Question 4.9

**> data<-read.table('abrasion.txt', header=T)**

#The matrix X'X

**> A = matrix(c(30,2108,5414,2108,152422,376562,5414,376562, 1015780), nrow = 3, ncol = 3)**

#The inverse of X'X, (X'X)^-1

**> A.inv <- solve(A)**

**> A.inv**

[,1] [,2] [,3]

[1,] 2.863941746 -2.254472e-02 -6.906905e-03

[2,] -0.022544720 2.554376e-04 2.546717e-05

[3,] -0.006906905 2.546717e-05 2.835655e-05

#The matrix X'y

**> B = matrix(c(5263,346867,921939), nrow = 3, ncol = 1)**

#Least square of estimates beta hat, (X'X)^-1X'y

**> A.inv%\*%B**

[,1]

[1,] 885.161109

[2,] -6.570830

[3,] -1.374312

**> summary(model)**

Call:

lm(formula = data$Loss ~ data$x1 + data$x2)

Residuals:

Min 1Q Median 3Q Max

-79.385 -14.608 3.816 19.755 65.981

Coefficients:

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

(Intercept) 885.1611 61.7516 14.334 3.84e-14 \*\*\*

data$x1 -6.5708 0.5832 -11.267 1.03e-11 \*\*\*

data$x2 -1.3743 0.1943 -7.073 1.32e-07 \*\*\*

---

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

Residual standard error: 36.49 on 27 degrees of freedom

Multiple R-squared: 0.8402, Adjusted R-squared: 0.8284

F-statistic: 71 on 2 and 27 DF, p-value: 1.767e-11

**> summary(model)$r.squared**

[1] 0.8402314

(a)

The (X’X)^-1 has computed in the code above.

The least square estimates for β0 is 885.1611. And its standard error is 61.7516.

The least square estimates for β1 is -6.5708. And its standard error is 0.5832.

The least square estimates for β2 is -1.3743. And its standard error is 0.1943.

(b)

The regression model is y = 885.1611 – 6.5708x1 – 1.3743x2.

The p-value of β0 is 3.84e-14 under the null hypothesis H0: β0 = 0 when β1 and β2 are in the model. The p-value is very small, so we have strong evidence to against H0. β0 is important to this model.

The p-value of β1 is 1.03e-11 under the null hypothesis H0: β1 = 0 when β0 and β2 are in the model. The p-value is very small, so we have strong evidence to against H0. β1 is important to this model.

The p-value of β2 is 1.32e-07 under the null hypothesis H0: β2 = 0 when β0 and β1 are in the model. The p-value is very small, so we have strong evidence to against H0. β2 is important to this model.

(c)

According to the R code above, the coefficient of variation R^2 is 0.8402. It means that 84.02% of variations are explained by the model. The value of R^2 is high, so that the independent variables help to determine the dependent variables sufficiently.

Question 4.11

**> data<-read.table('hospital.txt', header=T)**

**> model<-lm(data$Y ~ data$X1 + data$X2 + data$X3 + data$X4)**

**> summary(model)**

Call:

lm(formula = data$Y ~ data$X1 + data$X2 + data$X3 + data$X4)

Residuals:

Min 1Q Median 3Q Max

-61.022 -26.919 -7.331 27.707 76.854

Coefficients:

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

(Intercept) 2.953e+02 4.018e+01 7.350 2.4e-06 \*\*\*

data$X1 -4.808e+02 1.504e+02 -3.197 0.00600 \*\*

data$X2 -8.294e+02 1.965e+02 -4.222 0.00074 \*\*\*

data$X3 7.936e-03 3.554e-03 2.233 0.04121 \*

data$X4 2.360e+00 7.616e-01 3.099 0.00733 \*\*

---

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

Residual standard error: 46.77 on 15 degrees of freedom

Multiple R-squared: 0.8826, Adjusted R-squared: 0.8512

F-statistic: 28.18 on 4 and 15 DF, p-value: 8.046e-07

(a)

The R code for the regression model is given above.

