Estimation and Hypothesis Test

Guilherme J. M. Rosa
University of Wisconsin-Madison

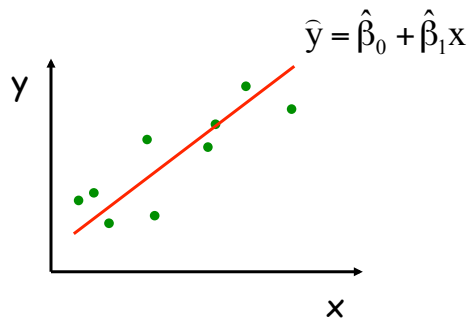## Simple Linear Regression

Data pairs

X	Y
x_1	y_1
x_2	y_2
$\vdots$	$\vdots$
x_n	y_n

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

$$e_i \overset{ind}{\sim} (0, \sigma^2)$$

$$\begin{cases} E[y_i] = \beta_0 + \beta_1 x_i \\ Var[y_i] = \sigma^2 \end{cases}$$

$$\widehat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Scatter plot of the data points (x_i, y_i) with the estimated regression line of y as a function of x.

Least Squares Estimation

- How to find 'good' estimates of the model parameters β_0 and β_1?

- Least squares solution: estimators that minimize the Residual Sum of Squares (RSS):

$$RSS = \sum_{i=1}^{n}[\hat{e}_i]^2 = \sum_{i=1}^{n}[y_i - \hat{y}_i]^2 = \sum_{i=1}^{n}[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

$$\begin{cases} \dfrac{\partial RSS}{\partial \hat{\beta}_0} = 2\sum_{i=1}^{n}[y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i](-1) \\[2em] \dfrac{\partial RSS}{\partial \hat{\beta}_1} = 2\sum_{i=1}^{n}[y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i](-x_i) \end{cases} \Rightarrow \begin{cases} \sum_{i=1}^{n}y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n}x_i = 0 \\[2em] \sum_{i=1}^{n}x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n}x_i - \hat{\beta}_1 \sum_{i=1}^{n}x_i^2 = 0 \end{cases}$$

Least Squares Estimation

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \quad \text{and} \quad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

$$S_{xx} = \sum_{i=1}^{n}x_i^2 - \frac{1}{n}\left[\sum_{i=1}^{n}x_i\right]^2 \qquad S_{xy} = \sum_{i=1}^{n}y_i x_i - \frac{1}{n}\sum_{i=1}^{n}x_i \sum_{i=1}^{n}y_i$$

- Unbiased estimators, i.e. $E[\hat{\beta}_0] = \beta_0$ and $E[\hat{\beta}_1] = \beta_1$

- In addition, as $E[RSS] = \sum_{i=1}^{n}E[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 = (n-2)\sigma^2$

  an unbiased estimator of σ^2 is: $s^2 = \dfrac{1}{n-2}\sum_{i=1}^{n}(\hat{e}_i)^2$

Example

Data: Information on phenotypes and genotypes for a QTL affecting a specific quantitative trait.

QTL Genotype	Phenotype (8 individuals per group)
CC	95.9, 108.0, 96.5, 92.9 101.0, 94.5, 93.7, 89.8
CT	101.2, 103.9, 85.9, 109.4 105.7, 98.4, 84.1, 103.1
TT	117.1, 95.2, 106.4, 104.7 92.5, 123.9, 97.8, 100.5

Suppose that two alleles segregate in co-dominance in this locus; infer the phenotypic mean of the heterozygous individuals, as well as the gene effect (allele substitution) and the residual variance (environment + polygenes).

---

linear reg

Example

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad \begin{cases} y_i \text{: phenotype of individual i} \\ x_i \text{: QTL genotype of individual i} \\ i = 1, 2, …, 24 \text{ (index for individual)} \end{cases}$$

Parameterization choice:
$\begin{cases} \beta_0 \text{: expected phenotypic value of heterozygous} \\ \beta_1 \text{: allelic substitution effect} \\ x_i = -1, 0 \text{ and } 1 \text{ for QTL} = CC, CT \text{ and } TT \end{cases}$

$$\sum_{i=1}^{24} x_i = 0.0 \qquad \sum_{i=1}^{n} x_i^2 = 16.0$$

$$\sum_{i=1}^{24} y_i = 2402.1 \qquad \sum_{i=1}^{24} y_i x_i = 65.8$$

⇨

$$S_{xx} = 16.0$$
$$S_{xy} = 65.8$$

⇨

$$\begin{cases} \hat{\beta}_0 = 100.09 \\ \hat{\beta}_1 = 4.11 \\ \hat{\sigma}^2 = 75.45 \end{cases}$$

Maximum Likelihood

$$y_1, y_2, \ldots, y_n \overset{iid}{\sim} p(y_i \mid \theta_1, \theta_2, \ldots, \theta_k)$$

Likelihood Function:
$$L(\theta \mid \mathbf{y}) = p(\mathbf{y} \mid \theta) \overset{ind}{=} \prod_{i=1}^{n} p(y_i \mid \theta_1, \theta_2, \ldots, \theta_k)$$

Log-Likelihood Function:
$$l(\theta \mid \mathbf{y}) = \log L(\theta \mid \mathbf{y}) = \sum_{i=1}^{n} \log p(y_i \mid \theta)$$

$$\hat{\theta} = MLE(\theta) \quad \Rightarrow \quad L(\hat{\theta} \mid \mathbf{y}) \geq L(\theta \mid \mathbf{y}), \quad \text{any} \quad \theta \in \Theta$$

parameter space

Maximum Likelihood

Finding the maximum of $L(\theta \mid \mathbf{y})$:

$$\frac{\partial}{\partial \theta_i} L(\theta \mid \mathbf{y}) = 0 \quad \text{(solutions are possible candidates)}$$

$$\left. \frac{\partial^2}{\partial \theta_i^2} L(\theta \mid \mathbf{y}) \right|_{\theta_i = \hat{\theta}_i} < 0 \quad \text{(maximum)}$$

Check also the boundaries
of the parameter space!!

