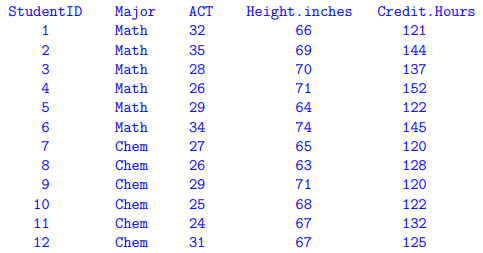
**STAT5120, Final Exam, Allen Baumgarten**

1. [40 POINTS] ACME University administrators want to predict the total number of credit hours a student will complete based on their majors in college, their ACT (standardized admission test) scores, their high school GPAs, and their heights in inches. Data on twelve randomly sampled students from last year are below.



(a) [5 POINTS] Write down the model matrix for these data (the X matrix) for finding the least squares coefficient estimates for a linear model for predicting the total number of credit hours a student will take based on major (Math or Chem), ACT score, and height in inches.

Our X matrix would be:

One’s X1 X2 X3

1 Math 32 66

1 Math 35 69

1 Math 28 70

1 Math 26 71

1 Math 29 64

1 Math 34 74

1 Chem 27 65

1 Chem 26 63

1 Chem 29 71

1 Chem 25 68

1 Chem 24 67

1 Chem 31 67

(b) [5 POINTS] Write down the Y vector for these data used to find the least-squares coefficient

estimates.The Y vector would be:

121

144

137

152

122

145

120

128

120

122

132

125

Based on this output, AND based on common sense, which variable(s) do you suggest we keep in the model and which could possibly be discarded? Explain. Remember our goal is to be able to predict the number of credit hours a student will take in college.

Using R, we enter this data into the console and run the lm() function to obtain a basic least squares model as follows:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 24.0662 59.4368 0.405 0.6962

question1$MajorMath 11.1286 6.1092 1.822 0.1060

question1$ACT -0.7349 0.8985 -0.818 0.4371

question1$Inches 1.7996 0.8827 2.039 0.0758 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

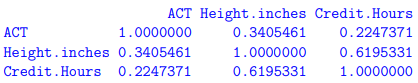
Residual standard error: 8.7 on 8 degrees of freedom

Multiple R-squared: 0.5646, Adjusted R-squared: 0.4014

F-statistic: 3.458 on 3 and 8 DF, p-value: 0.07128

My common sense (or lack thereof) notwithstanding, examination of our preliminary results above shows that there is only somewhat statistically significant predictor credit ours based on its somewhat low p-value. This significant predictor is to keep Major. The predictor Height in Inches has a lower p-value but height, insofar as we know, as no bearing on credit hours taken (unless those credit hours for the college basketball team). We should keep Major as its p-value is somewhat significant at .106

(d) [5 POINTS] The pair-wise correlations for all three variables, including the response, are



Based on this, are there signs of collinearity in the predictor variables? Explain.

Yes, there may be somewhat moderate collinearity between these predictors, specifically, between Height in Inches and Credit Hours at .6195

(e) [5 POINTS] Use the full model output from R in part (c) above to predict the total number of credit hours a student will take in college who is 70 inches tall, had an ACT score of 30, and who majors in math. Using the full model, we would say the predicted outcome for y would be based on:

y = 24.07 + 11.13(Major) + -.735(ACT) + 1.8(Inches)

y = 24.07 + 11.13(Major) + -.735(30) + 1.8(70)

y = 24.07 + 11.13 – 22.05 + 126

y = 139.15

(f) [5 POINTS] Use the full model output from R in part (c) above to predict the total number of credit hours a student will take in college who is 66 inches tall, had an ACT score of 28, and who majors in chemistry.

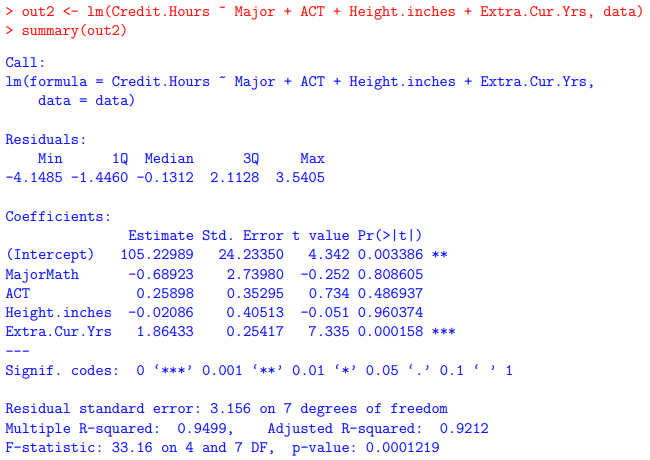
y = 24.07 + 11.13(Major) + -.735(ACT) + 1.8(Inches)

y = 24.07 + 11.13(Chem) + -.735(28) + 1.8(60)

y = 24.07 + 0 – 20.58 + 108

y = 111.49

(g) [5 POINTS] A new variable, the number of extra curricular activities in high school (measured in years) is discovered, and the output for a new model is given below:

