1 Introduction

Background, scope and notation

- This course focuses on the analysis of relationships between measurements made on groups of subjects or objects. For example, the measurements might be the heights or weights and the ages of boys and girls, or the yield of plants under various growing conditions. We use the terms response, outcome or dependent variable for measurements that are free to vary in response to other variables called explanatory variables or predictor variables or independent variables. Symbolically

\[ \text{Response} \sim \text{predictor variables} \]

- Types of variables:
  - Nominal classifications: binomial: two categories multinomial: more than two categories
  - Ordinal classifications: there is some natural order or ranking between the categories: e.g., young, middle aged, old.
  - Continuous measurements: where observations may, at least in theory, fall anywhere on a continuum: e.g., weight, length or time.

- The time until a specific event occurs, such as the failure of an electronic component; the length of time from a known starting point is called the failure time.
Nominal and ordinal data are sometimes called categorical or discrete variables and the numbers of observations, counts or frequencies in each category are usually recorded.

For continuous data the individual measurements are recorded. The term quantitative is often used for a variable measured on a continuous scale and the term qualitative for nominal and sometimes for ordinal measurements.

A qualitative, explanatory variable is called a factor and its categories are called the levels for the factor.

A quantitative explanatory variable is sometimes called a covariate.

Notation: Random variables: Upper case letters: \( X, Y, Z, \ldots \)
Realizations: Lower case letters: \( x, y, z, \ldots \)
Parameters: Greek letters: \( \alpha, \beta, \ldots \)
Estimates: Greek letters with hats: \( \hat{\alpha}, \hat{\beta}, \ldots \)
Vectors: bold faced letters: \( \mathbf{X}, \mathbf{a}, \ldots \)
Density: (probability mass function or probability density function): \( f(y; \theta) \) or \( f(y|\theta) \) where \( \theta \) is a parameter vector.

### Common Distributions

**Normal:** \( Y \sim N(\mu, \sigma^2) \) with pdf

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

- \( Y_i \sim N(\mu_i, \sigma_i^2), i = 1, \ldots, n \)

- Cov \((Y_i, Y_j) = \rho_{ij}\sigma_i\sigma_j, \ i, j = 1, \ldots, n.\)

  Equivalent to Multivariate Normal Distribution (MVN):

\[
\mathbf{Y} \sim N(\mu, \mathbf{V}), \quad V_{ij} = \rho_{ij}\sigma_i\sigma_j
\]

- If \( Y_1, \ldots, Y_n \) are independent with \( Y_i \sim N(\mu_i, \sigma_i^2) \), then \( a_1Y_1 + \ldots + a_nY_n \sim N(\sum a_i\mu_i, \sum a_i^2\sigma_i^2). \)

### Chi-square Distribution

- Central Chi-square with \( n \) degrees of freedom: If \( Z_1, \ldots, Z_n \sim iid N(0, 1) \), then \( X^2 = Z_1^2 + \ldots + Z_n^2 \sim \chi^2(n) \). \( E(X^2) = n \) and \( Var(X^2) = 2n. \)

- If \( Y_1, \ldots, Y_n \) are independent with \( Y_i \sim N(\mu_i, \sigma_i^2) \), then

\[
X^2 = \sum_i (Y_i - \mu_i)^2/\sigma_i^2 \sim \chi^2(n).
\]

- Noncentral Chi-square with \( n \) degrees of freedom and non-centrality parameter \( \lambda: \chi^2(n, \lambda) \): If \( Y_1, \ldots, Y_n \) are independent with \( Y_i \sim N(\mu_i, 1) \), then

\[
X^2 = \sum_i Y_i^2 \sim \chi^2(n, \lambda), \quad \lambda = \mu_1^2 + \ldots + \mu_n^2.
\]

Here \( E(X^2) = n + \lambda \) and \( Var(X^2) = 2n + 4\lambda. \) (HMK: Prove it).
• If \( Y \sim \text{MVN}(\mu, V) \), then 
\[
(Y - \mu)^\top V^{-1} (Y - \mu) \sim \chi^2(n), \quad Y^\top V^{-1} Y \sim \chi^2(n, \lambda),
\]
where \( V \) has inverse \( V^{-1} \), 
\[
\lambda = \mu^\top V^{-1} \mu.
\]

• If \( X_1^2, ..., X_m^2 \) are independent with \( X_i^2 \sim \chi^2(n_i, \lambda_i) \) for each \( i \), then
\[
\sum_{i=1}^{m} X_i^2 \sim \chi^2\left( \sum_{i} n_i, \sum_{i} \lambda_i \right).
\]
(HMK: Prove it).

