W271 Assignment 3

Contents


```
library(tidyverse)
library(stargazer)
library(package=car)
```

Instructions

Here are some resources that may come in handy as you work on this assignment:

- Access the most updated version of the assignment on the course's GitHub organization.
- Complete your assignments using iSchool DataHub.
- Submit your assignment to Gradescope.

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

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.

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

```
##   customerID gender SeniorCitizen Partner Dependents tenure PhoneService
## 1 7590-VHVEG Female             0     Yes         No      1           No
## 2 5575-GNVDE   Male             0      No         No     34          Yes
## 3 3668-QPYBK   Male             0      No         No      2          Yes
## 4 7795-CFOCW   Male             0      No         No     45           No
## 5 9237-HQITU Female             0      No         No      2          Yes
## 6 9305-CDSKC Female             0      No         No      8          Yes
##      MultipleLines InternetService OnlineSecurity OnlineBackup DeviceProtection
## 1 No phone service             DSL             No          Yes               No
## 2               No             DSL            Yes           No              Yes
## 3               No             DSL            Yes          Yes               No
## 4 No phone service             DSL            Yes           No              Yes
## 5               No     Fiber optic             No           No               No
## 6              Yes     Fiber optic             No           No              Yes
```

```
##   TechSupport StreamingTV StreamingMovies       Contract PaperlessBilling
## 1          No          No              No Month-to-month              Yes
## 2          No          No              No       One year               No
## 3          No          No              No Month-to-month              Yes
## 4         Yes          No              No       One year               No
## 5          No          No              No Month-to-month              Yes
## 6          No         Yes             Yes Month-to-month              Yes
##               PaymentMethod MonthlyCharges TotalCharges Churn
## 1          Electronic check          29.85        29.85    No
## 2              Mailed check          56.95      1889.50    No
## 3              Mailed check          53.85       108.15   Yes
## 4 Bank transfer (automatic)          42.30      1840.75    No
## 5          Electronic check          70.70       151.65   Yes
## 6          Electronic check          99.65       820.50   Yes
```

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.

```
dim(telcom_churn)
```

```
## [1] 7043   21
```

```
str(telcom_churn)
```

```
## 'data.frame':    7043 obs. of  21 variables:
##  $ customerID      : chr  "7590-VHVEG" "5575-GNVDE" "3668-QPYBK" "7795-CFOCW" ...
##  $ gender          : chr  "Female" "Male" "Male" "Male" ...
##  $ SeniorCitizen   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Partner         : chr  "Yes" "No" "No" "No" ...
##  $ Dependents      : chr  "No" "No" "No" "No" ...
##  $ tenure          : int  1 34 2 45 2 8 22 10 28 62 ...
##  $ PhoneService    : chr  "No" "Yes" "Yes" "No" ...
##  $ MultipleLines   : chr  "No phone service" "No" "No" "No phone service" ...
##  $ InternetService : chr  "DSL" "DSL" "DSL" "DSL" ...
##  $ OnlineSecurity  : chr  "No" "Yes" "Yes" "Yes" ...
##  $ OnlineBackup    : chr  "Yes" "No" "Yes" "No" ...
##  $ DeviceProtection: chr  "No" "Yes" "No" "Yes" ...
##  $ TechSupport     : chr  "No" "No" "No" "Yes" ...
##  $ StreamingTV     : chr  "No" "No" "No" "No" ...
##  $ StreamingMovies : chr  "No" "No" "No" "No" ...
##  $ Contract        : chr  "Month-to-month" "One year" "Month-to-month" "One year" ...
##  $ PaperlessBilling: chr  "Yes" "No" "Yes" "No" ...
##  $ PaymentMethod   : chr  "Electronic check" "Mailed check" "Mailed check" "Bank transfer (automatic)
##  $ MonthlyCharges  : num  29.9 57 53.9 42.3 70.7 ...
##  $ TotalCharges    : num  29.9 1889.5 108.2 1840.8 151.7 ...
##  $ Churn           : chr  "No" "No" "Yes" "No" ...
```

```
table(is.na(telcom_churn))
```

```
##
##  FALSE   TRUE
## 147892     11
```

```
colSums(is.na(telcom_churn))
```

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

There are 11 NA values in TotalCharges and no NA values in any other column. Since only a small amount of NA values are missing compaed to the total number of records, we can drop the rows with NA values.

```
telcom_churn <- na.omit(telcom_churn)
dim(telcom_churn)
```

```
## [1] 7032   21
```

```
colSums(is.na(telcom_churn))
```

