# Modelling Continuous Response Variables with Cumulative Probability Models

*Jules Lanari-Collard*

MATH 470 Final Report

McGill University

Montréal, Québec, Canada

December 20, 2024

---

Supervised by Dr. José Correa

# Abstract

Continuous data is itself ordered and hence ordinal regression models can be fitted to continuous data. This report studies the use of Cumulative Probability Models (CPMs), a particular class of ordinal regression models, for modelling continuous data. In particular, we examine an application for where there are missing data caused by detection limits. Chapter 1 reviews basic theory and notation for Generalised Linear Models, and Chapter 2 then studies the theory behind CPMs and the adaptations necessary to fit them to continuous data. Chapter 3 is a case study, fitting a CPM to data on Trace Organic Contaminants in Canadian lakes and discussing interpretations and diagnostics for the resulting model.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction to GLMs

## 1.1 Linear Models

Before diving into generalised linear models (GLMs), and the special case on which this report focuses, it is important to establish some foundational notation and theory.

We observe a (data)set $\mathcal{D}_n = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_n, y_n)\}$, with the observed response $y_i \in \mathbb{R}$ and covariates $\boldsymbol{x}_i \in \mathbb{R}^p$. The linear model considers the conditional expectation of the response random variable $Y_i$ given the covariates $\boldsymbol{X}_i$ as a linear function of the covariates:

$$\mathbb{E}[Y_i | \boldsymbol{X}_i = \boldsymbol{x}_i] = \boldsymbol{x}_i^\top \boldsymbol{\beta} \tag{1.1}$$

To consider the sample as a whole, we define the *design matrix* $\boldsymbol{X}$:

$$\boldsymbol{X} = \begin{bmatrix} — & \boldsymbol{x}_1^\top & — \\ & \vdots & \\ — & \boldsymbol{x}_n^\top & — \end{bmatrix} \in \mathbb{R}^{n \times p} \tag{1.2}$$

And now the full model can be expressed as a function of the observed sample[1]:

$$\mathbb{E}[\boldsymbol{Y} | \boldsymbol{X} = \boldsymbol{x}] = \boldsymbol{X}\boldsymbol{\beta} \tag{1.3}$$

A probabilistic model is constructed by removing the expectation operator and incorporating *residual error* random variables $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n)^\top \in \mathbb{R}^n$, writing:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1.4}$$

where $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}_n$ and $\mathbb{V}[\boldsymbol{\epsilon}] = \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top] = \boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ a symmetric positive definite matrix. For inference, a Gaussian Linear Model is commonly assumed, where $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_n$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \boldsymbol{I}_n)$.

Estimation of $\boldsymbol{\beta}$ generally depends on the assumptions made on the structure of $\boldsymbol{\Sigma}$. Assuming a full rank homoscedastic (i.e. $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_n$) linear model, the ordinary least squares (OLS) estimator $\hat{\boldsymbol{\beta}}_{ols}$ is the **best unbiased estimator** (BUE) of $\boldsymbol{\beta}$ [1]. In the case of

[1] Treating covariates as fixed constants, as opposed to stochastic.

heteroscedasticity (i.e. general $\mathbf{\Sigma}$), (feasible) generalised least squares and its particular case weighted least squares are popular alternatives. There exists widespread and thorough literature on these methods.

## 1.2 The Exponential Family

Generalised linear models apply to random variables within a particular family of distributions, known as the exponential-dispersion family.

**Definition 1** (Exponential Family)**.** The distribution of a random variable $Y$ is in the exponential family with parameters $\theta$ and $\phi$ if its probability density (or mass) function can be written in the form:

$$f_Y(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

with known functions $a$, $b$ and $c$.

$\theta$ is known as the **canonical parameter** and $\phi$ as the **dispersion parameter**. If $\phi$ is unknown, the model is the **exponential-dispersion family**.

The exponential family has some useful properties, namely:

$$\mathbb{E}[Y] = \dot{b}(\theta) = \frac{\partial}{\partial\theta} b(\theta) =: \mu$$
$$\mathbb{V}[Y] = a(\phi)\ddot{b}(\theta) = a(\phi) \cdot \left( \frac{\partial^2}{\partial\theta^2} b(\theta) \right) = a(\phi) V(\mu)$$

**Example 1** (Poisson Model)**.** Let $Y \sim \mathcal{P}(\lambda)$. Then:

$$f_Y(y;\lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$$
$$= \exp\{y \log\lambda - \lambda - \log y!\} \mathbf{1}_{\{y\geq 0\}, y\in\mathbb{Z}^+}$$

Setting $\theta = \log\lambda$, and $\phi = 1$, we have:

$$f_Y(y;\lambda) = \exp\left\{ \frac{y\theta - e^\theta}{\phi} - \log y! \right\}$$

So the Poisson distribution is in the exponential family, with $a(\phi) = \phi$ and $b(\theta) = e^\theta$. We can easily compute the expectation and variance:

$$\mathbb{E}[Y] = \frac{\partial}{\partial\theta} e^\theta = e^\theta = e^{\log\lambda} = \lambda (=: \mu)$$
$$\mathbb{V}[Y] = \phi \cdot \frac{\partial^2}{\partial\theta^2} e^\theta = e^{\log\lambda} = \lambda \qquad (\implies V(\mu) = \mu)$$

For this report, the Bernoulli distribution is of particular interest:

**Example 2** (Bernoulli Model)**.** Let $Y \sim \mathcal{B}(\pi)$. Then:

$$\begin{aligned} f_Y(y;\pi) &= \pi^y(1-\pi)^{1-y}\mathbf{1}_{y\in\{0,1\}} \\ &= \exp\{y\log\pi + (1-y)\log(1-\pi)\} \\ &= \exp\left\{y\log\frac{\pi}{1-\pi} + \log(1-\pi)\right\} \end{aligned}$$

Setting $\theta = \log\frac{\pi}{1-\pi}$ and $\phi = 1$ we see that:

$$f_Y(y;\pi) = \exp\left\{\frac{y\theta - \log(1+e^\theta)}{\phi}\right\}$$

So the Bernoulli distribution is an exponential family, with $a(\phi) = \phi$ and $b(\theta) = \log(1+e^\theta)$. We also notice that:

$$\begin{aligned} \mathbb{E}[Y] &= \dot{b}(\theta) = \frac{e^\theta}{1+e^\theta} = \pi (=: \mu) \\ \mathbb{V}[Y] &= a(\phi)\ddot{b}(\theta) = \frac{e^\theta}{(1+e^\theta)^2} = \pi(1-\pi) \qquad (\implies V(\mu) = \mu(1-\mu)) \end{aligned}$$

## 1.3 Link Functions

For regular linear models, we modelled the conditional expectation of the response $Y$ as a linear combination of the covariates and a parameter $\boldsymbol{\beta} \in \mathbb{R}^p$, also known as a **linear predictor**.

