CS&SS/STAT/SOC 536: Regression for Binary Outcomes

Adrian Dobra
adobra@uw.edu

1 Why simple linear regression is not an option?

We assume that the response variable y belongs to $\{0,1\}$. Once again, we assume we have only one explanatory variable x. It follows that

$$\begin{aligned}
E[y|x] &= 1 \cdot P(y=1|x) + 0 \cdot P(y=0|x) = P(y=1|x),\\
E[y^2|x] &= 1^2 \cdot P(y=1|x) + 0^2 \cdot P(y=0|x) = P(y=1|x),\\
Var[y|x] &= E[y^2|x] - (E[y|x])^2 = P(y=1|x)(1-P(y=1|x)).
\end{aligned}$$

The above relationships say that $P(y=1|x)$ is the quantity that needs to be modeled. As an aside, we have $P(y=1|x) = 1 - P(y=0|x)$.

Now let's say we make use of the linear regression model:

$$y = \beta_0 + \beta_1 x + \epsilon, \quad \epsilon \sim N(0,\sigma^2).$$

Under this model, we have

$$\begin{aligned}
E[y|x] &= \beta_0 + \beta_1 x,\\
Var[y|x] &= \sigma^2.
\end{aligned}$$

This implies

$$\begin{aligned}
P(y=1|x) &= \beta_0 + \beta_1 x,\\
P(y=1|x)(1-P(y=1|x)) &= \sigma^2.
\end{aligned}$$

The first equation does not make sense since $P(y=1|x) \in (0,1)$, while $\beta_0 + \beta_1 x$ cannot always be positive and smaller than 1 for all values of x. Moreover, the second equation implies:

$$\beta_0 + \beta_1 x = \frac{1}{2} \pm \sigma,$$

which should hold for any value of x. Again, this is impossible.

2 Probit and Logit: two latent variable models

We introduce a new latent variable y^*. A latent variable is not part of your data. It is only part of your modeling framework. We model the relationship between the latent variable y^* and the explanatory variable x using a simple linear regression:

$$y^* = \beta_0 + \beta_1 x + \epsilon.$$

The relationship between the binary outcome y and y^* is a function of the sign of y^*:

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

The *probit* regression model is obtained if we assume that the errors follow a standard normal distribution, i.e. $\epsilon \sim N(0,1)$. You may notice that we did not assume that the variance of the errors is unknown, i.e. $\epsilon \sim N(0,\sigma^2)$. Assuming up-front that $\sigma = 1$ is necessary to make the probit regression model identifiable. This means that the estimates of the model parameters are unique. The form of the probit model does not allow us to estimate σ. You can see this by writing:

$$\sigma y^* = \sigma\beta_0 + \sigma\beta_1 x + \sigma\epsilon.$$

The relationship between y and y^* will remain the same for any value of σ, hence σ cannot be estimated. As such, it has to be assumed fixed and a common choice is taking $\sigma = 1$.

The *logit* regression model is obtained if we assume that the errors follow a *logistic* distribution. In order to understand the relationship between probit and logit regression, we need to take a closer look at the distribution of the errors.

The density and the CDF of $N(0,1)$ are:

$$\phi(t) = \frac{1}{\sqrt{2\pi}} \exp(-t^2/2), \quad \Phi(t) = \int\limits_{-\infty}^{t} \phi(z)dz.$$

The logistic density and CDF are:

$$f(t) = \frac{\exp(t)}{[1+\exp(t)]^2}, \quad F(t) = \int\limits_{-\infty}^{t} f(z)dz = \frac{\exp(t)}{1+\exp(t)}.$$

When the errors are distributed $N(0,1)$, we have $E[\epsilon] = 0$ and $Var[\epsilon] = 1$. On the other hand, simple calculations show that, when the errors follow a logistic distribution, we have $E[\epsilon] = 0$ and $Var[\epsilon] = \frac{\pi^2}{3}$.

