Final

Adon Rosen

Date: 2020-05-08

Question 1

You are interviewing for a job and you will be asked to teach a course in linear models. To ascertain your competence and vision for the course, you are asked to give the 10 most important ideas, principles, or practical matters you have learned in this course. To help, make sure to not simply list 10 things - be sure to justify your ranking of the 10 most important things ... why you chose the ones you did and why did you rank them as you did? At least 2000 words for a sufficient answer

1. Independence of observations

2. Multivariate Normality

3. Homoskedastic error distribution

4. Model building techniques

5. Multicollinearity

6. Model building

7. Model Comparison

8. Outlier detection

9. Partial and semi-partial correlation

10. Mediation & Causal modeling

Like any good statistical test, the inferences that can be made are limited by the data provided by the experiment being performed. This is why the highest priority component when describing linear regression models includes the assumptions of linear models, and potential limitations of the inferential procedure, which are points one through five listed above, beginning with the independence of the observations. This point is made in most parametric statistical examinations, for instance the one group t-test assumes a random sample, normally distributed data, and equality of variance. Hidden within the assumption of random sampling is the importance of independence amongst the observations. In respect to linear regression, independence amongst the observations protects against autocorrelation being introduced into the predictions. Autocorrelation is a phenomenon where residual values display relationships with themselves. When an autocorrelation structure is present, this reduces standard error estimates inflating estimated significance of parameters, which is

something that can be easily protected against by employing simple random sampling. The next listed topic is that of multivariate normality, another issue that when dealt with improperly can reduce the ability to make statistically accurate tests.

Multivariate assumptions are usually tested visually using a qq-plot. A qq-plot displays the relationship between the standardized residuals and the theoretical quantiles. The theoretical quantiles are taken from a z-distribution allowing for quantiles to be obtained from a theoretically normal distribution. When multivariate normality is met, these distributions should be near identical and follow an identity line (x=y). When multivariate normality assumptions are met, it ensures that the statistical power of the underlying linear model is maximized. In contrast, when multivariate normality is not met it reduces the power of the test, explicitly, it reduces the ability to find an effect when one is present. Such patterns have received extensive exploration as they are detrimental when these effects are present, for instance Hopkins and Clay explored how variance in the samples' multivariate relationship effect inferences being made, suggesting alterations in type I and type II error rates (1963). This research has continued to be of significant interest to researchers and similar studies are still being performed well into the information era see for example a recent study exploring statistical procedures to report multivariate normality (Zhou & Shao, 2014). This proliferation of methods to analyze multivariate normality, and explore the impacts of these assumptions underscores the importance this carries into linear regression. Building off of these explorations, analysis of multivariate normality allows researchers to explore the underlying distribution of effects which can vary in shape.

Linear models can take many forms, the general assumption is that the relationship between the independent and dependent variables follows a linear trend. That is the relationship between the variables can be modeled using a straight line. However when deviations from these assumptions occur it alters the previous assumption of multivariate normality thus hindering the ability for inferences to be made statistically. This concern can be introduced in anticipated manners such as when the dependent variable is not normally distributed such as for binary outcomes, count outcomes, or highly skewed rare event prediction. However these trends can also occur when data takes unanticipated nonlinear form including quadratic and cubic shapes. For the former, these relationships can be modeled using a variety of link functions that work by imposing a distribution on the error term. Take for instance a binary outcome using a logit link, these transform the data, through exponentiation, such that the outcome of the model are now propabilities constrained between 0 and 1. Other such links exist, for instance a poisson distribution works for count occurances (i.e. when data take the form of frequencies of 0, 1, 2, or more events), a gamma function can be used to model exponential response data (i.e. a dependent variable that displays extreme skew). For the latter case, when data follow a quadratic or cubic trend, transformations can be made to the predictors, such as scaling and then squaring or cubing the predictors and then modeling these relationships again with an identity link function. These issues are usually flagged following the same methodology as for the multivariate assumption: using a qq plot, a final issue that needs to be satisfied for linear models which can further be analyzed by examining trends in the residuals is homoscedasticity.