(b)

The t value for β1 is -3.197 and the p value is 0.006. We can see the p-value is small, which means we have strong evidence against the null hypothesis H0: β1=0 when other variables in the model. Hence, the percentage of beds in for-profit hospitals is significant to the model.

(c)

I will test whether there exists the multicollinearity between the variables. Because x2 and x3 are both interested in finding number of people and the number of people may be influenced by the population size of a state.

(d)

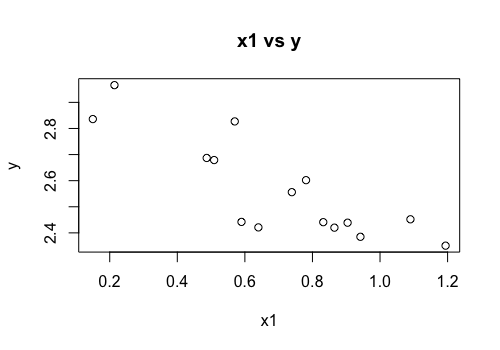
Rather than selecting 10 states at random, I may select exactly 10 states and choosing an same proportion of people in each state. For example, we may choose 10% of people in each state randomly. In this case, we can promise that each state has enough samples and the result will be more representative.

Question 4.15



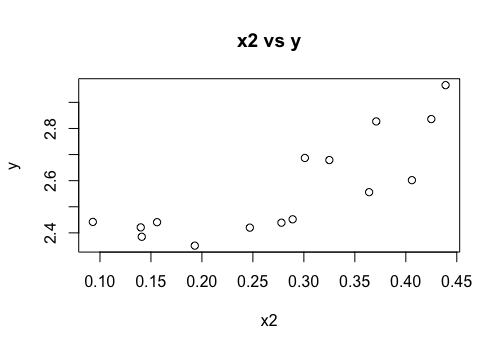
**data<-read.table('silkw.txt', header=T)**

**plot(data$X1, data$Y, xlab = 'x1', ylab = 'y', main = 'x1 vs y')**



**Comments:** The data plot x1 vs y shows strong negative linear relationship between x1 and y.

**plot(data$X2, data$Y, xlab = 'x2', ylab = 'y', main = 'x2 vs y')**

****

**Comments:** The data plot x2 vs y shows fairly strong positive linear relationship between x2 and y.

(b)  **model <- lm(data$Y ~ data$X1+data$X2)**

**summary(model)**

Call:

lm(formula = data$Y ~ data$X1 + data$X2)

Residuals:

Min 1Q Median 3Q Max

-0.071886 -0.049513 -0.001636 0.030004 0.129092

Coefficients:

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

(Intercept) 2.58810 0.08349 31.001 7.98e-13 \*\*\*

data$X1 -0.37802 0.06630 -5.701 9.89e-05 \*\*\*

data$X2 0.87677 0.17234 5.087 0.000267 \*\*\*

---

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

Residual standard error: 0.06263 on 12 degrees of freedom

Multiple R-squared: 0.9079, Adjusted R-squared: 0.8926

F-statistic: 59.16 on 2 and 12 DF, p-value: 6.093e-07

**Conclusion**:

The lease square estimates for β0 is 2.58810.

The lease square estimates for β1 is -0.37802.

The lease square estimates for β2 is 0.87677.

Hence, the fitted equation for this model is y = 2.58810 – 0.37802x1 + 0.87677x2.

(c)

**> anova(model)**

Analysis of Variance Table

Response: data$Y

Df Sum Sq Mean Sq F value Pr(>F)

data$X1 1 0.36266 0.36266 92.448 5.464e-07 \*\*\*

data$X2 1 0.10153 0.10153 25.881 0.0002674 \*\*\*

Residuals 12 0.04707 0.00392

---

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

**> summary(model)$adj.r.squared**

[1] 0.8925799

The p-value of β1 is 9.89e-05 under the null hypothesis H0: β1 = 0 when β0 and β2 are in the model. The p-value is very small, so we have strong evidence to against H0. β1 is important to this model.

