Introduction to Simple Linear Regression

.

Zhaohu (Jonathan) Fan

Department of Operations, Business Analytics and Information Systems

Carl H. Lindner College of Business

University of Cincinnati

Februrary 3 , 2023

Previous experiences with regression

- I have heard of regression before
- I have used regression before in a class or at work
- I know what adjusted R^2 means

Main topics

- Simple linear regression (SLR)
- Least squares (LS) estimation
- Fitting SLR models in R with the `lm()` function

Regression: What is it?

- Simply: The most widely used statistical tool for understanding relationships among variables
- The relationship is expressed in the form of an equation or a model connecting the outcome to the factors

Regression in business

- Optimal portfolio choice:
  - Predict the future joint distribution of asset returns
  - Construct an optimal portfolio (choose weights)
- Determining price and marketing strategy:
  - Estimate the effect of price and advertisement on sales
  - Decide what is optimal price and ad campaign

Regression in everything

- Straight prediction questions:
  - What price should I charge for my car?
  - What will the interest rates be next month?
- Explanation and understanding:
  - Does your income increase if you get an Master Degree?
  - Is my advertising campaign working?

Example: Predicting House Prices

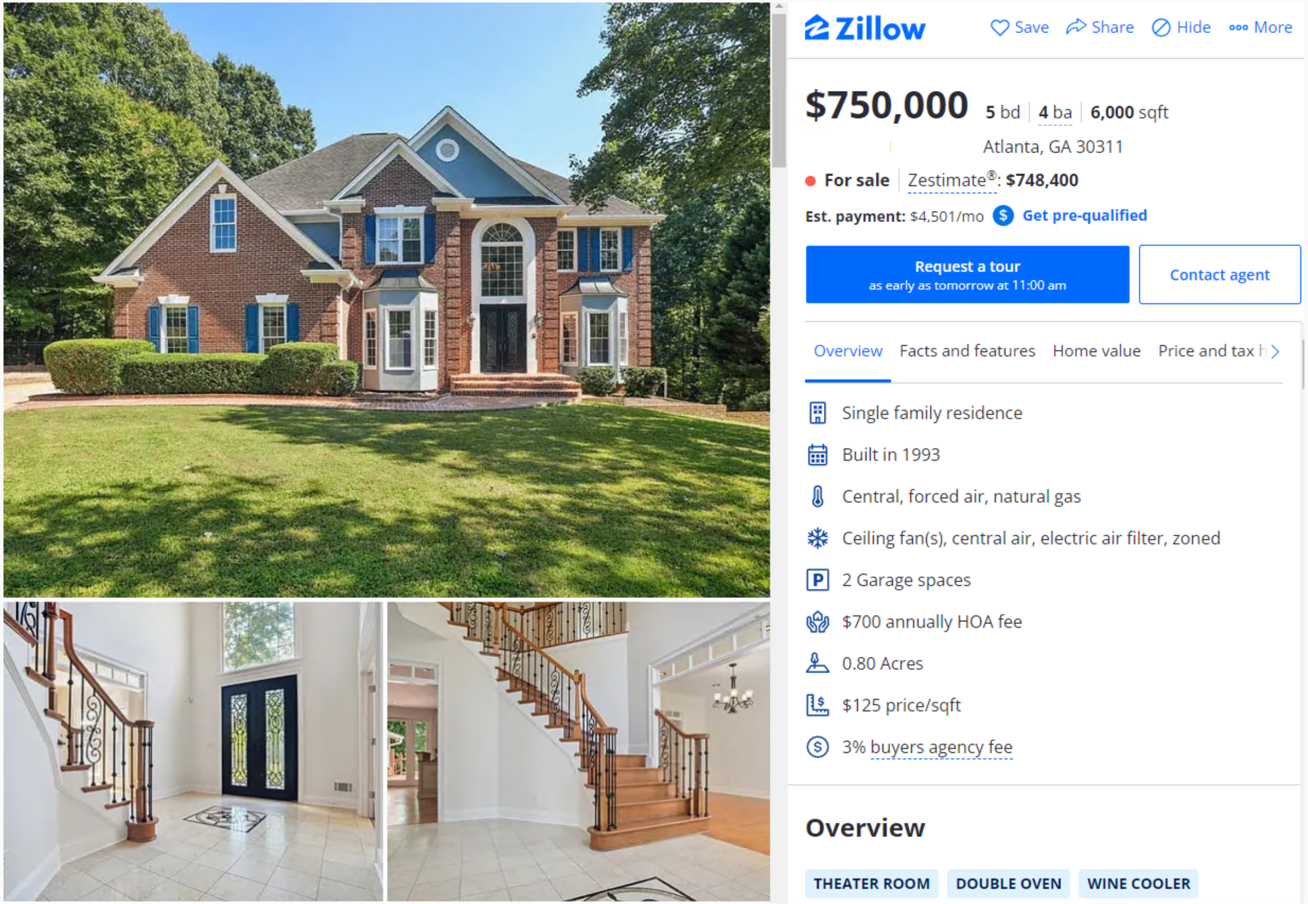


Image of Atlanta, GA by Zillow

Example: Predicting House Prices

Problem:

- Predict market price based on observed characteristics

Solution:

- Look at property sales data where we know the price and some observed characteristics.
- Build a decision rule that predicts price as a function of the observed characteristics.

Action:

- We have to define the variables of interest and develop a specific quantitative measure of these variables

What characteristics should we use?

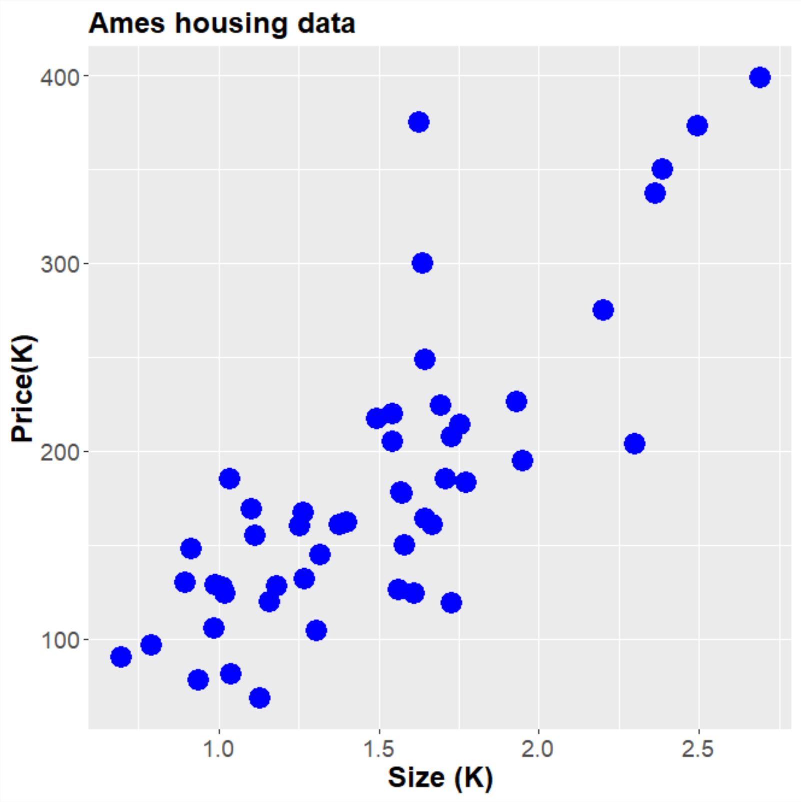
- Many factors or variables affect the price of a house
  - size of house
  - number of baths
  - garage
  - size of land
  - location, etc.
- Easy to quantify price and size but what about other variables such as location, aesthetics, workmanship, etc?

Simple linear regression (SLR)

- To keep things super simple, let's focus only on size of the house.
- The variable that we use to guide prediction is the explanatory (or input) variable, and this is labelled
    - X=size of house (e.g. thousands of square feet)
- The value that we seek to predict is called the dependent (or output) variable, and we denote this as
    - Y=price of house (e.g. thousands of dollars)

