**BIOM 505: Biostatistical Methods II**

**HOMEWORK 2 (Linear Regression) [Due March 6, 2018]**

**Dr. Fares Qeadan: fqeadan@salud.unm.edu**

|  |
| --- |
| **PART I (MULTIPLE CHOICE AND FILLIN THE BLANK QUESTIONS):** |

**Question 1 [1 points]:**

Multiple regression analysis involves building models for relating dependent variable y to \_\_\_\_\_\_\_\_\_\_or more independent variables.

**Answer:**

**Question 2 [1 points]:**

Many statisticians recommend \_\_\_\_\_\_\_\_\_\_ for an assessment of model validity and usefulness. These include plotting the residuals, standardized or studentized residuals on the vertical axis versus the independent variable xi or fitted values $\hat{y}$i on the horizontal axis.

**Answer:**

**Question 3 [1 points]:**

Which of the following statements are not true?

1. Provided that the model is correct, no residual plot should exhibit distinct patterns.
2. Provided that the model is correct, the residuals should be randomly distributed about 0 according to a normal distribution, so all but a very few standardized residuals should lie between -2 and +2 ( i.e., all but a few residuals are within 2 standard deviations of their expected value 0 ).
3. If we plot the fitted or predicted values on the vertical axis versus the actual values $\hat{y}$i on the horizontal axis, and the plot yields points close to the 60o line, then the estimated regression function gives accurate predictions of the values actually observed.

**Answer:**

**Question 4 [1 points]:**

Quite frequently, residual plots as well as other plots of the data will suggest some difficulties or abnormality in the data. Which of the following statements are not considered difficulties?

1. A nonlinear probabilistic relationship between x and y is appropriate.
2. The variance of the error term ε (and of Y) is a constant σ2.
3. The error term ε does not have a normal distribution.
4. The selected model fits the data well except for very few discrepant or outlying data values, which may have greatly influenced the choice of the best-fit function.
5. One or more relevant independent variables have been omitted from the model

**Answer:**

**Question 5 [1 points]:**

A multiple regression model has

1. One independent variable.
2. Two dependent variables
3. Two or more dependent variables.
4. Two or more independent variables.
5. One independent variable and one independent variable.

**Answer:**

**Question 6 [1 points]:**

In multiple regression models, the error term ε is assumed to have:

1. A mean of 1.
2. A standard deviation of 1.
3. A variance of 0.
4. Negative values.
5. Normal distribution.

**Answer:**

**Question 7 [1 points]:**

The coefficient of multiple determination R is

1. SSE/SST
2. SST/SSE
3. 1-SSE/SST
4. 1-SST/SSE
5. ( SSE + SST ) /2

**Answer:**

**Question 8 [1 points]:**

The adjusted coefficient of multiple determination is adjusted for

1. The value of the error term ε
2. The number of dependent variables in the model
3. The number of parameters in the model
4. The number of outliers
5. The level of significance α

**Answer:**

**Question 9 [1 points]:**

Which of the following statements are not true?

1. The way to incorporate a qualitative (categorical) variable with three possible categories into a regression model is to define a single-numerical variable with coded values such as 0, 1, and 2 corresponding to the three categories.
2. Incorporating a categorical variable with c possible categories into a multiple regression model requires the use of c-1 indicator variables.
3. The positive square root of the coefficient of multiple determination is called the multiple correlation coefficient R.

**Answer:**

**Question 10 [1 points]:**

Which of the following statements are true?

1. The proportion of total variation explained by the multiple regression model is R2=1-SSE/SST; the coefficient of multiple determination.
2. The coefficient of multiple determination R2 is often adjusted for the number of parameters (k+1) in the model by the formula R2a = [(n-1) R2 –k] / [(n – (k+1)].
3. With multivariate data, there is no preliminary picture analogous to a scatter plot to indicate whether a particular multiple regression model will be judged useful.
4. The model utility test in multiple regression involves testing H0: β1 = β2 =….. = βk =0 versus: H1: at least one βi≠ 0 (i = 1, 2, ……, k).
5. All of the above statements are true.

