Linear Regression Assumptions

We've covered in detail the practicalities of specifying, fitting, and inspecting a linear regression model.

Now let's turn to the mathematical and statistical details of linear regression.

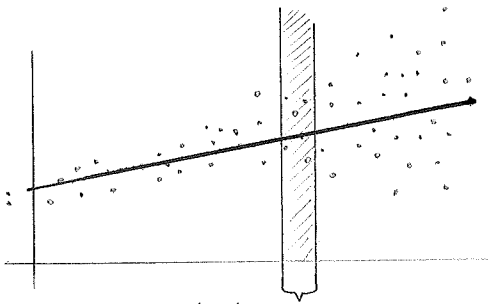
Basic Assumption: Linearity

To say much of anything, we need to be sure that our model **fits the data well!**

Pretty much everything we have done so far is to ensure that this is, to the best of our ability, true.

In detail, we assume our data is generated like this:

$$Y \mid X_1, X_2, \cdots, X_K \sim \beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K + \varepsilon$$



The residuals in any small band should approximately average to zero.

Where:

$X_1, X_2, \cdots, X_K$ are the features used in the model

$Y \mid X_1, \cdots, X_K$ is the value of Y given the known values of $X_1, \cdots, X_K$

ε is a random component.

Our **only** assumption about ε is that $E[\varepsilon \mid X_1, X_2, \cdots, X_K] = 0$.

When we have data $\{(x_{i1}, x_{i2}, \cdots, x_{ik}, y_i)\}$ we **estimate the parameters** by minimizing the residual squared error:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_i \left(y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}) \right)^2 \right\}$$

We want to know: **what properties does $\hat{\beta}$ enjoy?**

Desirable Properties of $\hat{\beta}$

Here are some properties we may wish for our estimate $\hat{\beta}$:

Existence: The optimization problem defining $\hat{\beta}$ has one, and only one, solution.

Consistency: $\hat{\beta} \longrightarrow \beta$ as $n \longrightarrow \infty$. I.e. as we collect more data, our estimate becomes more accurate.

Unbiasedness: $E[\hat{\beta}] = \beta$. I.e. no matter what sized data set we have, if we take the average $\hat{\beta}$ across many datasets, this average value is the correct answer.

Efficiency: $\operatorname{Var}(\hat{\beta}) \leq \operatorname{Var}(\tilde{\beta})$ for any $\tilde{\beta}$ created with a different algorithm.

Distributional Knowledge: We know the sampling distribution of $\hat{\beta}$.

To guarantee these properties, we need to make various assumptions about the data X, y, and the random noise ε.

Existence:

It would be nice if solutions exist, and are unique.

For example, if we use the same feature twice $x_1, x_2 = X$, then all of these equations are the same:

$$\left.\begin{aligned} y &= 4x_1 \\ y &= 3x_1 + x_2 \\ y &= 2x_1 + 2x_2 \\ y &= x_1 + 3x_2 \\ y &= 4x_2 \end{aligned}\right\} \text{ There are an infinite \# of possibilities.}$$

This is, more or less, all that can go wrong.

Assumption: Linear Independence:

Assume the columns of the feature matrix $\mathbb{X}$ are linearly independent.

Under this assumption, the parameter estimates exist, are unique, and are solutions to the system of linear equations:

$$\mathbb{X}^t \mathbb{X} \hat{\beta} = \mathbb{X} y$$

Note: If columns are almost linearly independent (ie, highly correlated), then the estimates $\hat{\beta}$ become very unstable (sensitive to the data).

Consistency

We would like collecting more data to **always improve** our estimate of $\hat{\beta}$.

This can fail if the y's are dependent. I.e. if we observe the same data again and again, we do not really get more information, so our estimates do not improve.

Assumption: Independence of Errors

Asume the values of ε are independent of one another

$\varepsilon_1, \varepsilon_2, \varepsilon_3, \ldots$ are all independent.

Then $\hat{\beta} \to \beta$ as $n \to \infty$.

I.e. the estimated regression line converges to the true line as we collect more data.

Failure in practice

- Estimating cancer rates in the u.s. by sampling only people living in a toxic waste dump.
- Estimating the rate I land kickflips by sampling on only one day.
- Estimating the effect of studying math by studying the same 10 people over multiple nights.

Distribution of Parameter Estimates

The most information we could possibly have is if we knew the distribution of $\hat{\beta}$. This would allow us to compute probabilities of events involving $\hat{\beta}$, run hypothesis tests, etc...

Assumption: Distribution of errors

Assume the distribution of the errors ε is normal

$$\varepsilon_1, \varepsilon_2, \varepsilon_3, \dots \sim \text{Normal}(0, \sigma)$$

Equivalently, assume the conditional distribution of Y given X is normal

$$y \mid X_1, \dots, X_k \sim \text{Normal}(\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k, \sigma)$$

This is a constant, i.e., does not depend on X. This is often called homoskedacity.

Then the distribution of $\hat{\beta}$ is also normal

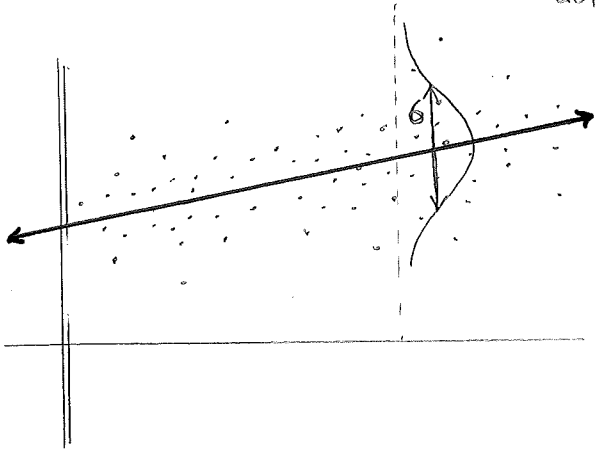
$$\hat{\beta} \sim \text{Normal}(\beta, \sigma^2 (X^t X)^{-1})$$

This is a multivariate normal distribution when there is more than one parameter.

and consequently the distribution of each coefficient is normal as well

$$\hat{\beta}_j \sim \text{Normal}\left(\beta_j, \frac{\sigma^2}{\underbrace{\vec{x}_j \cdot \vec{x}_j}}\right)$$

dot product of j^{th} feature with itself.



Note: This assumption is often much too restrictive for most applications.

It is **only needed** if you wish to run hypothesis tests on coefficients.