Homework 2: Stopping Distance

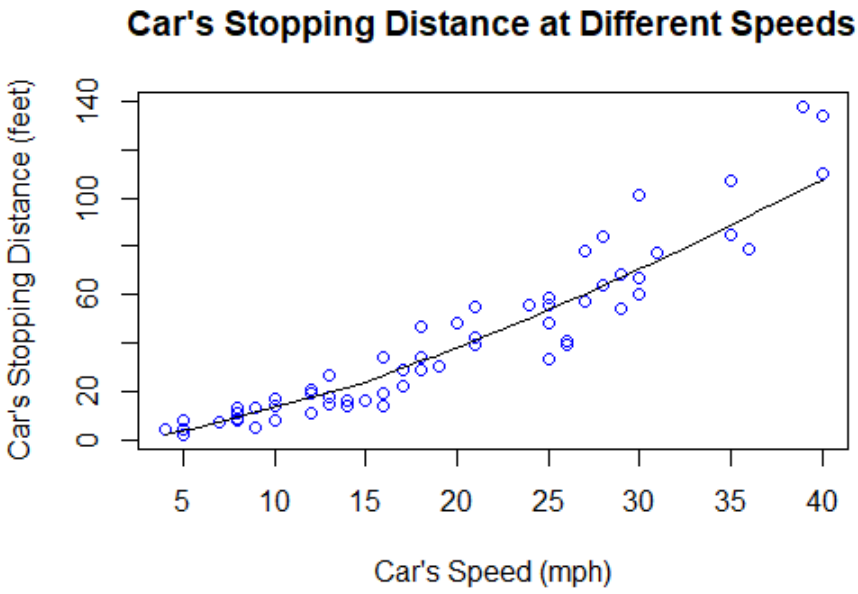
Kevin Toney

September 30, 2017

Question 1. The main purpose of this study is to compare the distance (in feet) required for a car to stop on a rural road to the car's speed. We must make this comparison so that your agency can better determine speed limits, in miles per hour, on this road to protect the residents who may be on the road, and to help drivers get to their destinations in an adequate amount of time. Studying the relationship between speed and stopping distance will also help us make predictions on how long it will take a car to stop in order to avoid a pedestrian or obstacle.

A statistical model that may answer your questions is simple linear regression. The purpose of linear regression is to determine the relationship between two factors, and use the relationship to predict one of their values. This model will help us study and predict what the stopping distance of a car will be if a car is going a certain miles per hour.

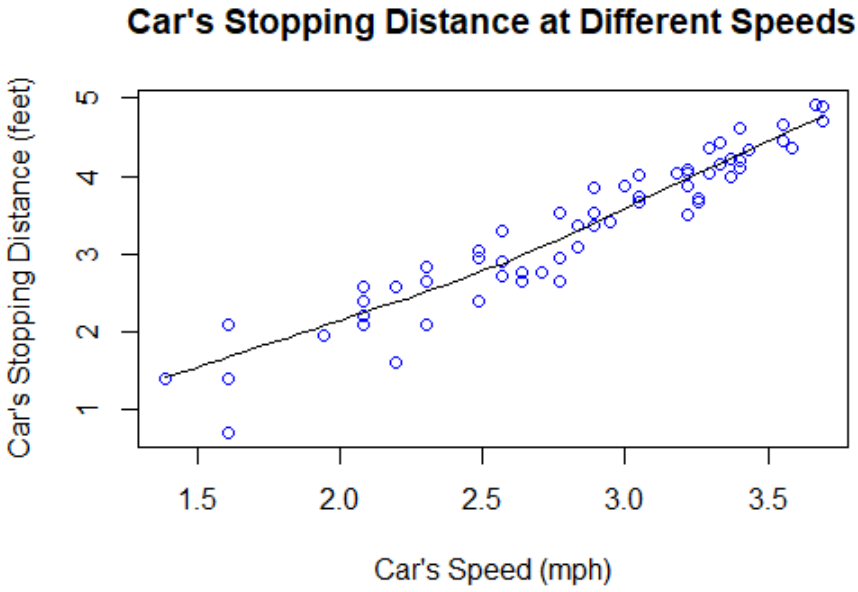
Question 2. Is simple linear regression appropriate for this situation? Let's start by seeing a plot of a car's stopping distance required to stop on a certain rural road in your area, due to many different speeds.



