ISYE 6414 Final Exam Review

Terms in this set (111)

Least Square Elimination (LSE) cannot be applied to GLM models.	False - it is applicable but does not use data distribution information fully.
In multiple linear regression with idd and equal variance, the least squares estimation of regression coefficients are always unbiased.	True - the least squares estimates are BLUE (Best Linear Unbiased Estimates) in multiple linear regression.
Maximum Likelihood Estimation is not applicable for simple linear regression and multiple linear regression.	False - In SLR and MLR, the SLE and MLE are the same with normal idd data.
The backward elimination requires a pre-set probability of type II error	False - Type I error
The first degree of freedom in the F distribution for any of the three procedures in stepwise is always equal to one	True

ISYE 6414 Final Exam Review

MLE is used for the GLMs for handling complicated link function modeling in the X-Y relationship.	True
In the GLMs the link function cannot be a non linear regression.	False - It can be linear, non linear, or parametric
When the p-value of the slope estimate in the SLR is small the r-squared becomes smaller too.	False - When P value is small, the model fits become more significant and R squared become larger.
In GLMs the main reason one does not use LSE to estimate model parameters is the potential constrained in the parameters.	False - The potential constraint in the parameters of GLMs is handled by the link function.
The R-squared and adjusted R-squared are not appropriate model comparisons for non linear regression but are for linear regression models.	TRUE - The underlying assumption of R-squared calculations is that you are fitting a linear model.
The decision in using ANOVA table for testing whether a model is significant depends on the normal distribution of the response variable	True
When the data may not be normally distributed, AIC is more appropriate for variable selection than adjusted R-squared	True
The slope of a linear regression equation is an example of a correlation coefficient.	False - the correlation coefficient is the r value. Will have the same + or - sign as the slope.

In multiple linear regression, as the value of R-squared increases, the relationship between predictors becomes stronger	False - r squared measures how much variability is explained by the model, NOT how strong the predictors are.
When dealing with a multiple linear regression model, an adjusted R-squared can be greater than the corresponding unadjusted R-Squared value.	False - the adjusted rsquared value take the number and types of predictors into account. It is lower than the r squared value.
In a multiple regression problem, a quantitative input variable x is replaced by x - mean(x). The R-squared for the fitted model will be the same	True
The estimated coefficients of a regression line is positive, when the coefficient of determination is positive.	False - r squared is always positive.
If the outcome variable is quantitative and all explanatory variables take values 0 or 1, a logistic regression model is most appropriate.	False - More research is necessary to determine the correct model.
After fitting a logistic regression model, a plot of residuals versus fitted values is useful for checking if model assumptions are violated.	False - for logistic regression use deviance residuals.

In a greenhouse experiment with several predictors, the response variable is the number of seeds that germinate out of 60 that are planted with different treatment combinations. A Poisson regression model is most appropriate for modeling this data	False - poisson regression models rate or count data.
For Poisson regression, we can reduce type I errors of identifying statistical significance in the regression coefficients by increasing the sample size.	True
Both LASSO and ridge regression always provide greater residual sum of squares than that of simple multiple linear regression.	True
If data on (Y, X) are available at only two values of X, then the model Y = \beta_1 X + \beta_2 X^2 + \epsilon provides a better fit than Y = \beta_0 + \beta_1 X + \epsilon.	False - nothing to determine of a quadratic model is necessary or required.
If the Cook's distance for any particular observation is greater than one, that data point is definitely a record error and thus needs to be discarded.	False - must see a comparison of data points. Is 1 too large?

ISYE 6414 Final Exam Review

We can use residual analysis to conclusively determine the assumption of independence	False - we can only determine uncorrelated errors.
It is possible to apply logistic regression when the response variable Y has 3 classes.	True
. A correlation coefficient close to 1 is evidence of a cause-and-effect relationship between the two variables.	False- cause and effect can only be determined by a well designed experiment.
Multiplying a variable by 10 in LASSO regression, decreases the chance that the coefficient of this variable is nonzero.	False - I am not sure why anyone would think this would be true.
In regression inference, the 99% confidence interval of coefficient \beta_0 is always wider than the 95% confidence interval of \beta_1.	False- can only compare beta1 with beta1 and beta0 with beta0
The regression coefficients for the Poisson regression model can be estimated in exact/closed form.	False - MLE is NOT closed form.
Mean square error is commonly used in statistics to obtain estimators that may be biased, but less uncertain than unbiased ones. And that's preferred.	True

