W271 Assignment 3 Submission

Huibin Chang

Contents


```r
library(tidyverse)
```

1 Customer churn study: Part-3 (100 Points)

In my submitted homework, each question is separated by a pagebreak, and there is a subtitle indicating the start of the answer for each question.

In the last two homework assignments, you initiated modeling a binary variable and used logistic regression to study the churn tendencies of customers.

Now, in Part-3, we're going to explore different interactions, transformations, and categorical explanatory variables to create a more comprehensive model.

```r
telcom_churn <- read.csv("./data/Telco_Customer_Churn.csv", header=T,na.strings=c("","NA"))
```

For the remainder of this section, pay particular attention to all variables.

1.1 Data Preprocessing (5 Points)

In this section, Convert variables as needed, and manage any missing values.

1.1.1 Answer

1.1.1.1 Handling missing values In the code block below, I first check missing values for each column in the data frame. Since there are only 11 missing values in **TotalCharges**, I dropped these missing values because the number is small compared to the entire sample.

1.1.1.2 Data type Categorical and binary variables are first checked for unique values then turned into R factor class.

1.1.1.3 Process data and sanity check I saved the processed and cleaned data to **telcom_churn_clean** and use glimpse() and summary to look at the dataframe and the distribution of variables. **glimpse()** are commented out later to declutter the output.

```
library(tidyverse)
#glimpse(telcom_churn)

telcom_churn %>%
  summarize(across(everything(), ~ sum(is.na(.))))
```

```
##   customerID gender SeniorCitizen Partner Dependents tenure PhoneService
## 1          0      0             0       0          0      0            0
##   MultipleLines InternetService OnlineSecurity OnlineBackup DeviceProtection
## 1             0               0              0            0                0
##   TechSupport StreamingTV StreamingMovies Contract PaperlessBilling
## 1           0           0               0        0                0
##   PaymentMethod MonthlyCharges TotalCharges Churn
## 1             0              0           11     0
```

```
# Check unique values of categorical variables
cate_vars <- c("Churn", "SeniorCitizen", "gender")
telcom_churn %>%
  reframe(across(all_of(cate_vars), unique))
```

```
##   Churn SeniorCitizen gender
## 1    No             0 Female
## 2   Yes             1   Male
```

```
unique(telcom_churn$Churn)
```

```
## [1] "No"  "Yes"
```

```
unique(telcom_churn$SeniorCitizen)
```

```
## [1] 0 1
```

```
class(telcom_churn$SeniorCitizen)
```

```
## [1] "integer"
```

```
telcom_churn_clean <-
  telcom_churn %>%
    filter(!is.na(TotalCharges)) %>%
    mutate(across(all_of(cate_vars), as.factor))
```

```r
# glimpse(telcom_churn_clean)
summary(telcom_churn_clean)
```

```
##   customerID           gender     SeniorCitizen   Partner
##  Length:7032        Female:3483   0:5890        Length:7032
##  Class :character   Male  :3549   1:1142        Class :character
##  Mode  :character                               Mode  :character
##
##
##
##   Dependents            tenure      PhoneService       MultipleLines
##  Length:7032        Min.   : 1.00   Length:7032        Length:7032
##  Class :character   1st Qu.: 9.00   Class :character   Class :character
##  Mode  :character   Median :29.00   Mode  :character   Mode  :character
##                     Mean   :32.42
##                     3rd Qu.:55.00
##                     Max.   :72.00
##  InternetService    OnlineSecurity     OnlineBackup       DeviceProtection
##  Length:7032        Length:7032        Length:7032        Length:7032
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  TechSupport        StreamingTV        StreamingMovies      Contract
##  Length:7032        Length:7032        Length:7032        Length:7032
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  PaperlessBilling   PaymentMethod      MonthlyCharges    TotalCharges
##  Length:7032        Length:7032        Min.   : 18.25   Min.   :  18.8
##  Class :character   Class :character   1st Qu.: 35.59   1st Qu.: 401.4
##  Mode  :character   Mode  :character   Median : 70.35   Median :1397.5
##                                        Mean   : 64.80   Mean   :2283.3
##                                        3rd Qu.: 89.86   3rd Qu.:3794.7
##                                        Max.   :118.75   Max.   :8684.8
##  Churn
##  No :5163
##  Yes:1869
##
##
##
##
```

