DATA605 Homework 11

Keith Colella

2023-11-12

```
library(tidyverse)
```

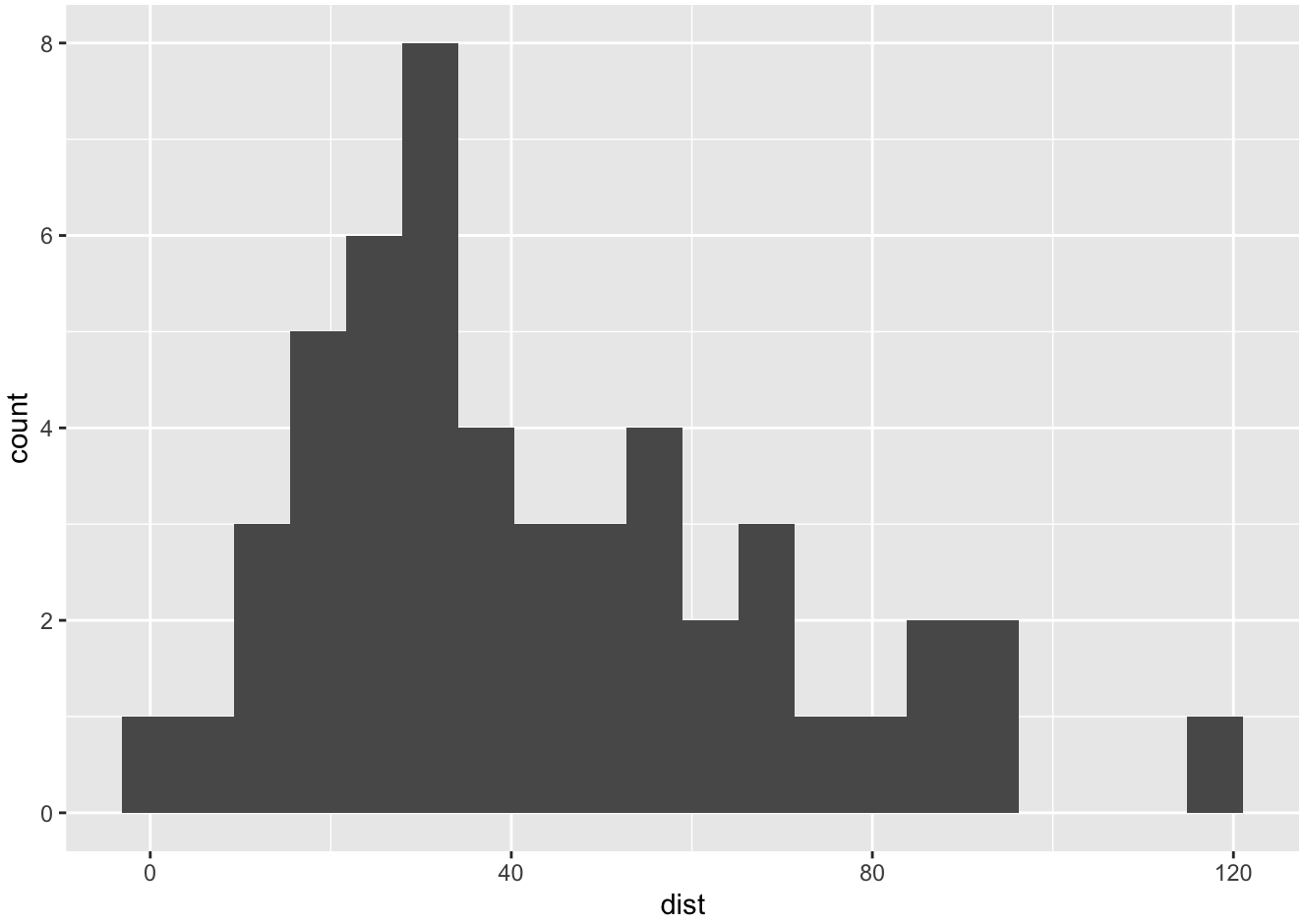
Assignment

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).

