Stopping Distance

Sam Reeves

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 Model

```r
library(ggplot2)
library(cowplot)
library(lmtest)

cars <- datasets::cars

(model1 <- lm(dist ~ speed, cars))
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Coefficients:
## (Intercept)        speed
##     -17.579        3.932
```

$$E(y|x) = \beta_0 + \beta_1 X + \epsilon$$

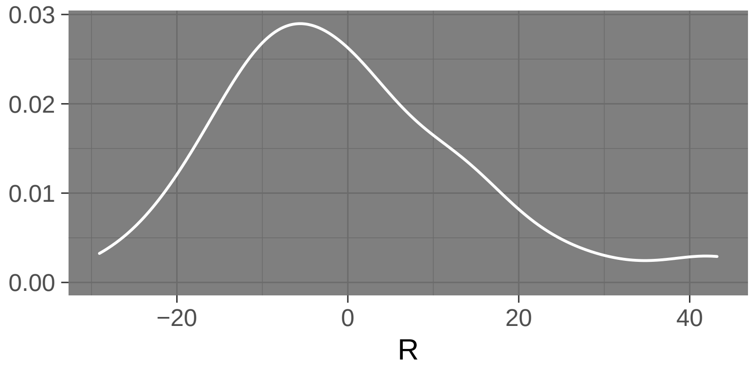
```
p1 <- ggplot(model1, aes(x = model1$residuals)) +
  geom_density(color = "white") +
  labs(x = "R",
       y = NULL,) +
  theme_dark()

p2 <- ggplot(model1,
       aes(model1$residuals^2, breaks=50)) +
  geom_histogram(fill = "white") +
  geom_density(aes(model1$residuals)) +
  theme_dark() +
  labs(x = "R^2",
       y = NULL)
```

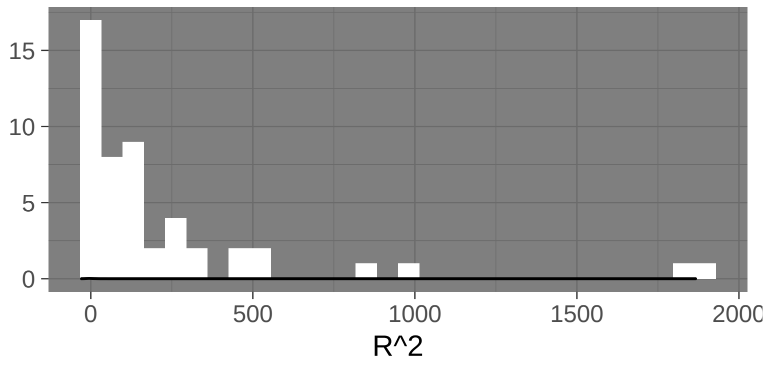
Original Data

We can see now how the residuals shape up around a fit line:

```
p1
```



```
p2
```



The original data's residuals are skewed right, but appear fairly normally distributed.

Removing the outliers progressively and making two new linear models with points whose R^2 are below thresholds 1000 and 400:

```r
which(model1$residuals^2 > 1000)
```

```
## 23 49
## 23 49
```

```r
no_outliers_1000 <- cars[-c(23,49), ]
model2 <- lm(dist ~ speed, no_outliers_1000)

which(model1$residuals^2 > 350)
```

```
## 22 23 24 34 35 36 39 45 49
## 22 23 24 34 35 36 39 45 49
```

