Homework 4

Jared Andreatta

2025-04-07

Problem 3

a.

b.

```
gpa <- 4.0
iq <- 110
pred <- 50 + 20*gpa +.07*iq + 35 + .01 * iq * gpa - 10 * gpa
cat("Predicted salary: $", pred, "thousand")
```

```
## Predicted salary: $ 137.1 thousand
```

c.

False. The magnitude of the coefficient is not particularly useful in consideration of its statistical significance, especially for larger values of predictors, like **GPA*IQ**. It's more helpful to conduct an actual hypothesis test to determine the probability of the coefficient being equal to 0.

Problem 10

a.

```
library(ISLR2)
```

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

```
data("Carseats")
df <- Carseats
head(df)
```

```
##   Sales CompPrice Income Advertising Population Price ShelveLoc Age Education
## 1  9.50       138     73          11        276   120       Bad  42        17
## 2 11.22       111     48          16        260    83      Good  65        10
## 3 10.06       113     35          10        269    80    Medium  59        12
## 4  7.40       117    100           4        466    97    Medium  55        14
## 5  4.15       141     64           3        340   128       Bad  38        13
## 6 10.81       124    113          13        501    72       Bad  78        16
##   Urban  US
## 1   Yes Yes
## 2   Yes Yes
## 3   Yes Yes
## 4   Yes Yes
## 5   Yes  No
## 6    No Yes
```

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

```
confint(lm.fit)
```

```
##                   2.5 %      97.5 %
## (Intercept) 11.76359670 14.32334118
## Price       -0.06476419 -0.04415351
## UrbanYes    -0.55597316  0.51214085
## USYes        0.69130419  1.70984121
```

b.

Price - For a $1 increase in price, we expect a ~54 decrease in unit sales for a store, on average. Urban - If the store is in an Urban neighborhood, we expect a ~22 decrease in unit sales for a store, on average. However, it is important to note that this coefficient has a very small t-statistic and a very large p-value, indicating that it is most likely not statistically significant to this model. US - For a store that is in the US, we expect the store to sell ~1201 more units than a store not in the US, on average.

c.

$$\widehat{Sales} = 13.043 - .054 * Price - .022 * Urban + 1.201 * US$$

where **Urban**, **US** are binary indicator variables.

d.

We can reject the null hypothesis for **Price** and **US**, but not **Urban**.

e.

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

According to the ANOVA table, adding the **Urban** variable to the regression has little to no effect on the fit of the model, which is indicated by the F-statistic. Additionally, the adjusted R^2 remain mostly the same, which means that the explained variance of the models is mostly unchanged after adding the predictor.

```
anova(lm.fit,lm.fit2)
```

```
## Analysis of Variance Table
##
## Model 1: Sales ~ Price + Urban + US
## Model 2: Sales ~ Price + US
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    396 2420.8
## 2    397 2420.9 -1  -0.03979 0.0065 0.9357
```

g.

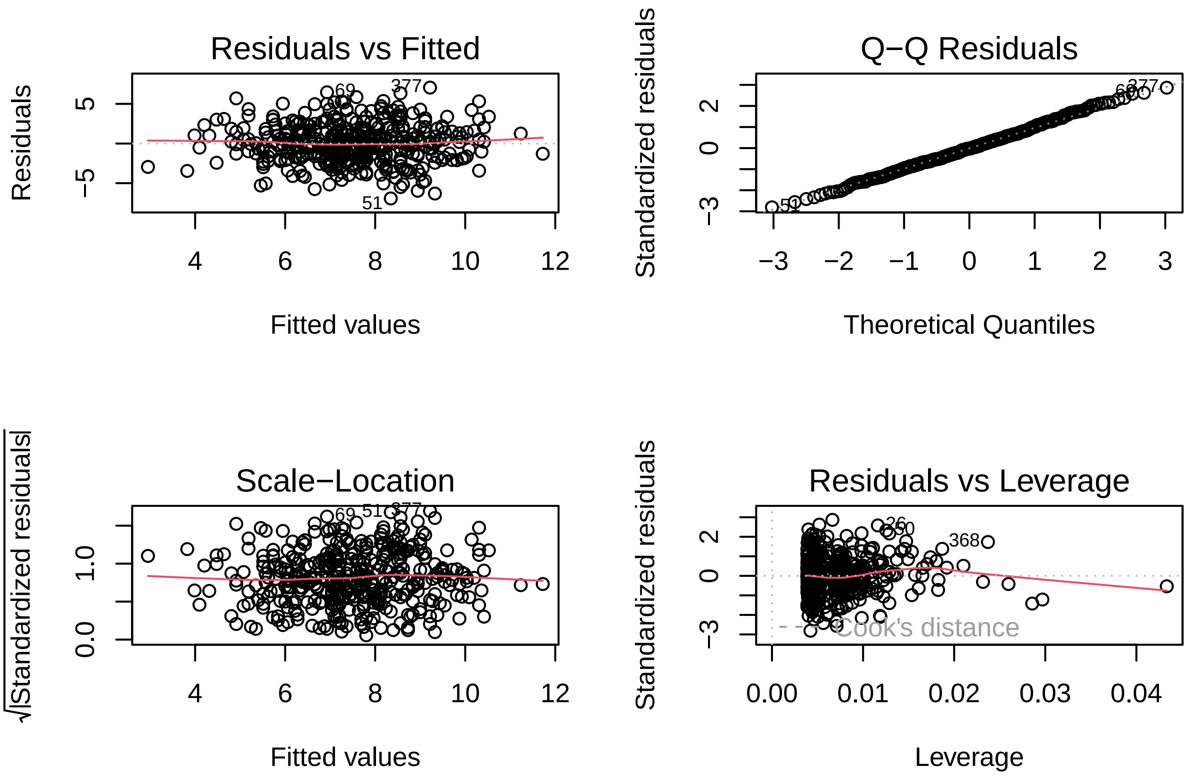
```r
confint(lm.fit2)
```

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

f.