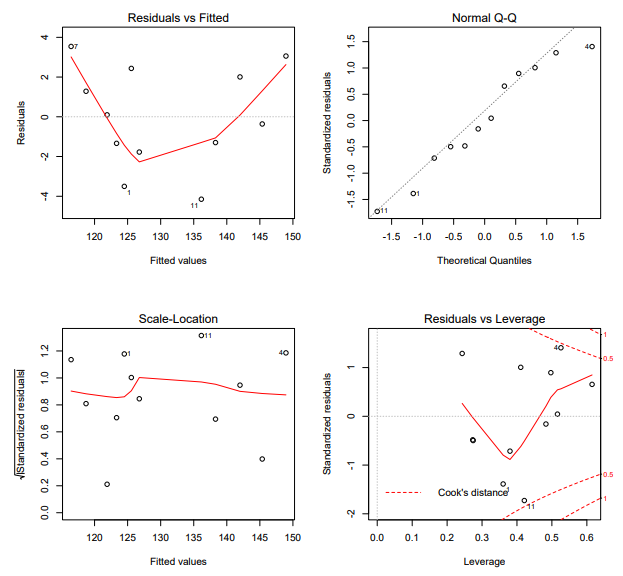


Remark on which variables now seem important, what has changed and why. Specifically address

(1) the change in p-values for the t-tests, The t-tests and corresponding p-value changed dramatically for MajorMath variable. Earlier it was somewhat significant with a p-value of .106 but now has change to a p-value of .808, indicating no significance. The t-test and p-value remain mostly unchanged for the ACT variable while the Height in Inches variable climbed from a formerly somewhat significant p-value to a quite insignificant p-value of .96. The final new variable has the significant p-value.

(2) the new Multiple/Adjusted R-squared values, and (3) the values of the beta estimates. Also (4) which (if any) variables would you suggest dropping from the model? The former multiple and adjusted R-square statistics were somewhat helpful in accounting for about 56% and 40% of the variation in y, respectively. Now, in this new model they are extremely high, accounting for the variation in y by 95% and 92%, respectively. I would say based on this new data, we should drop the first three available predictors and keep the final new predictor Extra.Cur.Yrs with a p-value of .000158.

(h) [5 POINTS] Use the diagnostic plots below to remark on whether or not the assumptions have been adequately met. If anything seems wrong, be sure to suggest some kind of appropriate remedial measure.



Looking at these four plots, our Residuals vs. Fitted values shows signs of a curvilinear relationship between the predictors and the outcome variable. A transformation to straighten this out would be suggested. Perhaps a square-root transformation would help. The Normal QQ plot also shows an outlier in the upper right area of the plot. This outlier should be investigated and if it truly belongs in the data, perhaps a transformation to center and standardize the data may be in order.

2. [5 POINTS] Write down the general matrix equation for finding the least-squares coefficient estimates. Don’t write any data values here: write the matrix formula.

(Y-Xb)T(Y-Xb)

3. [5 POINTS] Write down the general matrix equation for the hat matrix H in terms of the X matrix.

XT \* -XT

4. [5 POINTS] Write down the matrix equation for the residuals in terms of the hat matrix H and the matrix Y.

(Y-Xb)T - (Y-Xb)

5. [5 POINTS] Write down the formula for the linear correlation coefficient r when there is just one predictor variable.

Σ (x - x̅)(y - y̅) / (x - x̅)2

6. [5 POINTS] Write down the simple little formula (in terms of r and sample standard deviations) for the simple linear regression slope estimate when there is just one predictor variable.

We need for a linear model 4 things to derive our two beta coefficients:

1. x̅
2. y̅
3. (x - x̅)(y - y̅)
4. (x - x̅)2

Now, the betas are:

1. β1 = [ (x - x̅)(y - y̅) ] / (x - x̅)2
2. β0 = y̅ + β1(x̅)

7. [5 POINTS] What are the assumptions that need to be checked when constructing a linear model?

We assume that the errors (residuals) have a mean of zero with a standard deviation s2. We also assume independent random data.

