B.2 Regression for Categorical Response

Sources and Objectives

This lesson comes from Frees sections 11.1-11.2 and is designed to address SRM Syllabus learning objective 2(d).

Binary Response Variables and Regression

Bernoulli Random Variable (a.k.a. Binary)

$$Y = \begin{cases} 0 & \text{with prob. } 1-\pi \\ 1 & \text{with prob. } \pi \end{cases} \longrightarrow E[Y] = \pi \quad \text{Var}(Y) = \pi(1-\pi)$$

Ex. $Y_i = 1$ if customer i defaults on credit account. Suppose that for each customer we also know $X_i =$ customer i's account balance.

Plot of Y vs. X: points at height 0 or 1 at horizontal locations X.

Statistical Learning model: $Y_i = f(X_i) + \epsilon$

Classification: Aims to predict Y_i. $\hat{f}(X_i) = 0$ or 1, $e_i^2 = 0$ or 1

Regression: Pretend Y is quantitative, try to predict frequency

$$Y_i = f(X_i) + \epsilon \longrightarrow \mathrm{E}[Y_i] = \pi_i = f(X_i)$$

Challenges for Regression

Regression for $Y_i = f(X_i) + \epsilon \longrightarrow E[Y_i] = \pi_i = f(X_i)$ faces challenges.

- $0 \leqslant \pi_i \leqslant 1$, so $0 \leqslant f(X_i) \leqslant 1$ is needed to make sense of $\hat{\pi} = \hat{f}(X_i)$ or to even fit model using MLE
- $Y_i = 0$ or 1, so $0 \leqslant f(X_i) \leqslant 1 \longrightarrow \epsilon$ is not normally distributed, typical analysis using the e_i is not useful
- $\text{Var}(Y_i)$ depends on π_i. If π_i depends on $X_i \longrightarrow$ heteroscedasticity, model estimates better not depend on homoskedasticity.

The Null Model

The Null Model ignores X_i, so $f(X_i) = \beta_0$

$\hat{\pi}_i = \hat{f}(X_i) = \hat{\beta}_0$ for all i.

Least squares: Computations same as when Y is continuous.

Minimize $\sum(y_i - \hat{\beta}_0)^2 \longrightarrow \hat{\beta}_0 = \bar{Y}$ = proportion of 1's in the sample.

The MLE of $\hat{\beta}_0$ is also $\bar{Y}$.

As before, if $\pi_i = \pi = \beta_0$ is the correct model, then

$$\hat{\beta}_0 \approx N\left(\pi, \text{Var}(Y)/n\right) = N\left(\pi, \frac{\pi(1-\pi)}{n}\right)$$

Of course, in most cases we will expect π_i to depend on X_i, so we will need other models.

Linear Probability Model

First attempt: *Linear Probability model* $\text{E}[Y_i] = \beta_0 + \beta_1 X_i$

Advantages:

- Computing least squares $\hat{\beta}_0, \hat{\beta}_1$ is easy
- Interpretation: β_1 is the amount π increases with each unit increase in X

Main drawback: If $\beta_1 \neq 0$, $\beta_0 + \beta_1 X$ won't always be in $[0, 1]$

If $0 < \beta_0 + \beta_1 X_i < 1$ over the training data, then least squares $\hat{\beta}_0, \hat{\beta}_1$ are unbiased, but their variances are not optimal, due to heteroscedasticity.

Specification error is common as well, since π usually cannot be linear in X over the entire theoretical range of X.

Upshot: There are times for this model, but usually there's a better choice.

Dealing with Heteroscedasticity

In the cases where the linear probability model is preferred, there are two options to treat heteroscedasticity:

1. Compute standard errors in a way that incorporates heteroscedasticity - *heteroskedastic robust standard errors*
2. Use two-stage weighted least squares
   - Use least squares to estimate parameters.
   - Use variance estimates from the results to rescale the observations so that they all have estimated variance 1.
   - Run least squares again on the rescaled observations.

We will talk more about what the R-routines are doing in these cases later on.

Exercise 1

A model is to be built to compute the expected number of defaults for a set of loans. In the training data set, there are 4 individuals with a balance of $10,000, of which 2 defaulted, 3 with a balance of $12,000, of which 1 defaulted, and 1 with a balance of $16,000, which also defaulted. The null model is computed for this training data, and is applied to a new set of loans which consists of 7 individuals with a balance of $11,000, 10 with a balance of $14,000, and 3 with a balance of $15,000. How many loans are expected to default under the null model?

Exercise 1

A model is to be built to compute the expected number of defaults for a set of loans. In the training data set, there are 4 individuals with a balance of \$10,000, of which 2 defaulted, 3 with a balance of \$12,000, of which 1 defaulted, and 1 with a balance of \$16,000, which also defaulted. The null model is computed for this training data, and is applied to a new set of loans which consists of 7 individuals with a balance of \$11,000, 10 with a balance of \$14,000, and 3 with a balance of \$15,000. How many loans are expected to default under the null model?

The null model ignores the balance, and predicts that every loan will default with probability $\hat{\beta}_0 = \bar{Y} = \dfrac{\text{4 defaults}}{\text{8 total loans}} = 0.5$.

The test data set has 20 total (independent, we assume) loans, so the number that default, according to the null model, will be the sum of 20 Bernoulli random variables with probability $\pi = 0.5$. The expected value of this Binomial(20,0.5) random variable is $20 \cdot 0.5 = \boxed{10}$.

Exercise 2

A linear probability model to compute the probability of defaults as a function of loan balance in thousands of dollars, resulting in $\hat{\beta}_0 = -0.2667$ and $\hat{\beta}_1 = 0.067$. If this model is applied to a new set of loans which consists of 7 individuals with a balance of \$11,000, 10 with a balance of \$14,000, and 3 with a balance of \$15,000. How many loans are expected to default under the linear probability model?

Exercise 2

A linear probability model to compute the probability of defaults as a function of loan balance in thousands of dollars, resulting in $\hat{\beta}_0 = -0.2667$ and $\hat{\beta}_1 = 0.067$. If this model is applied to a new set of loans which consists of 7 individuals with a balance of \$11,000, 10 with a balance of \$14,000, and 3 with a balance of \$15,000. How many loans are expected to default under the linear probability model?

Predicted probability of default is different for each loan amount.

For each loan of \$11K, $\hat{\pi} = -0.2667 + 0.067 \cdot 11 = 0.4703$

For each loan of \$14K, $\hat{\pi} = -0.2667 + 0.067 \cdot 14 = 0.6713$

For each loan of \$15K, $\hat{\pi} = -0.2667 + 0.067 \cdot 15 = 0.7383$

$$\text{E}[\text{defaults}] = 7 \cdot 0.4703 + 10 \cdot 0.6713 + 3 \cdot 0.7383 = \boxed{12.2}$$