**Material Assess2ment Review – Summer 2022**

**Identify type of variable (Quantitative, Ordinal or Categorical)**

Quantitative: numeric variables that represent measurable quantities and can be expressed with numbers.

Categorical: qualitative variables that represent categories or groups.

Ordinal – logical ordering

Nominal: no logical ordering

**Dummy coding for Categorical (Effects Coding versus Reference Coding)**

Reference-level coding is similar to one-hot encoding except one of the levels of the attribute, called the **reference level**, is omitted

Effects coding: Effects coding is useful for obtaining a more general comparative interpretation when you have approximately equal sample sizes across each level of your categorical attribute.

Effects coding still requires a reference level,

Here, the reference level is left out in the sense that no comparison is made between it and the overall mean

reference level receives values of -1 rather than 0

**Difference between treating Ordinal Variable as Quantitative versus Categorical**

Levels are given values if treated Quant

* Optimal scaling( ord variables need not be equally spaced levels in term of target)
* Target level coding- using information from response variable

Levels become dummy variables if treated Cataeg

**Different measures of spread (impact of skewness or outliers on these measures**)

**Centre**

Mean – Abf value, affected by outliers

Median- middle value, 50Th percentile, unaffected by outliers

Mode – most frequent(categorical/ nominal)

**Location**

First quartile- 25th percentile; Median – 5oth, 3rd – 75TH ( p% of data is below it)

**Spread**

Range: max-min;IQR: 3rdquartile-firstquartile ( middle 50%)

Var/standard deviation- dispersion of data around mean;n-1 makes it unbiased

use standard deviation when you have mean but not median

**Shape**

Modality- unimodal or bimodal;Skew- Right/left;Kurtosis – leptokurtic(tails) +ve , platykurtic(middle) -ve, normal = 3

**Characteristics of a Normal distribution**

* Symmetric, asymptotic x axis, unimodal, bell shaped, mean= median = mode, kurtosis =3, skew=0. Fully defined by mean and standard deviation

**Empirical Rule** (68-95-99) 1,2,3 SD

**When graphs should be used for distributions/relationships**

(for example, histograms illustrate the distribution of a quantitative variable;

barplots illustrate distribution of a categorical variable;

boxplots can illustrate relationships between a categorical and quantitative variable)

**QQ plots to assess normality (identify if there is skewness or kurtosis in a QQ plot).**

A normal probability plot graphs the sorted data values against the values that one would expect if the same number of observations came from a theoretical normal distribution.

**Compare observed quantiles against theoretrical quantiles of a normal distribution with the same mean and variance.**The resulting image would look close to a straight line if the data was generated by a normal distribution. Strong deviations from a straight line indicate that the data distribution is not normal.

A diagram of a graph

Description automatically generated

**What values are used to create a boxplot**

Distribution- skewness- spread

Needs a factor(categorical) and continuous

**How to identify potential outliers**

For symmetric/normal – more than 3 s.d from mean

Less than 1.5IQR below Quartile 1

More than 1.5IQR above Quar 3

**Statistics versus Parameters and Sample versus Population**

Sample – reprsenatitive of population and calculate sample statistics- estimate of pop parameter counter parts

Paramaeter – population characteristics

**Central Limit Theorem**

* If sample size is large enough, distribution of sample means is approx. normal, irrespective of pop dist shape

**When do sample means follow a normal distribution?**

Pop – normal dist, sample size atleast 50

**Standard error versus Standard deviation**

Standard error – measures of estimate variability of sample means( if u were to resample data and calculate new sample average many times, how much variability you might expect in ur results)(s/sqrt(n)

Stand deviation – of samples, measure of variability in data(s)

**Sampling distribution of sample mean ( mean and s/sqrt(n))**

Confidence Interval = Xbar +- t(s/sqrt(n))

 The critical value (t) is chosen based on the desired confidence level and the degrees of freedom, which is typically (n - 1) for the sample mean.

T varies with sample size and confidence

t- no.of standard errors from mean

t dist used when standard deviation is unknow

**Confidence interval and terminology (confidence, MOE, standard error)**

% of confidence about true parameter

range of values within which u are x% confident that true population exists

A confidence interval is associated with a confidence level, a percentage, which indicates the strength of our confidence that the interval created actually captured the true parameter.

MOE

Range within which true pop parameter lies with certain level of confidence

It quantifies the uncertainty around the point estimate (e.g., sample mean) and is typically calculated as half of the width of the confidence interval.

