Multiple Linear Regression, Model Comparison

Introducing a multiple linear regression model.

Sean Kang

Recap Simple linear regression

Equation of the SLR line

A simple linear model (with a slope and intercept) is of the form

$$\hat{y} = \theta_0 + \theta_1 x$$

Note, we have two parameters now. For simplicity's sake, we will instead say (for now): (lab)

$$\hat{y} = a + bx$$

We call this the **simple linear regression** model.

How do we know our fit?
RMSE and R^2

Evaluating models

What are some ways to determine if our model was a good fit to our data?

- Look at MSE or RMSE or R^2
- When our MSE or RMSE is very close to zero, and when R2 is closer to 1
- These are metrics to measure the fitness of our model

Root Mean Squared Error (RMSE)

$$\text{RMSE}(y, \hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Root mean squared error is defined as being the square root of the mean squared difference between predictions and their true values.

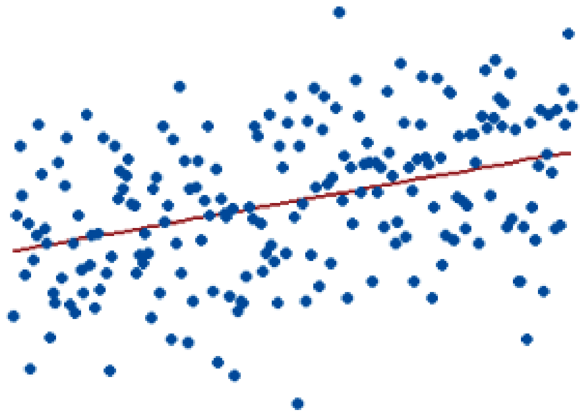
- It is the square root of MSE, which is the average loss that we've been minimizing to determine optimal model parameters.
- RMSE is in the same units as y.
- A lower RMSE indicates more "accurate" predictions.
    - Lower average loss across the dataset.
- Does not penalize as much as MSE due to square root

R SQUARE – R^2

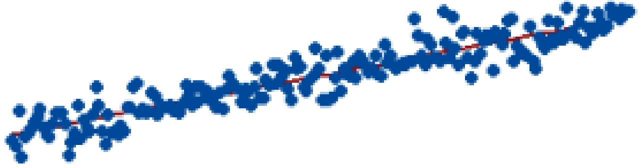
- It is a measurement of how good model is.
- It measures the relationship between the model and the dependent variables
- AKA
  - Coefficient of determination
  - Coefficient of multiple determination (for multiple regression)

Interpreting the Values of R Square

- 1 – Perfect Fit – suspicious
- ~ 0.9 – Very good
- < 0.7 – Not Great
- <0.4 – Terrible
- <0 Model makes no sense

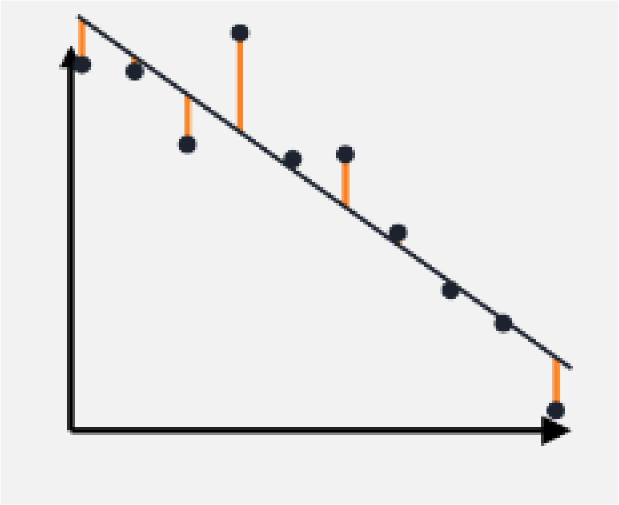


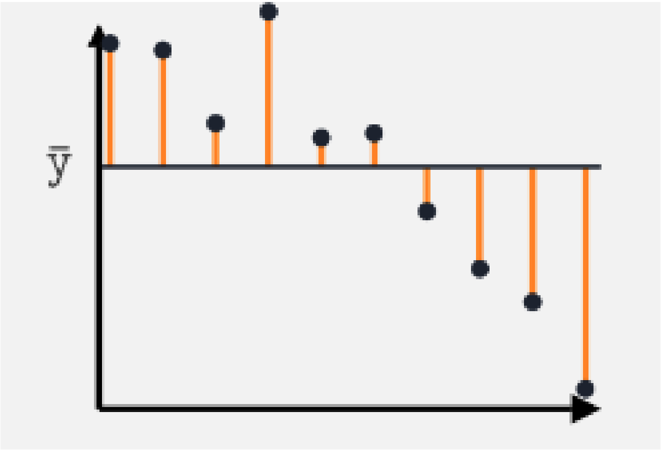
0.15



0.85

Y – hat versus Y- average





How to Calculate R Square

$$R^2 = \frac{\text{total variance} - \text{unexplained variance}}{\text{total variance}}$$

