STA135 Final Project

Spring 2020 | Professor Xiaodong Li

Haolin Li | 913838107 | June 8, 2020

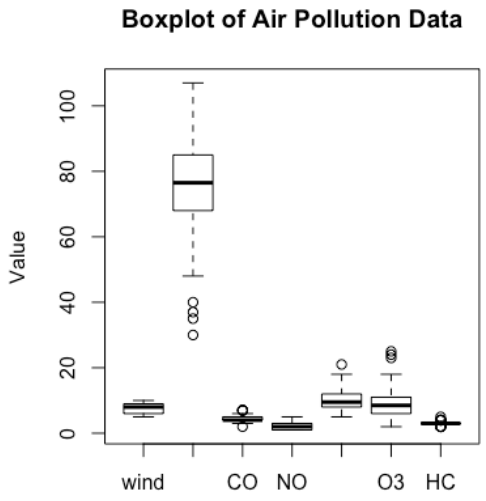
Dataset 1: Air Pollution Dataset ("T1-5.DAT)

Introduction

The goal of this dataset is to conduct the principal component analysis on the air pollution data. We want to see which features of the air pollution impact the most among all. I will generate a scree plot as well as showing table of loadings to support my results.

Summary

This dataset is about the air pollution variables in Los Angeles. There are 42 observations/measurements of the variables, and there are 7 variables representing different type of air pollutions, they are wind, solar radiation, CO, NO, NO2, O3 and HC. Below is the boxplot of all air pollution variables in the data. From the boxplot, we can clearly observe that solar radiation has the highest value of air pollution among 7 pollutions.

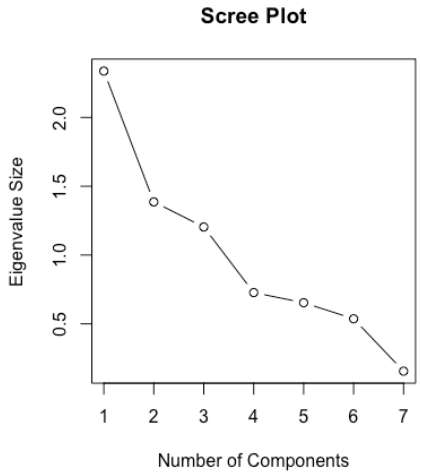


Analysis

From the table below, we can see the proportion of total variance for each principal component. We have obtained 7 principal components, each of these explains a percentage of the total variation in the dataset, where pc1 33.38% of the total variance, and pc2 explains 19.8% of the total variance and so on. We can see that the first 3 principal components cover 70.38% of the total variance, which means more than two-third of the information in the dataset can be explained by the first 3 principal components.

```
> summary(air.pc, loadings = T)
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4     Comp.5     Comp.6
Standard deviation     1.5286539 1.1772853 1.0972994 0.8526937 0.80837896 0.73259047
Proportion of Variance 0.3338261 0.1980001 0.1720094 0.1038695 0.09335379 0.07666983
Cumulative Proportion  0.3338261 0.5318262 0.7038356 0.8077051 0.90105889 0.97772872
                           Comp.7
Standard deviation     0.39484041
Proportion of Variance 0.02227128
Cumulative Proportion  1.00000000
```

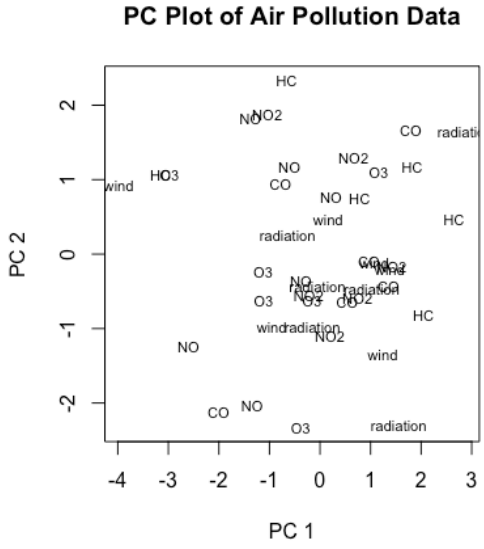
We can also know which principal component explains the most information by generating a scree plot. From the scree plot shown below, we can see that principal components are created in order of the amount of variation they cover: PC1 captures the most variation, then is PC2 and so on. Each principal component contributes at least some information.



