Lab Assignment: Multiple Linear Regression & Logistic Regression

Christian Acosta

August 1, 2016

Knit a Word file from this R Markdown file for the following exercises. Submit the R markdown file and resulting Word file via D2L Dropbox.

## Exercise 1

A personnel officer in a governmental agency administered three newly developed aptitude tests to a random sample of 25 applicants for entry-level positions in the agency. For the purpose of the study, all 25 applicants were accepted for positions irrespective of their test scores. After a probationary period, each applicant was rated for proficiency on the job.

The scores on the three tests (x1, x2, x3) and the job proficiency score (y) for the 25 employees are in the file JobProf.rda (load JobProf from DS705data)

(Based on an exercise from Applied Linear Statistical Models, 5th ed. by Kutner, Nachtsheim, Neter, & Li)

### Part 1a

Create a scatterplot matrix and the correlation matrix for all of the variables in the data set.

Do any aptitude test scores appear to be linearly related to the proficiency score? Do any relationships appear to be quadratic? Do any aptitude scores appear to be linearly related to each other?

### Answer 1a

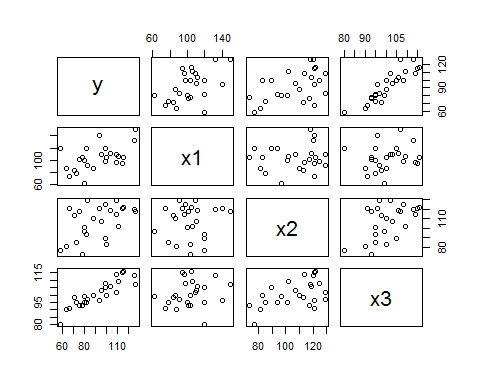
require(DS705data)

## Loading required package: DS705data

data("JobProf")  
head(JobProf)

## y x1 x2 x3  
## 1 88 86 110 100  
## 2 80 62 97 99  
## 3 96 110 107 103  
## 4 76 101 117 93  
## 5 80 100 101 95  
## 6 73 78 85 95

attach(JobProf)  
#JobProf  
pairs(JobProf)



cor(JobProf)

## y x1 x2 x3  
## y 1.0000000 0.5144107 0.4970057 0.8970645  
## x1 0.5144107 1.0000000 0.1022689 0.1807692  
## x2 0.4970057 0.1022689 1.0000000 0.5190448  
## x3 0.8970645 0.1807692 0.5190448 1.0000000

There appears to be heavy correlation between x3 and the proficiency score (correlation value of 0.897). x2 and x3 may be linearly related, but more investigation is require. No relationships appear quadratic.

### Part 1b

Obtain the model summary for the model composed of the three first-order terms and the three cross-product interaction terms (using the centered variables):

Also use R to compute the VIF for each term in the model. Are any of the VIFs over 10?

This model is suffering from the effects of collinearity, which inflates the standard errors of the estimated coefficients.

What do you notice about the overall model p-value (from the F-statistic) and the individual p-values for each term in the model? Does it make sense that the overall model shows statistical significance but no individual term does?

### Answer 1b

#B0 + B1x1 + B2x2 + b3x3 + b4x1x2 + b5x1x3 + b6x2x3 + E  
proficiencyModel <- lm(y ~ x1 + x2 + x3 + x1:x2 + x1:x3 + x2:x3, data = JobProf)  
summary(proficiencyModel)

##   
## Call:  
## lm(formula = y ~ x1 + x2 + x3 + x1:x2 + x1:x3 + x2:x3, data = JobProf)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.513 -3.408 -1.082 2.548 11.593   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -48.965067 142.039396 -0.345 0.734  
## x1 -0.580916 0.820429 -0.708 0.488  
## x2 -0.174913 0.905654 -0.193 0.849  
## x3 1.443371 1.495901 0.965 0.347  
## x1:x2 0.004012 0.004341 0.924 0.368  
## x1:x3 0.004959 0.008893 0.558 0.584  
## x2:x3 -0.002015 0.008399 -0.240 0.813  
##   
## Residual standard error: 5.431 on 18 degrees of freedom  
## Multiple R-squared: 0.9414, Adjusted R-squared: 0.9218   
## F-statistic: 48.17 on 6 and 18 DF, p-value: 4.042e-10

#install.packages("fmsb")  
require(fmsb)

## Loading required package: fmsb

## Warning: package 'fmsb' was built under R version 3.1.3

require(MASS)

## Loading required package: MASS

#install.packages("clusterGeneration")  
require(clusterGeneration)

## Loading required package: clusterGeneration

## Warning: package 'clusterGeneration' was built under R version 3.1.3

#install.packages("HH")  
require(HH)

## Loading required package: HH

## Warning: package 'HH' was built under R version 3.1.3

## Loading required package: lattice

## Loading required package: grid

## Loading required package: latticeExtra

## Warning: package 'latticeExtra' was built under R version 3.1.3

## Loading required package: RColorBrewer

## Warning: package 'RColorBrewer' was built under R version 3.1.3

## Loading required package: multcomp

## Warning: package 'multcomp' was built under R version 3.1.3

## Loading required package: mvtnorm

## Warning: package 'mvtnorm' was built under R version 3.1.3

## Loading required package: survival

## Loading required package: splines

## Loading required package: TH.data

## Warning: package 'TH.data' was built under R version 3.1.3

##   
## Attaching package: 'TH.data'

## The following object is masked from 'package:MASS':  
##   
## geyser

## Loading required package: gridExtra

## Warning: package 'gridExtra' was built under R version 3.1.3

VIF(proficiencyModel) #Returns VIF for entire model.