Unexplained Variance

$$= \frac{1}{m-1}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2$$

Total Variance

$$= \frac{1}{m-1}\sum_{i=1}^{m}(y_i - \bar{y})^2$$

Another way to calculate R^2

We define the **R^2** value as the square of the **correlation** between the true y and predicted y. This is also referred to as the **coefficient of determination**.

$$R^2 = [r(y, \hat{y})]^2$$

Since it is the square of a correlation coefficient (which ranged between -1 and 1), R^2 ranges between 0 and 1. Another way of expressing R^2, in linear models that have an intercept term, is

$$R^2 = \frac{\text{variance of fitted values}}{\text{variance of } y} = \frac{\sigma_{\hat{y}}^2}{\sigma_y^2}$$

Thus, we can interpret R^2 as the **proportion of variance** in our true y that our **fitted values** (predictions) capture, or "the proportion of variance that the **model explains**."

R^2

- As we add more features, our fitted values tend to become closer and closer to our actual y values. Thus, R^2 increases.
  - The simple model (AST only) explains 45.7% of the variance in the true y.
  - The AST & 3PA model explains 60.9%.
- Adding more features doesn't always mean our model is better, though!

$$R^2 = \frac{\text{variance of fitted values}}{\text{variance of } y}$$

$$\text{predicted PTS} = 3.98 + 2.4 \cdot \text{AST}$$

$R^2 = 0.457$

$$\text{predicted PTS} = 2.163 + 1.64 \cdot \text{AST} + 1.26 \cdot \text{3PA}$$

$R^2 = 0.609$

Multiple linear regression

MLR is an extension of Simple Linear Regression.

Adding independent variables

Estimated multiple regression model with two features (and thus, three parameters), is of the form

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

Adding More Independent Variables

- Adding more variables does not always mean a better fitting model.
- Does not always make predictions better
- More variables can explain more variations
- The best way:
    - Pick the best input variables for the model
- Problems:
    - Overfitting – add variance, or can explain for variation in output but does not add much prediction to the model.
    - Multicollinearity – independent variables are dependent on each other, not just the output variable

Idea Example

- Housing Price – the output dependent variable
- Input Independent Variables
  - Age of house
  - Square foot space of the house
  - Ratings of the high school
  - Great view from backyard
  - Interior Renovations

Bad Example

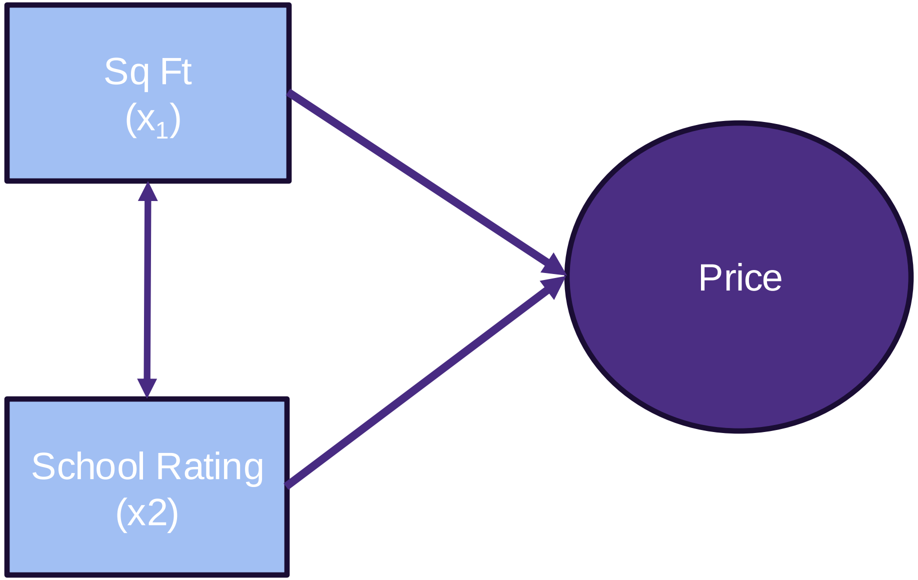
- Housing Price – the output dependent variable
- Input Independent Variables
  - Num of bed and bath
  - Square foot space of the house
  - Master bedroom on main floor