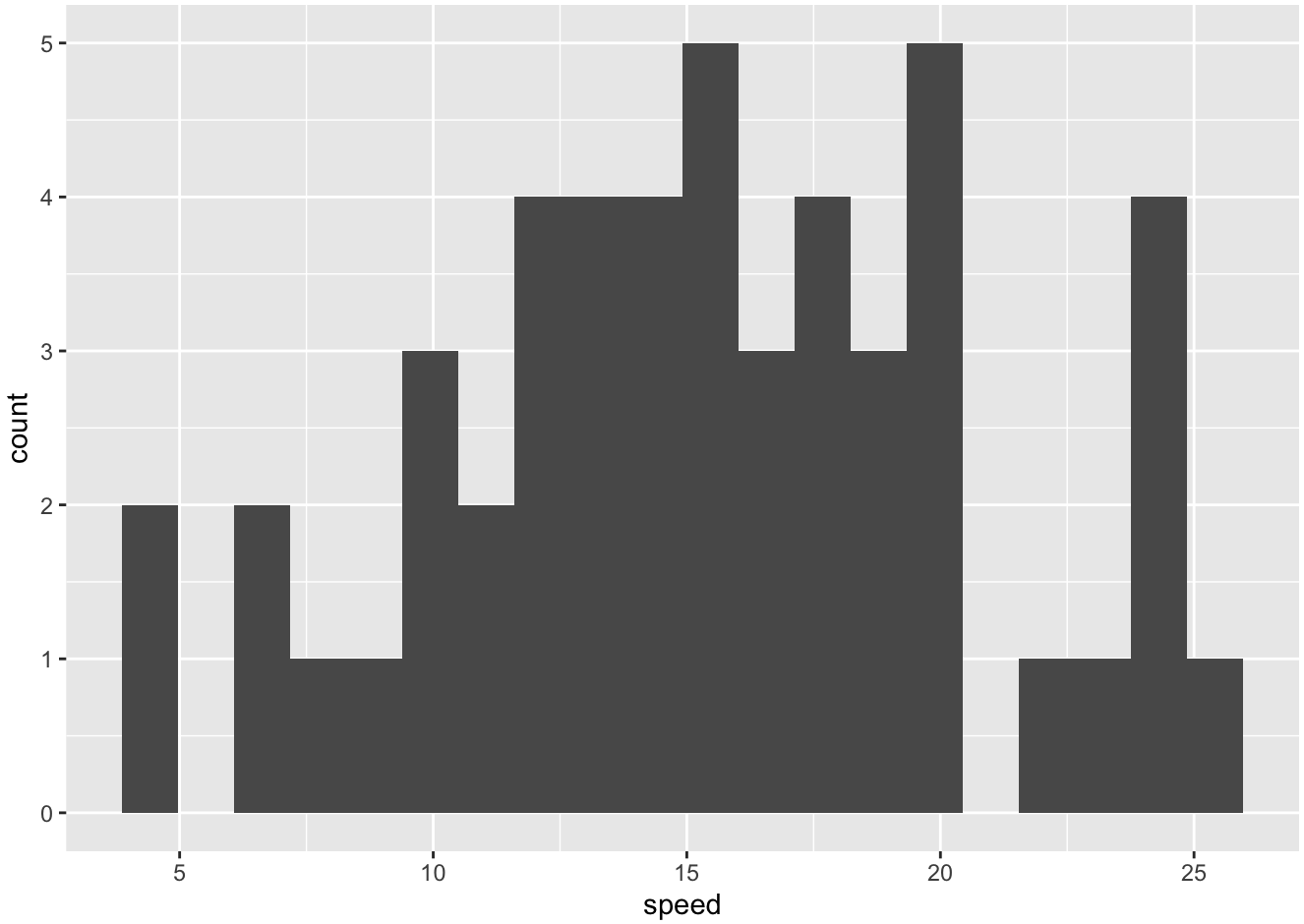
Response

We’ll start by loading in our dataset. Before we move to actual modeling, we’ll do some exploratory data analysis, plotting the distribution of each variable, along with a scatterplot examining the variables’ relationship.

