MA50260 Statistical Modelling

Lecture 13: GLM Diagnostics and Ordinal Regression

Ilaria Bussoli

March 19, 2024

Varying Exposure and Overdispersion

A Poisson regression model assumes

1. Observations are coming from equivalent populations,
2. The mean and variance are the same.

Is this reasonable in all situations?

Varying Exposure and Overdispersion

A Poisson regression model assumes

1. Observations are coming from equivalent populations,
2. The mean and variance are the same.

Is this reasonable in all situations? ***NO!***

1. Population sizes may differ (varying exposure),
2. The mean and variance may be different.

Varying Exposure and Overdispersion

A Poisson regression model assumes

1. Observations are coming from equivalent populations,
2. The mean and variance are the same.

Is this reasonable in all situations? ***NO!***

1. Population sizes may differ (varying exposure),
2. The mean and variance may be different.

We can address these aspects by

1. Including an offset term,
2. Using a quasi-Poisson or negative-binomial model.

GLM Diagnostics - Residuals

There are two types of residuals:

Pearson Residuals

$$r_i^P = \frac{(y_i - \hat{\mu}_i)}{\sqrt{V(\hat{\mu}_i)}},$$

with zero mean and variance ϕ.

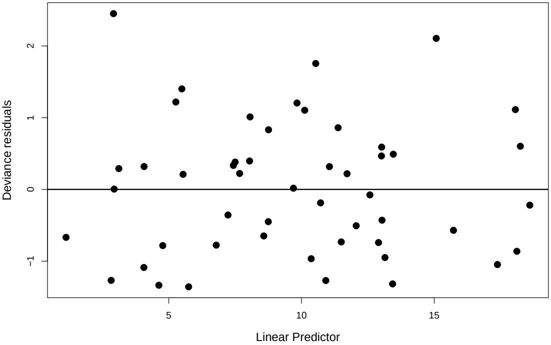
Deviance Residuals

$$r_i^D = \sqrt{D_i}\,\mathrm{sign}(y_i - \hat{\mu}_i).$$

To assess model fit, we compare the residuals to the Normal$(0, \phi)$ distribution, in particular, if ϕ is known.

GLM Diagnostics - Plotting Residuals

- For non-normal GLMs, the deviance residuals as a set are more nearly normal than the Pearson's residuals.
- The residuals should not display any trend in mean or variance when plotted against the fitted values, or the explanatory variables.



Leverage and Influence

Leverages and influence can be defined similarly to the Normal linear model case.

The hat matrix is now defined as

$$\mathbf{H} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{1/2},$$

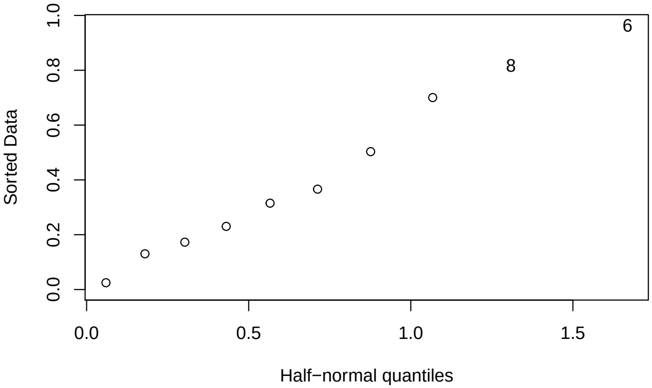
where $\mathbf{W}$ is the diagonal matrix in the IRWLS approach.

We can also again examine sensitivity of the model via Cook's distance.

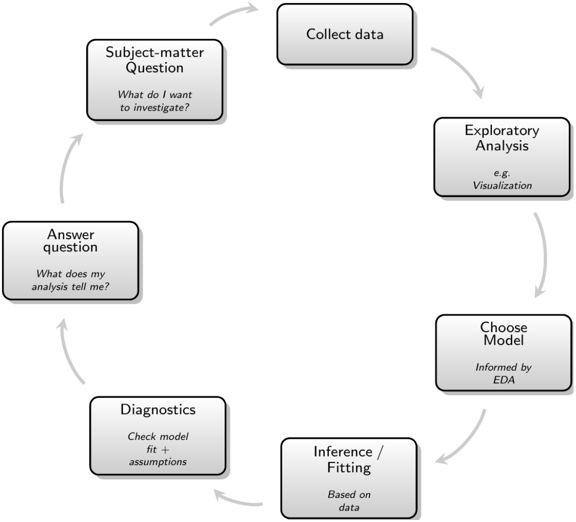
Outlier Detection

Half-normal quantile plots can be used to look for outliers.

These plots examine a sorted set of (positive) model quantities against the quantiles of the half-normal distribution.

