HW4

Jon Griffith and Lauren Quesada

2025-04-07

10

```r
library(ISLR2)
```

```
## Warning: package 'ISLR2' was built under R version 4.4.3
```

(a)

```r
df <- Carseats

fit <- lm(Sales ~ Price + Urban + US, data=df)
summary(fit)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Urban + US, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9206 -1.6220 -0.0564  1.5786  7.0581
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
## Price       -0.054459   0.005242 -10.389  < 2e-16 ***
## UrbanYes    -0.021916   0.271650  -0.081    0.936
## USYes        1.200573   0.259042   4.635 4.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.472 on 396 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2335
## F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

(b)

The coefficient for the intercept, $\hat{\beta}_0 = 13.043$, can be interpreted as the predicted sales, on average, for a carseat that costs \$0, is not in an Urban environment, and is not in the US, is 13,043 carseats.

The coefficient for 'Price', $\hat{\beta}_1 = -0.054$, can be interpreted as for each dollar increase in Price, you can expect the number of carseats sold to go down by approximately 54 units, with the other variables held constant.

The coefficient for 'UrbanYes', $\hat{\beta}_2 = -0.022$, can be interpreted as stores located in urban areas sell, on average, approximately 22 fewer units than stores located in rural areas, with the other variables held constant.

However, since this term has a high p-value, meaning it is not statistically significant, we should discount this variable's effect on the response.

The coefficient for 'USYes', $\hat{\beta}_3 = 1.2$, can be interpreted as stores located in the US sell, on average, approximately 1200 more units than stores not located in the US, with the other variables held constant.

The 'Price' and 'US' variables are both statistically significant while the 'Urban' variable is not.

(c)

$Sales = 13.043 - 0.054(\text{Price}) - 0.022(I_{Urban='Yes'}) + 1.2(I_{US='Yes'})$

(d)

We can reject the null hypothesis for the 'Price' and 'US' variables, based on the significance codes provided in the model summary.

(e)

```r
fit2 <- lm(Sales ~ Price + US, data = df)
summary(fit2)
```

```
##
## Call:
## lm(formula = Sales ~ Price + US, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9269 -1.6286 -0.0574  1.5766  7.0515
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
## Price       -0.05448    0.00523 -10.416  < 2e-16 ***
## USYes        1.19964    0.25846   4.641 4.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.469 on 397 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2354
## F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-16
```

(f)

RSE: From model (a) to (e), the RSE decreased from 2.472 to 2.469.

R2: From model (a) to (e), the R2 remained the same. The adjusted-R2 however, increased from 0.2335 to 0.2354.

F-statistic: From model (a) to (e), the F-statistic increased from 41.52 to 62.43, and both models returned statistically significant F-statistics.

Based on these metrics, model (e) fits the data better than model (a), admittedly by a small margin. Neither suggest a particularly strong fit between the estimated linear model and the data.

Since the first model has an adjusted R-squared of 0.2335 while the second model has an adjusted R-squared of 0.2354, this may be the best metric to suggest a better fit for the second model with one less variable. This also means that each model accounts for about 23.54 percent of variation in the predictions for sales.

Overall, whether this is a good or bad fit is subjective, but it seems like a poor fit since the goal here is presumably to make business decisions based on predicted unit sales. If these predictions aren't very accurate, you would not want to make important decisions based on these models.

(g)

```
confint(fit2, level=0.95)
```

```
##                   2.5 %      97.5 %
## (Intercept) 11.79032020 14.27126531
## Price       -0.06475984 -0.04419543
## USYes        0.69151957  1.70776632
```

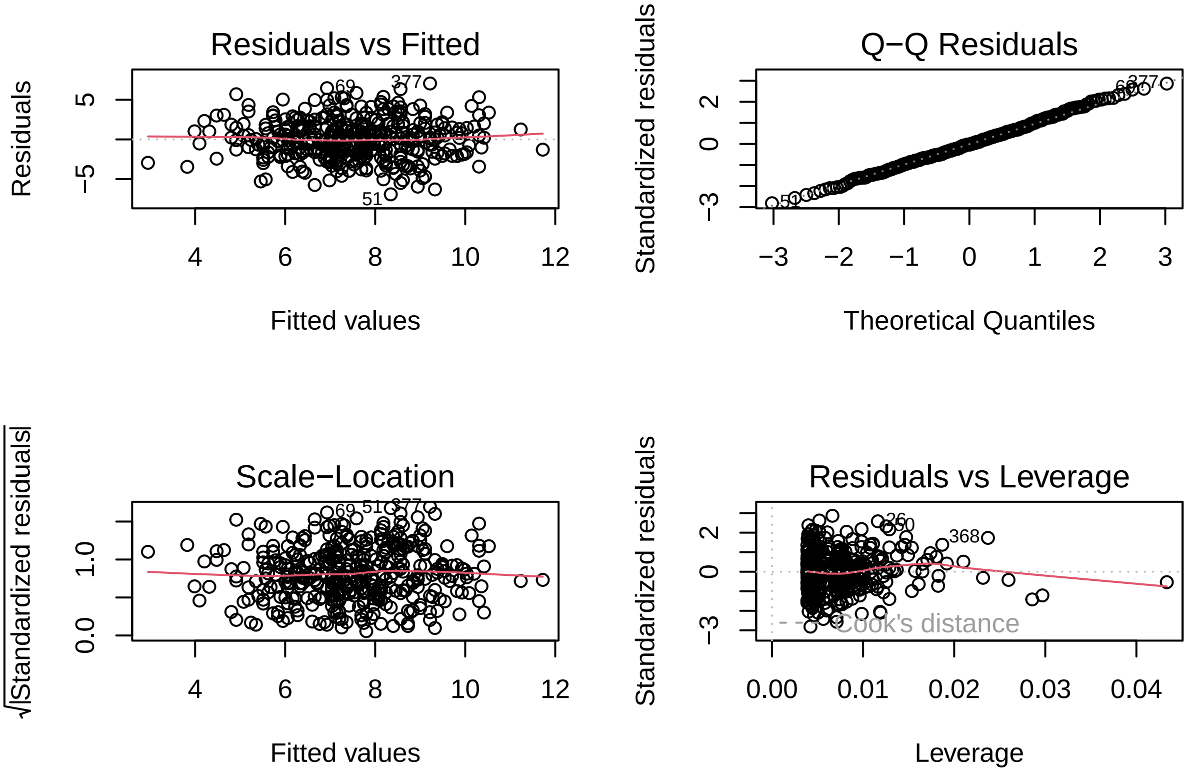
The 95% confidence interval for β_0, the intercept coefficient, is approx. $(11.79, 14.27)$.

The 95% confidence interval for β_1, the coefficient for the predictor Price, is approx. $(-0.06, -0.04)$.

The 95% confidence interval for β_2, the coefficient for the predictor US, is approx. $(0.69, 1.71)$.

(h)

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



