Applied Statistics II Study Guide

Topics

- Modeling Strategy
- Simple Linear Regression
    - Indicator Variables
- Multiple Linear Regression
    - Parallel Lines vs Interaction Models (Additive vs Multiplicative)
- Model Comparison
    - Full vs Reduced Model (SS F-Test)
    - Variable Selection
        - Forwards
        - Backwards
        - Stepwise
    - Identifying Multi-collinearity
    - Meaures of Fit
        - Mallow's Cp
        - Bayesian Information Criterion (BIC)
        - Akaike Information Criterion (AIC)
    - Identifying Influencers and Outliers
        - Leverage
        - Studentized Residuals
        - Cook's Distance
        - Partial Residuals
- Two-Way ANOVA
- Multi-factor Studies without Replication
- Time Series Analysis
    - Serial Correlation vs Autocorrelation
    - Partial Autocorrelation Function (PACF) plot
- Multivariate Analysis
    - Multivariate Responses
    - Hotelling's T
    - Two-Sample t^2-Test
    - Multivariate Analysis of Variance (MANOVA)
    - Traces
        - Pillai's
        - Roy's Greatest Root
        - Wilks' Test Statistic aka Wilks' Lambda
        - Lawley-Hotelling Statistic

Modeling Strategy

0. Define questions of interest.
1. Explore the data
2. Formulate an inferential model
3. Check the model

a. If appropriate, fit a richer model i.e. with interactions or curvature
b. examine residuals
c. See if extra terms can be dropped
d. If model not okay, GOTO 1

4. Infer the answers to the questions of interest using appropriate inferential tools
5. Presentation - communicate results

Generally want to start with a model which * can answer questions of interest * includes confounding variables * captures important relationships and be willing to make adjustments as you go

Definitions

Treatment: Specific values of an explanatory variable in a regression setting. Think ANOVA

Analysis of Covariance (ANACOVA): A Model with one continuous variable and one categorical variable. This is synonymous with a parallel lines model.

Saturated Model: Most complicated Model that can possibly be fit

block: Factor going in that we know will affect the response.

Multi-colinearity: Multiple predictors that are highly correlated with each other. Tends to inflate standard errors which drives t-ratios down and p-values up. Does not affect predicted values but kills inference.

Measurement Unit: What is the object I am taking a measurement on?

Sampling Unit: A unit that is randomly selected from a population that I am taking a measurement on.

Experimental Unit: A unit that is being experimented on.

Multivariate response: Instead of a single value, the outcome (or response) is a vector of values

Repeated Measure: A special kind of multivariate response where the same variable is measured several times on each sampling or experimental unit. Unlike Replicates which are multiple experimental at each combo of treatments

Longitudinal Studies: Response measured on each unit at multiple times. Units randomized to treatmeents at start

Crossover Experiments: Response measure on each unit after each treatment

Simple Linear Regression (SLR)

Definition: A single response as a linear function of a single explanatory variable

$\mu\{\mathrm{Y} \mid \mathrm{X}\} = \beta_0 + \beta_1\mathrm{X}$

Multiple Linear Regression (MLR)

Definition: A single response as a linear function of many explanatory variables (plus some assumptions)

An understanding of MLR can be extended to other model types - Mixed/Hierarchical Models - β's aren't fixed numbers but have distributions - Quantile Regression - modeling for a quantile (0-1) instead of a mean - Generalized Linear Models (GLM) - Possion, Logistic Regression: subpopulations aren't normal. Instead a parameter of their distribution is modeled (i.e. Lambda in a Poisson Distribution) - Lasso and Ridge Regression - estimate β's in a way that penalizes them for being big - Generalized Additive Models - non-linear left hand side of the equation plus penalization. Smooths response along locations

Coefficients

All Coefficients are linear. A model is considered linear if it can be written as a sum of terms:

β_1 * f(x)

where f(x) does not involve β's

examples

$\beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_3$

$\beta_0 + \beta_1 * X_1 + \beta_2 * X_1$ˆ2

$\beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_1 * X_2$

$\beta_0 + \beta_1 * X_1 + \beta_2 * \log(X_2)$

bad examples

$\beta_0 + \beta_1 * X_1 + \beta_2 * X_2$ˆ$\beta_3$

($\beta_0 + \beta_1 * X_1$) / ($\beta_2 * X_2$)

- Confidence Interval applies to **mean**
- Prediction Interval applies to **a single response**

Error = ϵ = Y - ($\beta_0 + \beta_1$X)

Parallel Lines Models

Definition: A MLR Model where the explanatory variables do not interact with each other but still have an effect on the response.

A parallel Lines Model is synonymous with Analysis of Covariance (ANACOVA). When toggling indicator variables in a model, the slope and intercept changes.

Interaction

Definition: The multiplicative effect multiple variables. Interaction is the effect of one variable as it varies across different levels of other variables. Two variables are said to interact if the effect of one variable on the mean response depends on the other variable.

Interpretation

Use the word **associated** when intepreting Observational studies. It implies that causation cannot be established

The **effect of an explanatory variable** is the change in mean response when the explanatory variable is increased by 1 unit, holding all other vars constant.

Often multiple parameterizations of the same model will be used to answer all the question of interest.

Occam's Razor: Sometimes the simplest explanation is the best.

Transformations

Given: 1. Scatterplot Funnels **and** observations are positive. 2. Data is right-skewed **and** observations are positive.

When: 1. Residuals are graphed 2. Summary Scatterplots are graphed

Then: Apply Log Transformation on X

Results: residuals should be more linear. The effect is now multiplicative on the median instead of an additive on the mean (when interpreting the result)

Extra Sum of Squares F-Test

Assumptions

1. Constant Spread
2. Response Variables normally distributed around mean
3. Observations are independent

H_0: $\beta_1 = 0$ (Reduced Model) H_A: β_1 != 0

How different is the Full model from the reduced model? Most F-tests in class are some form of an Extra Sum of Squares which compares a Full and Reduced model.

Extra Sum of Squares = ESS = SS_R - SS_F p = # of params

F-Statistic = (ESS / p) / $\hat{sigma}_F^2$

Other Statistics

Coefficient of Determination (R^2)

Proportion of variance in the response explained by explanatory variables. Adding variables always increases r^2, regardless of whether or not they are important.

Adjusted R^2 increases as additional variables explain more variance than expected by chance. It may penalize you for adding in a bad field.

R^2 = 1 - RSS / TotalSS p = # of estimated parameters Adj R^2 = 1 - ((RSS/n) - p) / (TotalSS/n - 1)

Indicators

X = 0 | 1

In $\beta_0 + \beta_1$X, β_1 is the difference between X = 1 and X = 0. More generally, each indicator variable represents the difference in mean between the indicatorand the baseline variable.

Guidelines

With K categories, you need K - 1 indicator variables. The category without an indicator variable **becomes** the baseline category.

Variable Notation - Indicator = CAPITALIZED - Continuous = CamelCase

A regression with a single indicator variable is a **Two-Sample T-Test**.

Outliers and Influencers

An observation is said to be influential if the fitted model depends unduly on its value. For example, removing it changes the estimate of parameters greatly, changes conclusions, or changes which terms are indluded in the model. Least squares estimators are not robust to outliers. Identify outliers early on so you don't end up tailoring a model to fit a few unusual observations.

Note: Outliers tend to be influential but not always.

Leverage h_i

Measures the distance of the observation from the average explanatory values (after taking correlation in account). Leverage values are the diagonal values of the Hat Matrix

- High leverage = unusual combination of explanatory values = possibility to be influential
- Typically on the extreme of X's for SLR - indicates a case occupies a position in the X-space that is not densely populated

Considered possibility to be influential if

h_i > 2p / n

where - p = # of parameters/coefficients in the model

Studentized Residuals $studres_i$

Residual divided by its expected variation with the expected variation being a mix of MSE and leverage.