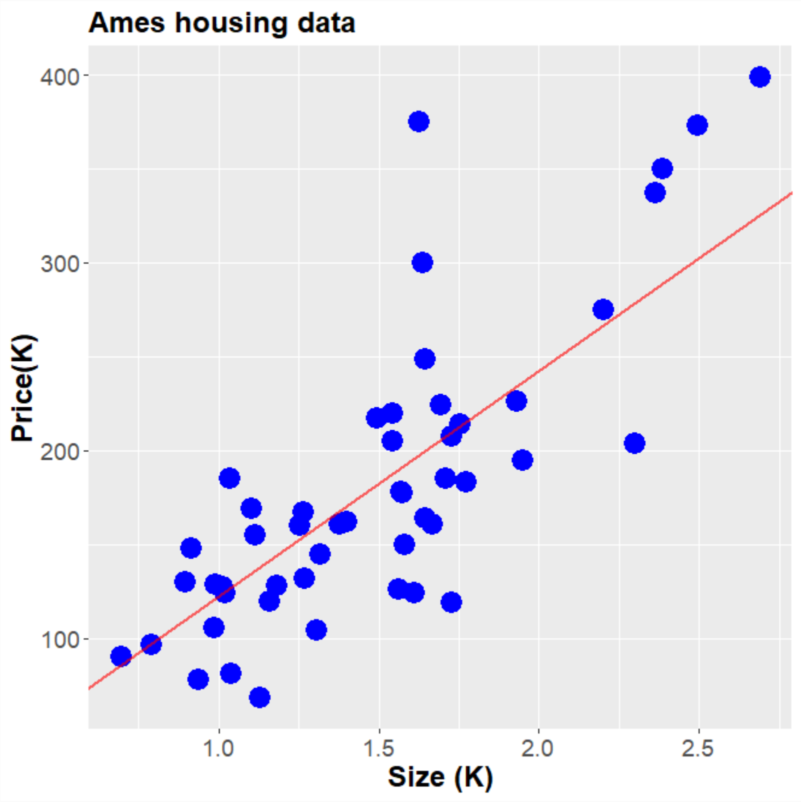
Example: Ames housing data

- Appears to be a linear relationship: as size goes up, price goes up.



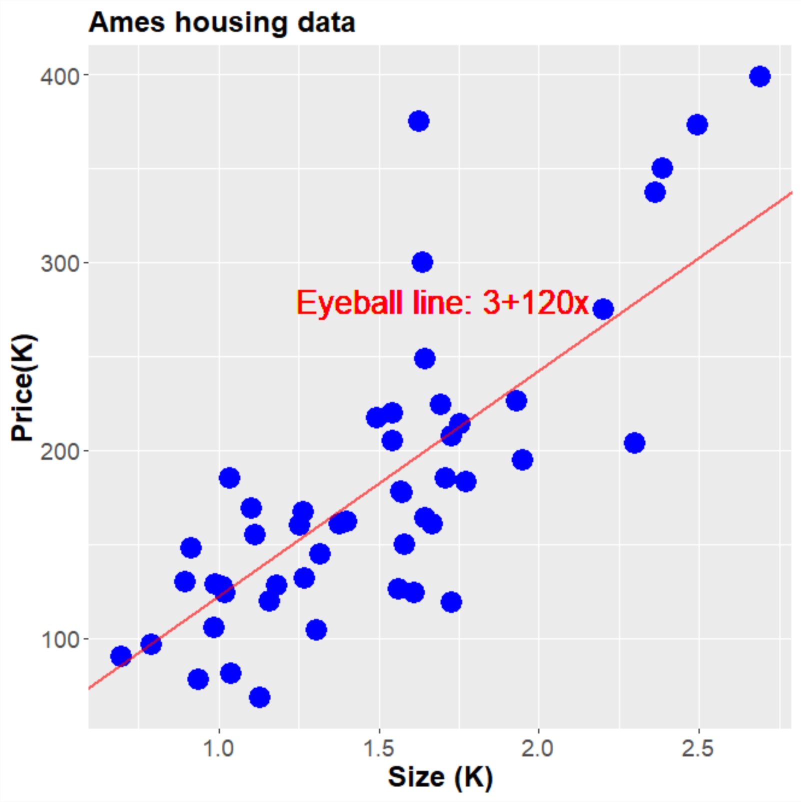
“Eyeball” method

- Appears to be a linear relationship: as size goes up, price goes up.
- Fitting a line by the “eyeball” method:



“Eyeball” method

- Appears to be a linear relationship: as size goes up, price goes up.
- Fitting a line by the “eyeball” method:



Linear prediction

- Recall that the equation of a line is:

$$Y = b_0 + b_1 X$$

where b_0 is the intercept and b_1 is the slope.

- The intercept value is in units of Y ($1,000).

- The slope is in units of Y per units of X ($1,000/1,000 sq ft).

Interpretation of coefficients

- Recall that the equation of a line is:

$$\text{Price of house} = 3 + 120 \times \text{size of house}$$

- **Slope** is 120:
    - The average price of a house increases by an estimated $ 120 for every square feet increase in size.
- **Intercept** is 3:
    - The average price of a house when 0 square feet of a house.
- **Does interpreting the intercept make sense in this problem?**

What is a good line?