```
#plot(fit2, which=5)
```

Observations 69, 377, and 51 are potential outliers, though they fall within the ± 3 studentized residuals rule of thumb. Further investigation should be taken with these three points. It also appears there may be a potential high leverage observation as identified in the Resids vs. Leverage plot, though it does not cross the threshold for Cook's distance so again, this would be a good point to investigate but doesn't immediately stand out as a problem point.

13

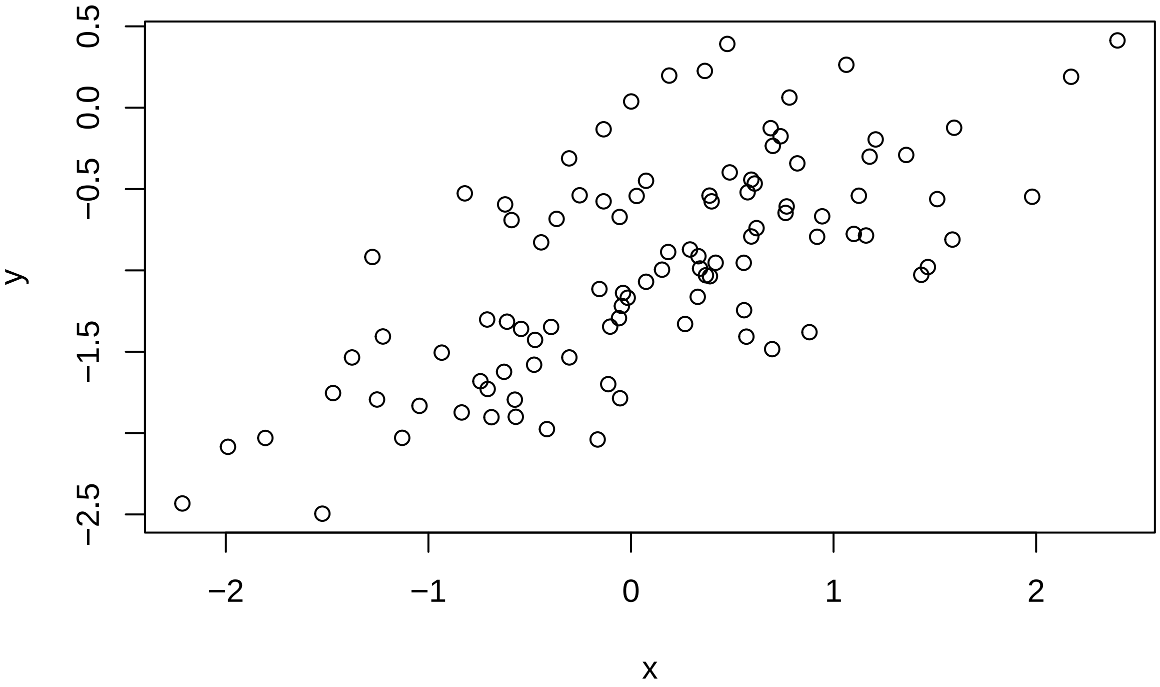
(a) - (c)

```
set.seed(1)

x <- rnorm(100)
eps <- rnorm(100, 0, sd=sqrt(0.25))
y <- -1 + 0.5*x + eps
```

Vector 'y' has a length of 100. For this model, $\beta_0 = -1$ and $\beta_1 = 0.5$.

(d)

```
plot(x, y)
```



We observe a positive linear relationship between x and y, which accurately reflects what we know the true function to equal.

(e)