Based on loadings, same as shown in the table below, the first principal component contains information of all the air pollutions. The second principal component is mainly determined by solar radiation, O3, NO and HC. While, the third principal component is mainly determined by win, HC and NO.

```
Loadings:
          Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
wind       0.237  0.278  0.643  0.173  0.561  0.224  0.241
radiation -0.206 -0.527  0.224  0.778 -0.156
CO        -0.551        -0.114         0.573  0.110 -0.585
NO        -0.378  0.435 -0.407  0.291         0.450  0.461
NO2       -0.498  0.200  0.197                -0.745  0.338
O3        -0.325 -0.567  0.160 -0.508         0.331  0.417
HC        -0.319  0.308  0.541 -0.143 -0.566  0.266 -0.314
```

From the PC plot shown below, we can see that all the air pollutions are grouping near the center of the plot, such as NO, O3, solar radiation and CO. From this plot, we can get a sense of which air pollution variables contribute more to determine the principal component.



Conclusion

From results of the analysis above, we can conclude that the air pollution dataset can be explained with the first 3 principal components as they together cover more than two-third of the information of the dataset. Then we can also conclude which air pollution contributes more to determine each principal component.

Dataset2: Lizard Data ("T6-7.DAT")

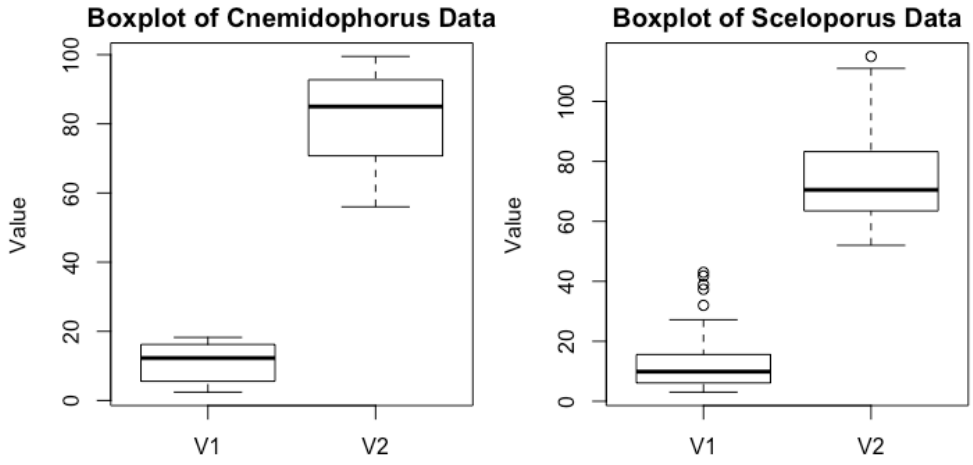
Introduction

The goal of this dataset is to conduct principal component analysis to the lizard data for two genera, Cnemidophorus and Sceloporus. I will derive the Hotelling's T^2 for the two samples, and then plot the confidence region. I will also use the simultaneous confidence intervals to check the significant components.

Summary

The dataset contains lizards from two genera, they are called Cnemidophorus and Sceloporus, denoted as C and S respectively in the dataset. There are three columns in the dataset, the first indicates the mass of the lizards, the second one indicates the snout-vent length of the lizard, and the third one which is only 0 and 1, where o is identified as the Cnemidophorus and 1 is identified as the Sceloporus. There are total of 60 observations in the dataset, where there are 20 Cnemidophorus and 40 Sceloporus.

Below are the boxplots of each lizard with V1 = mass and V2 = snout-vent length. The boxplots below show how well the data is distributed in the dataset. We can see that the means of mass of each lizard are 10.875 and 13.84. And there are outliers in the mass and snout-vent length of Sceloporus.



