

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?

The AIC score is 3233.

```
AIC(mod_1)
```

```
[1] 3233
```

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

The BIC score is 3239.417.

```
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].

The log odds is computed as follows and the value is -2.0383, which is the same as the intercept in the model.

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

```
[1] -2.0383
```

2. Add duration as a predictor to the model.

```
mod_2 <- glm(y ~ duration, data = bankmarketing, family = binomial(link = 'logit'))
summary(mod_2)
```

Call:

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

Deviance Residuals:

Min	1Q	Median	3Q	Max
-3.8683	-0.4303	-0.3548	-0.3106	2.5264

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-3.2559346	0.0845767	-38.50	<2e-16
duration	0.0035496	0.0001714	20.71	<2e-16

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 3231.0 on 4520 degrees of freedom
Residual deviance: 2701.8 on 4519 degrees of freedom
AIC: 2705.8

Number of Fisher Scoring iterations: 5

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

The AIC score drops by 527.2476 after *duration* is added to the model.

```
AIC(mod_2) - AIC(mod_1)
```

```
[1] -527.2476
```

(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?

The small p-value implies that the variable *duration* is significant.

(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?

The positive coefficient suggests that customer with larger duration tend to have 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)?

The probability is 0.037 if a client with 0 duration.

```
predict(mod_2, data.frame(duration = 0), type = 'response')
```

```
1
0.03711422
```

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.

Based on our model, $\ln(\frac{p}{1-p}) = \hat{\beta}_0 + \hat{\beta}_1 * duration$, if the estimated probability is 0.5, then we have $\hat{\beta}_0 + \hat{\beta}_1 * duration = 0$. Therefore, the duration at the halfway point is $-\frac{\hat{\beta}_0}{\hat{\beta}_1}$, and the value in this model is 917.2802.

```
d = unname(-mod_2$coefficients[1] / mod_2$coefficients[2])
d
```

```
[1] 917.2802
```

We can as well predict the estimated probability when duration is 917.2802. As expected, the probability is 0.5.

```
predict(mod_2, data.frame(duration = d), type = 'response')
```

```
1
0.5
```

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

A straight forward way to compute the slope of a function is to find its derivative. Also, the derivative of the model, $P(y = 1) = \frac{1}{1+e^{-(\beta_0+\beta_1*X)}}$, is $\beta_1 * \frac{e^{-(\beta_0+\beta_1*X)}}{(1+e^{-(\beta_0+\beta_1*X)})^2} = \beta_1 * P(y = 1) * (1 - P(y = 1))$. According to this formula, we know the slope at the halfway point ($P(y = 1) = 0.5$) is 0.0008873882.

```
unname(mod_2$coefficients[2]) * 0.5 * (1 - 0.5)
```

```
[1] 0.0008873882
```

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

```
mod_3 <- glm(y ~ age + marital + duration, data = bankmarketing,
             family = binomial(link = 'logit'))
summary(mod_3)
```

Call:

```
glm(formula = y ~ age + marital + duration, family = binomial(link = "logit"),
    data = bankmarketing)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-3.9035	-0.4360	-0.3542	-0.2943	2.6029

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-3.818323	0.211710	-18.036	< 2e-16
age	0.022285	0.005068	4.397	0.00001095
maritaldivorced	-0.230097	0.177829	-1.294	0.196
maritalmarried	-0.568236	0.127543	-4.455	0.00000838
duration	0.003534	0.000172	20.548	< 2e-16

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 3231.0 on 4520 degrees of freedom
Residual deviance: 2670.8 on 4516 degrees of freedom
AIC: 2680.8

Number of Fisher Scoring iterations: 5

(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.