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

The next step is to convert the categorical features to factors.

```
telcom_churn$SeniorCitizen <- factor(telcom_churn$SeniorCitizen, c(0,1),
                                     labels=c('No', 'Yes'), ordered = is.ordered(telcom_churn))
telcom_churn$Churn <- factor(telcom_churn$Churn)
telcom_churn$gender <- factor(telcom_churn$gender)
str(telcom_churn)
```

```
## 'data.frame':    7032 obs. of  21 variables:
##  $ customerID      : chr  "7590-VHVEG" "5575-GNVDE" "3668-QPYBK" "7795-CFOCW" ...
##  $ gender          : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 1 2 1 1 2 ...
##  $ SeniorCitizen   : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Partner         : chr  "Yes" "No" "No" "No" ...
##  $ Dependents      : chr  "No" "No" "No" "No" ...
##  $ tenure          : int  1 34 2 45 2 8 22 10 28 62 ...
##  $ PhoneService    : chr  "No" "Yes" "Yes" "No" ...
##  $ MultipleLines   : chr  "No phone service" "No" "No" "No phone service" ...
##  $ InternetService : chr  "DSL" "DSL" "DSL" "DSL" ...
##  $ OnlineSecurity  : chr  "No" "Yes" "Yes" "Yes" ...
```

```
##  $ OnlineBackup    : chr  "Yes" "No" "Yes" "No" ...
##  $ DeviceProtection: chr  "No" "Yes" "No" "Yes" ...
##  $ TechSupport     : chr  "No" "No" "No" "Yes" ...
##  $ StreamingTV     : chr  "No" "No" "No" "No" ...
##  $ StreamingMovies : chr  "No" "No" "No" "No" ...
##  $ Contract        : chr  "Month-to-month" "One year" "Month-to-month" "One year" ...
##  $ PaperlessBilling: chr  "Yes" "No" "Yes" "No" ...
##  $ PaymentMethod   : chr  "Electronic check" "Mailed check" "Mailed check" "Bank transfer (automatic)
##  $ MonthlyCharges  : num  29.9 57 53.9 42.3 70.7 ...
##  $ TotalCharges    : num  29.9 1889.5 108.2 1840.8 151.7 ...
##  $ Churn           : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1 1 2 1 ...
##  - attr(*, "na.action")= 'omit' Named int [1:11] 489 754 937 1083 1341 3332 3827 4381 5219 6671 ...
##   ..- attr(*, "names")= chr [1:11] "489" "754" "937" "1083" ...
```

```
head(telcom_churn)
```

```
##   customerID gender SeniorCitizen Partner Dependents tenure PhoneService
## 1 7590-VHVEG Female             No     Yes         No      1           No
## 2 5575-GNVDE   Male             No      No         No     34          Yes
## 3 3668-QPYBK   Male             No      No         No      2          Yes
## 4 7795-CFOCW   Male             No      No         No     45           No
## 5 9237-HQITU Female             No      No         No      2          Yes
## 6 9305-CDSKC Female             No      No         No      8          Yes
##      MultipleLines InternetService OnlineSecurity OnlineBackup DeviceProtection
## 1 No phone service             DSL             No          Yes               No
## 2               No             DSL            Yes           No              Yes
## 3               No             DSL            Yes          Yes               No
## 4 No phone service             DSL            Yes           No              Yes
## 5               No     Fiber optic             No           No               No
## 6              Yes     Fiber optic             No           No              Yes
##   TechSupport StreamingTV StreamingMovies       Contract PaperlessBilling
## 1          No          No              No Month-to-month              Yes
## 2          No          No              No       One year               No
## 3          No          No              No Month-to-month              Yes
## 4         Yes          No              No       One year               No
## 5          No          No              No Month-to-month              Yes
## 6          No         Yes             Yes Month-to-month              Yes
##               PaymentMethod MonthlyCharges TotalCharges Churn
## 1          Electronic check          29.85        29.85    No
## 2              Mailed check          56.95      1889.50    No
## 3              Mailed check          53.85       108.15   Yes
## 4 Bank transfer (automatic)          42.30      1840.75    No
## 5          Electronic check          70.70       151.65   Yes
## 6          Electronic check          99.65       820.50   Yes
```

```
summary(telcom_churn)
```

```
##   customerID          gender     SeniorCitizen   Partner
##  Length:7032        Female:3483   No :5890      Length:7032
##  Class :character   Male  :3549   Yes: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.

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

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

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

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

