W271 Assignment 2

Emanuel Mejía

Contents


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

```
## Warning: package 'mcprofile' was built under R version 4.1.3
```

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.

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

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.

We can first check the type of data we have in each column:

```
sapply(telcom_churn, function(x) typeof(x))
```

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

Since we'll be manually doing certain computations we'll turn the Churn column to have integers instead:

```
telcom_churn["Churn"][telcom_churn["Churn"] == "No"] <- 0
telcom_churn["Churn"][telcom_churn["Churn"] == "Yes"] <- 1

telcom_churn$Churn <- as.integer(telcom_churn$Churn)
```

We can also check the count of missing values for every column in our data:

```
sapply(telcom_churn, function(x) sum(is.na(x)))
```

```
##       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
```

We'll be using the Total Charges variable, which is the only column with missing values. Since we have enough observations we can drop the ones with NA.

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

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** $L(\alpha, \beta|Data)$.

$$\begin{aligned}
L(\alpha, \beta|Data) &= L(\alpha, \beta|y_1, y_2, ..., y_n) \\
&= \prod_{i=1}^{n} p(\alpha, \beta)_i^{y_i} \left(1 - p(\alpha, \beta)_i\right)^{1-y_i} \\
&= \prod_{i=1}^{n} \left[\left(\frac{e^{\alpha+\beta*T_i}}{1+e^{\alpha+\beta*T_i}} \right)^{y_i} \left(1 - \frac{e^{\alpha+\beta*T_i}}{1+e^{\alpha+\beta*T_i}} \right)^{1-y_i} \right] \qquad \text{with } Ti = Tenure \text{ for obs } i
\end{aligned}$$

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.

From the result above we can compute

$$\begin{aligned}
-log(L(\alpha, \beta|Data)) &= -log\left(\prod_{i=1}^{n} \left[\left(\frac{e^{\alpha+\beta*T_i}}{1+e^{\alpha+\beta*T_i}} \right)^{y_i} \left(1 - \frac{e^{\alpha+\beta*T_i}}{1+e^{\alpha+\beta*T_i}} \right)^{1-y_i} \right] \right) \\
&= -\sum_{i=1}^{n} y_i log\left(\frac{e^{\alpha+\beta*T_i}}{1+e^{\alpha+\beta*T_i}} \right) + (1-y_i) log\left(1 - \frac{e^{\alpha+\beta*T_i}}{1+e^{\alpha+\beta*T_i}} \right) \\
&= -\sum_{i=1}^{n} y_i log\left(\frac{e^{\alpha+\beta*T_i}}{1+e^{\alpha+\beta*T_i}} \right) + (1-y_i) log\left(\frac{1}{1+e^{\alpha+\beta*T_i}} \right) \\
&= -\sum_{i=1}^{n} y_i log\left(\frac{e^{\alpha+\beta*T_i}}{1+e^{\alpha+\beta*T_i}} \right) + log\left(\frac{1}{1+e^{\alpha+\beta*T_i}} \right) - y_i log\left(\frac{1}{1+e^{\alpha+\beta*T_i}} \right) \\
&= -\sum_{i=1}^{n} y_i \left[log\left(\frac{e^{\alpha+\beta*T_i}}{1+e^{\alpha+\beta*T_i}} \right) - log\left(\frac{1}{1+e^{\alpha+\beta*T_i}} \right) \right] + log\left(\frac{1}{1+e^{\alpha+\beta*T_i}} \right) \\
&= -\sum_{i=1}^{n} y_i \left[log\left(\frac{e^{\alpha+\beta*T_i}}{1+e^{\alpha+\beta*T_i}} \left(1+e^{\alpha+\beta*T_i}\right) \right) \right] + log\left(\frac{1}{1+e^{\alpha+\beta*T_i}} \right) \\
&= -\sum_{i=1}^{n} y_i log\left(e^{\alpha+\beta*T_i} \right) + log\left(\frac{1}{1+e^{\alpha+\beta*T_i}} \right)
\end{aligned}$$

