**STAT 501 – Final Exam Solutions**

1. **(10x2 = 20 points)** State which of the following statements are true and which are false. For the statements that are false, explain why they are false.
2. In weighted least squares, the “ideal” weights to use, in theory, will equalσi, the population standard deviation of the ith error term. **False. In weighted least squares, the “ideal” weights to use, in theory, will equal 1/σ2.**
3. Weighted least squares is used to adjust for autocorrelated errors. **False. Weighted least squares is used to adjust for non-constant variance.**
4. In the context of time series, suppose we wish to estimate the relationship between the variables and In these transformations, r represents the correlation between εt and εt-1, where these ε-values are errors from the regression between yt and xt. **True.**
5. A very large value of Cook's Di for a data point indicates that if the point is removed, the overall set of fits for the regression will likely change. **True.**
6. Suppose that X1 and X2 have been used to predict Y; then, X3 is added to the model. The value of R2 could possibly decrease due to adding X3. **False. The value of R2 can never decrease, but only increase or stay the same in such a situation.**
7. In binary logistic regression, the success event logit function (log odds of success) is the negative of the failure event logit function (log odds of failure). **True.**
8. In simple linear regression analysis, the width of a prediction interval for a future response of *Y* based on a single predictor *X* does not vary with the value of *X*. **FALSE – The prediction interval width increases as the deviation () increases.**
9. The error terms in a regression with autoregressive AR(1) errors have zero mean. **True.**
10. In model selection, the *MSE* (or *S*) criterion minimizes confidence/prediction interval widths. **True.**
11. In model selection, the *PRESS* criterion evaluates model unbiasedness. **FALSE: the Cp < p criterion should be checked to evaluate model unbiasedness.**
12. **(3+4+3+3 = 13 points)** The following ANOVA table is abstracted from a regression run for the model: *Y* = *β*0 + *β*1 *X*1 + *β*2 *X*2 + *β*3 *X*3 + … + *β*10 *X*10 + ε.

|  |  |  |
| --- | --- | --- |
| ***Source*** | ***DF*** | ***SS*** |
| ***Regression*** | **10** | 110.53 |
| ***Residual Error*** | **28** | **39.47** |
| ***Total*** | 38 | 150 |

|  |  |  |
| --- | --- | --- |
| ***Source*** | ***DF*** | ***Seq SS*** |
| ***X1*** | 1 | 0.10 |
| ***X2*** | 1 | 40 |
| ***X3*** | 1 | 1.0 |
| ***X4*** | 1 | 55 |
| ***X5*** | 1 | 2.5 |
| ***X6*** | 1 | 0.08 |
| ***X7*** | 1 | 6.5 |
| ***X8*** | 1 | 4.0 |
| ***X9*** | 1 | 0.4 |
| ***X10*** | 1 | 0.95 |

* 1. Calculate the three missing values in the upper table.

**See above.**

* 1. For the 10-predictor model, perform a hypothesis test at significance level 0.05 to determine whether predictors *X7*, *X8*, *X9*,and *X10* are significantly linearly related to Y upon controlling for the remaining predictors *X1*-*X6* using a general linear *F* test. Write the null and alternative hypotheses, the value of the test statistic, the decision rule, and the conclusion. *[Note: an F-distribution table is provided on the last page of the exam.]*

**H0: β7 = β8 = β9 = β10 = 0 *vs.* Ha: At least one of β7, β8, β9, β10 ≠ 0.**

**From the above ANOVA table,**

**F = ((6.5+4+0.4+0.95)/4)/(39.47)/28) = 2.9625/1.4096 = 2.10 < 2.71 (F0.95, 4, 28).**

**Therefore we fail to reject H0 in favor of Ha and conclude that none of the predictors *X7*, *X8*, *X9*, and *X10* are significantly linearly related to Y upon controlling for predictors *X1*-*X6*.**

* 1. Later it was decided to consider a regression of Y on the **first 4 predictors ONLY**. Use information from both tables to calculate adjusted R2 for the model with only the first 4 predictors?