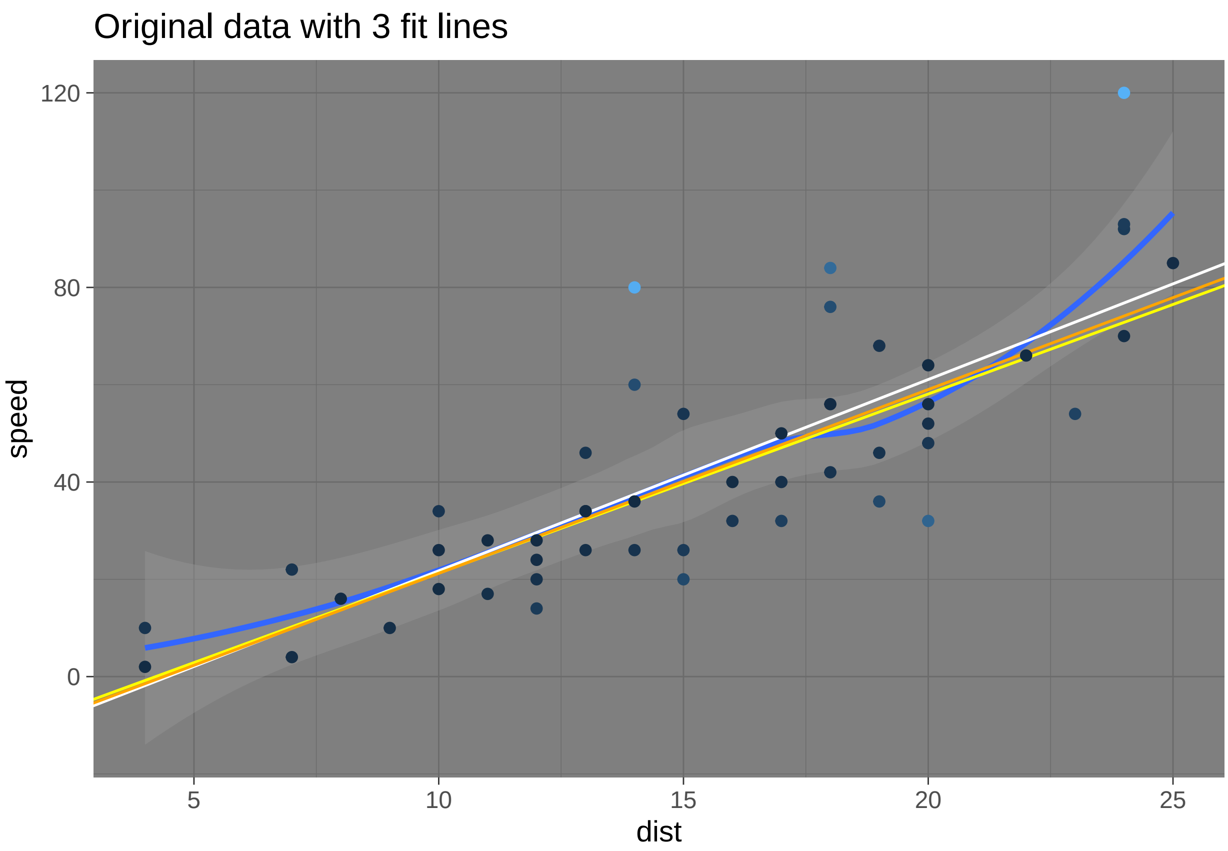
```r
no_outliers_400 <- no_outliers_1000[-c(22,23,24, 34,35,36,39,45,49),]
model3 <- lm(dist ~ speed, no_outliers_400)
```

Visualization

```r
p3 <- ggplot(cars, aes(x = cars[,1],
                 y = cars[,2],
                 color = model1$residuals^2)) +
  geom_smooth() +
  geom_abline(intercept = model1$coefficients[1],
              slope = model1$coefficients[2],
              color = "white") +
  geom_abline(intercept = model2$coefficients[1],
              slope = model2$coefficients[2],
              color = "yellow") +
  geom_abline(intercept = model3$coefficients[1],
              slope = model3$coefficients[2],
              color = "orange") +
  geom_point() +
  theme_dark() +
  xlab(NULL) +
  ylab(NULL) +
  theme(legend.position = "none") +
  labs(title = "Original data with 3 fit lines",
       x = "dist",
       y = "speed")

p3
```

```
## ‘geom_smooth()‘ using method = ’loess’ and formula ’y ~ x’
```



```
model1$coefficients
```

```
## (Intercept)       speed
##  -17.579095    3.932409
```

```
model2$coefficients
```

```
## (Intercept)       speed
##  -15.533643    3.681168
```

```
model3$coefficients
```

```
## (Intercept)       speed
##  -16.576459    3.779011
```

Evaluation

Just visually, stopping distance is very obviously well-approximated by a fit line from 20mph to around 60mph or 70 mph. There are many factors involving the health of a vehicle or its weight, however, the stopping distance is generally described:

