QTM 220 HW #8

Author

Veronica Vargas

QTM 220 HW #8

Exercise #1

```
library(tidyverse)
```

```
Warning: package 'tidyverse' was built under R version 4.3.3

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✓ dplyr     1.1.3     ✓ readr     2.1.4
✓ forcats   1.0.0     ✓ stringr   1.5.0
✓ ggplot2   3.4.3     ✓ tibble    3.2.1
✓ lubridate 1.9.2     ✓ tidyr     1.3.0
✓ purrr     1.0.2
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✕ dplyr::filter() masks stats::filter()
✕ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(experimentr)
library(ggplot2)
library(mosaic)
```

```
Warning: package 'mosaic' was built under R version 4.3.3

Registered S3 method overwritten by 'mosaic':
  method                           from
  fortify.SpatialPolygonsDataFrame ggplot2

The 'mosaic' package masks several functions from core packages in order to add
additional features.  The original behavior of these functions should not be affected by this.

Attaching package: 'mosaic'

The following object is masked from 'package:Matrix':

    mean

The following objects are masked from 'package:dplyr':

    count, do, tally

The following object is masked from 'package:purrr':

    cross

The following object is masked from 'package:ggplot2':

    stat

The following objects are masked from 'package:stats':
```

```
    binom.test, cor, cor.test, cov, fivenum, IQR, median, prop.test,
    quantile, sd, t.test, var

The following objects are masked from 'package:base':

    max, mean, min, prod, range, sample, sum
```

```
library(datasets)
library(leaps)
```

```
Warning: package 'leaps' was built under R version 4.3.3
```

```
library(caret)
```

```
Warning: package 'caret' was built under R version 4.3.3


Attaching package: 'caret'

The following object is masked from 'package:mosaic':

    dotPlot

The following object is masked from 'package:purrr':

    lift
```

```
library(ISLR2)
```

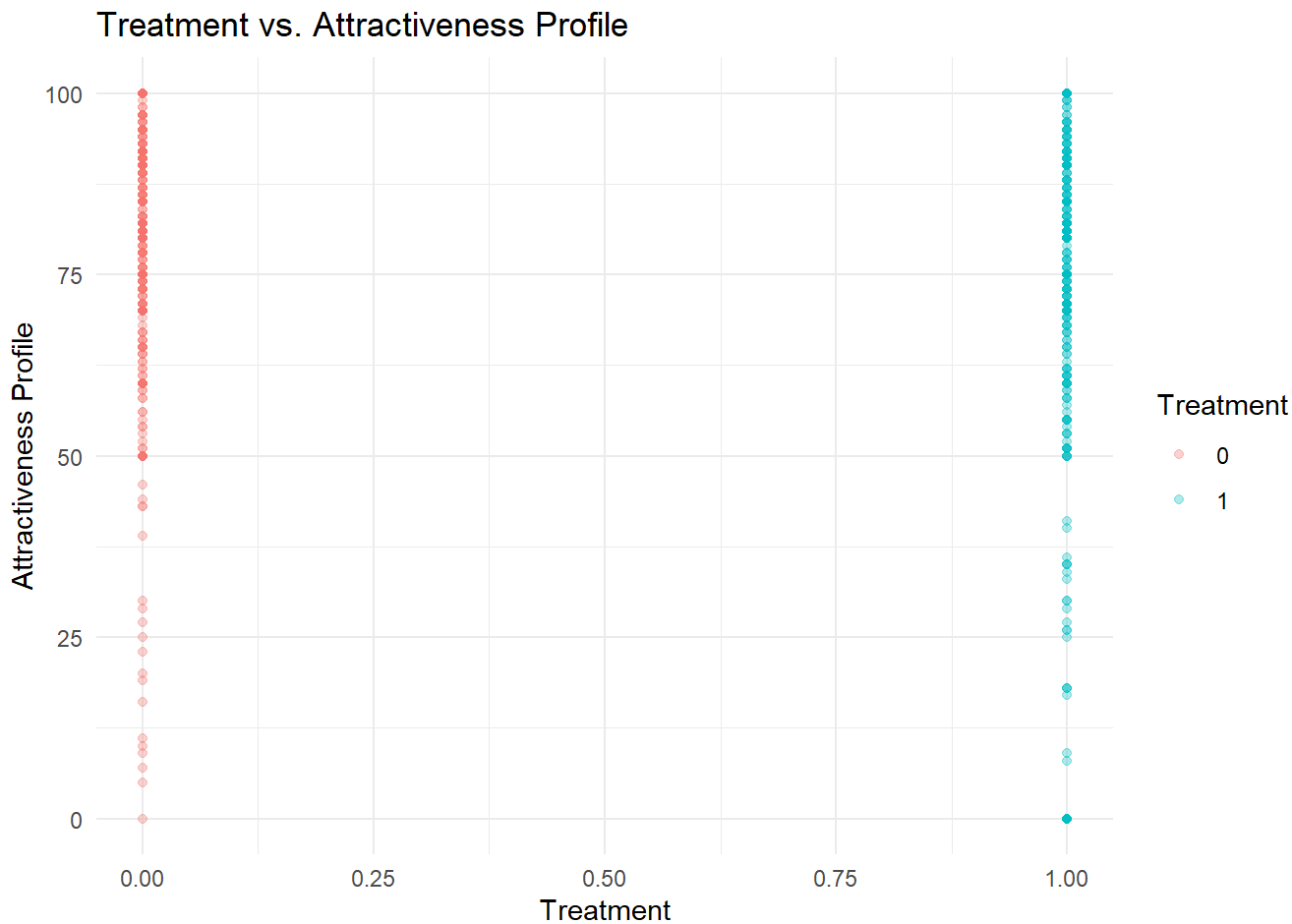
```
Warning: package 'ISLR2' was built under R version 4.3.3
```

```
data("easton")
head(easton)
```

```
  attractiveness_score age male republican treatment_republican_profile
1                   58  30    0          1                            1
2                   55  27    1          0                            1
3                   99  57    1          0                            1
4                   61  38    0          0                            1
5                   92  26    0          0                            0
6                   67  38    0          0                            0
```

(a) Scatterplot

```
ggplot(easton, aes(x = treatment_republican_profile, y = attractiveness_score)) +
  geom_point(aes(x = treatment_republican_profile, y = attractiveness_score, color =
         factor(treatment_republican_profile)),
             alpha = 0.3) +
  labs(
    title = "Treatment vs. Attractiveness Profile",
    x = "Treatment",
    y = "Attractiveness Profile",
    color = "Treatment") +
  theme_minimal()
```



(b) Average Treatment Effect (ATE)

```
mean(easton$attractiveness_score[easton$treatment_republican_profile == 1]) -
         mean(easton$attractiveness_score[easton$treatment_republican_profile == 0])
```

```
[1] -2.787521
```

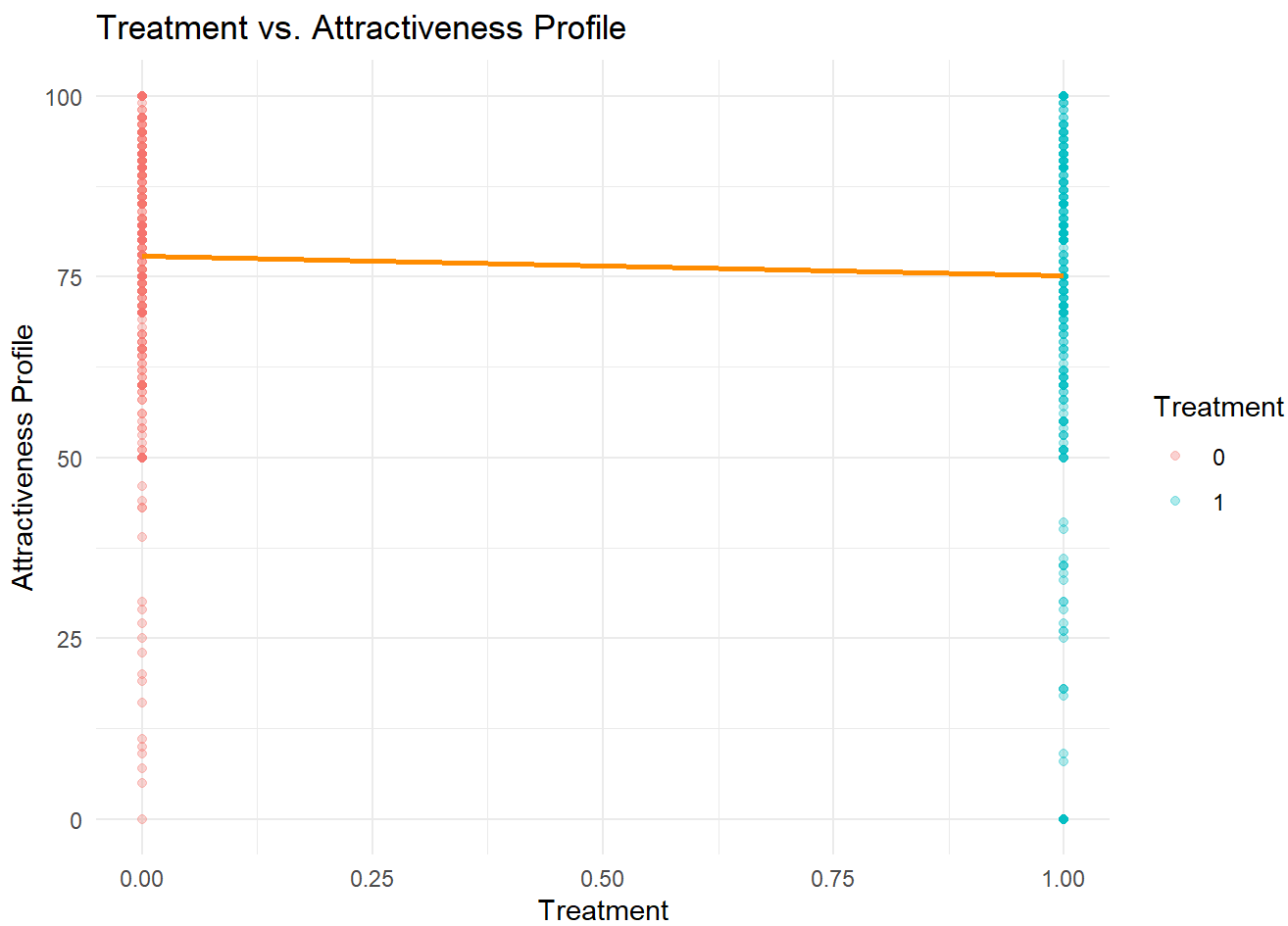
(c) Linear Regression

```
model <- lm(attractiveness_score ~ treatment_republican_profile, data = easton)
summary(model)
```

```
Call:
lm(formula = attractiveness_score ~ treatment_republican_profile,
    data = easton)

Residuals:
    Min      1Q  Median      3Q     Max
-77.870  -7.083   4.130  14.130  24.917

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                     77.870      1.064  73.208   <2e-16 ***
treatment_republican_profile    -2.788      1.503  -1.854   0.0641 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 20.24 on 723 degrees of freedom
Multiple R-squared:  0.004733,  Adjusted R-squared:  0.003357
F-statistic: 3.439 on 1 and 723 DF,  p-value: 0.0641

predicted_easton <- data.frame(attractiveness_score =  predict(model), treatment_republican_profile =
         easton$treatment_republican_profile)
```

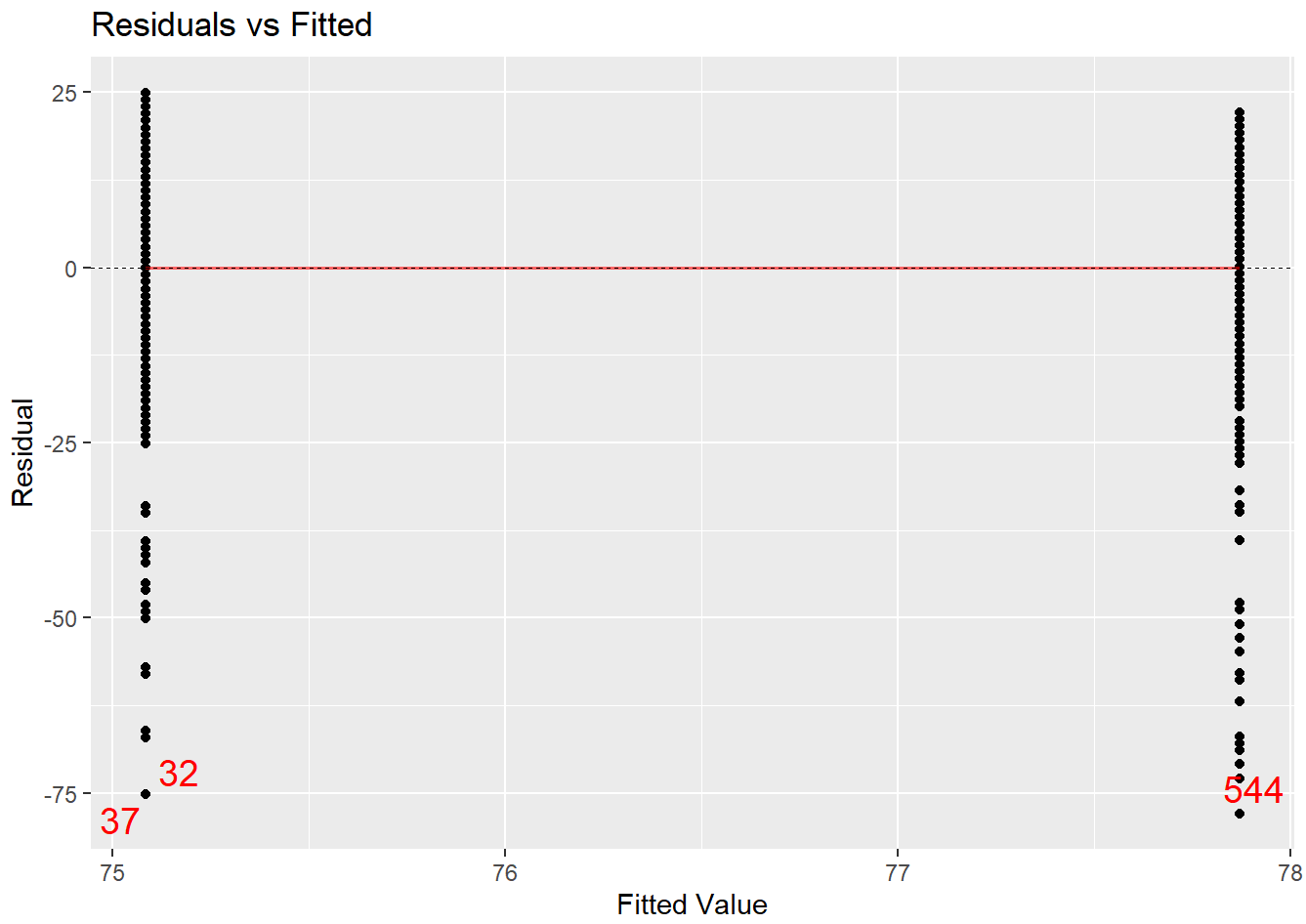
```
ggplot(easton, aes(x = treatment_republican_profile, y = attractiveness_score)) +
  geom_point(aes(x = treatment_republican_profile, y = attractiveness_score, color =
         factor(treatment_republican_profile)),
             alpha = 0.3) +
geom_line(data = predicted_easton,
            aes(x = treatment_republican_profile, y = attractiveness_score),color='darkorange', lwd=
         1) +
  labs(
    title = "Treatment vs. Attractiveness Profile",
    x = "Treatment",
    y = "Attractiveness Profile",
    color = "Treatment") +
  theme_minimal()
```



