CMDA-3654

Homework 6

Eduardo Salvador

Due as a .pdf upload

Instructions:

Delete the Instructions section from your write-up!!

I have given you this assignment as an .Rmd (R Markdown) file.

- Change the name of the file to: `Lastname_Firstname_CMDA_3654_HW6.Rmd`, and your output should therefore match but with a `.pdf` extension.
- You need to edit the R Markdown file by filling in the chunks appropriately with your code. Output will be generated automatically when you compile the document.
- You also need to add your own text before and after the chunks to explain what you are doing or to interpret the output.
- Feel free to add additional chunks if needed. I **will not** be providing assignments to you like this for the entire semester, just long enough for you to learn how to do it for yourself.

Required: The final product that you turn in must be a .pdf file.

- You can Knit this document directly to a PDF if you have LaTeX installed (which is preferred).
- If you absolutely can't get LaTeX installed and/or working, then you can compile to a .html first, by clicking on the arrow button next to knit and selecting Knit to HTML.
- You must then print you .html file to a .pdf by using first opening it in a web browser and then printing to a .pdf

---

Problem 1: [30 pts] Exploring Relationships between variables.

Load the `DatasaurusDozen.tsv` file into R.

This data consists of x and y observations for **13 sub-datasets** that have the following names:

`dino`, `away`, `h_lines`, `v_lines`, `x_shape`, `star`, `high_lines`, `dots`, `circle`, `bullseye`, `slant_up`, `slant_down`, `wide_lines`

a. Use `dplyr` functions to summarize each dataset in the following way: Compute the mean for x, mean for y, sd for x, sd for y, and the correlation coefficient between x and y. **Please round your answers to 2 decimal places.** The answers should be returned automatically in a tibble. Use `kable()` or `pandoc.table()` (use results='asis' in chunk definition if using `pandoc.table()`) or some other function to make nicely formatted table of your results.

```
library(tidyverse)
library(dplyr)
#Reading the file
dinof<-read.table("/Users/eduardosalvador/Desktop/FINAL\ Spring\ Semester\ 2021/CMDA\ /Assignments/HW6/Datasau
#Looking at the first 6 rows of dataframe
head(dinof)
```

```
  dataset       x       y
1    dino 55.3846 97.1795
2    dino 51.5385 96.0256
3    dino 46.1538 94.4872
4    dino 42.8205 91.4103
5    dino 40.7692 88.3333
6    dino 38.7179 84.8718
```

```
#Using dplyr to summerize mean,sd and correlation
dinof %>%
  group_by(dinof$dataset) %>%
  summarise("x_mean"=mean(x),"y_mean"=mean(y),"x_sd"=sd(x),"y_sd"=sd(y),"correlation_coef"=round(cor(x,y),2))
```

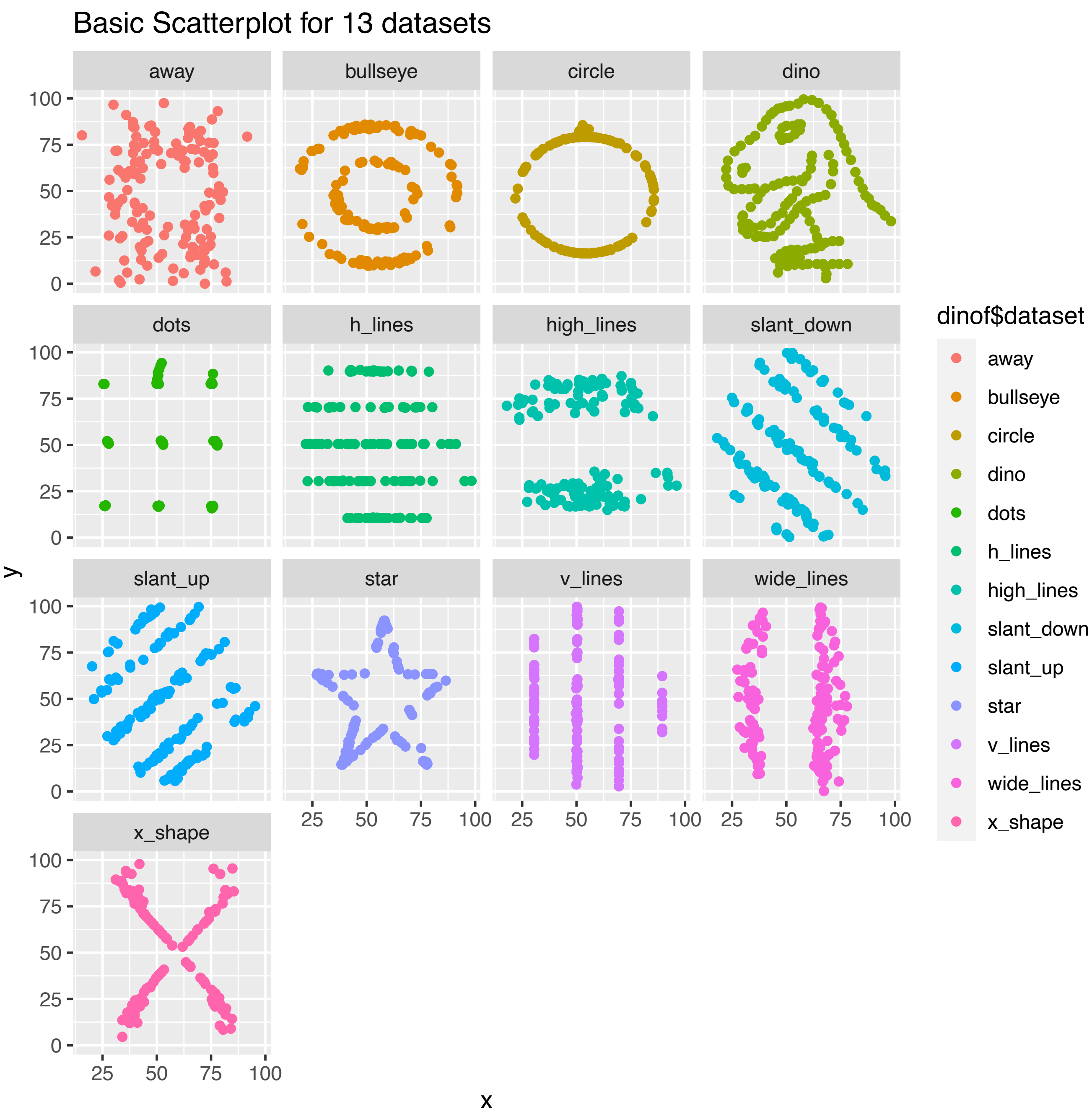
```
# A tibble: 13 x 6
   `dinof$dataset` x_mean y_mean  x_sd  y_sd correlation_coef
   <chr>            <dbl>  <dbl> <dbl> <dbl>            <dbl>
 1 away              54.3   47.8  16.8  26.9            -0.06
 2 bullseye          54.3   47.8  16.8  26.9            -0.07
 3 circle            54.3   47.8  16.8  26.9            -0.07
 4 dino              54.3   47.8  16.8  26.9            -0.06
 5 dots              54.3   47.8  16.8  26.9            -0.06
 6 h_lines           54.3   47.8  16.8  26.9            -0.06
 7 high_lines        54.3   47.8  16.8  26.9            -0.07
 8 slant_down        54.3   47.8  16.8  26.9            -0.07
 9 slant_up          54.3   47.8  16.8  26.9            -0.07
10 star              54.3   47.8  16.8  26.9            -0.06
11 v_lines           54.3   47.8  16.8  26.9            -0.07
12 wide_lines        54.3   47.8  16.8  26.9            -0.07
13 x_shape           54.3   47.8  16.8  26.9            -0.07
```

b. What does the numerical summaries tell you about the data in the 12 different data sets? In particular, does the correlation coefficient provide you with much information about the relationship between x and y?

```
## The 12 different data sets have the same mean for x and y. Furthermore, they also have the same standard di
```

c. Now make a basic scatterplot of x and y for the 13 different datasets. Use a different color for each dataset. My best advice is to simply use `ggplot()` with `facet_wrap()`, as this can be done in a singe line.

```
#Created scatterplot using ggplot (to color everything, use color function)
ggplot(dinof,aes(x,y,color=dinof$dataset))+facet_wrap(vars(dinof$dataset))+ggtitle("Basic Scatterplot for 13 d
```



d. How does your interpretation about the relationships between x and y change after seeing the plots?

```
##The points on each of the graphs appears to match the tittle of the dataset.
```

e. What lesson can be learned here?

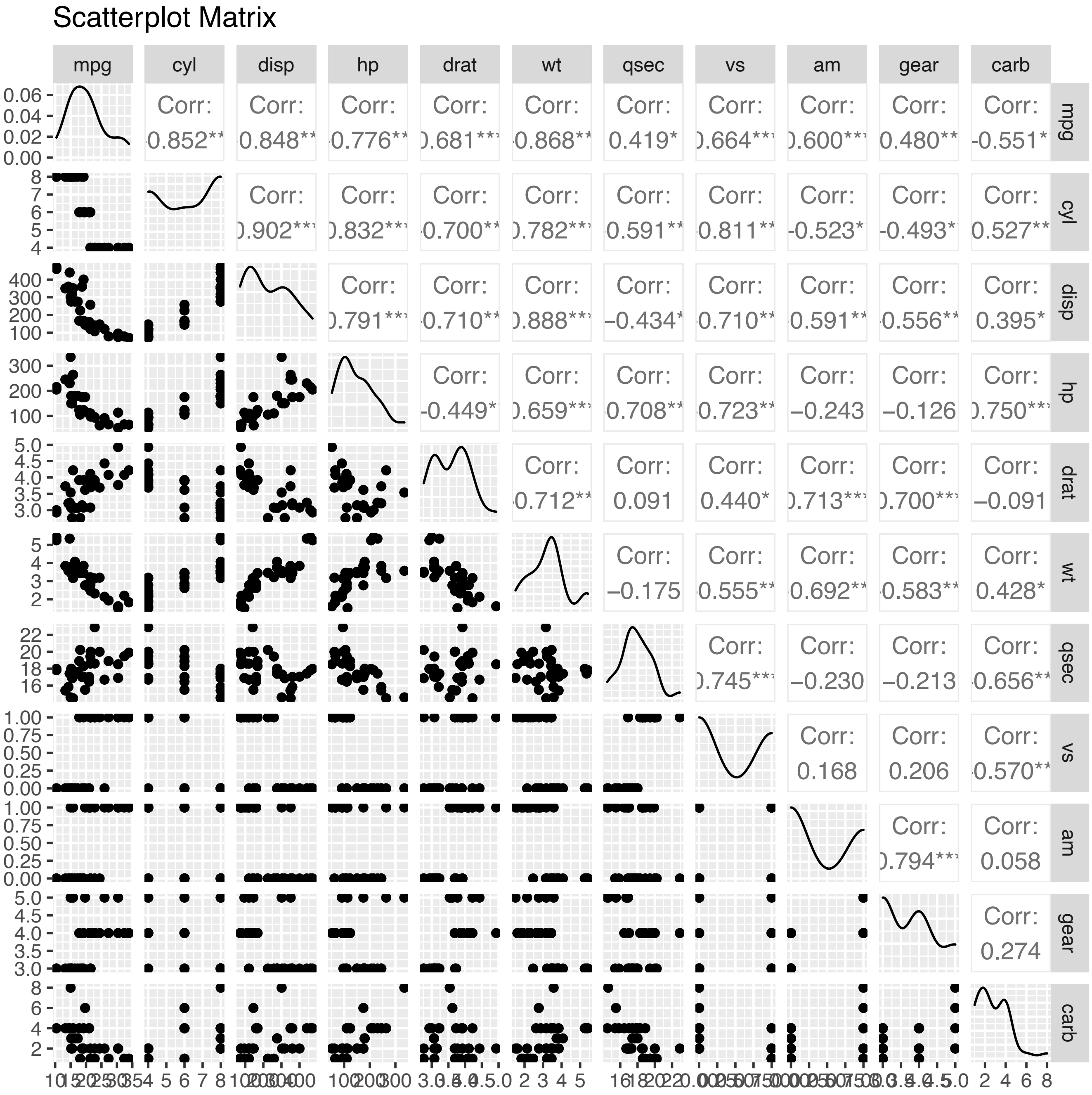
```
##That some graphs are useless in the sense of making sense of the data since, even tho the mean and sd are th
```

