# CIV1538 Assignment 1: Binary Models

Tiggy Chen 1004400784

27 February 2024

# 1 Introduction

The following report describes the process used to develop two binary models for peoples' ranking of local cycling infrastructure quality. The two models are:

- A **binary probit model** to predict whether someone ranks the overall quality of all cycling facilities as non-positive (>2 on a scale of 1-6, from best to worst).
- A **binary logit model** to predict whether someone ranks the quality of local on-street cycling lanes as non-excellent (>1 on a scale of 1-6, from best to worst).

For each model, this report will discuss the model formulation, choice of explanatory variables, best model selection, and resulting estimated parameter values.

# 2 Binary Probit Model

## 2.1 Model Formulation

Both the binary probit model and binary logistic model follow the same general binary model formulation, shown below in Equation 1, where the binary outcome $y_i$ is predicted based on a latent variable, $y_i^*$. The latent variable, whose typical model formulation is linear in parameter and shown in Equation 2, is associated with the probability of the binary outcome being a "success" ($y_i = 1$).

$$y_i = \begin{cases} 1, & y_i^* > 0 \\ 0, & y_i^* \leq 0 \end{cases} \tag{1}$$

$$y_i^* = \beta_0 + \sum \beta_j x_j + \epsilon_i \tag{2}$$

The binary probit and logit models differ in their assumptions regarding the latent variable. In the binary probit model, it is assumed that the latent variable's error term follows a normal distribution. The latent variable parameters can therefore be estimated using maximum likelihood estimation, where likelihoods are calculated using a normal probability density function.

## 2.2 Explanatory Variables

Recall that, for the binary probit model, the response variable of interest is whether people rank quality of local cycling facilities greater than two (one representing best quality and six representing worst), i.e. whether they have non-positive outlooks on local cycling facilities.

The list of all explanatory variables that were included for consideration in the binary probit model, along with rationale, is as follows:

- **Sex**: Males are often more confident cyclists and may perceive available facilities more positively than females.
- **Age**: Younger adults tend to be more confident than older adults and more likely to be more comfortable with lower-quality lanes.
- **Presence of children in households**: People living in households with children (<15 years) may be more concerned about their children's safety on cycling facilities, leading to less tolerance for lower-quality facilities.
- **Number of people in household who bike for various purposes**: People and households who are more willing to bike for various purposes may have more favourable attitudes to local facilities.
- **Access to motor vehicle**: People forced to bike due to lack of alternatives, or who enjoy biking enough to not own cars, may have more extreme attitudes.
- **Time from nearest major bike path or bike lane**: Accessibility to facilities may affect perception of overall quality of facilities.
- **Length of bike facilities available in neighbourhood**: Availability of facilities may affect perception of overall quality.

A variety of combinations of explanatory variables were tested to find statistically significant variables leading to a model with the best fit, or a model with high McFadden pseudo R-squared, low Akaike Information Criterion (AIC) score, and low log-likelihood ratio (LLR).

## 2.3 Model Selection

The following binary probit model was estimated using the discrete probit model of the Python's StatsModel library. All Python code used to produce the model is provided in Appendix A.

An overview of the best binary probit model's estimated parameters and fit statistics is provided in Table 1. The model fit is quite poor: the pseudo-$R^2$ is only 0.039 (a "good" score is typically $\geq$0.2) and the model AIC is not significantly better than the

null model AIC. However, all variables are statistically significant ($p < 0.05$) and the LLR p-value is relatively low, suggesting further model truncation is not necessary.

Table 1: Binary Probit Model Summary

| **Variable** | **Estimate** | **Z-statistic** | **P-value** |
|---|---|---|---|
| Intercept | 0.4933 | 3.557 | 0.000 |
| Number of household members who bike to work | 0.2962 | 2.023 | 0.043 |
| Children in household (1=yes, 0=no) | 0.4451 | 2.250 | 0.024 |
| **Model Fit Statistics** | **Best Model** | **Null Model** | |
| AIC | 243.35 | 249.03 | |
| Log-likelihood | -118.68 | -123.51 | |
| Log-likehood ratio (p-value) | 0.961 (0.0079) | | |
| McFadden's pseudo $R^2$ | 0.039 | 1.084e-10 | |

The estimated parameters generally correspond with expected relationships. Since the response variable represents whether respondents ranked quality of local cycling facilities non-positively, the positive parameter estimates indicate that households containing children or greater numbers of cycling commuters tend to rank cycling facilities' quality more negatively. People who bike to work are more likely to be "captive" cyclists rather than "choice" cyclists and may therefore rank facility quality lower, or be more likely to bike even when facility quality is lower. Similarly, households with children are more likely to be concerned about their children's safety, which would lead them to rank facility quality more harshly.

# 3 Binary Logit Model

## 3.1 Model Formulation

The general formulation of a binary logit model is the same as a binary probit model and also follows Equations 1 and 2. However, whereas the binary probit model assumes that the latent variable's error term follows a normal distribution, the binary logit model instead assumes that the error term follows a logit distribution. Therefore, parameter estimation is conducted using maximum likelihood estimation where likelihoods are calculated using the probability density function of a logit distribution.

## 3.2 Explanatory Variables

Recall that, for the binary logit model, the response variable of interest is whether people rank quality of local on-street cycling lane facilities to be greater than one (one

representing best quality and six representing worst), i.e. whether they think that local on-street cycling lane facilities are less than excellent.

The list all explanatory variables included for consideration is the same for the binary logit model as for the binary probit model and can be viewed in Section 2.2. Expected relationships between explanatory variables and the response variable are generally unchanged.

As with the binary probit model, a variety of combinations of explanatory variables were tested to find statistically significant variables leading to a model with good fit.

## 3.3 Model Selection

The following binary logit model was estimated using the discrete logit model in Python's StatsModel library. All Python code used to produce the model is provided in Appendix A.

An overview of the best binary logit model, including its estimated parameters and fit statistics, is provided in Table 2. Note that unlike the binary probit model, no combination of variables was found such that all variables included in the model were statistically significant ($p < 0.05$). Additionally, the model fit is quite poor, as the model AIC is quite close to the null model AIC and the model pseudo $R^2$ is only 0.06. However, the LLR p-value is quite low

Table 2: Binary Logit Model Summary

| **Variable** | **Estimate** | **Z-statistic** | **P-value** |
|---|---|---|---|
| Intercept | 1.9839 | 5.249 | 0.000 |
| Number of household members who bike to work | 1.0844 | 1.895 | 0.058 |
| Sex (1=male, 0=female) | 0.8559 | 1.400 | 0.161 |
| **Model Fit Statistics** | **Best Model** | **Null Model** | |
| AIC | 106.33 | 109.09 | |
| Log-likelihood | -50.166 | -53.545 | |
| Log-likehood ratio (p-value) | 0.937 (0.03408) | | |
| McFadden's pseudo $R^2$ | 0.06311 | 6.419e-11 | |

Again, the model shows that an increase in the number of household members who bike to work is associated with lower rankings in quality of local on-street cycling facilities. The model also suggests that male riders are more likely to rank quality of local on-street cycling facilities lower. This may be because male cyclists, who tend to be more confident, are more willing to ride on lower-quality facilities; therefore, lower-quality facilities have greater proportions of male riders compared to higher-quality facilities, which female riders are more willing to use.