Exercises: Generalized Linear Model

Valentina Zangirolami - valentina.zangirolami@unimib.it

December 21, 2023

(Referring to the theoretical parts: 20, 21, 22, 23, 24, 25, 26)

1 Exercise 1

To meet competition or cope with economic slowdowns, corporations sometimes undertake a "reduction in force" (RIF), in which substantial numbers of employees are terminated. Federal and various state laws require that employees be treated equally regardless of their age. In particular, employees over the age of 40 years are in a "protected" class, and many allegations of discrimination focus on comparing employees over 40 with their younger coworkers. Here are the data for a recent RIF:

Terminated	Over40: No	Over40: Yes
Yes	17	71
No	564	835

Exercise 1.1

Choose an appropriate response variable and then a regression model. Justify your answer. ~~Compute~~ Further, find the estimated regression model.

Our interest lies in understanding the relationship between the variable "Terminated" and "Over40". Therefore, an appropriate response variable can be "Terminated".

In this case, Terminated is a binary variable and we can use logistic regression. Then, we want to consider logit regression such that

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_1 + \beta_2 D_i \qquad \text{where } D_i = \begin{cases} 1, & \text{if Over40 = Yes} \\ 0, & \text{otherwise} \end{cases}$$

Assumption: $Y_i \sim Bernoulli(\pi_i)$ [Y_i = Terminated$_i$]

We can rewrite the model as

$$\pi_i = \frac{e^{\beta_1 + \beta_2 D_i}}{1 + e^{\beta_1 + \beta_2 D_i}}$$

In this case, our covariate is even a dummy. Then, we know

- $(\pi_i | D_i = 1) = \mathbb{P}(y_i = 1 | D_i = 1) = \frac{e^{\beta_1 + \beta_2}}{1 + e^{\beta_1 + \beta_2}}$ and

$$(1 - \pi_i | D_i = 1) = \mathbb{P}(y_i = 0 | D_i = 1) = \frac{1}{1 + e^{\beta_1 + \beta_2}}$$

$$\Rightarrow \frac{\pi_i}{1 - \pi_i} | D_i = 1 = e^{\beta_1 + \beta_2} \quad \text{(ODDS)}$$

- $(\pi_i | D_i = 0) = \mathbb{P}(y_i = 1 | D_i = 0) = \frac{e^{\beta_1}}{1 + e^{\beta_1}}$ and $(1 - \pi_i | D_i = 0) = \mathbb{P}(y_i = 0 | D_i = 0) = \frac{1}{1 + e^{\beta_1}}$

$$\Rightarrow \frac{\pi_i}{1 - \pi_i} | D_i = 0 = e^{\beta_1} \quad \text{(ODDS)}$$

In this case we can find in easy way our estimates:

- $e^{\beta_1} = \frac{17}{564} = 0.03014184 \Rightarrow \hat{\beta}_1 = -3.501841$

- $e^{\beta_1 + \beta_2} = \frac{71}{835} = 0.08502994 \Rightarrow \hat{\beta}_2 = \log(0.08502994) + 3.501841 = 1.037089$

Then, the estimated logit model is

$$\log(\text{odds}_i) = -3.501841 + 1.037089 D_i$$

Exercise 1.3

Software gives the estimated slope $\hat{\beta}_2 = 1.0371$ and its standard error $SE(\hat{\beta}_2) = 0.2755$. Transform the results to the odds scale. Summarize the results and write a short conclusion.

With confidence intervals

From the calculation in ex. 1.1, we can consider the odds ratio which corresponds to

$$\frac{\left(\frac{\pi_i}{1-\pi_i} \mid D_i = 1\right)}{\left(\frac{\pi_i}{1-\pi_i} \mid D_i = 0\right)} = e^{\beta_2}$$

and in this case

$$e^{\hat{\beta}_2} = 2.82$$

→ The odds of under 40 are multiplied by ~~coefficient~~ 2.82 to have the odds for over 40.

→ Employees over 40 are 2.82 times more likely to be terminated than those under 40.

We can provide also confidence intervals to summarise our result.

