consider the following models (in matrix form).



p.11.a. The sum of the speeds of the 50 observations is . For model 0, obtain: 



p.11.b. For Models 01, 02, and 012, you obtain the following  and *MSE*012





p.11.c. Obtain the Sequential and Partial sums of squares for X1 and X2, and their corresponding F-statistics.



QB.12. Consider the general linear test  where **K’** has q≤ p’ linearly independent rows.

p.12.a. Derive the mean vector and variance-covariance matrix of .

p.12.b. Show that  are independent. Hint: Write *Q* = **Y’AY**.

QB.13. A study was conducted, relating an abrasivity index measure (Y) to p = 4 predictors: UCS (X1), BTS (X2), and two brittleness indices: B1 (X3) and B3 (X4) in a sample of igneous rocks. The model fit is given below along with computations.

p.13.a. Compute SSE, dfE and MSE

p.13.b. Compute SSR, dfR and MSR





p.13.c. Test H0:  (abrasivity index is not associated with any of the predictors)

Test Statistic \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Rejection Region \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ P > < .05

p.13.c. Test H0:  (The coefficients for the two brittleness indices are equal)

**K’ =**  **m =**

Test Statistic \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Rejection Region \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ P > < .05

QB.14. Regression models were fit, relating various crime rates for U.S. states to a set of 25 predictors. The researchers fit the full model with all 25 predictors (say X1,…,X25) and then the best 4 predictor model (say X1,…,X4). For the outcome Total Crime, the authors report the following coefficients of multiple determination.



**Part C: Models with Qualitative Variables and Interactions**

QC.1. A linear regression model is fit, relating apartment rental prices (Y, in $100) to square footage for for 5 apartments in each of 4 luxury neighborhoods (all apartments were built in the same decade). We consider the following 3 models, where X1 is the square footage (100s of ft2); X2 =1 if neighborhood A, 0 otherwise; X3=1 if neighborhood B, 0 otherwise; and X4=1 if neighborhood C, 0 otherwise.



The ANOVA Tables from each model are given below.



p.1.a. Test whether the “square footage effect” is the same for each neighborhood by completing the following parts (homogeneity of regressions):

p.1.a.i. H0: HA:

p.1.a.ii. Conduct the test

Test Statistic \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Rejection Region \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

p.1.b. Assuming no interaction between neighborhood and square footage, test whether the neighborhoods have different means, controlling for square footage by completing the following parts (homogeneity of regressions):

p.1.b.i. H0: HA:

p.1.b.ii. Conduct the test

Test Statistic \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Rejection Region \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

QC.2.Write the (full rank, additive) multiple regression equation for determining if the linear relationship of Y = response time as a function of X = strength of signal has the same slope for three groups. Define all variables.

QC.3. A regression model is fit, relating time to complete a task (Y, in minutes) to nationality of the team (X1=1 if US, 0 if non-US) and complexity (X2, on a TACOM scale) for nuclear power plant operators. The model fit is:





p.3.a. Test whether the slopes (with respect to complexity scores) are equivalent for US and non-US power plants.

H0: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ HA: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Test Stat: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ P-Value: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

p.3.b. Give the estimated mean time to complete a task with complexity of X2 = 5 for US and non-US plants.

US: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ non-US: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

p.3.c. Compute a 95% Confidence Interval for the difference of the means estimated in p.3.b.

QC.4. Regression models were fit, relating height (Y, in mm) to hand length (X1, in mm), foot length (X2, in mm) and gender (X3=1 if male, 0 if female) based on a sample of 80 males and 75 females. Consider these 4 models:





p.4.a. Confirm the equivalence of the regression coefficients (but not standard errors) based on the appropriate models (Hint: set up the fitted equations based on the two models):

Females:

Males:

p.4.b. Test H0: = 0 (No interactions between Hand and Gender or Foot and Gender).

Test Statistic: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Rejection Region: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ p-value **>** or **<** 0.05?

p.4.c. Use Bartlett’s Test to test whether the error variances among the individual regressions are equal:



Test Statistic B = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Rejection Region: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ p-value **>** or **<** 0.05?

p.4.d. What fraction of the total variation in height is explained by the set of predictors: hand length, foot length, and gender (but no interactions)?

p.4.e. Compute the standard deviations among the 80 Male heights and among the 75 Female heights (ignoring hand and foot length).

