Homework 4

Brady Lamson

```
gpa <- read.table("../datasets/gpa.txt", header = T)
```

Problem 4.14

Part A

```
no_int_lm <- lm(gpa ~ 0 + act.score, data = gpa)
no_int_lm |> summary()
```

```

Call:
lm(formula = gpa ~ 0 + act.score, data = gpa)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.0276 -0.2737  0.1077  0.4754  2.1820 

Coefficients:
          Estimate Std. Error t value Pr(>|t|)    
act.score 0.121643   0.002637   46.13   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7257 on 119 degrees of freedom
Multiple R-squared:  0.947, Adjusted R-squared:  0.9466 
F-statistic:  2128 on 1 and 119 DF,  p-value: < 2.2e-16
```

$$Y = 0 - 3.0276X_1 + \epsilon$$

Part B

```
conf <- confint(no_int_lm, level = .95)
glue::glue(
    "With 95% confidence, the slope of beta_1 is in
    ({conf[1] |> round(3)}, {conf[2] |> round(3)})
    "
)
```

```
With 95% confidence, the slope of beta_1 is in
(0.116, 0.127)
```

As the confidence interval does not contain 0, there is significant evidence to indicate that, in the no intercept model, there is a positive relationship between act scores and college freshman gpa.

Part C

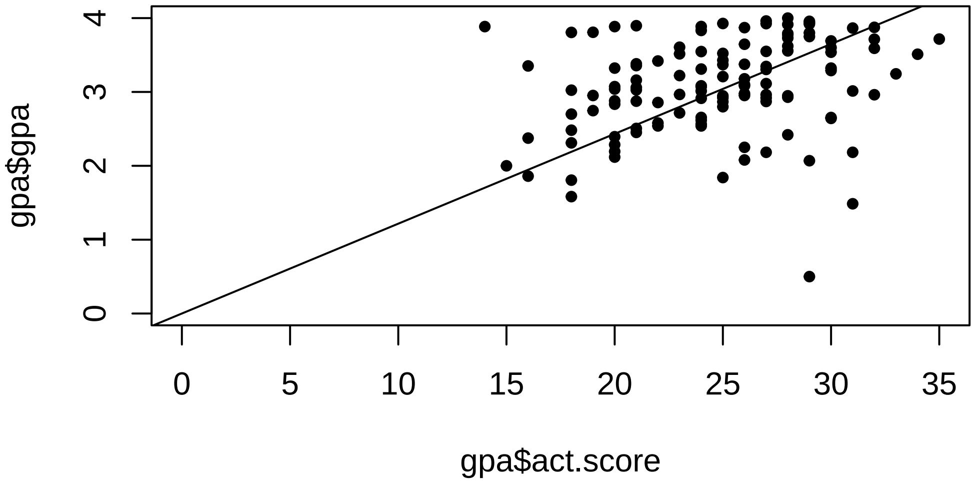
```
my.new.data <- data.frame(act.score = 30)
predict(no_int_lm, newdata = my.new.data, interval = "confidence",
level = 0.95)
```

```
       fit      lwr      upr
1 3.649287 3.492647 3.805928
```

Problem 4.15

Part A

```
plot(
    x = gpa$act.score, y = gpa$gpa, pch=20,
    xlim = c(0,35), ylim = c(0,4)
)
abline(no_int_lm)
```



The linear regression though the origin does not appear to be a good fit.

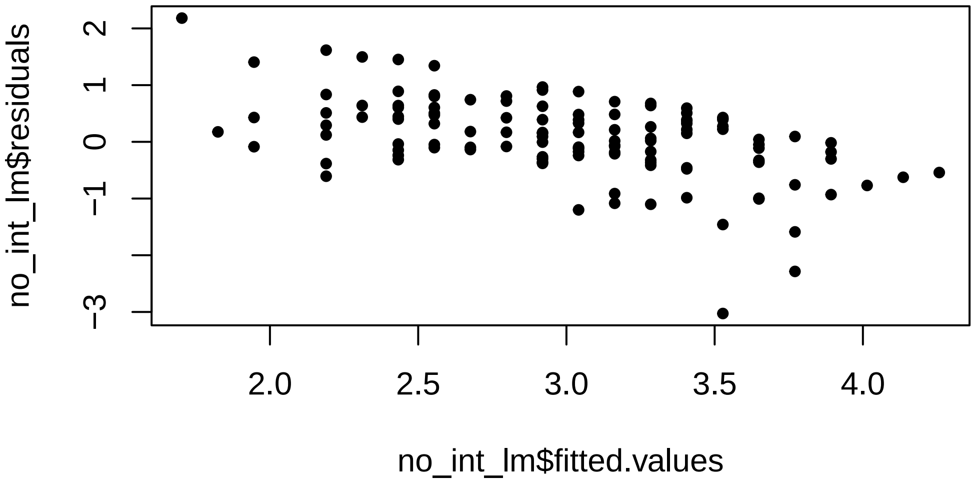
Part B

```r
sum_resid <- no_int_lm$residuals |> sum()

glue::glue("The sum of the residuals is {sum_resid |> round(3)}")
```

```
The sum of the residuals is 7.972
```