Let's write the logit regression model:

$$y^{*L} = \beta_0^L + \beta_1^L x + \epsilon^L, \quad \epsilon^L \sim f(\cdot),$$

and the probit regression model:

$$y^{*P} = \beta_0^P + \beta_1^P x + \epsilon^P, \quad \epsilon^P \sim \phi(\cdot).$$

We have $E[\epsilon^L] = E[\epsilon^P] = 0$. On the other hand, $Var[\epsilon^L] = \frac{\pi^2}{3} Var[\epsilon^P]$, which implies

$$\epsilon^L \approx \frac{\pi}{\sqrt{3}} \epsilon^P.$$

and

$$\beta_0^L \approx \frac{\pi}{\sqrt{3}} \beta_0^P, \quad \beta_1^L \approx \frac{\pi}{\sqrt{3}} \beta_1^P.$$

In the previous section we argued that the quantity we will model is $P(y = 1|x)$. If we work with the logit with probit model, we have

$$\begin{aligned} P(y = 1|x) &= P(y^* > 0|x), \\ &= P(\beta_0 + \beta_1 x + \epsilon > 0|x), \\ &= P(\epsilon > -(\beta_0 + \beta_1 x)), \\ &= P(\epsilon < \beta_0 + \beta_1 x), \\ &= \begin{cases} \Phi(\beta_0^P + \beta_1^P x), & \text{for PROBIT} \\ F(\beta_0^L + \beta_1^L x), & \text{for LOGIT} \end{cases} \end{aligned}$$

If we look back at the form of the logistic CDF, we see that:

$$P(y = 1|x) = F(\beta_0^L + \beta_1^L x) = \frac{\exp(\beta_0^L + \beta_1^L x)}{1 + \exp(\beta_0^L + \beta_1^L x)}. \tag{1}$$

We define the *logit* function:

$$\mathit{logit} : (0, 1) \longrightarrow (-\infty, +\infty), \quad \mathit{logit}(p) = \log \frac{p}{1 - p}.$$

Its inverse is:

$$\mathit{logit}^{-1} : (-\infty, +\infty) \longrightarrow (0, 1), \quad \mathit{logit}^{-1}(x) = \frac{\exp(x)}{1 + \exp(x)}.$$

Remark that $\mathit{logit}^{-1}(x) = F(x)$, the logistic CDF. With this notation, equation (1) becomes:

$$P(y = 1|x) = \mathit{logit}^{-1}(\beta_0^L + \beta_1^L x),$$

or, equivalently:

$$\mathit{logit}(P(y = 1|x)) = \beta_0^L + \beta_1^L x.$$

Yet another equivalent form is the logistic regression equation:

$$\log \frac{P(y = 1|x)}{P(y = 0|x)} = \beta_0^L + \beta_1^L x.$$

3 Maximum Likelihood Estimation

We assume to have observed the independent samples $(y_1, x_1), \ldots, (y_n, x_n)$. The probit and the logistic regression models have two parameters: $\theta = (\beta_0, \beta_1)$ that need to be estimated from the data. Our model assumptions say that each say that each y_i follows a Bernoulli distribution with probability of success $P(y_i = 1|x_i)$:

$$y_i \sim Ber(P(y = 1|x_i)).$$

Since the samples are assumed to be independent, the likelihood is:

$$L(\theta) = \prod_{i=1}^{n} [P(y_i = 1|x_i)]^{y_i} [1 - P(y_i = 1|x_i)]^{1-y_i}.$$

The log-likelihood is:

$$l(\theta) = \sum_{i=1}^{n} (y_i \log P(y_i = 1|x_i) + (1 - y_i) \log[1 - P(y_i = 1|x_i)]).$$

We denote:

$$\pi_i = P(y_i = 1|x_i) = logit^{-1}(\beta_0 + \beta_1 x_i) \in (0, 1).$$

Thus the log-likelihood becomes:

$$l(\theta) = \sum_{i=1}^{n} (y_i \log \pi_i + (1 - y_i) \log[1 - \pi_i]).$$

Simple calculations show that

$$\begin{aligned} \frac{\partial l(\theta)}{\partial \beta_0} &= \sum_{i=1}^{n} [y_i - \pi_i], \\ \frac{\partial l(\theta)}{\partial \beta_1} &= \sum_{i=1}^{n} [y_i x_i - \pi_i x_i]. \end{aligned}$$

Since the log-likelihood $l(\theta)$ is concave, we maximize it by solving the system of equations:

$$\frac{\partial l(\theta)}{\partial \beta_0} = 0, \quad \frac{\partial l(\theta)}{\partial \beta_1} = 0.$$

The MLEs $\widehat{\theta} = (\widehat{\beta_0}, \widehat{\beta_1})$ should be determined by solving

$$\begin{aligned} \sum_{i=1}^{n} y_i &= \sum_{i=1}^{n} \pi_i, \\ \sum_{i=1}^{n} y_i x_i &= \sum_{i=1}^{n} \pi_i x_i. \end{aligned}$$