**Interpret Confidence Intervals**

Provides info about variability around statistic

Its confidence not probability

important nuance to remember that the population mean is a fixed number. The source of randomness in our estimation is our sample.

When we construct a 95% confidence interval, we are claiming that, upon repetition of` the sampling and interval calculation process, we expect 95% of our created intervals to contain the population mean

**Hypothesis testing**

**Terminology and definitions (Null Hypothesis, Alternative Hypothesis, Type I errors, Type II errors, significance level, test statistic, p-value, how to state conclusions**

**Identify null and alternative hypotheses for different situations**

**When to reject the null hypothesis**

If true pop value is sig different from assumed value

**Null** – initial assumption , Ho

We analyze data and determine whether our observations are likely enough( beyond some threshold we set before hand) given assumption of null hypothesis

If we determine that observations are unlikely- we reject null hypo, in favor of alternative hypothesis

Alpha-

threshold of sig level that we used to determine how much evidence is required to reject null hyp is a prop alpha. how often we are incorrectly rejecting null hypothesis ( type 1 error)

* We make a null hypothesis and assume its true
* As P Value gets smaller and smaller - more evidence for alternate hypothesis
* Trying to get sufficient evidence to reject null hypothesis

Pvalue : The p-value is the area under the null distribution that represents the probability that we observed something as extreme or more extreme than we did (assuming the truth of the null hypothesis).

If our p-value is less than our confidence level, α , we have enough evidence to reject the null hypothesis in favor of the alternative.

A type I error (false-positive) occurs if an investigator rejects a null hypothesis that is actually true in the population; a type II error (false-negative) occurs if the investigator fails to reject a null hypothesis that is actually false in the population.

We have evidence to support the alternative hypothesis, and the result is considered statistically significant.

"failing to reject" the null hypothesis does not mean we accept it as true; it merely means we don't have enough evidence to reject it based on the data at hand.

**Test statistic:**

A test statistic is a numerical value calculated from sample data in hypothesis testing. It measures how far the sample data deviates from what would be expected under the null hypothesis. The choice of test statistic depends on the type of hypothesis test being performed and the nature of the data. Common examples include t-statistic, z-score, chi-square statistic, and F-statistic.

**Perform one sample (state hypotheses, know assumptions, make conclusions, read output from R, test statistic, calculate p-value)**

To test mean of a continuous variable equal to assumed value

H0 u = u0 , students t statistic(average score of students in the exam is equal to 75.)

Ass: we assume that the sample is a simple random sample, the data is normally distributed, and the population standard deviation is unknown.

DOF= n-1,P, test, CI 95%, sample estimate

**Relationship between a two-tailed hypothesis test and confidence intervals**

Inside CI- fail to reject H0 and outside reject.

**Perform two sample (state hypotheses, know assumptions, make conclusions, read output from R, test statistic, calculate p-value, when to use unequal variance t-test versus pooled t-test….how to test for this)**

If we have a hypothesis about a difference in the means of two groups of observations, a two-sample t-test can tell us whether that difference is statistically significant.

Ass- independent obs, each group normally dist, equal variance for each group

 1st1  assumption is not easily checked unless the data is generated over time (time-series) and is instead generally implied by careful data collect and the application of domain expertise

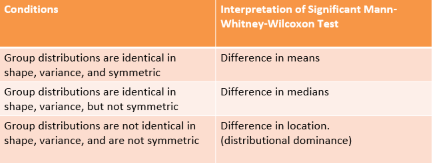
If the 2nd  assumption is not met, one must opt for a nonparametric test like the Mann-Whitney-U test (also called the Mann–Whitney–Wilcoxon or the Wilcoxon rank-sum test).

* QQplot or Shapiro-Wilk test

3rd3  assumption of equal variances is not met, we simply add the option var.equal=F to the t.test() function to use the Welch or Satterthwaite approximation to degrees of freedom (it’s becoming increasingly common for practitioners to use this option even when variances are equal).

* F test ( pop are normally dist; Ho equal var

**Perform Wilcoxon Rank sum (what it is and what is it testing, read output and make conclusion)**

****

When norm fails – conclusion will be diff in medians

**Explanatory versus Predictive Modeling**

**A white background with black text

Description automatically generated**

**Expl- (description, is xi related to yi)**

In explanatory modeling, you foremost want a model that is simple to interpret and doesn’t have too many input variables. It’s common to avoid many polynomial and interaction terms for explanatory models. While the error rates on holdout data will still be useful reporting metrics for explanatory models, it will be more important to craft the model for ease of interpretation.