There is no particularly strong evidence supporting outliers in the dataset, however, based off of the leverage plot, there may be presence of a high-leverage point.

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



Problem 13

a, b, and c.

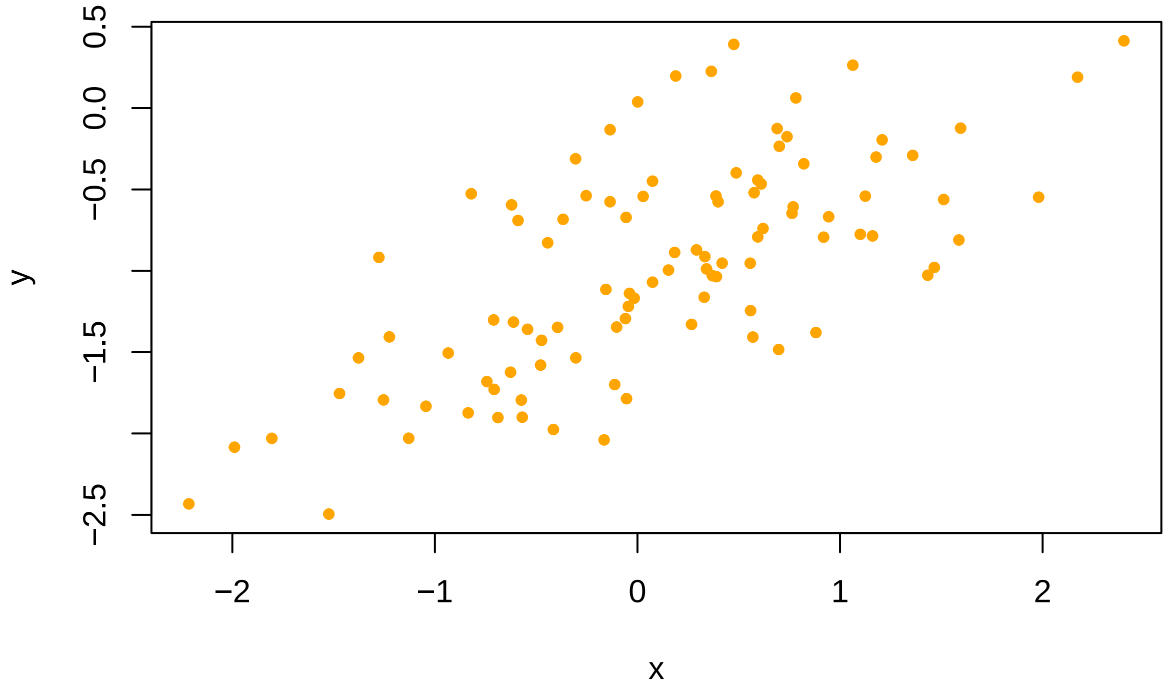
y has a length of 100, $\beta_0 = -1$, and $\beta_1 = .5$.

```r
set.seed(1)

# a.
x <- rnorm(100,0,1)

# b.
eps <- rnorm(100,0,sqrt(.25))

# c.
y = -1 + .5*x + eps
```

d.

Based off of the plot, there is a positive linear relationship between x and y.

```r
plot(x,y,pch=20,col="orange")
```