Regression models are only appropriate for continuous response variables.	False - logistic and poisson model probability and rate
The assumptions in logistic regression are - Linearity, Independence of response variable, and the link function is the logit function.	True - linearity is measured through the link, , the g of the probability of success and the predicted variable.
The log odds function, also called the logit function, which is the log of the ratio between the probability of a success and the probability of a failure	True
In logistic regression we interpret the Betas in terms of the response variable.	False - we interpret it in terms of the odds of success or the log odds of success
In logistic regression we have an additional error term to estimate.	False - there is not error term in logistic regression.
The least square estimation for the standard regression model is equivalent with Maximum Likelihood Estimation, under the assumption of normality.	True
The variance estimator in logistic regression has a closed form expression.	False - use statistical software to obtain the variance-co-variance matrix
We can use the z value to determine if a coefficient is equal to zero in logistic regression.	True - z value = (Beta-0)/(SE of Beta)

In testing for a subset of coefficients in logistic regression the null hypothesis is that the coefficient is equal to zero	True
Like standard linear regression we can use the F test to test for overall regression in logistic regression.	False - It's 1-pchisq(null deviance-residual deviance, DFnull-DFresidual)
For logistic regression we can define residuals for evaluating model goodness of fit for models with and without replication.	False - can only be with replication under the assumption that Yi is binary and n1 is greater than 1
The deviance residuals are the signed square root of the log-likelihood evaluated at the saturated model	True
From the binomial approximation with a normal distribution using the central limit theorem, the Pearson residuals have an approximately standard chi-squared distribution.	False - Normal distribution
Visual Analytics for logistic regression Normal probability plot of residuals Residuals vs predictors Logit of success rate vs predictors	True Normal probability plot of residuals - Normality Residuals vs predictors - Linearity/Independence Logit of success rate vs predictors - Linearity
Under the null hypothesis of good fit for logistic regression, the test statistic has a Chi-Square distribution with n- p- 1 degrees of freedom	True - don't forget, we want large P values

For the testing procedure for subsets of coefficients, we compare the likelihood of a reduced model versus a full model. This is a goodness of fit test	False - it provides inference of the predictive power of the model
Predictive power means that the predicting variables predict the data even if one or more of the assumptions do not hold.	True
One reason why the logistic model may not fit is the relationship between logit of the expected probability and predictors might be multiplicative, rather than additive	True
In logistic regression for goodness of fit, we can only use the Pearson residuals.	False - we can use Pearson or Deviance.
An indication that a higher order non linear relationship better fits the data is that the dummy variables are all, or nearly all, statistically significant	True
Simpson's Paradox - the reversal of association when looking at marginal vs conditional relationships	True
Classification is nothing else than prediction of binary responses.	True
We cannot use the training error rate as an estimate of the true error classification error rate because it is	False - biased downward

Random sampling is computationally more expensive than the K-fold cross validation, with no clear advantage in terms of the accuracy of the estimation classification error rate.	True
Leave on out cross validation is preferred	False - K fold is preferred.
The larger K is, the larger the number of folds, the less bias the estimate of the classification the error is but has higher variability.	True
In Poisson regression underlying assumption is that the response variable has a Poisson distribution, or responses could be wait times, or exponential distribution	True
The g link function is also called the canonical link function.	True - which means that parameter estimates under logistic regression are fully efficient and tests on those parameters are better behaved for small samples.
Poisson distribution, the variance is equal to the expectation. Thus, the variance is not constant	True
For Poisson regression we estimate the expectation of the log response variable.	False - we estimate the log of the expectation of the response variable.
Standard linear regression could be used to model Poisson regression using the variance stabilizing transformation sqrt(mu-3/8) if the number of counts	True - the number of counts can be small - then use Poisson

In Poisson Regression we do not interpret beta with respect to the response variable but with respect to the ratio of the rate.	True
In Poisson regression we model the error term	False - there is no error term
One problem with fitting a normal regression model to Poisson data is the departure from the assumption of constant variance	True
Event rates can be calculated as events per units of varying size, this unit of size is called exposure	True
The estimators for the regression coefficients in the Poisson regression are biased.	False - they are unbiased
To perform hypothesis testing for Poisson, we can use again the approximate normal sampling distribution, also called the Wald test	True - Wald Test also used with logistic regression
Hypothesis testing for Poisson regression can be done on small sample sizes	False - Approximation of normal distribution needs large sample sizes, so does hypothesis testing.
For large sample size data, the distribution of the test statistic, assuming the null hypothesis, is a chi-squared distribution	True

