week_3_exercises

Lindsey Greenhill

2/10/2021

Question 1

Part a

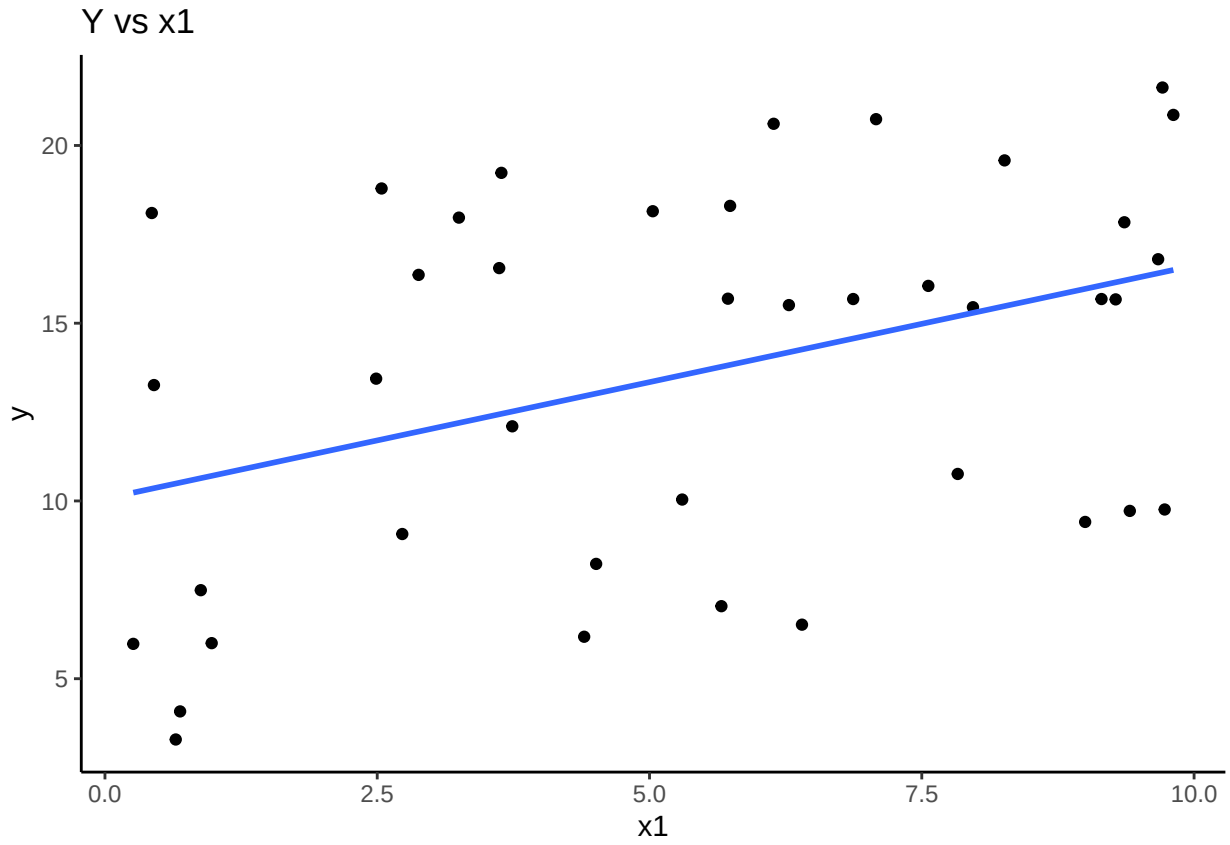
```
## lm(formula = y ~ x1 + x2, data = df_train)
##             coef.est coef.se
## (Intercept) 1.32     0.39
## x1          0.51     0.05
## x2          0.81     0.02
## ---
## n = 40, k = 3
## residual sd = 0.90, R-Squared = 0.97
```