4

Example 1: $y_1, y_2, \ldots, y_n \overset{iid}{\sim} \text{Bin}(n_i, p)$

$$L(p \mid \mathbf{y}) = \prod_{i=1}^{n} \Pr(y_i \mid p) \propto p^{\sum y_i} (1-p)^{\sum (n_i - y_i)}$$

$$l(p \mid \mathbf{y}) \propto \sum y_i \log p + \sum (n_i - y_i) \log(1-p)$$

$$\frac{d}{dp} l(p \mid \mathbf{y}) = \frac{1}{p} \sum y_i - \frac{1}{(1-p)} \sum (n_i - y_i)$$

$$\frac{1}{\hat{p}} \sum y_i - \frac{1}{(1-\hat{p})} \sum (n_i - y_i) = 0 \implies \boxed{\hat{p} = \frac{\sum y_i}{\sum n_i}}$$

$$\text{Check:} \quad \frac{d^2}{dp^2} l(p \mid \mathbf{y})\Big|_{\hat{p}} < 0$$

Example 2: $y_1, y_2, \ldots, y_n \overset{iid}{\sim} N(\mu, \sigma^2)$

$$L(\mu, \sigma^2 \mid \mathbf{y}) = \prod_{i=1}^{n} p(y_i \mid \mu, \sigma^2) \propto (\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right\}$$

$$l(\mu, \sigma^2 \mid \mathbf{y}) \propto -\frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2$$

$$\begin{cases} \dfrac{\partial}{\partial \mu} l(\mu, \sigma^2 \mid \mathbf{y}) = \dfrac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \mu) \\[3mm] \dfrac{\partial}{\partial \sigma^2} l(\mu, \sigma^2 \mid \mathbf{y}) = -\dfrac{n}{2\sigma^2} + \dfrac{1}{2\sigma^4} \sum_{i=1}^{n} (y_i - \mu)^2 \end{cases} \implies \boxed{\begin{aligned} \hat{\mu} &= \frac{1}{n} \sum_{i=1}^{n} y_i \\[3mm] \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\mu})^2 \end{aligned}}$$

Example 3: $\begin{bmatrix} y_{i1} \\ y_{i2} \end{bmatrix} \overset{iid}{\sim} N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \right)$

$$\rho = \frac{\sigma_{12}}{\sqrt{\sigma_1^2 \sigma_2^2}}$$

$$p(y_{i1}, y_{i2} \mid \theta) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}$$

$$\times \exp\left\{ -\frac{1}{2(1-\rho^2)}\left[\frac{(y_{i1}-\mu_1)^2}{\sigma_1^2} + \frac{(y_{i2}-\mu_2)^2}{\sigma_2^2} - 2\rho\frac{(y_{i1}-\mu_1)(y_{i2}-\mu_2)}{\sigma_1\sigma_2} \right] \right\}$$

$$\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} y_{ij}$$

$$\hat{\rho} = \frac{\sum_{i=1}^{n}(y_{i1}-\hat{\mu}_1)(y_{i2}-\hat{\mu}_2)}{\sqrt{\sum_{i=1}^{n}(y_{i1}-\hat{\mu}_1)^2 \sum_{i=1}^{n}(y_{i2}-\hat{\mu}_2)^2}}$$

$$\hat{\sigma}_j^2 = \frac{1}{n}\sum_{i=1}^{n}(y_{ij}-\hat{\mu}_j)^2$$

Hypothesis Testing

⇨ Likelihood Ratio Test (LRT)

$$y_1, y_2, \ldots, y_n \overset{iid}{\sim} p(y_i \mid \theta)$$

$$L(\theta \mid \mathbf{y}) = p(\mathbf{y} \mid \theta) \overset{ind}{=} \prod_{i=1}^{n} p(y_i \mid \theta)$$

Suppose: $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \notin \Theta_0$

Restricted (Θ_0) maximization

$$\text{LRT} = \frac{\max_{\Theta_0} L(\theta \mid \mathbf{y})}{\max_{\Theta} L(\theta \mid \mathbf{y})}$$

Unrestricted maximization

$0 \le \text{LRT} \le 1$

Hypothesis Testing

Let: $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$

So Θ_0 represents a unique value (θ_0)

$$LRT = \frac{L(\theta_0 \mid \mathbf{y})}{L(\hat{\theta} \mid \mathbf{y})}$$

Critical Region: LRT < c

How to choose the cutoff value c ?

Hypothesis Testing

	Accept H_0	Reject H_0
H_0 is true	$1 - \alpha$	α
H_0 is false	β	$1 - \beta$

Type I Error (Significance Level)

Type II Error

Power

➡ Standard approach:

① Specify an acceptable type I error rate (α)

② Seek tests that minimize the type II error rate (β), i.e., maximize power ($1 - \beta$)

Type I and Type II Errors

You're pregnant!

