midterm2stat632

Jessica Grover

4/19/2022

# Problem 1

# a. Using BWT as the response variable and only AGE, LWT, RACE, and FTV for predictors, create a scatterplot matrix of the data and comment on the relationship among the variables.

library(ggplot2)  
library(car)

## Loading required package: carData

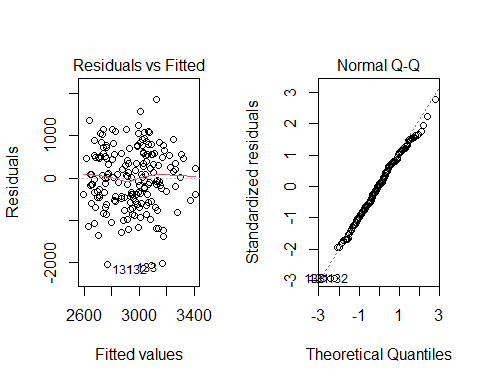
library(MASS)

Reading the file

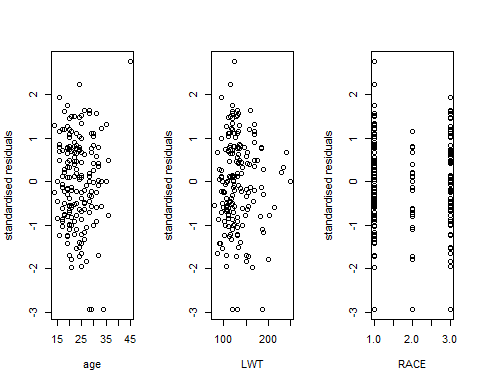
data1<-read.table("lowbwt.dat",header = T)  
head(data1)

## ID LOW AGE LWT RACE SMOKE PTL HT UI FTV BWT  
## 1 85 0 19 182 2 0 0 0 1 0 2523  
## 2 86 0 33 155 3 0 0 0 0 3 2551  
## 3 87 0 20 105 1 1 0 0 0 1 2557  
## 4 88 0 21 108 1 1 0 0 1 2 2594  
## 5 89 0 18 107 1 1 0 0 1 0 2600  
## 6 91 0 21 124 3 0 0 0 0 0 2622

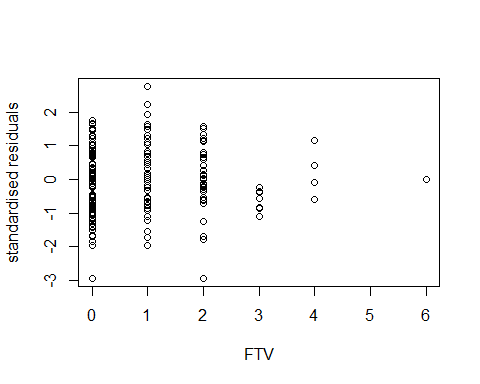
lm1 <- lm(BWT~ AGE+LWT+RACE+FTV,data=data1)  
par(mfrow=c(1,2))  
plot(lm1,1:2)



par(mfrow=c(1,3))  
plot(data1$AGE,rstandard(lm1),xlab="age",ylab="standardised residuals")  
plot(data1$LWT,rstandard(lm1),xlab="LWT",ylab="standardised residuals")  
plot(data1$RACE,rstandard(lm1),xlab="RACE",ylab="standardised residuals")

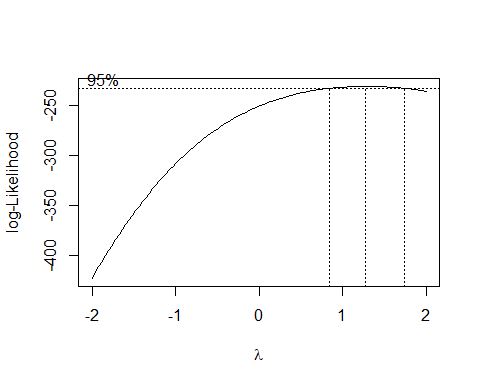


plot(data1$FTV,rstandard(lm1),xlab="FTV",ylab="standardised residuals")

 The above plot shows that there are some points which are not on the line. The residual vs fitted plot shows that it has constant variance.