The upward curve in the plot suggests that the required stopping distance increases as a car's speed increases. The data points seem to follow an upward pattern similar to a line. To confirm my observations, I used a metric called correlation. Correlation measures the strength and direction of a linear relationship. A correlation of positive 1 is a perfect score for an upward linear relationship; a negative 1 is a perfect score for a downward linear relationship. The correlation is 0.93, which is close to a score of 1. Therefore, the relationship is positive and linear. I confirmed that the required stopping distance of a car on a certain rural road increases as the speed does.

Despite all of these metrics and observations, I am not justified in using simple linear regression. First off, there are four points on the graph which seem to be anomalies. Moreover, there are four requirements our data must fulfill in order to perform linear regression and see the relationship between the two factors as they are. After my assessments, I determined the data did not fulfill all of the requirements. I will talk about these requirements more in depth later in this report.

Question 3. In order to more fully fulfill these four requirements, I needed to transform each data point by something called a natural log. After transforming each data point, we get the below plot.

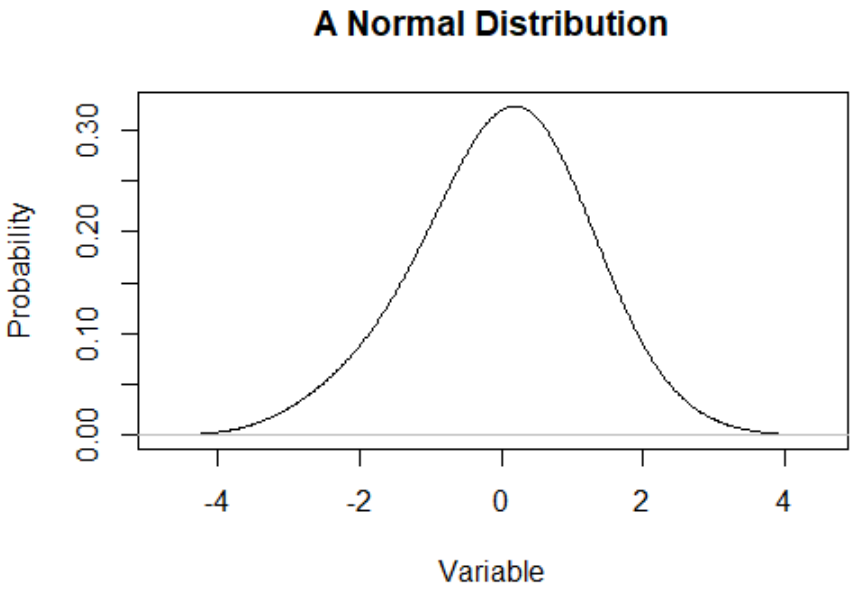


Now, the graph seems to show more of a linear relationship than before. The previous plot followed a curve. Yes, there are still some anomalies, especially at the beginning for the graph. Nonetheless, the requirements for simple linear regression are met more fully if I transform all of the data by the natural log (ln). At the end of the study, I will transform the data back to its original form so we can interpret my predictions.

The linear regression model we will use to predict the required stopping distance of a car, in relation to the car's speed is:
$ln(Distance_i) \sim N(\beta_0 + \beta_1 * ln(Speed_i), \sigma^2)$

i = a certain number of car from 1 to 62.

N = The normal distribution, which follows a bell shaped curve similar to the one below.



β_0 = The required stopping distance of a car, on average, if the car's speed was 0 mph.

