Homework 4

JP Zamanillo

9/27/2020

1.

a.

```r
set.seed(1)
x1 <- runif(100)
x2 <- 0.5 * x1 + rnorm(100) / 10
y <- 2 + 2 * x2 + rnorm(100)
```

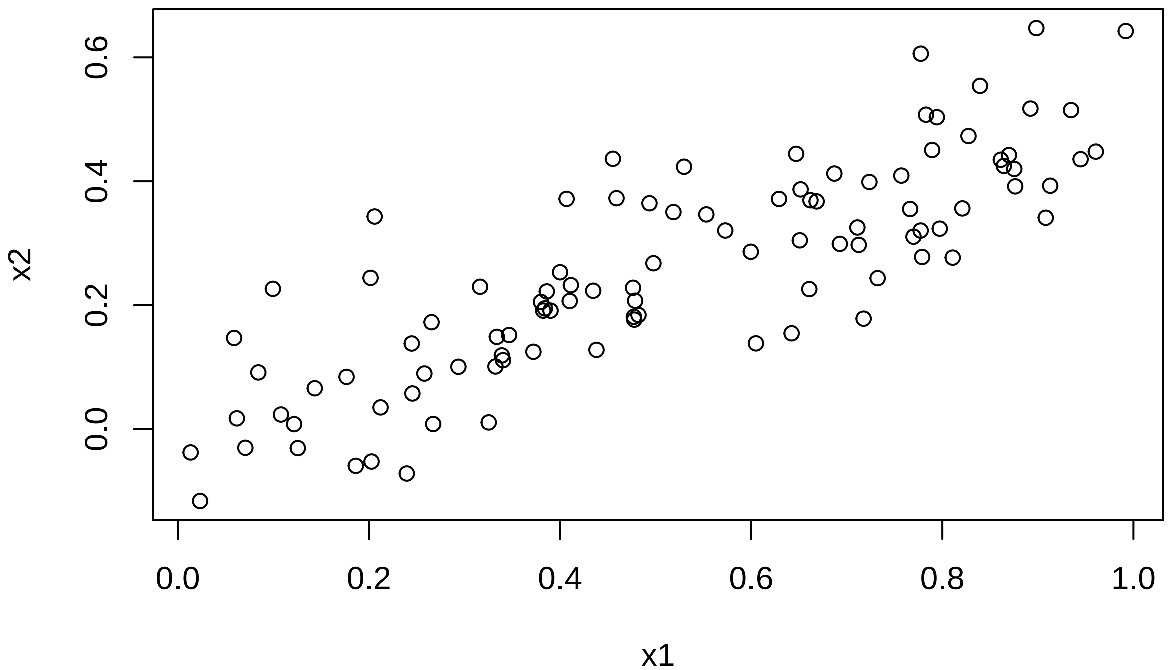
$Y = 2 + 2X_1 + 0.3X_2 + \epsilon$

b.

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

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

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



c.

```r
lm.fit <- lm(y ~ x1 + x2)
summary(lm.fit)
```

```
##
## 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           -0.5604     0.7212  -0.777   0.4390
## x2            2.7097     1.1337   2.390   0.0188 *
## ---
## 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:  0.0987, Adjusted R-squared:  0.08012
## F-statistic: 5.311 on 2 and 97 DF,  p-value: 0.006473
```

We can reject the null hypothesis that $\beta_2 = 0$ based on a p-value of 0.0188. However, with this current model, we fail to reject the null hypothesis that $\beta_1 = 0$ based on a p-value of 0.4390.

d.

```r
lm.fit.x1<- lm(y ~ x1)
summary(lm.fit.x1)
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.00241 -0.66755 -0.09282  0.71984  2.78124
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.0819     0.2365   8.804 4.75e-14 ***
## x1            0.8790     0.4061   2.164   0.0329 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.081 on 98 degrees of freedom
## Multiple R-squared:  0.04562,    Adjusted R-squared:  0.03588
## F-statistic: 4.685 on 1 and 98 DF,  p-value: 0.03286
```

We can reject the null hypothesis that $\beta_1 = 0$ based on a p-value of 0.0329 for a model that does not contain x2.