Homoscedasticity exists in multivariate relationships when error terms display equal variance across the entire distribution. This point, while distinct from independence of observations, builds off of the same logic. When error terms do display heteroscedasticity, it means that variance-covariance relationships are not consistent across the entire random distribution. Put in another way this occurs when sizes of the residuals change as the values of the predicted outcomes changes. This is typically due to some potential autocorrelation introduced in the data. An example of these which potentially distinguishes this from independence of the observations is when multiple dependent observations are being studied from individuals within the breath of one study. An explicit sample is when questions from a psychological battery are collapsed into themselves (i.e. sum scores are used). This maintains independence of observations across the population, but introduces a plausible intraclass correlation among the repeated measures. This, however, does create a similar effect as briefly described, when data are not homoscedastic, standard error estimates are likely to be reduced, again, this will inflate the absolute statistical estimate inflating the relative importance of predictors when standard errors are reduced. It is however, interesting to note, this does not affect the estimation of the coefficient, it impacts the statistical weight given to a coefficient (i.e. the p-value). One final issue that has increased in frequency given the rise of "big data" that also masks the relative variable importance includes multicollinearity.

Big data has allowed for an expansion of both the "n" parameter (sample size), as well as the "p" parameter (predictors). While this increase in degrees of freedom from growing sample sizes is readily desirable and allows for more predictors in a model, it does introduce interpretation issues. By introducing more predictors it affords the opportunity for predictors which share a large proportion of variance between themselves and the dependent variable to be introduced. When such instances occur a model is said to have multicollinearity issues. When such present in the data it makes the inferential statistics much more difficult to work with: true relationships between the independent variable and dependent variable may flip, and standard error point estimates will be largely increased yielding more instances of type II errors. Multicollinearity can be assessed by exploring these two effects among other manners, is the expected relationship between the variables flipped? This can be observed by looking for large changes in the correlation, and partial correlation estimates. An inflation in the standard error estimates can be analyzed by several qualitative steps, for instance models that have very high coefficient of determinations but no significant predictors, and expected significant effects (those with large bivariate relationships) display insignificant relationships. When such instances occur it makes it extremely difficult to interpret relationships between variables. This can be troublesome when inferential statistics are sought after, or when attempting to build the most parsimonious model.

Inferential statistics when used in concert with linear modeling, allows researchers to distinguish the likelihood of a coefficient obtaining the magnitude of the observed effect when the null hypothesis is true. The canonical null hypothesis for testing coefficients is to state that the relationship is 0. This logic can be used to trim down models, and is employed in techniques such as stepwise model building. This encourages points six and seven, techniques to infer the model that is being fitted is one that yields a desirable fit. Stepwise modeling is performed by finding the relationship between a set of predictors, and a dependent variable with the strongest absolute magnitude. It is an automated procedure which allows for models to be built with little to no input from the researcher. Model building has become an ever growing demand as modeling transition from inferential to predictive practices. Predictive modeling involves creating a model which predicts the outcome to the best degree; whereas, inferential seeks to find significant relationships between the independent and dependent variables. The goal of model building is to find the absolute best fitting model given the current data. This goal has long alluded practitioners of statistics given the fact that a linear model, in its raw form, can not lose its predictive nature when non influential variables are introduced. However, these nuisance variables bias the predictive nature of the model when introduced to unseen data. Which is something stepwise regression model building seeks to perform in a relatively unguided manner. Inputs to stepwise regression algorithms typically include input and output p-values which guide inclusion and exclusion of variables. Newer advances to model building, driven predominantly by the expansion of big data, have grown to include more automatic one step algorithms for variable selection, an example includes the LASSO algorithm (Tibshirani, 1996) . Lasso involves tuning a shrinkage parameter, "lambda", in order to reduce coefficients of negligible importance to 0. This procedure mimics that of stepwise regression in that the end goal is to obtain the most parsimonious model via variable selection to encourage out-of-sample prediction estimates. While building the most parsimonious model is highly desirable, this procedure does not need to be devoid of inferential statistics, model comparison is something that can be performed in sync with finding the best fitting model.

