PSTAT 126 HW #3

Mujie Wang

04/27/2018

```
#Problem 1
#(a)
library(faraway)
data(prostate)

lpsa = prostate$lpsa
lcavol = prostate$lcavol

fit = lm(lpsa ~ lcavol)
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: lpsa
##           Df Sum Sq Mean Sq F value    Pr(>F)
## lcavol     1 69.003  69.003  111.27 < 2.2e-16 ***
## Residuals 95 58.915   0.620
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#(b)
dim(anova(fit))
```

```
## [1] 2 5
```

```
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: lpsa
##           Df Sum Sq Mean Sq F value    Pr(>F)
## lcavol     1 69.003  69.003  111.27 < 2.2e-16 ***
## Residuals 95 58.915   0.620
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
ssr = anova(fit)[1,2]
ssr
```

```
## [1] 69.00283
```

```
sse = anova(fit)[2,2]
sse
```

```
## [1] 58.91476
```

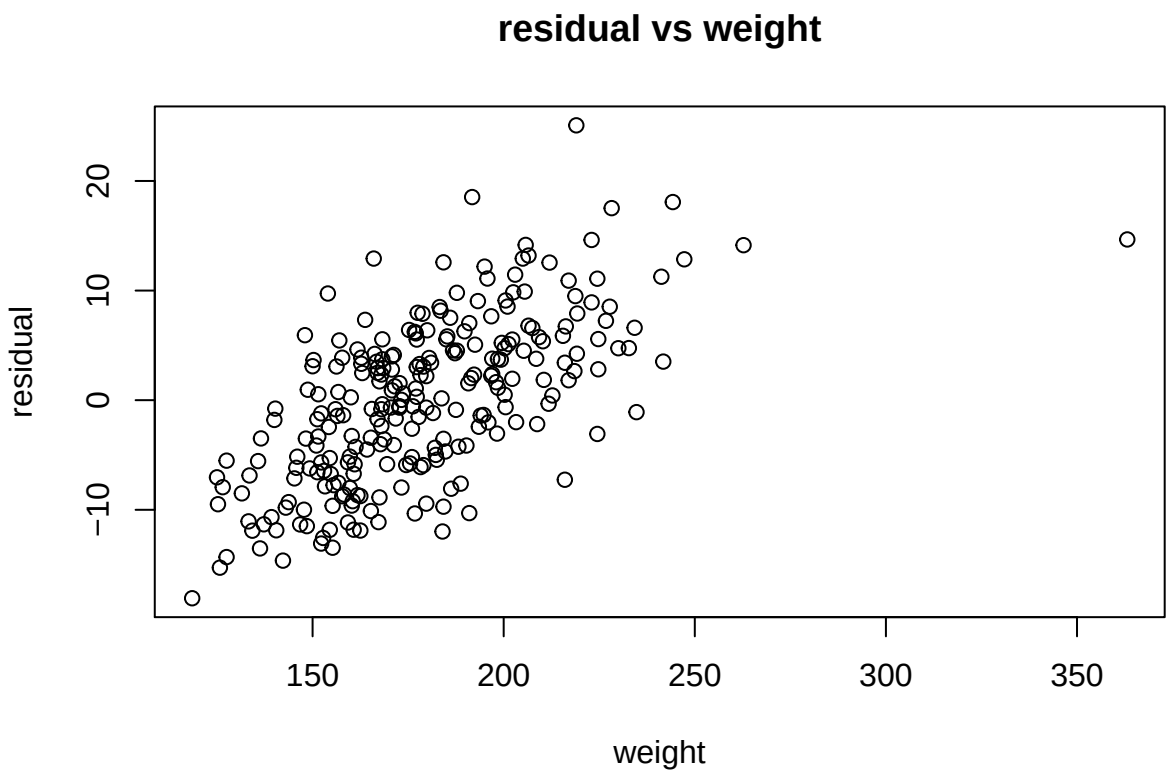
```
ssto = sse + ssr
ssto
```

```
## [1] 127.9176
```

```r
#From the anova table, ssr = 69.003 and it is explained by regression.
#sse = 58.915 and which is unexplained.
#The total variability in lpsa ssto = ssr + sse = 127.918.

#(c)
# Alternative hypothesis test: (H0: not equal 0)
#Decision rule: if p-value < a, we reject H0
#otherwise, we fail to reject H0; if the test statistic
#F* > F(1-a,1,n-2), then we reject H0; otherwise, we fail to reject H0.
#Conclusion: by the ANOVA table, the p-value for the
#F-test is extremely close to zero.
#We can reject H0: B1 =0 given that a =0.05,
#and conclude that there is a signi cant linear relationship between lpsa and lcavol.

#Problem 2
#(a)
library(faraway)
data(prostate)

age = fat$age
brozek = fat$brozek
fit = lm(brozek ~ age)

#(b)
r2 = summary(fit)$r.squared
r2
```

```
## [1] 0.08362132
```

