HW11-Data605

Cesar Espitia

April 22, 2018

Assignment

Using the “cars” dataset in R, build a linear model for stopping distance as a function of speed and replicate the analysis of your textbook chapter 3 (visualization, quality evaluation of the model, and residual analysis.)

The following is the histogram for the speed.

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

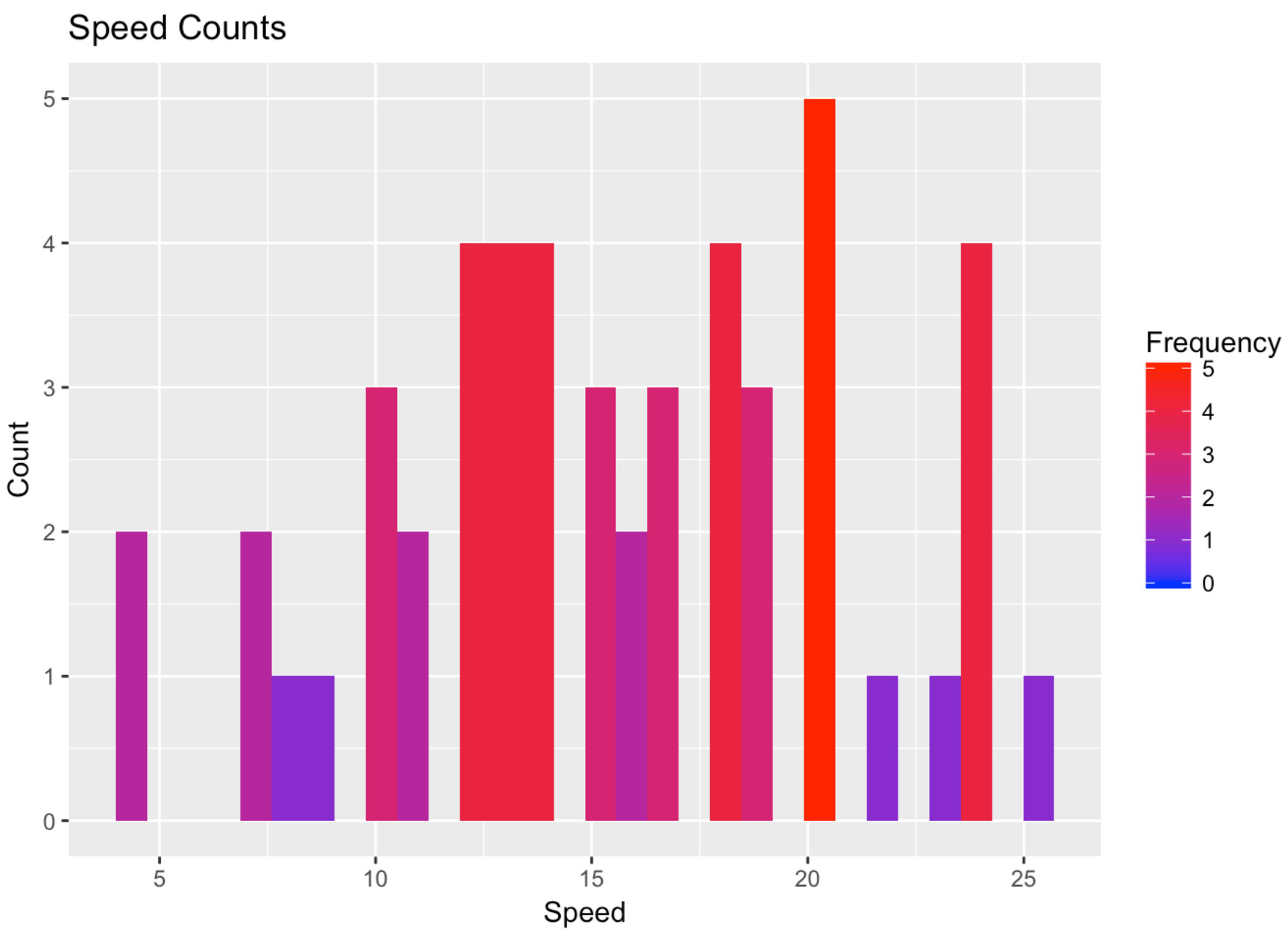
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
df = arrange(cars, speed)

