hw7_yw3204

NAME: Yuhao Wang, UNI: yw3204

11/10/2018

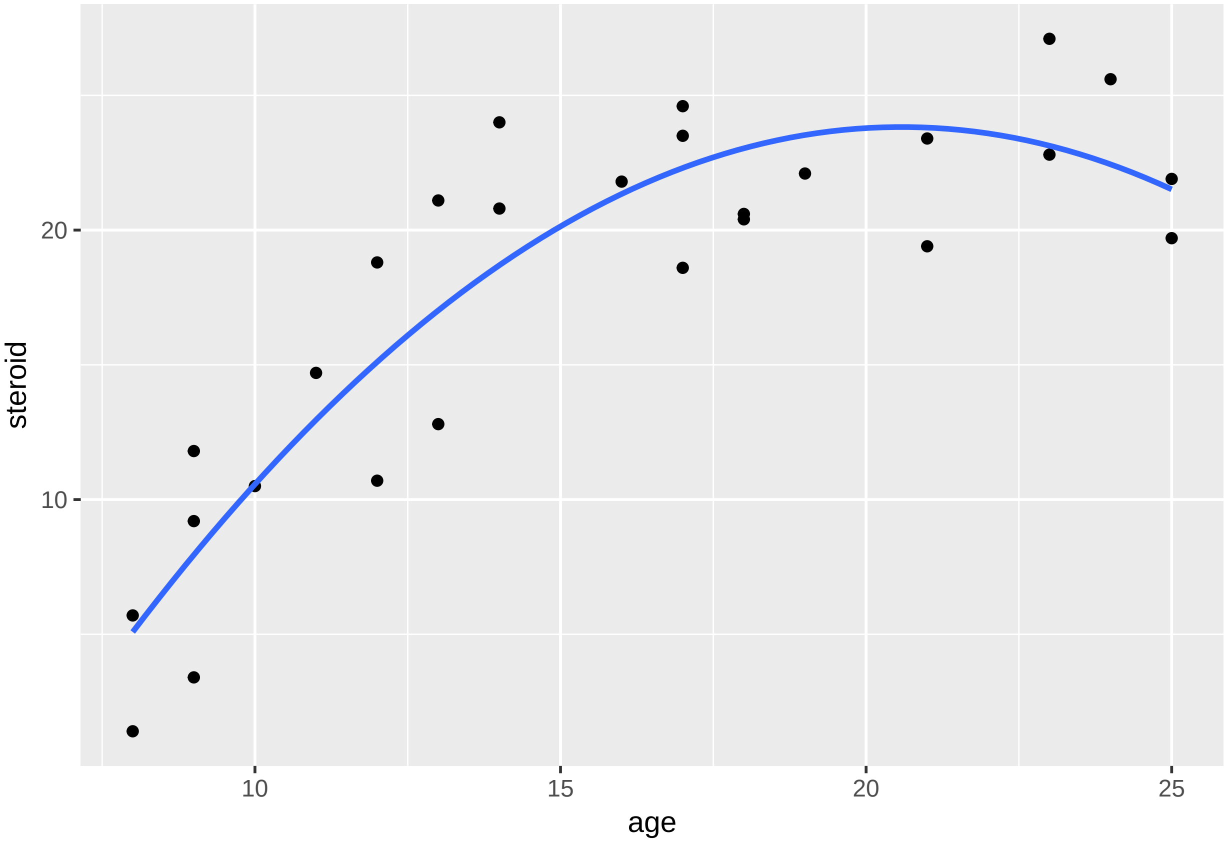
8.6

a

```
steroid <- read.table("CH08PR06.txt")
names(steroid) <- c("Y", "X1")
steroid$X2 <- steroid$X1^2

lm1 <- lm(Y~X1+X2, steroid)
#coef(lm1)

ggplot(steroid, aes(x = X1, y = Y)) + geom_point() +
  geom_smooth(method = 'lm', formula = y ~ poly(x, 2), se= FALSE) +
  labs(x = "age", y = "steroid")
```