8. [15 POINTS] Complete the following lack-of-fit ANOVA table:

Source df SS MS F\* p-value

Regression 2 34.783 17.3915 10.52 .021

Residual 3 4.957 1.6523

Lack-of-Fit 13 39.74

Pure Error 16 2.110

Total 21 41.85

9. [5 POINTS] Which is used in the case of nonconstant error variance when the errors are uncorrelated?

(a) Durbin-Watson test (b) Shapiro-Wilk test (c) weighted least-squares (d) none of these

weighted least-squares

10. [5 POINTS] Which is to be used when errors are found to be correlated, and can also help mitigate nonconstant error variance?

(a) Durbin-Watson (b) Shapiro-Wilk (c) generalized least-squares (d) none of these

none of these

11. [5 POINTS] How can you check for correlated residuals?

(a) Durbin-Watson (b) run cor() on ‘neighboring’ residuals (c) plot residuals vs. fits (d) all of these

plot residuals vs. fits

12. [5 POINTS] How can you check for normality of residuals?

(a) Anderson-Darling (ad.test) (b) Shapiro-Wilk (shapiro.test) (c) normal probability plot (qqnorm) (d) all of these

all of these

13. [5 POINTS] How can you check for equal variance of the residuals?

(a) plot residuals vs. fits; (b) Levene’s test on a partitioning of the data; (c) regress p |residuals| vs. predictor variable and check p-value associated with the slope; (d) all of these.

all of these

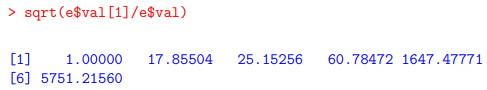
14. [5 POINTS] True or False: Cook distances measure the changes in fitted values when removing

data points and so are used to identify potential influential observations. TRUE

15. [20 POINTS] The four problems on the next page refer to the following R output.

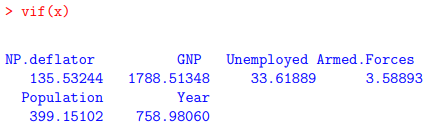
(a) [5 POINTS] In the output on the previous page, circle the condition numbers.

These are the condition numbers:



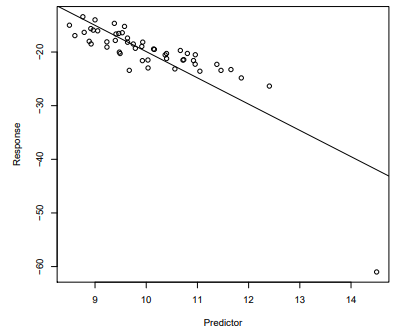
(b) [5 POINTS] Do the condition numbers seem to indicate anything about some of the variables in particular? Explain. Other than the fact that they are in good condition (sorry) the final two numbers of 1647 and 5751 indicate collinearity in the data.

(c) [5 POINTS] In the output on the previous page, draw a rectangle around the variance inflation factors. These are the variance inflation factors:

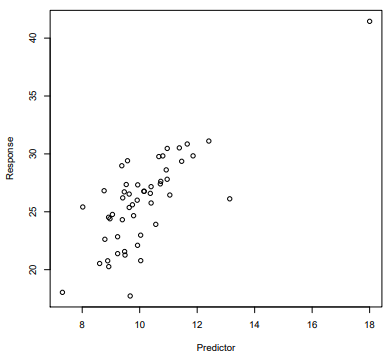


(d) [5 POINTS] Do the variance inflation factors seem to indicate anything about some of the variables in particular? Explain. Yes, the GNP has outliers greatly influencing the model.

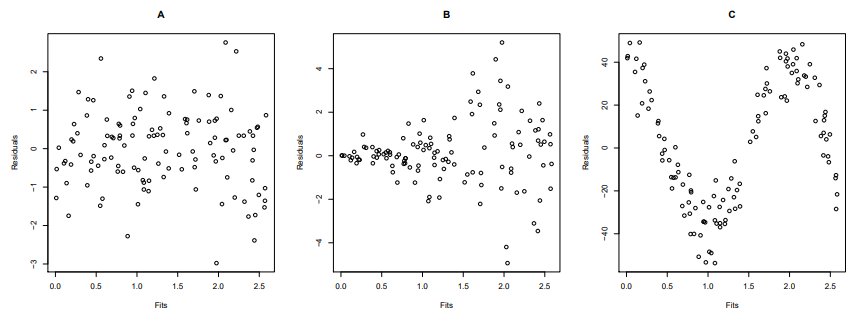
16. [5 POINTS] Circle the best candidate for an influential observation.



17. [5 POINTS] Circle the best candidate for a leverage point.



