HW 5

Jared Yu

November 9, 2018

Read in the flu file.

```
library(readxl)
flu <- read_excel('flu.xlsx')
flu
```

```
## # A tibble: 159 x 6
##     Shot   Age `Health Awareness` Gender X__1  X__2
##    <dbl> <dbl>              <dbl>  <dbl> <lgl> <chr>
##  1    0.   59.                 52.     0. NA    <NA>
##  2    0.   61.                 55.     1. NA    Gender: 0=female, 1=male
##  3    1.   82.                 51.     0. NA    Shot: 0=no flu shot, 1=rec~
##  4    0.   51.                 70.     0. NA    <NA>
##  5    0.   53.                 70.     0. NA    <NA>
##  6    0.   62.                 49.     1. NA    <NA>
##  7    0.   51.                 69.     1. NA    <NA>
##  8    0.   70.                 54.     1. NA    <NA>
##  9    0.   71.                 65.     1. NA    <NA>
## 10    0.   55.                 58.     1. NA    <NA>
## # ... with 149 more rows
```

1 a) Below is the calculation for the MLE's of β_0 and β_1

```
flu.fit = glm(Shot~Age, family=binomial(), data=flu)
flu.fit
```

```
##
## Call:  glm(formula = Shot ~ Age, family = binomial(), data = flu)
##
## Coefficients:
## (Intercept)          Age
##     -8.7433       0.1087
##
## Degrees of Freedom: 158 Total (i.e. Null);  157 Residual
## Null Deviance:       134.9
## Residual Deviance: 116.3     AIC: 120.3
```

The MLE for β_0 is:

```
flu.fit$coefficients[1]
```

```
## (Intercept)
##    -8.74326
```

The MLE for β_1 is:

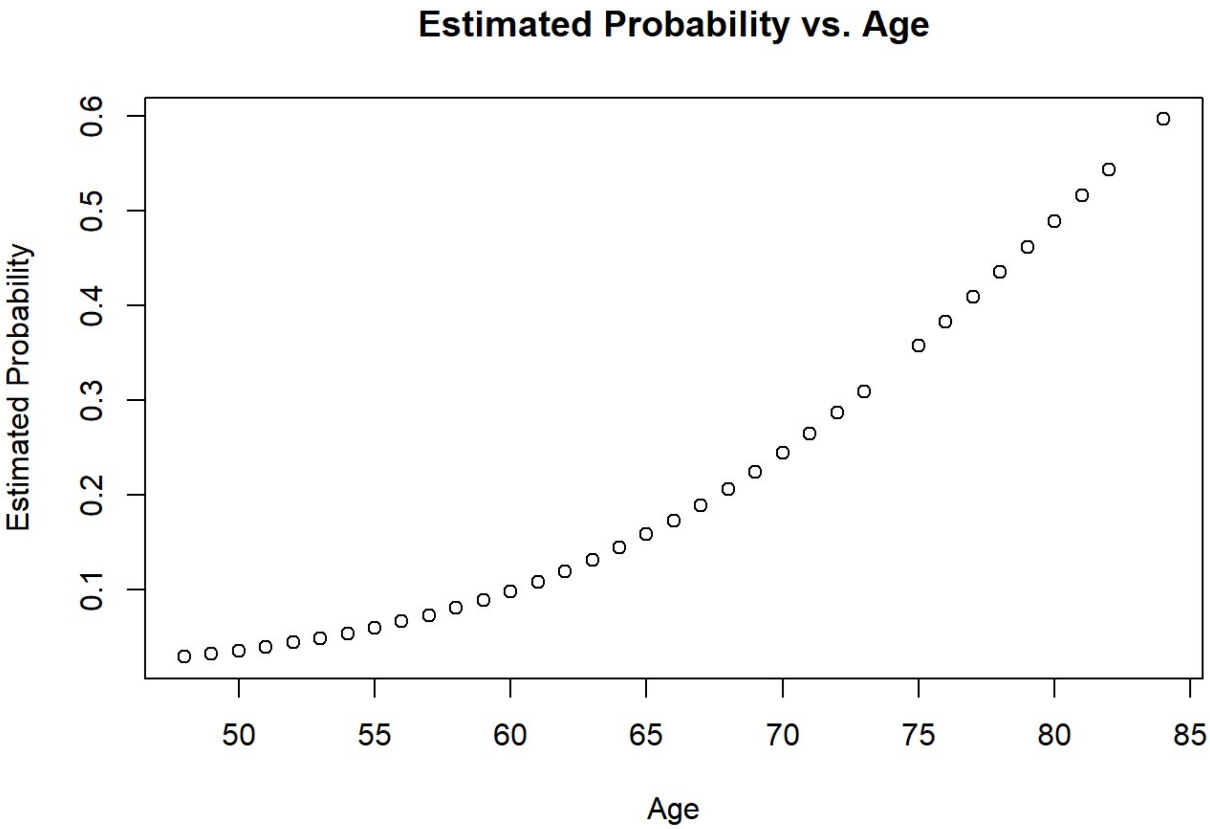
```
flu.fit$coefficients[2]
```

```
##       Age
## 0.1087366
```