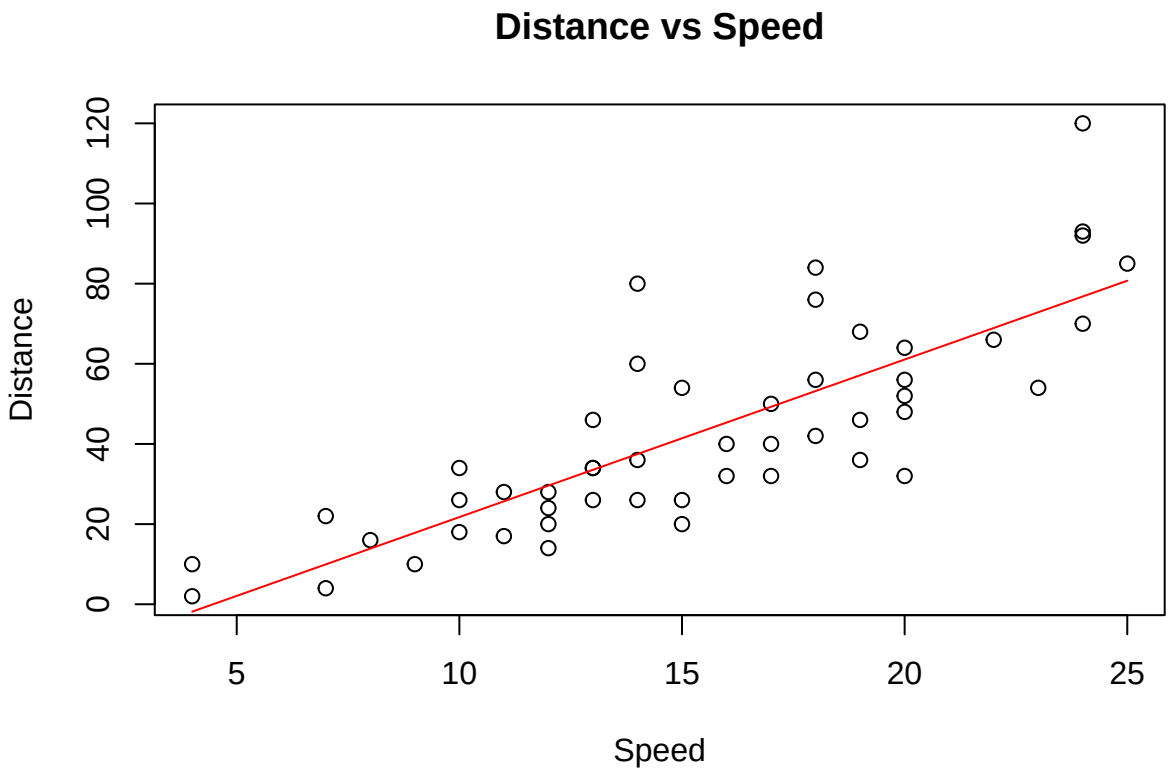
```r
#Coefficient of determination R2 = 0.083621324. About 8.4% variability
#in the response brozek is reduced (explained) by the predictor age.

#(c)
weight = fat$weight
e = resid(fit)
plot(weight,e, xlab = 'weight', ylab = 'residual', main = 'residual vs weight')
```



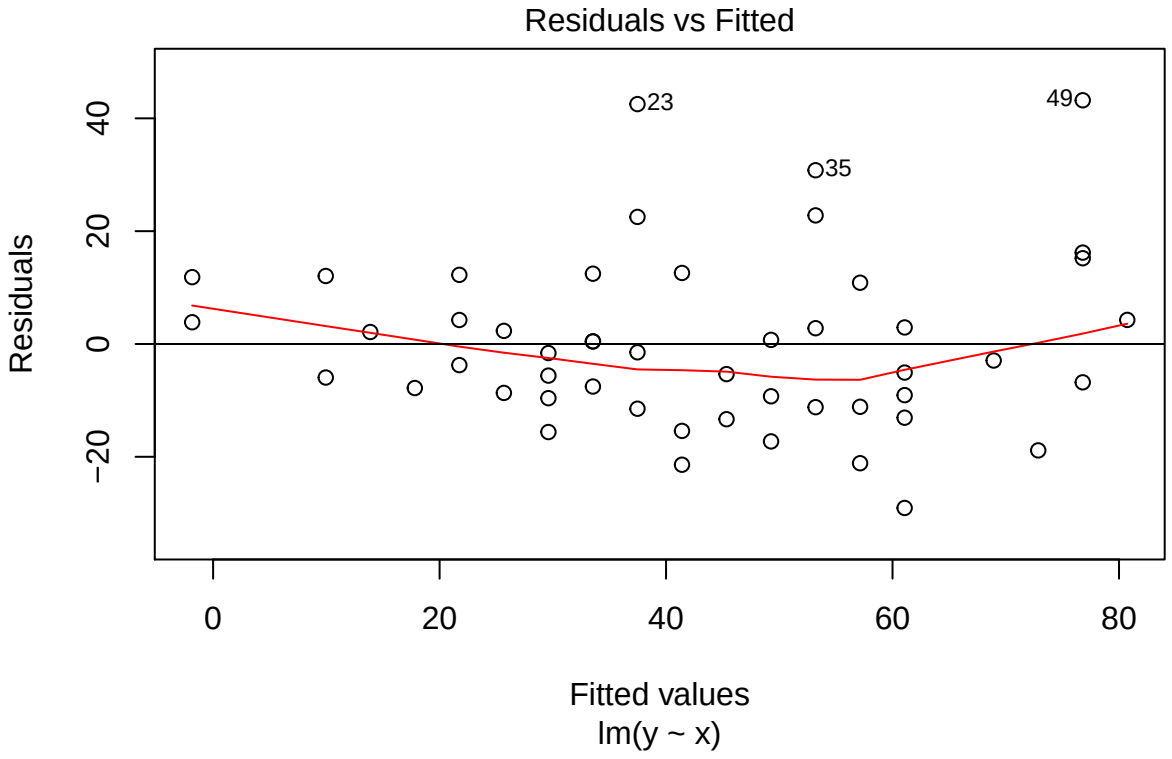

```
#(d)
fit_new = lm(brozek ~ age + weight)
r2_new = summary(fit_new)$r.squared
r2_new
```

```
## [1] 0.4641771
```