You're not pregnant.

Type I Error

Type II Error

Hypothesis Testing

⇨ Log-Likelihood Ratio Test

$$\text{logLRT} = -2\log\left[\frac{L(\theta_0 \mid \mathbf{y})}{L(\hat{\theta} \mid \mathbf{y})}\right]$$

$$-2\log\left[\frac{L(\theta_0 \mid \mathbf{y})}{L(\hat{\theta} \mid \mathbf{y})}\right] \sim \chi^2_\phi$$

φ: degrees of freedom

Difference in dimension of the spaces

The Randomization test

➡ The basic idea is attractively simple and free of mathematical assumptions

Experiment

Suppose:

Trt 1	Trt 2
y_{11}	y_{21}
y_{12}	y_{22}
⋮	⋮
y_{1n1}	y_{2n2}
$\overline{y}_1 \pm s_1$	$\overline{y}_2 \pm s_2$

From distribution F

From distribution G

$$H_0 : \ F = G \quad \text{vs.} \quad H_1 : \ F \neq G$$

The Randomization test

① Define the statistics (e.g. $t = \dfrac{\overline{y}_1 - \overline{y}_2}{se}$) and calculate its value for the data set (call it t*)

② Combine the $n_1 + n_2$ observations

③ Take a sample of size n_1 without replacement (to represent the Group C)

① The remaining n_2 observations constitute the Group T

② Compute the value of t (call it t_i) and repeat the process a large number (B) of times

③ P-value: $p = \Sigma \ I(t_{(i)} \geq t^*)/B$

The Randomization test

Experiment

Trt 1	Trt 2
y_{11}	y_{21}
y_{12}	y_{22}
$\vdots$	$\vdots$
y_{1n1}	y_{2n2}
$\bar{y}_1 \pm s_1$	$\bar{y}_2 \pm s_2$

$$t^* = \frac{\bar{y}_1 - \bar{y}_2}{se}$$

Permutation 1

'T1'	'T2'
y_{12}	y_{1n1}
y_{25}	y_{23}
$\vdots$	$\vdots$
y_{2n2}	y_{11}
$\bar{y}_1 \pm s_1$	$\bar{y}_2 \pm s_2$

Permutation 2

'T1'	'T2'
y_{21}	y_{27}
y_{11}	y_{1n1}
$\vdots$	$\vdots$
y_{12}	y_{2n2}
$\bar{y}_1 \pm s_1$	$\bar{y}_2 \pm s_2$

...

Permutation B

'T1'	'T2'
y_{1n1}	y_{13}
y_{14}	y_{22}
$\vdots$	$\vdots$
y_{21}	y_{2n2}
$\bar{y}_1 \pm s_1$	$\bar{y}_2 \pm s_2$

$$t_{(1)} < t_{(2)} < \ldots < t_{(B)}$$

P-value: $p = \Sigma\, I(t_{(i)} \geq t^*)/B$

Linear Models

Guilherme J. M. Rosa
University of Wisconsin-Madison

General Linear Model
(Fixed Effects Model)

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

responses   residuals

design/incidence matrix (known)   overall mean + fixed effects parameters

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, I_n\sigma^2) \ \rightarrow \ \varepsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$$

⇨ Fixed effect: levels included in the study represent all levels about which inference is to be made. Fixed effects models: models containing only fixed effects

---

Example 1

Experiment to compare growth performance of pigs under two experimental groups (Control and Treatment), with three replications each.

Control	Treatment
53	61
46	66
58	57

Model:

$$y_{ij} = \mu + \delta_i + e_{ij}$$

y_{ij}: weight gain of pig j of group i

μ: constant; general mean

δ_i: effect of group i

e_{ij}: residual term

Matrix Notation

Control	Treatment
53	61
46	66
58	57

$$
\begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{bmatrix} = \begin{bmatrix} 53 \\ 46 \\ 58 \\ 61 \\ 66 \\ 57 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \delta_1 \\ \delta_2 \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{21} \\ e_{21} \\ e_{23} \end{bmatrix}
$$

Alternative Parameterizations

⇨ Equivalent models with different parameterizations

For example, if the average weight gain in each group is expressed as $\mu_i = \mu + \delta_i$, the model becomes:

$$
\begin{bmatrix} 53 \\ 46 \\ 58 \\ 61 \\ 66 \\ 57 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{21} \\ e_{21} \\ e_{23} \end{bmatrix}
$$

Alternatively, the model can be expressed in terms of the average weight gain of the Control (μ_1) and the difference on weight gain between the two groups ($\tau = \mu_2 - \mu_1$):

$$
\begin{bmatrix} 53 \\ 46 \\ 58 \\ 61 \\ 66 \\ 57 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \tau \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{21} \\ e_{21} \\ e_{23} \end{bmatrix}
$$

Example 2

Flowering time (days, log scale) of *Brassica napus* according to genotype in specific locus, such as a candidate gene

Genotype		
qq	Qq	QQ
3.4	2.9	3.1
3.7	2.5	2.6
3.2		

Model: $y_{ij} = \mu_i + e_{ij}$

y_{ij}: flowering time of replication j (j = 1,…, n_i) of genotype i (i = qq, Qq and QQ)

μ_i: expected flowering time of plants of genotype i

e_{ij}: residual (environment and polygenic effects)