1.2 Estimate a logistic regression (10 Points)

Estimate the following binary logistic regressions and report the results in a table using stargazer package.

$$Churn = \beta_0 + \beta_1 tenure + \beta_2 MonthlyCharges + \beta_3 TotalCharges + \beta_4 SeniorCitizen + \beta_5 gender + e \quad \text{(Model 1)}$$

$$\begin{aligned} Churn = \beta_0 &+ \beta_1 tenure + \beta_2 MonthlyCharges + \beta_3 TotalCharges + \beta_4 SeniorCitizen + \beta_5 gender \\ &+ \beta_6 tenure^2 + \beta_7 MonthlyCharges^2 + \beta_8 TotalCharges^2 + e \end{aligned} \quad \text{(Model 2)}$$

$$\begin{aligned} Churn = \beta_0 &+ \beta_1 tenure + \beta_2 MonthlyCharges + \beta_3 TotalCharges + \beta_4 SeniorCitizen + \beta_5 gender \\ &+ \beta_6 tenure^2 + \beta_7 MonthlyCharges^2 + \beta_8 TotalCharges^2 \\ &+ \beta_9 SeniorCitizen \times tenure + \beta_{10} SeniorCitizen \times MonthlyCharges \\ &+ \beta_{11} SeniorCitizen \times TotalCharges + \beta_{12} gender \times tenure \\ &+ \beta_{13} gender \times MonthlyCharges + \beta_{14} gender \times TotalCharges + e \end{aligned} \quad \text{(Model 3)}$$

- where $SeniorCitizen \times MonthlyCharges$ denotes the interaction between `SeniorCitizen` and `MonthlyCharges` variables.

1.2.1 Answer

In the code block below, the three models are estimated using **glm()** and the results are reported using **stargazer()**.

```
#install.packages("stargazer")
library(stargazer)
```

```
##
## Please cite as:

##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```
model_1 <- glm(formula = Churn ~ tenure + MonthlyCharges + TotalCharges + SeniorCitizen + gender,
               data = telcom_churn_clean,
               family = binomial(link = "logit"))

# summary(model_1)

model_2 <- glm(formula = Churn ~ tenure + MonthlyCharges + TotalCharges + SeniorCitizen + gender
                          + I(tenure^2) + I(MonthlyCharges^2) + I(TotalCharges^2),
               data = telcom_churn_clean,
               family = binomial(link = "logit"))

# summary(model_2)


model_3 <- glm(formula = Churn ~ tenure + MonthlyCharges + TotalCharges + SeniorCitizen + gender
                          + I(tenure^2) + I(MonthlyCharges^2) + I(TotalCharges^2)
                          + SeniorCitizen:tenure + SeniorCitizen:MonthlyCharges
                          + SeniorCitizen:TotalCharges + gender:tenure
                          + gender:MonthlyCharges + gender:TotalCharges,
               data = telcom_churn_clean,
               family = binomial(link = "logit"))

# summary(model_3)
```

```r
stargazer(model_1, model_2, model_3,
          type = "text",
          title = "Comparison of Logistic Regression Models",
          dep.var.labels = "Customer Churn",
          covariate.labels = c("Tenure", "Monthly Charges", "Total Charges",
                               "Senior Citizen (Yes)", "Gender (Male)",
                               "Tenure Squared", "Monthly Charges Squared",
                               "Total Charges Squared",
                               "Senior Citizen * Tenure",
                               "Senior Citizen * Monthly Charges",
                               "Senior Citizen * Total Charges",
                               "Gender * Tenure",
                               "Gender * Monthly Charges",
                               "Gender * Total Charges"))
```

```
## 
## Comparison of Logistic Regression Models
## ===================================================================
##                                         Dependent variable:        
##                                  ----------------------------------
##                                            Customer Churn          
##                                     (1)         (2)         (3)    
## -------------------------------------------------------------------
## Tenure                           -0.068***   -0.125***   -0.123*** 
##                                   (0.005)     (0.013)     (0.014)  
##                                                                    
## Monthly Charges                   0.028***    0.023***    0.024*** 
##                                   (0.002)     (0.007)     (0.007)  
##                                                                    
## Total Charges                     0.0002**    0.001***    0.001*** 
##                                   (0.0001)    (0.0002)    (0.0002) 
##                                                                    
## Senior Citizen (Yes)              0.630***    0.634***    1.477*** 
##                                   (0.079)     (0.080)     (0.399)  
##                                                                    
## Gender (Male)                      -0.004      -0.007      0.247   
##                                   (0.062)     (0.063)     (0.235)  
##                                                                    
## Tenure Squared                                0.001***    0.001*** 
##                                               (0.0001)    (0.0001) 
##                                                                    
## Monthly Charges Squared                        0.00003     0.0001  
##                                               (0.0001)    (0.0001) 
##                                                                    
## Total Charges Squared                       -0.00000*** -0.00000***
##                                               (0.00000)   (0.00000)
##                                                                    
## Senior Citizen * Tenure                                    0.013   
##                                                           (0.013)  
##                                                                    
## Senior Citizen * Monthly Charges                         -0.013**  
##                                                           (0.005)  
```

```
##
## Senior Citizen * Total Charges                                   -0.0001
##                                                                  (0.0002)
##
## Gender * Tenure                                                  -0.010
##                                                                  (0.010)
##
## Gender * Monthly Charges                                         -0.006*
##                                                                  (0.003)
##
## Gender * Total Charges                                           0.0002*
##                                                                  (0.0001)
##
## Constant                          -1.581***   -1.241***   -1.358***
##                                    (0.122)     (0.201)     (0.236)
##
## --------------------------------------------------------------------
## Observations                        7,032       7,032       7,032
## Log Likelihood                    -3,156.802 -3,138.899  -3,126.703
## Akaike Inf. Crit.                 6,325.604   6,295.799   6,283.406
## ====================================================================
## Note:                                   *p<0.1; **p<0.05; ***p<0.01
```