**Answer:**

**Question 11 [1 points]:**

A first-order no-interaction model has the form $\hat{Y}$ = 5 + 3X1 + 2X2. As X1 increases by 1-unit, while holding X2 fixed, then Y will be expected to

1. increase by 10
2. increase by 5
3. increase by 3
4. decrease by 3
5. decrease by 6

**Answer:**

**Question 12 [1 points]:**

Incorporating a categorical variable with 4 possible categories into a multiple regression model requires the use of

1. 4 indicator variables
2. 3 indicator variables
3. 2 indicator variables
4. 1 indicator variable
5. no indicator variables at all

**Answer:**

**Question 13 [1 points]:**

The transformation \_\_\_\_\_\_\_\_\_\_ is used to linearize the function y =α +β ⋅ log(x)

**Answer:**

**Question 14 [1 points]:**

The transformation \_\_\_\_\_\_\_\_\_\_ of the dependent variable y and the transformation \_\_\_\_\_\_\_\_\_\_ of the independent variable x are used to linearize the power function y =αxβ.

**Answer:**

**Question 15 [1 points]:**

The kth -degree polynomial regression model equation is Y = β0 +β1X +β2X2 +....+βkXk +ε, where ε is a normally distributed random variable with E (ε) = \_\_\_\_\_\_\_\_\_\_\_ and Var (ε) = \_\_\_\_\_\_\_\_\_\_\_.

**Answer:**

**Question 16 [1 points]:**

The regression coefficient β2 in the multiple regression model Y = β0 +β1X1 +β2X2 +…. +βkXk +ε is interpreted as the expected change in \_\_\_\_\_\_\_\_\_\_\_associated with a 1-unit increase in \_\_\_\_\_\_\_\_\_\_\_, while\_\_\_\_\_\_\_\_\_\_\_ are held fixed.

**Answer:**

**Question 17 [1 points]:**

A dichotomous variable, one with just two possible categories, can be incorporated into a regression model via a \_\_\_\_\_\_\_\_\_\_\_ or \_\_\_\_\_\_\_\_\_\_ variable x whose possible values 0 and 1 to indicate which category is relevant for any particular observations.

**Answer:**

**Question 18 [1 points]:**

A multiple regression model with k predictors will include \_\_\_\_\_\_\_\_\_\_ regression parameters, because β0 will always be included.

**Answer:**

**Question 19 [1 points]:**

In many multiple regression data sets, the predictors X1, X2, ..., Xk are highly interdependent. When one of the Xis values can be predicted very well from the other predictor values, for at least one predictor, the data is said to exhibit \_\_\_\_\_\_\_\_\_\_.

**Answer:**

**Question 20 [1 points]:**

Which of the following statements are not true?

1. The exponential function y =αeβ x is intrinsically linear.
2. The power function y =αxβ can be linearized by the transformations y′ = log(y) and x′ = log(x).
3. The function y =α +γ xβ is intrinsically linear.

**Answer:**

**Question 21 [1 points]:**

Which of the following statements are true?

1. In general, it is not only permissible for some independent or predictor variables to the mathematical functions of others, but also of often highly desirable in the sense that the resulting model may be much more successful in explaining variation in y than any model without such predictors.
2. Polynomial regression is indeed a specific case of multiple regression.
3. The coefficient βk in the multiple regression model Y = β0 +β1X1 +β2X2 +…. +βkXk +ε is interpreted as the expected change in Y with a 1-unit increase in Xk , when X1, X2, …, Xk-1 are held fixed.
4. All of the above statements are true.
5. None of the above statements are true.

**Answer:**

**Question 22 [1 points]:**

Which of the following statements are true?

