Exam 2

Spencer Swartz

April 11, 2017

## Problem 1

People who are concerned about their health may prefer hot dogs that are low in sodium and calories. The data file contains sample data on the sodium and calories contained in each of 54 major hot dog brands. The hot dogs are classified by type: beef, poultry, and meat.

The data file called hotdogs.rda contains the sodium and calorie content for random samples of each type of hot dog. This data set is included in the DS705data package.

### Part 1a

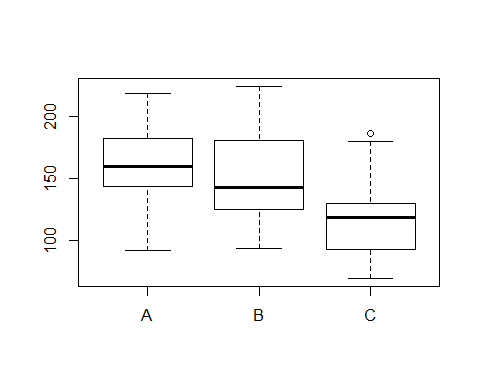
Make boxplots for the variable calories for the 3 types of hotdogs. Describe the 3 boxplots and the suitability of these samples for conducting analysis of variance.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1a -|-|-|-|-|-|-|-|-|-|-|-

require(DS705data)

## Loading required package: DS705data

data('hotdogs')  
boxplot(hotdogs$calories~hotdogs$type)



It looks as if the response variables in each poulation are normally distributed and have equal variance, assuming that the samples are selected randomly and independently, then these sets are sutables for conducting anova.

## [1] "Score for 1a : 0 / 3"

### Part 1b

Conduct an analysis of variance test (the standard one that assumes normality and equal variance) to compare population mean calorie counts for these three types of hot dogs. (i) State the null and alternative hypotheses, (ii) use R to compute the test statistic and p-value, and (iii) write a conclusion in context at .

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1b -|-|-|-|-|-|-|-|-|-|-|-

not all of the population mean calories are the same.

l.model <- lm( calories ~ type, data = hotdogs)  
test <- anova(l.model)  
test

## Analysis of Variance Table  
##   
## Response: calories  
## Df Sum Sq Mean Sq F value Pr(>F)   
## type 2 15425 7712.5 6.2784 0.003651 \*\*  
## Residuals 51 62649 1228.4   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

1. Reject at ( = 0.0037). There is significant evidence to show that the population mean calories are different between types.

## [1] "Score for 1b : 0 / 8"

### Part 1c

Follow up with a Tukey-Kramer multiple comparison procedure and control the experimentwise error rate at . Write an interpretation for your multiple comparison output in the context of the problem.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1c -|-|-|-|-|-|-|-|-|-|-|-

TukeyHSD(aov(calories ~ type, data = hotdogs),conf.level = .9)

## Tukey multiple comparisons of means  
## 90% family-wise confidence level  
##   
## Fit: aov(formula = calories ~ type, data = hotdogs)  
##   
## $type  
## diff lwr upr p adj  
## B-A -9.144118 -33.41772 15.129490 0.7102899  
## C-A -39.673529 -63.94714 -15.399922 0.0033799  
## C-B -30.529412 -55.76791 -5.290917 0.0371484

Hotdog Type C mean calories is between 5.3 and 55.8 more than Type B. Hotdog Type C mean calories is between 15.4 and 63.9 more than Type A. And hotdog Type B mean calories is between 15.1 less than and 33.4 more than Type B.

## [1] "Score for 1c : 0 / 5"

### Part 1d

As part of a vigorous road test for independent random samples of size 20 for 4 different brands of tires, the tread wear was measured for comparison. The data frame treadwear.rda contains the resulting data.

Begin by exploring the sample means and standard deviations for each brand and looking at a boxplot comparison. That is, find the sample means and standard deviations and produce a boxplots of the tread wear measures for the four brands of tires.

Conduct hypothesis tests for the normality of the tread wear for each brand using a 5% level of significance in each case.

Also test for the homogeneity of the variances using .

Comment on the results of each.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1d -|-|-|-|-|-|-|-|-|-|-|-

data("treadwear")  
mean(treadwear$wear[treadwear$brand=="A"])

## [1] 576.3017

mean(treadwear$wear[treadwear$brand=="B"])

## [1] 671.1044

mean(treadwear$wear[treadwear$brand=="C"])

## [1] 825.987

mean(treadwear$wear[treadwear$brand=="D"])

## [1] 853.2877

sd(treadwear$wear[treadwear$brand=="A"])

## [1] 148.1037

sd(treadwear$wear[treadwear$brand=="B"])

## [1] 111.2945

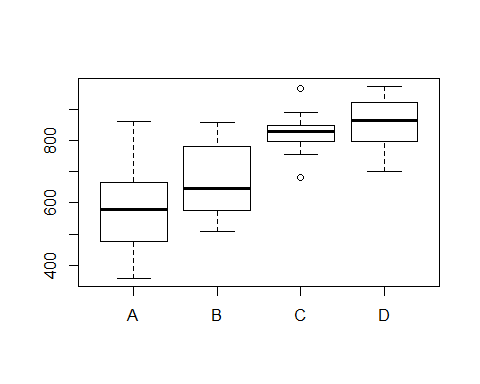
sd(treadwear$wear[treadwear$brand=="C"])

