w271 HW2 Huibin Chang 2025 Fall

Notes

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.

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. 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.

1. Answer

First import the tidyverse library and take a look at the variables, dimensions, and data type:

```r
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.2     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.1.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```r
glimpse(telcom_churn)
```

```
Rows: 7,043
Columns: 21
$ customerID       <chr> "7590-VHVEG", "5575-GNVDE", "3668-QPYBK", "7795-
CFOCW~
$ gender           <chr> "Female", "Male", "Male", "Male", "Female", "Female",~
$ SeniorCitizen    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ Partner          <chr> "Yes", "No", "No", "No", "No", "No", "No", "No", "Yes~
$ Dependents       <chr> "No", "No", "No", "No", "No", "No", "Yes", "No", "No"~
$ tenure           <int> 1, 34, 2, 45, 2, 8, 22, 10, 28, 62, 13, 16, 58, 49, 2~
$ PhoneService     <chr> "No", "Yes", "Yes", "No", "Yes", "Yes", "Yes", "No", ~
$ MultipleLines    <chr> "No phone service", "No", "No", "No phone service", "~
$ InternetService  <chr> "DSL", "DSL", "DSL", "DSL", "Fiber optic", "Fiber opt~
$ OnlineSecurity   <chr> "No", "Yes", "Yes", "Yes", "No", "No", "No", "Yes", "~
$ OnlineBackup     <chr> "Yes", "No", "Yes", "No", "No", "No", "Yes", "No", "N~
```

```
$ DeviceProtection <chr> "No", "Yes", "No", "Yes", "No", "Yes", "No", "No", "Y~
$ TechSupport      <chr> "No", "No", "No", "Yes", "No", "No", "No", "No", "Yes~
$ StreamingTV      <chr> "No", "No", "No", "No", "No", "Yes", "Yes", "No", "Ye~
$ StreamingMovies  <chr> "No", "No", "No", "No", "No", "Yes", "No", "No", "Yes~
$ Contract         <chr> "Month-to-month", "One year", "Month-to-month", "One ~
$ PaperlessBilling <chr> "Yes", "No", "Yes", "No", "Yes", "Yes", "Yes", "No", ~
$ PaymentMethod    <chr> "Electronic check", "Mailed check", "Mailed check", "~
$ MonthlyCharges   <dbl> 29.85, 56.95, 53.85, 42.30, 70.70, 99.65, 89.10, 29.7~
$ TotalCharges     <dbl> 29.85, 1889.50, 108.15, 1840.75, 151.65, 820.50, 1949~
$ Churn            <chr> "No", "No", "Yes", "No", "Yes", "Yes", "No", "No", "Y~
```

Then clean up the data.

- Select relevant variables.
- Create an integer variable indicating churn (for ease of analysis).
- Filter for (and drop) NA values.

Findings: There are 11 NAs for TotalCharges.

```
telcom_churn <- telcom_churn %>%
  select(Churn, tenure, MonthlyCharges, TotalCharges) %>%
  mutate(
    Churn = factor(Churn, levels = c("No","Yes")),
    churn_int = case_when(
      Churn %in% c("Yes","yes", 1, "1", TRUE) ~ 1L,
      Churn %in% c("No", "no",  0, "0", FALSE) ~ 0L,
      TRUE ~ NA_integer_
    ),
  )

telcom_churn_clean <- telcom_churn %>%
  filter(!is.na(churn_int) &
         !is.na(tenure) &
         !is.na(MonthlyCharges) &
         !is.na(TotalCharges))
```

Now take a look at the distribution of the cleaned data:

```
summary(telcom_churn_clean)
```

```
 Churn          tenure      MonthlyCharges    TotalCharges      churn_int
 No :5163   Min.   : 1.00   Min.   : 18.25   Min.   :  18.8   Min.   :0.0000
 Yes:1869   1st Qu.: 9.00   1st Qu.: 35.59   1st Qu.: 401.4   1st Qu.:0.0000
            Median :29.00   Median : 70.35   Median :1397.5   Median :0.0000
            Mean   :32.42   Mean   : 64.80   Mean   :2283.3   Mean   :0.2658
            3rd Qu.:55.00   3rd Qu.: 89.86   3rd Qu.:3794.7   3rd Qu.:1.0000
            Max.   :72.00   Max.   :118.75   Max.   :8684.8   Max.   :1.0000
```

This looks good, and I proceed with this version of cleaned data.

