MP4

STAT-340 Applied Regression Methods

Yerim Oh

Load Packages

The following R code loads packages needed in this assignment.

Problem 1: Adapted from ISLR Example 4.6.

Suppose we collect data for a group of students in a statistics class with variables

- X_1 = hours studied
- X_2 = undergrad GPA
- Y = receive an A in this class ("Yes" or "No")

We fit a logistic regression model and produce estimated coefficients, $\hat{\beta}_0 = -6$, $\hat{\beta}_1 = 0.05$, and $\hat{\beta}_2 = 1$.

(a) What is the interpretation of the coefficient estimate $\hat{\beta}_1 = 0.05$, in terms of the odds of getting an A?

For an increase of 1 hour of studying, we expect the odds of the student to receive an A in this class to increase by a factor of $\exp(0.05) = 1.05127$.

(b) Estimate the *probability* that a student who studies for 40 hours and has an undergrad GPA of 3.5 gets an A in the class.

$$\hat{Pr}(Y = 1|X_1 = 40, X_2 = 3.5) = \frac{e^{-6+0.05\cdot 40+1\cdot 3.5}}{1 + e^{-6+0.05\cdot 40+1\cdot 3.5}} = 0.3775$$

(c) Estimate the *probability* that a student who studies for 41 hours and has an undergrad GPA of 3.5 gets an A in the class.

$$\hat{Pr}(Y = 1|X_1 = 41, X_2 = 3.5) = \frac{e^{-6+0.05\cdot 41+1\cdot 3.5}}{1 + e^{-6+0.05\cdot 41+1\cdot 3.5}} = 0.3894$$

(d) By using your answer to part (b) and the definition of odds, estimate the *odds* that a student who studies for the class for 40 hours and has an undergrad GPA of 3.5 gets an A in the class. Do this again for the *odds* that a student who studies for the class for 41 hours and has an undergrad GPA of 3.5 gets an A in the class, using your answer to part (c). Verify that the interpretation you gave in part (a) holds in this example.

- estimated *odds* that a student who studies for the class for 40 hours and has an undergrad GPA of 3.5 gets an A in the class:

$$\frac{p(Y = 1|X_1 = 40, X_2 = 3.5)}{1 - p(Y = 1|X_1 = 40, X_2 = 3.5)} = \frac{0.3775407}{1 - 0.3775407} = 0.6065307$$
$$e^{-6+0.05\cdot 40+1\cdot 3.5} = 0.6065307$$

- estimated *odds* that a student who studies for the class for 41 hours and has an undergrad GPA of 3.5 gets an A in the class:

$$\frac{p(Y=1|X_1=41,X_2=3.5)}{1-p(Y=1|X_1=41,X_2=3.5)} = \frac{0.3893608}{1-0.3893608} = 0.6376282$$
$$e^{-6+0.05\cdot 41+1\cdot 3.5} = 0.6376282$$

- $\frac{0.6376282}{0.6065307} = 1.051271 = \exp(0.05)$

Therefore, the interpretation in part (a) holds, because in a multiple logistic regression, the ratio represents the change in odds of the response variable for a one-unit increase in a explanatory variable, while holding all other variables in the model constant.

(e) Suppose a student has an undergrad GPA of 3.5. How many hours would they need to study for us to estimate that there is a probability of 0.5 that they will get an A in the class?

$$\begin{aligned}
\hat{Pr}(Y=1|X_1=x,X_2=3.5) &= \frac{e^{-6+0.05\cdot x+1\cdot 3.5}}{1+e^{-6+0.05\cdot x+1\cdot 3.5}} = 0.5 \\
&= \frac{e^{-2.5+0.05\cdot x}}{1+e^{-2.5+0.05\cdot}} \\
= \frac{e^{-2.5}\cdot e^{0.05\cdot x}}{1+e^{-2.5}\cdot e^{0.05\cdot x}} &= \frac{1}{2} \\
e^{-2.5}\cdot e^{0.05\cdot x} &= \frac{1}{2}(1+e^{-2.5}\cdot e^{0.05\cdot x}) \\
2\cdot(e^{-2.5}\cdot e^{0.05\cdot x}) &= 1+e^{-2.5}\cdot e^{0.05\cdot x} \\
2\cdot(e^{-2.5}\cdot e^{0.05\cdot x})-(e^{-2.5}\cdot e^{0.05\cdot x}) &= 1 \\
e^{-2.5}\cdot e^{0.05\cdot x} &= 1 \\
e^{0.05\cdot x} &= \frac{1}{e^{-2.5}} = e^{2.5} \\
0.05\cdot x &= 2.5 \\
x &= 50
\end{aligned}$$

A student with an undergrad GPA of 3.5 needs to study 50 hours for a probability of 0.5 that they will get an A in the class.

Problem 2: Metabolic syndrome

Metabolic syndrome is a cluster of conditions that occur together, increasing your risk of heart disease, stroke and type 2 diabetes. These conditions include increased blood pressure, high blood sugar, excess body fat around the waist, and abnormal cholesterol or triglyceride levels. For the purpose of this assignment, we want to explore the association of hypertension status, weight and total cholesterol with the result of screening fasting blood sugar. We have the following variables:

- `cat_fbs` : fasting blood sugar coded as `Normal`, `DM` (diabetes mellitus), and `IFG` (impaired fasting glucose).
- `totchol`: Total Cholesterol (mmol/L).
- `hptc`: Hypertension Status (`Yes` or `No`).
- `weight`: Body weight measures in kilogram.