• If \( Y_1, ..., Y_n \) are ind. with \( Y_i \sim \text{MVN}_p(0, V) \), then
\[
S = \sum Y_i Y_i^\top \sim W(V, n),
\]
where \( W \) is the Wishart Distribution, a generalization of the Chisquare distribution. (HMK: One-page summary of the Wishart distribution).

**Student and F Distributions**

- **t distribution:** If \( Z \sim N(0, 1) \), \( X^2 \sim \chi^2(n) \) and \( Z \perp X^2 \), then
\[
T = Z/(X^2/n)^{1/2} \sim t(n).
\]

- **F distribution:** \( X_1^2 \sim \chi^2(n), \ X_2^2 \sim \chi^2(m) \) and \( X_1^2 \perp X_2^2 \), then
\[
F = \frac{X_1^2/n}{X_2^2/m} \sim F(n, m).
\]

- Relationship between \( t \) and \( F \): \( T^2 \sim F(1, n) \).

- Relationships among distribution.

**Quadratic Forms**

- A quadratic form is a polynomial expression in which each term has degree 2.
  E.g. \( X_1^2 + X_2^2 + X_1X_2 \).

- Let \( A = (a_{ij}) \) be a symmetric matrix: \( a_{ij} = a_{ji} \) for all \( i, j = 1, ..., n \). Then \( Y^\top AY = \sum_i \sum_j a_{ij} y_i y_j \) is a quadratic form in the \( y_i \)'s. And \( (Y - \mu)^\top A(Y - \mu) \) is a quadratic form in the terms \( y_i - \mu \) but not in the \( y_i \)'s.

- Quadratic form \( Y^\top AY \) and \( A \) are said to be **positive definite** if \( Y^\top AY > 0 \) whenever the elements of \( Y \) are not all zero. A necessary and sufficient condition for positive definiteness is that all the determinants are positive.
• If \( A \) is positive definite, then \( A^{-1} \) exists and has a square root \( A^{1/2} \):

\[
A^{T/2}A^{1/2} = A.
\]

• The rank of the matrix \( A \) is called the degrees of freedom of the quadratic form \( Q = \mathbf{Y}^T \mathbf{A} \mathbf{Y} \).

• Cochran Theorem: Let \( Y_1, ..., Y_n \) are independent and each \( Y_i \) has \( sN(0, \sigma^2) \) and \( Q = Y_1^2 + ... + Y_n^2 \). If \( Q_1, ..., Q_k \) are q.f. of \( Y_1, ..., Y_n \), and

\[
Q = Q_1 + ... + Q_k,
\]

with \( \text{Rank}(Q_i) = m_i \), then \( Q_1, ..., Q_k \) are independent and \( Q_i/\sigma^2 \sim \chi^2(m_i) \) for \( i = 1, ..., m \) iff \( m_1 + ... + m_k = n \).

• If \( X_1^2 \sim \chi^2(m) \) and \( X_2^2 \sim \chi^2(n) \), then \( X^2 = X_1^2 - X_2^2 \sim \chi^2(m-n) \) if \( X^2 \geq 0 \) and \( m > n \).

HMK

1. Write one-page summary of the Wishart distribution.

2. If \( X^2 = \sum Y_i^2 \sim \chi^2(n, \lambda) \) with \( \lambda \geq 0 \), then \( E(X^2) = n + \lambda \) and \( \text{Var}(X^2) = 2n + 4\lambda \).

3. If \( X_1^2, ..., X_m^2 \) are independent with \( X_i^2 \sim \chi^2(n_i, \lambda_i) \) for each \( i \), then

\[
\sum_{i=1}^{m} X_i^2 \sim \chi^2 \left( \sum_{i} n_i, \sum_{i} \lambda_i \right).
\]

4. If \( \Sigma \) is an \( n \times n \) covariance matrix, then \( \Sigma \) is semi-positive definite. Under what conditions \( \Sigma \) is positive definite?