**Definition 2** (Linear Predictor)**.** A linear predictor $\eta$ is a linear combination of $p$ covariates and a finite parameter $\boldsymbol{\beta} \in \mathbb{R}^p$. We write:

$$\eta = \boldsymbol{x}^\top \boldsymbol{\beta} = \sum_{j=1}^{p} x_j \beta_j$$

A generalised linear model takes a more flexible form than the standard linear model, modelling the conditional expectation of $Y$ as a function of a linear predictor. More formally, we let $\mu(\boldsymbol{x}) = \mathbb{E}[\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}]$, with $Y$ an exponential-dispersion family, and supposing that $\mu$ depends on a finite parameter $\boldsymbol{\beta}$, writing $\mu = \mu(\boldsymbol{x}; \boldsymbol{\beta}) = \mathbb{E}[\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}]$. We generalise the relationship between $\mu$ and the linear predictor $\eta$ by introducing a **link function** $g$:

$$g(\mu) = \eta = \boldsymbol{x}^\top \boldsymbol{\beta} \tag{1.5}$$

Many link functions are available, and their selection is a modelling choice[2] [2].

**Example 3** (Logit Link)**.** Recalling from Example 2 that the Bernoulli distribution is an exponential family, consider $Y \sim \mathcal{B}(\pi)$. Unlike linear regression, where $\mu$ takes values over all $\mathbb{R}$, in this case we should restrict $\mu$ to the interval $(0, 1)$ since $0 < \mathbb{E}[Y] < 1$. First notice

[2]Notice that standard linear regression is now a special case of the GLM, where $g(\mu) = \mu$.

that $0 < \frac{e^t}{e^t+1} < 1 \; \forall t \in \mathbb{R}$. Now we write:

$$\mu = g^{-1}(\eta) = \frac{e^\eta}{e^\eta + 1} \in (0,1)$$
$$\iff g(\mu) = \log\left(\frac{\mu}{1-\mu}\right) = \eta$$

This particular link function $g(t) = \log\left(\frac{t}{1-t}\right)$ is termed the **logit link**, and the resulting model is known as **logistic regression**. When restricting to $(0,1)$, any CDF can be used, since by definition they map from their support to $(0,1)$. In particular, the **probit link** uses the inverse standard normal CDF, with $g(t) = \Phi^{-1}(t) \implies \mu = \Phi(\eta)$.

# Chapter 2

# Cumulative Probability Models

## 2.1 Binary Logistic Regression

Let $Y \sim \mathcal{B}(\pi)$. Noticing that $\mu = \mathbb{E}[\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}] = \pi$, we can express the logistic regression model as follows:

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \boldsymbol{x}^\top \boldsymbol{\beta} \tag{2.1}$$

The primary issue with logistic regression, and with GLMs overall, is that of interpretability. The estimated coefficients no longer have a linear relationship with the response variable, and their meaning can vary with the different link functions. For interpretation of logistic regression models, the definition of **odds** and by consequence **odds ratios** is helpful.

**Definition 3** (Odds)**.** For a given probability $p \in (0, 1)$, we define the odds $O$ as[1]:

$$O = \frac{p}{1-p}$$

This allows the model to be written in terms of (log-)odds:

$$\text{logit}(\pi) = \log(O) = \boldsymbol{x}^\top \boldsymbol{\beta} \tag{2.2}$$

It follows that a unit increase in predictor $X_j$ implies an increase in log-odds by $\beta_j$, or equivalently an increase in odds by a factor of $e^{\beta_j}$, known as the **odds ratio** [3]. To see this, we consider a unit increase in $X_j$, setting $\boldsymbol{x} := (x_1, \cdots, x_p)^\top$ and $\boldsymbol{x'} := (x_1, \cdots, x_j + 1, \cdots, x_p)^\top$. Now:

$$\begin{aligned}
&\log(O|\boldsymbol{x'}) = \boldsymbol{x'}\boldsymbol{\beta} \\
\implies &O|\boldsymbol{x'} = \exp(x_1\beta_1) \cdots \exp((x_j+1)\beta_j) \cdots \exp(x_p\beta_p) \\
&\qquad = \exp(\beta_j) \cdot \underbrace{\Pi_{i=1}^{p} \exp(x_i\beta_i)}_{=O|\boldsymbol{x}} \\
\implies &e^{\beta_j} = \frac{O|\boldsymbol{x'}}{O|\boldsymbol{x}}
\end{aligned}$$

[1]Note that this is a one-to-one mapping; from the odds we can recover the probability and vice-versa.

## 2.2 Cumulative Probability Models

Cumulative Probability Models (CPMs) are a class of ordinal regression models which extend the above binary logistic regression model to ordinal response variables [4].

Letting $Y$ be an ordinal random variable with levels $\ell = 1, \cdots, L$, for a link function $g$ we write the CPM as:

$$g(\mathbb{P}(Y \leq \ell)) = \zeta_\ell - \boldsymbol{X}\boldsymbol{\beta} \qquad \ell = 1, \cdots, L-1 \tag{2.3}$$

We focus on the particular case with the logit link, referred to as a Proportional Odds Model (POM):

$$\log\left(\frac{\mathbb{P}(Y \leq \ell)}{\mathbb{P}(Y > \ell)}\right) = \zeta_\ell - \boldsymbol{X}\boldsymbol{\beta} \qquad \ell = 1, \cdots, L-1 \tag{2.4}$$

For each level $\ell$, the model has a different intercept $\zeta_\ell$, representing the log-odds of $Y \leq \ell$ when all predictors are 0. The interpretation is the similar to the binary logistic regression model; $e^{-\beta_j}$ is the odds ratio for $Y \leq \ell$ between unit differences in $X_j$.