18. [20 POINTS] The following three plots were made in conjunction with three different linear models of the form lm(y ∼ x1 + x2).



Use the plots to answer these questions.

(a) [5 POINTS] Which plot (A, B, or C) most indicates constant error variance?

(a) A (b) B (c) C

(b) [5 POINTS] Which plot (A, B, or C) most indicates the error variance is increasing in the fitted values?

(a) A (b) B (c) C

(c) [5 POINTS] Which plot (A, B, or C) most indicates a nonlinear model might be more appropriate than the one made with lm(y ∼ x1 + x2)?

(a) A (b) B (c) C

(d) [5 POINTS] How would you recommend correcting for the problem in plot B? Weighted Least squares might be an option, or a Box-Cox transformation.

19. [5 POINTS] When doing principle components analysis (PCA) or principle components regression (PCR), why is it typically important to standardize the predictor variables? This is used to correct for unequal variance between the predictors or other problems.

20. [5 POINTS] What is the objective function to be minimized in ridge regression?

λ|e|

21. [5 POINTS] What is the objective function that is to be minimized in lasso regression?

-λ|e|

22. [5 POINTS] How are the tuning parameters found when doing ridge and lasso regression? We minimize the lambda function

23. [5 POINTS] Match the following link functions to their names:

 This is the Probit link

 This is the complimentary log-log

 This is the logit function

24. [5 POINTS] Match the following goodness-of-fit measures to their names.

 This is adjusted R2

 This is SSE

 This is Mallow’s Cp

 This is BIC

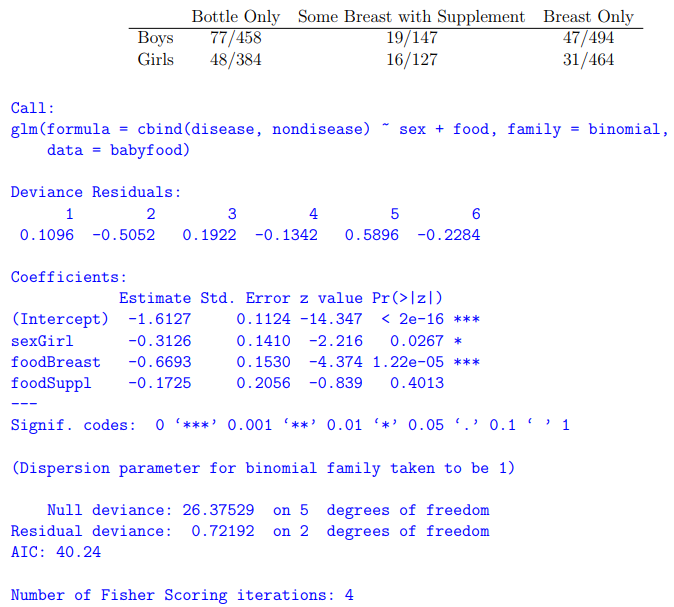
 This is AIC

25. [5 POINTS] What is the null deviance and what should you do if it is large? The null deviance is the deviance of the null model with no predictors, only the intercept term included. If it is large, our model does not fit well.

26. [5 POINTS] What is the residual deviance and what should you do if it is large? The residual deviance is the fit of the model errors based on a chi-square distribution with n-1 df. If it is much larger than the 1, the model may have problems.

27. [5 POINTS] If the response is truly a binomial random variable, and the ni are relatively large, what is the approximate distribution of the deviance? A chi-square distribution.

28. [20 POINTS] Data and R output on the incidence of respiratory disease in infants by sex and feeding method are below.



(a) [5 POINTS] Use the output to predict the probability that a baby girl who only bottle feeds will contract a respiratory disease.

Y = -1.61 -.3126(x) -.6693(x) - .1725(x)

Y = 0.199 - .732 - .512 - .84

All things being equal, a baby girl who feeds has a -.0732 odds of contract respiratory diseases

(b) [5 POINTS] Use the output to predict the probability that a baby boy who only breast feeds will contract a respiratory disease.

