STA 303 H1S / 1002 HS – Winter 2010
Test
February 25, 2010

LAST NAME:________________________ FIRST NAME:______________________________

STUDENT NUMBER:____

ENROLLED IN: (circle one)   STA 303   STA 1002

INSTRUCTIONS:

- Time: 90 minutes
- Aids allowed: calculator.
- Some formulae are on the last page (page 10).
- Total points: 45

1a	1bcd	2	3a	3b(i,ii,iii,iv)	3b(v,vi,vii)	4

1. A manufacturing facility needs to be able to switch from one type of package to another quickly to react to changes in orders. Consultants have developed a new method of changing the production line and used it to produce a sample of 48 change-over times (in minutes). Also available is an independent sample of 72 change-over times (in minutes) for the existing method. Does the mean change-over time differ between the two methods?

   Here is some output from SAS for these data.

```
                                 The GLM Procedure

                              Class Level Information

                        Class         Levels    Values
                        method             2    Existing New

                       Number of Observations Read         120
                       Number of Observations Used         120

Dependent Variable: changeover

                                         Sum of
Source                        DF        Squares    Mean Square   F Value   Pr > F

Model                          1     290.068056     290.068056      5.08   0.0260
Error                        118    6736.923611      57.092573
Corrected Total              119    7026.991667

            R-Square     Coeff Var      Root MSE    changeover Mean
            0.041279      45.54071      7.555963           16.59167


Source                        DF      Type I SS    Mean Square   F Value   Pr > F
method                         1    290.0680556    290.0680556      5.08   0.0260

Source                        DF    Type III SS    Mean Square   F Value   Pr > F
method                         1    290.0680556    290.0680556      5.08   0.0260

                                                    Standard
 Parameter                   Estimate                  Error    t Value    Pr > |t|

 Intercept                14.68750000 B           1.09060928      13.47      <.0001
 method    Existing        3.17361111 B           1.40797053       2.25      0.0260
 method    New             0.00000000 B            .                .         .

NOTE: The X'X matrix has been found to be singular, and a generalized inverse
      was used to solve the normal equations.  Terms whose estimates are
      followed by the letter 'B' are not uniquely estimable.
```

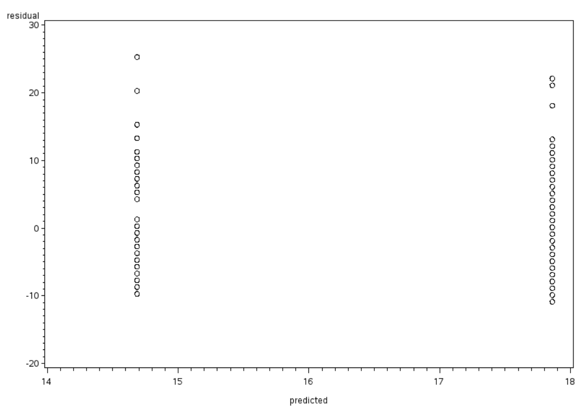
   (a) (1 mark) Is there evidence of a difference in the means of change-over time between the two methods? Explain.

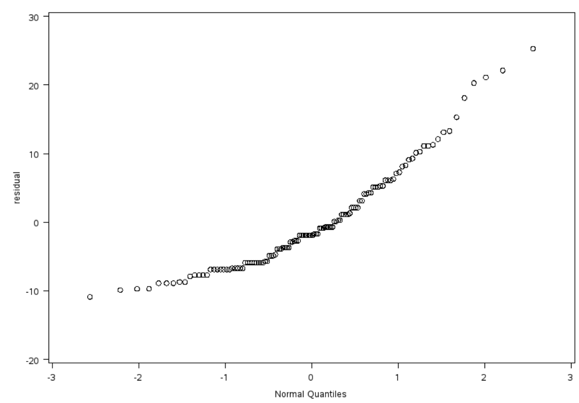
(b) (2 marks) What are the means of the 48 change-over times from the new method and the 72 change-over times from the existing method?