1. The forward selection method, an alternative to the backward elimination method, starts with no predictors in the model and consider fitting in turn the model with only X1 , only X2 ,….., and finally only Xk .
2. The stepwise procedure most widely used is a combination of forward selection (FS) method and backward elimination (BE) method.
3. The stepwise procedure starts by adding variables to the model, but after each addition it examines those variables previously entered to see whether any is a candidate for elimination.
4. All of the above statements are true.
5. None of the above statements are true

**Answer:**

**Question 23 [1 points]:**

Which of the following statements are true?

1. The idea behind the stepwise procedure is that with forward selection, a single variable may be more strongly related to y than either of two or more other variables individually, but the combination of those variables may make the single variable subsequently redundant.
2. When the predictors X1, X2, ..., Xk are highly interdependent, the data is said to exhibit multicollinearity.
3. There is unfortunately no consensus among statisticians as to what remedies are appropriate when sever multicollinearity is present. One possibility involves continuing to use a model that includes all the predictors but estimating parameters by using something other than least squares (Ridge regression as an example).
4. All of the above statements are true.
5. None of the above statements are true.

**Answer:**

**Question 24 [1 points]:**

Which of the following shouldn’t be used as a criterion for model selection:

1. LRT
2. AIC
3. BIC
4. R2
5. Radj2
6. Mallows' Cp

**Answer:**

**Question 25 [1 points]:**

For the model Y = β0 +β1X1 +β2X2 +β3X1X2 +ε, if β3 was found to be significant then the main effects shouldn’t be interpreted?

1. Yes
2. No

**Answer:**

**PART II (SAS):**

**Question 1 [45 points (3 points for each part)]:**

Data were collected on 252 men. The body density dataset includes the following 15 variables listed from left to right:

1. Density determined from underwater weighing
2. Percent body fat from Siri's (1956) equation
3. Age (years)
4. Weight (kg)
5. Height (cm)
6. Neck circumference (cm)
7. Chest circumference (cm)
8. Abdomen 2 circumference (cm)
9. Hip circumference (cm)
10. Thigh circumference (cm)
11. Knee circumference (cm)
12. Ankle circumference (cm)
13. Biceps (extended) circumference (cm)
14. Forearm circumference (cm)
15. Wrist circumference (cm)

For analysis purposes, six predictor variables were considered:

1. Age
2. Weight
3. Height
4. Chest
5. Hip
6. Forearm

The response variable of interest is BodyFat. The goal of the study was to determine the relationship between body fat and the six independent variables. Data could be downloaded from: <http://www.mathalpha.com/BIOM505II/bodyfat.sas7bdat>

**FULL MODEL:**

1. From the full model, one obtains F (6, 245) = 54.49 and R2=0.5716. Show the ANOVA and Parameters’ estimates Tables for the full model and show how the given overall F-value and R2 can be obtained from the MS and SS? Interpret R2?

**Answer:**

1. Construct the matrix of scatterplots between the dependent variable and all independent variables and comment on the linearity assumption of the conducted multiple linear regression model?

**Answer:**

1. Use the very same graph in (b) to comment on the presence of multicollinearity in these data? Please confirm your answer by looking at the VIF values from the conducted model? Was the assumption of lack of multicollinearity violated in this model? If it did, what remedial measure will you take (show it)?

**Answer:**

1. Conduct some residuals diagnostics tests and use graphs to comment on the possible violation of the two assumptions of constancy of the errors variance and normality of the error terms in the Full model? (use the Standardized residuals) Desired graphs would be a histogram with fitted normal curve, QQ-PLOT, and residuals against the fitted values.

**Answer:**

1. Does the full model suffer from outlying or influential points? For influential outliers, use the cut-off 6/n instead of 4/n.

**Answer:**

**REDUCED MODEL:**

1. What reduced model is suggested by the backward selection method and forward selection method respectively? Please write down the prediction equation.

**Answer:**

1. What is $\hat{β}\_{0}$ for the reduced model and does it have any meaningful interpretation in this particular example? Please interpret the other parameters’ estimates of the reduced model?