Given the conditions, the predictive probability is 0.0734.

```
predict(mod_3, data.frame(age = mean(bankmarketing$age),
                             marital = 'married',
                             duration = mean(bankmarketing$duration)),
        type = 'response')
```

```
1
0.07335345
```

(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;

The estimated probability is increased to 0.0825 as the duration is 300.

```
predict(mod_3, data.frame(age = mean(bankmarketing$age),
                             marital = 'married',
                             duration = 300),
        type = 'response')
```

```
1
0.08249455
```

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

The estimated probability is increased to 0.0911 if the age is one standard deviation above the mean value.

```
predict(mod_3, data.frame(age = mean(bankmarketing$age) +
                                sd(bankmarketing$age),
                            marital = 'married',
                            duration = mean(bankmarketing$duration)),
        type = 'response')
```

```
1
0.09107385
```

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!

The following table indicates that when duration is increased by one unit, the change of odds for the event y is 'yes' is $\exp(0.00277087 + 0.00032618 \times \text{campaign})$. On the other hand, if number of campaign is increased by one, the change of odds is $\exp(-0.27602084 + 0.00032618 \times \text{duration})$. Because of the small p-value, we can conclude that there exists a significant interaction between the two variables.

```
mod_4 <- glm(y ~ duration + campaign + duration * campaign,
             data = bankmarketing,
             family = binomial(link = 'logit'))
summary(mod_4)
```

Call:

```
glm(formula = y ~ duration + campaign + duration * campaign,
    family = binomial(link = "logit"), data = bankmarketing)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-3.7907	-0.4464	-0.3683	-0.2786	2.7341

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.61917477	0.14105246	-18.569	< 2e-16
duration	0.00277087	0.00026692	10.381	< 2e-16
campaign	-0.27602084	0.05590684	-4.937	0.000000793
duration:campaign	0.00032618	0.00009033	3.611	0.000305

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 3231.0 on 4520 degrees of freedom
Residual deviance: 2664.1 on 4517 degrees of freedom
AIC: 2672.1

Number of Fisher Scoring iterations: 6

6. (2 points) Center the variables *duration* and *campaign* and re-build the model built in question 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!

As the table below shows, the coefficient of the interaction term remains the same although the variables are centered. And the effects plot shows that as the number of campaign increased by 10, the coefficient of variable *duration* will increase by 0.005.

```
c_duration <- scale(bankmarketing$duration, scale = FALSE)
c_campaign <- scale(bankmarketing$campaign, scale = FALSE)

mod_5 <- glm(bankmarketing$y ~ c_duration + c_campaign + c_duration * c_campaign,
             family = binomial(link = 'logit'))
summary(mod_5)
```

Call:

```
glm(formula = bankmarketing$y ~ c_duration + c_campaign + c_duration *
     c_campaign, family = binomial(link = "logit"))
```

Deviance Residuals:

	Min	1Q	Median	3Q	Max
	-3.7907	-0.4464	-0.3683	-0.2786	2.7341

Coefficients:

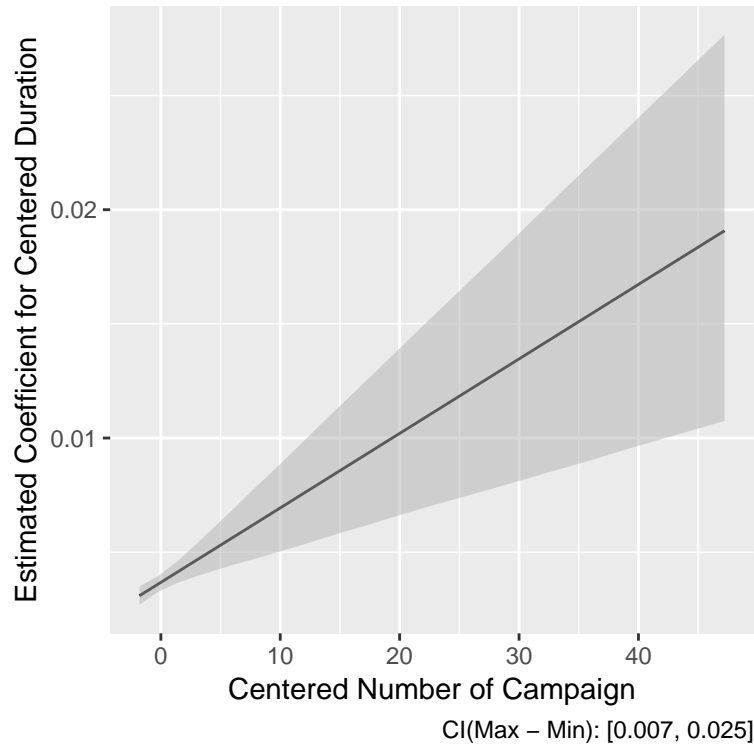
	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.41834174	0.06419453	-37.672	< 2e-16
c_duration	0.00368210	0.00018080	20.366	< 2e-16
c_campaign	-0.18992117	0.03835754	-4.951	0.000000737
c_duration:c_campaign	0.00032618	0.00009033	3.611	0.000305

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 3231.0 on 4520 degrees of freedom
 Residual deviance: 2664.1 on 4517 degrees of freedom
 AIC: 2672.1

Number of Fisher Scoring iterations: 6

