|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum Sq** | **Mean Sq** | **F Value** | B1 Is based onHo : B1 = ?? | **B0= Intercept** | **\*low variance means that data points are close to the mean and close together where as high variance means data points are far from mean and eachother** | **\*\* 95% CI means – “We are 95% sure that the**  **interval contents are true”** |
| **Model** | **1 = DF** | **MSS** |  |  | X bar = Mean  Xi  = X given | **\*\* Adding any addition predictor to model will make SSE decrease** | Dipat = (1 = Yes , 0 = no , 3 = maybe …)  **When you do not have dipat you must add error :** | **}1/2 (EXP OF WHOLE EQUATION)** |
| **Error** |  | **SSE =RSS** | **E(MSE) =** |  | **Will always go up when predictors are added – doesn’t mean predictors are useful** | **n = Observations**  **n-1 = total obs – 1** | **Linear Regression Assumptions**  **Linearity = b + xy**  **Uncorrelated error**  **Constant Variance or Errors**  **Normal Distribution of errors** | **}1/2 (EXP OF WHOLE EQUATION)** |
| **Corrected Total** |  | **SST** | **MST** |  | **\*\* if n increases the impact of a individual outlier decreases** |  |  | **SSE will always decrease when more predictors are added,**  **even when no significance to Y -- i.e. adding more predictors**  **will always reduce error term but doesn’t make it a**  **better model** |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Log (C \* D) = Log(C) + Log(D)** | **Log b(x) = y**  **X =** |  | **is used for multivariate (many X) – it has penelties for adding useless predictors so its more useful then general R2** | Parameter estimate in multivariate compare one another. Example: Catholic has a Parameter Est. of 12.9 for high cholesterol the interp. would be: On average cholesterol increases 12.9 units in areas that are mostly catholic vs. areas where they are mostly non catholic | \*\*Multivariate – means you have more then one X factor \*\*  Multivariate **DF = P -1 (P = # of predictors)** opposed to number of obs | When P value < .05 then significant and you reject the null hypothesis. |

|  |  |  |  |
| --- | --- | --- | --- |
| **LOGISTIC REGRESSION**  **FORMULAS / INTERPERTATIONS**  **\*\* ALWAYS ASSOCIATE W S-CURVE (BECOMES LINEAR WHEN NAT LOG IS TAKEN)\*\*** | **The point estimates on Logistic Regression Model can be calculated to OR**    **1-OR = x(100) = % increase/decrease of that variable** | **Interpretation:**  **For Log Odds: When there is a 1 unit increase in X, the log odds will increase by**  **ODDS RATI/(ODDS)With a 1 unit increase in x (x+1) the odds with be multiplied by**  **For Linear Regression: with a 1 unit increase in X, Y will increase by**  **Linear W. Interactions: After controlling for Mother’s IQ, children whose mothers completed HS have IQs, which are 5.95 (Bcoeff ) points greater on avg, then children whose mothers did not complete HS. The difference appears to be significant, w P-Value of .0074** | **\*\*\* In your interpretation for odds ratio make sure your write it as: Those who are in the “less active” vs. “very active” group will have a 44% higher odds in contracting diabetes, after holding all other variables constant** |

|  |  |  |  |
| --- | --- | --- | --- |
| Probability =  **WILL ALWAYS = BETWEEN 0 - 1** | Make sure that you include the intercept and any coeff. That the question refers to. Remember that variables will carry a value (0 or 1) based on what category they are in i.e. “female =1 male =2”  \*\*CONDITONAL LOGISTICE REGESSION MODEL IS USED TO MATCH DATA | \*\* WHEN ADDING INTERACTION TERMS TO ORGINAL MODEL\*\*  \*\* When talking about interactions & catagory you must add them to equations |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Polynomial Regression is when the relationship between X and Y is curvilinear, a simple approach is to add powers of X to the regression function  -Useful when the true relationship between X & Y is a polynomial function, is some unknown nonlinear function and X & are highly correlated (multicollinarity)  \*\*IN POLY. REGRESSION DO NOT TAKE OUT UNOBSERVED CALUES OF X – CAUSES TERRIBLE PREDICTIONS  \*\*F – TESTS ARE USED IN POLY. REGRESSION | | | | \*\*\*\* POLYNOMINAL REGRESSION GRAPH (looks like an upside down hook) | | |
| **Multi-Colliniartiy** (UNDESIRED EFFECT) – 2 or more predictors are highly correlated (meaning they are giving the same information)  \*\* PROBLEMS **1)** Estimated regression function is unstable (meaning that we could get widely different estimates for coefficients dues to small amounts of sampling variability **2)** strong correlation between independent variable **3)** Makes some variables insignificant when they really are significant **4)** misleadingly inflates the SE | | | | Multicollinarity is measured by is VIF When then there is multicolliniarity present  **The interpretation** when VIF = 5.16 is “SE of the VIF for agriculture is inflated by a factor of 5.16 , since the there is no evidence of multicoliniarity”  **Multi-Coliniarity can be useful –** correlated predictors can provide the data with a very good prediction fit. If prediction is the concern of the study, then one may no care about a strong relationship btwn variables | | |
| ROC curve is a plot of sensitivity (true positive) VS. 1-Specificity (false positive rate)  X Axis = Sensitivity; Y Axis = 1 – Specificity  \*\* Gives us the true positive rate and false positive rate for every possible rate for every possible cutoff between 0 and 1 | | | | ROC: The area under the curve determines if the classifier is useful or not **(the higher the area, the more useful the predictor is** | | |
