Model Fitting I

Carolina A. de Lima Salge
Assistant Professor
Terry College of Business
University of Georgia

Business Intelligence
Spring 2021

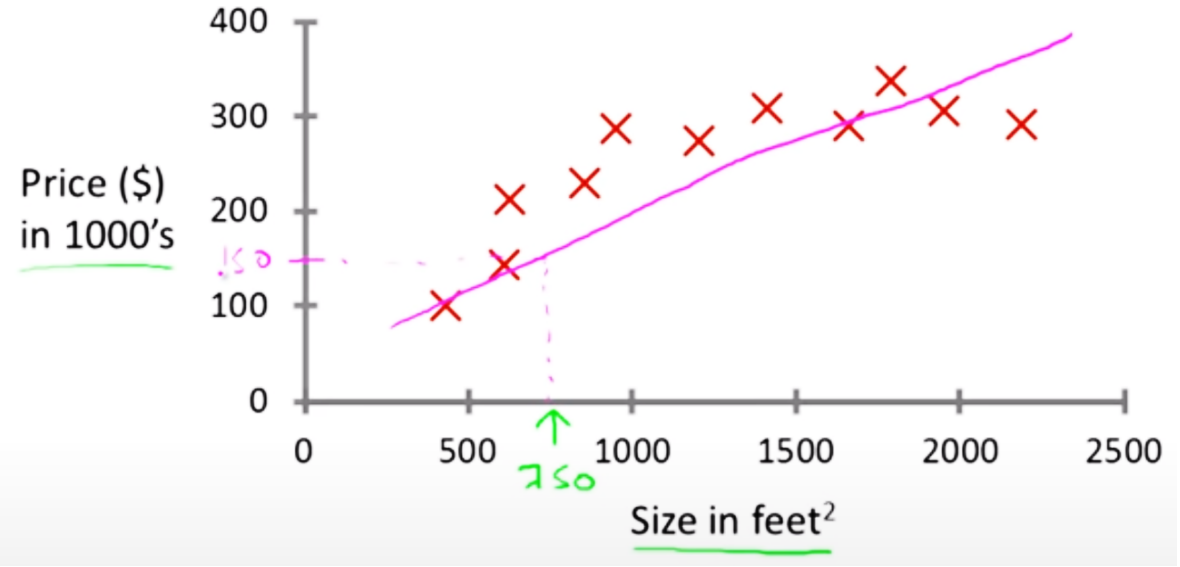
Terry College of Business
UNIVERSITY OF GEORGIA

```
Call:
lm(formula = Profit ~ Market * MarketSize + Market * COGS, data = CoffeeChain)

Residuals:
    Min      1Q  Median      3Q     Max
-771.87  -20.64    1.65   21.73  721.75

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                               -4.26570    4.79625  -0.889  0.37385
MarketEast                                30.68239    7.83224   3.917 9.09e-05 ***
MarketSouth                               17.89732   11.51494   1.554  0.12020
MarketWest                                48.99987    9.08179   5.395 7.21e-08 ***
MarketSizeSmall Market                   -14.88927    4.84672  -3.072  0.00214 **
COGS                                       0.96574    0.03719  25.964  < 2e-16 ***
MarketEast:MarketSizeSmall Market         -8.20585    8.25011  -0.995  0.31997
MarketSouth:MarketSizeSmall Market       -13.73362    9.74931  -1.409  0.15900
MarketWest:MarketSizeSmall Market        -21.48502    8.15131  -2.636  0.00843 **
MarketEast:COGS                           -0.39202    0.05883  -6.664 3.02e-11 ***
MarketSouth:COGS                          -0.10598    0.09318  -1.137  0.25544
MarketWest:COGS                           -0.55016    0.05187 -10.606  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 87.96 on 4236 degrees of freedom
Multiple R-squared:  0.254,     Adjusted R-squared:  0.2521
F-statistic: 131.1 on 11 and 4236 DF,  p-value: < 2.2e-16
```

Linear Regression

Suppose you want to predict house prices, and you have some data about the price of a house (in thousands of $) over size (sqft)

Estimate numeric value (e.g., with a linear regression)

