DATA 609: Homework 2

The Modeling Process, Proportionality, and Geometric Similarity

Aaron Grzasko

August 31, 2017

```r
library(dplyr)
library(tidyr)
library(knitr)
library(ggplot2)
library(latex2exp)
```

Page 69: #12

Scenario: A company with a fleet of trucks faces increasing maintenance costs as the age and mileage of the trucks increase.

Identify a problem worth studying:

Management must decide if it is in the company's best financial interest to A) continue operating with its aging fleet or B) sell the current fleet and purchase (or lease) newer vehicles.

The company could use a net present value, financial model to determine the least costly option:

$$NPV = PV\ Revenues - PV\ Costs$$

List the variables that affect the behavior you have identified

- interest rate and terms for new vehicle financing
- credit worthiness of company
- expected future maintenance costs for current and new fleets
- expected timing of future maintenance costs for the current and new fleets.
- inflation rate impacting both future maintenance costs and new car selling prices.
- sales price for new vehicles.
- likely selling price/salvage value for current fleet
- sales taxes
- depreciation impacting value of current fleet
- insurance costs associated with both new and old fleets
- discount rate for calculating net present values for new vs. current fleets
- fuel efficiencies and other quality improvements associated with newer fleet
- year, make, and model of vehicles

Which variables would be neglected completely?

- Rather than try to model various cost components, the company may attempt to model a generic category, "costs", for both scenarios under consideration.

- It may not be feasible (given data constraints) to model costs by vehicle make and model. This simplification is justified if costs by make and model are relatively homogeneous.

- For simplicity, all costs and revenues may be assumed to occur at each year's midpoint.

Which might be considered as constants initially?

- Inflation and loan interest rates may be treated as constants in a deterministic system; or varied within a reasonable range as part of a larger scenario analysis
- Similarly, the discount rate used in present value calculations may be treated as a constant.

Can you identify any submodels you would want to study in detail?

- The modeling of costs by age and mileage of fleet vehicles will likely require a separate submodel.

- It's also possible to create separate submodels for inflation and interest rates; however, the analysis in question probably does not require this level of sophistication.

Identify any data you would want collected.

- historical cost and maintenance records for the company's fleet
- survey of current loan rates and terms from multiple vendors

- If more rigorous modeling is desirable, CPI data from BLS and yield curve information could be collected.

Page 79: #11

Determine whether the data set supports the stated proportionality model:

$$y \propto x^3$$

```r
y <- c(0, 1, 2, 6, 14, 24, 37, 58, 82, 114)
x <- seq(1, 10)

mydf <- data.frame(y = y, x = x)
```

