Simple linear Regression and Bayesian Linear Regression

Kavinda Uthsuka

3/29/2022

Importing the dataset

library(BAS)

## Warning: package 'BAS' was built under R version 4.0.5

data(bodyfat)  
summary(bodyfat)

## Density Bodyfat Age Weight   
## Min. :0.995 Min. : 0.00 Min. :22.00 Min. :118.5   
## 1st Qu.:1.041 1st Qu.:12.47 1st Qu.:35.75 1st Qu.:159.0   
## Median :1.055 Median :19.20 Median :43.00 Median :176.5   
## Mean :1.056 Mean :19.15 Mean :44.88 Mean :178.9   
## 3rd Qu.:1.070 3rd Qu.:25.30 3rd Qu.:54.00 3rd Qu.:197.0   
## Max. :1.109 Max. :47.50 Max. :81.00 Max. :363.1   
## Height Neck Chest Abdomen   
## Min. :29.50 Min. :31.10 Min. : 79.30 Min. : 69.40   
## 1st Qu.:68.25 1st Qu.:36.40 1st Qu.: 94.35 1st Qu.: 84.58   
## Median :70.00 Median :38.00 Median : 99.65 Median : 90.95   
## Mean :70.15 Mean :37.99 Mean :100.82 Mean : 92.56   
## 3rd Qu.:72.25 3rd Qu.:39.42 3rd Qu.:105.38 3rd Qu.: 99.33   
## Max. :77.75 Max. :51.20 Max. :136.20 Max. :148.10   
## Hip Thigh Knee Ankle Biceps   
## Min. : 85.0 Min. :47.20 Min. :33.00 Min. :19.1 Min. :24.80   
## 1st Qu.: 95.5 1st Qu.:56.00 1st Qu.:36.98 1st Qu.:22.0 1st Qu.:30.20   
## Median : 99.3 Median :59.00 Median :38.50 Median :22.8 Median :32.05   
## Mean : 99.9 Mean :59.41 Mean :38.59 Mean :23.1 Mean :32.27   
## 3rd Qu.:103.5 3rd Qu.:62.35 3rd Qu.:39.92 3rd Qu.:24.0 3rd Qu.:34.33   
## Max. :147.7 Max. :87.30 Max. :49.10 Max. :33.9 Max. :45.00   
## Forearm Wrist   
## Min. :21.00 Min. :15.80   
## 1st Qu.:27.30 1st Qu.:17.60   
## Median :28.70 Median :18.30   
## Mean :28.66 Mean :18.23   
## 3rd Qu.:30.00 3rd Qu.:18.80   
## Max. :34.90 Max. :21.40

We will construct a Bayesian model of simple linear regression, which uses Abdomen to predict response variable bodyfat.

Fit frequentist OLS linear regression

bodyfat.lm = lm(Bodyfat ~ Abdomen,data = bodyfat)  
summary(bodyfat.lm)

##   
## Call:  
## lm(formula = Bodyfat ~ Abdomen, data = bodyfat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -19.0160 -3.7557 0.0554 3.4215 12.9007   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -39.28018 2.66034 -14.77 <2e-16 \*\*\*  
## Abdomen 0.63130 0.02855 22.11 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.877 on 250 degrees of freedom  
## Multiple R-squared: 0.6617, Adjusted R-squared: 0.6603   
## F-statistic: 488.9 on 1 and 250 DF, p-value: < 2.2e-16

Regression line: Bodyfat = -39.28018 + 0.63130\*Abdomen

Bodyfat will increase 0.6313 units when every additional increment of 1cm of Abdomen.

