DATA 609 Assignment 2

Dan Smilowitz

February 9, 2017

Section 2.1, Problem 12

Problem: How can the company reduce its maintenance costs as trucks' ages and mileages increase?

Variables: The variables that would likely affect the solution to the problem above are listed below, with some variables held constant initially and others excluded entirely:

- Age of company's trucks
- Mileage of company's trucks
- Frequency of truck usage
- Number of company trucks *(constant)*
- Traffic conditions (city/highway) of trucks *(neglect)*
- Average route distance *(constant)*
- Types of maintenance service *(constant)*
- Type of truck *(constant)*

Submodels: Before developing a full model, I would want to study the validity of the following submodels:

- Age $\propto$ cost
- Mileage $\propto$ cost
- Age $\propto$ frequency
- Age $\propto$ mileage

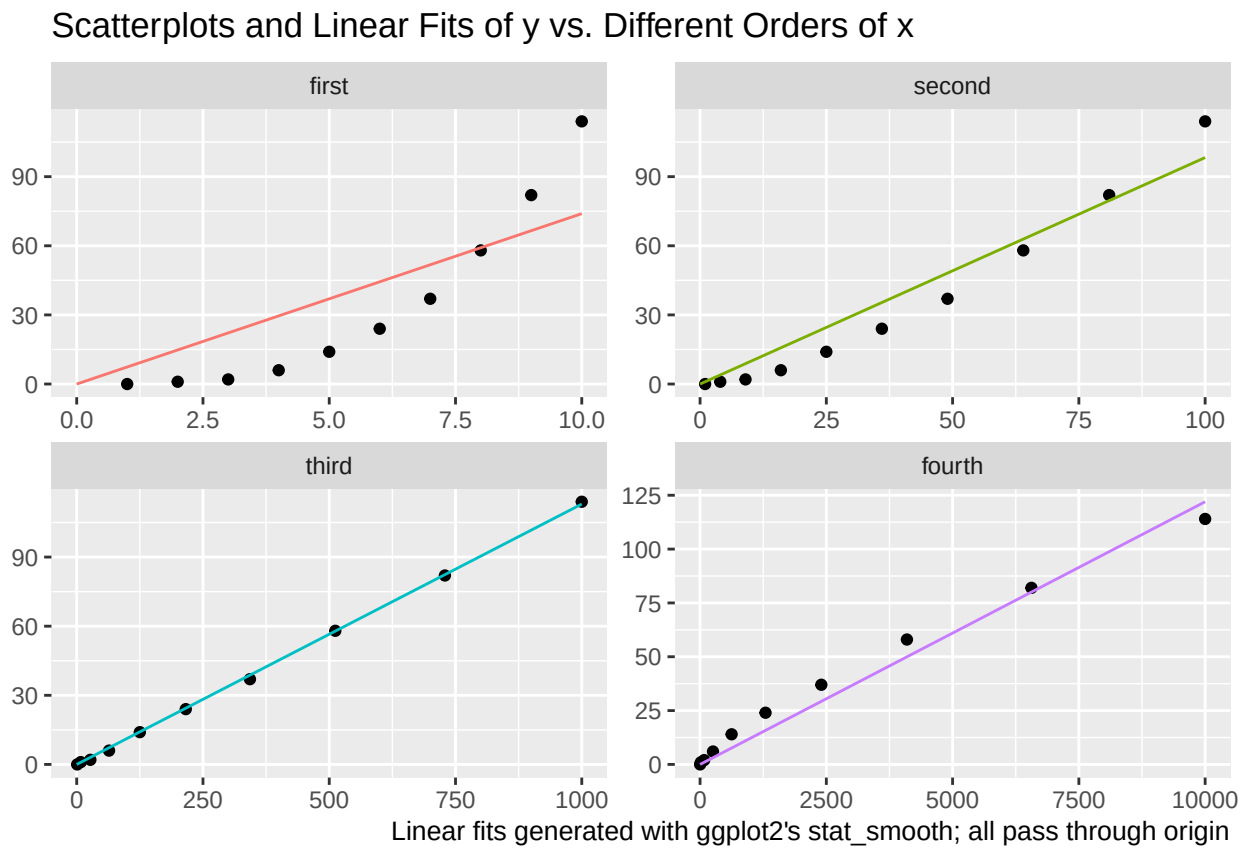
Data Collected: In order to develop the model, the following data would need to be collected:

- Vehicle inventory
  - Age
  - Mileage
- Dispatch log
  - Truck used
  - Route distance
- Maintenance log
  - Type of service
  - Cost of service

Section 2.2, Problem 11

```
# read in data
y_prop <- c(0, 1, 2, 6, 14, 24, 37, 58, 82, 114)
x_prop <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)

library(tidyverse)
# create data frame of y & x, x^2, x^3, x^4
df_prop <- data.frame(y_prop, first = x_prop, second = x_prop^2,
                      third = x_prop^3, fourth = x_prop^4) %>%
  gather(order, value, -y_prop)
df_prop$order <- factor(df_prop$order, levels = c('first', 'second', 'third', 'fourth'))
# draw scatterplots & linear fits
ggplot(df_prop, aes(y = y_prop, x = value)) + geom_point() +
  stat_smooth(method = 'lm', formula = y ~ 0 + x,
              se = FALSE, lwd = 0.5, fullrange = TRUE,
              aes(col = order), show.legend = FALSE) +
  facet_wrap(~order, scales = 'free') + expand_limits(x = 0) +
  scale_x_continuous(NULL) + scale_y_continuous(NULL) +
  labs(title = 'Scatterplots and Linear Fits of y vs. Different Orders of x',
       caption = "Linear fits generated with ggplot2's stat_smooth; all pass through origin")
```



Based on the scatterplots and linear fits presented above, the data set appears to follow the proportionality $y \propto x^3$ – the plot of y vs the third-order term x^3 follows a linear relationship quite well, especially as compared to the first, second, and fourth order terms.

Section 2.3, Project 4

It is reasonable to assume that the board feet of lumber L that a tree yields is proportional to its volume:

$$L \propto V$$

The volume of a right-circular cylinder is given by the equation $V = \pi r^2 h$, where r and h represent the radius and height, respectively.

Same Height

If the trees are the same height, then h is a constant. Since $L \propto V$ and $r^2 \propto d^2$, it can be said that