**SSE for 4-predictor model = 150 – (0.10+40+1+55) = 150 – 96.1 = 53.9.**

**MSE for 4-predictor model = 53.9 / (39 – 5) = 53.9/34.**

**Adj R2 for 4-predictor model = 1 – 38(53.9/34)/150 = 0.598.**

* 1. Given only the information in both tables, is it possible to test whether *X1* and *X3* can be dropped from the 4-predictor model? Give a brief argument supporting your answer. *[You do not have to do a test, even if one is possible.]*

**Given the current information it is not possible to test whether X1 and X3 can be dropped from the model. The sequential SS are given in the form SS(X1), SS(X2|X1), SS(X3|X1, X2) and SS(X4|X1, X2, X3). To test whether X1 and X3 may be dropped from the model, we would require SS(X1, X3|X2, X4) which is not available.**

1. **(4x3=12 points)** Answer the following questions showing calculations and/or brief justification.
   1. If Correlation(X, Y) = 0.853, what proportion of variation in Y is *not* explained by the regression of Y on X?

**R2 = (0.853)2 = 0.7276 = 72.76% of total variation in Y is explained by regression, so 100-72.76 = 27.24% is not explained.**

* 1. Consider regressing Y on X1 and X2. If SSR(X1 | X2) = SSR(X2 | X1), what can you say about the correlation between X1 and X2?

**X1 and X2 are uncorrelated. The order in which X1 and X2 enter the model is not important only if they are uncorrelated!**

* 1. Consider regressing Y on 7 predictors. When all 7 predictors are included in the model, R2 = 70% and adjusted R2 = 60%. When only the 5 most significant predictors are included in the model, R2 = 65% and adjusted R2 = 62%. Of these two competing models, which one will you prefer and why?

**Preferred model will be the second one since adjusted R2 is higher and it is more parsimonious (less complicated with fewer predictors, but almost same amount of total variability is being explained).**

* 1. The figure below gives the sample PACF for a time series data of monthly sales (in thousands of dollars). Use the PACF plot to propose an appropriate time series model for yt = sales during the tth month.



**The plot indicates that k=2. Therefore the appropriate model is**

**E(yt) = β0 + β1 yt-1 + β2 yt-2.**

1. **(4+8+4+4 = 20 points)** Data from a local supermarket revealed that the deli usage of customers depends on their grocery bill and also on the time of shopping. To understand the link between these variables, a logistic regression model was fitted based on data from 890 sales records, which yielded the following.

Source DF Adj Dev Adj Mean Chi-Square P-Value

Regression 2 17.532 8.7660 17.53 0.000

bill 1 10.824 10.8241 10.82 0.001

lunch 1 5.549 5.5489 5.55 0.018

Error 887 534.290 0.6024

Total 889 551.822

Odds Ratio 95% CI

Bill 1.0760 (1.0305, 1.1236)

Odds Ratio for lunch=1 relative to lunch=0

Odds Ratio 95% CI

0.5771 (0.3651, 0.9124)

Here is the estimated probability of deli usage, bill is the amount of the grocery bill and lunch is a binary variable that equals 1 for a store visit at lunchtime and 0 for a store visit at other times.

* 1. Is their any statistical evidence that shopping time is related to the odds of deli usage, and if so, does lunchtime have a higher odds of deli usage than a visit at other times?

**Yes, the p-value of the Chi-square statistic for lunch is 0.018 indicating that shopping time is a statistically significant effect on odds of deli usage. However, lunchtime seems to have lower odds of deli usage as the 95% confidence interval of odds ratio comparing lunchtime to others is entirely less than 1.**

* 1. Write the regression equations for estimating the logit of the probability of:
     1. deli usage in terms of both predictors bill and lunch;

**Ln =**

* + 1. deli usage for lunchtime shoppers;

**Ln =**

* + 1. deli usage for others;

**Ln =**

* + 1. NO deli usage in terms of both predictors bill and lunch.

**Ln =**

* 1. Use your answer in (b)(ii) above to find the value of bill at which the probability of using the deli is 0.80 for a lunchtime shopper.

**Ln =**

**bill = 76.5**

* 1. Write a sentence that interprets the coefficient estimate for the predictor variable bill.

**The slope estimate of the variable bill is 0.0733 and e0.0733 = 1.0760 is its estimated odds ratio indicating that the odds of deli usage are multiplied by 1.0760 for every dollar increase in bill.**

1. **(5+15 = 20 points)** The linear regression results below are for a response variable, Y, and three predictor variables, X1, X2, X3, for a dataset of ten observations. Assume that the intercept term is β0 and the coefficients for the predictor variables are β1, β2 and β3, respectively.
   1. Interpret the t-test result provided for X2 in the Minitab output given below.

