Lab Assignment: Multiple Linear Regression

Spencer Swartz

March 26, 2017

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

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

require(HH)

## Loading required package: HH

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

## Loading required package: lattice

## Loading required package: grid

## Loading required package: latticeExtra

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

## Loading required package: RColorBrewer

## Loading required package: multcomp

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

## Loading required package: mvtnorm

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

## Loading required package: survival

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

## Loading required package: TH.data

## Warning: package 'TH.data' was built under R version 3.3.3

## Loading required package: MASS

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

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

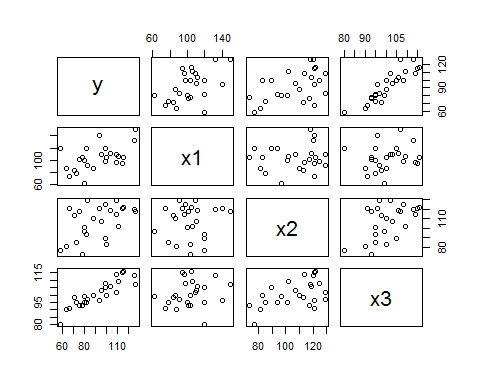
## Loading required package: gridExtra

require(leaps)

## Loading required package: leaps

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

data(JobProf)  
pairs(y~x1+x2+x3,data=JobProf)



mat <- cbind(JobProf$y,JobProf$x1,JobProf$x2,JobProf$x3)  
cor(mat)

## [,1] [,2] [,3] [,4]  
## [1,] 1.0000000 0.5144107 0.4970057 0.8970645  
## [2,] 0.5144107 1.0000000 0.1022689 0.1807692  
## [3,] 0.4970057 0.1022689 1.0000000 0.5190448  
## [4,] 0.8970645 0.1807692 0.5190448 1.0000000

Aptitude test 3 looks to be the most linearly related to the proficiently scored. Beyond that it looks as if they may be a slight correlation between multiple tests but nothing very strong.

### 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

fit <- lm(y~x1+x2+x3+x1:x2+x1:x3+x2:x3,JobProf)  
summary(fit)

##   
## 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

vif(fit)

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

The overall p-value seems to show that the model is of good fit, but when looking at the p-values for each individual coefficient not one is significant. This is likely a cause of collinearity and that claim is supported in the VIF test output.

### 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 <- JobProf$x1- mean(JobProf$x1)  
cx2 <- JobProf$x2- mean(JobProf$x2)  
cx3 <- JobProf$x3- mean(JobProf$x3)  
cx2sq <- 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

newfit <- lm(y~cx1+cx2+cx3+cx1:cx2+cx1:cx3+cx2:cx3,JobProf)  
summary(newfit)

##   
## 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(newfit)

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

Now that some of the coefficients p-value are significant, it now makes more since that the overall pvalues is below .05. the VIF terms seem to show that the problem of colinearity has been fixed.

### 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

rednewfit <- lm(y~cx1+cx2+cx3,JobProf)  
summary(rednewfit)

##   
## 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

anova(rednewfit,newfit)

## Analysis of Variance Table  
##   
## Model 1: y ~ cx1 + cx2 + cx3  
## Model 2: y ~ cx1 + cx2 + cx3 + cx1:cx2 + cx1:cx3 + cx2:cx3  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 21 596.72   
## 2 18 530.86 3 65.861 0.7444 0.5395

All of the coefficients are 0. At least one of the coefficients is not 0.

At = .05 we do not reject the that all of the coefficients are 0.

### 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

sqfit <- lm(y~cx1+cx2+cx3+cx2sq,JobProf)  
summary(sqfit)

##   
## Call:  
## lm(formula = y ~ cx1 + cx2 + cx3 + cx2sq, 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 \*\*\*  
## cx2sq 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

vif(sqfit)

## cx1 cx2 cx3 cx2sq   
## 1.053967 1.645200 1.525984 1.582072

At the = 0.05 significance level the quadratic term should not be retained in the model.