### 2.2.1 Proportional Odds Assumption

Notice that the coefficients $\boldsymbol{\beta}$ are invariant with $\ell$. In other words, we assume that the coefficient associated with each predictor $X_j$ is the same for all $\ell$ [5]. Writing the odds for $Y \leq \ell$ as $\mathcal{O}_\ell := \frac{\mathbb{P}(Y \leq \ell)}{\mathbb{P}(Y > \ell)}$, we say:

**Definition 4** (Proportional Odds Assumption)**.** For any two predictor realisations $\boldsymbol{x} \neq \boldsymbol{x}' \in \mathbb{R}^p$, the resulting odds ratio is invariant with $\ell$:

$$\frac{\mathcal{O}_1|\boldsymbol{x}'}{\mathcal{O}_1|\boldsymbol{x}} = \frac{\mathcal{O}_2|\boldsymbol{x}'}{\mathcal{O}_2|\boldsymbol{x}} = \cdots = \frac{\mathcal{O}_{L-1}|\boldsymbol{x}'}{\mathcal{O}_{L-1}|\boldsymbol{x}}$$

An important point to consider is that we make no assumption on the particular distribution of $Y$, only that its distribution is such that the proportional odds assumption is satisfied[2].

## 2.3 Applications to Continuous Data

Continuous data can also be seen as ordinal [4], and as such ordinal regression models can be fit to continuous outcomes [3]. Although the transformation from continuous to ordinal data entails a loss of information, Liu et al. [4] studied applications of ordinal CPMs to continuous data, and proposed numerous advantages to this approach:

- CPMs only incorporate order information from response variables, rendering them invariant to monotonic transformations of outcomes and robust against outliers.
- CPMs directly model the CDF, allowing means and quantiles to be derived directly from the model.

[2]The GLM is still valid; occurrences of the event $Y \leq \ell$ are Bernoulli *under the proportional odds assumption.*

- Since only order information is used, CPMs can "handle any orderable response", such as mixed continuous and ordinal distributions.

Finally, the authors point out that alternative regression methods (for example the GLMs outlined in Chapter 1) usually require an assumption of the distribution of the outcome variable, or a transformation thereof. CPMs, as discussed in section 2.2, are not so rigid, and instead allow the outcome distribution to be estimated.

### 2.3.1 Motivation

We now review the motivation for this application as outlined by Liu et al. [4]. We first consider a more general property of the CPM.

**Example 4** (Equivalence of Linear Transformation Model with CPM)**.** We model an observed outcome variable $Y$ as a monotonic transformation of a latent variable $Y^*$ by an unknown, but strictly increasing, function $H$, writing $Y = H(Y^*)$. Furthermore, we model $Y^*$ as a linear combination of predictor variables $\boldsymbol{X}$ with parameters $\boldsymbol{\beta}$, with an error term $\epsilon \sim F_\epsilon$ ($F_\epsilon$ specified). We write the semi-parametric **linear transformation model**:

$$Y = H(\boldsymbol{X}\boldsymbol{\beta} + \epsilon) \tag{2.5}$$

Now consider the conditional CDF of $Y$:

$$\begin{aligned}\mathbb{P}(Y \leq y|\boldsymbol{X}) &= \mathbb{P}[H(\boldsymbol{X}\boldsymbol{\beta} + \epsilon) \leq y|\boldsymbol{X}] \\ &= \mathbb{P}[\epsilon < H^{-1}(y) - \boldsymbol{X}\boldsymbol{\beta}|\boldsymbol{X}] \\ &= F_\epsilon[H^{-1}(y) - \boldsymbol{X}\boldsymbol{\beta}]\end{aligned}$$

Letting $G := F_\epsilon^{-1}$ and $\zeta(y) = H^{-1}(y)$, we obtain:

$$G[\mathbb{P}(Y \leq y|\boldsymbol{X})] = \zeta(y) - \boldsymbol{X}\boldsymbol{\beta}$$

with $G$ a link function and $\zeta$ an intercept function. Considering the observed data $\{y_1, \cdots, y_n\}$, we recover a CPM as in equation (2.3):

$$G[\mathbb{P}(Y \leq y_i|\boldsymbol{X})] = \zeta(y_i) - \boldsymbol{X}\boldsymbol{\beta} \qquad i = 1, \cdots, n \tag{2.6}$$

This is a rather general result, but notice in particular that when $H$ is a step function, $Y$ is a discrete ordinal variable. So when seeking to apply an ordinal CPM to a continuous response, we can view $Y^*$ as the underlying continuous variable, from which we generate an ordinal response $Y$ by transforming with $H(\cdot)$.

**Example 5** (Ordinal Transformation and its Associated CPM)**.** An ordinal response $Y$ with $K$ levels (denoted $C_1, \cdots, C_K$ in ascending order) can be generated by a step function $H$

from the continuous response $Y^*$ as follows:

$$Y = H(Y^*) := C_j \text{ where } \alpha_{j-1} < Y^* \leq \alpha_j \qquad j = 1, \cdots, K \tag{2.7}$$

with $-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_K = \infty$, and we write its associated CPM as:

$$G[\mathbb{P}(Y \leq C_k | \boldsymbol{X})] = \zeta_k - \boldsymbol{X}\boldsymbol{\beta} \qquad k = 1, \cdots, K-1 \tag{2.8}$$

Given that $H$ is strictly increasing, we can then rewrite the model in a more interpretable way; with respect to the underlying continuous response $Y^*$:

$$G[\mathbb{P}(Y^* \leq \alpha_k | \boldsymbol{X})] = \zeta_k - \boldsymbol{X}\boldsymbol{\beta} \qquad k = 1, \cdots, K-1 \tag{2.9}$$

We will focus on the associated POM, allowing interpretation of coefficients in terms of log-odds and odds ratios:

$$\log\left(\frac{\mathbb{P}(Y^* \leq \alpha_k)}{\mathbb{P}(Y^* > \alpha_k)}\right) = \zeta_k - \boldsymbol{X}\boldsymbol{\beta} \qquad k = 1, \cdots, K-1 \tag{2.10}$$

To retain as much information as possible, it is common practice (as implemented in software) to set the sequence of thresholds $(\alpha_k)_{k \in [K]}$ to be the unique values of $Y$.