```
fit_e <- lm(y ~ x)
summary(fit_e)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.93842 -0.30688 -0.06975  0.26970  1.17309
##
## Coefficients:
```

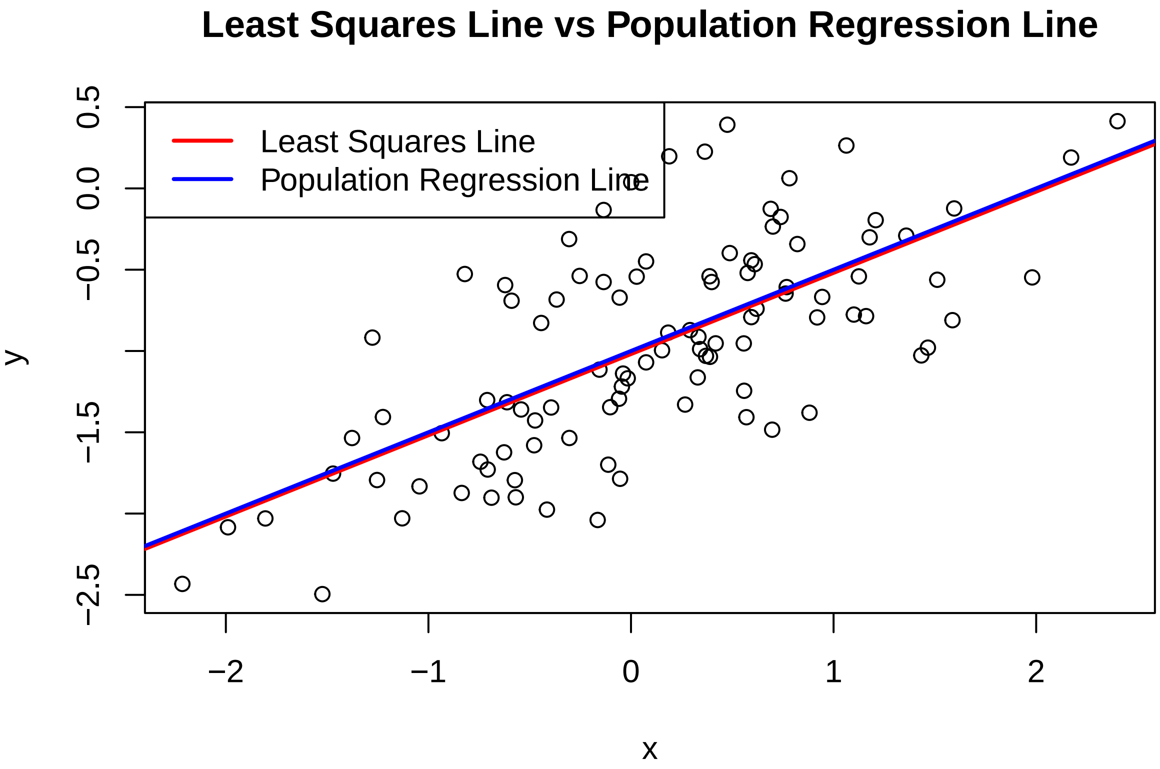
```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.01885    0.04849 -21.010  < 2e-16 ***
## x            0.49947    0.05386   9.273 4.58e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4814 on 98 degrees of freedom
## Multiple R-squared:  0.4674, Adjusted R-squared:  0.4619
## F-statistic: 85.99 on 1 and 98 DF,  p-value: 4.583e-15
```

The model has a statistically significant $\hat{\beta}_0 = -1.01885$ for the intercept and $\hat{\beta}_1 = 0.49947$ for 'x', which accurately reflects the true coefficients of $\beta_0 = -1$ and $\beta_1 = 0.5$. We also see that the R-squared is approximately 0.4674 which suggests that 'x' accounts for approximately 46.7 percent of the explained variance for the predictions of 'y'.

With the near zero p-value for x and an F-statistic of 38.64, this further suggests that x is a statistically significant variable.

(f)

```r
plot(x,y,
     main='Least Squares Line vs Population Regression Line')
abline(fit_e, col='red', lwd=2)
abline(a=-1, b=0.5, col='blue', lwd=2)
legend('topleft', c('Least Squares Line', 'Population Regression Line'), col=c('red', 'blue'), lwd=2)
```



(g)

