Homework 7 Solutions
401 Fall 2001
Brian Junker and Kimberly Sellers

Thursday November 29, 2001

There is one problem on this homework set (YEAH!). It is the ICU Problem from the DASL site.

1. Perform a logit regression of vital status on the following predictors: AGE, SER, CRN, INF, CPR, SYS, TYP, PO2, CRE, LOC. Which of these predictors appear significant?

   The regression summary is below.

```
> logit1_glm(STA~AGE+ SER+ CRN+ INF+ CPR+ SYS+ TYP+PO2+ CRE+ LOC,
     data=icu.dat, family=binomial)
> summary(logit1)

Call: glm(formula = STA ~ AGE + SER + CRN + INF + CPR + SYS + TYP +
PO2 + CRE + LOC, family = binomial, data = icu.dat)
Deviance Residuals:
       Min         1Q     Median          3Q      Max
 -2.432629 -0.6381858 -0.3598685 -0.2200567 2.487251

Coefficients:
                   Value  Std. Error    t value
(Intercept) -9.155354525 3.026165147 -3.0253982
        AGE  0.029011242 0.012361101  2.3469788
        SER -0.114523150 0.495403084 -0.2311717
        CRN  0.456541979 0.663690465  0.6878839
        INF  0.205290553 0.463537169  0.4428783
        CPR  0.530504822 0.845952945  0.6271091
        SYS -0.008988061 0.007004207 -1.2832374
        TYP  1.790629681 0.825673693  2.1686893
        PO2  0.079684019 0.675732974  0.1179223
        CRE  0.522866394 0.837659332  0.6241993
        LOC  1.744695870 0.577356842  3.0218675

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 200.161 on 199 degrees of freedom
Residual Deviance: 146.6754 on 189 degrees of freedom
```

To determine which of these are significant, we need to compare them against the critical t value at $\alpha = 0.05$ with $N - K = 200 - 11 = 189$ degrees of freedom. This is 1.96. The variables which are significant are the variables whose t-values are larger than 1.96 in absolute value. These variables include the intercept, AGE (age of the patient in years), TYP (Type of admission 0 = elective, 1 = emegerency), LOC (level of consciousness at admission (0 = no coma or stupor, 1 = deep stupor, 2= coma).

2. Compare your results from above with the smaller model considering AGE, SYS, TYP, LOC. Are these predictors significant? How does this model compare with the model above?

   The results from the regression are below.

```
> logit2_glm(STA~AGE+ SYS+ TYP+LOC, data=icu.dat, family=binomial)
> summary(logit2)

Call: glm(formula = STA ~ AGE + SYS + TYP + LOC, family = binomial,
    data = icu.dat)
Deviance Residuals:
       Min         1Q     Median        3Q      Max
 -2.473428 -0.6214165 -0.3816336 -0.214436 2.450215

Coefficients:
                  Value  Std. Error   t value
(Intercept) -8.028492449 2.058680615 -3.899824
        AGE  0.032145549 0.011715349  2.743883
        SYS -0.009375171 0.006486795 -1.445270
        TYP  2.063883482 0.763355062  2.703701
        LOC  1.821664633 0.556695249  3.272283

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 200.161 on 199 degrees of freedom

Residual Deviance: 148.8141 on 195 degrees of freedom

Number of Fisher Scoring Iterations: 5
```

   In this model, we see what variables are significant by comparing the t-values to the critical values of a t distribution with $N - 5 = 195$ degrees of freedom. Again, the critical value is 1.96. From this, we can see that all the variables are significant except for SYS. This agrees with the model above.

   To do a more formal statistical test, we compare the -2 Log Likelihoods. These are the residual deviance numbers from the regression printouts. Thus, I compare 148.81-146.67 = 2.14. This number is a chi-squared random variable with (195-189 = 6) degrees of freedom. The p-value for this is (1-pchisq(2.14, 6) = 0.906) 0.9, which indicates that we can not reject the null hypothesis that the (SER, CRN, INF, CPR, PO2, CRE) variables all have zero coefficients. Thus, the two variables are not statistically different.