The estimated logistic regression function is therefore: $log(\frac{\pi}{1-\pi}) = -8.74326 + 0.1087366X$ Where π is probability of getting a flu shot and X is the age.

1 b) Plot the estimated probability against age

```
plot(flu$Age, flu.fit$fitted.values, xlab='Age', ylab='Estimated Probability', main='E
stimated Probability vs. Age')
```



It is apparent from the plot that the estimated probability is related to the age. As age gets higher, the estimated probability of getting a flu shot will also increases.

1 c) Estimate the probability that a 55 y.o. will get a flu shot within a 95% confidence interval.

```
test.55 = data.frame(55)
colnames(test.55) = 'Age'
prob = predict(flu.fit, test.55, type="response")
result = predict(flu.fit, test.55, se.fit = TRUE)

confidence_interval = c(result$fit-1.96*result$se.fit, result$fit+1.96*result$se.fit)
link_function = function(x){
  return(exp(x)/(1+exp(x)))
}
c(link_function(confidence_interval[1]),link_function(confidence_interval[2]))
```

```
##          1          1
## 0.02754806 0.12329397
```

The confidence interval is therefore (0.02754806, 0.12329397), where there is a 2.75% - 12.33% that a 55 y.o. will get a flu shot.

1 d) Below is the calculation for $exp(\hat{\beta}_1)$

```
exp(flu.fit$coefficients[2])
```

```
##      Age
## 1.114869
```

The odds of getting a flu shot at age X + 1 is about 1.114869 times the odds of getting the flu at age X.

2 a) An estimate of the age when the probability of a flu shot is 0.5.

```
logit = function(x){
  return(log(x/(1-x)))
}
age_0.5 = (logit(0.5)-flu.fit$coefficients[1])/flu.fit$coefficients[2]
age_0.5
```

```
## (Intercept)
##    80.40767
```

The estimate age at which there is a 50% chance of getting a flu shot is 80.40767.

2 b) It wouldn't make sense because the range of the data doesn't include 15. Therefore a prediction of that number can't be properly modeled. $\hat{\beta}_0$ is when $\hat{\beta}_1 = 0$, so there is no practical meaning for this case.

2 c) A 95% confidence interval for

```
summary(flu.fit)
```

β_1

```
##
## Call:
## glm(formula = Shot ~ Age, family = binomial(), data = flu)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3472  -0.6008  -0.3688  -0.2683   2.5447
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.74326    1.85307  -4.718 2.38e-06 ***
## Age          0.10874    0.02737   3.973 7.08e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 134.94  on 158  degrees of freedom
## Residual deviance: 116.27  on 157  degrees of freedom
## AIC: 120.27
##
## Number of Fisher Scoring iterations: 5
```

```
c(0.10874-1.96*0.02737,0.10874+1.96*0.02737)
```

```
## [1] 0.0550948 0.1623852
```

A 95% confidence interval for β_1 is (0.0550948, 0.1623852). The confidence interval does not include zero, so the H_0 that $\beta_1 = 0$ is rejected at $\alpha = 0.05$. The conclusion is that the probability of getting a flu shot changes with age.