Males: SD = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Females: SD = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

QC.5. A study was conducted to determine whether having been exposed to an advertisement claiming a natural ingredient is contained in a perfume had an effect on subjects’ rating of the perfume’s scent. There were 112 subjects of which, 56 were exposed to the ad, and 56 were not. We fit the following regression model:





p.5.a. First, we fit a model with only an intercept term, what will  be (symbolically, do not write out a 112x112 matrix!)? Compute R(0).

P0 = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ R(0) = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

p.5.b. Compute 

p.5.c. Compute R(0 , 1) , R(0), and MSResidual

R(0 , 1) = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ R(0) = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ MSResidual = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

p.5.d. Use the t-test and the F-test to test H0: 0 vs HA≠

t-Statistic: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Rejection Region: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

F-Statistic: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Rejection Region: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

QC.6. A regression model is fit, relating Weight (Y, in pounds) to Gender (X1=1 if Male, 0 if Female) and Height (X2, in inches) among professional NBA and WNBA players. The model fit is:





p.6.a. Test whether the slopes (with respect to height) are equivalent for male and female pro basketball players.

H0: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ HA: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Test Stat: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ P-Value: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

p.6.b. Give the estimated mean weight for a player with height X2 = 72 inches for male and female pro basketball players.

Male \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Female: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

p.6.c. Compute a 95% Confidence Interval for the difference of the means estimated in p.6.b.

QC.7. A study measured Total Mercury levels (Y, in mg/g) in a sample of n=135 Kuwaiti men. The independent variables were: X1=1 if fisherman, 0 if not; X2 = Weight (kg); and X3 = # Fish Meals/Week. The matrix results are given below.



p.7.a Complete the following Analysis of Variance table.



p.7.b. Obtain a 95% Confidence Interval for the effect of being a fisherman on expected total Mercury, controlling for Weight and Fish Meals/Week.

p.7.c. What proportion of the variance in Total Mercury is “explained” by this set of predictors?

QC.8. A regression model is fit, relating weight (Y, in pounds) to height (X1 in inches) and gender (X2=1 if male, 0 if female) among a random sample of NBA/WNBA basketball players. The relationship between weight and height is fit first, separately for males and females, then combined in the model: 

p.8.a. Complete following table and test  using Bartlett’s Test.





Test Statistic \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Rejection Region \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

p.8.b. The following (partial tables) include the estimated coefficients and standard errors for the males and females separately as well as the combined model. Complete the tables.



p.8.c. A final model is fit among all players, relating Weight to Height without including gender or the interaction.

 Test 

Test Statistic \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Rejection Region \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

QC.9. A regression model was fit based on a sample of n=117 Black Holes. The response was Bolometric Luminosity (Y), with predictors: Black Hole Mass (X1) and Black Hole Type (X2 = 1 if Radio Quiet Quasar (RQQ), 0 if Radio Loud Quasar (RLQ)), and a cross-product term to allow for a possible interaction between Mass and Type. Note that there were 20 RQQ and 97 RLQ Black Holes.



p.9.a. Based on Model 1, give the fitted equations relating Bolometric Luminosity to Mass, seperately by Quasar Type.

RQQ: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ RLQ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

P.9.b. Test whether the true relationship between Bolometric Luminosity and Mass is the same for RQQs and RLQs.

H0: HA:

Test Statistic: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Rejection Region: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ P-value **>** or **<** .05

p.9.c. When the regressions are fit seperately, the fitted equations are the same as you should have in part p.7.a. The residual variances (MSE’s) for the models are: RQQ: sQ2 = 0.088 RLQ: sL2 = 0.309. Use Bartlett’s test to test whether the true variances are equal.