ggplot(data=df, aes(df$speed)) +
  geom_histogram(aes(fill = ..count..)) +
  scale_fill_gradient("Frequency", low = "blue", high = "red") +
  labs(title = "Speed Counts") +
  labs(x = "Speed") +
  labs(y = "Count")
```

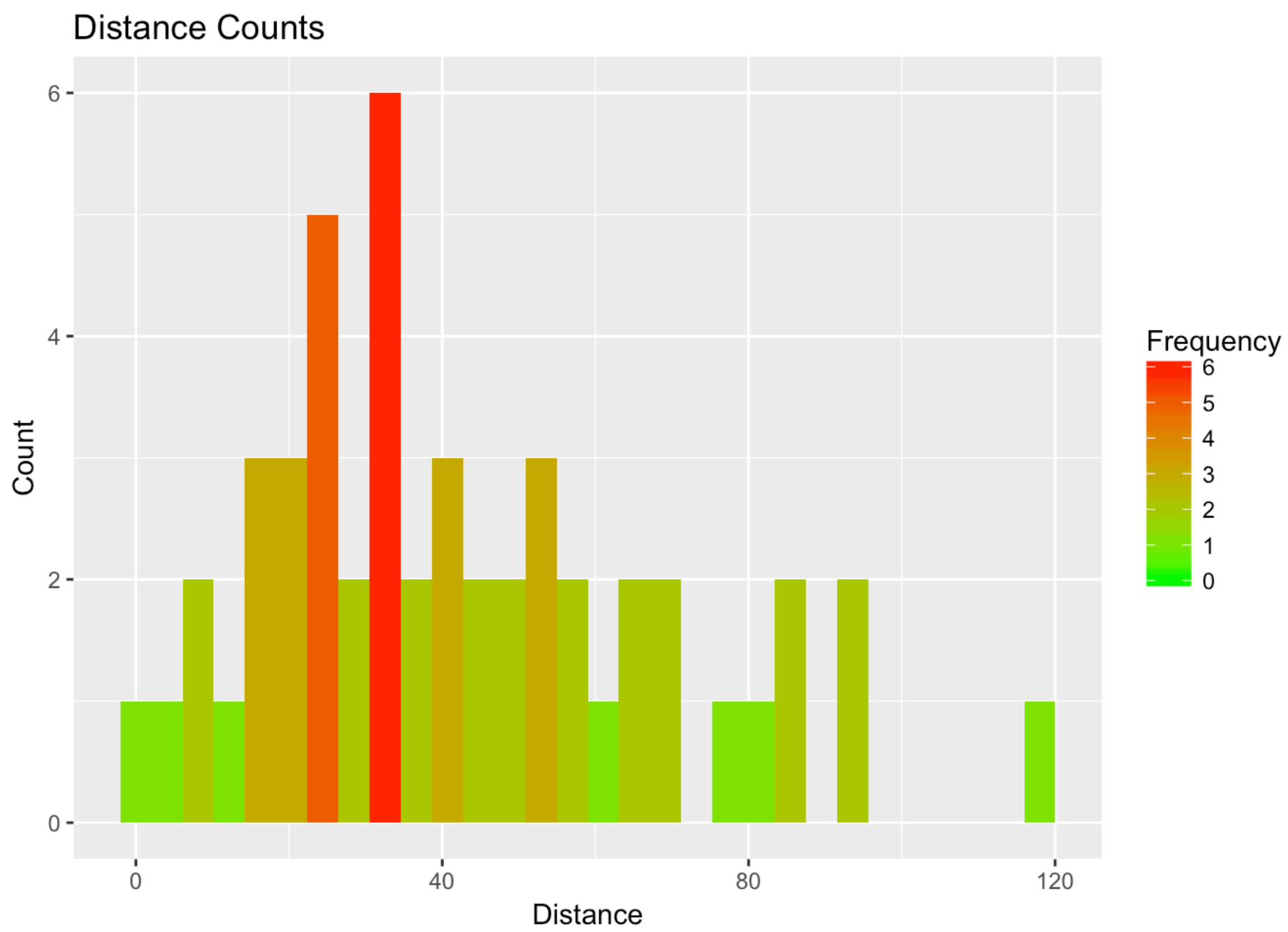
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The following is the histogram for distance.

