STA841 HW3

Dipesh Gautam

November 10, 2015

1 Problem 1

see attached

2 Problem 2

2.1

Given the null hypothesis that $log_2 LD_{50} = 5$, we fit the probit model below. This is restriction of the full probit model we fit in the HW2 problem 3. The summaries are shown for both full and the restricted model below.

```
Call:
glm(formula = cbind(dead, alive) ~ dose, family = binomial(link = probit),
    data = data)

Deviance Residuals:
       1         2         3         4         5         6
-0.96616   0.33030   0.00717   0.58395  -0.23383  -0.25443

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.5180     0.4219  -3.598 0.000321 ***
dose          0.5981     0.1391   4.300 1.71e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25.792  on 5  degrees of freedom
Residual deviance:  1.503  on 4  degrees of freedom
AIC: 16.966

Number of Fisher Scoring iterations: 4

Call:
glm(formula = cbind(dead, alive) ~ I(-5 + dose) - 1, family = binomial(link = probit),
    data = data)

Deviance Residuals:
      1        2        3        4        5        6
-1.8174  -0.2598   0.3638   1.8279   2.0924   2.9236

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
```

```
I(-5 + dose)  0.16118    0.06576   2.451   0.0142 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25.969  on 6  degrees of freedom
Residual deviance: 19.769  on 5  degrees of freedom
AIC: 33.232

Number of Fisher Scoring iterations: 4

Analysis of Deviance Table

Model 1: cbind(dead, alive) ~ I(-5 + dose) - 1
Model 2: cbind(dead, alive) ~ dose
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1         5     19.769
2         4      1.503  1   18.266 1.921e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Log-likelihood ratio statistic LR(5) can be computed by just taking twice the diffecence between unrestricted and restricted log maximum likelihood which was found to be: 18.266.
If the null hypothesis is correct, the approximated distribution of the difference will be χ^2_1. We used chisq test to get the p-value which was found to be: 1.92e-05.

2.2

Profile log-likelihood of $log_2 LD_{50}$ is plotted against $log_2 LD_{50}$ for a range of possible values below. We fitted different sub-model defined by different values of $log_2 LD_{50}$ and obtained the profile log-likelihood for each sub-model and the values are plotted in Fig 1.

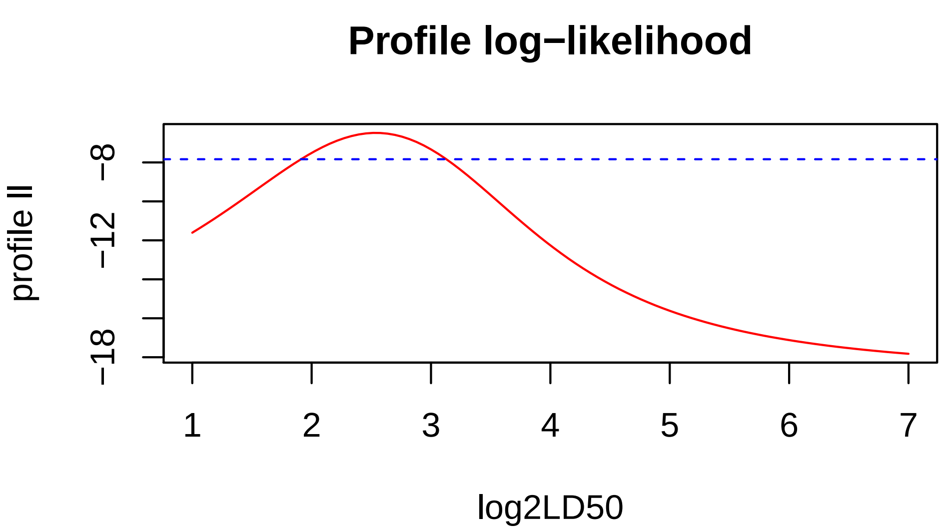

Figure 1:

Using the data and the plot, we also constructed profile likelihood 90% confidence set, which is [3.756, 8.701].

3 Problem 3

For this problem data from *cyl.txt* is used. The data has result from tossing of cylindrical ”dice” with same radius and different heights. We're interested in the probability that the dice lands on the side and not the ends.

3.1

For this part, we fit a simple logistic regression model, the summary of which is presented below.

