Homework 4

Jing Leng

October 2, 2014

1

a)

```
lm <- lm(lpsa~., prostate)
summary(lm)$coefficient
```

```
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  0.669337   1.296387  0.5163 6.069e-01
## lcavol       0.587022   0.087920  6.6767 2.111e-09
## lweight      0.454467   0.170012  2.6731 8.955e-03
## age         -0.019637   0.011173 -1.7576 8.229e-02
## lbph         0.107054   0.058449  1.8316 7.040e-02
## svi          0.766157   0.244309  3.1360 2.329e-03
## lcp         -0.105474   0.091013 -1.1589 2.496e-01
## gleason      0.045142   0.157465  0.2867 7.750e-01
## pgg45        0.004525   0.004421  1.0235 3.089e-01
```

```
confint(lm)["age",]
```

```
##     2.5 %    97.5 %
## -0.041841  0.002566
```

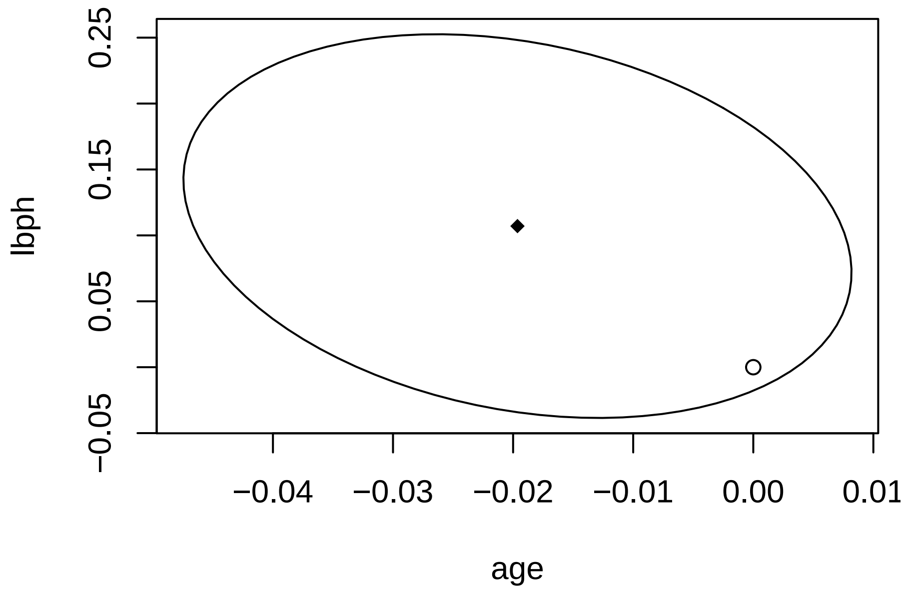
```
confint(lm, level = 0.9)["age",]
```

```
##       5 %      95 %
## -0.038210 -0.001064
```

0 is in the 95% confidence interval of parameter for `age`, we fail to reject the null hypothesis. The p-value for variable `age` is $0.08229 > 0.05$, therefore we fail to reject the null hypothesis.
0 is not in the 90% confidence interval of parameter for `age`, we reject the null hypothesis. The p-value $0.08229 < 0.1$, we reject the null hypothesis.

b)

```
library(ellipse)
plot(ellipse(lm, c("age", "lbph")),
     type="l")
points(lm$coef["age"],lm$coef["lbph"],pch=18)
points(0,0)
```



The null hypothesis is:

$$H_0 : \beta_{age} = \beta_{lbph} = 0$$

Since the origin is within the ellipse, we fail to reject the null hypothesis.

c)

```
newdata <- data.frame(lcavol = 1.44692, lweight = 3.62301, age = 65, lbph = 0.3001,
                      svi = 0, lcp = -0.79851, gleason = 7, pgg45 = 15)

predict(lm, newdata, interval = "confidence")
```

```
##     fit   lwr   upr
## 1 2.389 2.172 2.606
```

```
predict(lm, newdata, interval = "prediction")
```

```
##     fit    lwr   upr
## 1 2.389 0.9647 3.813
```