### 2.3.2 Undetected Data

The capacity or CPMs to deal with mixed types of ordinal and continuous distributions has useful applications to data with detection limits [4]. Such data occurs when measuring physical quantities where the measuring instrument has finite precision, so true values below a certain threshold (detection limit) are recorded as either missing or zeroes.

One existing technique for this problem involves categorising measurements as either 'undetectable' or 'detectable' and then fitting a logistic regression model, however this ignores the information from the distribution of observations above the detection limit. One can also impute the missing data with a particular value, however this makes assumptions about the values below the detection limit.

Alternatively, fitting a CPM does not make assumptions on the undetected measurements, whilst retaining the order information of the detected measurements. Setting $(\alpha_k)_{k\in[K]}$ to be the unique values of $Y$, we have $\alpha_1 = Y_{min}$, the lowest observed detected value of $Y$, thus mapping undetected values to the lowest level $C_1$, and then mapping the detected values to their respective levels $C_2, \cdots, C_{K-1}$.

# Chapter 3

# Case Study

## 3.1 Introduction to the Dataset

In this chapter we consider a use case for CPMs, using data collected on the presence of Trace Organic Contaminants (TrOCs) in Canadian lakes by Lahens et al. [6]. The authors investigated the presence of 54 separate contaminants representative of human activity in 290 different lakes across Canada[1]. The sampled lakes were in a wide range of locations, so the authors also collected data on surrounding land use, presence of waste-water treatment plants (WWTPs), lake depth, watershed slope and precipitation over the 7 day period prior to taking the measurements.

The total detected concentrations of TrOCs ranged from 0.23 ng/L to 2.22 $\mu$g/L. Within the available dataset, 32.7% of lakes were recorded as having no contaminants. Detection

[1]Only 284 available in the dataset.

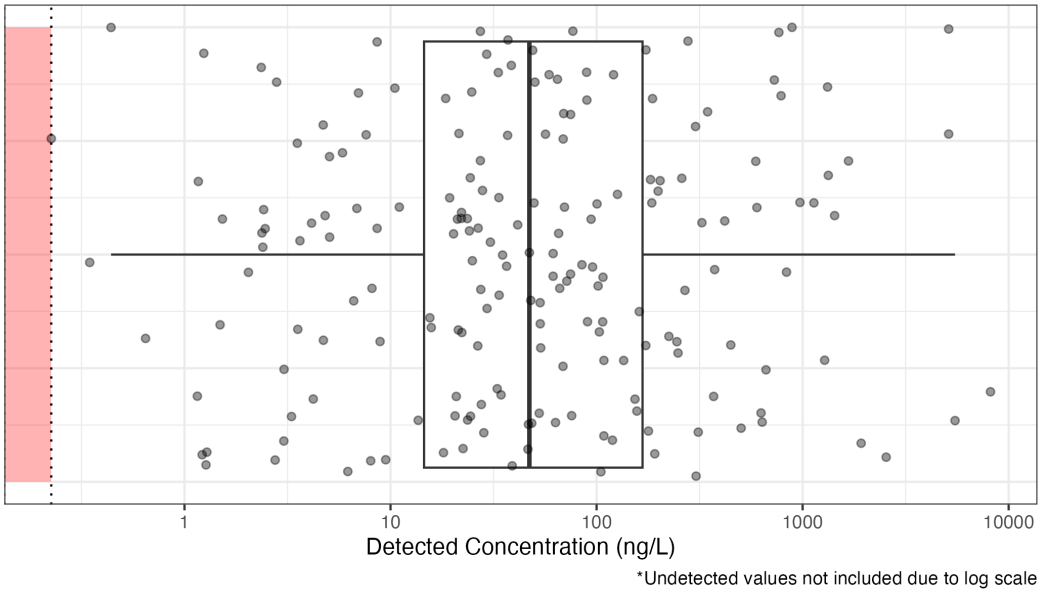


**Figure 3.1:** Observed Concentrations

limits are clearly a concern; it is likely that the measurements of 0 ng/L have a true value in the range $[0, 0.23)$ ng/L, and given their high prevalence in the dataset it is important to model them effectively. The hypothesised undetected range is highlighted in red in figure 3.1. As recommended by Liu et al. [4] and discussed in section 2.3.2, a CPM is useful in this case.

## 3.2 Modelling Contamination Levels with a POM

We seek to model the impact of land use and presence of WWTPs on the concentration of the measured TrOCs, using lake depth, watershed slope, precipitation, residence time, area ratio between a lake and its watershed, and day of the year as control variables. To make use

of as much information as possible, whilst accounting for the impact of undetected data, we can fit a CPM using the logit link, known as a Proportional Odds Model. The implemented model will have 191 separate coefficients for the 192 unique values of $Y$ (concentration of contaminants).

**Table 3.1:** Summary Table

| | |
|---|---|
| **Contaminant Density (ng/L)** | |
| Mean (SD) | 201 (770) |
| Median [Min, Max] | 15.6 [0, 8160] |
| **WWTP Presence** | |
| Yes | 264 (93%) |
| No | 20 (7%) |
| **Urban Land Use** | |
| Mean | 7.81% |
| Median [Min, Max] | 1.93% [0%, 82.6%] |
| **Agricultural Land Use** | |
| Mean | 8.05% |
| Median [Min, Max] | 0% [0%, 87.5%] |
| **Area Ratio** | |
| Mean (SD) | 0.0998 (0.104) |
| Median [Min, Max] | 0.064 [0, 0.605] |
| **Lake Depth (m)** | |
| Mean (SD) | 13.8 (17.4) |
| Median [Min, Max] | 7.95 [0.65, 138] |
| **Mean Watershed Slope (% Rise)** | |
| Mean (SD) | 8.48 (10.1) |
| Median [Min, Max] | 5.42 [0.0662, 62.4] |
| **Residence Time (days)** | |
| Mean (SD) | 2880 (8050) |
| Median [Min, Max] | 403 [0.1, 45500] |
| **7-day Precipitation (cm)** | |
| Mean (SD) | 2.05 (1.63) |
| Median [Min, Max] | 1.73 [0.002, 8.04] |

