DS6372 Midterm Guide

**Key Conceptual Idea One**

Goals of Statistics

To explain or predict

Explanatory modeling

* Testing of causal theory. A set of underlying factors measured by variable X that are assumed to cause an underlying effect measured by variable Y
* Observational studies-Are used to test for causality. There are usually strong theoretical arguments to be made before drawing this type of conclusion. Replicated studies is another way to help “validate” a causal relationship.
* Controlled experiments and randomization, allow for causal inferences or direct inference to a larger population to be made.

Predictive modeling

* The process of applying a statistical model or data mining algorithm to data for the purpose of predicting new or future observations
* The scientific value of predictive modeling….just a mere tool?
  + Some argue, that without proper study design, random sampling etc, there is nothing “scientific” about predictive modeling.
  + I’d argue that PM has a damn fine suite of tools, just as with any tool, it is applied in the right settings and with the right goals.

[PDF on Explain or Predict](https://arxiv.org/pdf/1101.0891.pdf)

**The bias variance trade off**

**Key Conceptual Idea 2**

In statistics and machine learning, the **bias–variance tradeoff** is the property of a model that the variance of the parameter estimates across samples can be reduced by increasing the bias in the estimated parameters. The **bias–variance dilemma** or **bias–variance problem** is the conflict in trying to simultaneously minimize these two sources of error that prevent supervised learning algorithms from generalizing beyond their training set:

* The *bias error* is an error from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting).
* The *variance* is an error from sensitivity to small fluctuations in the training set. High variance can cause an algorithm to model the random noise in the training data, rather than the intended outputs (overfitting).

The **bias–variance decomposition** is a way of analyzing a learning algorithm's expected generalization error with respect to a particular problem as a sum of three terms, the bias, variance, and a quantity called the *irreducible error*, resulting from noise in the problem itself.

**For clean and large enough dataset:**

1. The greater the model complexity the lower the estimation error, the better the approximation.
2. The greater the model complexity the lower the bias
3. The greater the model complexity the **lower** the variance

For noisy dataset:

1. Greater model complexity doesn’t mean lower estimation error
2. The greater the model complexity the lower the bias
3. The greater the model complexity the **higher** the variance

[Data camp bias variance trade off](https://www.datacamp.com/community/tutorials/tutorial-ridge-lasso-elastic-net)

**Key Conceptual Idea 1:**

* + Can you explain what the ASE outputs in SAS and R are telling us in regards to bias variance trade off?
* Chart, line chart

  Description automatically generated
* You use this plot to detect overfitting of the training data. For many types of regression (including the linear regression that is used in the examples in this section), the fit statistic shown in this plot is the average square error (ASE) evaluated separately for each data role. The ASE decreases monotonically on the training data as parameters are added to a model. However, the ASE on test and validation data usually starts increasing when overfitting occurs.

**Key Conceptual Idea 2:**

* + What does it mean for a model to suffer from bias? Variance?
* The *bias error* is an error from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting).
* The *variance* is an error from sensitivity to small fluctuations in the training set. High variance can cause an algorithm to model the random noise in the training data, rather than the intended outputs (overfitting).

**Key Conceptual Idea 3:**

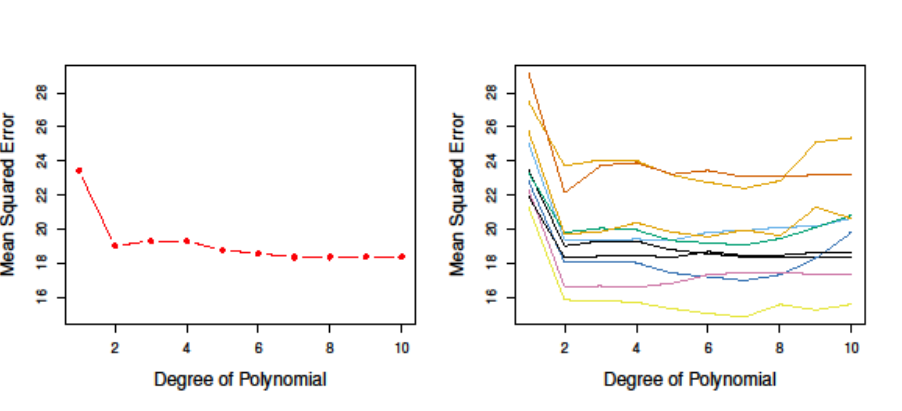
* + What sort of metrics and methods allow us to assess the bias-variance trade off when building models in the regression setting?
* Test ASE, CV ASE, Adj R^2, AIC, BIC (All penalize for many predictors)
* AIC look for small values, tends to select models with large p
* BIC more heavily penalizes models with large p
* ASE = sum(obs-pred)^2