### 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

regfit <- lm(y~x1+x2+x3,JobProf)  
summary(regfit)

##   
## Call:  
## lm(formula = y ~ x1 + x2 + x3, 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) -127.77378 12.88053 -9.920 2.23e-09 \*\*\*  
## x1 0.34813 0.05451 6.387 2.48e-06 \*\*\*  
## x2 0.04353 0.07362 0.591 0.561   
## x3 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

vif(regfit)

## x1 x2 x3   
## 1.033886 1.368890 1.400332

Yes, x2 would be the next term to be dropped at the level of significance.

### Part 1h

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

### Answer 1h

finalfit <- lm(y~x1+x3,JobProf)  
summary(finalfit)

##   
## Call:  
## lm(formula = y ~ x1 + x3, 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) -127.59569 12.68526 -10.06 1.09e-09 \*\*\*  
## x1 0.34846 0.05369 6.49 1.58e-06 \*\*\*  
## x3 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

vif(finalfit)

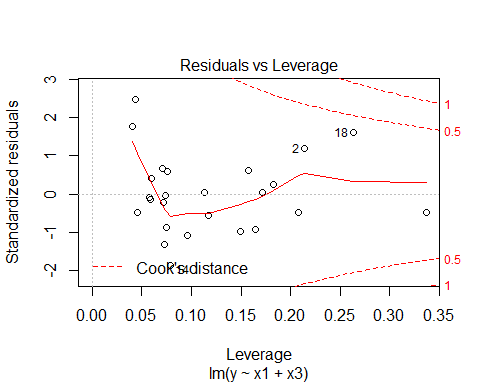
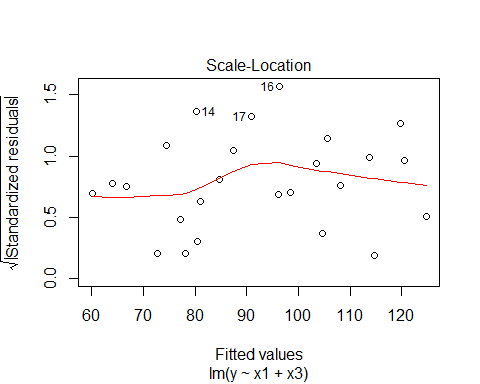
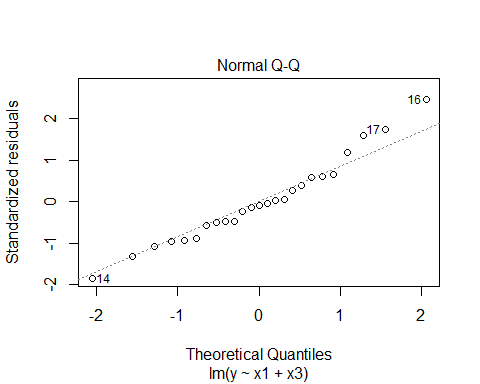
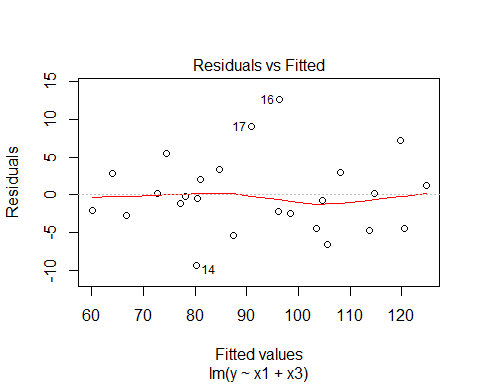
## x1 x3   
## 1.033781 1.033781

### 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

plot(finalfit)



It looks that the regression model is pretty good fit. It also looks that the model assumptions are satisfied.

### Part 1j

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

### Answer 1j

shapiro.test(finalfit$resid)

##   
## Shapiro-Wilk normality test  
##   
## data: finalfit$resid  
## W = 0.97113, p-value = 0.6738

Based on the p-value it looks as if the residuals come from a normal distribution.