## [1] 17.05534

vif(proficiencyModel) #Uses HH package and returns VIF for each individual term.

## x1 x2 x3 x1:x2 x1:x3 x2:x3   
## 225.6691 199.6007 142.7966 138.0512 368.6751 308.2454

There are zero terms that meet our standard p value requirements (0.05) despite the overall model's p-value at practically 0. This gives us indiciation that there may be a case of collinearity. VIF also shows as 17.05 for the entire model - which is extremely high. The VIF for each individual term is also extremely high.

### Part 1c

Many times, collinearity can be alleviated by centering the predictor variables. Center the predictor variables x1, x2, and x3 and create new variables to hold them (call them cx1, cx2, and cx3). Furthermore, create a quadratic term for the centered x2 variable.

### Answer 1c

cx1 <- x1-mean(x1)  
cx2 <- x2-mean(x2)  
cx3 <- x3-mean(x3)  
qx2 <- cx2\*cx2

### Part 1d

Now obtain the model summary for the model composed of the three first-order terms and the three cross-product interaction terms using the centered variables:

Use R to compute the VIF for each term in the model. Have the VIF values decreased after the variables are centered? What can you about the overall model p-value (from the F-statistic) and the individual p-values for each term in the model? Does this make more sense?

### Answer 1d

#B0 + B1x1 + B2x2 + b3x3 + b4x1x2 + b5x1x3 + b6x2x3 + E  
proficiencyModelCentered <- lm(y ~ cx1 + cx2 + cx3 + cx1:cx2 + cx1:cx3 + cx2:cx3, data = JobProf)  
summary(proficiencyModelCentered)

##   
## Call:  
## lm(formula = y ~ cx1 + cx2 + cx3 + cx1:cx2 + cx1:cx3 + cx2:cx3,   
## data = JobProf)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.513 -3.408 -1.082 2.548 11.593   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 92.060813 1.325366 69.461 < 2e-16 \*\*\*  
## cx1 0.347097 0.057934 5.991 1.15e-05 \*\*\*  
## cx2 0.036629 0.083585 0.438 0.666   
## cx3 1.740924 0.151386 11.500 9.98e-10 \*\*\*  
## cx1:cx2 0.004012 0.004341 0.924 0.368   
## cx1:cx3 0.004959 0.008893 0.558 0.584   
## cx2:cx3 -0.002015 0.008399 -0.240 0.813   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.431 on 18 degrees of freedom  
## Multiple R-squared: 0.9414, Adjusted R-squared: 0.9218   
## F-statistic: 48.17 on 6 and 18 DF, p-value: 4.042e-10

vif(proficiencyModelCentered) #vif values much lower.

## cx1 cx2 cx3 cx1:cx2 cx1:cx3 cx2:cx3   
## 1.125258 1.700164 1.462463 1.293315 1.432634 1.456335

VIF values are much lower in the centered model. Despite the 'centering' of factors cx1, cx2, cx3 - the three interaction term's p-values all remained the same (but their VIF values decreased). I wonder why?

### Part 1e

Test the significance of all three coefficients for the interaction terms as a subset. Use a 5% level of significance. State and and provide the R output as well as a written conclusion.

Look back and check the individual p-values for the interactions terms from the previous model, how do they compare to the p-value when the interaction terms are tested together as a subset?

### Answer 1e

# Test the significance of all three coefficients for the interaction terms as a subset. Use a 5% level of significance.  
#Both are continuous variables.   
#1st value is y-intercept.  
#2nd value is slope of cx1 when cx2 = 0.  
#3rd value is slope of cx2 when cx1 = 0.  
#4th value is change in slope as one of two variables increases as the y value increases by 1.   
summary(lm(y ~ cx1\*cx2)) #p:0.22836 v. 0.368

##   
## Call:  
## lm(formula = y ~ cx1 \* cx2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -27.006 -11.240 1.552 10.098 21.696   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 91.75249 2.95622 31.037 < 2e-16 \*\*\*  
## cx1 0.41676 0.15048 2.770 0.01148 \*   
## cx2 0.50337 0.17409 2.892 0.00873 \*\*   
## cx1:cx2 0.01299 0.01047 1.241 0.22836   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14.67 on 21 degrees of freedom  
## Multiple R-squared: 0.5008, Adjusted R-squared: 0.4295   
## F-statistic: 7.022 on 3 and 21 DF, p-value: 0.001895

summary(lm(y ~ cx1\*cx3)) #p:0.22 v. 0.584

##   
## Call:  
## lm(formula = y ~ cx1 \* cx3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.499 -3.017 -0.712 2.236 12.867   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 91.913001 1.060875 86.639 < 2e-16 \*\*\*  
## cx1 0.356498 0.053361 6.681 1.30e-06 \*\*\*  
## cx3 1.789513 0.124330 14.393 2.38e-12 \*\*\*  
## cx1:cx3 0.009201 0.007281 1.264 0.22   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.181 on 21 degrees of freedom  
## Multiple R-squared: 0.9377, Adjusted R-squared: 0.9288   
## F-statistic: 105.4 on 3 and 21 DF, p-value: 8.04e-13

summary(lm(y ~ cx2\*cx3)) #p:0.679 v. 0.813

##   
## Call:  
## lm(formula = y ~ cx2 \* cx3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.7912 -6.7972 -0.4643 4.1050 22.1671   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 91.781384 2.075836 44.214 < 2e-16 \*\*\*  
## cx2 0.072727 0.138580 0.525 0.605   
## cx3 1.910886 0.246262 7.760 1.34e-07 \*\*\*  
## cx2:cx3 0.005486 0.013057 0.420 0.679   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 9.106 on 21 degrees of freedom  
## Multiple R-squared: 0.8077, Adjusted R-squared: 0.7802   
## F-statistic: 29.4 on 3 and 21 DF, p-value: 1.043e-07

Null Hypothesis (for all interaction terms): Interaction term (cx1, cx2 or cx3) is not a significant variable in the proficiency model.