## [1] 57.85366

sd(treadwear$wear[treadwear$brand=="D"])

## [1] 81.23036

boxplot(treadwear$wear~treadwear$brand)



shapiro.test(treadwear$wear[treadwear$brand=="A"])

##   
## Shapiro-Wilk normality test  
##   
## data: treadwear$wear[treadwear$brand == "A"]  
## W = 0.94655, p-value = 0.3177

shapiro.test(treadwear$wear[treadwear$brand=="B"])

##   
## Shapiro-Wilk normality test  
##   
## data: treadwear$wear[treadwear$brand == "B"]  
## W = 0.92027, p-value = 0.1003

shapiro.test(treadwear$wear[treadwear$brand=="C"])

##   
## Shapiro-Wilk normality test  
##   
## data: treadwear$wear[treadwear$brand == "C"]  
## W = 0.95389, p-value = 0.4301

shapiro.test(treadwear$wear[treadwear$brand=="D"])

##   
## Shapiro-Wilk normality test  
##   
## data: treadwear$wear[treadwear$brand == "D"]  
## W = 0.95129, p-value = 0.3871

Based on the Shapiro Wilk tests, we cannot reject that the sample comes from a population which has a normal distrobution for any of the diferent tire brands. (pvalues in alphabetical order: 0.3177, 0.1003, 0.4301, 0.3871)

## [1] "Score for 1d : 0 / 6"

### Part 1e

What is the most appropriate inference procedure to compare population mean tread wear for these four brands of tires? Perform this procedure.

1. State the null and alternative hypotheses, (ii) use R to compute the test statistic and p-value, and (iii) write a conclusion in context at .

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1e -|-|-|-|-|-|-|-|-|-|-|-

Replace this text with your answer here.

(i)

not all of the population mean wear are the same.

oneway.test(wear ~ brand, data = treadwear, var.equal=FALSE)

##   
## One-way analysis of means (not assuming equal variances)  
##   
## data: wear and brand  
## F = 27.201, num df = 3.000, denom df = 40.197, p-value = 8.988e-10

1. Reject at ( = 0). There is significant evidence to show that the population mean tread wears are different.

## [1] "Score for 1e : 0 / 8"

### Part 1f

Conduct the most appropriate multiple comparisons procedure which brands have significantly different tread wear. Use a familywise error rate of . Use complete sentences to interpret the results in the context of the problem.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1f -|-|-|-|-|-|-|-|-|-|-|-

onewayComp(wear~brand,data = treadwear,var.equal=FALSE,adjust='one.step',alpha=0.05)

## $call  
## onewayComp(formula = wear ~ brand, data = treadwear, alpha = 0.05,   
## var.equal = FALSE, adjust = "one.step")  
##   
## $comp  
## diff lwr upr t p p adj  
## B-A 94.80270 -16.87538 206.48079 2.288520 2.820971e-02 1.097707e-01  
## C-A 249.68530 151.79715 347.57345 7.022705 2.502251e-07 4.666289e-09  
## D-A 276.98604 174.17864 379.79344 7.333286 4.024639e-08 1.139764e-09  
## C-B 154.88259 78.39766 231.36753 5.522105 6.246844e-06 2.652978e-06  
## D-B 182.18334 99.06516 265.30152 5.913167 1.029246e-06 5.337504e-07  
## D-C 27.30075 -32.89664 87.49813 1.224272 2.291848e-01 6.134985e-01  
## rej H\_0  
## B-A 0  
## C-A 1  
## D-A 1  
## C-B 1  
## D-B 1  
## D-C 0  
##   
## $pairw  
##   
## Pairwise comparisons using t tests with unpooled SD   
##   
## data: wear and brand   
##   
## A B C   
## B 0.11 - -   
## C 4.7e-09 2.7e-06 -   
## D 1.1e-09 5.3e-07 0.61  
##   
## P value adjustment method: one.step

We are 95% confident that the population mean tread wear of brand C is 151.8 to 347.6 more than that of brand A, and 78.4 to 231.4 more than that of Brand B. Similarly the population mean tread wear of brand D is 174.2 to 379.8 more than that of Brand A, and 99.1 to 265.3 more than that of brand B. There appear to be no other significant population mean differences.

## [1] "Score for 1f : 0 / 5"

## Problem 2

This dataset contains the prices of ladies' diamond rings and the carat size of their diamond stones. The rings are made with gold of 20 carats purity and are each mounted with a single diamond stone. The data was presented in a newspaper advertisement suggesting the use of simple linear regression to relate the prices of diamond rings to the carats of their diamond stones.

The data is in the file diamond.rda and is included in the DS705data package.

