HW3

Jingshi

2/8/2017

Question 4

a)

The training RSS is expected to be lower for the cubic regression than it is for the linear regression. This is due to the Bias-variance tradeoff. Although the true relationship is linear, as the model complexity increase, the training RSS is expected to decrease.

b)

The test RSS is expected to be lower for the linear regression than it is for the cubic regression. The cubic regression fit training data better, but worse for the test data, since it is biased (the true relationship is linear).

c)

The training RSS is expected to be lower for the cubic regression than it is for the linear regression. This is due to the Bias-variance tradeoff. Although the true relationship is not linear, as the model complexity increase, the training RSS is expected to decrease.

d)

Since we don't know the true relationship between X and Y (except for non-linear), we don't know whether it is quadratic, cubic, or any other polynomials, we cannot make a conclusion about RSS of test data set.

Question 9

c)

```r
library(ISLR)
auto<-lm(mpg~.-name, data=Auto)
summary(auto)
```

```
##
## Call:
## lm(formula = mpg ~ . - name, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5903 -2.1565 -0.1169  1.8690 13.0604
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
## cylinders     -0.493376   0.323282  -1.526  0.12780
## displacement   0.019896   0.007515   2.647  0.00844 **
## horsepower    -0.016951   0.013787  -1.230  0.21963
## weight        -0.006474   0.000652  -9.929  < 2e-16 ***
## acceleration   0.080576   0.098845   0.815  0.41548
## year           0.750773   0.050973  14.729  < 2e-16 ***
## origin         1.426141   0.278136   5.127 4.67e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.328 on 384 degrees of freedom
## Multiple R-squared:  0.8215, Adjusted R-squared:  0.8182
## F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

i)

According to the output, the F-statistic (252.4) is high and the p-value (< 2.2e-16) is much less than any level of α. Therefore, the relationship between predictors and response is statistically significant. Moreover, the adjusted R-squared (0.8182) is high, so the relationship is strong.

ii)

The predictors that are statistically significant are displacement, weight, year and origin due to their p-values are much lower than $\alpha = 0.05$.

iii)

When model year increase by 1 year, mpg is expected to incerase by 0.750773, keeping other variables constant.

e)

```r
interaction1<-lm(mpg~displacement*weight, data=Auto)
summary(interaction1)
```

```
##
## Call:
## lm(formula = mpg ~ displacement * weight, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.8664  -2.4801  -0.3355   1.8071  17.9429
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          5.372e+01  1.940e+00  27.697  < 2e-16 ***
## displacement        -7.831e-02  1.131e-02  -6.922 1.85e-11 ***
## weight              -8.931e-03  8.474e-04 -10.539  < 2e-16 ***
## displacement:weight  1.744e-05  2.789e-06   6.253 1.06e-09 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.097 on 388 degrees of freedom
## Multiple R-squared:  0.7265, Adjusted R-squared:  0.7244
## F-statistic: 343.6 on 3 and 388 DF,  p-value: < 2.2e-16
```

The interaction between displacement and weight appears to be significant because the p-value (1.06e-09) of the interaction term is much less than $\alpha = 0.05$. Additionally, the above model is statistically significant because the F-statistic (343.6) is very high and the p-value ($< 2.2e-16$) is much lower than $\alpha = 0.05$.

```r
interaction2<-lm(mpg~year*displacement, data=Auto)
summary(interaction2)
```

```
##
## Call:
## lm(formula = mpg ~ year * displacement, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.8530  -2.4250  -0.2234   2.0823  16.9933
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -7.288e+01  8.368e+00  -8.709  < 2e-16 ***
## year               1.408e+00  1.102e-01  12.779  < 2e-16 ***
## displacement       2.523e-01  4.059e-02   6.216 1.32e-09 ***
## year:displacement -4.080e-03  5.453e-04  -7.482 4.96e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.729 on 388 degrees of freedom
## Multiple R-squared:  0.7735, Adjusted R-squared:  0.7718
## F-statistic: 441.7 on 3 and 388 DF,  p-value: < 2.2e-16
```

The interaction between year and displacement appears to be significant because the p-value (4.96e-13) of the interaction term is much less than $\alpha = 0.05$. Additionally, the above model is statistically significant because the F-statistic (441.7) is very high and the p-value ($< 2.2e-16$) is much lower than $\alpha = 0.05$.

