# ECON3389 Machine Learning in Economics

## Module 2.2
## Classification

Alberto Cappello

Department of Economics, Boston College

Fall 2024

# Overview

Agenda:

- Poisson Regression on count data
- Discriminant Analysis

Readings:

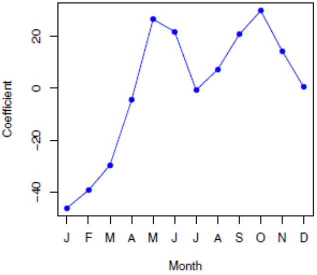
- ISLR sections 4.1, 4.2, 4.3

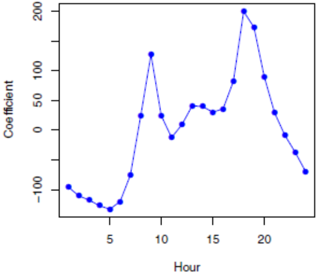
# Count data

- So far our outcome variable $Y$ has always been assumed to be either quantitative (continuous) or qualitative
- But now we want to quantify a relationship where the outcome $Y$ is a count of events?
  - The number of hourly users of a bike sharing program in Washington, DC
  - The number of phone calls I get in a day
  - The number of dinner customers at a restuarant on a Tuesday evening
- The outcome variable in count data only takes on non-negative integer values. It is the count of events within a specified interval of time
- *Bikeshare Data*
  - **Y:** Number of hourly users of a bike sharing program in Washington D.C.
  - **X:** month of the year, hour of the day, workingday (1 if it is neither a weekend nor a holiday), temperature, weather (clear, misty, light rain, heavy rain)

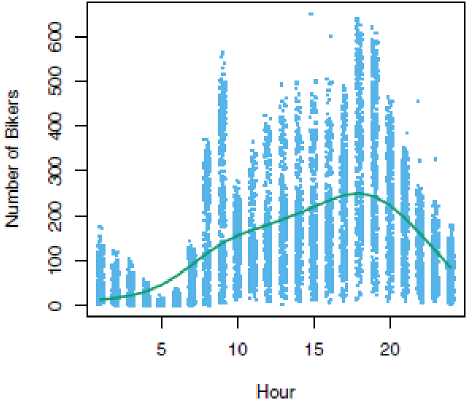
# Basic Linear Regression

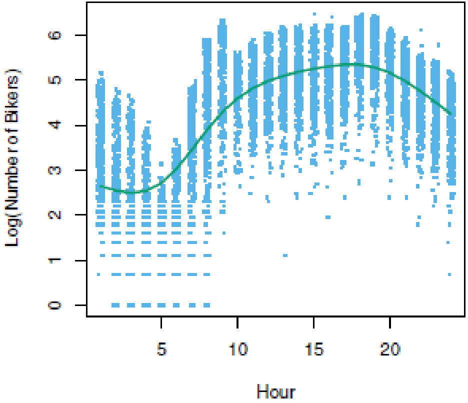
| | Coefficient | Std. error | $z$-statistic | $p$-value |
|---|---|---|---|---|
| `Intercept` | 73.60 | 5.13 | 14.34 | 0.00 |
| `workingday` | 1.27 | 1.78 | 0.71 | 0.48 |
| `temp` | 157.21 | 10.26 | 15.32 | 0.00 |
| `weathersit[cloudy/misty]` | -12.89 | 1.96 | -6.56 | 0.00 |
| `weathersit[light rain/snow]` | -66.49 | 2.97 | -22.43 | 0.00 |
| `weathersit[heavy rain/snow]` | -109.75 | 76.67 | -1.43 | 0.15 |



# Higher Mean and Variance



# Problems with Linear regression

- **Problem 1:** 9.6% of the predicted values are negative
- **Problem 2:** The linear model assumes

$$Y = X\beta + \epsilon$$
$$E(Y|X) = X\beta$$

  where $\epsilon$ is mean 0 and variance $\sigma^2$ (constant)
- However, in the data $\sigma^2$ should be a function of $X$
- **Problem 3:** The response variable ($Y$) is integer-valued. But under a linear model, the predicted Y is necessarily continuous valued. Thus, the integer nature of the response *bikers* suggests that a linear regression model is not entirely satisfactory for this data set.