Extract coefficient

beta = coef(bodyfat.lm)  
beta

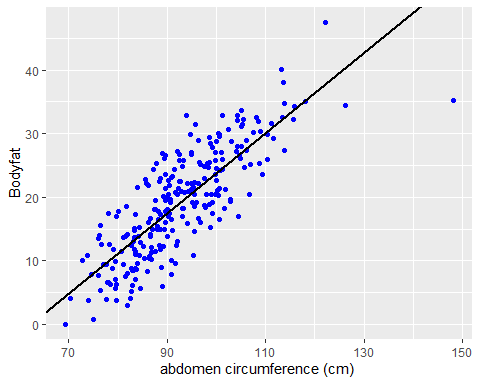
## (Intercept) Abdomen   
## -39.2801847 0.6313044

Visualize regression line on the scatter plot

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.0.5

ggplot(data = bodyfat,aes(x=Abdomen,y=Bodyfat))+  
 geom\_point(color = 'blue')+  
 geom\_abline(intercept = beta[1],slope=beta[2],size=1)+  
 xlab("abdomen circumference (cm) ")



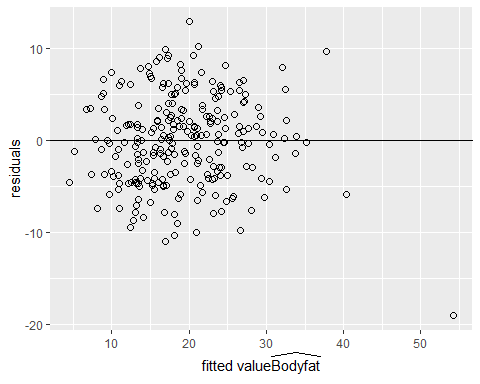
Calculate the Mean Squared Error (MSE)

resid = residuals(bodyfat.lm)  
n = length(resid)  
  
MSE = 1/(n-2) \* sum((resid ^2))  
MSE

## [1] 23.78985

We apply the scatterplot of residuals versus fittedvalues, which provides an additonal isual check of the model adequacy.

#Combine resiaudals and fitted values into a data frame  
result = data.frame(fitted\_values = fitted.values(bodyfat.lm), residuals = residuals(bodyfat.lm))  
  
#load library and plot residuals versus fitted values  
ggplot(data = result,aes(x= fitted\_values,y=residuals))+  
 geom\_point(pch = 1,size = 2) +   
 geom\_abline(intercept = 0,slope = 0)+  
 xlab(expression(paste("fitted value",widehat(Bodyfat))))+  
 ylab("residuals")

 There is one outlier point. Therefore we have to deal with it.

Find the observation with the largest fitted value.

which.max(as.vector(fitted.values(bodyfat.lm)))

## [1] 39

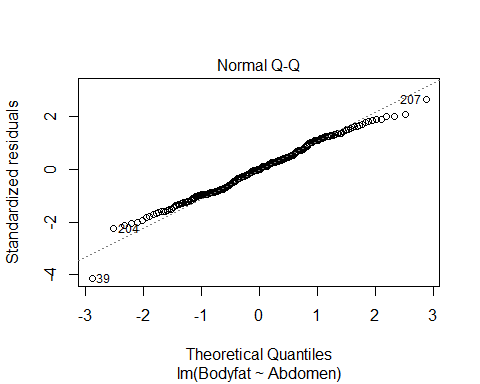
Shows this observation has the largest Abdomen

which.max(bodyfat$Abdomen)

## [1] 39

Let’s plot normal probability plot of the residuals for check the assumption of normally distributed errors

plot(bodyfat.lm,which = 2)



Credible intervals for slope beta and y intercept alpha

output = summary(bodyfat.lm)$coef[,1:2]  
output

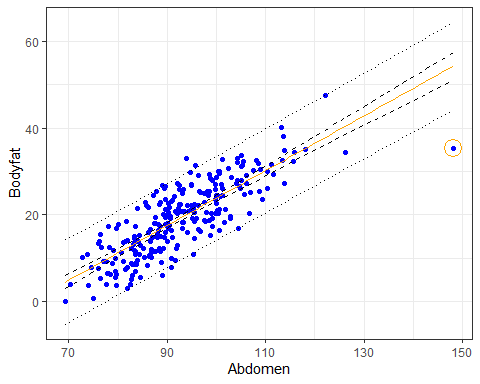
## Estimate Std. Error  
## (Intercept) -39.2801847 2.66033696  
## Abdomen 0.6313044 0.02855067

out = cbind(output,confint(bodyfat.lm))  
colnames(out) = c("posterior mean","posterior std","2.5","97.5")  
round(out,2)

