W271 Assignment 2

Contents

```r
library(tidyverse)
```

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-2 (100 Points)

In the previous homework assignment, you began modeling a binary variable using customer churn data from a telecommunications company to analyze churn tendencies among senior and non-senior customers.

Now, in Part-2 of the homework, we will delve into regression techniques to develop a more comprehensive model for the telecom company. This model will provide insights into the reasons why customers may choose to discontinue their services.

```r
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
```

Churn dataset consists of 21 variables and 7043 observations. The customer variables are provided below:

For the remainder of this section, pay particular attention to `Churn`, `tenure`, `MonthlyCharges`, and `TotalCharges`.

1.1 Data Preprocessing (5 Points)

In this section, review the data structure to ensure the correct data types for variables of interest, convert variables as necessary, and address any missing values.

```r
#Convert Churn variable from character to binary int - 0 or 1.

telcom_churn <- telcom_churn %>% mutate(Churn = case_when(Churn == 'Yes' ~ 1,
                                         Churn == 'No' ~ 0,
                                         TRUE ~ NA))

#telcom_churn$Churn <- factor(telcom_churn$Churn)
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     0
## 2              Mailed check          56.95      1889.50     0
## 3              Mailed check          53.85       108.15     1
## 4 Bank transfer (automatic)          42.30      1840.75     0
## 5          Electronic check          70.70       151.65     1
## 6          Electronic check          99.65       820.50     1
```

```r
summary(telcom_churn)
```

```
##   customerID           gender          SeniorCitizen      Partner
##  Length:7043        Length:7043        Min.   :0.0000   Length:7043
##  Class :character   Class :character   1st Qu.:0.0000   Class :character
##  Mode  :character   Mode  :character   Median :0.0000   Mode  :character
##                                        Mean   :0.1621
##                                        3rd Qu.:0.0000
##                                        Max.   :1.0000
##
##   Dependents            tenure      PhoneService       MultipleLines
##  Length:7043        Min.   : 0.00   Length:7043        Length:7043
##  Class :character   1st Qu.: 9.00   Class :character   Class :character
##  Mode  :character   Median :29.00   Mode  :character   Mode  :character
##                     Mean   :32.37
##                     3rd Qu.:55.00
##                     Max.   :72.00
##
##  InternetService    OnlineSecurity     OnlineBackup       DeviceProtection
##  Length:7043        Length:7043        Length:7043        Length:7043
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  TechSupport        StreamingTV        StreamingMovies      Contract
##  Length:7043        Length:7043        Length:7043        Length:7043
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  PaperlessBilling   PaymentMethod      MonthlyCharges    TotalCharges
##  Length:7043        Length:7043        Min.   : 18.25   Min.   :  18.8
##  Class :character   Class :character   1st Qu.: 35.50   1st Qu.: 401.4
##  Mode  :character   Mode  :character   Median : 70.35   Median :1397.5
##                                        Mean   : 64.76   Mean   :2283.3
##                                        3rd Qu.: 89.85   3rd Qu.:3794.7
##                                        Max.   :118.75   Max.   :8684.8
##                                                         NA's   :11
##      Churn
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.2654
##  3rd Qu.:1.0000
##  Max.   :1.0000
##
```

```
any(is.na(telcom_churn$TotalCharges.))
```

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

```
any(is.na(telcom_churn$Churn))
```

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

1.2 Maximum Likelihood (15 Points)

Let's build off of the maximum likelihood model of a binomial distribution from lecture and apply it to the churn data set.

Our objective is to estimate the probability of a customer churning based on their `tenure` with the company. While we will use logistic regression in subsequent sections, here, we will focus on the maximum likelihood approach.

Suppose that we can express the probability of a customer churning as a function of tenure in the following form (you should recognize this as the connection between log odds and probability from the lecture):

$$P(Churn) = P(\alpha, \beta) = \frac{e^{\alpha+\beta*Tenure}}{1+e^{\alpha+\beta*Tenure}}$$

Using this and assuming the number of churned customers in the data set follows a binomial distribution with parameters n and $p(\alpha, \beta)$, **write down the likelihood function** $L(\alpha, \beta|Data)$.