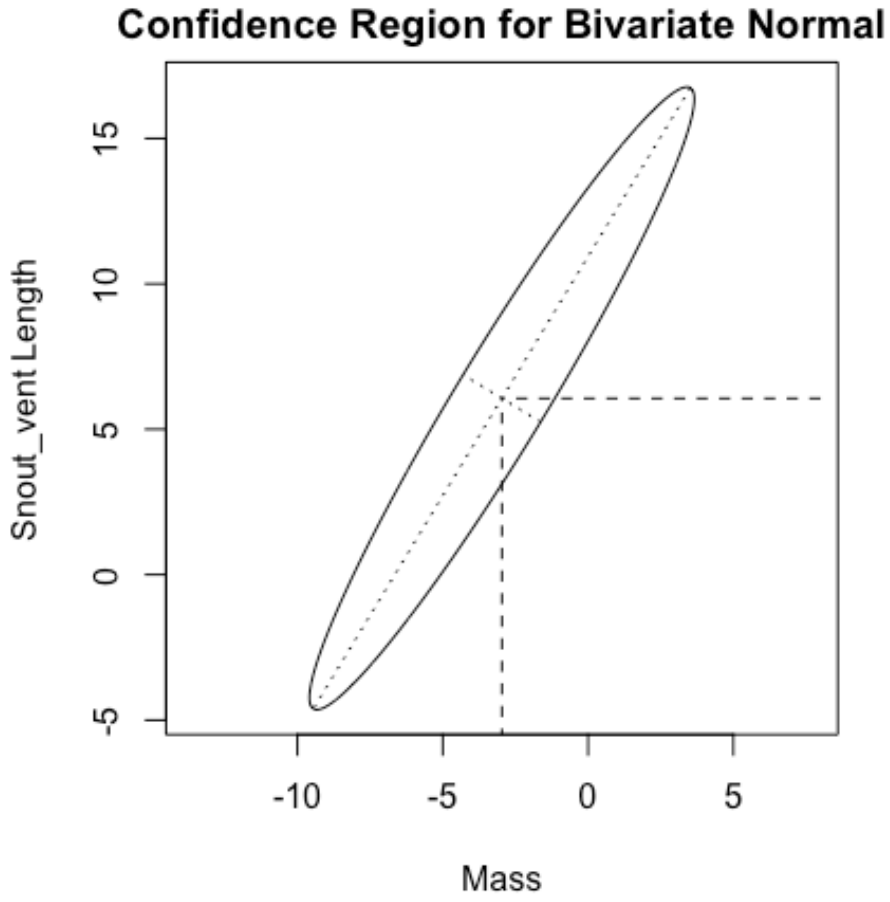
TWO-SAMPLE TEST

Analysis

Hotelling's T^2 for the two samples:

From the Hotelling's T^2, we obtained that T^2 = 85.7604. And the critical value is 6.4285. Since the T^2 is larger than the critical value, we reject H0 at the level of 0.05. We can say that we are 95% confident that the two samples are not the same.

Below is the confidence region for the difference between the two samples. The center of the confidence ellipse is (-2.9652, 6.0625) with axes of directions $\begin{bmatrix}0.5194\\0.8545\end{bmatrix}$ and $\begin{bmatrix}-0.8545\\0.5914\end{bmatrix}$ .The value of c is 2.5355. The half axis length is 12.4878 and 1.5607.



Since we rejected the null hypothesis, we use the simultaneous confidence interval to check the significant components.

The 95% simultaneous confidence intervals for u11 – u21 and u12 – u22 are (-9.5868, 3.6565) and (-4.6394, 16.7644).

The 95% Bonferroni simultaneous confidence intervals are (-8.9748, 3.0444) and (-3.6501, 15.7751).

Conclusion

From comparing the Hotelling's T^2 with the critical value, we reject the null hypothesis that the two samples are the same. We say that the two samples are different. When we use the

simultaneous confidence interval to check the significant componenets, both component-wise simultaneous confidence intervals contain 0, which means that there is significant difference in the two means.

LINEAR DISCRIMINANT ANALYSIS

Analysis

For the linear discriminant analysis, I will conduct a confusion matrix regarding the classification results. I will also calculate the apparent error rate of the classification results.

The confusion matrix of the predicted results is:

	Cnemidophorus = 0	Sceloporus = 1
Cnemidophorus = 0	18	2
Sceloporus = 1	1	39

To interpret the confusion matrix:

18 is the number of items correctly classified as Cnemidophorus, which is 0 in the dataset.

39 is the number of items correctly classified as Sceloporus, which is 1 in the dataset.

2 is the number of items misclassied as Sceloporus when it's actually Cnemidophorus.

1 is the number of items misclassied as Cnemidophorus when it's actually Sceloporus.

The apparent error rate is (2+1)/60 = 0.05 = 5%

Conclusion

From the confusion matrix above, it shows the results of the classification of the two type of lizards. Out of 20 Cnemidophorus lizards, 18 were classified correctly. Out of 40 Sceloporus lizards, 39 were classified correctly. The apparent erro rate is only 5% which is very low. It all reveals that there is a significant different between the two lizards in terms of mass and snout-vent length.