## posterior mean posterior std 2.5 97.5  
## (Intercept) -39.28 2.66 -44.52 -34.04  
## Abdomen 0.63 0.03 0.58 0.69

Let’s plot the prediction intervals

#construct the current prediction  
alpha = bodyfat.lm$coefficients[1]  
beta = bodyfat.lm$coefficients[2]  
new\_x = seq(min(bodyfat$Abdomen),max(bodyfat$Abdomen), length.out = 100)  
  
y\_hat = alpha + beta\*new\_x  
#Get lower and upper bounds for mean  
ymean = data.frame(predict(bodyfat.lm, newdata = data.frame(Abdomen = new\_x), interval = "confidence",level = 0.95))  
  
#Get lower and upper bounds for prediction  
ypred = data.frame(predict(bodyfat.lm, newdata = data.frame(Abdomen = new\_x), interval = "prediction",level = 0.95))  
  
output = data.frame(x = new\_x,y\_hat = y\_hat,ymean\_lwr = ymean$lwr , ymean\_upr = ymean$upr, ypred\_lwr = ypred$lwr , ypred\_upr = ypred$upr)  
  
#Extract potential outlier data point  
outlier = data.frame(x = bodyfat$Abdomen[39],y=bodyfat$Bodyfat[39])  
  
#scatterplot of original  
plot1 = ggplot(data = bodyfat, aes(x=Abdomen,y=Bodyfat)) + geom\_point(color = "blue")  
  
# Add bounds of mean and prediction  
plot2 = plot1 + geom\_line(data=output,aes(x=new\_x,y=y\_hat,color="first"),lty = 1)+  
 geom\_line(data=output,aes(x=new\_x,y=ymean\_lwr,lty = "second")) +  
 geom\_line(data=output,aes(x=new\_x,y=ymean\_upr,lty = "second")) +  
 geom\_line(data=output,aes(x=new\_x,y=ypred\_upr,lty = "third")) +  
 geom\_line(data=output,aes(x=new\_x,y=ypred\_lwr,lty = "third")) +   
 scale\_colour\_manual(values = c("orange"),labels = "Posterior mean",name = "") +  
 scale\_linetype\_manual(values = c(2,3),labels = c("95% CI for mean","95% CI for predictions"),name = "")+  
 theme\_bw()+  
 theme(legend.position = c(1,0),legend.justification = c(1,5,0))  
#Identify potential outlier  
plot2 + geom\_point(data = outlier , aes(x=x,y=y),color="orange",pch = 1,cex=6)



pred.39 = predict(bodyfat.lm,newdata = bodyfat[39,], interval = "prediction", level = 0.95)  
out = cbind(bodyfat[39,]$Abdomen,pred.39)  
colnames(out) = c("abdomen","prediction","lower","upper")  
out

## abdomen prediction lower upper  
## 39 148.1 54.21599 44.0967 64.33528

The bayesian posterior distribution results of alpha and beta show that under the reference prior, the posterior credible intervals are numerically equivalent to the cionfidence intervals from the classical frequentist OLS analysis.

Since the credible intervals are numerically the same as the confidence intervals, we use the lm function to obtain the OLS estimates and construct the credible intervals of alpha and beta.

output1 = summary(bodyfat.lm)$coef[,1:2]  
output1

## Estimate Std. Error  
## (Intercept) -39.2801847 2.66033696  
## Abdomen 0.6313044 0.02855067

out1 =cbind(output1,confint(bodyfat.lm))  
colnames(out1) = c("posterior mean","posterior std","2.5","97.5")  
round(out1,2)

## posterior mean posterior std 2.5 97.5  
## (Intercept) -39.28 2.66 -44.52 -34.04  
## Abdomen 0.63 0.03 0.58 0.69

We believe that there is 95% chance that body fat will increase by 58% up to 69% for every additional 10 centimeter increase in the waist circumference.