```r
mod1 <- glm(formula = Churn ~ tenure + MonthlyCharges + TotalCharges + SeniorCitizen +
              gender, family=binomial(link=logit), data=telcom_churn)

mod2 <- glm(formula = Churn ~ tenure + MonthlyCharges + TotalCharges + SeniorCitizen +
              gender + I(tenure^2) + I(MonthlyCharges^2) + I(TotalCharges^2),
            family=binomial(link=logit), data=telcom_churn)

mod3 <- 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, family=binomial(link=logit), data=telcom_churn)


stargazer(mod1, mod2, mod3, type="text", omit.stat="f",
          star.cutoffs=c(0.05, 0.01, 0.001),
          title="Table of Estimated Relationships between Variables and Log Odds of Churn")
```

```
##
## Table of Estimated Relationships between Variables and Log Odds of Churn
## ================================================================
##                                        Dependent variable:
##                                    -----------------------------
##                                               Churn
##                                      (1)          (2)          (3)
## ----------------------------------------------------------------
## tenure                             -0.068***    -0.125***    -0.123***
##                                     (0.005)      (0.013)      (0.014)
##
## MonthlyCharges                      0.028***     0.023***     0.024***
##                                     (0.002)      (0.007)      (0.007)
##
## TotalCharges                        0.0002*      0.001***     0.001**
##                                     (0.0001)     (0.0002)     (0.0002)
##
## SeniorCitizenYes                    0.630***     0.634***     1.477***
##                                     (0.079)      (0.080)      (0.399)
##
## genderMale                           -0.004       -0.007       0.247
##                                     (0.062)      (0.063)      (0.235)
##
## I(tenure2)                                       0.001***     0.001***
##                                                  (0.0001)     (0.0001)
##
## I(MonthlyCharges2)                                0.00003      0.0001
##                                                  (0.0001)     (0.0001)
##
## I(TotalCharges2)                                -0.00000***  -0.00000***
##                                                  (0.00000)    (0.00000)
##
## tenure:SeniorCitizenYes                                       0.013
##                                                              (0.013)
##
## MonthlyCharges:SeniorCitizenYes                              -0.013*
```

```
##                                                          (0.005)
##
## TotalCharges:SeniorCitizenYes                            -0.0001
##                                                         (0.0002)
##
## tenure:genderMale                                         -0.010
##                                                           (0.010)
##
## MonthlyCharges:genderMale                                 -0.006
##                                                           (0.003)
##
## TotalCharges:genderMale                                   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<0.05; **p<0.01; ***p<0.001
```

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.

```
mod1 <- glm(formula = Churn ~ tenure + MonthlyCharges + TotalCharges + SeniorCitizen +
              gender, family=binomial(link=logit), data=telcom_churn)

Anova(mod1, test="LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Churn
##                LR Chisq Df Pr(>Chisq)
## tenure          192.288  1  < 2.2e-16 ***
## MonthlyCharges  289.800  1  < 2.2e-16 ***
## TotalCharges      6.021  1    0.01414 *
## SeniorCitizen    62.612  1  2.517e-15 ***
## gender            0.004  1    0.94700
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using the Anova function from the car package, the Likelihood Ratio Test indicates that the p-values for the effect of tenure, MonthlyCharges, and SeniorCitizen are highly statistically significant, meaning that these variables are important given that the other variables are in the model. Additionally, the effect of TotalCharges on customer churn has a p-value of 0.01414, which is less than a cutoff value of 0.05. Hence, there is evidence that TotalCharges is important, given that the other variables are in the model. We notice that the effect of gender on customer churn has a p-value of 0.94700, which means that we do not have strong evidence that gender is important, given that the other variables are in the model.