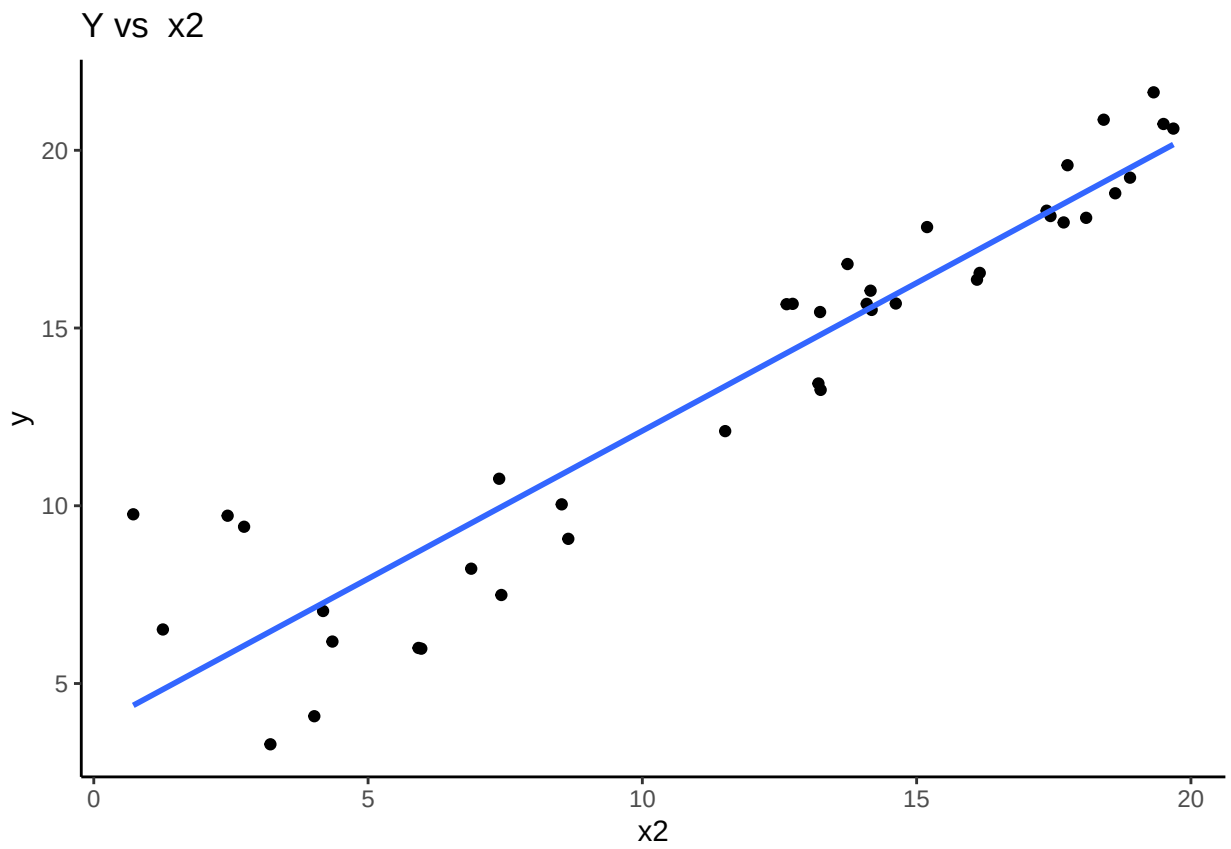
The intercept is equal to 1.31, meaning that the average value of y when x1 and x2 are both 0 is 1.13. The coefficient for x1 is .514, meaning that for every 1 unit increase in x1, y increases by .514 on average, holding x2 constant. The coefficient for x2 is .806, meaning that for every 1 unit increase in x2, y increases by .806 on average, holding x1 constant.

The R squared value is .97, meaning that about 97% of the variance is explained by the model