**Answer:**

**COMPARE FULL MODEL WITH REDUCED MODEL**

1. State the R2 for the full model and R2 for the reduced model? Is the R2 for the full model larger than the R2 for the reduced model? Why? Does the reduced model has a larger adjusted R2 than the full model? Which model is more adequate?

**Answer:**

**CONFOUNDING VARIABLES**

1. Let’s consider the reduced model and assume that Hip is the primary predictor variable for Body Fat. An investigator decided to add Ankle circumference (cm) to the reduced model even though it wasn’t significant by claiming that it’s a possible confounder. Is the claim of this investigator justified? [hint: run the reduced model with and without Ankle and study the change happens to the parameter coefficient for Hip]

**Answer:**

**INTERACTION**

1. Firstly, please construct a new variable and call it **BMI** according to the formula: **Weight/(Height/100)^2**

Secondly, create a dummy variable for BMI and call it BMI\_dummy such that it takes on the value 1 if the BMI>=25 and the value 0 if the BMI<25 [in here, 1 indicates overweight/obese and 0 indicates underweight/Normal]. What is the distribution of the newly constructed variable (show the contingency table)

**Answer:**

1. Conduct a multiple regression model to predict Body Fat by using the predictors: Chest, Age, Height, Hip, Ankle, and BMI\_dummy. Provide the ANOVA table and Parameters’ estimates table for this model. What is the prediction equation for this model?

**Answer:**

1. What is the predicted mean Body Fat for a randomly selected subject with the following values:

**Chest Age Height Hip Ankle BMI\_dummy**

99.1 35 167.64 99.2 21.7 1

**Answer:**

1. Please interpret the coefficient 2.869 for the variable BMI\_dummy?

**Answer:**

1. Considering the very same model in (k), please check whether BMI\_dummy has a significant interaction with Hip? Provide the ANOVA and Parameters’ estimates tables.

**Answer:**

1. Give an interpretation for the interaction term? Does your interpretation support the finding that a wide hip circumference is a protective factor among overweight/obese subjects? Explain

**Answer:**

1. Is the model with interaction more adequate than the model without interaction?

**Answer:**

**PART III (SAS): Question 1 [25 points]:**

Linear regression was applied to a large data set having age and weight as covariates. The estimated coefficients for these two variables and their standard errors are as follows:

|  |  |  |
| --- | --- | --- |
| **Covariate** | **Est. Coefficient** | **Est. Standard Error** |
| Age | 1.43 | 0.46 |
| Weight | 25.9 | 31.0 |

**Part- 1a.** Can we reject the null hypothesis that the associated parameter equals zero for either of these variables? Justify your answer.

**[HINT:** Recall that the 95% CI. for β is $\hat{β}$±tn-p\*SE($\hat{β}$) where p is the number of parameters in the model not counting Ϭ2. For large enough sample size (i.e. n≥30), tn-p=2.**]**

**Part- 1b.** Can we infer anything about the biologic significance of these variables from the magnitudes of the estimated coefficients? Justify your answer.

**[HINT:** You may need to make a hypothetical assumption regarding the scale of the DV as a change of 5 units in Y if Y ranges from 0 to 10 is not the same as a change of 5 units in Y if Y ranges from 0 to 10000; in the former the change is biologically more significant**]**

**Using the funding data please answer the following questions:**

Data could be downloaded from: <http://www.mathalpha.com/BIOM505II/funding.sas7bdat>

**Part 2**. Explore the relationship between dollars and the other covariates. Fit a model that you feel best captures this relationship.

**[HINT**: for more information about the dataset and the variables please visit: <http://www.nejm.org/doi/full/10.1056/NEJM199906173402406>. In here your DV and IVs are:

DV: *dollars*

IVs: *incid, preval, hospdays, mort, yrslost, disabil***]**

**Part 3.** -Perform ***forward linear regression*** of log(dollars) against log(incid), log(prevail), log(hospdays), log(mort), log(yrslost) and log(disabil). Use entry of 0.1 (i.e. significance level of α=0.10 to allow the variable in the model). Which covariates are selected by this procedure?