- High Studentized Residual = observation far from the fitted line

Considered potential outlier if - abs(studres_i) > 2 (some people say 3 per the empirical rule)

Cook's Distance D_i

Effect on estimated parameters when the observation is dropped out. In other words, the effect on the regression model when the ith case is moved.

- High Cook's distance = influential on parameter estimates = changes regression estimates

Considered influential if - D_i > 1

Partial Residuals

Sometimes you want to look at the relationship between an explanatory variable and the response, after taking account the other vars. This relationship is always relative to an explanatory variable.

Residual = Obs - $\hat{mu}$

Partial = Obs - (rest of parameters. i.e b_0 + b_1 * explanatoryVar)

Two-Way ANOVA

Review on One-Way ANOVA

- 1 response variable
- 1 grouping variable with many levels

H_0: All means are the same (Reduced Model) H_A: At least one mean is different (Full Model)

Example

Full Model:

$\mu\{\%$ women | Judge$\}$ = JUDGE = $\beta_0 + \beta_1$A + β_2B + β_3C + β_4D + β_5E + β_6F

Reduced Model:

$\mu\{\%$ women | Judge$\}$ = JUDGE = β_0

Two-Way ANOVA

- 1 response variable
- 2 grouping variables with many levels

This is also a MLR with two categorical variables. Typically use **low** dimension data when designing a Two-Way ANOVA experiment.

Balanced Design

Each row/column or factor combinations have the same number of observations. Treatment combination means indicate a different conclusion when the data is balanced/unbalanced.

Parallel Lines / Additive Model

$\mu\{\}$ = FACTOR1 + FACTOR2 (I + J - 1) parameters

Non-additive model

$\mu\{\}$ = FACTOR1 + FACTOR2 + FACTOR1 * FACTOR2 (I * J) parameters

Note: There are a variety of ways the means can behave in a non-additive way

Steps

1. Fit the Model and check for significance
    - Numerator DF = p_full - p_reduced = (I - 1)(J - 1)
    - Denominator DF = n - p_full = n - I * J
    - F-Test with (I -1)(J - 1) and n - (I * J) DF
2. Check for transformations and outliers
3. Refine Model
4. Run a SS F-Test against previous Model and check for significance

If we fit a full model and there are no degrees of freedom left over, then we **cannot** fit the model. There will be no DF left to estimate the variance so none of our statistical tools will work.

If we want to make all pairwise comparisons between signs we should adjust for multiple comparisons Tukey-Kramer for pairwise comparisons

```r
library(agricolae)
HSD.test(lm.add.log, "Sign", console=T)
```

Type 1 SS: Sequential SS (variables enter the model in order listed) Type 3 SS: Marginal SS (What happens if variable in question is the last variable entering the model)

If the design is balanced, both are the same.

If proportion is in the response, then a transformation is needed

Multifactor Studies

Replication

Replicates: Multiple measurements at a specific combination of explanatory variable values. Replicates need to be independent applications of the same treatment.

Pseudo Replication: replication needs to be at the level of experimental unit (items randomly assigned to treatment).

When Designing Experiments

- If interactions are of interest, then replicate!
- If experimental units are expensive, you can sometimes gain more by reducing variability than increasing replicates
- Think of important sources of variation when designing experiments
- Continuous variables assume that you have some rate of change between the variables. Categorical variables do not make that assumption

With

- Allow a “model free” estimate of variation. i.e. Lack of Fit F-Tests for any model
- Allows the analyst to attempt to avoid overfitting a model

Without

- Assume some interactions don’t exist
- Treat numerical factors as continuous, not categorical

Without replicates we rely on our model being adequate & using the residuals to estimate variance - The deviation from each observation is used to estimate the variance of a group - If saturated model is fitted, there are no degrees of freedom left for estimation (its considered a perfect fit) and thus it is overfit

Variable Selection

Process of taking a large number of explanatory variables and selecting only a few to be in the regression model.

Strategy: find a subset of good models, then restrict attention to those that follow good practice.