# b. Run a Box-Cox analysis of the response using the full additive model (no interactions). What transformation is suggested?

boxcox(lm1,lambda=seq(-2,2,by=0.05))



summary(powerTransform(lm1))

## bcPower Transformation to Normality   
## Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd  
## Y1 1.2765 1 0.8315 1.7215  
##   
## Likelihood ratio test that transformation parameter is equal to 0  
## (log transformation)  
## LRT df pval  
## LR test, lambda = (0) 39.40588 1 3.4426e-10  
##   
## Likelihood ratio test that no transformation is needed  
## LRT df pval  
## LR test, lambda = (1) 1.545638 1 0.21378

It is clear that value of lambda is 1, so no transformation is required.

# c. Use step-wise selection to determine the best model according to the AIC criteria. Report the model and the AIC associated with it.

Using the step function,

nullmodel <- lm(BWT~1,data=data1)  
fullmodel <- lm(BWT~ AGE+LWT+RACE+FTV,data=data1)  
step(nullmodel,scope = list(lower=nullmodel,upper=fullmodel),direction = "forward")

## Start: AIC=2492.66  
## BWT ~ 1  
##   
## Df Sum of Sq RSS AIC  
## + RACE 1 3846362 96070691 2487.2  
## + LWT 1 3448881 96468171 2488.0  
## <none> 99917053 2492.7  
## + AGE 1 806927 99110126 2493.1  
## + FTV 1 339165 99577887 2494.0  
##   
## Step: AIC=2487.24  
## BWT ~ RACE  
##   
## Df Sum of Sq RSS AIC  
## + LWT 1 2417227 93653463 2484.4  
## <none> 96070691 2487.2  
## + AGE 1 322514 95748177 2488.6  
## + FTV 1 153208 95917483 2488.9  
##   
## Step: AIC=2484.43  
## BWT ~ RACE + LWT  
##   
## Df Sum of Sq RSS AIC  
## <none> 93653463 2484.4  
## + AGE 1 108497 93544966 2486.2  
## + FTV 1 38457 93615006 2486.3

##   
## Call:  
## lm(formula = BWT ~ RACE + LWT, data = data1)  
##   
## Coefficients:  
## (Intercept) RACE LWT   
## 2706.05 -135.09 3.76

n <- length(data1$BWT)  
step(nullmodel,scope = list(lower=nullmodel,upper=fullmodel),direction = "forward",k=log(n))

## Start: AIC=2495.9  
## BWT ~ 1  
##   
## Df Sum of Sq RSS AIC  
## + RACE 1 3846362 96070691 2493.7  
## + LWT 1 3448881 96468171 2494.5  
## <none> 99917053 2495.9  
## + AGE 1 806927 99110126 2499.6  
## + FTV 1 339165 99577887 2500.5  
##   
## Step: AIC=2493.73  
## BWT ~ RACE  
##   
## Df Sum of Sq RSS AIC  
## <none> 96070691 2493.7  
## + LWT 1 2417227 93653463 2494.2  
## + AGE 1 322514 95748177 2498.3  
## + FTV 1 153208 95917483 2498.7

##   
## Call:  
## lm(formula = BWT ~ RACE, data = data1)  
##   
## Coefficients:  
## (Intercept) RACE   
## 3232.3 -155.8

reducedModel <- step(nullmodel,scope = list(lower=nullmodel,upper=fullmodel),direction = "forward",trace=0)  
summary(reducedModel)

##   
## Call:  
## lm(formula = BWT ~ RACE + LWT, data = data1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2042.93 -462.13 57.21 517.05 1956.60   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2706.053 266.780 10.143 <2e-16 \*\*\*  
## RACE -135.092 57.137 -2.364 0.0191 \*   
## LWT 3.760 1.716 2.191 0.0297 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 709.6 on 186 degrees of freedom  
## Multiple R-squared: 0.06269, Adjusted R-squared: 0.05261   
## F-statistic: 6.22 on 2 and 186 DF, p-value: 0.002428

We are using reduced model as it is much better for AIC. It is seen that race and lwt are less than 0.05 so they are significant. Lwt has the positive relationship with bwt. race has the negative relationship with birth which is weight in grams. The predictors are 6.2% of response of bwt.