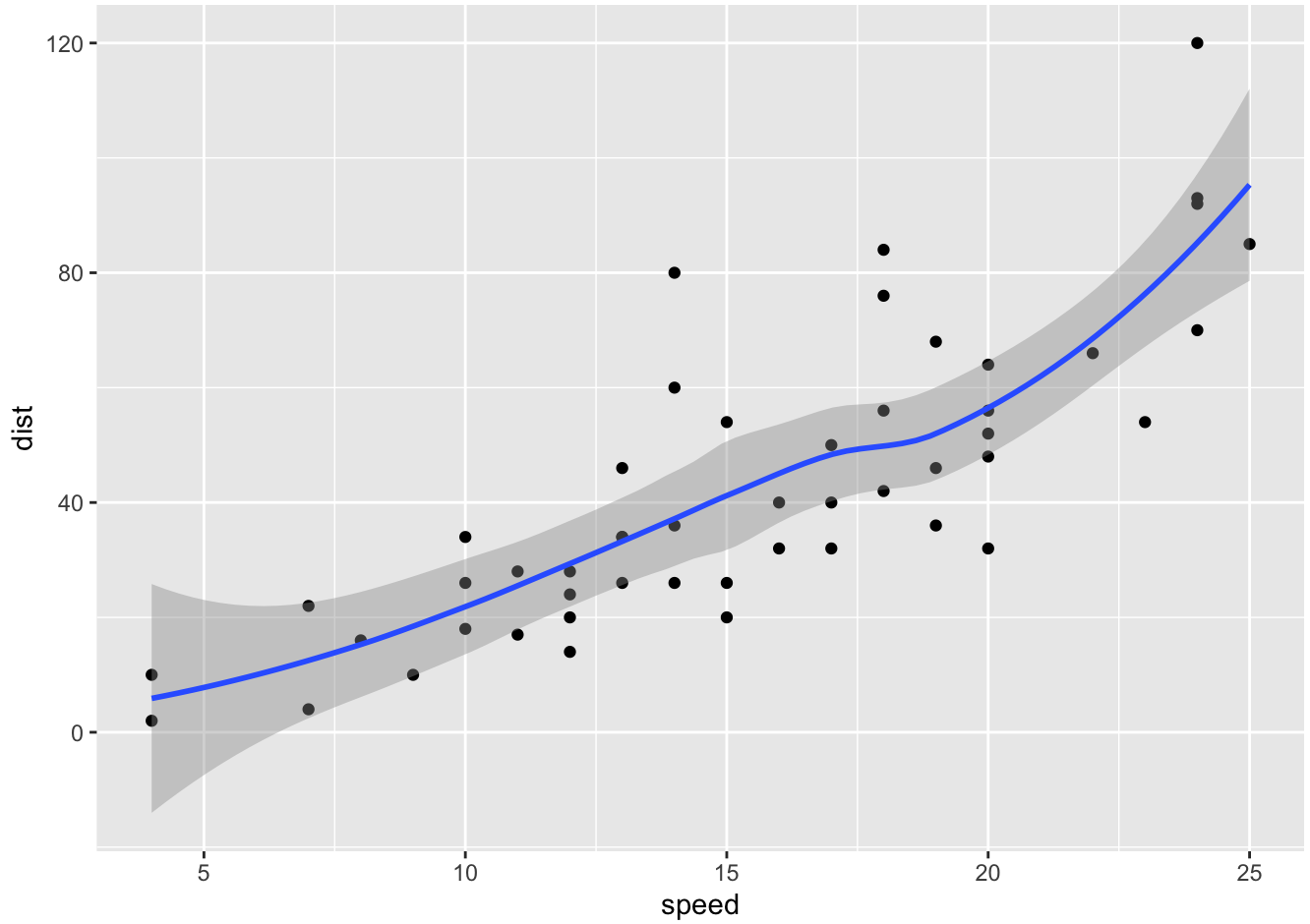
```
data(cars)

cars %>%
  ggplot(aes(dist)) +
  geom_histogram(bins = 20)
```




```
cars %>%
  ggplot(aes(speed)) +
  geom_histogram(bins = 20)
```




```
cars %>%
  ggplot(aes(speed, dist)) +
  geom_point() +
  geom_smooth(formula = 'y ~ x', method = 'loess')
```



The two variables appear non-normal. This doesn't necessary prove problematic for building a linear model. What's more important is the presence of a linear relationship between the two. And our scatter plot (and LOESS line) seem to indicate a relatively clear linear relationship.