Part b

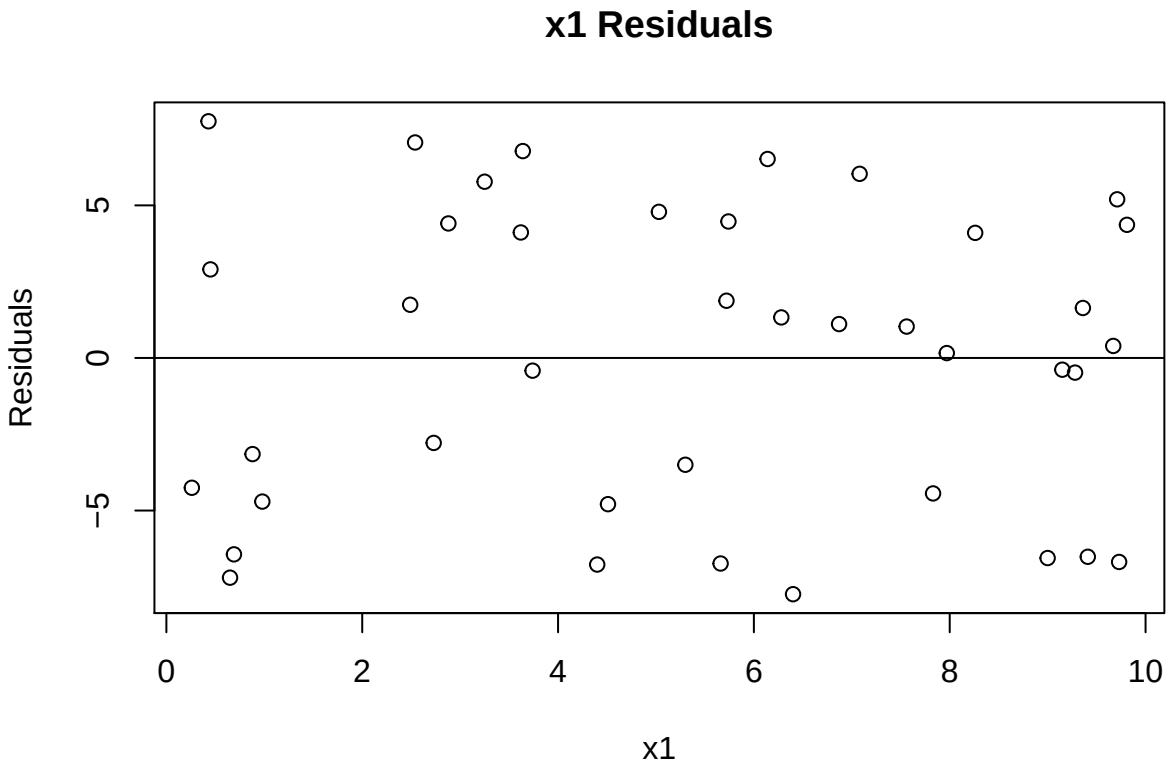
The graphs below visualize the relationship between y and x1 and x2. The blue line is the least squared regression line.

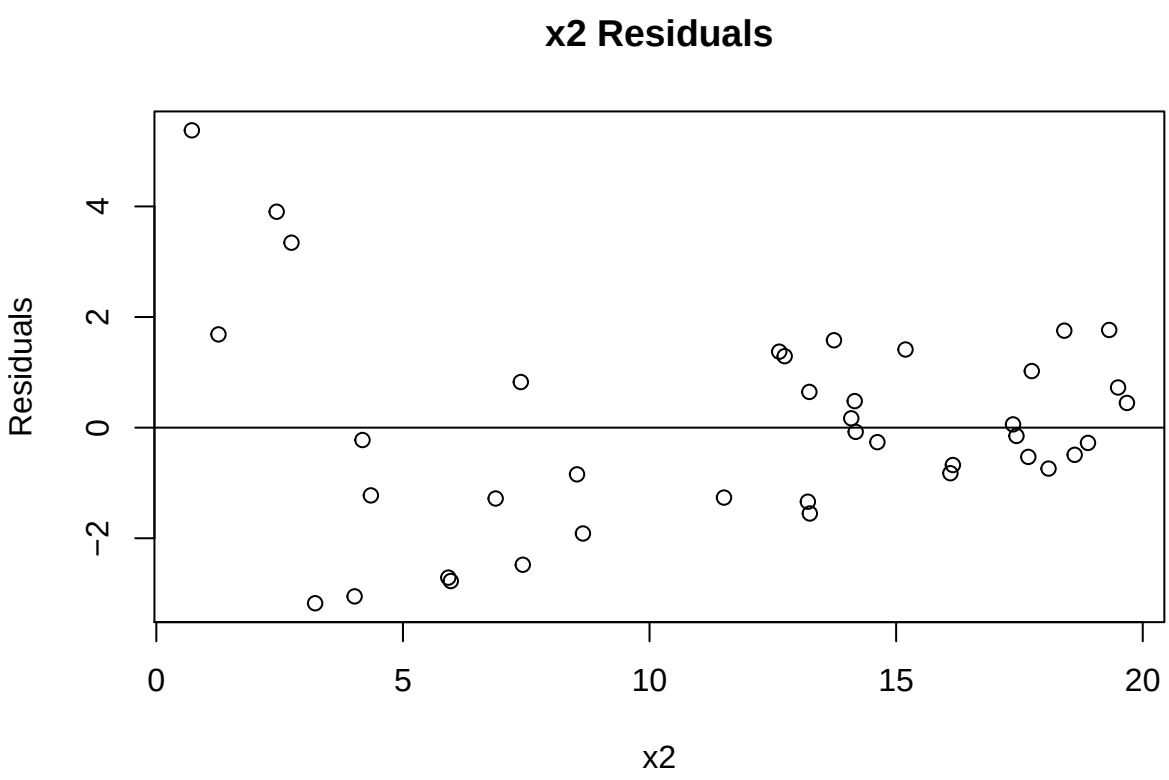



Y vs x2
y
x2
0
5
10
15
20
5
10
15
20

Part c

The plots below show the residuals of the two above models plotted against the observed values from the data frame. For the x1 plot, the residuals seem to be relatively evenly spread throughout the x1 values with no clear patterns emerging. For the x2 plot, the residuals seems to be larger for the x2 values closer to 0. However, the residuals for x2 seem to be smaller on average than the residuals for x1.