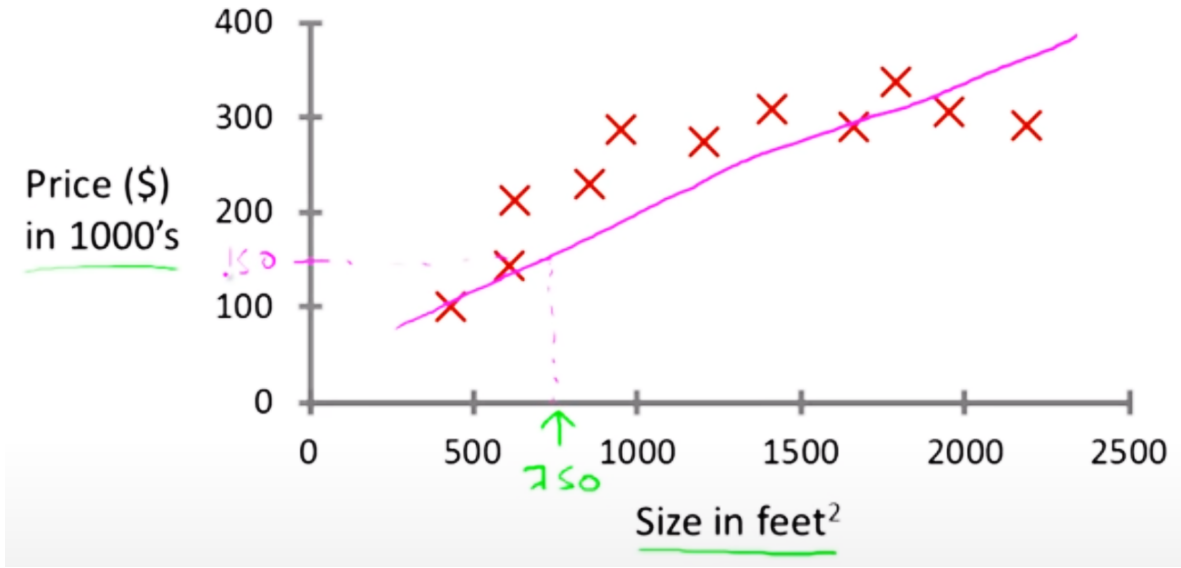


Linear Regression

The model function

$$y = \alpha + \beta x$$

y = target
α = y-intercept
β = slope
x = predictor

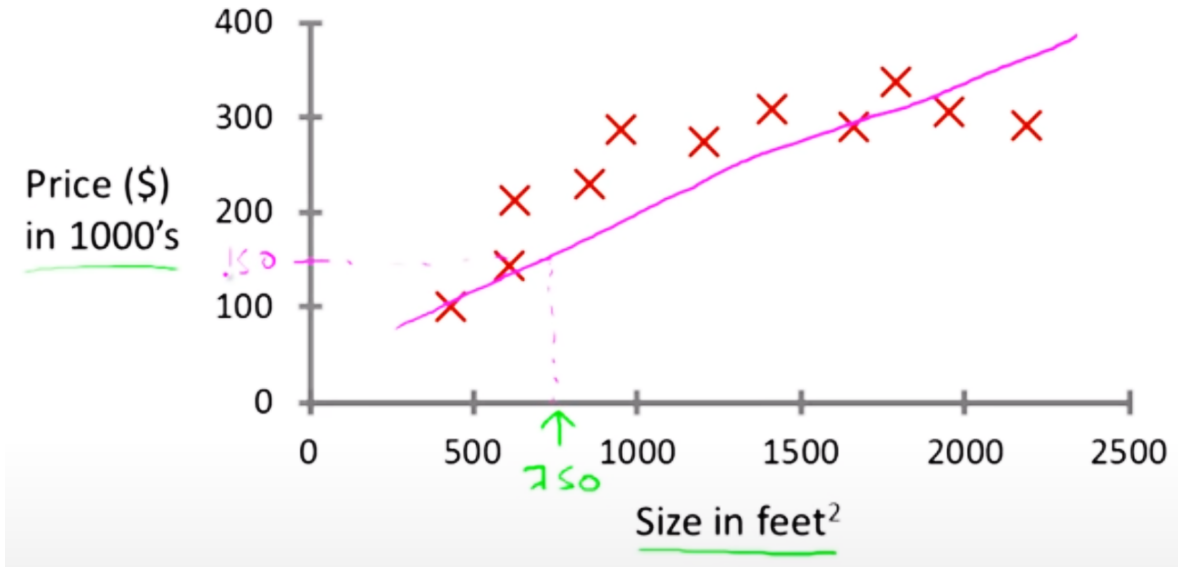


Linear Regression

Fitted line minimizes the sum or mean of the squares of the errors

$$y = \alpha + \beta x$$

y = target
α = y-intercept
β = slope
x = predictor



Also known as ordinary least squares (OLS) regression – very popular!

Traditional Use

Explanatory modeling

- The goal is to explain the relationship between predictors (independent variables) and target (dependent variable)

Model Evaluation

Fit the data well and understand the amount of variance explained as well as the statistical significance of each predictor

- Evaluation (of goodness of fit) involves the use of (Adjusted) R-squared and p-values

Significance

What does it mean for a predictor to be statistically significant?

- In practice, a p-value < 0.05

A measure of the probability that the observed effect was due to random chance. A p < 0.05 means we are 95% confident the result is not a mistake (i.e., not driven by randomness)

Significance

What does it mean for a predictor to be statistically significant?