It is not possible to solve this system explicitly the way we did in the case of simple linear regression, hence we will need to employ an iterative numerical proceduce such as Newton-Raphson (SEE HANDOUT!) to find the MLEs $\widehat{\theta}$. To this end, we need to compute the Hessian of $l(\theta)$:

$$H(\theta) = H(\beta_0, \beta_1) = \left[\begin{array}{cc} \frac{\partial^2 l(\theta)}{\partial \beta_0^2} & \frac{\partial^2 l(\theta)}{\partial \beta_0 \partial \beta_1} \\ \frac{\partial^2 l(\theta)}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 l(\theta)}{\partial \beta_1^2} \end{array} \right].$$

We have

$$\begin{aligned} \frac{\partial^2 l(\theta)}{\partial \beta_0^2} &= -\sum_{i=1}^{n} \pi_i (1 - \pi_i), \\ \frac{\partial^2 l(\theta)}{\partial \beta_0 \partial \beta_1} = \frac{\partial^2 l(\theta)}{\partial \beta_1 \partial \beta_0} &= -\sum_{i=1}^{n} \pi_i (1 - \pi_i) x_i, \\ \frac{\partial^2 l(\theta)}{\partial \beta_1^2} &= -\sum_{i=1}^{n} \pi_i (1 - \pi_i) x_i^2. \end{aligned}$$

The Newton-Raphson algorithm starts with some initial values $\theta^{(0)} = (\beta_0^{(0)}, \beta_1^{(0)})$. Typically $\theta^{(0)} = 0$ should work fine. At iteration k, the current values $\theta^{(k)} = (\beta_0^{(k)}, \beta_1^{(k)})$ are updated as follows:

$$\theta^{(k+1)} = \theta^{(k)} - H^{-1}(\theta^{(k)}) \left(\begin{array}{c} \frac{\partial l(\theta^{(k)})}{\partial \beta_0} \\ \frac{\partial l(\theta^{(k)})}{\partial \beta_1} \end{array} \right).$$

After running the updates for a certain number of iterations (need to check convergence!), the last updated values $\theta^{(k)}$ will be the MLEs $\widehat{\theta}$.

4 Interpreting the coefficients of a logistic regression

The interpretation of the coefficients of a logistic regression cannot be done in the same way we interpreted the coefficients of a simple linear regression. For example, if the explanatory variable x is continuous, we would try to obtain an interpretation of β_1 by calculating the partial derivative:

$$\frac{\partial P(y=1|x)}{\partial x} = P(y=1|x) \cdot P(y=0|x) \cdot \beta_1.$$

There is no easy way to express in words the meaning of the expression above (at least as far as I can tell!). When x is binary, the corresponding quantity is

$$P(y=1|x=1) - P(y=1|x=0).$$

It is a lot more convenient to define the odds:

$$\Omega(x) = \frac{P(y=1|x)}{P(y=0|x)}.$$

With this notation, the logistic regression is written as:

$$\log \Omega(x) = \beta_0 + \beta_1 x.$$

Now, if x is continuous, the partial derivative of the log odds is

$$\frac{\partial \log \Omega(x)}{\partial x} = \beta_1.$$

If x is binary, we have

$$\log \frac{\Omega(x+1)}{\Omega(x)} = \beta_1.$$

The interpretation of β_1 is as follows:

> *For a unit change in x, the odds $\Omega(x)$ are expected to change by a factor of* $\exp(\beta_1)$.

5 Variable Selection

One of the most important problems you will need to deal with is the question of which variables do you want to include in your model? Consider the multivariate logistic regression model:

$$\mathcal{M}: \quad \log \frac{P(y=1|x_1,\ldots,x_p)}{P(y=0|x_1,\ldots,x_p)} = \beta_0 + \beta_1 \cdot x_1 + \ldots + \beta_p \cdot x_p.$$

We want to decide between $\mathcal{M}'$ and another logistic model that involves the additional variables $x_{p+1},\ldots,x_{p'}$:

$$\mathcal{M}': \quad \log \frac{P(y=1|x_1,\ldots,x_p,\ldots,x_{p'})}{P(y=0|x_1,\ldots,x_p,\ldots,x_{p'})} = \beta_0 + \beta_1 \cdot x_1 + \ldots + \beta_p \cdot x_p + \beta_{p+1} \cdot x_{p+1} + \ldots + \beta_{p'} \cdot x_{p'}.$$

Note that the additional terms might be quadratic terms or interactions between two or more variables, e.g.