### Part 1k

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

### Answer 1k

bptest(finalfit)

##   
## studentized Breusch-Pagan test  
##   
## data: finalfit  
## BP = 0.25783, df = 2, p-value = 0.879

At a 5% level of significance, this test indicates that the error variances are constant, breaking the model assumption.

### Part 1l

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

### Answer 1l

dwtest(finalfit)

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

At a 5% level of significance, this test indicates that the residuals are serially correlated, breaking the model assumption.

### Part 1m

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

### Answer 1m

confint(finalfit)

## 2.5 % 97.5 %  
## (Intercept) -153.903314 -101.2880611  
## x1 0.237103 0.4598121  
## x3 1.567966 2.0784446

with 95% confidence we can assume that the y intercept is between -153.90 and -101.29. with 95% confidence we can assume that the x1 coefficient is between 0.24 and 0.46. with 95% confidence we can assume that the x3 coefficient is between 1.57 and 2.08.

### 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

newdata <- data.frame(x1=99,x2=112,x3=105)  
predict.lm(finalfit,newdata,interval="confidence")

## fit lwr upr  
## 1 98.33819 95.82481 100.8516

The 95% prediction interval for a subject with aptitude scores of and is between 95.82 and 100.85.

## 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.

### 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)  
Engfit <- lm(y~x1+x2,English)  
summary(Engfit)

##   
## Call:  
## lm(formula = y ~ x1 + 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 \*\*\*  
## x1 61.400 3.114 19.719 <2e-16 \*\*\*  
## 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

vif(Engfit)

## x1 x2   
## 1.333333 1.333333

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

cx4 <- English$x4- mean(English$x4)  
Engfit <- lm(y~cx4+x1+x2+x1:cx4+x2:cx4,English)  
summary(Engfit)

##   
## Call:  
## lm(formula = y ~ cx4 + x1 + x2 + x1:cx4 + x2:cx4, 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 \*\*\*  
## cx4 0.1220 0.2983 0.409 0.6841   
## x1 59.9319 2.3011 26.045 < 2e-16 \*\*\*  
## x2 4.2308 2.2997 1.840 0.0713 .   
## cx4:x1 1.7797 0.4039 4.407 5.02e-05 \*\*\*  
## cx4: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

vif(Engfit)

## cx4 x1 x2 cx4:x1 cx4:x2   
## 3.356921 1.344551 1.342912 2.225276 2.139974

:

:

:

:

## 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. 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 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

### 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

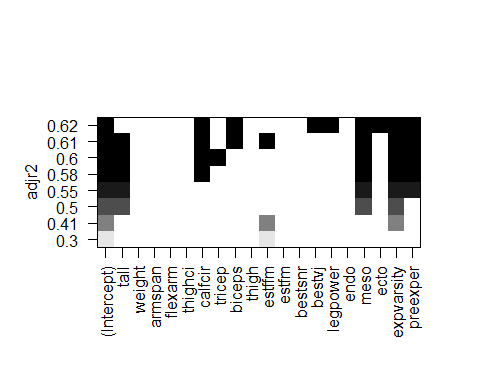
allmods <- regsubsets(racetime~.,nvmax=8, data=rowtime)  
summary(allmods)