Poisson Assumptions - log transformation of the rate is a linear combination of the predicting variables, the response variables are independently observed, the link function g is the log function	True - remember, NO ERROR TERM
Overdispersion is when the variability of the response variable is larger than estimated by the model	True
The gam() function is a non-parametric test to determine what transformation is best.	True
The deviance and pearson residuals are normally distributed	TRUE - the residual deviances are chi square distributed
Model with many predictors have high bias but low variance.	False - low bias and high variance
When the objective is to explain the relationship to the response, one might consider including predicting variables which are correlated	True - But this should be avoided for prediction
Variable selection addresses multicolinearity, high dimensionaltiy, and prediction vs explanatory prediction	TRUE
The variables chosen for prediction and the variables chosen for explanatory objectives will be the same	False

Variable selection is not special, it is affected by highly correlated variables	TRUE
Confounding variable is a variable that influences both the dependent variable and independent variable	True
Explanatory variable is one that explains changes in the response variable	TRUE
Predicting variable is used in regression to predict the outcome of another variable.	True
It is good practice apply variable selection without understanding the problem at hand to reduce bias.	False - always understand the problem at hand to better select variables for the model.
When a statistically insignificant variable is discarded from the model, there is little change in the other predictors statistical significance.	False - it is possible that when a predictor is discarded, the statistical significance of other variables will change.
We can do a partial F test to determine if variable selection is necessary.	True
When selecting variables for a model, one needs also to consider the research hypothesis, as well as any potential confounding variables to control for	True

We would like to have a prediction with low uncertainty for new settings. This means that we're willing to give up some bias to reduce the variability in the prediction.	True
Generally models with covariance have high bias but low variance	False - they have low bias but high variance.
A measure of the bias-variance tradeoff is the prediction risk	TRUE
To estimate prediction risk we compute the prediction risk for the observed data and take the sum of squared differences between fitted values for sub model S and the observed values.	True - this is called training risk and it is a biased estimate of prediction risk
The larger the number of variables in the model, the larger the training risk.	False - the larger the number of variables in a model the lower the training risk.
The Mallow's CP complexity penalty is two times the size of the model (the number of variables in the submodel) times the estimated variance divided by n.	True
AIC looks just like the Mallow's Cp except that the variance is the true variance and not its estimate.	True

ISYE 6414 Final Exam Review

Another criteria for variable selection is cross validation which is a direct measure of explanatory power.	False - Predictive power
Stepwise is a heuristic search	TRUE it is also a greedy search that does not guarantee to find the best score
If p is larger than n, stepwise is feasible	TRUE - for forward, but not backward
Forward stepwise will select larger models than backward.	False - it will typically select smaller models especially if p is large
Mallow's CP is useful when there are no control variables.	TRUE
The overall regression F-statistic tests the null hypothesis that	the coefficients are equal to zero
The test of subset of coefficients tests the null hypothesis that	discarded variables have coefficients equal to zero.
Goodness of fit tests the null hypothesis that	the model fits the data
The prediction risk is the sum between the irreducible error and the mean square error	True
There is never a situation where a complex model is best.	False - there are situations where a complex model is best

L0 penalty, which is the number of nonzero regression coefficients	True - not feasible for a large number of predicting variables as requires fitting all models
L1 penalty will force many betas, many regression coefficients to be 0s	True - is equal to the sum of the absolute values of the regression coefficients to be penalized
L2 does not perform variable selection	True - is equal to the sum of the squared regression coefficients to be penalized and does not do variable selection
L2 penalty term measures sparsity	False - L1 penalty measures sparsity. L2 removes the limitation on variable selection
The estimated regression coefficients from Lasso are less efficient than those provided by the ordinary least squares	True
Where p the number of predictors is larger than n the number of observationsthe Lasso selects, at most, n variables	True when p is greater than n, lasso will select n variables at the most
If there is a high correlation between variables, Lasso will select both.	False lasso will select 1