We can now move ahead with the actual model.

```
model <- lm(dist ~ speed, data = cars)

summary(model)
```

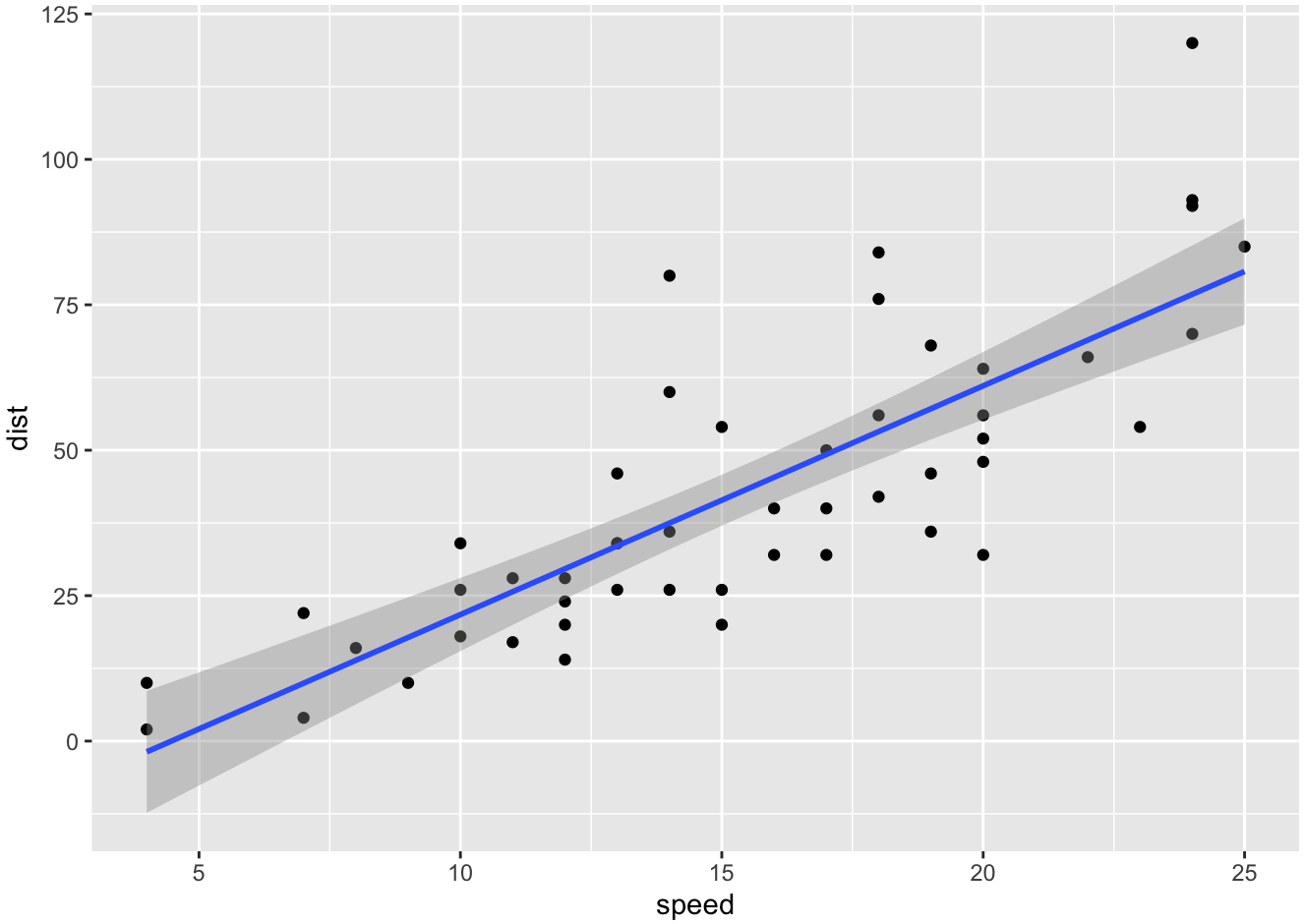
```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
## speed         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
```

