**STAT 40001/MA59800 Statistical Computing Fall 2017**

**Lab -20**

1. How does the cost of a movie depend on its length? Data on the cost (millions of dollars) and the running time (minutes) for major release films in one recent year are provided in the Blackboard.
2. Draw a scatter plot of Time vs. Budget. Also choose different colors to display MPAA Rating.

> data = read.csv('C:\\Users\\wu1114\\Desktop\\Movie\_ratings.csv')

> head(data)

Run.Time Budget Rating

1 91 1.000 PG

2 104 1.000 R

3 97 1.100 R

4 93 1.125 R

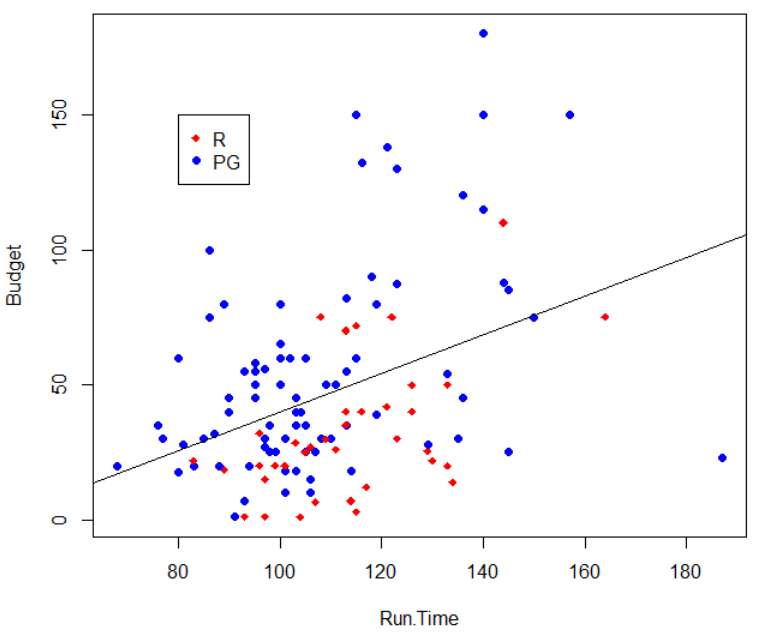
5 115 2.800 R

6 107 6.500 R

> attach(data)

> plot(Run.Time,Budget,col = ifelse(Rating=="R","red","blue"),pch=ifelse(Rating=="R",18,19))

> legend(80,150,pch = c(18,19),col = c("red","blue"),c("R","PG"))



1. Fit a regression model with indicator variable and write out the regression model.

> model = lm(Budget~Run.Time)

> model

Call:

lm(formula = Budget ~ Run.Time)

Coefficients:

(Intercept) Run.Time

-31.3869 0.7144

> model2 = lm(Budget~Run.Time+Rating)

> model2

Call:

lm(formula = Budget ~ Run.Time + Rating)

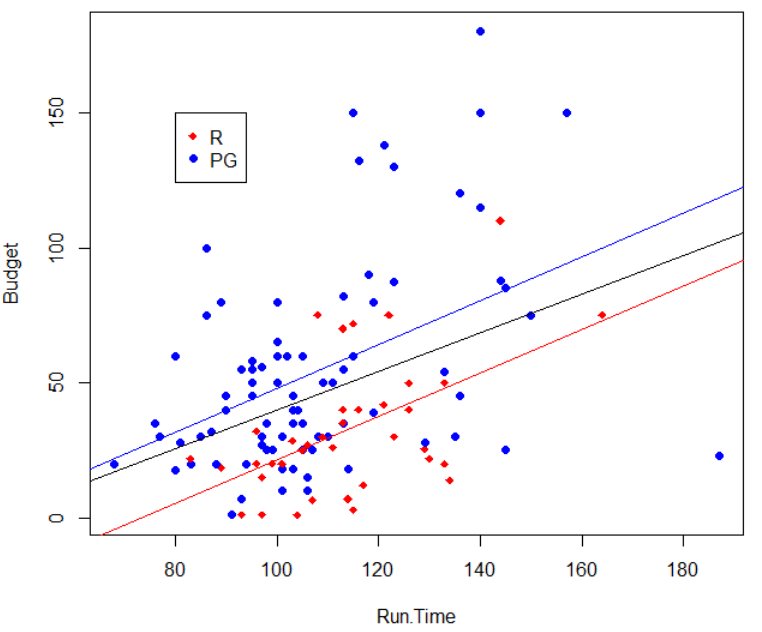
Coefficients:

(Intercept) Run.Time RatingR

-32.8774 0.8029 -25.8851

> abline(-32.877,0.8092,col="blue")

> abline(-58.765,0.8029,col="red")



1. The percentage of a person’s weight that is made up of body fat is often used as an indicator of health and fitness. However, accurate methods of measuring percent body fat are difficult to implement. One method involves immersing the body in water to estimate its density and then applying a formula to estimate percent body fat. An alternative is to develop a model for percent body fat that is based on body characteristics such as height and weight that are easy to measure. The dataset *BodyFat* in *Lock5withR* contains such measurements for a sample of 100 men.
2. Import the data and Access the variable names included in the dataset.

library(BodyFat)

> names(BodyFat)

[1] "Bodyfat" "Age" "Weight" "Height" "Neck" "Chest" "Abdomen"

[8] "Ankle" "Biceps" "Wrist"

1. Fit a model to predict Bodyfat using Height and Weight. Comment on whether either of the predictors appears to be important in the model.