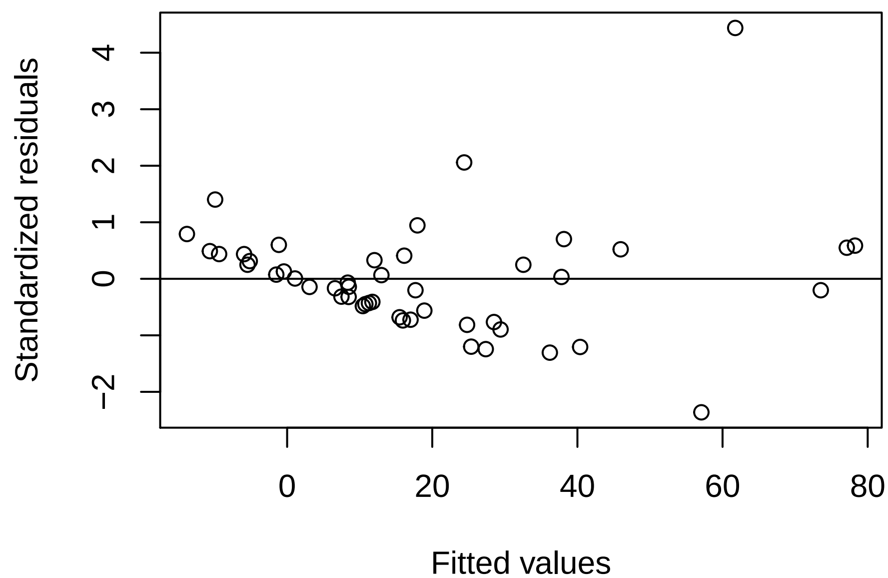
The confidence interval for mean response is (2.17, 2.61), the prediction interval for new response is (0.96, 3.81).

2

a)

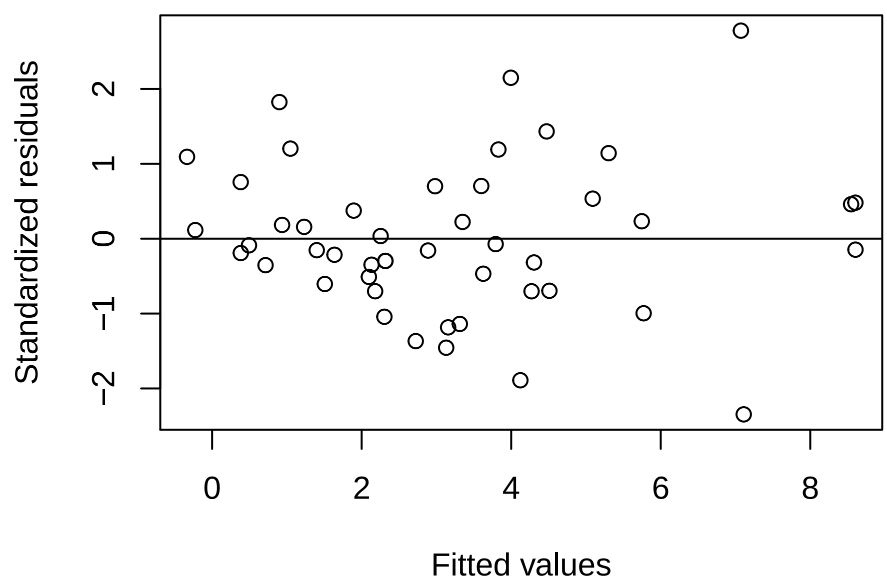
```
lm2 <- lm(gamble ~ ., data = teengamb)
st_res <- rstandard(lm2)
plot(fitted.values(lm2),st_res,
```

```
      xlab="Fitted values",
      ylab="Standardized residuals")
abline(h = 0)
```



The constant variance assumption does not hold. We transform the response variable into square root of `gamble`.