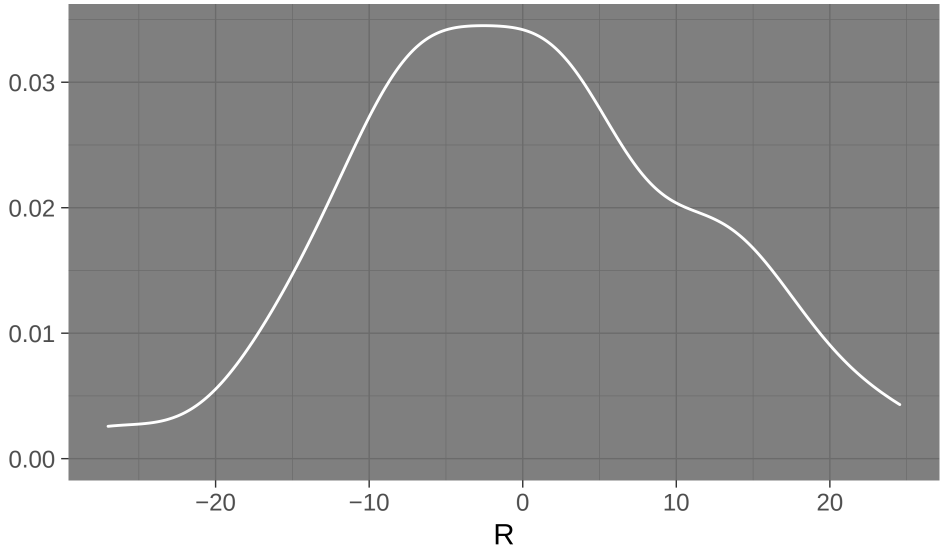
$$dist = \frac{speed^2}{2 \times friction \times gravity)}$$

$$dist = k \times speed^2$$

It stands to reason this would be a good approximation.

```
p4 <- ggplot(model3, aes(x = model3$residuals)) +
  geom_density(color = "white") +
  labs(x = "R",
       y = NULL,) +
  theme_dark()

p4
```



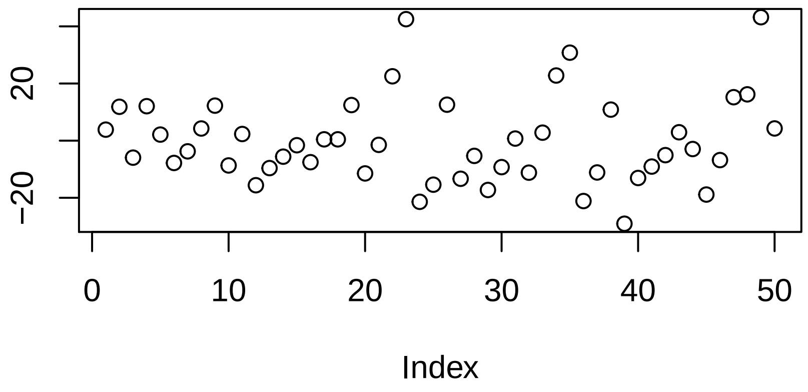
The cherrypicked points are nearly normally distrbuted around zero.

Residual Analysis

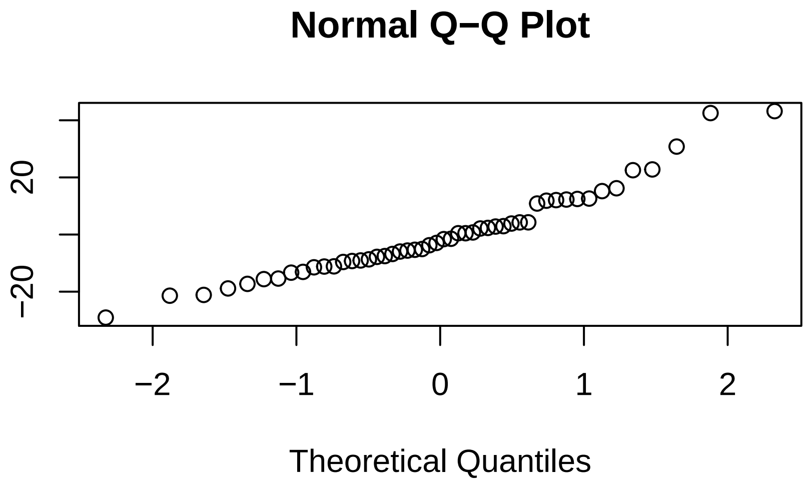
The observations are independent, but the variance of residuals increases slightly with X. This also coincides perfectly with my personal knowledge of car accidents. I think the reader will relate.