Problem 2: [30 pts] Linear Regression

Consider the `mtcars` dataset. Say we want to build a linear regression model that predicts `mpg`, using any subset of the other variables as predictors.

a. Begin by creating a scatterplot matrix between mpg and all other predictors. Report the correlations as well in either the upper or lower half of the scatterplot matrix.

```
library(GGally)
#Create scatterplot with ggpairs
ggpairs(mtcars)+ggtitle("Scatterplot Matrix")
```



b. What are the three variables most highly correlated with mpg?

```
##The variables most highly correlated with mpg are wt (.868), then cyl (.852) and displacement (.848)
```

c. Fit three simple linear regression models using your previous three variables/predictors. Report summaries for the models. Which model would you choose and why?

```
#Simple linear regression with mpg~wt
slrmwt<-lm(mpg ~ wt,data=mtcars)
summary(slrmwt)
```

```

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
```

```
#Simple linear regression with mpg~cyl
slrmcyl<-lm(mpg ~ cyl ,data=mtcars)
summary(slrmcyl)
```

```

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

Residuals:
    Min      1Q  Median      3Q     Max 
-4.9814 -2.1185  0.2217  1.0717  7.5186 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  37.8846     2.0738   18.27  < 2e-16 ***
cyl          -2.8758     0.3224   -8.92 6.11e-10 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 3.206 on 30 degrees of freedom
Multiple R-squared:  0.7262,    Adjusted R-squared:  0.7171 
F-statistic: 79.56 on 1 and 30 DF,  p-value: 6.113e-10
```

```
#Simple linear regression with mpg~disp
slrmdisp<-lm(mpg ~ disp,data=mtcars)
summary(slrmdisp)
```

```

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

Residuals:
    Min      1Q  Median      3Q     Max 
-4.8922 -2.2022 -0.9631  1.6272  7.2305 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 29.599855   1.229720  24.070  < 2e-16 ***
disp        -0.041215   0.004712  -8.747 9.38e-10 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 3.251 on 30 degrees of freedom
```

```
Multiple R-squared:  0.7183,    Adjusted R-squared:  0.709 
F-statistic: 76.51 on 1 and 30 DF,  p-value: 9.38e-10
```

```
##I would use the simple linear regression model with the variable wt since it gave the largest R^2 (0.7528) o
```

d. Create a multiple linear regression (MLR) model using `stepAIC()` to identify the best subset of predictors from all of the variables in `mtcars` (obviously `mpg` is still the response variable). Report these predictors, and a summary of the model these predictors produced.

```
library(MASS)
#Null model
modfit0<-lm(mpg~1,data=mtcars)
#Model with all of the predictors variables (a . calls for all of the variables in the dataframe)
modfit_full<-lm(mpg~.,data=mtcars)
#Using stepwise regression
modelfit_best<-stepAIC(modfit0,trace=0,scope = list(lower=modfit0,upper=modfit_full),direction = "both")
summary(modelfit_best)
```

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

Residuals:
    Min      1Q  Median      3Q     Max 
-3.9290 -1.5598 -0.5311  1.1850  5.8986 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 38.75179    1.78686  21.687  < 2e-16 ***
wt          -3.16697    0.74058  -4.276 0.000199 ***
cyl         -0.94162    0.55092  -1.709 0.098480 .  
hp          -0.01804    0.01188  -1.519 0.140015    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.512 on 28 degrees of freedom
Multiple R-squared:  0.8431,    Adjusted R-squared:  0.8263 
F-statistic: 50.17 on 3 and 28 DF,  p-value: 2.184e-11
```

```
##The best subset of predictors are wt, cyl and hp with a Multiple R^2 of 0.8431 and Adjusted R^2 of 0.8263 wh
```

e. Compare your MLR model to your three simple linar regression models earlier. Are any of those predictors in your MLR model? Are the coefficients the same for those predictors? If not, explain what may have caused the change.

```
##The predictors/variables wt and cyl were in my simple linear regression model but, hp wasn't, instead I thou
```

Problem 3: [30 pts] More Linear Regression

Sometimes your dataset is rather small, but you see that a simple linear regression is not appropriate so you try harder to fit a more complicated model. This is an example of such a situation.

A poultry scientist was studying various dietary additives to increase the rate at which chickens gain weight. One of the potential additives was studied by creating a new diet that consisted of a standard basal diet supplemented with varying amounts of the additive (0, 20, 40, 60, 80, and 100 grams). There were 60 chicks available for the study. Each of the six diets was randomly assigned to 10 chicks. At the end of 4 weeks, the feed efficiency ratio, feed consumed (gm) to weight gain (gm), was obtained for the 60 chicks. The experiment was also concerned with the effects of high levels of copper in the chick feed. Five of the 10 chicks in each level of the feed additive received 400 ppm of copper, while the remaining five chicks received no copper.

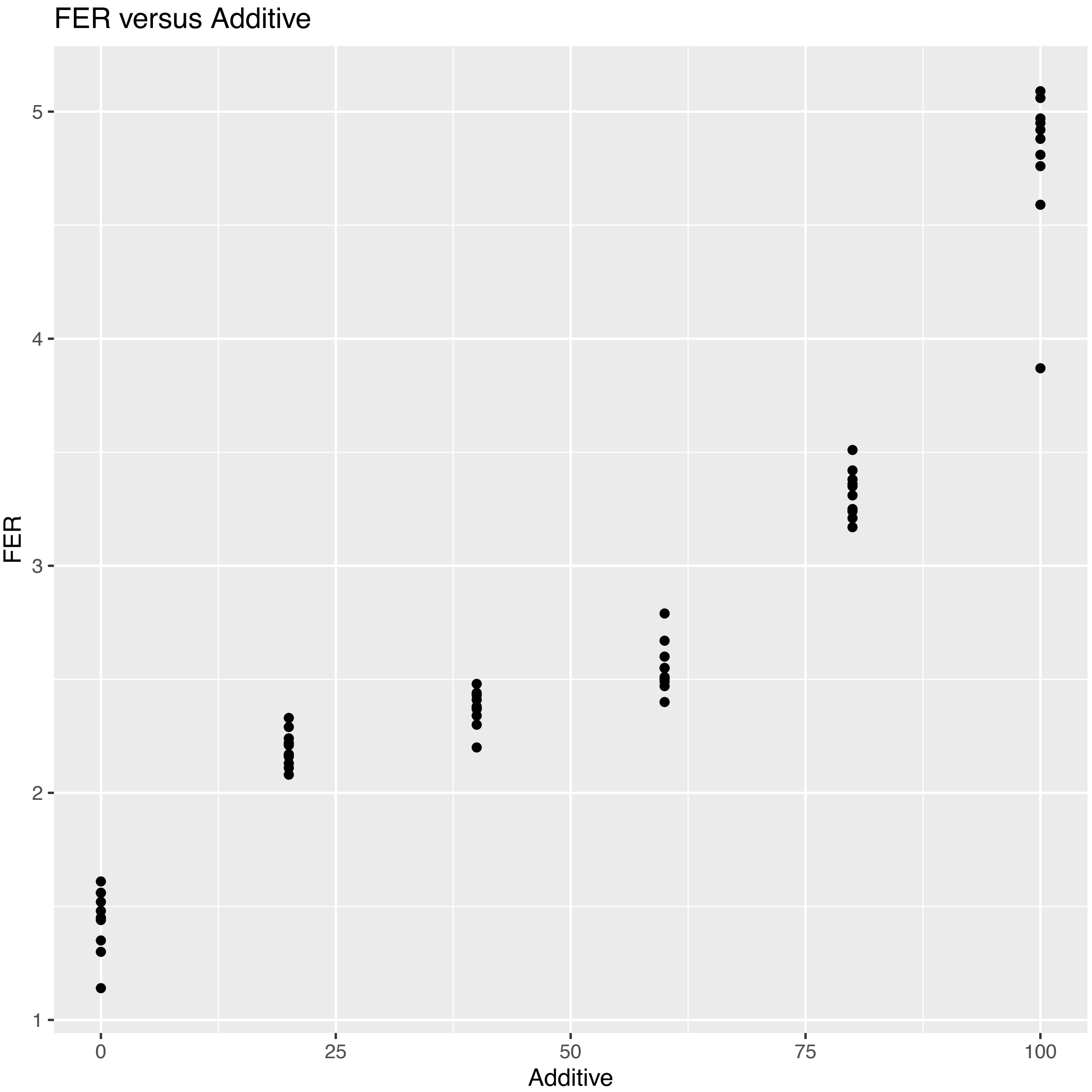
The data is contained in the `chicken.csv` data file.

a. In order to explore the relationship between feed efficiency ratio (FER) and feed additive (A), plot the FER versus A.