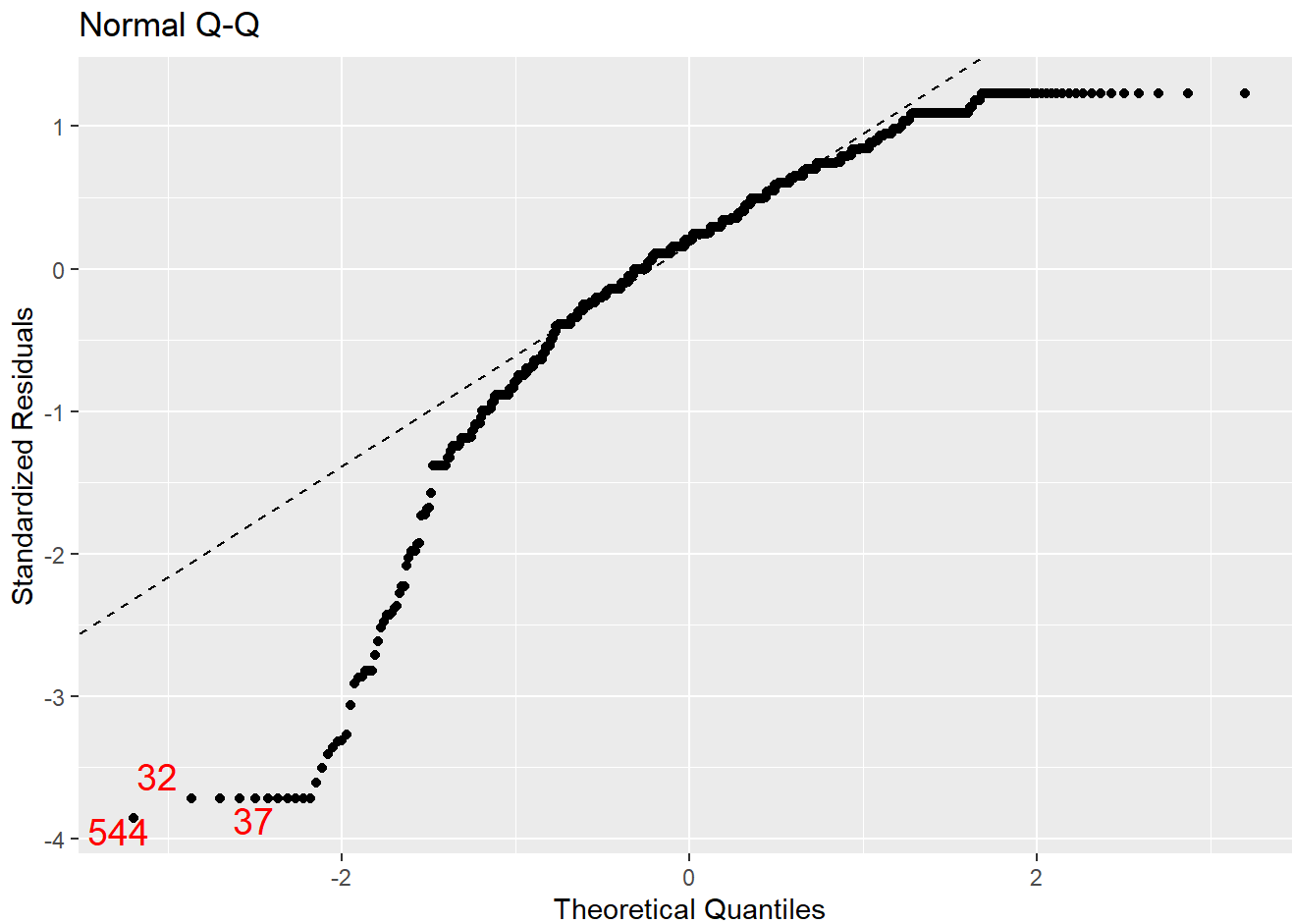
The estimates from my linear regression are the same as those calculated by sub-sample means in the average treatment effect (ATE). Furthermore, while the ATE measures the difference in sub-sample means, the slope (or coefficient beta) of the line of best fit is roughly the same as the ATE.

(d) Diagnostic Plots

```
mplot(model, which = 1:2)

[[1]]

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

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: pseudoinverse used at 75.069

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: neighborhood radius 2.8015

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: reciprocal condition number 0

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: There are other near singularities as well. 7.8482

Warning in predLoess(object$y, object$x, newx = if (is.null(newdata)) object$x
else if (is.data.frame(newdata))
as.matrix(model.frame(delete.response(terms(object)), : pseudoinverse used at
75.069

Warning in predLoess(object$y, object$x, newx = if (is.null(newdata)) object$x
else if (is.data.frame(newdata))
as.matrix(model.frame(delete.response(terms(object)), : neighborhood radius
2.8015

Warning in predLoess(object$y, object$x, newx = if (is.null(newdata)) object$x
else if (is.data.frame(newdata))
as.matrix(model.frame(delete.response(terms(object)), : reciprocal condition
number 0

Warning in predLoess(object$y, object$x, newx = if (is.null(newdata)) object$x
else if (is.data.frame(newdata))
as.matrix(model.frame(delete.response(terms(object)), : There are other near
singularities as well. 7.8482
```

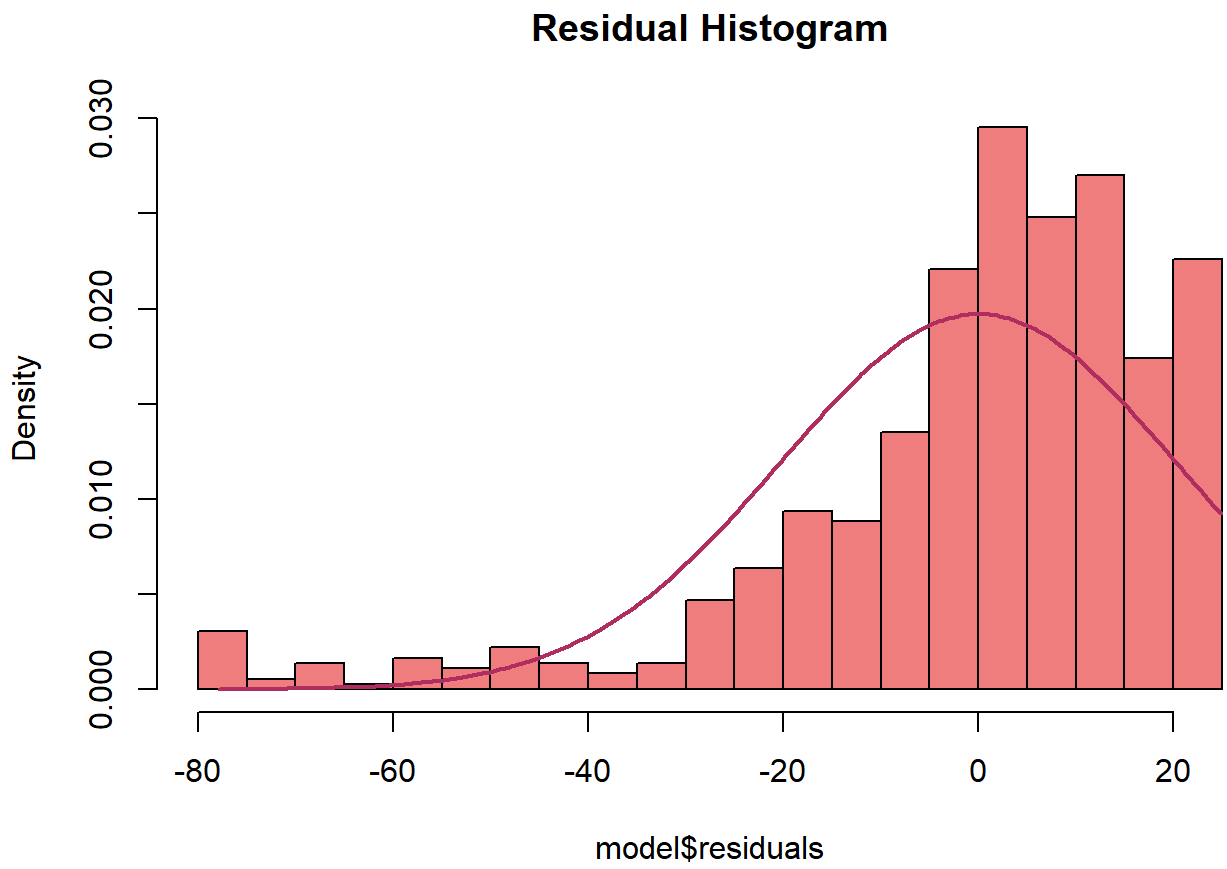

Residuals vs Fitted
Residual
Fitted Value
25
0
-25
-50
-75
75
76
77
78
32
37
544

```
[[2]]
```



By looking at the residuals vs fitted plot, we see that the points are generally centered around 0, as according to the line of best fit. Furthermore, looking at the Normal Q-Q plot, we see that the tails diverge from the line at either extreme. Therefore, while this model meets the assumption of mean zero, it does not meet the assumption of homoscedasticity by looking at both models respectively.

```
hist(model$residuals, prob = TRUE, breaks = 20, col = "lightcoral", main = "Residual Histogram")

grid = sort(model$residuals)
lines(grid,
      dnorm(grid,
            mean = mean(model$residuals),
            sd = sd(model$residuals)),
      col = 'maroon', lwd = 2 )
```



Finally, looking at the residual histogram plot, we see that the histogram follows the density line. Therefore, we can suppose that the residuals are normally distributed.

(e) Conditional Average Treatment Effect (CATE)

```
cate <- easton %>%
  group_by(republican) %>%
  summarise(
    N_Treated = sum(treatment_republican_profile == 1),
    N_Control = sum(treatment_republican_profile == 0),
    Mean_Treated = mean(attractiveness_score[treatment_republican_profile == 1]),
    Mean_Control = mean(attractiveness_score[treatment_republican_profile == 0]),
    CATE = Mean_Treated - Mean_Control
  ) %>%
  ungroup()

print(cate)

# A tibble: 2 × 6
  republican N_Treated N_Control Mean_Treated Mean_Control  CATE
       <int>     <int>     <int>        <dbl>        <dbl> <dbl>
1          0       236       236         71.9         78.6 -6.70
2          1       127       126         81.0         76.5  4.50
```

