Midterm I

February 22, 2017

Notes:

1. You may use one sheet of notes and a calculator. If you do not have a calculator work out expressions as far as you can.
2. Distributions are at the end of the exam.
3. Most parts do not depend on earlier parts of the problem, so if you are stuck please move on to the next section. Partial credit will be given if an answer does depend on an earlier part that was incorrect and the later problem was worked correctly after accounting for the previous error.
4. You do not have to re-derive well known results unless asked to, but if you do use specific results from class or Theorems please state them where used in your explanations.
5. The amount of space is not always an indication of the expected length of an answer. In general brief answers to all questions are better than more detailed responses to half the problems, so please use your time wisely.
6. Partial credit will be given where appropriate, although no points will be given for simply restating the problem.
7. Please cross out any work that you do not want to be graded, so that there is one solution. If you need additional space, you may use the backs of pages, but please indicate that you have done so.
8. In signing your name below, you agree to abide by the Duke Honor Code and will not receive or give help from/to anyone else.

Name:_____

Score:_____

1. Circle True or False

   (a) **[True or False]:** Overdispersion is a problem when there is more variation than the model can explain and indicates lack of fit.

   (b) **[True or False]:** Extremely large absolute values of parameter estimates and large standard errors (or NA's) leading to estimated probabilities that are 0 or 1 are indications of perfect or quasi-separation in logistic regression.

   (c) **[True or False]:**The residual deviance always decreases or stays the same as more predictors are added to a model.

   (d) **[True or False]:** When using training and test data, the model that has the smallest AIC in the training data will have the smallest RMSE in the test data.

   (e) **[True or False]:** perfect or quasi separation occurs only when there is collinearity among the predictors

   (f) **[True or False]:** The best AIC model on a training set will have fewer predictors than the best BIC model.

   (g) **[True or False]:** The negative binomial model is appropriate for data that shows under-dispersion relative to a Poisson regression model

   (h) **[True or False]:** Leave-One-Out Cross validation is a special case of k-fold cross validation

   (i) **[True or False]:** Points with leverage close to 1 will almost always lead to cases that are identified as outliers using studentized residuals.

   (j) **[True or False]:** A large value for the Sum of Squares error in Gaussian regression is an indication of lack of fit.

   (k) **[True or False]:** Choosing the model with the smallest out of sample RMSE ensures that there will not be any lack of fit.

   (l) **[True or False]:** Cases with high leverage values will always have either large Cook's distance or will be outliers and should be removed.

   (m) **[True or False]:** For comparing two models using coverage in predictive intervals, we should also select the model that has the narrowest prediction intervals.

   (n) **[True or False]:** For Poisson or logistic regression, the residual deviance has an approximate Chi-squared distribution with $p + 1$ degrees of freedom (here there are p predictors plus an intercept).

   (o) **[True or False]:** If the model has an R^2 above 0.95 then we can be assured that the regression model is adequate.

   (p) **[True or False]:** Any regression model with a small R^2 value (say less than 0.20) suggests that the model does not fit the data.

   (q) **[True or False]:** A Cook's distance greater than 1 suggests that the case is an outlier and should be removed from the analysis.

   (r) **[True or False]:** logistic or binary regression should be used whenever we have predictors that are indicator variables.

   (s) **[True or False]:** Step-wise selection will always identify the best model in terms of AIC or BIC when it is not possible to enumerate all models.

   (t) **[True or False]:** The absolute value of the coefficients is a good measure of the importance of a variable.

2. Suppose we have a data set with five predictors, $X_1 = GPA$,$X_2 = IQ$, $X_3 = Gender$, with 1 if the individual is a female, $X_4 = GPA : IQ$ and $X_5 = GPA : Gender$ and a response variable that is the starting salary after graduation in thousands of dollars. Roughly 68% of the population has an IQ between 85 and 115 and 70 and 130 represents about 95% of the population. GPA ranges from 0 to 4. After model fitting we obtain the following estimates: $\hat{\beta}_0 = 50$, $\hat{\beta}_1 = 20$, $\hat{\beta}_2 = -0.07$, $\hat{\beta}_3 = 35$, $\hat{\beta}_4 = 0.01$ and $\hat{\beta}_5 = -10$.

   (a) Which answer is correct and why?

      i. For a fixed GPA and IQ, males earn more on average than females.

      ii. For a fixed GPA and IQ, females earn more on average than males.

      iii. For a fixed GPA and IQ, males earn more on average than females, provided that the GPA is high enough.

      iv. For a fixed GPA and IQ, females earn more on average than males, provided that the GPA is high enough.