```
#Read the file
chicken_df<-read.csv("/Users/eduardosalvador/Desktop/FINAL\ Spring\ Semester\ 2021/CMDA\ /Assignments/HW6/chic
#Look at table
head(chicken_df)
```

```
   FER Additive Copper
1 1.30        0      0
2 1.35        0      0
3 1.44        0      0
4 1.52        0      0
5 1.56        0      0
6 1.61        0    400
```

```
#Used ggplot to plot FER vs Additive
ggplot(chicken_df,aes(Additive,FER)) + ggtitle("FER versus Additive")+ geom_point()
```



b. What type of regression appears most appropriate?

```
##The regression type most appropiate for FER versus Additive appears to be the Polynomial Regression
```

c. Fit first-order, quadratic, and cubic regression models to the data. Which regression equation provides the best fit to the data? Justify your answer using evidence based upon plots and relevant summaries.

```
library(pander)
library(olsrr)
#First-orfer regression model
f_model<-lm(Additive~FER,data=chicken_df)
summary(f_model)

Call:
lm(formula = Additive ~ FER, data = chicken_df)
```

```
Residuals:
     Min       1Q   Median       3Q      Max 
-18.5200 -12.0502  -0.5643  13.0942  21.2142 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -32.352      4.679  -6.914 4.08e-09 ***
FER           29.641      1.572  18.856  < 2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 13.01 on 58 degrees of freedom
Multiple R-squared:  0.8598,    Adjusted R-squared:  0.8573 
F-statistic: 355.6 on 1 and 58 DF,  p-value: < 2.2e-16
```

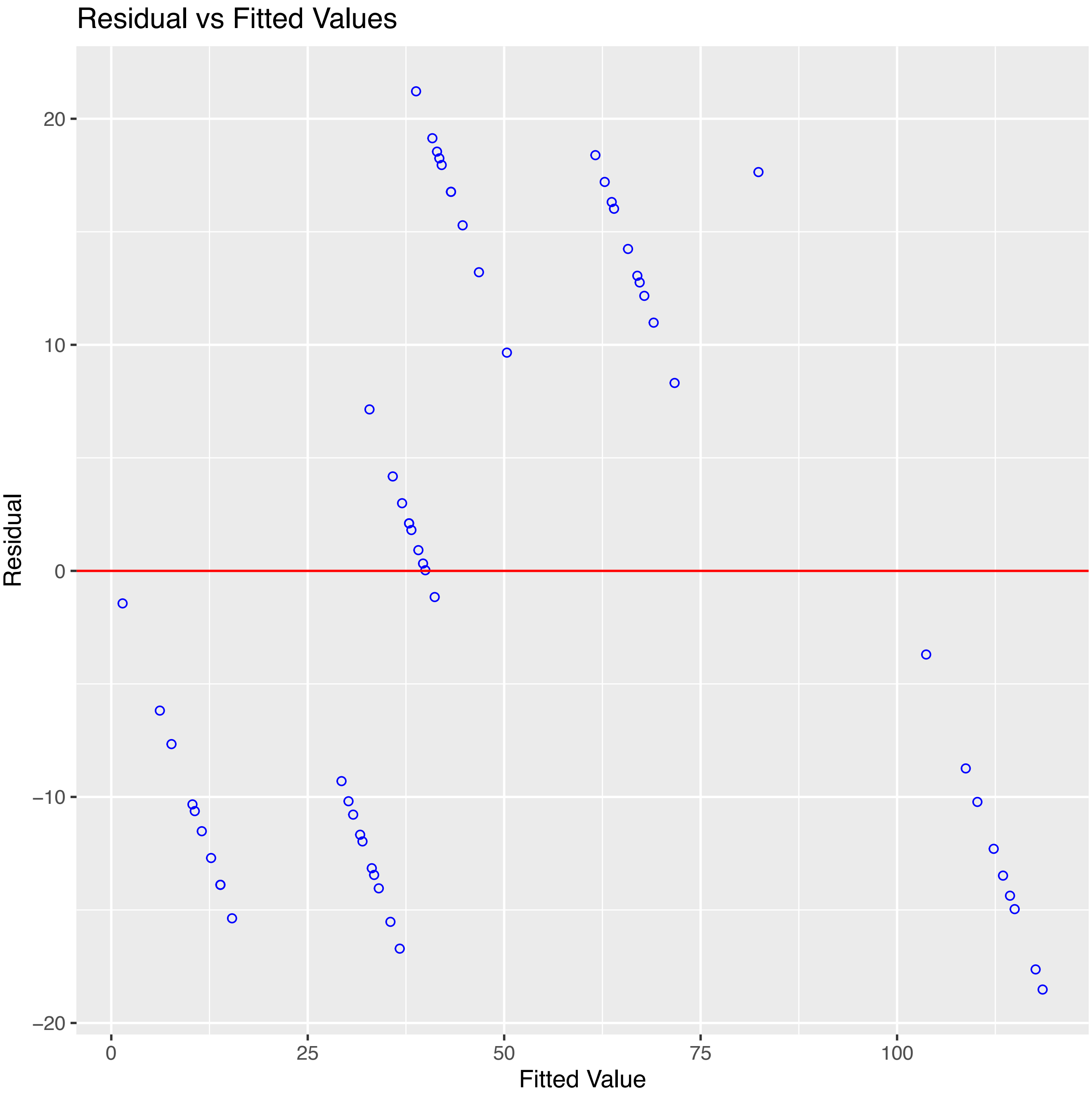
```
f_model%>%summary%>% pander()
```

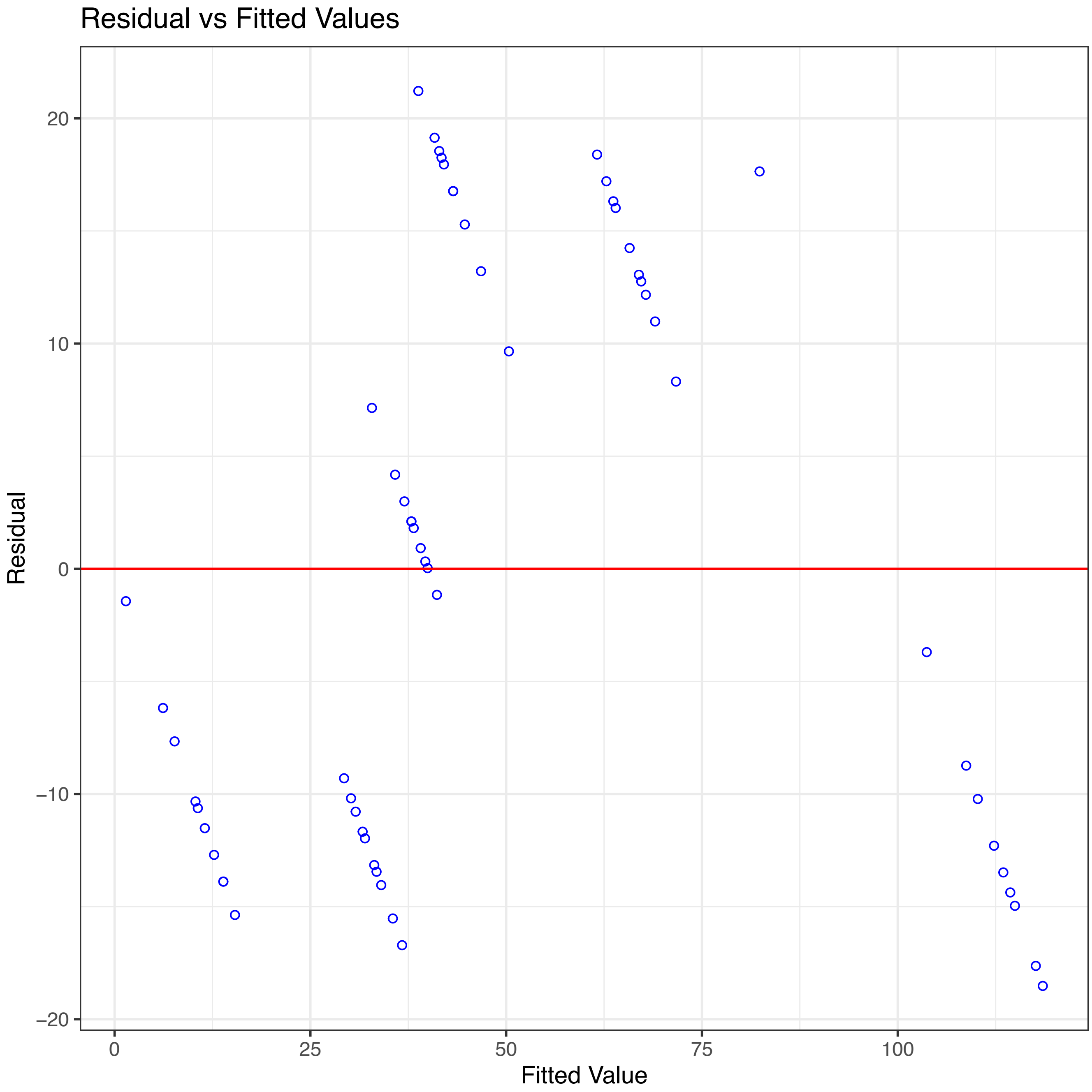
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-32.35	4.679	-6.914	4.084e-09
FER	29.64	1.572	18.86	2.04e-26

Table 2: Fitting linear model: Additive ~ FER

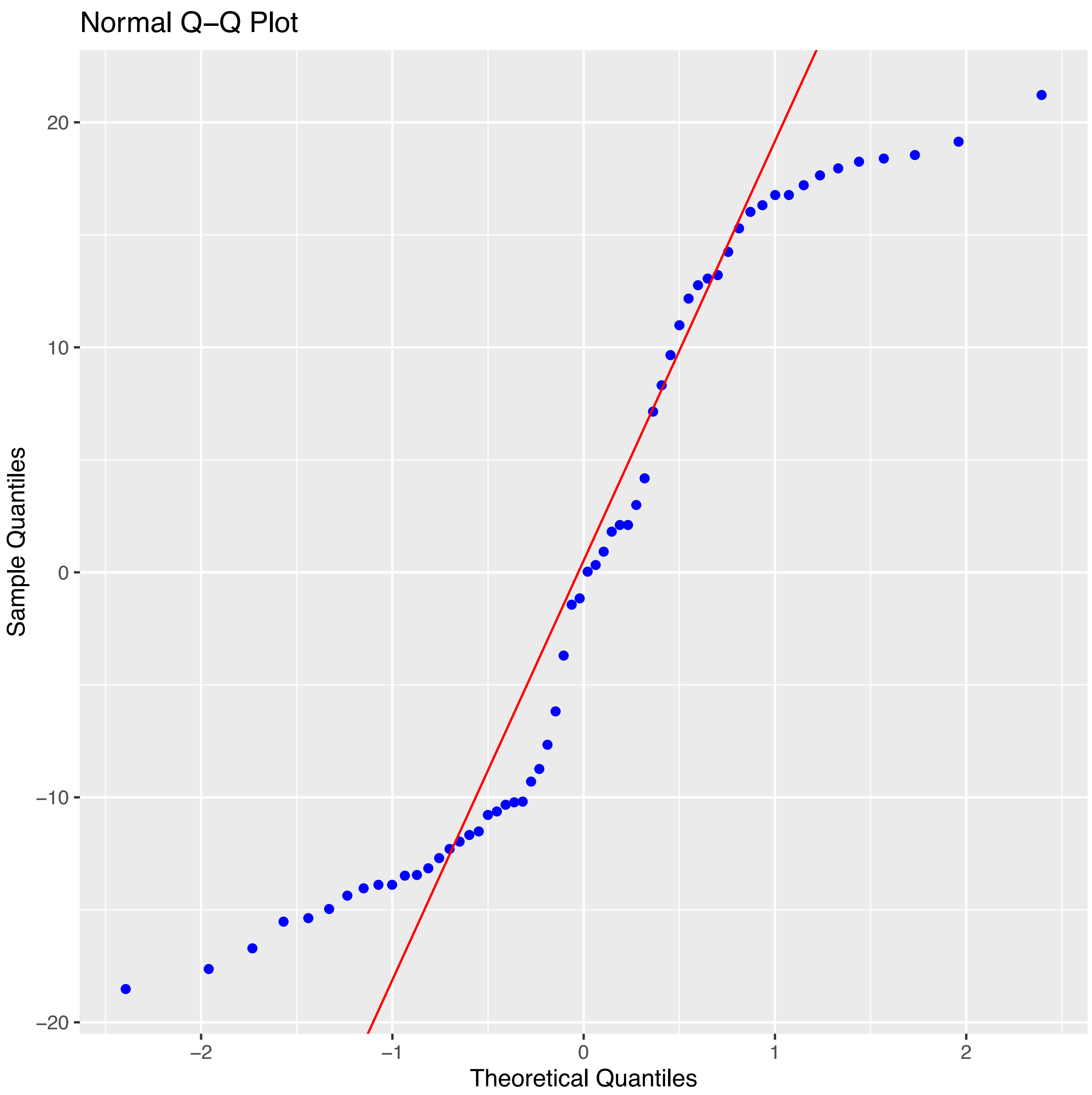
Observations	Residual Std. Error	R^2	Adjusted R^2
60	13.01	0.8598	0.8573