**Pred;** xi predict yi, prediction.

you are most interested in how much error your model has on holdout data, that is, validation or test data.

**What is Honest Assessment?**

Honest assessment in modeling involves unbiased evaluation of model performance, considering data quality, validation techniques, and potential limitations for reliable and meaningful prediction

If model captures nuances of data they are built(train), when these patterns don’t hold on test and validation model performance suffers. We say model doesn’t generalize wello **rmodel is overfit**

* We want to make sure model generalizable
* Not just good models of training , can predict well on out of sample data
* 2/3 train, 1/3 test (50-40-20,70-20-10,not enough cross validation)

**What the training data, validation data and test data is used for (and when it should be used)**

Training data is used to build the model's parameters and learn patterns from the data; validation data is used to tune hyperparameters and assess model performance during development; test data is used to evaluate the final model's generalization to new, unseen data.

**When to create them**

After DEA, ( dist,outliers,missing values). Before checking any relations split into train, validation and test

**EDA (univariate and bivariate)** and graphs to do this ( we shouldn’t declare a statistical relation without a formal test, this is preliminary analysis)

**Asscociation:** exp value of one variable changes at diff levels of another variable.

**Linear association**: when general shape of a scatter plot of 2 variablespotrays a straight line

Continuous vs Categorical Variables: Bar Graphs, Histograms, and side by side Box-Plots

Bar chart only reporting

For Explorartory purpose- freq histogram/ density histogram

For further exploration we can create box plots

**ANOVA (what it is, what it is testing, assumptions, hypotheses, make conclusions, number of predictions, setup…what type of response variable and what type of explanatory variable)**

One-way ANOVA aims to determine whether there is a difference in the mean of a continuous attribute across levels of a categorical attribute.ANOVA with a binary input variable is mathematically identical to the two-sample t-test

Anova – F test

Ass:

1. The observations are independent
2. The model residuals are normally distributed
3. The variances for each group are equal ( T- pooled Var, F- welch Anova)

Hyp: means are equal,means are not equal(atleast)

One hot coding cause multicollinearity- so we need to have k-1 variables

* β0 0 represents the mean of reference group, group d.
* βa,βb,βc , ,  all represent the difference in the respective group means compared to the reference level. Positive values thus reflect a group mean that is higher than the reference group, and negative values reflect a group mean lower than the reference group.
* ε  is called the error

no. of predictors= npoof levels

no.of Beta coeff without Bo = levels-1

lm() or aov()- that lm() will also provide the user with the coefficient of determinination, R2 2, which tells you how much of the variation in y  is accounted for by your categorical input.

Levene test- to test variance

Ho Var are equal( assumes normality)

Fligner test doesn’t require normality

**Welch’s ANOVA (when to use it) :** when equal var fails( similar to 2 sample t test)

**Kruskal-Wallis (when to use it):** non para test, when normality fails.

Dist of xx is diff for diff levels of YY ( when var also fails Kruskal wallis)

**Tukey-Kramer and Dunnett’s post hoc test (why do we do them, what are they used for…also understand experiment-wise error rate and why important in multiple hypothesis testing)**

* ANova we know atleast one group is diff, to know which group is diff we do post hoc
* To compare pairwise ( 2 sample t tests); multiple hypothesis compounds error
* No.of comaprisions is (n!/(n-2)!2!)
* Experiment wise error rate 1-(1-alpha)^n

To control exp wise error rate - **Tukey-Kramer and Dunnett’s**

**Tukey-Kramer –** all pairwise comparison , eerate alpaha (overly conservative**)**

**Dunnets –** designed to correct L-1 tests ( L no.of levels in control). Designed interested in comparison to a contraol/placebo**.**

fewer than all comp , eerate <alpaha

**Pearson’s correlation (properties and hypothesis test…how does this test relate to test of slope in SLR) Correlation versus Causation versus slope in SLR**

To formally test the (linear) relationship between two continuous attributes, we introduce Pearson correlation, commonly referred to as simply correlation. Correlation is a number between -1 and 1 which measures the strength of a linear relationship between two continuous attributes.

designed to detect linear associations between variables. Even when a correlation between two variables is 0, the two variables may still have a very clear association, whether it be quadratic, cyclical, or some other nonlinear pattern of association.