```
Call:
glm(formula = cbind(y, n - y) ~ height, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2826  -0.6224   0.2124   1.1587   3.0209

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.3580     0.4622  -9.429   <2e-16 ***
height        6.4919     0.6349  10.225   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 258.43  on 74  degrees of freedom
Residual deviance: 116.55  on 73  degrees of freedom
AIC: 227.67

Number of Fisher Scoring iterations: 4
```

After we fit the model, we were able to estimate the height for which the cylindrical dice have probability of 1/3 of landing on their side.
The estimated height using logistic regression is: 0.565.

3.2

Similarly, we fit a probit model, the summary of which is presented below.

```
Call:
glm(formula = cbind(y, n - y) ~ height, family = binomial(link = "probit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3001  -0.5737   0.2599   1.1494   2.9969

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.5982     0.2608  -9.963   <2e-16 ***
height        3.8559     0.3509  10.989   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 258.43  on 74  degrees of freedom
Residual deviance: 117.12  on 73  degrees of freedom
AIC: 228.25

Number of Fisher Scoring iterations: 4
```

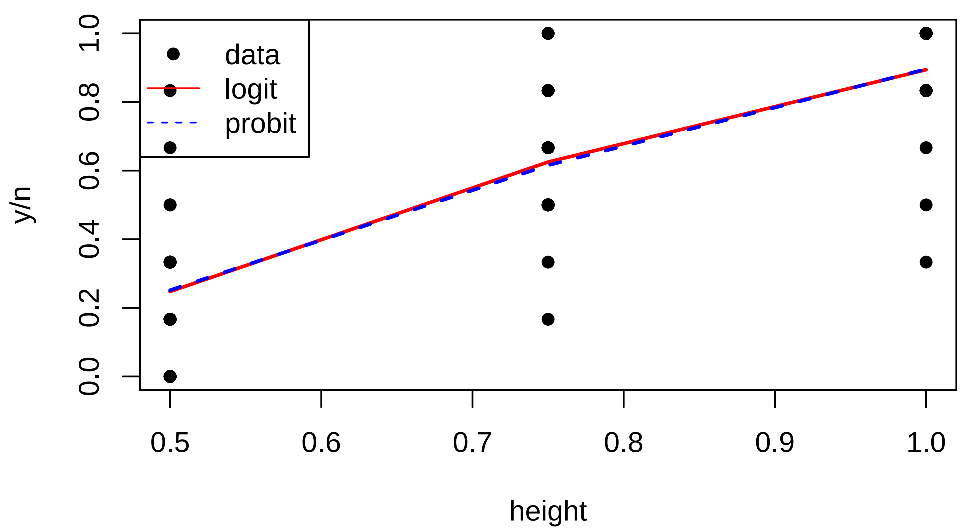

Figure 2:

Estimated height using probit model for which the cylindrical dice have probability of 1/3 of landing on their side is: 0.538.
The two binary regressions along with the sample data are plotted in Fig. 2

3.3

We used bootstrap to construct a confidence interval for the difference between critical height estimated by logistic regression and the one estimated by probit model.

The 95% confidence interval is found to be: [-0.113, -0.083].

4 Problem 4

To start with, we fit a logistic regression model with all of the given covariates. The result is shown below.

```
Call:
glm(formula = cbind(credit, 1 - credit) ~ currentBalance + duration +
    paymentPrevious + use + maritalStatusGender, family = binomial(link = "logit"),
    data = data4)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5547  -0.8292   0.4505   0.7753   2.2362

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -0.975241   0.550119  -1.773  0.07626 .
currentBalance2       0.508081   0.194308   2.615  0.00893 **
currentBalance3       1.103670   0.343806   3.210  0.00133 **
currentBalance4       1.832536   0.212407   8.627  < 2e-16 ***
duration             -0.042406   0.006848  -6.193 5.91e-10 ***
paymentPrevious1      0.115967   0.494030   0.235  0.81441
paymentPrevious2      1.014170   0.391629   2.590  0.00961 **
paymentPrevious3      1.043826   0.449800   2.321  0.02031 *
paymentPrevious4      1.661510   0.413059   4.022 5.76e-05 ***
use1                  1.394731   0.333920   4.177 2.96e-05 ***
use2                  0.552933   0.237390   2.329  0.01985 *
use3                  0.802997   0.224140   3.583  0.00034 ***
use4                  0.550642   0.729603   0.755  0.45042
use5                  0.108902   0.507979   0.214  0.83025
use6                 -0.309443   0.369446  -0.838  0.40226
use8                  1.825148   1.135049   1.608  0.10784
use9                  0.634231   0.309687   2.048  0.04056 *
use10                 1.025893   0.712423   1.440  0.14987
maritalStatusGender2  0.087982   0.351280   0.250  0.80223
maritalStatusGender3  0.580052   0.342061   1.696  0.08993 .
maritalStatusGender4  0.279581   0.417534   0.670  0.50311
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1221.7  on 999  degrees of freedom
Residual deviance:  980.2  on 979  degrees of freedom
AIC: 1022.2