Concepts * There are different approaches * Compare models with model selection criteria - AIC - BIC * Generally a few good models are considered, not just the “best model”

Problems * Can’t trust inference after variable selection - Why? Because we only include significant variables * Model selection criteria are subject to variability

Legitimate uses * Adjust for a large set of explanatory variables - Large # of variables to account for but not of direct interest. Do variable selection on just these variables * Prediction - Want a simple model purely to predict mean response. Do not care about interpreting results

Illegitimates Uses * Fishing for explanations - Which variables are important? Variable selection will not uncover some “true” model. The best model in one sample wont often be the best in another * Interpretation of included variables is dangerous - Inclusion depends on what other variables are being considered (particularly if they are correlated)

Stepwise Methods

Add or Remove one variable at a time. Only looks at a subset of all possible models.

Forward Selection

Start with an intercept term. Test each term for inclusion, include the “best” (smallest p-value from F-test) one. Repeat until no term passes our threshold

Backward Selection

Start with a Full Model. Test each term for deletion, delete the “worst” one (Biggest Value from F-Test). Repeat until no term fails our criteria.

Measures of fit

If number of params are the same, we prefer the model with small RSS If different, we want to balance smaller RSS with fewer parameters - RSS always gets smaller if you add another parameter

Common model selection criteria

Guiding Principles * Models shouldn’t include quadratic terms if they don’t include the linear one * Models shouldn’t include interaction terms if they don’t include the main effects * Lowest number of parameters for Mallow C__p and lowest BIC ideal * AIC/BIC contain some measure of variation in the model which assigns a penalty to worthless variables

Mallows’ C__p stat

C_p = (RSS / Var_full) - n + 2p

if C_p < P, then model potentially will have no problems with bias as long as the full model has none.

Bayesian Information Criterion (BIC)

BIC = n * log(RSS / n) + log(n) * (p + 1) * smaller the value, the better the fit (includes negative value. we want larger negative values) * R's BIC function uses a different formula

Akaike Information Criterion (AIC)

AIC = n * log(RSS / n) + 2 * (p + 1)

- smaller the better
- R's AIC function also uses a different formula

Multicollinearity

Multiple predictors that are highly correlated with each other. Tends to inflate standard errors which drives t-ratios down and p-values up. Does not affect predicted values but kills inference. When two variables are correlated, information is redundant which causes multi-collinearity.

When including quadratic terms in a model, center the quadratic terms to remove correlation with linear terms.

Centering a variable

- Polynomial - s^2 => (s - mean(s))^2
- Interaction - a*s => (a - mean(a))(s - mean(s))

Time Series Analysis

In Time Series Analysis, the assumption of independence is no longer true. Time Series techniques apply to both Temporal (Time) and Spatial (Region) Data.

Serial Correlation a.k.a Autocorrelation

Often when measurements are made at a adjacent points, there is a correlation

Can adjust for skewness by centering Run: Number of consecutive observations above or below the mean. Easier to visualize graphically

Positive Serial Correlation * an observation on one side of the mean tends to be followed by another observation on the same side of the mean * makes actual SE much larger

Negative serial correlation: * an observation on one side of the mean tends to be followed by another observation on the opposide side of the mean

Two Solutions 1. Adjust SE to be more appropriate 2. Filter variables to remove correlation

For both, you need to estimate the extent of the correlation (and make an assumption about its structure)

More advanced methods explicitly model the correlation * Time series analysis * Longitudinal data (Panel data in economics)

Adjusted SE on the sample average where r1 is the first serial correlation coeffcient. Appropriate under the autoregressive model of order 1, denoted as AR(1) * series is measured at equally spaced times * let v be the long run series mean, then mean{Y_t - v | past history} =…

$SE_{\bar{y}}$ = sqrt((1 + r1)/(1 - r1)) * s / sqrt(n)

** Used for two-sample T's or ANOVA where we need to account for serial correlation

Filter Variables

Review the slides. Paid attention in class

examine for serial correlation in the residuals, not the raw response

AR(1) => Autoregression Model with ORder 1 (i.e. Lag 1)

Testing for Serial Correlation

Is AR(1) Model adequate? - Primary tool is the PACF (Partial Autocorrelation function) plot

Large Sample Test

Z = r1 * sqrt(n)

If there is no serial correlation, Z has a normal distr. * only appropriate when n > 100

Runs Test

- Count how many runs there are and compare to how many we would expect by chance alone with no serial correlation
- Simple non-parametric test

Multivarite Analysis

Instead of a single value, the outcome (or response) is a vector of values.