I'll highlight some key takeaways from our summary:

1. Speed appears to be a statistically significant predictor of stopping distance. Our p-value for the t-test of variable significance is near-zero.
2. The model overall appears statistically significant, based on the near-zero p-value for our F-test. This comes as no suprise, given that our only predictor variable is significant.
3. Speed is able to explain a significant portion of the variance of stopping distance, as evidenced by the R^2 metric of 0.6511.
4. The β coefficient indicates that, for each 1 mph increase in speed, the stopping distance would be expected to increase by 3.9324 feet (we can find details on the units by calling `?cars`).
5. The model should provide us reasonably precise predictions, based on the relatively small standard error of 0.4155 feet.

Let's examine how our predictions match up with actual values visually.

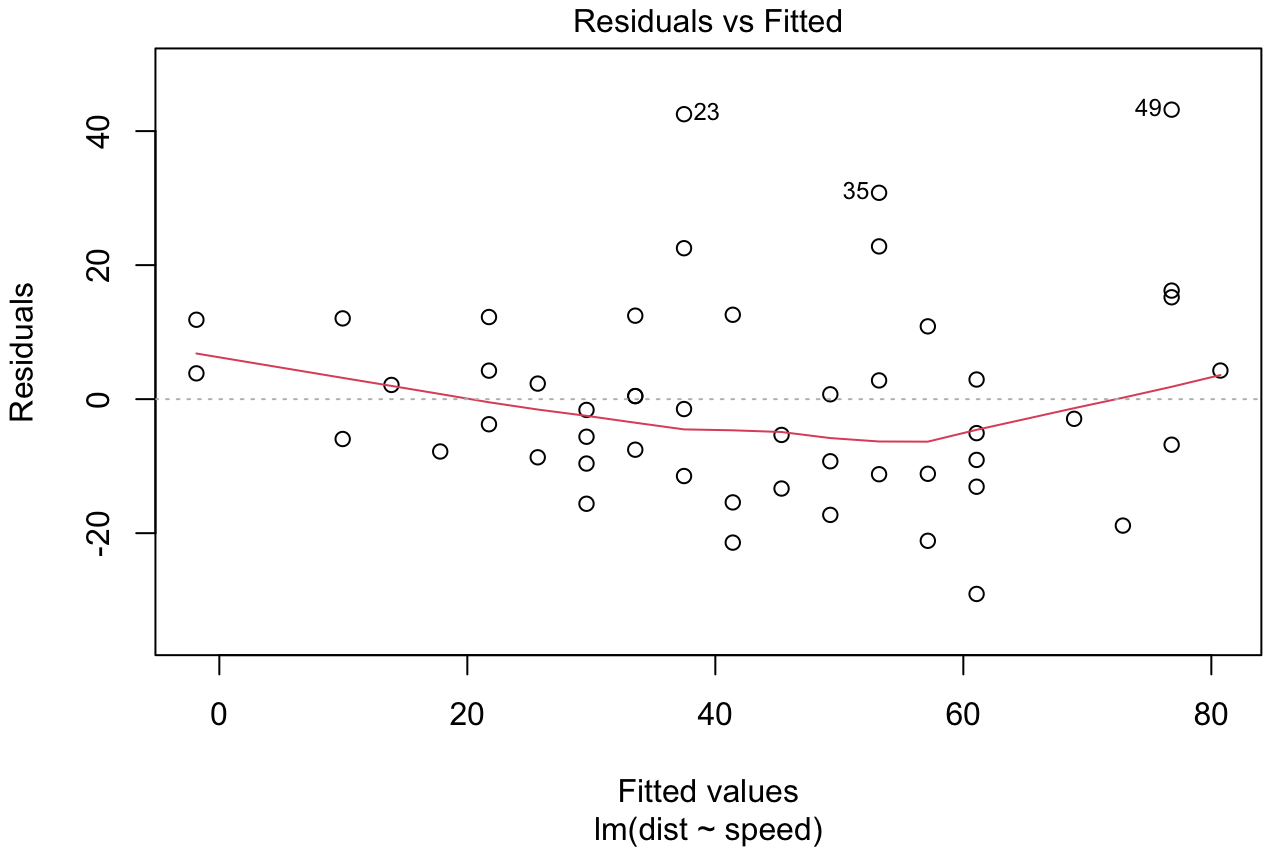
```
cars %>%
  ggplot(aes(speed, dist)) +
  geom_point() +
  geom_smooth(formula = 'y ~ x', method = 'lm', se = TRUE)
```