Model comparison has been lauded as a fix to the limitations of null hypothesis significance testing. For instance Rodgers indicates that null hypothesis significance testing was "put on trial" for several factors including: 1) ability to reject the null given ample n 2) inability to support the alternative when the null is rejected 3) inability to provide support for the null 4) the backward nature of the null hypothesis significance test, and 5) significance does not imply practicality (Rodgers, 2010). Rodgers goes on to describe a solution to these limitations in model comparison. Model comparison involves shifting the focus away from the significance of predictors within the model, but comparing models at large for improvement in prediction. Such comparisons allow for more meta approaches to be taken when experiments are performed. For example, rather than detailing a t-statistic within a model, it requires researchers to question the current pedagogical model as the null hypothesis, and see if a modification to this practice yields a significant increase in the entire model fit. This obviates and appeases the criticisms that null hypothesis significance testing has been accused of being guilty. Such practices should be highly encouraged, however, this point is placed lower in the list as it requires good models to have been trained prior to comparing the models, underscoring the

"garbage in, garbage out" ideology. One further claim is made when assessing models, the importance of individual data points.

Outliers are of extreme interest for some disciplines of psychology, for example abnormal psychology seeks to study only outliers when compared to normative development. That is why dealing with outliers needs to be performed with extreme delicacy as psychological research is performed. Outliers for linear regression models can greatly influence model fit, and the direction of coefficients. Assessment of model fit has already been introduced as a critical topic from points two through five. Outliers are introduced as point seven as they are of extreme importance but must be assessed after the validity of the data, validity of the model, and also after the validity of the claims being made are already valid. However outliers can impact all of these claims which is why it is critical to explore the influence that individual observations may contribute to overall model fit. Explorations typically can be performed in qualitative manners (i.e. graphing the data and observing any "weird" trends) and quantitative measures (i.e. observation leverage). A more intuitive approach to observe relative influence is to calculate studentized residuals, the t value given the observation, and the residual distribution which provides an intuitive estimate of the probability of observing such an estimate. Larger magnitudes are more improbable. When such instances occur model fit may be grossly misestimated given the veracity of the observation. Researchers need to be extremely cautious about what to do when such observations are present, if the observation is valid then is it not worth preserving? If the observation is not valid, is there a confound or nuisance variable that is driving the effect that needs to be controlled for? Such effects can be controlled for by partialling out nuisance variables thus introducing the next topic.

Partial and semi-partial correlations are bivariate relationships whilst controlling for other covariates. This may seem oxymoronic at first glance, how can a relationship be bivariate if other variables are introduced? But the nature of the relationship is powerful and is the engine for multiple regression and ANCOVA modeling, and therefore all of the extensions for these modeling techniques. Briefly, these correlations represent the relationship when controlling for other covariates in either one side (semi-partial) or both sides (partial) of the relationship. This provides insights into relationships of variables when potential confounds are removed. An example of this effect is what is the relationship between academic knowledge and height after controlling for age. At first glance, academic knowledge and height may be shown to be positively and strongly correlated , however by removing the influence that age has on both of these relationships it will reduce and possibly remove the relationships. Such analyses are typically performed through structural models known as mediation analysis, which is the final point listed above.

A common place phrase is "correlation is not causation", and while there is truth to this saying, casual modeling is a technique which questions this platitude. The final point, that of mediation analyses, speaks to the power of regression modeling and is the last point on the list as it is critical that all of the previous points are strictly observed and protected against prior to mediation modeling being applied. Mediation modeling allows for causal statements to be inferred, is the relationship between A and Y driven by variable B? For instance, in the previous example, the relationship between height and academic knowledge can be thought to be fully mediated through thus allowing researchers to state that age "causes" academic knowledge to increase. Obviously this isn't true, but in the sense of the variables available to the example this statement can be made.