The confidence interval for the odds ratio can be obtained by

$$\left(e^{\beta_2 - z_{1-\frac{\alpha}{2}} SE(\beta_2)},\ e^{\beta_2 + z_{1-\frac{\alpha}{2}} SE(\beta_2)}\right)$$

Then, in this case, we have

$z_{1-\frac{\alpha}{2}} = 1.96 \qquad 1-\alpha = 0.95$

- $e^{\hat{\beta}_2 - 1.96\, SE(\hat{\beta}_2)} = e^{1.0371 - 1.96 \cdot 0.2755} = 1.644$

- $e^{\hat{\beta}_2 + 1.96\, SE(\hat{\beta}_2)} = e^{1.0371 + 1.96 \cdot 0.2755} = 4.840$

- $IC(0.95) = (1.644, 4.840)$

Because the interval does not contain 1, the results are significant at the 5% significance level.

Exercise 1.4

If additional explanatory variables were available, for example, a performance evaluation, how would you use this information to study the RIF?

We could use the additional variables in the logistic regression model to account for their effects before assessing if age has an effect.

Or we can just add the additional explanatory variables into the previous model.

2 Exercise 2

The acquisition literature suggests that takeovers occur either due to conflicts between managers and shareholders or to create a new entity that exceeds the sum of its previously separate components. Other research has offered managerial hubris as a third option, but it has not been studied empirically. Recently, some researchers revisited acquisitions over a 10-year period in the Australian financial system. A measure of CEO overconfidence was based on the CEO's level of media exposure, and a measure of dominance was based on the CEO's remuneration relative to the firm's total assets. They then used logistic regression to see whether CEO overconfidence and dominance were positively related to the probability of at least one acquisition in a year. To help isolate the effects of CEO hubris, the model included explanatory variables of firm characteristics and other potentially important factors in the decision to acquire. The following table summarizes the results for the two key explanatory variables:

Covariates	Estimates	SE
Overconfidence	0.0878	0.0402
Dominance	1.5067	0.0057

Exercise 2.1

Write the estimated regression model and interpret the coefficients estimates.

In this case, our ~~variable~~ response variable corresponds to

$$Y_i = \begin{cases} 1, & \text{if the firm made at least one acquisition} \\ 0, & \text{otherwise} \end{cases}$$

Therefore, we can use logistic regression using the logit link function (which corresponds to the canonical one).

The estimated regression model is

$$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = \hat{\beta}_1 + 0.0878 \text{ Overconfidence}_i + 1.5067 \text{ Dominance}_i$$

Interpretation of regression coefficients:

- the log odds increases by 0.0878 if overconfidence increases of 1 units, while keeping constant the other covariates.
- the log odds increases by 1.5067 if Dominance increases of 1 units, while keeping constant the other covariates fixed.

Exercise 2.2

Perform the significance tests and determine whether the variables are significant at the 0.05 level.

Significance test for $\hat{\beta}_2$

Hypothesis

$$\alpha \begin{cases} H_0: \beta_2 = 0 \\ H_1: \beta_2 \neq 0 \end{cases}$$

Test statistic

$$z_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{j(\hat{\beta})^{-1}_{jj}}} \overset{H_0}{\sim} N(0,1)$$

where $j(\hat{\beta}) = X^T U(\hat{\beta}) X$

$U(\hat{\beta}) = \text{diag}\{\hat{\pi}_1(1-\hat{\pi}_1), \ldots, \hat{\pi}_m(1-\hat{\pi}_m)\}$

Therefore, the observed test statistic

$$z_2^{obs} = \frac{0.0878}{0.0402} = 2.18408$$

p-value

$$\alpha^{obs} = P_{H_0}(|z_j| \geq |z^{obs}|) = 2(1 - \Phi(|z_j^{obs}|))$$

knowing the quantile : $z_{0.9855} = 2.18408$

$$\alpha^{obs} = 2 \cdot 0.01448 = 0.02896$$

$\Rightarrow$ We reject H_0 at 5% significance level, thus the coefficient β_2 is significant.