```r
interaction3<-lm(mpg~weight*year, data=Auto)
summary(interaction3)
```

```
##
## Call:
## lm(formula = mpg ~ weight * year, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0397 -1.9956 -0.0983  1.6525 12.9896
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.105e+02  1.295e+01  -8.531 3.30e-16 ***
## weight       2.755e-02  4.413e-03   6.242 1.14e-09 ***
## year         2.040e+00  1.718e-01  11.876  < 2e-16 ***
## weight:year -4.579e-04  5.907e-05  -7.752 8.02e-14 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.193 on 388 degrees of freedom
## Multiple R-squared:  0.8339, Adjusted R-squared:  0.8326
## F-statistic: 649.3 on 3 and 388 DF,  p-value: < 2.2e-16
```

The interaction between weight and year appears to be significant because the p-value (8.02e-14) of the interaction term is much less than $\alpha = 0.05$. Additionally, the above model is statistically significant because the F-statistic (649.3) is very high and the p-value ($<$ 2.2e-16) is much lower than $\alpha = 0.05$.

```r
interaction4<-lm(mpg~weight*origin, data=Auto)
summary(interaction4)
```

```
##
## Call:
## lm(formula = mpg ~ weight * origin, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.4126  -2.8476  -0.4004   2.1815  15.5139
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   38.8991363  2.2031615  17.656  < 2e-16 ***
## weight        -0.0055411  0.0007845  -7.064 7.56e-12 ***
## origin         4.1312744  1.4980510   2.758  0.00609 **
## weight:origin -0.0012729  0.0006248  -2.037  0.04230 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.255 on 388 degrees of freedom
## Multiple R-squared:  0.7051, Adjusted R-squared:  0.7028
## F-statistic: 309.3 on 3 and 388 DF,  p-value: < 2.2e-16
```

The interaction between weight and origin appears to be significant because the p-value (0.04230) of the interaction term is less than $\alpha = 0.05$. Additionally, the above model is statistically significant because the F-statistic (309.3) is high and the p-value ($<$ 2.2e-16) is much lower than $\alpha = 0.05$.

f)

```r
transformation1<-lm(mpg~I(displacement^2) + log(weight)+log(year) +origin, data = Auto)
summary(transformation1)
```

```
##
## Call:
## lm(formula = mpg ~ I(displacement^2) + log(weight) + log(year) +
##     origin, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.8384 -1.8241 -0.0329  1.6253 12.8660
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)       -7.008e+01  1.712e+01  -4.094 5.17e-05 ***
## I(displacement^2)  2.202e-05  6.670e-06   3.301  0.00105 **
## log(weight)       -2.235e+01  1.187e+00 -18.825  < 2e-16 ***
## log(year)          6.217e+01  3.531e+00  17.609  < 2e-16 ***
## origin             7.777e-01  2.456e-01   3.166  0.00167 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.102 on 387 degrees of freedom
## Multiple R-squared:  0.8437, Adjusted R-squared:  0.8421
## F-statistic: 522.3 on 4 and 387 DF,  p-value: < 2.2e-16
```

The model is statistically significant because the F-stsatistic (522.3) is high and p-value (< 2.2e-16) for the F test is very low. All predictors in the model are statistically significant because their p-values are much lower than $\alpha = 0.05$.

```r
transformation2<-lm(mpg~I(displacement^2) + log(weight)+I(year^2) + origin, data = Auto)
summary(transformation2)
```

```
##
## Call:
## lm(formula = mpg ~ I(displacement^2) + log(weight) + I(year^2) +
##     origin, data = Auto)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.741 -1.837 -0.036  1.685 12.826
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.660e+02  9.247e+00  17.952  < 2e-16 ***
## I(displacement^2)  2.129e-05  6.584e-06   3.234  0.00133 **
## log(weight)       -2.215e+01  1.174e+00 -18.868  < 2e-16 ***
## I(year^2)          5.446e-03  3.024e-04  18.009  < 2e-16 ***
## origin             7.829e-01  2.432e-01   3.219  0.00139 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.07 on 387 degrees of freedom
## Multiple R-squared:  0.8468, Adjusted R-squared:  0.8453
## F-statistic: 534.9 on 4 and 387 DF,  p-value: < 2.2e-16
```

The model is statistically significant because the F-stsatistic (534.9) is high and p-value (< 2.2e-16) for the F test is very low. All predictors in the model are statistically significant because their p-values are much lower than $\alpha = 0.05$.