Test Statistic: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Rejection Region: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ P-value **>** or **<** .05

**Part D: Models with Curvature and Response Surfaces**

QD.1. A second-order response surface is fit with 2 independent variables (including all main effects, cross-product, and squared terms) and n=20 observations. Give the degrees of freedom for regression and residual, as well as the rejection region for testing H0: E{Y} = 0

df(Regression) = \_\_\_\_\_\_\_\_ df(Residual) = \_\_\_\_\_\_\_\_\_\_ Rejection Region: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

QD.2. A regression model is fit, relating the number of breeding pairs of penguins to the year, over a period of years. The researchers use *Y*= log10(# breeding pairs) and *X* = (Year – mean(Year)) . They fit 3 Models:





p.2.a. Compute R(0), R(0, ), R(0, 2), R(0), and R(0, ) (use 4 decimal places)

R(0) = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ R(0, ) = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ R(0, 2)\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

R(0) = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ R(0, ) = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

p.2.b. Compute the fitted values and residuals for the following years, for each model:



QD.3. An experiment to study the effect of temperature (*x*) on the yield of a chemical reaction (*Y*), was conducted. There was a total of n = 30 experimental runs, each using one of 2 catalysts (*z*=0 if catalyst 1, *z*=1 if catalyst 2). There were 5 evenly-spaced temperatures, coded as *x* = -2, -1, 0, +1, +2. There were 3 replicates per temperature/catalyst. The model fit was:



You are given the following results: 

p.3.a. Test whether there is evidence of difference in catalysts, controlling for temperature.

H0: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ HA: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Test Stat: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Rej. Region: \_\_\_\_\_\_\_\_\_\_\_\_\_\_

p.3.b. Can we conclude that the relationship is not linear? Obtain a 95% Confidence Interval for the relevant parameter, and interpret.

Confidence Interval \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Conclude that the relation is linear? **Yes** or **No**

p.3.c. Obtain the estimated mean yield when catalyst 2 is used and at the standard temperature (*x* = 0), and compute a 95% CI for the mean.

Point Estimate: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 95% CI: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

p.3.d. At what (centered) temperature do you estimate the yield to be maximized?

QD.4. A response surface was fit, relating (coded) Nitrogen (*XN*), Phosphorous (*XP*) and Number of Days (*XD*) on the percent crude oil removed from an experimental oil spill (*Y*). The following 3 models were fit, based on *n* = 20 experimental spills:



p.4.a. Use Models 1 and 2 to test whether any of the interaction terms are significant, after controlling for main effects: 

Test Statistic \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Rejection Region \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Reject H0? **Yes** or **No**

p.4.b. Use Models 2 and 3 to test whether any of the quadratic terms are significant, after controlling for main effects and interactions: 

Test Statistic \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Rejection Region \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Reject H0? **Yes** or **No**

p.4.c. The coded and actual levels are given below. The model was fit based on the coded values (-1, 0, 1) and several axial points.



Give the actual levels, corresponding to the models’ intercepts: Nit = \_\_\_\_\_\_\_, Phos = \_\_\_\_\_\_\_, Days = \_\_\_\_\_\_

QD.5. A study related Personal Best Shot Put distance (Y, in meters) to best preseason power clean lift (X, in kilograms). The following models were fit, based on a sample of n = 24 male collegiate shot putters:



p.5.a. Use Model 2 to test H0: ****(Y is not related to X)

Test Statistic\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Rejection Region: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Reject H0? **Yes** or **No**

p.5.b. Use Models 1 and 2 to test H0: **(Y is linearly related to X)

Test Statistic: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Rejection Region: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Reject H0? **Yes** or **No**

p.5.c. Give an estimate of the level of *X* is that maximizes *E*{*Y*}.

X\* = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

QD.6. A study related Freight Volume (Y) in Shanghai to GDP (X1) and Fixed Investment (X2) over a period of n = 11 years. The authors fit the following 3 models:





p.6.a. Compute SSTotalCorrected

p.6.b. Compute SSRegression and SSResidual for each model.

p.6.c. Compute 

p.6.d. Test H0: 11 = 0 vs HA: 11 ≠ 0 (Note there are 2 ways of doing this).

p.6.e. What proportion of the variation in Y that is not explained by X1 is explained by X2?