The fitted line is $Y = -26.33 + 4.87 * X - 0.12 * X^2$. It appears to be a good fit according to the plot. And the coefficient of determination is 0.8143 which further consolidates the conclution.

b.

```r
summary(lm1)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2, data = steroid)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5463 -2.5369  0.3868  2.1973  5.3020
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -26.32541    5.88154  -4.476 0.000157 ***
## X1            4.87357    0.77515   6.287 1.69e-06 ***
## X2           -0.11840    0.02347  -5.045 3.71e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.153 on 24 degrees of freedom
## Multiple R-squared:  0.8143, Adjusted R-squared:  0.7989
## F-statistic: 52.63 on 2 and 24 DF,  p-value: 1.678e-09
```

```r
qf(0.99, 2, 24)
```

```
## [1] 5.613591
```

```r
1-pf(52.63, 2, 24)
```

```
## [1] 1.678669e-09
```

The alternative H_a is $\beta_1 \neq 0$ or $\beta_2 \neq 0$. The decision rule is when the calculated F-statistic is greater than the theoretical $F(0.99, 2, 24)$, we reject the null. Accoring to the summary above, the calculated F is 52.63, which is greater than $F(0.99, 2, 24) = 5.61$., we thus accept the alternative. The P-value is $1.68 * 10^{-9}$, which is significantly smaller than 0.01 and thus further consolidates our conclusion.

c.

```r
# Working-Hotelling
W <- sqrt(2*qf(0.99, 2, 25))
# Bonferroni
B <- qt(1 - 0.01/6, 25)

# anova(lm1) # to find mse
mse <- 9.939167
des_X <- matrix(nrow = 27, ncol = 3)
des_X[, 1] <- 1
des_X[, 2] <- steroid[, 2]
des_X[, 3] <- steroid[, 3]

cov_b <- mse * solve(t(des_X) %*% des_X)
b <- as.numeric(coef(lm1))
```

```r
for(i in c(10, 15, 20)) {
  ci <- c()
  y_h <- sum(b * c(1, i, i^2))
  s_h <- t(c(1, i, i^2)) %*% cov_b %*% c(1, i, i^2)
  s_h <- sqrt(s_h)
  ci[1] <- y_h - s_h * B
  ci[2] <- y_h + s_h * B
  print(ci)
}
```

```
## [1]  7.573032 13.567381
## [1] 17.24158 23.03426
## [1] 21.00352 26.56763
```

We calculatev the joint interval estimates based on the Bonferroni procedure because its intervals are narrower. And according to the result above, we say the joint intervals are $(7.57, 13.57)$, $(17.24, 23.03)$ and $(21.00, 26.57)$ seperately for age 10, 15 and 20. It means the probabilty that the mean responses at the three levels fall into their respective intervals simultaneously is 0.99.

d.

```r
t <- qt(0.995, 25)
y_h <- sum(b * c(1, 10, 100))
s_pred <- sqrt(mse * (1 + t(c(1, 10, 100)) %*% solve(t(des_X) %*% des_X) %*% c(1, 10, 100)))
ci <- c()
ci[1] <- y_h - s_pred * t
ci[2] <- y_h + s_pred * t
ci
```

```
## [1]  1.412876 19.727537
```

The interval is $(1.41, 19.73)$, which means the new observed steroid level at age 15 will fall into the it with probability 0.99.

e.

```r
lm2 <- lm(Y~X1, steroid)
#anova(lm2)
# sse(x1) = 491.53
#anova(lm1)
# sse(x1, x2) = 238.54

F_star <- (491.53 - 238.54) / 238.54 * 24
F_star
```

```
## [1] 25.45384
```

```r
qf(0.99, 1, 24)
```

```
## [1] 7.822871
```

The alternatibve is $\beta_2 \neq 0$. And the decision rule is when the clacilated F-statistic is greater than the theoretical F, we accept the alternative. And based on the result above, we reject the null because the calculated F-statistis, which is 25.45, is greater than $F(0.99, 1, 24) = 7.82$.

f.

