Chap3 HW

Tanuj Guha

2023-04-03

##Question 3.26

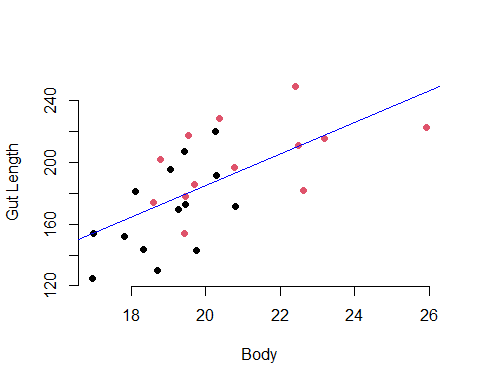
Loading the dataset.

tadpole = read.csv("Tadpoles.csv")  
summary(tadpole)

## Treatment Body GutLength MouthpartDamage   
## Length:27 Min. :16.95 Min. :124.9 Min. :0.4140   
## Class :character 1st Qu.:18.75 1st Qu.:162.1 1st Qu.:0.5890   
## Mode :character Median :19.46 Median :181.6 Median :0.6790   
## Mean :19.94 Mean :184.2 Mean :0.6587   
## 3rd Qu.:20.58 3rd Qu.:208.9 3rd Qu.:0.7140   
## Max. :25.92 Max. :249.3 Max. :0.8570

3.26.a

x <- tadpole$Body  
y <- tadpole$GutLength  
  
plot(x, y, col=as.factor(tadpole$Treatment),  
 xlab = "Body", ylab = "Gut Length",  
 pch = 19, frame = FALSE)  
abline(lm(y ~ x, data = tadpole), col = "blue")



3.26.b

m = lm(GutLength ~ Body, data = tadpole)  
summary(m)

##   
## Call:  
## lm(formula = GutLength ~ Body, data = tadpole)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -41.575 -22.245 0.027 17.815 39.998   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -20.764 49.002 -0.424 0.675384   
## Body 10.280 2.445 4.204 0.000293 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 24.83 on 25 degrees of freedom  
## Multiple R-squared: 0.4141, Adjusted R-squared: 0.3907   
## F-statistic: 17.67 on 1 and 25 DF, p-value: 0.0002931

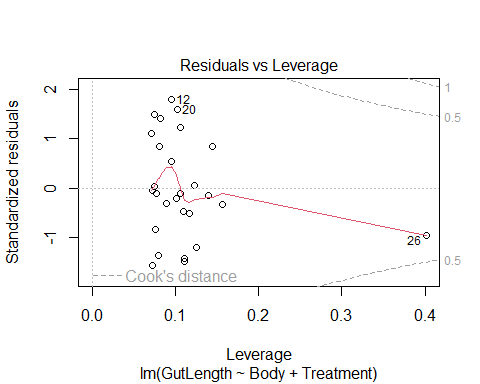
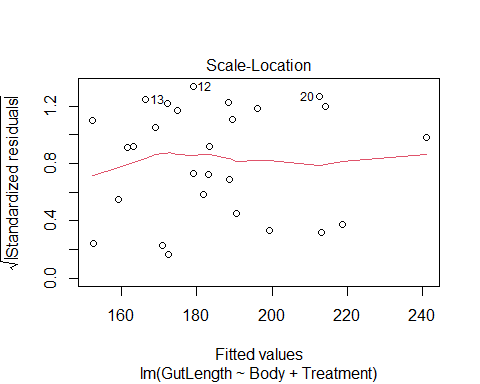
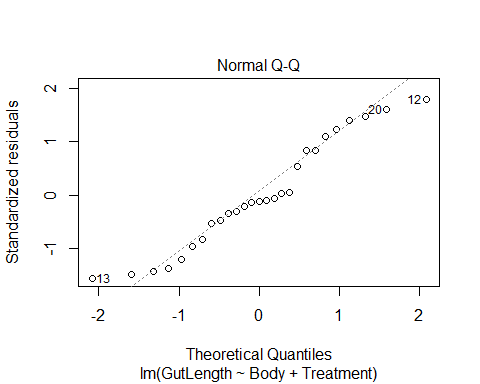
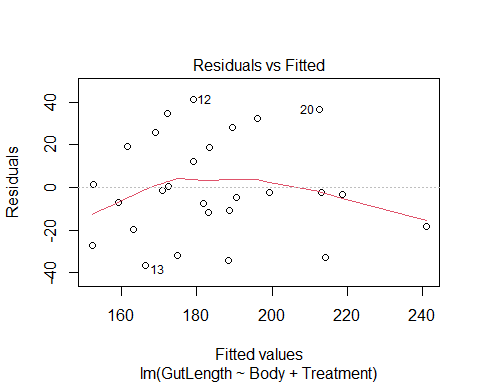
Here we see that the R^2 value is ~0.41. Hence there is no strong linear relationship between GutLength and Body.