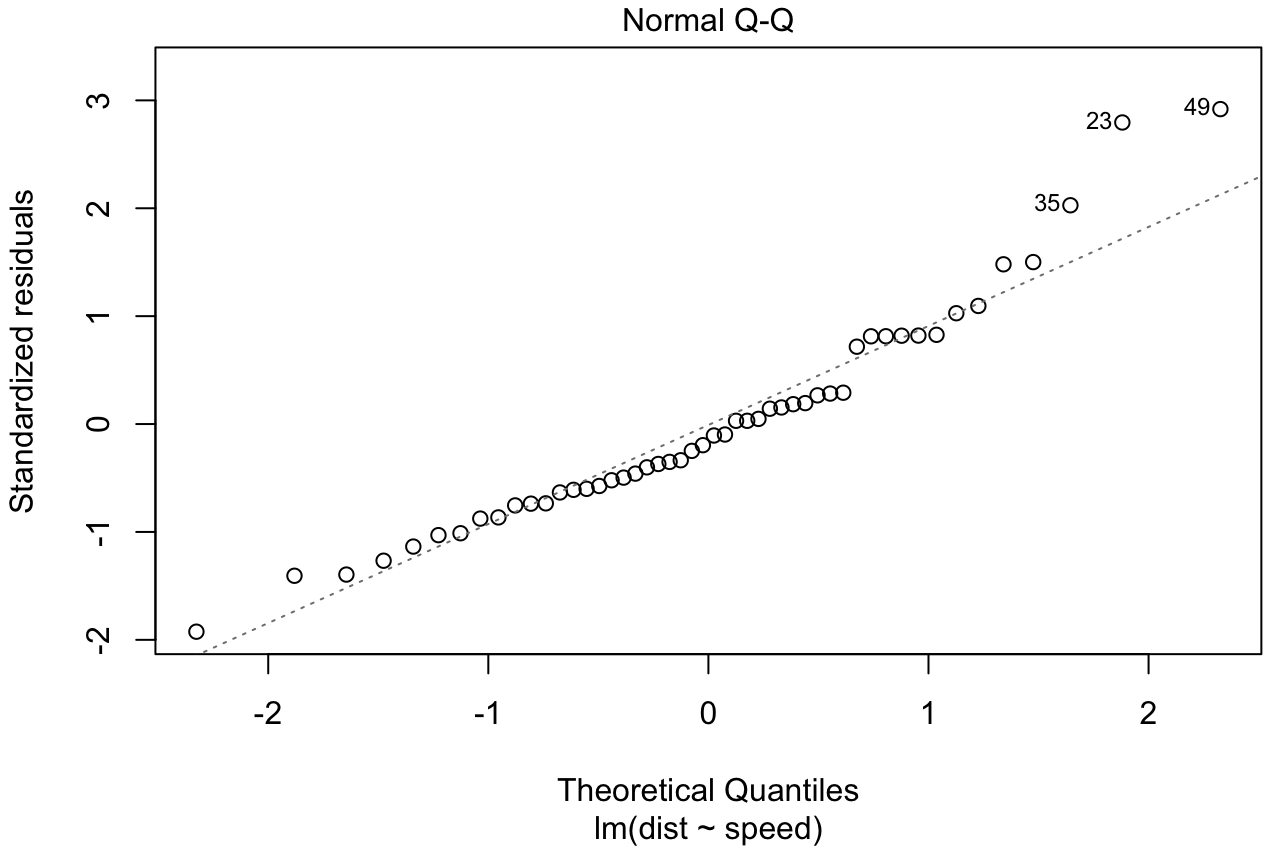


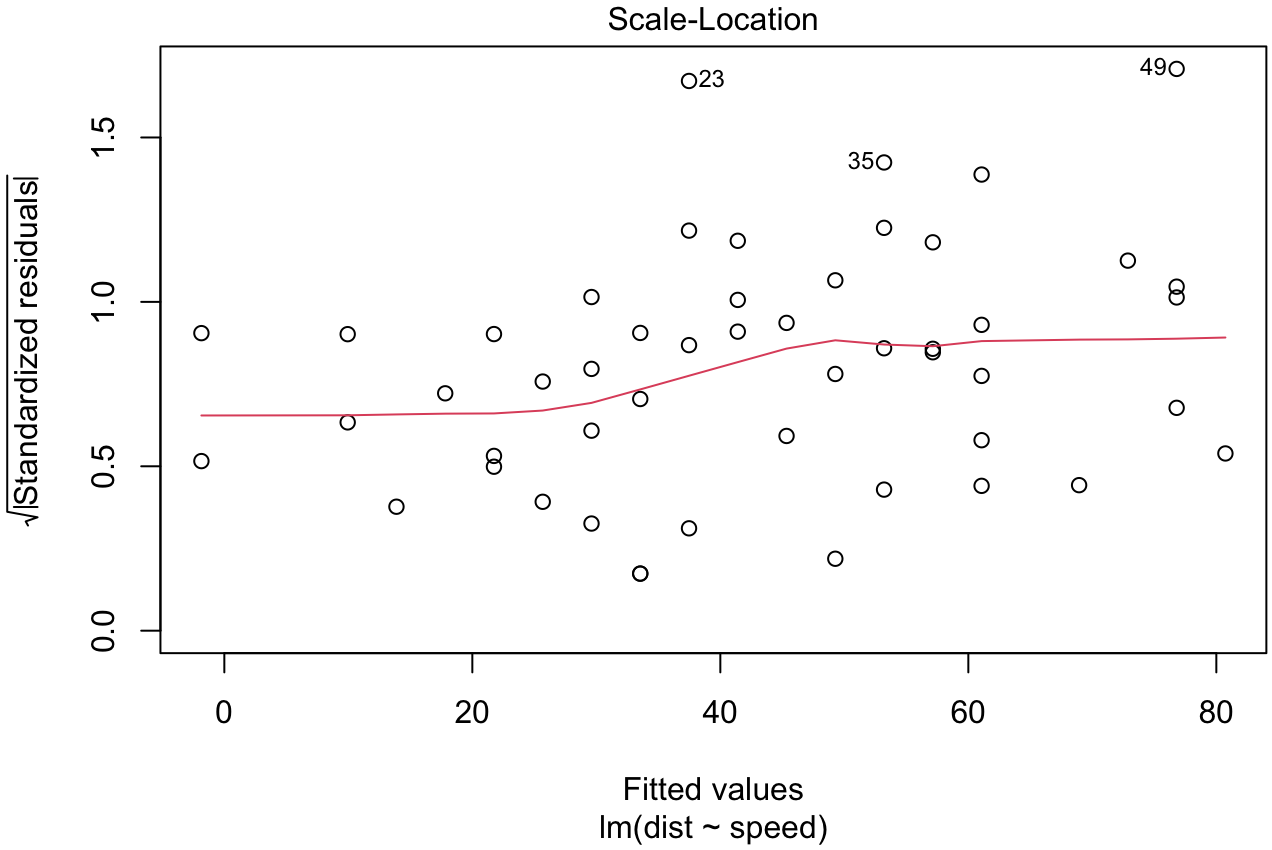
Our predictions seem to match relatively well to actual values.