3. From part (2), preform a final logit regression using only the predictors that were found to be significant in part (2). Interpret these coefficients.

The results from the regression are below.

```
> logit3_glm(STA~AGE+TYP+LOC, data=icu.dat, family=binomial)
> summary(logit3)

Call: glm(formula = STA ~ AGE + TYP + LOC, family = binomial, data = icu.dat)
Deviance Residuals:
       Min          1Q     Median          3Q     Max
 -2.570309 -0.6692762 -0.3652437 -0.2189821 2.48935

Coefficients:
                  Value Std. Error   t value
(Intercept) -9.54344153 1.80847942 -5.277053
        AGE  0.03291119 0.01178783  2.791963
        TYP  2.18838611 0.76025187  2.878501
        LOC  1.83445330 0.51595680  3.555440

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 200.161 on 199 degrees of freedom

Residual Deviance: 151.0171 on 196 degrees of freedom

Number of Fisher Scoring Iterations: 5
```

Here, I compare the t-value to the critical value of a t distribution with $N - 4 = 196$ degrees of freedom. AGain, the cirtical value is 1.96. I can see that all of the variables are significant. The interpretation of the variables is as follows.

- Intercept
  When AGE, TYP and LOC are zero, the odds of favoring STA=1 (the patient dies) are $e^{-9.54} = 7.191685e - 05$. The probability that the patient dies is 7.191168e-05 ($= 1/(1 + exp(9.54))$).
- AGE
  For each one unit increase in AGE multiplies the odds favoring the patient dying by $e^{0.0329} = 1.033447$. In other words, the odds of the patient dying change by $100(e^{0.0329} - 1) = 3.344719$ for each one unit increase in AGE.
- TYP
  For each one unit increase in TYP (i.e. if type goes from elective to emergency) multiplies the odds favoring the patient dying by $e^{2.18} = 8.84630$. In other words, the odds of the patient dying change by $100(e^{2.18} - 1) = 784.6306$ if the TYP chnages from elective to emergency.
- LOC
  For each one unit increase in LOC multiplies the odds favoring the patient dying by $e^{1.83} = 6.233887$. In other words, the odds of the patient dying change by $100(e^{6.233} - 1) = 523.3883$ for each one unit increase in LOC. Recall that LOC is the kevel of consciousness at admission (0 = no coma or stupor, 1= deep stupor, 2 = coma) .

4. Use the model in part (3) to determine the survival probability of a patient who is an emergency admint in a deep stupor at age 73.

   To do this, we note that $\hat{P} = P(STA = 1) = 1/(1 + e^{-L})$ where L is the value obtained from the logistic regression. From part (3), we find that the value of L is
   $L = -9.54 + 0.03 * AGE + 2.18 * TYP + 1.83 * LOC$.
   Thus, our value of L is $-9.54 + 0.03 * 73 + 2.18 * 1 + 1.83 * 1 = -3.34$. Thus, we find that $P(STA = 1) = 1/(1 + e^{3.34}) = 0.03422416$. Recall that STA is coded as 0 for lived and 1 for died. Thus, the probability that the patient survives is $1 - \hat{P}$ = 1-0.034 = 0.966.

5. Construct a conditional effect plot for the relation between age and survival probability, with separate cureves for

   - an emergency admit that is not comatose or in a deep stupor
   - an elective admit that is not comatose or in a deep stupor

   Label the curves and interpret your results.

   The conditional effect plot is in figure 1. From this plot, we can see that as age increase, so does the probability of death for both emegerency and elective admits that are neither comatose nor in a deep stupor. We can also see that emergency admits have an overall higher probability of dying than do elective admits that are neither comatose nor in a deep stupor.