Dataset 3: Real-Estate Data ("T7-1.DAT")

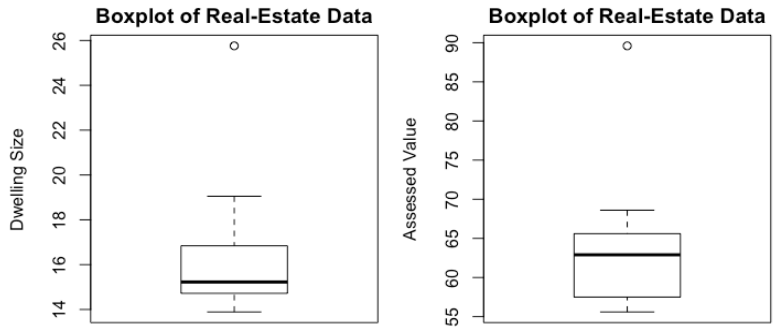
Introduction

The goal of analysis is to fit a regression model to the real-estate data. While fitting the linear model in this dataset, I will also find the features of the model such as the least square estimate, the R-squared statistics and more. At the end, I will find the 95% confidence interval for the mean response and prediction interval.

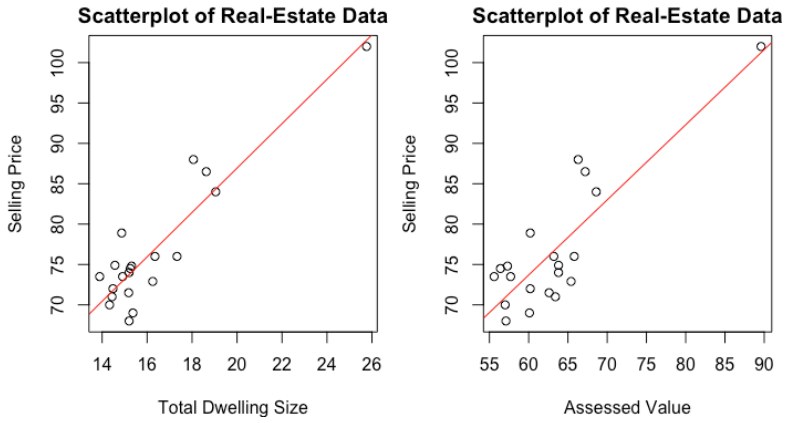
Summary

The dataset contains two explanatory variables and a response variable, the explanatory variables are the dwelling size and the assessed value of the houses, where the response variables is the selling price of the house. There are 20 observations in the dataset.

Below is the summary of the data, including the distribution of the variables and their relationships between the response variable.



The Boxplots above show how well the data is distributed in the dataset. There are outliers in both plots, which for the dwelling size is 25.76 and for the assessed value is 89.6. For the mean of both variables, we have 16.222 and 63.065 for the dwelling size and the assessed value, respectively.



The scatterplots above show the relationship between the dwelling size and the selling price, as well as the assessed value and the selling price. We can say that both the dwelling size and the assessed value have a linear relationship with the selling price.

Analysis

Below are the implementations of the analysis based on the homework. Implementation of the equations can be found in R code.

The least square estimates of the dataset are: $\begin{bmatrix} 30.967 \\ 2.634 \\ 0.045 \end{bmatrix}$.

The R-squared statistic is 0.834.

The $\hat{\sigma}$^2 (sigma_hat_square) is 12.059.

The estimated covariance of $\hat{\beta}$ is $\begin{bmatrix} 62.129 & 3.068 & -1.765 \\ 3.068 & 0.617 & -0.207 \\ -1.765 & -0.207 & 0.081 \end{bmatrix}$.

The 95% confidence interval for β_1 is: (0.977, 4.292).

The 95% confidence interval for β_2 is: (-0.556, 0.647).

Let C = $\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, here testing H_0: $\beta_1 = \beta_2 = 0$ at the level of 0.05. The F-test value 1032.875 and the critical value in F distribution is 86.617. Then we can say that since the F-test value is larger than the critical value, we reject H_0. β_1 and β_2 are not equal to 0.

The confidence interval for the mean response given $\vec{z}_0$=$\begin{bmatrix} 1 \\ 16.222 \\ 63.065 \end{bmatrix}$ which 16.222 and 63.065 is the mean of the variables dwelling size and assessed value. We are conducting the 95% confidence interval, the result is [74.9118, 78.1882].