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

Perform a likelihood ratio test to assess the hypothesis that $\beta_6 = 0$, $\beta_7 = 0$, and $\beta_8 = 0$ 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`.

```
mod2_H0 <- glm(formula = Churn ~ tenure + MonthlyCharges + TotalCharges + SeniorCitizen +
                 gender, family=binomial(link=logit), data=telcom_churn)
mod2_Ha <- glm(formula = Churn ~ tenure + MonthlyCharges + TotalCharges + SeniorCitizen +
                 gender + I(tenure^2) + I(MonthlyCharges^2) + I(TotalCharges^2),
               family=binomial(link=logit), data=telcom_churn)

anova(mod2_H0, mod2_Ha, test = "Chisq")
```

```
## 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
```

Using anova, we performed an LRT to assess the null hypothesis that $\beta_6 = \beta_7 = \beta_8 = 0$ and the alternate hypothesis that at least one of the betas is not 0. From the results above, the p-value is 8.232e-08 using a Chi-Squared approximation. Since this result is highly statistically significant underneath a p-value cutoff of 0.001, we reject the null hypothesis. Therefore, there is strong evidence that at least one of the quadratic transformations of tenure, MonthyCharges, and TotalCharges are important, given that tenure, MonthyCharges, TotalCharges, SeniorCitizen, and gender are in the model.

```
mod3_H0 <- glm(formula = Churn ~ tenure + MonthlyCharges + TotalCharges + SeniorCitizen +
                 gender + SeniorCitizen:tenure + SeniorCitizen:MonthlyCharges +
                 SeniorCitizen:TotalCharges + gender:tenure + gender:MonthlyCharges +
                 gender:TotalCharges, family=binomial(link=logit), data=telcom_churn)

mod3_Ha <- 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, family=binomial(link=logit), data=telcom_churn)

anova(mod3_H0, mod3_Ha, test = "Chisq")
```

```
## 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
```

Using anova, we performed an LRT to assess the null hypothesis that $\beta_6 = \beta_7 = \beta_8 = 0$ and the alternate hypothesis that at least one of the betas is not 0. From the results above, the p-value is 4.958e-07 using a Chi-Squared approximation. Since this result is highly statistically significant underneath a p-value cutoff of 0.001, we reject the null hypothesis. Therefore, there is strong evidence that at least one of the quadratic transformations of tenure, MonthyCharges, and TotalCharges are important, given that tenure, MonthyCharges, TotalCharges, SeniorCitizen, and gender are in the model. The result also assumes that there is an interaction between SeniorCitizen and tenure, SeniorCitizen and MonthlyCharges, SeniorCitizen and TotalCharges, gender and tenure, gender and MonthlyCharges, as well as gender and TotalCharges.

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.

```
mod3_H0 <- 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)

mod3_Ha <- 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, family=binomial(link=logit), data=telcom_churn)


anova(mod3_H0, mod3_Ha, test = "Chisq")
```

```
## 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
```

Using the anova function to perform the LRT, we assess null hypothesis that the coefficients of gender and its interaction terms are all 0, as well as the alternative hypothesis that at least one of the coefficients of gender and its interaction terms are non-zero. From the results, we obtain a deviance value of 9.5332 and a p-value of 0.04907. Since the p-value is slightly less than the 0.05 significance level, we reject the null hypothesis that the coefficients of gender and its interaction terms are all 0. Hence, we have marginal evidence that gender has an effect on churn.

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.

```
Anova(mod3, test="LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Churn
##                               LR Chisq Df Pr(>Chisq)
## tenure                         104.850  1  < 2.2e-16 ***
## MonthlyCharges                  11.405  1  0.0007324 ***
## TotalCharges                    14.810  1  0.0001189 ***
## SeniorCitizen                   63.458  1  1.638e-15 ***
## gender                           0.026  1  0.8727620
## I(tenure^2)                     31.383  1  2.118e-08 ***
## I(MonthlyCharges^2)              1.435  1  0.2309311
## I(TotalCharges^2)               16.528  1  4.795e-05 ***
## tenure:SeniorCitizen             0.961  1  0.3270439
## MonthlyCharges:SeniorCitizen     5.645  1  0.0175095 *
## TotalCharges:SeniorCitizen       0.261  1  0.6093226
## tenure:gender                    0.829  1  0.3626033
## MonthlyCharges:gender            3.407  1  0.0649308 .
## TotalCharges:gender              3.193  1  0.0739758 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the Anova function above, we see that only the variables tenure, MonthlyCharges, TotalCharges, SeniorCitizen, tenure squared, TotalCharges squared, and the interaction variable between MonthlyCharges and SeniorCitizen are significant.