5. If \( X_1^2 \sim \chi^2(m) \) and \( X_2^2 \sim \chi^2(n) \), then \( X^2 = X_1^2 - X_2^2 \sim \chi^2(m-n) \) if \( X^2 \geq 0 \) and \( m > n \).


Maximum likelihood estimation

• Let \( \mathbf{Y} = (Y_1, ..., Y_n)^T \) denote a random vector and let the joint density function of the \( Y_i \)'s be \( f(y; \theta) \) which depends on the vector of parameters \( \theta = (\theta_1, ..., \theta_p)^T \). Usually \( f(y; \theta) = f(y_1; \theta) ... f(y_n; \theta) \).

• The likelihood function \( L(\theta; \mathbf{y}) \) is algebraically the same as the joint probability density function \( f(\mathbf{y}; \theta) \): \( L(\theta; \mathbf{y}) = f(\mathbf{y}; \theta) \). But the change in notation reflects a shift of emphasis from the r.v. \( \mathbf{y} \), with \( \theta \) fixed, to the parameters \( \theta \) with \( \mathbf{y} \) fixed.
Let $\Theta$ denote the set of all possible values of the parameter vector $\theta$; $\Theta$ is called the parameter space.

The maximum likelihood estimator (MLE) of $\theta$ is the value $\hat{\theta}$ which maximizes the likelihood function, that is

$$L(\hat{\theta}; y) \geq L(\theta; y), \quad \forall \theta \in \Theta.$$ 

Equivalently, the MLE $\hat{\theta}$ is the value which maximizes the log-likelihood function:

$$\ell(\hat{\theta}; y) = \log L(\hat{\theta}; y).$$

So

$$\ell(\hat{\theta}; y) \geq \ell(\theta; y), \quad \theta \in \Theta.$$ 

Does the MLE exist? Unique? How to find it? Explicit formula? Numerical solution? Using packages such as R, S plus, SAS, etc.

If $\ell$ has derivative w.r.t. $\theta$, then we may find the MLE via solving the score equation:

$$S(\theta; y) = \frac{\partial}{\partial \theta} \ell(\theta; y) = 0,$$

where $S$ is called the score function.

ARE the solutions the maxima? Check the second derivatives (matrix)

$$\frac{\partial^2}{\partial \theta \partial \theta} \ell(\theta; y) < 0$$

evaluated at $\theta = \hat{\theta}$ is negative definite.

Invariance Principle If $g(\theta)$ is a function of the parameters $\theta$, then the maximum likelihood estimator of $g(\theta)$ is $g(\hat{\theta})$.

Example: Poisson distribution

Let $Y_1, ..., Y_n \sim \text{iid Poi}(\theta)$ with $\theta > 0$:

$$f(y_i; \theta) = e^{-\theta} \frac{\theta^{y_i}}{y_i!}, \quad y_i = 0, 1, 2, ...$$

The log-likelihood function is

$$\ell(\theta; y) = \log L(\theta; y) = \sum \log f(y_i; \theta)$$

$$= -n\theta + \sum y_i \log \theta - \sum \log(y_i!).$$

The score equation is

$$S(\theta; y) = \frac{\partial}{\partial \theta} \ell(\theta; y) = -n + \theta^{-1} \sum y_i = 0.$$
• The solution is \( \hat{\theta} = \bar{y} \).

• Since

\[
\frac{\partial^2}{\partial \theta \partial \theta^T} \ell(\theta; y) = -\hat{\theta}^2 \sum y_i < 0,
\]

\( \hat{\theta} \) is the MLE.

**Least Squares Estimation**

• Let \( Y_1, ..., Y_n \) be independent random variables with expected values \( \mu_1, ..., \mu_n \). Suppose that the \( \mu_i \)'s are functions of the parameter vector:

\[
\mu_i = E(Y_i) = \mu_i(\beta), \quad \beta \in \mathbb{R}^p, \ p < n.
\]

• The least squares estimates (LSE) \( \hat{\beta}_n \) is the minimizers of

\[
S_n(\beta) = \sum_{i=1}^{n} (y_i - \mu_i(\beta))^2,
\]

• One way to find it is by taking derivative and set it equal zero: \( \frac{\partial}{\partial \beta} S_n(\beta) = 0 \).

