Regression Primer

Variable Selection

Boruta Algorithm: The Boruta algorithm is a wrapper built around the random forest classification algorithm. It tries to capture all the important, interesting features you might have in your dataset with respect to an outcome variable. ***Feature Selection***

Mallows CP: help you choose between multiple regression models. It helps you strike an important balance with the number of predictors in the model. Mallows' Cp compares the precision and bias of the full model to models with a subset of the predictors.

Usually, you should look for models where Mallows' Cp is small and close to the number of predictors in the model plus the constant (p). A small Mallows' Cp value indicates that the model is relatively precise (has small variance) in estimating the true regression coefficients and predicting future responses. ***Model Selection***

Descriptive Analysis

- Use data visualization tools: histograms for predictors and dependent variable. Use qqplot to understand the distribution of the variables.
- After the initial data visualization, run a correlation plot amongst the predictor variables to understand the correlation.
    - When independent variables are highly correlated, change in one variable would cause change to another and so the model results fluctuate significantly. The model results will be unstable and vary a lot given a small change in the data or model
    - Highly correlated predictors can lead to collinearity issues, and this can greatly increase the model variance, especially in the context of regression. In some cases, there could be relationships between multiple predictor variables, and this is called multicollinearity. Having correlated variables will result in unnecessarily complex models with more than necessary predictor variables.
    - From a data collection point of view, spending time and money for collecting correlated variables could be a waste of effort. In terms of linear regression or the models that are based on regression, the collinearity problem is more severe because it creates unstable models where statistical inference becomes difficult or unreliable.
    - **On the other hand, correlation between variables may not be a problem for the predictive performance if the correlation structure in the training and the future tests data sets are the same.** However, more often, correlated structures within the training set might lead to overfitting.

- - Plot a density plot: Density plots are a variation of Histograms. It charts the values from a selected column as equally binned distributions. It uses kernel smoothing to smoothen out the noise. Thus, the plots are smooth across bins and are not affected by the number of bins created, which helps create a more defined distribution shape.

- Identify non-linearities from visualizations and suggest appropriate transformations.
  - When a residual plot reveals a data set to be nonlinear, it is often possible to "transform" the raw data to make it more linear. This allows us to use linear regression techniques more effectively with nonlinear data.
  - **Box Cox Transformations**: is a transformation of **non-normal dependent variables** into a normal shape. Normality is an important assumption for many statistical techniques; if your data isn't normal, applying a Box-Cox means that you are able to run a broader number of tests.
    - Use the confidence interval to determine whether a transformation is appropriate, as follows:
      - For the Box-Cox transformation, a λ value of 1 is equivalent to using the original data. Therefore, if the confidence interval for the optimal λ includes 1, then no transformation is necessary.
      - If the confidence interval for λ does not include 1, a transformation is appropriate.
    - If lambda is close to 0 or 0 implement a log transformation.

- Dealing with outliers, influential and high leverage
  - First look at residual with respect to dependent variables
  - **Q-Q Plot** This plot shows the studentized residuals vs. the corresponding quantiles of a t-distribution with N – K - 2 degrees of freedom, and allows us to identify outliers (std. residuals & 2 are considered large).
  - **Bar Plot of Cook's** distance to detect observations **that strongly influence fitted values of the model**. Cook's distance was introduced by American statistician R Dennis Cook in 1977. It is used to identify influential data points. It depends on both the residual and leverage i.e it takes it account both the x value and y value of the observation. (**Anything greater than 1 gets the boot)**
  - **DFBETA** measures the difference in each parameter estimate with and without the influential point. There is a DFBETA for each data point i.e if there are n observations and k variables, there will be n∗kn∗k DFBETAs. **In general, large values of DFBETAS indicate observations that are influential in estimating a given parameter.**
    - **For influential and high leverage observations.** For example, high leverage observations appear the in added variable plots as points horizontally distant from the rest of the data

- **DFFITS** Plot Proposed by Welsch and Kuh (1977). It is the scaled difference between the ith fitted value obtained from the full data and the ith fitted value obtained by deleting the ith observation. DFFIT - difference in fits, is used to identify influential data points. It quantifies the number of standard deviations that the fitted value changes when the ith data point is omitted.

- **Predictor Transformations**
  - Component Residual Plots
    - When there are strong non-linear relations between predictors, the component-plus-residual plots may not work, in this case, we instead look at CERES plots
  - Combining conditional Expectations and Residuals

- **Omitted Variable Bias**
  - The Omitted Variable Bias quantifies the impact of omitting a relevant (e.g., statistically and/or economically significant) variable in a regression model.

- **Irrelevant Variables**
  - Typically, including irrelevant variables into the model has the effect of increasing the variance of the other variables, and therefore, potentially rendering the respective coefficient estimates statistically insignificant.