```
ols_plot_resid_fit(f_model)+theme_bw()
```


Residual vs Fitted Values
Residual
Fitted Value
20
10
0
−10
−20
0
25
50
75
100



```
ols_plot_resid_qq(f_model)
```



```
ols_test_normality(f_model)
```

```
-----------------------------------------------
       Test             Statistic       pvalue
-----------------------------------------------
Shapiro-Wilk              0.8902          1e-04
Kolmogorov-Smirnov        0.1686         0.0661
Cramer-von Mises          4.7343         0.0000
Anderson-Darling          2.3759         0.0000
-----------------------------------------------
```

```
#Quadratic regression model
q_model<-lm(Additive~FER+I(FER^2),data=chicken_df)
summary(q_model)
```

```
Call:
```

```
lm(formula = Additive ~ FER + I(FER^2), data = chicken_df)

Residuals:
    Min      1Q  Median      3Q     Max 
-20.756  -3.341  -1.026   5.986  21.251 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -100.639      9.362  -10.75 2.48e-15 ***
FER           78.217      6.333   12.35  < 2e-16 ***
I(FER^2)      -7.525      0.966   -7.79 1.54e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.134 on 57 degrees of freedom
Multiple R-squared:  0.9321,    Adjusted R-squared:  0.9297 
F-statistic:   391 on 2 and 57 DF,  p-value: < 2.2e-16
```

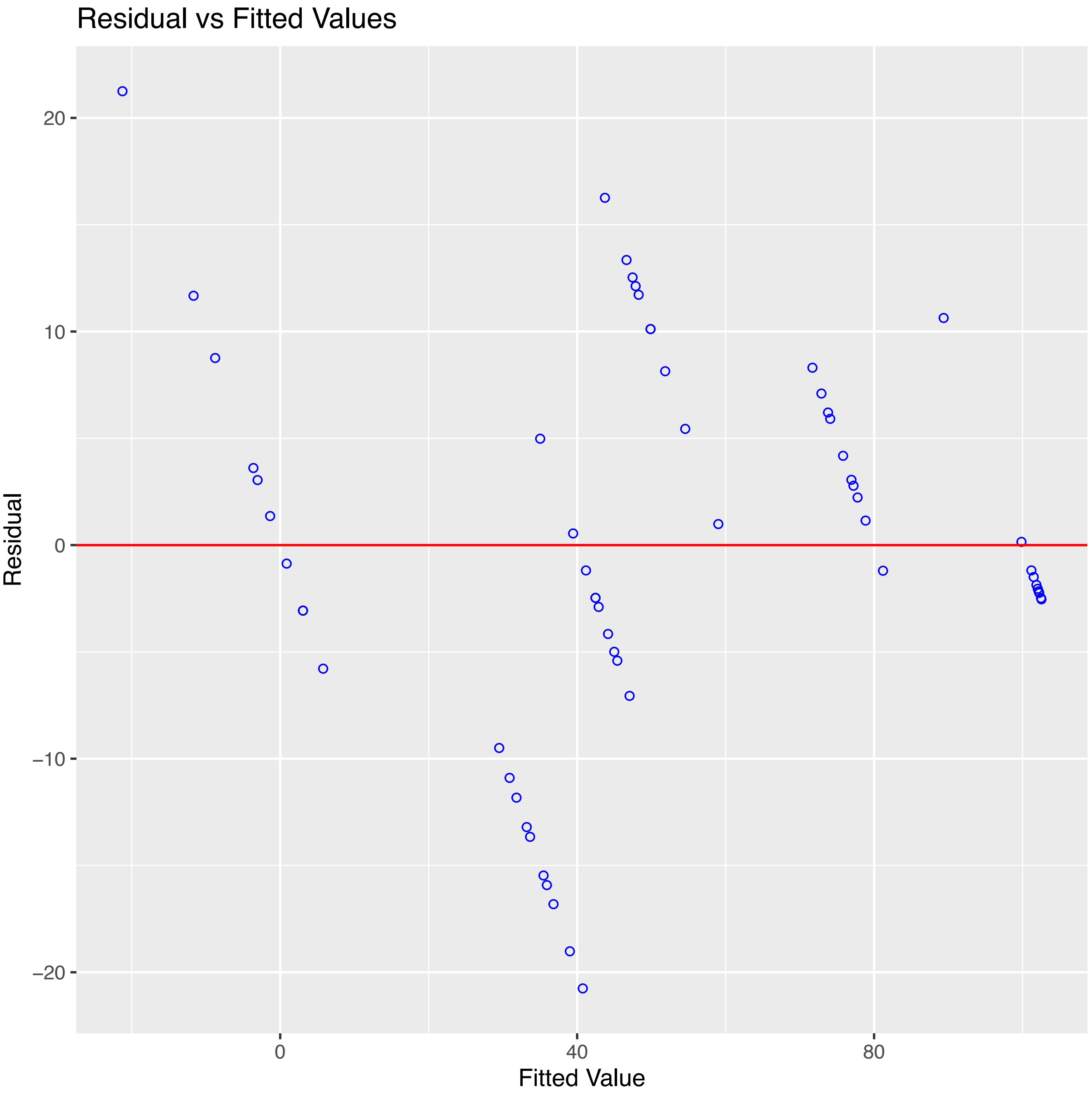
```
q_model%>%summary%>% pander()
```

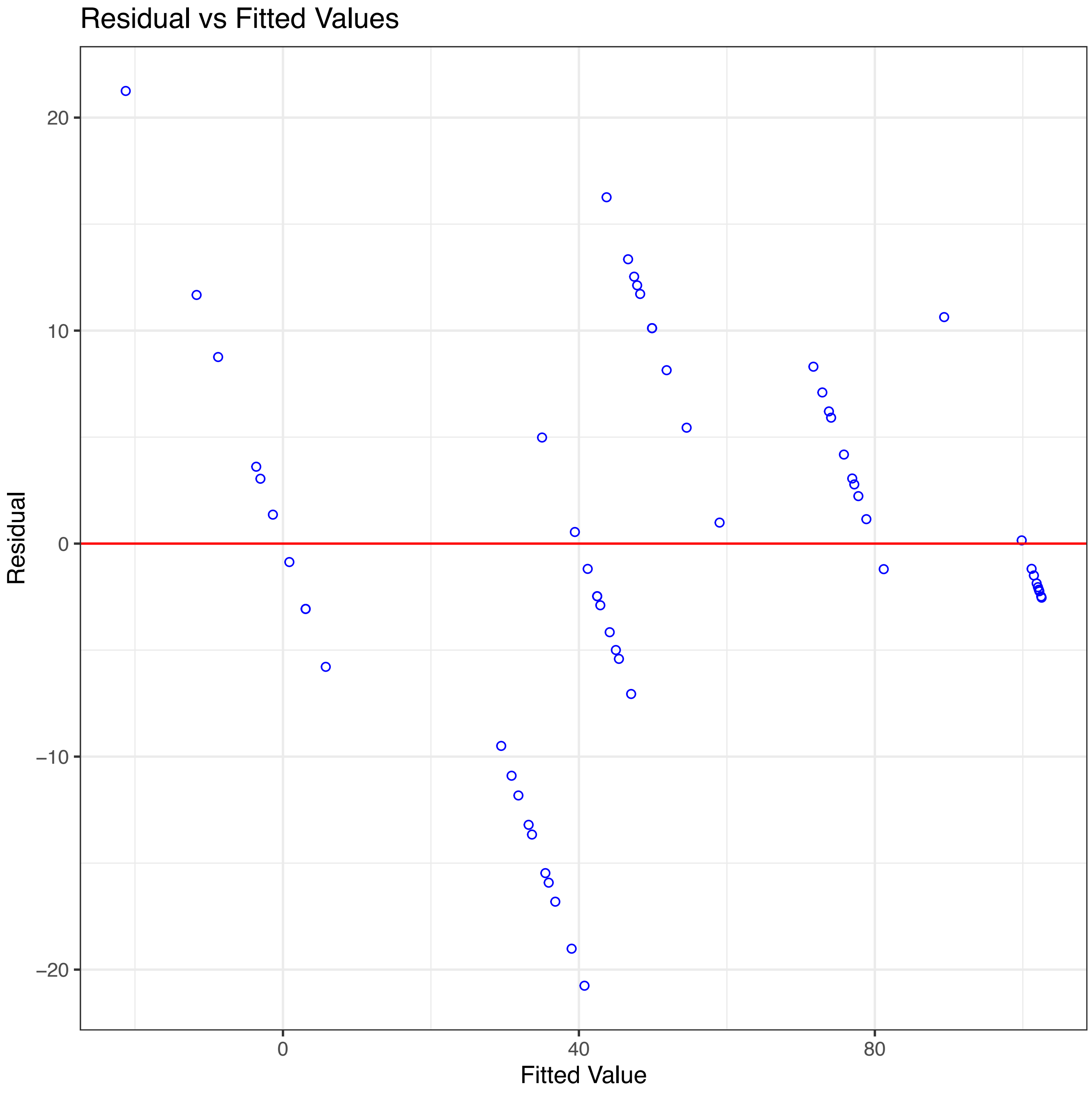
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-100.6	9.362	-10.75	2.483e-15
FER	78.22	6.333	12.35	9.43e-18
I(FER^2)	-7.525	0.966	-7.79	1.538e-10

Table 4: Fitting linear model: Additive ~ FER + I(FER^2)

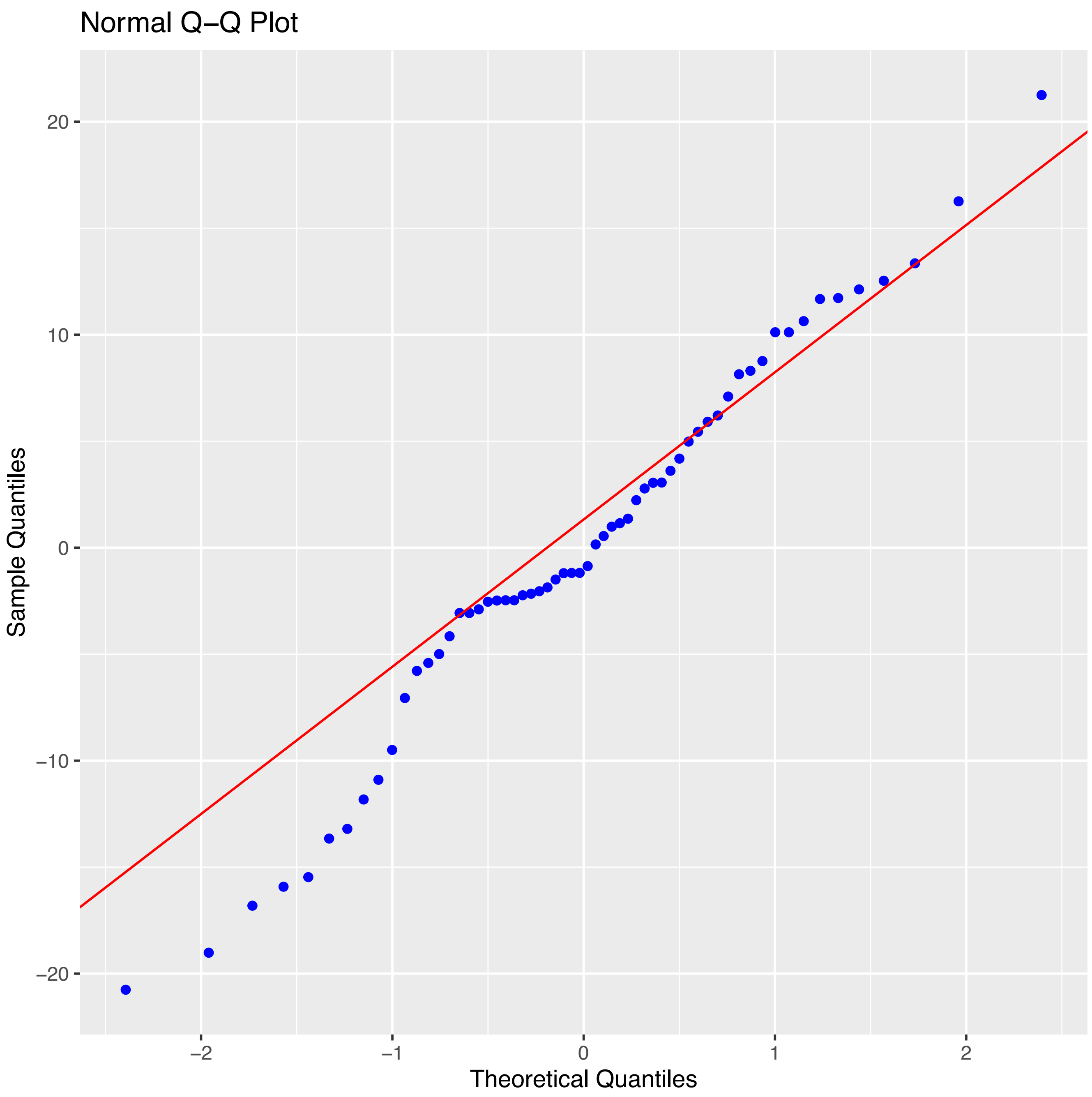
Observations	Residual Std. Error	R^2	Adjusted R^2
60	9.134	0.9321	0.9297