```r
fit4 <- lm(y ~ x + I(x^2))
summary(fit4)
```

```
##
## Call:
## lm(formula = y ~ x + I(x^2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.98252 -0.31270 -0.06441  0.29014  1.13500
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.97164    0.05883 -16.517  < 2e-16 ***
## x            0.50858    0.05399   9.420  2.4e-15 ***
## I(x^2)      -0.05946    0.04238  -1.403    0.164
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.479 on 97 degrees of freedom
## Multiple R-squared:  0.4779, Adjusted R-squared:  0.4672
## F-statistic:  44.4 on 2 and 97 DF,  p-value: 2.038e-14
```

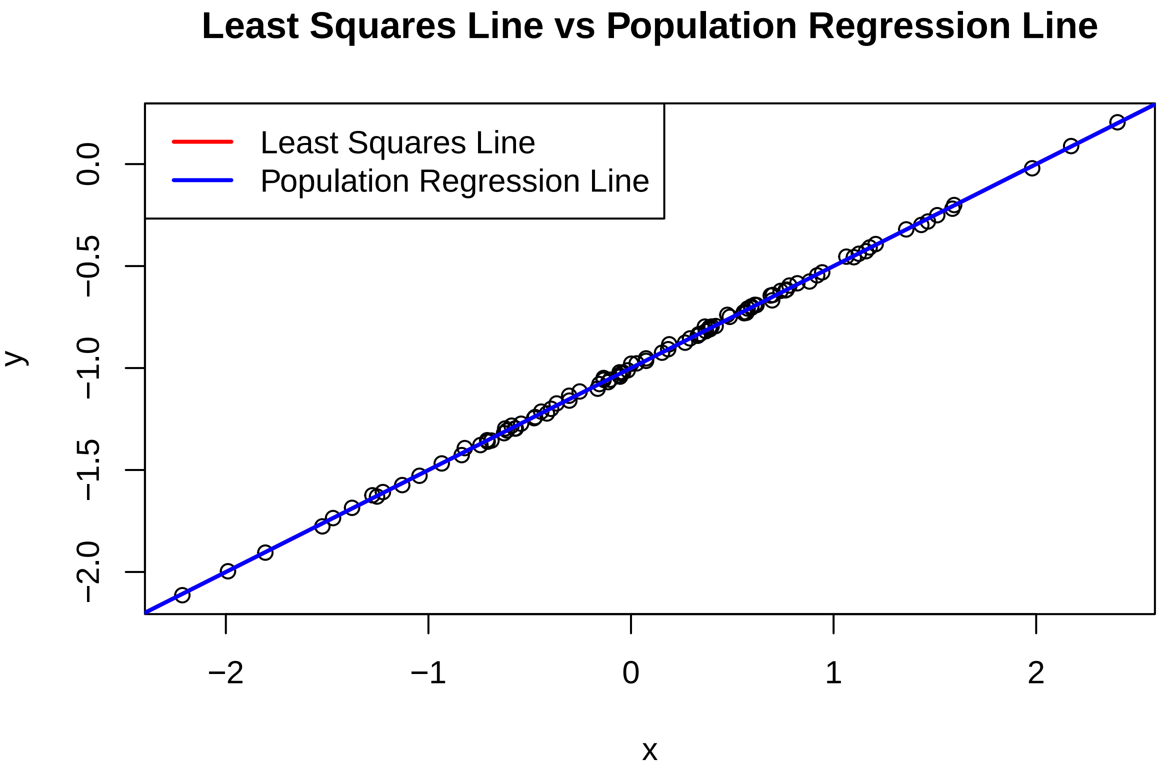
There is no evidence that the quadratic term improves the fit of the model. We conclude this based on a high p-value that doesn't beat any standard threshold, therefore we fail to reject the null hypothesis $H_0 : \beta_2 = 0$. We also only observe a marginal increase in the adjusted R-squared as further evidence that this does not do much to improve the fit. This makes sense since the true equation is not quadratic.

(h)

```
set.seed(1)

x <- rnorm(100)
eps <- rnorm(100, 0, sd=0.01)
y <- -1 + 0.5*x + eps

fit_h <- lm(y ~ x)

summary(fit_h)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.018768 -0.006138 -0.001395  0.005394  0.023462
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.0003769  0.0009699 -1031.5   <2e-16 ***
## x            0.4999894  0.0010773   464.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.009628 on 98 degrees of freedom
## Multiple R-squared:  0.9995, Adjusted R-squared:  0.9995
## F-statistic: 2.154e+05 on 1 and 98 DF,  p-value: < 2.2e-16
```

```
plot(x,y,
     main='Least Squares Line vs Population Regression Line')
abline(fit_h, col='red', lwd=2)
abline(a=-1, b=0.5, col='blue', lwd=2)
legend('topleft', c('Least Squares Line', 'Population Regression Line'), col=c('red', 'blue'), lwd=2)
```



We see that we have similar values for both $\hat{\beta}_0$ and $\hat{\beta}_1$ which very closely approximate their true values. The variation is now even more explained by this model with an increase in proportion comparable to the decrease in variance for the true function. From the plot, we see that the OLS line is an even better approximation of the population regression line than we saw in the previous model, which makes sense since the data points are tighter around the line due to lower random error.

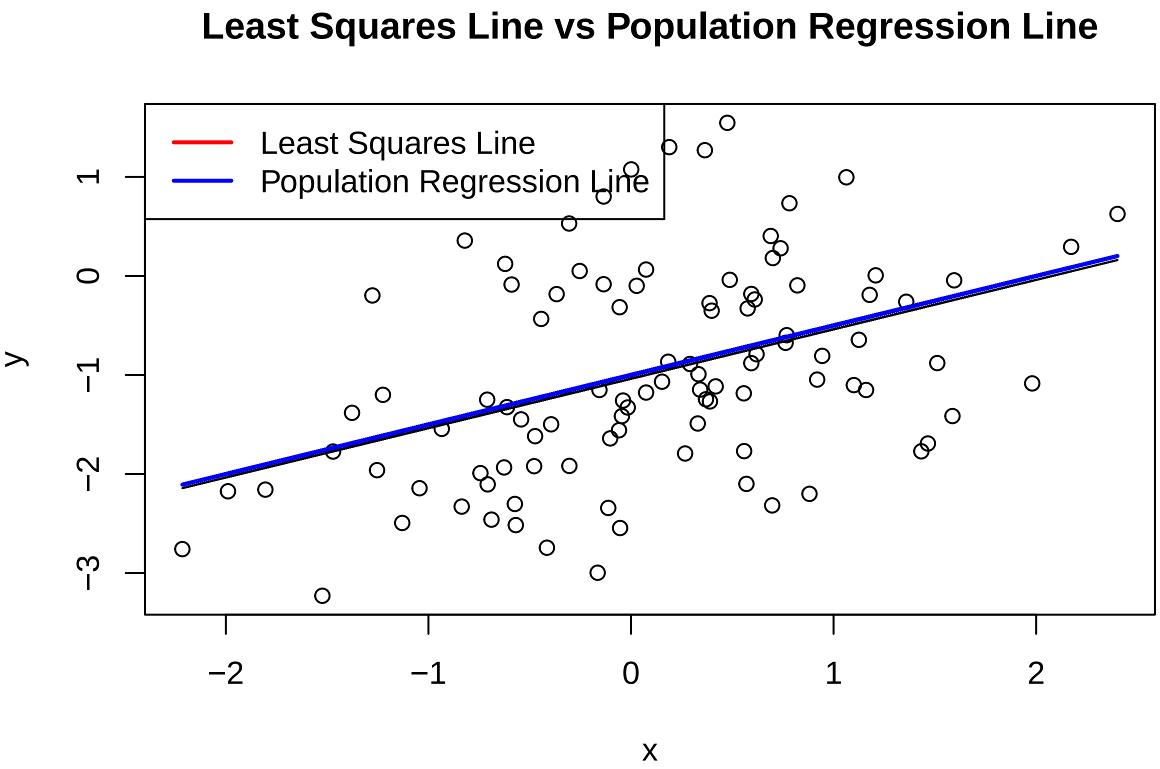
(i)