Can we do better than the eyeball method?

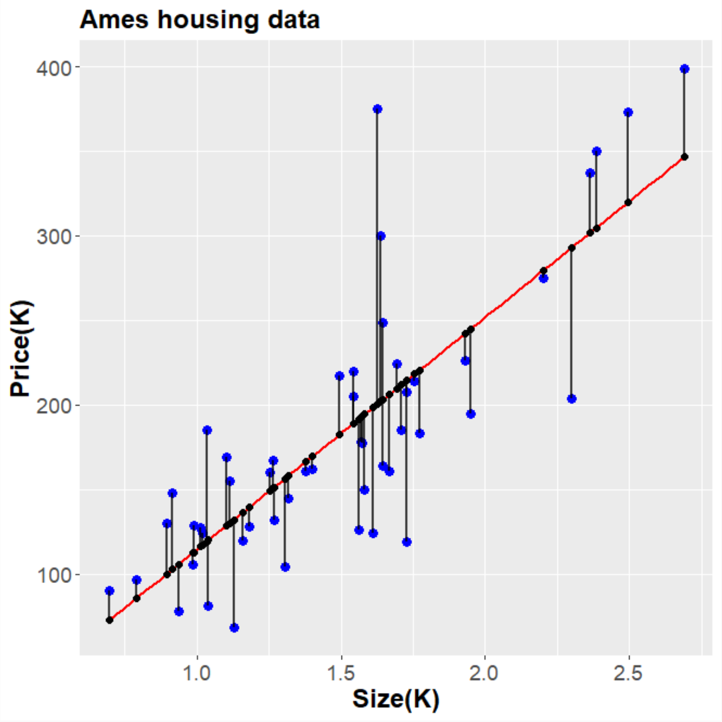
- We desire a strategy for estimating the slope and intercept parameters in the model $\hat{Y} = b_0 + b_1 X$.

That involves

- choosing a criteria, i.e., quantifying how good a line is
- and matching that with a solution i.e., finding the best line subject to that criteria.

A reasonable goal is to minimize the size of all residuals:

- Residual errors e_i is the distance from the observed value to the red solid line to $e_i = (Y_i - \hat{Y}_i)$.
- The red solid line is our predictions or fitted values: $\hat{Y}_i = b_0 + b_1 X_i$.



Least Squares (LS)

The line fitting process:

- Give weights to all of the residuals (positive and negative), .e.g e_i^2
- Trade-off between moving closer to some points and at the same time moving away from other points.
- Least square choose b_0 and b_1 to minimize

$$\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{N} (Y_i - [b_0 + b_1 X_i])^2$$

R's built-in lm() function

- The `lm()` function can be used to fit the SLR model (or any LM for that matter!)
    - In R, type `?lm` to view the associated documentation/help page

- The statement `lm(y ~ x, data = df)` fits an SLR model by regressing `y` on `x`, where `y` and `x` are columns in `df`

- To suppress the intercept term, use `y ~ x - 1` (not often necessary)

Example: Ames housing data

Fit an SLR model to the Ames housing dataa using `price` as the response and `size` as the predictor and interpret the estimated coefficients.

Code

```
set.seed(750)  # for reproducibility
data(ames, package = "modeldata") # Loac
ames$Price ← ames$Sale_Price / 1000  #
ames$Size ← ames$Gr_Liv_Area / 1000  #
ids ← sample.int(nrow(ames), size = 50)
ames.trn ← ames[ids, ]  # training (or
fit ← lm(Price ~ Size, data = ames.trn)
summary(fit)  # print a more verbose sun
```

Output

```
Call:
lm(formula = Price ~ Size, data = ames.trn)

Residuals:
    Min      1Q  Median      3Q     Max
-95.591 -27.706  -5.042  28.520 174.538

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -22.45      23.33  -0.962    0.341
Size          137.18      14.96   9.173 3.96e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘

Residual standard error: 48.36 on 48 degrees of freedom
Multiple R-squared:  0.6367,	Adjusted R-squared:  0.629
F-statistic: 84.14 on 1 and 48 DF,  p-value: 3.958e-12
```

Example: Ames housing data

The estimated model is:

$$\text{Price of house} = -22.45 + 137.18 \times \text{size of house}$$

Slope is 137.18

- The average price of a house increases by an estimated $137.18 for every square feet increase in size

Intercept is -22.45

- $\mathrm{E}\left(\text{Price}|\text{Size} = 0\right) = -22.45$
- Does interpreting the intercept make sense in this problem?