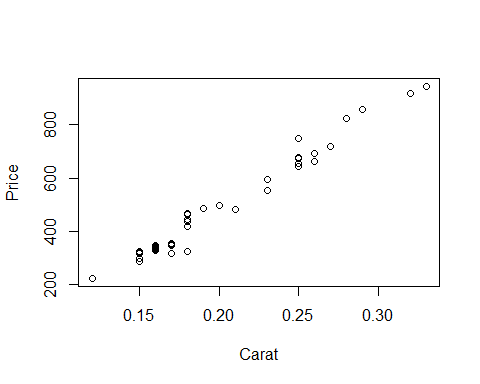
### Part 2a

Does it appear that a linear model is at least possibly a plausible model for predicting the price from carats of the diamonds for these rings?

Begin by creating a scatterplot and comment on the suitability of a linear regression model.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2a -|-|-|-|-|-|-|-|-|-|-|-

data("diamond")  
diamond = diamond[-42,]  
carat <- diamond$carat  
price <- diamond$price  
plot(carat,price,xlab="Carat",ylab="Price",main="")



Replace this text with your answer here.

## [1] "Score for 2a : 0 / 2"

### Part 2b

Obtain the estimated slope and y-intercept for the estimated regression equation and write the equation in the form pricecarats (only with and replaced with the numerical estimates from your R output).

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2b -|-|-|-|-|-|-|-|-|-|-|-

linear.model <- lm(price~carat,data=diamond)  
b0 <- linear.model$coef[1]  
b1 <- linear.model$coef[2]  
b0

## (Intercept)   
## -250.5681

b1

## carat   
## 3671.397

Replace the ## symbols with your slope and intercept

price = ## + ## carat

## [1] "Score for 2b : 0 / 5"

### Part 2c

Compute the sample Pearson correlation coefficient and test whether or not the population Pearson correlation coefficient between price and carat is zero using a 1% level of significance. (i) State the null and alternative hypotheses, (ii) test statistic, (iii) the p-value, and (iv) conclusion.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2c -|-|-|-|-|-|-|-|-|-|-|-

cor.test(price,carat, alpha=.01)

##   
## Pearson's product-moment correlation  
##   
## data: price and carat  
## t = 42.116, df = 45, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.9776332 0.9930866  
## sample estimates:  
## cor   
## 0.9875512

Sample correlation:

(i) true correlation is equal to 0

true correlation is not equal to 0

## [1] "Score for 2c : 0 / 8"

### Part 2d

Provide a 95% confidence interval to estimate the slope of the regression equation and interpret the interval in the context of the application (do not us the word âslopeâ in your interpretation).

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2d -|-|-|-|-|-|-|-|-|-|-|-

cor.test(price,carat, alpha=.01)$conf.int

## [1] 0.9776332 0.9930866  
## attr(,"conf.level")  
## [1] 0.95

We are 95% confident the population correlation between diamond type and carat lies in the range 0.978 to 0.993.

## [1] "Score for 2d : 0 / 5"

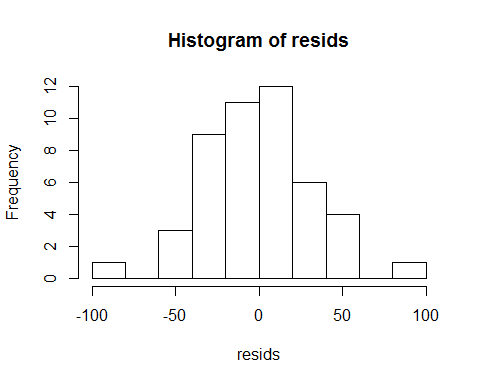
### Part 2e

Check to see if the linear regression model assumptions are reasonable for this data.

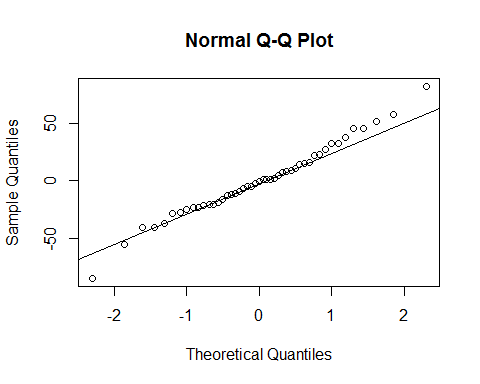
(Step 1) Are the residuals normal? Construct a histogram, normal probability plot, and boxplot of the residuals and perform a Shapiro-Wilk test for normality.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2e.1 -|-|-|-|-|-|-|-|-|-|-|-

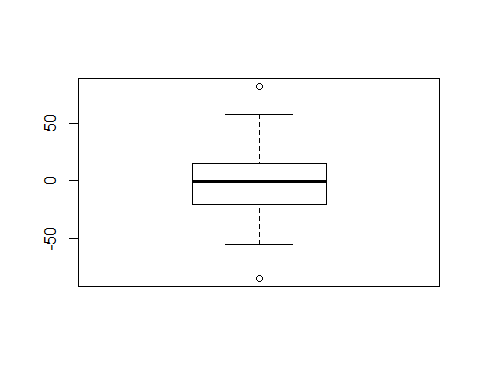
resids <- linear.model$resid  
  
hist(resids)



qqnorm(resids)  
qqline(resids)



boxplot(resids)



shapiro.test(resids)