β_1 = The rate at which the stopping distance of a car increases if the speed rises by 1 mph.

σ^2 = The average amount the required stopping distances are from the average.

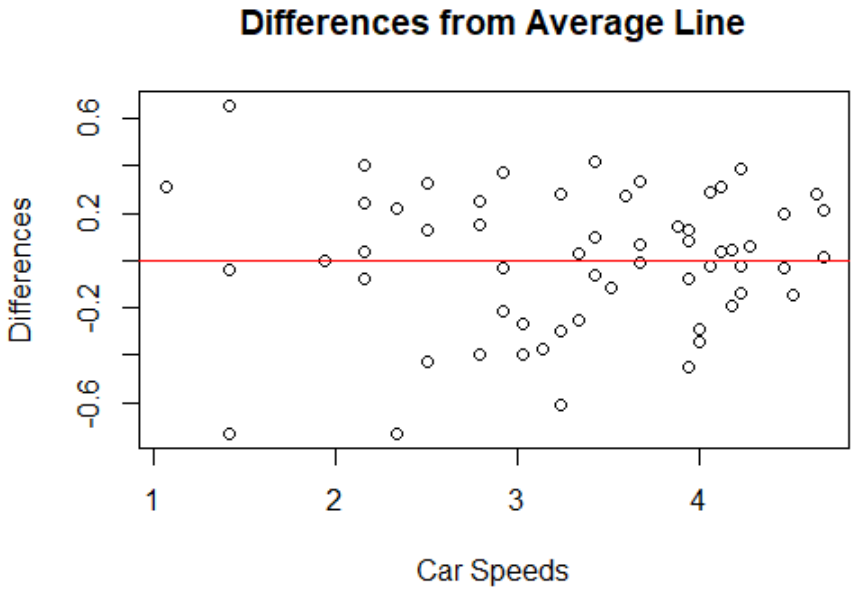
I assumed each recorded stopping distance doesn't affect other recorded distances. Moreover, I assumed the distance's variability follows a normal distribution similar to the graph on the top of the page. Also, I assumed σ^2 of the stopping distances is equal enough all across the average line created by the linear equation $\beta_0 + \beta_1 * ln(Speed_i)$. Finally, I assumed the relationship between the two variables are linear.

In the transformed data, the model fulfills the assumptions and requirements enough. I will now be able to use the linear equation to determine the relationship between speed and stopping distance. I will also be able to predict a stopping distance if a car is going a certain mph.

Question 4. Our model's requirements/assumptions are met because of these reasons:

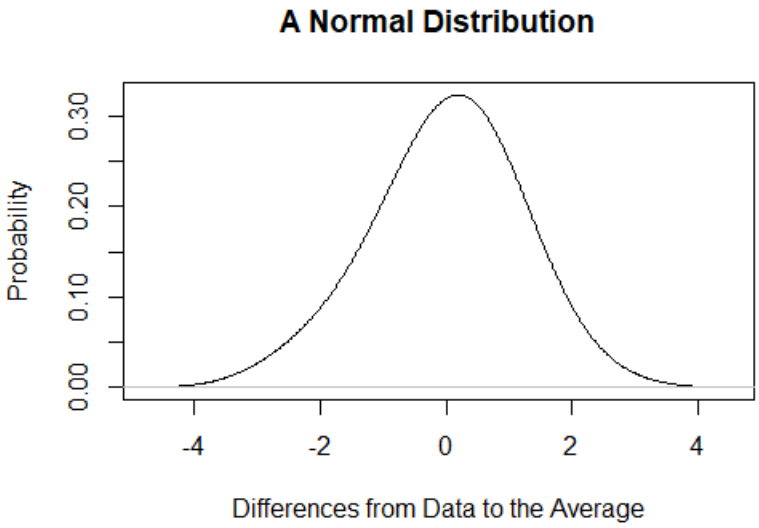
- The correlation score showed a stronger linear relationship than before the transformation.