## e. The approximated hat values are very close to the true values.

```r
fit_sim <- lm(y~x)
summary(fit_sim)
```

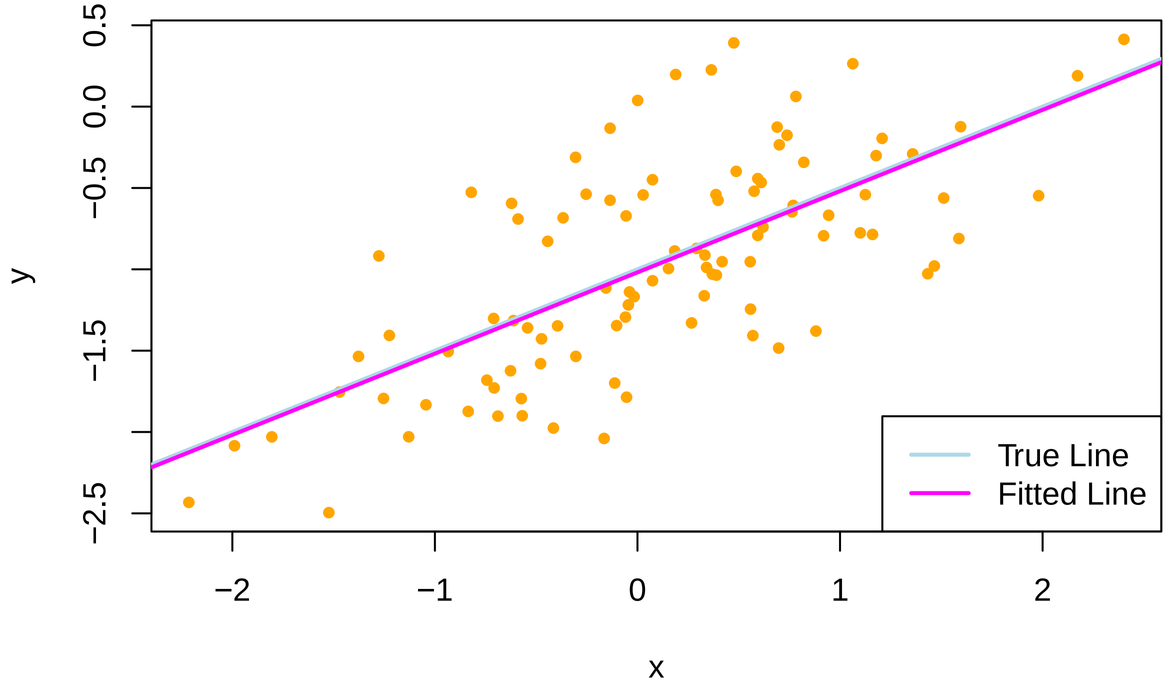
```
##
## Call:
```

```
## 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
```

```
coef(fit_sim)
```

```
## (Intercept)           x
##  -1.0188463   0.4994698
```

```
plot(x,y,pch=20,col="orange")
abline(-1,.5, col="lightblue", lwd=2)
abline(fit_sim, col="magenta", lwd=2)
legend("bottomright",
       legend=c("True Line", "Fitted Line"),
       col=c("lightblue","magenta"),
       lwd=2)
```



## g. After adding the quadratic term, we can see directly that the quadratic term is (likely) not statistically significant to the model. The adjusted R^2 value rises nominally from adding the quadratic term, and the anova table shows little evidence of the new term having any significant effect on the model.

```
fit_quad <- lm(y ~ poly(x,2))
summary(fit_quad)
```

```
##
## Call:
## lm(formula = y ~ poly(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)  -0.9645     0.0479 -20.134  < 2e-16 ***
## poly(x, 2)1   4.4638     0.4790   9.319 3.97e-15 ***
## poly(x, 2)2  -0.6720     0.4790  -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
```

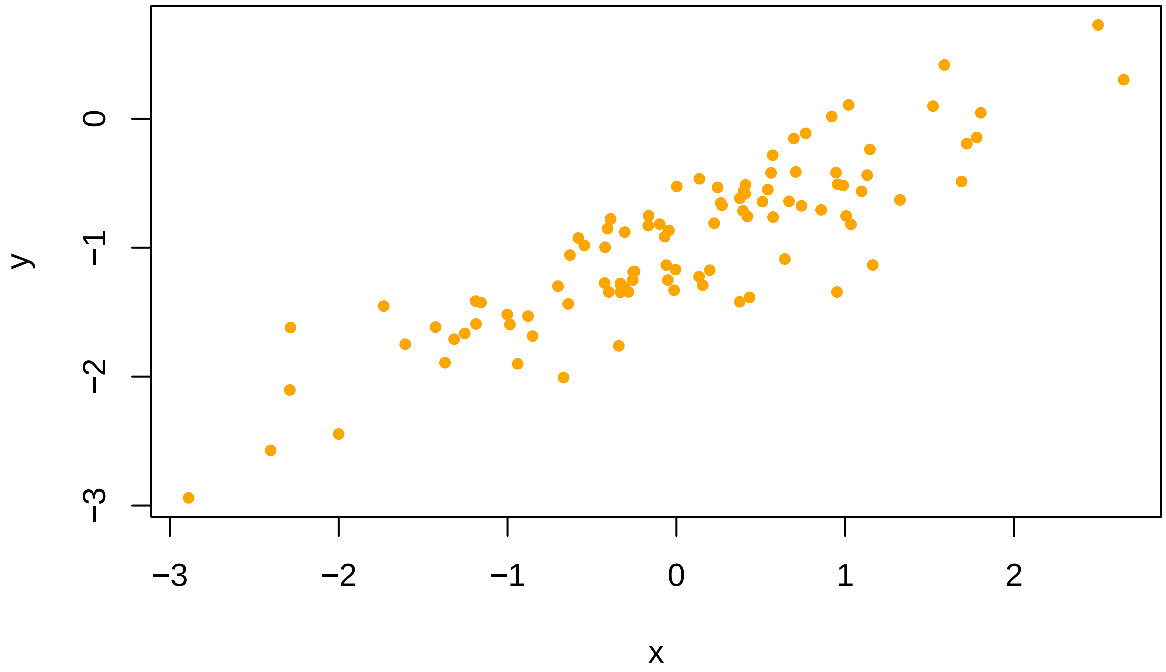
```r
anova(fit_sim, fit_quad)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x
## Model 2: y ~ poly(x, 2)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     98 22.709
## 2     97 22.257  1   0.45163 1.9682 0.1638
```

h.