1.3 Test a hypothesis: linear effects (15 Points)

Using `Model 1`, test the hypothesis of linear effects of variables on customer churn using a likelihood ratio test.

1.3.1 Answer

The model is

$$P_i = \frac{e^V}{1 + e^V}$$

Hypothesis testing:

$$H_o : V = constant + e$$

$$H_a : V \text{ as specified model 1}$$

In answering this part, I first did the likelihood ratio test explicitly, then I used the deviance values given for free by the fitted model to verify that the log-likelihood (used to compute the asymptotically χ^2) is correct.

To do that, I first estimated the **null model** which regress the dependent variable on a constant term only. Then I use the anova() method to carry out the likelihood-ratio test. In the test print-out, it displays the χ^2 statistic, which I then compare to a manually computed value to double-check.

The test statistic is 1829.751 (checked in two ways shown in the code block), with degree of freedom equaling 5, and its corresponding p-value is approximately 0.

```r
# Method 1
model_null <- glm(formula = Churn ~ 1,
                  data = telcom_churn_clean,
                  family = binomial(link = "logit"))

summary(model_null)
```

```
##
## Call:
## glm(formula = Churn ~ 1, family = binomial(link = "logit"), data = telcom_churn_clean)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.016      0.027  -37.64   <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: 8143.4  on 7031  degrees of freedom
## Residual deviance: 8143.4  on 7031  degrees of freedom
## AIC: 8145.4
##
## Number of Fisher Scoring iterations: 4
```

```r
anova(model_null, model_1, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: Churn ~ 1
```

```
## Model 2: Churn ~ tenure + MonthlyCharges + TotalCharges + SeniorCitizen +
##     gender
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      7031     8143.4
## 2      7026     6313.6  5   1829.8 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Method 2
#summary(model_1)
deviance(model_1)
```

```
## [1] 6313.604
```

```r
g2 <- deviance(model_null) - deviance(model_1)
g2
```

```
## [1] 1829.751
```

1.4 Test a hypothesis: Non linear effect (15 Points)

Perform a likelihood ratio test to assess the hypothesis that $\beta_6 = 0$ or $\beta_7 = 0$ or $\beta_8 = 0$ within the context of `Model 2`. Interpret the implications of this test result in the context of the estimated `Model 2`.

Then, test the same hypothesis in `Model 3` using a likelihood ratio test. Interpret what this test result means in the context of a model like what you have estimated in `Model 3`.