##   
## Shapiro-Wilk normality test  
##   
## data: resids  
## W = 0.98604, p-value = 0.8406

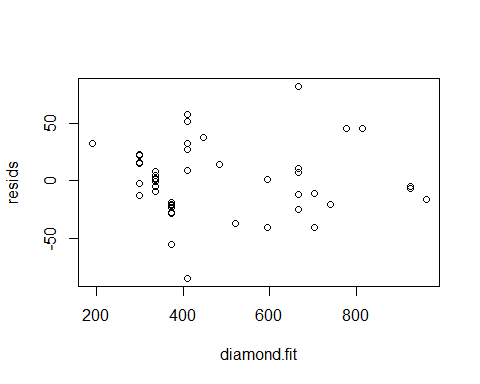
Replace this text with your answer here

## [1] "Score for 2e.1 : 0 / 3"

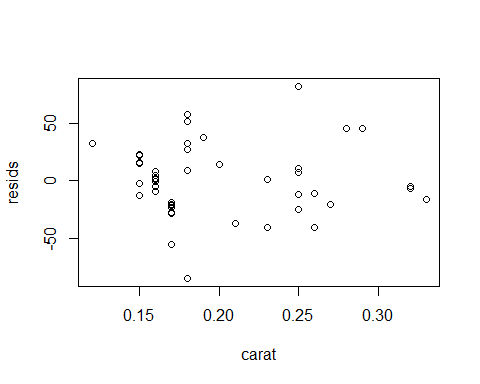
(Step 2) Plot the residuals against the fitted values. Does the equal variances assumption seem reasonable? Does the linear regression line seem to be a good fit?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2e.2 -|-|-|-|-|-|-|-|-|-|-|-

diamond.fit <- linear.model$fitted.values  
plot(diamond.fit,resids)



plot(carat,resids)



Replace this text with your answer here

## [1] "Score for 2e.2 : 0 / 3"

(Step 3) Perform the Bruesch-Pagan test for equal variances of the residuals. What does the test tell you?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2e.3 -|-|-|-|-|-|-|-|-|-|-|-

require(lmtest)

## Loading required package: lmtest

## Warning: package 'lmtest' was built under R version 3.3.3

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 3.3.3

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

bptest(linear.model)

##   
## studentized Breusch-Pagan test  
##   
## data: linear.model  
## BP = 0.18208, df = 1, p-value = 0.6696

Replace this text with your answer here

## [1] "Score for 2e.3 : 0 / 3"

### Part 2f

Calculate and interpret the coefficient of determination (same as ).

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2f -|-|-|-|-|-|-|-|-|-|-|-

out <- summary(linear.model)  
r.sq <- out$r.squared  
r.sq.adj <- out$adj.r.squared  
r.sq.adj

## [1] 0.9747075

r.sq

## [1] 0.9752573

The coefficient of determination, .975 suggests that 97.5% of the total variation in diamond price is explained by the linear relationship between price and carat. We can also say that 2.5% of the total variation in heart rates is not explained by the linear relationship.

## [1] "Score for 2f : 0 / 5"

### Part 2g

Should the regression equation obtained for price and carats be used for making predictions? Explain your answer.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2g -|-|-|-|-|-|-|-|-|-|-|-

Replace this text with your answer here.

## [1] "Score for 2g : 0 / 2"

### Part 2h

What would be the straightforward interpretation of the y-intercept in this regression model? Does it make sense here? Why would this not be appropriate as a stand-alone interpretation for this scenario? (hint: what is extrapolation?)

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2g -|-|-|-|-|-|-|-|-|-|-|-

Replace this text with your answer here.

## [1] "Score for 2h : 0 / 3"

### Part 2i

Create 95% prediction and confidence limits for the population mean price for the carats given in the sample data and plot them along with a scatterplot of the data.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2i -|-|-|-|-|-|-|-|-|-|-|-

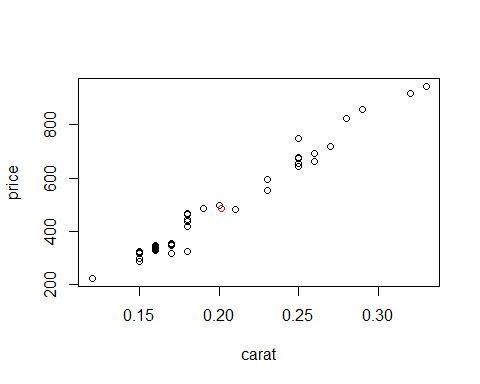
x <- data.frame( carat = mean(carat) )  
predict( linear.model, x, interval="confidence")

## fit lwr upr  
## 1 487.617 478.3884 496.8457

predict( linear.model, x, interval="prediction")