QD.7. Show that for simple regression, when we have *n*i observations at the ith distinct level of X, the Pure error sum of squares can be written as



p.7.a. An experiment was conducted to study the relationship of between yield from a chemical reaction (y) and the reaction temperature (x). The following data were obtained from n=12 runs. The fitted equation based on n=12 runs was Y-hat = 92.68-0.15x. Complete the table by filling in values for X=100.



p.7.b. Conduct the Lack-of-Fit F-test by completing the following table: (H0: Linear Model is appropriate)



p.7.c. Do you reject the hypothesis that a linear fit is appropriate at the 0.05 significance level? **Yes / No**

p.7.d. Based on the same dataset, a quadratic model is fit based on the original (non-centered) X values:

E(Y) = 0+1X+2X2

* Give the fitted value for X=100
* Compute simultaneous 95% CIs for 1 and 2 based on Bonferroni’s adjustment.

****

* Based on your simultaneous CIs give an approximate confidence interval for the X value where Y is maximized.

QD.8. A response surface model related yield of methyl ester from waste canola oil (Y) to 3 factors: Time (X1, in minutes (15,30,45)), Temperature (X2, in degrees Celsius (240,255,270)) and Methanol/Oil Ratio (X3 (1,1.5,2)).

Two models are considered are given below, along with the partial ANOVA tables, based on n=19 cases.



p.8.a. Complete the ANOVA Tables.



p.8.b. Test whether all terms that include X3 can be excluded from the model.

Null Hypothesis:

Test Statistic: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Rejection Region: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

p.8.c. The fitted equation for model 2 can be written as follows. Solve for the value **x\*** that maximizes the response (in terms of **B**1 and **B**2)?



p.8.d. The estimated regression coefficients, **B**1 and **B**2 and **B**2-1 are given below. Obtain the optimum levels of Temperature (X1) and Time (X2).



QD.9. A second-order response surface is to be fit with 4 predictors. How many experimental runs will be needed so that the full model will have 20 Error degrees of freedom?

**Part E: Model Building**

QE.1. A regression model is to be fit, relating mean January High temperature (Y) to 3 potential predictors (ELEVation, LATitude, and LONGitude). The following results are obtained:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  | *df* | *SS* | *MS* |
|  | *b* | *SE(b)* | *t Stat* | *P-value* | Regression | 1 | 2311.213 | 2311.213 |
| Intercept | 60.94146 | 0.363461 | 167.6699 | 0 | Residual | 367 | 8428.007 | 22.9646 |
| ELEV | -0.00188 | 0.000187 | -10.0321 | 4.32E-21 | Total | 368 | 10739.22 |  |
|  |  |  |  |  |  |  |  |  |
| Intercept | 129.2194 | 1.1704 | 110.4105 | 0.0000 | Regression | 1 | 9768.30 | 9768.30 |
| LAT | -2.2656 | 0.0373 | -60.7646 | 0.0000 | Residual | 367 | 970.92 | 2.65 |
|  |  |  |  |  |  |  |  |  |
| Intercept | 87.8721 | 9.7737 | 8.9906 | 0.0000 | Regression | 1 | 261.75 | 261.75 |
| LONG | -0.2990 | 0.0988 | -3.0280 | 0.0026 | Residual | 367 | 10477.47 | 28.55 |
|  |  |  |  |  |  |  |  |  |
| Intercept | 129.5893 | 1.2998 | 99.7002 | 0.0000 | Regression | 2 | 9769.44 | 4884.72 |
| ELEV | 0.0000 | 0.0001 | 0.6563 | 0.5120 | Residual | 366 | 969.78 | 2.65 |
| LAT | -2.2796 | 0.0430 | -53.0544 | 0.0000 |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| Intercept | 105.6653 | 6.5092 | 16.2332 | 0.0000 | Regression | 2 | 854.10 | 427.05 |
| ELEV | 0.0023 | 0.0001 | 16.2047 | 0.0000 | Residual | 366 | 1049.01 | 2.87 |
| LONG | -0.7846 | 0.0676 | -11.6090 | 0.0000 |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| Intercept | 117.5506 | 2.9467 | 39.8925 | 0.0000 | Regression | 2 | 9814.92 | 4907.46 |
| LAT | -2.3027 | 0.0374 | -61.5047 | 0.0000 | Residual | 366 | 924.30 | 2.53 |
| LONG | 0.1297 | 0.0302 | 4.2967 | 0.0000 |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| Intercept | 57.9594 | 7.2895 | 7.9510 | 0.0000 | Regression | 3 | 9976.42 | 3325.47 |
| ELEV | -0.0014 | 0.0002 | -8.7908 | 0.0000 | Residual | 365 | 762.80 | 2.09 |
| LAT | -2.0491 | 0.0446 | -45.9090 | 0.0000 | Total | 368 | 10739.22 |  |
| LONG | 0.6718 | 0.0675 | 9.9520 | 0.0000 |  |  |  |  |