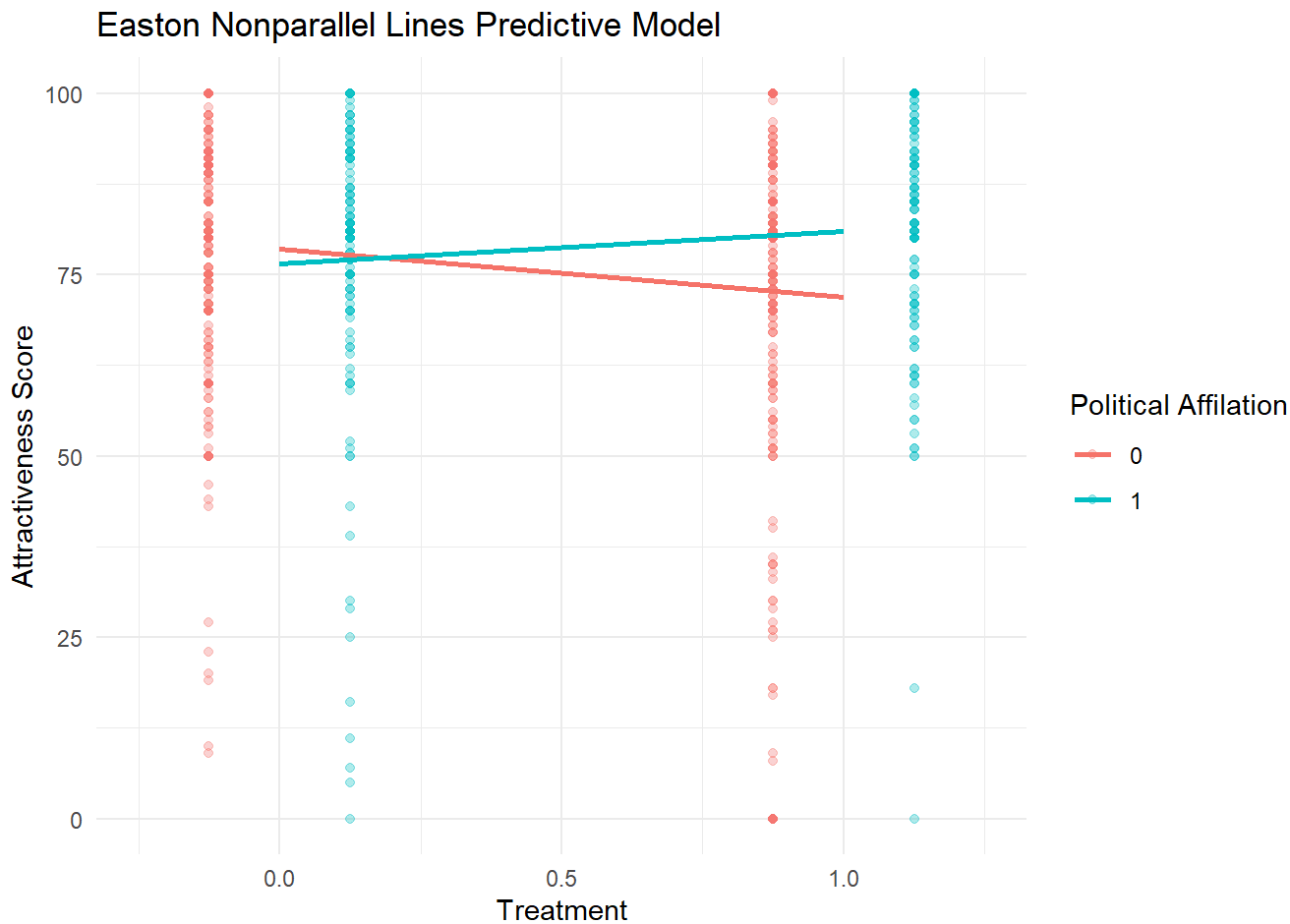
(f) Linear Regression Scatter Plot w/ Interaction

```
model <- lm(attractiveness_score ~ treatment_republican_profile + republican +
         treatment_republican_profile*republican, data = easton)
summary(model)
```

```
Call:
lm(formula = attractiveness_score ~ treatment_republican_profile +
    republican + treatment_republican_profile * republican, data = easton)

Residuals:
    Min      1Q  Median      3Q     Max
-81.024  -7.886   3.476  13.114  28.114

Coefficients:
                                       Estimate Std. Error t value Pr(>|t|)
(Intercept)                              78.589      1.303  60.315  < 2e-16
treatment_republican_profile             -6.703      1.843  -3.638 0.000295
republican                               -2.065      2.209  -0.935 0.350057
treatment_republican_profile:republican  11.203      3.119   3.592 0.000351

(Intercept)                             ***
treatment_republican_profile            ***
republican
treatment_republican_profile:republican ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.02 on 721 degrees of freedom
Multiple R-squared:  0.02908,   Adjusted R-squared:  0.02504
F-statistic: 7.199 on 3 and 721 DF,  p-value: 9.142e-05
```

```
republican_seq <- seq(min(easton$republican), max(easton$republican), by = 1)

pred_data <- expand.grid(
  republican = republican_seq,
  treatment_republican_profile = c(0, 1))

pred_data$treatment_republican_profile <- as.numeric(pred_data$treatment_republican_profile)

pred_data$predicted_score <- predict(model, newdata = pred_data)

ggplot() +
  geom_point(data = easton, aes(x = treatment_republican_profile, y = attractiveness_score, color =
         factor(republican)),
             alpha = 0.3, position = position_dodge(width = 0.5)) +
  geom_line(data = pred_data, aes(x = treatment_republican_profile, y = predicted_score,
                                  color = factor(republican)), size = 1) +
  labs(
    title = "Easton Nonparallel Lines Predictive Model",
    x = "Treatment",
    y = "Attractiveness Score",
    color = "Political Affilation"
  ) +
  theme_minimal()
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
ℹ Please use `linewidth` instead.
```



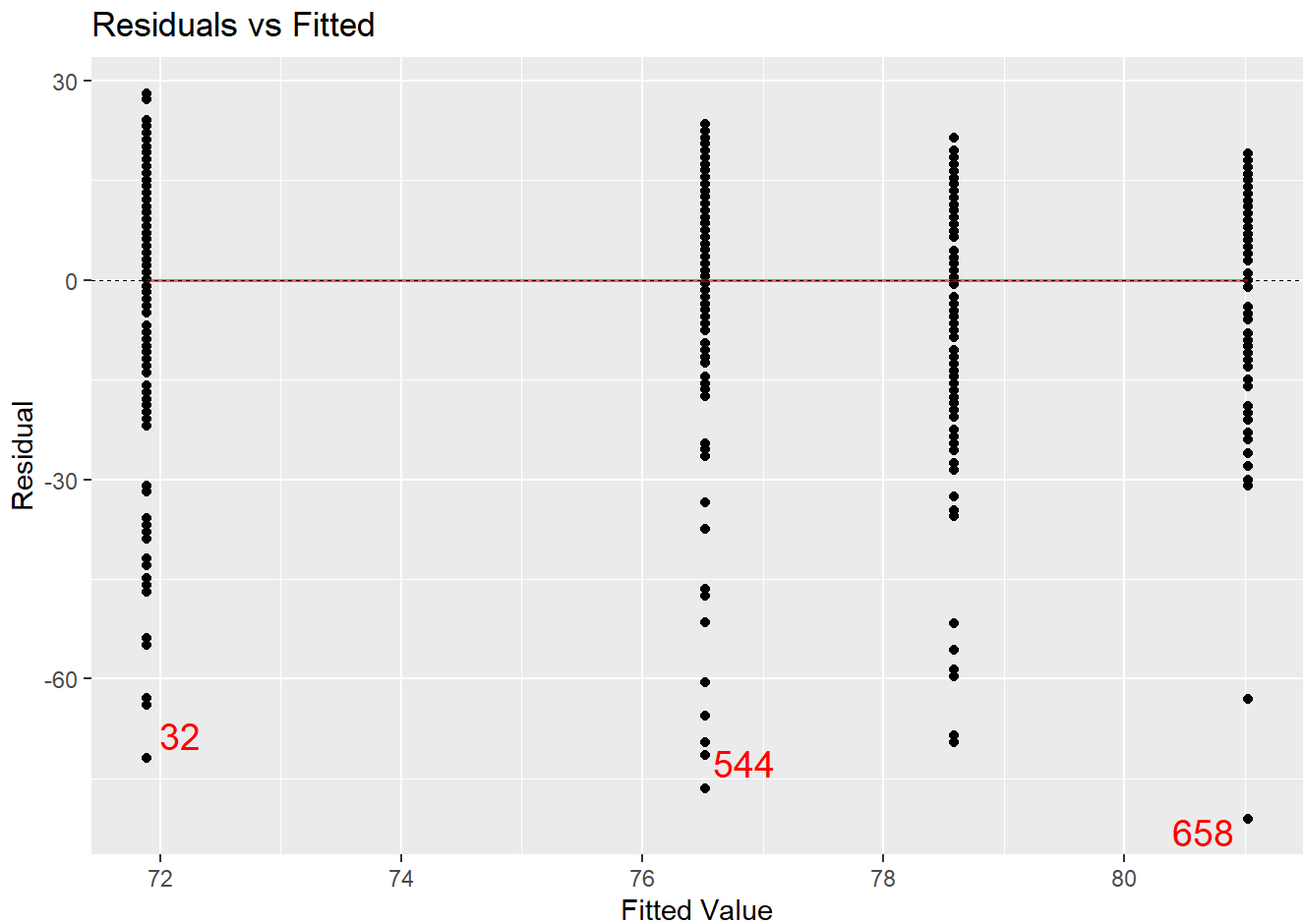
The CATEs are the difference between the values estimated by the linear regression conditioning for each political affiliation.

(g) Diagnostics Plots

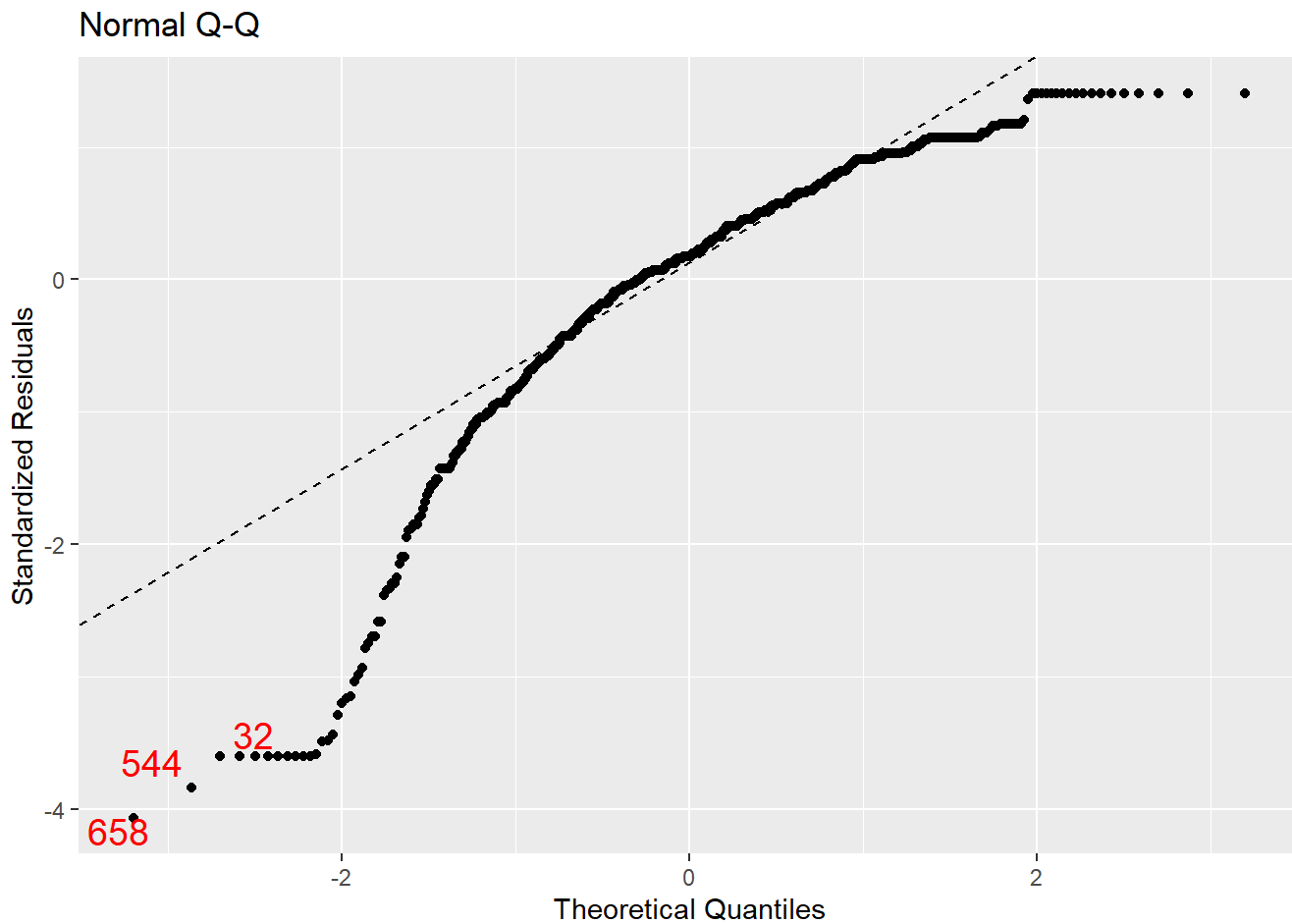
```
mplot(model, which = 1:2)
```

```
[[1]]

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

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: pseudoinverse used at 71.84

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: neighborhood radius 6.7491

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: reciprocal condition number 2.3644e-16

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: There are other near singularities as well. 44.935

Warning in predLoess(object$y, object$x, newx = if (is.null(newdata)) object$x
else if (is.data.frame(newdata))
as.matrix(model.frame(delete.response(terms(object)), : pseudoinverse used at
71.84

Warning in predLoess(object$y, object$x, newx = if (is.null(newdata)) object$x
else if (is.data.frame(newdata))
```

```
as.matrix(model.frame(delete.response(terms(object)), : neighborhood radius
6.7491

Warning in predLoess(object$y, object$x, newx = if (is.null(newdata)) object$x
else if (is.data.frame(newdata))
as.matrix(model.frame(delete.response(terms(object)), : reciprocal condition
number 2.3644e-16

Warning in predLoess(object$y, object$x, newx = if (is.null(newdata)) object$x
else if (is.data.frame(newdata))
as.matrix(model.frame(delete.response(terms(object)), : There are other near
singularities as well. 44.935
```