> model = lm(BodyFat~Height+Weight)

> summary(model)

Call:

lm(formula = Bodyfat ~ Height + Weight)

Residuals:

Min 1Q Median 3Q Max

-12.7697 -3.9527 -0.5364 4.0473 13.2829

Coefficients:

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

(Intercept) 71.48247 16.20086 4.412 2.65e-05 \*\*\*

Height -1.33568 0.25891 -5.159 1.32e-06 \*\*\*

Weight 0.23156 0.02382 9.721 5.36e-16 \*\*\*

---

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

Residual standard error: 5.754 on 97 degrees of freedom

Multiple R-squared: 0.494, Adjusted R-squared: 0.4836

F-statistic: 47.35 on 2 and 97 DF, p-value: 4.48e-15

The Pr(>|t|) of Weight is 5.36e-16, which is less than Height, so Weight is more important.

1. Add Abdomen as a third predictor to the model (b) and repeat the assessment of the effectiveness of each predictor.

> model2 = update(model,.~.+Abdomen)

> summary(model2)

Call:

lm(formula = Bodyfat ~ Height + Weight + Abdomen)

Residuals:

Min 1Q Median 3Q Max

-9.5219 -2.9969 0.0378 2.8933 9.2859

Coefficients:

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

(Intercept) -56.1329 18.1372 -3.095 0.002580 \*\*

Height 0.1018 0.2444 0.417 0.677750

Weight -0.1756 0.0472 -3.720 0.000335 \*\*\*

Abdomen 1.0747 0.1158 9.279 5.27e-15 \*\*\*

---

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

Residual standard error: 4.199 on 96 degrees of freedom

Multiple R-squared: 0.7332, Adjusted R-squared: 0.7249

F-statistic: 87.96 on 3 and 96 DF, p-value: < 2.2e-16

The Pr(>|t|) of Abdomen is 5.27e-15, which is less than Height and Weight, so Abdomen is more important.

1. Interpret the coefficient of Abdomen you get in the model part (c)

Y\_hat = 71.48 + 1.0747\*Abdomen + 0.1018\*Height – 0.1756\*Weight

If given two data, their Height and Weight are the same, if the Abdomen of one data is one more than another, then the body fat will increase 1.0747

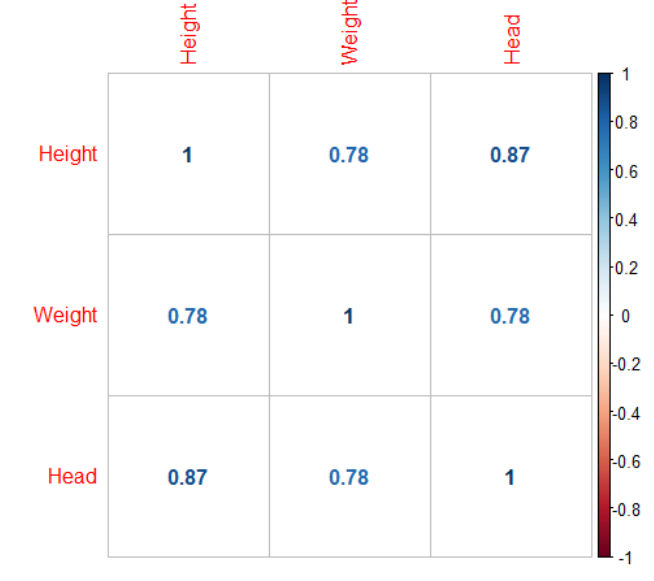
1. A pediatrician wants to determine the relation that may exist between a child‘s head circumference (in centimeters), height (in inches), and weight (in ounces). She randomly selects 14 three-year-old children from her practice and obtains the following data:

|  |  |  |
| --- | --- | --- |
| Height | Weight | Head Circumference |
| 30 | 339 | 47 |
| 26.25 | 267 | 42 |
| 25 | 289 | 43 |
| 27 | 332 | 44.5 |
| 27.5 | 272 | 44 |
| 24.5 | 214 | 40.5 |
| 27.75 | 311 | 44 |
| 25 | 259 | 41.5 |
| 28 | 298 | 46 |
| 27.25 | 288 | 44 |
| 26 | 277 | 44 |
| 27.25 | 292 | 44.5 |
| 27 | 302 | 42.5 |
| 28.25 | 336 | 44.5 |

1. Construct a correlation matrix. Is there any reason to be concerned with multicollinearity?

> correlation = cor(data)

> corrplot(corelation,method="number")



Through the matrix we can see that the correlation between every two variables are very high, so they are multicollinearity.

1. Find the least-squares regression equation with the response variable, head circumference.

> attach(data)

> model = lm(Head~Weight+Height)

> model

Call:

lm(formula = Head ~ Weight + Height)

Coefficients:

(Intercept) Weight Height

18.82425 0.01281 0.78634

Head\_hat = 18.82425 + 0.01281\*Weight + 0.78634\*Height

1. Construct 95% confidence and prediction intervals for the head circumference of a child whose height is 27.5 inches and whose weight is 285 ounces.

> predict(model,data.frame(Height=27.5,Weight=285),interval = 'pred',level = 0.95)

fit lwr upr

1 44.09898 42.05886 46.1391

> predict(model,data.frame(Height=27.5,Weight=285),interval = 'confidence',level = 0.95)

fit lwr upr

1 44.09898 43.39962 44.79834

1. Perform the residual analysis of the model.

Plot(model)