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 0.208 0.374 0.55 0.599

x1 0.946 0.293 3.23 0.018 1.09

x2 0.697 0.223 3.13 0.020 1.38

x3 0.301 0.506 0.60 0.573 1.31

**Each t-test considers the question of whether the variable is needed, given that all other variables will remain in the model. The sample coefficient for *X2* achieves statistical significance (p-value = 0.020). This indicates that it is useful as a predictor of *Y* in a model including *X1* and *X3*.**

* 1. Use the following Minitab output, which gives the ANOVA tables for several models, to answer the questions below.

\* Regression Equation: y = 0.208 + 0.946 x1 + 0.697 x2 + 0.301 x3

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

Regression 3 31.63 10.543 11.01 0.007

Error 6 5.75 0.958

Total 9 37.38

\* Regression Equation: y = 0.191 + 0.914 x1 + 0.761 x2

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

Regression 2 31.29 15.645 17.99 0.002

Error 7 6.09 0.870

Total 9 37.38

\* Regression Equation: y = 0.966 + 1.172 x1

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

Regression 1 16.75 16.753 6.50 0.034

Error 8 20.62 2.578

Total 9 37.38

* + 1. Test, at α = 0.05, H0: β2 = β3 = 0 vs. Ha: at least one of β2, β3 are not equal to 0.

**From the Minitab output, we see that SSE(F) = 5.75 with df = 6, and SSE(R) = 20.62 with df = 8.**

**Since the observed F = 7.76 > F0.95(2, 6) = 5.14, we reject H0. It is not reasonable to remove both X2 and X3 from the model.**

* + 1. Test, at α = 0.05, H0: β3 = 0 vs. Ha: β3 is not equal to 0. *[Make sure you use the ANOVA tables to answer this question, although you may use the Coefficient Table in part (a) to check your answer.]*

**From the Minitab output, we see that SSE(F) = 5.75 with df = 6 and SSE(R) = 6.09 with df = 7.**

**Since the observed F = 0.35 < F0.95(1, 6) = 5.99, we do not reject H0 and conclude *X3* can be dropped from the model.**

* + 1. Calculate R2Y,2|1 (note this is a partial R2) and interpret the value.

**The Minitab output give us SSE(X1) = 20.62, and SSE(X1, X2) = 6.09. So,**

**Thus X2 explains 70.5% of the variation in Y that could not be explained by X1.**

1. **(5x3 = 15 points)** In an experiment, a researcher compares three different metal alloys (say, A, B, and C) used to weld pipes. For each alloy, 10 welds are made. The response variable is strength of the weld (Y). In addition to the type of alloy (A, B, or C), a quantitative predictor, diameter of weld (X), will be used in a regression model for predicting Y.
   1. Write a population regression model (with no interaction) for predicting Y using X and alloy (A, B, or C) as predictors. Clearly define any necessary indicator variables.

**Y = Response variable**

**X = Diameter of weld**

**Z1 = 1 if metal alloy is A, 0 otherwise**

**Z2 = 1 if metal alloy is B, 0 otherwise**

**Note that there are 3 types of alloy, hence two indicator variables are sufficient.**

**The model is given by**

* 1. Explain precisely what each regression coefficient measures in the model that you wrote for part (a).

**β0: Intercept term when alloy used is C.**

**β1: The increase in the average weld strength when alloy used is A, compared to that of alloy C. (Note that for alloy A, intercept term is β0 + β1)**

**β2: The increase in the average weld strength when alloy used is B, compared to that of alloy C. (Note that for alloy B, intercept term is β0 + β2)**

**β3: The increase in average weld strength per unit increase in weld diameter.**

* 1. What null hypothesis would be tested to determine whether there are differences among the alloys? Write the hypothesis in terms of the regression coefficients in part (a).

**Null hypothesis: There is no difference in strength among the three alloys against the alternative that at least one alloy differs from the others.**

**H0: β1 = β2 = 0 against Ha: at least one of β1 and β2 is not equal to 0**

* 1. Explain how you would carry out the test of hypothesis described in part (c). Do not forget to write down the degrees of freedom for the test you propose.

**Full model:**

**Reduced Model: ,**

**Test Statistic:**

**Reject if (For F-distribution, numerator df = 2 and denominator df = 26).**

* 1. What term(s) should be added to the model to create a model with interactions? Write down the model.

**The interaction model would include interaction terms of alloy and weld diameter. Hence the model will be E(Y) = β0 + β1 Z1 + β2 Z2 + β3 X + β4 X Z1 + β5 X Z2.**