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

2 Answer

$$\begin{aligned}
L(\alpha, \beta \mid \text{Data}) &= \mathbb{P}(x_1, x_2, \dots, x_n \mid \alpha, \beta) && (1)\\
&= \prod_{i=1}^{n} \mathbb{P}(x_i \mid \alpha, \beta) && (2)\\
&= \prod_{i=1}^{n} \left(\frac{e^{\alpha + \beta t_i}}{1 + e^{\alpha + \beta t_i}} \right)^{y_i} \left(\frac{1}{1 + e^{\alpha + \beta t_i}} \right)^{1 - y_i} && (3)\\
&= \prod_{i=1}^{n} \frac{\exp(y_i(\alpha + \beta t_i))}{1 + e^{\alpha + \beta t_i}}, && (4)
\end{aligned}$$

where t_i is the value of **tenure** of the ith observation, and $y_i = 1$ indicates the ith observation churned and $y_i = 0$ indicates otherwise.

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.

3. Answer

In implementing the function in R, I have used a numerical trick to shift the argument to the exponential function by the max v_i (defined below) in each choice situation, which doesn't change choice probabilities; in this case the max is either v_i or 0.

$$v_i = \alpha + \beta t_i, \qquad p_i \equiv \mathbb{P}(y_i = 1 \mid t_i) = \frac{e^{v_i}}{1 + e^{v_i}} \tag{5}$$

$$\ell(\alpha, \beta) = \sum_{i=1}^{n} \Big[y_i \log p_i + (1 - y_i) \log(1 - p_i) \Big] \tag{6}$$

$$= \sum_{i=1}^{n} \Big[y_i v_i - \log(1 + e^{v_i}) \Big] \tag{7}$$

$$\mathrm{NLL}(\alpha, \beta) = -\ell(\alpha, \beta) = \sum_{i=1}^{n} \Big[\log(1 + e^{v_i}) - y_i v_i \Big]. \tag{8}$$

```r
# Negative log-likelihood for logistic model: churn ~ tenure
nll_logistic_tenure <- function(par, df) {
  alpha <- par[1]
  beta  <- par[2]
  v     <- alpha + beta * df$tenure

  # Shift the argment to the exponential function to prevent overflow
  log1pexp <- log1p(exp(-abs(v))) + pmax(v, 0)

  # NLL = sum[ log(1+exp(v)) - y*v ]
  sum(log1pexp - df$churn_int * v)
}
```

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.

4. Answer

```
# run optimization to minimize the negative log-likelihood
# and assign the output to mle_fit
mle_fit <- optim(
  par = c(0, 0),
  fn  = nll_logistic_tenure,
  df  = telcom_churn_clean,
  # This is needed for the MLE sample variance
  hessian = TRUE
)

# Print estimates
mle_fit$par
```

```
[1]  0.03758525 -0.03901693
```

```
# Print negative log-likelihood value
mle_fit$value
```

```
[1] 3588.13
```

Comment

Above are the solutions to the optimization and I have also printed the minimized negative log-likelihood function value in case I need it for later.

The solutions will be used to compare in Part 6 to the estimates produced by the logistic regression.

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?