**Table 3.2:** Log-Odds Coefficient Estimates (95% CI)

| | Estimate | S.E. | Wald Z | Lower CI | Upper CI | p |
|---|---|---|---|---|---|---|
| WWTP Presence | 1.94 | 0.41 | 4.74 | 1.14 | 2.74 | <0.0001 |
| Urban Prop. | 4.58 | 0.76 | 5.99 | 3.08 | 6.07 | <0.0001 |
| Agriculture Prop. | 2.34 | 0.67 | 3.49 | 1.03 | 3.66 | 0.0005 |
| Area Ratio | -1.14 | 1.24 | -0.92 | -3.57 | 1.29 | 0.3583 |
| log(Lake Depth) | -0.31 | 0.13 | -2.43 | -0.57 | -0.06 | 0.015 |
| log(Mean Slope) | -0.28 | 0.15 | -1.91 | -0.56 | 0.01 | 0.0562 |
| log(Residence Time) | 0.01 | 0.06 | 0.13 | -0.11 | 0.13 | 0.8968 |
| Precipitation | 0.24 | 0.07 | 3.49 | 0.10 | 0.37 | 0.0005 |
| Day | 0.01 | 0.01 | 1.65 | 0.00 | 0.02 | 0.0991 |

Summary statistics for each variable are shown in table 3.1, and plots are included in appendix A.1. Since some variables exhibit a highly skewed distribution (see figures A.3, A.5 and A.6), a log-transform was applied to lake depth, watershed slope and residence time.

Due to absence of discharge, the authors were unable to record residence time for 6 lakes, for which a value of 45500 days (corresponding to the highest observed value in the dataset) was imputed.

The Proportional Odds Model was fitted using the `orm` function from the `rms` package [7] in *R*, yielding the output in table 3.2, with the corresponding odds ratios reported in table 3.3.

### Interpretation

Due to the implementation in software, the estimated coefficients correspond to exceedance probabilities; that is the log-odds of $Y \geq y$, and also are inversed in sign. The reformulated

model as implemented in software is written in Equation (3.1).

$$\log\left(\frac{\mathbb{P}(Y \geq \ell)}{\mathbb{P}(Y < \ell)}\right) = \zeta_\ell + \boldsymbol{X}\boldsymbol{\beta} \qquad \ell = 1, \cdots, L-1 \tag{3.1}$$

The simplest example is the presence of WWTPs, being a binary variable. The model estimates that the presence of a WWTP increases the log-odds of $Y \geq y$ by 1.94. In other words, the odds of exceedance increase by a factor (odds ratio) of $e^{1.94} \approx 6.95$. For the land use variables, the model estimates that an increase of 10% in urban land use (e.g. 30% to 40%) increases the odds of exceedance by a factor of 1.58, compared to an odds ratio of 1.26 for an identical increase in agricultural land use. All three predictor variables are statistically significant.

**Table 3.3:** Odds Ratio Estimates (95% CI)

| | Estimate | Lower CI | Upper CI |
|---|---|---|---|
| WWTP Presence | 6.95 | 3.12 | 15.49 |
| Urban Prop. | 97.13 | 21.73 | 434.19 |
| Agriculture Prop. | 10.43 | 2.80 | 38.84 |
| Area Ratio | 0.32 | 0.03 | 3.64 |
| log(Lake Depth) | 0.73 | 0.57 | 0.94 |
| log(Mean Slope) | 0.76 | 0.57 | 1.01 |
| log(Residence Time) | 1.01 | 0.89 | 1.14 |
| Precipitation | 1.27 | 1.11 | 1.45 |
| Day | 1.01 | 1.00 | 1.02 |

For a binary logistic regression, the odds ratio represents the change in the odds of some *outcome*, whereas for a for an ordinal POM the odds ratio represents a change in odds of *more*

*severe levels* of the outcome [5]. When applied to continuous data, an odds ratio greater than 1 represents a greater probability of a higher outcome (in this case contamination levels), and due to the proportional odds assumption, this increase is assumed to be constant across all outcome values. For example, we see that the presence of a WWTP increases the odds of contamination levels greater than 0.5 ng/L by a factor of 1.94, but the same factor applies to odds of contamination levels greater than 2 $\mu$g/L. This assumption is difficult to verify and is indeed unlikely to hold for extreme levels of the outcome variable [5].

## 3.3 Diagnostics

Shepherd et al. developed 'probability-scale residuals' [8] for diagnostics of GLMs where the "expectation of the fitted distribution cannot be calculated", demonstrating in particular their utility for ordinal CPMs. Generally for well-specified models, probability-scale residuals are uniformly distributed, however this is not the case if the outcome distribution is a mixture of discrete and continuous distributions, as is the case in this example [4]. However, they do still have expectation 0 under properly specified models, so residual-by-predictor plots (as in figure 3.2) are useful for visual diagnostics [4][2]. Residual plots for the remaining predictors are included in Appendix A.2.

In their simulations, Liu et al. [4] found that CPMs and logistic regression models

[2]Probability-scale residuals can be easily calculated with the `PResiduals` package [9] in *R*. See Appendix B.2 for implementation examples with a CPM.

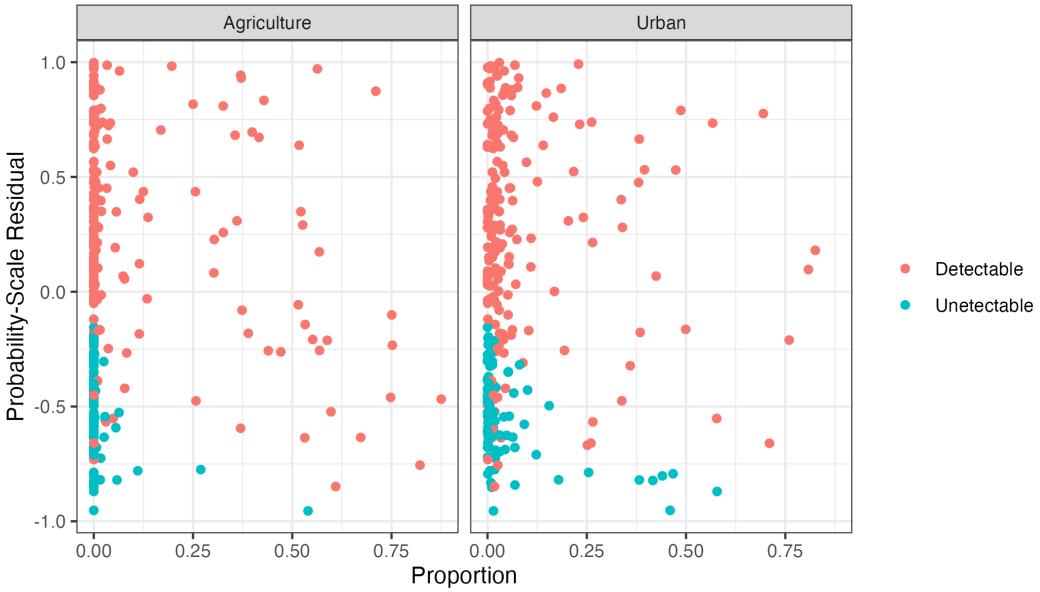


**Figure 3.2:** Residual-by-Predictor Plot for Land Use

perform very similarly when estimating the probability of detection, however CPMs exhibited narrower confidence intervals when estimating the probability of exceedance of a given threshold (above the detection limit). The authors note that the gains in efficiency are reduced as the proportion of undetectable measurements increases.

