Stat 608 Midterm 2 Study Guide (First Draft)

Covers Handouts 4-7 and Homeworks 4 & 5

**STAT 608 Chapter 4: Weighted Least Squares:**

* In Chapter3 we transformed Y and/or X to overcome nonconstant error variance
* In this chapter we **use Weighted Least Squares to overcome nonconstant error variance**

The use of Weighted Least Squares:

* Yi is the *average or median* of ni observations.
* Yi is the sum of ni observations. ; Ex: =>
* =>
* Note: (1) the weights must be known (2) **WLS is sensitive to outliers.** If outliers present, use transformations. (3) If the model for the var isn’t clear, use transformations

Weighted Least Squares Model: . If large => i.e. point is really close to reg. line. ith point has more influence on our estimates. If small => is large. Points are far from the regression line and the ith point has less of an influence on the regression line.

* Covariance Matrix: Weight matrix

Derivation of WLS Coefficients: To estimate the coefs. we Minimize wrt. =>

Special Case: Multiple Measurements at each X: let B/c multiple obs. at each X we can estimate

* Problems: For k unique x values we must estimate k unique vars () and we don't take this extra variability into account when computing which makes our p-values too small and our CI's too narrow. Also, in general we don’t have multiple measurements at each value of X.

Case 2: X-continuous: Consider the LS reg mod. Multiply both sides by where the =

**STAT 608 Chapter 5: Multiple Linear Regression**

Two-Way Anova: When we have more than 1 categorical variable; ANCOVA: Mix of quant. and cat. vars; Including an interaction allows us to have model with separate intercepts and slopes depending on the value of one of the categorical variables.

Simpsons Paradox (confounding): The effect of one var. on the response depends heavily on the value of another variable

Table

Description automatically generatedInteractions: An **interaction** between two EVs exists when the effect of one EV (X1) on Y depends on the value of another EV (X2). NOTE: variables X1 and X2 may be **independent**, yet and interaction between the two may still exist.

Multiple regression: (1) Always conduct F-test: ;

(2) Test if each variable is significant: ->



* CI: **NOTE:** if we conduct p separate t-tests, our overall Type I error rate increases.



* When EVs are highly correlated, important EVs may become insignificant and/or the coef. to have the wrong sign

Model Reduction: (1) **Partial F-Test**: Suppose we have model: (can we drop k EVs) where k is the # of params removed from the full model.

* Under . (2) \*See example

R^2adj: . Beware that when used to compare models, is biased towards adding too many irrelevant predictor variables

The mean function might not be modeled correctly because: (1) We didn’t add variables we should have (simpsons paradox) (2) We added variables we shouldn’t have (correlated predictors) (3) We didn’t consider interactions (4) We didn’t consider polynomial terms.

**STAT 608 Chapter 6: Regression Diagnostics for Multiple Regression**

(1) Draw Scatterplots of the data: (i) Std. Resid. plots, (ii) marginal mod. plots (3) inv. resp. plots (4) plots for constant variance

(2) Identify leverage points and other outliers

(3) Assess relationships between EVs: (1) AVPs (2) VIF (3) (4) Forward, backward, stepwise, AIC, BIC selection (CHP 7)

What Makes a Valid Model? (1) Mean modeled appropriately (2) independent obs. (SRS or experiment) (3) Constant error var. (4) Scatterplots or resid. vs any EV or don’t have patterns (5) Predictor variables somewhat independent (6) Non-normality only matters for inference when n is small or PI's are desired)

Model Checking: When a valid model has been fit, plots of the residuals against X or will have (1) a random scatter of points (2) have constant var. as horizontal axis increases

* Any pattern to the residual plots indicate the model is NOT valid
* If then residual plots can help us determine the function g(). We can't use residual plots to tell us what part of the model has been misspecified unless those two conditions are met.

Leverage: recall that where is the leverage of the ith point. That is, each pred. y is a linear combination of all the values of Y in the data set. As with SLR, if any of the values is much larger than the others, it means that single obs. may be influencing the model much more than the others.

* High Leverage =>

Cooks D: High Influence (bad lev.) Di > , but also look for gaps in Cook's D and not just whether values exceed the suggested cut-off. For p EVs:

Marginal Model Plots: Assessing Mean: To find out if a SLR models E[Y|X] adequately we can fit a nonparametric estimator like loess and compare the curves

* If we have p ≥ 2 predictors it's not as straightforward. In general we'd need to make (p+1)-dimensional plots (not possible for p>2). Instead, we use **Marginal Model Plots.** Basically, we do the above for each of our p predictors, if all of them match then our model is a good fit.

Transformations (Box-Cox: Approach 1): (1) Transform all the X's to multivariate normal (2) Transform Y given X's so that the residuals are as normal as possible

Transformations (Using Logs for % Effects): ; for a 1 unit ↑ in X2 our model predicts a 100\* ∆ in Y. for a 1% ↑ in X1 … 100\* ∆ in Y

Multicollinearity: when we have strong correlation between X's. (1) Reg. Coefs. may have the wrong sign. (2) many of the X's may not be statistically significant when the overall F-test is highly significant. When the X's are highly correlated, the columns of **X** are close to being linear combinations of one another. \*\*\*The param. estimates become unstable, variance of the coef. est. become large and p-values become inflated.

Added Variable Plots: Goal: Find out whether the variable Z adds anything to the model after X has already been added.