⇨ The expected phenotypic values μ_i, however, can be expressed as a function of the additive and dominant effects
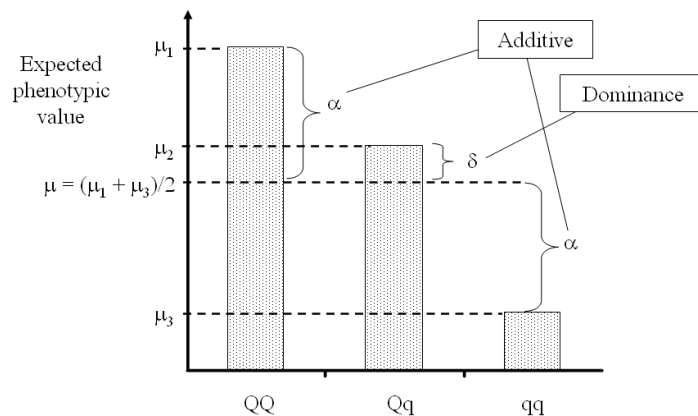
Expected phenotypic value according to the genotype on a specific locus.

The model can be written then as:

$$y_{ij} = \mu + x_{ij}\alpha + (1 - |x_{ij}|)\delta + e_{ij}$$

μ: constant (mid-point flowering time between homozygous genotypes)

x_{ij}: indicator variable (genotype), coded as -1, 0 and 1 for genotypes qq, Qq and QQ

α and β: additive and dominance effects

In matrix notation:

$$
\begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{bmatrix}
=
\begin{bmatrix} 3.4 \\ 3.7 \\ 3.2 \\ 2.9 \\ 2.5 \\ 3.1 \\ 2.6 \end{bmatrix}
=
\begin{bmatrix} 1 & -1 & 0 \\ 1 & -1 & 0 \\ 1 & -1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix}
\begin{bmatrix} \mu \\ \alpha \\ \delta \end{bmatrix}
+
\begin{bmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{21} \\ e_{22} \\ e_{31} \\ e_{32} \end{bmatrix}
$$

More on the LS Methodology

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{\varepsilon} \sim (\mathbf{0}, I_n\sigma^2) \quad \rightarrow \quad \varepsilon_i \overset{iid}{\sim} (0, \sigma^2)$$

An estimate $(\hat{\boldsymbol{\beta}})$ of the vector β can be obtained by the method of least-squares which, as discussed before, aims to minimize the residual sum of squares, given (in matrix notation) by:

$$\text{RSS} = \sum_{i=1}^{n}(\hat{\varepsilon}_i)^2 = \hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

Taking the derivatives and equating to zero, it can be shown that the least-squares estimator of β is:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

➲ It is shown that $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$ and $\text{Var}[\hat{\boldsymbol{\beta}}] = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$

Example

```
# Example with candidate gene

# vector of phenotypes
y<-matrix(c(3.4,3.7,3.2,2.9,2.5,3.1,2.6),nrow=7)

# incidence matrix (genotypes)
X<-matrix(c(1,1,1,1,1,1,1,-1,-1,-1,0,0,1,1,0,0,0,1,1,0,0),nrow=7)


#Matrix calculations
XX<-crossprod(X,X)
Xy<-t(X) %*% y

b <- solve(XX) %*% Xy
```

qtl

Example

```
# R code for the example with a QTL affecting a specific quantitative trait

y<-matrix(c(95.9, 108.0, 96.5, 92.9, 101.0, 94.5, 93.7, 89.8,
       101.2, 103.9, 85.9, 109.4, 105.7, 98.4, 84.1, 103.1,
       117.1, 95.2, 106.4, 104.7, 92.5, 123.9, 97.8, 100.5), nrow=24)

X<-matrix(c(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
   -1, -1, -1, -1, -1, -1, -1, -1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1), nrow=24)

b <- solve(t(X) %*% X) %*% t(X) %*% y

# Using package lm

qtl <- X[,2]
reg <- lm(y ~ qtl)
summary(reg)
```

Candidate gene

More on the LS Methodology

The estimator $\hat{\boldsymbol{\beta}}_{OLS} = \hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ is called ordinary least squares (OLS) estimator, and it is indicated only in situations with homoscedastic and uncorrelated residuals

If the residual variance is heterogeneous (i.e., $\mathrm{Var}(\boldsymbol{\varepsilon}_i) = \boldsymbol{\sigma}_i^2 = w_i\boldsymbol{\sigma}^2$), the residual variance matrix can be expressed as $\mathrm{Var}(\boldsymbol{\varepsilon}) = \mathbf{W}\boldsymbol{\sigma}^2$, where **W** is a diagonal matrix with the elements w_i, a better estimator of β is given by: $\hat{\boldsymbol{\beta}}_{WLS} = (\mathbf{X}^T\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{-1}\mathbf{y}$

which is generally referred to as weighted least squares (WLS) estimator.