```
library(interplot)
interplot(m = mod_5, var1 = 'c_duration', var2 = 'c_campaign') +
  xlab("Centered Number of Campaign") +
  ylab("Estimated Coefficient for Centered Duration")
```



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?

Under significance level of 5%, the significant variables are *housing*, *loan*, *contact (if unknown)*, *duration*, *campaign*, and *poutcome (if success)*.

The coefficients of these variables seem valid. For the variables with a positive sign: duration, poutcome (if success), if a client's last contract was longer, it is more possible for him/her to have a term deposit since he/she might be satisfied with contract. We can also understand that a positive previous campaign outcome will lead to higher chance for a client to make a term deposit.

On the other hand, if people have a housing or personal load, it is anticipatable that they would not make a term deposit. Also, if the contact communication type is unknown, we may assume that the campaign was not made, and hence the client did not subscribe a term deposit. Interestingly, a higher number of campaign contacts will prevent people from subscribing a term deposit in that people might feel irritated if they had frequent contacts with campaign representatives.

```
mod_6 <- glm(y ~ . - day - month, data = bankmarketing,
             family = binomial(link = 'logit'))
summary(mod_6)
```

Call:

```
glm(formula = y ~ . - day - month, family = binomial(link = "logit"),
    data = bankmarketing)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-4.0779	-0.4067	-0.2785	-0.1705	3.0457

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.58481120	0.49297812	-5.243	0.000000157766
age	0.00051541	0.00687649	0.075	0.940252
jobblue-collar	-0.45713374	0.23660305	-1.932	0.053351
jobentrepreneur	-0.46289495	0.37353608	-1.239	0.215262
jobhousemaid	-0.32234424	0.39517541	-0.816	0.414672
jobmanagement	-0.11365818	0.23572561	-0.482	0.629690
jobretired	0.58415128	0.29855600	1.957	0.050396
jobself-employed	-0.24456557	0.34018308	-0.719	0.472188
jobservices	-0.23224216	0.26610659	-0.873	0.382804
jobstudent	0.51174726	0.36139116	1.416	0.156761
jobtechnician	-0.26279749	0.22354603	-1.176	0.239761
jobunemployed	-0.69747294	0.41220863	-1.692	0.090639
jobunknown	0.46787823	0.54422750	0.860	0.389948
maritaldivorced	0.19815965	0.19766206	1.003	0.316094
maritalmarried	-0.23249919	0.14359896	-1.619	0.105429
educationsecondary	0.07954357	0.19583232	0.406	0.684609
educationtertiary	0.36044330	0.22596527	1.595	0.110684
educationunknown	-0.34886940	0.33917035	-1.029	0.303669
defaultyes	0.46738306	0.42123528	1.110	0.267192
balance	0.00000330	0.00001756	0.188	0.850967
housingyes	-0.47768772	0.12435061	-3.841	0.000122
loanyes	-0.76137581	0.19372118	-3.930	0.000084851850
contacttelephone	-0.03948148	0.22245467	-0.177	0.859131
contactunknown	-1.09294023	0.17621602	-6.202	0.000000000557
duration	0.00404244	0.00019582	20.644	< 2e-16
campaign	-0.07430257	0.02700051	-2.752	0.005925
pdays	-0.00014347	0.00093454	-0.154	0.877991
previous	-0.00473025	0.03806701	-0.124	0.901109
poutcomesuccess	2.40156770	0.26266846	9.143	< 2e-16
poutcomeother	0.46819766	0.25950506	1.804	0.071201
poutcomeunknown	-0.28114500	0.30276181	-0.929	0.353096