1.3 Write and compute the log-likelihood (10 Points)

Find the **negative log likelihood** and write an R function to calculate it given inputs of alpha and beta and using the churn data.

```
negativeLogL <- function(params, x, Y) {
  pi.hat <- exp(params[1] + params[2]* x) / (1 + exp(params[1] + params[2]*x))
  -1 * sum(Y * log(pi.hat) + (1 - Y) * log(1 - pi.hat))
}
```

1.4 Compute the MLE of parameters (10 Points)

Use the optim function to **find the MLE of alpha and beta on the churn data**. You can use starting values of 0 for both parameters. Note that optim by default finds the minimum, so you can use the negative log likelihood directly.

```
params <- c(0,0)
mod.fit.optim <- optim(par = params, fn = negativeLogL, hessian=TRUE, x=telcom_churn$tenure, Y=telcom_ch
names(mod.fit.optim)
```

```
## [1] "par"         "value"       "counts"      "convergence" "message"
## [6] "hessian"
```

```
mod.fit.optim$par
```

```
## [1]  0.02766712 -0.03877677
```

The MLE of alpha is 0.0277 and the MLE of beta is -0.0388.

1.5 Calculate a confidence interval (10 Points)

Again using the optim function, find the **variance of the MLE estimates** (hint use hessian = TRUE in optim) for alpha and beta. Calculate a **95% confidence interval** for each parameter. Are they statistically different than zero?

```
cov_matrix <- solve(mod.fit.optim$hessian)
cov_matrix
```

```
##               [,1]          [,2]
## [1,]  1.782206e-03 -4.323216e-05
## [2,] -4.323216e-05  1.973784e-06
```

```
var.alpha <- cov_matrix[1, 1]
var.alpha
```

```
## [1] 0.001782206
```

```
var.beta <- cov_matrix[2, 2]
var.beta
```

```
## [1] 1.973784e-06
```

```
a <- 0.05
alpha_estimate <- mod.fit.optim$par[1]
alpha_estimate
```

```
## [1] 0.02766712
```

```
alpha_estimate + qnorm(p = c( a /2, 1- a /2)) * sqrt(var.alpha)
```

```
## [1] -0.05507508  0.11040931
```

Using a 95% Wald Confidence interval, with 95% confidence the true value of the parameter alpha lies in the interval -0.05507508 and 0.11040931. Since the interval contains 0, alpha is not statistically different than 0.

```
a <- 0.05
beta_estimate <- mod.fit.optim$par[2]
beta_estimate
```

```
## [1] -0.03877677
```

```
beta_estimate + qnorm (p = c(a /2, 1- a/2)) * sqrt(var.beta)
```

```
## [1] -0.04153035 -0.03602319
```

Using a 95% Wald Confidence interval, with 95% confidence the true value of the parameter beta lies in the interval -0.04153035 and -0.03602319. Since the interval does not contain 0, beta is statistically different than 0.

1.6 Model comparison (10 Points)

Estimate a logistic regression model with `tenure` as the independent variable. Compare **MLE of alpha and beta to the output of the logistic regression**. What do you notice? Can you think of why this is the case? (Think about the connection between MLE of regression coefficients and linear regression)

```
logRegFit <- glm(formula = Churn ~ tenure, family = binomial(link=logit), data=telcom_churn)
logRegFit
```

```
##
## Call:  glm(formula = Churn ~ tenure, family = binomial(link = logit),
##     data = telcom_churn)
```

```
##
## Coefficients:
## (Intercept)       tenure
##     0.02731     -0.03877
##
## Degrees of Freedom: 7042 Total (i.e. Null);  7041 Residual
## Null Deviance:       8150
## Residual Deviance: 7192  AIC: 7196
```

```
mod.fit.optim
```

```
## $par
## [1]  0.02766712 -0.03877677
##
## $value
## [1] 3595.934
##
## $counts
## function gradient
##       28        6
##
## $convergence
## [1] 0
##
## $message
## NULL
##
## $hessian
##           [,1]       [,2]
## [1,]  1197.198   26222.46
## [2,] 26222.455 1080996.58
```

Using the glm function, estimating a logistic regression model with tenure as the independent variable yields the parameter estimates for alpha as 0.0277 and beta as -0.0388. These estimates are similar to those produced using the optim() function, with minor differences due to different convergent criteria between the methods produced by glm() and optim().