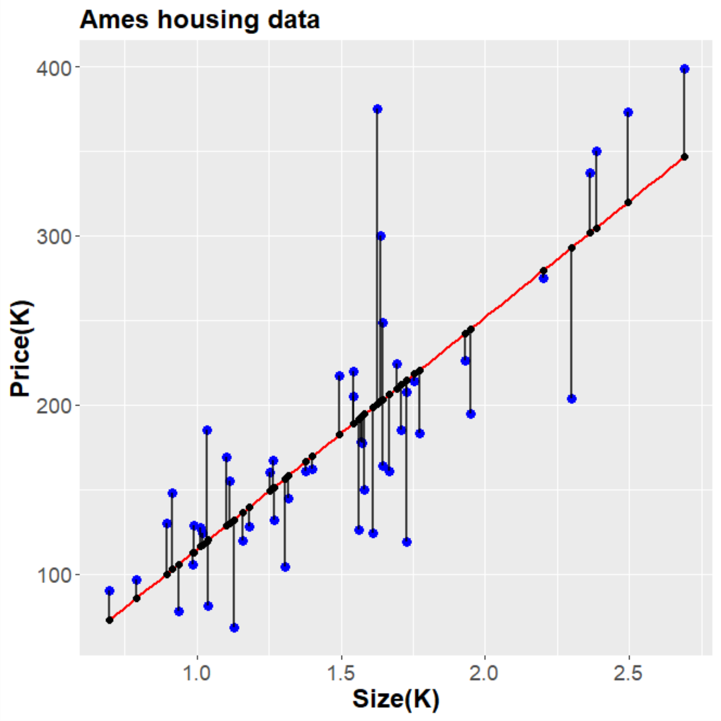
R^2 is 62.9%

- 62.9% of the price of house variation explained by size of house

The fitted LS line

```
tail(cbind(ames.trn$Price, "fitted_values" = fitted(fit)))
##        fitted_values
## 45 150      194.5631
## 46 226      242.1637
## 47 161      166.5789
## 48 106      112.9426
## 49 132      151.4894
## 50 373      320.0805
```

Residuals: $Y_i - \hat{Y}_i$



Code

```
ggplot(ames.trn, aes(x = Size, y = Price)) +
geom_point(size =3,color="blue")+
geom_smooth(method = "lm", formula = y ~ x,
  se = FALSE, alpha = 0.5, color="red") +
geom_segment(aes(x = Size, y = fitted(fit),
  xend = Size, yend = Price),
  alpha = 0.75, size=1, col = "black") +
geom_point(aes(x  = Size, y = fitted(fit)), color = "
  labs(x = "Size (K)",
       y = "Price (K)",
  title = "Ames housing data")+
  theme(axis.title = element_text(face="bold"))+
  theme(axis.title.y = element_text(face="bold"))+
  theme(text = element_text(size = 18))+
  theme(plot.title = element_text(face="bold", size=1
```

Steps in a regression analysis

1. State the problem
2. Data collection
3. Model fitting & estimation (this class)
   3.1 Model specification (linear? logistic?)
   3.2 Select potentially relevant variables
   3.3 Model fitting (least squares)
   3.4 Model validation and criticism
   3.5 Back to 3.1? Back to 2?
4. Answering the posed questions
   - But that oversimplifies a bit;
     - it is more iterative, and can be more art than science

Thank you!

Zhaohu(Jonathan) Fan, PhD Candidate in Business Analytics
fanzh@ucmail.uc.edu

Simple linear regression

- Data: $\{(X_i, Y_i)\}_{i=1}^{n}$
- Model: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
  - Y_i is a continuous response
  - X_i is a continuous predictor
  - β_0 is the intercept of the regression line
  - β_1 is the slope of the regression line
  - $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

More examples of statistical relationships

- Simple linear regression: $Y = \beta_0 + \beta_1 X + \epsilon$
- Multiple linear regression: $Y = \beta_0 + \sum_{i=1}^{p} \beta_p X_p + \epsilon$
- Polynomial regression: $Y = \beta_0 + \sum_{i=1}^{p} \beta_p X^p + \epsilon$
- Nonlinear regression: $Y = \frac{\beta_1 X}{(\beta_2 + X)} + \epsilon$
- and more.