Read in the Bottle Return file and rename the columns

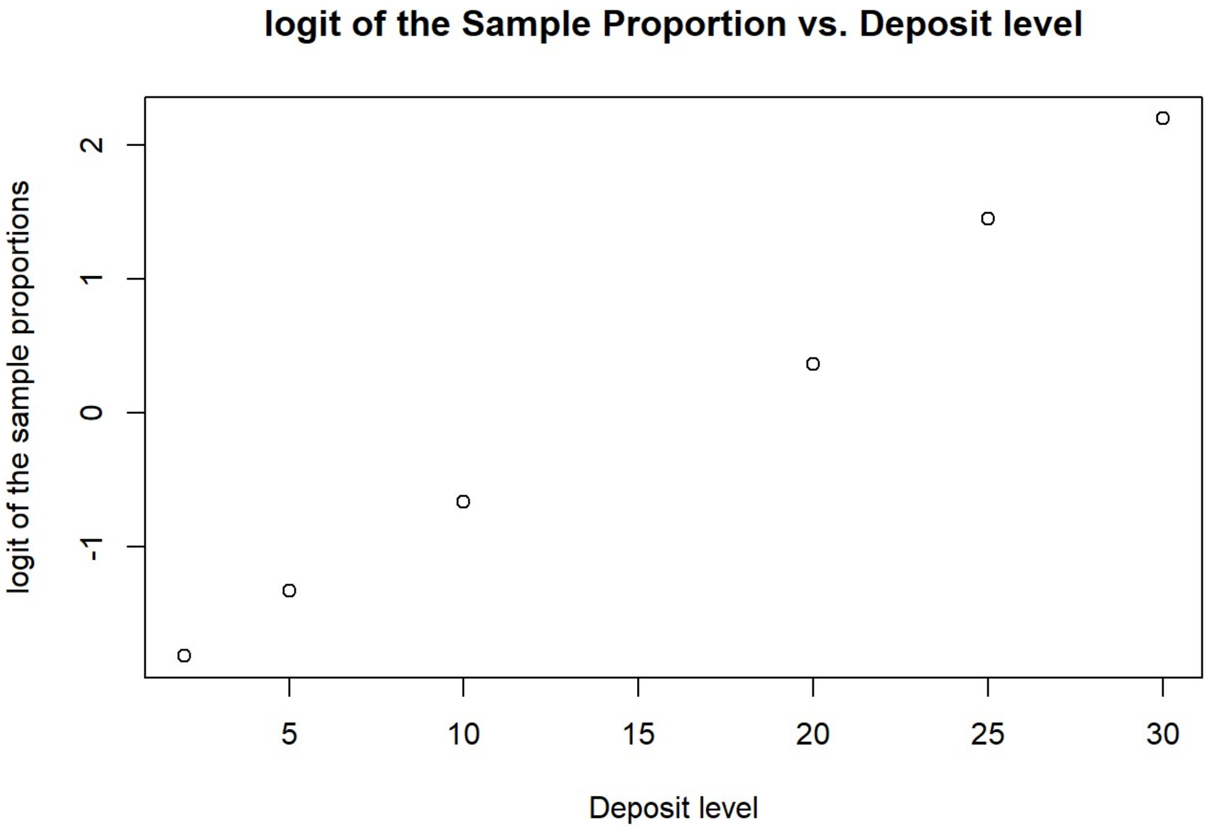
```
bottle <- read_excel('BottleReturn.xlsx')
colnames(bottle) <- c('Number_Returned', 'Number_Sold', 'Deposit_level')
bottle$Number_Not_Returned = 100 - bottle$Number_Returned
bottle
```

```
## # A tibble: 6 x 4
##   Number_Returned Number_Sold Deposit_level Number_Not_Returned
##             <dbl>       <dbl>         <dbl>               <dbl>
## 1             14.        100.            2.                 86.
## 2             21.        100.            5.                 79.
## 3             34.        100.           10.                 66.
## 4             59.        100.           20.                 41.
## 5             81.        100.           25.                 19.
## 6             90.        100.           30.                 10.
```