# Bibliography

[1] B. E. Hansen, "A modern gauss–markov theorem," *Econometrica*, vol. 90, no. 3, pp. 1283–1294, 2022. DOI: https://doi.org/10.3982/ECTA19255.

[2] L. Raymond-Belzile, "Course notes for math 523 - generalised linear models," McGill University, 2022.

[3] F. E. Harrell Jr, *Regression Modelling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*, 2nd ed. Springer, 2015.

[4] Q. Liu, B. E. Shepherd, C. Li, and F. E. Harrell Jr, "Modeling continuous response variables using ordinal regression," *Stat Med.*, 2017.

[5] R. W. Nahhas, *Introduction to Regression Methods for Public Health Using R*. Routledge & CRC Press, 2024.

[6] L. Lahens, H. Cabana, Y. Huot, and P. A. Segura, "Trace organic contaminants in lake waters: Occurrence and environmental risk assessment at the national scale in

canada," *Environmental Pollution*, vol. 347, p. 123 764, 2024, ISSN: 0269-7491. DOI: https://doi.org/10.1016/j.envpol.2024.123764.

[7] F. E. Harrell Jr, *Rms: Regression modeling strategies*, R package version 6.8-2, 2024. [Online]. Available: https://CRAN.R-project.org/package=rms.

[8] B. E. Shepherd, C. Li, and Q. Liu, "Probability-scale residuals for continuous, discrete, and censored data," *Can J Stat.*, 2016.

[9] Q. Liu, B. Shepherd, and C. Li, "PResiduals: An R package for residual analysis using probability-scale residuals," *Journal of Statistical Software*, vol. 94, no. 12, pp. 1–27, 2020. DOI: 10.18637/jss.v094.i12.