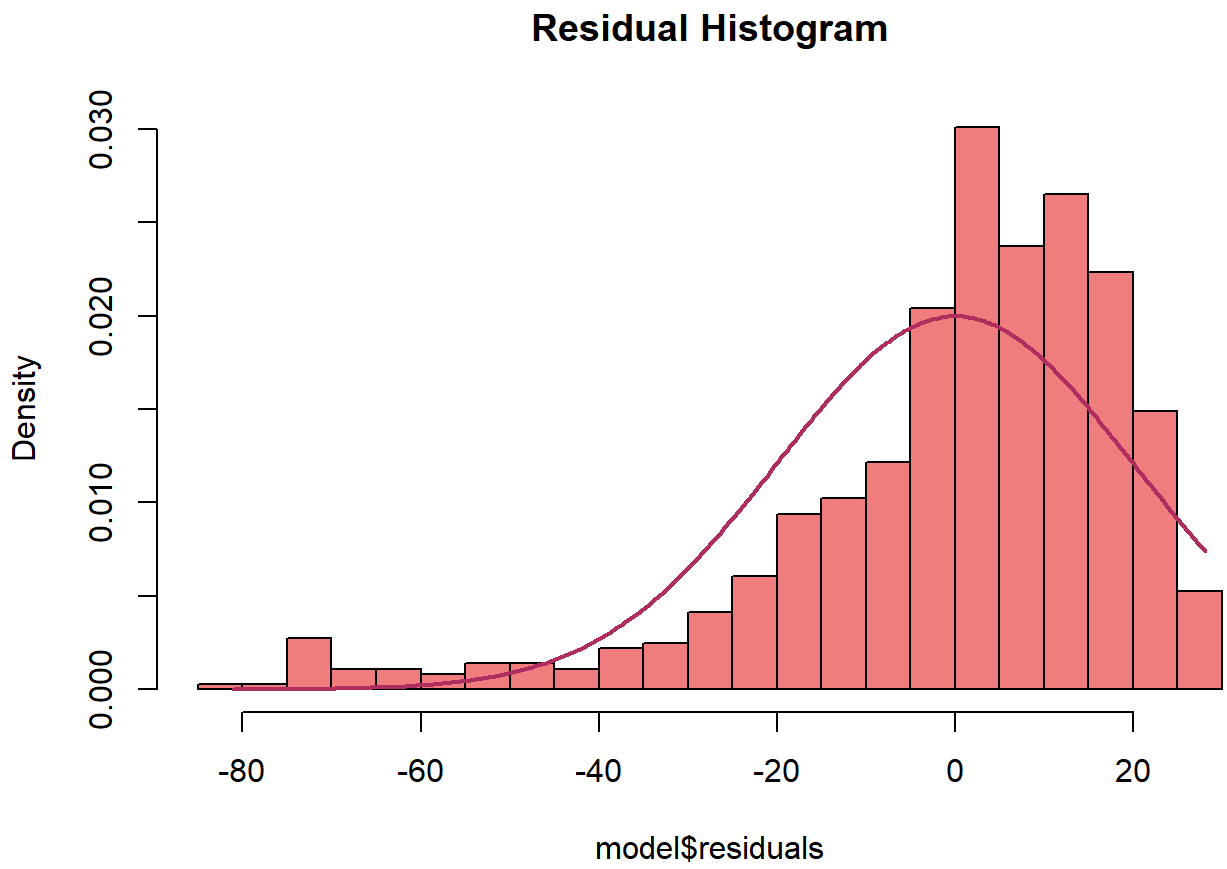



```
[[2]]
```



Looking at the Normal Q-Q plot, we see that the tails diverge from the line at either extreme, so the model does not meet the assumption of homoscedasticity. Also, looking at the Residuals vs Fitted plot, we see that the line of best fit is centered around zero, meaning that the model meets the mean zero assumption.

```
hist(model$residuals, prob = TRUE, breaks = 20, col = "lightcoral", main = "Residual Histogram")

grid = sort(model$residuals)
lines(grid,
      dnorm(grid,
            mean = mean(model$residuals),
            sd = sd(model$residuals)),
      col = 'maroon', lwd = 2 )
```



Finally, by looking at the Residual Histogram, we see that the histogram does not accurately follow the density line. Therefore, we can assume that the residuals of the model are not normally distributed.

(h) Linear Regression

Before looking at the estimates, I expect the coefficients on my model to be different from those in (f). This is because age is definitely a factor that some people may find attractive or unattractive. Therefore, it will likely affect the relationship between treatment and attractiveness score. I do think that the standard errors from my coefficients will be different from (f). It's possible that while age affects the relationship of interest, it might also add noise and affect the variance.

```
model <- lm(attractiveness_score ~ treatment_republican_profile + republican + age +
         treatment_republican_profile*republican, data = easton)
summary(model)
```

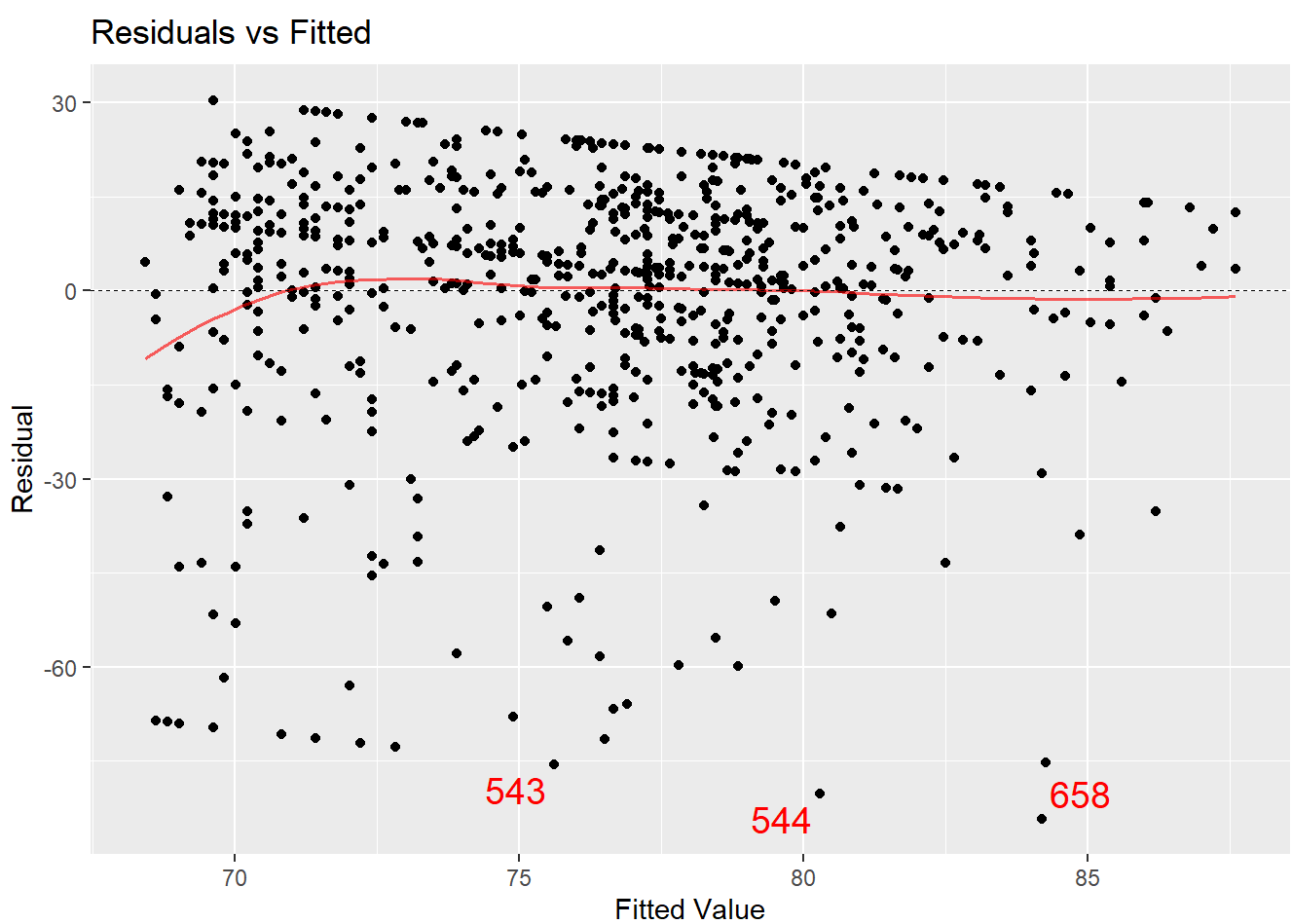
```
Call:
lm(formula = attractiveness_score ~ treatment_republican_profile +
    republican + age + treatment_republican_profile * republican,
    data = easton)

Residuals:
    Min      1Q  Median      3Q     Max
-84.197  -7.453   4.003  13.188  30.388

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                              71.4536     2.5847  27.645  < 2e-16
treatment_republican_profile             -6.8415     1.8316  -3.735 0.000202
```

```
republican                                -2.5597     2.2001  -1.163 0.245027
age                                        0.2000     0.0627   3.190 0.001486
treatment_republican_profile:republican   11.3447     3.1000   3.660 0.000271

(Intercept)                              ***
treatment_republican_profile             ***
republican
age                                      **
treatment_republican_profile:republican ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.89 on 720 degrees of freedom
Multiple R-squared:  0.04261,	Adjusted R-squared:  0.03729
F-statistic: 8.012 on 4 and 720 DF,  p-value: 2.54e-06
```

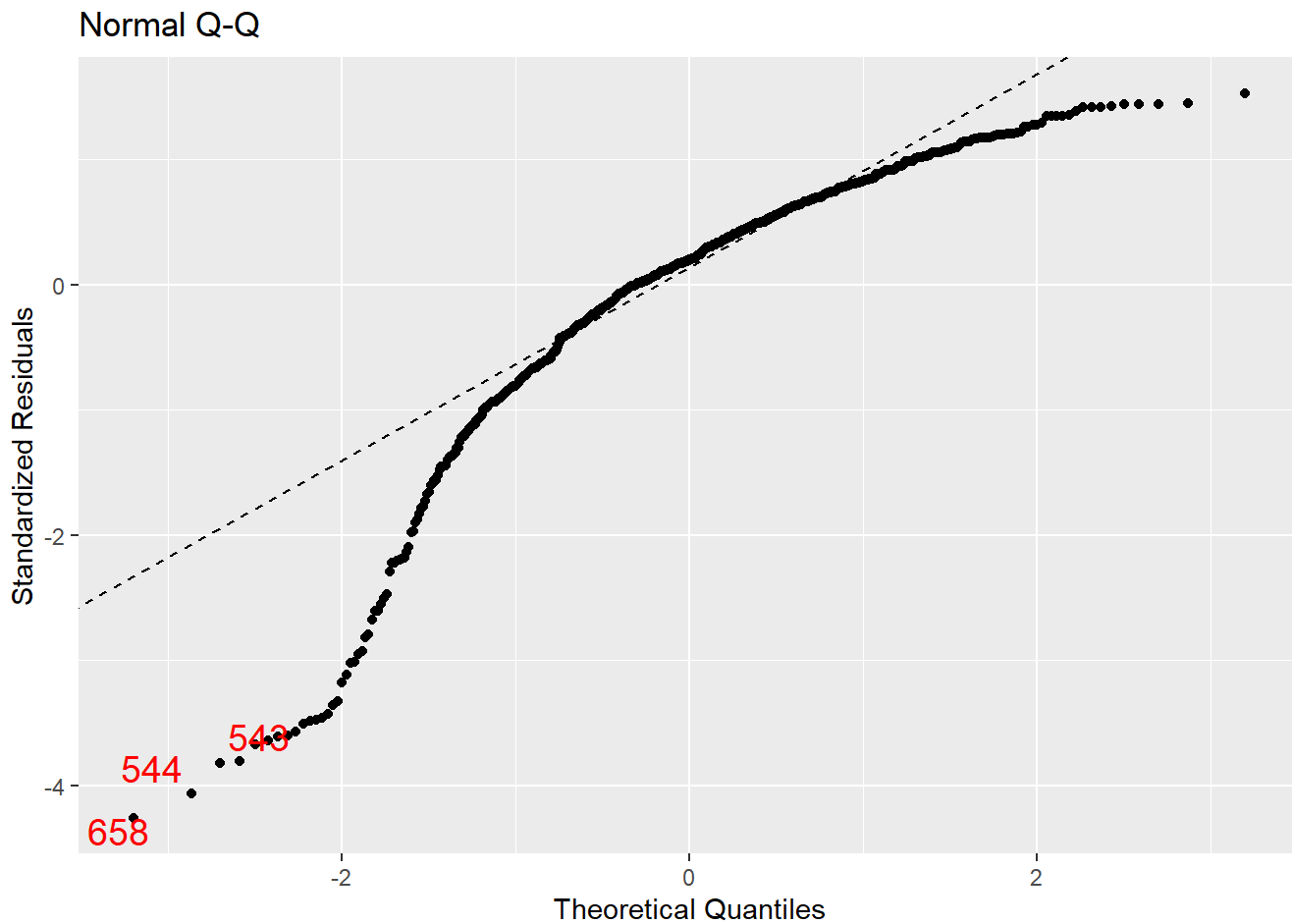
After looking at the results, the coefficients in this model are roughly the same as those in the previous model. This means that political affiliation likely does not affect the relationship between treatment and attractiveness score. Additionally, the standard errors in this model were very similar to the standard errors in the previous model. This is because the covariate of age likely has little to do with the relationship between treatment effect and attractiveness score.

```
mplot(model, which = 1:2)
```

```
[[1]]

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

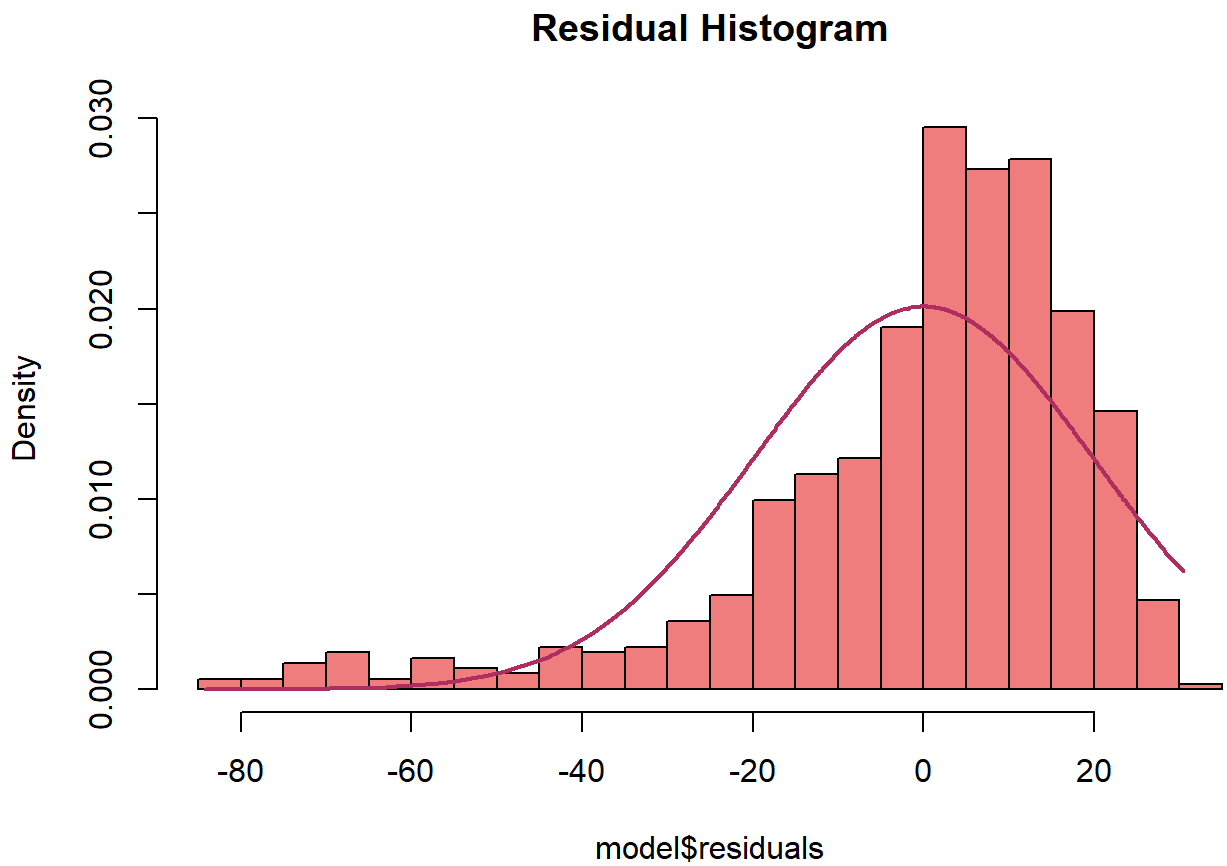



```
[[2]]
```



By looking at the Normal Q-Q plot, we see that we see that the tails diverge from the line at either extreme, so the model does not meet the assumption of homoscedasticity. Furthermore, looking at the Residuals vs Fitted model, we see that the points are scattered as random clouds of points falling within an area representing a horizontal band. Therefore, while the points are incredibly scattered they meet the assumption of mean zero as according to the line of best fit.

