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]
- Faraway - Linear Models with R [LM]

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, quadratic forms, and maximum likelihood estimation and least squares. Probability distributions related to the Normal , 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:

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

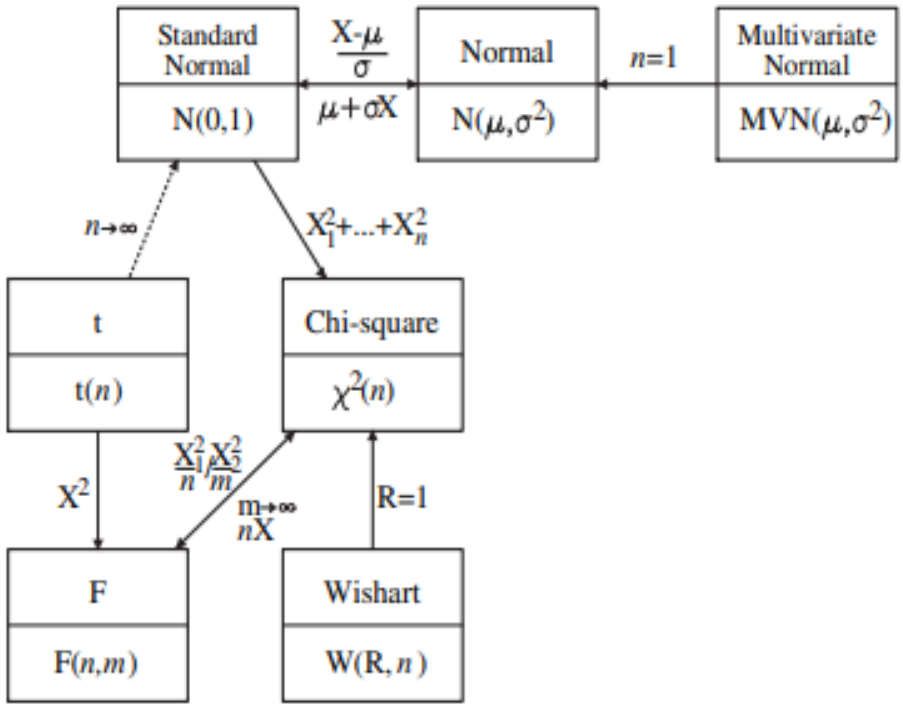
- The **non-central F distribution** is defined as the ratio of two independent RVs, each divided by its degrees of freedom, with:

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

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.

A nice graphic summarizing the relationships of these distributions comes from page 10:

We summarize the above relationships in Figure 1.1. In later chapters we add to this diagram and a more extensive diagram involving most of the distributions used in this book is given in the Appendix. Asymptotic relationships are shown using dotted lines and transformations using solid lines. For more details see Leemis (1986) from which this diagram was developed.



Chapter 2. Model Fitting

This is a fairly short chapter, it provides some exmaples related to statistical modeling and some principles here and there. A quick summary and review of topics will be discussed here.

- Chi-square goodness of fit statistic:

$$X^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$

- Solution to the **Normal equations** for model $E(Y_{jk}) = \alpha_j + \beta_j x_{jk}$:

$$\hat{\beta}_j = \frac{K \sum_k x_{jk} y_{jk} - (\sum_k x_{jk})(\sum_k y_{jk})}{K \sum_k x_{jk}^2 - (\sum_k x_{jk})^2}, \; \hat{\alpha}_j = \overline{y}_j - \hat{\beta}_j \overline{x}_j$$

- Modeling Principles:
  - Look at the scale and type of measurement (continuous vs categorical). Look at the shape of the distribution with graphical methods: frequency tables, dot plots, histograms, etc. Look at association between variables, using summaries such as scatter plots or box plots grouped by factors.
  - Models have two components (here a univariate response is assumed): a probability distribution of Y with $Y \sim \mathbb{D}(\mu, \sigma^2)$, and an equation linkiing the expected value of Y with a linear combination of the explanatory variables:

    $$g[E(Y)] = \beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m$$

    Where the above equation is known as the **linear component**.
  - Parameter estimation is then done, usually through MLE or least squares in the Frequentist perspective, or through alternative Bayesian methods.
  - Residuals important for checking assumptions of model fits, should assume residuals ε follow distribution $\varepsilon \sim N(0, \sigma^2)$ where $\sigma^2 \in \mathbb{R}$. You also want to check for normality using a **Normal probability plot**. You also want to fit residual against:
    * Explanatory variables - this should determine whether or not there is patterns in the residuals across the features, any given pattern suggest the addition of other variables.
    * Fitted values $\hat{y}_i$ - this will tell us if we have **homosecdasticity** (costant variance in the residuals). [LM] page 74 has a nice example of this.
    * Order of measure - this could be time, spatial order, or other sequential measures. This is done to assure we have some independence among the residuals. From [LM] we can also conduct the Durbin-Watson test to check for correlation among residuals, i.e. the correlations are independent.

– **The law of parsimony** or **Occam's razor** refers to the principle that no more causes should be assumed than will account for the effect, all that to say that simpler models should be preferred. Hypothesis testing can help identify a good model with differing hypotheses, but model interpretability results from point estimates and confidence intervals over p-values.

Some functions in R [LM]:

Here we take code from [LM] pages 76-83.

– Fitting a linear model then plotting residuals against fitted values:

```
data(savings,package="faraway")
lmod <- lm(sr ~ pop15+pop75+dpi+ddpi,savings)
plot(fitted(lmod),residuals(lmod),xlab="Fitted",ylab="Residuals")
abline(h=0)
```

– QQ Plots for normality:

```
lmod <- lm(sr ~ pop15+pop75+dpi+ddpi,savings)
qqnorm(residuals(lmod),ylab="Residuals",main="")
qqline(residuals(lmod))
```

– Durbin-Watson statistic:

```
require(lmtest)
dwtest(nhtemp ~ wusa + jasper + westgreen + chesapeake +
tornetrask + urals + mongolia + tasman, data=globwarm)
```

Chapter 3.

The **link function** g satisfies $g(\mu_i) = \mathbf{x}_i^\intercal \boldsymbol{\beta}$ for the nonlinear relation $E(Y_i) = \mu_i$.
Exponential family of distributions: $f(y;\theta) = s(y)t(\theta)e^{a(y)b(\theta)} = \exp\left(a(y)b(\theta) + c(\theta) + d(y)\right)$

- First form is most common one, second one used when we have $s(y) = \exp\left(d(y)\right), t(\theta) = \exp\left(c(\theta)\right)$.
- If $a(y) = y$, district is said to be in **canonical** (or standard) form, with $b(\theta)$ sometimes called the natural parameter of the distribution.
- Table of most common distributions:

Distribution	Natural Parameter	c	d
Poisson	$\log\theta$	-θ	-$\log y!$
Normal	$\frac{\mu}{\sigma^2}$	$-\frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)$	$-\frac{y^2}{2\sigma^2}$
Binomial	$\log(\frac{\pi}{1-\pi})$	$n\log(1-\pi)$	$\log\binom{n}{y}$

Chapter 4.