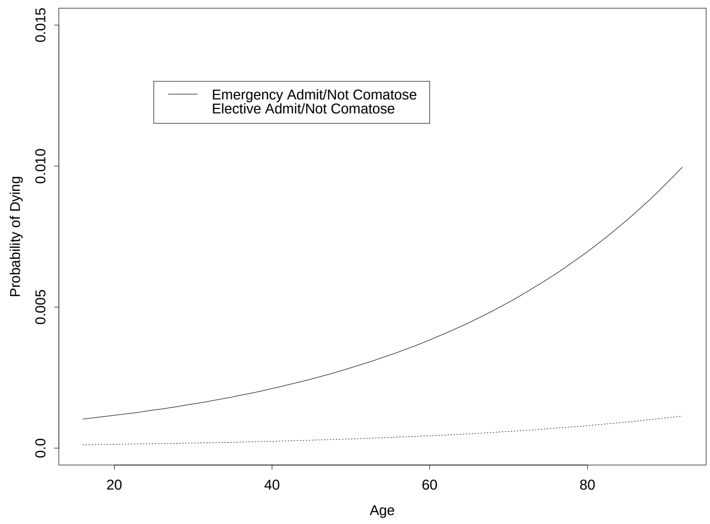

Figure 1: Conditional Effect Plot

6. Construct the following diagnostic graphs and interpret your results.

   This problem was dropped from the homework.

7. Are there any other factors that you would include in the model? If so, what are they? If not, explain. How did you come to this conclusinon? Summarize your procedure.

   There are many methods to do this. First, I regressed all of the variables to see which ones were significant.

```
> logit4_glm(STA~AGE + SEX + RAC + SER + CAN + CRN + INF + CPR + SYS +
     HRA + PRE + TYP + FRA + PO2 + PH + PCO + BIC + CRE + LOC,
     data=icu.dat, family=binomial)
> summary(logit4)

Call: glm(formula = STA ~ AGE + SEX + RAC + SER + CAN + CRN + INF +
CPR + SYS + HRA + PRE + TYP + FRA + PO2 + PH + PCO + BIC + CRE
+ LOC, family = binomial,  data = icu.dat)
Deviance Residuals:
       Min         1Q     Median         3Q     Max
 -1.844384 -0.6041056 -0.2529607 -0.1151384 2.87473

Coefficients:
                    Value  Std. Error      t value
(Intercept) -14.489673266 4.238017886 -3.418974071
        AGE   0.052426547 0.017179465  3.051698564
        SEX  -0.554900057 0.499566039 -1.110764172
        RAC  -0.004552407 0.526441371 -0.008647511
        SER  -0.545280964 0.573984791 -0.949992008
        CAN   2.757361894 0.973729438  2.831753653
        CRN  -0.101860156 0.760597448 -0.133921243
        INF  -0.056021626 0.532955567 -0.105115003
        CPR   0.977737959 0.980567617  0.997114265
        SYS  -0.011542707 0.007818187 -1.476391804
        HRA  -0.003641320 0.009316317 -0.390854049
        PRE   0.929235317 0.626889670  1.482294830
        TYP   2.743712482 0.985149172  2.785073125
        FRA   1.151902788 0.996882186  1.155505440
        PO2   0.386604516 0.848148612  0.455821669
         PH   2.414851861 1.223555166  1.973635459
        PCO  -3.172145775 1.373154688 -2.310115388
        BIC  -0.793681034 0.913148404 -0.869169820
        CRE   0.233314024 1.072219947  0.217599034
        LOC   2.705721398 0.745173331  3.630996017

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 200.161 on 199 degrees of freedom

Residual Deviance: 129.5997 on 180 degrees of freedom

Number of Fisher Scoring Iterations: 5
```

From this, I then kept only the variables whose t-values were greater than 1 in absolute value. (the cut-off is 1.96, however removing so many variables at once can sometimes lead to problems).