Significance test for $\hat{\beta}_3$

~~system~~ Hypothesis

$$\begin{cases} H_0: \beta_3 = 0 \\ H_1: \beta_3 \neq 0 \end{cases}$$

Test statistic

$$z_3^{obs} = \frac{1.5067}{0.0057} = 264.333$$

P-VALUE

$\alpha^{obs} = 2 \cdot (1 - \Phi(|z_3^{obs}|)) \approx 0$ we reject H_0

Exercise 2.3

Estimate the odds ratio for each variable and construct a 95% confidence interval.

(For simplicity, we call the variable ~~overconfidence~~ confidence as X_2 while the latter as X_3)

ODDS RATIO regarding Overconfidence

Let consider

$$\log \frac{\pi_0}{1-\pi_0} = \beta_1 + \beta_2 x_2 + \beta_3 x_3$$

and

$$\log \frac{\pi_1}{1-\pi_1} = \beta_1 + \beta_2(x_2+1) + \beta_3 x_3 = \beta_1 + \beta_2 + \beta_2 x_2 + \beta_3 x_3$$

And the odds ratio corresponds to

$$\frac{\log\left(\frac{\pi_1}{1-\pi_1}\right)}{\log\left(\frac{\pi_0}{1-\pi_0}\right)} = \beta_2 \Rightarrow \boxed{\frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_0}{1-\pi_0}} = e^{\beta_2}}$$

Hence the ~~est~~ value of the odds ratio is $1.09177 = \widehat{odds_2}$

→ Interpretation: An increase of one unit in the overconfidence measure is associated 1.1-fold increase in the odds

ODDS RATIO regarding Dominance

Now, we need to consider

$$\log \frac{\pi_0}{1-\pi_0} = \beta_1 + \beta_2 x_2 + \beta_3 x_3$$

and

$$\log \frac{\pi_1}{1-\pi_1} = \beta_1 + \beta_2 x_2 + \beta_3(x_3+1) = \beta_1 + \beta_2 x_2 + \beta_3 + \beta_3 x_3$$

Therefore, the odds ratio is

$$\frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_0}{1-\pi_0}} = e^{\beta_3} \Rightarrow \widehat{odds_3} = e^{\widehat{\beta_3}} = 4.511817$$

→ Interpretation: An increase of one unit in the dominance measure is associated 4.5-fold increase in the odds.

Confidence intervals

[ODDS2]

The confidence interval for the odds ratio (related to overconfidence) can be obtained as

$$\left(e^{\hat{\beta}_2 - z_{1-\frac{\alpha}{2}} SE(\hat{\beta}_2)}, e^{\hat{\beta}_2 + z_{1-\frac{\alpha}{2}} SE(\hat{\beta}_2)}\right)$$

Given $z_{0.95} = 1.96$, the estimated C.I. is

(1.009049, 1.181272)

[ODDS3]

The confidence interval for the odds ratio (related to Dominance) can be obtained by

$$\left(e^{\hat{\beta}_3 - z_{1-\alpha/2} SE(\hat{\beta}_3)}, e^{\hat{\beta}_3 + z_{1-\frac{\alpha}{2}} SE(\hat{\beta}_3)}\right)$$

and hence corresponds to (4.461692, 4.562506)

3 Exercise 3

Let consider a dataset on the number of research articles published by 915 graduate students in biochemistry PhD programs. The variables for this dataframe are

- **art**: count of articles produced during last 3 years of PhD
- **fem**: factor indicating gender of student, with levels Men and Women
- **mar**: factor indicating marital status of student, with levels Single and Married
- **kid5**: number of children aged 5 or younger
- **phd**: prestige of PhD department
- **ment**: count of articles produced by PhD mentor during last 3 years

Exercise 3.1
Choose an appropriate response variable and then a regression model. Justify your answer.

In this case, the most appropriate response variable is **art**.
Given the nature of our variable, we assume $\underset{(art)_i}{Y_i} \sim \text{Poisson}(\mu_i)$.

- $\mu_i = \exp\{x_i^T \underline{\beta}\}$

In this case,

$$\log \mu_i = \beta_1 + \beta_2 D_{1i} + \beta_3 D_{2i} + \beta_4 \, kid5 + \beta_5 \, phd + \beta_6 \, ment$$

where

$$D_{1i} = \begin{cases} 1, & \text{if fem = Women} \\ 0, & \text{otherwise} \end{cases}$$

and

$$D_{2i} = \begin{cases} 1, & \text{if mar = Married} \\ 0, & \text{otherwise} \end{cases}$$

We assume $Y_i \sim \text{Poisson}(\mu_i)$, since $(art)_i$ is a count variable.

