**STAT5120—Week 3 Homework, Allen Baumgarten**

1. Open the GPA/ACT data from CH03PR03.txt. This data set is the dataset from Chapter 1 on GPA vs. ACT, but includes observations on two additional variables, namely intelligence test scores (third column) and high school class rank percentile (fourth column). We want to know which of the three explanatory variables (ACT, intelligence test score, high school class rank percentile) can best be used to make a linear model for predicting GPA. So you will build and compare three simple linear regression models for

1. GPA vs. ACT

2. GPA vs. intelligence test score

3. GPA vs. class rank percentile

So for each of these three cases, do the following:

(a) obtain the linear model (lm(y ~ x)) and output (summary()) and compare the R2 values.

For Model 1, GPA vs. ACT Scores:

Coefficients:

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

(Intercept) 2.11405 0.32089 6.588 1.3e-09 \*\*\*

ch03pr03[, 2] 0.03883 0.01277 3.040 0.00292 \*\*

---

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

Residual standard error: 0.6231 on 118 degrees of freedom

Multiple R-squared: 0.07262, Adjusted R-squared: 0.06476

F-statistic: 9.24 on 1 and 118 DF, p-value: 0.002917

For Model 2, GPA vs. Intelligence Scores:

Coefficients:

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

(Intercept) -1.873921 0.345709 -5.421 3.2e-07 \*\*\*

ch03pr03[, 3] 0.041944 0.002915 14.389 < 2e-16 \*\*\*

---

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

Residual standard error: 0.3899 on 118 degrees of freedom

Multiple R-squared: 0.637, Adjusted R-squared: 0.6339

F-statistic: 207 on 1 and 118 DF, p-value: < 2.2e-16

For Model 3, GPA vs. Class Rank Percentile:

Coefficients:

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

(Intercept) 2.306901 0.185497 12.436 < 2e-16 \*\*\*

ch03pr03[, 4] 0.010417 0.002406 4.329 3.15e-05 \*\*\*

---

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

Residual standard error: 0.6011 on 118 degrees of freedom

Multiple R-squared: 0.1371, Adjusted R-squared: 0.1298

F-statistic: 18.74 on 1 and 118 DF, p-value: 3.153e-05

Comparing the R2 statistics, we find the following results and see that Model 2 has the highest one:

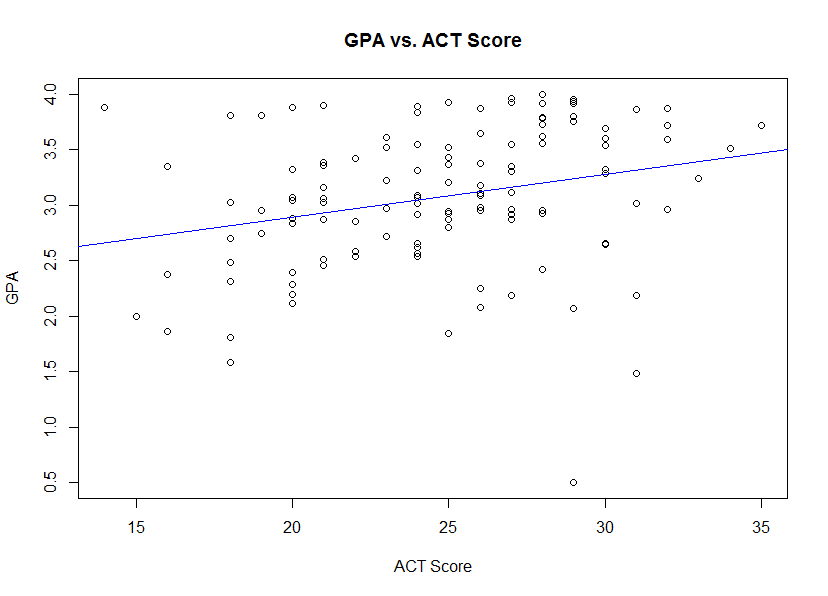
Model 1 R2: 0.07262

Model 2 R2: 0.637

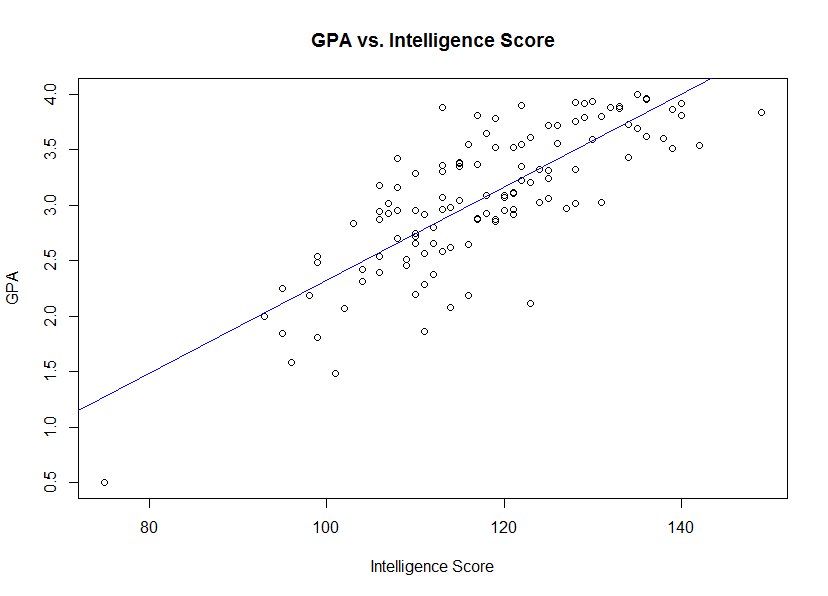
Model 3 R2: 0.1371

(b) make scatter plots that include the regression lines. Identify any potential outliers and influential observations. Decide whether or not to remove them before moving on.

