Ph.D. Qualifying Examination: Applied Probability and Statistics

May 28, 2010

Exam C, Topics in Statistics

1. The Poisson regression model consists of count response $y_1,..., y_n$ that follow independent Poisson distribution with mean $EY_i = \mu_i = e^{\mathbf{x}_i'\boldsymbol{\beta}}$, i= 1, 2, …, n, where $\mathbf{x_i'} = [1, x_{i1},..., x_{ip}]$, respectively.

   a. Derive the deviance for the Poisson regression model.
   b. What is the asymptotic distribution of this deviance in 1 a.? Explain.

2. Consider the linear model described by

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{X}$ is the $n \times p$ design matrix of full rank p and $\boldsymbol{\beta}$ is the $p \times 1$ vector of parameters. Assume that the linear model satisfies normality, independence, and homoscedasticity. Consider the residuals e_i , i =1,2, …., n.

   (a) Derive the variance and covariance of the residual in terms of the elements of the hat matrix.
   (b) Define the studentized residuals.
   (c) Explain how the studentized residual and certain elements of the hat matrix are used to analyze the least squares fit.

3. Consider the two-way factorial design with each cell as described below:

$$E\left(Y_{ijk}\right) = \mu_{ij} = \mu + \alpha_i + \beta_j \, , \begin{cases} i = 1, 2, ..., I, \\ j = 1, 2, ..., J, \end{cases}$$

with α_i's effects of factor A at I levels $\left(\sum_{i=1}^{I} \alpha_i = 0\right)$ and the β_j's effects of factor B at J levels $\left(\sum_{j=1}^{J} \beta_j = 0\right)$. Under the assumptions of normality, independence, and homoscedasticity, show that $S^2_{\text{Resid}} \big/ \sigma^2$, where

$$S^2_{\text{Resid}} = \sum_{i=1}^{I} \sum_{j=1}^{J} \left(Y_{ij} - \overline{Y}_{i+} - \overline{Y}_{+j} + \overline{Y}_{++}\right)^2,$$

has a Chi-square distribution with (I -1)(J -1) degrees of freedom.

4. The Gauss-Newton procedure for computation of regression coefficients in a nonlinear regression model with nonhomogeneous variance involves computing

$$\mathbf{b} = \mathbf{b_0} + \left(\mathbf{D'V^{-1}D}\right)^{\mathbf{-1}} \mathbf{D'V^{-1}(y - \boldsymbol{\mu}_0)},$$

where $\mathbf{V} = diag\left\{\sigma_i^2\right\}$ and $\mathbf{D}$ is the matrix of derivatives of the type $\partial\boldsymbol{\mu}/\partial\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$ is a starting value. This procedure can be adopted to compute estimates of regression coefficients for logistic regression model iteratively and using re-weighted least squares. Show that the above equation reduces to

$$\mathbf{b} = \mathbf{b_0} + \left(\mathbf{X'VX}\right)^{\mathbf{-1}} \mathbf{X'(y - \boldsymbol{\mu}_0)}$$

for the logistic regression model, where $\mathbf{X}$ is the $m \times p$ design matrix of full rank *p*.

5. An experiment is run on an operating chemical process and the objective of this experiment is to study the effects of four variables on the yield of a chemical yield. A 2^4 factorial design is used and the order of the 16 runs is randomly assigned. For each variable we use (-) and (+) to denote its two levels.

The four variables are in the following table.

Table 1:

Variable Number/Level	**-**	+
1: Concentration of CO2 (%) (C)	5	10
2. Temperature (T) (°C)	20	40
3. Catalyst (K)	A	B
4. Agitation speed (A) (rpm)	10	20

Table 2:

Order of Runs/ Variable Number	**1**	**2**	**3**	**4**	**Observed yield**
13	-	-	-	-	0.4
5	+	-	-	-	0.4
9	-	+	-	-	0.3
16	+	+	-	-	0.3
11	-	-	+	-	0.6
15	+	-	+	-	0.55
10	-	+	+	-	0.3

8	+	+	+	-	0.3
3	-	-	-	+	0.60
14	+	-	-	+	0.60
7	-	+	-	+	0.55
1	+	+	-	+	0.50
4	-	-	+	+	0.80
12	+	-	+	+	0.75
6	-	+	+	+	0.55
2	+	+	+	+	0.55

(a) Using table 2, please estimate the interaction between temperature and concentration of C02.

Use the following table for the estimates of the three-factor and four-factor interactions to answer part (b) below.

Effect	**Estimates**
C*T*K	0.019
C*T*K	-0.0063
C*K*A	0.0063
T*K*A	0.0063
C*T*K*A	0.0063

(b) Assume all the three-factor and four-factor interactions to be zeros; that is, the observed three-factor and higher order interactions are due to error (noise). Please estimate the error variance (assume the error variance are the same under every experimental condition) and construct a 95% confidence interval for interaction between temperature and concentration of C02 (Upper 2.5% points of t-distribution: $t_{2,0.025} = 4.302$; $t_{3,0.025} = 3.182$; $t_{4,0.025} = 2.776$; $t_{5,0.025} = 2.57$; $t_{6,0.025} = 2.447$)

6. Paint used for marking lanes on highways must be very durable. In one trial paint from three different suppliers, labeled A, B, and C were tested on four different highways (sites), denoted as 1, 2, 3, 4. The objective of this trial is to test whether paint from different suppliers (treatment groups) has significantly different wear. Different sites have different levels of traffic and weather; therefore the trial was conducted using a randomized block design to reduce the effect of this nuisance variable (blocking variable).

After the experiments are done, the average wear of paints was shown below.

Table 3:

Sites/ Suppliers	**A**	**B**	**C**
1	70	60	55
2	80	65	65
3	70	50	60
4	60	60	58

Table 4: the ANOVA table

	Sum of Squares	**Degrees of Freedom**	**Mean Squares**	**F values**
Suppliers	316.5		(1)	
Site	218.91			
Error (Residuals)	688.25		(2)	

The model is as follows:

$Y_{it} = \mu + B_i + T_t + e_{it}$, where Y_{it} is the wear (response variable), B_i is the i th site effect (blocking effect), T_t is the effect for the t th supplier group (treatment effect), and e_{it}, i=1,…,4 and t=1, 2, 3 are independent errors and normally distributed with mean 0 and variance σ^2 (Note: $\sum_{t=1}^{3} T_t = 0$ and $\sum_{i=1}^{4} B_i = 0$).

Note: some information that you may need to use:

The grand average is 62.75.

Upper 5% points for F distribution: $F_{2,6,0.05}$ =5.14, $F_{3,6,0.05}$ =4.76, $F_{4,6,0.05}$ =4.53, $F_{2,12,0.05}$ =3.89, $F_{3,12,0.05}$ =3.49, $F_{4,12,0.05}$ =3.26. Upper 2.5% points of t distribution: $t_{2,0.025}$ = 4.302; $t_{3,0.025}$ =3.182; $t_{4,0.025}$ = 2.776; $t_{5,0.025}$ = 2.57; $t_{6,0.025}$ = 2.447)

Please answer the following questions.

(a) Please fill out (1) and (2) in Table 4.

(b) Please conduct an F test for testing whether there is any significant blocking effect ($H_0: B_1 = B_2 = B_3 = B_4 = 0$) at significance level=0.05.

(c) Let $\mu_1 = \mu + T_1$ and $\mu_2 = \mu + T_2$. Calculate the 95% confidence interval for $\mu_1 - \mu_2$ (the mean difference between the two treatment groups).