3 a) A plot the logit of the sample proportions $P_i = Y_i/n_i$ against X_i

```
bottle.fit = glm(cbind(Number_Returned,Number_Not_Returned)~Deposit_level, data=bottl
e, family=binomial())

plot(bottle$Deposit_level, logit(bottle$Number_Returned/(bottle$Number_Returned+bottle
$Number_Not_Returned)), ylab='logit of the sample proportions', xlab='Deposit level',
main='logit of the Sample Proportion vs. Deposit level')
```



The relationship in the logit is very close to a linear trend. So it seems to support the analyst’s belief that a logistic model is appropriate in modeling the probability of return as a function of deposit level.

3 b) The calculation for the MLE of β_0 and β_1 are below. MLE for β_0 :

```
bottle.fit$coefficients[1]
```

```
## (Intercept)
##   -2.080671
```

MLE for β_1 :

```
bottle.fit$coefficients[2]
```

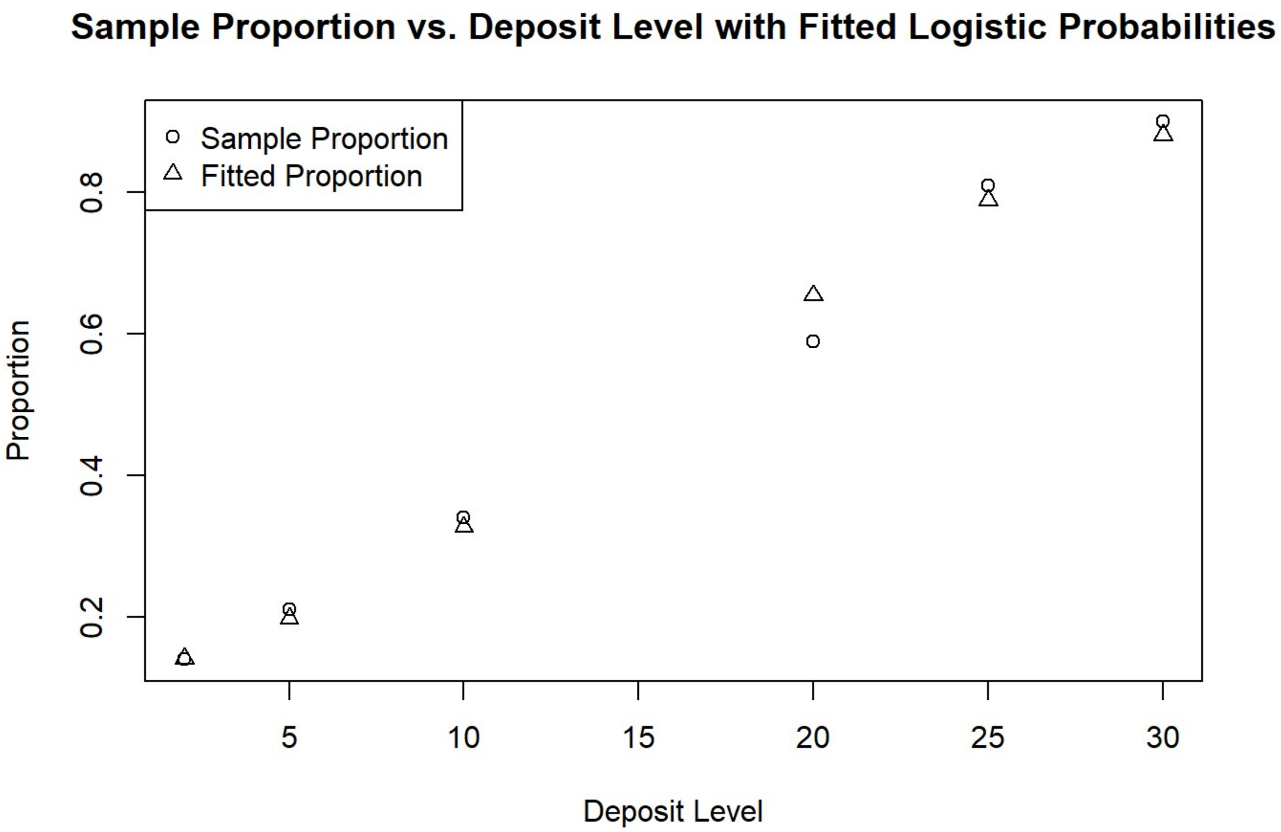
```
## Deposit_level
##     0.1359899
```