After we redo the process with a lower variance ($\epsilon \sim N(0, 0.1)$), we can see that there is a significant improvement in the amount of variance explained by the model compared to the model with higher variance; the previous model had an adjusted R^2 value of approximately .46, while the lower variance model has a value of .76.

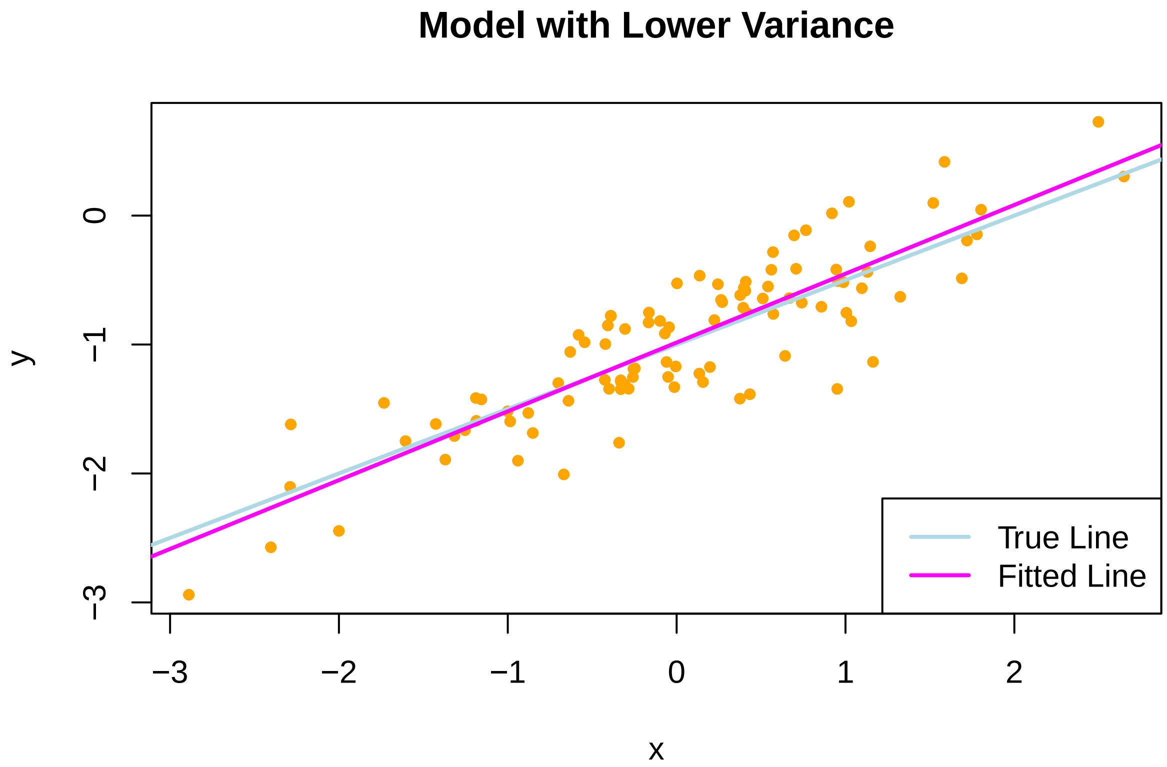
```r
x <- rnorm(100,0,1)
eps <- rnorm(100,0,sqrt(.1))
y = -1 + .5*x + eps

plot(x, y, pch=20, col="orange")
```



```r
fit_sim2 <- lm(y ~ x)
summary(fit_sim2)
```

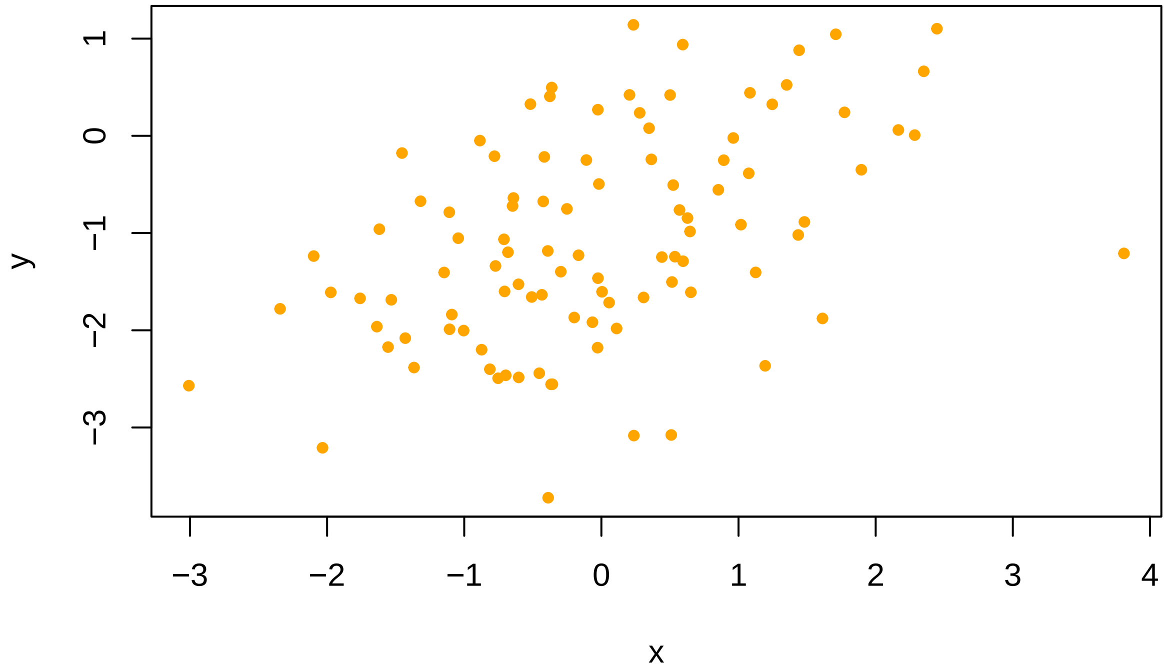
```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.86703 -0.17753 -0.00553  0.21495  0.58452
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.98468    0.03134  -31.42   <2e-16 ***
## x            0.53359    0.03044   17.53   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.3133 on 98 degrees of freedom
## Multiple R-squared:  0.7582, Adjusted R-squared:  0.7557
## F-statistic: 307.3 on 1 and 98 DF,  p-value: < 2.2e-16
```

```r
plot(x,y,pch=20,col="orange", main="Model with Lower Variance")
abline(-1,.5, col="lightblue", lwd=2)
abline(fit_sim2, col="magenta", lwd=2)
legend("bottomright",
       legend=c("True Line", "Fitted Line"),
       col=c("lightblue","magenta"),
       lwd=2)
```