$$
\begin{aligned}
&= -\sum_{i=1}^{n} y_i \left(\alpha + \beta * T_i\right) + log\left(\frac{1}{1+e^{\alpha+\beta*T_i}}\right) \\
&= -\sum_{i=1}^{n} y_i log\left(e^{\alpha+\beta*T_i}\right) - log\left(1+e^{\alpha+\beta*T_i}\right) \\
&= -\sum_{i=1}^{n} y_i \left(\alpha + \beta * T_i\right) - log\left(1+e^{\alpha+\beta*T_i}\right) \\
&= \sum_{i=1}^{n} log\left(1+e^{\alpha+\beta*T_i}\right) - y_i \left(\alpha + \beta * T_i\right)
\end{aligned}
$$

Now if we put it in R:

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

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.

```
mod.fit.optim <- optim(
  par = c(0,0),
  fn = neglogL,
  hessian = T,
  x = telcom_churn$tenure,
  Y = as.numeric(telcom_churn$Churn),
  method = "BFGS"
)
```

We can test that our model indeed converged as follows:

```
mod.fit.optim$convergence
```

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

And get the parameters:

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

```
## [1]  0.03765195 -0.03901939
```

So now we know:

$\hat{\alpha}$ = 0.0376519

$\hat{\beta}$ = -0.0390194

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?

We first find the Var-Cov Matrix as follows:

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

```
##                [,1]          [,2]
## [1,]  1.790616e-03 -4.349281e-05
## [2,] -4.349281e-05  1.983769e-06
```

So we have the following variances:

$\widehat{Var}(\hat{\alpha}) = 0.0017906$

$\widehat{Var}(\hat{\beta}) = 1.9837693 \times 10^{-6}$

And now we'll compute the following Wald confidence intervals for each parameter:

```
alpha.wci <- mod.fit.optim$par[1] + qnorm(p = c(0.025, 0.975)) * sqrt(mod.optim.vcov[1,1])
alpha.wci
```

```
## [1] -0.04528525  0.12058914
```

```
beta.wci <- mod.fit.optim$par[2] + qnorm(p = c(0.025, 0.975)) * sqrt(mod.optim.vcov[2,2])
beta.wci
```

```
## [1] -0.04177992 -0.03625885
```

Which means that:

- The 95% Wald confidence interval for α is -0.0453 $< \alpha <$ 0.1206, so it seems that it's **not statistically different than zero** (since 0 is part of the interval). Nevertheless, this is our Intercept coefficient, so we might just use it anyways.
- The 95% Wald confidence interval for β (tenure parameter) is -0.0418 $< \beta <$ -0.0363, so it seems that is indeed **statistically different than zero** (since 0 is not part of the interval). Meaning that in a hypothesis test, we would reject a *null hypothesis* of the form $H_0 : \beta = 0$, and we would include Tenure in our model.

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)

```
mod.fit <- glm(
  formula = as.numeric(Churn) ~ tenure,
  family = binomial(link = logit),
  data = telcom_churn
)
summary(mod.fit)
```

```
##
## Call:
## glm(formula = as.numeric(Churn) ~ tenure, family = binomial(link = logit),
##     data = telcom_churn)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1767  -0.8401  -0.4788   1.1781   2.3800
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.037299   0.042319   0.881    0.378
## tenure      -0.039010   0.001409 -27.691   <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: 7176.3  on 7030  degrees of freedom
## AIC: 7180.3
##
## Number of Fisher Scoring iterations: 4
```

The parameters α and β are the same in both methods, the standard error and the variance are also consistent:

$\widehat{Var}(\hat{\alpha}) = 0.0017906 \implies SE = \sqrt{\widehat{Var}(\hat{\alpha})} = 0.0423157$

$\widehat{Var}(\hat{\beta}) = 1.9837693 \times 10^{-6} \implies SE = \sqrt{\widehat{Var}(\hat{\beta})} = 0.0014085$