```
ols_plot_resid_fit(q_model)+theme_bw()
```


Residual vs Fitted Values
Residual
20
10
0
−10
−20
0
40
80
Fitted Value



```
ols_plot_resid_qq(q_model)
```



```
ols_test_normality(q_model)

-----------------------------------------------
       Test             Statistic       pvalue
-----------------------------------------------
Shapiro-Wilk              0.9808          0.4619
Kolmogorov-Smirnov        0.1163          0.3916
Cramer-von Mises          4.7082          0.0000
Anderson-Darling          0.4826          0.2222
-----------------------------------------------

#Cubic regression model
c_model<-lm(Additive~FER+I(FER^2)+I(FER^3),data=chicken_df)
summary(c_model)

Call:
```

```
lm(formula = Additive ~ FER + I(FER^2) + I(FER^3), data = chicken_df)

Residuals:
     Min       1Q   Median       3Q      Max 
-17.3941  -4.2858   0.0314   4.0400  19.2826 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -12.7719    22.7726  -0.561 0.577139    
FER         -26.4251    25.8822  -1.021 0.311656    
I(FER^2)     30.0682     9.1187   3.297 0.001698 ** 
I(FER^3)     -4.0714     0.9833  -4.141 0.000118 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.063 on 56 degrees of freedom
Multiple R-squared:  0.948, Adjusted R-squared:  0.9452 
F-statistic: 340.3 on 3 and 56 DF,  p-value: < 2.2e-16
```

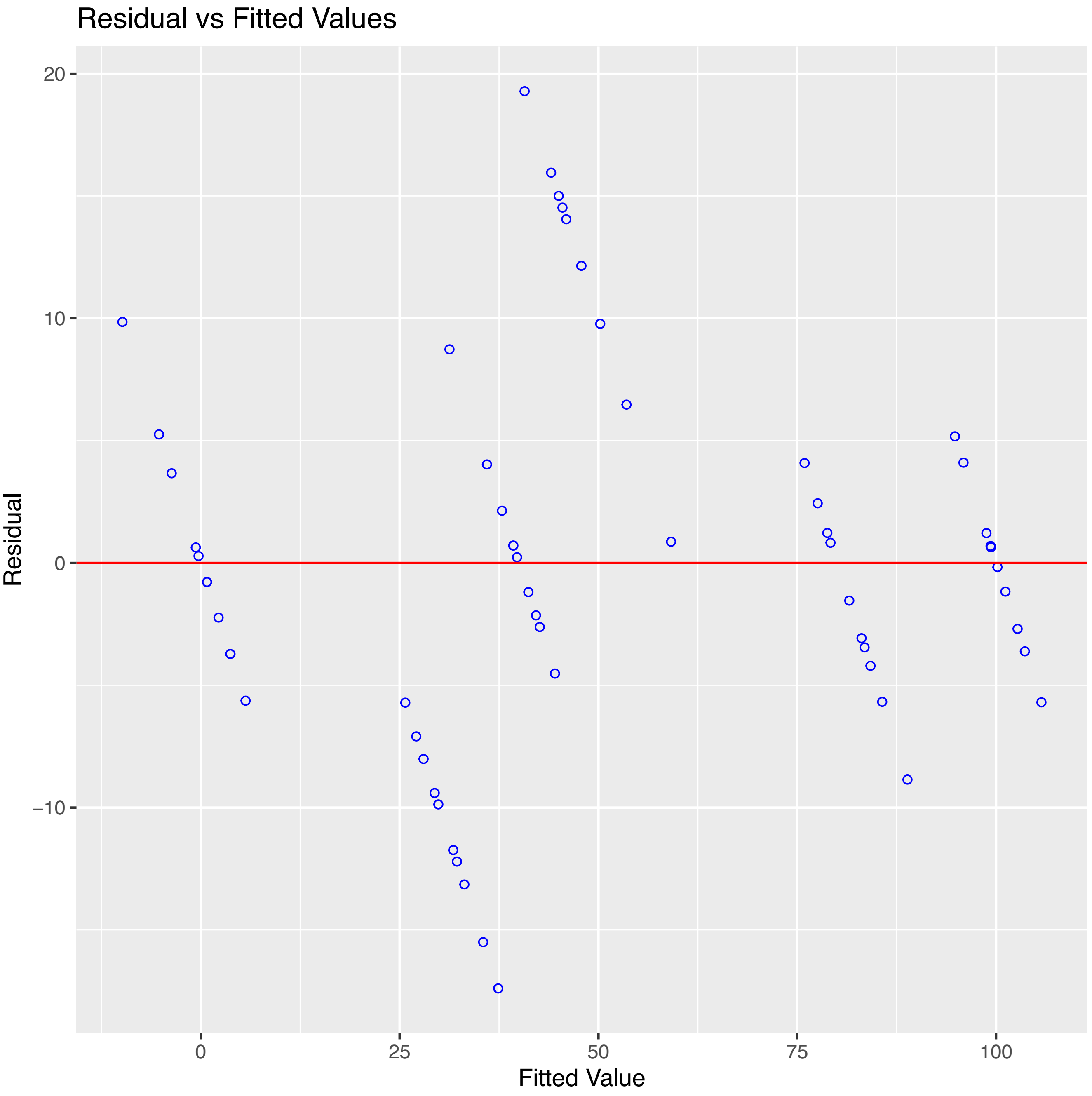
```
c_model%>%summary%>% pander()
```

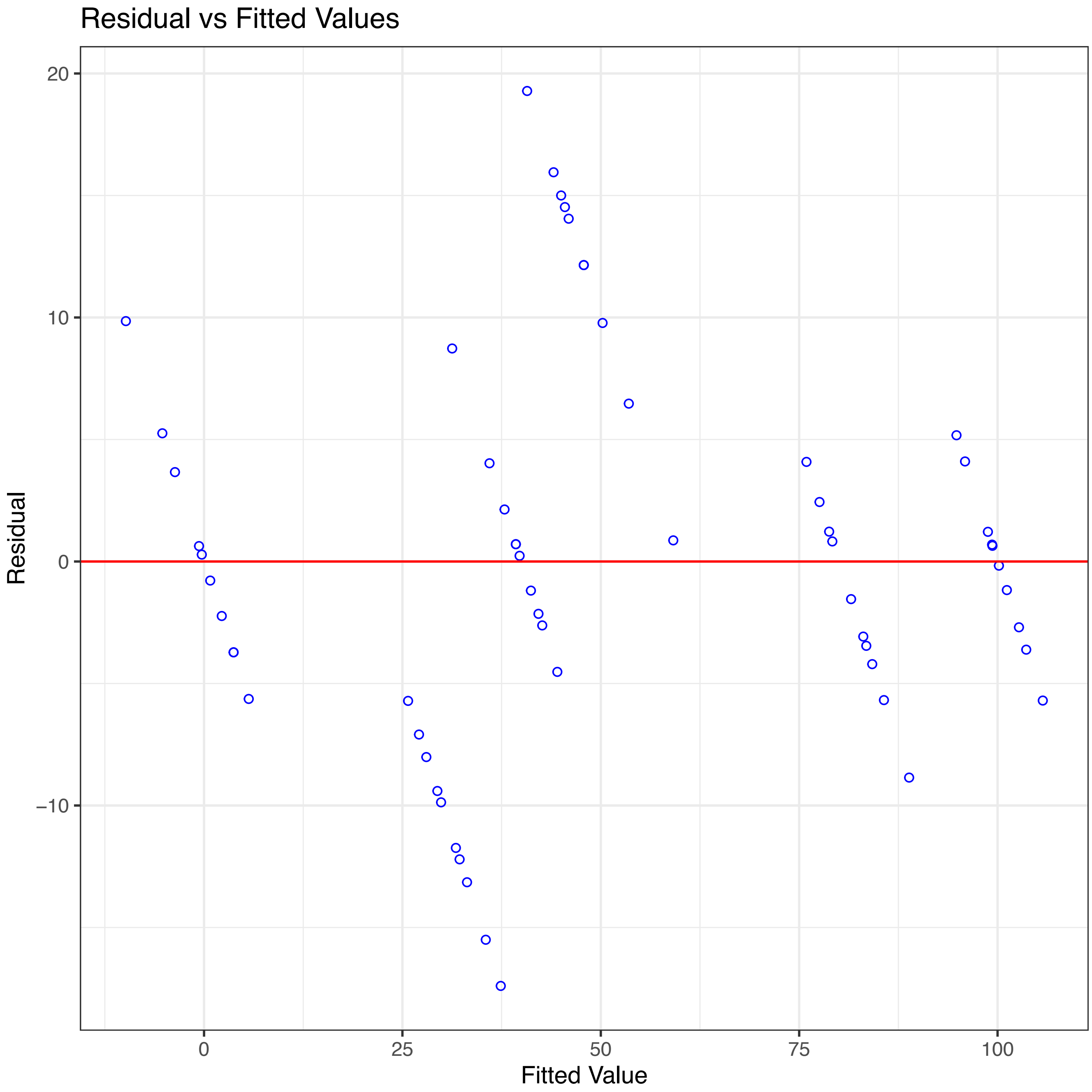
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-12.77	22.77	-0.5608	0.5771
FER	-26.43	25.88	-1.021	0.3117
I(FER^2)	30.07	9.119	3.297	0.001698
I(FER^3)	-4.071	0.9833	-4.141	0.0001178