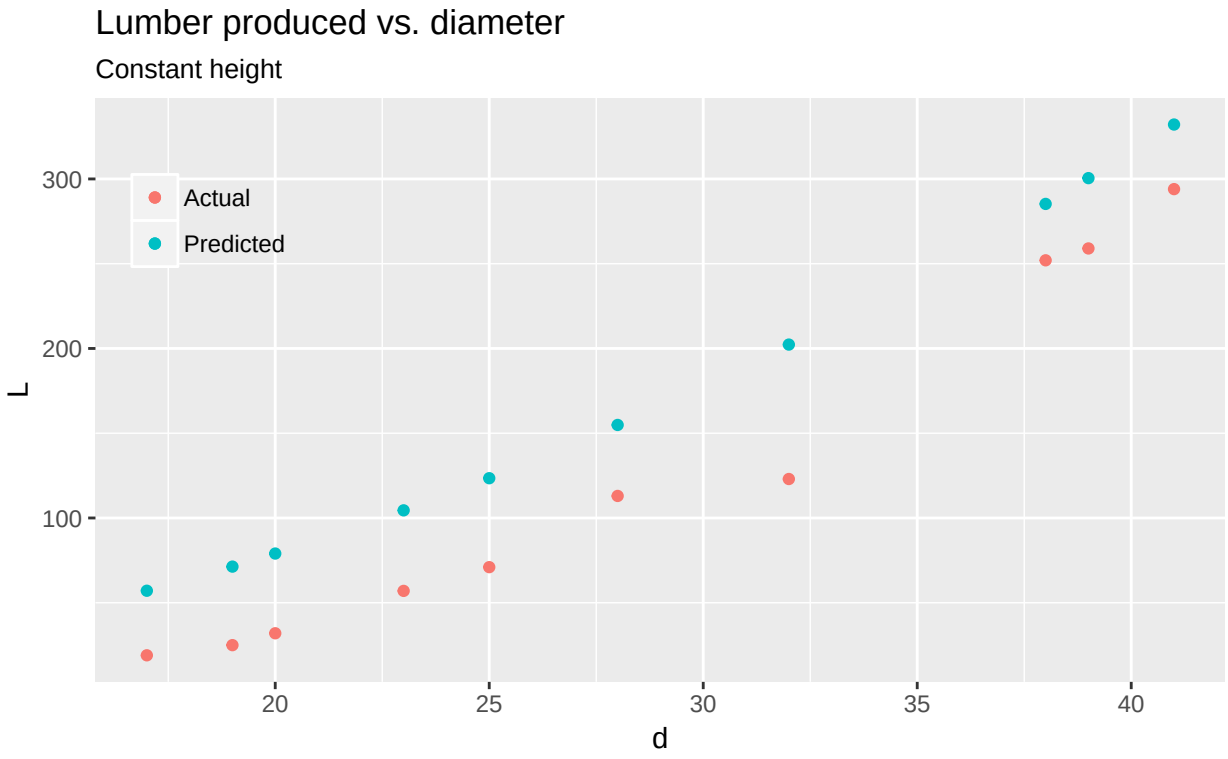
$$L = k_1 d^2$$

To calculate the value of k_1, the slope between L and d^2 is calculated using the first and last points:

$$k_1 = \frac{294 - 19}{41^2 - 17^2} = 0.1976$$

```
d <- c(17, 19, 20, 23, 25, 28, 32, 38, 39, 41)
L <- c(19, 25, 32, 57, 71, 113, 123, 252, 259, 294)
k1 <- (294 - 19)/(41^2 - 17^2)

df_tree <- data.frame(L, d) %>% mutate(same = d^2 * k1)
ggplot(df_tree, aes(x = d)) +
  geom_point(aes(y = L, col = 'Actual')) +
  geom_point(aes(y = same, col = 'Predicted')) +
  scale_color_discrete(NULL) +
  theme(legend.position = c(0.1, 0.8), legend.background = element_blank()) +
  labs(title = 'Lumber produced vs. diameter',
       subtitle = 'Constant height')
```



Varying Height

If the trees' heights are proportional to their diameters, then $h \propto d$. Since $L \propto V$ and $r^2 \propto d^2$, it can be said that

