Stat 415/615, Lab 8. Introduction to Logistic Regression

Jun Lu

Stat 415/615 Regression, 2023

Contents

Comments and explanations are not included here. We'll discuss them in class.

1 Coupon Effectiveness (Binomial response, counts)

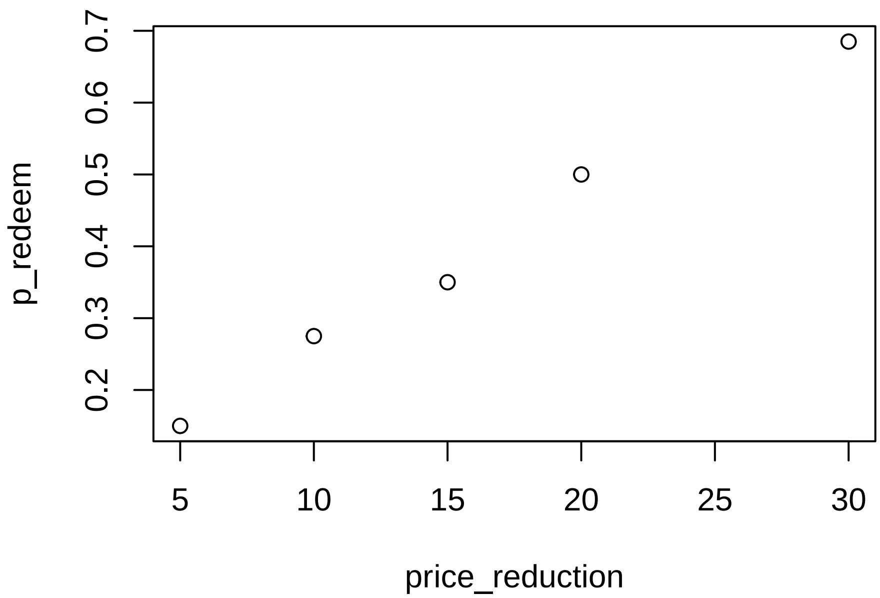
Refer to the Coupon Effectiveness example from textbook p.569. The study is to evaluate how the amount of price deduction (X) may affect the households' purchasing choice. The data set is in file CH14TA02.txt and CouponEffectiveness.sav.

```
coupon <- read.table("../DataSets/CH14TA02.txt", header=F)
colnames(coupon) <- c("price_reduction", "n_house", "n_redeem", "p_redeem")
coupon
```

```
##   price_reduction n_house n_redeem p_redeem
## 1               5     200       30    0.150
## 2              10     200       55    0.275
## 3              15     200       70    0.350
## 4              20     200      100    0.500
## 5              30     200      137    0.685
```

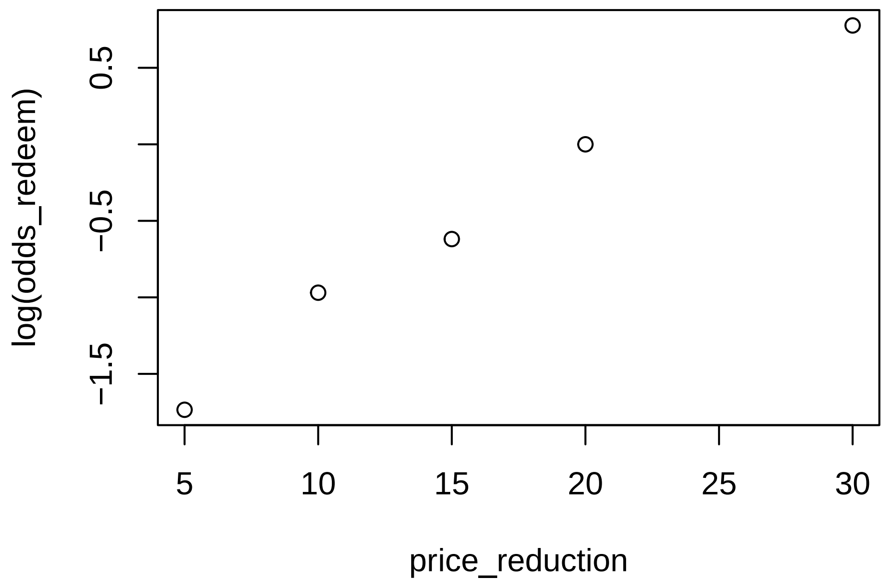
1.1 Plot the proportion of redeemed coupons against the price deduction and comment.

```
plot(p_redeem ~ price_reduction, data=coupon)
```