Multicollinearity

- We understand that leaving certain variables out of our model can potentially introduce bias into our OLS estimators It therefore may be tempting to include any and all possible variables to avoid introducing bias into our models. This is a poor solution that may create an entirely new problem in our model - inflating our standard errors or variance estimates
  - **VIF:** The variance Inflation Factor was developed as a measure of how much the variance of Var(βj) is inflated by the correlation between and our other explanatory variables
    - If we rely only on VIF we may end up omitting relevant variables and end up biasing our regression estimates by leaving out important variables
  - **Correlation Plots:**
  - **Compare multiple models to understand effects**

Model Building & Selection

- Conduct F-tests to check whether independent variables are jointly significant
- ***Ramsay RESET***
    - The Ramsey RESET test is a general test of functional form misspecification
- **AIC and BIC**
    - AIC and BIC are both measures of model fit that penalize the use of many parameters in a model. Using these measures, we hope to avoid over fitting by not including too many parameters in our model
        - BIC will penalize high dimensional models more than AIC
- **Cross-Validation**
    - K-fold cross validation is a technique to evaluate how well a given model predicts out of sample data
    - The algorithm is as follows:
        - Shuffle data and split into K groups
        - Leave one group out, train the model on the rest
        - Predict the group that was left out
        - Compute evaluation score (mean squared error, RMSE, accuracy, etc.)
        - Store the score for later analysis
        - We can then pick the model that, for example, on average has the lowest error or variance

---

Heteroskedasticity

- The good news is that heteroscedasticity does not bias our OLS estimates or affect their consistency. R^2 is also not affected.
- Heteroscedasticity will, however, invalidate any inference made on the basis of standard errors that were estimated in the presence of heteroscedasticity:
    - Confidence intervals
    - T-statistics
    - F-statistics

- *Detecting Heteroscedasticity*
    - **Breusch-Pagan**: null-hypothesis → homoscedasticity
    - **White-Test**: white test for heteroscedasticity fundamentally is implemented by regressing the squared residuals on all our original regressors, their cross products and squares
        - This specification will capture more types of heteroscedastcity, but also may detect specification errors in the original regression (like should have been included)

    - **The Goldfeld-Quandt** Test: test allows us to test whether the variance between groups is equal.

- *Approaches to mitigate*
    - **Robust Standard Errors:** use HC0, HC1, HC3 or HC3
    - **Generalized Weighted Least Squares:**
        - **Known Form of Variance**
        - **Unknown form of variance**

Qualitative and Limited Dependent Variable Models

- Logit and probit models are appropriate when attempting to model a dichotomous dependent variable, e.g. yes/no, agree/disagree, like/dislike, etc. The problems with utilizing the familiar linear regression line are most easily understood visually
- Dealing with data where response variable is binary or a multi-level variable (choices – 1, 2, 3 etc.)
- Building the model
  - For binary model – choose a baseline group (y = 0) (y = 1 commuting by car, y = 0 commuting by bus)
  - Predictor x: (commuting time by bus – computing time by car)
  - Heteroskedasticity is present in the model as variance is not constant

Linear Probability Model

Problems with the linear probability model (LPM):

1. Heteroskedasticity: can be fixed by using the "robust" option in Stata. Not a big deal.

2. Possible to get y_hat < 0 or y_hat > 1. This makes no sense—you can't have a probability below 0 or above 1. This is a fundamental problem with the LPM that we can't patch up.

Logit and Probit Model

- The logit model uses something called the cumulative distribution function of the logistic distribution. The probit model uses something called the cumulative distribution function of the standard normal distribution to define f(∗). Both functions will take any number and rescale it to fall between 0 and 1. Hence, whatever α + βx equals, it can be transformed by the function to yield a predicted probability
- suggests using standard normal distribution to compute probabilities.
- Trade off: estimation is not as user friendly as the linear model.

Instrumental Variables

- Conversely, if x and e are correlated, then Cov(x, e) is not 0 and we can show that E(e|x) is not 0.
- Explanatory variables that are correlated with the error term are called endogenous variables
- In a perfect world looking at OLS, you want x causes y. If possible, uncover some degree of causation.
- If X causes Y, beta is the correctly estimated causal effect.
- Confounding variable – (mediating the observed relationship) something effecting both X and Y. Endogeneity. (X and errors are correlated, but X and e should not be correlated). This can be checked by plotting residuals against X values.
- For example, there might be relationship between, wrinkles and grey hair. However, age is mitigating the effect of of X and Y
- An instrument, Z induces change in the covariate in X, but does not affect Y
- Z must be correlated to X
- Z and Y should not be correlated – exclusion restriction. (How much of that relationship between X and Y is caused by something else)

How is this carried out.

Two Stage Least Squares

1) X ~ Z, save predicted values
2) Take original equation and replace (Y ~ Xhat)
   a. The beta identified in stage 2 will be the approximate effect

Partial Correlation

- R(X,Y) = language skills, toe size
- R(X,Y) = .40
- R(X,Z) = 0.55, R(Y,Z) = 0.65 (Z = age)
- R((X|Y), Z) = 0.07 – Z explains the relationship between language and toe size

Note on R2

- Unfortunately, R2 can be negative when based on IV estimates.
- Therefore, the use of measures like R2 outside the context of the least squares estimation should be avoided.