1.4.1 Answer

The model is

$$P_i = \frac{e^V}{1+e^V}$$

1.4.1.1 Test of Model 2 $\beta_i = 0$ for $i = 6, 7, 8$, so the **restricted model** is

$$Churn = \beta_0 + \beta_1 tenure + \beta_2 MonthlyCharges + \beta_3 TotalCharges + \beta_4 SeniorCitizen + \beta_5 gender + e \quad \text{(Restricted Model 2)}$$

Hypothesis

$$H_o : \text{restricted model 2}$$

$$H_a : \text{As specified in (full) model 2}$$

Note that the restricted model_2 is the same as model_1 but I explicitly coded the restricted model_2 for clarity.

Similar to the last part, I first did the likelihood ratio test by explicitly specifying a null-hypothesis model, doing the test; then verify by manually compute the χ^2 statistic.

```
model_2_restricted <- glm(formula = Churn ~ tenure + MonthlyCharges + TotalCharges + SeniorCitizen + gen
                          data = telcom_churn_clean,
                          family = binomial(link = "logit"))

#summary(model_2_restricted)

anova(model_2_restricted, model_2, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: Churn ~ tenure + MonthlyCharges + TotalCharges + SeniorCitizen +
##     gender
## Model 2: Churn ~ tenure + MonthlyCharges + TotalCharges + SeniorCitizen +
##     gender + I(tenure^2) + I(MonthlyCharges^2) + I(TotalCharges^2)
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      7026     6313.6
## 2      7023     6277.8  3   35.806 8.232e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Double check
#anova(model_1, model_2, test = "LRT")

# Triple check by manually compute the difference of deviances
deviance(model_1) - deviance(model_2)
```

```
## [1] 35.80555
```

Interpretation Based on the likelihood ratio test, we can reject the null (that the restricted model 2 is the true model) in favor of the alternative (full model 2).

1.4.1.2 Test of Model 3

Again I have specified the restricted model (null hypothesis), did the test, then verified using manually computed χ^2 statistic.

$\beta_i = 0$ for $i = 6, 7, 8$, so the **restricted model** is

$$
\begin{aligned}
Churn = {} & \beta_0 + \beta_1 tenure + \beta_2 MonthlyCharges + \beta_3 TotalCharges + \beta_4 SeniorCitizen + \beta_5 gender \\
& + \beta_6 tenure^2 + \beta_7 MonthlyCharges^2 + \beta_8 TotalCharges^2 \\
& + \beta_9 SeniorCitizen \times tenure + \beta_{10} SeniorCitizen \times MonthlyCharges \\
& + \beta_{11} SeniorCitizen \times TotalCharges + \beta_{12} gender \times tenure \\
& + \beta_{13} gender \times MonthlyCharges + \beta_{14} gender \times TotalCharges + e
\end{aligned} \tag{Model 3}
$$

```r
model_3_restricted <- glm(formula = Churn ~ tenure + MonthlyCharges + TotalCharges + SeniorCitizen + ger
                          + SeniorCitizen:tenure + SeniorCitizen:MonthlyCharges
                          + SeniorCitizen:TotalCharges + gender:tenure
                          + gender:MonthlyCharges + gender:TotalCharges,
              data = telcom_churn_clean,
              family = binomial(link = "logit"))

#summary(model_3_restricted)

anova(model_3_restricted, model_3, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: Churn ~ tenure + MonthlyCharges + TotalCharges + SeniorCitizen +
##     gender + SeniorCitizen:tenure + SeniorCitizen:MonthlyCharges +
##     SeniorCitizen:TotalCharges + gender:tenure + gender:MonthlyCharges +
##     gender:TotalCharges
## Model 2: Churn ~ tenure + MonthlyCharges + TotalCharges + SeniorCitizen +
##     gender + I(tenure^2) + I(MonthlyCharges^2) + I(TotalCharges^2) +
##     SeniorCitizen:tenure + SeniorCitizen:MonthlyCharges + SeniorCitizen:TotalCharges +
##     gender:tenure + gender:MonthlyCharges + gender:TotalCharges
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      7020     6285.5
## 2      7017     6253.4  3   32.111 4.958e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
deviance(model_3_restricted) - deviance(model_3)
```

```
## [1] 32.11138
```

Interpretation Based on the likelihood ratio test, we can reject the null (the restricted model 3 is the true model) in favor of the full model 3.

1.5 Test a hypothesis: Total effect of gender (15 Points)

Test the hypothesis that `gender` has no effect on the likelihood of churn, in `Model 3`, using a likelihood ratio test.