p.1.a. Based on Stepwise Regression with SLS=SLE=0.05, what will be the sequence of models selected and the final model. Give BRIEFLY the reason for each step.

p.1.b. Compute SBC and Cp for the 3 2-variable models and the 3-variable model (Cp for the 3-variable model will be 4 by definition). Based on each criteria which model is selected?



ELEV, LAT Cp= \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ SBC = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

ELEV, LONG Cp= \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ SBC = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

ELEV, LAT Cp= \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ SBC = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

ELEV, LAT, LONG Cp= \_\_\_\_ SBC = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

p.1.c. Which model will have the highest adjusted-R2?

p.1.d. Give the Sequential and Partial sums of squares for each variable (for the ordering: ELEV, LAT, LONG) by completing the following table:



QE.2. A study looked at the relationship between stack loss (Y, a measure of ammonia escaping a process), and 3 protential predictors: airflow (Air), cooling temperature (temp), and acid concentration (acid).

E(Y) = 0 + AirAir +TempTemp +AcidAcid

p.2.a Complete the following table where:





p.2.b. Which model is best by each of the following criteria? Why do you choose that model for that criteria?

p.2.b.i. Adjusted-R2:

p.2.b.ii. Cp:

p.2.b.iii. AIC:

p.2.c. Give the following sums of squares:

p.2.c.i. R(Air | Intercept, Temp, Acid):

p.2.c.ii. R(Temp, Acid | Intercept, Air)

p.2.d. Test H0: Temp = Acid = 0 versus HA: Temp and/or Acid ≠ 0 at the  = 0.05 significance level:

p.2.d.i. Test Statistic:

p.2.d.ii. Rejection Region:

QE.3. A potentially cubic regression model is fit, relating Y to X. We get the following fits for all possible models:



p.3.a. We fit a Stepwise Regression Model with SLE = SLS = 0.20

p.3.a.i. What variable is entered at Step 1? Why?

p.3.a.ii. What happens at Step 2? Why?

p.3.a.iii. What happens at Step 3? Why?

QE.4a. True or False: In all possible regressions, the model chosen based on R2-Adj criterion will always give the model as the MS(Residual) criterion. **True** or **False**

QE.4b. True or False: In Stepwise Regression, it is possible for a predictor to enter a model at an early stage, then be removed at a later stage. **True** or **False**

QE.5. All possible regressions are fit among models containing 3 potential independent variables (X1= quay cranes/berth, X2 = terminal (yard) cranes/berth, and X3 = berth length. The response is the Throughput/berth (Y in 1000s of TEU). The models are based on a sample n=15 Chinese ports.

 

p.5.a. Compute SBC for the model with X1 as the only predictor.

p.5.b. Compute Adjusted-R2 for the model with X1 and X3.

p.5.c. Compute Cp for the model with X2 and X3.

p.5.d. Will Cp choose model (X2,X3) or model (X1,X2,X3)? Why?

QE.6. A series of models were fit, relating Average January High Temperature (Y, in degrees F) to Elevation (X1, in 100s ft above sea level), and Latitude (degrees North Lat) for n = 369 weather stations in Texas. Latitude and Elevation were centered in the regression models.





p.6.a. Complete the table.

p.6.b. Based on each criteria, which model do you choose?

Cp: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ AIC: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ SBC: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

QE.7. Regression models are fit, relating price of Compact Hybrid Cars (Y, in $1000s) to Acceleration (X1, in km/hour/sec) and Miles per Gallon (max of Gas and Electric mph) for n=25 models from years 2009-2013.

Consider the following models with Residual Sums of Squares for each model (SSTotalCorr = 2196)



Note:



p.7.a. Compute Cp for Model 1.

p.7.b. Compute AIC for model 2

p.7.c. Compute SBC for models 3 and 4. Which model is preferred based on that criteria?