## fit lwr upr  
## 1 487.617 423.6792 551.5549

plot(carat,price)  
points(mean(carat),487.611, col='red')



## [1] "Score for 2i : 0 / 5"

## Problem 3

Blood type is classified as âA, B, or ABâ, or O. In addition, blood can be classified as Rh+ or Rh -. In a survey of 500 randomly selected individuals, a phlebotomist obtained the results shown in the table below.

|  |  |  |  |
| --- | --- | --- | --- |
| Rh Factor | A, B, or AB | O | Total |
| Rh+ | 226 | 198 | 424 |
| Rh- | 46 | 30 | 76 |
| Total | 272 | 228 | 500 |

### Part 3a

Conduct the appropriate test of significance to test the following research question âIs Rh factor associated with blood type?â Use a 5% level of significance and include all parts of the test.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 3a -|-|-|-|-|-|-|-|-|-|-|-

Rhpos <- c(226,198)  
Rhneg <- c(46,30)  
BloodTable <- rbind(Rhpos,Rhneg)  
dimnames(BloodTable) <- list( Rh = c("pos","neg"),Type =c ("A,B,AB","O") )  
BloodTable

## Type  
## Rh A,B,AB O  
## pos 226 198  
## neg 46 30

out <- chisq.test(BloodTable,correct=FALSE)  
out

##   
## Pearson's Chi-squared test  
##   
## data: BloodTable  
## X-squared = 1.356, df = 1, p-value = 0.2442

Blood type and Rh factor are independent.

Blood type and Rh factor are dependent or associated.

Do not reject at ( = .2442). There is not enough evidence to show that there is an association between the Blood type and Rh factor.

## [1] "Score for 3a : 0 / 8"

### Part 3b

Compute and interpret the odds ratio of having Type O blood for Rh+ compared to Rh-.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 3b -|-|-|-|-|-|-|-|-|-|-|-

oddsR <-(198/226)/(30/46)  
oddsR

## [1] 1.343363

100\*(abs(1-round(oddsR,2)))

## [1] 34

Replace this text with your answer here.

## [1] "Score for 3b : 0 / 5"

### Part 3c

Construct and interpret a 90% confidence interval for the population proportion of people with Type O blood who are Rh-.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 3c -|-|-|-|-|-|-|-|-|-|-|-

out <- prop.test(30,228,correct=FALSE, conf.level = .9)  
out

##   
## 1-sample proportions test without continuity correction  
##   
## data: 30 out of 228, null probability 0.5  
## X-squared = 123.79, df = 1, p-value < 2.2e-16  
## alternative hypothesis: true p is not equal to 0.5  
## 90 percent confidence interval:  
## 0.09903906 0.17275997  
## sample estimates:  
## p   
## 0.1315789

The population proportion of patients who have blood type O and have Rh- is between 0.099 and 0.173.

## [1] "Score for 3c : 0 / 5"

## Problem 4

The carinate dove shell is a yellowish to brownish colored smooth shell found along shallow water coastal areas in California and Mexico. A study was conducted to determine if the shell height of the carinate dove shell could be accurately predicted and to identify the independent variables needed to do so. Data was collected for a random sample of 30 of these gastropods and 8 variables that researchers thought might be good predictors were recorded.

The shell heights (in mm) are labeled in the file as Y and the potential predictor variables are simply named as X1, X2, X3, X4, X5, X6, X7, and X8. Independent variables X1 through X7 are quantitative while X8 is categorical.

The data is in the file shells.rda and is included in the DS705data package.

### Part 4a

Use stepwise model selection with AIC as the stepping criterion and direction = "both" to identify the best first-order model for predicting shell height (Y).

Identify the predictor variables in the final model as well as the AIC for that model.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 4a -|-|-|-|-|-|-|-|-|-|-|-

data("shells")  
lm.full = lm(Y~., data=shells) # regress y on everything in data set  
step(lm.full,direction="both")