$$x_1^2 \quad x_1 \cdot x_2 \quad x_3 \cdot x_{10} \cdot x_{11}.$$

We denote by $|\mathcal{M}|$ the *dimension* of model $|\mathcal{M}|$ given by the number of free parameters of $\mathcal{M}$. For example, $|\mathcal{M}| = p+1$ since $\mathcal{M}$ has $p+1$ parameters, while $|\mathcal{M}'| = p'+1$. We also introduce the *deviance* of $\mathcal{M}$ to be

$$D(\mathcal{M}) = -2l(\mathcal{M}),$$

where $l(\mathcal{M})$ is the maximum value of the log-likelihood of model $\mathcal{M}$. That is, $l(\mathcal{M})$ is the log-likelihood evaluated at the MLEs of the regression parameters. Since $\mathcal{M}$ is obtained from $\mathcal{M}'$ by setting $\beta_{p+1} = \ldots = \beta_{p'} = 0$, it should be obvious that

$$l(\mathcal{M}) \leq l(\mathcal{M}'),$$

which implies that

$$D(\mathcal{M}) \geq D(\mathcal{M}').$$

This relationship holds between any two models that are nested within each other.

Deciding between $\mathcal{M}$ and $\mathcal{M}'$ is done by testing the null hypothesis:

$$H_0: \quad \beta_{p+1} = \ldots = \beta_{p'} = 0.$$

We consider the likelihood ratio statistic defined as the difference in the deviances of $\mathcal{M}$ and $\mathcal{M}'$:

$$G^2(\mathcal{M}|\mathcal{M}') = D(\mathcal{M}) - D(\mathcal{M}')$$

It can be shown that *asymptotically* (that is, as the sample size n goes to ∞) $G^2(\mathcal{M}|\mathcal{M}')$ follows a Chi-square distribution with $|\mathcal{M}'| - |\mathcal{M}|$ degrees of freedom. Therefore the p-value corresponding with H_0 is

$$P(\chi_{p'-p} \geq G^2(\mathcal{M}|\mathcal{M}')).$$

If the p-value is greater than 0.05 (or 0.01, your choice), we *fail* to reject H_0. In other words, we favor $\mathcal{M}$ against the more complex model $\mathcal{M}'$. If the p-value is smaller than 0.05 (or 0.01), we *reject* H_0 and favor $\mathcal{M}'$ against the simpler model $\mathcal{M}$.

The major drawback of hypothesis testing is that you can only compare two nested models. For example, you can compare a regression that involves the intercept and two explanatory variables:

$$\mathcal{M}_1: \quad \log \frac{P(y=1|x_1,x_2)}{P(y=0|x_1,x_2)} = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2,$$

against the model that involves only the intercept:

$$\mathcal{M}_2: \quad \log \frac{P(y=1)}{P(y=0)} = \beta_0.$$

But how can you compare $\mathcal{M}_1$ against

$$\mathcal{M}_3: \quad \log \frac{P(y=1|x_3)}{P(y=0|x_3)} = \beta_0 + \beta_3 \cdot x_3.$$

It is unclear which null hypothesis one would need to set up and which test statistic one would need to use. To this end, statisticians have prosed *information theoretic* measures to assess model fit. The most popular measures are *Akaike's information criterion* (AIC) and *Bayesian information criterion* (BIC):

$$\begin{aligned} AIC(\mathcal{M}) &= D(\mathcal{M}) + 2 \cdot |\mathcal{M}|, \\ BIC(\mathcal{M}) &= D(\mathcal{M}) + \log(n) \cdot |\mathcal{M}|. \end{aligned}$$

Models with *smaller* values of AIC or BIC are preferred. Why? The first part of AIC and BIC involves the deviance of the model. This part will *always* decrease as you add more and more parameters/variables to the model (i.e., the fit of the model can be improved with the addition of *any* variables, even if they are pure noise). The second part of AIC and BIC penalizes for the complexity of a model. That is, this second part will *always* increase as you add more and more parameters/variables to the model. The first and second parts put together create a balance between fit and complexity. These two information theoretic measures will help you decide between any two models even if they are not nested within each other.