Furthermore, in situations with a general residual variance-covariance matrix **V**, including correlated residuals, a generalized least squares (GLS) estimator $\hat{\boldsymbol{\beta}}_{GLS} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$ is obtained by minimizing the generalized sum of squares, given by: $\mathrm{GSS} = \boldsymbol{\varepsilon}^T\mathbf{V}^{-1}\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$

Example

Suppose an experiment on a completely randomized design to compare three diets on beef cattle growth. The results in terms of weight gain are given below :

Diets		
A	B	C
106	84	92
99	99	99
97	89	85
104	80	91
99	82	89
105		92
95		

Model: $y_{ij} = \mu_i + e_{ij}$

y_{ij}: weight gain observed on animal j of diet i

μ_i: mean of diet i

e_{ij}: residual term

i = 1, 2, 3 (Diets A, B and C)

j = 1, 2, …, n_i

(n_1 = 7, n_2 = 5, n_3 = 6)

LS Estimates

diets

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{15} \\ y_{16} \\ y_{17} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{24} \\ y_{25} \\ y_{31} \\ y_{32} \\ y_{33} \\ y_{34} \\ y_{35} \\ y_{36} \end{bmatrix} = \begin{bmatrix} 106 \\ 99 \\ 97 \\ 104 \\ 99 \\ 105 \\ 95 \\ 84 \\ 99 \\ 89 \\ 80 \\ 82 \\ 92 \\ 99 \\ 85 \\ 91 \\ 89 \\ 92 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{14} \\ e_{15} \\ e_{16} \\ e_{17} \\ e_{21} \\ e_{22} \\ e_{23} \\ e_{24} \\ e_{25} \\ e_{31} \\ e_{32} \\ e_{33} \\ e_{34} \\ e_{35} \\ e_{36} \end{bmatrix}$$

$$\mathbf{X}^{\mathrm{T}}\mathbf{X} = \begin{bmatrix} 7 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 6 \end{bmatrix}$$

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \begin{bmatrix} 1/7 & 0 & 0 \\ 0 & 1/5 & 0 \\ 0 & 0 & 1/6 \end{bmatrix}$$

$$\mathbf{X}^{\mathrm{T}}\mathbf{y} = \begin{bmatrix} 705 \\ 434 \\ 548 \end{bmatrix}$$

$$\hat{\boldsymbol{\mu}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \begin{bmatrix} 1/7 & 0 & 0 \\ 0 & 1/5 & 0 \\ 0 & 0 & 1/6 \end{bmatrix} \begin{bmatrix} 705 \\ 434 \\ 548 \end{bmatrix}$$

$$\hat{\boldsymbol{\mu}} = \begin{bmatrix} 705/7 \\ 434/5 \\ 548/6 \end{bmatrix} = \begin{bmatrix} 100.7 \\ 86.8 \\ 91.3 \end{bmatrix}$$

Maximum Likelihood Estimation

Likelihood Function: any function of the model parameters that is proportional to the density function of the data

Hence, to use a likelihood-based approach for estimating model parameters, some extra assumptions must be made regarding the distribution of the data

In the case of the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, if the residuals are assumed normally distributed with mean vector zero and variance-covariance matrix $\mathbf{V}$, i.e. $\boldsymbol{\varepsilon} \sim \mathrm{MVN}(\mathbf{0}, \mathbf{V})$, the response vector $\mathbf{y}$ is also normally distributed, with expectation $\mathrm{E}[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}$ and variance $\mathrm{Var}[\mathbf{y}] = \mathbf{V}$

Maximum Likelihood Estimation

The distribution of **y** has a density function given by:

$$p(\mathbf{y} \mid \boldsymbol{\beta}, V) = (2\pi)^{-n/2} \mid \mathbf{V} \mid^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}$$

so that the likelihood and the log-likelihood functions can be expressed respectively as:

$$L(\boldsymbol{\beta}, V) \propto \mid \mathbf{V} \mid^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}$$

and

$$l(\boldsymbol{\beta}, V) = \log[L(\boldsymbol{\beta}, V)] \propto -\frac{1}{2}\log \mid \mathbf{V} \mid -\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Maximum Likelihood Estimation

Assuming **V** known, the likelihood equations for **β** are given by taking the first derivatives of l(β,**V**) with respect to **β** and equating it to zero:

$$\frac{\partial l(\boldsymbol{\beta}, V)}{\partial \boldsymbol{\beta}} \equiv \frac{\partial}{\partial \boldsymbol{\beta}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0$$

from which the following system of equations is obtained:

$$\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y}$$

The maximum likelihood estimator (MLE) for **β** is given then by:

$$\mathrm{MLE}(\boldsymbol{\beta}) = \hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y}$$

Maximum Likelihood Estimation

If the inverse of $\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}$ does not exist, a generalized inverse $(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^-$ can be used to obtain a solution for the system of likelihood equations:

$$\boldsymbol{\beta}^0 = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^-\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$$

Note: Under normality the MLE coincides with the GLS estimator discussed previously. Similarly, in situations in which the matrix **V** is diagonal, or when **V** can be represented as $\mathbf{V} = \mathbf{I}_n\sigma^2$, the MLE coincides with the WLS and the OLS estimators, respectively

Maximum Likelihood Estimation

The expectation and the variance-covariance matrix of the MLE are given by:

$$E[\hat{\boldsymbol{\beta}}] = E[(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}] = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}E[\mathbf{y}] = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

$$\mathrm{Var}[\hat{\boldsymbol{\beta}}] = \mathrm{Var}[(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}] = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathrm{Var}[\mathbf{y}]\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}$$

$$= (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{V}\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}$$

As $\hat{\boldsymbol{\beta}}$ is a linear combination of the response vector **y**, we have that $\hat{\boldsymbol{\beta}} \sim \mathrm{MVN}(\boldsymbol{\beta}, (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1})$, from which confidence intervals (regions) and hypothesis testing regarding any (set of) element(s) of β can be easily obtained