```r
transformation3<-lm(mpg~I(displacement^2) + log(weight)+sqrt(year) + origin, data = Auto)
summary(transformation3)
```

```
##
## Call:
## lm(formula = mpg ~ I(displacement^2) + log(weight) + sqrt(year) +
##     origin, data = Auto)
##
## Residuals:
##    Min     1Q Median     3Q    Max
```

```
## -9.812 -1.834 -0.051  1.633 12.854
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        7.391e+01  1.113e+01   6.640 1.06e-10 ***
## I(displacement^2)  2.185e-05  6.647e-06   3.287  0.00111 **
## log(weight)       -2.231e+01  1.184e+00 -18.840  < 2e-16 ***
## sqrt(year)         1.432e+01  8.083e-01  17.716  < 2e-16 ***
## origin             7.790e-01  2.450e-01   3.180  0.00159 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.093 on 387 degrees of freedom
## Multiple R-squared:  0.8445, Adjusted R-squared:  0.8429
## F-statistic: 525.6 on 4 and 387 DF,  p-value: < 2.2e-16
```

The model is statistically significant because the F-stsatistic (525.6) is high and p-value (< 2.2e-16) for the F test is very low. All predictors in the model are statistically significant because their p-values are much lower than $\alpha = 0.05$.

```r
transformation4<-lm(mpg~I(displacement^2) + sqrt(weight)+log(year) + origin, data = Auto)
summary(transformation4)
```

```
##
## Call:
## lm(formula = mpg ~ I(displacement^2) + sqrt(weight) + log(year) +
##     origin, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.8498 -1.9600  0.0498  1.6963 12.9031
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -2.011e+02  1.570e+01 -12.811  < 2e-16 ***
## I(displacement^2)  2.839e-05  7.257e-06   3.912 0.000108 ***
## sqrt(weight)      -8.401e-01  4.710e-02 -17.835  < 2e-16 ***
## log(year)          6.169e+01  3.619e+00  17.046  < 2e-16 ***
## origin             9.640e-01  2.494e-01   3.866 0.000130 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.18 on 387 degrees of freedom
## Multiple R-squared:  0.8357, Adjusted R-squared:  0.834
## F-statistic:   492 on 4 and 387 DF,  p-value: < 2.2e-16
```

The model is statistically significant because the F-stsatistic (492) is high and p-value (< 2.2e-16) for the F test is very low. All predictors in the model are statistically significant because their p-values are much lower than $\alpha = 0.05$.

Question 14

a)

```
set.seed(1)
x1=runif(100)
x2=0.5*x1+rnorm(100)/10
y=2+2*x1+0.3*x2+rnorm(100)
```

The form of the linear model:

$$y = \beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2 + \epsilon$$

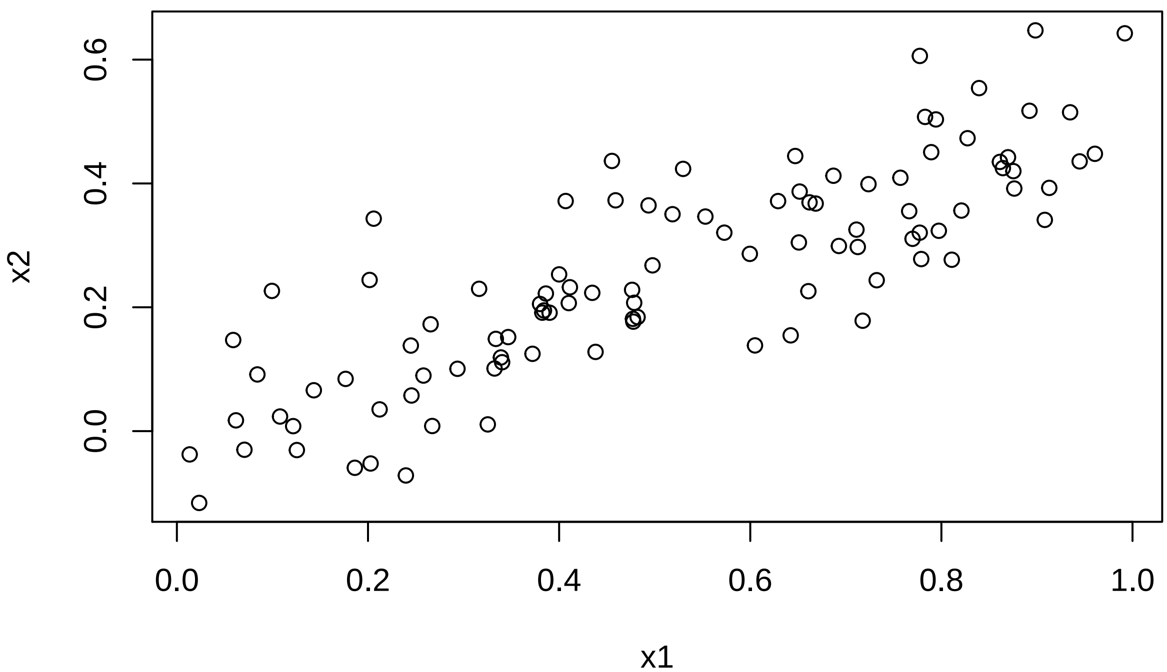
The regression coefficients are β_0 $(the\ intercept) = 2$, $\beta_1 = 2$ and $\beta_2 = 0.3$.

b)