5. Answer

```
# assign estimates
alpha_hat <- mle_fit$par[1]
beta_hat  <- mle_fit$par[2]

# Invert Hessian to get covariance matrix/Fisher info
cov_matrix <- solve(mle_fit$hessian)

# Get the sampling SE
se <- sqrt(diag(cov_matrix))

# Now 95% Wald CIs for alpha and beta
q = qnorm(0.975)
q
```

```
[1] 1.959964
```

```
ci_alpha <- c(alpha_hat - q*se[1], alpha_hat + q*se[1])
ci_beta  <- c(beta_hat  - q*se[2], beta_hat  + q*se[2])

# Print the estimates and their CIs
list(
  alpha_hat = alpha_hat, ci_alpha = ci_alpha,
  beta_hat  = beta_hat,  ci_beta  = ci_beta
)
```

```
$alpha_hat
[1] 0.03758525

$ci_alpha
[1] -0.04535178  0.12052228

$beta_hat
```

```
[1] -0.03901693

$ci_beta
[1] -0.04177741 -0.03625645
```

Comment

I printed out the solutions and the Wald confidence intervals constructed similar to that shown in the async. Note that I use the asymptotic properties of the MLE estimate and plugged in the $(1-\alpha/2)$th quantile of a standard Gaussian.

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)

6. Answer

Run the logistic regression:

```
logit_mod <- glm(churn_int ~ tenure,
                 data = telcom_churn_clean,
                 family = binomial)

summary(logit_mod)$coefficients
```

```
               Estimate  Std. Error     z value      Pr(>|z|)
(Intercept)  0.03729859 0.042319295   0.8813613  3.781223e-01
tenure      -0.03900965 0.001408723 -27.6914926 8.839967e-169
```

Comment:

What do I notice?

- The logistic regression coefficients are found by maximum likelihood, so the glm() method should give the same (or very close) results to what we did in Part 5.
- β is very close: -0.0390169 in the optim() approach and -0.0390096 in the glm() approach. Moreover, the p-value is very small so that the coefficient itself is statistically significant (single variable hypothesis testing).
- α is statistically **insignificant** and also close in the two approaches. The optim() gives 0.0375852 and the glm() gives 0.0372986. Notice that the CI for both contains 0.
- Connection: in linear regression, the MLEs are the same (closed-form) as the method of moments estimates. In the logistic regression, there is no closed-form solution for the coefficients and they are found by numeric methods.

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.

7. Answer

```r
logit_mod_ext <- glm(churn_int ~ tenure + MonthlyCharges + TotalCharges,
                     data = telcom_churn_clean,
                     family = binomial(link = "logit"))

summary(logit_mod_ext)
```

```
Call:
glm(formula = churn_int ~ tenure + MonthlyCharges + TotalCharges,
    family = binomial(link = "logit"), data = telcom_churn_clean)

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
```

Interpretation

Recall that $\ln \frac{\pi}{1-\pi} = \mathbf{x} \cdot \mathbf{beta}$, and $odds = \frac{\pi}{1-\pi} = exp(\mathbf{x} \cdot \mathbf{beta})$ so a c units change in the independent variable x, **keeping all other independent variables fixed**, will lead to the

odds of churn to change by $e^{c\beta}$ times, where β is the coefficient of x .

Therefore we could interpret:

- First notice that TotalCharges exhibits certain degree of colinearity with MonthlyCharges and tenure, which may explain it's relatively large p-value.
- A unit increase in tenure (assuming measured in months) changes the odds of churn by $e^{-0.067}$ times. Makes intuitive sense.
- A unit (measured in dollar) increase in either MonthlyCharges or TotalCharges increase the odds of churn by $e^{0.03}$ and $e^{0.00015}$ times, respectively. This makes sense as demand elasticity with respect to price is negative, and it is more elastic for monthly charges than total charges which also reflects tenure.

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.

8. Answer

The full model is already fitted in Part 7. To do the likelihood ratio (LR) test for **each** independent variable, we can fit the restricted models and use anova(), or we can use Anova() from the car package as shown in the async.

Carrying out the LR test using Anova()

```
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


Analysis of Deviance Table (Type II tests)

Response: churn_int
               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
```

As shown in the test results, for each individual variable (keeping other vars constant):

- Both tenure and MonthlyCharges have very large test statistic values so that we can easily reject the null that the variable coefficient is zero.