```
set.seed(1)

x <- rnorm(100)
eps <- rnorm(100, 0, 1)
y <- -1 + 0.5*x + eps

fit_i <- lm(y ~ x)

summary(fit_i)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -1.8768 -0.6138 -0.1395  0.5394  2.3462
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.03769    0.09699 -10.699  < 2e-16 ***
## x            0.49894    0.10773   4.632 1.12e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9628 on 98 degrees of freedom
## Multiple R-squared:  0.1796, Adjusted R-squared:  0.1712
## F-statistic: 21.45 on 1 and 98 DF,  p-value: 1.117e-05
```

```
plot(x,y,
     main='Least Squares Line vs Population Regression Line')
lines(x, fit_i$fitted.values)
lines(x, -1 + 0.5*x, col='blue', lwd=2)
legend('topleft', c('Least Squares Line', 'Population Regression Line'), col=c('red', 'blue'), lwd=2)
```



We once again see similar coefficient estimates for $\hat{\beta}_0$ and $\hat{\beta}_1$ as the previous two models, which closely approximates the true values. We also see a comparable decrease in R^2 proportional to variance increase in the error term. As the error term variance goes up, we see that we now only explain approximately 17.96. For the plot, we now see that the lines can be differentiated moreso than all previous models meaning that the OLS fit is slightly less accurate (though still very accurate) than the previous models. This makes sense with the increase in the variance for the error term resulting in a larger spread of data points.

```
cat("Second fit with e ~ N(0, 0.001) \n")
```

```
## Second fit with e ~ N(0, 0.001)
```

```
confint(fit_h)
```

```
##                   2.5 %     97.5 %
## (Intercept) -1.0023016 -0.9984522
## x            0.4978516  0.5021272
```

```
cat('\n')
```

```
cat("First fit with e ~ N(0, 0.25) \n")
```

```
## First fit with e ~ N(0, 0.25)
```

```
confint(fit_e)
```

```
##                   2.5 %     97.5 %
## (Intercept) -1.1150804 -0.9226122
## x            0.3925794  0.6063602
```