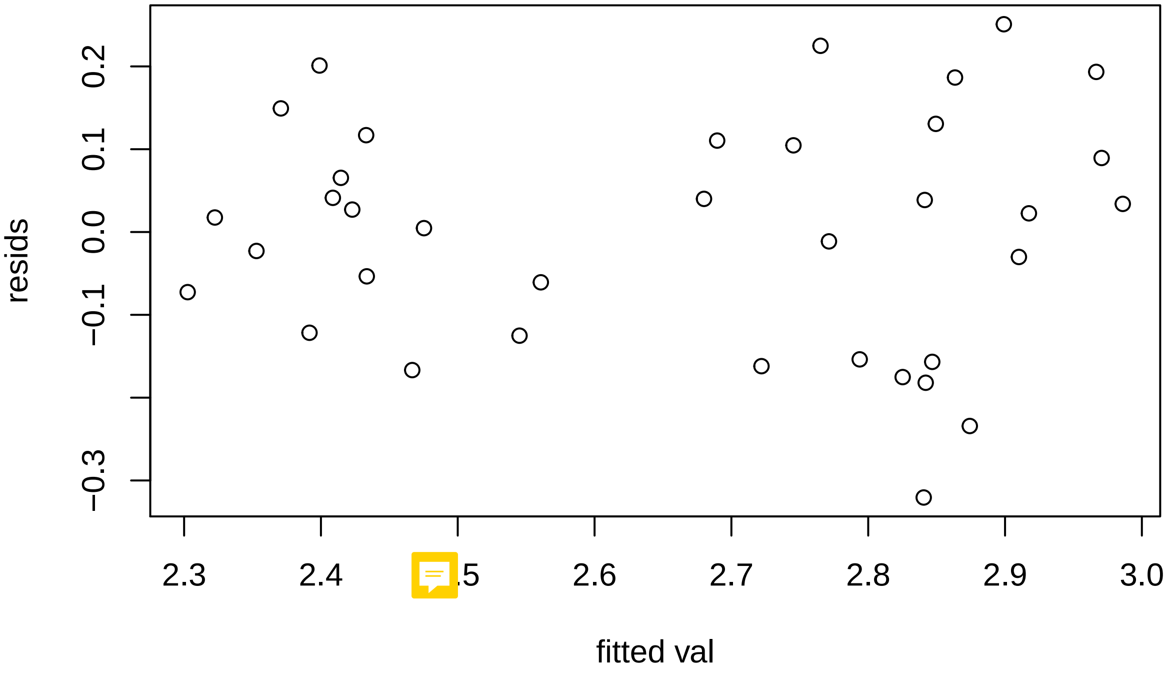
The fitted line is $Y = -26.33 + 4.87 * X - 0.12 * X^2$.

8.42

a.

```r
market <- read.table("APPENC03.txt")
market <- market[, c(2, 3, 4, 5, 6, 8)]
names(market) <- c("Y", "X1", "X2", "X3", "X4", "X5")
market$X5 <- ifelse(market$X5 == 2000, 1, 0)

lm_market <- lm(Y~X1+X2+X3+X4+X5, market)
plot(fitted(lm_market), resid(lm_market), xlab = "fitted val", ylab = "resids")
```



```r
#summary(lm_market)
# R^2 0.7181
```

According to the $R^2 = 0.72$, the model fits pretty well.

```r
lm_sec <- lm(Y~X1+X2+X3+X4+X5+X1^2+X2^2+X1*X2, market)
#anova(lm_market)
# sse 0.68961
#anova(lm_sec)
# sse 0.68872

F_star <- ((0.68961 - 0.68872) / 3) / (0.68872 / 27)
# 0.01163027
qf(0.95, 3, 27)
```

```
## [1] 2.960351
```

```
# 2.960351
```