## i. As we can see, the fitted regression line deviates from the true regression line, both in the intercept and the slope, which can be seen in the chart and in the regression results. Additionally, the R^2 value drops significantly, which is to be expected, as the data is more noisy for this model.

```
x <- rnorm(100,0,1)
eps <- rnorm(100,0,1)
y = -1 + .5*x + eps

plot(x, y, pch=20, col="orange")
```

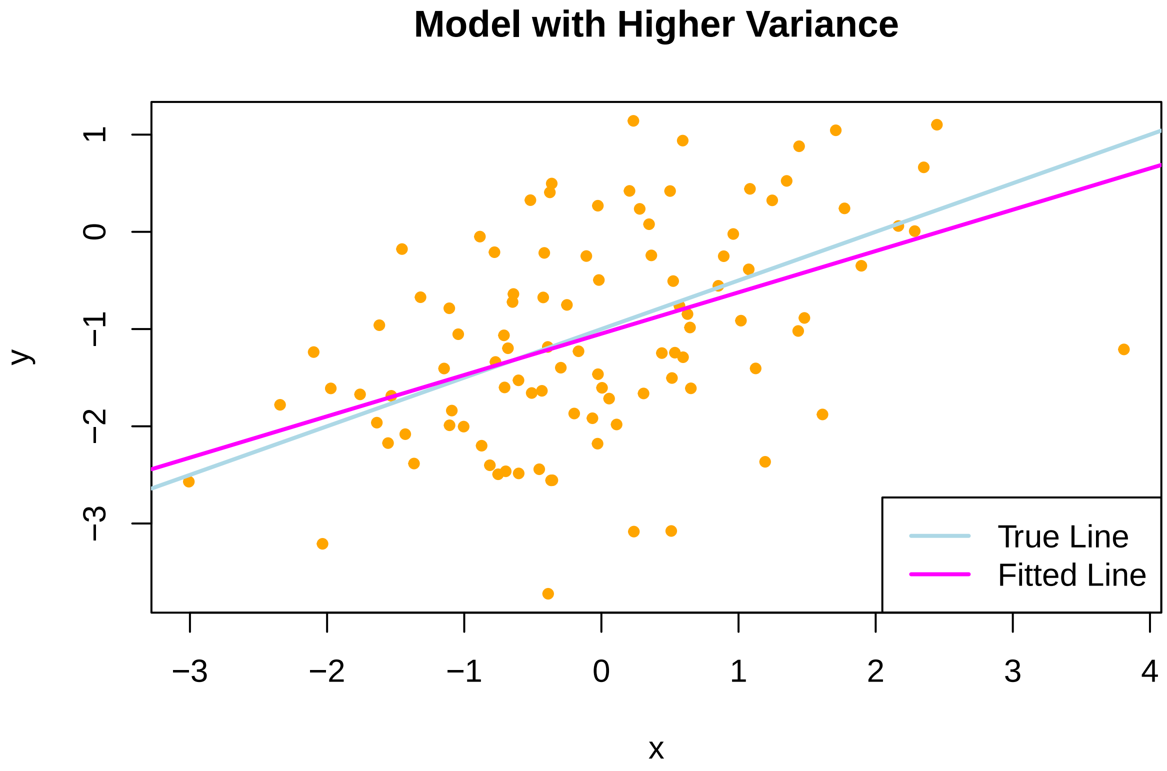


```
fit_sim3 <- lm(y ~ x)
summary(fit_sim3)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.51014 -0.60549  0.02065  0.70483  2.08980
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.04745    0.09676 -10.825  < 2e-16 ***
## x            0.42505    0.08310   5.115 1.56e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9671 on 98 degrees of freedom
## Multiple R-squared:  0.2107, Adjusted R-squared:  0.2027
## F-statistic: 26.16 on 1 and 98 DF,  p-value: 1.56e-06
```

```
plot(x,y,pch=20,col="orange", main="Model with Higher Variance")
abline(-1,.5, col="lightblue", lwd=2)
abline(fit_sim3, col="magenta", lwd=2)
```

```
legend("bottomright",
       legend=c("True Line", "Fitted Line"),
       col=c("lightblue","magenta"),
       lwd=2)
```