**Part 4** – Repeat part 3 using ***backward model selection***. Use the same thresholds for removal (i.e. significance level of α=0.10 to keep the variable in the model). Do you get the same model as part 3?

**Part 4.5** – Repeat part 3 using ***stepwise model selection***. Use entry of 0.2 and removal of .10. Do you get the same model as part 3?

**Part 5** – Regress log(dollars) against the covariates from the model chosen in part 4.5. Do you get the same parameter estimates? If not why not?

**[HINT**: the parameters’ estimates table you get from the model selection procedure don’t reflect the parameters’ estimates table of the final model; that table includes the variable that was lastly removed and thus you need to re-run the model without the selection option while including the selected variables**]**

**Part 6** – Regress log(dollars) against log(hospdays), log(mort), log(yrslost) and log(disabil).

1. What bounds should contain 95% of the studentized residuals under this regression?

**Answer:** -2 and 2

1. Draw a scatter of the studentized residuals against expected log(dollars).
2. Draw the loess regression curve of the studentized residuals against the expected values. Draw horizontal lines at zero and 95% for the studentized residuals.

**[HINT**: please use sgplot for this purpose**]**

1. What does this graph tell you about the quality and fit of this model to these data?

**Part 7**. In the model from part 6, calculate the Δβ influence statistic for log(mort).

1. List the values of this statistic along with the disease name, studentized residual, and leverage for all diseases for which the absolute value of this statistic is greater than 0.5.

**[HINT**: use DFBETAS not DFFITS. To get DFBETAS:

1. please use the following SAS statement before proc reg:

ods output outputstatistics=outstats;

1. merge **outstats** and **outreg1** together by the *id* variable; this could be done by firstly renaming the *observation* variable in the outstats dataset to *id* as follows:

data outstats;

set outstats;

rename observation=id;

run;

and secondly merging the two datasets as:

data regstats;

merge outreg1 outstats;

by id;

run;

**]**

1. Which disease has the largest influence on the log (mort) parameter estimate?
2. How many standardized errors does this data point shift the log(mort) parameter estimate?
3. How big is its studentized residual?

**Part 8.**

1. Draw scatter plots of log(dollars) against the other covariates in the model from part Six
2. Identify the disease in these plots that had the most influence on the log(mort) in part seven.

**[HINT**: use datalabel=id option as:

SCATTER y=……… x=………/datalabel=id;**]**

1. Does it appear to be particularly influential in any of these scatter plots?

**Part 9**. Repeat the regression from Part 6 excluding the observation on perinatal conditions.

1. Compare your coefficient estimates with those from part 6.
2. What is the change in the coefficient for log(mort) that results from deleting this disease?
3. Explain the difference as a percentage change and as a difference in the standard errors.

**Part 10**. Perform influence analysis on the other covariates in the model from Part 6 [use the model from part 6]

1. Are there any observations that you feel should be dropped from the analysis?
2. Do you think a simpler model might be more appropriate for these data?

**Part 11**. Regress log(dollars) against log(disabil) and log(hopsdays).

1. What is the estimated expected amount of research funds budgeted for a disease that causes a million hospital-days a year and the loss of a million disability adjusted life-years? [**HINT:** a million hospital-days a year means hopsdays=1000 and hence loghopsdays= 6.907755 using the natural logarithm]
2. Calculate a 95% confidence interval for this expected value?
3. Calculate a 95% prediction interval for the funding that would be provided for a new disease that causes a million hospital-days a year and the loss of a million disability-adjusted life-years.

[**HINT**: use clm cli and remember to exponentiate back]

**Part 12**. In part 11, give an interpretation for the parameter estimate of disability adjusted life-years lost while keeping the number of hospital-days constant.

**Part 13.** Perform an influence analysis on the model from question 11.Is this analysis more reassuring than the one that you preformed in Part 10? Justify your answer.