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 3231.0 on 4520 degrees of freedom
Residual deviance: 2277.2 on 4490 degrees of freedom
AIC: 2339.2

Number of Fisher Scoring iterations: 6

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

If we consider a significance level of 5%, the significant predictors of the final model are *job* (if retired), *housing*, *loan*, *contact* (if unknown), *duration*, *campaign*, and *poutcome* (if success). Additionally, the model's AIC score is 2330.4.

```
mod_7 <- step(mod_6, direction = 'both')
```



```
summary(mod_7)
```

Call:

```
glm(formula = y ~ job + marital + education + housing + loan +  
     contact + duration + campaign + poutcome, family = binomial(link = "logit"),  
     data = bankmarketing)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-4.0745	-0.4072	-0.2788	-0.1704	3.0496

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.6139848	0.3255815	-8.029	9.85e-16
jobblue-collar	-0.4494496	0.2363066	-1.902	0.05717
jobentrepreneur	-0.4300067	0.3706045	-1.160	0.24593
jobhousemaid	-0.3032897	0.3920882	-0.774	0.43921
jobmanagement	-0.1039637	0.2354379	-0.442	0.65880
jobretired	0.6054857	0.2691559	2.250	0.02448
jobself-employed	-0.2343157	0.3400443	-0.689	0.49078
jobservices	-0.2283782	0.2659796	-0.859	0.39054
jobstudent	0.5139966	0.3550967	1.447	0.14776
jobtechnician	-0.2509853	0.2231478	-1.125	0.26070
jobunemployed	-0.6949974	0.4126048	-1.684	0.09210
jobunknown	0.4676864	0.5433718	0.861	0.38940
maritaldivorced	0.2094087	0.1861087	1.125	0.26051
maritalmarried	-0.2297354	0.1329582	-1.728	0.08401
educationsecondary	0.0846584	0.1943381	0.436	0.66311
educationtertiary	0.3604550	0.2236398	1.612	0.10701
educationunknown	-0.3377539	0.3378899	-1.000	0.31751
housingyes	-0.4801920	0.1217720	-3.943	8.03e-05
loanyes	-0.7579438	0.1933619	-3.920	8.86e-05
contacttelephone	-0.0411671	0.2194125	-0.188	0.85117
contactunknown	-1.0960326	0.1761025	-6.224	4.85e-10
duration	0.0040386	0.0001956	20.650	< 2e-16
campaign	-0.0749988	0.0269881	-2.779	0.00545
poutcomesuccess	2.4100962	0.2559920	9.415	< 2e-16
poutcomeother	0.4772387	0.2573995	1.854	0.06373
poutcomeunknown	-0.2275282	0.1732319	-1.313	0.18904

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 3231.0 on 4520 degrees of freedom
Residual deviance: 2278.4 on 4495 degrees of freedom
AIC: 2330.4

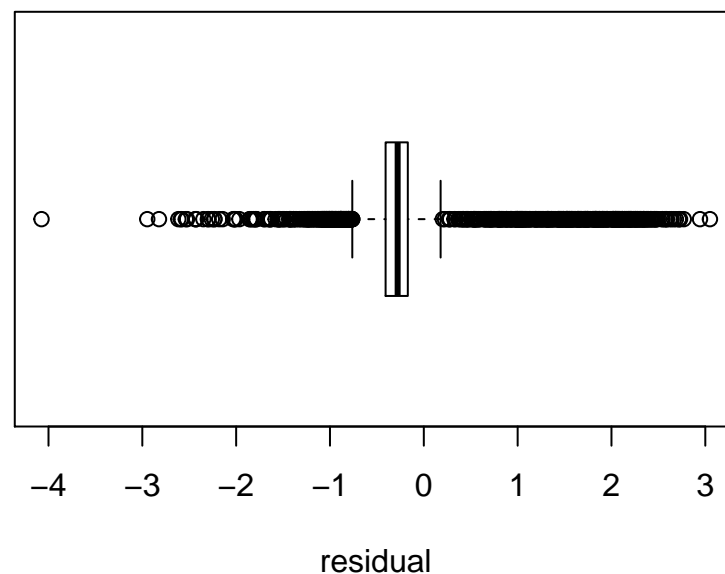
Number of Fisher Scoring iterations: 6

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?

The boxplot shows the distribution of the residuals of the model obtained in Task 8. In this task, we define outliers as the residuals with a absolute value greater than 2.8. By doing so, a new model could be obtained after we removed the outliers. The AIC score was reduced from 2330.4 to 2277.3; the model was improved by removing the outliers.