QE.8. A model is fit relating January Mean Temperature (Y, in Fahrenheit) to Elevation (X1, in 100s of feet above sea level) and Latitude (X2 = Degrees North Latitude – 30) for a random sample of n = 29 weather stations in Texas. Two models are fit:



The matrix form for Model 1 is given below.



p.8.a. Compute  the Error and Regression Sums of Squares, and the estimate of  for Model 1.



p.8.b. Compute a 95% Confidence Interval for 2

p.8.c. For Model 2, we get the following results. Obtain the fitted value for a location with an Elevation of 500 feet above sea level and at 30 degrees North Latitude based on each model. Note the units with which X1 and X2 been “operationalized” with.



Model 1 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Model 2 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

p.8.d. Test 

Test Statistic \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Rejection Region \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

p.8.e. Compute AIC for each model. Which model do you select based on the AIC criteria. 

QE.9. Over Michael Jordan’s (Pro Basketball player, not UCBerkley Stat/CS Professor) career, he played 15 seasons. A plot of his average Points per 48 Minutes (regulation game) versus season is given below.

Consider the following orders of polynomials models (treating his seasons as a random sample of all seasons he could have played over the same ages / physical conditions): 2nd, 4th, 5th, 6th .

Note: 

Complete the following table. Which order model is selected by Cp? by BIC? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_



QE.10. A study relating shipping fuel use (Y, tons/day) to speed (X, knots) for container ships travelling between Tokyo and Xiamen was fit by the following regression models based on n = 20 experimental runs.



p.10.a. Complete the following table.



Note:



p.10.b. Which model would you choose based on the 3 criteria? Cp \_\_\_\_\_\_\_\_\_\_\_\_ AIC \_\_\_\_\_\_\_\_\_\_ BIC\_\_\_\_\_\_\_\_\_\_\_\_

**Part F: Multicollinearity**

QF.1. True or False: When the independent variables have been set up in a controlled experiment to be uncorrelated among themselves, the Variance Inflation Factor for each predictor will be 0. **True** or **False**

QF.2. A regression model is fit, relating January mean temperature (Y) to ELEVation, LATitude, and LONGitude for n=369 weather stations in Texas (the data are aggregated over a period of years). The following table gives the Regression and Residual sums of squares for each model. All models contain an intercept.



p.2.a. Compute R(LONG), R(LONG|ELEV), R(LONG|LAT), R(LONG|ELEV,LAT)

p.2.b. Based on the simple linear regression, relating Y to LONG: E{Y} = 0 + LONGLONG, test

H0: LONG = 0 HA: LONG ≠ 0

Test Statistic: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Rejection Region: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

p.2.c. Based on the simple linear regression, relating Y to ELEV, LAT, LONG:

E{Y} = 0 + ELEVELEV+ LATLAT + LONGLONG, test H0: LONG = 0 HA: LONG ≠ 0

Test Statistic: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Rejection Region: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

p.2.d. The Coefficient of Determination, when LONG is regressed on ELEV and LAT is 0.843. Compute the Variance Inflation Factor for LONG.

**Part G: Weighted and Generalized Least Squares**

QG.1. A regression model is fit, where X is the dose individuals receive and Y is a measure of therapeutic response. The following table gives the sample size, mean, and variance for each level of dose (the overall mean is 25). Consider the simple linear regression model Y=X



p.1.a. Obtain the weighted least squares estimate of 

p.1.b. Obtain the fitted values in the original scale: 

p.1.c. Test whether the relationship between dose and response is linear (=0.05).

