Final Exam

Statistics 207
Winter Quarter, 2015

PLEASE SHOW ALL WORK

1. I. A random sample of s subjects was chosen for a clinical study spanning k months in order to investigate the effects of a certain blood pressure lowering drug. On month j, each subject was given a dose X_j of the drug, $j = 1, \ldots, k$. Let Y_{ij} be the average daily blood pressure (of patient i) in the j^{th} month. Let $t_j = j$, $j = 1, \ldots, k$ be the months. A reasonable model for this study seems to be

$$Y_{ij} = \mu + \rho_i + \gamma t_j + \varepsilon_{ij}, j = 1, \ldots, k, i = 1, \ldots, s,$$

where μ and γ are constants, ε_{ij}'s are iid $N(0, \sigma^2)$, ρ_i's are iid $N(0, \sigma_\rho^2)$ (random subject effect), and $\{\rho_i\}$ and $\{\varepsilon_{ij}\}$ are mutually independent.
(a) Explicitly obtain $E(Y_{ij})$, $Var(Y_{ij})$, $Cov(Y_{ij}, Y_{ij'})$ and $Corr(Y_{ij}, Y_{ij'})$ when $j \neq j'$.
(b) The experimenter is also interested in the following random slope model which is of the form

$$Y_{ij} = \mu + \rho_i + \gamma t_j + \gamma_{1i} t_j + \varepsilon_{ij},$$

where ε_{ij}'s are iid $N(0, \sigma^2)$, ρ_i's are iid $N(0, \sigma_\rho^2)$, γ_{i1}'s are iid $N(0, \sigma_1^2)$, and $\{\rho_i\}, \{\gamma_{1i}\}$ and $\{\varepsilon_{ij}\}$ are mutually independent.

Explicitly obtain $E(Y_{ij})$, $Var(Y_{ij})$ and $Cov(Y_{ij}, Y_{ij'})$ and $Corr(Y_{ij}, Y_{ij'})$ when $j \neq j'$.

II. The number of auto insurance claims is often modeled as a Poisson random variable. A random sample of n drivers over the age of 40 was taken, and let Y_i be the number of auto insurance claims made by the i^{th} person in the last ten years and let X_i be the person's age. If $Y_i \sim$Poisson(μ_i), then the mean μ_i is modeled by $\mu_i = \exp(\beta_0 + \beta_1 X_i)$, where β_0 and β_1 are unknown and needs to be estimated form the data (X_i, Y_i),$i = 1, \ldots, n$. Let L be the likelihood and l be the log-likelihood, i.e., $l = \log L$. [Recall that the pmf of a Poisson random variable Y with mean μ is $f(y) = e^{-\mu}\mu^y/y!$].
(a) Show that

$$l(\beta_0, \beta_1) = \sum(\beta_0 + \beta_1 X_i)Y_i - \sum \exp(\beta_0 + \beta_1 X_i) - \sum \log(Y_i!).$$

(b) In order to obtain the maximum likelihood estimators, the log-likelihood $l(\beta_0, \beta_1)$ has to be maximized with respect to β_0 and β_1. The likelihood equations are obtained when the derivatives of $l(\beta_0, \beta_1)$ with respect to β_0 and β_1 are set to zero. Show that the likelihood equations are

$$\sum \mu_i = \sum Y_i \text{ and } \sum X_i \mu_i = \sum X_i Y_i,$$

where μ_i is of the form $\mu_i = \exp(\beta_0 + \beta_1 X_i)$.

III. In a ridge regression problem, all the variables have been standardized and the model is $Y = X\beta + \varepsilon$, where X is 30×3. The eigenvalues of the matrix $X^T X$ are 25, 3 and 1. Let $\hat{\beta}(k)$ be the ridge estimate of β when k is the penalty parameter. Thus $X\hat{\beta}(k)$ is the ridge estimator of $X\beta$. Let $\beta(k) = E[\hat{\beta}(k)]$, $L_1(k) = E[||X\hat{\beta}(k) - X\beta(k)||^2]$ and $L_2(k) = ||X\beta(k) - X\beta||^2$. The quantities $L_1(k)$ and $L_2(k)$ can be regarded as the variance and bias-square when estimating $X\beta$ by $X\hat{\beta}(k)$ (with the loss $||X\hat{\beta}(k) - X\beta||^2$). The following table given you the values of L_1 and L_2 for $k = 0, 2.5, 5, 7.5$ and 10.

k	0	2.5	5	7.5	10
$L_1(k)$	12	4.82	3.45	2.75	2.29
$L_2(k)$	0	1.06	2.15	3.22	4.27

Of the five values of the penalty parameter k, which one is the most appropriate for estimating $X\beta$ by a ridge estimator? Is the ridge estimate of $X\beta$ using this "most appropriate value of k" better than $X\hat{\beta}$, where $\hat{\beta}$ is the least squares estimate of β? Explain your answers.