The prediction interval for Y_0 given $\vec{z}_0$=$\begin{bmatrix} 1 \\ 16.222 \\ 63.065 \end{bmatrix}$ is [69.0427, 84.0573].

Conclusion

From the least square estimates, the fitted equation is $\hat{y} = 30.967 + 2.634z_1 + 0.045z_2$. With R-squared statistic equals to 0.834, we can say that the data shows a strong regression relationship. Despite in the hypothesis testing we say that β_1 and β_2 are not equal to 0, but in the confidence interval of β_2, (-0.556, 0.647), includes 0, then the regression model can be considered to drop the assessed value variable. Overall, with the dwelling size and assessed value of the houses, we can predict the selling price of the houses.

Appendix

```
#PCA
air = read.table("~/Desktop/T1-5.dat")
par(mfrow=c(1,1))
names(air) = c("wind","radia","CO","NO","NO2","O3","HC")
names = c("wind","radia","CO","NO","NO2","O3","HC")
#proportion of variance
air.pc = princomp(air, cor=T)
summary(air.pc, loadings = T)
(state.pc$sdev)^2

#Scree plot
plot(1:(length(air.pc$sdev)),  (air.pc$sdev)^2, type='b',
    main="Scree Plot", xlab="Number of Components", ylab="Eigenvalue Size")

# PC plot
par(pty="s")
plot(air.pc$scores[,1], air.pc$scores[,2],
    xlab="PC 1", ylab="PC 2", type ='n', lwd=2, main = "PC Plot of Air Pollution Data")
# labeling points with air pollution names:
text(air.pc$scores[,1], air.pc$scores[,2], labels=names, cex=0.7, lwd=2)

#About the data box plot
boxplot(air, main = " Boxplot of Air Pollution Data ", ylab = "Value")
#Two Sample Test
lizard = read.table("~/Desktop/T6-7.dat")
C = lizard[1:20,1:2]
```

```
S = lizard[21:60,1:2]

# about the data
boxplot(C, main = " Boxplot of Cnemidophorus Data ", ylab = "Value")
boxplot(S, main = " Boxplot of Sceloporus Data ", ylab = "Value")

# hotelling t^2
n = c(20,40)
p = 2
xmean1 = colMeans(C)
xmean2 = colMeans(S)
d = xmean1 - xmean2
S1 = var(C)
S2 = var(S)
Sp = ((n[1]-1)*S1+(n[2]-1)*S2)/(sum(n)-2)
t2 = t(d)%*%solve(sum(1/n)*Sp)%*%d
t2

alpha<-0.05
cval <- (sum(n)-2)*p/(sum(n)-p-1)*qf(1-alpha,p,sum(n)-p-1)
cval

#since t2 is larger than c.v., we reject H0.

#Confidence interval
es<-eigen(sum(1/n)*Sp)
```

```
e1<-es$vec %*% diag(sqrt(es$val))

r1<-sqrt(cval)

theta<-seq(0,2*pi,len=250)

v1<-cbind(r1*cos(theta), r1*sin(theta))

pts<-t(d-(e1%*%t(v1)))

plot(pts,type="l",main="Confidence Region for Bivariate
Normal",xlab="Sepal.length",ylab="Sepal.width",asp=1)

segments(0,d[2],d[1],d[2],lty=2) # highlight the center

segments(d[1],0,d[1],d[2],lty=2)


th2<-c(0,pi/2,pi,3*pi/2,2*pi)   #adding the axis

v2<-cbind(r1*cos(th2), r1*sin(th2))

pts2<-t(d-(e1%*%t(v2)))

segments(pts2[3,1],pts2[3,2],pts2[1,1],pts2[1,2],lty=3)

segments(pts2[2,1],pts2[2,2],pts2[4,1],pts2[4,2],lty=3)

# center is (-2.9652, 6.0625)

# c = 2.5355

# direction = [0.5194, 0.8545],[-0.8545, 0.5194]

# length 12.4878, 1.5607


#Bonferroni simultaneous confidence intervals

wd.b<- qt(1-alpha/(2*p),n[1]+n[2]-2) *sqrt(diag(Sp)*sum(1/n))