```
cor (x1,x2)
```

```
## [1] 0.8351212
```

```
plot(x1,x2)
```



As the output shows, the correlation between x1 and x2 is 0.8351212 which is high. The scatterplot shows that there is a strong positive linear relationship between x1 and x2.

c)

```
model1<-lm(y~x1+x2)
summary(model1)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8311 -0.7273 -0.0537  0.6338  2.3359
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1305     0.2319   9.188 7.61e-15 ***
## x1            1.4396     0.7212   1.996   0.0487 *
## x2            1.0097     1.1337   0.891   0.3754
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.056 on 97 degrees of freedom
## Multiple R-squared:  0.2088, Adjusted R-squared:  0.1925
## F-statistic:  12.8 on 2 and 97 DF,  p-value: 1.164e-05
```

According to the output, $\hat{\beta_0} = 2.1305$, $\hat{\beta_1} = 1.4396$, $\hat{\beta_2} = 1.0097$. $\hat{\beta_0}$ is slightly larger than β_0. $\hat{\beta_1}$ is less than β_1. $\hat{\beta_2}$ is larger than β_2. The model does not fit quite well because the adjusted R-squared (0.1925) is small.

The null hypothesis $H_0 : \beta_1 = 0$ can be rejected because the p-value for x1 is 0.0487 which is less than $\alpha = 0.5$.

The null hypothesis $H_0 : \beta_2 = 0$ cannot be rejected because the p-value for x2 is 0.3754 which is larger than $\alpha = 0.5$.

d)

```r
model2<-lm(y~x1)
summary(model2)
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.89495 -0.66874 -0.07785  0.59221  2.45560
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1124     0.2307   9.155 8.27e-15 ***
## x1            1.9759     0.3963   4.986 2.66e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.055 on 98 degrees of freedom
## Multiple R-squared:  0.2024, Adjusted R-squared:  0.1942
## F-statistic: 24.86 on 1 and 98 DF,  p-value: 2.661e-06
```

According to the output, $\hat{\beta_0} = 2.1124$, $\hat{\beta_1} = 1.9759$. Compared with the previous modle in (c), this model ($\hat{y} = 2.1124 + 1.9759 \times x_1$) has a higher F-statistic and smaller p-value of F test, it has a higher adjusted

R-squred. Overall, it fits better.

The null hypothesis $H_0 : \beta_1 = 0$ can be rejected because the p-value for x1 is 2.66×10^{-6} which is much less than $\alpha = 0.5$.

e)