**Key Conceptual Idea 4:**

* + How are training and testing error metrics affected by this tradeoff.
* A model with high variance may represent the data set accurately but could lead to overfitting to noisy or otherwise unrepresentative training data. In comparison, a model with high bias may underfit the training data due to a simpler model that overlooks regularities in the data.
* A model that exhibits small variance and high bias will underfit the target, while a model with high variance and little bias will overfit the target.

**Key Conceptual Idea 5:**

* + Advantages of CV over Train/Test splitting (see unit 5 discussion)
* Criticism of the train/test split approach alone
* Its possible to ”game the system” by considering multiple train/test splits, build your models, and then report the best test ASE using the split that produced it. Then reporting that your predictions will be within that type of error range on future data sets
* This approach biases the test ASE into looking too small
* It really helps no one. It can make you look bad if your model doesn’t perform as well in the future. People may make decisions using your model that produces error that they can’t live with in reality.
* It can impact model assessment, but is not much of an issue with model selection (see graph)
* If we do this to the nth degree (which we typically do), then the test data set, is really helping you train your model. You should start thinking about it that it is just an extension of the training data set, and the ASE that comes from it could be biased too low.
* In comes the validation set. You leave it out. Never look at it. Build your models using train/test. Make your final decision. Then ASSESS how well your model performs on the Validation set. It it is low and in the realistic range you are looking for, then you have validated your model as working as you intend.
* You can cheat the validation set as well of course by rinse a repeating the issue. You keep building models with the train/test and then compare the validation result and then based on all 3 you make your model choice. Then you haven’t validated that model yet. You’ll need another data set to do that.



* + A note on ASE
* Computing the test ASE metric can serve 2 functions:
* Model Selection: To assess the bias variance trade off of competing models (both within the same method and between different methods)
* Model Assessment: To use the ASE or Root ASE, as direct indication of how well your model is going to predict future values from the same population your training data is coming from

Key Conceptual Idea: Page 2, #1

* Cross Validation? What is its purpose? Explain like you’re teaching it, how K-fold CV works? What are some advantages of k-fold over a train/test split approach?
* When we’re building a machine learning model using some data, we often split our data into training and validation/test sets. The training set is used to train the model, and the validation/test set is used to validate it on data it has never seen before. The classic approach is to do a simple 80%-20% split, sometimes with different values like 70%-30% or 90%-10%. In cross-validation, we do more than one split. We can do 3, 5, 10 or any K number of splits. Those splits called Folds, and there are many strategies we can create these folds with.