Table 6: Fitting linear model: Additive ~ FER + I(FER^2) + I(FER^3)

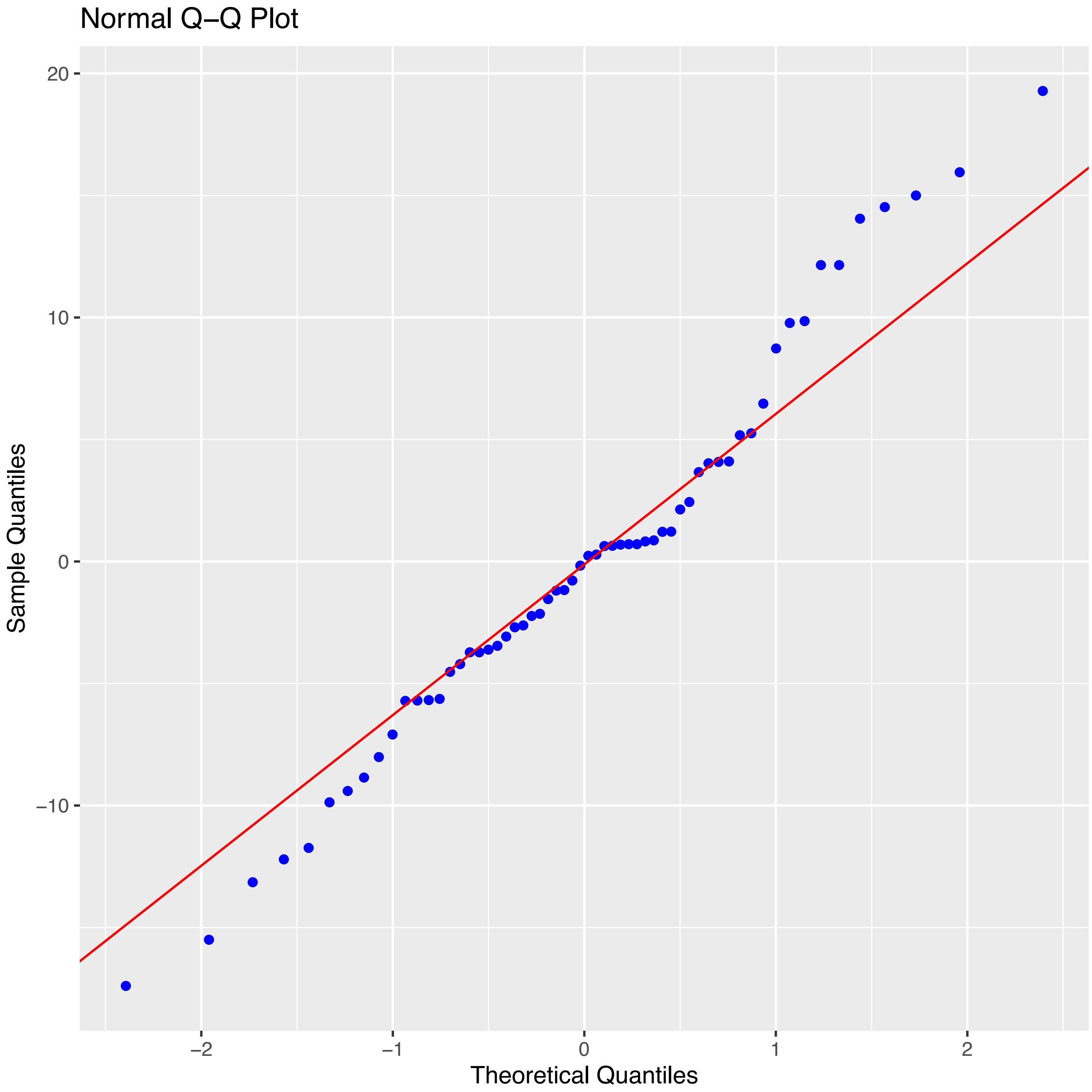
Observations	Residual Std. Error	R^2	Adjusted R^2
60	8.063	0.948	0.9452

```
ols_plot_resid_fit(c_model)+theme_bw()
```


Residual vs Fitted Values
Residual
Fitted Value
20
10
0
−10
0
25
50
75
100



```
ols_plot_resid_qq(c_model)
```



```
ols_test_normality(c_model)
```

```
-----------------------------------------------
       Test             Statistic       pvalue
-----------------------------------------------
Shapiro-Wilk              0.9773         0.3236
Kolmogorov-Smirnov        0.1214         0.3390
Cramer-von Mises          3.634          0.0000
Anderson-Darling          0.5943         0.1180
-----------------------------------------------
```

```
##I think cubic model provides the best fit since it has the least residual error and the line has the closest
```

d. Is there anything peculiar about any of the data values? Provide an explanation of what may have happened. (Hint: Look at regression diagnostics like plots of the residuals versus the fitted values (or x), plot the leverages, or plot some measure of influence.)

```
##There appears to be a pattern aliggning in the residuals vs fitted values which is very peculiar since it is
```

e. Using your best polynomial model from (b) & (c). Fit a new model that includes the linear addition of copper and display the estimate table. Does Copper provide a significant improvement to the fit? Carry out an F-test that compares the Full model that contains Copper and the reduced model that has your polynomial model fit on the additive only. Discuss the results.