(c) (3 marks) Explain, in the context of this problem, the meaning of the following note produced by SAS:

```
NOTE: The X'X matrix has been found to be singular, and a generalized inverse
      was used to solve the normal equations.  Terms whose estimates are
      followed by the letter 'B' are not uniquely estimable.
```

(d) (3 marks) Below are a plot of the residuals versus the predicted values and a normal quantile plot of residuals. What do you conclude from them?





2. An alternative formulation of the model that could have been used in question 1 is

$$Y_{gi} = \theta_g + \epsilon_{gi}, \quad g = 1, 2$$

where Y_{gi} is the change-over time for the ith observation using the gth method and ϵ_{gi} are random errors. By the method of least squares, the estimates of θ_g are found by minimizing

$$\sum_{g=1}^{2}\sum_{i=1}^{n_g}(Y_{gi} - \theta_g)^2$$

with respect to θ_1, θ_2.

(a) (2 marks) Find the least squares estimates of θ_1 and θ_2.

(b) (2 marks) How are θ_1 and θ_2 related to the parameters of the model fit in question 1?

3. A book on baseball uses regression analysis to compare the success of 30 Major League Baseball teams. One relationship the author considers is the linear relationship between market size (that is, the population, in millions, of the city associated with each team (variable name: `population`)) and the number of times the team made the playoffs in the 10 seasons between 1995 and 2004 (variable name: `appearances`). The author found that "it is hard to find much correlation between market size and success in making the playoffs. The relationship is quite weak."

   (a) (2 marks) The author's comments are about a linear regression analysis that was carried out. Indicate two concerns that potentially threaten the validity of this analysis.

   (b) Some SAS output for an appropriate logistic regression analysis is given below and on the next page. A few numbers have been replaced by letters.

```
                         The LOGISTIC Procedure

                            Model Information
            Data Set                        WORK.A
            Response Variable (Events)      appearances
            Response Variable (Trials)      n
            Model                           binary logit
            Optimization Technique          Fisher's scoring

                Number of Observations Read          30
                Number of Observations Used          30
                Sum of Frequencies Read             300
                Sum of Frequencies Used             300

                             Response Profile
                   Ordered     Binary           Total
                     Value     Outcome      Frequency
                         1     Event               80
                         2     Nonevent           220

                        Model Convergence Status
            Convergence criterion (GCONV=1E-8) satisfied.

                 Deviance Goodness-of-Fit Statistic
     Criterion          Value        DF     Value/DF      Pr > ChiSq
     Deviance        116.2229       (A)       4.1508          <.0001

               Number of events/trials observations: 30
```

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	349.949	(B)
SC	353.653	351.483
-2 Log L	347.949	340.075

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	(C)	1	0.0050

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.4584	0.2110	47.7649	<.0001
population	1	0.0781	0.0275	8.0534	0.0045

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
population	1.081	1.024	(D)

Obs	team	DevResid	Pearson Resid
1	Mets	-1.92105	-1.85370
2	Yankees	3.76061	3.20643
3	Angels	-1.22434	-1.16810
4	Dodgers	-0.52485	-0.51634
5	Cubs	-0.85685	-0.82115
6	WhiteSox	-1.65667	-1.49836
7	Phillies	-2.48767	-1.90432
8	Rangers	0.29713	0.30201
9	Marlins	-0.41610	-0.40514
10	Astros	1.68376	1.81046
11	BlueJays	-2.40481	-1.83112
12	Tigers	-2.38611	-1.81475
13	RedSox	1.72103	1.85669
14	Braves	5.30552	5.55467
15	Athletic	1.09465	1.15770
16	Giants	0.98205	1.02942
17	Expos	-2.30392	-1.74343
18	Diamondb	0.50449	0.52033
19	Mariners	1.21489	1.29822
20	Twins	0.53480	0.55290
21	Padres	-0.18949	-0.18692
22	Cardinal	1.91378	2.10315
23	Orioles	-0.16301	-0.16108
24	Pirates	-2.22632	-1.67701
25	DevilRay	-2.22363	-1.67472
26	Rockies	-0.97242	-0.89234
27	Indians	2.62450	2.95412
28	Reds	-0.95668	-0.87852
29	Royals	-2.18087	-1.63850
30	Brewers	-2.15560	-1.61721

i. (5 marks) Give the values of the missing numbers. ((D) is worth 2 marks.)