The alternative is not all the coefficients of the second terms are 0. And the decision rule is when the F-statistic is greater than the theoretical $F(0.95, 3, 27)$, we reject the null. Accoding to the result, since the F-statistic 0.012 is smaller than $F(0.95, 3, 27) = 2.96$, we accept the null. Thus, not all the second terms should be included.

c.

```
lm_drop <- lm(Y~X1+X3+X4, market)
#anova(lm_drop)
#0.71795
#anova(lm_market)
# sse 0.68961

F_star_d <- ((0.71795 - 0.68961) / 2) / (0.68961 / 30)
# 0.6164354
qf(0.95, 2, 30)
```

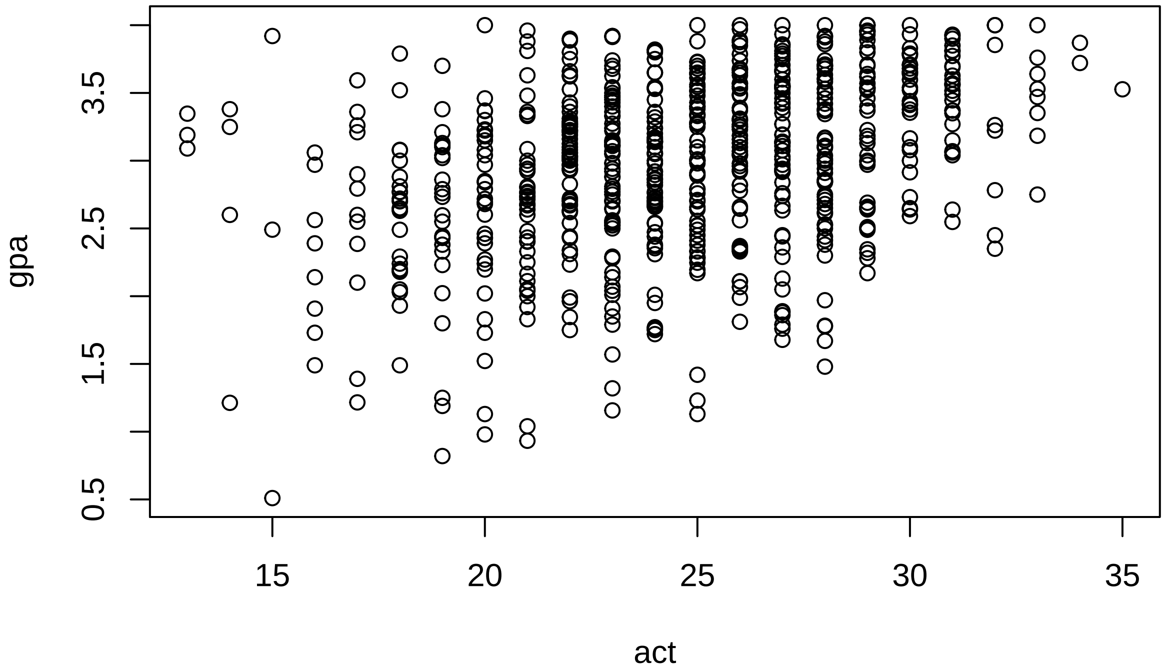
```
## [1] 3.31583
```

Here, the null is $\beta_2 = 0$ and $\beta_5 = 0$. So the alternative is not all these coefficients are 0. The rule is when the F-statistic is greater than $F(0.95, 2, 30)$, we accept the alternative. Since the F-statistic is 0.62 and thus smaller than $F(0.95, 2, 30) = 3.32$, we accept the null that we can drop X_2 and X_5.