## j. The 95% confidence set significantly tightens for the reduced variance models, while the model with increased variance produces a much wider confidence set. The noisier the data becomes, the the estimates of the parameters become less confident and the confidence intervals widen.

```
confint(fit_sim) # Original
```

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

```
confint(fit_sim2) # Less noisy
```

```
##                  2.5 %     97.5 %
## (Intercept) -1.0468683 -0.9224893
## x            0.4731823  0.5939962
```

```
confint(fit_sim3) # More noisy
```

```
##                  2.5 %     97.5 %
## (Intercept) -1.2394772 -0.8554276
## x            0.2601391  0.5899632
```

Problem 14

a.

$$Y = 2 + 2x_1 + 0.3x_2 + \epsilon$$

where $\epsilon \sim N(0,1)$. The coefficients are $\beta_0 = 2$, $\beta_1 = 2$, and $\beta_2 = 0.3$

```
set.seed(1)
x1 <- runif(100)
x2 <- .5 * x1 + rnorm(100) /10
y <- 2 + 2*x1 + .3*x2 + rnorm(100)
```

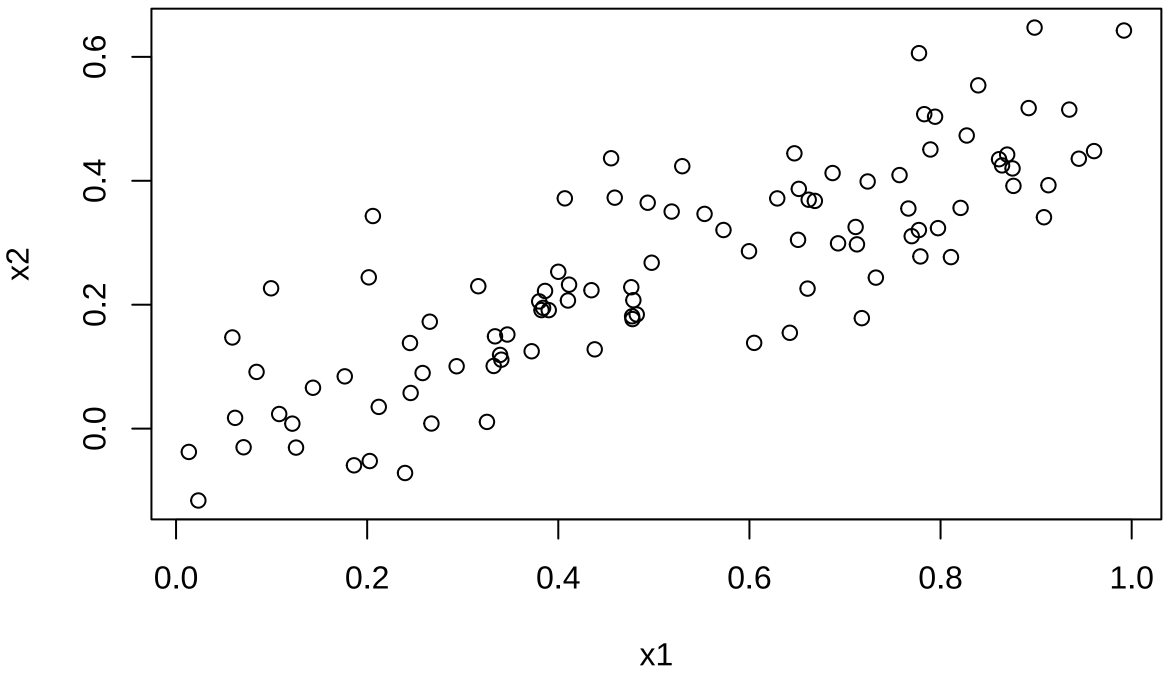
b.

x_1 and x_2 have a Pearson correlation coefficient of 0.835.

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

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

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



c. $\hat{\beta}_0 = 2.1305$, $\hat{\beta}_1 = 1.4396$, and $\hat{\beta}_2 = 1.0097$. The first two coefficients are relatively close to their true values, but this is not true for $\hat{\beta}_2$. We can reject the null hypothesis for β_1, but we fail to reject for β_2.

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

```
##
## 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
```

d.

The estimate for β_1 is much closer to the true value and is significant on the $>99.9\%$ level, with the same being true for the intercept. We can reject the null hypothesis that $\beta_1 = 0$.

```r
yhat2 <- lm(y~x1)
summary(yhat2)
```

```
##
## 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
```

e.

The results are similar as above. We can reject the null hypothesis again for β_1.

```
yhat3 <- lm(y~x2)
summary(yhat3)
```

```
##
## 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
```

f.

These regressions do not produce conflicting results. In the model where the coefficients are jointly estimated for x_1 and x_2, the estimate for x_2 is not significant. Yet, when we estimate the coefficients for these variables separately, they are both significant. While these results do not necessarily conflict, they arise questions about other problems, such as multicollinearity. This makes sense since these variables are highly correlated and x_2 is distributed dependently on x_1.

g.