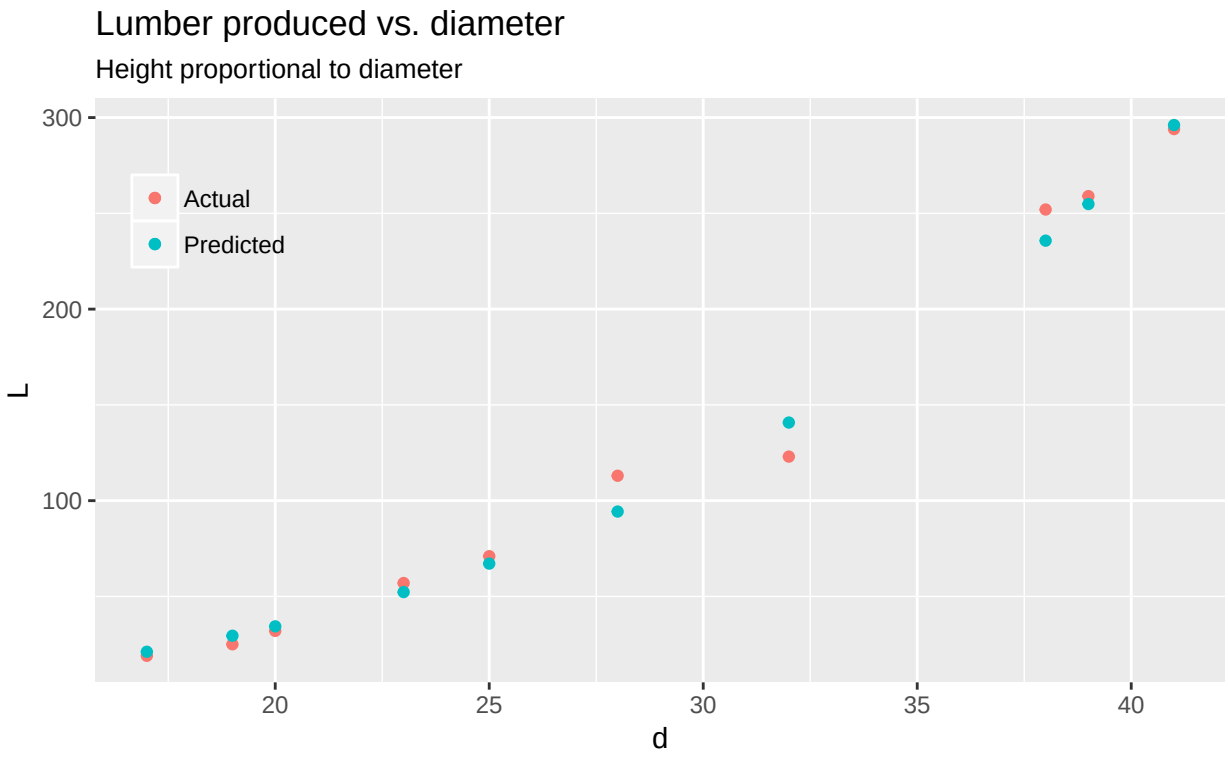
$$L = k_2 d^3$$

k_2 is calculated similarly as above:

$$k_2 = \frac{294 - 19}{41^3 - 17^3} = 0.0043$$

```
k2 <- (294 - 19)/(41^3 - 17^3)

df_tree <- df_tree %>% mutate(vary = d^3 * k2)
ggplot(df_tree, aes(x = d)) +
  geom_point(aes(y = L, col = 'Actual')) +
  geom_point(aes(y = vary, col = 'Predicted')) +
  scale_color_discrete(NULL) +
  theme(legend.position = c(0.1, 0.8), legend.background = element_blank()) +
  labs(title = 'Lumber produced vs. diameter',
       subtitle = 'Height proportional to diameter')
```



Model Comparison

The second model, in which height is assumed to be proportional to diameter, appears to be a better model. The predicted values appear to match the actual values far better than those for the first model. Intuitively, this makes sense — trees have varying heights, and it seems reasonable that trees will grow up (in height) as they grow out (in diameter).

Section 2.4, Problem 3

The analysis mentioned wind and air densities, but there are a number of other weather factors that may affect driving habits or drag forces, which may have an effect on fuel mileage. Additionally, the topography and amount of turning required may also affect the mileage – going uphill or making more turns will decrease gas mileage. The number of passengers in a vehicle will change the total weight that has to be transported, which will increase the friction that must be overcome to move the car – more passengers will decrease gas mileage. Finally, many modern cars have different driving modes, such as "Sport" or "Eco" which change the way the car's transmission and fuel injection behave – modification of these settings would likely affect gas mileage.

Section 2.5, Problem 2

For people whose accurately-measured body fat percentage is known, the collection of additional body measurements often associated with attractiveness should be collected: height; weight; and the circumference & length of the waist, chest, hips, legs, and arms. This would allow for a more accurate calculation of each person's volume V. Using the calculated body fat percentage, the volume of the inner and outer layers V_in and V_{out} could be estimated – these could then be used to check the validity of the assumption of constant outer & inner core densities