- Consider the graph below:



There is no pattern in the data points, such as a wave, line, or clump. Additionally, no chunks of points are only above or below the point in succession. Therefore, each stopping distance seems to not be affected by others.

- I ran a test called the KS test to see if the differences of each point from the average followed a normal distribution. The test gave me a metric of 0.5781, which isn't small enough to prove that the differences do not follow a normal distribution (the bell curve). Please consider the graph below for a visual representation of these differences.



Our statistical model explains 90% of the stopping distance's variability. This score shows the model is optimal for finding a relationship between stopping distance and speed. On average, my model of the data, when it is transformed, tends to predict lower stopping distances than the distances on the average line. If I was to make predictions over and over again, the car's predicted stopping distance, on average, would be off of the average line by 10.35 feet. Therefore, our predictions have error associated with them. Your agency needs to be aware of the error while you are considering different speed limits, but also please be aware the accuracy of the predictions is satisfactory.

Question 5. The model, or the linear equation that provides the predicted average line of the data with the least amount of error is below:
$ln(Distance_i) \sim N(\hat{\beta}_0 + \hat{\beta}_1 * ln(Speed_i), \hat{\sigma}^2)$

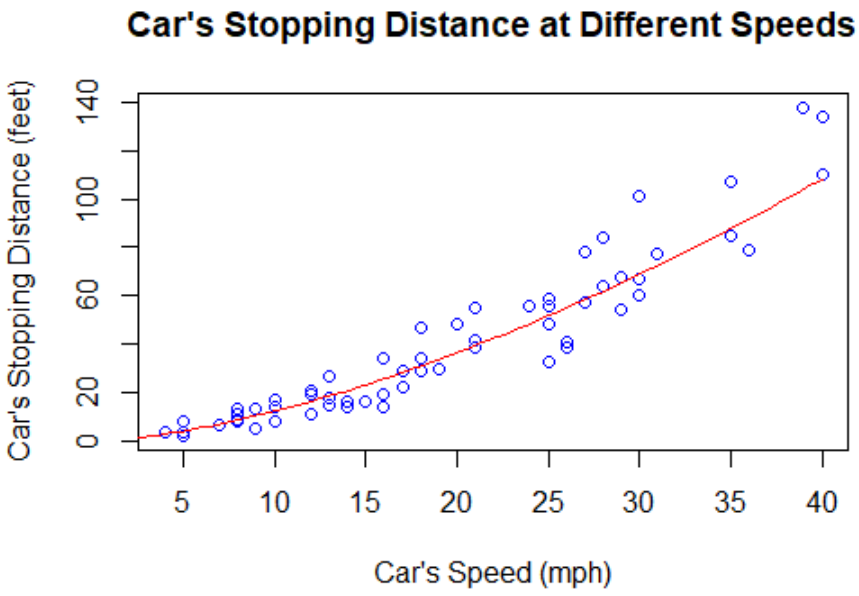
$i = 1$ to the number of recorded stopping distances (62).

$\hat{\beta}_0$ = The predicted required stopping distance, given the car's speed equals 0 mph.

$\hat{\beta}_1$ = The predicted amount of the predicted required stopping distance will change if the car's speed rises by 1 mph.

$\hat{\sigma}^2$ = The average distance of the predicted required stopping distance of the car from the average line.

The stopping distance data with the predicted mean line, which has the least amount of error, is plotted below:



The red line shows the average line of the required stopping distance in relation to the car's speed in mph. This line is more similar to a curve, but it has the least amount of prediction error. Therefore, I am confident the required stopping distances will follow this curve if I predict stopping distances when a car is going between 5 and 40 mph.

Question 6. The statistical model we are using to determine the relationship between a car's speed and stopping distance predicts that a car will need 87.6 feet, on average, to stop fully if it is going 35 mph at the beginning.