- Can also look at confidence intervals

You can claim statistical significance (i.e., reject the null hypothesis) when the CI does not include zero

Example

Predict wage from education

Years of Education	Wage
16	52,000
18	65,000
16	45,000
21	80,000
14	40,000
12	50,000
…	…

Example

Values of the predictor

Years of Education	Wage
16	52,000
18	65,000
16	45,000
21	80,000
14	40,000
12	50,000
…	…

Example

Values of the target

Years of Education	Wage
16	52,000
18	65,000
16	45,000
21	80,000
14	40,000
12	50,000
…	…

Example

Find values of α and β that best fit y – try and get my predicted value of y as close as possible to the actual value of y

$$y = \alpha + \beta x$$

y = wage
α = wage-intercept
β = slope
x = years of education

Years of Education	Wage
16	52,000
18	65,000
16	45,000
21	80,000
14	40,000
12	50,000
…	…

Interpretation

Find values of α and β that best fit y – try and get my predicted value of y as close as possible to the actual value of y

$$y = \alpha + \beta x$$

y = wage

α = wage-intercept indicates the value of the target (wage) when all predictors are zero

β = slope indicates how much the target (wage) changes when the predictor (years of education) changes

x = years of education

R^2 = percentage of variance in the target explained by the predictors, ranges from 0 to 1

Traditional OLS Regression in R

Use the the `lm()` function

CoffeeChain dataset – recall, *the goal is to understand the relationship between predictors and target*

```
library(tidyverse)
library(readxl)

CoffeeChain <- read_excel("CoffeeChain.xlsx")
```

Questions

Does the coffee chain tend to make more money in small or large markets?

What region is the most profitable?

In what region is there the biggest difference between large and small markets?

In which regions does COGS lead to the most profit?

Code and Model

```
m1 <- lm(Profit ~ Market * MarketSize +
Market * COGS, data = CoffeeChain)


summary(m1)
```

```
Call:
lm(formula = Profit ~ Market * MarketSize + Market * COGS, data = CoffeeChain)

Residuals:
    Min      1Q  Median      3Q     Max
-771.87  -20.64    1.65   21.73  721.75

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       -4.26570    4.79625  -0.889  0.37385
MarketEast                        30.68239    7.83224   3.917 9.09e-05 ***
MarketSouth                       17.89732   11.51494   1.554  0.12020
MarketWest                        48.99987    9.08179   5.395 7.21e-08 ***
MarketSizeSmall Market           -14.88927    4.84672  -3.072  0.00214 **
COGS                               0.96574    0.03719  25.964  < 2e-16 ***
MarketEast:MarketSizeSmall Market  -8.20585    8.25011  -0.995  0.31997
MarketSouth:MarketSizeSmall Market -13.73362    9.74931  -1.409  0.15900
MarketWest:MarketSizeSmall Market  -21.48502    8.15131  -2.636  0.00843 **
MarketEast:COGS                    -0.39202    0.05883  -6.664 3.02e-11 ***
MarketSouth:COGS                   -0.10598    0.09318  -1.137  0.25544
MarketWest:COGS                    -0.55016    0.05187 -10.606  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 87.96 on 4236 degrees of freedom
Multiple R-squared:  0.254,     Adjusted R-squared:  0.2521
F-statistic: 131.1 on 11 and 4236 DF,  p-value: < 2.2e-16
```

Questions

Does the coffee chain tend to make more money in small or large markets?

- **Larger Markets**

The baseline market is the Larger Market, which when compared to the Small Market, has a higher profit (negative coefficient for MarketSizeSmall Market), while controlling for the effect of other variables (e.g., COGS). The difference is statistically significant (**)

```
Call:
lm(formula = Profit ~ Market * MarketSize + Market * COGS, data = CoffeeChain)

Residuals:
    Min      1Q  Median      3Q     Max
-771.87  -20.64    1.65   21.73  721.75

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       -4.26570    4.79625  -0.889  0.37385
MarketEast                        30.68239    7.83224   3.917 9.09e-05 ***
MarketSouth                       17.89732   11.51494   1.554  0.12020
MarketWest                        48.99987    9.08179   5.395 7.21e-08 ***
MarketSizeSmall Market           -14.88927    4.84672  -3.072  0.00214 **
COGS                               0.96574    0.03719  25.964  < 2e-16 ***
MarketEast:MarketSizeSmall Market  -8.20585    8.25011  -0.995  0.31997
MarketSouth:MarketSizeSmall Market -13.73362    9.74931  -1.409  0.15900
MarketWest:MarketSizeSmall Market  -21.48502    8.15131  -2.636  0.00843 **
MarketEast:COGS                    -0.39202    0.05883  -6.664 3.02e-11 ***
MarketSouth:COGS                   -0.10598    0.09318  -1.137  0.25544
MarketWest:COGS                    -0.55016    0.05187 -10.606  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 87.96 on 4236 degrees of freedom
Multiple R-squared:  0.254,     Adjusted R-squared:  0.2521
F-statistic: 131.1 on 11 and 4236 DF,  p-value: < 2.2e-16
```

Questions

What region is the most profitable?