Alternative Hypothesis (for all interaction terms): Interaction term (cx1, cx2 or cx3) belongs in the proficieny model.

Conclusion: FAIL TO REJECT null hypothesis at alpha = 0.05. There is not enough evidence to suggest the interaction terms cx1:cx2 (p=.228), cx1:cx3 (p=.22), cx2:cx3 (p=.679) belong in the proficiency model.

When tested individually, the interaction models exhibited a lower p value as compared to being tested together (comparisons are in the code comment above).

### Part 1f

Drop the interaction terms from the model and fit the following model with the quadratic term for :

Should the quadratic term be retained in the model at a 5% level of significance?

### Answer 1f

proficiencyMCQ <- lm(y ~ cx1 + cx2 + cx3 + qx2, data = JobProf)  
summary(proficiencyMCQ)

##   
## Call:  
## lm(formula = y ~ cx1 + cx2 + cx3 + qx2, data = JobProf)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.852 -2.724 -0.918 1.956 10.071   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 90.899655 1.734388 52.410 < 2e-16 \*\*\*  
## cx1 0.340887 0.055159 6.180 4.89e-06 \*\*\*  
## cx2 0.075087 0.080889 0.928 0.364   
## cx3 1.820764 0.152130 11.968 1.42e-10 \*\*\*  
## qx2 0.004530 0.004759 0.952 0.353   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.343 on 20 degrees of freedom  
## Multiple R-squared: 0.9369, Adjusted R-squared: 0.9243   
## F-statistic: 74.3 on 4 and 20 DF, p-value: 1.03e-11

The quadratic term should NOT be retained in the model given p-value of 0.353.

### Part 1g

Drop the quadratic term and fit the model with only the original uncentered variables:

Are there any other terms that should be dropped from the model using the criteria of a 5% level of significance?

### Answer 1g

proficiencySimple <- lm(y ~ cx1 + cx2 + cx3, data = JobProf)  
summary(proficiencySimple)

##   
## Call:  
## lm(formula = y ~ cx1 + cx2 + cx3, data = JobProf)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.7517 -3.0371 -0.4618 1.8358 11.7315   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 92.20000 1.06612 86.482 < 2e-16 \*\*\*  
## cx1 0.34813 0.05451 6.387 2.48e-06 \*\*\*  
## cx2 0.04353 0.07362 0.591 0.561   
## cx3 1.77921 0.14541 12.236 5.08e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.331 on 21 degrees of freedom  
## Multiple R-squared: 0.9341, Adjusted R-squared: 0.9247   
## F-statistic: 99.21 on 3 and 21 DF, p-value: 1.457e-12

Term cx2 should be dropped from the model given p value of 0.561.

### Part 1h

Fit the final model for predicting the proficiency score for the population of all employees for this government agency.

### Answer 1h

proficiencyFinal <- lm(y ~ cx1 + cx3, data = JobProf)  
summary(proficiencyFinal)

##   
## Call:  
## lm(formula = y ~ cx1 + cx3, data = JobProf)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.3489 -2.8086 -0.4546 2.8981 12.6469   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 92.20000 1.05024 87.79 < 2e-16 \*\*\*  
## cx1 0.34846 0.05369 6.49 1.58e-06 \*\*\*  
## cx3 1.82321 0.12307 14.81 6.31e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.251 on 22 degrees of freedom  
## Multiple R-squared: 0.933, Adjusted R-squared: 0.9269   
## F-statistic: 153.2 on 2 and 22 DF, p-value: 1.222e-13

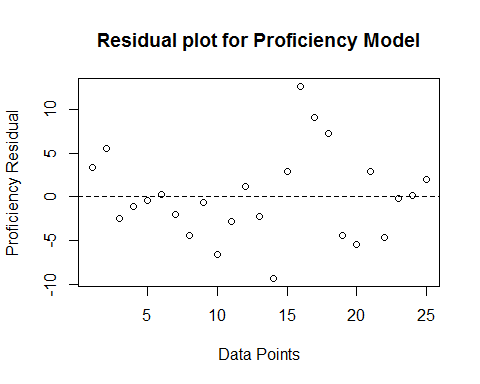
The final model only included cx1 and cx3 with an adjusted R-squared value of 0.9269 meaning ~92.7% variation in y can be explained by the predictor variables.

### Part 1i

Obtain the residuals for your final model and evaluate the residual plots using the "plot" function. Does the regression line appear to be a good fit? Does a visual inspection indicate that the model assumptions appear to be satisfied? Comment.

