Linear Regression

There are four main assumptions of a linear regression model:

1) The first assumption is linearity. That is, the relationship between X and Y must be linear.
2) The second assumption is independence of errors. There is not a relationship between the residuals and the Y variable; in other words, Y is independent of errors.
3) The third assumption is equality of variance or homoscedasticity. This means that The variance of the residuals is the same for all values of X.
4) The final assumption is the normality of residuals, this requires the residuals to be approximately normally distributed.

The main use of linear regression is for prediction of a dependent variable based on known variables of one or more independent variables. Other uses of linear regression include determining the relationship between a dependent and independent variable(s)—that is the strength and direction of the relationship—and a second use is trend analysis, the goal here being to understand the trend of a variable overtime. Given its simplicity and interpretability, linear regression is a great tool to use in various situations. However, there are situations where this model is not optimal, such as when the relationship between variables is *not* linear, in that case we will want to consider non-linear methods or polynomial regression instead. Moreover, linear regression deals with continuous variables, and classification problems require discrete (and typically binary) values as the outcome. In these situations, we may consider employing logistic or probit regression methods.

Logistic Regression

Much like ordinary least squares (OLS), using logistic regression to make inferences requires model assumptions.

1) The response variable is dichotomous (two possible responses) or the sum of dichotomous responses.
2) Logistic regression assumes the observations to be independent of each other and independent of repetitive measurement.
3) The errors in logistic regression are not assumed to be normally distributed because the response variable is binary, violating the assumptions of normality.
4) Both logistic regression and linear regression have common assumptions:
    a. A linear relationship between the explanatory variables and the response variable.
        i. The log of the odds ratio, log(p/1-p), must be a linear function of x.
    b. Homoscedasticity between the residuals.

Logistic regression is typically used when the dependent variable is a binary or dichotomous variable. Therefore, the main uses of a logistic regression is to find the probability of a binary event occurring, and to tackle issues of classification. This regression method is preferable to other methods of predicting a binary outcome such as a probit regression. the choice between the logit and probit models is largely one of convenience since the substantive results are generally indistinguishable. For instance, the interpretation of betas is more intuitive in a logistic regression than in a probit regression logistic regression, a one unit change in X1 is associated with a $\beta 1$ change in the log odds of 'success', all else being equal. On the other hand, a probit would be a change of $\beta 1$ z's. To convert these into predicted probabilities, you can pass them through the normal CDF, or look them up on a z-table. For this reason a logistic regression can be preferable to a probit regression model. Additionally, logistic regression can be better in the presence of “extreme independent variables.” These are variables where one particularly

large or small value will overwhelmingly often determine whether the dependent variable is a 0 or a 1, overriding the effects of most other variables.

Probit Regression

The probit regression assumptions are similar to that of a logistic regression with some minor changes:

1) The outcome is binary. The response variable is dichotomous (two possible responses) or the sum of dichotomous responses.
2) The probit (aka z-score) of the outcome and independent variable have a linear relationship.
3) The normality of residuals assumption requires the residuals to be approximately normally distributed.
4) The final assumption is independence of errors. There is not a relationship between the residuals and the Y variable; in other words, Y is independent of errors.

Probit regression is used to model dichotomous or binary outcome variables. In the probit model, the inverse standard normal distribution of the probability is modeled as a linear combination of the predictors. Therefore, the main uses of a probit regression is to find the probability of a binary event occurring, and to tackle issues of classification. The probit model estimates the probability a value will fall into one of the two possible binary (i.e. unit) outcomes. Like logistic regression, probit regression does not require an assumption of homoskedasticity, but it does assume that the errors are normally distributed, unlike logistic regression. The main reason probit regression is to do some more complicated modelling on a Normal latent variable -- e.g. panel data models, or multivariate outcomes, or measurement error in predictors. In other words, probit regression works better in cases of "random effects models" with moderate or large sample sizes; random effects models with moderate size data sets are improved generally by selecting the probit link.

Poisson Regression

Using a Poisson regression to make inferences requires model assumptions:

1) The response variable is a count per unit of time or space, described by a Poisson distribution. These counts are positive integers (i.e. 0,1, 2,…k).
2) The observations in the dataset should be independent of each other. This ensures that the model's residuals are not correlated.
3) Since counts must follow a Poisson Distribution, the mean of a Poisson random variable must be equal to its variance, this is the assumption of equidispersion.
4) The log of the mean rate, $\log(\lambda)(s)$ must be a linear function of x.

Poisson regression is for modeling count variables. It can be used to answer questions such as what factors can predict the frequency of an event. The Poisson distribution is unique in that its mean and its variance are equal. In terms of when to use, and when its preferable to use a Poisson regression-- the Poisson regression model is a great model to reach for anytime you need a simple baseline model for count data. The Poisson regression model is simpler than other count-based regression models like zero-inflated Poisson, negative binomial, and zero-inflated negative binomial and it has the least parameters to fit. Additionally, since the Poisson regression is more simple and has less parameters to estimate, its preferable to use when you're working with a smaller sample size.

Negative Binomial Regression

Negative binomial regression shares many common assumptions with Poisson regression, such as linearity in model parameters, independence of individual observations, and the multiplicative effects of independent variables:

1) The observations in the dataset should be independent of each other. This ensures that the model's residuals are not correlated.
2) The relationship between the dependent variable (count data) and the independent variables should be linear. The model assumes that the effect of each independent variable on the log count rate is constant across all levels of that variable.
3) This is a fundamental characteristic the model addresses, assuming that the variance is greater than the mean in the count data, this is the assumption of overdispersion. It's essential to confirm that the model adequately handles overdispersion.

Negative Binomial regression is for modeling count variables, usually for over dispersed count outcome variables–that is when the conditional variance exceeds the conditional mean. Overdispersion means that the variance of the response is greater than what's assumed by the model. It can be considered as a generalization of Poisson regression since it has the same mean structure as Poisson regression, and it has an extra parameter to model the over-dispersion—it offers greater flexibility in model fitting. If the conditional distribution of the outcome variable is over-dispersed, the confidence intervals for the Negative binomial regression are likely to be narrower as compared to those from a Poisson regression model—therefore this is when we would choose a Negative Binomial regression over a Poisson.

<u>Zero-inflation Poisson Regression</u>

1) The response variable is a count per unit of time or space, described by a Poisson distribution. These counts are positive integers (i.e. 0,1, 2,…k).
2) The observations in the dataset should be independent of each other. This ensures that the model's residuals are not correlated.
3) Since counts must follow a Poisson Distribution, the mean of a Poisson random variable must be equal to its variance, this is the assumption of equidispersion
4) The log of the mean rate, log(λ)(s) must be a linear function of x.
5) The model assumes that the excess zeros are generated by a separate process from the count data; the excess zeros are generated by a separate process from the count values and that the excess zeros can be modeled independently.

Zero-inflated Poisson regression is used to model count data that has an excess of zero counts. Thus, the zip model has two parts, a Poisson count model and the logit model for predicting excess zeros. This preferable to use, than say a Poisson regression, if the distribution of your outcome variable is zero-inflated— if the distribution of your outcome variable is zero-inflated, then you should expect to see an excessive number of zeros. So this is in a situation where we have data sets that contain more number of zero valued counts than what one would expect to observe using the traditional model's probability distribution, so it's a situation where we need to account for the presence of extra zeros. It's important to note that we are in a situation where the data are not over-dispersed, i.e. when variance is not much larger than the mean and we have an excessive number of zeros.