1.5.1 Answer

Test that there is no effect of gender. So the restricted version of **Model 3** is

$$\begin{aligned}
Churn = \beta_0 + \beta_1 tenure + \beta_2 MonthlyCharges + \beta_3 TotalCharges + \beta_4 SeniorCitizen + 0 \qquad &\text{(Restricted Model 3 - drop} \\
+ \beta_6 tenure^2 + \beta_7 MonthlyCharges^2 + \beta_8 TotalCharges^2 & \\
+ \beta_9 SeniorCitizen \times tenure + \beta_{10} SeniorCitizen \times MonthlyCharges & \\
+ \beta_{11} SeniorCitizen \times TotalCharges + 0 & \\
+ e &
\end{aligned}$$

Hence the null hypothesis, H_o is given by the restricted model. The alternative hypothesis is the full model where coefficents of gender (linear, squared, and interaction terms) are not zero.

Interpretation of the results

The p-value is 0.049 (chisq stat 9.53 with degree of freedom 4). So it is marginally significant at $\alpha = 0.05$, but not statistically significant if we lower the value of α.

Hence the **marginally** reject the null that gender (linear, squared, interactions) are all zero, but the evidence is very weak. In practice, I would be more conservative and adopt a smaller α, hence not reject the null.

```r
model_3_no_gender <- glm(formula = Churn ~ tenure + MonthlyCharges + TotalCharges + SeniorCitizen
                         + I(tenure^2) + I(MonthlyCharges^2) + I(TotalCharges^2)
                         + SeniorCitizen:tenure + SeniorCitizen:MonthlyCharges
                         + SeniorCitizen:TotalCharges,
               data = telcom_churn_clean,
               family = binomial(link = "logit"))

summary(model_3_no_gender)
```

```
##
## Call:
## glm(formula = Churn ~ tenure + MonthlyCharges + TotalCharges +
##     SeniorCitizen + I(tenure^2) + I(MonthlyCharges^2) + I(TotalCharges^2) +
##     SeniorCitizen:tenure + SeniorCitizen:MonthlyCharges + SeniorCitizen:TotalCharges,
##     family = binomial(link = "logit"), data = telcom_churn_clean)
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -1.240e+00  1.991e-01  -6.229 4.69e-10 ***
## tenure                        -1.284e-01  1.311e-02  -9.791  < 2e-16 ***
## MonthlyCharges                 2.141e-02  6.620e-03   3.234 0.001222 **
## TotalCharges                   6.078e-04  1.620e-04   3.752 0.000176 ***
## SeniorCitizen1                 1.474e+00  3.994e-01   3.691 0.000223 ***
## I(tenure^2)                    8.231e-04  1.442e-04   5.708 1.14e-08 ***
## I(MonthlyCharges^2)            6.380e-05  5.626e-05   1.134 0.256827
## I(TotalCharges^2)             -6.273e-08  1.584e-08  -3.961 7.47e-05 ***
## tenure:SeniorCitizen1          1.302e-02  1.323e-02   0.984 0.325276
## MonthlyCharges:SeniorCitizen1 -1.268e-02  5.353e-03  -2.369 0.017856 *
## TotalCharges:SeniorCitizen1   -7.967e-05  1.534e-04  -0.519 0.603414
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8143.4  on 7031  degrees of freedom
## Residual deviance: 6262.9  on 7021  degrees of freedom
## AIC: 6284.9
##
## Number of Fisher Scoring iterations: 6
```

```r
anova(model_3_no_gender, model_3, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: Churn ~ tenure + MonthlyCharges + TotalCharges + SeniorCitizen +
##     I(tenure^2) + I(MonthlyCharges^2) + I(TotalCharges^2) + SeniorCitizen:tenure +
##     SeniorCitizen:MonthlyCharges + SeniorCitizen:TotalCharges
## Model 2: Churn ~ tenure + MonthlyCharges + TotalCharges + SeniorCitizen +
##     gender + I(tenure^2) + I(MonthlyCharges^2) + I(TotalCharges^2) +
##     SeniorCitizen:tenure + SeniorCitizen:MonthlyCharges + SeniorCitizen:TotalCharges +
##     gender:tenure + gender:MonthlyCharges + gender:TotalCharges
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      7021     6262.9
## 2      7017     6253.4  4   9.5332  0.04907 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
deviance(model_3_no_gender) - deviance(model_3)
```

```
## [1] 9.533195
```

1.6 Senior V.S. non-senior customers (20 Points)

Estimate a new model, `Model 4`, by excluding all insignificant variables from `Model 3`. Then, predict how the likelihood of churn changes for senior customers compared to non-senior customers, while keeping `tenure`, `MonthlyCharges`, and `TotalCharges` at their average values.