### Answer 1i

proficiency.res <- resid(proficiencyFinal)  
plot(proficiency.res, main = "Residual plot for Proficiency Model", xlab = "Data Points", ylab = "Proficiency Residual"); abline(h=0, lty='dashed')



Judging by the visual plot, there is no reason to believe that any model assumptions are incorrect (of course, more investigation can be done). Visually speaking: 1. Errors appear to have mean 0. 2. Error variance appears fairly similar above and below 0 on the residual plot. 3. Errors are independant from each other. 4. Errors appear normally distributed (with slight potential for a 'diamon' shape).

### Part 1j

Perform a Shapiro-Wilk test for normality. Use . Comment on the results.

### Answer 1j

shapiro.test(proficiency.res)

##   
## Shapiro-Wilk normality test  
##   
## data: proficiency.res  
## W = 0.9711, p-value = 0.6738

Null Hypothesis: Residual data is normal. Alternative Hypothesis: Residual data is not normal.

Conclusion: Fail to reject null hypothesis of Shapiro-Wilk test (alpha = 0.05).

Residual data appears to be normal (p = 0.6738).

### Part 1k

Perform a Bruesch-Pagan test for homogeneity of variance among the residuals. Use . Comment on the results.

### Answer 1k

require(lmtest)

## Loading required package: lmtest

## Warning: package 'lmtest' was built under R version 3.1.3

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 3.1.3

##   
## Attaching package: 'zoo'

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

bptest(proficiencyFinal)

##   
## studentized Breusch-Pagan test  
##   
## data: proficiencyFinal  
## BP = 0.2578, df = 2, p-value = 0.879

Null Hypothesis: There are equal variances in the residuals of the final proficiency model.

Alternative Hypothesis: There are unequal variances in the residuals of the final proficiency model.

Conclusion: Fail to reject null (alpha = 0.05).

Fail the reject null hypothesis of equal variances. There is not enough evidence to suggest that residual variances are unequal (p = 0.879).

### Part 1l

Perform a Durbin-Watson test for serial correlation the residuals. Use . Comment on the results.

### Answer 1l

dwtest(proficiencyFinal)

##   
## Durbin-Watson test  
##   
## data: proficiencyFinal  
## DW = 1.2807, p-value = 0.03426  
## alternative hypothesis: true autocorrelation is greater than 0

Null Hypothesis: There are no autocorrelations (alpha = 0.05). Alternative Hypothesis: There are autocorrelations (alpha = 0.05).

Conclusion: There is enough evidence to REJECT the null hypothesis at alpha = 0.05.

The results of the Durbin-Watson test show evidence suggesting that there are autocorrelations in the proficiency model (p = 0.03426).

### Part 1m

Obtain a 95% confidence interval for and interpret it in the context of this problem.

### Answer 1m

confint(proficiencyFinal, 'cx3', level=.95)

## 2.5 % 97.5 %  
## cx3 1.567966 2.078445

confint(proficiencyFinal)

## 2.5 % 97.5 %  
## (Intercept) 90.021926 94.3780738  
## cx1 0.237103 0.4598121  
## cx3 1.567966 2.0784446

We are 95% confident that the population mean proficiency score increases 1.568 to 2.078 points for each one point rise on test 3.

### Part 1n

Construct a 95% prediction interval for a randomly selected employee with aptitude scores of and to forecast their proficiency rating at the end of the probationary period. Write an interpretation for the interval in the context of this problem.

### Answer 1n

# x2 is not part of the final model.   
#since the final model has mean subtracted, have to do some manual calculation.   
empTest1 <- 99-mean(x1); empTest1

## [1] -4.36

empTest3 <- 105-mean(x3); empTest3

## [1] 4.2

testEmployee <- data.frame(cx1 = empTest1, cx3 = empTest3)  
predict(proficiencyFinal, testEmployee, interval = "prediction")

## fit lwr upr  
## 1 98.33819 87.16155 109.5148

We are 95% confident that, for a test 1 score of 99 and a test 3 score of 105, the proficiency level will be between 87.16 and 109.51.

## Exercise 2

Consider the scenario from Exercises 12.5 and 12.7 on page 725 of Ott's textbook. There are two categorical variables (Method and Gender) and one quantitative variable (index of English proficiency prior to the program). See the textbook for details on how the qualitative variables are coded using indicator variables.

Load English from DS705data for the data set.

### Part 2a

Use data in the file English.rda to estimate the coefficients for the model in Exercise 12.5:

Obtain the estimated intercept and coefficients and state the estimated mean English proficiency scores for each of the 3 methods of teaching English as a second language.

### Answer 2a

data(English)  
summary(English)

## x1 x2 x3 x4   
## Min. :0.0000 Min. :0.0000 Min. :0.0 Min. :32.00   
## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0 1st Qu.:41.00   
## Median :0.0000 Median :0.0000 Median :0.5 Median :45.00   
## Mean :0.3333 Mean :0.3333 Mean :0.5 Mean :45.08   
## 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0 3rd Qu.:49.00   
## Max. :1.0000 Max. :1.0000 Max. :1.0 Max. :64.00   
## y   
## Min. : 30.00   
## 1st Qu.: 43.75   
## Median : 53.00   
## Mean : 66.53   
## 3rd Qu.: 96.25   
## Max. :141.00

head(English) #x1, x2, x3, x4, y

## x1 x2 x3 x4 y  
## 1 0 0 0 48 45  
## 2 0 0 0 41 40  
## 3 0 0 0 46 45  
## 4 0 0 0 50 46  
## 5 0 0 0 53 40  
## 6 0 0 0 49 38