# Appendix A

# Supplementary Plots

## A.1 Summary Plots

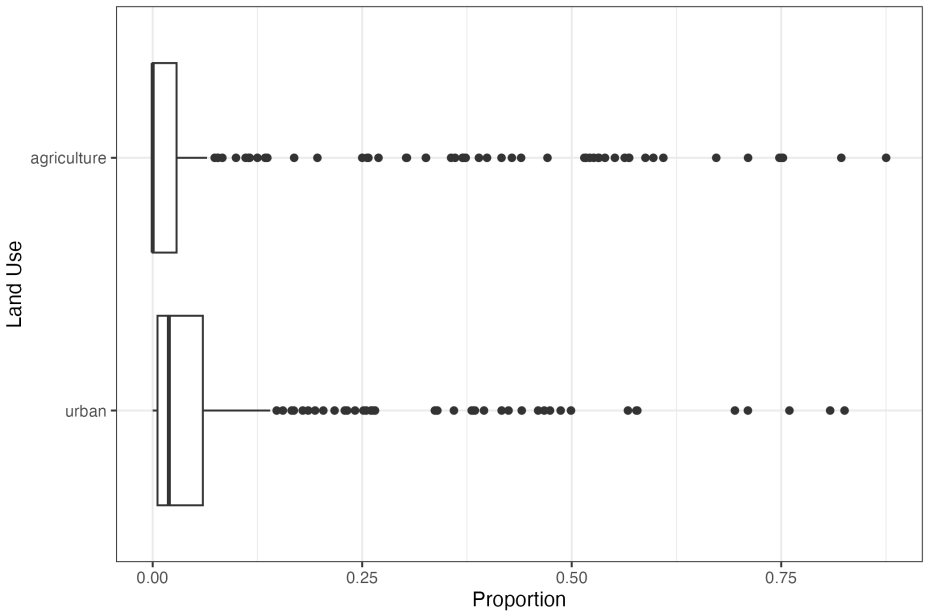


**Figure A.1:** Land Use Summary

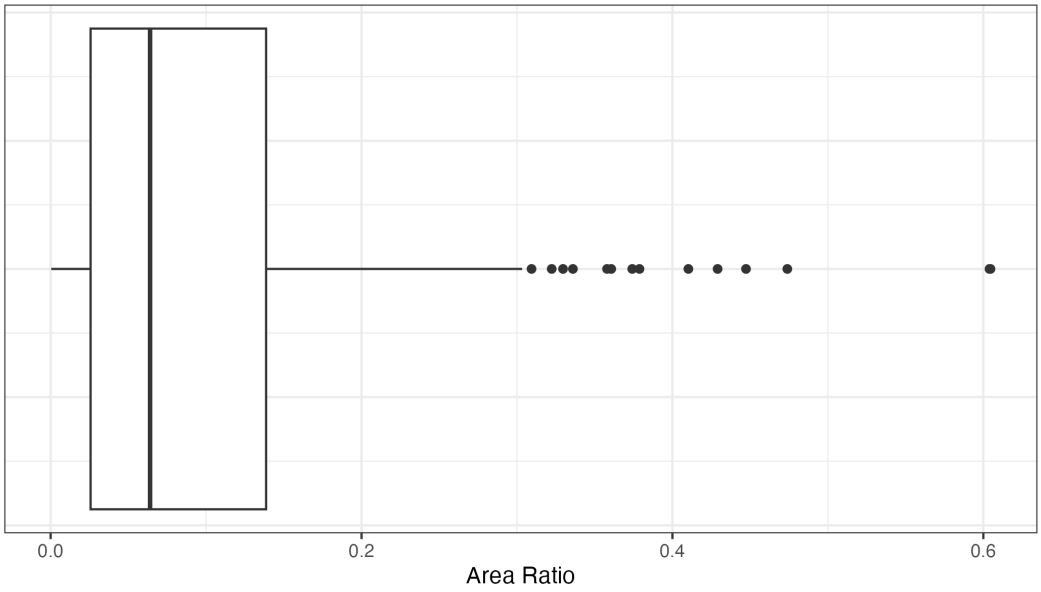


**Figure A.2:** Area Ratio Summary

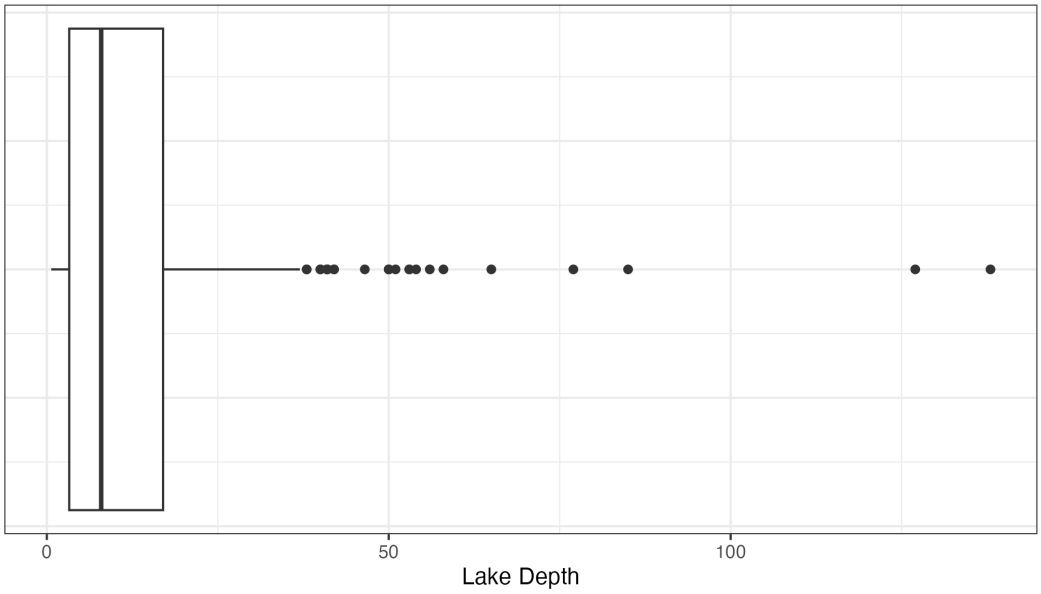


**Figure A.3:** Lake Depth Summary

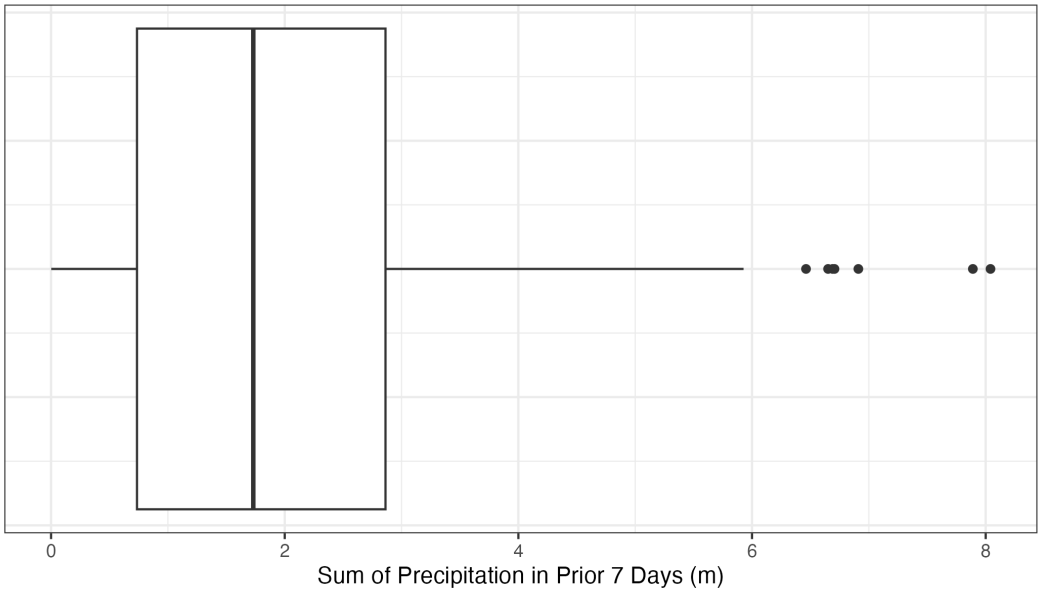


**Figure A.4:** Precipitation of Preceding 7 Days Summary

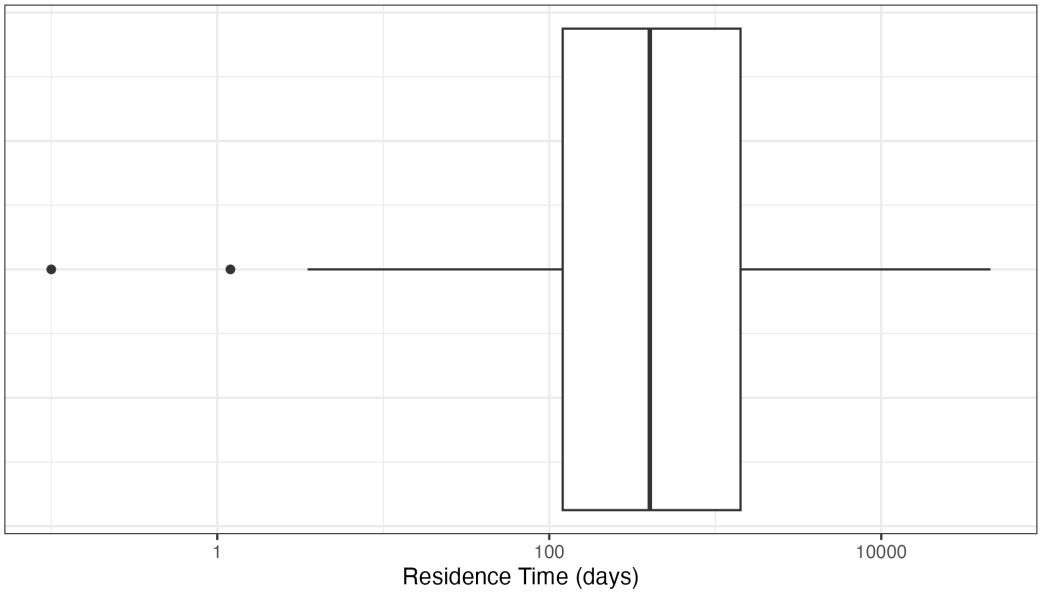


**Figure A.5:** Residence Time Summary

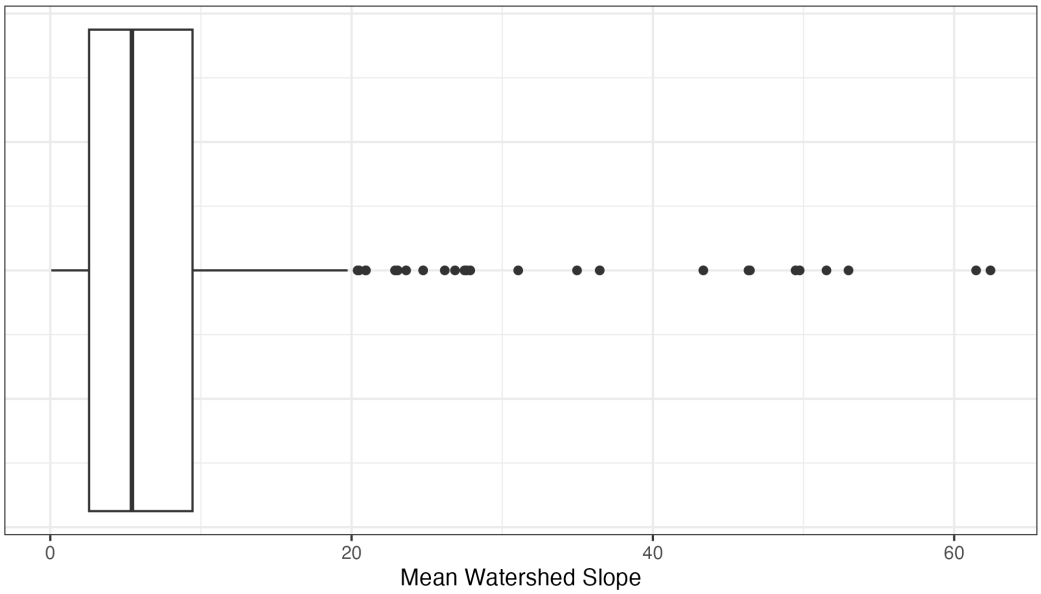


**Figure A.6:** Mean Watershed Slope Summary

## A.2 Residual-by-Predictor Plots for TrOC Data

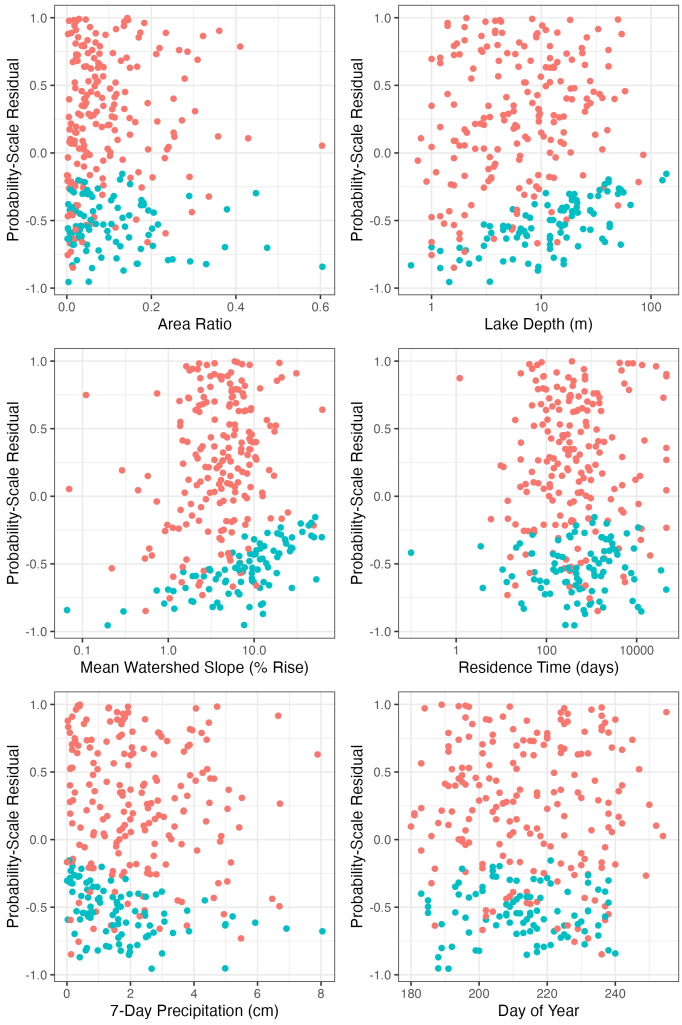


**Figure A.7:** Residual-by-Predictor Plots for fitted POM

# Appendix B

# *R* Code for Chapter 3

## B.1 Fitting a CPM

```
library(rms) # see https://cran.r-project.org/package=rms
orm.logit <- orm(sum_overall ~ wwtps_pres + urban_frac + agriculture_frac +
                 area_ratio + log(lake_depth) + log(mean_slope) + log(res_time)
               + precipM + day , data = data, x=TRUE, y=TRUE)
print(orm.logit) # View model output


# Confidence Intervals
ci <- confint(orm.logit)
confidence_intervals <- tibble(variable = rownames(ci),
```

```
    estimate = orm.logit$coefficients,
    lower_bound = ci[,1],
    upper_bound = ci[,2]
)

# Odds Ratios
odds_ratios <- confidence_intervals %>%
    mutate(odds_ratio = exp(estimate),
        lower_or_ci = exp(lower_bound),
        upper_or_ci = exp(upper_bound)) %>%
    select(variable, odds_ratio, lower_or_ci, upper_or_ci)
```

## B.2 Probability-Scale Residuals

```
library(PResiduals) # see https://cran.r-project.org/package=PResiduals
residuals <- presid(orm.logit)
data <- data %>% mutate(resid = residuals)
```