# d. Perform a partial F−test to see if you can simultaneously drop the variables that were dropped in part (c). Does this agree with the results of the step-wise selection?

Using anova is partial f-test

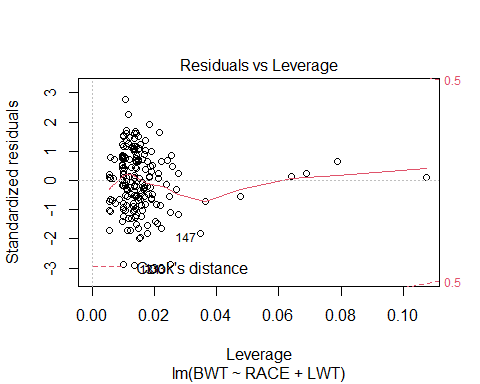
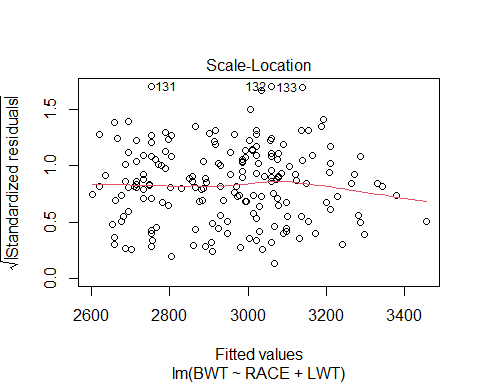
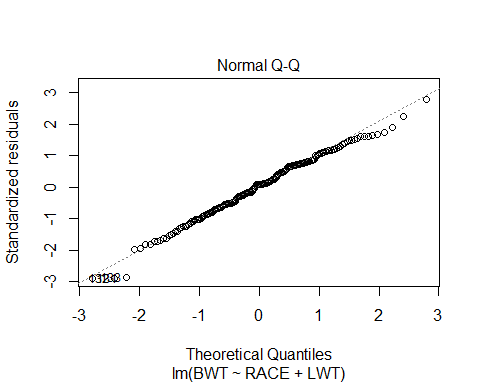
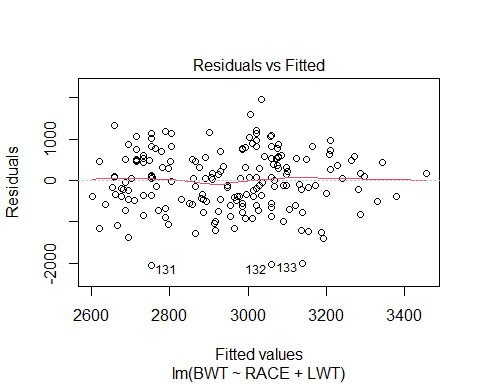
anova(reducedModel,lm1)

## Analysis of Variance Table  
##   
## Model 1: BWT ~ RACE + LWT  
## Model 2: BWT ~ AGE + LWT + RACE + FTV  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 186 93653463   
## 2 184 93526169 2 127294 0.1252 0.8824

We fail to reject null hypothesis so reduced model is better than the full model.

# e. Check the usual assumptions for your final model.

plot(reducedModel)

 There is normal distribution. There appears to be the linear association. This satisfies the conditions of normality, constant variance and linearity.

# f. If someone is not white or black, 35 years old, had 2 visits during the first trimester, and weighed 136 pounds at the last menstrual period, what is estimated birthweight of the baby, using your model?

predict(lm1,data.frame(RACE=3,AGE=35,FTV=2,LWT=136),data=data1)

## 1   
## 2878.533

# g. Reality check. Is your model useful? Why or why not?

summary(reducedModel)

##   
## Call:  
## lm(formula = BWT ~ RACE + LWT, data = data1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2042.93 -462.13 57.21 517.05 1956.60   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2706.053 266.780 10.143 <2e-16 \*\*\*  
## RACE -135.092 57.137 -2.364 0.0191 \*   
## LWT 3.760 1.716 2.191 0.0297 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 709.6 on 186 degrees of freedom  
## Multiple R-squared: 0.06269, Adjusted R-squared: 0.05261   
## F-statistic: 6.22 on 2 and 186 DF, p-value: 0.002428

As the model satisfies the normality assumptions, we can say it is useful.