Cis.b<-cbind(d-wd.b,d+wd.b)

cat("95% Bonferroni simultaneous confidence interval","\n")

Cis.b

# simultaneous confidence intervals

wd<-sqrt(cval*diag(Sp)*sum(1/n))
```

```
Cis<-cbind(d-wd,d+wd)

cat("95% simultaneous confidence interval","\n")

Cis


#LDA

# compute pooled estimate for the covariance matrix:

w<-solve(Sp)%*%(xmean1-xmean2)

w0<-(xmean1+xmean2)%*%w/2


library(MASS)

names(lizard) = c("mass","svl","group")

lda.obj = lda(group~., data = lizard)

plda = predict(object = lda.obj, newdata = lizard)

#confusion matrix

table(lizard$group,plda$class)

#Multiple linear regression

estate = read.table("~/Desktop/T7-1.DAT")

# about the data

par(mfrow=c(1,2))

boxplot(estate$V1, main = " Boxplot of Real-Estate Data ", ylab = "Dwelling Size")

boxplot(estate$V2, main = " Boxplot of Real-Estate Data ", ylab = "Assessed Value")


par(mfrow=c(1,2))

plot(estate$V1, estate$V3, main = "Scatterplot of Real-Estate Data", xlab = "Total Dwelling
Size", ylab = "Selling Price")

abline(lm(estate$V3~estate$V1), col="red")
```

```
plot(estate$V2, estate$V3, main = "Scatterplot of Real-Estate Data", xlab = "Assessed Value",
ylab = "Selling Price")

abline(lm(estate$V3~estate$V2), col="red")


n = length(estate[,1])

Y = estate[,3]

Z = cbind(rep(1,n),as.matrix(estate[,1:2]))

r <- dim(Z)[2]-1


# least square estimates

beta_hat = solve(t(Z)%*%Z) %*% t(Z) %*% Y

beta_hat


# R^2 statistic

R_square <- 1 - sum((Y - Z%*%beta_hat)^2)/sum((Y-mean(Y))^2)

R_square


# sigma_hat_square

sigma_hat_square <- sum((Y - Z%*%beta_hat)^2)/(n-r-1)

sigma_hat_square


# estimated covariance of hat{beta}

sigma_hat_square * solve(t(Z)%*%Z)


# F-test

# H_0: beta_1 = beta_2 = 0
```

```
C <- matrix(c(0,0,1,0,0,1),2,3)

alpha = 0.05

df_1 <- qr(C)$rank # df_1: rank of matrix C

q = 0

r - q


Omega_22 = C%*%solve(t(Z)%*%Z)%*%t(C)

f_stat <- (t(C%*%beta_hat)%*%solve(Omega_22)%*%(C%*%beta_hat))

f_stat


cval_f <- qf(1-alpha, 2, n-r-1)

cval_f*(r - q)*sigma_hat_square


# confidence interval for z_0^T beta (correct)


z_0 <- c(1, mean(estate$V1), mean(estate$V2))


cat('[',

   z_0%*%beta_hat - sqrt(sigma_hat_square)*sqrt(t(z_0)%*%solve(t(Z)%*%Z)%*%z_0)*qt(1-
alpha/2, n-r-1),

   ',',

   z_0%*%beta_hat + sqrt(sigma_hat_square)*sqrt(t(z_0)%*%solve(t(Z)%*%Z)%*%z_0)*qt(1-
alpha/2, n-r-1),

   ']')


# prediction interval for Y_0 = z_0^T beta + epsilon_0
```

```
cat('[',
   z_0%*%beta_hat -
sqrt(sigma_hat_square)*sqrt(1+t(z_0)%*%solve(t(Z)%*%Z)%*%z_0)*qt(1-alpha/2, n-r-1),
   ',',
   z_0%*%beta_hat +
sqrt(sigma_hat_square)*sqrt(1+t(z_0)%*%solve(t(Z)%*%Z)%*%z_0)*qt(1-alpha/2, n-r-1),
   ']')


# confidence interval for beta_j
j = 1 #change of value for beta1 and beta2
alpha <- 0.05
cval_t <- qt(1-alpha/2, n-r-1)
cval_t
value1 = beta_hat[j+1] - sqrt(sigma_hat_square)*sqrt(solve(t(Z)%*%Z)[j+1,j+1])*cval_t
value2 = beta_hat[j+1] + sqrt(sigma_hat_square)*sqrt(solve(t(Z)%*%Z)[j+1,j+1])*cval_t
c(value1,value2)
```