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.

```
mod.fit.full <- glm(
  formula = as.numeric(Churn) ~ tenure + MonthlyCharges + TotalCharges,
  family = binomial(link = logit),
  data = telcom_churn
)
summary(mod.fit.full)
```

```
##
## Call:
## glm(formula = as.numeric(Churn) ~ tenure + MonthlyCharges + TotalCharges,
##     family = binomial(link = logit), data = telcom_churn)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8474  -0.7316  -0.4042   0.8036   3.1441
##
```

```
## 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
## AIC: 6384.2
##
## Number of Fisher Scoring iterations: 6
```

So our logistic regression model is the following:

$logit(\hat{\pi})$ = -1.5988 + -0.0671 $tenure$+ 0.0302 $MonthlyCharges$+ 0.000145 $TotalCharges$

Which basically means the following:

- Given that the other variables are in the model we have enough evidence that there's a negative relationship between `tenure` and π: The higher the `tenure` the lower the probability of `Churn`.
- Given that the other variables are in the model we have enough evidence that there's a positive relationship between `MonthlyCharges` and π: The higher the `MonthlyCharges` the higher the probability of `Churn`.
- Given that the other variables are in the model we have marginal evidence that there's We have a positive relationship between `TotalCharges` and π: The higher the `TotalCharges` the higher the probability of `Churn`.

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.

We can do these tests with one of two methods, the first one is by using the Anova function as follows:

```
Anova.tests <- Anova(mod.fit.full, test = "LR")
LR.chisq <- Anova.tests[["LR Chisq"]]
p.val.chisq <- Anova.tests[["Pr(>Chisq)"]]
Anova.tests
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: as.numeric(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 second one is by computing them through the "anova" function. Atlhough we need to do it one by one, as follows:

```
mod.fit.H0_ten <- glm(
  formula = as.numeric(Churn) ~ MonthlyCharges + TotalCharges,
  family = binomial(link = logit),
  data = telcom_churn
)

mod.fit.H0_Mon <- glm(
  formula = as.numeric(Churn) ~ tenure + TotalCharges,
  family = binomial(link = logit),
  data = telcom_churn
)

mod.fit.H0_Tot <- glm(
  formula = as.numeric(Churn) ~ tenure + MonthlyCharges,
  family = binomial(link = logit),
  data = telcom_churn
)
```

```
anova(mod.fit.H0_ten, mod.fit.full, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: as.numeric(Churn) ~ MonthlyCharges + TotalCharges
## Model 2: as.numeric(Churn) ~ tenure + MonthlyCharges + TotalCharges
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      7029     6566.8
## 2      7028     6376.2  1   190.56 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(mod.fit.H0_Mon, mod.fit.full, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: as.numeric(Churn) ~ tenure + TotalCharges
## Model 2: as.numeric(Churn) ~ tenure + MonthlyCharges + TotalCharges
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      7029     6719.0
## 2      7028     6376.2  1   342.74 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(mod.fit.H0_Tot, mod.fit.full, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: as.numeric(Churn) ~ tenure + MonthlyCharges
## Model 2: as.numeric(Churn) ~ tenure + MonthlyCharges + TotalCharges
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
```

```
## 1      7029     6381.9
## 2      7028     6376.2  1   5.6672  0.01728 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

These results are consistent with the ones obtained in question 1.7:

- For the test of `tenure` with $H_0 : \beta_1 = 0$ VS $H_a : \beta_1 \neq 0$ we obtain $-2log(\Delta) = 190.5617$ and a p-value of $\mathbb{P}(A > 190.5617\ ) = 2.3973662 \times 10^{-43}$ meaning that we have enough evidence that `tenure` is important to include in the model (given that the other variables are in the model as well).
- For the test of `MonthlyCharges` with $H_0 : \beta_2 = 0$ VS $H_a : \beta_2 \neq 0$ we obtain $-2log(\Delta) = 342.7401$ and a p-value of $\mathbb{P}(A > 342.7401\ ) = 1.614873 \times 10^{-76}$ meaning that we have enough evidence that `MonthlyCharges` is important to include in the model (given that the other variables are in the model as well).
- For the test of `TotalCharges` with $H_0 : \beta_3 = 0$ VS $H_a : \beta_3 \neq 0$ we obtain $-2log(\Delta) = 5.6672$ and a p-value of $\mathbb{P}(A > 5.6672\ ) = 0.0172846$ meaning that we have marginal evidence that `TotalCharges` is important to include in the model (given that the other variables are in the model as well). So, depending on the confidence level we select we might include this variable or not (with $\alpha = 0.05$ we would include it, and we wouldn't include it if using $\alpha = 0.01$)

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.