(a) Fit a model to explain the fasting blood sugar (`cat_fbs`) with the other variables. Print out the model summary.

```
## Call:
```

```
## nnet::multinom(formula = .outcome ~ ., data = dat, decay = param$decay,
##     trace = FALSE)
##
## Coefficients:
##     (Intercept)   totchol    hptyes     weight
## IFG   -5.112735 0.2393748 0.8672457 0.02199345
## DM    -4.907901 0.2772382 0.9000338 0.02267585
##
## Std. Errors:
##     (Intercept)    totchol    hptyes      weight
## IFG   0.3663222 0.04239126 0.1455646 0.003710277
## DM    0.3037717 0.03505572 0.1204522 0.003073929
##
## Residual Deviance: 5476.397
## AIC: 5492.397
```

(b) Interprete the coefficients $\hat{\beta}_{IFG,totchol}$ and $\hat{\beta}_{DM,totchol}$ in context.

$\hat{\beta}_{IFG,totchol} = 0.2394$: For an increase of 1 mmol/L total cholesterol, we expect the odds of impaired fasting glucose versus normal fasting blood sugar to increase by a factor of $\exp(0.2394) = 1.2705$, while the other variables are held constant.

$\hat{\beta}_{DM,totchol} = 0.2772$: For an increase of 1 mmol/L total cholesterol, we expect the odds of diabetes mellitus versus normal fasting blood sugar to increase by a factor of $\exp(0.2772) = 1.3194$, while the other variables are held constant.

(c) Calculate the probability of having impaired fasting glucose (`IFG`) when totchol = 6, hptyes = yes (hypertension), and weight = 75kg.

$$
\begin{aligned}
&Pr(\text{IFG}|\text{totchol} = 6, \text{hptyes} = 1, \text{weight} = 75) \\
&= \frac{e^{-5.1127+0.2394\cdot 6+0.8672\cdot 1+0.02199\cdot 75}}{1 + e^{-5.1127+0.2394\cdot 6+0.8672\cdot 1+0.02199\cdot 75} + e^{-4.9079+0.2772\cdot 6+0.9000\cdot 1+0.02268\cdot 75}} \\
&= \frac{0.3135332}{1 + 0.3135332 + 0.525345} = 0.1705024
\end{aligned}
$$

(d) Calculate the probability of a normal fasting blood sugar (`Normal`) when totchol = 6, hptyes = yes (hypertension), and weight = 75kg.

$$
\begin{aligned}
&Pr(\text{Normal}|\text{totchol} = 6, \text{hptyes} = 1, \text{weight} = 75) \\
&= \frac{1}{1 + e^{-5.1127+0.2394\cdot 6+0.8672\cdot 1+0.02199\cdot 75} + e^{-4.9079+0.2772\cdot 6+0.9000\cdot 1+0.02268\cdot 75}} \\
&= \frac{1}{1 + 0.3135332 + 0.525345} = 0.5438098
\end{aligned}
$$

(e) <u>Extra credit</u>: By definition, under multinomial regression, we have that

$$\text{odds}(Y = k \mid X) = \frac{p(Y = k \mid X)}{p(Y = K \mid X)}.$$

Show that:

$$e^{\beta_{kr}} = \frac{\text{odds}\left(Y = k \mid X_r + 1, \boldsymbol{X}_{-r}\right)}{\text{odds}\left(Y = k \mid X_r, \boldsymbol{X}_{-r}\right)},$$

where $\text{odds}(Y = k \mid X) = \frac{p(Y=k|X)}{p(Y=K|X)}$ **and** $\boldsymbol{X}_{-r}$ **denotes all predictors except** X_r**.**

$$\begin{aligned}
\text{odds}(Y = k \mid X) &= \frac{p(Y = k \mid X)}{p(Y = K \mid X)} \\
&= \frac{\frac{e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kp}x_p}}{1+\sum_{j=1}^{K-1} e^{\beta_{j0}+\beta_{j1}x_1+\cdots+\beta_{jp}x_p}}}{\frac{1}{1+\sum_{j=1}^{K-1} e^{\beta_{j0}+\beta_{j1}x_1+\cdots+\beta_{jp}x_p}}} = \frac{e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kp}x_p}}{1} \\
&= e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kp}x_p}
\end{aligned}$$

$$\text{odds}(Y = k \mid X_r, X_{-r}) = e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kr}x_r+\cdots+\beta_{kp}x_p}$$

$$\begin{aligned}
\text{odds}(Y = k \mid X_r + 1, X_{-r}) &= e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kr}(x_r+1)+\cdots+\beta_{kp}x_p} \\
&= e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kr}x_r+\beta_{kr}+\cdots+\beta_{kp}x_p} \\
&= \left(e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kr}x_r+\cdots+\beta_{kp}x_p}\right) \cdot e^{\beta_{kr}}
\end{aligned}$$

$$\begin{aligned}
\frac{\text{odds}(Y = k \mid X_r + 1, X_{-r})}{\text{odds}(Y = k \mid X_r, X_{-r})} &= \frac{\left(e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kr}x_r+\cdots+\beta_{kp}x_p}\right) \cdot e^{\beta_{kr}}}{e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kr}x_r+\cdots+\beta_{kp}x_p}} \\
&= e^{\beta_{kr}}
\end{aligned}$$