Linear Regression

Review of linear regression model

Terms

- X_1 - X_p are predictors. Typically non-random because they assume no variance (i.e., measured 100% exactly right).
- Y_i is response variable.
- e random variable with mean 0 that is the noise (error of other discrepancies) present in a model.

Basic

- If there are multiple quantitative predictors, the model is fitting a hyper-plane in multi-dimensional space.
- We can estimate p + 1 parameters where the "+1" is the intercept.

Least squares

- Least Squares is a method for estimating the unknown parameters β0,β1,...,βp where we minimize the sum of the squared errors (the differences between the estimated values Y^i and the actual values Yi in our *training* data Yi).
  - $$MSE = E(\hat{y}_i - y_i)^2$$
  - With one equation for each parameter, each equation is a restriction which reduces the MSE degrees of freedom to n–(p+1).

f test

- non-simmetrical distribution
- We also use the LINE assumptions to support the F-test which allows us to test multiple variables for significance at once (vs t-test which only lets you test 1)
- Can compare variance in two models (e.g., in a full + reduced model)

T-test

- Now that we have some estimates for our parameters, we need to check if the variables have predictive value in our model. One way to do that is with the t-test.
- you assume null hypothesis of 0 effect bc if we picked a number (e.g., 41) we would need to KNOW the distribution of this variable. But we already know the distribution if the variable is 0.
- normal distribution

Regression output

- the *Estimate* is the β^k for each parameter. This is the average increase in the response variable associated with a one unit increase in the predictor variable, assuming all other predictor variables are held constant.
- the *Std. Error* is the square root of the estimated variance for the parameter.
- the t *value* is the ratio of Estimat.eStd.Error
- Pr>|t| is the result from checking the t distribution with the degrees of freedom n−(p+1). Here that means df = 392−(4+1)=387 since there are four variables plus the slope parameter.
- **Residual standard error**: The average distance observed values are from the regression line. Smaller is better.
- **Multiple R-Squared**: (aka the coefficient of determination). This is the proportion of the variance in the Y that can be explained by the predictors.
- **Adjusted R-squared**: A version of R-squared based on a penalty for the number of predictors. Useful for comparing models with different numbers of variables.
- **F-statistic**: The ratio of MSR/MSE=SSR/pSSE/(n−(p+1).
  - Use this to test the value of the overall model based on a Null hypothesis of all $\beta k=0$, i.e., is there at least one variable with a $\beta k\neq 0$.
  - Note SSR is the sum of the squared difference between the y^i and y^- or SSR=Σ(y^i−y^-)2.
- p-value: This is the p-value for the F-statistic with df of p and n−(p+1).

4 assumptions of linear regression

- **Linearity**: The response and predictor variables have a linear relationship.
- **Independence**: the predictor variables are independent of each other — no multicollinearity.
- **Normal**: The residuals have a Normal distribution with mean 0 and equal variance for all values of X.
- **Uncorrelated Errors**: The residuals are uncorrelated with each other.

Goodness of fit

- Goodness of Fit or Lack-of-Fit is a discussion about over-fitting and under-fitting.

  It is about testing if a linear model (the first LINE assumption) is a good fit to explain the relationship between Y and X as expressed in

  (3.22)$Y\longrightarrow=X\beta\longrightarrow+\epsilon\longrightarrow$

  In Linear Regression, the Null hypothesis $H0:\beta\longrightarrow=0$ asserts there is no relationship between Y and X versus an alternative of there is a linear trend or component to the relationship between Y and X.

- The Alternative model is a **Saturated Regression Model**, the **most** general model we can create, with **maximum flexibility**.

  We will estimate the saturated g(X) based on the values of X.

  - For a **single** value of xi, with **multiple** (k) responses yi,k, the most general model is $yi,k=g(xi)+\epsilon$
  - We can estimate g(xi) by using the average of the yi,k. This is also the least squares estimate, yxi^{-}.

MISC

iid = identically, independently distributed

for 3 level categorical variable: 1 predictor, but uses three degrees of freedom (since one level is reference category)

for Auto data analysis

- If you remove intercept, the estimates are based on if y = 0 (vs y = the intercept). Reduces a degree of freedom and hence making it under-fit the data.