Exercise 3.2

Complete the following table.

Coefficients	Estimates	SE	Z-obs	P-value
β_1	0.304617	0.102981	2.958	0.0031
β_2	-0.224594	0.054613	-4.113	3.92e-05
β_3	0.155243	0.06136	2.53019	0.0114
β_4	-0.18487	0.040127	-4.607	4.0852e-06
β_5	0.012824	0.026397	0.4858	0.6271
β_6	0.025543	0.002006	12.733	< 2e-16

Then, interpret the coefficient β_6.

inference about β_2

system of hypothesis: $\begin{cases} H_0: \beta_2 = 0 \\ H_1: \beta_2 \neq 0 \end{cases}$

$$z_2^{obs} = \frac{-0.224594}{0.054613} = -4.11264$$

(Recall that $z_2 \overset{H_0}{\sim} N(0,1)$)

p-value:

$$\alpha^{obs} = \mathbb{P}_{H_0}(|z_2| \geq |z_2^{obs}|) = 2(1-\Phi(4.11264)) \approx 0$$ we reject Ho

inference about β_3

Since $\alpha^{obs} = 0.0114 = 2(1-\Phi(|z_3^{obs}|))$ [we reject Ho at 5% sign. level]

$$1-\Phi(|z_3^{obs}|) = \frac{0.0114}{2} = 0.0057$$

$$z_3^{obs} = 2.530192$$

Now, we can find $SE(\hat{\beta}_3) = \frac{\hat{\beta}_3}{z_3^{obs}} = \frac{0.155243}{2.530192} = 0.06135$

inference about β_4

$$\hat{\beta}_4 = z_4^{obs} \cdot SE(\hat{\beta}_4) = -0.1848651$$

$$\alpha^{obs} = 2(1-\Phi(\cancel{0.040127}\ 4.607)) \approx 0$$ We reject Ho

inference about β_5

$\alpha^{obs} = 0.6271 = 2 \cdot (1 - \Phi(|t_5^{obs}|))$ $\Rightarrow t_5^{obs} = 0.4858127$

$\hat{\beta}_5 = t_5^{obs} \cdot SE(\hat{\beta}_5) = 0.4858127 \cdot 0.026397 = 0.012824$

Exercise 3.3

Knowing that the null deviance is equal to 1817.4 while the residual deviance is equal to 1634.4. Perform a statistical test about the overall significance. Specify the hypothesis, the test statistic and the p-value.

System of hypothesis

$$\begin{cases} H_0: \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0 \\ H_1 = \bar{H_0} \end{cases}$$

test statistic

$$W = 2 \log \frac{\hat{L}(\text{model})}{\tilde{L}(\text{null})} = 2\{\hat{\ell}(\text{model}) - \tilde{\ell}(\text{null})\} \overset{H_0}{\sim} \chi^2_5$$

We know that

$$2\{\hat{\ell}(\text{model}) - \tilde{\ell}(\text{null})\} = 2\{\hat{\ell}(\text{model}) + \tilde{\ell}(\text{saturated}) - \tilde{\ell}(\text{saturated}) - \tilde{\ell}(\text{null})\}$$

$$= 2\{[\tilde{\ell}(\text{saturated}) - \tilde{\ell}(\text{null})] - [\tilde{\ell}(\text{saturated}) - \hat{\ell}(\text{model})]\} =$$

$$= D(\text{null}) - D(\text{model})$$

$$w^{obs} = 1817.4 - 1634.4 = 183$$

$$\alpha^{obs} = \mathbb{P}(W \geq w^{obs}) = 1 - \mathbb{P}(W < w^{obs}) \approx 0$$ We reject H_0

Exercise 3.4

Knowing the value of the following quantity

$$\sum_{i=1}^{n} y_i \log(\hat{\mu}_i) - \hat{\mu}_i = -642.0261$$

Find the log-likelihood of the saturated model. ~~Can we perform a test about the~~ Further, interpret the results in terms of goodness of fit? ~~Justify your answer and interpret the results.~~