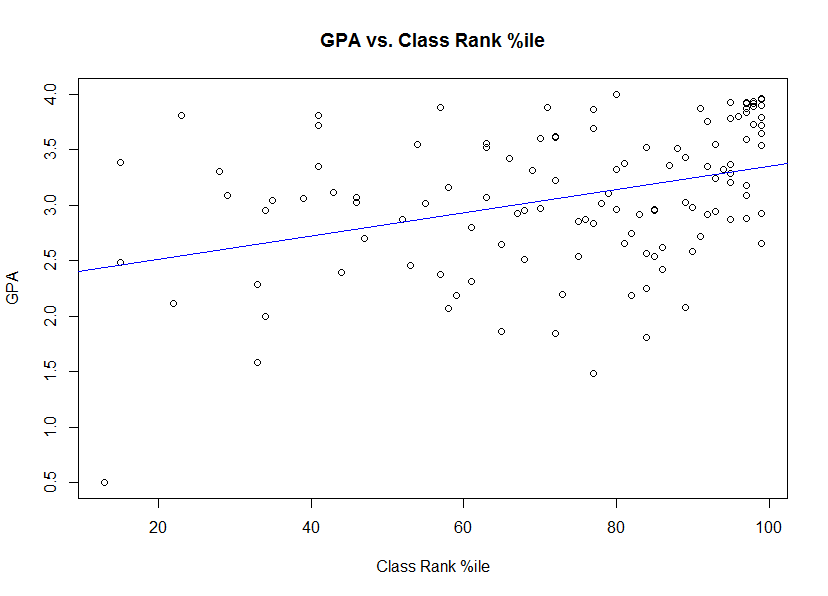
Scatter plot for Model 1:



Scatter plot for Model 2:



Scatter plot for Model 3:



At least one outlier point was observed, viz., in Model 1 with an ACT Score of only 27 or 28 and a correspondingly low GPA score.

(c) check for normality of the residuals with Shapiro-Wilk tests (shapiro.test()).

(d) make boxplots and histograms of the residuals to help check for normality.

(e) make normal probability plots for the residuals (to check for normality).

(f) Split the data set into two groups: students with ACT scores less than 26 and students with ACT scores at least 26. Run Levene's test for equality of variance of the residuals (leveneTest() requires the car package) on the response for these two groups. Remark on what you observe. The command subset() can be useful for splitting data sets.

(g) Repeat part (f) for the intelligence test score model, splitting at < 120 and 120. Do the residual variances for the two groups appear to differ?

(g) plot the residuals vs. the predictor variable values to check for equality of variance.

(h) Remark on which of the three explanatory variables seems to be the most useful in building a linear model for predicting first-year GPA.

2. Adapted from ALSM 3.15. A chemist wanted to model the evolution of a solution concentration over time. To do this, she randomly assigned three solutions to measure after one hour, three solutions to measure after three hours, three to measure after five hours, three to measure after seven hours, and three to measure after nine hours. The data are in CH03PR15.

(a) Find the equation of the least-squares regression line.

(b) Run the F-test for lack of linear fit. Use the significance level α = 0.025. State your p-value and provide your conclusion.

(c) When a lack-of-linear fit test indicates there is a lack of linear fit, does it suggest exactly what kind of function would be appropriate? Explain.

3. Adapted from ALSM 3.16. Use the data from the previous problem (CH03PR15).

(a) Make a scatterplot of the data, with concentration as the response variable. Based on the scatterplot, what kind of data transformation do you suggest to adjust for the non-constant variance and/or non-linearity?

(b) Use the log-likelihood method in R we used for selecting the Box-Cox, and apply the transformation to the data. Then build a new regression model (using lm()) for the transformed data, make a fitted-line plot, and discuss how the new relationship compares with the old one.

(c) Apply the log10 transformation and get the new regression line equation. Plot this model on a scatterplot of the transformed data. Compare the results of the Box-Cox transformation with those of the log10 transformation. Which do you like better and why?

(d) Plot the residuals against the fitted values. Also make a normal probability plot. What do these plots indicate?

(e) Express your estimated regression functions in the original units.

4. Adapted from ALSM 3.17. A marketing manager studied annual product sales figures over a ten year period. The data (years and sales in thousands of units) are in the file CH03PR17.

(a) Make a scatterplot. Is the linearity assumption reasonable?