The p-value of β2 is 0.000267 under the null hypothesis H0: β2 = 0 when β0 and β1 are in the model. The p-value is very small, so we have strong evidence to against H0. β2 is important to this model.

(d)

The adjusted R^2 for model consists of x1 and x2 is 0.893, it means 89.3% variations is explained by the model. The adjusted R^2 is very high in this model. Both independent variables help to determine dependent variables nicely.

To find which is the better predictor of log (survival time), we can construct the model with x1 and x2 separately and compare their adjusted R^2.

**Model with x1 only:**

**> modelonlyx1 <- lm(data$Y ~ data$X1)**

**> summary(modelonlyx1)**

Call:

lm(formula = data$Y ~ data$X1)

Residuals:

Min 1Q Median 3Q Max

-0.185607 -0.051481 0.002821 0.067218 0.188409

Coefficients:

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

(Intercept) 2.95166 0.07367 40.065 5.25e-15 \*\*\*

data$X1 -0.54924 0.09751 -5.633 8.16e-05 \*\*\*

---

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

Residual standard error: 0.1069 on 13 degrees of freedom

Multiple R-squared: 0.7093, Adjusted R-squared: 0.687

F-statistic: 31.73 on 1 and 13 DF, p-value: 8.164e-05

**> summary(modelonlyx1)$adj.r.squared**

[1] 0.6869866

The adjusted R^2 for model of x1 only is 0.687, it means 68.7% of variations are explained by this model. The adjusted R^2 for model with x1 only is fairly high.

**Model with x2 only:**

**> modelonlyx2 <- lm(data$Y ~ data$X2)**

**> summary(modelonlyx2)**

Call:

lm(formula = data$Y ~ data$X2)

Residuals:

Min 1Q Median 3Q Max

-0.14119 -0.11629 0.04170 0.07746 0.17741

Coefficients:

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

(Intercept) 2.1847 0.0820 26.644 9.92e-13 \*\*\*

data$X2 1.3756 0.2747 5.007 0.00024 \*\*\*

---

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

Residual standard error: 0.1159 on 13 degrees of freedom

Multiple R-squared: 0.6585, Adjusted R-squared: 0.6322

F-statistic: 25.07 on 1 and 13 DF, p-value: 0.00024

**> summary(modelonlyx2)$adj.r.squared**

[1] 0.6322409

The adjusted R^2 for model of X2 only is 0.632, it means 63.2% of variations are explained by the model.

**Conclusion**:

We find that the adjusted R^2 for model with x2 only is less than the adjusted R^2 for model with x1 only. The variations are explained more adequately with x1. Therefore, we consider x1 be a better predictor for log (survival time).

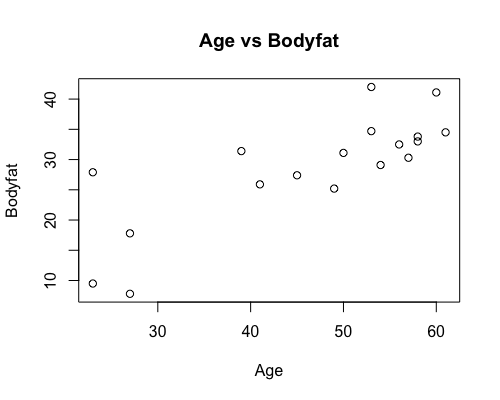
(e)

Since the adjusted R^2 for model consists of both x1 and x2 is much higher than the model with one of the independent variables, so I prefer the model with both predictors.