Log-likelihood of the saturated model

We know that

$$D(\text{model}) = 2\{\tilde{\ell}(\text{saturated}) - \hat{\ell}(\text{model})\} = 1634.4$$

and

$$\hat{\ell}(\text{model}) = \hat{\ell}(\mu) = \sum_{i=1}^{n} y_i \log \hat{\mu}_i - \hat{\mu}_i = -642.0261$$

Hence,

$$\tilde{\ell}(\text{saturated}) = \frac{1634.4}{2} + -642.0261 = 175.1739$$

Goodness of fit

Two ways to interpret results:

- $\hat{\ell}(\text{model})$ should not be "too far" from $\tilde{\ell}(\text{saturated})$
- $D(\text{model})$ should be less than $n-p$

In this case, $n-p = 915-6 = 909$ and $D(\text{model}) = 1634.4$.

Hence, the model is not good enough.

4 Exercise 4

A researcher is interested in how variables, such as **GRE** (Graduate Record Exam scores), **GPA** (grade point average) and prestige of the undergraduate institution (**Rank**), effect admission into graduate school. The response variable (**Admit**), admit/don't admit, is a binary variable. **Rank** takes on the values 1 through 4. The total number of observations is 400.

Exercise 4.1

Write the equation of the regression model using probit. Which other kind of model can we use?

PROBIT LINK FUNCTION within logistic regression

Model assumptions

- $Y_i \sim \text{Bernoulli}(\pi_i)$
- $\eta_i = \beta_1 + \beta_2 GRE_i + \beta_3 GPA_i + \beta_4 Rank_i{:}2 + \beta_5 Rank_i{:}3 + \beta_6 Rank_i{:}4$
- $g(\pi_i) = \Phi^{-1}(\pi_i) = \eta_i$ where $\Rightarrow \pi_i = \Phi(\eta_i)$ where Φ is the CDF of a Gaussian distribution

Basically, we need to assume Y_i is obtained from Y_i^* as

$$Y_i = \begin{cases} 1 & \text{if } Y_i^* > K \\ 0 & \text{if } Y_i^* \leq K \end{cases}$$

where $Y_i^* \overset{\perp}{\sim} N(\eta_i, 1)$

Hence, in this case we have

$$Y_i^* = \beta_1 + \beta_2 GRE + \beta_3 GPA + \beta_4 D_{1i} + \beta_5 D_{2i} + \beta_6 D_{3i} + \varepsilon_i \qquad \varepsilon_i \overset{i.i.d.}{\sim} N(0,1)$$

where $D_{1i} = \begin{cases} 1, & \text{if } Rank_i = 2 \\ 0, & \text{otherwise} \end{cases}$ $\quad D_{3i} = \begin{cases} 1, & \text{if } Rank_i = 4 \\ 0, & \text{otherwise} \end{cases}$

$D_{2i} = \begin{cases} 1, & \text{if } Rank_i = 3 \\ 0, & \text{otherwise} \end{cases}$

Instead of the probit, we can use the logit link function.

Exercise 4.2

Knowing the estimates of regression coefficients using a probit model, such that

Coefficients	Estimates
β_1	-2.38684
β_2	0.00138
β_3	0.47773
β_4	-0.41540
β_5	-0.81214
β_6	-0.93590

Interpret them.

① $\hat{\beta}_2 = 0.00138$

An increase of ~~[illegible]~~ GRE score increases the predicted probability of admission.

② $\hat{\beta}_3 = 0.47773$

An increase of GPA score increases the predicted probability of admission.

③ $\hat{\beta}_4 = -0.41540$

~~[illegible]~~ If the prestige of the undergraduate institution corresponds to the value 2, the predicted probability decreases

④ $\hat{\beta}_5 = -0.81214$

If the prestige of the undergraduate institution corresponds to the value 3, the predicted probability decreases

⑤ $\hat{\beta}_6 = -0.93590$

If the prestige of the undergraduate institution corresponds to the value 4, the predicted probability decreases.

Exercise 4.3

Now, let consider another appropriate regression model within these data (should have already been specified in ex. 4.1). Write the regression model.