```
ggplot(data=df, aes(df$dist)) +
  geom_histogram(aes(fill = ..count..)) +
  scale_fill_gradient("Frequency", low = "green", high = "red") +
  labs(title = "Distance Counts") +
  labs(x = "Distance") +
  labs(y = "Count")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The following is the correlation of the dataframe.

```
Correlation <- function() {
  x = cars$speed
  y = cars$dist
  c = cor(x, y)
  print (c)
  return (c)
}

d = Correlation()
```

```
## [1] 0.8068949
```

The following is the linear model calculations.

```
  model = lm (df$dist ~ df$speed, data = df)
  modelsum = summary(model)
  print(modelsum)
```

```
##
## Call:
## lm(formula = df$dist ~ df$speed, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
## df$speed      3.9324     0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

```
  model$coefficients[2]
```

```
## df$speed
## 3.932409
```

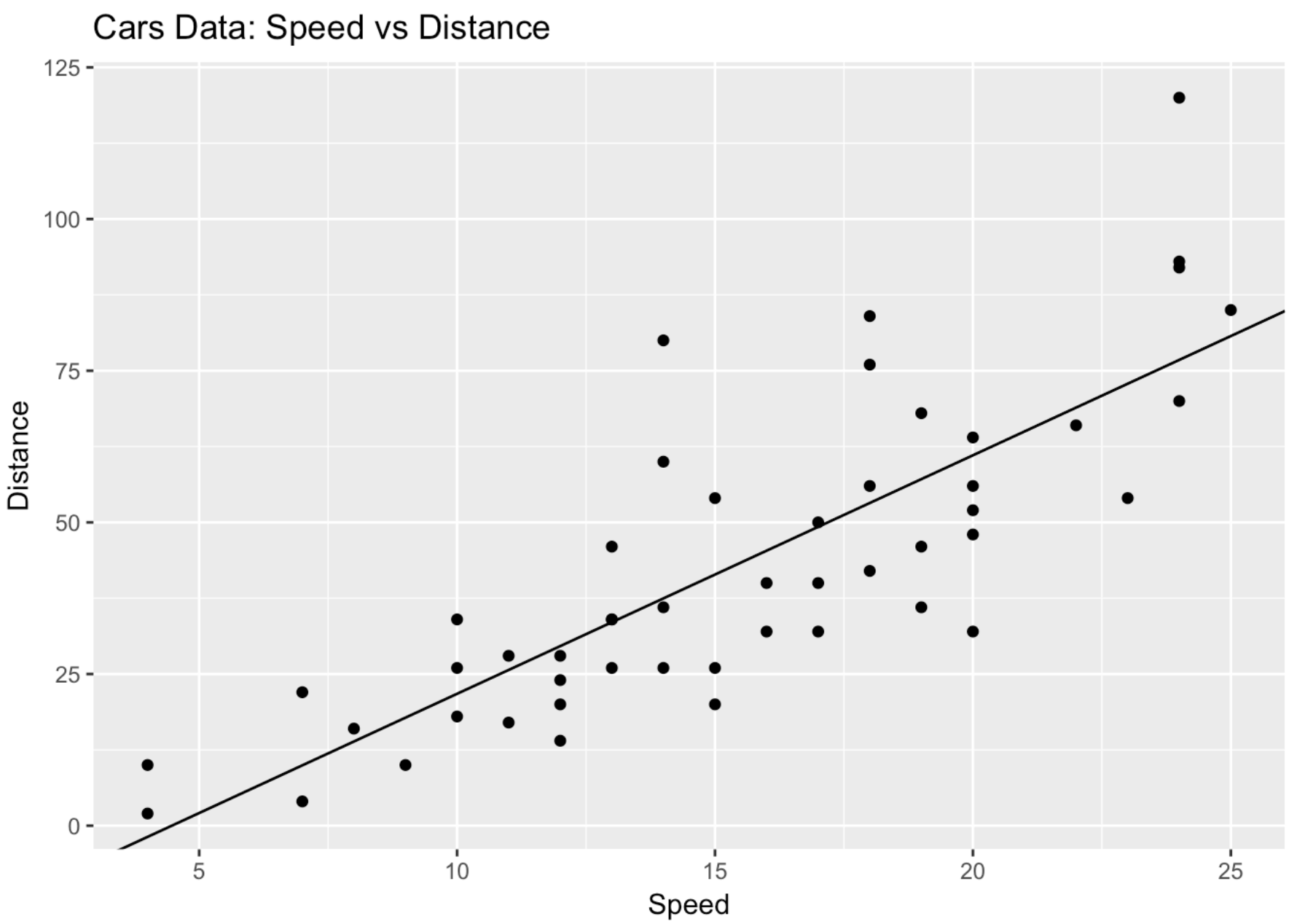
```
  model$coefficients[1]
```

```
## (Intercept)
##   -17.57909
```

The linear model is distance’ = -57909 + 3.932409

The following is the plot of the data and the linear model.