Model1 on the original dataset

```
plot(model1$residuals, ylab='')
```

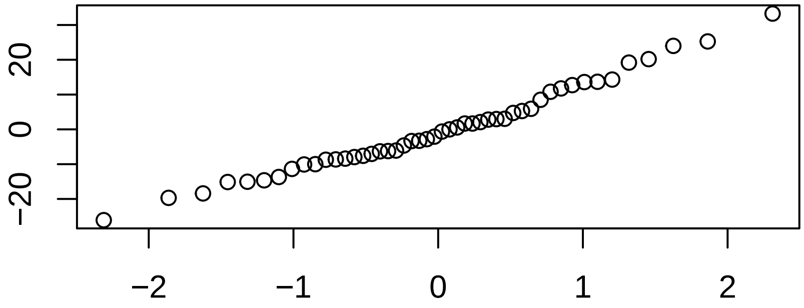


```
qqnorm(model1$residuals, ylab='')
```

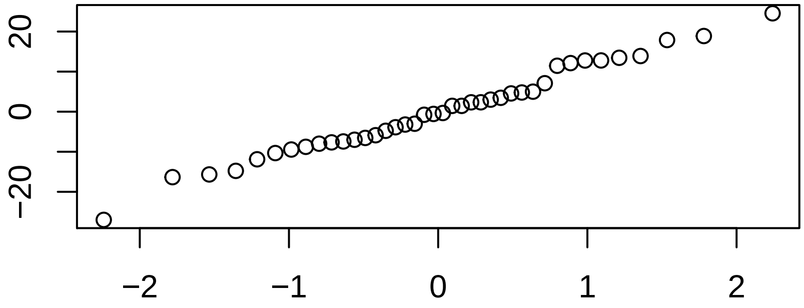


Models 2 and 3 on the progressively trimmed datasets

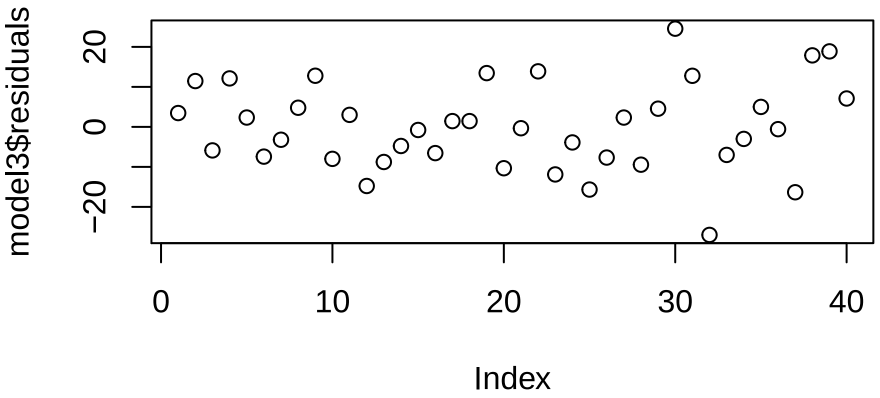
```
qqnorm(model2$residuals,
       main="", xlab="", ylab="")
```



```
qqnorm(model3$residuals,
       main="", xlab="", ylab="")
```



```
plot(model3$residuals)
```



Cherrypicking the data has left us with highly normally distributed residuals. This means the variance is consistent enough across the data, that we have passed all three tests for the linear model.

```
bptest(model1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model1
## BP = 3.2149, df = 1, p-value = 0.07297
```

```
bptest(model3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model3
## BP = 4.7082, df = 1, p-value = 0.03002
```

The third model also has a p-value below 0.05, while the model for the original data is above. The linear model for the original data is not strong enough to rule out h_o in this case. However, the third model is probably a solid approximation for for the data with outliers thrown out. Values tested for $R^2 < 350$ and $R^2 > 350$ did not pass this test.