The estimated logistic regression function is therefore: $log(\frac{\pi}{1-\pi}) = -2.080671 + 0.1359899X$.

Where π is the probability that a bottle is returned, and X is the deposit level.

3 c)

```
plot(bottle$Deposit_level, bottle$Number_Returned/(bottle$Number_Returned+bottle$Numbe
r_Not_Returned),
     xlab='Deposit Level', ylab='Proportion', main='Sample Proportion vs. Deposit Leve
l with Fitted Logistic Probabilities')
points(bottle$Deposit_level, bottle.fit$fitted.values, pch=2)
legend("topleft", c("Sample Proportion", "Fitted Proportion"), pch = c(1,2))
```



From the plot it is apparent that the fitted logistic model closely follows the sample data.

3 d)

```
res.D = residuals(bottle.fit, type='deviance')
res.D
```

```
##           1           2           3           4           5           6 
## -0.02291418  0.30645498  0.27142876 -1.34026362  0.52057855  0.61053789
```

```
res.P = residuals(bottle.fit, type='pearson')
res.P
```

```
##           1           2           3           4           5           6 
## -0.02289609  0.30879827  0.27231498 -1.35761098  0.51401972  0.59556810
```

```
res.P.standard = rstandard(bottle.fit)
res.P.standard
```

```
##           1           2           3           4           5           6
## -0.02860569  0.37833633  0.32128976 -1.58867565  0.64689426  0.76969001
```

```
res.D.standard = rstudent(bottle.fit)
res.D.standard
```

```
##          1          2          3          4          5          6
## -0.0285976  0.3793337  0.3215904 -1.5946306  0.6440339  0.7627469
```

All the standardized Pearson and deviance residuals remain within ± 2 . This indicates that there are no outliers for the fitted model.

4 a) An estimate of whether the bottle will be returned when the deposit is 15 cents.

```
test.15 = data.frame(15)
colnames(test.15) = 'Deposit_level'
predict(bottle.fit, test.15, type='response')
```

```
##         1
## 0.4897958
```

The probability that the bottle will be returned when the deposit is 15 cents is 0.4897958.

4 b) Estimate the amount of deposit for which 75% of the bottles are expected to be returned.

```
bottles_0.75 = (logit(0.75)-bottle.fit$coefficients[1])/bottle.fit$coefficients[2]
bottles_0.75
```

```
## (Intercept)
##    23.37882
```

When the deposit is 23.37882, 75% of the bottles are expected to be returned.

4 c) Obtain a 95% confidence interval for β_1 and interpret it. Convert this interval into one for the odds ratio. Interpret this interval.

```
summary(bottle.fit)
```

```
##
## Call:
## glm(formula = cbind(Number_Returned, Number_Not_Returned) ~ Deposit_level,
##     family = binomial(), data = bottle)
##
## Deviance Residuals:
##        1         2         3         4         5         6
## -0.02291   0.30645   0.27143  -1.34026   0.52058   0.61054
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -2.08067    0.18988  -10.96   <2e-16 ***
## Deposit_level  0.13599    0.01068   12.74   <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: 222.1459  on 5  degrees of freedom
## Residual deviance:   2.6082  on 4  degrees of freedom
## AIC: 34.218
##
## Number of Fisher Scoring iterations: 3
```

```
c(0.13599-1.96*0.01068,0.13599+1.96*0.01068)
```

```
## [1] 0.1150572 0.1569228
```