Chart, diagram

Description automatically generated

**Multiple Linear Regression (Specifics)**

* Make sure you know what the ASE is. How is it computed. What’s the difference between train ASE and test ASE?
  + You use this plot to detect overfitting of the training data. For many types of regression (including the linear regression that is used in the examples in this section), the fit statistic shown in this plot is the average square error (ASE) evaluated separately for each data role. The ASE decreases monotonically on the training data as parameters are added to a model. However, the ASE on test and validation data usually starts increasing when overfitting occurs.
* Comfortable with model adequacy diagnostics
  + Adjusted R^2, AIC, BIC, ASE (CVPRESS, TEST ASE, TRAIN ASE)
  + Adjusted R^2, higher better
  + AIC,BIC,ASE, lower is better
* Comfortable with interpretation of regression coefficients
  + [Interpreting Regression Coefficient Estimates](http://www.cazaar.com/ta/econ113/interpreting-beta)
  + I will not give any exotic interpretations like log(x) or log(y)=b0+b1\*log(x). You will get standard stuff or a simple log(y) transformation and that is it.
* LASSO vs OLS. How does the penalty work?...What does large penalty values do to the regression model? What if the penalty is 0?
  + Lasso, or Least Absolute Shrinkage and Selection Operator, is quite similar conceptually to ridge regression. It also adds a penalty for non-zero coefficients, but unlike ridge regression which penalizes sum of squared coefficients (the so-called L2 penalty), lasso penalizes the sum of their absolute values (L1 penalty). As a result, for high values of λ, many coefficients are exactly zeroed under lasso, which is never the case in ridge regression.
  + Large penalty, the coefficients get squeezed closer to zero
  + It should be expected that the lasso and OLS fits diverge, particularly when collinearities exist; lasso is performing a feature selection (finding a parsimonious model), while OLS will fit whatever you ask of it (always producing a more complex model than the lasso).
* Importance of feature selection. When could feature selection/LASSO actually hurt you in the model building process? When would it be most beneficial?
  + Feature selection can hurt you when all predictors are beneficial.
  + It is desirable to reduce the number of input variables to both reduce the computational cost of modeling and, in some cases, to improve the performance of the model.
* Know assumptions and how to check them.
  + First, multiple linear regression requires the relationship between the independent and dependent variables to be linear.
  + Second, the multiple linear regression analysis requires that the errors between observed and predicted values (i.e., the residuals of the regression) should be normally distributed. This assumption may be checked by looking at a histogram or a Q-Q-Plot.  Normality can also be checked with a goodness of fit test (e.g., the Kolmogorov-Smirnov test), though this test must be conducted on the residuals themselves.
  + Third, multiple linear regression assumes that there is no multicollinearity in the data.  Multicollinearity occurs when the independent variables are too highly correlated with each other.
    - Multicollinearity may be checked multiple ways:
    - 1) Correlation matrix – When computing a matrix of Pearson’s bivariate correlations among all independent variables, the magnitude of the correlation coefficients should be less than .80.
    - 2) Variance Inflation Factor (VIF) – The VIFs of the linear regression indicate the degree that the variances in the regression estimates are increased due to multicollinearity. VIF values higher than 10 indicate that multicollinearity is a problem.
    - If multicollinearity is found in the data, one possible solution is to center the data.  To center the data, subtract the mean score from each observation for each independent variable. However, the simplest solution is to identify the variables causing multicollinearity issues (i.e., through correlations or VIF values) and removing those variables from the regression.
    - The last assumption of multiple linear regression is homoscedasticity.  A scatterplot of residuals versus predicted values is good way to check for homoscedasticity.  There should be no clear pattern in the distribution; if there is a cone-shaped pattern (as shown below), the data is heteroscedastic.

Shape, arrow

Description automatically generatedChart, scatter chart

Description automatically generated

* + - If the data are heteroscedastic, a non-linear data transformation or addition of a quadratic term might fix the problem.
  + Other concern: Outliers and Leverage- check through residual diagnostics
* MLR workflow

1. Summary statistics, scatterplots, boxplots, etc
   1. Outlier detection, transformation of nonlinear trends, removing variables that are unnecessary, redundant, or simply don’t make sense
2. Model building (Partly depends on the goal, Unit 2.6 & 2.7 is good for this)
   1. Including key predictors of interest and other potential variables that could contribute
   2. For many predictors use a selection approach
3. Assumption checking, residual diagnostics, checking model fit, multicollinearity
4. Possibly repeat 2-3 as needed to address issues
   * Note: If you reduce your model based on individual t-tests, those t-tests are only valid if the assumption are appropriate
5. Hypothesis testing
   * High level: Overall F-tests (Null: All coeffs are 0)
   * Low level: Individual t-tests
6. Interpretation of regression coefficients, CI’s

When predictors are uncorrelated

* + For a 1 unit increase in X there is a β increase in the response holding the other predictors fixed

When predictors are correlated

* + Its harder to come up with a nice statement because you can’t hold the other predictors fixed because when one goes up the others must go up or down because they are correlated
  + One way to describe this situation is just to refer that the particular predictor is statistically significant after adjusting for the other predictors in the model. Then provide a CI for the regression coefficient.