2. In a trial to investigate if Rogaine (Minoxidil) promotes hair growth, four women were randomly selected and given Rogaine over a five week period. Hair gains were monitored for five weeks 0, 8, 16, 24 and 32. A summary of the data is given below

Source	df	SS	MS	F	p-val
Subject		54091			
Week		11683			
Error					
Total		72014			

Grand mean: 204.75, Subject means: 283.0, 182.2, 213.2, 140.6

Mean hair gains over weeks:164.50, 189.75, 227.00, 217.00, 225.50.

(a) Write down an appropriate model for analyzing this data. Explain the terms in the model and the assumptions on them.

(b) Complete the ANOVA table.

(c) Obtain an estimate of the proportion of variability in hair gain that can be explained by variability among the subjects (women).

(d) Use Bonferroni method to construct simultaneous 90% confidence intervals for the changes $\mu_2 - \mu_1, \mu_3 - \mu_2$ and $\mu_4 - \mu_3$. Explain the results. [Mean hair gains at weeks 0, 8, 16, 24 and 32 are denoted by $\mu_1, \ldots, \mu_5$.]

(e) Some people believe that it is important to compare the average hair gain in weeks 8 and 16 to the average gain in weeks 8 through 32. Let $\theta = (\mu_2 + \mu_3 + \mu_4 + \mu_5)/4 - (\mu_2 + \mu_3)/2$. Carry out a test to decide if $H_0 : \theta = 0$ vs $H_1 : \theta > 0$ at level $\alpha = 0.05$. State your conclusion.

(f) The data set whose summary is given above is a part of a larger study where in addition to the 4 women who were given Rogaine, four other women were also chosen at random and were given a placebo (control group). Hair gains were also recorded for the subjects in the control group at weeks 0, 8, 16, 24 and 32. Write down a model that can be used to analyze the bigger data set (containing women on Rogaine and on placebo). Carefully explain the terms in the model and the assumptions associated with them.

3. From the record of last 10 years' of applicants to graduate programs in a large university, a random sample 400 files was taken. From each file, the following were recorded: Y=admission status (1=admitted, 0=de-

nied), X_1=GRE score (divided by 100 for numerical convenience), X_2=undergraduate GPA, X_3=prestige of applicant's undergraduate institution (1=prestigious, 0=not prestigious). A logistic regression on the probability of admittance was run with five predictors: X_1, X_2, X_3, $X_4 = X_1X_3$ and $X_5 = X_2X_3$. A summary is given below

Coefficients:	Intercept	X_1	X_2	X_3	X_4	X_5
Estimate	-5.35281	0.27727	0.70273	0.56111	-0.07821	0.24549
Std. Error	1.78781	0.17456	0.55005	2.27438	0.22343	0.68547

Null deviance: 499.98 on 399 degrees of freedom

Residual deviance: 463.19 on 394 degrees of freedom.

(a) As part of a model building process, one may need to remove one (or more) predictor from the model. If you decide to delete only one variable from this logistic regression, which one is the best candidate for deletion? Explain your answer.

(b) Another logistic regression has also been fitted with variables X_1, X_2 and X_3 and a summary is given below.

Coefficients:	Intercept	X_1	X_2	X_3
Estimate	-5.61799	0.2280	0.86406	0.93832
Std. Error	1.12206	0.1085	0.32756	0.23295

Null deviance: 499.98 on 399 degrees of freedom

Residual deviance: 463.37 on 396 degrees of freedom.

It is of interest to know if the two interaction terms (i.e., X_4 and X_5) can be dropped from the model given in part (a). Write down the appropriate null and the alternative hypotheses and carry out a test at a level of significance $\alpha = 0.05$. Find the p-value of your test. [Please find tightest bounds on the p-value using the statistical table.]

For parts (c), (d) and (e), use the model given in part (b).

(c) Obtain simultaneous 95% confidence intervals for the beta parameters associated with variables X_1, X_2 and X_3.

(d) Estimate the probability π of getting admitted when $X_1 = 6.5$, $X_2 = 3.2$ and $X_3 = 1$. You are given the information that $SE(\hat{\pi}) = 0.04122$. Also estimate the odds θ of admission when $(X_1, X_2, X_3) = (6.5, 3.2, 1)$ and obtain a 95% confidence interval for θ.

(e) Let θ_1 be the odds of getting admission when $(X_1, X_2, X_3) = (6.5, 3.2, 1)$ and θ_0 be the odds of getting admission when $(X_1, X_2, X_3) = (6.5, 3.2, 0)$. Estimate the odds ratio θ_1/θ_0 and construct a 95% confidence interval for it.

(f) Once again let θ_1 be the odds of getting admitted when GRE=X_1,GPA=X_2 and $X_3 = 1$ and θ_0 be the odds when GRE=X_1,GPA=X_2 and $X_3 = 0$. Obtain an expression for the odds ratio θ_1/θ_0 in terms of X_1, X_2 and the beta parameters **when the model contains all the five variables** $X_1, \ldots, X_5$. Use this expression to find the conditions on the beta parameters under which the odds ratio θ_1/θ_0 does not depend on the values of X_1 and X_2? [There is no need to carry out any numerical calculation for this part.]