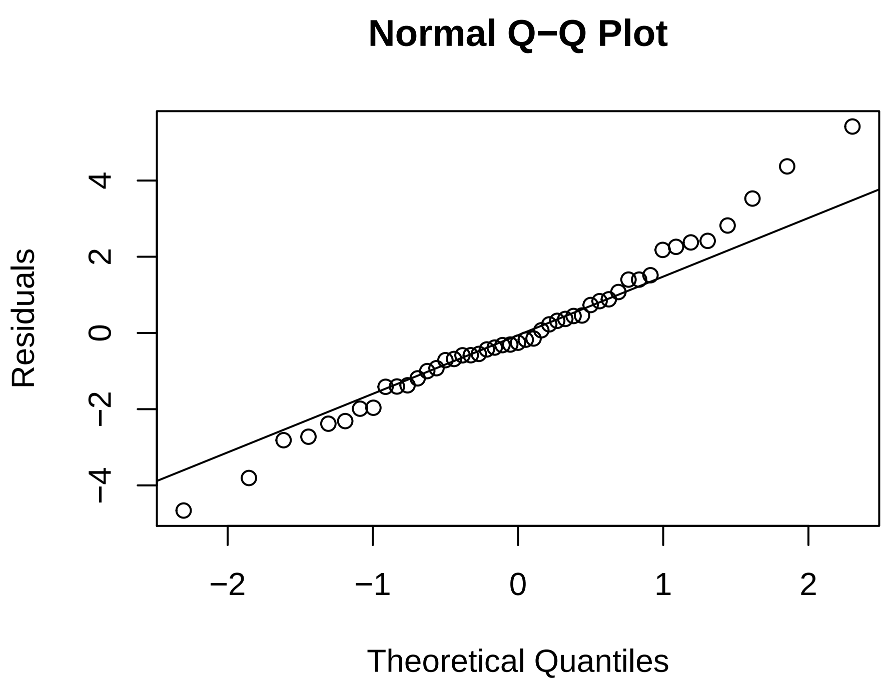
```
newgamble <- sqrt(teengamb$gamble)
teengamb$newgamble <- newgamble
lm3 <- lm(newgamble ~ . - gamble, data = teengamb)

st_res <- rstandard(lm3)
plot(fitted.values(lm3),st_res,
      xlab="Fitted values",
      ylab="Standardized residuals")
abline(h = 0)
```



The constant variance assumption holds. We will continue with the new model.

b)

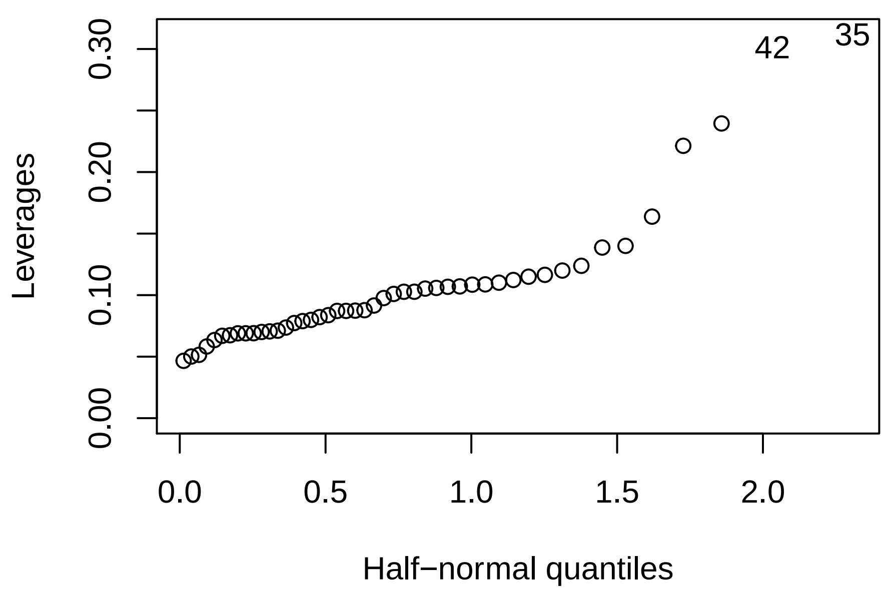
```r
qqnorm(lm3$residual, ylab="Residuals")
qqline(lm3$residual)
```