```
res <- residuals(mod_7)
boxplot(res, horizontal = TRUE,
        main = 'Boxplot of Residuals',
        xlab = 'residual')
```

Boxplot of Residuals



```
bankmarketing_2 <- subset(bankmarketing, !abs(res) > 2.8)
mod_8 <- glm(formula = y ~ job + marital + education +
             housing + loan + contact + duration + campaign +
             poutcome,
             family = binomial(link = "logit"),
             data = bankmarketing_2)
summary(mod_8)
```

Call:

```
glm(formula = y ~ job + marital + education + housing + loan +
    contact + duration + campaign + poutcome, family = binomial(link = "logit"),
    data = bankmarketing_2)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.7266	-0.3998	-0.2701	-0.1596	2.8264

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.6136506	0.3288497	-7.948	0.0000000000000019
jobblue-collar	-0.4776930	0.2399265	-1.991	0.04648
jobentrepreneur	-0.4600744	0.3759706	-1.224	0.22107
jobhousemaid	-0.4766016	0.4060914	-1.174	0.24054
jobmanagement	-0.1126178	0.2382413	-0.473	0.63642
jobretired	0.5803876	0.2726283	2.129	0.03327
jobself-employed	-0.2566454	0.3446276	-0.745	0.45645
jobservices	-0.2348513	0.2698859	-0.870	0.38420
jobstudent	0.5006036	0.3587605	1.395	0.16290
jobtechnician	-0.2632163	0.2262383	-1.163	0.24465
jobunemployed	-0.6014264	0.4059984	-1.481	0.13851
jobunknown	0.4494780	0.5505753	0.816	0.41428
maritaldivorced	0.2019291	0.1884468	1.072	0.28392
maritalmarried	-0.2381778	0.1343801	-1.772	0.07632
educationsecondary	0.0377348	0.1960466	0.192	0.84737
educationtertiary	0.3306011	0.2255193	1.466	0.14266
educationunknown	-0.3908434	0.3418143	-1.143	0.25286
housingyes	-0.5076566	0.1233136	-4.117	0.0000384180755829
loanyes	-0.8159104	0.1981440	-4.118	0.0000382565339868
contacttelephone	-0.0469475	0.2224582	-0.211	0.83286
contactunknown	-1.1777561	0.1811436	-6.502	0.0000000000793740
duration	0.0042669	0.0002019	21.134	< 2e-16
campaign	-0.0764423	0.0273166	-2.798	0.00514
poutcomesuccess	2.4585831	0.2587358	9.502	< 2e-16
poutcomeother	0.4752446	0.2601408	1.827	0.06772
poutcomeunknown	-0.2319531	0.1749554	-1.326	0.18491

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 3221.6 on 4515 degrees of freedom
Residual deviance: 2225.3 on 4490 degrees of freedom
AIC: 2277.3

Number of Fisher Scoring iterations: 6