```
hist(model$residuals, prob = TRUE, breaks = 20, col = "lightcoral", main = "Residual Histogram")

grid = sort(model$residuals)
lines(grid,
      dnorm(grid,
            mean = mean(model$residuals),
            sd = sd(model$residuals)),
      col = 'maroon', lwd = 2 )
```



Finally, by looking at the Residual Histogram, we see that the histogram does not follow the density line. Therefore, we can suppose that the residuals are not normally distributed.

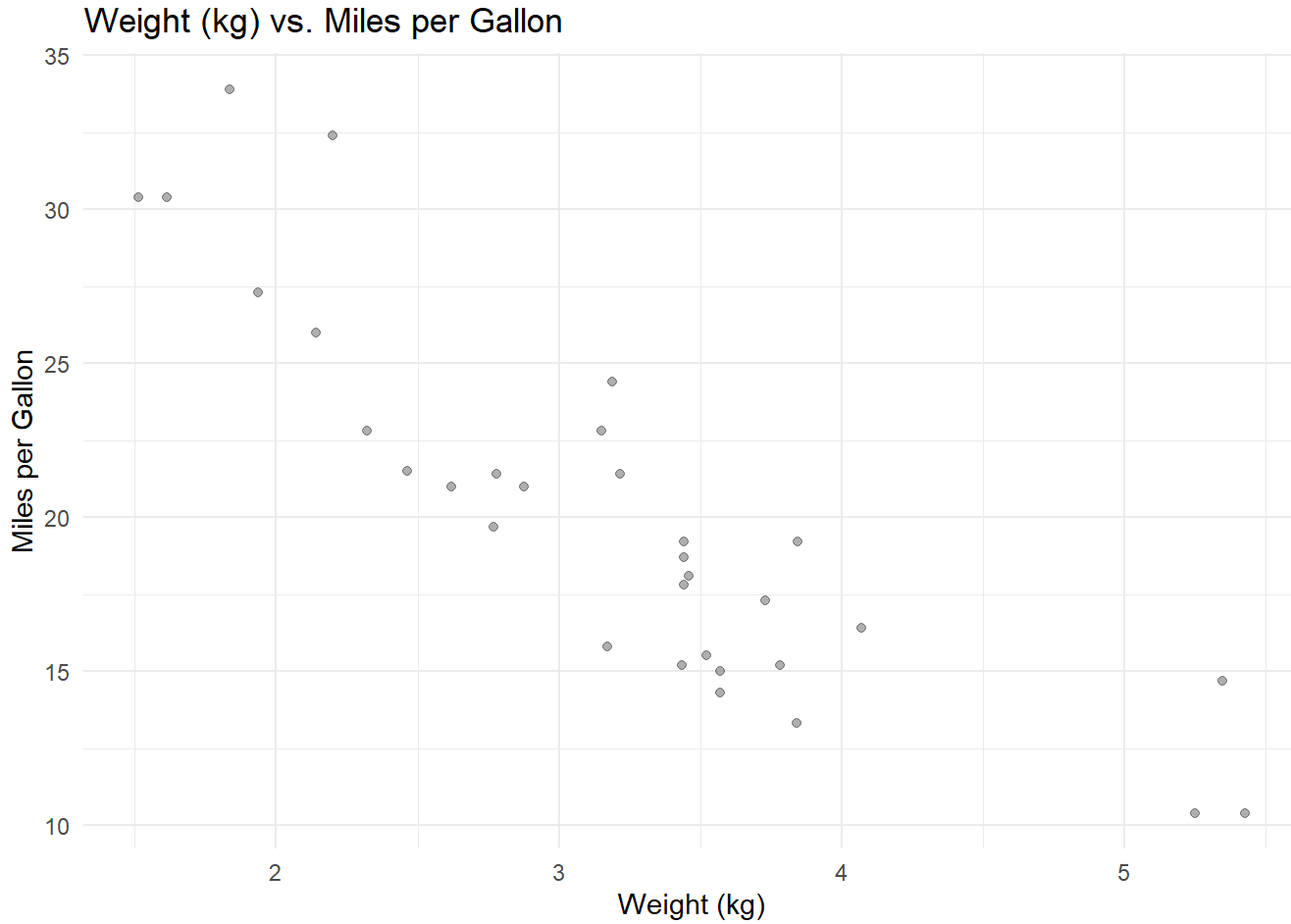
Exercise #2

```
data("mtcars")
head(mtcars)
```

```
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

(a) Scatter Plot

```
ggplot(mtcars, aes(x = wt, y = mpg)) +
  geom_point(aes(x = wt, y = mpg), alpha = 0.3) +
  labs(
    title = "Weight (kg) vs. Miles per Gallon",
    x = "Weight (kg)",
    y = "Miles per Gallon") +
  theme_minimal()
```



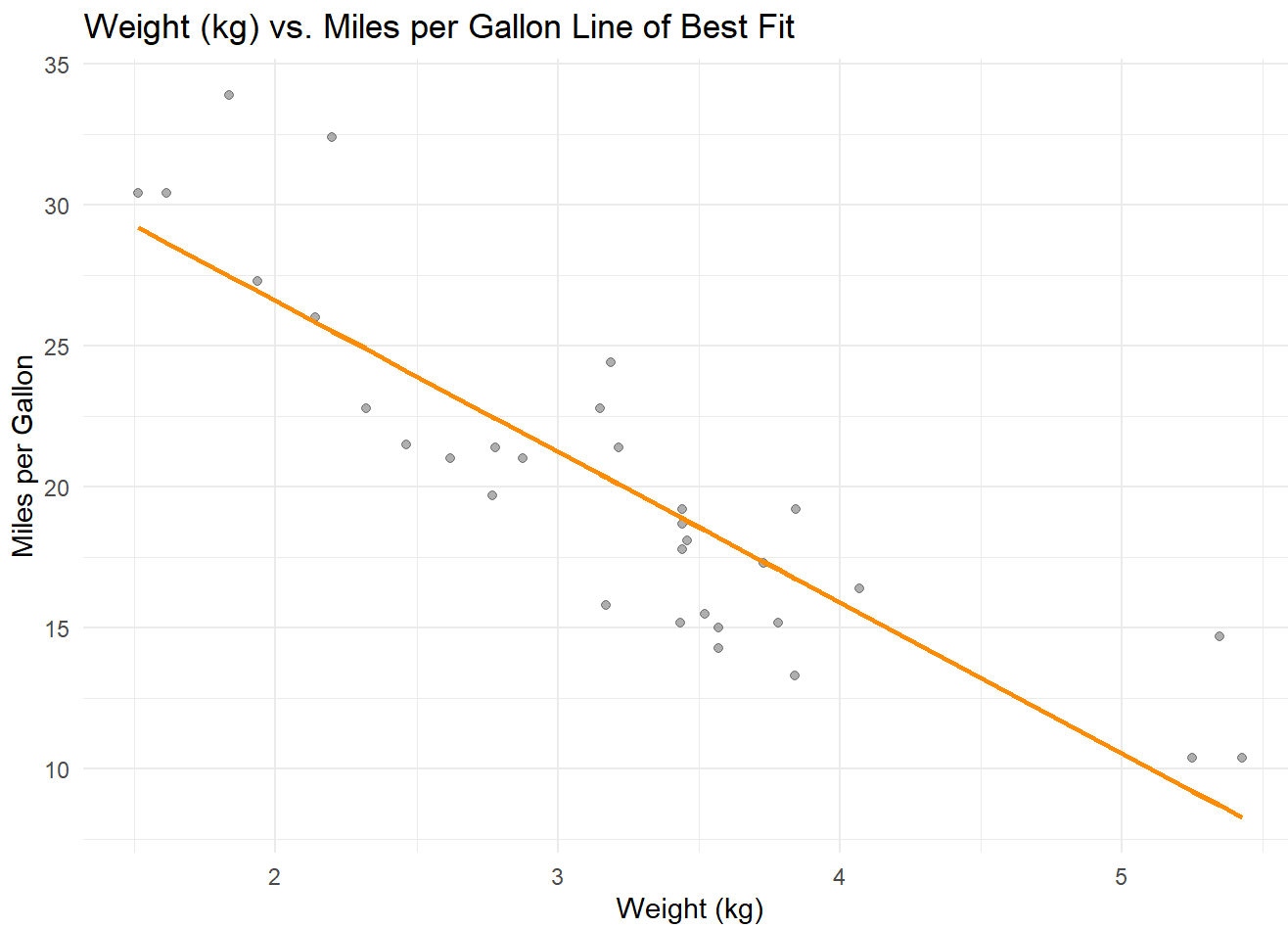
(b) Linear Regression Scatter Plot

```
model <- lm(mpg ~ wt, data = mtcars)
summary(model)
```

```
Call:
lm(formula = mpg ~ wt, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.5432 -2.3647 -0.1252  1.4096  6.8727 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
wt           -5.3445     0.5591  -9.559 1.29e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.046 on 30 degrees of freedom
Multiple R-squared:  0.7528,	Adjusted R-squared:  0.7446 
F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

```
predicted_mtcars <- data.frame(mpg =  predict(model), wt = mtcars$wt)

ggplot(mtcars, aes(x = wt, y = mpg)) +
  geom_point(aes(x = wt, y = mpg), alpha = 0.3) +
geom_line(data = predicted_mtcars,
            aes(x = wt, y = mpg),color='darkorange', lwd= 1) +
```