- The LR test for TotalCharges has a p-value of 0.017. So we **can** reject the null hypothesis (that the $\beta_{TotalCharges} = 0$) at $\alpha = 0.05$ but **cannot** reject at $\alpha = 0.01$ or smaller.
- Again, notice that TotalCharges exhibits certain degree of colinearity with MonthlyCharges and tenure, which may explain it's relatively large p-value.

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.

9. Answer

Get the estimates and the (sampling) standard errors

```r
coefs <- coef(logit_mod_ext)
coef_monthly <- coefs["MonthlyCharges"]

sd_monthly <- sd(telcom_churn_clean$MonthlyCharges)
```

So, the MLE for the coefficient for MonthlyCharges is 0.0301997 and the standard deviation in the (cleaned) sample is 30.0859739.

Effect of change in MonthlyCharges on odds of churn.

The odds of churn will increase by $e^{\sigma_{MonthlyCharges} \cdot \beta_{MonthlyCharges}}$ (= 2.4808166) times when MonthlyCharges increases by one standard deviation ($30).

Wald CI

First recall the Wald CI for $c\beta$ is

$$c\hat{\beta} \pm c \cdot Z_{1-\frac{\alpha}{2}} \cdot \sqrt{Var(\hat{\beta}_{MonthlyCharges})}$$

```r
vc <- vcov(logit_mod_ext)
se_monthly <- sqrt(vc["MonthlyCharges", "MonthlyCharges"])

ci_lower  <- exp((coef_monthly - q*se_monthly) * sd_monthly)
ci_upper <- exp((coef_monthly + q*se_monthly) * sd_monthly)

OR_monthly <- exp(coef_monthly * sd_monthly)

OR_ci <- c(OR = OR_monthly, lower = ci_lower, upper = ci_upper)
```

- c is 30.0859739.

- $\hat{\beta}_{MonthlyCharges}$ is 0.0301997.
- $\sqrt{Var(\hat{\beta}_{MonthlyCharges})}$ is 0.0017174.
- Then I exponentiate the CI for $c\hat{\beta}$ to get the Wald CI for odds ratio.
- The Wald CI for the odds ratio is: (2.2418866, 2.7452107) and recall the estimate is 2.4808166.

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`.

10. Answer

Sanity check on the average values of the independent vars:

- Sample mean of tenure: 32.4217861.
- Sample mean of MonthlyCharges: 64.7982082.
- Sample mean of TotalCharges: 2283.3004408.
- Makes sense.

Calling the mcprofile package

```r
library(mcprofile)

avg_tenure = mean(telcom_churn_clean$tenure)
avg_MonthlyCharges = mean(telcom_churn_clean$MonthlyCharges)
avg_TotalCharges = mean(telcom_churn_clean$TotalCharges, na.rm=TRUE)

K <- matrix(data = c(1,
                     avg_tenure,
                     avg_MonthlyCharges,
                     avg_TotalCharges),
            nrow = 1)

colnames(K) <- names(coef(logit_mod_ext))

linear_combo <- mcprofile(object = logit_mod_ext,
                          CM = K)

ci_logit_profile <- confint(object = linear_combo, level = 0.95)
ci_logit_profile
```

```
   mcprofile - Confidence Intervals

level:        0.95
```

```
adjustment:  single-step

   Estimate lower upper
C1    -1.49 -1.57  -1.4
```

```r
names(ci_logit_profile)
```

```
[1] "estimate"    "confint"     "CM"          "quant"       "alternative"
[6] "level"       "adjust"
```

```r
churn_prob_ci <- exp(ci_logit_profile$confint) / (1 + exp(ci_logit_profile$confint))
```

Comment

So the confidence interval (using profile likelihood ratio) for the probability of churn, given average sample values of the independent variables, is between 0.171823387843509 and 0.197231887362678, providing a somewhat tight CI for the estimated probability.