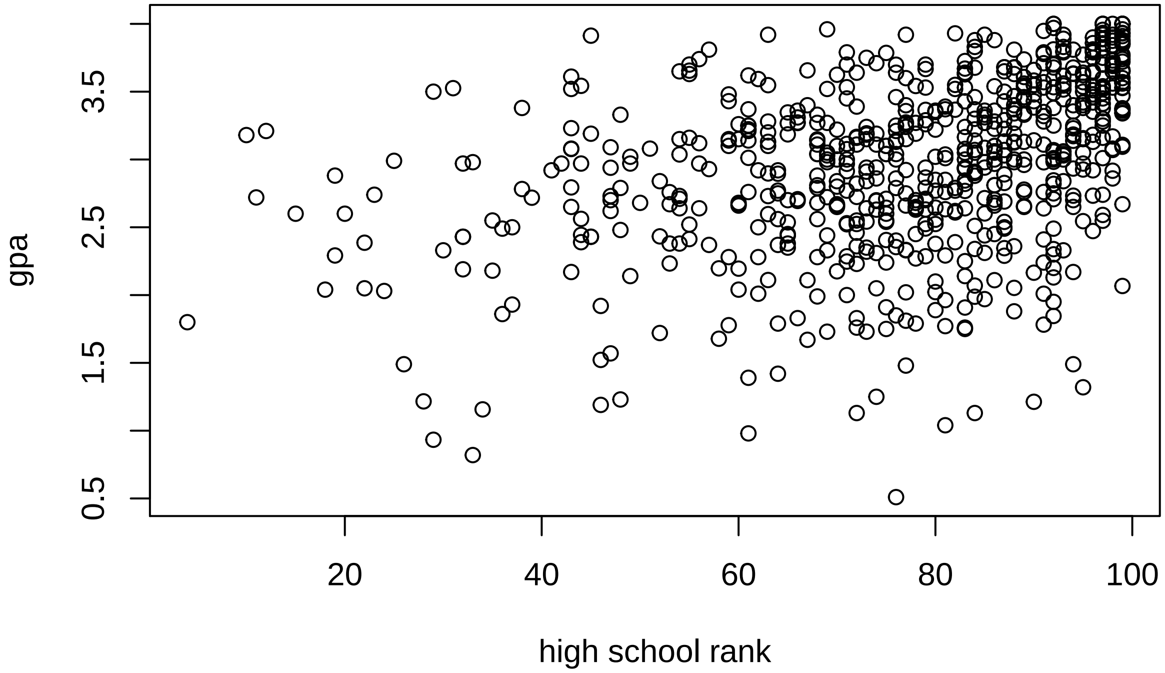
8.43

```
admission <- read.table("APPENC04.txt")
admission <- admission[, c(2, 3, 4, 5)]
names(admission) <- c("Y", "X1","X2","X3")
admission$X31 <- ifelse(admission$X3 == 1996, 1, 0)
admission$X32 <- ifelse(admission$X3 == 1997, 1, 0)
admission$X33 <- ifelse(admission$X3 == 1998, 1, 0)
admission$X34 <- ifelse(admission$X3 == 1999, 1, 0)
admission <- admission[, -4]

plot(admission$X2, admission$Y, ylab = "gpa", xlab = "act")
```



```
plot(admission$X1, admission$Y, ylab = "gpa", xlab = "high school rank")
```



```
set.seed(12)
ind_train <- sample(c(1:705), 600)
train <- admission[ind_train, ]
test <- admission[-ind_train, ]

lm_order1 <- lm(Y~X1+X2+X31+X32+X33+X34, train)
lm_order2 <- lm(Y~X1+X2+X31+X32+X33+X34+X1^2+X2^2+X1*X2, train)

pre1 <- predict(lm_order1, test[, -1])
```

```
pre2 <- predict(lm_order2, test[, -1])

err1 <- sum((pre1 - test$Y)^2) / nrow(test)
err2 <- sum((pre2 - test$Y)^2) / nrow(test)

err1
```

```
## [1] 0.3512574
```

```
err2
```

```
## [1] 0.3421515
```