Prep Work

- Correlations
- Scatter Plots
- Simple Regressions

Let's Start

- Dependent Variable: Housing Price (Estimated)
- Independent Variable:
  - Square Foot of the living space
  - Ratings of the High School



If you add more variables

- The number of potential collinearity relationships between the independent variables exists and must be analyzed.
- Some independent variables are better than others
- Some have no contribution to the dependent variable

General notation

Our models can be expressed as a function $\hat{y} = f_\theta(x)$ of an input variable, x.

Constant model: $f_\theta(x) = \theta$

Type equation here.

Simple linear regression model: $f_\theta(x) = \theta_0 + \theta_1 x$

Multiple linear regression model: $f_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_p x_p$

Intercept

Coefficients

Variables

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

Interpretation

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

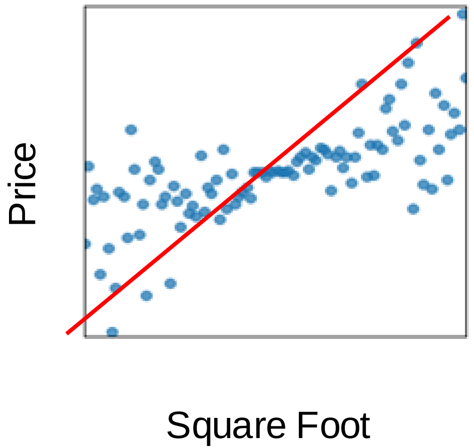
x_1 = square foot
x_2 = high school rating
$\hat{y}$ = price in $1000

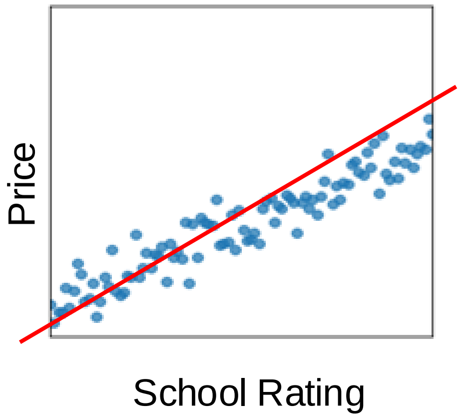
$$\hat{y} = 5 + 13x_1 + 6x_2$$

1 square foot of space would translate to 13k towards the price of the house, keeping the school rating constant.

If the high school rating is an ordinal from 1-10 where 10 is the best and 1 is the lowest, a school rating of 1 would only contribute 6k towards the price of the house.

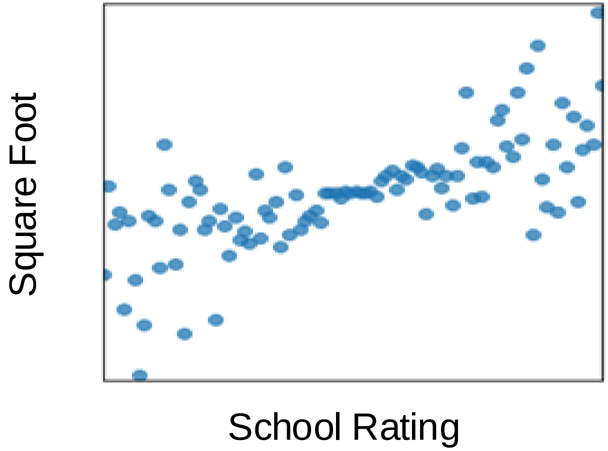
Some Scatter Plots of IV to DV





Both square footage and school ratings have a highly correlated to Price.

Some Scatter Plots of IV to IV



These two IV have a visual correlation relationship. We will just note that for now. We can calculate the actual correlation coefficient in the lab.

Summary

Summary

- We now know of three models, .
  - The constant model, .
  - The simple linear regression model .
  - The multiple linear regression model ..
- We looked at the correlation coefficient, r, and studied its properties.
- We solved for the optimal parameters for the simple linear model by hand, by minimizing average squared loss (MSE) or algebra in prior lecture.
- We introduced the notion of a feature, and how we can have multiple in our models.
- We discussed the R^2 coefficient and RMSE as methods of evaluating the quality of a linear model.