#x1 - 1 if method 2, 0 if otherwise.  
#x2 - 1 if Method 3, 0 if otherwise.  
#x3 - Assumption is Gender.  
#x4 - Pre-test Score  
#y - dependant variable (post score)  
simpleEnglishModel <- lm(English$y ~ English$x1 + English$x2, data = English)  
summary(simpleEnglishModel)

##   
## Call:  
## lm(formula = English$y ~ English$x1 + English$x2, data = English)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.150 -5.713 -0.225 4.850 34.850   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 44.750 2.202 20.325 <2e-16 \*\*\*  
## English$x1 61.400 3.114 19.719 <2e-16 \*\*\*  
## English$x2 3.950 3.114 1.269 0.21   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 9.847 on 57 degrees of freedom  
## Multiple R-squared: 0.8953, Adjusted R-squared: 0.8916   
## F-statistic: 243.6 on 2 and 57 DF, p-value: < 2.2e-16

#y = 44.75 + 61.4(x1) + 3.95(x2)

Replace the ## symbols with the parameter estimates:

y = 44.75 + 61.4 + 3.95

State the estimated mean English proficiency scores for each of the 3 methods:

Estimated mean for Method 1 = 44.75 Estimated mean for Method 2 = 106.15 Estimated mean for Method 3 = 48.7

### Part 2b

Before fitting the model of Exercise 12.7, create a centered variable for x4 (call it cx4).

Fit the model for Exercise 12.7 using the centered variable x4c:

Using the estimated coefficients, write three separate estimated models, one for each method, relating the scores after 3 months in the program (y) to the index score prior to starting the program ().

### Answer 2b

# Create centered variable cx4 off of x4.  
English$cx4 <- English$x4-mean(English$x4) #loaded directly into the data frame.   
#y = b0 + b1cx4 + b2x1 + b3x2 + b5x1cx4 + b6x2cx4 + E  
complexEnglishModel <- lm(English$y ~ English$cx4 + English$x1 + English$x2 + English$cx4\*English$x1 + English$cx4\*English$x2, data = English)  
summary(complexEnglishModel)

##   
## Call:  
## lm(formula = English$y ~ English$cx4 + English$x1 + English$x2 +   
## English$cx4 \* English$x1 + English$cx4 \* English$x2, data = English)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.845 -4.696 -0.110 4.178 19.470   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 44.7602 1.6205 27.621 < 2e-16 \*\*\*  
## English$cx4 0.1220 0.2983 0.409 0.6841   
## English$x1 59.9319 2.3011 26.045 < 2e-16 \*\*\*  
## English$x2 4.2308 2.2997 1.840 0.0713 .   
## English$cx4:English$x1 1.7797 0.4039 4.407 5.02e-05 \*\*\*  
## English$cx4:English$x2 0.3038 0.4104 0.740 0.4624   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.246 on 54 degrees of freedom  
## Multiple R-squared: 0.9463, Adjusted R-squared: 0.9413   
## F-statistic: 190.2 on 5 and 54 DF, p-value: < 2.2e-16

OVERALL --

Method 1 -- both x1 and x2 are zeroed out.

Method 2 -- x1 = 1, x2 = 0.

Method 3 -- x1 = 0, x2 = 1.

## Exercise 3

Ninety members (aged = 18.1 â 23.4 years) of three Division I womenâs intercollegiate rowing teams (National Collegiate Athletic Association) within the Big Ten Conference volunteered to participate in a study to predict race time for female collegiate rowers from nineteen physical characteristics.

Data is in the file rowtime.rda (load from DS705data). The race times are in the variable named "racetime".

### Part 3a

Load the data and use head(rowtime) to see the other variable names and the first 6 values of each.

### Answer 3a

data(rowtime)  
head(rowtime) #racetime holds ttc.

## racetime tall weight armspan flexarm thighci calfcir tricep biceps  
## 1 470.3 171.5 86.7 172.085 34.2 65.5 40.4 21 19  
## 2 469.2 167.8 72.6 155.575 31.2 59.4 39.5 24 11  
## 3 509.0 169.3 69.4 167.000 31.0 57.5 39.0 22 19  
## 4 516.0 157.8 58.6 158.115 29.5 54.0 37.0 19 12  
## 5 465.0 172.0 72.8 175.895 33.0 55.0 38.0 21 7  
## 6 480.5 176.2 71.7 170.815 32.5 54.5 37.0 17 7  
## thigh estffm estfm bestsnr bestvj legpower endo meso ecto  
## 1 29 66.53111 20.14889 43 21 139.90643 6.84670 4.02678 0.29427  
## 2 34 54.41205 18.17795 25 16 102.26945 6.09077 4.66443 1.00103  
## 3 35 52.14987 17.25013 41 17 100.78434 5.78748 3.88055 1.57270  
## 4 13 47.25539 11.33461 44 13 72.96047 5.75961 4.20958 1.20026  
## 5 23 59.45383 13.31617 49 18 108.74211 4.84827 4.92608 1.61281  
## 6 29 56.61784 15.11216 39 15 97.84882 4.38835 3.24785 2.49913  
## expvarsity preexper  
## 1 0 1  
## 2 0 0  
## 3 0 0  
## 4 0 0  
## 5 0 0  
## 6 0 0

attach(rowtime)

### Part 3b

Use the **regsubsets** function to find the "best" model for predicting the response variable rowtime with up to 8 of the 19 predictor variables in the data set. Produce the summary and the plot for the best single models with up to 8 predictors according to .