We first treat the “year” variable as a categorical variable and introduce 4 indicator variables instead. And we are only interested in comparing the first order regression and the second order regression here. That’s to say, we’d like to test whether or not the coefficients of the second order terms involing only quantitative variables are all 0. Then, we split the data set into training set and test set which consist of 700 and 105 observations seperately. According to the result, the difference in test error of the first order model and second order model are quite small. And we thus tend to accept the null. To be more rigorous, we do the following F-test with $\alpha = 0.01$.

```
lm_1st <- lm(Y~X1+X2+X31+X32+X33+X34, admission)
lm_2nd <- lm(Y~X1+X2+X31+X32+X33+X34+X1^2+X2^2+X1*X2, admission)

#anova(lm_1st)
# sse 224.742
#anova(lm_2nd)
# sse 219.753

F_star <- ((224.742 - 219.753) / 3) / (219.753 / 695)

F_star
```

```
## [1] 5.259473
```

```
qf(0.99, 3, 695)
```

```
## [1] 3.809898
```

On the contrary, according to the above F-test, we conclude that not all coefficients of the second term are 0. Thus, we finally choose model 2 as our model. To use the model, we could simply plug in new students’ data and take the predicted gpa as a reference. As for its predictbility, it has been shown before that it has a test error of 0.34 which means we should take more account into consideration when making the admission decision.

10.9

a.

```
brand <- read.table("CH06PR05.txt")
names(brand) <- c("Y", "X1", "X2")

X <- matrix(nrow = 16, ncol = 3)
X[, 1] <- 1
X[, 2] <- brand[, 2]
X[, 3] <- brand[, 3]
```

```r
H <- X %*% solve(t(X) %*% X) %*% t(X)
h <- diag(H)

lm_brand <- lm(Y~X1+X2, brand)
e <- as.numeric(resid(lm_brand))
#anova(lm_brand) # obtain sse
sse <- 94.30

stu_del_t <- c()
for(i in c(1:16)) {
  # by the algebraic result, page 396
  stu_del_t[i] <- e[i] * sqrt(13 / (sse * (1-h[i]) - e[i]^2))
}
stu_del_t
```

```
##  [1] -0.04252322  0.06379037 -1.41615612  1.44262038 -0.38192921
##  [6] -0.69205630 -0.79848708  0.52521749  0.48405702 -0.62904090
## [11]  1.89745968  1.01777124 -1.18620007 -2.18858703  1.55056231
## [16]  0.25576262
```

```r
bonfrroni_t <- qt(1 - 0.1/(2*16), 13)
abs(stu_del_t) > bonfrroni_t
```

```
##  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [12] FALSE FALSE FALSE FALSE FALSE
```

The studentized deleted residuals are provided above. The rule is when the absolute studentized deleted residuals are greater than $t(1 - 0.1/(2 * 16), 13)$, we identify it as an outlying Y. And by the result, we conclude that there is no outlier w.r.t Y.

b.

```r
h
```

```
##  [1] 0.2375 0.2375 0.2375 0.2375 0.1375 0.1375 0.1375 0.1375 0.1375 0.1375
## [11] 0.1375 0.1375 0.2375 0.2375 0.2375 0.2375
```

The diagnal elements only take value from 0.2375 and 0.1375. It might be because the pattern of the predictors which behaves quite regularly.

c.