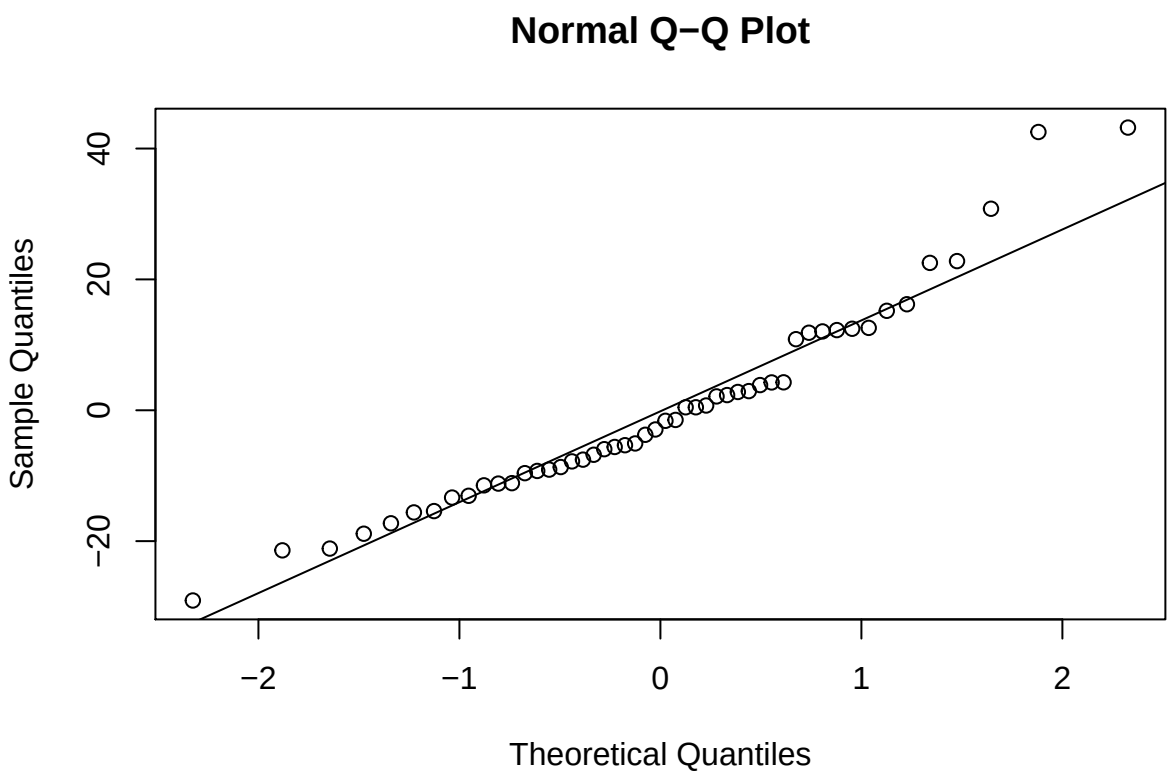
```
#Problem 3
#(a)
library(datasets)
library(MASS)

data(cars)

x = cars$speed
y = cars$dist

fit = lm(y~x)
plot(x,y, xlab = "Speed", ylab = "Distance", main = "Distance vs Speed")
#the linear line shows that there is a positive relationship between
lines(x, fitted(fit), col = "RED")
```




```
#(b)
yhat = fitted(fit)
e = y - yhat
plot(fit, which = 1)
abline(h=0)
```



```r
#sccording to the plot, we can see that there are more dots under the line and the dots are more disper
#therefore, it indicated non-constant error variance.

#(c)
qqnorm(e)
qqline(e)
```




```r
#according to the plot, the line is almost to linear relationship but it still has some outliers.

#(d)
shapiro.test(e)$p
```

```
## [1] 0.02152458
```

the residual of the linear fit in part a is normally distributed, but the vriance is not. In shapiro test, we conclude that p-value is 0.02152 which is smaller than our confidence level 0.05, so we should reject that errors are normally distributed.