QG.2. A regression model is fit, where X is the dose individuals receive and Y is a measure of therapeutic response. The following table gives the sample size, mean, and variance for each level of dose (the overall mean is 25). Consider the simple linear regression model Y=X



p.2.a. Obtain the weighted least squares estimate of 

p.2.a.1*W* =

p.2.a.ii. *X*\* =

p.2.a.iii. *Y*\* =

p.2.a.iv. 

p.2.b. Obtain the fitted values in the original scale: 

QG.3. A study related the spread in shotgun pellets (Y, sqrt(AREA)) to distance shot (X, in feet) for a particular shotgun and bullet type (different from the in-class example).

p.3.a. Based on the residuals versus fitted plot (Plot 2 on page 1), which violation(s) if any is/are violated?

i) Linearity ii) Equal Variances iii) Independence

p.3.b. We wish to fit an Estimated Weighted Least Squares Regression of the following model:

The data are given below, give the form of **W** used in estimating 





p.3.c. From the results below, complete the following table. (SSE\* based on transformed residuals)



QG.4. A regression is fit, relating Gross Profits (Y, in $100M) to amount bet in Slot Machines (X1, in $1B) and amount bet on table games (X2, in $1B) for all Atlantic City casinos annually for 1978-2012 (n=35). The results for the regression coefficients and their standard errors are given below, for the model fit by Ordinary Least Squares.

 

p.4.a. Compute the test statistic and give the rejection region for testing H0: 1 = 0 vs HA: 1 ≠ 0

Test Statistic: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Rejection Region: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

p.4.b. Compute a 95% Confidence Interval for 2.

p.4.c. The Durbin-Watson test results in strongly rejecting the null hypothesis H0:  = 0 (See plot 2, page 1). A transformation, that when applied to the X-matrix and Y-vector produces uncorrelated errors: :



Compute the first 2 and last rows of the estimated transformed vector and matrix: **T-1Y** and **T-1X**

p.4.d. The results from the estimated Generalized least squares fit are given below. Repeat parts p.3.a. and p.3.b. using them.



****

QG.6. A study compared surveyed velocities measured for n = 9 rock glaciers in Western Canada (Y, in meters/year) and their long term velocities, based on glacier length and age (X, in meters/year). The number of surveyed points on the glaciers varies, and **Y is the mean of those measurements**. Assume the variances of the individual points are all equal, and equal to 2. Note the “effective” sample size in 9 here, since we are working with the 9 means as the data.



p.6.a. Give the variance of Y5 (as a function of 2) :

p.6.b. We want to obtain the Weighted Least Squares Estimator: 

Set up the **W** matrix (note that **X** also has a column of 1s for the intercept).

p.6.c. Based on the Ordinary Least Squares (middle portion), and Weighted Least Squares (right-side portion), obtain estimated standard errors for  and 95% Confidence Intervals for 1:

OLS: Standard Error: \_\_\_\_\_\_\_\_\_\_ CI:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

WLS: Standard Error: \_\_\_\_\_\_\_\_ CI:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

QG.7. A study is conducted, relating dose of a weight loss drug (X) to weight reduction (Y). There were n = 4 doses, with varying numbers of subjects per dose. The doses were X = 1, 2, 3, 4, and the sample sizes were r = 6, 8, 2, 4, respectively. A simple linear regression model, relating Y to X, with the error variance for individual subjects being 2. Observations are independent across subjects.

p.7.a. Set up the matrix computations to obtain the variance-covariance matrix for the weighted least squares estimator as a function of 2. All numbers, no symbols.

p.7.b. Set up the matrix computations to obtain the variance-covariance matrix for the ordinary least squares estimator as a function of 2. All numbers, no symbols.

QG.7. An economic history report studied the relationship between marginal product of labour (MPL) and estimated population of sampled communities (Pop) over an n=28 decade period from 1250-1529. Due to theoretical reasons, the model was fit on the log scale, with Y=ln(MPL) and X=ln(Pop). The author fit **weighted least squares**, where the **weight used was the number of communities (C)** which were used for population estimates by decade (the more communities used, the better the overall estimate, and C ranged from 2 to 14). The model fit was: E{**Y**} = **X**.



The following results are obtained for the transformed **Y**\* and **X**\*. Complete the following table that would be obtained by standard software packages that fit WLS. Hint: in the unweighted case, 



QG.8. A regression model, relating mean temperature (Y, in F) to Year-1946 (X) for Years 1946-2014 has a Durbin-Watson statistic of DW = 0.628. Two regression models were fit, one with ordinary least squares (OLS) and one with estimated generalized least squares (GLS) with an AR1 process for errors.

Which of the following sets of estimates and standard errors do you believe are OLS and GLS?