Residuals can be seen as normaly distributed.

c)

```r
halfnorm(influence(lm3)$hat, nlab = 2, ylab="Leverages")
```



```r
teengamb[c(42, 35), ]
```

```
##    sex status income verbal gamble newgamble
## 42   0     61   15.0      9   69.7     8.349
## 35   0     28    1.5      1   14.1     3.755
```

Number 42 and number 35 has the largest leverages.

d)

```r
ti <- rstudent(lm3)
max(abs(ti))
```

```
## [1] 3.037
```

```r
which(abs(ti) == max(abs(ti)))
```

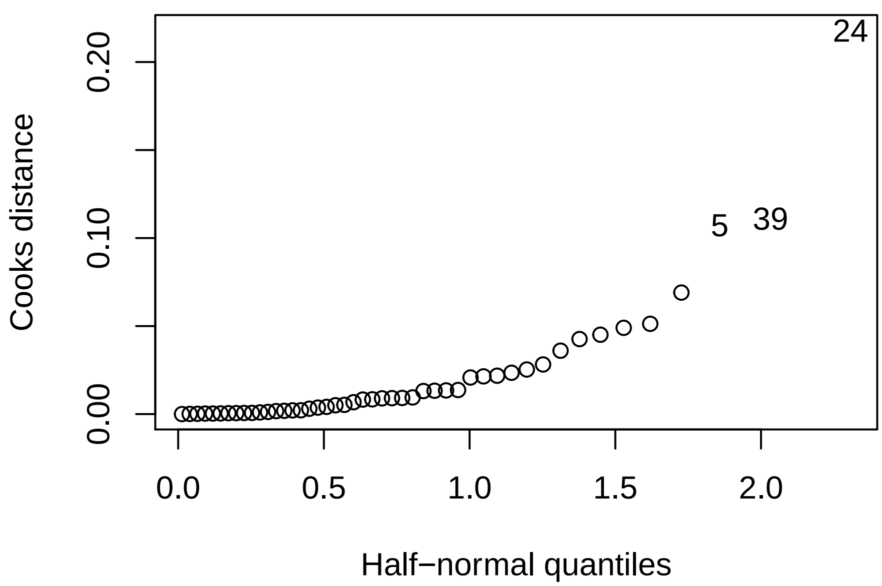
```
## 24
## 24
```

```r
p <- 2*(1-pt(max(abs(ti)), df=47-4-1))
thres <- 0.05/47
```

The p-value $= 0.0041 > 0.0011$, we fail to reject the null hypothsis. Therefore we cannot name a outlier from the data.

e)

```r
cook <- cooks.distance(lm3)
halfnorm(cook, nlab = 3, ylab="Cooks distance")
```



number 5, 39 and 24 are suspected to be influential points.

```
lm4 <- lm(newgamble ~ . - gamble, data = teengamb[-24, ])
predict(lm4, teengamb[24,], interval ="prediction")
```

```
##      fit   lwr   upr
## 24 6.307 2.195 10.42
```

```
teengamb[24,]$newgamble
```

```
## [1] 12.49
```

```
lm5 <- lm(newgamble ~ . - gamble, data = teengamb[-39, ])
predict(lm5, teengamb[39,], interval ="prediction")
```

```
##     fit   lwr   upr
## 39 7.58 3.414 11.75
```

```
teengamb[39,]$newgamble
```

```
## [1] 2.449
```

```
lm6 <- lm(newgamble ~ . - gamble, data = teengamb[-5, ])
predict(lm6, teengamb[5,], interval ="prediction")
```

```
##      fit    lwr   upr
## 5 0.3318 -4.073 4.736
```

```
teengamb[5,]$newgamble
```

```
## [1] 4.427
```

The real response for 24 and 39 are not in the respective prediction intervals in models excluding them. Therefore 24 and 39 are influential points. The real response for 5 is in the prediction interval, so it is not influential.