```
mod4$coefficients
```

```
##                      (Intercept)                           tenure
##                    -1.404039e+00                    -1.270122e-01
##                   MonthlyCharges                     TotalCharges
##                     2.815008e-02                     6.208794e-04
##                 SeniorCitizenYes                      I(tenure^2)
##                     1.548575e+00                     8.020748e-04
##                I(TotalCharges^2) MonthlyCharges:SeniorCitizenYes
##                    -6.058805e-08                    -1.146578e-02
```

```
beta0 <- mod4$coefficients[1]
beta0
```

```
## (Intercept)
##   -1.404039
```

```
beta1 <- mod4$coefficients[2]
beta1
```

```
##     tenure
## -0.1270122
```

```
beta2 <- mod4$coefficients[3]
beta2
```

```
## MonthlyCharges
##     0.02815008
```

```
beta3 <- mod4$coefficients[4]
beta3
```

```
## TotalCharges
## 0.0006208794
```

```
beta4 <- mod4$coefficients[5]
beta4
```

```
## SeniorCitizenYes
##         1.548575
```

```
beta5 <- mod4$coefficients[6]
beta5
```

```
##  I(tenure^2)
## 0.0008020748
```

```
beta6 <- mod4$coefficients[7]
beta6
```

```
## I(TotalCharges^2)
##     -6.058805e-08
```

```
beta7 <- mod4$coefficients[8]
beta7
```

```
## MonthlyCharges:SeniorCitizenYes
##                      -0.01146578
```

With the coefficients identified from model 4, we want to assess the change in the likelihood of churn for senior customers compared to non-senior customers. This can be accomplished with odds ratios.

```
avgMonthyCharges <- mean(telcom_churn$MonthlyCharges)
avgMonthyCharges
```

```
## [1] 64.79821
```

```
OR <- exp(beta0 + beta4 + beta7 * avgMonthyCharges) / exp(beta0)
OR
```

```
## (Intercept)
##    2.238069
```

```
exp(beta4 + beta7 * avgMonthyCharges)
```

```
## SeniorCitizenYes
##         2.238069
```

The result shows us that the estimated odds of churning are about 2.24 times as large for senior customers than for non-senior customers when we hold tenure, MonthlyCharges, and TotalCharges at their average values.

```
1 / exp(beta4 + beta7 * avgMonthyCharges)
```

```
## SeniorCitizenYes
##        0.4468137
```

Alternatively, the estimated odds of churning are 0.4468137 times as large for non-seniors than for senior customers when we hold tenure, MonthlyCharges, and TotalCharges at their average values.

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

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

```
predict_data_1 <- data.frame(tenure=55, MonthlyCharges=89.86,
                             TotalCharges=3794.7, SeniorCitizen="No")
predict_data_2 <- data.frame(tenure=29.00, MonthlyCharges=18.25,
                             TotalCharges=401.4, SeniorCitizen="Yes")

logit_pred_1 <- predict(mod4, newdata=predict_data_1, type="link", se.fit=TRUE)
logit_pred_2 <- predict(mod4, newdata=predict_data_2, type="link", se.fit=TRUE)

logit_mod_1 <- logit_pred_1$fit
logit_mod_2  <- logit_pred_2$fit

logit_se_1 <- logit_pred_1$se.fit
logit_se_2 <- logit_pred_2$se.fit

ci_logit_lower_1 <- logit_mod_1 - 1.96 * logit_se_1
ci_logit_upper_1 <- logit_mod_1 + 1.96 * logit_se_1

ci_logit_lower_2 <- logit_mod_2 - 1.96 * logit_se_2
ci_logit_upper_2 <- logit_mod_2 + 1.96 * logit_se_2

ci_prob_lower_1 <- exp(ci_logit_lower_1) / (1 + exp(ci_logit_lower_1))
ci_prob_upper_1 <- exp(ci_logit_upper_1) / (1 + exp(ci_logit_upper_1))

ci_prob_lower_2 <- exp(ci_logit_lower_2) / (1 + exp(ci_logit_lower_2))
ci_prob_upper_2 <- exp(ci_logit_upper_2) / (1 + exp(ci_logit_upper_2))
```

```
c(ci_prob_lower_1, ci_prob_upper_1)
```

```
##         1         1
## 0.1056106 0.1462704
```

The 95% Wald CI for the probability of Churn for profile 1 is 0.1056106 and 0.1462704.

```
c(ci_prob_lower_2, ci_prob_upper_2)
```

```
##         1         1
## 0.0562602 0.1393393
```

The 95% Wald CI for the probability of Churn for profile 2 is 0.0562602 and 0.1393393.