I then preformed the second regression below.

```
> logit5_glm(STA~AGE+SEX+CAN+CPR+SYS+PRE+TYP+FRA+PH+PCO+LOC,
      data=icu.dat, family=binomial)
> summary(logit5)

Call: glm(formula = STA ~ AGE + SEX + CAN + CPR + SYS +  PRE + TYP +
FRA + PH + PCO + LOC, family = binomial, data = icu.dat)
Deviance Residuals:
       Min          1Q     Median          3Q      Max
 -1.842783 -0.5948605 -0.2598861 -0.1202714 2.757431

Coefficients:
                   Value  Std. Error   t value
(Intercept) -16.00327944 3.676014412 -4.353432
        AGE   0.04937599 0.015207680  3.246780
        SEX  -0.57475644 0.484689736 -1.185823
        CAN   2.67471302 0.916122502  2.919602
        CPR   1.01549052 0.902105735  1.125689
        SYS  -0.01024318 0.007097126 -1.443285
        PRE   0.90423907 0.582311240  1.552845
        TYP   2.91230757 0.926938933  3.141855
        FRA   1.01604971 0.940583097  1.080234
         PH   2.00119450 0.951043528  2.104209
        PCO  -2.64042792 1.138488759 -2.319239
        LOC   2.49893678 0.669560167  3.732206

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 200.161 on 199 degrees of freedom

Residual Deviance: 131.3039 on 188 degrees of freedom

Number of Fisher Scoring Iterations: 5
```

I next check to see if these two regressions are statistically different. (131.30-129.60 = 1.7 on 188-180 = 8 d.f. with a p-value of 0.988869). Thus, the two regressions are statistically the same. Let me see if I can remove some more. Now I'll try the 1.96 cuttoff.

```
> logit6_glm(STA~AGE+CAN+TYP+PH+PCO+LOC, data=icu.dat, family=binomial)
> summary(logit6)

Call: glm(formula = STA ~ AGE + CAN + TYP + PH + PCO + LOC, family =
 binomial, data  = icu.dat)
Deviance Residuals:
       Min          1Q     Median          3Q      Max
 -2.116047 -0.6276995 -0.2980968 -0.1700174 2.534939

Coefficients:
```

```
                  Value Std. Error   t value
(Intercept) -13.65820876  2.7608621 -4.947081
        AGE   0.04018224  0.0130946  3.068611
        CAN   2.14663557  0.8435854  2.544657
        TYP   2.81578023  0.8895847  3.165275
         PH   1.76829630  0.8542301  2.070047
        PCO  -2.13252583  0.9874111 -2.159714
        LOC   2.30889143  0.5745017  4.018946

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 200.161 on 199 degrees of freedom

Residual Deviance: 139.0279 on 193 degrees of freedom

Number of Fisher Scoring Iterations: 5
```

Here, I see that all of the t-values are now significant. Comparing this model to the model with all the variables, I have that the chis-square statistic is 139.02-129.60= 9.42 on 13 degrees of freedom, for a p-value of 1-pchisq(9.42, 13)=0.74. Thus, it is not statistically different from the model with all of the variables.

For the first reduced model above (logit5), I find that the chisquare statistic is 139.02-131.3=7.72 on 193-188=5 d.f. I find that the p-value is 1-pchisq(7.72, 5)=0.17 which is a bit lower, but still not statistically significant.

Finally, let me now compare this model to the model found in part (3). I get the chi-square statistic of 151.01-139.02 = 11.99 on 196-193=3 degrees of freedom. The p-value is 1-pchisq(11.99, 3) = 0.007, which is statistically significant. Thus, I will conclude that the above model (logit6) is the model I would use. This adds variables CAN(Is cancer part of the problem?), PH (PH from initial blood gasses), and PCO (PCO2 from initial blood gasses).

Note that for a complete analysis, the model assumptions would need to be checked. Because the logitinf() command only handles 7 variables, I did not use it for any of the above regressions. For the analysis to be complete though, they should be checked.