# Logistic Regression Exmaple

**1.** Suppose that the marketing department for a credit card company is about to embark on a campaign to convince existing holders of the company's standard credit card to upgrade to one of the company's premium cards for a nominal annual fee. The major decision facing the marketing department concerns which of the existing standard credit card holders should be targeted for the campaign. Data available from a sample of 30 credit card holders who were contacted during last year's campaign indicate the following information: whether the credit card holder upgraded for the standard to a premium card (`premium`: 0 = no, 1 = yes), the total amount of credit card purchases (in thousands of dollars) using the company's credit card in the 1 year prior to the campaign (`spend`), and whether the credit card holder possessed additional credit cards (which involved an extra cost) for other members of the household (`addcard`: 0 = no, 1 = yes). The data can be found in `credit.csv`.

```
credit <- read.csv("credit.csv")
library(ggplot2)
library(broom)
library(boot)
library(pROC)
```

```
## Warning: package 'pROC' was built under R version 3.1.3
```

```
## Type 'citation("pROC")' for a citation.
##
## Attaching package: 'pROC'
##
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
head(credit)
```

```
##   premium   spend addcard
## 1       0 32.1007       0
## 2       1 34.3706       1
## 3       0  4.8749       0
## 4       0  8.1263       0
## 5       0 12.9783       0
## 6       0 16.0471       0
```

[**a**] Fit a logistic regression model to predict the probability of a credit card holder upgrading to a premium card based on the total amount of credit card purchases using the company's credit card in the 1 year prior to the campaign and whether the credit card holder possessed additional credit cards for other members of the household. Is total amount of credit card purchases an important indicator of choice to upgrade? Is possession of additional credit cards an important indicator of choice to upgrade?

```
fit <- glm(premium ~ spend + addcard, data = credit, family = binomial)
tidy(fit)
```

```
##          term   estimate  std.error statistic    p.value
## 1 (Intercept) -6.9398388 2.94722315 -2.354704 0.01853745
## 2       spend  0.1394685 0.06806604  2.049018 0.04046036
## 3     addcard  2.7743352 1.19269821  2.326100 0.02001322
```

```
glance(fit)
```

```
##   null.deviance df.null    logLik     AIC      BIC deviance df.residual
## 1      41.05391      29 -10.03845 26.0769 30.28049  20.0769          27
```

[**b**] Predict the probability that a credit card holder will upgrade to a premium card if he/she has purchased additional cards for members of the household, and used the company's card to charge $36,000 last year.

```
predict(fit, data.frame(spend = 36, addcard = 1), type = "response")
```

```
##         1
## 0.7016911
```

[**c**] Predict the probability that a credit card holder will upgrade to a premium card if he/she has NOT purchased additional cards for members of the household, and used the company's card to charge $36,000 last year.

```
predict(fit, data.frame(spend = 36, addcard = 0), type = "response")
```

```
##         1
## 0.1279763
```

[**d**] Predict the probability that a credit card holder will upgrade to a premium card if he/she has purchased additional cards for members of the household, and used the company's card to charge $50,000 last year.

```
predict(fit, data.frame(spend = 50, addcard = 1), type = "response")
```

```
##         1
## 0.9431025
```

[**e**] At the 0.10 level of significance, is there evidence that a logistic regression model that uses the total amount of credit card purchases using the company's credit card in the 1 year prior to the campaign and whether the credit card holder possessed additional credit cards for other members of the household to predict probability of upgrading to a premium card is a good fitting model? What is the p-value for this test?

```r
with(fit, pchisq(deviance, df.residual, lower.tail = FALSE))
```

```
## [1] 0.8275135
```

$$Deviance = -2\log(likelihood)$$

$$P(Y_i = 0) = 1 - p_i \qquad P(Y_i = 1) = p_i$$

$$likelihood = \prod_{i=1}^{n} p_i^{Y_i}(1 - p_i)^{1-Y_i}$$

Deviance is approximately $\chi^2_{n-p}$, where $p$ is the number of parameters.

$Deviance_{reduced} - Deviance_{full} \sim \chi^2_{q-q}$ where $q$ is the number of parameters in the reduced model.

```r
with(fit, null.deviance - deviance)
```

```
## [1] 20.977
```

```r
with(fit, df.null - df.residual)
```

```
## [1] 2
```

```r
with(fit, pchisq(null.deviance - deviance, df.null - df.residual, lower.tail = FALSE))
```

```
## [1] 2.785488e-05
```