LOGIT LINK FUNCTION within logistic regression

Model assumptions

- $Y_i \sim \text{Bernoulli}(\pi_i)$
- $\eta_i = \beta_1 + \beta_2 GRE_i + \beta_3 GPA_i + \beta_4 D_{1i} + \beta_5 D_{2i} + \beta_6 D_{3i}$
- $g(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i$

Hence, we can write our regression model as

$$\pi_i = \frac{\exp\left(\beta_1 + \beta_2 GRE_i + \beta_3 GPA_i + \beta_4 D_{1i} + \beta_5 D_{2i} + \beta_6 D_{3i}\right)}{1 + \exp\left(\beta_1 + \beta_2 GRE_i + \beta_3 GPA_i + \beta_4 D_{1i} + \beta_5 D_{2i} + \beta_6 D_{3i}\right)}$$

Exercise 4.4

We are interested in performing a test for comparing models. Let consider the previous model as full model, and the restricted model as a model which just includes **Rank**. Knowing the value of the observed test statistic, $w^{obs} = 16.449$, specify the hypothesis, the theoretical test statistic and p-value. Discuss about the result.

FULL MODEL (M_6):

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_1 + \beta_2\, GRE_i + \beta_3\, GPA_i + \beta_4 D_{1i} + \beta_5 D_{2i} + \beta_6 D_{3i}$$

RESTRICTED MODEL (M_4):

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \gamma_1 + \gamma_2 D_{1i} + \gamma_3 D_{2i} + \gamma_4 D_{3i}$$

System of hypothesis

$$\begin{cases} H_0: \beta_2 = \beta_3 = 0 \\ H_1: \overline{H_0} \end{cases}$$

Test statistic

$$W = 2 \log \frac{\hat{L}(\text{model})}{\tilde{L}(\text{restricted})} \overset{H_0}{\sim} \chi^2_{p-p_0}$$

(p: number of coeff. of the model / p_0: number coeff of restricted)

In this case, $w^{obs} = 16.449$

Hence, the p-value is

$$\alpha^{obs} = P_{H_0}(W \geq w^{obs}) \overset{(*)}{\approx} 0.00026$$

we reject H_0

(*) knowing the quantile of χ^2_2, s.t. $w_{2;0.99974} = 16.509$

Exercise 4.5

Knowing the deviance of the previous restricted model $D(restricted) = 458.52$, and the deviance of the null model $D(null) = 499.98$, perform a test about overall significance by specifying the hypothesis, the test statistic and the p-value.

System of hypothesis

$$\begin{cases} H_0: \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0 \\ H_1: \overline{H_0} \end{cases}$$

TEST STATISTIC

$$W = 2 \log \frac{\hat{L}(model)}{\tilde{L}(null)} \overset{H_0}{\sim} \chi^2_{p-1}$$

We know

$$W = 2 \cancel{\log} [\hat{\ell}(model) - \tilde{\ell}(null)] = D(null) - D(model)$$

and we need to find D(model).

Since in the previous exercise, we used

$$W^{obs}_{4.4} = 16.449$$

and knowing $D(restricted) = 458.52$,

$$D(model) = D(restricted) - W^{obs}_{4.4} = 458.52 - 16.449 = 442.071$$

Hence, the W^{obs} for ~~the~~ test about overall significance corresponds to

$$W^{obs} = 499.98 - 442.071 = 57.909$$

P-value:

$$\alpha^{obs} = \mathbb{P}_{H_0}(W \geq W^{obs}) \approx 0$$ We reject H_0

Exercise 4.6

Referring to the full model, we know

- $\hat{\beta}_3 = 0.804038$
- $z^{obs} = 2.423$

Interpret the value of $\hat{\beta}_3$. Perform the test of significance and discuss about the result.

INTERPRETATION of $\hat{\beta}_3$

The log odds increases by 0.804038 if the grade point average increases of 1 units while keeping the other variables fixed.

TEST OF SIGNIFICANCE

System of hypothesis

$$\begin{cases} H_0: \beta_3 = 0 \\ H_1: \beta_3 \neq 0 \end{cases}$$

TEST STATISTIC

$z^{obs} = 2.423$

(If we want to find $SE(\hat{\beta}_3) = \dfrac{0.804038}{2.423} = 0.3318$)

p-value =

$\alpha^{obs} = 2(1 - \Phi(2.423)) \overset{*}{\approx} 0.015$

We reject Ho at 5% sign. level.