```
  labs(
    title = "Weight (kg) vs. Miles per Gallon Line of Best Fit",
    x = "Weight (kg)",
    y = "Miles per Gallon") +
  theme_minimal()
```



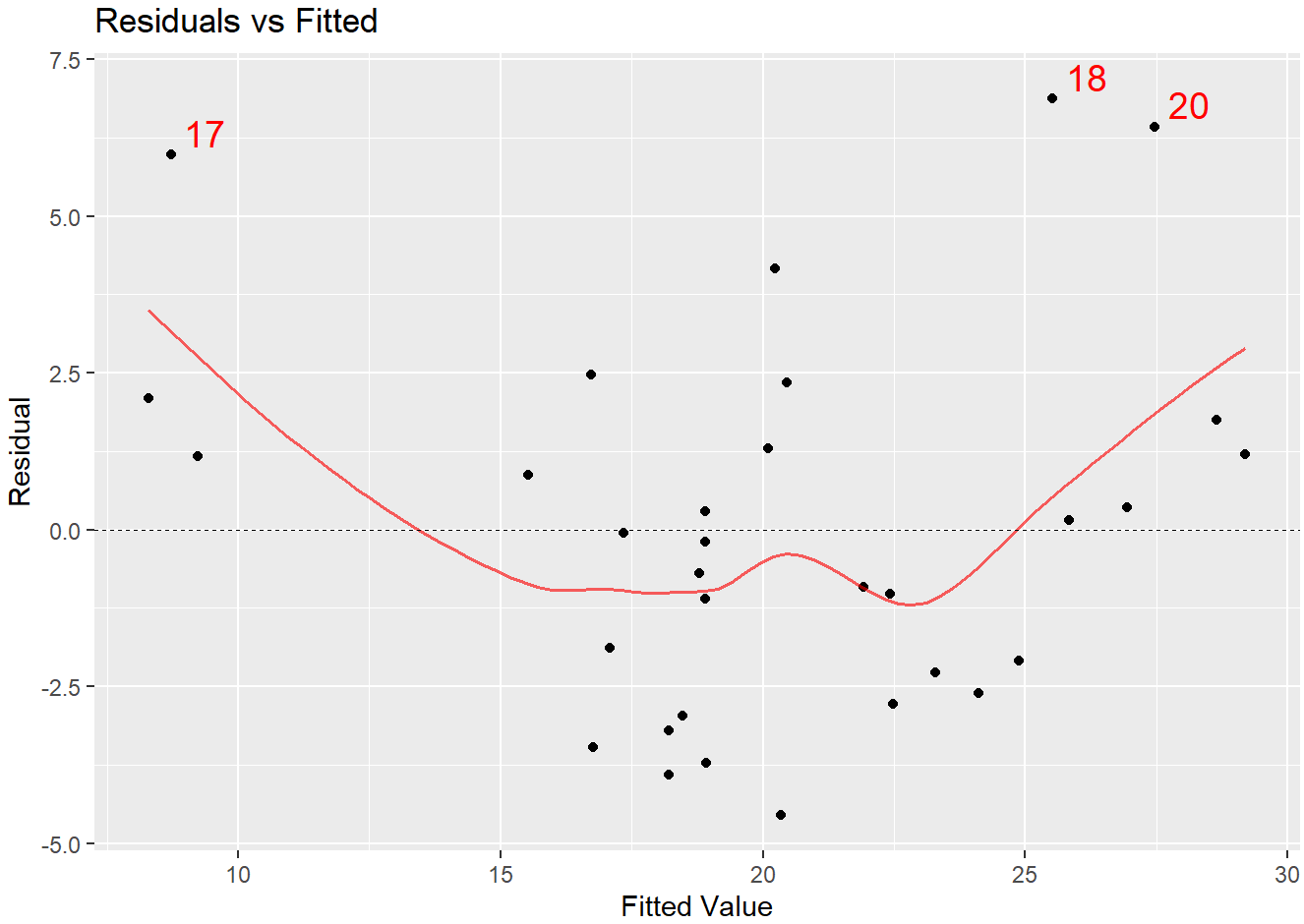
When looking at the linear regression, the first coefficient of 37.2851 is the intercept of the line. Therefore, this is the baseline value of the model. The following coefficient of -5.3445 is the slope of the line of best fit. In other words, on average the miles per gallon decreases by -5.3445 for each additional unit of weight (presumably kg).

(c) Diagnostics Plots

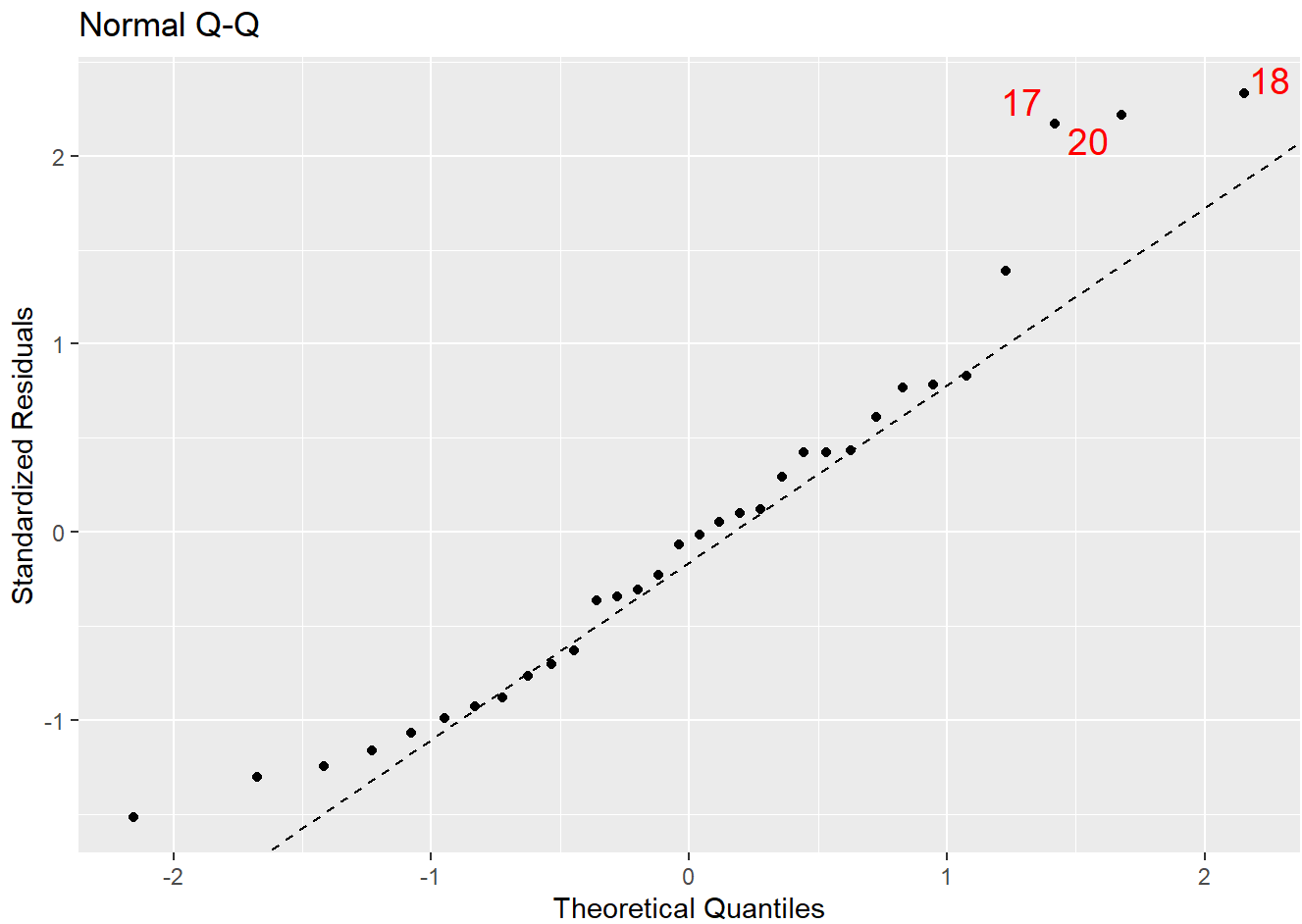
```
mplot(model, which = 1:2)
```

```
[[1]]
```

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

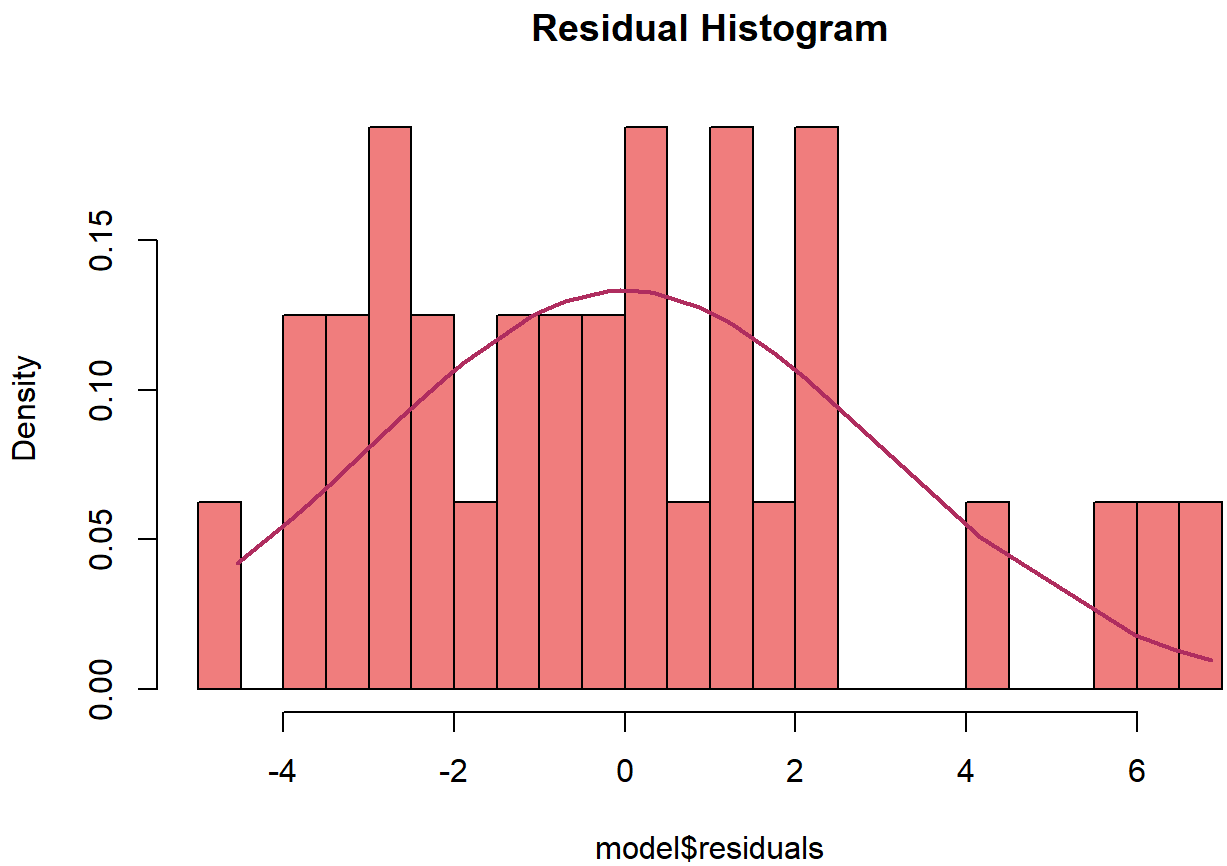

Residuals vs Fitted
Residual
Fitted Value
7.5
5.0
2.5
0.0
-2.5
-5.0
10
15
20
25
30
17
18
20

```
[[2]]
```



When looking at the Normal Q-Q plot, we see that the points are following the line with the exception of a few outliers. Therefore, this model meets the assumption of homoscedasticity if we chose to exclude those points. Furthermore, looking at the Residuals vs Fitted plot, we see that the lines are not all around 0. Therefore, this model does not meet the assumption of mean zero.

```
hist(model$residuals, prob = TRUE, breaks = 20, col = "lightcoral", main = "Residual Histogram")

grid = sort(model$residuals)
lines(grid,
      dnorm(grid,
            mean = mean(model$residuals),
            sd = sd(model$residuals)),
      col = 'maroon', lwd = 2 )
```



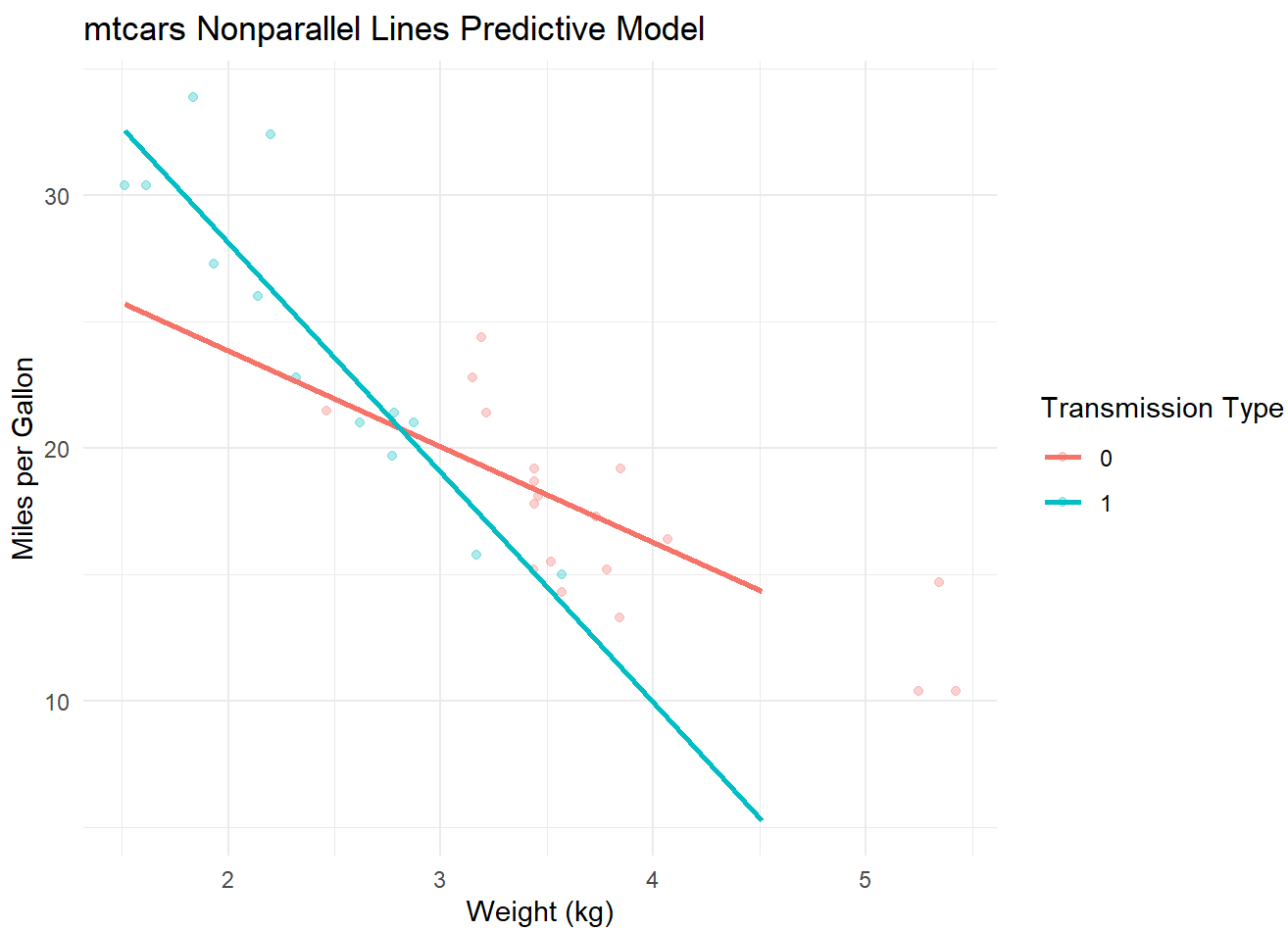
Finally, looking at the Residual Histogram, we see that the histogram does not follow the density line. Therefore we can suppose that the residuals are not normally distributed. The graph is closer to a uniform distribution.

(d) Linear Regression Scatter Plot w/ Interaction

```
model <- lm(mpg ~ wt + am + wt*am, data = mtcars)
summary(model)
```

```
Call:
lm(formula = mpg ~ wt + am + wt * am, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.6004 -1.5446 -0.5325  0.9012  6.0909 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  31.4161     3.0201  10.402 4.00e-11 ***
wt           -3.7859     0.7856  -4.819 4.55e-05 ***
am           14.8784     4.2640   3.489  0.00162 ** 
wt:am        -5.2984     1.4447  -3.667  0.00102 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.591 on 28 degrees of freedom
Multiple R-squared:  0.833, Adjusted R-squared:  0.8151 
F-statistic: 46.57 on 3 and 28 DF,  p-value: 5.209e-11
```