```
#First-order regression model using Addivive, FER and Copper
new_modfit<-lm(Additive~FER+Copper,data=chicken_df)
summary(new_modfit)
```

```
Call:
lm(formula = Additive ~ FER + Copper, data = chicken_df)

Residuals:
     Min       1Q   Median       3Q      Max 
-18.9476 -12.1329   0.2289  13.5052  20.0136 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -31.256799   4.938201  -6.330 4.13e-08 ***
FER          29.684656   1.579631  18.792  < 2e-16 ***
Copper       -0.006085   0.008439  -0.721    0.474    
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 13.06 on 57 degrees of freedom
Multiple R-squared:  0.861, Adjusted R-squared:  0.8561 
F-statistic: 176.6 on 2 and 57 DF,  p-value: < 2.2e-16
```

```
summary(f_model)
```

```
Call:
lm(formula = Additive ~ FER, data = chicken_df)

Residuals:
     Min       1Q   Median       3Q      Max 
-18.5200 -12.0502  -0.5643  13.0942  21.2142 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -32.352      4.679  -6.914 4.08e-09 ***
FER           29.641      1.572  18.856  < 2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 13.01 on 58 degrees of freedom
Multiple R-squared:  0.8598,    Adjusted R-squared:  0.8573 
F-statistic: 355.6 on 1 and 58 DF,  p-value: < 2.2e-16
```

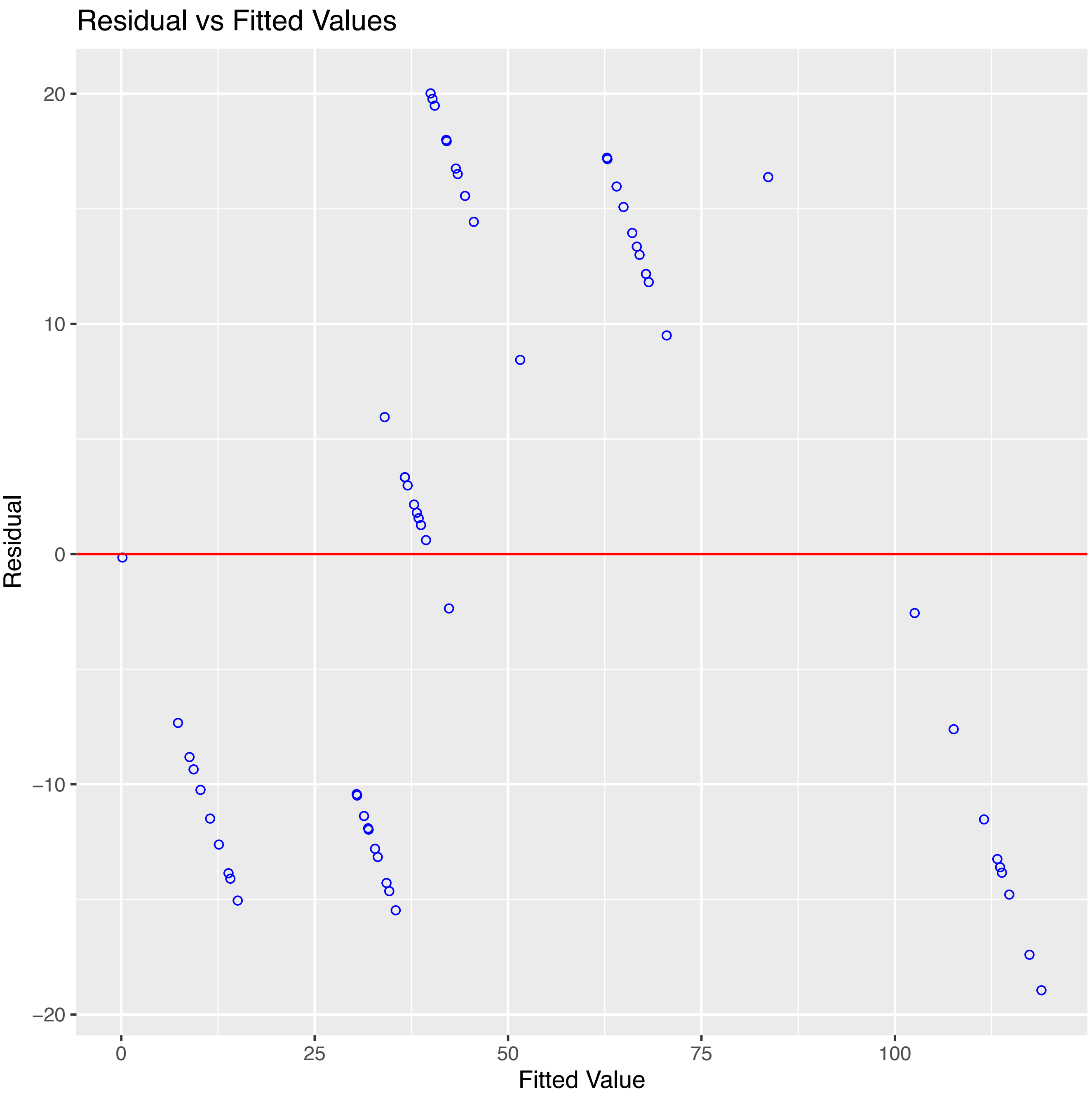
```
new_modfit%>%summary%>% pander()
```

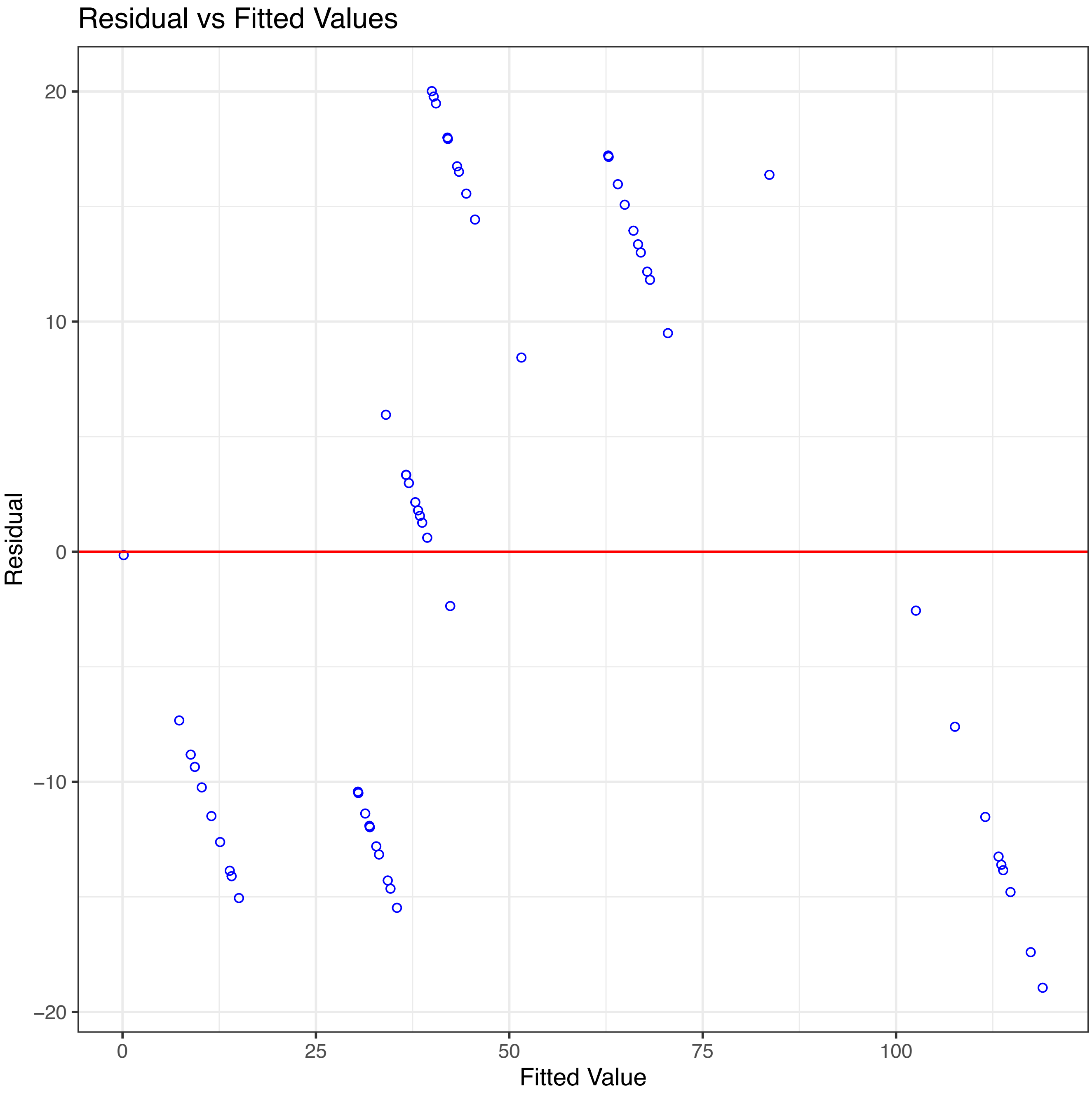
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-31.26	4.938	-6.33	4.127e-08
FER	29.68	1.58	18.79	4.238e-26
Copper	-0.006085	0.008439	-0.7211	0.4738

Table 8: Fitting linear model: Additive ~ FER + Copper

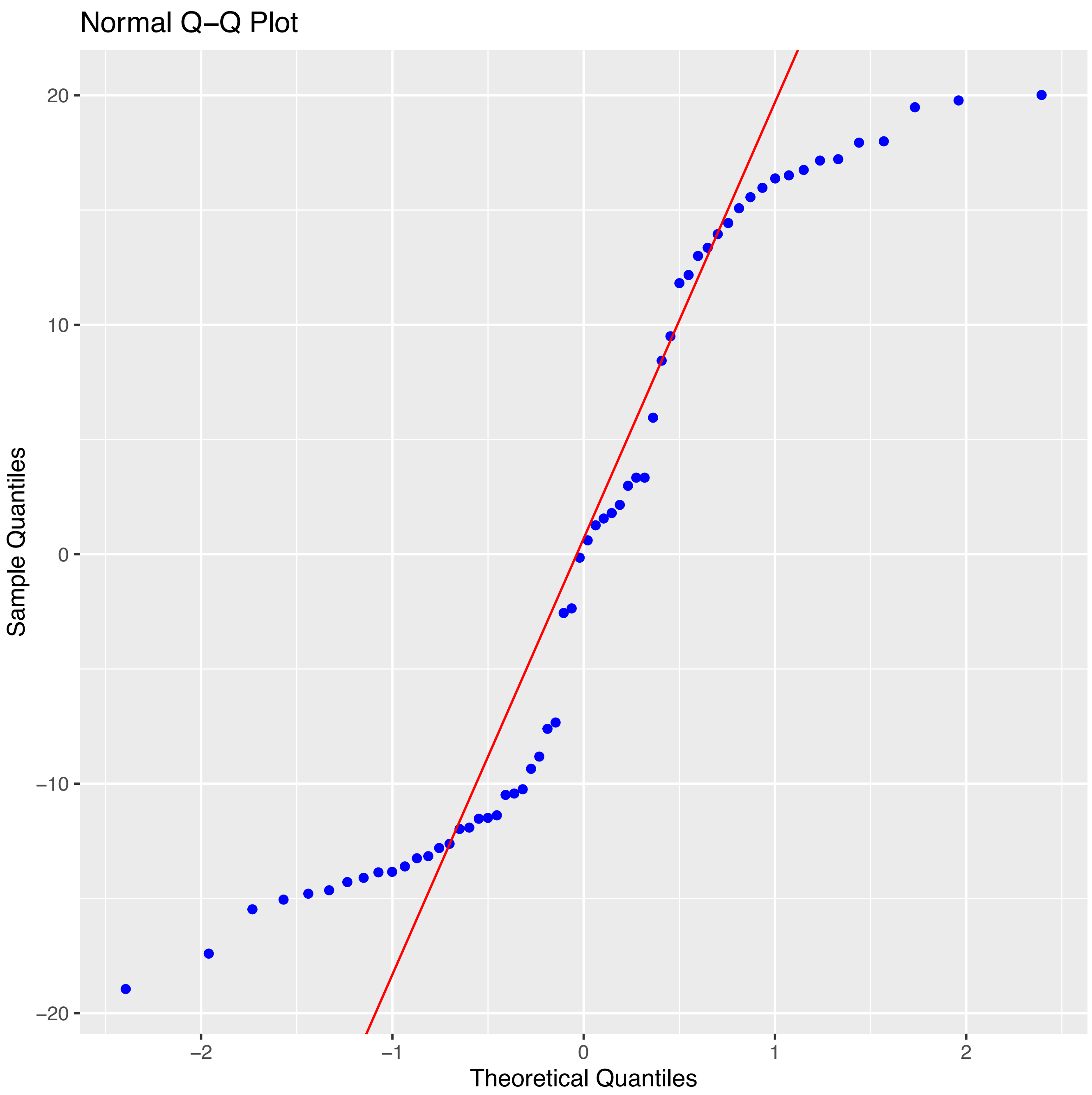
Observations	Residual Std. Error	R^2	Adjusted R^2
60	13.06	0.861	0.8561