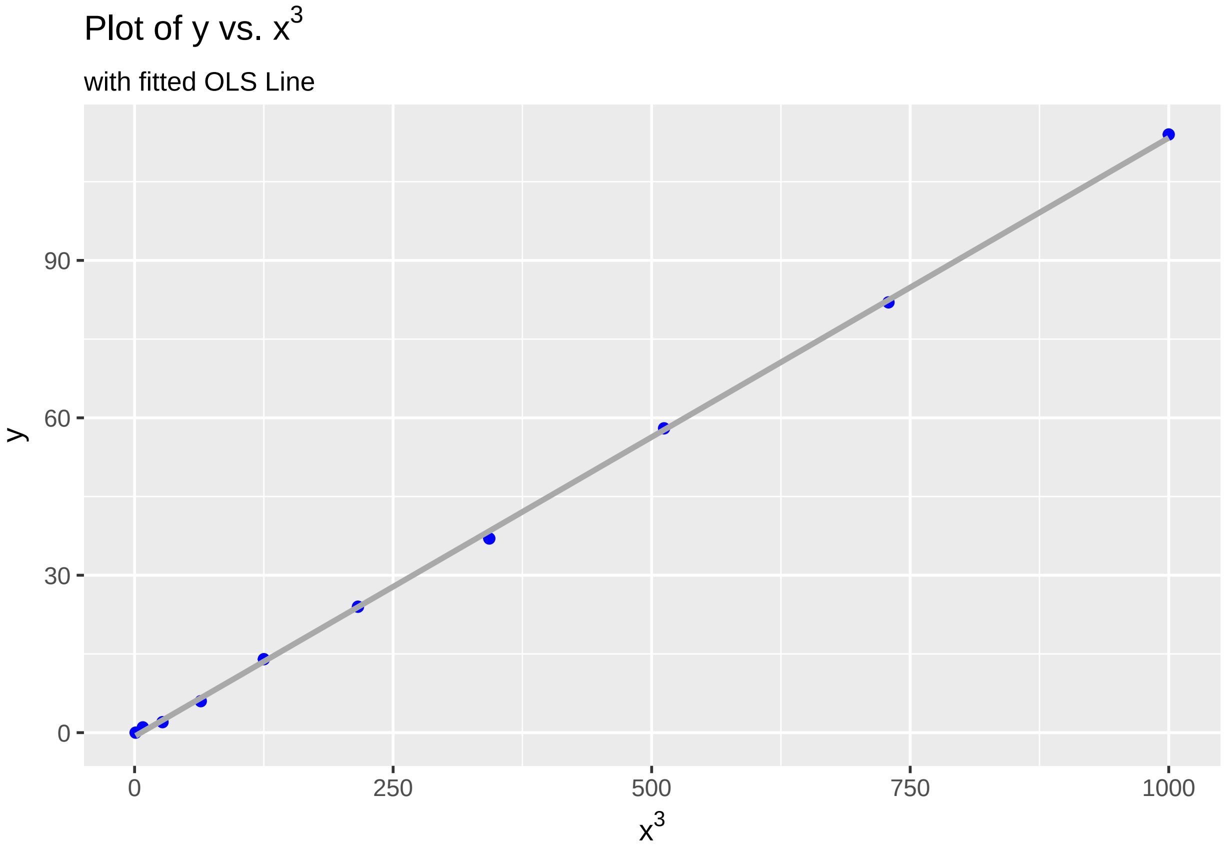
y	0	1	2	6	14	24	37	58	82	114
x	1	2	3	4	5	6	7	8	9	10

We can determine whether the proportionality model is reasonable by plotting y values against $x^3$. If the model is a good fit, then a fitted straight line passing through the origin should provide a reasonable approximation.

```r
# add column in dataframe to represent x^3
mydf$x3 = x^3

# ols line with y intercept set to zero
mylm <- lm(y ~ x3, data = mydf)

# plot of y vs. x^3 with fitted OLS line
g <- ggplot(mydf, aes(x^3, y)) + geom_point(col = "blue")
g <- g + xlab(TeX("$x^3$")) + ggtitle(TeX("Plot of y vs. $x^3$"))
g <- g + labs(subtitle = "with fitted OLS Line")
g + geom_smooth(method = "lm", color = "darkgray", se = FALSE)
```



Let’s look at the model output for the fitted line:

```r
summary(mylm)
```

```

Call:
lm(formula = y ~ x3, data = mydf)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.4200 -0.4326  0.1844  0.5571  0.7949 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.7074611  0.3196464  -2.213   0.0578 .  
x3           0.1140743  0.0007186 158.736 2.78e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7411 on 8 degrees of freedom
Multiple R-squared:  0.9997,    Adjusted R-squared:  0.9996 
F-statistic: 2.52e+04 on 1 and 8 DF,  p-value: 2.775e-15
```

The proposed proportionality model appears to provide a good approximation:

- The R^2 statistic for our fitted, OLS model is 0.9997, which is indicative of a strong model fit.
- While the model’s estimate of the y-intercept is not zero, the relatively high p-value for our estimate suggests that we do not have compelling evidence to reject the null hypothesis of a zero value for the

intercept.