Which independent variables are in the best model with 8 predictors when the is the criterion for selection?

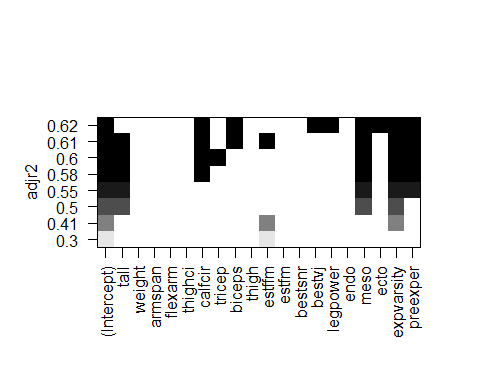
### Answer 3b

#?regsubsets  
require(leaps)

## Loading required package: leaps

## Warning: package 'leaps' was built under R version 3.1.3

require(MASS)  
regmodel <- regsubsets(racetime~., data=rowtime, nvmax = 8)  
plot(regmodel, scale="adjr2")



#calfcir, biceps, bestvj, legpower, meso, ecto, expvarsity, preexper with adj r^2 of .62.

The eigh best predictors for selection (based on adjusted R^2) are: 1. Calf Circumference. 2. Biceps Strength. 3. Best Vertical Jump. 4. Leg Power. 5. Meso body type. 6. Ecto body type. 7. Varsity Experience. 8. Previous Experience.

### Part 3c

Use the **step** function with forward selection to find the "best" model for predicting the response variable rowtime.

Which independent variables are in the model selected? What is the AIC value for this model?

### Answer 3c

null <- lm(racetime~1, data=rowtime)  
full <- lm(racetime~., data=rowtime)  
step(null, scope=list(lower=null, upper=full, direction="forward"))