Specification Tests

- Hausman Tests to test which instrumental variables are best in our model
- The null hypothesis is H0: Cov(x, e) = 0 against the alternative
  - H1: Cov(x, e) is not equal to 0
- Step 1:
  - Run regressor and instrumental variables
  - Estimate residuals
- Step 2:
  - Use residuals in original model along with x (which was tested for endogeneity)
  - Estimate this “artificial regression” by least squares, and employ the usual t-test for the hypothesis of significance
    - H0 : d = 0 i.e., no correlation between x and e
    - H1 : d is not equal to 0 i.e., correlation between x and e

---

Simultaneous Equation Model

- When two equations are solved at the same time, there is bound to be endogeneity between the variables
- **Reduced Form Equations**
  - The reduced form of a model expresses each y variable only in terms of the exogenous variables, X.

The Failure of Least Squares Estimation

- The least squares estimator of parameters in a structural simultaneous equation is biased and inconsistent because of the correlation between the random error and the endogenous variables on the right-hand side of the equation

Identification Problem

- It is the absence of variables in one equation that are present in another equation that makes parameter estimation possible

Necessary Condition for Identification

- In a system of M simultaneous equations, which jointly determine the values of M endogenous variables, at least M- 1 variables must be absent from an equation for estimation of its parameters to be possible

2SLS for Simultaneous Equations

- Least squares estimation of the reduced-form equation for P and the calculation of its predicted value P_hat
- Least squares estimation of the structural equation in which the right-hand-side endogenous variable P is replaced by its predicted value P_hat

The properties of the two-stage least squares estimator are:
– The 2SLS estimator is a biased estimator, but it is consistent
– In large samples the 2SLS estimator is approximately normally distributed
– The variances and covariances of the 2SLS estimator are unknown in small samples, but for large samples we have expressions for them that we can use as approximations
– If you obtain 2SLS estimates by applying two least squares regressions using ordinary least squares regression software, the standard errors and t-values reported in the second regression are not correct for the 2SLS estimator

---

Panel Data Models

- A panel of data consists of a group of cross-sectional units (people, households, firms, states, countries) who are observed over time
- Denote the number of cross-sectional units (individuals) by N
- Denote the number of time periods in which we observe them as T
- Different ways of describing panel data sets:
    - Long and narrow:
        - "Long" describes the time dimension and "narrow" implies a relatively small number of cross sectional units
    - Short and wide
        - There are many individuals observed over a relatively short period of time
    - Long and wide
        - Both N and T are relatively large

Pooled Model

- A pooled model is one where the data on different individuals are simply pooled together with no provision for individual differences that might lead to different coefficients
- Applying pooled least squares in a way that ignores the panel nature of the data is restrictive in a number of ways
- The first unrealistic assumption that we consider is the lack of correlation between errors corresponding to the same individual

First Difference

- Not controlling for fixed effects will often lead to omitted variable bias when we are trying to make inferences
- For example, if we are trying to estimate the effect of an alcohol excise tax on traffic deaths, a naive analysis may inadvertently find that these taxes increase traffic deaths How? States with excise taxes on alcohol may have imposed such policies in response to high traffic deaths.
- However, if we can control for certain fixed or slow changing factors specific to each city (demographics, number of alcohol serving establishments, and even road infrastructure such as traffic lights or street lamps) we may be able to derive a more reliable estimate of the relationship!

Fixed Effects Model

- Individual intercepts are included to "control" for individual-specific, time-invariant characteristics.
- Two methods for fixed effects model
  - *Dummy Variable Estimator*
    - One way to estimate the model is to include an intercept dummy variable (indicator variable) for each individual
  - *Within Estimator*
    - Using the dummy variable approach is not feasible when N is large
    - Take original model and do a time average
    - Subtract original model and time average model
    - Usually we are most interested in the coefficients of the explanatory variables and not the individual intercept parameters

Note of First Difference vs Fixed Effects

- First differences is equivalent to fixed effects when t=2
- When the errors are serially uncorrelated, fixed effects is more efficient than first differencing (and the standard errors reported from fixed effects are valid)

Random Effects Model

- In the random effects model, we assume that all individual differences are captured by the intercept parameters
- But we also recognize that the individuals in our sample were randomly selected, and thus we treat the individual differences as random rather than fixed, as we did in the fixed-effects dummy variable model

- Random individual differences can be included in our model by specifying the intercept parameters to consist of a fixed part that represents the population average and random individual differences from the population average.

Comparing Fixed Effects and Random Effects Model

- If random effects are present, then the random effects estimator is preferred for several reasons:
  - The random effects estimator takes into account the random sampling process by which the data were obtained
- The random effects estimator permits us to estimate the effects of variables that are individually time-invariant
- The random effects estimator is a generalized least squares estimation procedure, and the fixed effects estimator is a least squares estimator
- If the random error is correlated with any of the right-hand-side explanatory variables in a random effects model, then the least squares and GLS estimators of the parameters are biased and inconsistent

Hausman Taylor Test

- For the Hausman test, the null hypothesis is that the preferred model is the Random Effects model vs. the alternative that the the Fixed Effects model is better.