Page 94: #4

Lumber cutters wish to use readily available measurements to estimate the number of board feet of lumber in a tree. Assume they measure the diameter of the tree in inches at waist height. Develop a model that predicts board feet as a function of diameter in inches. Use the following data for your test:

```
x <- c(17, 19, 20, 23, 25, 28, 32, 38, 39, 41)  # diameter
y <- c(19, 25, 32, 57, 71, 113, 123, 252, 259, 294)  # board height
mydf <- data.frame(x = x, y = y)
```

x	17	19	20	23	25	28	32	38	39	41
y	19	25	32	57	71	113	123	252	259	294

The variable x is the diameter of a ponderosa pine in inches, and y is the number of board feet divided by 10.

A: Consider two separate assumptions, allowing each to lead to a model. Completely analyze each model.

(i): Assume that all trees are right-circular cylinders and are approximately the same height.

Tree volume is approximated by:

$$V = \pi r^2 h$$

Since we're invoking geometric similarity, and assuming height is constant, then Volume is proportional to any characteristic dimension, l, squared:

$$V \propto l^2$$

Tree diameter, d, is a characteristic dimension; so substitute this specific variable in our model as follows:

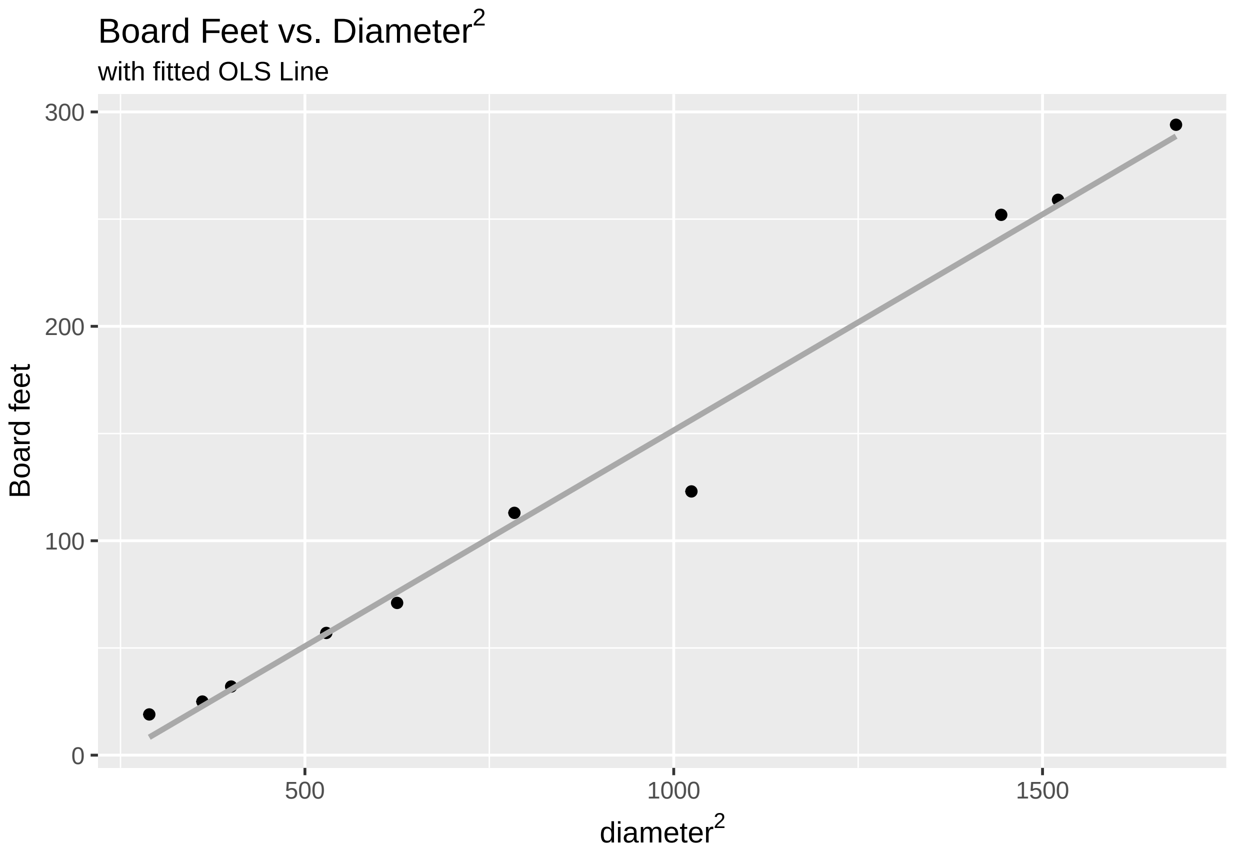
$$V \propto d^2$$

If we assume the number of board feet produced, B, is proportional to tree volume, then our final model becomes:

$$B \propto d^2$$

```
# fit linear model y = Bx^2 + A
mydf$x2 <- mydf$x^2
mylm <- lm(y ~ x2, data = mydf)

# plot diammeter squared model
g <- ggplot(mydf, aes(x2, y)) + geom_point()
g <- g + xlab(TeX("$diameter^2$")) + ylab("Board feet")
g <- g + ggtitle(TeX("Board Feet vs. $Diameter^2$"))
g <- g + labs(subtitle = "with fitted OLS Line")
g + geom_smooth(method = "lm", color = "darkgray", se = FALSE)
```



```r
summary(mylm)
```

```
Call:
lm(formula = y ~ x2, data = mydf)

Residuals:
    Min      1Q  Median      3Q     Max 
-33.358   0.568   2.357   5.247  11.064 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -49.851357   8.566936  -5.819 0.000396 ***
x2            0.201376   0.008596  23.426 1.17e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.42 on 8 degrees of freedom
Multiple R-squared:  0.9856,    Adjusted R-squared:  0.9838 
F-statistic: 548.8 on 1 and 8 DF,  p-value: 1.172e-08
```

The R^2 of the OLS model indicates a strong linear relationship between board feet and the square of tree diameter. However, the intercept of the OLS line is significantly different than zero–refer to the low p-value associated with the intercept estimate. Therefore, the proportionality model–with it's implied y intercept of zero–may not be the most appropriate model.