A Repeated Measure is a special kind of multivariate response where the same variable is measured several times on each sampling or experimental unit. This is different from a replcates which are multiple responses measured at each experimental unit.

Strategies

1. Single univariate analysis on a summary of the multivariate response

- average, min, max, slope, etc

2. Separate univariate analyses on several summaries

- only if uncorrelated and don't need to adjust for making many comparisons (i.e. Bonferroni)

3. Multivariate analysis on several summaries

- Hotelling's T^2 - multivariate T-Test. mean vector of μ_1 == mean vector of μ_2

4. Treat subject (units) as a factor

- If multiple measurements on one individual are independent (i.e. chimpanzee & signs study)

Hotelling’s Tˆ2

Hotelling’s Tˆ2 extends the “two-sample t-test” to multiple “two-sample t-tests” on different response variables. In other words, the samples/groups/populations are the same in each, but the response variable is different.

For correlated response variables, the confidence region may be an ellipse which is hard to compute and present. An ellipse is the best description of our joint confidence, but hotelling’s adjusted conf int. guarantee at least 95% conf.

Hotelling’s Tˆ2 Adjustment adjusts the univariate confidence intervals to conservatively approximate the ellipse Hotelling’s Tˆ2 statistic provides a joint test for both parameters at once

Tˆ2 = (t__1ˆ2 + t__2ˆ2 - 2 * r * t__1 * t__2) / 1 - rˆ2

where r is the sample correlation between the two responses, and t__n is the sample t-statistic

Tˆ2 can be transformed into an F distribution

F = ((n__1 + n__2 - 3) / (2 * (n__1 + n__2 - 2))) * Tˆ2 - 2 and n__1 + n__2 - 3 DF

H_0: the difference in means between all group comparisons is 0

H_A: at least one mean group comparison is non-zero

Per typical C.I.

estimate +- multiplier * SE

multiplier = sqrt((2 * (n - 1) / (n - 2)) * qf(.95, 2, n- 2)

Box’s M-Test

Checks for the equality of covariance matrices for each group. **Sensitive to departures from normality**

Mardia’s Test

Checks for multivariate normality.

Traces

A Trace is a sum of the diagonal of a matrices. These are used in the calculation of multivariate test statistics

Comparing them

- Wilks’ Lambda - most widely used
- if k > 2, each test stat can take on different values & one test is not usually superior than the others in all circumstances
- Roy’s theta is not recommended in any situation unless all mus are collinear under standard MANOVA assumptions
- all tests robust to nonnormal pops exhibiting skewness or positive kurtosis.
- Pillai’s stat is superior to other when there is a heterogeneity of covariance matrices
- Wilks lambda can be used except when there is severe heterogeneity of covariance matrices
- Most MANOVA software programs calculate all 3 and reach the same conclusions
- When they don’t, dig deeper

Pillai’s

Multiplying eigenvalues of variance-covariance matrices

sum(1 - s, lambda_i / 1 + lambda_i) s = min(k - 1, p)

Roy’s Test

θ = lambda_i / 1 + lambda_i

lambda_i = largest eigenvalue of Eˆ-1 * H

Wilk’s Test Statistic (Wilk’s Lambda)

Lambda = det(E) / det(E + H)

Liklihood Ratio Test. Analagous to an F-Test

Intraclass Correlation

MSB - MSE / (MSBB + (n + 1) * MSE) n = # of observations per group

if intraclass correlation < 0.5, then obs are not strongly correlated