Number of Fisher Scoring iterations: 5
```

We have a Residual deviance of 980.197 on 979 degrees of freedom. We are fairly satisfied with this model as the ratio is close to 1.

We then fit a probit model with the same covariates. Summary is given below and we see that both the probit and logit models behave similarly and there's no gain from probit model.

```
Call:
glm(formula = cbind(credit, 1 - credit) ~ currentBalance + duration +
```

```
    paymentPrevious + use + maritalStatusGender, family = binomial(link = "probit"),
    data = data4)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6349  -0.8480   0.4486   0.7860   2.2342

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -0.532191   0.326309  -1.631 0.102903
currentBalance2       0.314881   0.116952   2.692 0.007094 **
currentBalance3       0.655792   0.199580   3.286 0.001017 **
currentBalance4       1.070966   0.120157   8.913  < 2e-16 ***
duration             -0.024938   0.003988  -6.253 4.03e-10 ***
paymentPrevious1      0.055405   0.292568   0.189 0.849800
paymentPrevious2      0.586619   0.231193   2.537 0.011169 *
paymentPrevious3      0.591788   0.264736   2.235 0.025392 *
paymentPrevious4      0.972122   0.242434   4.010 6.08e-05 ***
use1                  0.837387   0.190971   4.385 1.16e-05 ***
use2                  0.312661   0.140026   2.233 0.025556 *
use3                  0.466341   0.130483   3.574 0.000352 ***
use4                  0.363506   0.427882   0.850 0.395578
use5                  0.039271   0.303021   0.130 0.896885
use6                 -0.190677   0.216768  -0.880 0.379056
use8                  1.056011   0.616825   1.712 0.086895 .
use9                  0.340559   0.180152   1.890 0.058705 .
use10                 0.548818   0.419432   1.308 0.190711
maritalStatusGender2  0.025402   0.209916   0.121 0.903684
maritalStatusGender3  0.313079   0.204064   1.534 0.124976
maritalStatusGender4  0.149198   0.248093   0.601 0.547587
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1221.73  on 999  degrees of freedom
Residual deviance:  979.47  on 979  degrees of freedom
AIC: 1021.5

Number of Fisher Scoring iterations: 5

Analysis of Deviance Table

Model 1: cbind(credit, 1 - credit) ~ currentBalance + duration + paymentPrevious +
    use + maritalStatusGender
Model 2: cbind(credit, 1 - credit) ~ currentBalance + duration + paymentPrevious +
    use + maritalStatusGender
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       979     980.20
2       979     979.47  0  0.72265
```

To look for any evidence of overdispersion, we fit a quasi-likelihood model with beta-binomial like variance, which is summarized below.

```
Quasi-likelihood generalized linear model
-----------------------------------------
```

```
quasibin(formula = cbind(credit, 1 - credit) ~ currentBalance +
    duration + paymentPrevious + use + maritalStatusGender, data = data4,
    link = "logit")

Fixed-effect coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           -0.9752     0.5501 -1.7728   0.0763
currentBalance2        0.5081     0.1943  2.6148   0.0089
currentBalance3        1.1037     0.3438  3.2102   0.0013
currentBalance4        1.8325     0.2124  8.6275   < 1e-4
duration              -0.0424     0.0068 -6.1928   < 1e-4
paymentPrevious1       0.1160     0.4940  0.2347   0.8144
paymentPrevious2       1.0142     0.3916  2.5896   0.0096
paymentPrevious3       1.0438     0.4498  2.3206   0.0203
paymentPrevious4       1.6615     0.4131  4.0225   < 1e-4
use1                   1.3947     0.3339  4.1768   < 1e-4
use2                   0.5529     0.2374  2.3292   0.0198
use3                   0.8030     0.2241  3.5826   0.0003
use4                   0.5506     0.7296  0.7547   0.4504
use5                   0.1089     0.5080  0.2144   0.8302
use6                  -0.3094     0.3694 -0.8376   0.4023
use8                   1.8251     1.1350  1.6080   0.1078
use9                   0.6342     0.3097  2.0480   0.0406
use10                  1.0259     0.7124  1.4400   0.1499
maritalStatusGender2   0.0880     0.3513  0.2505   0.8022
maritalStatusGender3   0.5801     0.3421  1.6958   0.0899
maritalStatusGender4   0.2796     0.4175  0.6696   0.5031