3.26.c

m1 = lm(GutLength ~ Body + Treatment, data = tadpole)  
summary(m1)

##   
## Call:  
## lm(formula = GutLength ~ Body + Treatment, data = tadpole)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -36.462 -15.006 -2.545 19.117 41.298   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 15.47 53.81 0.288 0.77618   
## Body 8.07 2.82 2.862 0.00859 \*\*  
## TreatmentControl 16.28 11.03 1.476 0.15286   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 24.26 on 24 degrees of freedom  
## Multiple R-squared: 0.4629, Adjusted R-squared: 0.4181   
## F-statistic: 10.34 on 2 and 24 DF, p-value: 0.0005762

plot(m1)

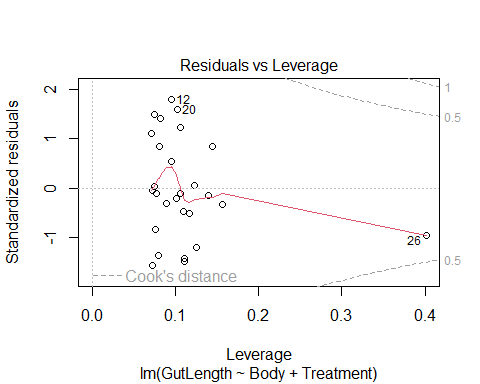
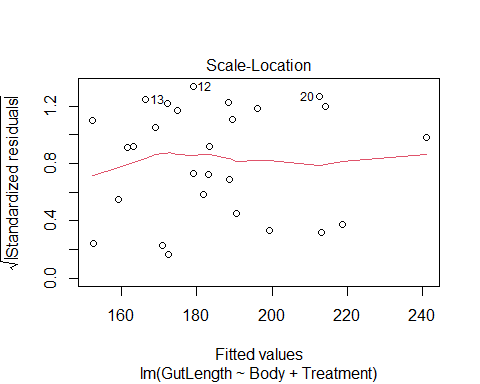
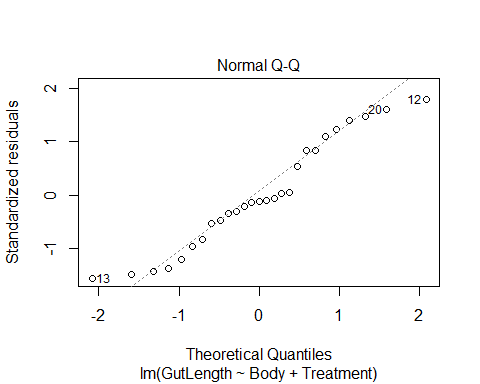
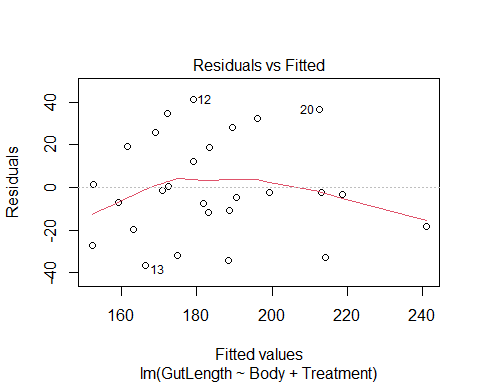


3.26.d

m2 = lm(GutLength ~ Body + Treatment + MouthpartDamage, data = tadpole)  
summary(m1)