In summary the provided list ranges from experimental design, to model fit, to model inferences, to model building, and finally to claims that can be made from a model. This list attempts to build a hierarchy of importance when working begging with the limitations of linear models, to the inferences that can be made using linear models. In summary, it is critical for researchers to plot relationships among their data, and for well designed studies to be followed through in order for inferences to be made.

References

Hopkins, J. W., & Clay, P. P. F. (1963). Some Empirical Distributions of Bivariate T 2 and Homoscedasticity Criterion M Under Unequal Variance and Leptokurtosis. Journal of the American Statistical Association, 58(304), 1048–1053. https://doi.org/10.1080/01621459.1963.10480686

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. American Psychologist, 65(1), 1–12. https://doi.org/10.1037/a0018326

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological), 58(1), 267–288. JSTOR.

Zhou, M., & Shao, Y. (2014). A Powerful Test for Multivariate Normality. Journal of Applied Statistics, 41(2), 351–363. https://doi.org/10.1080/02664763.2013.839637

Question 2

a.

The numerator degrees of freedom is: 19

The denominator degrees of freedom is: 3

The numerator degrees of freedom reflects the values consumed when training the model, that is these are the values that are estimated from the sample to be population level effects. The denominator degrees of freedom reflect the elements that remain free in the data after training the model

b.

The n for this model was: 23

c.

The root MSE is: 10.28945

This value reflects the conditional standard deviation, or the standard deviation of the residual values.

d.

The dependent variables mean value is: 61.34783

e.

The best estimate for the PRE is 0.6210; this is the PRE when taking in to account the number of variables used to train the model.

f.

The value of the intercept term is 162.87590

This value reflects the predicted outcome when all independent variables are 0.

g.

The variable with the greatest predictive influence when accounting for covariates is x3

This can be observed in various ways, perhaps the most convincing is the t value, which is the largest t value for any predictor other than the intercept, furthermore the standardized coefficient is also the largest, suggesting the strongest conditioned relationship exists between the DV and x3.

h.

A parameter estimate reflects the change in the DV in one unit change of the IV.

For instance, using the raw coefficients, as x3 changes from 3 to 4; there is a decrease in y of -1.21032 units.

i.

A semi-partial correlation is the relationship between two variables when variance from one or more variables is removed from one of the variables of interest. When this relationship is squared it gives the coefficient of determination after controlling from any shared variance from one of the predictors.

A partial correlation describes the relationship between two variables when both variables have other confounds partialed out from the relationship.

The difference is that for a semi partial the IV's have confounds regressed out, whereas for the partial both the IV and DVs have effects partialed out.

A potential explanation for why these values have large differences in their values is that the relationship between X3 and Y is dependent on removing the covariance of Y with X1 and X2.

Question 3

Model DF: 3

Model Sum of Squares: 312.4490415

Model Mean Square: 104.1496805

Model F value: 11.4165597

Error DF: 18

Error Sum of Squares: 385.6685013

Corrected Total DF: 21

Root MSE: 10.2053751

Type 1 Anova Table

Prof DF: 1

Prof Type I SS: 252.2422458

Prof Mean Square: 252.2422458

Sex DF: 1

Sex Type I SS: 72.7990279

Sex Mean Square: 72.7990279

PROF∗SEX DF: 1

PROF∗SEX Type I SS: 0.6337764

PROF∗SEX Mean Square: 0.6337764

PROF∗SEX F Value: 0.0694726

Type II Anova Table

PROF DF: 1

SEX DF: 1

PROF∗SEX DF: 1

PROF∗SEX Type II SS: 0.6337719

PROF∗SEX F Value: 0.0694726

Type III Anova Table

PROF DF: 1

SEX DF: 1

PROF∗SEX DF: 1

PROF∗SEX Type III SS: 0.6337764

PROF∗SEX Mean Square: 0.6337764

PROF∗SEX F Value: 0.0694726