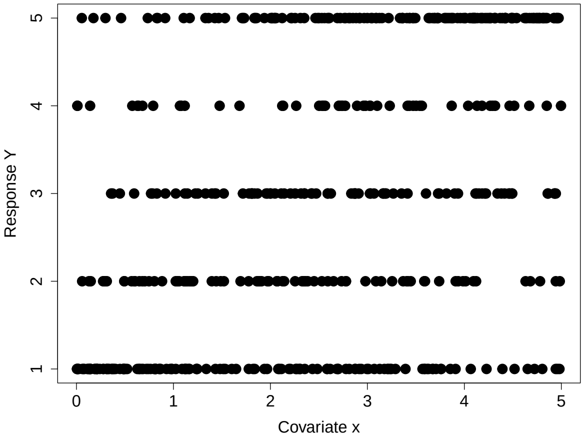


What have we achieved?



Modelling Categorical Variables (I)

Suppose we have a categorical variable Y, whose levels have a natural ordering.



Modelling Categorical Variables (II)

We use a **categorical distribution** to model Y.

If Y has K levels, then

$$\mathbb{P}(Y = k) = p_k \qquad (k = 1, \ldots, K),$$

with $p_1 + \cdots + p_K = 1$.

Modelling Categorical Variables (II)

We use a **categorical distribution** to model Y.

If Y has K levels, then

$$\mathbb{P}(Y = k) = p_k \qquad (k = 1, \ldots, K),$$

with $p_1 + \cdots + p_K = 1$.

We thus have to model $p_1, \ldots, p_{K-1}$ conditional on $\mathbf{x}$.

This framework is called **ordinal regression**.

The Ordinal Logistic Regression Model (I)

Let $F_k = p_1 + p_2 + \cdots + p_k = \mathbb{P}(Y \leq k)$.

The Ordinal Logistic Regression Model (I)

Let $F_k = p_1 + p_2 + \cdots + p_k = \mathbb{P}(Y \leq k)$.

Then we can define a logistic regression model for F_k with

$$\log\left(\frac{F_k}{1 - F_k}\right) = \eta,$$

where $\eta = \mathbf{x}^T \underline{\beta}$.

The Ordinal Logistic Regression Model (I)

Let $F_k = p_1 + p_2 + \cdots + p_k = \mathbb{P}(Y \leq k)$.

Then we can define a logistic regression model for F_k with

$$\log\left(\frac{F_k}{1 - F_k}\right) = \eta,$$

where $\eta = \mathbf{x}^T \underline{\beta}$.

We could consider each level $k = 1, \ldots, K - 1$ and estimate separate models for $F_1, \ldots, F_{k-1}$.

The Ordinal Logistic Regression Model (II)

Then

$$p_k = \mathbb{P}(Y \leq k) - \mathbb{P}(Y \leq k-1) = F_k - F_{k-1} \qquad (k = 1, \ldots, K),$$

where $F_0 = 0$ and $F_K = 1$.

The Ordinal Logistic Regression Model (II)

Then

$$p_k = \mathbb{P}(Y \leq k) - \mathbb{P}(Y \leq k-1) = F_k - F_{k-1} \qquad (k = 1, \ldots, K),$$

where $F_0 = 0$ and $F_K = 1$.

Lines may cross $\Rightarrow$ Contradiction

Instead, we define the ordinal logistic regression model

$$\log\left(\frac{F_k}{1 - F_k}\right) = \alpha_k + \mathbf{x}^{\mathrm{T}}\underline{\beta},$$

where $\alpha_1 < \alpha_2 < \cdots < \alpha_{K-1}$.

The Ordinal Logistic Regression Model (II)

Then

$$p_k = \mathbb{P}(Y \leq k) - \mathbb{P}(Y \leq k-1) = F_k - F_{k-1} \qquad (k = 1, \ldots, K),$$

where $F_0 = 0$ and $F_K = 1$.

Lines may cross $\Rightarrow$ Contradiction

Instead, we define the ordinal logistic regression model

$$\log\left(\frac{F_k}{1-F_k}\right) = \alpha_k + \mathbf{x}^{\mathrm{T}}\underline{\beta},$$

where $\alpha_1 < \alpha_2 < \cdots < \alpha_{K-1}$.

This model requires the proportional odds assumption.

Example

For the data considered at the beginning, we estimate

```
## Call:
## polr(formula = y ~ x, Hess = TRUE)
##
## Coefficients:
##    Value Std. Error t value
## x 0.5834     0.0613   9.518
##
## Intercepts:
##     Value   Std. Error t value
## 1|2  0.3663  0.1682     2.1781
## 2|3  1.1897  0.1752     6.7902
## 3|4  1.9744  0.1892    10.4370
## 4|5  2.4996  0.2004    12.4723
##
## Residual Deviance: 1442.982
## AIC: 1452.982
```