## Subset selection object  
## Call: regsubsets.formula(racetime ~ ., nvmax = 8, data = rowtime)  
## 19 Variables (and intercept)  
## Forced in Forced out  
## tall FALSE FALSE  
## weight FALSE FALSE  
## armspan FALSE FALSE  
## flexarm FALSE FALSE  
## thighci FALSE FALSE  
## calfcir FALSE FALSE  
## tricep FALSE FALSE  
## biceps FALSE FALSE  
## thigh FALSE FALSE  
## estffm FALSE FALSE  
## estfm FALSE FALSE  
## bestsnr FALSE FALSE  
## bestvj FALSE FALSE  
## legpower FALSE FALSE  
## endo FALSE FALSE  
## meso FALSE FALSE  
## ecto FALSE FALSE  
## expvarsity FALSE FALSE  
## preexper FALSE FALSE  
## 1 subsets of each size up to 8  
## Selection Algorithm: exhaustive  
## tall weight armspan flexarm thighci calfcir tricep biceps thigh  
## 1 ( 1 ) " " " " " " " " " " " " " " " " " "   
## 2 ( 1 ) " " " " " " " " " " " " " " " " " "   
## 3 ( 1 ) "\*" " " " " " " " " " " " " " " " "   
## 4 ( 1 ) "\*" " " " " " " " " " " " " " " " "   
## 5 ( 1 ) "\*" " " " " " " " " "\*" " " " " " "   
## 6 ( 1 ) "\*" " " " " " " " " "\*" "\*" " " " "   
## 7 ( 1 ) "\*" " " " " " " " " "\*" " " "\*" " "   
## 8 ( 1 ) " " " " " " " " " " "\*" " " "\*" " "   
## estffm estfm bestsnr bestvj legpower endo meso ecto expvarsity  
## 1 ( 1 ) "\*" " " " " " " " " " " " " " " " "   
## 2 ( 1 ) "\*" " " " " " " " " " " " " " " "\*"   
## 3 ( 1 ) " " " " " " " " " " " " "\*" " " "\*"   
## 4 ( 1 ) " " " " " " " " " " " " "\*" " " "\*"   
## 5 ( 1 ) " " " " " " " " " " " " "\*" " " "\*"   
## 6 ( 1 ) " " " " " " " " " " " " "\*" " " "\*"   
## 7 ( 1 ) "\*" " " " " " " " " " " "\*" " " "\*"   
## 8 ( 1 ) " " " " " " "\*" "\*" " " "\*" "\*" "\*"   
## preexper  
## 1 ( 1 ) " "   
## 2 ( 1 ) " "   
## 3 ( 1 ) " "   
## 4 ( 1 ) "\*"   
## 5 ( 1 ) "\*"   
## 6 ( 1 ) "\*"   
## 7 ( 1 ) "\*"   
## 8 ( 1 ) "\*"

summary(allmods)$adjr2

## [1] 0.3020824 0.4089748 0.4967972 0.5546830 0.5846293 0.5970301 0.6096559  
## [8] 0.6162122

plot(allmods, scale="adjr2")



The selected variables are as follows: intercept calfcir biceps bestvj legpower meso ecto expvarsity preexper

### Part 3c

Use the **step** function with backward selection to find the "best" model for predicting the response variable rowtime. Recall that the formula structure y~. will produce the model using y as the response variable and all other variables in the data set as the predictors; in this set racetime is the response(not y) and all other variables are potential predictors.

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

### Answer 3c

full <- lm(racetime~.,data=rowtime)  
step(full,direction="backward")