The estimation of variance and covariance parameters will be discussed later

Maximum Likelihood Estimation

�’ **Note:** In the case of the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$,
with $\boldsymbol{\varepsilon} \sim \mathrm{MVN}(\mathbf{0}, \mathbf{I}\sigma^2)$, it can be shown that:

$$\begin{cases} \hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \quad \rightarrow \quad \hat{\boldsymbol{\beta}} \sim \mathrm{N}(\boldsymbol{\beta}, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2) \\[2mm] \hat{\sigma}^2 = \dfrac{1}{n}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \dfrac{1}{n} \| \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \|^2 \end{cases}$$

$$\hat{\sigma}^2 \sim \sigma^2 \frac{\chi^2_{(n-k)}}{n} \qquad \left(E[\hat{\sigma}^2] = \frac{n-k}{n}\sigma^2 \right)$$

$$\tilde{\sigma}^2 = s^2 = \frac{n}{n-k}\hat{\sigma}^2 = \frac{1}{n-k} \| \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \|^2 \quad \rightarrow \quad \tilde{\sigma}^2 \sim \sigma^2 \frac{\chi^2_{(n-k)}}{n-k}$$

Test for Independence
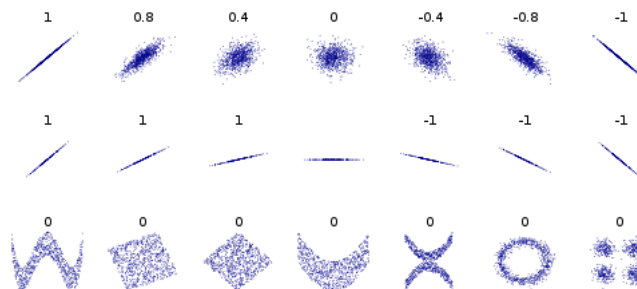
Guilherme J. M. Rosa
University of Wisconsin-Madison

Test for Independence

"I can't move in with my parents. They moved in with my grandparents."

Correlation and Dependence

If two variables are independent, their corresponding Pearson's correlation coefficient is 0, but the converse is not true. However, in the special case of joint multivariate normal distribution, uncorrelatedness is equivalent to independence.

Testing for Independence

- Two random variables are independent if the realization of one does not affect the probability distribution of the other
- The concept of independence extends to dealing with collections of more than two events or random variables (pairwise or mutually independent)
- Testing for independence: chisquare-based test (multinomial variables), correlation test (multivariate normal), …
- More generally: mutual information (parametric, semiparametric and permutation tests)

Multinomial Data

Variable A, with I categories: A_1, A_2,...,A_I
Variable B, with J categories: B_1, B_2,...,B_J

(I x J) Contingency Table

	B_1	B_2	$\cdots$	B_J	Totals
A_1	n_{11}	n_{12}	$\cdots$	n_{1J}	n_{1+}
A_2	n_{21}	n_{22}	$\cdots$	n_{2J}	n_{2+}
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
A_I	n_{I1}	n_{I2}	$\cdots$	n_{IJ}	n_{I+}
Totals	n_{+1}	n_{+2}	$\cdots$	n_{+J}	n_{++}

n_{ij}: observed count in category A_iB_j

$n_{i+} = n_{i1} + ... + n_{iJ}$ and $n_{+j} = n_{1j} + ... + n_{Ij}$

(I x J) Contingency Tables

n_{ij}: observed frequencies

n_{i+} and n_{+j}: marginal frequencies

Under the assumption of independency between A and B:

Expected frequencies: $E_{ij} = \dfrac{n_{i+} \times n_{+j}}{n_{++}}$

$\varphi = (I-1)(J-1)$
degrees of freedom

Statistical test for Independence:

Pearson chi-squared: $X^2 = \displaystyle\sum_{i=1}^{I} \sum_{j=1}^{J} \dfrac{(n_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{\varphi}$

Deviance statistics (LRT): $G^2 = 2\displaystyle\sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \log\left(\dfrac{n_{ij}}{E_{ij}}\right)$

Three-Way Tables

Variable A, with I categories: $A_1, A_2, ..., A_I$
Variable B, with J categories: $B_1, B_2, ..., B_J$
Variable C, with K categories: $C_1, C_2, ..., C_K$

		C		
A	B	C_1	$\cdots$	C_K
A_1	B_1	n_{111}	$\cdots$	n_{11K}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
A_1	B_J	n_{1J1}	$\cdots$	n_{1JK}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
A_I	B_1	n_{I11}	$\cdots$	n_{I1K}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
A_I	B_J	n_{IJ1}	$\cdots$	n_{IJK}

Saturated Model:

$\mu_{ijk} = E(n_{ijk}) = n_{+++}\hat{\pi}_{ijk}$

$= n_{+++} \dfrac{n_{ijk}}{n_{+++}}$

Example

A: Defendant's Race
B: Victim's Race
C: Death Penalty

Defendant's Race	Victim's Race	Death Penalty	
		Yes	No
White	White	19	132
	Black	0	9
Black	White	11	52
	Black	6	97

Is there an association between death penalty, defendant's race and victim's race?

