Introduction to Linear Regression

Problem Motivation

Q: How to make predictions?

- Ex. Predict home selling price based on square feet, location, number of bedrooms, etc.
- Ex. Predict pageviews based on day of week, product category, etc.

A: Popular method is Linear Regression

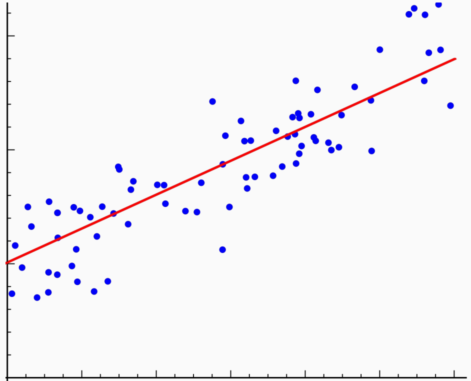
Basic Formulation

$$E[HomePrice] = \beta_0 + \beta_1 SquareFeet + \beta_2 NumBedrooms$$

Linear - target is predicted by linear combination of features

Regression - target is continuous (as opposed to categorical)

Linear regression assumes the target variable, on average, equals a weighted sum of its features



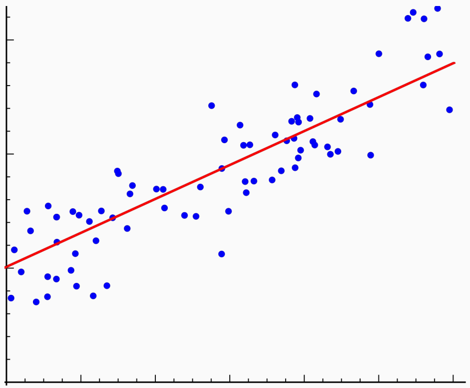
Basic Formulation

Start by assuming linear model:

$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

Based on data, estimate beta coefficients:

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X_1$$



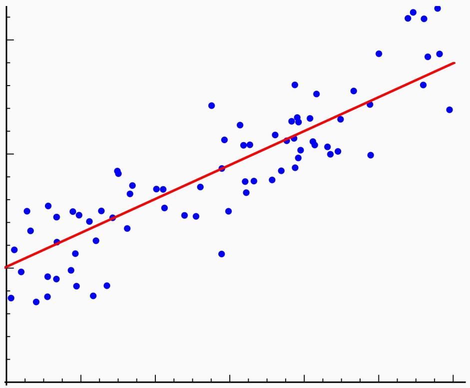
Basic Formulation

Simple Linear Regression

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X_1$$

Multiple Linear Regression

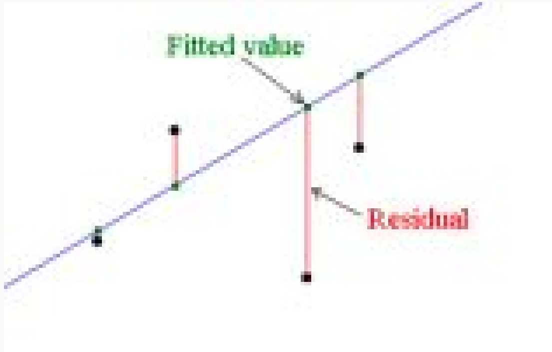
$$\widehat{Y} = \widehat{\beta}_0 + \sum_{i=1}^{p} \widehat{\beta}_i x_i$$



Estimating Coefficients

- Beta coefficients are estimated to minimize the *squared* error
- Error term ε, aka the "residual," represents difference between predictions, and is assumed to be i.i.d ~ $N(0, \sigma^2)$

$$Y = \beta_0 + \beta_1 X_1 + \epsilon \rightarrow \widehat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X_1$$



Estimating Coefficients

Cost Function

Choose betas which minimize “residual sum of squares”

$$RSS = \sum_{i=1}^{n} (y_i - \widehat{y})^2$$

$$RSS = \sum_{i=1}^{n} \left(y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_1)\right)^2$$

Estimating Coefficients

Matrix Form

Basic model: $Y_{n \times 1} = X_{n \times p} B_{p \times 1} + \epsilon_{n \times 1}$