If your agency reduced the speed limit to 30 mph, and a car was truly going 30 mph, they would need 68.8 feet, on average, to fully stop. This would be 18.8 feet difference, which would account for a 21.5 percent reduction from a speed limit of 35 mph.

Since this rural road passes through a neighborhood with many homes, I suggest lowering the speed limit to 30 mph. This neighborhood will likely have lots of children who may cross the road spontaneously. Moreover, the rural neighborhood may attract deer, especially if the neighborhood is located near mountains. An 18 foot difference in stopping distances is significant enough to protect children, pedestrians, and drivers if a car needs to stop as quick as possible. Please set the road's speed limit as 30 mph.

A R Code:

```
rm(list=ls())

stop.dat <- read.table("Fall 2017/STAT 330/Homework/Homework2/
stopping_distances.txt", header=T, sep="")
#the dependent variable is the car's stopping distance in feet
#the explanatory variable is the car's speed in mph.

speed <- stop.dat$Speed
distance <- stop.dat$Distance

plot(speed, distance, col="blue", xlab="Car's Speed (mph)",
     ylab="Car's Stopping Distance (feet)",
     main="Car's Stopping Distance at Different Speeds")
#plot the data
scatter.smooth(speed, distance, col="blue", xlab="Car's Speed (mph)",
     ylab="Car's Stopping Distance (feet)",
     main="Car's Stopping Distance at Different Speeds")
cov(speed, distance)
cor(speed, distance)
#according to the plot, covariance, and correlation, there is a
#linear relationship strong enough to perform simple linear regression.
#covariance and correlation show a positive relationship.
#a car's stopping distance will increase as the car's speed
#increases by one mph.

linear_model <- lm(Distance ~ Speed, data=stop.dat)
summary(linear_model)

####Test L-I-N-E assumptions.
```

```
linear_model.res <- resid(linear_model)
#residuals vs. fitted values plot
#The plot doesn't fully support equal variance all along
#the line of best fit. The variance of the last few points
#seems larger than the points at the very beginning.
#Also, the first 10 points follow a pattern that approaches the line
#at 0. Was the data collected with the same test subject,
#who could learn or anticipate needing to stop?
plot(linear_model$fitted.values, linear_model.res,
     ylab = "Residuals", xlab = "Car Speeds")
abline(0,0, col="red")

library(lmtest)
bptest(linear_model)
#look at these results to test homoskedasicty
#but don't trust them fully.
#according to the BP test, which has a p-value of 0.00014,
#the data doesn't have equal variance.

library(MASS)
st.resids <- stdres(linear_model)
#these are the standardized residuals.

cooks.distance(linear_model)
#what is the cooks distance, again?

which(cooks.distance(linear_model)>4/62)
#there are four data points that are too far,
#according to the cooks distance.
ks.test(st.resids, "pnorm")
#the test shows that the residuals follow a normal distribution.
hist(st.resids)
plot(density(st.resids, adjust = 2),
xlab="Differences from Data to the Average",
     ylab="Probability", main="A Normal Distribution")
#the histogram of the standardized residuals support
#this conclusion.



#try a log transformation on both variables
scatter.smooth(log(speed), log(distance), col="blue",
xlab="Car's Speed (mph)",
               ylab="Car's Stopping Distance (feet)",
```

```
          main="Car's Stopping Distance at Different Speeds")
linear_model_log <- lm(log(Distance) ~ log(Speed), data=stop.dat)
summary(linear_model_log)

log_linear_model.res <- resid(linear_model_log)
#residuals vs. fitted values plot
#The plot doesn't fully support equal variance all along
#the line of best fit. The variance of the first few points
#seems larger than the points at the very end
#The independence seems to be better. No chunks
#of data are above or below.
plot(linear_model_log$fitted.values, log_linear_model.res,