1.2 Transform the proportion into log(odds), and plot it against the price deduction and comment.

```r
coupon$odds_redeem <- coupon$p_redeem/(1-coupon$p_redeem)
plot(log(odds_redeem) ~ price_reduction, data=coupon)
```



1.3 Fit a linear regression model on log(odds) vs (price reduction). Plot the residuals and comment.

```r
coupon.lm <- lm(log(odds_redeem)~price_reduction, data=coupon)
summary(coupon.lm)
```

```
##
## Call:
## lm(formula = log(odds_redeem) ~ price_reduction, data = coupon)
##
## Residuals:
##        1        2        3        4        5
## -0.14344  0.12998 -0.01144  0.11581 -0.09091
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -2.082943   0.132751  -15.69 0.000563 ***
## price_reduction  0.098356   0.007308   13.46 0.000887 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1406 on 3 degrees of freedom
```

```
## Multiple R-squared:  0.9837, Adjusted R-squared:  0.9783
## F-statistic: 181.2 on 1 and 3 DF,  p-value: 0.0008868
```

```
plot(coupon$price_reduction, coupon.lm$residuals)
```

1.4 Discuss: Why don't we just fit a linear regression between log(odds) and X?

1.5 Analyze the data using logistic regression. (Note that this the Binomial data)

There are two ways to specify the model for Binomial logistic model in R.

```
coupon.logi1<-glm(n_redeem/n_house ~ price_reduction,
                  family=binomial(link=logit), weight=n_house,
                  data=coupon)
summary(coupon.logi1)
```

```
##
```

```
## Call:
## glm(formula = n_redeem/n_house ~ price_reduction, family = binomial(link = logit),
##     data = coupon, weights = n_house)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -2.044348   0.160977  -12.70   <2e-16 ***
## price_reduction  0.096834   0.008549   11.33   <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: 149.4627  on 4  degrees of freedom
## Residual deviance:   2.1668  on 3  degrees of freedom
## AIC: 33.793
##
## Number of Fisher Scoring iterations: 3
```

```
coupon.logi2<-glm(cbind(n_redeem, n_house - n_redeem) ~ price_reduction,
                  family=binomial(link=logit),
                  data=coupon)
summary(coupon.logi2)
```

```
##
## Call:
## glm(formula = cbind(n_redeem, n_house - n_redeem) ~ price_reduction,
##     family = binomial(link = logit), data = coupon)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -2.044348   0.160977  -12.70   <2e-16 ***
## price_reduction  0.096834   0.008549   11.33   <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: 149.4627  on 4  degrees of freedom
## Residual deviance:   2.1668  on 3  degrees of freedom
## AIC: 33.793
##
## Number of Fisher Scoring iterations: 3
```

1.6 Interpret the estimated logistic regression model (slope) in the context of the problem.

```
coupon.logi1
```

```
##
## Call:  glm(formula = n_redeem/n_house ~ price_reduction, family = binomial(link = logit),
##     data = coupon, weights = n_house)
##
## Coefficients:
##     (Intercept)  price_reduction
##        -2.04435          0.09683
##
## Degrees of Freedom: 4 Total (i.e. Null);  3 Residual
## Null Deviance:       149.5
## Residual Deviance: 2.167     AIC: 33.79
```

2 Problem 2: Disease outbreak (0-1 response)

Refer to Appendix C, Data Set C.10. In a health study to investigate an epidemic outbreak of a disease that is spread by mosquitoes, individuals were randomly sampled within two sectors in a city to determine if the person had recently contracted the disease under study. Data provides information about 196 persons selected in a probability sample within two sectors in a city. Data file `APPENC10.txt` is on Blackboard. Variables are:

- ID
- Age of person in years
- Socioeconomic status: 1= upper; 2 = middle; 3 = lower
- Sector: 1 = sector 1; 2 = sector 2
- Disease status: 1 = with disease; 0 = without
- Savings account: 1 = has savings account; 0 = without savings account