   (b) Predict the salary for a male with IQ of 110 and a GPA of 4.0

   (c) True or False: Since the coefficient for the GPA IQ interaction term is very small there is little evidence of an interaction effect. (Justify your answer)

3. Consider the output from fitting a logistic regression to the credit card default data, where default is a binary variable (Yes=Default, No = did not default on debt), Student is an indicator of being a student (Student = Yes), balance is the average credit card balance after making monthly payments, and income is the income of the customer. (output on next page)

   (a) Based on these results is there any indication that the model has lack of fit? Explain.

   (b) Provide an interpretation of the coefficient for Student in terms of odds of defaulting that the manager at the credit company can understand.

   (c) Calculate an approximate 95% confidence interval ($Z_{.975} = 1.92$) for $\exp \beta_{Student}$ and provide a one sentence interpretation for the manager.

   (d) The credit card manager seems surprised that coefficient for income is not statistically different from zero. Can one conclude that income is unrelated to the probability of default based on the p-value? Explain why or why not.

```
Call:
glm(formula = default ~ ., family = binomial, data = Default)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4691  -0.1418  -0.0557  -0.0203   3.7383

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
income       3.033e-06  8.203e-06   0.370  0.71152
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2920.6  on 9999  degrees of freedom
Residual deviance: 1571.5  on 9996  degrees of freedom
AIC: 1579.5

Number of Fisher Scoring iterations: 8
```

4. Suppose that given λ, my response Y has a Poisson distribution with mean λ,

$$Y \mid \lambda \sim P(\lambda) \tag{1}$$

$$p(y \mid \lambda) = \frac{\lambda^y e^{-\lambda}}{y!} \tag{2}$$

$$\mathsf{E}[Y \mid \lambda] = \lambda \tag{3}$$

and that λ has a Gamma distribution:

$$\lambda \mid \mu, \theta \sim G(\theta, \theta/\mu) \tag{4}$$

$$p(\lambda \mid \mu, \theta) = \frac{(\theta/\mu)^\theta}{\Gamma(\theta)} \lambda^{\theta-1} e^{-\lambda\theta/\mu} \tag{5}$$

$$\tag{6}$$

(a) Find the mean of Y as a function of μ and θ. (show)

(b) Find the variance of Y as a function of μ and θ (show work)

(c) Explain how this model can handle overdispersion. Can it model data that exhibit under-dispersion?

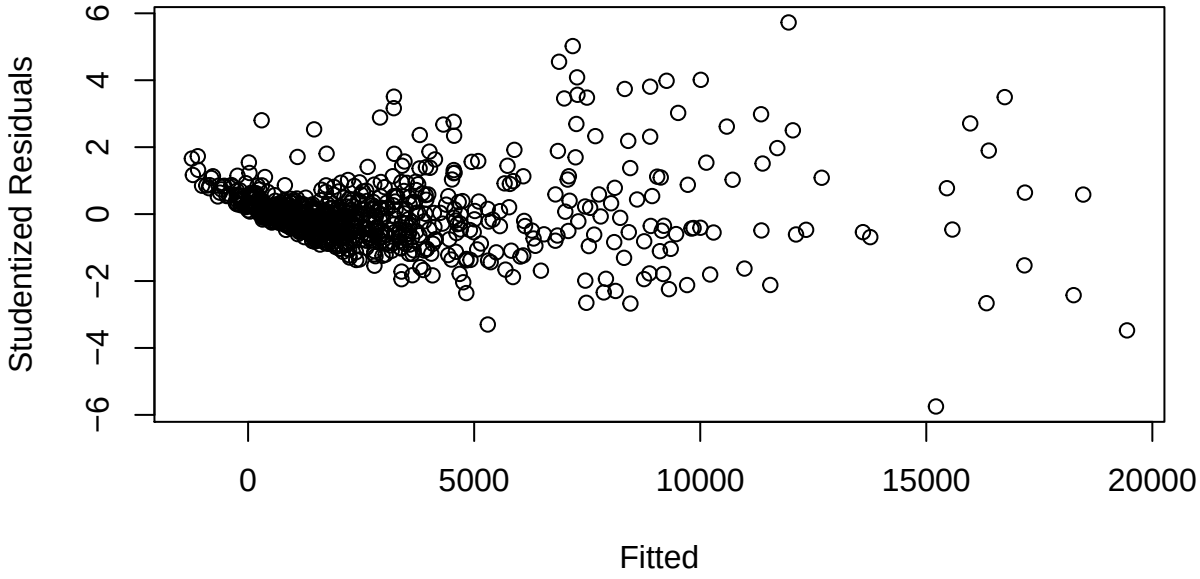
5. Based on the following diagnostic plot below, which of the following are true?