• Check out it is the LSE!

**Weighted LSE**

• Now suppose that the \( Y_i \)'s have unequal variances \( \sigma_i^2 \). Then desirable to minimize the weighted sum:

\[
S_n(\beta) = \sum_{i=1}^{n} w_i [y_i - \mu_i(\beta)]^2
\]

where the weights are \( w_i = 1/\sigma_i^2 \).

• let \( Y = [Y_1, ..., Y_n]^T \) denote a r.v. with mean vector \( \mu = [\mu_1, ..., \mu_n]^T \) and variance-covariance matrix \( \mathbf{V} \). Then the weighted least squares estimator is obtained by minimizing

\[
S_n(\beta) = (Y - \mu)^T \mathbf{V}^{-1} (Y - \mu).
\]

• Example: Tropical cyclones: Page 15.

2 Model Fitting

2.1 Introduction

Introduction
Model specification – a model is specified in two parts: an equation linking the response and explanatory variables and the probability distribution of the response variable.

Estimation of the parameters of the model.

Checking the adequacy of the model – how well it fits or summarizes the data.

Inference – calculating confidence intervals and testing hypotheses about the parameters in the model and interpreting the results.

2.2 Examples

Example: Chronic medical conditions

Data from the Australian Longitudinal Study on Women’s Health (Brown et al., 1996) show that women who live in country areas tend to have fewer consultations with general practitioners (family physicians) than women who live near a wider range of health services.

It is not clear whether this is because they are healthier or because structural factors, such as shortage of doctors, higher costs of visits and longer distances to travel, act as barriers to the use of general practitioner (GP) services.

Table 1 shows the numbers of chronic medical conditions reported by samples of women living in large country towns (town group) or in more rural areas (country group) in New South Wales, Australia. All the women were aged 70-75 years, had the same socio-economic status and had three or fewer GP visits during 1996.

The question of interest is: do women who have similar levels of use of GP services in the two groups have the same need as indicated by their number of chronic medical conditions?
• How to model count data? Poisson distribution? Check the data if mean $\approx$ variance.

• $Y_{jk} = \text{number of conditions for } k\text{th woman in } j\text{th group}$, where $j = 1$ for town group and $j = 2$ for country group and $k = 1, \ldots, K_j$ with $K_1 = 26$ and $K_2 = 23$.

• Naturally, $Y_{1k} \sim \text{Poi}(\theta_1)$ and $Y_{2k} \sim \text{Poi}(\theta_2)$, where $\theta_1$ and $\theta_2$ are parameters.

• The question of interest can be formulated as a test of the null hypothesis $H_0 : \theta_1 = \theta_2$ against the alternative hypothesis $H_1 : \theta_1 \neq \theta_2$.

• How to test the hypothesis? There are several ways.

• Under the null, $\theta_1 = \theta_2 = \theta$ and combine the two groups and find the estimate $\hat{\theta}$. Under the alternative, find the estimates $\hat{\theta}_1, \hat{\theta}_2$. If the maximized likelihoods are not “close”, then we should reject the null. (Can we look at the estimates of the parameters of the two groups?)

• How to estimate? MLE. Under $H_0$, the MLE of $\theta$ is the sample mean $\hat{\theta} = \bar{y}.$ = 1.184 with the maximized log-likelihood $\hat{\ell}_0 = \ell_0(\hat{\theta}) = -68.386$. Under $H_1$, the MLE’s are $\hat{\theta}_1 = 1.423$ and $\theta_2 = 0.913$ with $\hat{\ell}_1 = \ell_1(\hat{\theta}_1) = -67.0230$. Since the two likelihoods are so “close”, we most possibly do not reject the null.

• Observe $\hat{\ell}_0 \leq \hat{\ell}_1$. This is actually true in general. Why?

• Even though $\hat{\ell}_1 - \hat{\ell}_0 = 1.3638$ is “small”, we still have to test whether this is significant. What is the distribution of the difference? We’ll see it is Chisquare.

Residuals

• For $Y \sim \text{Poi}(\theta)$, $E(Y) = \text{Var}(Y) = \theta$, so we estimate $E(Y)$ by $\hat{\theta}$, which is called the fitted value of $Y$.

