Homework 3

Shun-Lung Chang, Dilip Hiremath

```
load('data/bankmarketing.Rdata')
```

1. Create a logistic regression model to predict the efficiency of the marketing campaign using y as dependent variable and no predictor, just the intercept.

```
mod_1 <- glm(y ~ 1, data = bankmarketing, family = binomial(link = "logit"))
summary(mod_1)
```

```
Call:
glm(formula = y ~ 1, family = binomial(link = "logit"), data = bankmarketing)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.4949  -0.4949  -0.4949  -0.4949   2.0788

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.03830    0.04658  -43.76   <2e-16

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3231  on 4520  degrees of freedom
Residual deviance: 3231  on 4520  degrees of freedom
AIC: 3233

Number of Fisher Scoring iterations: 4
```

(a) (half a point) How large is the AIC score for this model?

```
AIC(mod_1)
```

```
[1] 3233
```

(b) (half a point) How large is the BIC score for this model?

```
BIC(mod_1)
```

```
[1] 3239.417
```

(c) (1 point) Compute the log odds for the mean response and compare it to the coefficient estimate for the intercept in this model [hint: Be careful about the internal numeric coding of the variable y in the data set].

```
log(sum(bankmarketing$y == 'yes') / sum(bankmarketing$y == 'no'))
```

```
[1] -2.0383
```

2. Add duration as a predictor to the model.

(a) (half a point) In comparison to the naive model, by how much has the AIC changed?

(b) (half a point) Is last contact duration (duration) a significant predictor in this model for modeling the probability of subscribing to a term deposit?

(c) (half a point) With growing duration of the last contact are clients less likely or are they more likely to subscribe to a term deposit?

(d) (half a point) According to the second model, what is the estimated probability of subscribing to a term deposit for a client who immediately terminated the last contact (i.e. duration equals 0)?

3. Using the second model with duration as predictor,

(a) (1 point) compute the halfway point, i.e the last contact duration at which the estimated probability of subscribing to a term deposit equals 0.5

(b) (1 point) Compute the slope of the tangent to the regression curve at the halfway point.

4. Compute a logistic regression model for subscribing to a term deposit using age, marital and duration as predictors.

(a) (1 point) Calculate the predictive probability of subscribing to a term deposit for a married client at the mean values of the numeric predictors.

(b) In the above model, calculate the effect on the probability of subscribing when keeping all other predictors constant and

i. (half a point) changing duration from the mean score to 300 seconds;

ii. (half a point) changing age from the mean score to one standard deviation above the mean score.

5. (2 points) Compute a logistic regression model for the decision to subscribe to a term deposit using duration, campaign, and the interaction between the two. How do you interpret the regression coefficients? Are these interpretations meaningful? Give reasons for your answer!

6. (2 points) Center the variables duration and campaign and re-build the model built in ques- tion 5. How do you interpret the regression coefficients? Are these interpretations meaningful? Give reasons for your answer! Draw the effects plot for this model and interpret!

7. (2 points) Create a logistic regression model using y as dependent variable and all available predictors, except day and month. Which predictors are significant? Do the estimated coefficients make common sense to you?

8. (2 points) Starting with the model of Question 7 use the automatic backward/forward selec- tion method to derive a suitable model. Report the significant predictors and the AIC score of the resulting model.

9. (2 points) Draw a box plot of the residuals and look for extreme outliers. Remove the outlier and re-run the model you have obtained in Question 8. Which changes in the model are to be noted?