In the first model, there is strong evidence that the new point is a high-leverage point, which is indicated in the bottom right leverage plot for the diagnostic. For the x_1 model, it seems to be an outlier, whereas for the x_2 model, there doesn't seem to be sufficiently strong evidence to support that it is an outlier or a high-leverage point.

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

yhat <- lm(y~x1+x2)
yhat2 <- lm(y~x1)
yhat3 <- lm(y~x2)

summary(yhat)
```

```
##
## 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
```

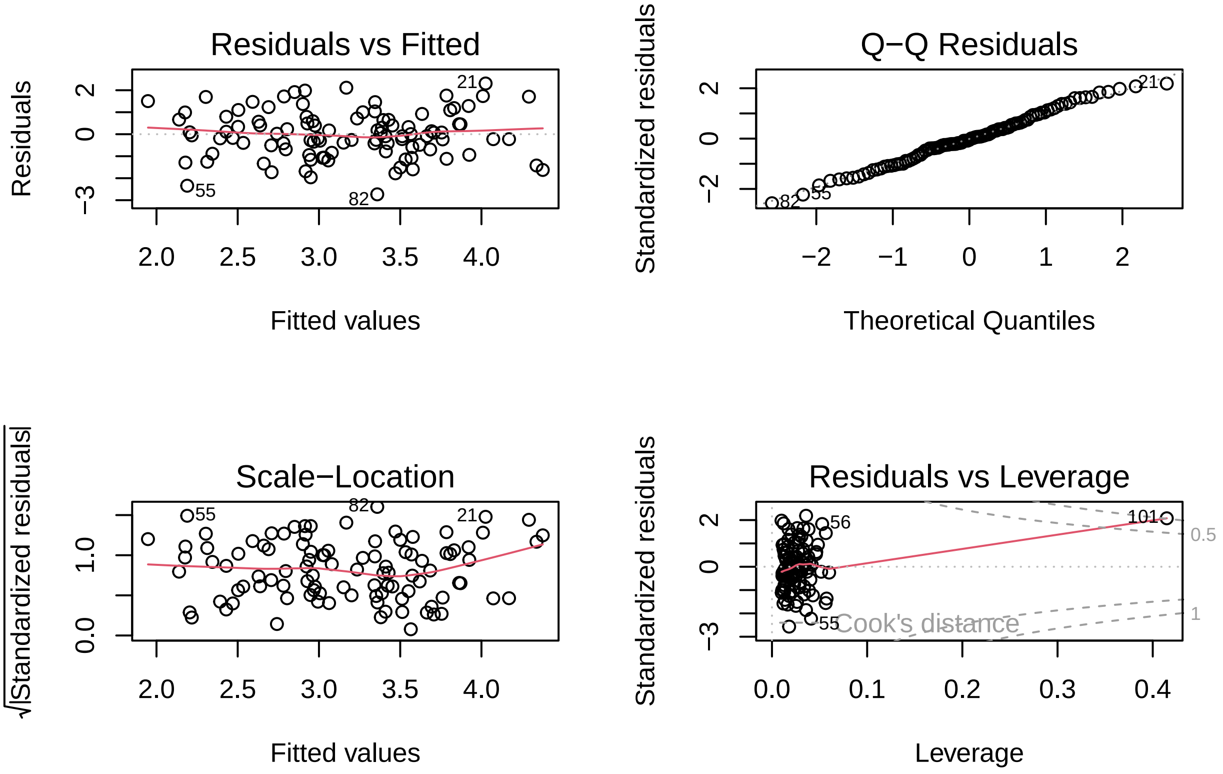
```
summary(yhat2)
```

```
##
## 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
```

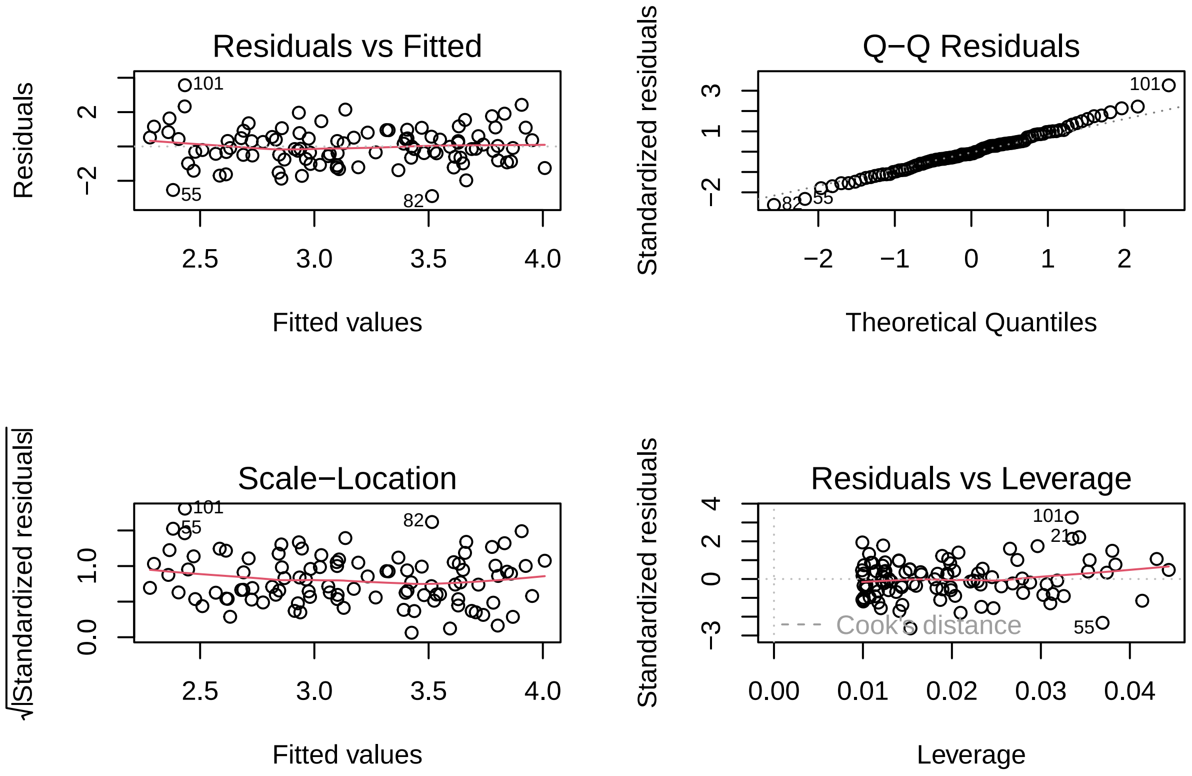
```
summary(yhat3)
```