## Start: AIC=575.99  
## racetime ~ 1  
##   
## Df Sum of Sq RSS AIC  
## + estffm 1 16415.9 36552 544.60  
## + tall 1 13189.7 39778 552.21  
## + weight 1 12987.5 39980 552.67  
## + legpower 1 8478.5 44489 562.29  
## + expvarsity 1 7731.5 45236 563.79  
## + flexarm 1 7190.1 45777 564.86  
## + preexper 1 5346.5 47621 568.41  
## + thighci 1 4806.2 48161 569.43  
## + estfm 1 4288.2 48679 570.39  
## + armspan 1 3492.1 49476 571.85  
## + calfcir 1 2072.4 50895 574.39  
## <none> 52968 575.99  
## + meso 1 932.8 52035 576.39  
## + bestvj 1 203.9 52764 577.64  
## + ecto 1 110.7 52857 577.80  
## + thigh 1 102.8 52865 577.81  
## + biceps 1 76.4 52891 577.86  
## + bestsnr 1 49.6 52918 577.90  
## + tricep 1 42.6 52925 577.91  
## + endo 1 12.6 52955 577.97  
##   
## Step: AIC=544.6  
## racetime ~ estffm  
##   
## Df Sum of Sq RSS AIC  
## + expvarsity 1 5950.0 30602 530.61  
## + biceps 1 4411.1 32141 535.03  
## + preexper 1 3700.7 32851 536.99  
## + ecto 1 3276.0 33276 538.15  
## + tall 1 3072.6 33479 538.70  
## + endo 1 2677.3 33874 539.75  
## + calfcir 1 2326.6 34225 540.68  
## + tricep 1 2238.5 34313 540.91  
## + estfm 1 1274.1 35278 543.41  
## + weight 1 1270.1 35282 543.42  
## <none> 36552 544.60  
## + thighci 1 762.4 35789 544.70  
## + meso 1 705.0 35847 544.85  
## + flexarm 1 690.1 35862 544.89  
## + thigh 1 433.5 36118 545.53  
## + legpower 1 311.0 36241 545.83  
## + bestvj 1 53.8 36498 546.47  
## + armspan 1 53.4 36498 546.47  
## + bestsnr 1 3.1 36549 546.59  
## - estffm 1 16415.9 52968 575.99  
##   
## Step: AIC=530.61  
## racetime ~ estffm + expvarsity  
##   
## Df Sum of Sq RSS AIC  
## + tall 1 3673.9 26928 521.10  
## + preexper 1 3345.3 27256 522.19  
## + calfcir 1 3235.8 27366 522.55  
## + ecto 1 3153.8 27448 522.82  
## + biceps 1 3060.2 27542 523.13  
## + legpower 1 1245.3 29356 528.87  
## + endo 1 1169.4 29432 529.10  
## + bestvj 1 1086.4 29515 529.36  
## + meso 1 992.0 29610 529.64  
## + tricep 1 681.0 29921 530.58  
## <none> 30602 530.61  
## + estfm 1 521.4 30080 531.06  
## + weight 1 516.1 30086 531.08  
## + flexarm 1 419.8 30182 531.37  
## + thighci 1 311.9 30290 531.69  
## + armspan 1 292.3 30309 531.75  
## + thigh 1 74.2 30528 532.39  
## + bestsnr 1 0.7 30601 532.61  
## - expvarsity 1 5950.0 36552 544.60  
## - estffm 1 14634.4 45236 563.79  
##   
## Step: AIC=521.1  
## racetime ~ estffm + expvarsity + tall  
##   
## Df Sum of Sq RSS AIC  
## + preexper 1 3687.9 23240 509.84  
## + bestvj 1 1256.1 25672 518.80  
## + meso 1 1247.8 25680 518.83  
## + calfcir 1 1212.2 25716 518.95  
## + legpower 1 1144.1 25784 519.19  
## + biceps 1 1141.2 25787 519.20  
## <none> 26928 521.10  
## + tricep 1 276.8 26651 522.17  
## + armspan 1 239.2 26689 522.30  
## + endo 1 198.0 26730 522.44  
## + estfm 1 123.0 26805 522.69  
## + weight 1 119.8 26808 522.70  
## + bestsnr 1 42.8 26885 522.96  
## + ecto 1 28.0 26900 523.01  
## + flexarm 1 19.4 26908 523.03  
## + thigh 1 8.9 26919 523.07  
## + thighci 1 7.7 26920 523.07  
## - tall 1 3673.9 30602 530.61  
## - estffm 1 4923.3 31851 534.21  
## - expvarsity 1 6551.3 33479 538.70  
##   
## Step: AIC=509.84  
## racetime ~ estffm + expvarsity + tall + preexper  
##   
## Df Sum of Sq RSS AIC  
## + biceps 1 1637.2 21603 505.27  
## + meso 1 827.2 22413 508.58  
## + calfcir 1 810.2 22430 508.65  
## + tricep 1 582.5 22657 509.56  
## <none> 23240 509.84  
## + legpower 1 479.6 22760 509.97  
## + bestvj 1 437.9 22802 510.13  
## + estfm 1 383.2 22857 510.35  
## + weight 1 382.7 22857 510.35  
## + endo 1 343.9 22896 510.50  
## + ecto 1 200.9 23039 511.06  
## + thigh 1 183.4 23057 511.13  
## + armspan 1 167.4 23073 511.19  
## + bestsnr 1 27.6 23212 511.74  
## + thighci 1 14.5 23225 511.79  
## + flexarm 1 0.5 23239 511.84  
## - preexper 1 3687.9 26928 521.10  
## - tall 1 4016.5 27256 522.19  
## - estffm 1 4022.2 27262 522.21  
## - expvarsity 1 6190.7 29431 529.10  
##   
## Step: AIC=505.27  
## racetime ~ estffm + expvarsity + tall + preexper + biceps  
##   
## Df Sum of Sq RSS AIC  
## + meso 1 951.9 20651 503.21  
## + bestvj 1 523.9 21079 505.06  
## + calfcir 1 519.6 21083 505.08  
## <none> 21603 505.27  
## + legpower 1 263.5 21339 506.16  
## + armspan 1 153.4 21449 506.63  
## + thighci 1 153.1 21450 506.63  
## + flexarm 1 96.6 21506 506.87  
## + bestsnr 1 36.2 21567 507.12  
## + endo 1 32.9 21570 507.13  
## + weight 1 32.5 21570 507.13  
## + estfm 1 32.1 21571 507.13  
## + thigh 1 25.4 21577 507.16  
## + tricep 1 7.8 21595 507.24  
## + ecto 1 1.6 21601 507.26  
## - biceps 1 1637.2 23240 509.84  
## - tall 1 1754.6 23357 510.30  
## - preexper 1 4183.8 25787 519.20  
## - expvarsity 1 4779.8 26383 521.26  
## - estffm 1 5600.4 27203 524.01  
##   
## Step: AIC=503.21  
## racetime ~ estffm + expvarsity + tall + preexper + biceps + meso  
##   
## Df Sum of Sq RSS AIC  
## + calfcir 1 1601.4 19049 497.95  
## + bestvj 1 746.5 19904 501.90  
## + legpower 1 553.0 20098 502.77  
## <none> 20651 503.21  
## - estffm 1 640.0 21291 503.96  
## + tricep 1 174.8 20476 504.45  
## + ecto 1 107.3 20544 504.74  
## + armspan 1 74.7 20576 504.89  
## + bestsnr 1 42.1 20609 505.03  
## + thighci 1 36.6 20614 505.05  
## + thigh 1 9.7 20641 505.17  
## + estfm 1 7.5 20643 505.18  
## + weight 1 6.8 20644 505.18  
## + flexarm 1 4.5 20646 505.19  
## + endo 1 3.1 20648 505.20  
## - meso 1 951.9 21603 505.27  
## - biceps 1 1761.9 22413 508.58  
## - tall 1 2450.9 23102 511.31  
## - preexper 1 3730.4 24381 516.16  
## - expvarsity 1 4687.5 25338 519.62  
##   
## Step: AIC=497.95  
## racetime ~ estffm + expvarsity + tall + preexper + biceps + meso +   
## calfcir  
##   
## Df Sum of Sq RSS AIC  
## + bestvj 1 481.3 18568 497.65  
## <none> 19049 497.95  
## + legpower 1 286.2 18763 498.59  
## + flexarm 1 278.2 18771 498.62  
## + thighci 1 257.7 18792 498.72  
## + armspan 1 196.4 18853 499.02  
## + tricep 1 130.8 18919 499.33  
## + bestsnr 1 122.4 18927 499.37  
## + ecto 1 35.4 19014 499.78  
## + thigh 1 25.9 19024 499.83  
## + weight 1 10.0 19039 499.90  
## + estfm 1 9.6 19040 499.90  
## + endo 1 1.2 19048 499.94  
## - estffm 1 980.1 20030 500.46  
## - biceps 1 1299.9 20349 501.89  
## - calfcir 1 1601.4 20651 503.21  
## - meso 1 2033.7 21083 505.08  
## - preexper 1 2774.2 21824 508.18  
## - tall 1 3005.4 22055 509.13  
## - expvarsity 1 5313.5 24363 518.09  
##   
## Step: AIC=497.65  
## racetime ~ estffm + expvarsity + tall + preexper + biceps + meso +   
## calfcir + bestvj  
##   
## Df Sum of Sq RSS AIC  
## <none> 18568 497.65  
## - bestvj 1 481.3 19049 497.95  
## + armspan 1 233.7 18334 498.51  
## + thighci 1 224.4 18344 498.55  
## + tricep 1 180.0 18388 498.77  
## + flexarm 1 155.8 18412 498.89  
## + legpower 1 146.0 18422 498.93  
## + bestsnr 1 38.3 18530 499.46  
## + ecto 1 28.3 18540 499.51  
## + endo 1 9.6 18559 499.60  
## + estfm 1 4.5 18564 499.62  
## + weight 1 4.3 18564 499.62  
## + thigh 1 0.0 18568 499.65  
## - estffm 1 929.2 19497 500.04  
## - calfcir 1 1336.2 19904 501.90  
## - biceps 1 1412.9 19981 502.25  
## - preexper 1 2127.0 20695 505.41  
## - meso 1 2152.9 20721 505.52  
## - tall 1 3209.4 21778 509.99  
## - expvarsity 1 5793.6 24362 520.09