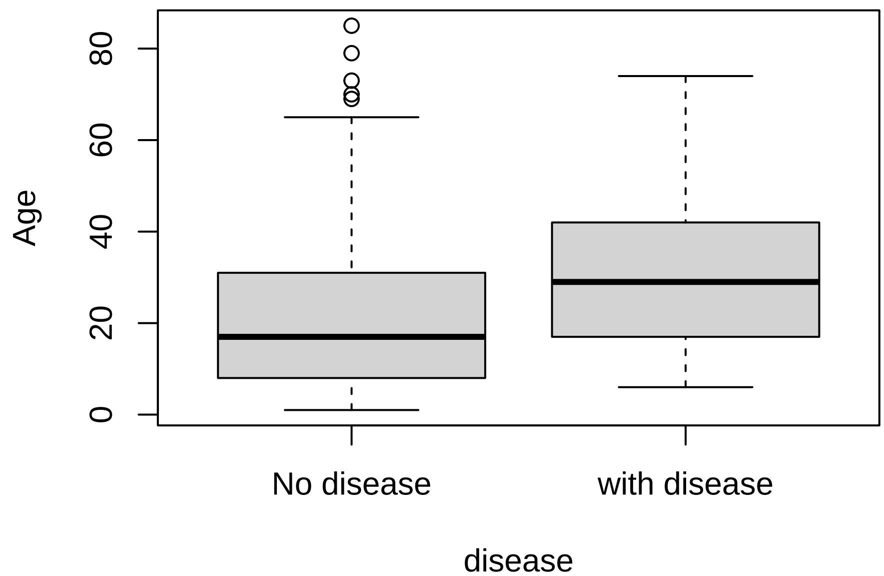
The primary purpose of the study was to assess the strength of the association between each of the predictor variables and the probability of a person having contracted the disease.

What will you do to analyze this data? Lay out a plan and consider "what questions may be of interest in this context?"

```r
outbreak <- read.table("../DataSets/APPENC10.txt", header=F)
colnames(outbreak) <- c("ID", "age", "ses", "sec", "disease", "savings")
summary(outbreak)
```

```
##        ID              age             ses             sec       
##  Min.   :  1.00   Min.   : 1.00   Min.   :1.000   Min.   :1.000  
##  1st Qu.: 49.75   1st Qu.:10.75   1st Qu.:1.000   1st Qu.:1.000  
##  Median : 98.50   Median :21.00   Median :2.000   Median :1.000  
##  Mean   : 98.50   Mean   :25.18   Mean   :1.964   Mean   :1.403  
##  3rd Qu.:147.25   3rd Qu.:35.00   3rd Qu.:3.000   3rd Qu.:2.000  
##  Max.   :196.00   Max.   :85.00   Max.   :3.000   Max.   :2.000  
##     disease          savings      
##  Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.:0.0000  
##  Median :0.0000   Median :1.0000  
##  Mean   :0.2908   Mean   :0.5459  
##  3rd Qu.:1.0000   3rd Qu.:1.0000  
##  Max.   :1.0000   Max.   :1.0000  
```

```r
boxplot(age~disease,names=c("No disease","with disease"),
        ylab="Age", data=outbreak)
```




```r
outbreak.logi<-glm(disease ~ age + as.factor(ses) + as.factor(sec),
                   family=binomial(link=logit), data=outbreak)
summary(outbreak.logi)
```

```
##
## Call:
## glm(formula = disease ~ age + as.factor(ses) + as.factor(sec),
##     family = binomial(link = logit), data = outbreak)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -2.293933   0.436769  -5.252  1.5e-07 ***
## age               0.026991   0.008675   3.111 0.001862 **
## as.factor(ses)2   0.044609   0.432490   0.103 0.917849
## as.factor(ses)3   0.253433   0.405532   0.625 0.532011
## as.factor(sec)2   1.243630   0.352271   3.530 0.000415 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 236.33  on 195  degrees of freedom
## Residual deviance: 211.22  on 191  degrees of freedom
## AIC: 221.22
##
## Number of Fisher Scoring iterations: 3
```

—— **This is the end of Lab 8.** ——