* Hypo H0 is exactly zero. It doesn’t mean relation is practically sig
* Can be skewed by outliers
* 2 sample t test ( 4 assunptions)
* Pairwise correlations are arranged in a correlation matrix
* Corr doesn’t implay causation
* 2 variables can have same same corr but diff linear equation; its is not = to slope of SLR

**SLR (write out equation, make predictions, calculate residuals, interpret slope and intercept, interpret confidence intervals, do hypothesis test and interpret, assumptions)**

* variables share a linear relationship, the next question is natural: what is that relationship? How much,on average, should we expect one variable to change as the other changes by a single unit?
* Simple linear regression answers this question by creating a linear equation that best represents the relationship in the sense that it minimizes the squared error between the observed data and the model predictions (i.e. the sum of the squared residuals).
* *Ordinary Least Squares seeks to minimize the sum of squared residuals or sum of squared error*

a simple linear regression serves two purposes:

1. to predict the expected value of y  for each value of x  and
2. to explain *how* y  is expected to change for a unit change in x

As x increases by 1 unit, slope tells how much y increase

Pop parameters are unknown true relations.

Sums of squares of regression + sums of squares for error = sums squares of total.

|  |  |  |  |
| --- | --- | --- | --- |
| source | degress of freedom | Sumof squares | Meansquares |
| Regression | Parameter ( degree of freedom)p-1 ( simple LR 1) | SSR | SSR/2= mSR |
| Error | n-p (for SLR n-2) | SSE | SSE/(n-2) = mSE |

**For SLR- global F test and parameter tstet and test of pearson coeff are equivalent**

**H0 beta 1 = 0**

**4 Assump :**

**Resid – normally dist, equal variance, independent, linearity of mean**

**A white paper with black text

Description automatically generated**

**-----------------------------------------------------------------------------------------------------------------**

**Two-Way ANOVA (and n-way ANOVA)**

One-Way ANOVA model, we used a single categorical predictor variable with k  levels to predict our continuous target variable. Now we will generalize this model to include n  categorical variables that each have different numbers of levels

The F values are calculated the exact same way as described before with the mean square for each variable divided by the mean square error.

same F test approach is also used, just for each one of the variables. Each variable’s test has a null hypothesis assuming all categories have the same mean. The alternative for each test is that at least one category’s mean is different.

**Interactions:I** f the relationship between a predictor and target variable changed depending on the value of another predictor variable.with the interaction effect, (αβ)ij( ) , as the multiplication of the two variables involved in the interaction. Interactions can occur between more than two variables as well. Interactions are good to evaluate as they can mask the effects of individual variables. If our interaction was significant (say a 0.02 p-value instead) then we would keep it in our model, but here we would remove the interaction term from our model and rerun the analysis.

**Sliced ANOVA** : With interactions present in ANOVA models, post-hoc tests might get overwhelming in trying to find where differences exist. To help guide the exploration of post-hoc tests with interactions, we can do slicing.

Performs F test for means for one variable with level of another variable

Assumptions: indept ons, equal Var, norm ( both for errors)

When norm and var not met: Unfortunately, the Kruskal-Wallis approach is not applicable to n -Way ANOVA where n>1 >1. These approaches would need more non-parametric versions of multiple regression

2 types of data analysis**. Obs/retrospective, controlled exp**

**blocking ANOVA** model is the exact same as the Two-Way ANOVA model. The variable that identifies which sector (block) an observation is in serves as another variable in the mode

**Differences in MLR and SLR**

This model has the predictor variables x1,x2,...,xk 1, 2,...,  trying to either explain or predict the target variable y . The intercept, β0 0, still gives the expected value of y , when all of the predictor variables take a value of 0. With the addition of multiple predictors, the interpretation of the slope coefficients change slightly. The slopes, β1,β2,...,βk 1, 2,..., , give the expected change in y  for a one unit change in the respective predictor variable, holding all other predictor variables constant. The random error term, ε , is the error between our predicted value, ^y=^β0+^β1x1+⋯+^βkxk ^= ^0+ ^1 1+⋯+ ^ , and our actual value of y .

**What does LINEAR mean in linear regression**

linear in linear regression doesn’t have to deal with the visualization of the fitted plane (or line in two dimensions), but instead refers to the linear combination of variables. A linear combination is an expression constructed from a set of terms by multiplying each term by a constant and adding the results.

**Global F-test** : Global testing entire model - all variables at a time. Multiple t tests for multiple parameters. To test whether model is adequate to predict y.