# Poisson Distribution

- Suppose the random variable $Y$ takes on nonnegative integer values, $Y \in \{0, 1, 2, 3..\}$. If $Y$ follows the Poisson distribution, then

$$Pr(Y = k) = \frac{\exp(-\lambda).\lambda^k}{k!} \quad \text{for k} = \{0,\ 1,\ 2,\ 3..\}$$
$$\lambda > 0$$
$$E(Y) = Var(Y) = \lambda$$

- If Y follows the Poisson distribution, then the larger the mean of Y , the larger is its variance
- The Poisson distribution is typically used to model counts. This is a natural choice because counts, like the Poisson distribution, take on nonnegative integer values
- Occurrence of each individual event is independent of each other

# Poisson Regression: Idea

- Let Y denote the number of users of the bike sharing program during a particular hour of the day, under a particular set of weather conditions, and during a particular month of the year
- Suppose $E(Y) = \lambda = 5$
- $Pr(Y = 0) = \frac{\exp(-5)5^0}{0!} = 0.0067$
- $Pr(Y = 2) = \frac{\exp(-5)5^2}{2!} = 0.084$
- We want the mean number of users of the bike sharing program ($\lambda$) to vary as a function of the hour of the day, the month of the year, the weather conditions
- Modeling the rate parameter

$$\ln(\lambda(X_1, X_2, ..., X_p)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$
$$\lambda(X_1, X_2, ..., X_p) = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)$$

- This ensures that the rate parameter is always positive

# Poisson Regression Model

- Poisson Regression model

$$Pr(Y = k|X_1, X_2, ...X_p) = \frac{\exp(-\lambda(X_1, X_2, ...X_p)).\lambda(X_1, X_2, ...X_p)^k}{k!} \quad \text{for k} = (0,1,2,...)$$

$$\ln(\lambda(X_1, X_2, ..., X_p)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

$$\lambda(X_1, X_2, ..., X_p) = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)$$

- We estimate the parameters of the Poisson regression model using MLE
- Remember, each observation in the data is the number of bikeshare users in a particular day of the year ($Y$) which has some covariates ($X$) like day of the week and weather on that day
- We maximize the model predicted likelihood of observing the $Y = k$ value in our data for each data point as a function of the covariates

$$\ell(\beta_0, \beta_1, \beta_2, ..\beta_p) = \prod_{i=1}^{n} Pr(Y_i = k|x_1, x_2, ...x_p) = \prod_{i=1}^{n} \frac{\exp(-\lambda(x_1, x_2, ...x_p))\lambda(x_1, x_2, ...x_p)^k}{k!}$$

# Poisson Regression Model: Interpretation

| | Coefficient | Std. error | $z$-statistic | $p$-value |
|---|---|---|---|---|
| `Intercept` | 4.12 | 0.01 | 683.96 | 0.00 |
| `workingday` | 0.01 | 0.00 | 7.5 | 0.00 |
| `temp` | 0.79 | 0.01 | 68.43 | 0.00 |
| `weathersit[cloudy/misty]` | -0.08 | 0.00 | -34.53 | 0.00 |
| `weathersit[light rain/snow]` | -0.58 | 0.00 | -141.91 | 0.00 |
| `weathersit[heavy rain/snow]` | -0.93 | 0.17 | -5.55 | 0.00 |

- An increase in $X_j$ by one unit is associated with a change in $E(Y) = \lambda$ by a factor of $exp(\beta_j)$
- A change in weather from clear to cloudy skies is associated with a change in mean bike usage by a factor of $\exp(-0.08) = 0.923$
- On average, only 92.3% as many people will use bikes when it is cloudy relative to when it is clear
- If the weather worsens further and it begins to rain, then the mean bike usage will further change by a factor of $\exp(-0.5) = 0.607$, i.e. on average only 60.7% as many people will use bikes when it is rainy relative to when it is cloudy

# Poisson Regression Model: Characteristics

- Mean variance relationship : by modeling bike usage with a Poisson regression, we implicitly assume that mean bike usage in a given hour equals the variance of bike usage during that hour. Thus, the Poisson regression model is able to handle the mean-variance relationship seen in the *Bikeshare* data in a way that the linear regression model is not
- Negative predictions : There are no negative predictions using the Poisson regression model. This is because the Poisson model itself only allows for nonnegative values