## Start: AIC=-116.2  
## Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8  
##   
## Df Sum of Sq RSS AIC  
## - X8 1 0.000007 0.34233 -118.20  
## - X3 1 0.001085 0.34340 -118.10  
## - X5 1 0.001152 0.34347 -118.10  
## <none> 0.34232 -116.20  
## - X7 1 0.041077 0.38340 -114.80  
## - X4 1 0.084530 0.42685 -111.58  
## - X2 1 0.103328 0.44565 -110.28  
## - X6 1 0.181019 0.52334 -105.46  
## - X1 1 0.226388 0.56871 -102.97  
##   
## Step: AIC=-118.2  
## Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7  
##   
## Df Sum of Sq RSS AIC  
## - X3 1 0.001154 0.34348 -120.09  
## - X5 1 0.001197 0.34352 -120.09  
## <none> 0.34233 -118.20  
## - X7 1 0.042323 0.38465 -116.70  
## + X8 1 0.000007 0.34232 -116.20  
## - X4 1 0.091675 0.43400 -113.08  
## - X2 1 0.105735 0.44806 -112.12  
## - X6 1 0.189571 0.53190 -106.97  
## - X1 1 0.309263 0.65159 -100.89  
##   
## Step: AIC=-120.09  
## Y ~ X1 + X2 + X4 + X5 + X6 + X7  
##   
## Df Sum of Sq RSS AIC  
## - X5 1 0.00116 0.34464 -121.99  
## <none> 0.34348 -120.09  
## + X3 1 0.00115 0.34233 -118.20  
## + X8 1 0.00008 0.34340 -118.10  
## - X7 1 0.06623 0.40971 -116.81  
## - X4 1 0.11094 0.45442 -113.70  
## - X2 1 0.14822 0.49169 -111.33  
## - X6 1 0.20134 0.54482 -108.25  
## - X1 1 0.42107 0.76455 -98.09  
##   
## Step: AIC=-121.99  
## Y ~ X1 + X2 + X4 + X6 + X7  
##   
## Df Sum of Sq RSS AIC  
## <none> 0.34464 -121.993  
## + X5 1 0.00116 0.34348 -120.095  
## + X3 1 0.00112 0.34352 -120.091  
## + X8 1 0.00017 0.34448 -120.008  
## - X7 1 0.06623 0.41088 -118.720  
## - X4 1 0.12074 0.46538 -114.983  
## - X2 1 0.16170 0.50634 -112.452  
## - X6 1 0.20066 0.54531 -110.228  
## - X1 1 0.45355 0.79819 -98.798

##   
## Call:  
## lm(formula = Y ~ X1 + X2 + X4 + X6 + X7, data = shells)  
##   
## Coefficients:  
## (Intercept) X1 X2 X4 X6   
## 1.42718 1.15023 0.65317 0.60923 -0.06487   
## X7   
## 0.02636

Replace this text with your answer here.

## [1] "Score for 4a : 0 / 5"

### Part 4b

Compute the variance inflation factor for the final model from part 4a.

Does this model suffer from multicollinearity? Explain your answer.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 4b -|-|-|-|-|-|-|-|-|-|-|-

fit <- lm(formula = Y ~ X1 + X2 + X4 + X6 + X7, data = shells)  
summary(fit)

##   
## Call:  
## lm(formula = Y ~ X1 + X2 + X4 + X6 + X7, data = shells)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.200151 -0.080277 -0.007467 0.081058 0.173162   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.42718 0.73427 1.944 0.06375 .   
## X1 1.15023 0.20467 5.620 8.73e-06 \*\*\*  
## X2 0.65317 0.19465 3.356 0.00263 \*\*   
## X4 0.60923 0.21010 2.900 0.00787 \*\*   
## X6 -0.06487 0.01735 -3.738 0.00102 \*\*   
## X7 0.02636 0.01227 2.148 0.04205 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1198 on 24 degrees of freedom  
## Multiple R-squared: 0.9514, Adjusted R-squared: 0.9413   
## F-statistic: 94.05 on 5 and 24 DF, p-value: 5.821e-15

require(car)

## Loading required package: car

## Warning: package 'car' was built under R version 3.3.2

vif(fit)

## X1 X2 X4 X6 X7   
## 4.286121 3.804041 2.697711 4.428297 3.172575

The overall model p-value is very small at p=5.821e-15, and the individual p-values are all under 0.05 (the exception being the intercept at 0.064). This makes it seem as if the model does not suffer from collinerarity/multicollinearity.

VIFs are all less than 10 for all terms futher supporting the idea that this model is not suffering from collinerarity/multicollinearity.

## [1] "Score for 4b : 0 / 5"

### Part 4c

Let's define **Model B** as follows:

Y =

Fit **Model B** and compare the AIC of it to the model that the stepwise selection procedure identified as best in 4a, which you may refer to as **Model A**.

Which model is the better fitting model according to AIC? Explain your answer.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 4c -|-|-|-|-|-|-|-|-|-|-|-

xSqr <- shells$X2\*shells$X2  
fit2 <- lm(formula = Y ~ X1 + X2 + xSqr + X4 + X6, data = shells)  
summary(fit2)

##   
## Call:  
## lm(formula = Y ~ X1 + X2 + xSqr + X4 + X6, data = shells)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.17500 -0.05247 -0.00598 0.05188 0.21059   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.16461 2.31597 3.957 0.000587 \*\*\*  
## X1 1.11286 0.17872 6.227 1.96e-06 \*\*\*  
## X2 -5.45450 1.70283 -3.203 0.003811 \*\*   
## xSqr 1.14091 0.31981 3.567 0.001559 \*\*   
## X4 0.79526 0.16439 4.838 6.28e-05 \*\*\*  
## X6 -0.03961 0.01106 -3.581 0.001507 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1058 on 24 degrees of freedom  
## Multiple R-squared: 0.9622, Adjusted R-squared: 0.9543   
## F-statistic: 122.1 on 5 and 24 DF, p-value: 2.955e-16

AIC(fit2)

## [1] -42.34679

Model A's AIC is -121.99 and Model B's is -42.35, when looking at AIC as a criteria to pick a model smaller is better and Model A has the smaller AIC.