- **South**

```
library(emmeans)

emmeans(m1, ~ Market)
```

Market	emmean	SE	df	lower.CL	upper.CL
Central	69.8	2.40	4236	65.1	74.5
East	63.3	3.06	4236	57.3	69.3
South	71.9	4.01	4236	64.1	79.8
West	61.6	3.15	4236	55.5	67.8

Results are averaged over the levels of: MarketSize
Confidence level used: 0.95

The **emmeans** function calculates the mean profit for each region, while controlling for the effect of other variables included in the model

Questions

In what region is there the biggest difference between large and small markets?

- **West**

```
emmeans(m1, ~ MarketSize | Market)
```

By adding MarketSize, we now calculate the **mean profit for each region by market**, while controlling for the effect of other variables included in the model

```
Market = Central:
 MarketSize   emmean   SE   df lower.CL upper.CL
 Major Market   77.3 3.35 4236     70.7     83.8
 Small Market   62.4 3.47 4236     55.6     69.2

Market = East:
 MarketSize   emmean   SE   df lower.CL upper.CL
 Major Market   74.9 3.90 4236     67.2     82.5
 Small Market   51.8 5.08 4236     41.8     61.7

Market = South:
 MarketSize   emmean   SE   df lower.CL upper.CL
 Major Market   86.2 6.83 4236     72.8     99.6
 Small Market   57.6 4.61 4236     48.6     66.6

Market = West:
 MarketSize   emmean   SE   df lower.CL upper.CL
 Major Market   79.8 5.82 4236     68.4     91.2
 Small Market   43.4 2.72 4236     38.1     48.8

Confidence level used: 0.95
```

Questions

In which regions does COGS lead to the most profit?

- **South**

```
emmeans(m1, ~ COGS | Market)
```

```
Market = Central:
 COGS emmean   SE   df lower.CL upper.CL
 84.4   69.8 2.40 4236     65.1     74.5

Market = East:
 COGS emmean   SE   df lower.CL upper.CL
 84.4   63.3 3.06 4236     57.3     69.3

Market = South:
 COGS emmean   SE   df lower.CL upper.CL
 84.4   71.9 4.01 4236     64.1     79.8

Market = West:
 COGS emmean   SE   df lower.CL upper.CL
 84.4   61.6 3.15 4236     55.5     67.8

Results are averaged over the levels of: MarketSize
Confidence level used: 0.95
```

Other Observations

The model is significant and explains about 25.4% of variance in profit

Relationship between COGS and profit is positive and statistically significant

```
Call:
lm(formula = Profit ~ Market * MarketSize + Market * COGS, data = CoffeeChain)

Residuals:
    Min      1Q  Median      3Q     Max
-771.87  -20.64    1.65   21.73  721.75

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       -4.26570    4.79625  -0.889  0.37385
MarketEast                        30.68239    7.83224   3.917 9.09e-05 ***
MarketSouth                       17.89732   11.51494   1.554  0.12020
MarketWest                        48.99987    9.08179   5.395 7.21e-08 ***
MarketSizeSmall Market           -14.88927    4.84672  -3.072  0.00214 **
COGS                               0.96574    0.03719  25.964  < 2e-16 ***
MarketEast:MarketSizeSmall Market  -8.20585    8.25011  -0.995  0.31997
MarketSouth:MarketSizeSmall Market -13.73362    9.74931  -1.409  0.15900
MarketWest:MarketSizeSmall Market  -21.48502    8.15131  -2.636  0.00843 **
MarketEast:COGS                    -0.39202    0.05883  -6.664 3.02e-11 ***
MarketSouth:COGS                   -0.10598    0.09318  -1.137  0.25544
MarketWest:COGS                    -0.55016    0.05187 -10.606  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 87.96 on 4236 degrees of freedom
Multiple R-squared:  0.254,     Adjusted R-squared:  0.2521
F-statistic: 131.1 on 11 and 4236 DF,  p-value: < 2.2e-16
```

At-Home Exercises

Think about other variables in the dataset that, if added to the OLS regression model, could help increase the model's $\boldsymbol{R^2}$

Next, add such variables to the model and check how the results change. As you add these variables, pay attention to $\boldsymbol{R^2}$ and adjusted $\boldsymbol{R^2}$. Are they like one another or not?

Based on your modeling results, what recommendations do you have for the coffee chain?

Thank You!