##   
## Call:  
## lm(formula = GutLength ~ Body + Treatment, data = tadpole)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -36.462 -15.006 -2.545 19.117 41.298   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 15.47 53.81 0.288 0.77618   
## Body 8.07 2.82 2.862 0.00859 \*\*  
## TreatmentControl 16.28 11.03 1.476 0.15286   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 24.26 on 24 degrees of freedom  
## Multiple R-squared: 0.4629, Adjusted R-squared: 0.4181   
## F-statistic: 10.34 on 2 and 24 DF, p-value: 0.0005762

plot(m1)



##Interpretation 3.26: In this case there is not much discernible difference in r^2 values between the models. Residuals V Fitted plots for both these models appear as ‘clouds’ and as such do not seem to indicate the necessity for further action. Q-Q Plot looks better with the model that does include MouthpartDamage, which indicates that the residuals of the Model ‘m2’ follow the normal distribution ‘better’ than m1. — ##Question 3.32

Loading the dataset.

speed = read.csv("Speed.csv")  
summary(speed)

## Year FatalityRate StateControl   
## Min. :1987 Min. :1.360 Min. :0.000   
## 1st Qu.:1992 1st Qu.:1.510 1st Qu.:0.000   
## Median :1997 Median :1.650 Median :1.000   
## Mean :1997 Mean :1.715 Mean :0.619   
## 3rd Qu.:2002 3rd Qu.:1.750 3rd Qu.:1.000   
## Max. :2007 Max. :2.410 Max. :1.000

3.32.a

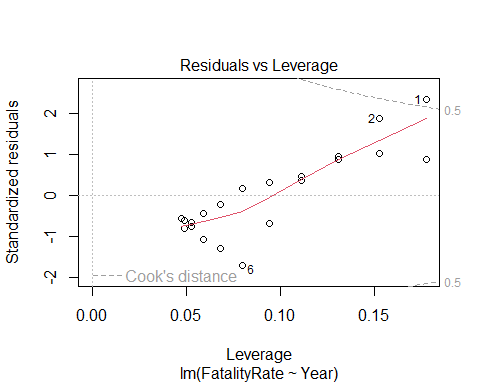
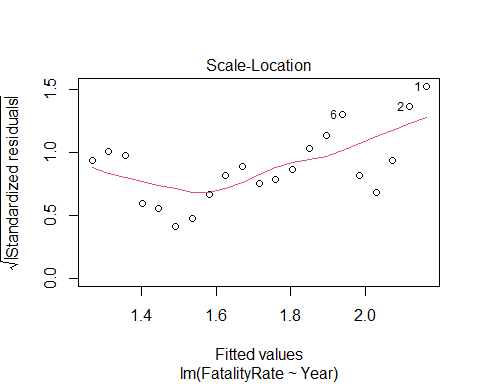
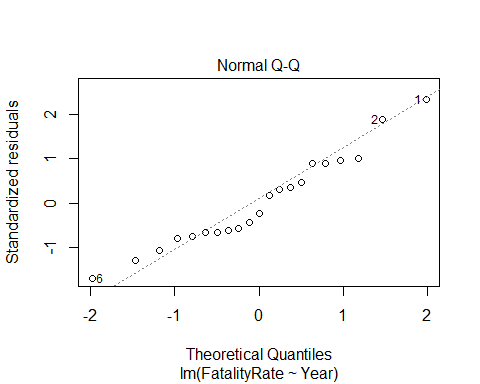
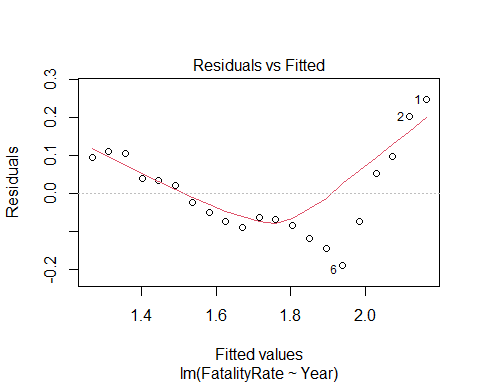
m3 = lm(FatalityRate ~ Year, data = speed)  
summary(m3)

##   
## Call:  
## lm(formula = FatalityRate ~ Year, data = speed)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.18959 -0.07550 -0.02576 0.09346 0.24606   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 91.320887 8.374227 10.9 1.28e-09 \*\*\*  
## Year -0.044870 0.004193 -10.7 1.75e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1164 on 19 degrees of freedom  
## Multiple R-squared: 0.8577, Adjusted R-squared: 0.8502   
## F-statistic: 114.5 on 1 and 19 DF, p-value: 1.75e-09

The least squares slope is -0.044870 fatalities per 100 / year.

