General Notes

Darian S. Martos

GLM Notes

These notes are a compilation of some self-studying done for better understanding GLMs, primarily logistic and Poisson models. When time permits, these notes will also include notes on survival analysis and Bayesian analysis.

This will include various topics including derivations, estimation, and inference. Prediction is also obviously helpful as well, with a greater focus on diagnostics for GLM. The material is predominantly applied.

Resources used (these are hyperlinks):

- Dobson - An Introduction to Generalized Linear Models [IGLM]
- Faraway - Extending the Linear Model [ELM]
- Agresti - Categorical Data Analysis [CDA]

The first source is the "primary" one, which is the base for much of the outline and the initial coverage. The second is where much of my code and applied perspective comes from. The rest of the resources are supplemental and referenced occasionally for a more comprehensive view of the various topics.

Chapter 1. Introduction

Most of this chapter covers different variables and terminology (i.e. ordinal and nominal variables, etc.). **Nominal variables** are essentially categorical variables (male/female, dead/alive, etc.) while **ordinal variables** are categorical variables that have some ordering or ranking. This table is particularly useful in deciding what models or strategies to use as a baseline:

Response (Chapter)	Explanatory Variables	Methods
Continuous (Ch. 6)	Binary	t-test
	Nominal, > 2 categories	Analysis of Variance
	Ordinal	Analysis of Variance
	Continuous	Multiple Regression
	Nominal and some continuous	Analysis of Covariance
	Categorical and continuous	Multiple Regression
Binary (Ch. 7)	Categorical	Contingency Tables, Logistic Regression
	Continuous	Logistic, Probit and Other Dose-Response Models
	Categorical and continuous	Logistic Regression
Nominal with > 2 categories (Ch. 8-9)	Nominal	Contingency Tables
	Categorical and continuous	Nominal Logistic Regression
Ordinal (Ch. 8)	Categorical and continuous	Ordinal Logistic Regression
Counts (Ch. 9)	Categorical	Log-Linear Models
	Categorical and continuous	Poisson Regression

Other topics in this chapter included the coverage of the rest of IGLM, notation, and quadratic forms. Probability distributions related to the Normal and parameter estimation are reviewed, which I detail more.

Probability Distributions Related to the Normal

1. The Normal Distribution

   - For $Y \sim N(\mu, \sigma^2)$, **the Normal distribution** has density:

   $$f(y; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

   - $\mu = 0, \sigma^2 = 1$ gives the standard Normal $Y \sim N(0, 1)$.
   - Let $Y_1, \ldots, Y_n \sim N(\mu_i, \sigma_i^2)$ with covariance of Y_i, Y_j defined as:

   $$Cov(Y_i, Y_j) = \rho_{ij}\sigma_i\sigma_j$$

   Let $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_n]^\intercal$ and let the variance-covariance matrix $\mathbf{V}$ have $\rho_{ij}\sigma_i\sigma_j$ for $i \neq j$. Then we have the **multivariate Normal distribution**, with $\mathbf{y} \sim MVN(\boldsymbol{\mu}, \mathbf{V})$ and $\mathbf{y} = [Y_1, \ldots, Y_n]^\intercal$.

- Let $Y_1, \ldots, Y_n \sim N(\mu_i, \sigma_i^2)$ be independent, then:

$$W = \sum_{i=1}^{n} a_i Y_i \sim N(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2)$$

2. Chi-Squared Distribution (only some properties are listed here):

   - The **central Chi-Squared distribution** with n degrees of freedom is the sum of squares of n independent RVs $Z_1, \ldots, Z_n \sim N(0, 1)$. It is denoted by:

$$X^2 = \sum_{i=1}^{n} Z_i^2 \sim \chi^2(n)$$

   - If $X^2 \sim \chi^2(n)$, then $E[X^2] = n$ and $Var(X^2) = 2n$.
   - If $Y_1, \ldots, Y_n \sim N(\mu_i, \sigma_i^2)$ are independent, then:

$$X^2 = \sum_{i=1}^{n} \left(\frac{Y_i - \mu_i}{\sigma_i} \right)^2 \sim \chi^2(n)$$

   - Let $Z_1, \ldots, Z_n$ be iid with $Z_i \sim N(0, 1)$ and let $Y_i = Z_i + \mu_i$, with at least one of the μ_i's nonzero. Then the distribution of:

$$\sum_{i=1}^{n} Y_i^2 = \sum (Z_i + \mu_i)^2 = \sum Z_i^2 + 2 \sum Z_i \mu_i + \sum \mu_i^2$$

     has larger mean $n + \lambda$ and larger variance $2n + 4\lambda$ that $\chi^2(n)$ where $\lambda = \sum \mu_i^2$. This is the **noncentral Chi-Squared distribution** with n degrees of freedom and **noncentrality parameter** λ. It is denoted by $\chi^2(n, \lambda)$.

3. t-distribution

   The **t-distribution** with n degrees of freedom is the ration of two independent RVs:

$$T = \frac{Z}{(X^2/n)^{\frac{1}{2}}}$$

   With $Z \sim N(0, 1), X^2 \sim \chi^2(n), Z \perp X^2$.

   Denoted $T \sim t(n)$.

4. F-distribution

   - The **central F-distribution** with n, m degrees of freedom is defined as the ratio of two independent central Chi-Squared RVs:

$$F = \frac{X_1^2}{n} / \frac{X_2^2}{m}$$

With $X_1^2 \sim \chi^2(n), X_2^2 \sim \chi^2(m), X_1^2 \perp X_2^2$.
Denoted $F \sim F(n, m)$.

- We also have that:

$$T^2 = \frac{Z^2}{1} / \frac{X^2}{n} \sim F(1, n)$$

- The **non-central F distribution** is defined as the ratio of two independent RVs, each divided by its degrees of freedom, with:

$$F = \frac{X_1^2}{n} / \frac{X_2^2}{m}$$

With $X_1^2 \sim \chi^2(n, \lambda), \lambda = \boldsymbol{\lambda}^\intercal \mathbf{V}^{-1} \boldsymbol{\mu}, X_2^2 \sim \chi^2(m), X_1^2 \perp X_2^2$.
The mean of a noncentral F-distribution is larger than the mean of central F-distribution with the same degrees of freedom.

Chapter 2.

Chapter 3.

Chapter 4.