## [1] "Score for 4c : 0 / 5"

### Part 4d

Compute the variance inflation factor for **Model B** from part 4c.

Does this model suffer from multicollinearity? Explain your answer.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 4d -|-|-|-|-|-|-|-|-|-|-|-

vif(fit2)

## X1 X2 xSqr X4 X6   
## 4.195143 373.677887 369.491380 2.119951 2.309547

It looks as if the model may suffer from multicollinearity as X2 and xSqr both have vif values above 10, 373.7 and 369.5 respectively.

## [1] "Score for 4d : 0 / 2"

### Part 4e

Center the variable X2 and compute the quadratic term associated with it (call them cX2 and cx2sq, respectively). We'll identify this as **Model C**. Compute the variance inflation factor for **Model C**.

Does this model suffer from multicollinearity? Explain your answer.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 4e -|-|-|-|-|-|-|-|-|-|-|-

cx2 <- shells$X2-mean(shells$X2)  
cx2sq <- cx2^2  
fit3 <- lm(formula = Y ~ X1 + cx2 + cx2sq + X4 + X6, data = shells)  
summary(fit3)

##   
## Call:  
## lm(formula = Y ~ X1 + cx2 + cx2sq + X4 + X6, data = shells)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.17500 -0.05247 -0.00598 0.05188 0.21059   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.70477 0.55462 4.877 5.68e-05 \*\*\*  
## X1 1.11286 0.17872 6.227 1.96e-06 \*\*\*  
## cx2 0.52081 0.17093 3.047 0.00555 \*\*   
## cx2sq 1.14091 0.31981 3.567 0.00156 \*\*   
## X4 0.79526 0.16439 4.838 6.28e-05 \*\*\*  
## X6 -0.03961 0.01106 -3.581 0.00151 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1058 on 24 degrees of freedom  
## Multiple R-squared: 0.9622, Adjusted R-squared: 0.9543   
## F-statistic: 122.1 on 5 and 24 DF, p-value: 2.955e-16

vif(fit3)

## X1 cx2 cx2sq X4 X6   
## 4.195143 3.765151 1.036090 2.119951 2.309547

No.

## [1] "Score for 4e : 0 / 2"

### Part 4f

Compare the adjusted R-squared for **Models A and C**.

Explain what adjusted R-squared measures and state which model is "better" according to this criterion.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 4f -|-|-|-|-|-|-|-|-|-|-|-

summary(fit)

##   
## Call:  
## lm(formula = Y ~ X1 + X2 + X4 + X6 + X7, data = shells)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.200151 -0.080277 -0.007467 0.081058 0.173162   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.42718 0.73427 1.944 0.06375 .   
## X1 1.15023 0.20467 5.620 8.73e-06 \*\*\*  
## X2 0.65317 0.19465 3.356 0.00263 \*\*   
## X4 0.60923 0.21010 2.900 0.00787 \*\*   
## X6 -0.06487 0.01735 -3.738 0.00102 \*\*   
## X7 0.02636 0.01227 2.148 0.04205 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1198 on 24 degrees of freedom  
## Multiple R-squared: 0.9514, Adjusted R-squared: 0.9413   
## F-statistic: 94.05 on 5 and 24 DF, p-value: 5.821e-15

summary(fit3)

##   
## Call:  
## lm(formula = Y ~ X1 + cx2 + cx2sq + X4 + X6, data = shells)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.17500 -0.05247 -0.00598 0.05188 0.21059   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.70477 0.55462 4.877 5.68e-05 \*\*\*  
## X1 1.11286 0.17872 6.227 1.96e-06 \*\*\*  
## cx2 0.52081 0.17093 3.047 0.00555 \*\*   
## cx2sq 1.14091 0.31981 3.567 0.00156 \*\*   
## X4 0.79526 0.16439 4.838 6.28e-05 \*\*\*  
## X6 -0.03961 0.01106 -3.581 0.00151 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1058 on 24 degrees of freedom  
## Multiple R-squared: 0.9622, Adjusted R-squared: 0.9543   
## F-statistic: 122.1 on 5 and 24 DF, p-value: 2.955e-16

Adjusted R-squared indicates how much a model explains the variability of the response data around its mean. So for model A it explains 94.1% and for model C it explains 95.4%, there for based on Adjusted R-squared model B is the better predictor model.

## [1] "Score for 4f : 0 / 2"

### Part 4g

Test the residuals of **Model C** for serial correlation. Use a 5% level of significance.

Describe the outcome of this test.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 4g -|-|-|-|-|-|-|-|-|-|-|-

dwtest(fit3)

##   
## Durbin-Watson test  
##   
## data: fit3  
## DW = 2.8267, p-value = 0.9895  
## alternative hypothesis: true autocorrelation is greater than 0

There is not sufficient evidence at to reject , that the residuals are not serially correlated (p=.9895).