(b) Apply the maximum likelihood Box-Cox method (like we did in the Trees example) to get an appropriate power transformation of the response (sales). What is the value of SSE in this case?

(c) Try using the square-root transformation and get a new regression line.

(d) Plot the regression line from the previous part on a scatterplot of the transformed data. Does this line seem to fit the transformed data well?

(e) Make a plot of the residuals vs. fits. Also make a normal probability plot. What do these plots indicate for your transformed data?

(f) Express the regression models in the original units.

5. The Blaisdell Company wanted to use industry sales to predict its sales. Adjusted quarterly sales data for 1998 – 2002 are in the ALSM data set CH12TA02. The first column of data are observations of Blaisdell's sales, and the second column contain industry sales. The very first column, the one with the row numbers, is a time index: 1 means first quarter of 1998, 2 means 2nd quarter of 1998, etc.

(a) Make a scatter plot of company sales using industry sales as the predictor (with R of course). Describe the apparent relationship between the two variables.

(b) Make a scatter plot of company sales versus the time index. You might have to create a new column for the time values or figure out how to reference the row numbers in R.

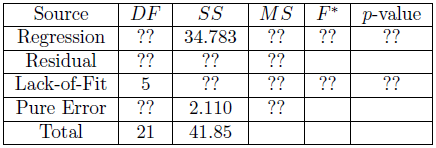
(c) Use R to run a Neumann-Durbin-Watson to check for autocorrelation of company sales over time:

H0 : = 0;

HA : 6= 0:

(d) How would you suggest to proceed in modelling the relationship between these variables?

6. Complete the following lack-of-fit ANOVA table:



APPENDIX: R SCRIPTS USED

**QUESTION 1:**

> library(xlsx)

> ch03pr03 <- read.xlsx("C:/Users/allen.baumgarten/Documents/CH03PR03.xlsx", sheetName = "Sheet1")

> ch03pr03\_reg1 <- lm(ch03pr03[,1] ~ ch03pr03[,2])

> ch03pr03\_reg2 <- lm(ch03pr03[,1] ~ ch03pr03[,3])

> ch03pr03\_reg3 <- lm(ch03pr03[,1] ~ ch03pr03[,4])

> summary(ch03pr03\_reg1 <- lm(ch03pr03[,1] ~ ch03pr03[,2]))

Call:

lm(formula = ch03pr03[, 1] ~ ch03pr03[, 2])

Residuals:

Min 1Q Median 3Q Max

-2.74004 -0.33827 0.04062 0.44064 1.22737

Coefficients:

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

(Intercept) 2.11405 0.32089 6.588 1.3e-09 \*\*\*

ch03pr03[, 2] 0.03883 0.01277 3.040 0.00292 \*\*

---

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

Residual standard error: 0.6231 on 118 degrees of freedom

Multiple R-squared: 0.07262, Adjusted R-squared: 0.06476

F-statistic: 9.24 on 1 and 118 DF, p-value: 0.002917

> summary(ch03pr03\_reg2 <- lm(ch03pr03[,1] ~ ch03pr03[,3]))

Call:

lm(formula = ch03pr03[, 1] ~ ch03pr03[, 3])

Residuals:

Min 1Q Median 3Q Max

-1.1672 -0.2402 -0.0225 0.2977 1.0193

Coefficients:

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

(Intercept) -1.873921 0.345709 -5.421 3.2e-07 \*\*\*

ch03pr03[, 3] 0.041944 0.002915 14.389 < 2e-16 \*\*\*

---

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

Residual standard error: 0.3899 on 118 degrees of freedom

Multiple R-squared: 0.637, Adjusted R-squared: 0.6339

F-statistic: 207 on 1 and 118 DF, p-value: < 2.2e-16

> summary(ch03pr03\_reg3 <- lm(ch03pr03[,1] ~ ch03pr03[,4]))

Call:

lm(formula = ch03pr03[, 1] ~ ch03pr03[, 4])

Residuals:

Min 1Q Median 3Q Max

-1.94233 -0.40879 0.05516 0.48679 1.25950

Coefficients:

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

(Intercept) 2.306901 0.185497 12.436 < 2e-16 \*\*\*

ch03pr03[, 4] 0.010417 0.002406 4.329 3.15e-05 \*\*\*

---

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

Residual standard error: 0.6011 on 118 degrees of freedom

Multiple R-squared: 0.1371, Adjusted R-squared: 0.1298

F-statistic: 18.74 on 1 and 118 DF, p-value: 3.153e-05

> plot(ch03pr03[,1] ~ ch03pr03[,2], main = "GPA vs. ACT Score", xlab="ACT Score", ylab="GPA")

> abline(ch03pr03\_reg1, col="blue")

> plot(ch03pr03[,1] ~ ch03pr03[,3], main = "GPA vs. Intelligence Score", xlab="Intelligence Score", ylab="GPA")

> abline(ch03pr03\_reg2, col="blue")

> plot(ch03pr03[,1] ~ ch03pr03[,4], main = "GPA vs. Class Rank %ile", xlab="Class Rank %ile", ylab="GPA")

> abline(ch03pr03\_reg3, col="blue")