```
wt_seq <- seq(min(mtcars$wt), max(mtcars$wt), by = 1)

pred_data <- expand.grid(
  wt = wt_seq,
  am = c(0,1))

pred_data$predicted_mpg <- predict(model, newdata = pred_data)

ggplot() +
  geom_point(data = mtcars, aes(x = wt, y = mpg, color = factor(am)),
             alpha = 0.3) +
  geom_line(data = pred_data, aes(x = wt, y = predicted_mpg, color = factor(am)), size = 1) +
  labs(
    title = "mtcars Nonparallel Lines Predictive Model",
    x = "Weight (kg)",
    y = "Miles per Gallon",
    color = "Transmission Type"
  ) +
  theme_minimal()
```

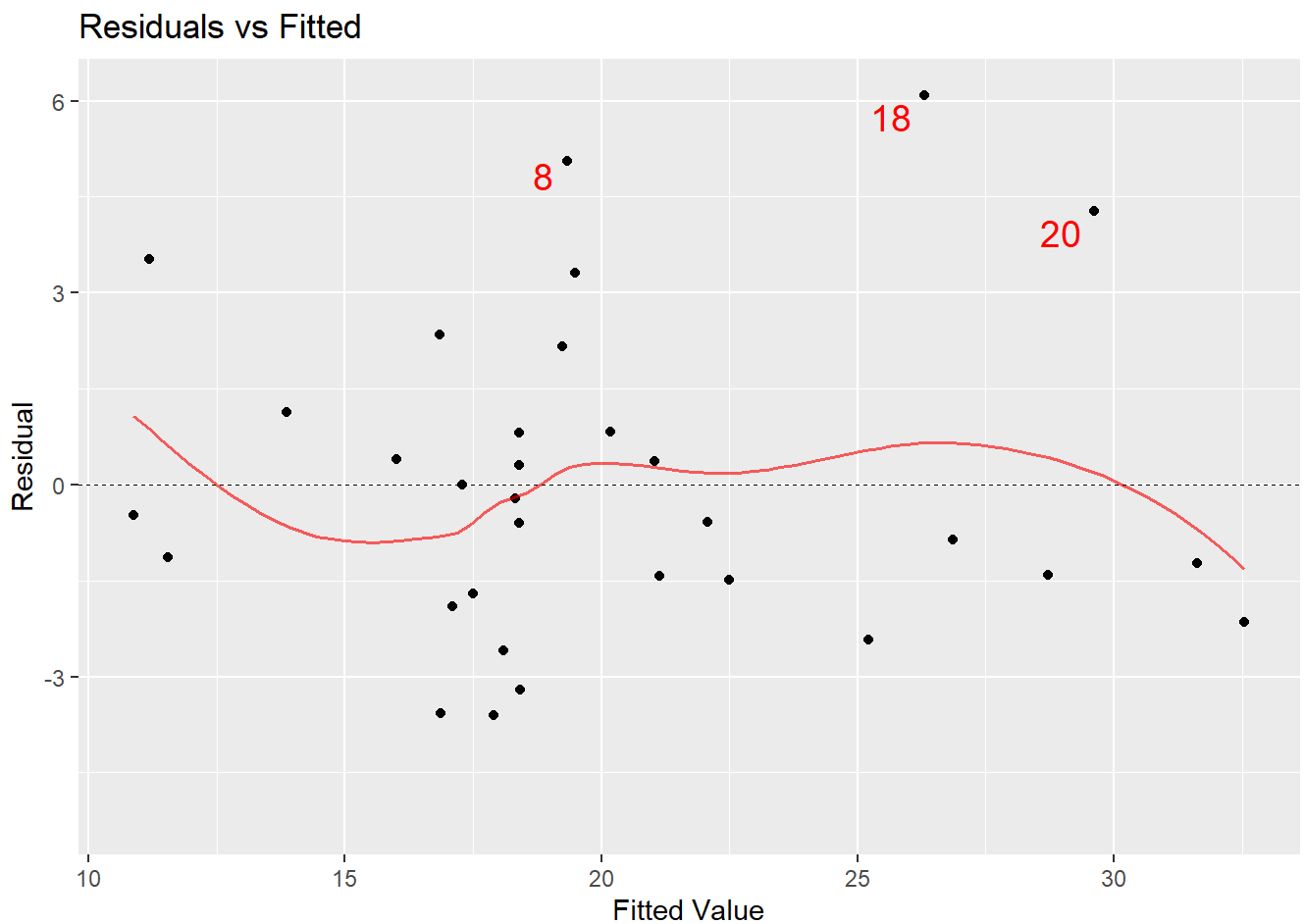


Looking at the nonparallel lines model, we see that the relationship between weight and miles per gallon is more negatively affected by the an automatic transmission type than a manual transmission type. This is because the slope for an automatic is steeper than the slope for the manual.

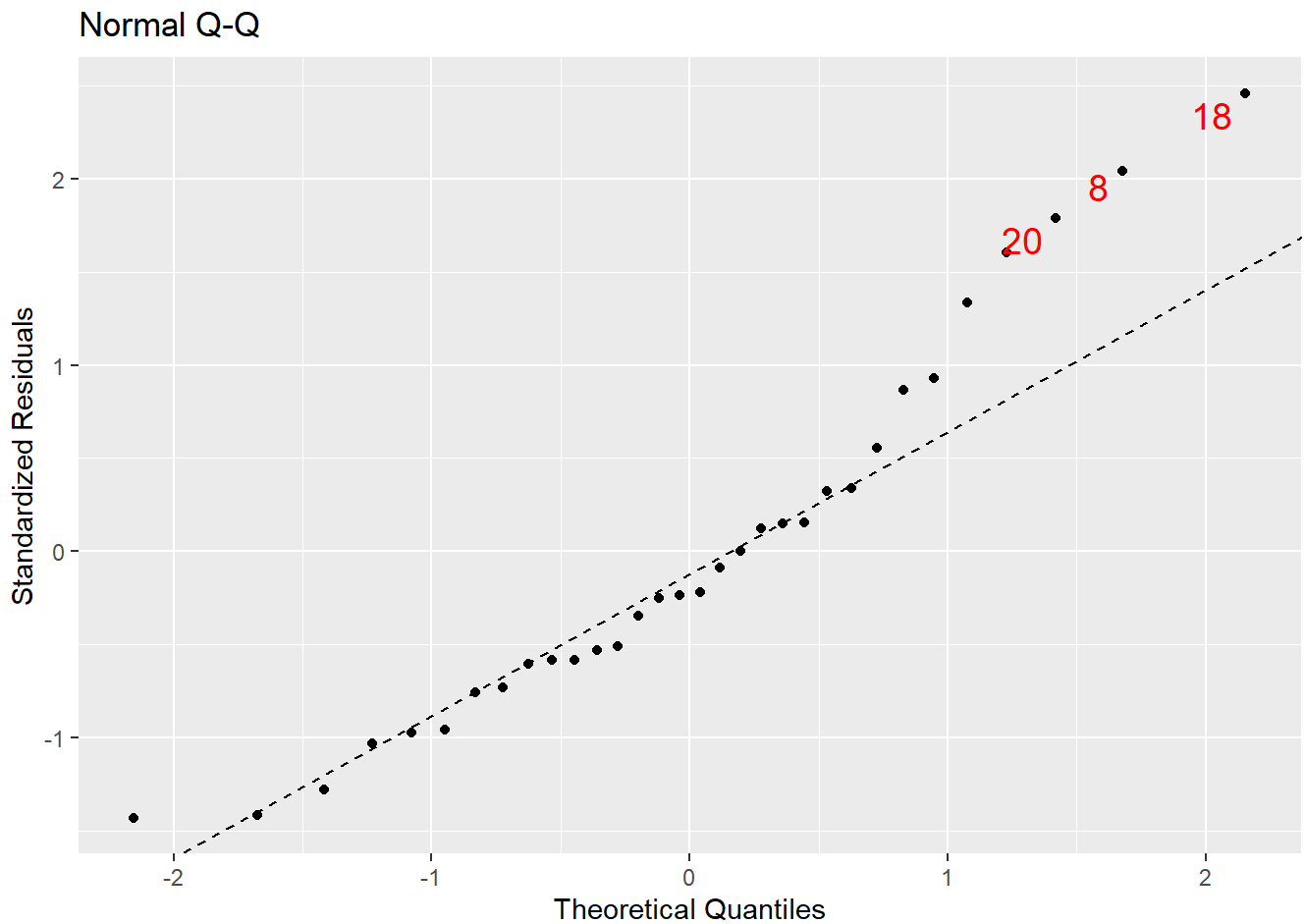
(e) Diagnostic Plots

```
mplot(model, which = 1:2)
```

```
[[1]]

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

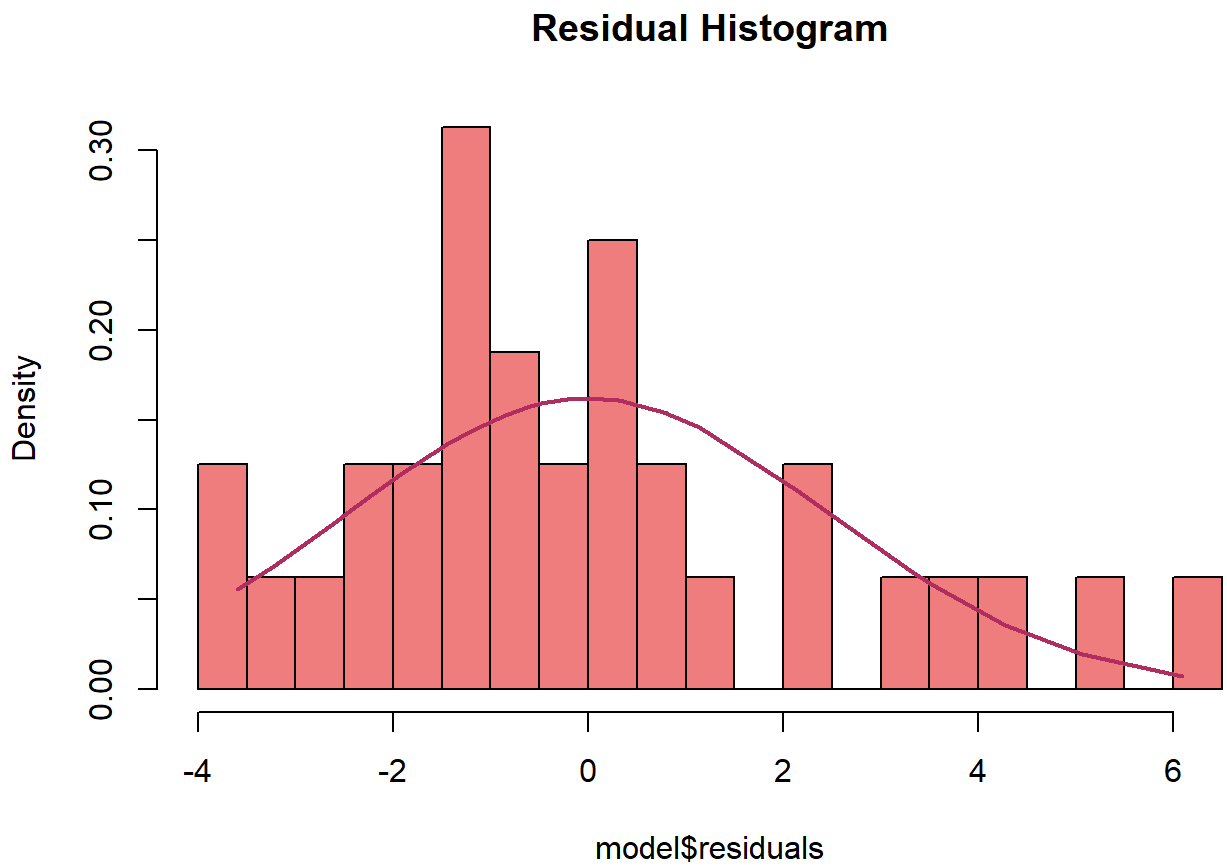



```
[[2]]
```



Looking at the Normal Q-Q plot, we see that the points follow the line but trail off at the extreme. Therefore, this model does not meet the assumption of homoscedasticity. Furthermore, looking at the Residuals vs Fitted Lines plot, we see that the points are generally close to zero, meaning that this model meets the mean zero assumption.

```
hist(model$residuals, prob = TRUE, breaks = 20, col = "lightcoral", main = "Residual Histogram")

grid = sort(model$residuals)
lines(grid,
      dnorm(grid,
            mean = mean(model$residuals),
            sd = sd(model$residuals)),
      col = 'maroon', lwd = 2 )
```



Finally, looking at the Residual Histogram plot, we see that the histogram does not follow the density line. Therefore, we can suppose that the residuals are not normally distributed.