## Start: AIC=512.17  
## racetime ~ tall + weight + armspan + flexarm + thighci + calfcir +   
## tricep + biceps + thigh + estffm + estfm + bestsnr + bestvj +   
## legpower + endo + meso + ecto + expvarsity + preexper  
##   
## Df Sum of Sq RSS AIC  
## - bestsnr 1 1.1 17090 510.18  
## - thigh 1 1.3 17090 510.18  
## - endo 1 4.4 17094 510.20  
## - tricep 1 10.7 17100 510.23  
## - ecto 1 65.4 17155 510.52  
## - weight 1 76.6 17166 510.58  
## - estffm 1 77.2 17166 510.58  
## - estfm 1 83.8 17173 510.62  
## - flexarm 1 155.8 17245 510.99  
## - armspan 1 227.7 17317 511.37  
## - thighci 1 286.5 17376 511.67  
## <none> 17089 512.17  
## - legpower 1 452.9 17542 512.53  
## - bestvj 1 569.5 17659 513.13  
## - biceps 1 572.4 17662 513.14  
## - tall 1 807.5 17897 514.33  
## - calfcir 1 1453.5 18543 517.52  
## - preexper 1 1748.1 18837 518.94  
## - meso 1 2299.2 19388 521.54  
## - expvarsity 1 5176.8 22266 533.99  
##   
## Step: AIC=510.18  
## racetime ~ tall + weight + armspan + flexarm + thighci + calfcir +   
## tricep + biceps + thigh + estffm + estfm + bestvj + legpower +   
## endo + meso + ecto + expvarsity + preexper  
##   
## Df Sum of Sq RSS AIC  
## - thigh 1 1.9 17092 508.19  
## - endo 1 3.7 17094 508.20  
## - tricep 1 9.6 17100 508.23  
## - ecto 1 67.8 17158 508.54  
## - weight 1 84.6 17175 508.62  
## - estffm 1 85.6 17176 508.63  
## - estfm 1 92.3 17183 508.67  
## - flexarm 1 159.8 17250 509.02  
## - armspan 1 234.9 17325 509.41  
## - thighci 1 285.5 17376 509.67  
## <none> 17090 510.18  
## - legpower 1 516.3 17607 510.86  
## - biceps 1 572.9 17663 511.15  
## - bestvj 1 662.2 17752 511.60  
## - tall 1 806.7 17897 512.33  
## - calfcir 1 1463.4 18554 515.57  
## - preexper 1 1766.9 18857 517.04  
## - meso 1 2303.2 19393 519.56  
## - expvarsity 1 5451.5 22542 533.10  
##   
## Step: AIC=508.19  
## racetime ~ tall + weight + armspan + flexarm + thighci + calfcir +   
## tricep + biceps + estffm + estfm + bestvj + legpower + endo +   
## meso + ecto + expvarsity + preexper  
##   
## Df Sum of Sq RSS AIC  
## - endo 1 8.0 17100 506.23  
## - tricep 1 11.9 17104 506.25  
## - ecto 1 66.0 17158 506.54  
## - weight 1 87.8 17180 506.65  
## - estffm 1 88.6 17181 506.66  
## - estfm 1 96.5 17189 506.70  
## - flexarm 1 161.7 17254 507.04  
## - armspan 1 239.4 17332 507.44  
## - thighci 1 290.6 17383 507.71  
## <none> 17092 508.19  
## - legpower 1 514.7 17607 508.86  
## - biceps 1 572.7 17665 509.16  
## - bestvj 1 662.2 17754 509.61  
## - tall 1 834.4 17927 510.48  
## - calfcir 1 1482.3 18574 513.68  
## - preexper 1 1770.3 18863 515.06  
## - meso 1 2302.7 19395 517.57  
## - expvarsity 1 5464.4 22557 