We have a model with a shape like:

$$logit(\pi) = \beta_0 + \beta_1 * tenure + \beta_2 * MonthlyCharges + \beta_3 * TotalCharges$$

We know that $OR = \frac{Odds_{x+c}}{Odds_x}$, and in this specific case, for `MonthlyCharges` we'll get the following:

$$\begin{aligned} OR &= \frac{Odds_{MonthlyCharges+c}}{Odds_{MonthlyCharges}} \\ &= \frac{e^{\beta_0+\beta_1*tenure+\beta_2*(MonthlyCharges+c)+\beta_3*TotalCharges}}{e^{\beta_0+\beta_1*tenure+\beta_2*MonthlyCharges+\beta_3*TotalCharges}} \\ &= e^{c\beta_2} \end{aligned}$$

Now, if we compute the standard deviation for `MonthlyCharges` we have the following:

```
sdMC <- sd(telcom_churn$MonthlyCharges)
sdMC
```

```
## [1] 30.08597
```

So $\sigma_{MonthlyCharges} = 30.09$ and from question 1.6 we know that $\beta_2 = 0.0302$. Therefore, the Odds Ratio for $c = \sigma_{MonthlyCharges}$ is:

```
OR_MC <- exp(sdMC * mod.fit.full$coefficients[3])
OR_MC
```

```
## MonthlyCharges
##       2.480817
```

And we have $OR = e^{\sigma_{MonthlyCharges}\beta_2} = 2.4808$, meaning that the odds of a customer churning change by 2.4808 times for every 30.09 ($\sigma_{MonthlyCharges}$)increase in `MonthlyCharges`. And the 95% Wald confidence interval for the OR would be:

```
beta2.ci <- confint.default(object = mod.fit.full, parm = "MonthlyCharges", level = 0.95)
OR.ci <- exp(beta2.ci * sdMC)
OR.ci
```

```
##                    2.5 %   97.5 %
## MonthlyCharges 2.241887 2.745211
```

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`.

We first compute all of the averages as follows:

```
avg_ten <- mean(telcom_churn$tenure)
avg_Mon <- mean(telcom_churn$MonthlyCharges)
avg_Tot <- mean(telcom_churn$TotalCharges)
```

Before calculating the profile likelihood CI, ee'll compute a 95% Wald Interval (just for comparison)

```
predict.data <- data.frame(tenure = avg_ten, MonthlyCharges = avg_Mon, TotalCharges = avg_Tot)
linear.pred <- predict(object = mod.fit.full, newdata = predict.data, type = "link", se = TRUE)
CI.lin.pred <- linear.pred$fit + qnorm(p = c(0.05/2, 1 - 0.05/2)) * linear.pred$se
CI.pi <- exp(CI.lin.pred) / (1+exp(CI.lin.pred))
CI.pi
```

```
## [1] 0.1720546 0.1974799
```

Now, to estimate the 95% Profile likelihood interval we'll compute -2log(Lambda) as follows:

```
K <- matrix(data = c(1,avg_ten, avg_Mon, avg_Tot), nrow = 1, ncol = 4)
linear.combo <- mcprofile(object = mod.fit.full, CM = K)
```

Finally creating the 95 % profile LR interval as follows:

```
ci.logit.profile <- confint(object = linear.combo, level = 0.95)
prob.ci <- exp(ci.logit.profile$confint)/ (1+exp(ci.logit.profile$confint))
prob.ci
```

```
##       lower     upper
## 1 0.1718234 0.1972319
```

Which is very similar to the Wald interval (because we have a very large sample size).