| Deviance = G (Deviance is a test statistic that helps compare the FULL & REDUCED model)  \*Interpertation: G compared to the X2 distribution, with p-value…. (determine significance) | | | | \*\*\*When comparing a **full and reduced model**, you much do:  **Deviance (G) calculation, df of each model = (# predictors -1) , and a P-Value** | | |
| THIS EQUATION IS USED TO CALCULATE CONFIDENCE INTERVAL FOR A SPECIFIC Bcoeff | | | | Variance Inflation Factor:  \*\*Kth variance inflation factor (VIFK) = | | |
| **Diagnostic Tests** | Hat Matrix:  \*\* Unusual values on predictors  K = # of predictors  n = # of observations  \*\* Those with highest variance has a leverage higher then the cutoff, these obs are high leverage | Cooks: Impact on regression predictors  \*Comparing cooks distance to F distribution will tell us how much impact the obs. has on the estimated regression coeff.  n = number of observations  Those predictors that have a cooks distance higher then the cutoff will have high influence on the fitted values of the model | | | DFBETAs: Estimates regression coefficients  **\*\*You will find the DFBETAS threshold with above equation – anything with a DFBETA that is higher then the calculated threshold will have a large influence on the respective regression coefficients** \*\* | alpha = .05, .02, .01 ect  M = total number of comparisons  \*TYPE 1 ERROR CONTROLLED BY ACCOUNTING FOR # of COMPARISONS  i.e. adding useless variables will be accounted for |
| \*\* Statistical model selection is often framed in terms of the *bias-variance tradeoff*  ***\*\* Low Bias but high variance means the estimated model may be overfit to the data and would change dramatically if re-estimated in another sample*** |
| **Studentized Residuals**: Observations lie very far from the estimated regression function | **Goodness of fit tests:**  \*\***Homer-Lemeshow**: Compares actual proportion of y =1 w the probabilities predicted by the model  **\*\*Person Chi-Square:** compares obs. data with expected data | Jackknife: is useful in analyzing the influence of particular observations on the regression function | | | OVER FITTING MODELS: situations when there are a lot of predictors & can get a good fit of sample data BUT the fit maybe poor for “out-of-sample” data from the same populations making the inferences & the predictors based on **the over fitted models useless** | \*\* Use F-TEST for group of predictors  (Polynomial regression model)  \*\*Use T-test for single predictors |
| **Example Questions**:  What are the estimated odds ratio (with respect to the odds of diabetes) associated with an extra 5 units of BMI and being in a physical activity group 4 rather then 2, holding other predictions constant? | | | **Answer to EX Question** | | | |
| **Forward selection:** automated algorithm for deciding which variables to include. Where you can add in variables  \*\*best to start with a model with **NO** possible predictors  **Drawbacks**: **1)** Multiple comparisons are not taken into accounts **2)** Significance of a given predictor may change when other predictors are added at later steps **3)** no guarantee of “best” model | | | **Backwards Selection**: you can take out variables that aren’t significant in the model  \*\*Best to start with a model with **ALL** possible predictors  **Drawbacks**: **1)** Multiple comparisons are not taken into accounts **2)** sometimes variables are dropped that would be significant when added to final reduced models **3)** no guarantee of “best” model | | | |
| **Stepwise Selection**: modification of forward modification (where we can remove predictors if they lose significance after adding new predictors. | | | **Model Selection:**  R2 can measure model performance (higher R2 scores are generally better but adding predictors will always increase the R2 – this doesn’t mean that it was a useful predictors just because R2 goes up)  Ra2 = adjusted R2 (takes into account the # of predictors) adding useless predictors here will reduce your Ra2 and it will show fault in your model. \*\*\*\*\*R2 and Ra2 equations are by ANOVA table at the top of page \*\*\*\*\*\*\* | | | |
| **Mallow’s Cp =** (low Cp means low bias and low variance: compare Cp to the other models and the lowest Cp value will be the best model for selection) | | |
| **Other Model Selection Measures**: measure the same ‘goodness of fit’ that Cp measures \*\***DON’T FORGET SSE = RSS**\* | | | **Likelihood Ratio Test:** degrees of freedom for the full and reduced model include B0 | | | |
| **Sensitivity & Specificity:** Suppose that for a given threshold, the sensitivity of a test is 80%, and the specificity of the test is 55%  \*\*80% of patients who have the disease will have a positive test results \*\*45% of patients who do not have the disease will have a positive test results  \*\*20% of patients who have the disease will have a negative test results \*\*55% of patients who do not have the disease will have a negative test results | | | | | | |
| **\*\*Matched Pair Case Control**: study is one in which cases – subjects exhibiting an outcome of interest, are matched with controls – subjects not exhibiting the outcome of interest  \*\*Matching is done on the basis of equal or similar values one or more risk factor through to be associated with the outcome  \*\* Purpose of matching is in a case control study is to completely balance the cases and control groups with respect to the matched variables, this makes it possible to estimate the effects of other variables of interest after controlling for the confounding effects of the matched variable  \*\*Matching confounders allows us to look at the effect of variables of interest within strata defined my matching variables | | | | | | |
| We want to test H0: B1=B2=B3 = 0 |  | | | | | |