Overdispersion parameter:
  phi
1e-04

Pearson's chi-squared goodness-of-fit statistic = 973.9329
```

We obtain a $\hat{\rho}$ value close to 0, indicating that we do not have overall overdispersion problem.

We then looked at interaction between our continuous variable given by duration of credit in months with a couple of covariates, namely, running account and payment of previous credits. Then we compare it with our original logistic model.

```
Call:
glm(formula = cbind(credit, 1 - credit) ~ currentBalance + duration +
    duration:currentBalance + duration:paymentPrevious + paymentPrevious +
    use + maritalStatusGender, family = binomial(link = "logit"),
    data = data4)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5645  -0.7887   0.4438   0.7382   2.2514

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               -1.74102    0.88237  -1.973 0.048482 *
currentBalance2            0.17760    0.41775   0.425 0.670748
currentBalance3            0.55766    0.74080   0.753 0.451583
currentBalance4            1.21641    0.44584   2.728 0.006365 **
```

```
duration                   -0.01649    0.02735  -0.603 0.546472
paymentPrevious1           -0.21119    0.99410  -0.212 0.831762
paymentPrevious2            2.18967    0.79171   2.766 0.005679 **
paymentPrevious3            1.59599    0.96114   1.661 0.096809 .
paymentPrevious4            3.38449    0.85017   3.981 6.86e-05 ***
use1                        1.53969    0.34784   4.426 9.58e-06 ***
use2                        0.60976    0.24349   2.504 0.012270 *
use3                        0.84897    0.22851   3.715 0.000203 ***
use4                        0.47389    0.72483   0.654 0.513246
use5                        0.17649    0.52314   0.337 0.735844
use6                       -0.24893    0.38026  -0.655 0.512703
use8                        2.15584    1.17009   1.842 0.065408 .
use9                        0.53036    0.31090   1.706 0.088027 .
use10                       0.76571    0.69400   1.103 0.269886
maritalStatusGender2        0.04926    0.35670   0.138 0.890170
maritalStatusGender3        0.56390    0.34659   1.627 0.103736
maritalStatusGender4        0.30880    0.42324   0.730 0.465628
currentBalance2:duration    0.01543    0.01663   0.928 0.353602
currentBalance3:duration    0.02984    0.03582   0.833 0.404872
currentBalance4:duration    0.03006    0.01801   1.669 0.095048 .
duration:paymentPrevious1   0.02139    0.03409   0.627 0.530382
duration:paymentPrevious2 -0.04627    0.02625  -1.763 0.077965 .
duration:paymentPrevious3 -0.02134    0.03140  -0.680 0.496654
duration:paymentPrevious4 -0.07223    0.02899  -2.492 0.012710 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1221.73  on 999  degrees of freedom
Residual deviance:  962.28  on 972  degrees of freedom
AIC: 1018.3

Number of Fisher Scoring iterations: 5

Analysis of Deviance Table

Model 1: cbind(credit, 1 - credit) ~ currentBalance + duration + paymentPrevious +
    use + maritalStatusGender
Model 2: cbind(credit, 1 - credit) ~ currentBalance + duration + duration:currentBalance +
    duration:paymentPrevious + paymentPrevious + use + maritalStatusGender
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       979     980.20
2       972     962.28  7   17.919  0.01234 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see that this model performs slightly better than our original model with Devianc of 17.919 on loss of 7 degrees of freedom.