Basic Confidence interval statement

* + We are 95% confident that the true regression coefficient β for predictor X is between (lower limit, upper limit).

1. Providing predicted values and prediction intervals
   * To do this in SAS all you need to do is include the values you want to predict at the end of your data set with a missing value “.” for the response. Option in SAS to print the result is CLI.
2. Include any concerns or faults with the analysis as currently conducted.

* If your data set is big enough, always leave some data out for a validation set or conduct a k-fold cross validation to assess over fitting concerns.
* Important Note: It is possible to over fit a test set. If you are using the test set to help decide what model is best, you have technically not “validated” your model. You would need a new data set to to then see how your model performs. (Why I don’t like kaggle competitions. Promotes bad habit that is often lost to the competitor desperate to improve their score.

**Two Way ANOVA Specifics**

1. 1. Be comfortable with the general workflow of analysis.
   * + Assumptions?

* **Assumption #1:** Your **dependent variable** should be measured at the **continuous** level (i.e., they are **interval** or **ratio** variables). Examples of **continuous variables** include revision time (measured in hours), intelligence (measured using IQ score), exam performance (measured from 0 to 100), weight (measured in kg), and so forth.
* **Assumption #2:** Your **two independent variables** should each consist of **two or more categorical**, **independent groups**. Example independent variables that meet this criterion include gender (2 groups: male or female), ethnicity (3 groups: Caucasian, African American and Hispanic), profession (5 groups: surgeon, doctor, nurse, dentist, therapist), and so forth.
* **Assumption #3:** You should have **independence of observations**, which means that there is no relationship between the observations in each group or between the groups themselves. For example, there must be different participants in each group with no participant being in more than one group. This is more of a study design issue than something you would test for, but it is an important assumption of the two-way ANOVA. If your study fails this assumption, you will need to use another statistical test instead of the two-way ANOVA (e.g., a repeated measures design).
* **Assumption #4:** There should be **no significant outliers**. Outliers are data points within your data that do not follow the usual pattern (e.g., in a study of 100 students' IQ scores, where the mean score was 108 with only a small variation between students, one student had a score of 156, which is very unusual, and may even put her in the top 1% of IQ scores globally). The problem with outliers is that they can have a negative effect on the two-way ANOVA, reducing the accuracy of your results.
* **Assumption #5:** Your **dependent variable** should be **approximately normally distributed for each combination of the groups of the two independent variables**. Whilst this sounds a little tricky, it is easily tested for using SPSS Statistics. Also, when we talk about the two-way ANOVA only requiring approximately normal data, this is because it is quite "robust" to violations of normality, meaning the assumption can be a little violated and still provide valid results. You can test for normality using the Shapiro-Wilk test for normality, which is easily tested for using SPSS Statistics. In addition to showing you how to do this in our enhanced two-way ANOVA guide, we also explain what you can do if your data fails this assumption (i.e., if it fails it more than a little bit).
* **Assumption #6:** There needs to be **homogeneity of variances for each combination of the groups of the two independent variables**. Again, whilst this sounds a little tricky, you can easily test this assumption in SPSS Statistics using Levene’s test for homogeneity of variances. In our enhanced two-way ANOVA guide, we (a) show you how to perform Levene’s test for homogeneity of variances in SPSS Statistics, (b) explain some of the things you will need to consider when interpreting your data, and (c) present possible ways to continue with your analysis if your data fails to meet this assumption.
  + - What is the purpose of the F-tests?
    - ANOVA uses the F-test to determine whether the variability between group means is larger than the variability of the observations within the groups. If that ratio is sufficiently large, you can conclude that not all the means are equal.
    - What is the purpose of the contrasts / individual comparisons
    - Simple idea:
    - Once an overall F-test is conducted and found significant and if the factor of interest has more than two levels, we would like to know which factor levels are different
    - Contrasts allow us to do this