1.7 Extended Model, with Linear Effects (10 Points)

Use the `Churn`, `tenure`, `MonthlyCharges`, and `TotalCharges` as independent variables in a logistic regression model for predicting a customer churning. Proceed to estimate the model and subsequently, interpret each of the indicator variables incorporated within the model.

```
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     0
## 2              Mailed check          56.95      1889.50     0
## 3              Mailed check          53.85       108.15     1
## 4 Bank transfer (automatic)          42.30      1840.75     0
## 5          Electronic check          70.70       151.65     1
## 6          Electronic check          99.65       820.50     1
```

```r
logRegFit <- glm(formula = Churn ~ tenure + MonthlyCharges + TotalCharges, family = binomial(link=logit
summary(logRegFit)
```

```
##
## Call:
## glm(formula = Churn ~ tenure + MonthlyCharges + TotalCharges,
##     family = binomial(link = logit), data = telcom_churn)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.599e+00  1.173e-01 -13.628   <2e-16 ***
## tenure         -6.711e-02  5.458e-03 -12.297   <2e-16 ***
## MonthlyCharges  3.020e-02  1.717e-03  17.585   <2e-16 ***
## TotalCharges    1.451e-04  6.144e-05   2.361   0.0182 *
## ---
## 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: 6376.2  on 7028  degrees of freedom
##   (11 observations deleted due to missingness)
## AIC: 6384.2
##
## Number of Fisher Scoring iterations: 6
```

The estimated coefficient for tenure is about -0.0671. This means that an increase in tenure by 1 unit is associated with a 0.067 decrease in the log odds, where the odds is the ratio of the probability churning over the probability not churning. The estimated coefficient for MonthyCharges is about 0.0302. This means that an increase in MonthyCharges by 1 unit is associated with a 0.0302 increase in the log odds, holding the other variables constant. The estimated coefficient for TotalCharges is about 0.00015. This means that an increase in TotalCharges by 1 unit is associated with a 0.00015 increase in the log odds, holding the other variables constant. All coefficients are statistically significant. Hence, tenure has a slight negative association with the probability of churning. MonthyCharges has a slight positive association with the probability of churning. TotalCharges has a slight positive correlation with the probability of churning.

1.8 Likelihood Ratio Tests (10 Points)

Perform likelihood ratio tests for all independent variables to evaluate their importance within the model. Discuss and interpret the results of these tests.

```
library(package=car)
```

```
## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

## The following object is masked from 'package:purrr':
##
##     some
```

```
logRegFit <- glm(formula = Churn ~ tenure + MonthlyCharges + TotalCharges, family = binomial(link=logit
Anova (logRegFit, test = "LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Churn
##                LR Chisq Df Pr(>Chisq)
## tenure           190.56  1    < 2e-16 ***
## MonthlyCharges   342.74  1    < 2e-16 ***
## TotalCharges       5.67  1    0.01728 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Anova() function with the test = "LR" argument was used to perform a Likelihood Ratio Test.

For the test of tenure with the null hypothesis stating that its coefficient is 0 and the alternate hypothesis stating that its coefficient is non-zero, we obtain a Chi-Squared statistic of 190.56 with p-value < 2e-16. Using a cutoff of 0.05, we would reject the null hypothesis that the coefficient of tenure is zero. Hence, there is strong evidence that tenure is important given that all the other independent variables are in the model.

For the test of MonthlyCharges with the null hypothesis stating that its coefficient is 0 and the alternate hypothesis stating that its coefficient is non-zero, we obtain a Chi-Squared statistic of 342.74 with p-value < 2e-16. Using a cutoff of 0.05, we would reject the null hypothesis that the coefficient of MonthlyCharges is zero. Hence, there is strong evidence that MonthlyCharges is important given that all the other independent variables are in the model.

For the test of TotalCharges with the null hypothesis stating that its coefficient is 0 and the alternate hypothesis stating that its coefficient is non-zero, we obtain a Chi-Squared statistic of 5.67 with p-value 0.017. Using a cutoff of 0.05, we would reject the null hypothesis that the coefficient of TotalCharges is zero. Hence, there is strong evidence that TotalCharges is important given that all the other independent variables are in the model.

1.9 Effect of change in Monthly payments (10 Points)

What is the effect of a standard deviation increase in `MonthlyCharges` on the odds of the customer getting churned? Also, calculate the Wald CI for the odds ratio.