The 95% confidence interval for β_1 is (0.1150572, 0.1569228). The confidence interval does not include zero, so we can reject the H_0 that $\beta_1 = 0$ at $\alpha = 0.05$. The conclusion is that the probabily of a bottle being returned changes with the deposit level.

```
confidence_interval2 = c(0.13599-1.96*0.01068,0.13599+1.96*0.01068)
c(link_function(confidence_interval2[1]),link_function(confidence_interval2[2]))
```

```
## [1] 0.5287326 0.5391504
```

The 95% confidence interval for the odds ratio is (0.5287326, 0.5391504). The odds ratio doesn't include 1, so the odds of number returned decreases by a factor between (0.5287326, 0.5391504) as the deposit level changes by 1 cent.

4 d) Conduct a likelihood ratio test to determine whether deposit level X is related to the probability is returned at level $\alpha = 0.05$. Write down the null and the alternative hypotheses, the decision rule, and state your conclusion. H_0 : Deposit level is not related to the probability that the bottle is returned at $\alpha = 0.05$. H_1 : Deposit level is related to the probability that the bottle is returned at $\alpha = 0.05$.

```
anova(bottle.fit,test='Chi')
```

```
## Analysis of Deviance Table
## 
## Model: binomial, link: logit
## 
## Response: cbind(Number_Returned, Number_Not_Returned)
## 
## Terms added sequentially (first to last)
## 
## 
##               Df Deviance Resid. Df Resid. Dev  Pr(>Chi)    
## NULL                              5    222.146              
## Deposit_level  1   219.54         4      2.608 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

log(LR) = D(smaller model) - D(larger model) ~ $\chi^2_p, p =$ df(larger model) - df(smaller model), where D is the deviance of the corresponding model.

4 d) (cont.)

```
bottle.fit$null.deviance-bottle.fit$deviance > qchisq(0.95,1)
```

```
## [1] TRUE
```

Decision rule is to reject the null hypothesis because p-value is too small. Conclusion is that the deposit level and probability that the bottle returned are related.

5 a) It is desired to carry out a likelihood ratio test to determine if a logistic linear model is appropriate for the data at level $\alpha = 0.05$. Write down the null and the alternative hypotheses, the decision rule, and state your conclusion. H_0 : The logistic linear model is appropriate for the data. H_1 : The logistic linear model is not appropriate for the data.

```
bottle.fit$deviance>qchisq(0.95,length(bottle.fit$fitted.values)-2)
```

```
## [1] FALSE
```

```
1-pchisq(bottle.fit$deviance,length(bottle.fit$fitted.values)-2)
```

```
## [1] 0.625375
```

The decision rule is to fail to reject the null hypothesis that the logistic linear model is appropriate for the data. The p-value for the test is 0.625375, which is quite large. The conclusion is that the logistic linear model is appropriate for the data.

5 b) The estimated expected counts under the null hypothesis are given below.

```
expected_bottle = (bottle$Number_Returned + bottle$Number_Not_Returned)*bottle.fit$fit
ted.values
expected_bottle
```

```
##        1        2        3        4        5        6 
## 14.07964 19.77016 32.72230 65.45561 78.90281 88.06948
```

5 c)

```
expected_bottle2 = (bottle$Number_Returned + bottle$Number_Not_Returned)*(1 - bottle.f
it$fitted.values)

sum((bottle$Number_Returned - expected_bottle)^2 / expected_bottle) + sum((bottle$Numb
er_Not_Returned - expected_bottle2)^2 / expected_bottle2)
```

```
## [1] 2.632061
```

The test statistic for Pearson’s chi-square test is 2.632061.

```
1 - pchisq(q = 2.632061, df = length(bottle.fit$fitted.values)-2)
```

```
## [1] 0.6211543
```

The p-value for the Pearson’s chi-square test is 0.6211543. This is much larger than 0.05, so the decision rule is to fail to reject the null hypothesis. The conclusion is that the model fits well for the data.