```
model3<-lm(y~x2)
summary(model3)
```

```
##
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.62687 -0.75156 -0.03598  0.72383  2.44890
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.3899     0.1949   12.26  < 2e-16 ***
## x2            2.8996     0.6330    4.58 1.37e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.072 on 98 degrees of freedom
## Multiple R-squared:  0.1763, Adjusted R-squared:  0.1679
## F-statistic: 20.98 on 1 and 98 DF,  p-value: 1.366e-05
```

As the output shows, the model is $\hat{y} = 2.3899 + 2.8996 \times x_1$. It fits worse than the model in (c) and (d) because its adjusted R-squared (0.1679) becomes smaller.

The null hypothesis $H_0 : \beta_1 = 0$ can be rejected because the p-value for x2 is 1.37×10^{-5} which is much less than $\alpha = 0.5$.

f)

No, they do not contradict with each other. The fact that x2 is a significant predictor in (e) and not significant in (c) is because x2 and x1 are highly correlated. So it is hard for the linear model in (c) to determine which predictor is truly associated with the response, y.

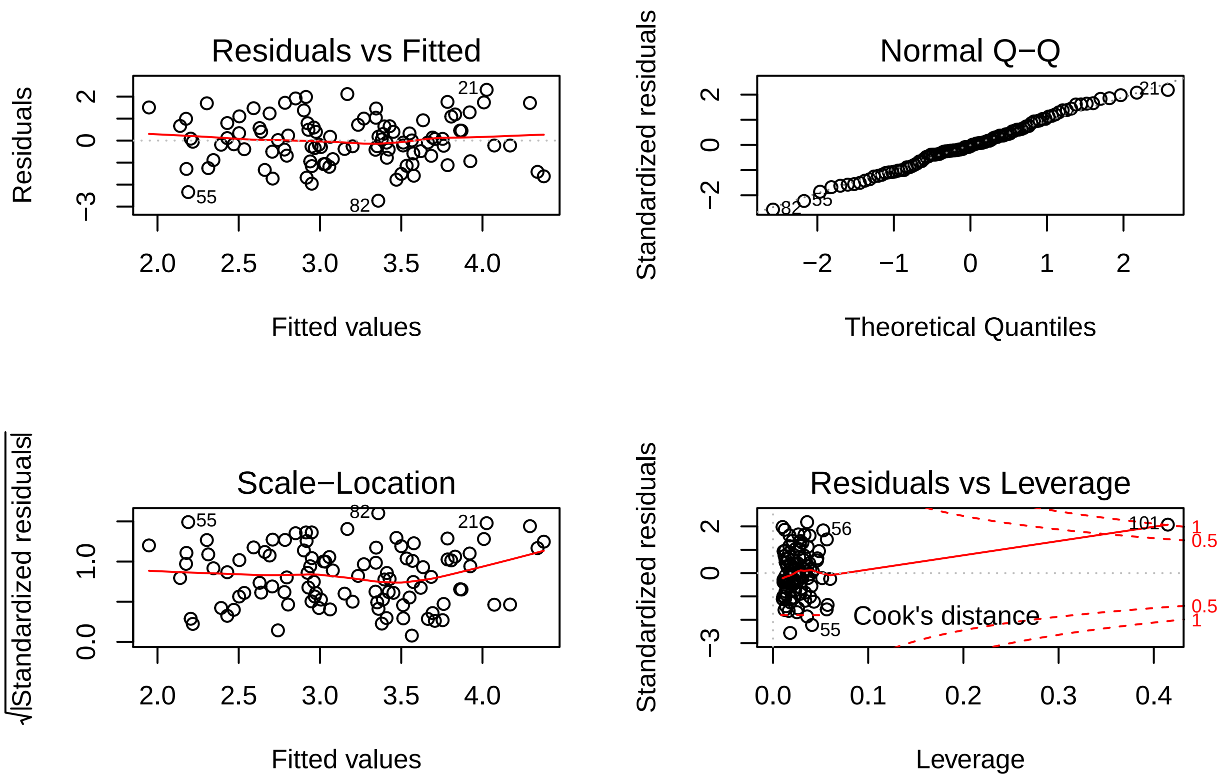
g)

```
x1=c(x1, 0.1)
x2=c(x2, 0.8)
y=c(y,6)
modelc<-lm(y~x1+x2)
modeld<-lm(y~x1)
modele<-lm(y~x2)
summary(modelc)
```

```
##
```

```
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.73348 -0.69318 -0.05263  0.66385  2.30619
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2267     0.2314   9.624 7.91e-16 ***
## x1            0.5394     0.5922   0.911  0.36458
## x2            2.5146     0.8977   2.801  0.00614 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.075 on 98 degrees of freedom
## Multiple R-squared:  0.2188, Adjusted R-squared:  0.2029
## F-statistic: 13.72 on 2 and 98 DF,  p-value: 5.564e-06
```