# Problem 2.

# a. Run a logistic regression and discuss the significance of the predictor variables.

lm2 <-glm(LOW~AGE+LWT+RACE+FTV,data = data1,family = "binomial")  
summary(lm2)

##   
## Call:  
## glm(formula = LOW ~ AGE + LWT + RACE + FTV, family = "binomial",   
## data = data1)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.1995 -0.8938 -0.7206 1.2946 1.9690   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.008923 1.136163 0.888 0.3745   
## AGE -0.032661 0.033374 -0.979 0.3278   
## LWT -0.011438 0.006262 -1.826 0.0678 .  
## RACE 0.230347 0.177409 1.298 0.1942   
## FTV -0.039010 0.165445 -0.236 0.8136   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 234.67 on 188 degrees of freedom  
## Residual deviance: 225.31 on 184 degrees of freedom  
## AIC: 235.31  
##   
## Number of Fisher Scoring iterations: 4

# b. Drop the predictors that are not significant and rerun the model. For any quantitative predictor variables, what do the sign of the estimated regression coefficients tell you about the relationship between these predictors and the probability of low birthweight?

As it is seen that FTV has the p-value is greater than 0.01 and is insignificant, so we decide to remove ftv from the model. Rerunning the model,

lm3<-glm(LOW~AGE+LWT+RACE,data = data1,family="binomial")  
summary(lm3)

##   
## Call:  
## glm(formula = LOW ~ AGE + LWT + RACE, family = "binomial", data = data1)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.2110 -0.8851 -0.7290 1.2719 1.9902   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.020118 1.133633 0.900 0.3682   
## AGE -0.034108 0.032814 -1.039 0.2986   
## LWT -0.011553 0.006234 -1.853 0.0639 .  
## RACE 0.233926 0.176639 1.324 0.1854   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 234.67 on 188 degrees of freedom  
## Residual deviance: 225.37 on 185 degrees of freedom  
## AIC: 233.37  
##   
## Number of Fisher Scoring iterations: 4

The p-value of age is greater than 0.01 and is insignificant, so removing age from the model.

lm4 <-glm(LOW~LWT+RACE,data = data1,family="binomial")  
summary(lm4)

##   
## Call:  
## glm(formula = LOW ~ LWT + RACE, family = "binomial", data = data1)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.1883 -0.9038 -0.7519 1.2743 1.9795   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.312346 0.905857 0.345 0.7302   
## LWT -0.012520 0.006196 -2.021 0.0433 \*  
## RACE 0.259138 0.174360 1.486 0.1372   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 234.67 on 188 degrees of freedom  
## Residual deviance: 226.48 on 186 degrees of freedom  
## AIC: 232.48  
##   
## Number of Fisher Scoring iterations: 4

Now, the p-value of race is 0.13 and is not significant so that is being dropped.

lm4<- glm(LOW~LWT,data=data1,family="binomial")  
summary(lm4)

##   
## Call:  
## glm(formula = LOW ~ LWT, family = "binomial", data = data1)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.0951 -0.9022 -0.8018 1.3609 1.9821   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.99831 0.78529 1.271 0.2036   
## LWT -0.01406 0.00617 -2.279 0.0227 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 234.67 on 188 degrees of freedom  
## Residual deviance: 228.69 on 187 degrees of freedom  
## AIC: 232.69  
##   
## Number of Fisher Scoring iterations: 4

# c. If someone is not white or black, 35 years old, had 2 visits during the first trimester, and weighed 136 pounds at the last menstrual period, what is estimated probability that the baby has low birthweight, using your model?

predict(lm2,data.frame(RACE=3,AGE=35,FTV=2,LWT=136),data=data1,type = "response")

## 1   
## 0.2541249

# d. Does your answer to question 2 part (c) seem consistent with your answer to question 1 part (f)? Explain.

From the outputs, it is consistent.