• The difference $Y - \hat{\theta}$ is called the residual, which is the basis for assessing model adequacy. A residual is usually standardized by dividing by its standard error. For the Poisson distribution an approximate standardized residual is

$$r = \frac{Y - \hat{\theta}}{\sqrt{\hat{\theta}}}.$$ 

The standardized residuals for models (2.1) and (2.2) are shown in Table 2.2 and Figure 2.1.
Approximate Chisquare Distribution

• The residuals can also be aggregated to produce summary statistics measuring the overall adequacy of the model. For example, for Poisson data denoted by independent random variables $Y_i$, provided that the expected values $\theta_i$’s are not too small, the standardized residuals $r_i = (Y_i - \hat{\theta}_i)/\sqrt{\hat{\theta}_i}$ approximately have the standard Normal distribution $N(0, 1)$, although they are not usually independent. Hence

$$\sum r_i^2 = \sum \frac{(Y_i - \hat{\theta}_i)^2}{\hat{\theta}_i} \approx \chi^2(m),$$  

(1)

It can be shown that for large samples, (1) is a good approximation with $m$ equal to the number of observations minus the number of parameters estimated in order to calculate the fitted values $\hat{\theta}_i$.

• (1) is the usual chisquare goodness of fit statistic for count data:

$$X^2 = \sum \frac{(o_i - e_i)^2}{e_i} \approx \chi^2(m),$$

where $o_i$ denotes the observed frequency and $e_i$ denotes the corresponding expected frequency.

• For the data on chronic medical conditions, for model under $H_0$,

$$r_i^2 = 6 * (-1.088)^2 + 10 * (-0.169)^2 + ... + 1 * 1.669^2 = 46.759.$$  

This value is consistent with $\sum r_i^2$ being an observation from the central chisquared distribution with $m = 23 + 26 - 1 = 48$ degrees of freedom.

Birthweight and gestational age

• pages 23-33.

HMK

• Describe the advantages and disadvantages of the maximum likelihood estimation and least squares estimation (LSE).

• Consider a simple linear regression model, $Y_i = \beta_0 + \beta_1x_i + \epsilon_i, i = 1, 2, 3$, where the random errors $\epsilon_i$ are independent with $\epsilon_i \sim N(0, \sigma_i^2), i = 1, 2, 3$. Give the weighted LSE $\hat{\beta}_0, \hat{\beta}_1$ and the LSE $\tilde{\beta}_0, \tilde{\beta}_1$. If all $\sigma_i$ are different, what are the consequences of using the LSE?

• Page 40, 2.1, 2.3, 2.5.
2.3 Some principles of statistical modelling

Exploratory data analysis

- The scale of measurement? Is it continuous or categorical? If it is categorical, how many categories does it have and are they nominal or ordinal?
- The shape of the distribution? This can be examined using frequency tables, dot plots, histograms and other graphical methods.
- How is it associated with other variables? Cross tabulations for categorical variables, scatter plots for continuous variables, side-by-side box plots for continuous scale measurements grouped according to the factor levels of a categorical variable, and other such summaries can help to identify patterns of association.

Model formulation

- What should be the distribution of the response $Y$?
- What link function appropriate?

Parameter estimation: MLE? LSE?

Model Adequacy

- Suppose $Y_i \sim N(\mu_i, \sigma^2)$. Then the fitted values $\hat{Y}_i = \hat{\mu}_i$ and the residual is $Y_i - \hat{Y}_i$ and the standardized residual is
  \[ r_i = (Y_i - \mu_i) / \hat{\sigma}. \]
  These residuals are approximately normally distributed with mean zero. Recall $\mu_i = \mu(x_i^T \beta)$. If the model is a good fit, then there should be no remaining pattern in the residual plot. This can also be checked by looking at the statistic
  \[ \sum (y_i - \hat{\mu}_i)^2 / \hat{\sigma}^2 \overset{\text{appr}}{\sim} \chi^2(m), \]
  where $m$ is the d.f.
- Suppose $Y_i \sim Pois(\theta)$. Then the approximate residual is
  \[ r_i = (Y_i - \hat{\theta}_i) / \sqrt{\hat{\theta}_i}. \]