5 Problem 5, 14.7 from Agresti

5.1

We fit the model

$$logit(\pi_{it}) = \beta_1 z_1 + \beta_2 z_2 + \beta_3 z3$$

where $z_t = 1$ for treatment t and 0 else, summary of which is given below:

```
Call:
glm(formula = cbind(hatched, total - hatched) ~ treatment - 1,
    family = binomial(link = "logit"), data = data5)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.8809  -0.4949  -0.0002   0.9663   3.1039

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
treatment1  -20.3852  1580.7240  -0.013   0.9897
treatment2    0.4055     0.1992   2.035   0.0418 *
treatment3   -0.2103     0.1963  -1.072   0.2839
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 232.261  on 21  degrees of freedom
Residual deviance:  81.317  on 18  degrees of freedom
AIC: 117.39

Number of Fisher Scoring iterations: 17
```

We see that th residual deviance is 81.317 on 18 degrees of freedom. So, we have a overdispersion problem.

Also, we see that $\hat{\beta}_1$ is -20.385. Since treatment 1 has no success for at all, $\hat{\beta}_1$ should be $-\infty$.

Since $\hat{\beta}_1$ is $-\infty$ becuase of zero probability for all in treatment 1, we can do a bayesian inference by putting beta prior spiked close to 0 for π_i for treatment 1. It'll help us avoid the negative infinity in posterior and make ML estimation possible. Cons of this remedy is that it'll induce bias and also that we're ignoring the case when it might acutally be 0.

5.2

To allow for overdispersion, we used the quasi-likelihood method with beta-binomial type variance. The summary of which is below.

```
Quasi-likelihood generalized linear model
-----------------------------------------
quasibin(formula = cbind(hatched, total - hatched) ~ treatment -
    1, data = data5, link = "logit")

Fixed-effect coefficients:
           Estimate Std. Error z value Pr(>|z|)
treatment1 -20.1868  2936.8224 -0.0069   0.9945
treatment2   0.3716     0.4074  0.9121   0.3617
treatment3  -0.3806     0.4077 -0.9334   0.3506
```

```
Overdispersion parameter:
   phi
0.2146

Pearson's chi-squared goodness-of-fit statistic = 18.0004
```

We see that the overdispersion parameter is 0.21, which is greater significantly greater than 0 which is an evidence of the presence of overdispersion in the data.
If we look at the SE values for both treatment 2 and treatment 3, we see that they are inflated when compared to the SE values we obtained in previous part with binomial maximum likelihood. This shows the evidence of overdispersion for both treatments 2 and 3.

5.3

Finally we fit the beta-binomial regression model on the data which allows for overdispersion. We fitted two different models with same ϕ for all groups and different ϕ_i for each treatment group. The summary for both is given below.

```
Beta-binomial model
-------------------
betabin(formula = cbind(hatched, total - hatched) ~ treatment -
    1, random = ~1, data = data5)

Convergence was obtained after 121 iterations.

Fixed-effect coefficients:
             Estimate Std. Error    z value Pr(> |z|)
treatment1 -1.980e+01  3.425e+03 -5.780e-03 9.954e-01
treatment2  2.514e-01  4.747e-01  5.297e-01 5.963e-01
treatment3 -5.285e-01  4.850e-01 -1.090e+00 2.759e-01

Overdispersion coefficients:
                 Estimate Std. Error   z value   Pr(> z)
phi.(Intercept) 3.305e-01  1.086e-01 3.044e+00 1.169e-03

Log-likelihood statistics
   Log-lik      nbpar    df res.   Deviance        AIC       AICc
-3.621e+01          4         17  4.235e+01  8.042e+01  8.292e+01

Beta-binomial model
-------------------
betabin(formula = cbind(hatched, total - hatched) ~ treatment -
    1, random = ~treatment, data = data5)

Convergence was obtained after 171 iterations.

Fixed-effect coefficients:
             Estimate Std. Error    z value Pr(> |z|)
treatment1 -2.009e+01  6.043e+03 -3.325e-03 9.973e-01
treatment2  2.483e-01  4.806e-01  5.165e-01 6.055e-01
treatment3 -5.285e-01  4.827e-01 -1.095e+00 2.735e-01

Overdispersion coefficients:
                Estimate Std. Error   z value   Pr(> z)
phi.treatment1 3.820e-01  2.594e+03 1.473e-04 4.999e-01
```

```
phi.treatment2 3.368e-01  1.582e-01 2.129e+00 1.664e-02
phi.treatment3 3.248e-01  1.490e-01 2.180e+00 1.465e-02

Log-likelihood statistics
   Log-lik      nbpar    df res.   Deviance        AIC       AICc
-3.621e+01          6         15  4.235e+01  8.442e+01  9.042e+01
```

We can conclude that treatment 2 allows for higher probability of hatched eggs than treatment 3.