UNIVERSITY OF MASSACHUSETTS
Department of Mathematics and Statistics
Applied Statistics
August 2016

Work all problems. 60 points are needed to pass at the Masters Level and 75 to pass at the Ph.D. level.

1. (28pts) A local health clinic sent fliers to its clients to encourage everyone to get a flu shot. In a pilot follow-up study, 155 clients were randomly selected and asked whether they actually received a flu shot. A client who received a flu shot was coded Y=1, and a client who did not receive a flu shot was coded Y=0. In addition, the following variables are available: X1 = age; X2 = health awareness index, with higher values indicating greater awareness; gender, with X3 = 1 for males and X3 = 0 for females. Partial data is shown below. The goal of the study is to model the probability of getting a flu shot as a function of X1, X2 and X3.

i	**1**	**2**	…	**154**	**155**
X1	59	61	…	67	57
X2	52	55	…	66	64
X3	0	1	…	0	0
Y	0	0	…	0	0

a) List three problems with using linear regression with ordinary least squares when the outcome variable is binary. If an assumption is violated, describe in detail the assumption and why the assumption is not met.

The following multiple logistic regression model was fit to the data:

Y_i are independent Bernoulli random variables with
expected values $\mathrm{E}\{\mathrm{Y_i}\} = \pi_i$, and

$$E\{Y_i\} = \pi_i = \frac{\exp(\beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_3 X3)}{1 + \exp(\beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_3 X3)}.$$

Relevant output from SAS PROC GENMOD is shown below and can be used to answer the following questions:

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Likelihood Ratio 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.5995	3.1660	-7.9341	4.5746	0.26	0.6134
x1	1	0.0701	0.0307	0.0114	0.1328	5.22	0.0223
x2	1	-0.0888	0.0363	-0.1648	-0.0207	5.97	0.0146
x3	1	0.3950	0.5308	-0.6352	1.4720	0.55	0.4568
Scale	0	1.0000	0.0000	1.0000	1.0000		

Note: The scale parameter was held fixed.

LR Statistics For Type 3 Analysis			
Source	DF	Chi-Square	Pr > ChiSq
x1	1	5.50	0.0190
x2	1	6.67	0.0098
x3	1	0.56	0.4538

b) State the estimated logistic regression function (i.e. the estimated logistic regression model).

c) Find $\exp(b_1), \exp(b_2), \exp(b_3)$.

d) Interpret $\exp(b_1), \exp(b_2), \exp(b_3)$.

e) Find the estimated log-odds of receiving a flu shot for i=2.

f) Find the corresponding estimate for the probability of receiving a flu shot for i=2.

g)
g.i) State the null and alternative hypotheses.

g.ii) Find the likelihood ratio G^2 test statistic.

g.iii) What degree of freedom is used for likelihood ratio G^2 test statistic?

g.iv) Find the p-value of the test.

g.v) Based on the p-value, should the null hypothesis be rejected? Why or why not?

g.vi) What do the above results tell you about β_1 and X1? Explain your answer.

2. (28pts) A large company is studying whether the quality of work score, based on supervisor evaluations, is different for female and male employees. They collected data on a random sample of *n* employees, with the following variables: Y=quality of work score (range 0-100, 100 is highest quality), X1=years of work experience, X2=1 if female, X2=0 if male. The company studied the following multiple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} * X_{i2} + \varepsilon_i; \quad i = 1,...,n,$$

where the ε_i are independent, Normally distributed, (Formula A)

$\mathrm{E}\{\varepsilon_i\} = 0,$ and $\mathrm{Var}\{\varepsilon_i\} = \sigma^2;$ in addition, ε_i are independent of X1.

a) The model was fit and the resulting parameter values are:

$$\beta_0 = 7.1, \beta_1 = 2.3, \beta_2 = 6.8, \beta_3 = -1.7.$$

Draw the mean regression function E(Y | X1, X2). Discuss the type of dependence between Y and (X1,X2) that is described by this regression model.

b) Consider the following statement: for every year of work experience there is no difference in the quality of work score between females and males. Set up a hypothesis based on one more parameters of the regression model that can be used to test this statement.

c) Based on the model given in Formula (A) above, how would you test your hypothesis of the previous part? State the test statistic and its distribution under the null hypothesis. What degrees of freedom are used?

d) Consider the following statement: The effect of years of work experience on the quality of work score does not depend on being female or male. Set up a hypothesis based on one more parameters of the regression model that can be used to test this statement.

e) Based on the model given in Formula (A) above, how would you test your hypothesis of the previous part? State the test statistic and its distribution under the null hypothesis. What degrees of freedom are used?

3. (16pts) Design a simulation study to test the robustness of the Wald confidence interval of regression coefficients to the assumption of normality of the random errors in linear regression. For example, one can generate data from a linear regression with known regression coefficients and random errors from a t distribution with 5 degrees of freedom. Then construct the 95% Wald confidence intervals for the regression coefficients, pretending the errors are normally distributed, and calculate the percentage of times the true coefficients are covered by the constructed confidence intervals. Robustness means the coverage percentage should be close to the nominal 95%. Write R code to implement the simulation.

4. (28pts) You have regressed Y on variables $X_1, X_2, ..., X_p$. Your colleague Bob has regressed Y on the variables $Z_1, Z_2, ..., Z_p$, where

$$Z_j = c_{j0} + \sum_{k=1}^{p} c_{jk} X_k$$

That is, Bob has applied a linear transformation to the predictors (but not to the response).

(a) Show that Bob's $n \times (p+1)$ design matrix $\mathbf{Z}$ is related to yours via

$$\mathbf{Z} = \mathbf{X}\mathbf{t}$$

for some $(p+1) \times (p+1)$ matrix $\mathbf{t}$; explain how the entries in $\mathbf{t}$ are related to Bob's coefficients c's.

(b) Using the hat matrices of the two regressions, show that your fitted values and Bob's fitted values are exactly equal, if $\mathbf{t}$ is invertible.

(c) Show that, if $\hat{\beta}$ is your vector of coefficients, and if $\mathbf{t}$ is invertible, then Bob's vector of coefficient estimates is exactly

$$\mathbf{t}^{-1}\hat{\beta}$$

(d) Is there any point to Bob's transformation of the predictor variables?