Homework 5

Stat 151A, Fall 2017

Due: November 28 at 12:30 pm

1. Download the two datasets train.csv and test.csv from `https://www.kaggle.com/c/titanic/data` (this is the competition *Titanic: Machine Learning from Disaster* from Kaggle). Based on the training data, build a reasonable model based on logistic regression for the survival status based on the explanatory variables (you can start with a basic model and subsequently either expand it using interactions etc. and/or perform model selection to remove some variables). Describe your model. Use your model to predict the survival status for the subjects present in the test dataset. Report your prediction accuracy score. (**1.75 points**).

2. In the logistic regression model, let $\hat{p}$ denote the vector of fitted probabilities. Show that $Y - \hat{p}$ is orthogonal to the columns of the X matrix. (**0.25 points**).

3. (**Do not use R for this problem**) Consider the frogs dataset which, briefly, consists of 212 sites of the Snowy Mountain area of New South Wales, Australia were surveyed for the species of the Southern Corroboree frog. The response variable, named *pres.abs*, takes the value 1 if frogs of this species were found at the site and 0 otherwise. The explanatory variables include *altitude*, *distance*, *NoOfPools*, *NoOfSites*, *avrain*, *meanmin* and *meanmax*. The dataset contains 212 observations and the response variable equals one for 79 observations and equals 0 for the rest. I fit a logistic regression model to the data via

```
frogs.glm <- glm(formula = pres.abs ~ log(distance) +
log(NoOfPools) + meanmin,
family = binomial, data = frogs)
summary(frogs.glm)
```

This gave me the following output:

```
Call:
glm(formula = pres.abs ~ log(distance) + log(NoOfPools) + meanmin,
family = binomial, data = frogs)

Deviance Residuals:
Min       1Q   Median       3Q      Max
```

```
-1.9642  -0.7657  -0.4619   0.8728   2.3219

Coefficients:
                 Estimate  Std.Error z value Pr(>|z|)
(Intercept)      0.6864     XXXXX    0.313   0.754146
log(distance)   -0.9050     XXXXX   -4.349   1.37e-05 ***
log(NoOfPools)   0.5027     0.2004   2.509   0.012102 *
meanmin          1.1153     0.3131   3.562   0.000369 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ

(Dispersion parameter for binomial family taken to be 1)

Null deviance: XXXXX  on XXX  degrees of freedom
Residual deviance: XXXXX  on XXX  degrees of freedom
AIC: 222.18

Number of Fisher Scoring iterations: 5
```

Also consider the following R code:

```
X = model.matrix(frogs.glm)
W = diag(frogs.glm$fitted.values*(1 - frogs.glm$fitted.values))
solve(t(X) %*% W %*% X)
```

which gave me the output

```
                (Intercept) log(distance) log(NoOfPools)     meanmin
(Intercept)      4.8038479  -0.363947754   -0.255928180 -0.49698440
log(distance)   -0.3639478   0.043313307    0.008053415  0.01562971
log(NoOfPools)  XXXXXXXXXX   0.008053415    0.040141698  0.02678507
meanmin         -0.4969844   0.015629708    0.026785069  XXXXXXXXX
```

(a) Fill the eight missing values in the above output giving appropriate reasons. (**2 points**)

(b) Suppose a new site is found where the values of the explanatory variables are

```
distance = 265      NoOfPools = 26       meanmin = 3.5
```

According to the logistic regression model, what is the predicted probability that Southern Corroboree frogs will be found at this site? (**0.5 points**).

(c) Suppose I add the variable *altitude* to the model. Would the residual deviance increase or decrease? Explain with reason. Would the null deviance increase or decrease? Explain with reason.(**0.5 points**).

4. Consider the usual regression data with binary response values $y_1, \ldots, y_n$ and explanatory variable values $x_{ij}, i = 1, \ldots, n$ and $j = 1, \ldots, p$. The response vector is Y and the matrix of explanatory variables is X. I wish to fit the logistic regression model to the data:

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \text{for } \; i = 1, \ldots, n$$

where $y_1, \ldots, y_n$ are independent random variables having the Bernoulli distribution with means $p_1, \ldots, p_n$.

   (a) Argue that the maximum likelihood estimates of $\beta_0, \beta_1, \ldots, \beta_p$ depend on Y only through the vector $X^T Y$. (**0.5 points**).

   (b) Let $\hat{p}$ denote the vector of fitted probabilities with components $\hat{p}_1, \ldots, \hat{p}_n$. Express $\hat{p}_i$ in terms of the MLE $\hat{\beta}_0, \ldots, \hat{\beta}_p$ and the explanatory variable values. (**0.5 points**).

   (c) Argue that the sum of the components of $\hat{p}$ equals the number of response values $y_1, \ldots, y_n$ that are equal to one. (**0.5 point**).

   (d) Express the residual deviance in terms of $y_1, \ldots, y_n$ and $\hat{p}_1, \ldots, \hat{p}_n$. (**0.5 points**).

5. (**Do not use R for this problem**) Consider the email spam data. This data consists of 4601 emails of which 1813 emails were identified as spam. The response variable, named *yesno*, takes the value y if the email is spam and n otherwise. The explanatory variables include *crl.tot* (total length of words in capitals), *dollar* (number of occurences of the \$ symbol), *bang* (number of occurences of the symbol), *money* (number of occurences of the word *money*), *n000* (number of occurences of the string *000*) and *make* (number of occurences of the word *make*). I fitted a logistic regression model to the data via

```
s = 0.001
M1 <- glm(yesno~ log(crl.tot) + log(dollar+s) + log(bang+s)
+log(money+s) + log(n000+s) + log(make+s),
family=binomial, data=spam)
summary(M1)
```

This gave me the following output:

```
Call:
glm(formula = yesno ~ log(crl.tot) + log(dollar + s) + log(bang +
s) + log(money + s) + log(n000 + s) + log(make + s), family = binomial,
data = spam)

Deviance Residuals:
Min       1Q   Median       3Q      Max
```

```
-3.1657  -0.4367  -0.2863   0.3609   2.7152

Coefficients:
                 Estimate  Std. Error z value
(Intercept)      4.11947    0.36342   XXXXXX
log(crl.tot)     0.30228    0.03693    8.185
log(dollar + s)  0.32586    0.02365   13.777
log(bang + s)    0.40984    0.01597   25.661
log(money + s)   XXXXXXX    0.02800   12.345
log(n000 + s)    0.18947    0.02931    6.463
log(make + s)   -0.11418    0.02206   -5.177
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ

(Dispersion parameter for binomial family taken to be 1)

Null deviance: XXXXXX  on XXXX  degrees of freedom
Residual deviance: 3245.1  on XXXX  degrees of freedom
AIC: XXXXXX

Number of Fisher Scoring iterations: 6
```

(a) Fill the six missing values in the above output giving appropriate reasons. (**2 points**)

(b) Suppose a new email comes in for which

crl.tot	dollar	bang	money	n000	make
157	0.868	2.894	0	0	0

According to the above logistic regression model, what is the predicted probability that this email is spam? (**0.5 points**).

(c) It may be noted that in the model $M1$, I took logarithms of the explanatory variables. I decided to fit another logistic regression model without taking logarithms of the explanatory variables:

```
M2 = glm(yesno~ crl.tot + dollar+ bang +money+n000 +make,
family=binomial, data=spam)
```

The residual deviance for this model turned out to be 4058.8. On the basis of this, which of the two models $M1$ and $M2$ would you use and why? (**0.5 points**).