The residuals do not sum to 0.

```r
plot(no_int_lm$fitted.values, no_int_lm$residuals, pch=20)
```



The residuals seem to trend downward as the fitted values increase.

Part C

```r
full_reg <- lm(gpa ~ act.score, data = gpa)

anova(no_int_lm, full_reg)
```

```
Analysis of Variance Table

Model 1: gpa ~ 0 + act.score
Model 2: gpa ~ act.score
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    119 62.670
2    118 45.818  1    16.852 43.401 1.304e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$
\begin{aligned}
H_0 &: E(Y) = \beta_0 + \beta_1 X \\
H_a &: E(Y) \neq \beta_0 + \beta_1 X \\
\alpha &= 0.005 \\
F &= 43.401 \\
p &\approx 1.304 \cdot 10^{-9}
\end{aligned}
$$

As $p < \alpha$, there is significant evidence to indicate the linear regression model is not sufficient to explain the relationship in the data.

Problem 5.5

```
finance <- read.table("../datasets/CH05PR05.txt", header = F)
colnames(finance) <- c("delinquent_loans", "num_companies")
fin_reg <- lm(delinquent_loans ~ num_companies, data = finance)
X <- model.matrix(fin_reg)
Y <- finance$delinquent_loans
```

Part 1

```
t(Y) %*% Y
```

```
     [,1]
[1,] 1259
```

Part 2

```
t(X) %*% X
```

```
              (Intercept) num_companies
(Intercept)             6            17
num_companies          17            55
```

```
t(X) %*% Y
```

```
              [,1]
(Intercept)     81
num_companies  261
```

Problem 5.13

```
solve(t(X) %*% X)
```

```
              (Intercept) num_companies
(Intercept)     1.3414634    -0.4146341
num_companies  -0.4146341     0.1463415
```

Problem 5.24

Part A

```
b <- solve(t(X) %*% X) %*% t(X) %*% Y
fitted_vec <- X %*% b
resid_vec <- Y - (X %*% b)

print("---[Part 1]---")
```

```
[1] "---[Part 1]---"
```

```
print("The vector of estimated regression coefficients is")
```

```
[1] "The vector of estimated regression coefficients is"
```

```
b
```

```
                   [,1]
(Intercept)   0.4390244
num_companies 4.6097561
```

```
print("---[Part 2]---")
```

```
[1] "---[Part 2]---"
```

```
print("The vector of residuals is")
```

```
[1] "The vector of residuals is"
```

```
resid_vec
```

```
        [,1]
1 -2.87804878
2 -0.04878049
3  0.34146341
4  0.73170732
5 -1.26829268
6  3.12195122
```

Part C