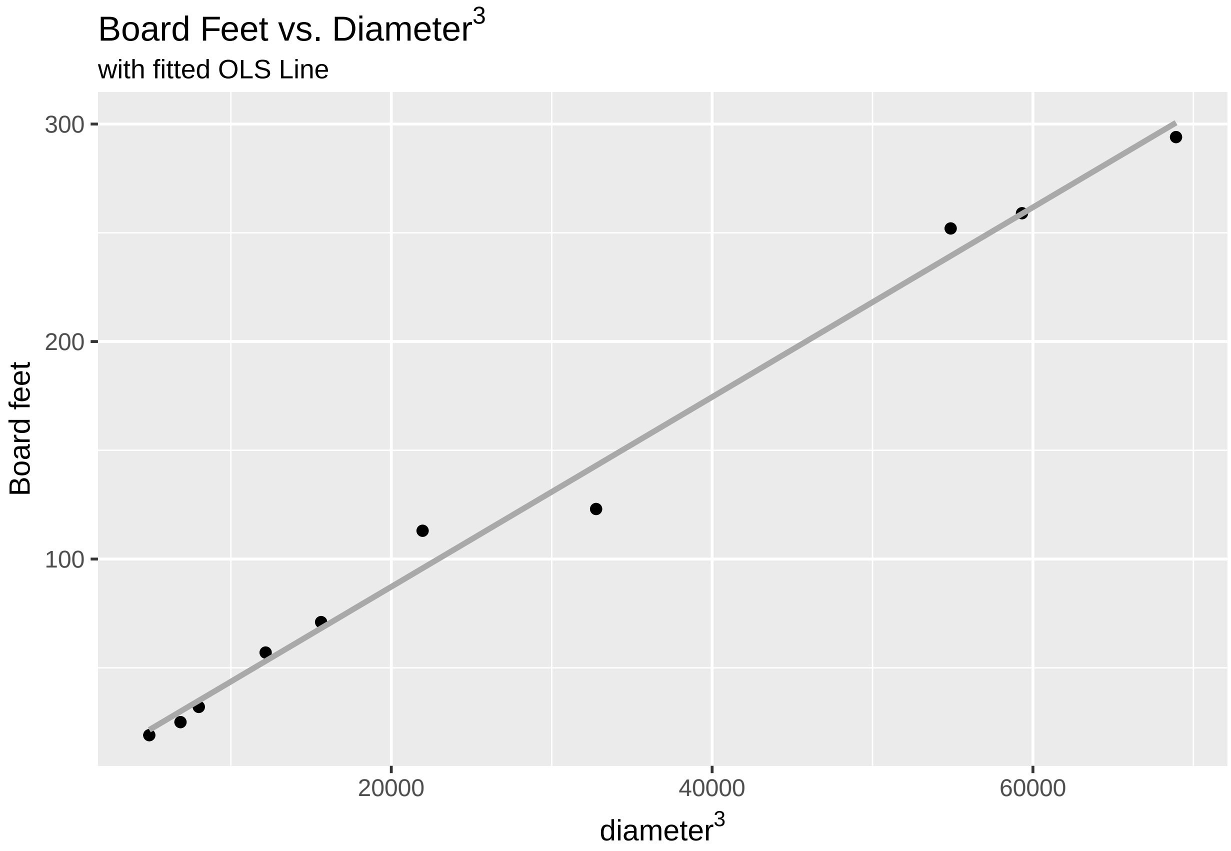
(ii): Assume that all trees are right-circular cylinders and that the height of the tree is proportional to the diameter.

With height proportional to diameter, our model becomes:

$$B \propto d^3$$

```r
# fit linear model y = Bx^3 + A
mydf$x3 <- mydf$x^3
mylm <- lm(y ~ x3, data = mydf)

# plot diammeter cubed model
g <- ggplot(mydf, aes(x3, y)) + geom_point()
g <- g + xlab(TeX("$diameter^3$")) + ylab("Board feet")
g <- g + ggtitle(TeX("Board Feet vs. $Diameter^3$"))
g <- g + labs(subtitle = "with fitted OLS Line")
g + geom_smooth(method = "lm", color = "darkgray", se = FALSE)
```



```r
# OLS model output
summary(mylm)
```

```

Call:
lm(formula = y ~ x3, data = mydf)

Residuals:
    Min      1Q  Median      3Q     Max 
-19.942  -4.434  -1.099   3.639  17.232 
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0237507  5.5460514   0.004    0.997
x3          0.0043615  0.0001517  28.748 2.32e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.96 on 8 degrees of freedom
Multiple R-squared:  0.9904,    Adjusted R-squared:  0.9892
F-statistic: 826.4 on 1 and 8 DF,  p-value: 2.319e-09
```

This model has a very attractive–albeit only slightly higher–R^2 value as compared to the previous model, which is generally associated with a strong model fit. In this model, the estimated intercept is not significantly different than zero–see the large p-value corresponding to the intercept parameter. Therefore, the proposed proportionality model based on diameter cubed is probably more appropriate than the squared diameter model.

(b): Which model appears to be better? Why? Justify your conclusions.

As discussed in the previous section, the model $B \propto d^3$ appears to be superior to the $B \propto d^2$ model. Our fitted OLS line, with its high R^2 and intercept parameter close to zero provide evidence that the proposed proportionality model is appropriate.

This latter model also employs assumptions that are more consistent with our casual observations. That is, individual trees exhibit wide variation in height, but tall trees tend to be associated associated with wider trunks.