e.

```r
lm.fit.x2 <- lm(y ~ x2)
summary(lm.fit.x2)
```

```
##
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.91065 -0.65771 -0.06083  0.65167  2.47408
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.0295     0.1916  10.590  < 2e-16 ***
## x2            1.9739     0.6224   3.172  0.00202 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.054 on 98 degrees of freedom
## Multiple R-squared:  0.09309,    Adjusted R-squared:  0.08384
## F-statistic: 10.06 on 1 and 98 DF,  p-value: 0.002024
```

In this model, we can reject the null hypothesis that $\beta_1 = 0$ based on a p-value of 0.00202.

f.

No the results do not contradict each other because multicollinearity exists between `x1` and `x2`, making it difficult to distinguish their effects on `y`.

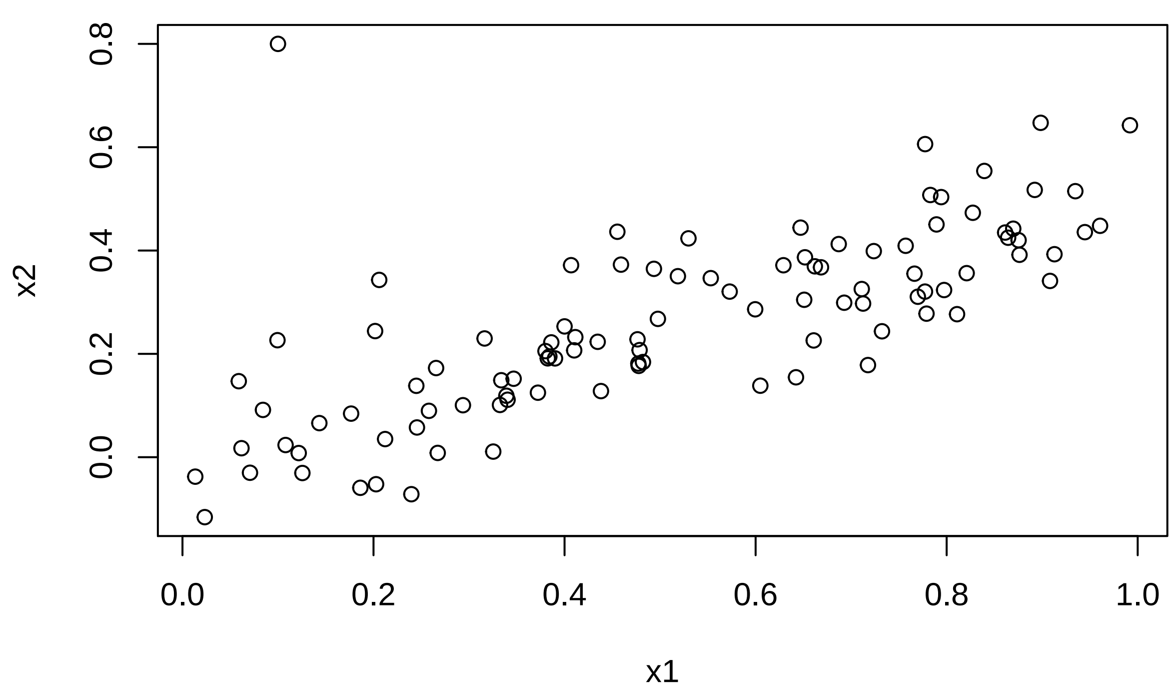
g.

```
x1 <- c(x1, 0.1)
x2 <- c(x2, 0.8)
y <- c(y, 6)
```