```r
par(mfrow=c(2,2))
plot(modelc)
```



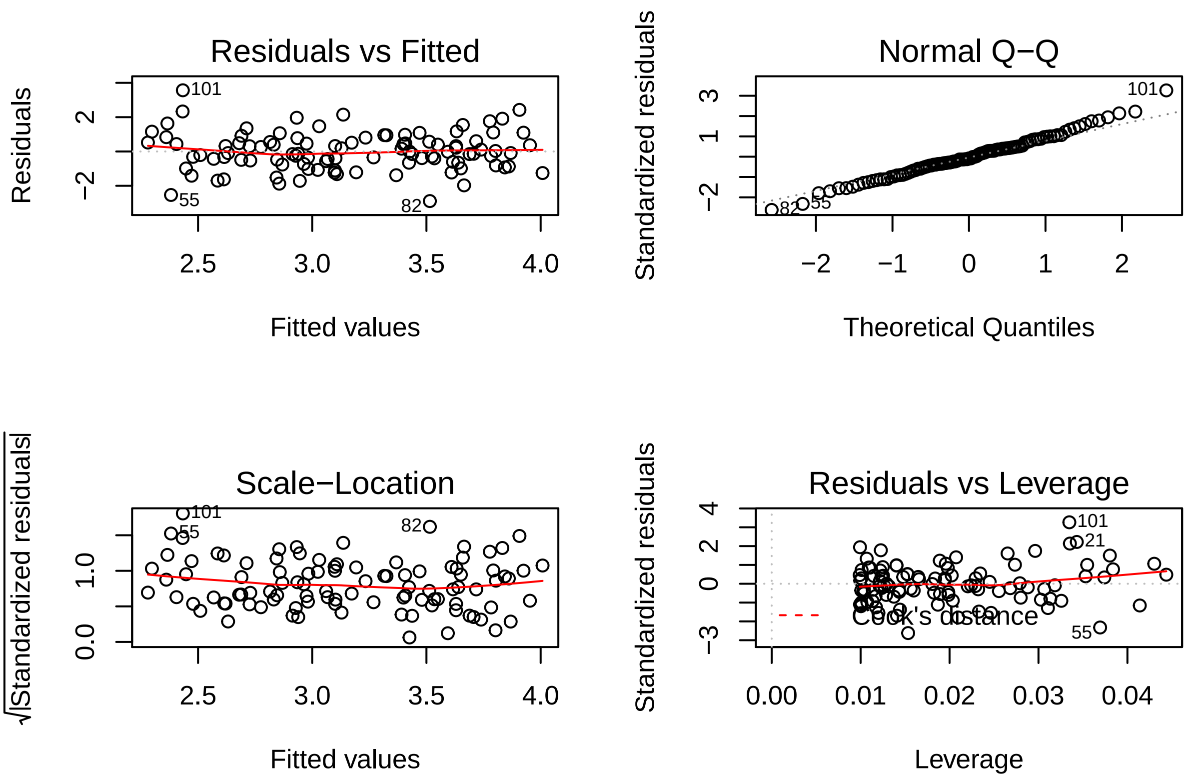
According to the output, when we fit the model in (c) with the new observation, $\hat{\beta}_0$ is about the same as it is in the modle in (c). $\hat{\beta}_1$ becomes much smaller and $\hat{\beta}_2$ becomes much larger. X1 becomes insignifiant since its p-value (0.36458) is much larger than $\alpha = 0.05$. X2 becomes significant since its p-value (0.00614) is much smaller than $\alpha = 0.05$.

The new observation doesn't appear to be an outlier, because its residual is within (-3, 3) in the residuals vs fitted plot. However, since it has cook's distance (about 1) and high leverage value (0.4, higher than 0.04(=4/n=4/101)), it is a high-leverage point.