Question 5.3

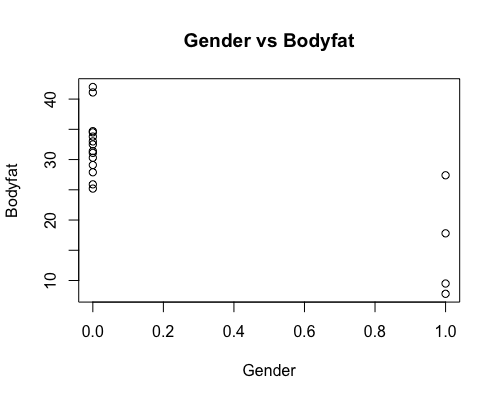
**> bodyfat<-read.table('bodyfat.txt', header=T)**

**> plot(bodyfat$Age, bodyfat$X.fat, xlab = 'Age', ylab = 'Bodyfat', main = 'Age vs Bodyfat')**

****

**Comment:** We can see there is a strong positive linear relationship between age and body fat. Although there is one obvious outlier.

**> plot(bodyfat$Gend, bodyfat$X.fat, xlab = 'Gender', ylab = 'Bodyfat', main = 'Gender vs Bodyfat')**

****

**Comment:** We can see there is strong negative linear relationship between gender and body fat when 0.0 represents female and 1.0 represents male.

**> model <- lm(bodyfat$X.fat~ bodyfat$Age + bodyfat$Gend)**

**> summary(model)**

Call:

lm(formula = bodyfat$X.fat ~ bodyfat$Age + bodyfat$Gend)

Residuals:

Min 1Q Median 3Q Max

-6.638 -3.455 -1.103 3.297 8.952

Coefficients:

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

(Intercept) 15.0708 6.2243 2.421 0.0286 \*

bodyfat$Age 0.3392 0.1196 2.835 0.0125 \*

bodyfat$Gend -9.7914 3.6966 -2.649 0.0182 \*

---

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

Residual standard error: 4.905 on 15 degrees of freedom

Multiple R-squared: 0.7461, Adjusted R-squared: 0.7123

F-statistic: 22.04 on 2 and 15 DF, p-value: 3.424e-05

**> summary(model)$r.squared**

[1] 0.7461359

**Comment:** The regression equation is y = 15.0708 + 0.3392x1 – 0.7014x2. The estimate of age is 0.3392 and the estimate of gender is -9.7914. And F-statistics is 22.04 with p value very small. It suggests that at least one of the variables is significant in the model. The multiple R-square in this model is 0.7461. It means 74.61% of variation is explained by the model.

Let’s include an interaction term for age and gender.

**> model2 <- lm(bodyfat$X.fat~ bodyfat$Age + bodyfat$Gend + bodyfat$Age \* bodyfat$Gend)**

**> summary(model2)**

Call:

lm(formula = bodyfat$X.fat ~ bodyfat$Age + bodyfat$Gend + bodyfat$Age \*

bodyfat$Gend)

Residuals:

Min 1Q Median 3Q Max

-6.6756 -2.8862 -0.2464 1.9100 9.1641

Coefficients:

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

(Intercept) 20.1116 6.2395 3.223 0.00613 \*\*

bodyfat$Age 0.2401 0.1204 1.994 0.06600 .

bodyfat$Gend -29.2692 10.4098 -2.812 0.01386 \*

bodyfat$Age:bodyfat$Gend 0.5725 0.2893 1.978 0.06790 .

---

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

Residual standard error: 4.488 on 14 degrees of freedom

Multiple R-squared: 0.8016, Adjusted R-squared: 0.7591

F-statistic: 18.86 on 3 and 14 DF, p-value: 3.455e-05

**> summary(model2)$r.squared**

[1] 0.8016062

**Comment:**

The regression equation now becomes y = 20.1166 + 0.2401x1 – 29.2692x2 + 0.5725x3. We can see the multiple R-squared now becomes 0.8016. It means 80.16% of variation is explained by this model. The multiple R-squared increases 4% compare to the model without an interaction term. Therefore, we can conclude that the model has improved.