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

```
## [1] 0.7392279
```

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

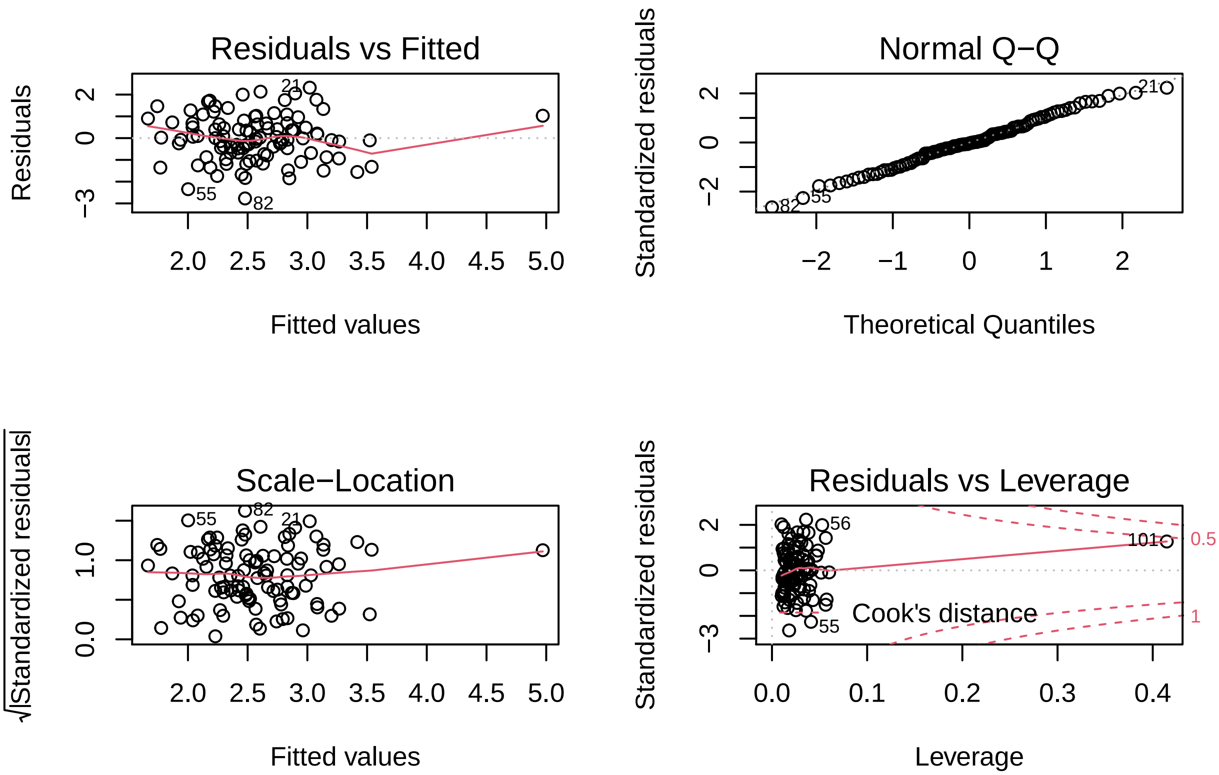


With the new observations, we see a lower correlation between `x1` and `x2`. It also looks like there's a serious outlier towards the top-left of the plot above, which corresponds to our new value.

```
lm.fit2 <- lm(y ~ x1 + x2)
summary(lm.fit2)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.77230 -0.68497 -0.03604  0.67478  2.31801
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1884     0.2281   9.595 9.18e-16 ***
## x1           -1.1027     0.5838  -1.889   0.0619 .
## x2            3.6163     0.8850   4.086 8.98e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.06 on 98 degrees of freedom
## Multiple R-squared:  0.1661, Adjusted R-squared:  0.1491
## F-statistic: 9.761 on 2 and 98 DF,  p-value: 0.0001363
```

```
par(mfrow = c(2, 2))
plot(lm.fit2)
```