(Agresti, 1990)

Marginal Tables

Defendant's Race x Death Penalty (A x C)

Defendant's Race	Death Penalty	
	Yes	No
White	19	141
Black	17	149

Marginal Odds Ratio:

$$\hat{\theta}_{AC} = \frac{19 \times 149}{17 \times 141} = 1.18$$

$$n_{i+k} = \sum_{j=1}^{J} n_{ijk}$$

Same partial tables for [Defendant's Race x Victim's Race] and [Victim's Race x Death Penalty]

Partial Tables

Conditional distribution, e.g. A x C tables
for particular values of B

Victim = White $(A \times C \mid B_1)$

Defendant's Race	Death Penalty	
	Yes	No
White	19	132
Black	11	52

Conditional Odds: $\hat{\theta}_{AC(B_1)} = \dfrac{19 \times 52}{11 \times 132} = 0.6804$

Partial Tables

Victim = Black $(A \times C \mid B_2)$

| Defendant's Race	Death Penalty	
	Yes	No
White	0	9
Black	6	97

Conditional Odds: $\hat{\theta}_{AC|B_2} = \dfrac{0.5 \times 97.5}{6.5 \times 9.5} = 0.7895$

Ad-hoc method with zero counts: add 0.5 to each cell value

Note: If conditional odds is different from 1 for
at least one level of B, we say that variables
A and C are conditionally associated.

Simpson's Paradox

Marginal and partial (conditional) associations
in opposite direction

A, B continuous
C binary
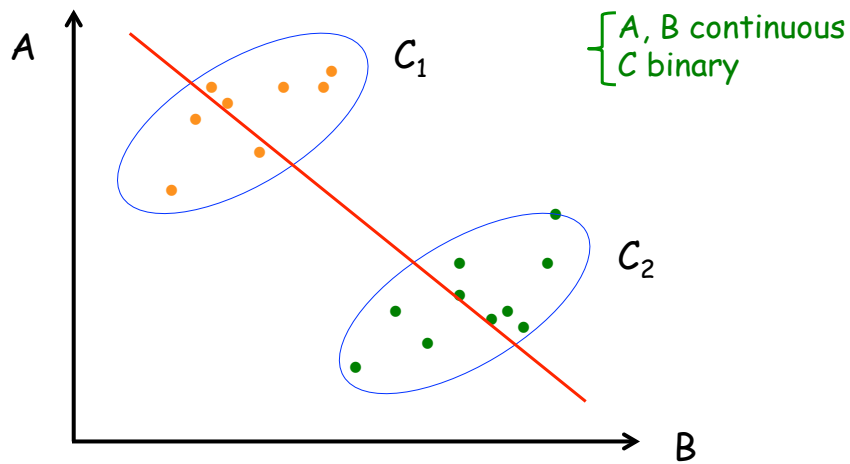
UC Berkeley Gender Bias Story

- One of the best-known examples of Simpson's paradox
- Potential gender bias among graduate school admissions at UC-Berkeley in 1973
- Admission figures showed that men applying were more likely than woman to be admitted:

Table 1. Decisions on applications to Graduate Division for fall 1973, by sex of applicant—naive aggregation. Expected frequencies are calculated from the marginal totals of the observed frequencies under the assumptions (1 and 2) given in the text. $N = 12,763$, $\chi^2 = 110.8$, d.f. $= 1$, $P = 0$ (18).

	Outcome				Difference	
Applicants	Observed		Expected			
	Admit	Deny	Admit	Deny	Admit	Deny
Men	3738	4704	3460.7	4981.3	277.3	− 277.3
Women	1494	2827	1771.3	2549.7	− 277.3	277.3

- Admission rates:
  Men: 3738/8442 = 0.443, Women: 1494/4321 = 0.346

UC Berkeley Gender Bias Story

- However, when examining the 85 individual departments, it appeared that 6 were significantly biased against men, whereas only 4 were significantly biased against women.
- In fact, the pooled and corrected data showed a small but statistically significant bias in favor of women.
- Data from the six largest departments:

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	373	6%	341	**7%**

UC Berkeley Gender Bias Story

- The research by Bickel et al (1975) concluded that women tended to apply to competitive departments with low rates of admission even among qualified applicants (such as English Department), whereas men tended to apply to less-competitive departments with high rates of admission among qualifies applicants (such as engineering and chemistry).

Bickel PJ, Hammel EA and O'Connell JW. Sex bias in graduate admissions: data from Berkeley. Measuring bias is harder than usually assumed, and the evidence is sometimes contrary to expectation. *Science* 187: 398-404, 1975.
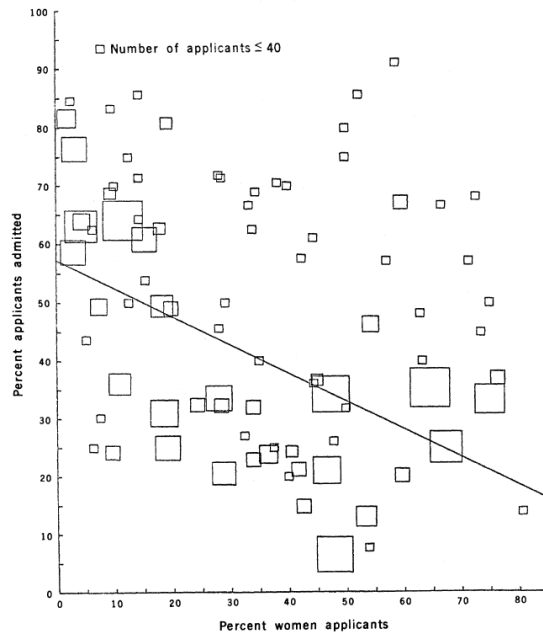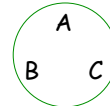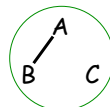
Fig. 1. Proportion of applicants that are women plotted against proportion of applicants admitted, in 85 departments. Size of box indicates relative number of applicants to the department.