```
cat('\n')
```

```
cat("Third fit with e ~ N(0, 1) \n")
```

```
## Third fit with e ~ N(0, 1)
```

```
confint(fit_i)
```

```
##                   2.5 %     97.5 %
## (Intercept) -1.2301607 -0.8452245
## x            0.2851588  0.7127204
```

We put the models in ascending order in terms of variance for the error term, with the second fit first, the original fit second, and the third fit third. We see that there is a positive correlation with variance in the error term and width of the confidence interval. That is, as the data set becomes more noisy, the range of values in the confidence interval becomes wider. This makes sense since the proportion of the variance explained by the model is also going down, meaning we become less confident in our ability to explain the variance in the response variable y, which also corresponds to a wider range of values for our confidence interval.

14

(a)

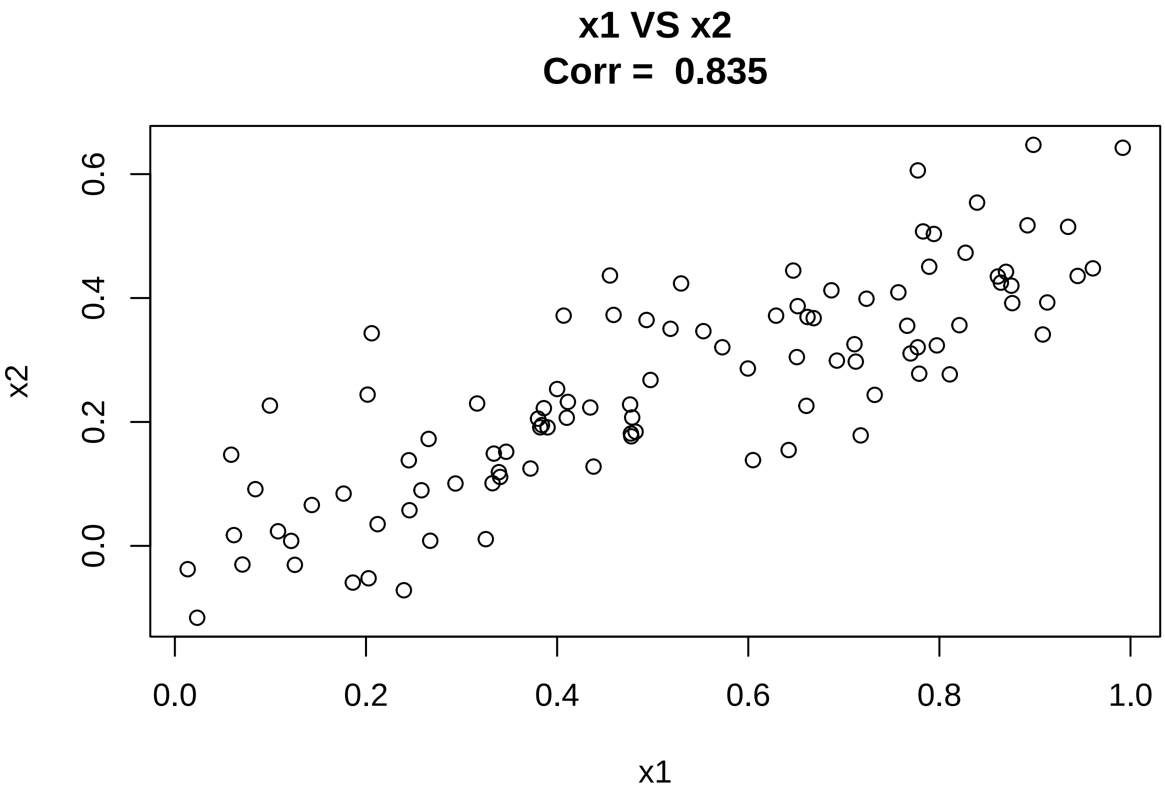
```
set.seed(1)
x1 <- runif(100)
x2 <- 0.5*x1 + rnorm(100) / 10
y <- 2 + 2*x1 + 0.3*x2 + rnorm(100)
```

The linear model has the form

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \sim N(0,1) \\
Y &= 2 + 2X_1 + 0.3X_2 + \epsilon
\end{aligned}
$$

(b)

```
plot(x1,x2,
     main = paste('x1 VS x2\nCorr = ', round(cor(x1,x2),3)))
```



(c)

```
fit_14c <- lm(y ~ x1 + x2)
summary(fit_14c)
```

```
##
## 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
```

Overall, we see that an additive linear model with variables X1 and X2 is not a good fit for the responses. The R2 value is 0.2, which is relatively low. The F-statistic is not very large (it is 12.8), and the RSE is close to 1 (it is 1.056).

$\hat{\beta}_0 \approx 2.13$, which is an overestimate of $\hat{\beta}_0 = 2.0$.

$\hat{\beta}_1 \approx 1.44$, which is an overestimate of $\hat{\beta}_1 = 2.0$.

$\hat{\beta}_2 \approx 1.01$, which is an overestimate of $\hat{\beta}_1 = 0.3$.

With a p-value of 0.0487, we can reject the null hypothesis $H0 : \hat{\beta}_1 = 0$ at the 0.05 significance level and suggest $\hat{\beta}_1 \neq 0$. With a p-value of 0.3754, we fail to reject the null hypothesis $H0 : \hat{\beta}_2 = 0$–which agrees with the earlier simplification of the linear model.

(d)

```
fit_14d <- lm(y ~ x1)
summary(fit_14d)
```

```
##
## 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
```

This fit has an even lower p-value for X_2 showing that we can still reject the null hypothesis, but this time with a coefficient that approximately matches the true β_1. It makes sense that the coefficient estimate would match the true value in this case since X_1 is independent of X_2.

(e)

```
fit_14e <- lm(y ~ x2)
summary(fit_14e)
```

```
##
## 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
```

We see that X_2 is now statistically significant meaning we can reject the null that $\beta_2 = 0$. However, the coefficient estimate for X_2 does not come close to the true value for β_2. This makes sense since it is highly correlated with and depends on X_1. It is therefore trying to signal the magnitude of the X_1 variable's impact on Y through X_2 since X_1 wasn't included in the model.

(f)

The results in (c)-(e) don't contradict each other since there is a high correlation between X_1 and X_2. When we include both in the model, we see that X_1 impact is being underestimated while X_2 is being overestimated, which can be attributed to collinearity. The second model has just X_1 which is independent of X_2, so we get its isolated impact on the response pretty accurately. The third model with just X_2 is statistically significant, but mainly because it is essentially a proxy for both the impact of X_1 and itself.