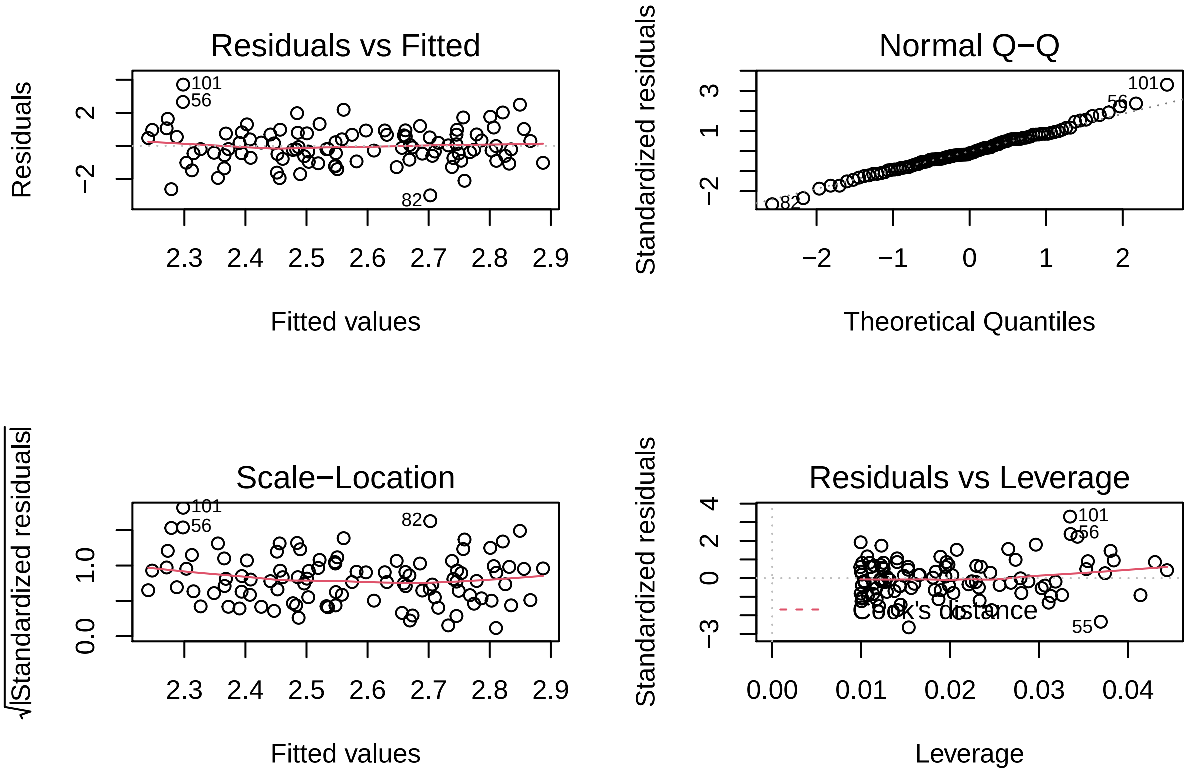


Our full model now shows that we fail to reject that $\beta_1 = 0$ with a p-value of 0.0619. We can still conclude that $\beta_2 \neq 0$ in this current model. The leverage plot suggests that our 101st observation, or our new value, acts as a high leverage point, but it does not exceed the 2 outlier threshold in regards to the Studentized residuals.

```
lm.fit3 <- lm(y ~ x1)
summary(lm.fit3)
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9970 -0.7260 -0.1236  0.6885  3.7020
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2319     0.2452   9.101 9.99e-15 ***
## x1            0.6608     0.4232   1.561    0.122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.14 on 99 degrees of freedom
## Multiple R-squared:  0.02403,    Adjusted R-squared:  0.01418
## F-statistic: 2.438 on 1 and 99 DF,  p-value: 0.1216
```

```r
par(mfrow = c(2, 2))
plot(lm.fit3)
```