No of variables are useful in predicting y

At least one variable is useful in predicting y

A math equations and numbers

Description automatically generated

**F-distribution :** bounded by 0, right skewed, num & Den DOF

Assumptions for MLR (adding no perfect collinearity)

* Linear combination
* errors are normally distributed
* errors have constant variance
* observations are independent
* No perfect collinearity- predict var are a perf linear relationship of each other

**Adjusted R^2**

calculation of **R^2** in a multiple linear regression is that the addition of any variable (useful or not) will never make the **R^2** decrease. In fact, it typically increases even with the addition of a useless variable. The reason is rather intuitive. When adding information to a regression model, your predictions can only get better, not worse. If a new predictor variable has no impact on the target variable, then the predictions can not get any worse than what they already were before the addition of the useless variable. Therefore, the SSE would never increase and R^2 would never decreae

Rsquare is amount of variation you can't explain .I'm multiplying everything I can't explain with (SSE/TSS)

If k is 1 As I add variables k value increases - everything I can't explain increases (n-1)/(nmodel -(k+1))

But SSE decereases when the added varaible is good

Adjusted Rsquare can go up or down as we add variables,Based on good or varibele.Based on math adjRsq is always small or equal to RsquareIf we reduce the number of variables and have same predictive power- we use model with less variables.Adjusted R square for comparison;Normal Rsquare for interpreting

**Categorical Predictor Variables** The interpretation of the coefficients (β ’s) of these variables in a regression model depend on the specific coding used. The predictions from the model, however, will remain the same regardless of the specific coding that is used.

Dummy Beta 1: the coefficient on each dummy variable is the average difference between that category and the reference category

**Effects coding :** the coefficient on each dummy variable is the average difference between that category and the average price across all homes ( more than one variable)\*\*\*\*\*\*\*\*8

Best Linear Unbiased Estimator

What is an unbiased estimator

What is meant by best

Regularized regression

Ridge, LASSO, Elastic Net

Compare and contrast

Cross-validation

**AIC, BIC (what they are, penalties ….and how affected by number of variables)**

Info criteria: to slect variables

Selection algorithm; evaluate models based on selection critereia

Stepwise selction – F,B,Step ; All reg selection( Rsq, adj Rsq, Mallow’s Cp)

CAUTION: you should NEVER just use the final model created from an automatic procedure! Always explore your data (both automatically selected and excluded variables), and use domain knowledge, diagnostics, and critical thought to decide on your final model.

AIC- large sample, leave one out cross validation ( P^2 – penalty)

BIC – it favours smaller models with fewer variables, penalize complexity more( plogn- penalty)

Smaller AIC/BIC better model; p – no.of parameters, n sample size

**Forward selection, backward elimination and stepwise selection (what they are and how they work, issues with these procedures) p-values and sample size**

Forward – systematically adds variable to base model( intercept only)

Step: starts with intercept model , adds and delets variables

if we didn't check assumptions- like normality,collenearity,…do may have biases

The larger the sample size, the smaller the p-values will be and the more likely it is that you will end up seeing many “significant” p-values due to the relationship between p-values and sample siz

How to assess assumptions in MLR (Misspecified model, lack of constant variance, lack of normality, correlated error terms)

A close-up of a test

Description automatically generated

1,2 residual plot ( random,

a quadratic shape or curvilinear pattern to the residuals would indicate that one of our input variables has a nonlinear relationship to the response and transformations should be made to that input accordingly.

Model Heirarachy : when high order terms- include lower order terms as well

When we have higher order, we can’t have interpretability

A white background with black text

Description automatically generated

**Lack of constant variance- spearman or residual plot**

**Spearmen will not detect if its non linear**

**A white background with black text

Description automatically generated**

How to identify potential outliers and influential points (what measures are used for each)

Multicollinearity (what it is and how to identify it)

-------------------------------------------------------------------------------------------------------------------------------

Nominal vs. Ordinal categorical variables

Association of categorical variables

Pearson’s Chi-square

Likelihood Ratio Chi-square

Mantel-Haenzsel Chi-square

Chi-square distribution

Assumptions to tests of association

Measures of association

Odds ratios, odds vs probability

Cramer’s V

Spearman’s correlation

Logistic regression for any type of categorical target variable

Why is least squares bad?

Logit link function

Interpretation on coefficients

Concordant, discordant, tied pairs

Variable selection