```
H <- X %*% solve(t(X) %*% X) %*% t(X)
print("The hat matrix is")
```

```
[1] "The hat matrix is"
```

```
print(H)
```

```
            1          2          3         4         5           6
1  0.36585366 -0.1463415 0.02439024 0.1951220 0.1951220  0.36585366
2 -0.14634146  0.6585366 0.39024390 0.1219512 0.1219512 -0.14634146
3  0.02439024  0.3902439 0.26829268 0.1463415 0.1463415  0.02439024
4  0.19512195  0.1219512 0.14634146 0.1707317 0.1707317  0.19512195
5  0.19512195  0.1219512 0.14634146 0.1707317 0.1707317  0.19512195
6  0.36585366 -0.1463415 0.02439024 0.1951220 0.1951220  0.36585366
```

Problem 6.2

Part A

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & X_{11}^2 \\ 1 & X_{21} & X_{22} & X_{21}^2 \\ 1 & X_{31} & X_{32} & X_{31}^2 \\ 1 & X_{41} & X_{42} & X_{41}^2 \\ 1 & X_{51} & X_{52} & X_{51}^2 \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

Part B

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & logX_{12} \\ 1 & X_{21} & logX_{22} \\ 1 & X_{31} & logX_{32} \\ 1 & X_{41} & logX_{42} \\ 1 & X_{51} & logX_{52} \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

Problem 6.3

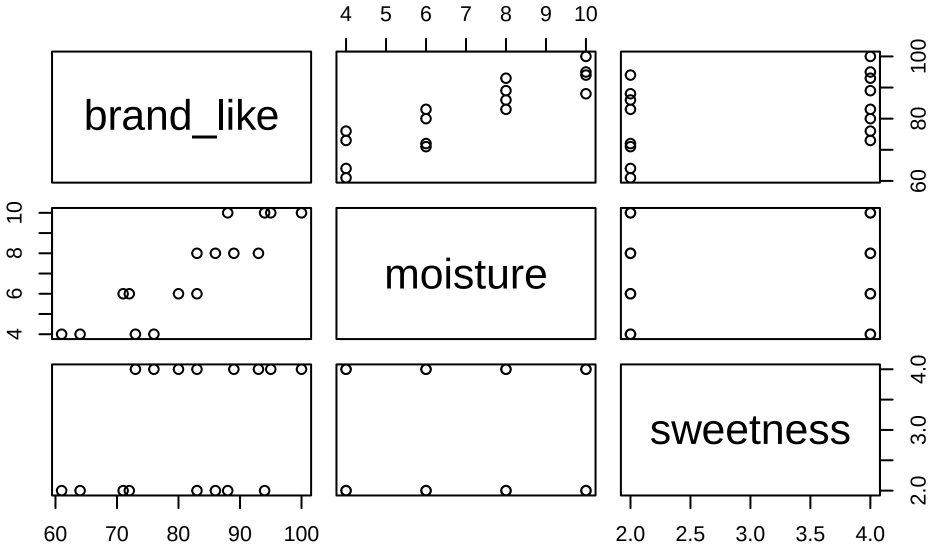
There are many reasons to potentially ignore insignificant predictors. The first is that R^2 is only one measure of the fit of a model and does not its utility. A high R^2 on its own is not enough to state that the model is a good fit. As well, models tend to become less useful as you use them to predict values outside of their scope. What this means is that as you add predictors, you add more dimensionality and a wider range of values you likely shouldn't be predicting on. Not only that, more predictors can hurt interpretation of a model. This cost may be worth it if the predictors are good, but if they add little to the predictive power of the model this can easily snowball into black box territory.

Problem 6.5

```r
brands <- read.table("../datasets/CH06PR05.txt")
colnames(brands) <- c("brand_like", "moisture", "sweetness")
```

Part A

```r
pairs(brands)
```



```r
cor(brands)
```

```
           brand_like  moisture sweetness
brand_like  1.0000000 0.8923929 0.3945807
moisture    0.8923929 1.0000000 0.0000000
sweetness   0.3945807 0.0000000 1.0000000
```

What we can see from this is that brand_like and moisture seem to have a relationship with eachother and are highly correlated.

Sweetness meanwhile has what seems to be a very non-linear relationship with brand like and less correlation, and is totally uncorrelated to moisture.

It’s worth examining if sweetness is work keeping in the model.

Part B

```r
brand_reg <-lm(brand_like ~ moisture + sweetness, data = brands)
brand_reg |> summary()
```

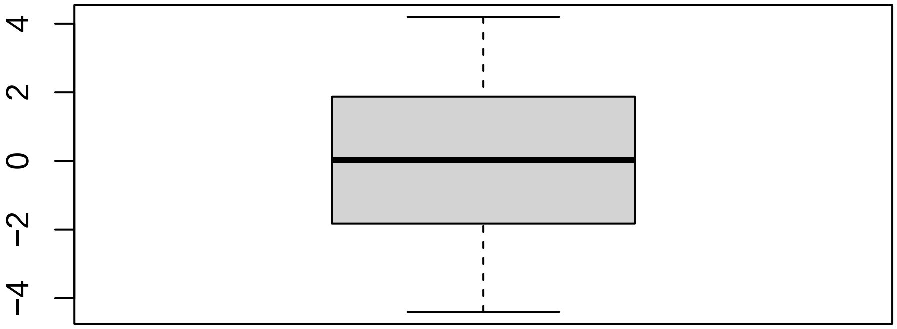
```

Call:
lm(formula = brand_like ~ moisture + sweetness, data = brands)