```
##
## 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
```

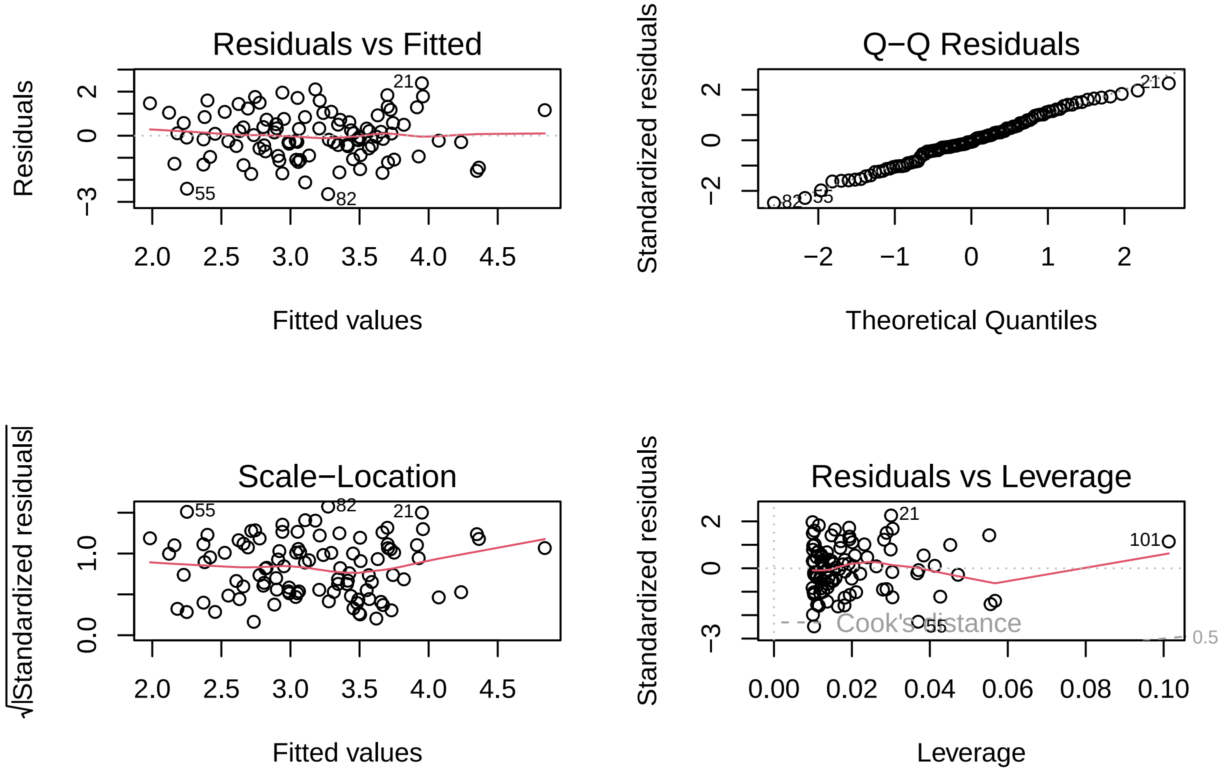
```r
par(mfrow=c(2,2))
plot(yhat)
```



```r
par(mfrow=c(2,2))
plot(yhat2)
```



```
par(mfrow=c(2,2))
plot(yhat3)
```


Residuals vs Fitted
Residuals
2
0
−3
21
55
82
2.0
2.5
3.0
3.5
4.0
4.5
Fitted values
Q−Q Residuals
Standardized residuals
2
0
−2
21
82
55
−2
−1
0
1
2
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
0.0
1.0
55
82
21
2.0
2.5
3.0
3.5
4.0
4.5
Fitted values
Residuals vs Leverage
Standardized residuals
2
0
−3
21
101
55
Cook's distance
0.5
0.00
0.02
0.04
0.06
0.08
0.10
Leverage