```r
fit_14f <- lm(y ~ x1*x2)
summary(fit_14f)
```

```
##
## Call:
## lm(formula = y ~ x1 * x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.79572 -0.67391 -0.05085  0.61296  2.29607
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2418     0.3167   7.079 2.39e-10 ***
## x1            1.1235     0.9467   1.187    0.238
## x2            0.3833     1.6603   0.231    0.818
## x1:x2         1.2497     2.4121   0.518    0.606
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.06 on 96 degrees of freedom
## Multiple R-squared:  0.211,  Adjusted R-squared:  0.1864
## F-statistic: 8.559 on 3 and 96 DF,  p-value: 4.296e-05
```

(g)

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

fitg1 <- lm(y~x1 + x2)
fitg2 <- lm(y~x1)
fitg3 <- lm(y~x2)
```

```r
summary(fitg1)
```

```
##
```

```
## 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
```

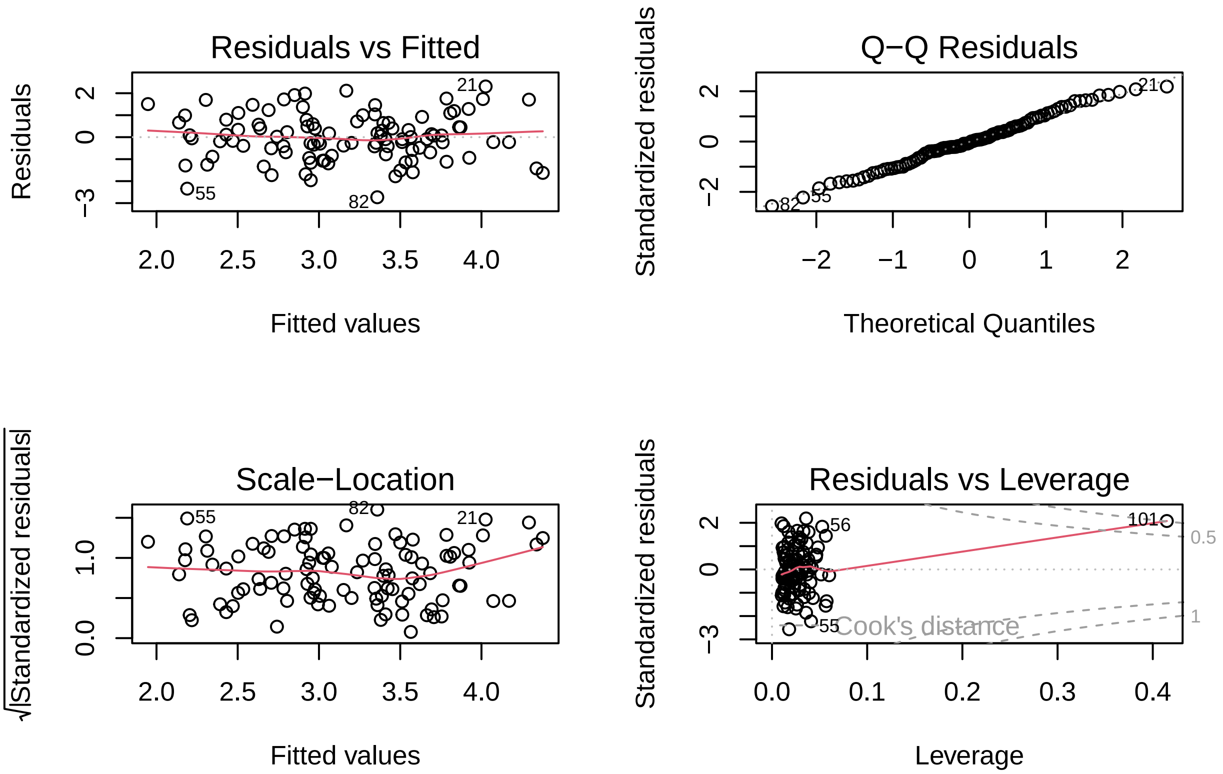
```
summary(fitg2)
```

```
##
## 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
```

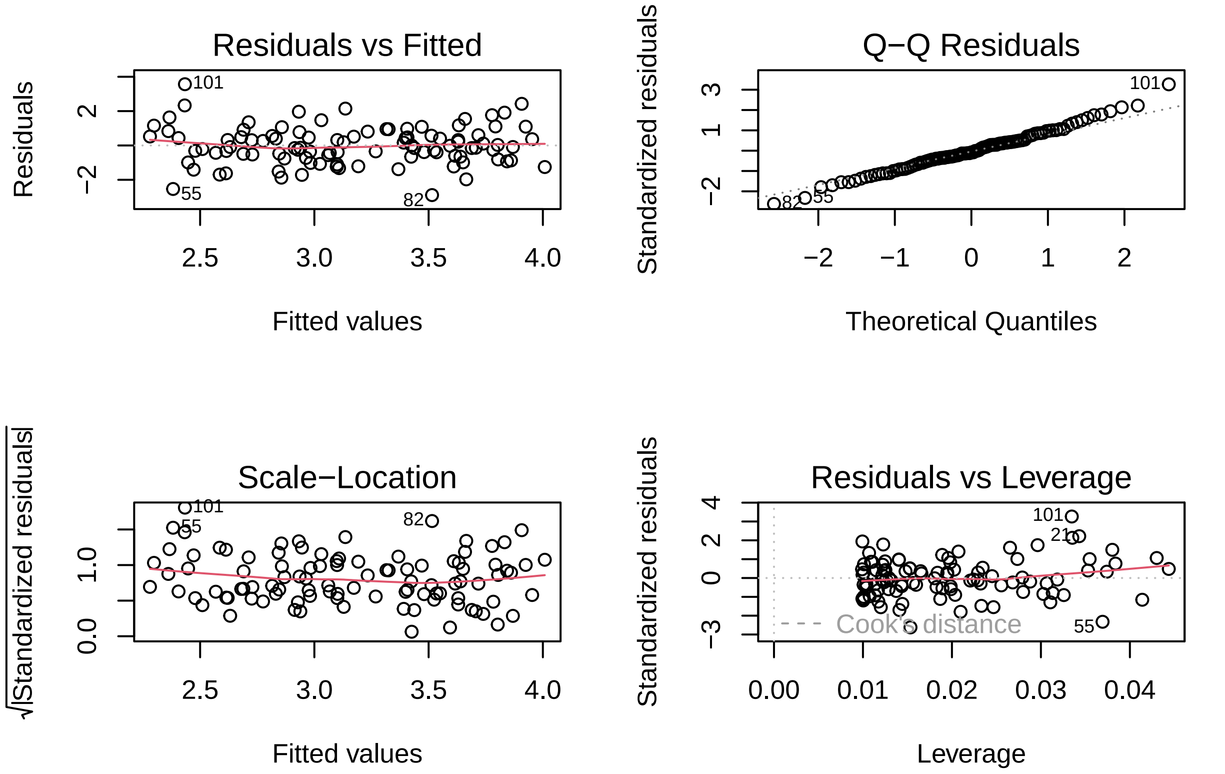
```
summary(fitg3)
```

```
##
## 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
```