Residuals:
   Min     1Q Median     3Q    Max 
-4.400 -1.762  0.025  1.587  4.200 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  37.6500     2.9961  12.566 1.20e-08 ***
moisture      4.4250     0.3011  14.695 1.78e-09 ***
sweetness     4.3750     0.6733   6.498 2.01e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.693 on 13 degrees of freedom
Multiple R-squared:  0.9521,	Adjusted R-squared:  0.9447 
F-statistic: 129.1 on 2 and 13 DF,  p-value: 2.658e-09
```

$$Y_i = 37.65 + 4.43X_1 + 4.37X_2$$

β_1 represents the change in brand likeness as moisture increases.

Part C

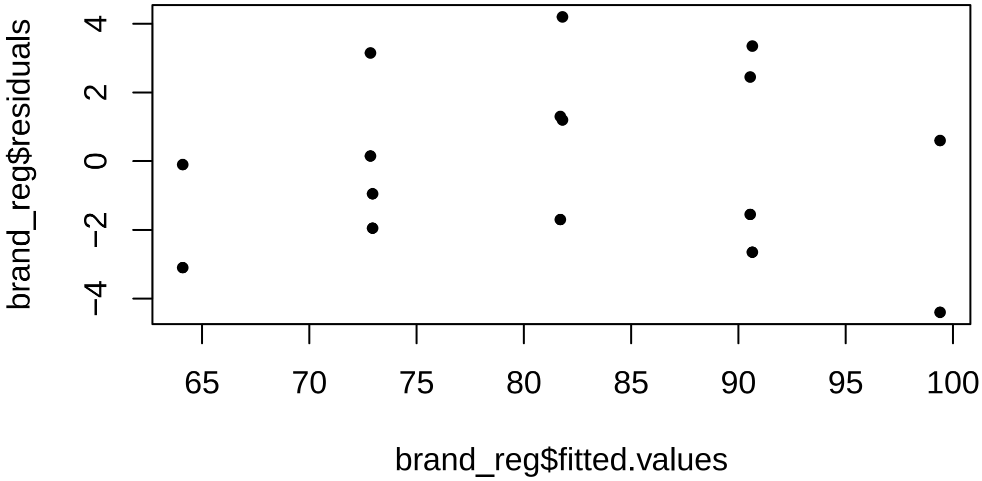
```r
boxplot(brand_reg$residuals)
```



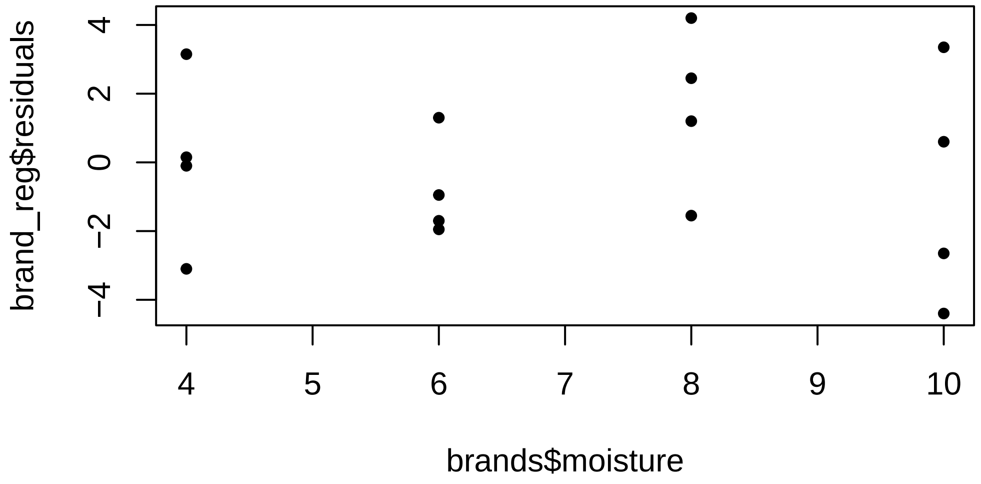
The residuals appear symmetric around 0 and don't seem to do anything weird in either direction. This is a good thing and may give us confidence that the residuals are normally distributed.

Part D

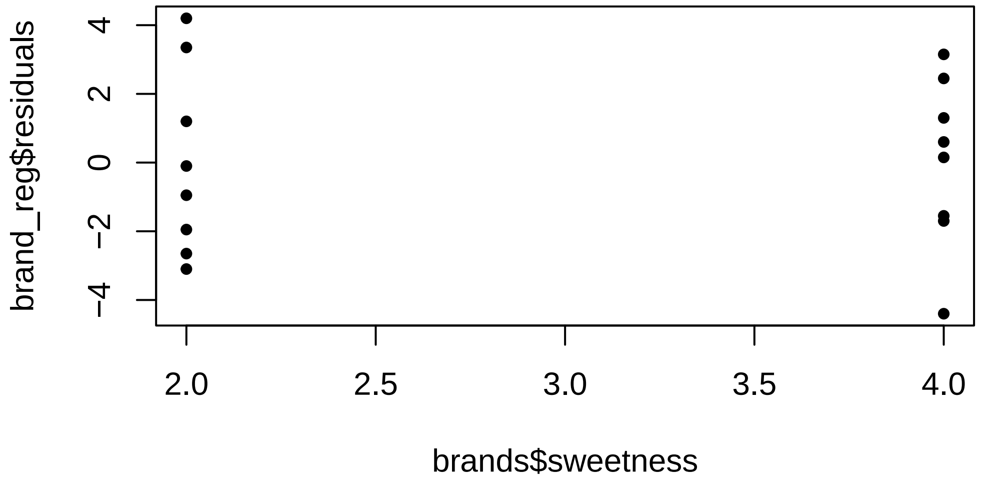
```
plot(x = brand_reg$fitted.values, y = brand_reg$residuals, pch=20)
```



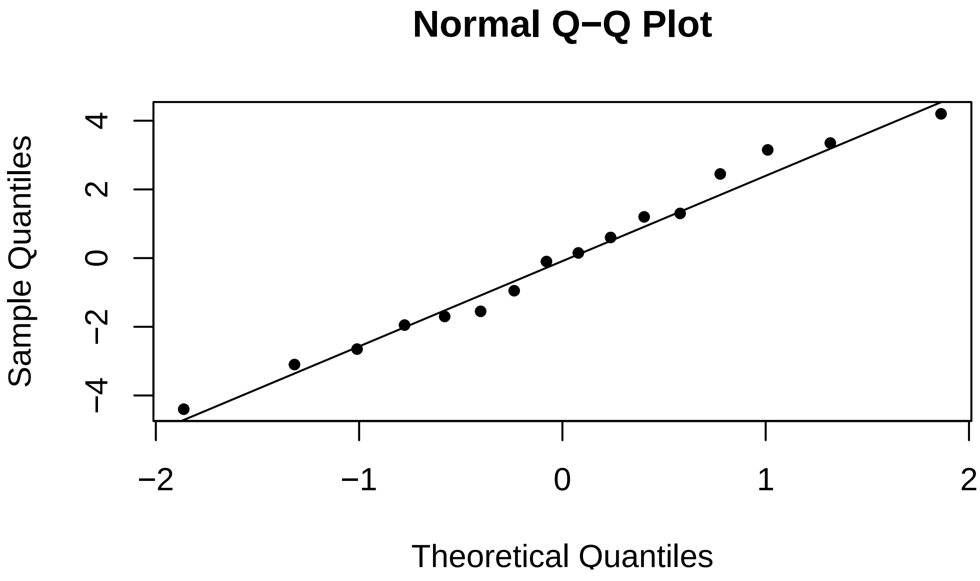
```
plot(x = brands$moisture, y = brand_reg$residuals, pch=20)
```