```r
summary(modeld)
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8897 -0.6556 -0.0909  0.5682  3.5665
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2569     0.2390   9.445 1.78e-15 ***
## x1            1.7657     0.4124   4.282 4.29e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.111 on 99 degrees of freedom
## Multiple R-squared:  0.1562, Adjusted R-squared:  0.1477
## F-statistic: 18.33 on 1 and 99 DF,  p-value: 4.295e-05
```

```r
par(mfrow=c(2,2))
plot(modeld)
```



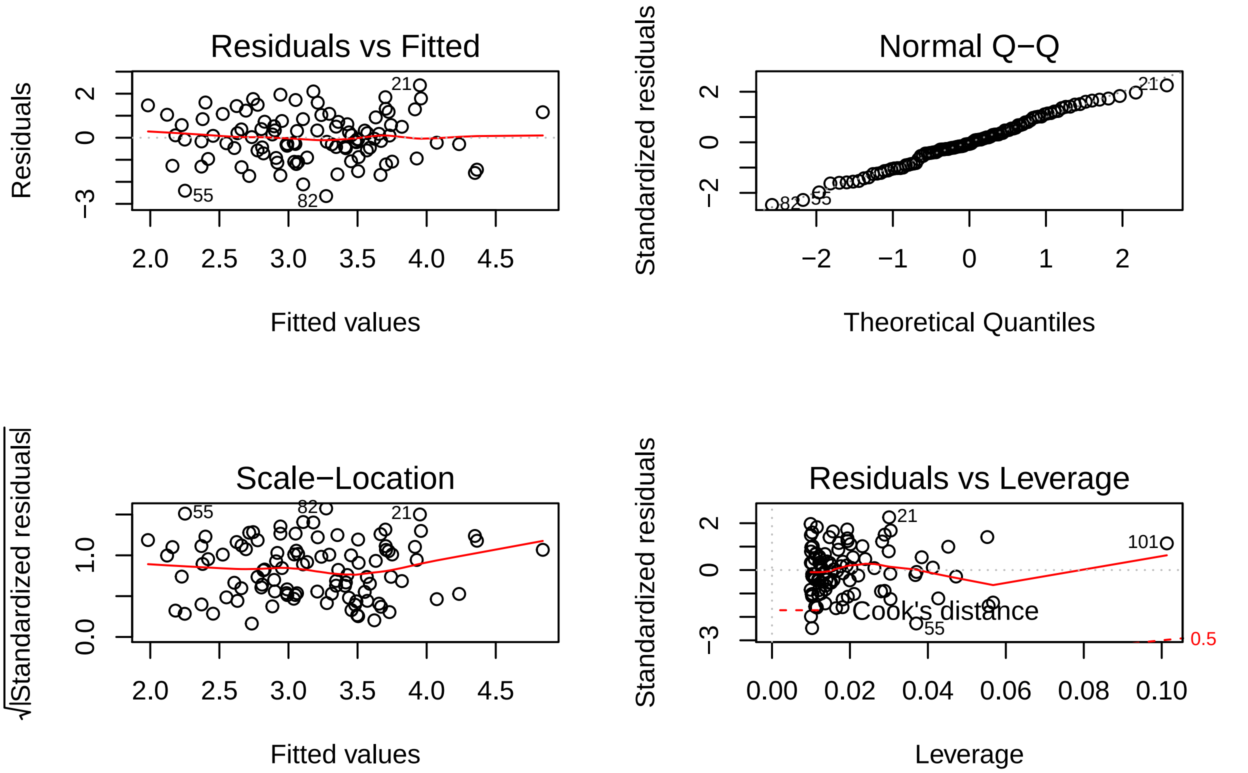
According to the output, when we fit the model in (d) with the new observation, the new model stays pretty much similar to the one in (d) but has a lower multiple R-squred and adjusted R-squared.

The new observation is an outlier, since the residula vs fitted plot shows that it has high residual (near 4). The new observation is not a high-leverage point, since its leverage value is less than 0.04 (= 4/n = 4/101).

```r
summary(modele)
```

```
##
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.64729 -0.71021 -0.06899  0.72699  2.38074
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.3451     0.1912  12.264  < 2e-16 ***
## x2            3.1190     0.6040   5.164 1.25e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.074 on 99 degrees of freedom
## Multiple R-squared:  0.2122, Adjusted R-squared:  0.2042
## F-statistic: 26.66 on 1 and 99 DF,  p-value: 1.253e-06
```

```r
par(mfrow=c(2,2))
plot(modele)
```



According to the output, when we fit the model in (e) with the new observation, the slop of x2 becomes slightly steeper than it is in the model in (e), the multiple R-squred and adjusted R-squred become lager which indicates a better fit.

The new observation is not an outlier, since its residual is within (-3,3) in the residual vs fitted plot. Its leverage value (about 0.1) is slightly high (higher than 0.04 (= 4/n = 4/101)) in the residuals vs leverage plot, so it is considered to be a high-leverage point.