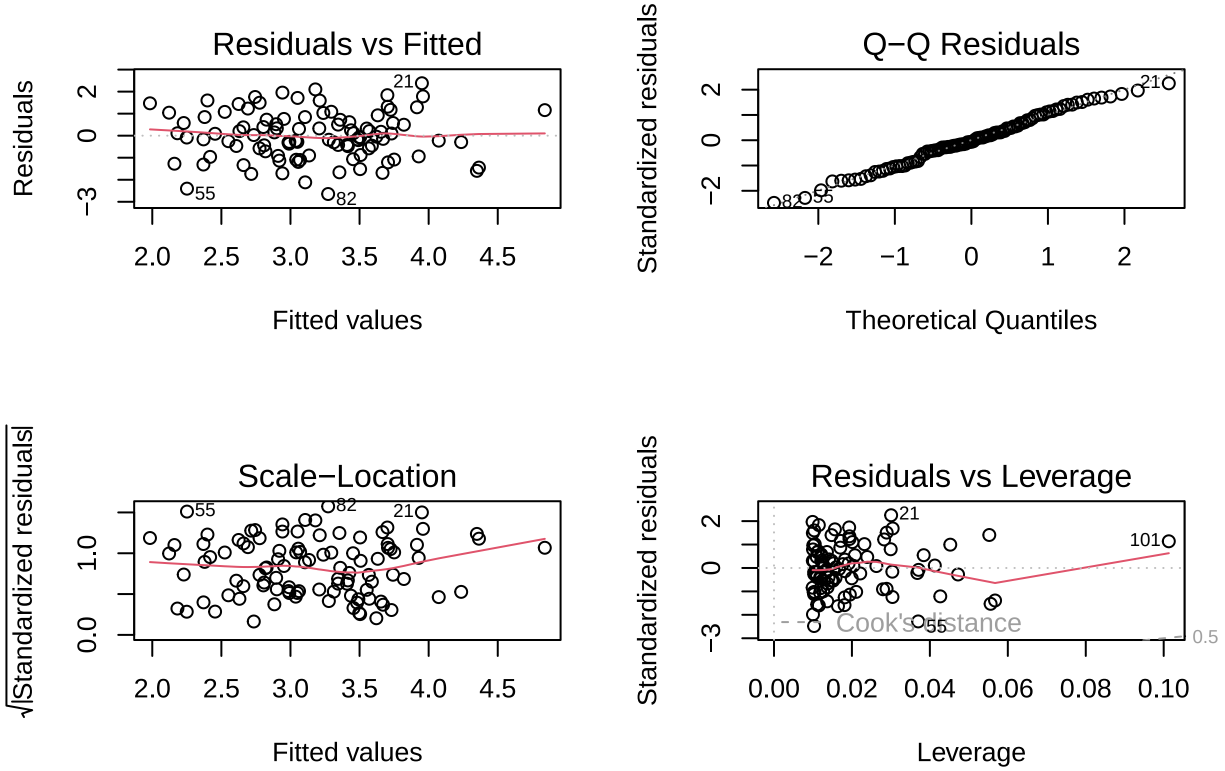
```r
par(mfrow=c(2,2))
plot(fitg1)
```



```r
par(mfrow=c(2,2))
plot(fitg2)
```



```r
par(mfrow=c(2,2))
plot(fitg3)
```



We see that the first model now has X_2 being statistically significant meaning we can reject the null hypothesis, while X_1 is now not statistically significant and we fail to reject the null in its case. Both of the coefficient estimates are still inaccurate.

For the second and third models, we see that both X_1 and X_2 are both statistically significant for each of their respective regressions, meaning we can reject the null in both cases. For the model with X_1, we see that it still comes close to its true β estimate, but not as close as it did before this new observation was introduced. The β estimate for X_2 is still way off for its respective model.

Looking at the residual plots for each model, we see that there is strong evidence across the board that the new observation is a high leverage point in the first model with both variables since it crosses the highest threshold for Cook's distance in the leverage plot.

The added observation does not show up as a high leverage point in either of the models with just X_1 or X_2. However, it does appear to have relatively high leverage for the third model with just X_2, just not enough to be a concern based on Cook's distance.

For the second model with just X_1, we have evidence that the additional point could be an outlier from looking at the QQ plot and the residuals vs fitted plots. However, we don't appear to have any evidence that the added observation is an outlier in the first or third model with both variables and just X_2, respectively.