531.16  
##   
## Step: AIC=506.23  
## racetime ~ tall + weight + armspan + flexarm + thighci + calfcir +   
## tricep + biceps + estffm + estfm + bestvj + legpower + meso +   
## ecto + expvarsity + preexper  
##   
## Df Sum of Sq RSS AIC  
## - tricep 1 6.4 17107 504.27  
## - ecto 1 61.9 17162 504.56  
## - weight 1 83.0 17183 504.67  
## - estffm 1 83.8 17184 504.67  
## - estfm 1 91.3 17191 504.71  
## - flexarm 1 181.0 17281 505.18  
## - armspan 1 235.7 17336 505.46  
## - thighci 1 282.8 17383 505.71  
## <none> 17100 506.23  
## - legpower 1 541.1 17641 507.04  
## - biceps 1 565.6 17666 507.16  
## - bestvj 1 695.8 17796 507.82  
## - tall 1 826.4 17927 508.48  
## - calfcir 1 1567.9 18668 512.13  
## - preexper 1 1763.4 18863 513.07  
## - meso 1 2295.3 19395 515.57  
## - expvarsity 1 5456.6 22557 529.16  
##   
## Step: AIC=504.27  
## racetime ~ tall + weight + armspan + flexarm + thighci + calfcir +   
## biceps + estffm + estfm + bestvj + legpower + meso + ecto +   
## expvarsity + preexper  
##   
## Df Sum of Sq RSS AIC  
## - ecto 1 72.5 17179 502.65  
## - weight 1 93.2 17200 502.76  
## - estffm 1 94.0 17200 502.76  
## - estfm 1 102.8 17209 502.81  
## - flexarm 1 220.2 17327 503.42  
## - armspan 1 242.2 17349 503.53  
## - thighci 1 279.9 17386 503.73  
## <none> 17107 504.27  
## - legpower 1 623.1 17730 505.49  
## - biceps 1 642.6 17749 505.59  
## - bestvj 1 782.5 17889 506.29  
## - tall 1 840.6 17947 506.58  
## - calfcir 1 1591.5 18698 510.27  
## - preexper 1 1758.8 18865 511.07  
## - meso 1 2290.4 19397 513.58  
## - expvarsity 1 5946.4 23053 529.12  
##   
## Step: AIC=502.65  
## racetime ~ tall + weight + armspan + flexarm + thighci + calfcir +   
## biceps + estffm + estfm + bestvj + legpower + meso + expvarsity +   
## preexper  
##   
## Df Sum of Sq RSS AIC  
## - weight 1 78.8 17258 501.06  
## - estffm 1 79.9 17259 501.06  
## - estfm 1 89.1 17268 501.11  
## - armspan 1 254.4 17433 501.97  
## - flexarm 1 265.2 17444 502.03  
## - thighci 1 304.3 17483 502.23  
## <none> 17179 502.65  
## - legpower 1 553.8 17733 503.50  
## - biceps 1 593.2 17772 503.70  
## - bestvj 1 715.5 17895 504.32  
## - calfcir 1 1644.1 18823 508.87  
## - preexper 1 1700.8 18880 509.14  
## - meso 1 2228.7 19408 511.62  
## - tall 1 3578.6 20758 517.68  
## - expvarsity 1 6028.6 23208 527.72  
##   
## Step: AIC=501.06  
## racetime ~ tall + armspan + flexarm + thighci + calfcir + biceps +   
## estffm + estfm + bestvj + legpower + meso + expvarsity +   
## preexper  
##   
## Df Sum of Sq RSS AIC  
## - estffm 1 9.8 17268 499.11  
## - armspan 1 238.2 17496 500.29  
## - flexarm 1 268.3 17526 500.45  
## - thighci 1 289.1 17547 