(f) Linear Regression Scatter Plot w/ Interaction

Before looking at the results, I expect the coefficients to be different from those in the previous model. This is because I expect miles per gallon to increase as horse power increases. Additionally, because I expect there to be a relationship between horse power and and miles per gallon, I expect the standard errors to differ.

```
model <- lm(mpg ~ wt + am + hp + wt*am, data = mtcars)
summary(model)
```

```
Call:
lm(formula = mpg ~ wt + am + hp + wt * am, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.0639 -1.3315 -0.9347  1.2180  5.0822 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 30.947333   2.723411  11.363 8.55e-12 ***
wt          -2.515586   0.844497  -2.979  0.00605 ** 
am          11.554813   4.023277   2.872  0.00784 ** 
hp          -0.026949   0.009796  -2.751  0.01048 *  
wt:am       -3.577910   1.442796  -2.480  0.01968 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

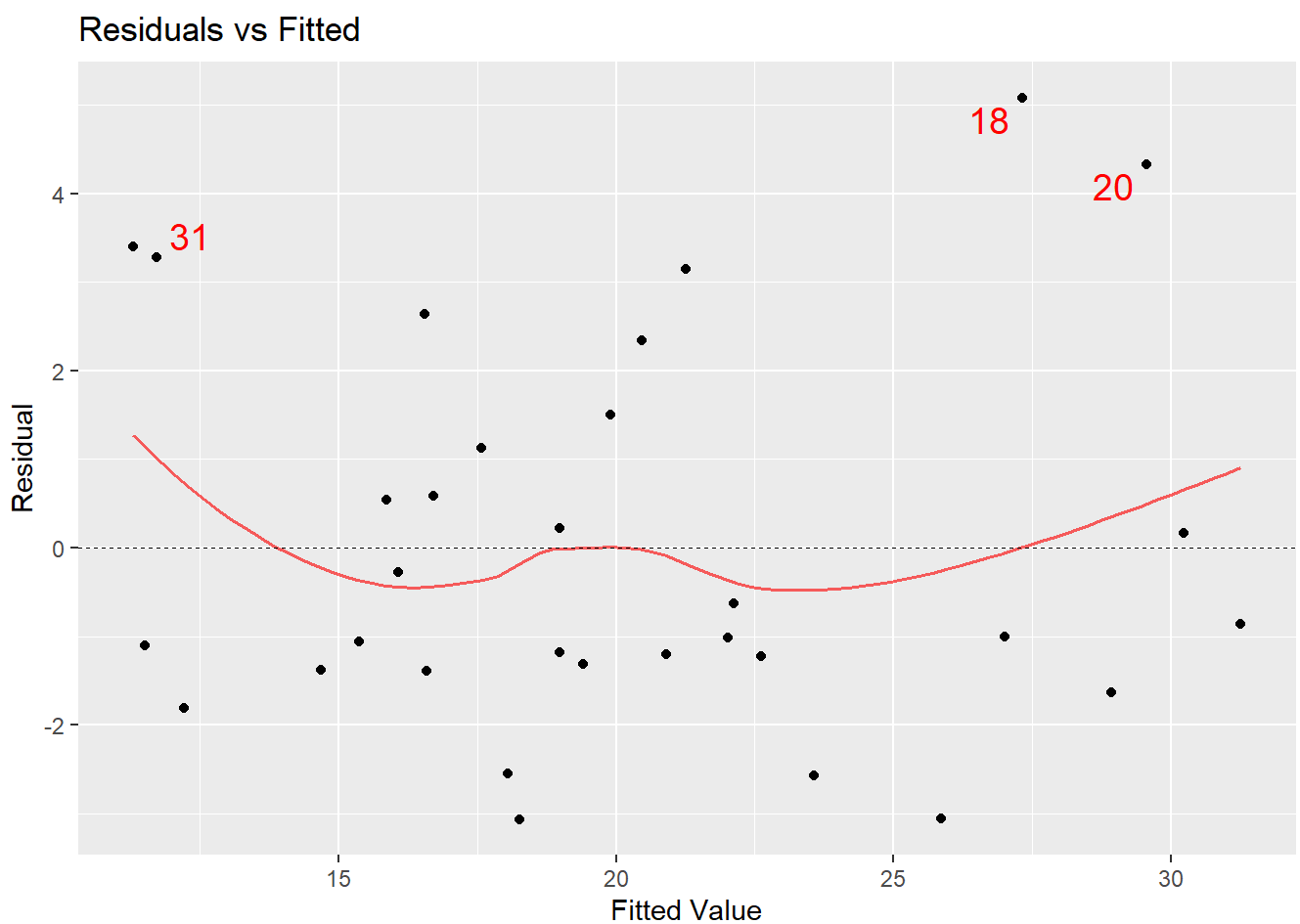
```
Residual standard error: 2.332 on 27 degrees of freedom
Multiple R-squared:  0.8696,    Adjusted R-squared:  0.8503
F-statistic: 45.01 on 4 and 27 DF,  p-value: 1.451e-11
```

After looking at the results, the intercept did not change by a lot, but the interaction variable changed and the am variable changed. Therefore, the horse power likely has a relationship with the transmission type which, in turn, effects the interaction variable. As such, the horse power variable likely has an indirect effect on the miles per gallon since the baseline did not feature significant change. Furthermore, the standard error did not change. This is likely because the horse power does not directly affect the other variables.

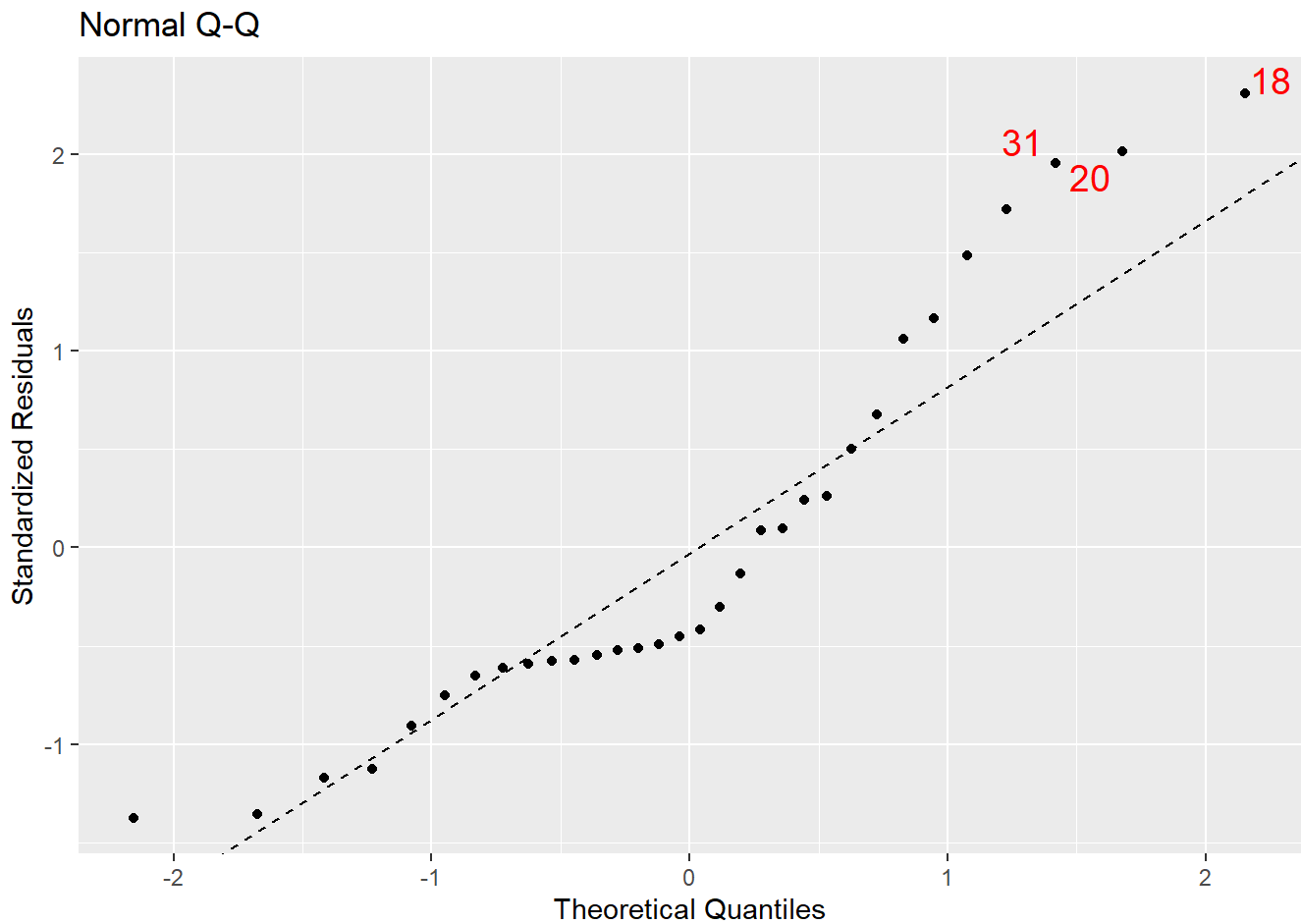
(g) Diagnostic Plots

```
mplot(model, which = 1:2)
```

```
[[1]]

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

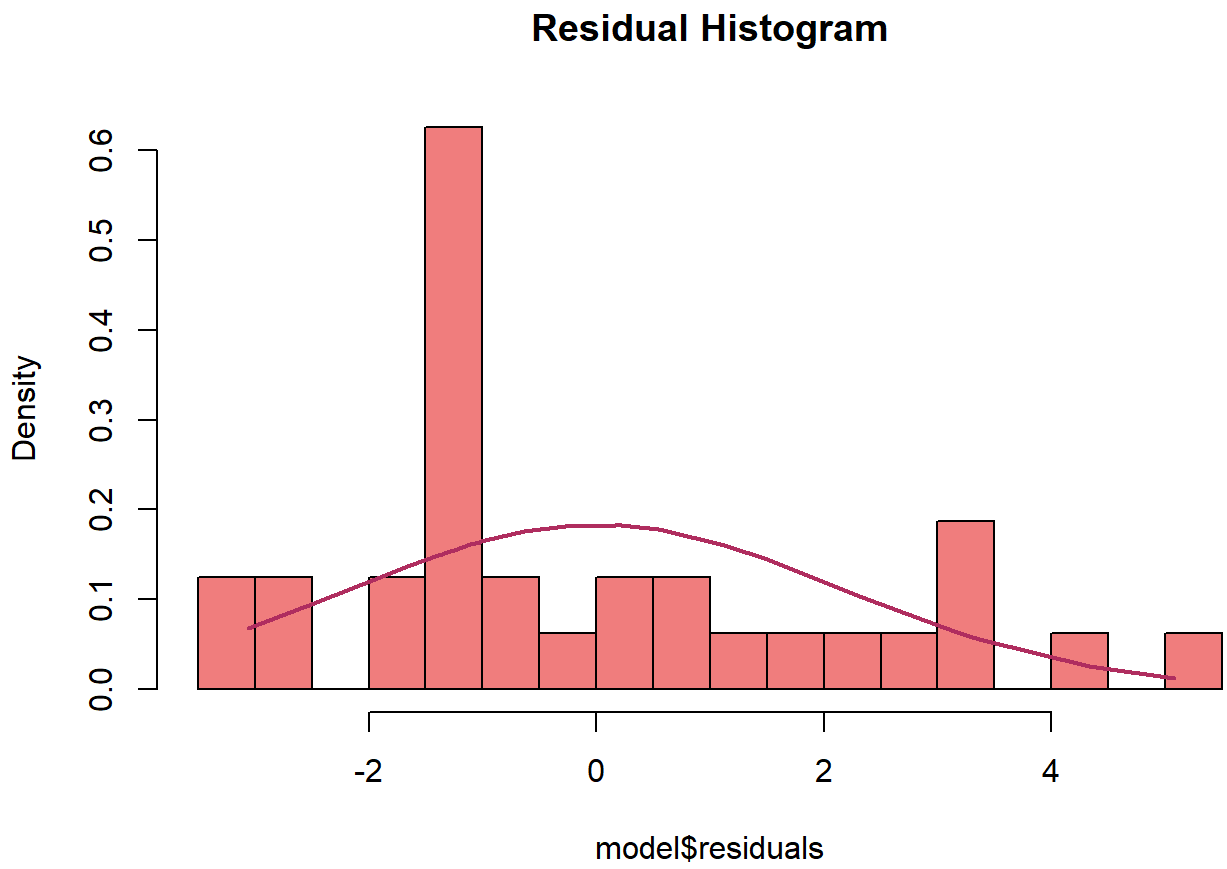



```
[[2]]
```



Looking at the Normal Q-Q plot, we see that the points do not follow the line. Therefore, this model does not meet the assumption of homoscedasticity. Furthermore, looking at the Residuals vs Fitted Plot, we see the points are generally around zero. Therefore, the model meets the mean zero assumption.

```
hist(model$residuals, prob = TRUE, breaks = 20, col = "lightcoral", main = "Residual Histogram")

grid = sort(model$residuals)
lines(grid,
      dnorm(grid,
            mean = mean(model$residuals),
            sd = sd(model$residuals)),
      col = 'maroon', lwd = 2 )
```



Finally looking at the Residual Histogram, we see that the histogram does not follow the density line. Therefore, we can suppose that the residuals are not normally distributed.

(h) Two Linear Models w/ 5-fold Validation

```
model <- lm(mpg ~ cyl + hp + cyl*hp, data = mtcars)
summary(model)
```

```
Call:
lm(formula = mpg ~ cyl + hp + cyl * hp, data = mtcars)

Residuals:
   Min     1Q Median     3Q    Max 
-4.778 -1.969 -0.228  1.403  6.491 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 50.751207   6.511686   7.794 1.72e-08 ***
cyl         -4.119140   0.988229  -4.168 0.000267 ***
hp          -0.170680   0.069102  -2.470 0.019870 *  
cyl:hp       0.019737   0.008811   2.240 0.033202 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.974 on 28 degrees of freedom
Multiple R-squared:  0.7801,	Adjusted R-squared:  0.7566 
F-statistic: 33.11 on 3 and 28 DF,  p-value: 2.386e-09
```

```
model <- lm(mpg ~ cyl + disp + cyl*disp, data = mtcars)
summary(model)
```

```
Call:
lm(formula = mpg ~ cyl + disp + cyl * disp, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.0809 -1.6054 -0.2948  1.0546  5.7981 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 49.037212   5.004636   9.798 1.51e-10 ***
cyl         -3.405244   0.840189  -4.053 0.000365 ***
disp        -0.145526   0.040002  -3.638 0.001099 ** 
cyl:disp     0.015854   0.004948   3.204 0.003369 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.66 on 28 degrees of freedom
Multiple R-squared:  0.8241,	Adjusted R-squared:  0.8052 
F-statistic: 43.72 on 3 and 28 DF,  p-value: 1.078e-10
```

For both models, the only variable I changed between them was the variable displacement and the horsepower. This is because I expect horsepower to indirectly affect the relationship between miles per gallon whereas displacement is arbitrary. Furthermore, the number of cylinders inversely correlates with fuel efficiency. Therefore, I expect miles per gallon to increase with the number of cylinders.

```
rss_summary <- function(data, lev = NULL, model = NULL) {
  residuals <- data$obs - data$pred
  rss <- sum(residuals^2)
  rmse <- sqrt(mean(residuals^2))
  return(c(RMSE = rmse, RSS = rss))
}

train_control_kfold <- trainControl(
  method = "cv",
  number = 5,
  summaryFunction = rss_summary,
  savePredictions = "final",
  classProbs = FALSE,
  allowParallel = FALSE)

# Train Model A: Parallel Lines Model
set.seed(123)
model_A_caret <- train(
  mpg ~ cyl + hp + cyl*hp,
  data = mtcars,
  method = "lm",
  trControl = train_control_kfold,
  metric = "RMSE")

# Train Model B: Nonparallel Lines Model
set.seed(123)
model_B_caret <- train(
  mpg ~ cyl + disp + cyl*disp,
  data = mtcars,
  method = "lm",
  trControl = train_control_kfold,
  metric = "RMSE")
```

```
model_A_caret$results
```

```
  intercept     RMSE      RSS   RMSESD    RSSSD
1      TRUE 3.148715 68.58221 1.041376 43.00386
```

```
model_B_caret$results
```

```
  intercept     RMSE      RSS    RMSESD    RSSSD
1      TRUE 3.062096 60.70225 0.6136897 20.43787
```

Looking at the RMSE values for each mode, I actually prefer model B. This is because the Mean Squared Error is less in model B than it is in model A. Meaning that horse power likely adds noise to the model and should not be included as a covariate.