1.6.1 Answer

Here, because if I do not drop the marginally significant variables (at $\alpha = 0.05$), I will have to deal with choosing/fixing a gender value for prediction.

Also because the language of the problem statement was not explict on the value of α, I dropped all insignificant variables at $\alpha = 0.01$

The resulting model, **model 4** therefore is:

$$
\begin{aligned}
Churn = \beta_0 &+ \beta_1 tenure + \beta_2 MonthlyCharges + \beta_3 TotalCharges + \beta_4 SeniorCitizen + 0 \\
&+ \beta_6 tenure^2 + 0 + \beta_8 TotalCharges^2 \\
&+ 0 + \beta_{10} SeniorCitizen \times MonthlyCharges \\
&+ 0 + 0 \\
&+ + e
\end{aligned} \tag{Model 4}
$$

Results The prediction results are printed out at the end of the following code chunk:

0.1477 for non-senior customers, and 0.2794 for senior customers.

```r
model_4 <- glm(formula = Churn ~ tenure + MonthlyCharges + TotalCharges + SeniorCitizen
                          + I(tenure^2)   + I(TotalCharges^2)
                          + SeniorCitizen:MonthlyCharges,
               data = telcom_churn_clean,
               family = binomial(link = "logit"))

summary(model_4)
```

```
##
## Call:
## glm(formula = Churn ~ tenure + MonthlyCharges + TotalCharges +
##     SeniorCitizen + I(tenure^2) + I(TotalCharges^2) + SeniorCitizen:MonthlyCharges,
##     family = binomial(link = "logit"), data = telcom_churn_clean)
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -1.404e+00  1.388e-01 -10.114  < 2e-16 ***
## tenure                       -1.270e-01  1.314e-02  -9.668  < 2e-16 ***
## MonthlyCharges                2.815e-02  2.165e-03  13.002  < 2e-16 ***
## TotalCharges                  6.209e-04  1.621e-04   3.831 0.000128 ***
## SeniorCitizen1                1.549e+00  2.839e-01   5.454 4.93e-08 ***
## I(tenure^2)                   8.021e-04  1.381e-04   5.807 6.34e-09 ***
## I(TotalCharges^2)            -6.059e-08  1.538e-08  -3.940 8.15e-05 ***
## MonthlyCharges:SeniorCitizen1 -1.147e-02  3.430e-03  -3.342 0.000831 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 8143.4  on 7031  degrees of freedom
## Residual deviance: 6267.1  on 7024  degrees of freedom
## AIC: 6283.1
##
## Number of Fisher Scoring iterations: 6
```

```r
# get averages
avg_tenure <- mean(telcom_churn_clean$tenure, na.rm = TRUE)
avg_monthly <- mean(telcom_churn_clean$MonthlyCharges, na.rm = TRUE)
avg_total <- mean(telcom_churn_clean$TotalCharges, na.rm = TRUE)

# 2. Predict for non-senior customer
prob_non_senior <- predict(
  model_4,
  newdata = data.frame(
    tenure = avg_tenure,
    MonthlyCharges = avg_monthly,
    TotalCharges = avg_total,
    SeniorCitizen = factor("0", levels = levels(telcom_churn_clean$SeniorCitizen))
  ),
  type = "response"
)

prob_senior <- predict(
  model_4,
  newdata = data.frame(
    tenure = avg_tenure,
    MonthlyCharges = avg_monthly,
    TotalCharges = avg_total,
    SeniorCitizen = factor("1", levels = levels(telcom_churn_clean$SeniorCitizen))
  ),
  type = "response"
)

# 4. Print results clearly
cat("Predicted churn probability (Non-senior):", round(prob_non_senior, 4), "\n")
```

```
## Predicted churn probability (Non-senior): 0.1477
```

```r
cat("Predicted churn probability (Senior):    ", round(prob_senior, 4), "\n")
```

```
## Predicted churn probability (Senior):     0.2794
```

```r
cat("Difference (Senior - Non-senior):        ", round(prob_senior - prob_non_senior, 4), "\n")
```

```
## Difference (Senior - Non-senior):        0.1317
```