500.55  
## <none> 17258 501.06  
## - legpower 1 499.5 17757 501.63  
## - estfm 1 548.7 17807 501.88  
## - biceps 1 586.3 17844 502.07  
## - bestvj 1 661.4 17919 502.44  
## - calfcir 1 1707.4 18965 507.55  
## - preexper 1 1889.5 19147 508.41  
## - meso 1 2159.5 19417 509.67  
## - tall 1 3502.6 20760 515.69  
## - expvarsity 1 5959.7 23217 525.76  
##   
## Step: AIC=499.11  
## racetime ~ tall + armspan + flexarm + thighci + calfcir + biceps +   
## estfm + bestvj + legpower + meso + expvarsity + preexper  
##   
## Df Sum of Sq RSS AIC  
## - armspan 1 248.4 17516 498.39  
## - thighci 1 286.5 17554 498.59  
## - flexarm 1 313.1 17581 498.73  
## <none> 17268 499.11  
## - biceps 1 599.0 17867 500.18  
## - estfm 1 831.1 18099 501.34  
## - legpower 1 1570.5 18838 504.94  
## - calfcir 1 1817.1 19085 506.11  
## - preexper 1 1889.0 19157 506.45  
## - bestvj 1 2062.9 19331 507.27  
## - meso 1 2150.4 19418 507.67  
## - tall 1 3592.7 20860 514.12  
## - expvarsity 1 6282.6 23550 525.04  
##   
## Step: AIC=498.39  
## racetime ~ tall + flexarm + thighci + calfcir + biceps + estfm +   
## bestvj + legpower + meso + expvarsity + preexper  
##   
## Df Sum of Sq RSS AIC  
## - flexarm 1 261.9 17778 497.73  
## <none> 17516 498.39  
## - thighci 1 400.0 17916 498.43  
## - biceps 1 721.4 18237 500.03  
## - estfm 1 723.8 18240 500.04  
## - legpower 1 1358.5 18874 503.12  
## - calfcir 1 1711.8 19228 504.79  
## - bestvj 1 1831.8 19348 505.35  
## - preexper 1 1966.2 19482 505.97  
## - meso 1 2104.6 19621 506.61  
## - tall 1 3345.4 20861 512.13  
## - expvarsity 1 6482.8 23999 524.73  
##   
## Step: AIC=497.73  
## racetime ~ tall + thighci + calfcir + biceps + estfm + bestvj +   
## legpower + meso + expvarsity + preexper  
##   
## Df Sum of Sq RSS AIC  
## - thighci 1 297.6 18075 497.22  
## <none> 17778 497.73  
## - estfm 1 656.5 18434 498.99  
## - biceps 1 876.1 18654 500.06  
## - legpower 1 1115.6 18893 501.21  
## - calfcir 1 1450.1 19228 502.79  
## - bestvj 1 1616.2 19394 503.56  
## - meso 1 1894.0 19672 504.84  
## - preexper 1 2036.0 19814 505.49  
## - tall 1 3099.5 20877 510.19  
## - expvarsity 1 6567.6 24345 524.03  
##   
## Step: AIC=497.22  
## racetime ~ tall + calfcir + biceps + estfm + bestvj + legpower +   
## meso + expvarsity + preexper  
##   
## Df Sum of Sq RSS AIC  
## <none> 18075 497.22  
## - estfm 1 433.1 18508 497.36  
## - biceps 1 868.7 18944 499.45  
## - calfcir 1 1260.9 19336 501.29  
## - legpower 1 1421.3 19497 502.04  
## - meso 1 1885.0 19960 504.15  
## - bestvj 1 1887.1 19962 504.16  
## - preexper 1 2168.0 20243 505.42  
## - tall 1 2926.1 21001 508.73  
## - expvarsity 1 6381.6 24457 522.44