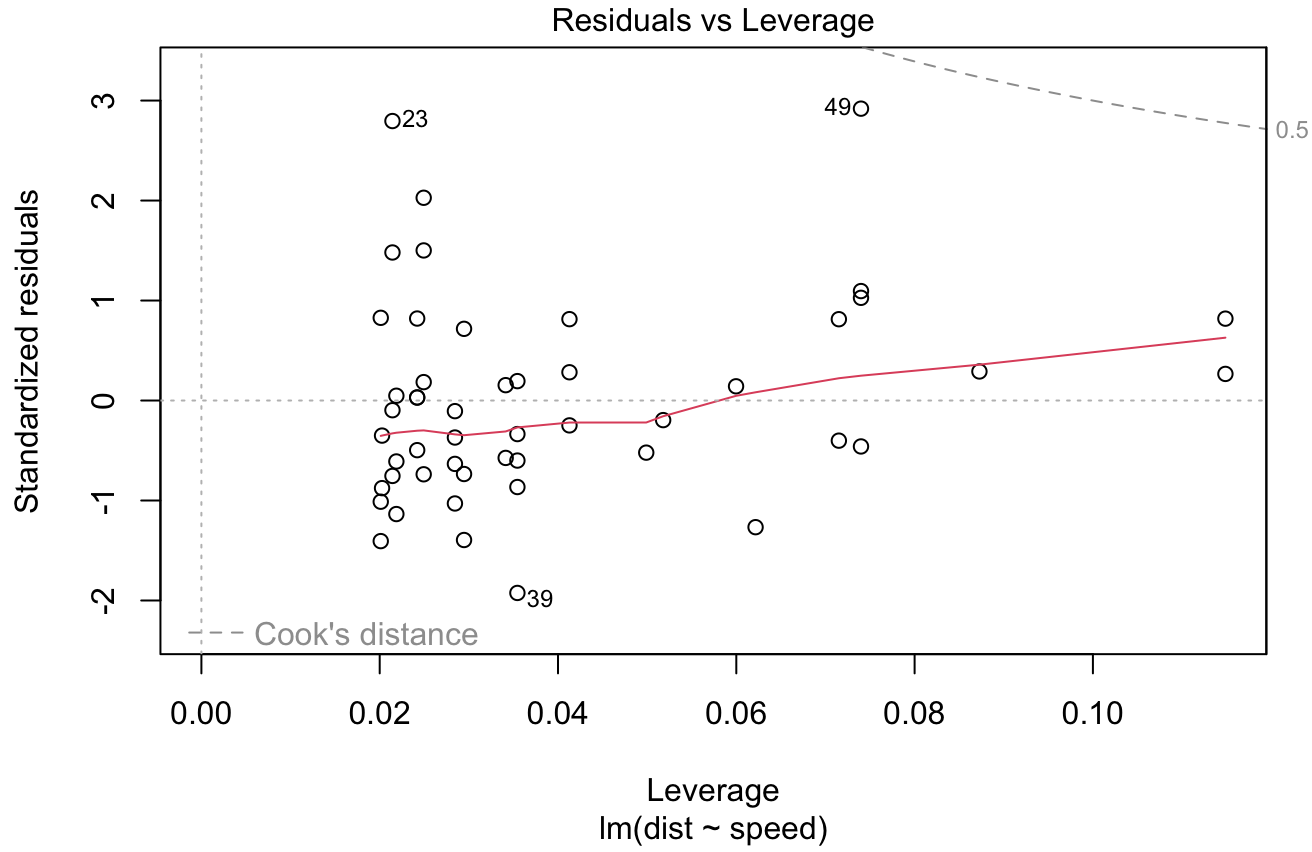
Finally, let's examine the residuals for any potential problems.

```
plot(model)
```


Residuals vs Fitted
Residuals
Fitted values
lm(dist ~ speed)
23
35
49
40
20
0
-20
0
20
40
60
80


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(dist ~ speed)
23
35
49
3
2
1
0
-1
-2
-2
-1
0
1
2


Scale-Location
√|Standardized residuals|
23
49
35
0.0
0.5
1.0
1.5
0
20
40
60
80
Fitted values
lm(dist ~ speed)


Residuals vs Leverage
Standardized residuals
23
49
39
0.5
Cook's distance
-2
-1
0
1
2
3
0.00
0.02
0.04
0.06
0.08
0.10
Leverage
lm(dist ~ speed)

Looking at fitted versus residuals plots, we see no evidence of heteroskedasticity or non-constant variance. The residuals seem randomly distributed around zero.

Looking at our QQ plot, the residuals seem relatively normally distributed. There are some outliers in the right tail, but overall, the residuals sit along the normal diagonal.

Finally, our leverage plot indicates that, for the most part, single observations do not hold undue influence over parameter estimates. We see some tail values that have higher leverage, but nothing too concerning.

Overall, our model seems to provide a good fit and key regression assumptions hold!