```
logRegFit <- glm(formula = Churn ~ tenure + MonthlyCharges + TotalCharges, family = binomial(link=logit
month_charge_coef <- coef(logRegFit)["MonthlyCharges"]
sd_monthyCharges <- sd(telcom_churn$MonthlyCharges)
```

```r
effect_odds <- exp(sd_monthyCharges * month_charge_coef)
effect_odds
```

```
## MonthlyCharges
##       2.481122
```

The odds of the customer churning increase by 2.48 times for every standard-deviation-unit increase in MonthyCharges.

```r
std_err_monthyCharges <- sqrt(diag(vcov(logRegFit))["MonthlyCharges"])

a <- 0.05
z <- qnorm(1-a/2)
lower_bound <- exp(sd_monthyCharges * (month_charge_coef - z * std_err_monthyCharges))
upper_bound <- exp(sd_monthyCharges * (month_charge_coef + z * std_err_monthyCharges))

cat("95% Wald Confidence Interval for the Odds Ratio of a Standard Deviation
Increase in MonthlyCharges:\n",
    "Lower Bound:", round(lower_bound, 4), "\n",
    "Upper Bound:", round(upper_bound, 4), "\n")
```

```
## 95% Wald Confidence Interval for the Odds Ratio of a Standard Deviation
## Increase in MonthlyCharges:
##  Lower Bound: 2.2421
##  Upper Bound: 2.7456
```

1.10 Confidence Interval for the Probability of Success (10 Points)

Estimate the 95% profile likelihood confidence interval for the probability of a customer getting churned, considering an average `tenure`, `MonthlyCharges`, and `TotalCharges`.

```r
mean_tenure <- mean(telcom_churn$tenure, na.rm = TRUE)
mean_monthlyCharges <- mean(telcom_churn$MonthlyCharges, na.rm = TRUE)
mean_totalCharges <- mean(telcom_churn$TotalCharges, na.rm = TRUE)


fitted_mod <- predict(logRegFit,
                      newdata = data.frame(tenure = mean_tenure,
                                           MonthlyCharges =
mean_monthlyCharges,
                                           TotalCharges =
mean_totalCharges),
                      type = "link", se.fit = TRUE)

lower_bound <- exp(fitted_mod$fit - z * fitted_mod$se.fit) / (1 +
exp(fitted_mod$fit - z * fitted_mod$se.fit))
upper_bound <- exp(fitted_mod$fit + z * fitted_mod$se.fit) / (1 +
exp(fitted_mod$fit + z * fitted_mod$se.fit))

cat("95% Wald Confidence Interval for the Probability of a Customer Churning (with
average tenure, MonthlyCharges, and TotalCharges):\n",
    "Lower Bound:", round(lower_bound, 4), "\n",
    "Upper Bound:", round(upper_bound, 4), "\n")
```

```
## 95% Wald Confidence Interval for the Probability of a Customer Churning (with
## average tenure, MonthlyCharges, and TotalCharges):
##  Lower Bound: 0.1724
##  Upper Bound: 0.1978
```

Using a 95% Wald CI, we are 95% confident that the true probability of an average customer churning would lie between 0.1724 and 0.1978.