Model Adequacy

- Normal probability plot can be used to check normality. In R, call `qqnorm`.
- Inference and interpretation.
2.4 Notation and coding for explanatory variables

Notation

- $Y \sim f(y; \theta)$ where $f$ belongs to the exponential family. $x_1, ..., x_m$ explanatory variables, $\beta_1, ..., \beta_m$ parameters. In vector form $Y = (Y_1, ..., Y_m)^T$, $x = (x_1, ..., x_m)^T$ (column vector) and $\beta = (\beta_1, ..., \beta_m)^T$.

$$g(E(Y)) = \mathbf{x}^T \beta = \beta^T \mathbf{x},$$

where $g(E(Y)) = (g(E(Y_1)), ..., g(E(Y_n)))^T$. Equivalently,

$$E(Y) = h(x^T \beta),$$

$g$ is the called a link function and $h$ an inverse link function.

- dummy variables, indicator variable, design matrix $X$, and linear component: $x^T \beta$.

Coding Examples

- Means for two groups
- Simple linear regression for two groups
- Alternative formulations for comparing the means of two groups
- Ordinal explanatory variables

3 Exponential Family and Generalized Linear Models

3.1 Introduction

Introduction

- The usual linear model is

$$Y_i = X_i^T \beta + \varepsilon_i, \quad \varepsilon_i \sim i.i.d. N(0, \sigma^2), \quad \sigma^2 > 0, \quad i = 1, ..., n.$$  

Equivalently, $Y_i$'s are indepedednt and satisfy

$$E(Y_i) = \mu_i = X_i^T \beta, \quad Y_i \sim N(\mu_i, \sigma^2), \quad i = 1, ..., n.$$  

- LM is not appropriate for many real life problems. Such examples includes when $Y$ denotes (1) binomial observation, (2) Poission observation.
Generalized Linear Models (GLM)

- Response variables have distributions other than the Normal distribution – they may even be categorical rather than continuous.

- Relationship between the response and explanatory variables need not be of the simple linear form:

  \[ g(\mu_i) = X_i^T \beta, \quad \mu_i = h(X_i^T \beta), \quad i = 1, \ldots, n \]

  where \( g \) is called a link function and \( h \) the inverse link function. It is monotone and smooth and usually known.

- More generally, semiparametric models:

  \[ g(\mu_i) = X_i^T \beta + \rho(Z_i), \quad i = 1, \ldots, n, \]

  including partially linear models, single-index models, GEE models, etc.

### 3.2 Exponential family

**Definition 3.1.** Let \( Y \) be a r.v. whose probability distribution depends on parameters \( \theta \in \mathbb{R} \) and \( \phi \). The distribution of \( Y \) belongs to the exponential family if the density (pmf or pdf) can be written as

\[
  f(y; \theta) = \exp \left\{ a(y) b(\theta) + c(\theta) + d(y; \phi) \right\}, \quad y \in \mathbb{R},
\]

where \( a, b, c, d, \nu \) are known functions.

- If \( a(y) = y \), the distribution is called in canonical form and the parameter \( \vartheta = b(\theta) \) is called the natural parameter.

- Other parameters \( \phi \) in addition to the parameters \( \theta \) of interests are called nuisance parameters.

**More General Exponential Family**

- A more general definition of the exponential family is

\[
  f(y; \theta) = \exp \left\{ \frac{a^T(y)b(\theta) + c(\theta)}{\nu(\phi)} + d(y; \phi) \right\}, \quad y \in \mathbb{R}^p,
\]

  where \( a, b, c, d, \nu \) are known functions and \( \theta \) is a parameter vector \( \phi \) is a nuisance parameter vector.

- Again if \( a(y) = y \), the canonical form is

\[
  f(y; \theta) = \exp \left\{ \frac{y^T b(\theta) + c(\theta)}{\nu(\phi)} + d(y; \phi) \right\}, \quad y \in \mathbb{R}^p,
\]

  where \( \theta \in \mathbb{R}^p \) with natural parameter \( \vartheta = b(\theta) \).
Examples