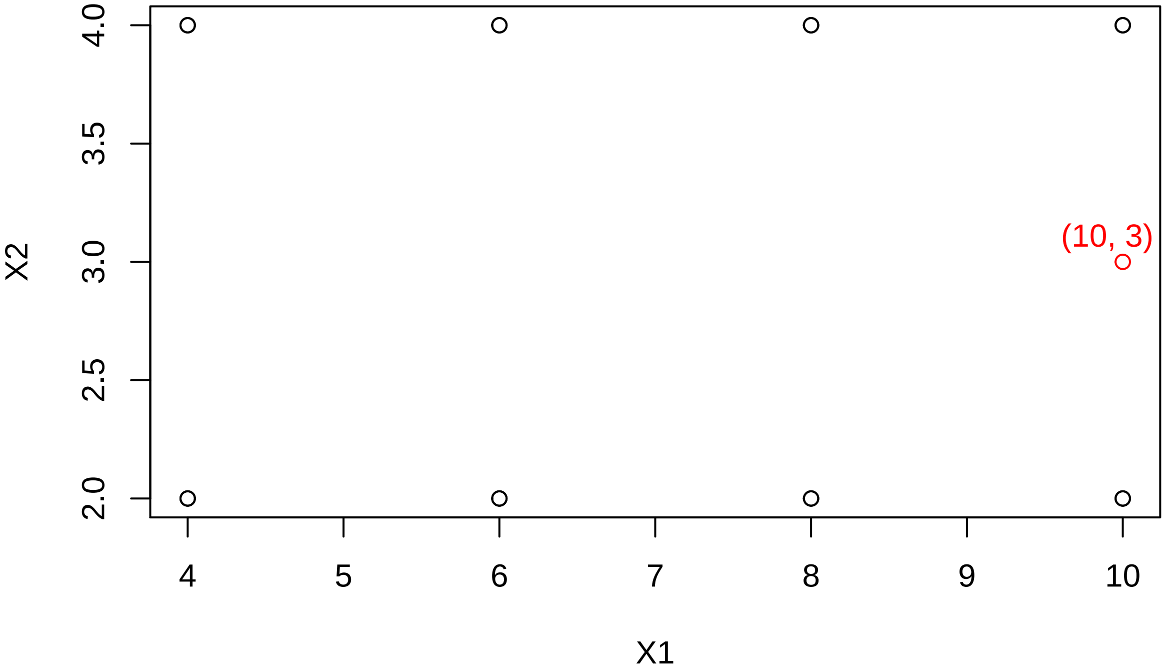
```r
h > 3 / 16 * 2
```

```
##  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [12] FALSE FALSE FALSE FALSE FALSE
```

According to the thumb of rule, when the diagnal is greater than $2 * \frac{p}{n} = 0.375$, we take the corresponding observation as an outlier w.r.t. predictors. Unfortunately, there is no high leverage observation here.

d.

```
plot(brand$X1, brand$X2, xlab = "X1", ylab = "X2")
points(10, 3, col = "red")
text(9.9, 3.1, "(10, 3)", col = "red")
```



```
h_new <- t(c(1, 10, 3)) %*% solve(t(X) %*% X) %*% c(1, 10, 3)
```

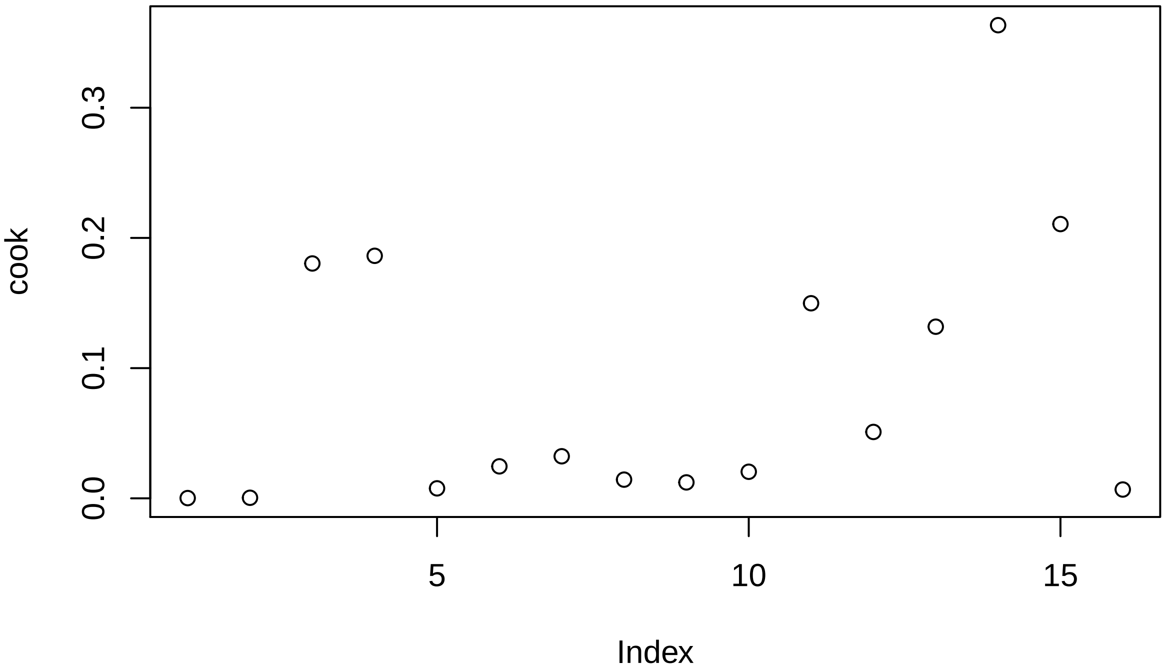
Visually, the prediction doesn't seem to involve an extrapolation beyond the range of data. Further, by calculation, the leverage of the new data is 0.175 which is in the range of the leverage of the data set and thus indicates no extrapolation is involved. Therefore, the coclusions agree with each other.