6 Residuals and Prediction

We assume that you have observed n samples involving a binary outcome and p explanatory variables:

$$(y^i, x_1^i, \ldots, x_p^i), \ldots i = 1, \ldots, n.$$

Let $(\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p)$ be the MLEs corresponding with the logistic regression model

$$\mathcal{M}: \quad \log \frac{P(y=1|x_1,\ldots,x_p)}{P(y=0|x_1,\ldots,x_p)} = \beta_0 + \beta_1 \cdot x_1 + \ldots + \beta_p \cdot x_p.$$

The *predicted* value at $(x_1, \ldots, x_p)$ is

$$\widehat{\pi}(x^1, \ldots, x^p) = logit^{-1}(\widehat{\beta}_0 + \widehat{\beta}_1 \cdot x^1 + \ldots + \widehat{\beta}_p \cdot x^p).$$

The *fitted* value corresponding with the i-th sample is the predicted value at $(x_1^i, \ldots, x_p^i)$:

$$\widehat{\pi}_i = \widehat{\pi}(x_1^i, \ldots, x_p^i).$$

Remember that the observed value y_i is binary, i.e. $y_i \in \{0,1\}$. On the other hand, the corresponding fitted value is $\widehat{\pi}_i \in (0,1)$ and represents the probability that the outcome takes value 1 at $(x_1^i, \ldots, x_p^i)$. You also remember that y_i was assumed to follow a Bernoulli distribution with probability of success (i.e., the probability of taking value 1) equal to $\widehat{\pi}_i$. This implies $E[y_i|x_1^i, \ldots, x_p^i] = \widehat{\pi}_i$ and $Var[y_i|x_1^i, \ldots, x_p^i] = \widehat{\pi}_i(1-\widehat{\pi}_i)$ (SEE THE FIRST PAGE OF THE HANDOUT).

The discrepancy between the observed and the fitted values associated with the i-th sample is quantified using the *Pearson residual*:

$$r_i = \frac{y_i - E[y_i|x_1^i, \ldots, x_p^i]}{\sqrt{Var[y_i|x_1^i, \ldots, x_p^i]}} = \frac{y_i - \widehat{\pi}_i}{\sqrt{\widehat{\pi}_i(1-\widehat{\pi}_i)}}.$$

Larger values indicate a higher discrepancy. In other words, the regression model performs *worse* for samples with larger residuals. It is good practice to create an *index plot* that has the observation number on the x-axis and the residuals on the y-axis. This means that you plot the points:

$$(i, r_i), \quad i = 1, \ldots, n.$$

If your model is appropriate for your data, all Pearson's residuals should be between -2 and 2 (WHY?). The samples whose residuals are larger than 2 in absolute value could indicate outliers or influential observations, etc. It is always a good idea to take a closer look at such samples where your model does not perform too well.

You can formally test whether the fit of your model is dominated by outliers by calculating the sum of squares of the Pearson residuals:

$$\sum_{i=1}^{n} r_i^2$$

It can be shown that asymptotically this quantity follows a Chi-square distribution with $n - |\mathcal{M}| = n - p - 1$ degrees of freedom. You decide that your model does not fit the data well if the p-value

$$P\left(\chi^2_{n-p-1} \geq \sum_{i=1}^{n} r_i^2\right).$$

is smaller than 0.05 (or 0.01). Your intuition should go like this:

$$\text{worse fit} \Leftrightarrow \text{larger residuals} \Leftrightarrow \text{ larger } \sum_{i=1}^{n} r_i^2 \Leftrightarrow \text{ smaller probability to be out in the tail of } \chi^2_{n-p-1}.$$

The Brier score is the sum of squares of the difference between the observed and the fitted values:

$$Brier(\mathcal{M}) = \sum_{i=1}^{n} (y_i - \widehat{\pi}_i)^2.$$

Larger values of the Brier score indicate a worse fit. The Brier score will improve (i.e., gets smaller) as the fit of $\mathcal{M}$ improves. It not recommended to use the Brier score as a model selection criterion since it does not penalize for increased complexity of your model (i.e., the addition of extra parameters). In order to make the Brier score reflect the predictive performance of your model, you need to perform *cross-validation*. This involves fitting the model with a part of your samples and predict the other samples. For example, leave-one-out cross validation means that you hold out one sample, calculate the MLEs of your regression parameters based on the remaining $n - 1$ samples, then obtain the predicted value for that sample based on these coefficient estimates. The procedure is repeated for each sample. At the end of this process you compute your Brier score, which will be higher than the Brier score associated with fitted values obtained by fitting your regression based on all the available samples.

Occasionally it does not suffice to report the fitted values $\widehat{\pi}_i$, $i = 1, 2, \ldots, n$. You could produce the fitted values $\widehat{y}_i \in \{0, 1\}$ as the most likely binary value indicated by $\widehat{\pi}_i$:

$$\widehat{y}_i = \begin{cases} 1, & \text{if } \widehat{\pi}_i \geq 0.5, \\ 0, & \text{if } \widehat{\pi}_i < 0.5. \end{cases}$$

The *prediction error* is defined as the number of samples correctly predicted (i.e., $y_i = \widehat{y}_i$) divided by the sample size n.