3.32.b

plot(m3)



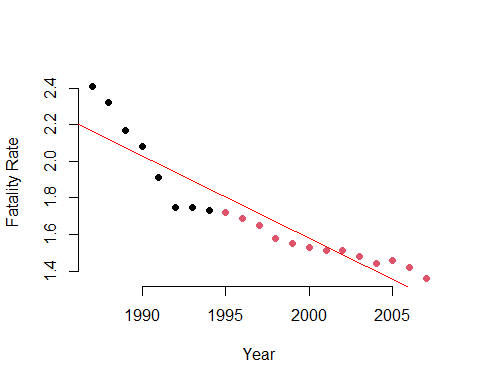
The resualds do not follow a random distribution.

3.32.c

m4 = lm(FatalityRate ~ Year + StateControl + Year\*StateControl, data = speed)  
summary(m4)

##   
## Call:  
## lm(formula = FatalityRate ~ Year + StateControl + Year \* StateControl,   
## data = speed)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.103571 -0.020769 0.004048 0.022473 0.091667   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.162e+02 1.303e+01 16.59 6.19e-12 \*\*\*  
## Year -1.076e-01 6.548e-03 -16.44 7.19e-12 \*\*\*  
## StateControl -1.614e+02 1.447e+01 -11.15 3.07e-09 \*\*\*  
## Year:StateControl 8.097e-02 7.264e-03 11.15 3.08e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.04243 on 17 degrees of freedom  
## Multiple R-squared: 0.9831, Adjusted R-squared: 0.9801   
## F-statistic: 329 on 3 and 17 DF, p-value: 2.998e-15

plot(speed$Year, speed$FatalityRate, col=as.factor(speed$StateControl),  
 xlab = "Year", ylab = "Fatality Rate",  
 pch = 19, frame = FALSE)  
abline(lm(speed$FatalityRate ~ speed$Year, data = speed), col = "red")



We can see that the variable StateControl changed in 1995. There is significant change in the relationship between fatality rate and year starting in 1995

3.32.d

speedOld = subset(speed , speed$Year < 1995, select = c(FatalityRate, Year))  
speedNew = subset(speed , speed$Year > 1994, select = c(FatalityRate, Year))  
m5 = lm(FatalityRate ~ Year, data = speedOld)  
m6 = lm(FatalityRate ~ Year, data = speedNew)  
coef(m5)

## (Intercept) Year   
## 216.230714 -0.107619

coef(m6)

## (Intercept) Year   
## 54.85412088 -0.02664835

Writing the actual equations, before and after 1995 respectively:

model\_equation <- function(model, ...) {  
 format\_args <- list(...)  
   
 model\_coeff <- model$coefficients  
 format\_args$x <- abs(model$coefficients)  
 model\_coeff\_sign <- sign(model\_coeff)  
 model\_coeff\_prefix <- case\_when(model\_coeff\_sign == -1 ~ " - ",  
 model\_coeff\_sign == 1 ~ " + ",  
 model\_coeff\_sign == 0 ~ " + ")  
 model\_eqn <- paste(strsplit(as.character(model$call$formula), "~")[[2]], # 'y'  
 "=",  
 paste(if\_else(model\_coeff[1]<0, "- ", ""),  
 do.call(format, format\_args)[1],  
 paste(model\_coeff\_prefix[-1],  
 do.call(format, format\_args)[-1],  
 " \* ",  
 names(model\_coeff[-1]),  
 sep = "", collapse = ""),  
 sep = ""))  
 return(model\_eqn)  
}  
model\_equation(m5, digits = 2)

## [1] "FatalityRate = 216.23 - 0.11 \* Year"

model\_equation(m6, digits = 2)

## [1] "FatalityRate = 54.854 - 0.027 \* Year"

Fatality rates decreased faster before 1995. we can say that there is evidence that transferring control of highway speeds to state governments from the federal government had an adverse affect on highway safety.