(A) = ____________________________

(B) = ____________________________

(C) = ____________________________

(D) = ____________________________

ii. (2 marks) Give the p-values for 2 tests with null hypothesis that the coefficient of population is 0.

iii. (2 marks) Explain what is being tested by the Deviance Goodness-of-Fit test.

iv. (2 marks) Explain in practical terms the interpretation of the estimated coefficient of `population`.

v. (2 marks) What population is associated with an estimated 50% chance of making the playoffs?

vi. (2 marks) What do you conclude from the residuals?

vii. (4 marks) Does the fitted model appear to be appropriate from the SAS output you are given? What else would you like to see to assess the appropriateness of the model?

4. A textile researcher is interested in how four different colours of dye affect the durability of fabrics. Because the effects of the dye may be different for different types of cloth, he applies each dye to five different kinds of cloth. For each kind of cloth, 24 fabric specimens are cut from a length of the cloth and the first six of the 24 specimens are dyed the first colour, the second six the second colour, etc. All 120 specimens are tested for durability, measured as the length of time for the fabric to break down under a stress.

   Explain how you would carry out the analysis on the resulting data. In particular, indicate:

   (a) (1 mark) The type of analysis (one-way analysis of variance, two-way analysis of variance, binary response logistic regression, or binomial response logistic regression) to be carried out.

   (b) (3 marks) The response variable and the explanatory variables as they will be entered into the model.

   (c) (5 marks) The test(s) you would carry out to evaluate effects of dye on the durability of the fabrics. For the test(s) indicate the null and alternative hypotheses and the probability distribution(s) (including the degrees of freedom) of the test statistic(s) under the null hypothesis.

   (d) (2 marks) Do you have any concerns about the validity of the tests? Why or why not?

Some formulae:

Pooled t-test

$$t_{obs} = \frac{\overline{y}_1 - \overline{y}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

Linear Regression

$$b_1 = \frac{\sum(X_i-\overline{X})(Y_i-\overline{Y})}{\sum(X_i-\overline{X})^2} = \frac{\sum X_iY_i - n\overline{XY}}{\sum X_i^2 - n\overline{X}^2} \qquad b_0 = \overline{Y} - b_1\overline{X}$$

One-way analysis of variance

$$\text{SSTO} = \sum_{i=1}^{N}(Y_i - \overline{Y})^2 \qquad \text{SSE} = \sum_{g=1}^{G}\sum_{(g)}(Y_i - \overline{Y}_g)^2$$

$$\text{SSR} = \sum_{g=1}^{G} n_g(\overline{Y}_g - \overline{Y})^2$$

Bernoulli and Binomial distributions

If $Y \sim \text{Bernoulli}(\pi)$
$\text{E}(Y) = \pi$, $\text{Var}(Y) = \pi(1-\pi)$

If $Y \sim \text{Binomial}(m, \pi)$
$\text{E}(Y) = m\pi$, $\text{Var}(Y) = m\pi(1-\pi)$

Logistic Regression with Binomial Response formulae

$$\text{Deviance} = 2\sum_{i=1}^{n}\{y_i\log(y_i) + (m_i - y_i)\log(m_i - y_1) - y_i\log(\hat{y}_i) + (m_i - y_i)\log(m_i - \hat{y}_1)\}$$

$$D_{res,i} = \text{sign}(y_i - m_i\hat{\pi}_i)\sqrt{2\left\{y_i\log\left(\frac{y_i}{m_i\hat{\pi}_i}\right) + (m_i - y_i)\log\left(\frac{m_i - y_i}{m_i - m_i\hat{\pi}_i}\right)\right\}}$$

$$P_{res,i} = \frac{y_i - m_i\hat{\pi}_i}{\sqrt{m_i\hat{\pi}_i(1-\hat{\pi}_i)}}$$

Model Fitting Criteria

$$\text{AIC} = -2\log(L) + 2(k+1) \qquad \text{SC} = -2\log(L) + (k+1)\log(N)$$