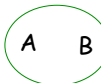# Models of Independence and Associations in 3-Way Tables

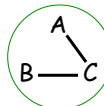1. Mutual (complete) independence (A,B,C)
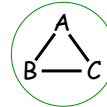
2. Joint independence (AB,C)

3. Marginal independence (A,B)

4. Conditional independence (AC,BC)

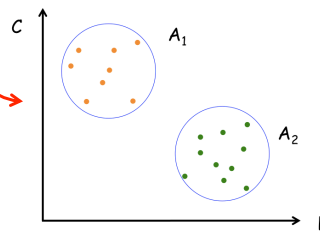5. Homogeneous association (AB,AC,BC)

Conditional Independence

- Three possible models with 3 variables:
    (AB,AC), (AB,BC) and (AC,BC)

- Take (AB,AC) as an example:
    $P(B = j, C = k \mid A = i) = P(B = j \mid A = i) \times P(C = k \mid A = i)$

- Odds-ratios: B x C for each level of A

$$\begin{cases} H_0: \theta_{BC(A=i)} = 1 \text{ for every } i \\ H_a: \text{at least one } \theta_{BC(A=i)} \neq 1 \end{cases}$$

- Under H_0, cell probabilities are equal to:

$$\begin{aligned} \pi_{ijk} &= P(A = i)P(B = j, C = k \mid A = i) \\ &= P(A = i)P(B = j \mid A = i)P(C = k \mid A = i) \\ &= \pi_{i++}\pi_{jli}\pi_{kli} \end{aligned}$$

- ML estimates: $\begin{cases} \hat{\pi}_{i++} = n_{i++} / n_{+++} \\ \hat{\pi}_{jli} = n_{ij+} / n_{i++} \\ \hat{\pi}_{kli} = n_{i+k} / n_{i++} \end{cases}$

- Estimated expected frequencies: $\hat{E}_{ijk} = \dfrac{n_{ij+} \times n_{i+k}}{n_{i++}}$

➜ Overall X^2 or G^2 by summing individual test statistics across levels of A. Degrees of freedom: I(J-1)(K-1). Notice: Conditional independence can be rejected if any of the partial tests is rejected.

More Complex Data

→ Alternative models for more complex data: log-linear model, logistic regression, generalized linear mixed models, etc.

Explanatory variables · Response variable

A	B	Response	
		Yes	No
A_1	B_1	2	0
	B2	0	1
A_2	B1	3	2
	B2	1	0

A	B	y
1	1	1
1	1	1
1	2	0
2	1	1
2	1	1
2	1	1
2	1	0
2	1	0
2	2	1

Generalized linear model: binary outcome, logit, $\eta = \mu + A + B$

Gaussian Distribution

- Marginal correlation : $\rho_{XY} = \dfrac{\sigma_{XY}}{\sqrt{\sigma_X^2 \sigma_Y^2}}$

Inference:

$$\hat{\rho}_{XY} = \frac{\sum_{i=1}^{n}(x_i - \hat{\mu}_X)(y_i - \hat{\mu}_Y)}{\sqrt{\sum_{i=1}^{n}(x_i - \hat{\mu}_X)^2 \sum_{i=1}^{n}(y_i - \hat{\mu}_Y)^2}} \rightarrow t = \hat{\rho}_{XY}\sqrt{\frac{n-2}{1-\hat{\rho}_{XY}^2}} \sim t_{(\varphi = n-2)}$$

Fisher transformation:

$$z = \operatorname{arctanh}(\hat{\rho}_{XY}) = \frac{1}{2}\ln\left(\frac{1+\hat{\rho}_{XY}}{1-\hat{\rho}_{XY}}\right) \rightarrow \sqrt{n-3} \times |z| \overset{H_0 : \rho = 0}{\sim} N(0,1)$$

Gaussian Distribution

- Partial correlation: $\rho_{XY \cdot Z} = \dfrac{\rho_{XY} - \rho_{XZ}\rho_{ZY}}{\sqrt{1 - \rho_{XZ}^2}\ \sqrt{1 - \rho_{ZY}^2}}$

Conditioning on multiple variables: $P = \Sigma^{-1} \rightarrow \rho_{y_i y_j \cdot \text{others}} = -\dfrac{p_{ij}}{\sqrt{p_{ii}p_{jj}}}$

where:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p} & \sigma_{2p} & \cdots & \sigma_p^2 \end{bmatrix}$$

and $\begin{cases} \sigma_i^2 : \text{ variances} \\ \\ \sigma_{ij} \quad = \rho_{ij}\sqrt{\sigma_i^2 \sigma_j^2} : \\ \qquad\qquad\qquad \text{covariances} \end{cases}$

Mutual Information

- Mutual information (MI): measures of the mutual dependence between two random variables X and Y, or the "amount of information" (in shannons, or bits) obtained about one variable, through the other variable.

- MI determines how similar the joint distribution p(X,Y) is to the products of factored marginal distribution p(X)p(Y).