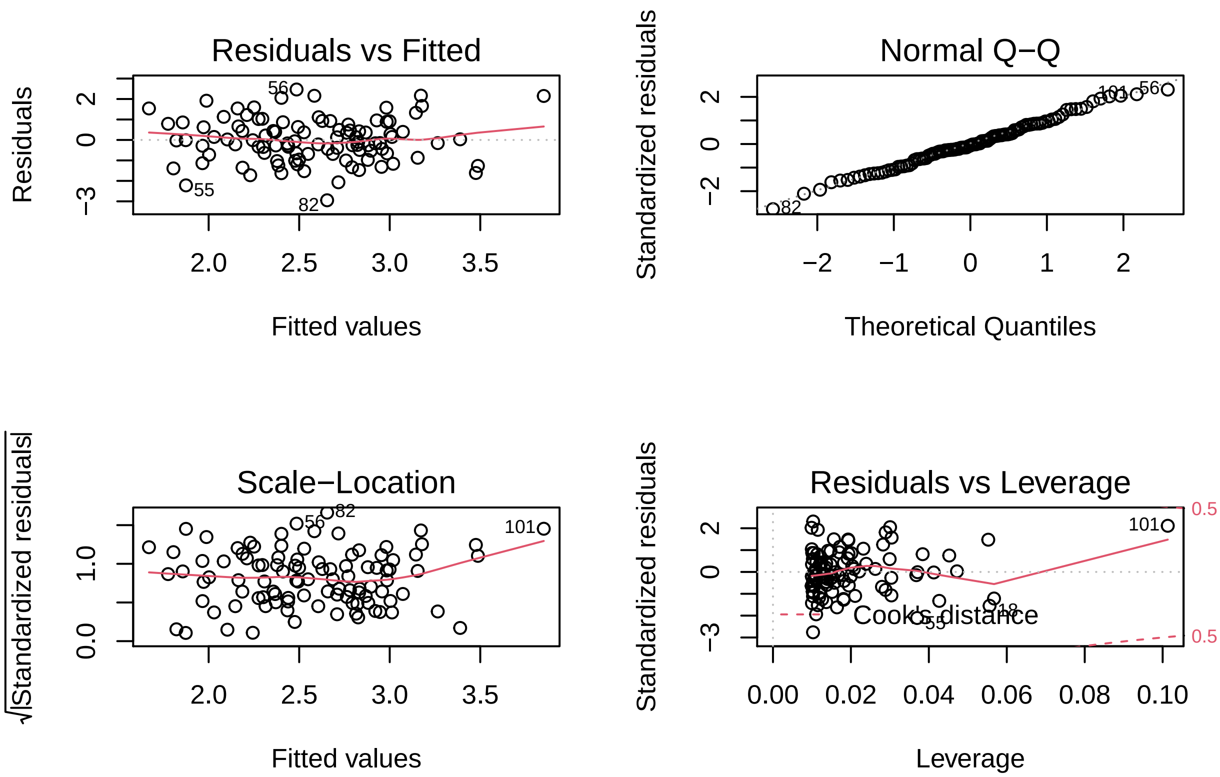
x1 alone no longer appears to have a relationship with y based on a p-value of 0.122. For this model, observation 101 acts as a both a serious outlier above a Studentized residual value of 2 and it also serves as a high leverage point according the leverage plot above.

```r
lm.fit4 <- lm(y ~ x2)
summary(lm.fit4)
```

```
##
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.94849 -0.68322 -0.06569  0.75209  2.46508
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.9464     0.1911  10.185  < 2e-16 ***
## x2            2.3806     0.6037   3.943  0.00015 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.073 on 99 degrees of freedom
```

```
## Multiple R-squared:  0.1358, Adjusted R-squared:  0.127
## F-statistic: 15.55 on 1 and 99 DF,  p-value: 0.00015
```

```r
par(mfrow = c(2, 2))
plot(lm.fit4)
```



x2 continues to show that a relationship exists with y based on a p-value of 0.00015. Just like the previous lm.fit3 model, observation 101 is both a high leverage point and outlier.

For lm.fit3, the slope of x1 is reduced compared to the previous iteration. lm.fit4 $beta_1$ estimate shows an increase of slope against y.

h.

Based on the outputs above:

- lm.fit2: 1.06
- lm.fit3: 1.14
- lm.fit4: 1.073

The full model, or lm.fit2, has the lowest standard error. This means that this model produces the most reliable estimates despite the lack of significance of x1.

i.

```r
library(car)
```

```r
vif(lm.fit)
```

```
##       x1       x2
## 3.304993 3.304993
```

```r
vif(lm.fit2)
```

```
##       x1       x2
## 2.204867 2.204867
```

As we see from the VIF calculations, our model with the outlier has less multicollinearity than our model without the outlier. Our model with the outlier performed better because the lower multiocollinearity amongst the predictors enabled us to better identify `x1` and `x2`'s effects on `y`.