* Interested in final model (2) **;** Baseline Model: (1)  and we want to add new EV. Z.
* AVP: Plot on the vertical axis the residuals from model (1) against the residuals from (3) . Resids. from (1) give the part of Y not predicted by X. Residuals from model (3) give the part of Z not predicted by X.
* (1) AVPs enable us to visually assess the additional effect of each predictor, ***after the others have been included in the model.*** (2) AVPs should display straight line relationships. If they don’t ***the model is misspecified***. (3) The slope from the AVP is the slope of the multiple linear regression model for that variable (i.e., (4) The scatter of the points in the AVP visually indicates which points are most influential in determining the estimate of .

Multicollinearity and VIF: Consider the multiple regression model . If The first fraction is called the jth VIF. We say our model has problems with multicollinearity if VIF > 5

Omitted Variables: **Spurious Correlation** is found when two variables being studied are related because both are related to a third variable currently omitted from the regression model. EX. # of ice cream cones sold and shark attacks are positively correlated. Weather is a **lurking variable.**

**STAT 608 Chapter 7: Variable Selection**

* Overspecified Model (contains irrelevant predictors): (1) MSE has few df, are inflated => larger p-values and wider confidence intervals b/c smaller
* Underspecified Model: (1) Reg. Coefs. and thus predictions are biased (2) arguably worse than overspecified model.

Forward, Backward and Stepwise Subsets: If there are m variables, there are 2m possible regression equations. If m is small enough, run all of them (all possible subsets)

* Backward Elimination: (1) All vars. Included in the model. The predictor with the largest insignificant p-value is eliminated. (2) The remaining m-1 variables are now in the model, again delete the variable with the largest insignificant p-value. (3) This process is repeated until all remaining variables are significant.
* Forward Selection: (1) No variables in the model. All m models with only one predictor are run. The predictor with the smallest p-value is entered into the model (as long as its significant), call this variable X1. (2) All models with X1 and one other Xi are run. Of the remaining predictors the one with the smallest (significant) p-value is added to the model. (3) This process is repeated until no more predictors are significant, given the others already in the model.
* Stepwise Subsets: (1) Choose , significance values to Enter and Remove preds. (2) Forward Step: No vars. In model. All models with 1 pred. are run and the pred. with the smallest p-value is added to the model as long as the p-value ≤ . Call this variable X1. (3) Forward Step: All models with X1 and one other Xi are run. Of the remaining preds. the one with the smallest p-value is entered as long as p-value ≤ . (4) Backwards Step: Check to see that the p-value of X1 is less than , if it is not, remove it. (5) Take another forward step, attempting to add a third variable. Continue taking backward and forward steps until adding an additional predictor does not yield a p-value below . **NOTE:** stepwise is a forward selection procedure, except variables can be removed.
* **MORE NOTES:** (1) These procedures only consider some of the predictors, so they do not necessarily find the model that fits the data the best among all possible subsets (i.e. they don’t consider interactions or polynomial terms). (2) They may not all produce the same final model, though they often do. (3) If covariance of all the parameters is 0, all three produce the same model. (4) These methods are prone to overfitting, but stiff criteria for adding or deleting variables can mitigate this problem.

Selection Criteria (1) R2-adj: Adding irrelevant predictors to a model will often increase R2 (never decreases when you add variables even if irrelevant.

* adjusts for the # of preds. in the model. **Choosing model that maximizes = choosing model that minimizes MSE**

Selection Criteria (2) AIC: based on maximum likelihood calculation. . **Only meant to compare sub-models to one another or to the full model, not models with different transformations.**

Selection Criteria (3) AICc: Corrects for bias when n small or p large compared to n (AIC tends to overfit; the penalty for model complexity isn’t strong enough. AICc converges to AIC as n increases.

Selection Criteria (4) BIC: **BIC favors simpler models than AIC** (when n ≥ 8, log(n) ≥ 2)

Selection Criteria (5) Mallows' Cp: Used unbiasedness as a criterion for choosing a model; assumes the full model is unbiased. . (1) Choose a model whose value is close to the # of params. in the model counting the intercept (err on the side of a smaller value of ). (2) Don't use to choose the full model; always equals p in that case. (3) If the full model contains a large # of insignificant variables, will be inflated. Then is NOT an appropriate method for choosing best model.

Comparison of Selection Criteria: (1) Typically use BIC. (2) R2-adj and tend toward over-fitting (3) is equivalent to AIC for linear models with normal errors (4) AIC chooses too complex models when n is large, BIC chooses too simple models when n is small. (5) Pro of AIC and AICc: They are "efficient." Asymptotically, the error in prediction form the model using AIC and AICc is no different from the error of the best model. NOT true of BIC. (6) Pro of BIC: The probability it selects the correct model is asymptotically 1. NOT true of AIC.

Comparison of Selection Procedures:

* All possible subsets: (1) if the # of preds. in the model is of fixed size p, all four criteria will choose the same model.
* Forward, Backward and Stepwise: using other information criteria (AIC,BIC) to select a model is equivalent to using p-values to add and remove variables; the difference is where the algorithm stops.

**Reminders:** (1) The regression coefs. obtained after variable selection are biased. (2) P-values from these models are generally much smaller than their true values.

LASSO: performs variable selection and parameter estimation simultaneously.

* When some variables have larger scales, they appear more important to this method; standardize (z-scores) or normalize (transform to [0,1] scale) to mitigate this effect.
* We can use our selection criteria above to choose the best LASSO model.
* When s is very large, LASSO is equivalent to LS. When s is small some of the coefficients are 0, effectively removing them from the model.