* If we know the following quantiles of gaussian distribution

$q_{0.993} = 2.457$

Exercise 4.7

Now, we would like to evaluate our models in terms of mis-classification. Let consider two logistic regression models with probit and logit link functions, involving all covariates. The confusion matrix for each model corresponds to

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 157  40
         1 116  87
```

Figure 1: Probit

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 254  97
         1  19  30
```

Figure 2: Logit

Compute the mis-classification error and interpret the results.

PROBIT

We can compute the accuracy as

$$\text{Accuracy} = \frac{157+87}{N} = \frac{157+87}{400} = 0.61$$

And the misclassification rate is equal to 0.39.
Hence, 39% of ~~[illegible]~~ outcomes are misclassified

LOGIT

$$\text{Accuracy} = \frac{254+30}{400} = 0.71$$

Mis-classification rate is equal to $1-0.71 = 0.29$

Hence, 29% of ~~[illegible]~~ outcomes are misclassified

The logit model better fit the data while then ~~[illegible]~~ using probit we need to fix a threeshold to obtain a binary prediction. Hence, the performance can be affected from this choice.

5 Exercise 5

A researcher reports an experiment on the toxicity to the tobacco budworm *Heliothis virescens* of doses of the pyrethroid *trans*-cypermethrin to which the moths were beginning to show resistance. Batches of 20 moths of each sex were exposed for three days to the pyrethroid and the number in each batch that were dead or knocked down was recorded. The results were

	Dose					
Sex	1	2	4	8	16	32
Male	1	4	9	13	18	20
Female	0	2	6	10	12	16

Exercise 5.1
Write an appropriate regression model using $\log_2(Dose)$ and *Sex*. Justify your answer.

In this case, we have grouped data which means binomial data.

Therefore, for each level of our covariates, the above table shows

$$S_{ij}^{*} = \sum_{k=1}^{N} \mathbb{1}(T_k = 1 \mid Dose_k = x_i, Sex = j)$$

where $j \in \{0,1\}$ and $x_i \in$ any levels of $\log_2(Dose)$

$$m_{ij} = \sum_{k=1}^{N} \mathbb{1}(Dose_k = x_i, Sex = j)$$

And we have

$$S_{i1} \overset{\perp}{\sim} Bin(m_{i1}, \pi_i = \pi(x_i))$$
$$S_{i2} \overset{\perp}{\sim} Bin(m_{i2}, \pi_i = \pi(x_i))$$

The regression model corresponds to

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_1 + \beta_2 \log_2(Dose_i) + \beta_3 D_{1i}$$

$$\text{where } D_{1i} = \begin{cases} 1, & \text{if Sex = Male} \\ 0, & \text{otherwise} \end{cases}$$

Exercise 5.2

Considering the following table

Covariates	Estimates	SE
Intercept	-3.4732	0.4685
Dose	1.0642	0.1311
Sex	1.1007	0.3558

Interpret the estimated regression coefficients and obtain a confidence interval at 95% confidence level.

INTERPRETATION of regression coefficients

- the log odds encrease by 1.0642 ~~for Dose~~ for every log Dose of pyrethroid, for male ~~or~~ female moths
- the log odds ~~encrease by 1.1007~~ for a male moth are 1.1007 times that for a female moths, given a fixed dose of pyrethroid

Confidence Intervals

~~$IC(1-\alpha)=(\hat{\beta}_j \pm z_{1-\frac{\alpha}{2}})$~~

$IC(1-\alpha) = (\hat{\beta}_j \pm z_{1-\frac{\alpha}{2}} SE(\hat{\beta}_j))$

β_2 : $IC(0.95) = (1.0642 - 1.96 \cdot 0.1311, 1.0642 + 1.96 \cdot 0.1311)$

$= (0.807244, 1.321156)$

β_3 : $IC(0.96) = (1.1007 - 1.96 \cdot 0.3558, 1.1007 + 1.96 \cdot 0.3558)$

$= (0.403332, 1.798068)$