## [1] "Score for 4g : 0 / 4"

### Part 4h

Using **Model C**, construct a 95% prediction interval for the shell height (Y) for a randomly selected shell with .

Write an interpretation for the interval in the context of this problem.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 4h -|-|-|-|-|-|-|-|-|-|-|-

newcx2 <- 2.4-mean(shells$X2)  
newcx2sq <- newcx2\*newcx2  
newdata <- data.frame(X1=3.6,cx2=newcx2,cx2sq=newcx2sq,X4=3.0,X6=48)  
predict.lm(fit3,newdata,interval="prediction")

## fit lwr upr  
## 1 7.136107 6.907404 7.36481

With 95% confidence, the predicted shell height (Y) for a randomly selected shell with X1=3.6, X2=2.4, X4=3.0, and X6=48, is between 6.907 and 7.365.

## [1] "Score for 4h : 0 / 5"

## Problem 5

A study on the Primary News Source for Americans was conducted using a random sample of 115 Americans in 2015. The results are shown below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | TV | Radio | Newspaper | Internet |
| Sample from 2015 | 38 | 20 | 15 | 42 |
| Distribution in 1995 | 45% | 18% | 16% | 21% |

Conduct the hypothesis test to determine if the distribution of Primary News Source for Americans is the same in 2015 as it was in 1995. Use . State your hypotheses, test statistic, df, p-value, and conclusions, including a practical conclusion in the context of the problem.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 5 -|-|-|-|-|-|-|-|-|-|-|-

dis95 = c(0.45, 0.18, 0.16, 0.21)  
Counts15 = c(38, 20, 15, 42)  
out <- chisq.test(x=Counts15,p=dis95)  
out

##   
## Chi-squared test for given probabilities  
##   
## data: Counts15  
## X-squared = 17.499, df = 3, p-value = 0.000558

:

: At least one proportion is different.

Reject at ( = 0.0006). There is sufficient evidence to show that the distribution of primary newssources in 2015 are different from the 1995 distribution.

## [1] "Score for 5 : 0 / 8"

## Problem 6

In an effort to make better cheese, a company has a random sample of 30 cheese consumers taste 30 specially prepared pieces of Australian cheddar cheese (1 piece for each person). Each subject rated the taste of their piece of cheese as âacceptableâ or ânot acceptable.â One variable measured was called ACETIC and was a quantitative variable ranging from 4.5 to 6.5 units. The other variable recorded was whether the person was a child or an adult

The data file called cheese.rda. This data set is included in the DS705data package.

### Part 6a

Fit the first order model for predicting whether or not the taste of the cheese is acceptable from the acetic value and also whether the person was a child or an adult.

At a 5% level of significance, should either variable be dropped from the model?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 6a -|-|-|-|-|-|-|-|-|-|-|-

data("cheese")  
taste.out <- glm(taste~acetic+person,data=cheese,family="binomial")  
summary(taste.out)

##   
## Call:  
## glm(formula = taste ~ acetic + person, family = "binomial", data = cheese)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.2245 -0.4998 -0.2002 0.3040 1.6066   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -18.709 8.517 -2.197 0.0280 \*  
## acetic 2.787 1.412 1.975 0.0483 \*  
## personAdult 3.096 1.371 2.258 0.0239 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 34.795 on 29 degrees of freedom  
## Residual deviance: 20.550 on 27 degrees of freedom  
## AIC: 26.55  
##   
## Number of Fisher Scoring iterations: 6

Both Remain.

## [1] "Score for 6a : 0 / 4"

### Part 6b

Convert the estimated coefficient of **acetic** to an odds ratio and interpret it in the context of the problem.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 6b -|-|-|-|-|-|-|-|-|-|-|-

exp(2.787)

## [1] 16.23225

The odds of having acceptable cheese increase by 16.23 for each unit increase in acetic.

## [1] "Score for 6b : 0 / 4"

### Part 6c

Compute the predicted probability of a child finding the taste of the cheese acceptable when the value for acetic is 6.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 6c -|-|-|-|-|-|-|-|-|-|-|-

newdata2 <- data.frame(acetic=6,person='Child')  
predict(taste.out, newdata2, type="response")

## 1   
## 0.1206732

## [1] "Score for 6c : 0 / 3"

### Part 6d

Compute a 95% confidence interval for the predicted probability of a child finding the taste of the cheese acceptable when the value for acetic is 6.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 6d -|-|-|-|-|-|-|-|-|-|-|-

out <- predict(taste.out, newdata2, se.fit=TRUE)  
C = .95 # define the level of confidence  
crit = qnorm(1-(1-C)/2) # get the appropriate critical value  
lower = exp(out$fit-crit\*out$se.fit)/(1+exp(out$fit-crit\*out$se.fit))  
upper = exp(out$fit+crit\*out$se.fit)/(1+exp(out$fit+crit\*out$se.fit))  
c(lower,upper)

## 1 1   
## 0.01591904 0.53793967

## [1] "Score for 6d : 0 / 5"

## [1] "Total points earned 0 out of 151"