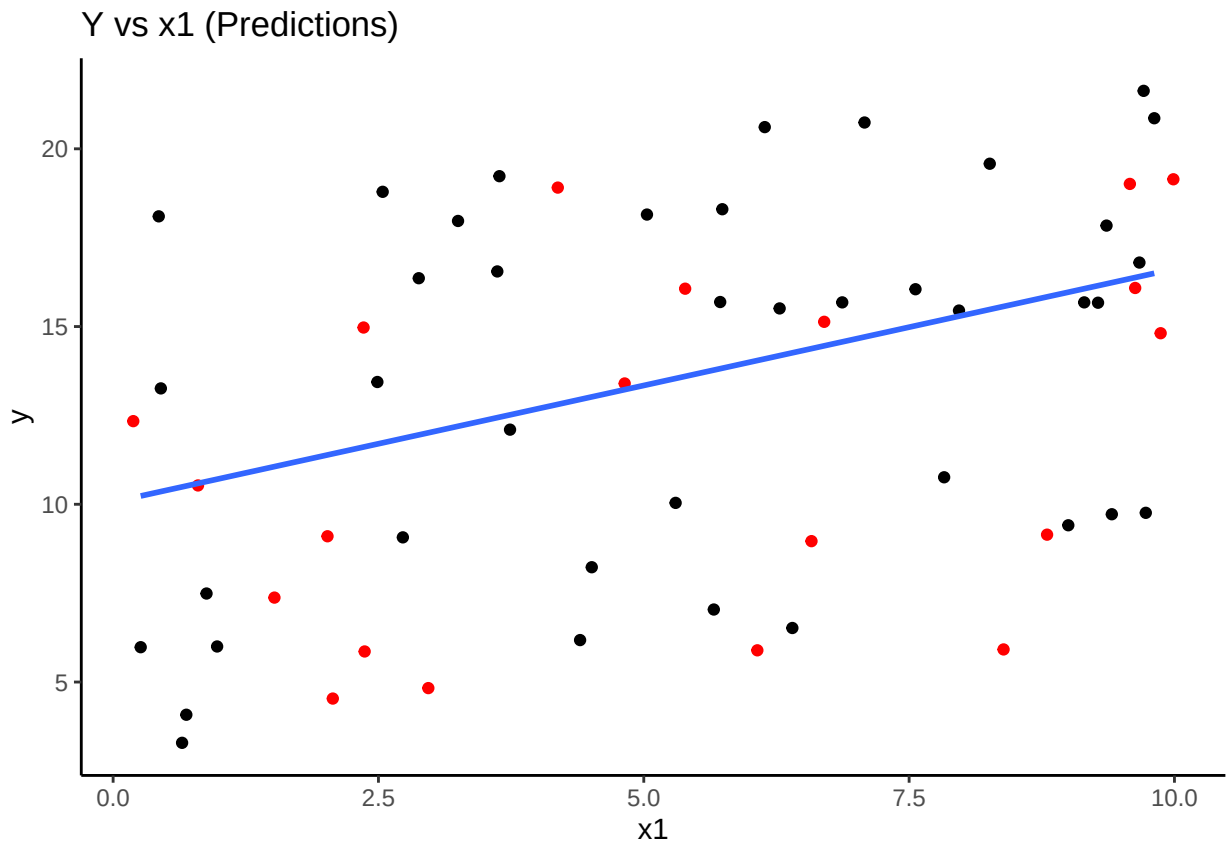



x2 Residuals
Residuals
4
2
0
−2
0
5
10
15
20
x2

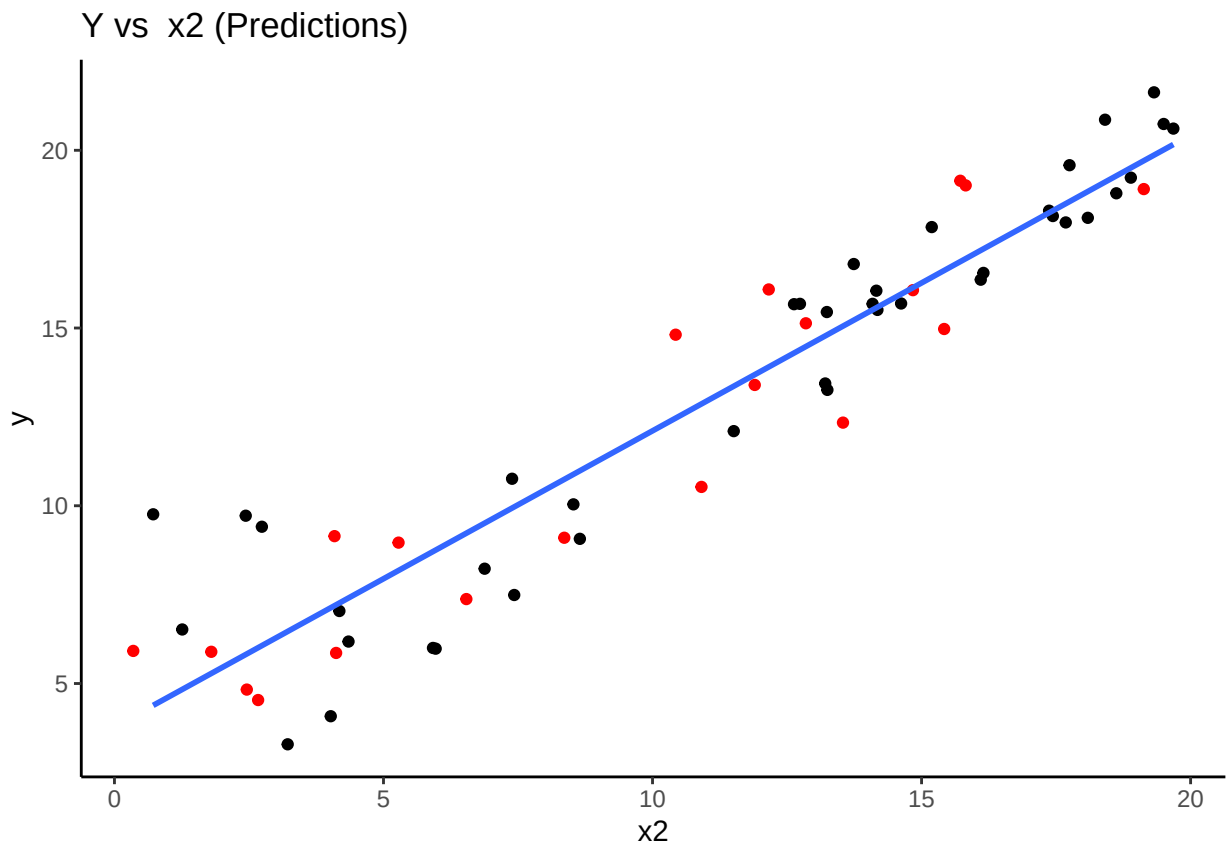
Part d

The plots below show the same regression lines shown in part b above. The red points are the predictions generated using the last 20 observations in the data frame. The predictions seem to follow the same general pattern and trend line as the data points used to generate the regression. Therefore, we should be relatively confident that these predictions are reasonable. With that being said, it is worth recognizing that the fits are not perfect representations of reality and the plot for x2 appears to be a closer fit than x1.

```
## `geom_smooth()` using formula 'y ~ x'
```




```
## `geom_smooth()` using formula 'y ~ x'
```


Y vs x2 (Predictions)
y
x2
0
5
10
15
20
5
10
15
20

Question 3

```
##
## Call:
## lm(formula = var1 ~ var2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6724 -0.6604 -0.0067  0.6302  3.6015
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.007457   0.030078   0.248    0.804
## var2        0.018581   0.029158   0.637    0.524
##
## Residual standard error: 0.9496 on 998 degrees of freedom
## Multiple R-squared:  0.0004067,  Adjusted R-squared:  -0.0005949
## F-statistic: 0.4061 on 1 and 998 DF,  p-value: 0.5241
```

Part a

The slope is not statistically significant, because it is within 2 standard deviations from 0 (logic taken from Gelman and Hill "if the absolute value of the z-score exceeds 2, the estimate is statistically significant") and there are also no stars in the model output. This makes sense because the variables are random and therefore we do not expect them to have any meaningful relationship.

Part b

Out of the 100 simulations, 3 estimates are statistically significant. I determined this by counting how many z scores were greater than 2 or less than -2 (logic taken from Gelman and Hill "if the absolute value of the z-score exceeds 2, the estimate is statistically significant").