```
ggplot(df, aes(speed, dist)) + geom_point(colour="black") +
    geom_abline(aes(slope=model$coefficients[2], intercept=model$coefficients[1])) +
    labs(title = "Cars Data: Speed vs Distance") +
    xlab("Speed") +
    ylab("Distance")
```

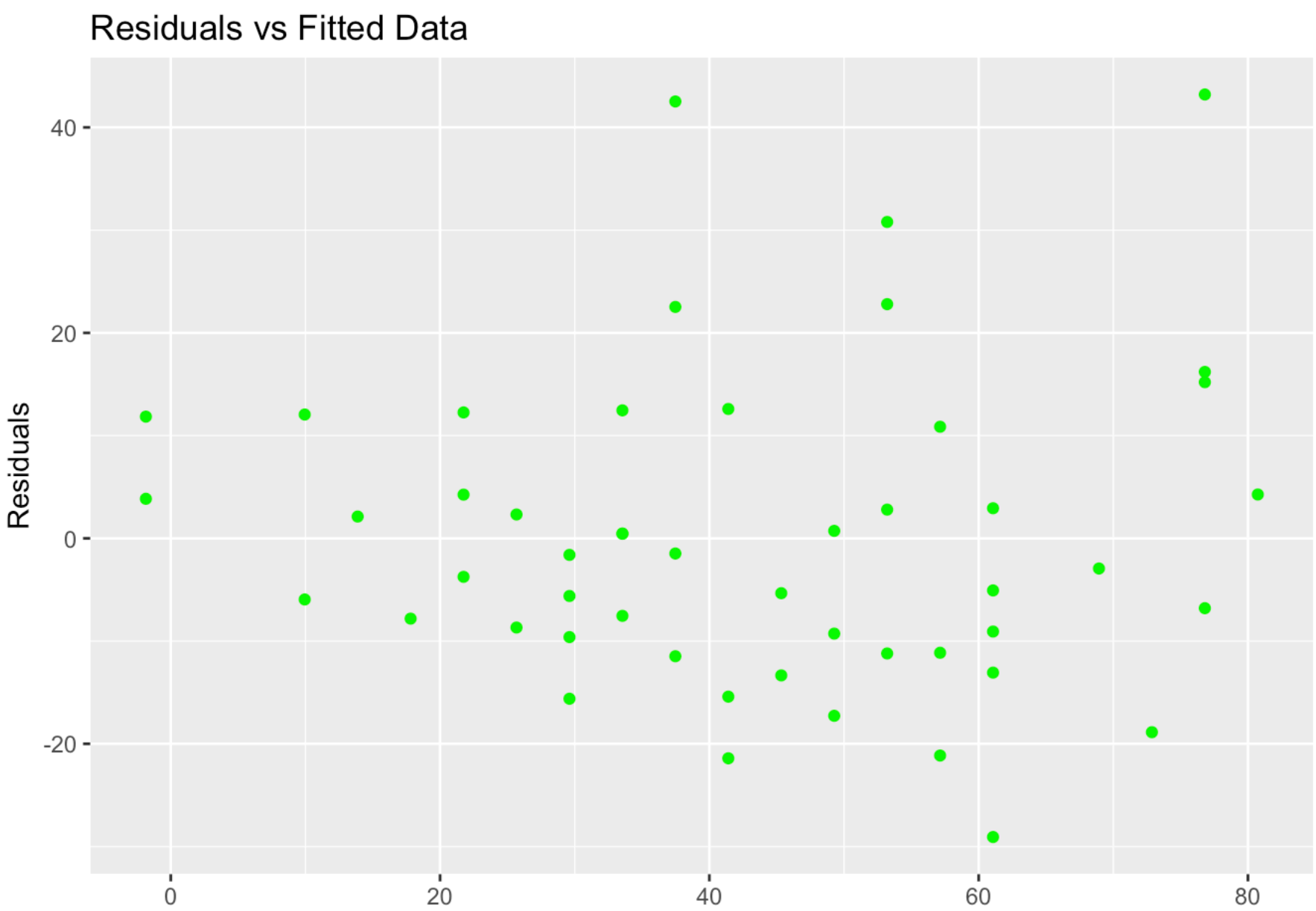


The model validity is:

Multiple R-squared: 0.6511, Adjusted R-squared: 0.6438

This model can only explain 2/3 of the data, meaning we could potentially develop better models if more variables were available.

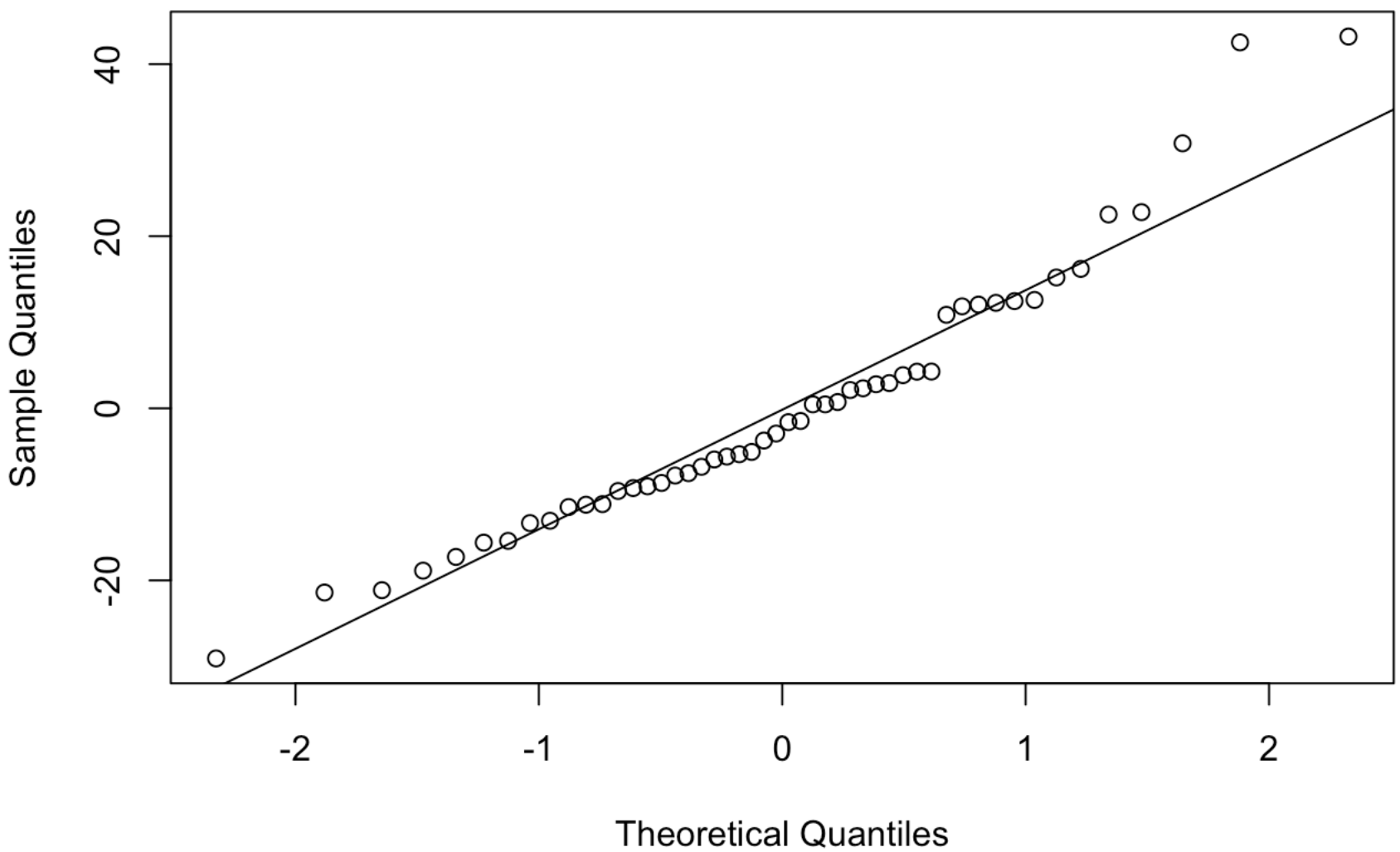
```
ggplot(model, aes(.fitted, .resid)) +
  geom_point(color = "green") +
  labs(title = "Residuals vs Fitted Data") +
  labs(x = "Model Values") +
  labs(y = "Residuals")
```



As you can see from the data there is no skewness or heteroskedasticity.

This can be further validated by using qq plots.

```
qqnorm(resid(model))
qqline(resid(model))
```



There may be some skewness on the right tail but its very minimal.

Overall, considering the data available the model has great correlation (>0.8) and better than average fit (R^2 ~ 2/3). There is likely other variables that are impacting the correlation between speed and distance that are not visible. Some items that come to mind may be road conditions, make and model of vehicle, time of day etc.