Introduction to Regression Modeling

Justin Pounders

Objectives

— Describe predictive and inferential modeling

— Describe the linear regression model

— Calculate and interpret a regression model

— Define, interpret and calculate loss/error metrics MSE and R^2

— Construct and interpret regression models in `sklean` and `statsmodels`

What is modeling?

— We can use models to conduct inference

— We can use models to make predictions

You have already built a lot of models in your life.

How long did it take you to get here today?

Model-Based Inference

— **Goal**: want precise relationships between variables

— Models are simplifications of reality that help us understand relationships

— What is the relationship between input X and output Y

What are some examples of inference from a mental model that would be valuable at work?

Model-Based Prediction

— **Goal**: estimate the value of a future or unknown variable in terms of currently-known data

— Predictions have value even if they are approximate

— Given input X, predict the value of output Y

What are some examples of prediction from a mental model that could be valuable?

So how can we build **a model?**

Say that you are given some (X, Y) data
- Y is fuel consumption (mpg)
- X is speed

Python Time

Go to notebook, "Naive/Baseline Prediction"

Simple Linear Regression

Simple Linear Regression

Assume a linear relationship between the dependent and independent variables:

$$Y = mX + b$$

where both X and Y are continuous.

What are some examples of continuous variables?

What are some examples of non-continuous variables?

Simple Linear Regression

$$Y = mX + b$$

— m is the slope

— b is the intercept

Example

Suppose my commute time to GA includes...

— 3 mins walking

— 5 mins per Marta stop

— 15 mins PCM shuttle

Model:

— Input: # Marta stops

— Output: commute time [mins]

Simple Linear Regression

Our model

$$Y = mX + b$$

is probably wrong.

Simple Linear Regression

A more realistic model may be

$$Y = mX + b + \epsilon$$

where ϵ is some error.

Simple Linear Regression

$$Y = mX + b + \epsilon$$

$$Y = b_1 X + b_0 + \epsilon$$

$$Y = \alpha + \beta X + \epsilon$$

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Simple Linear Regression

Consider:
- We have (X, Y) data
- Assume a linear model based on EDA
- What's the value of intercept (β_0) and slope (β_1)?

Simple Linear Regression

— Y is observed data

— $\hat{Y}$ is predicted value based on some β_0, β_1

$$Y \approx \hat{Y} = \beta_0 + \beta_1 X$$

Residual = error

$$\epsilon = Y - \hat{Y}$$

Simple Linear Regression

Residual sum of squares/sum of square errors:

$$RSS = SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Mean squared error:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

R^2: Coefficient of Determination

$$SS_{tot} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

$$SS_{res} = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Assumptions of Simple Linear Regression

1. **Linearity:** Y and X must have an approximately linear relationship.
2. **Independence:** Errors (residuals) ϵ_i and ϵ_j must be independent of one another for any $i \neq j$.
3. **Normality:** The errors (residuals) follow a Normal distribution.
4. **Equality of Variances** (Homoscedasticity of errors): The errors (residuals) should have a roughly consistent pattern, regardless of the value of X. (There should be no discernable relationship between X and the residuals.)

The mnemonic **LINE** is a useful way to remember these four assumptions.

Assumptions of Simple Linear Regression

If all four assumptions are true, the following holds:

$$Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma)$$