##   
## Call:  
## lm(formula = racetime ~ estffm + expvarsity + tall + preexper +   
## biceps + meso + calfcir + bestvj, data = rowtime)  
##   
## Coefficients:  
## (Intercept) estffm expvarsity tall preexper   
## 867.706 -1.299 -17.579 -2.404 -10.983   
## biceps meso calfcir bestvj   
## 1.155 -10.389 2.810 1.002

The AIC for the forward step model is 497.65. This includes the independent variables: estffm + expvarsity + tall + preexper + biceps + meso + calfcir + bestvj.

## Exercise 4

A study was conducted whereby the type of anesthetic (A or B), nausea after the surgery (Yes or No), the amount of pain medication taken during the recovery period, and age for a random sample of 72 patients undergoing reconstructive knee surgery.

The data is in the file anesthesia.rda (load from DS705data).

### Part 4a

Obtain the output from R (including the Wald tests for coefficients - so use "summary" function) for the logistic regression model with nausea as the dependent variable and the type of anesthetic (anesthetic) and the amount of pain medication (painmed) taken as the predictor variables.

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

a <- data("anesthesia")  
head(anesthesia) #age(#), painmed(#), anesthetic(A or B) to determine nausea.

## age painmed anesthetic nausea  
## 1 49.90137 53.49 A No  
## 2 70.23562 21.66 A No  
## 3 64.07123 86.50 A No  
## 4 73.31507 37.34 A No  
## 5 59.72603 87.00 A No  
## 6 80.60822 18.33 A No

attach(anesthesia)  
nauseaModel <- glm(nausea ~ anesthetic + painmed, family = "binomial")  
summary(nauseaModel)

##   
## Call:  
## glm(formula = nausea ~ anesthetic + painmed, family = "binomial")  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.53173 -0.59107 -0.08386 0.76926 1.94523   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.415567 0.870914 -2.774 0.005544 \*\*   
## anestheticB -0.831144 0.637377 -1.304 0.192232   
## painmed 0.033737 0.008843 3.815 0.000136 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 99.813 on 71 degrees of freedom  
## Residual deviance: 66.373 on 69 degrees of freedom  
## AIC: 72.373  
##   
## Number of Fisher Scoring iterations: 5

### Part 4b

What is the outcome of the hypothesis test that the coefficient of **anesthetic** is "zero" vs "not zero" at a 5% level of significance? (use the Wald test from the R output from the logistic regression you performed)

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

Null Hypothesis: Coefficients in the population are 0. Alternative Hypothesis: Coefficients in the population are different than 0.

Anesthetic A (p = .0055) appears to be significant at the 5% level while Anesthetic B (p=0.1922) does not. We can't be sure if the coefficient for Anesthetic B is different from zero.

### Part 4c

What is the outcome of the hypothesis test that the coefficient of **painmed** is "zero" vs "not zero" at a 5% level of significance? (use the Wald test from the R output from the logistic regression you performed)

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

Null Hypothesis: Coefficient for painmed is 0. Alternative Hypothesis: Coefficient for painmed is different than 0.

There is sufficient evidence to claim that the coefficient value for painmed is different than zero (p = 0.000136).

### Part 4d

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

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

exp(0.033737)

## [1] 1.034313

The odds of having nausea increases by 3.43% for each unit increase in the independent variable painmed.

### Part 4e

Compute the predicted probabilities of a reconstructive knee surgery patient having nausea in the recovery time after surgery for when 50 units of pain medication and anesthetic A are used.

Comment on this probability.

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

kneeData <- data.frame(anesthetic="A", painmed=50)  
predict(nauseaModel, kneeData, type="response")

## 1   
## 0.3254815

The probability of a patient experiencing nausea in recovery time after surgery when 50 units of pain medication and anesthetic A are used is 32.55%.