g.

```
cook <- c()
for(i in c(1:16)) {
  cook[i] <- e[i]^2 / (3 * sse / 13) * (h[i] / (1-h[i])^2)
}
cook
```

```
##  [1] 0.0001877130 0.0004223542 0.1803921815 0.1862582123 0.0076655286
##  [6] 0.0245466787 0.0322971439 0.0143542862 0.0122308711 0.0204060192
## [11] 0.1498281704 0.0509831969 0.1318214458 0.3634123447 0.2106609008
## [16] 0.0067576676
```

```
plot(cook)
```



```
pf(cook[14], 3, 13)
```

```
## [1] 0.2194682
```

According to the plot, we find the 14th observation has the largest Cook's distance and its corresponding F quantile is 0.21 which means the observation has a mild impact over the regression fit. The extent of the influence, however, may not be large enough to call for further remedial measures.

5.

Let $X_1^T, X_2^T, ..., X_n^T$ be the n rows of the design matrix or predictors of the n observations. Thus, finding the least square estimates with and without i_{th} observation is corresponding to the following two optimization problems.

$\hat{\beta} = argmin_{\beta}\Sigma_{j=1}^{n}(y_j - X_j^T * \beta)^2$

$\hat{\beta_{(i)}} = argmin_{\beta}\Sigma_{j\neq i}(y_j - X_j^T * \beta)^2$

Since the target functions are both polynomial functions w.r.t β, their first order partial derivatives should be 0 when they reach the minimum.

$\Sigma_{j=1}^{n} 2 * X_j * (y_j - X_j^T * \hat{\beta}) = 0$

$\Sigma_{j\neq i} 2 * X_j * (y_j - X_j^T * \hat{\beta_{(i)}}) = 0$

Comparing the two equations and undr the condition that $\hat{\beta} = \hat{\beta_{(i)}}$, we have $2 * X_i * (y_i - X_i^T * \hat{\beta}) = 0$ or $y_i = X_i^T * \hat{\beta}$ assuming X_i is not a 0 vector. Thus, the i_{th} observation is on the regression line.

On the onther hand, when i_{th} observation is on the regression line, the i_{th} term is 0 in the first equation above. Then, we have

$\Sigma_{j\neq i} 2 * X_j * (y_j - X_j^T * \hat{\beta}) = 0$

$\Sigma_{j\neq i} 2 * X_j * (y_j - X_j^T * \hat{\beta_{(i)}}) = 0$

These two systems of equations in terms of β are clearly the same, both of which have the same solutions

$\hat{\beta} = \hat{\beta_{(i)}} = (\Sigma_{j \neq i} X_j * X_j^T)^{-1} * (\Sigma_{j \neq i} X_j * y_j)$.

Therefore, we proved both sufficiency and necessity.