# Discriminant analysis

- Logistic regression involves directly modeling $\Pr(Y = k|X = x)$, the conditional distribution of the response $Y$, given the predictor(s) X.
- In DA, we model the distribution of the predictors X separately in each of the response classes (i.e. for each value of Y)
- Suppose the qualitative response variable Y can take on K possible distinct and unordered values
- Instead of directly modeling $\Pr(Y|X)$, model the distribution of $X$ in each of the $K$ classes separately, and then use *Bayes theorem* to flip things around and obtain $\Pr(Y|X)$:

$$\Pr(Y = k|X = x) = \frac{Pr(Y = k)Pr(X = x|Y = k)}{\sum_{k=1}^{K} Pr(Y = k)Pr(X = x|Y = k)} = \frac{\pi_k f_k(x)}{\sum_{k=1}^{K} \pi_k f_k(x)}$$

  where $f_k(x) = \Pr(X = x|Y = k)$ is the *density* for X in class $k$ and $\pi_k = \Pr(Y = k)$ is the *prior* probability for class $k$.
- $\Pr(Y = k|X = x)$ is called the *posterior probability*

# Discriminant analysis

- $\pi_k = \Pr(Y = k)$ *(prior)* is generally easy to compute. In a random sample, we simply compute the fraction of the training observations that belong to the kth class
- Estimating $f_k(x)$ is much more challenging
- In LDA we only have one predictor $X$. We assume that $f_k(x)$ is normally distributed

$$f_k(x|\mu_k, \sigma_k) = \frac{1}{\sigma_k \sqrt{2\pi}} \exp(-\frac{(x-\mu_k)^2}{2\sigma_k^2}) \tag{1}$$

  where $\mu_k$ and $\sigma_k^2$ are the mean and variance parameters for the kth class
- Once we estimate the class-specific $\mu_k$ and $\sigma_k^2$, we can calculate the $\Pr(Y = k|X = x)$ and assign the class with the highest probability

# Power of Baye's Theorem

- Steve: Very shy and withdrawn, invariably helpful but with very little interest in people or in the world of reality. A meek and tidy soul, he has a need for order and structure, and a passion for detail
- Is Steve a librarian or a farmer?
- Additional information: The ratio of librarians to farmers in the US is 1:19
- Using Baye's Theorem

$$\Pr(librarian|meek) = \frac{Pr(librarian)Pr(meek|librarian)}{Pr(librarian)Pr(meek|librarian) + Pr(farmer)Pr(meek|farmer)}$$

- Putting in the values

$$\Pr(librarian|meek) = \frac{0.05 * 0.4}{0.05 * 0.4 + 0.95 * 0.1} = 17.3\%$$

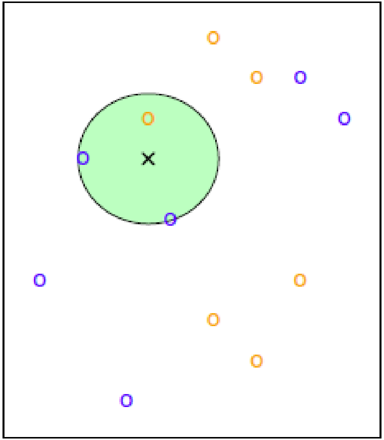
# K-nearest Neighbors

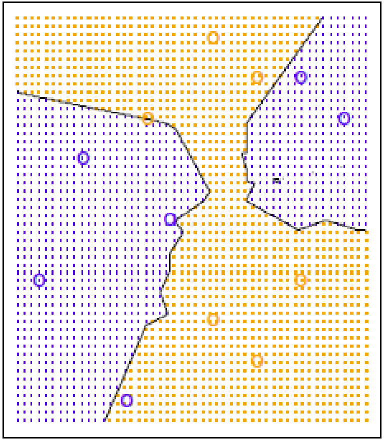
- For any given observation $x_0$ and a positive integer $K$, first identify $K$ points in the training data that are closest to $x_0$, denoted as $\mathcal{N}_0$. Then the conditional probability for class $j$ is calculated as

$$\Pr(Y = j|X = x_0) = \frac{1}{K}\sum_{i\in\mathcal{N}_0} \mathbb{I}(y_i = j)$$

  which is simply a fraction of points in $\mathcal{N}_0$ with response values equal to $j$.
- What happens to the bias vs variance trade-off as K goes up?
- KNN is an example of non-parametric supervised learning algorithm and as such places almost no restrictions on the nature of the data.
- It quickly loses its potency when the number of features in $X$ grows above 4-5 (too many points in high-dimensional space could be equally close to $x_0$). Thus it is often paired with other methods aimed at feature extraction and dimensionality reduction, such as *principal component analysis* (PCA).

# K-nearest Neighbors





# K-nearest Neighbors