Error: $\epsilon = (Y - XB)$

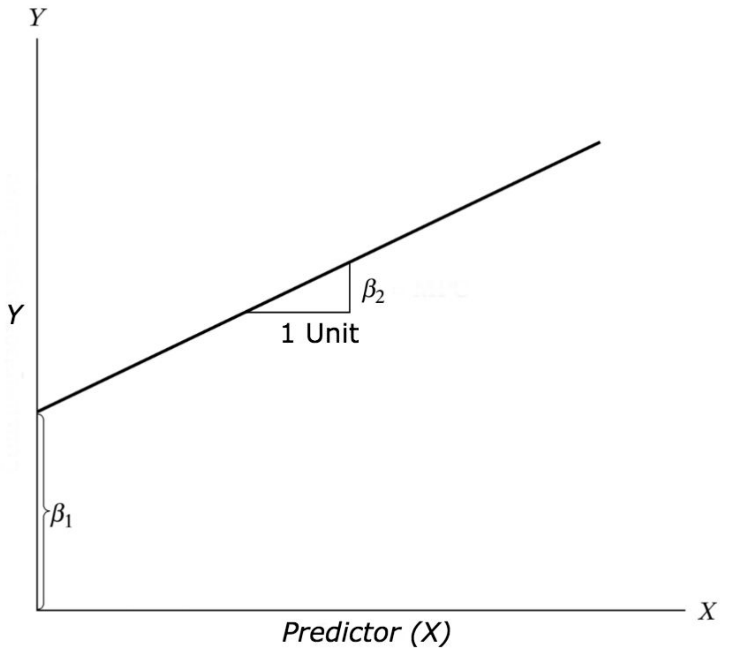
Betas that minimize TSS: $B = \left(X^T X\right)^{-1} X^T Y$

Interpreting Coefficients

One unit change in predictor →

beta change in target

Image credit: http://r-statistics.co/Linear-Regression.html



Model Evaluation

How do we know if a linear regression model is reliable?

1. R^2
2. Coefficient p-values
3. Coefficient confidence intervals
4. F statistic

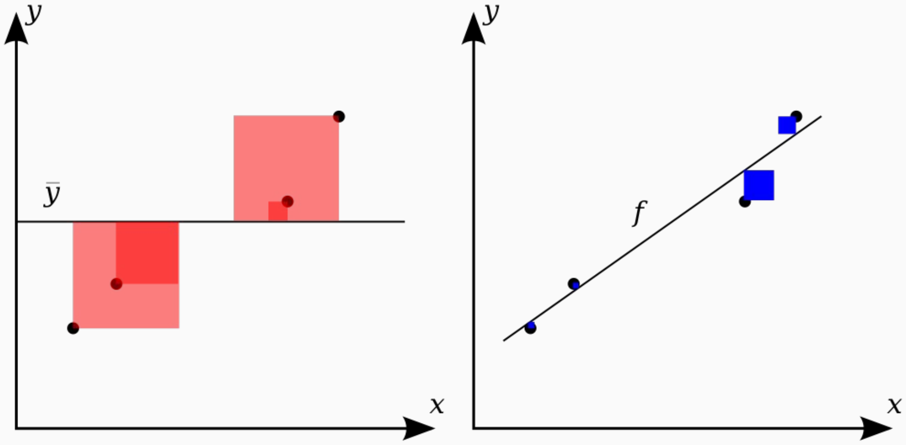
Model Evaluation - R^2

- compares the model with the mean
- interpreted as percent of variance explained by the model

$$R^2 = 1 - \frac{RSS}{TSS}$$

$$RSS = \sum_{i=1}^{n} (y_i - \widehat{y})^2$$

$$TSS = \sum_{i=1}^{n} (y_i - \overline{y})^2$$



Model Evaluation - R^2

- R^2 necessarily improves with the addition of each new feature (even if that features is irrelevant!)
- High R^2 by itself doesn't imply a good model

Model Evaluation - p-values and confidence intervals

- Beta coefficients have sampling distributions
- Can perform hypothesis test on coefficients

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\sigma^2 = \mathrm{Var}(\epsilon)$$

	Recall	Here
Setup Hypothesis	$H_0 : \mu = \mu_0 = 100$	$H_0 : \beta_1 = 0$ ← Test if X has effect on Y
Sample Statistic	$\bar{x}$	$\hat{\beta}_1$
Test Statistic	$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$	$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$
Confidence Interval	$(\bar{x} - t_{\alpha/2} * \frac{s}{\sqrt{n}},\ \bar{x} + t_{\alpha/2} * \frac{s}{\sqrt{n}})$	$[\hat{\beta}_1 - t_{\alpha/2} * SE(\hat{\beta}_1),\ \hat{\beta}_1 + t_{\alpha/2} * SE(\hat{\beta}_1)]$

Model Evaluation - F-test

Compares model with null model:

$$H_0 : \beta_i = 0 \ \forall i \text{ not including intercept}$$
$$H_1 : \beta_i \neq 0 \text{ for some } i$$

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} \sim F_{p,n-p-1}$$

Shortcoming: doesn't tell you which beta is unequal to zero.