- Any exponential family distribution can be reparametrized to be of canonical form with natural parameter: \( u = a(y) \) and \( \vartheta = b(\theta) \), so that
  \[
  f(u; \vartheta, \phi) = \exp \left\{ u^\top \vartheta - \tilde{b}(\vartheta) \nu(\phi) + \tilde{c}(u; \phi) \right\}, \quad u \in \mathbb{R}^q,
  \]
  where \( \phi \) is a nuisance parameter.
- Poisson Distribution
- Normal Distribution
- Binomial Distribution

3.3 Moments of the Exp Family

- Let
  \[
  \ell(\theta; y) = \log f(y; \theta) = a(y)b(\theta) + c(\theta) + d(y).
  \]
  Differentiating and interchanging differential and integral accross
  \[
  \int f(y; \theta) \, dy = 1,
  \]
  we obtain
  \[
  \int \frac{\partial}{\partial \theta} \ell(\theta; y) f(y; \theta) \, dy = 0.
  \]
  In the notation of expectation,
  \[
  E\left[ \frac{\partial}{\partial \theta} \ell(\theta; Y) \right] = 0
  \]
  For convenience, assume \( \nu(\phi) = 1 \) and \( \phi \) is known
- Since \( \frac{\partial}{\partial \theta} \ell = a(y)b'(\theta) + c'(\theta) \), it follows
  \[
  b'(\theta)E(a(Y)) + c'(\theta) = 0, \quad E(a(Y)) = -\frac{c'(\theta)}{b'(\theta)}.
  \]
- Differentiating further and with some algebra, we obtain
  \[
  Var(a(Y)) = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{b'(\theta)^3},
  \]
• In the canonical form, the density can be written as
  \[ f(y; \theta, \phi) = \exp \left\{ \frac{y \theta - b(\theta)}{a(\phi)} + c(y; \phi) \right\}, \]
  where \( c(y; \phi) \) does not depend on \( \theta \). (HMK) Show the nice formulas
  \[ E(Y) = b'(\theta), \quad Var(Y) = b''(\theta) a(\phi). \]

• (HMK) Let
  \[ f(y; \varphi) = \exp(\varphi y - b(\varphi)), \quad \varphi \in \Phi; \]
  be a density with respect to sigma-finite measure \( \nu \) (Namely, \( \int f(y; \varphi) \, d\nu(y) = 1 \) for all \( \varphi \in \Phi \), where \( \Phi \) is a nonempty open subset of \( \mathbb{R} \)). Derive
  \[ \int y^k \exp(\varphi y) \, d\nu(y) = \frac{\partial^k}{\partial \varphi^k} \exp(b(\varphi)), \quad k = 1, 2, \ldots \]
  From this derive the formulas for \( E(Y) \) and \( Var(Y) \).

Score Function
Recall the score function \( (U(\theta; y) = S(\theta; y)) \) is calculated by
  \[ U(\theta; y) = \frac{\partial}{\partial \theta} \ell(\theta; y) = a(y) b'(\theta) + c'(\theta). \]
Clearly, \( E(U(Y; \theta)) = b'(\theta) E(a(Y)) + c'(\theta) = b'(\theta) \frac{c(\theta)}{b(\theta)} + c'(\theta) = 0. \) Hence
\[ Var(U) = E(U^2(Y; \theta)) = b'(\theta)^2 Var(a(Y)) = \frac{b''(\theta) c'(\theta)}{b'(\theta)^2} - c''(\theta) \]
so that
\[ Var(U) = E(U^2) = -E(U'). \] (2)
This is an important relation and can be used in calculation. It is true in general under certain regularity conditions (HWK).

3.4 Generalized linear models
• \( Y_1, \ldots, Y_N \) are independent and each \( Y_i \) has a distribution from the exponentail family with density of canonical form
  \[ f(y_i; \theta_i) = \exp \left\{ \frac{y_i b(\theta_i) + c(\theta_i)}{\nu(\phi)} + d(y_i; \phi) \right\} \]
• \( x_1, \ldots, x_N \) are \( p \)-dim covariate vectors and \( \beta = (\beta_1, \ldots, \beta_p)^T \) are parameters with \( p < N \). In a GLM, the response \( Y_i \) and the covariate vector \( x_i \in \mathbb{R}^p \) satisfies the structural relationship
  \[ g(\mu_i) = x_i^T \beta, \quad i = 1, \ldots, n, \]
where \( \beta = (\beta_1, \ldots, \beta_p)^T \).