**Two Way ANOVA (Basic Flow)**

* What situation are we in?
  + One Factor of interest with a Block
    - Randomized Complete Block Design
  + Two Factors
    - Both Factors are of interest
* Plot data (visualize through mean profile plotting)
* Fit full saturated model with both factors and the interaction (nonadditive) term
* Diagnostics
  + Residuals
  + Normality, Independence, Constant Variance
  + Outliers (Don’t concern with leverage as these apply more to continuous explanatory variables)
* Testing
  + High level (ANOVA)
  + Contrasts
    - If nonaddtive (interaction) is significant then we reports differences of groups by factor levels
    - If not significant, we compare contrasts of the individual factor levels (very simple -> very complex)
    - Both situations call for multiple testing corrections

**Time Series Specifics**

* 1. What is the major pitfall of not appropriately identifying a data set that is time series?
     + When data is collected over time the independence assumption is violated.
     + What does this mean?
     + Estimates and standard errors are biased when modeling and conducting hypothesis tests
     + General methods treat your data as if you have more information than you actually have
     + 2 subjects with 5 measurements or 10 subjects with 1 measurement
     + If we don’t model correlation, can’t do a HT, p-value, t-test
  2. What is the major advantage of appropriately identifying a data set that is time series and applying a time series model?
     + Without complicating things too much
     + Correlated observations is tougher on hypothesis testing
     + It should be helpful to predict future values in time (they are correlated)
     + Forecasting a time series
     + Differences between forecasts from different models
     + Stationary - predictions regress towards a mean -
     + Predictions from nonstationary models do not
     + (explanatory variables, transformation to stationary (differencing, the “I” in ARIMA, etc)
  3. Comfortable with interpretation of ACF, PACF plots. Revisit the Durbin Watson test. We did not really cover it in class, but it may creep up in the midterm, revisit HW discussion.
     + Durbin Watson tests for serial correlation

A picture containing text

Description automatically generated

* 1. General understanding of a stationary time series possesses
     + How does this effect forecasts into the future?
     + Predictions regress towards a mean, nonstationary models don’t
  2. If a time series model include predictors, is the original time series stationary or not? What assumption does it break?
     + Explanatory variables are time series themselves (they should all be stationary)
     + Explanatory variables need to be lagged so that they can effectively be used in prediction of future outcomes in a natural way.
* Stationary TS
  1. Constant Mean
     + The time series may show cyclical behavior or general up and down behavior but it clearly is centered around a mean value
  2. Constant variance
     + Typically violated when low observations of the time series are less variable than higher observations
  3. Constant autocorrelations (correlation of lags doesn’t depend on where you are at in the time series)
     + This allows the estimation of the ACF
     + Wavelets and G-stationary models are alternatives when not true
  4. We have to have these assumptions met for ARIMA, AR, and MA models to be appropriately applied

ARIMA-Autoregressive integrated moving average. In statistics and econometrics, and in particular in time series analysis, an autoregressive integrated moving average model is a generalization of an autoregressive moving average model. Both of these models are fitted to time series data either to better understand the data or to predict future points in the series.

AR- Autoregressive. In statistics, econometrics and signal processing, an autoregressive model is a representation of a type of random process; as such, it is used to describe certain time-varying processes in nature, economics, etc.

MA- Moving Average. In time series analysis, the moving-average model, also known as moving-average process, is a common approach for modeling univariate time series. The moving-average model specifies that the output variable depends linearly on the current and various past values of a stochastic term.

Difference between AR and MA time series models?

This means that the **moving average**(**MA**) **model** does not uses the past forecasts to predict the future values whereas it uses the errors from the past forecasts. While, the **autoregressive model**(**AR**) uses the past forecasts to predict future values.