Linear Models

What is Statistical Learning?

Modeling

Recall: We are assuming that

$$(\text{output}) = f(\text{input}) + (\text{noise})$$

and we would like to estimate f.

Linear Model

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

Goal: Use observations $(x_1, y_1), \ldots, (x_n, y_n)$ to estimate β_0 and β_1.

Measures of success

What is the "best" choice of $\hat{\beta}_0$ and $\hat{\beta}_1$?

- The ones that are **statistically most justified**, under certain assumptions about Y and X?
- The ones that minimize the prediction error?
    - $|\hat{y}_i - y_i|$?
    - $(\hat{y}_i - y_i)^2$?
    - $(\hat{y}_i - y_i)^4$?
- The ones that minimize the prediction error *on new information*?
    - $|y_{new} - \hat{y}_{new}|$?
    - $(y_{new} - \hat{y}_{new})^2$?
    - $(y_{new} - \hat{y}_{new})^4$?

Maximum Likelihood Estimate (MLE)

Let's assume X is **Normally distributed** with some mean and variance.

Let's assume that Y *given* X is **Normally distributed** with a mean of $\beta_0 + \beta_1 X$ and some variance.

- Based on our data, what do we think are the mean and variance of X?
- Based on our data, what do we think are the mean and variance of Y?
- Based on our data, what is our *best guess* about β_0 and β_1?

What if Y was **Exponentially distributed**? **Poisson? Bernoulli**????

MSE/SSE/RSE/RMSE

The **residual** of a model is how much you "missed" by:

$y_i - \hat{y_i}$

The **squared residual**, or **squared error** is that squared:

$(y_i - \hat{y_i})^2$

MSE/SSE/RSE/RMSE

The **sum of squared error (SSE)** is all those added up:

sum over all i of $(y_i - \hat{y}_i)^2$

This is also called the **residual squared error (RSE)** or **residual sum of squares (RSS)**.

(Yes, those are all the same thing. Statistics is silly sometimes.)

MSE/SSE/RSE/RMSE

The **mean squared error (MSE)** is all those added up and then averaged:

sum over all i of $(y_i - \hat{y}_i)^2$ divided by n

The **root mean squared error (RMSE)** is the square root of the MSE:

sum over all i of $(y_i - \hat{y}_i)^2$ divided by n, then square root it

MSE/SSE/RSE/RMSE

What's the point of all this?

If we decide the "best" β_0 and β_1 are the ones that minimize the **MSE**...

that's exactly the same as minimizing the **RMSE**, the **RSS**, the **SSE**, the **RSE**!

These are all measuring "how far are our predictions from the truth, in squared distance?"

NOT the same as "absolute" error ($|y_i - \hat{y_i}|$)

Checkpoint:

Before any modeling analysis, you have to decide what your definition of the **BEST MODEL** is!

Linear model estimates

For linear models (but not *every* model!) the "best" model according to the (Normal) **MLE** and the "best" model according to the **MSE** agree!

Okay, so how do we calculate β_0 and β_1?

Make the computer do it!

Get the idea, not the math

What you DON'T need to know:

- What mathematical equations are used to compute β_0 and β_1.
- What statistical properties your estimates of β_0 and β_1 have.
- How the computer efficiently does these calculations.

What you DO need to know:

- How to make a computer do the calculations
- Which measure of model success lead to this estimation choice
- How to interpret the results in the real world.

Let's get started.