```
ols_plot_resid_fit(new_modfit)+theme_bw()
```





```
ols_plot_resid_qq(new_modfit)
```



```
ols_test_normality(new_modfit)
```

```
-----------------------------------------------
       Test             Statistic       pvalue  
-----------------------------------------------
Shapiro-Wilk              0.8835         0.0000 
Kolmogorov-Smirnov        0.1708         0.0604 
Cramer-von Mises          4.7681         0.0000 
Anderson-Darling          2.5735         0.0000 
-----------------------------------------------
```

```
#Copper does cluster the points more in the middle and at the ends on the line but statistically speaking, it
```

Problem 4: [10 pts] Linear Regression with Indicator Variables

Consider the data in smoking_birthweight.csv. This data contains 3 variables. The birth weight of a baby (Weight), the length of gestation (Gestation) in weeks, and the smoking status of the mother (Smoke). The smoking status of the mother in this case is coded as yes or no. This is a categorical variable (aka factor) with 2 categories (a binary variable). We could have coded the levels of this factor as an indicator variable using TRUE or FALSE, or equivalently 1 or 0, respectively.

a. Fit a first-order regression model with birth weight as the response variable and the gestation and smoking status as predictors. Write down the fitted regression model equation and interpret the regression coefficients. If you can do this, you should have no problem handling the extra credit.

```
library(dplyr)
##Read the file
smokebw<-read.csv("/Users/eduardosalvador/Desktop/FINAL\ Spring\ Semester\ 2021/CMDA\ /Assignments/HW6/smoking

#Looking at first six rows
head(smokebw)
```

```
  Weight Gestation Smoke
1   2940        38   yes
2   3130        38    no
3   2420        36   yes
4   2450        34    no
5   2760        39   yes
6   2440        35   yes
```

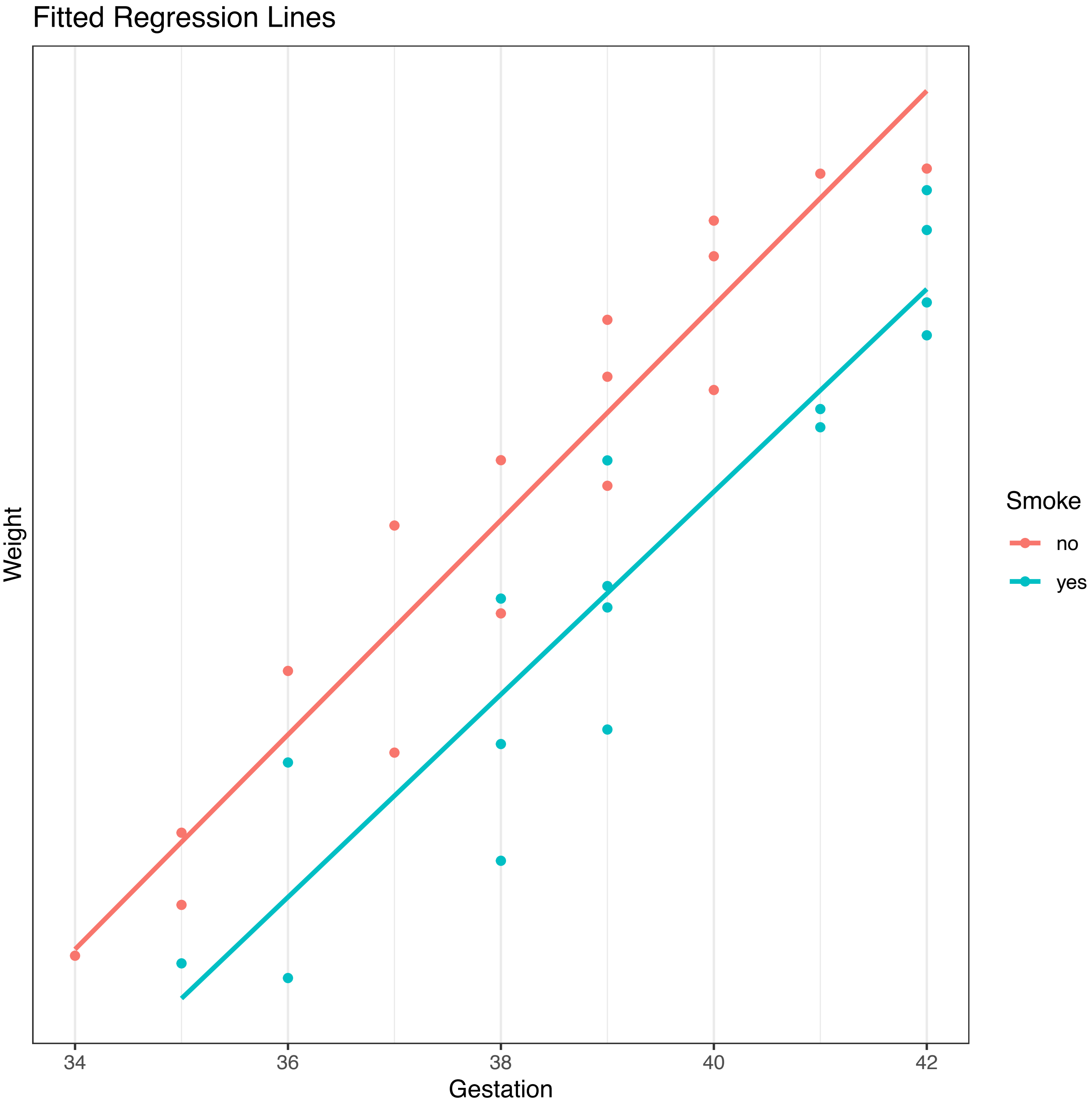
```
#Looking for person that is smoker
smokebw<-smokebw %>% mutate(mystatus=if_else(Smoke=="yes",1,0))

#Making a fitted generalized linear model for weight with mystatus and gestation
glm_smoke<-glm(Weight~mystatus+Gestation,data=smokebw)

summary(glm_smoke)
```

```

Call:
glm(formula = Weight ~ mystatus + Gestation, data = smokebw)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-223.693   -92.063    -9.365    79.663   197.507

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2389.573    349.206  -6.843 1.63e-07 ***
mystatus     -244.544     41.982  -5.825 2.58e-06 ***
Gestation     143.100      9.128  15.677 1.07e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 13347.24)

    Null deviance: 3735790  on 31  degrees of freedom
Residual deviance:  387070  on 29  degrees of freedom
AIC: 399.63

Number of Fisher Scoring iterations: 2
```

```
## My interpretation is that if mother were to smoke, the change in weight decreases by 244.544 with 41.982 er
```

?glm b. Plot the fitted regression lines (yes plural), why are there two?

```
library(ggplot2)
##GGplot to create fitted regression lines
ggplot(smokebw,aes(x=Gestation,y=Weight,color=Smoke))+theme_bw()+ggtitle("Fitted Regression Lines")+geom_point
```



```
##There are two lines because there is moms that smoke and no smoke
```

Problem 5: [15 pts Extra Credit] Parameter Interpretation with Indicator Variables

Recall that indicator variables, sometimes called "dummy" variables, are binary variables that indicate whether an event is recognized or not (i.e., `1` if `TRUE` `0` if `FALSE`). Suppose we have a data set of reported salaries and highest achieved education levels. Suppose the variables are as follows: `salary`, `noHS`, `highSchoolGrad`, `Assoc`, `Bach`, `Masters`, `Doctorate`, where the levels of education are either a `1` or `0` depending on whether that is the given observation's highest level of achieved education.

- Write down the multiple linear regression model. Specify which β_i are indicator variables.

```
##salary=$\beta_{0}$+$\beta_{1}$noHS+$\beta_{2}$highSchoolGrad+$\beta_{3}$Assoc+$\beta_{4}$Bach+$\beta_{5}$Mas
```

- Write interpretations for all of your model parameters, that is β_i, for $i \in \{0, 1, 2, 3, 4, 5\}$.

```
##$\beta_{}$ always increases the salary with each increase in education. For example, for highSchoolGrad, the
```

- Now assume we were to add another variable to this data set: an observation's `gender`. Write down this new model, and now interpret β_0.

```
##salary=$\beta_{0}$+$\beta_{1}$noHS+$\beta_{2}$highSchoolGrad+$\beta_{3}$Assoc+$\beta_{4}$Bach+$\beta_{5}$Mas
##$\beta_{0}$ is the salary that it doesn't matter your level of education or gender, it is the lowest amount
```