1.7 Construct a confidence interval (20 Points)

Use `Model 4` and construct the 95% wald confidence interval for the churn probability for the customers with the following profile:

- $tenure = 55.00$;
- $MonthlyCharges = 89.86$;
- $TotalCharges = 3794.7$;
- $SeniorCitizen = "No"$;

and

- $tenure = 29.00$;
- $MonthlyCharges = 18.25$;
- $TotalCharges = 401.4$;
- $SeniorCitizen = "Yes"$

1.7.0.1 Answer

The code chunk below predicted the probabilities using different methods (and results agree).

- First I created a data frame as the new data (for prediction).
- Then I used the fitted model to predict the linear predictor (at the link level).
- Next constructed Wald confidence intervals for the linear predictors.
- Lastly plug in the Wald CIs for the linear predictor to the non-linear map.

```
profiles <- data.frame(
  tenure = c(55, 29),
  MonthlyCharges = c(89.86, 18.25),
  TotalCharges = c(3794.7, 401.4),
  SeniorCitizen = factor(c("0","1"),
                         levels = levels(telcom_churn_clean$SeniorCitizen))
)

# Predict on logit scale with SEs
# So what I get here is the linear predictor for the 2 profiles
pred_link <- predict(model_4, newdata = profiles, type = "link", se.fit = TRUE)

pred_link$fit
```

```
##         1         2
## -1.950270 -2.320326
```

```
pred_link
```

```
## $fit
##         1         2
## -1.950270 -2.320326
##
## $se.fit
##          1          2
## 0.09495534 0.25486596
##
## $residual.scale
## [1] 1
```

```
# Wald 95% CI on logit scale
q <- qnorm(0.975)
lower_logit <- pred_link$fit - q * pred_link$se.fit
```

```
upper_logit <- pred_link$fit + q * pred_link$se.fit

# Transform to probability scale
prob_est  <- plogis(pred_link$fit)
prob_low  <- plogis(lower_logit)
prob_high <- plogis(upper_logit)

# Put results in a table
wald_ci <- data.frame(
  profile = c("Non-senior: tenure=55, MC=89.86, TC=3794.7",
              "Senior: tenure=29, MC=18.25, TC=401.4"),
  est_prob = round(prob_est, 4),
  lower95  = round(prob_low, 4),
  upper95  = round(prob_high, 4)
)

wald_ci
```

```
##                                      profile est_prob lower95 upper95
## 1 Non-senior: tenure=55, MC=89.86, TC=3794.7   0.1245  0.1056  0.1463
## 2      Senior: tenure=29, MC=18.25, TC=401.4   0.0895  0.0563  0.1393
```

```
linear_pred_profile_1 <- predict(object = model_4,
                                 newdata = profiles[1, ],
                                 type = "link",
                                 se = TRUE)
linear_pred_profile_1
```

1.7.0.2 This following code chunk does the same Wald confidence intervals but following the approach shown in the async, manually compute the linear predictors, get the CIs for the linear predictors, than map to probabilities. This is should give same/very similar numbers as the first approach.

```
## $fit
##        1
## -1.95027
##
## $se.fit
## [1] 0.09495534
##
## $residual.scale
## [1] 1
```

```
linear_pred_profile_1_ci_upper <- linear_pred_profile_1$fit + q * linear_pred_profile_1$se

linear_pred_profile_1_ci_lower <- linear_pred_profile_1$fit - q * linear_pred_profile_1$se

pi_hat_1 <- exp(linear_pred_profile_1$fit) / (1 + exp(linear_pred_profile_1$fit))
pi_hat_1
```

```
##         1
## 0.1245239
```

```
pi_hat_1_ci_upper <- exp(linear_pred_profile_1_ci_upper) / (1 + exp(linear_pred_profile_1_ci_upper))
pi_hat_1_ci_lower <- exp(linear_pred_profile_1_ci_lower) / (1 + exp(linear_pred_profile_1_ci_lower))
```

So in method 2, I manually calculated the CI (for probability) for profile 1 and it is exactly the same as the result form method 1. I don't have to repeat for method 1. Probability estimate for profile 1 is 0.1245239, and the 95% CI for this probability estimate is (0.1056109, 0.1462699).