   (a) [**True or False**] There is non-constant variance

   (b) [**True or False**] Only the response needs to be transformed

   (c) [**True or False**] The response and at least one of the predictors should be transformed

   (d) [**True or False**] If the Bonferroni adjustment to the t-value for testing for outliers is 4, we would conclude that there are no outliers.



Useful Distributions

Multivariate Normal

$$\begin{aligned}\mathbf{Y} \mid \boldsymbol{\mu}, \mathbf{V} &\sim \mathsf{N}(\boldsymbol{\mu}, \mathbf{V}) \\ p(\mathbf{Y}) &= (2\pi)^{-n/2}|V|^{-1/2}\exp\{-\frac{1}{2}(\mathbf{Y}-\boldsymbol{\mu})^T\mathbf{V}^{-1}(Y-\boldsymbol{\mu})\} \quad \mathbf{Y}, \boldsymbol{\mu} \in \mathbb{R}^n, \mathbf{V} > 0\end{aligned}$$

Poisson

$$\begin{aligned}Y \mid \lambda &\sim \text{Poi}(\lambda) \\ p(y) &= \frac{y^\lambda e^{-\lambda}}{y!}\end{aligned}$$

Bernoulli

$$\begin{aligned}Y \mid \lambda &\sim \mathsf{Ber}(\pi) \\ p(y) &= \pi^y(1-\pi)^{1-y}\end{aligned}$$

Gamma

$$\begin{aligned}Y \mid a, b &\sim \mathsf{Gamma}(a, b) \\ p(y) &= \frac{b^a}{\Gamma(a)}y^{a-1}e^{-by} \text{ for } a > 0, b > 0, y > 0 \\ \mathsf{E}[Y] &= a/b \\ \mathsf{Var}[Y] &= a/b^2\end{aligned}$$

Generalized Inverse Gaussian

$$\begin{aligned}Y \mid \mu, \nu, \xi &\sim \mathsf{InvGaussian}(\mu, \nu, \xi) \\ p(y) &= \frac{1}{2}K_\mu(\sqrt{\nu\xi})\left(\frac{\nu}{\xi}\right)^{\mu/2}y^{\mu-1}\exp\left\{-\frac{1}{2}\left(\nu y + \frac{\xi}{y}\right)\right\} \quad y > 0, \nu > 0, \xi > 0, \mu \in \mathbb{R}\end{aligned}$$

$K_\mu(\cdot)$ is a modified Bessel function of the second kind

Double Exponential

$$\begin{aligned}Y \mid \lambda &\sim \mathsf{DE}(\mu, \lambda) \\ p(y) &= \frac{\lambda}{2}\exp\left(-\lambda|Y-\mu|\right) \qquad y \in \mathbb{R}, \lambda > 0, \mu \in \mathbb{R}\end{aligned}$$

Student-t

$$\begin{aligned}Y \mid \mu, \sigma^2 &\sim \mathsf{St}(\nu, \mu, \sigma^2) \\ p(y) &= \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\sigma^2\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{1}{\nu}\left(\frac{y-\mu}{\sigma}\right)^2\right)^{-(\nu+1)/2} \qquad y \in \mathbb{R}, \mu \in \mathbb{R}, \sigma^2 > 0\end{aligned}$$

Multivariate Student-t

$$\begin{aligned}\mathbf{Y} \mid \boldsymbol{\mu}, \mathcal{S} &\sim \mathsf{St}(\nu, \boldsymbol{\mu}, \mathcal{S}) \\ p(\mathbf{Y}) &= \frac{\Gamma\left(\frac{\nu+p}{2}\right)}{(\nu\pi)^{p/2}|\mathcal{S}|^{1/2}\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{1}{\nu}(\mathbf{Y}-\boldsymbol{\mu})^T\mathcal{S}^{-1}(\mathbf{Y}-\boldsymbol{\mu})\right)^{-(\nu+p)/2} \qquad \mathbf{Y}, \boldsymbol{\mu} \in \mathbb{R}^p, \mathcal{S} > 0\end{aligned}$$