##   
## Call:  
## lm(formula = racetime ~ tall + calfcir + biceps + estfm + bestvj +   
## legpower + meso + expvarsity + preexper, data = rowtime)  
##   
## Coefficients:  
## (Intercept) tall calfcir biceps estfm   
## 797.881 -2.284 2.764 1.118 1.665   
## bestvj legpower meso expvarsity preexper   
## 5.101 -1.140 -9.786 -18.370 -11.096

The variables in this model are: tall + calfcir + biceps + estfm + bestvj + legpower + meso + expvarsity + preexper

The AIC for this model is: 497.22

### Part 3d

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 3d

lmnull <- lm(racetime~1,data=rowtime)  
step(lmnull,direction="forward",trace = 1, scope = ~tall + weight + armspan + flexarm + thighci + calfcir +   
 tricep + biceps + thigh + estffm + estfm + bestsnr + bestvj +   
 legpower + endo + meso + ecto + expvarsity + preexper)

## 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  
##   
## 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  
##   
## 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  
##   
## Step: AIC=509.84  
## racetime ~ estffm + expvarsity + tall + preexper  
##   
## Df Sum of Sq RSS AIC  
## + biceps 1 1637.17 21603 505.27  
## + meso 1 827.19 22413 508.58  
## + calfcir 1 810.18 22430 508.65  
## + tricep 1 582.46 22657 509.56  
## <none> 23240 509.84  
## + legpower 1 479.60 22760 509.97  
## + bestvj 1 437.89 22802 510.13  
## + estfm 1 383.16 22857 510.35  
## + weight 1 382.67 22857 510.35  
## + endo 1 343.89 22896 510.50  
## + ecto 1 200.92 23039 511.06  
## + thigh 1 183.37 23057 511.13  
## + armspan 1 167.39 23073 511.19  
## + bestsnr 1 27.65 23212 511.74  
## + thighci 1 14.47 23225 511.79  
## + flexarm 1 0.53 23239 511.84  
##   
## Step: AIC=505.27  
## racetime ~ estffm + expvarsity + tall + preexper + biceps  
##   
## Df Sum of Sq RSS AIC  
## + meso 1 951.94 20651 503.21  
## + bestvj 1 523.92 21079 505.06  
## + calfcir 1 519.62 21083 505.08  
## <none> 21603 505.27  
## + legpower 1 263.49 21339 506.16  
## + armspan 1 153.44 21449 506.63  
## + thighci 1 153.08 21450 506.63  
## + flexarm 1 96.62 21506 506.87  
## + bestsnr 1 36.23 21567 507.12  
## + endo 1 32.94 21570 507.13  
## + weight 1 32.50 21570 507.13  
## + estfm 1 32.12 21571 507.13  
## + thigh 1 25.43 21577 507.16  
## + tricep 1 7.76 21595 507.24  
## + ecto 1 1.60 21601 507.26  
##   
## Step: AIC=503.21  
## racetime ~ estffm + expvarsity + tall + preexper + biceps + meso  
##   
## Df Sum of Sq RSS AIC  
## + calfcir 1 1601.37 19049 497.95  
## + bestvj 1 746.45 19904 501.90  
## + legpower 1 553.04 20098 502.77  
## <none> 20651 503.21  
## + tricep 1 174.83 20476 504.45  
## + ecto 1 107.34 20544 504.74  
## + armspan 1 74.75 20576 504.89  
## + bestsnr 1 42.05 20609 505.03  
## + thighci 1 36.63 20614 505.05  
## + thigh 1 9.73 20641 505.17  
## + estfm 1 7.48 20643 505.18  
## + weight 1 6.80 20644 505.18  
## + flexarm 1 4.54 20646 505.19  
## + endo 1 3.15 20648 505.20  
##   
## Step: AIC=497.95  
## racetime ~ estffm + expvarsity + tall + preexper + biceps + meso +   
## calfcir  
##   
## Df Sum of Sq RSS AIC  
## + bestvj 1 481.32 18568 497.65  
## <none> 19049 497.95  
## + legpower 1 286.15 18763 498.59  
## + flexarm 1 278.17 18771 498.62  
## + thighci 1 257.74 18792 498.72  
## + armspan 1 196.38 18853 499.02  
## + tricep 1 130.75 18919 499.33  
## + bestsnr 1 122.38 18927 499.37  
## + ecto 1 35.40 19014 499.78  
## + thigh 1 25.90 19024 499.83  
## + weight 1 10.02 19039 499.90  
## + estfm 1 9.57 19040 499.90  
## + endo 1 1.16 19048 499.94  
##   
## Step: AIC=497.65  
## racetime ~ estffm + expvarsity + tall + preexper + biceps + meso +   
## calfcir + bestvj  
##   
## Df Sum of Sq RSS AIC  
## <none> 18568 497.65  
## + armspan 1 233.742 18334 498.51  
## + thighci 1 224.398 18344 498.55  
## + tricep 1 180.044 18388 498.77  
## + flexarm 1 155.779 18412 498.89  
## + legpower 1 145.988 18422 498.93  
## + bestsnr 1 38.290 18530 499.46  
## + ecto 1 28.280 18540 499.51  
## + endo 1 9.589 18559 499.60  
## + estfm 1 4.491 18564 499.62  
## + weight 1 4.321 18564 499.62  
## + thigh 1 0.008 18568 499.65

##   
## 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 variables in this model are: estffm + expvarsity + tall + preexper + biceps + meso + calfcir + bestvj

The AIC for this model is: 497.65

### Part 3e

Compute the AIC for the the best model with 8 predictors from the **regsubsets** function. How does it compare with the AIC for the two models produced by the backward and forward selection procedure?

Which model is the "best" according to the AIC? (remember, smaller is better for AIC)

### Answer 3e

bestallmods <- lm(racetime~calfcir+biceps+bestvj+legpower+meso+ecto+expvarsity+preexper,data=rowtime)  
AIC(bestallmods)

## [1] 754.7286

The backwards and forwards model have very similar AIC which are both better that that of the best model based on R^2 out of the regsubsets, which is close to twice the size of the stepwise models.