     ylab = "Differences", xlab = "Car Speeds",
     main = "Differences from Average Line")
abline(0,0, col="red")

library(lmtest)
bptest(linear_model_log)
#look at these results to test homoskedasicty
#but don't trust them fully.
#according to the BP test, which has a p-value of 0.0006,
#the data doesn't have equal variance.

library(MASS)
st.resids <- stdres(linear_model_log)
#these are the standardized residuals.

cooks.distance(linear_model_log)
#what is the cooks distance, again?

which(cooks.distance(linear_model_log)>4/62)
#there are five data points that are too far,
#according to the cooks distance.
ks.test(st.resids, "pnorm")
#the test shows that the residuals follow a normal distribution.
hist(st.resids)
#the histogram of the standardized residuals support
#this conclusion.

###the log transformation seems to improve our requirements.
###independence is now valid. Normality of residuals is still valid.
###Linearity seems improved.
###Equal variance still does qualify.

plot(log(speed), log(distance), col="blue", xlab="Car's Speed (mph)",
```

```
ylab="Car's Stopping Distance (feet)",
main="Car's Stopping Distance at Different Speeds")
abline(reg = linear_model_log, col="red")
legend(legend = "average line")

ticks <- seq(0, 40, length=100)
log.pred <- exp(predict.lm(linear_model_log, newdata =
data.frame(Speed= ticks, Distance=1)))
plot(speed, distance, col="blue", xlab="Car's Speed (mph)",
     ylab="Car's Stopping Distance (feet)",
     main="Car's Stopping Distance at Different Speeds")
lines(ticks, log.pred, col="red")

n.cv <- 200
bias <- rep(NA, n.cv)
rpmse <- rep(NA, n.cv)
#create two NULL vectors

for(i in 1:n.cv) {
  ## Step 1: split data into test and training sets
  adv.test <- sample(1:nrow(stop.dat), 8)
  test.data <- stop.dat[adv.test,]
  train.data <- stop.dat[-adv.test,]

  ## Step 2: Fit model to training data
  my.model <- lm(log(Distance) ~ log(Speed), data = train.data)
  #if I am using a predict.lm statement, I need to fit the model
  #using the column names of the data frame, not variables
  #I created before.
  ## Step 3: predict for test data
  test.preds <- exp(predict.lm(my.model, newdata = test.data))
  ## Step 4: calculate the bias and RPMSE
  bias[i] <- mean((test.preds - test.data$Distance))
  rpmse[i] <- sqrt(mean((test.preds - test.data$Distance)^2))
}

#create a histogram and find the mean for bias and rpmse
#rpmse = on average, my data is off of the mean regression line by --
#bias = my predictions tend to over predict (greater than 0)
#or under predict.
mean(bias)
hist(bias)

mean(rpmse) #on average, my data is off of the mean
regression line by 25.39327
```

```
hist(rpmse)

#predict the stopping distance of a car going 35 mph
exp(predict.lm(linear_model_log, newdata = data.frame(Speed= 35, Distance=1))

exp(predict.lm(linear_model_log, newdata = data.frame(Speed= 30, Distance=1))

#the statistical model we are using to determine the relationship between a
car's speed and stopping
#distance predicts that a car will need 87.6 feet, on average,
#to stop fully if it is going 35 mph at the beginning.

#If your agency reduced the speed limit to
30 mph, and a car was truly going 30 mph,
#they would need 68.8 feet, on average, to fully stop.
This would be 18.8 feet distance,
#which would account for a 21.5 percent reduction
from a speed limit of 35 mph.

#Since this rural road passes through a neighborhood with many homes,
#I suggest lowering the speed limit to 30 mph. This neighborhood will likely
#have lots of children who may cross the road spontaneously.
#Moreover, the rural neighborhood may attract deer, especially
#if the neighborhood is located near mountains. An 18 foot difference in
#stopping distances is significant enough to protect children,
#pedestrians, and drivers if a car needs to stop as quick as
#possible. Please set the road's speed limit as 30 mph.
```