Y = -1.61 -.3126(x) -.6693(x) - .1725(x)

Y = 0.199 - .732 - .512 - .84

All things being equal, a baby girl who feeds has a -.0732 odds of contract respiratory diseases

(c) [5 POINTS] Breast feeding reduces the odds of respiratory disease to % of that for bottle feeding. .0732

(d) [5 POINTS] Is the χ2 approximation valid for obtaining a p-value for the residual deviance in

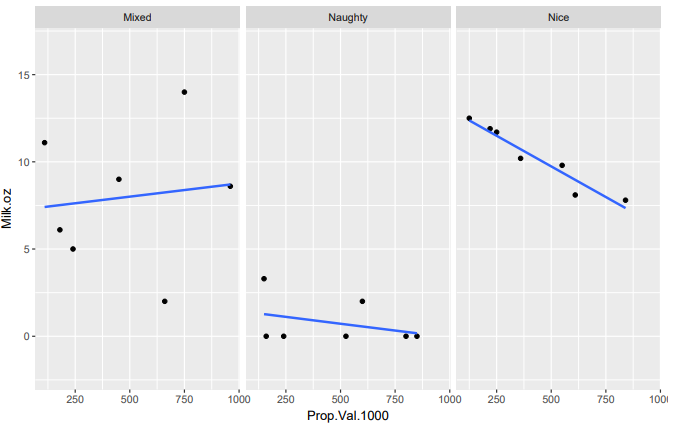
this case? Explain. In this case, yes: the null deviance is much higher than the residual deviance.

29. [20 POINTS] Santa Claus wants to estimate how many ounces of milk will be left at houses he visits based on (1) the assessed property value of the residence; (2) if the children of the household were naughty, nice, or some of each (mixed). Below are some data he took last year.

(a) [5 POINTS] According to the R output on the previous page, do the interaction terms seem statistically significant? Explain. The interaction terms for BehaviorNaughty and BehaviorNice are statistically significant with very small p-values.

(b) [5 POINTS] Now considering the following plot of the data, do you think the interaction terms

are significant? Explain. I would say that based on these nice plots from the ggplot2 package that Nice is significant.



(c) [5 POINTS] Use the model to predict the amount of milk left for Santa at a $500,000 home of a child who was partly naughty, and partly nice (mixed).

Y = 7.24 + .0015(500,000) – 5.74 + 6.01

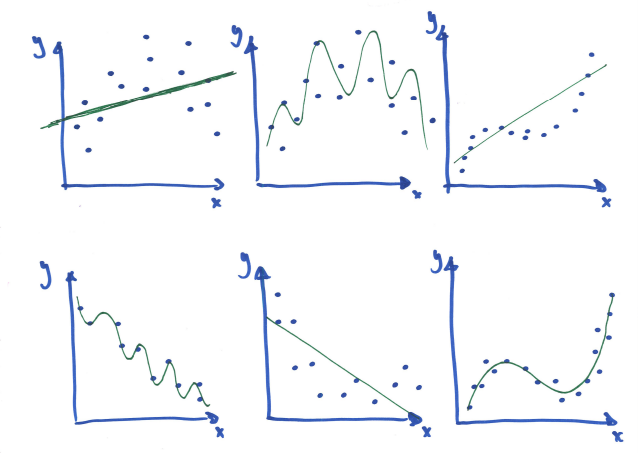
Y = 757.5

(d) [5 POINTS] Use the model to predict the amount of milk left for Santa at a $150,000 home of a nice child.

Y = 7.24 + .0015(150,000) + 6.01

Y= 238.25

30. [10 POINTS] Which of the following modeling scenarios look (i) like a high bias scenario; (ii) like a high variance scenario; (iii) just about right? Label each of the six plots with (i) or (ii) or (iii).



i

ii

iii

ii

iii

ii

31. [5 POINTS] What happens to the training error (such as SSE or MSE evaluated on the training set) as the model flexibility increases? It decreases.

32. [5 POINTS] What happens to the test or cross validation error (such as SSE or MSE evaluated on the test or cross validation set) as the model flexibility increases? It decreases.

APPENDIX: R SCRIPTS, MUSINGS ON THE MIAMI DOLPHINS, AND OTHER MINUTIAE

Question 1

> setwd("c:/Users/baumgaral/R")

> library(readxl)

> question1 <- read\_excel("c:/Users/baumgaral/Data/Practice\_Sets/question1.xlsx")

> lmod <- lm(question1$CreditHours ~ question1$Major + question1$ACT + question1$Inches)

> summary(lmod)

Call:

lm(formula = question1$CreditHours ~ question1$Major + question1$ACT +

question1$Inches)

Residuals:

Min 1Q Median 3Q Max

-10.5287 -6.3170 0.2087 5.7811 10.3511

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 24.0662 59.4368 0.405 0.6962

question1$MajorMath 11.1286 6.1092 1.822 0.1060

question1$ACT -0.7349 0.8985 -0.818 0.4371

question1$Inches 1.7996 0.8827 2.039 0.0758 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.7 on 8 degrees of freedom

Multiple R-squared: 0.5646, Adjusted R-squared: 0.4014

F-statistic: 3.458 on 3 and 8 DF, p-value: 0.07128