```
plot(x = brands$sweetness, y = brand_reg$residuals, pch=20)
```



```
qqnorm(brand_reg$residuals, pch=20)
qqline(brand_reg$residuals)
```



The residuals seem mostly constant when plot against the fitted values. There's a weird dip at the far left and right of the plot though which may be cause for concern.

The residuals plotted against moisture appear constant, nothing of note here.

The residuals plotted against sweetness seem constant as well, though its harder to gauge as there's only 2 values used in this data.

The normal probability plot approximately follows the line as well which is good.

Problem 6.6

Part A and B

For this we can refer to the summary output at the start of problem 6.

$$H_0 : \beta_1 = \beta_2 = 0 H_a : \beta_1 \ or \ \beta_2 \neq 0 \alpha = 0.01 F = 129.1 p = 2.01 \cdot 10^{-5}$$

As $p < \alpha$, there is significant evidence to indicate that brand like is influenced by moisture and sweetness. This means either β_1 or β_2 isn't 0.

problem 6.7

Part A

$$R^2 = .9521$$

This means that 95% of the variation in the data is explained by the regression model.

Part B

```r
ssto <- sum(
    (brands$brand_like - mean(brands$brand_like))^2
)

ssr <- sum(
    (brand_reg$fitted.values - mean(brands$brand_like))^2
)

ssr / ssto
```

```
[1] 0.952059
```

The simple and multiple determination coefficient are the same.

Problem 6.8

```r
new_data = data.frame(moisture = 5, sweetness = 4)
```

Part A

```r
predict(brand_reg, newdata = new_data, interval = "confidence", level = .99)
```

```
     fit      lwr      upr
1 77.275 73.88111 80.66889
```

Part B

```
predict(brand_reg, newdata = new_data, interval = "prediction", level = .99)
```

```
     fit      lwr      upr
1 77.275 68.48077 86.06923
```