Fundamental Probability and Statistics

"There are known knowns. These are things we know that we know. There are known unknowns. That is to say, there are things that we know we don't know. But there are also unknown unknowns. There are things we don't know we don't know," Donald Rumsfeld

Probability Theory

Reference: G.R. Grimmett and D.R. Stirzaker, *Probability and Random Processes*, Oxford Science Publications, 1997

Probability Space: $(\Omega, \mathcal{F}, P)$

Ω – Sample Space – set of all possible outcomes of an experiment

$\mathcal{F}$ – σ-field of subsets of Ω that contains all events of interest

$P : \mathcal{F} \to [0,1]$: probability or measure that satisfies

$$\text{(i)} \quad P(\emptyset) = 0$$

$$\text{(ii)} \quad P(\Omega) = 1$$

$$\text{(iii)} \quad A_i \in \mathcal{F} \text{ and } A_i \cap A_j = \emptyset \text{ implies } P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

Example: Toss possibly biased coin once

$$\Omega = \{H, T\}$$

$$\mathcal{F} = \{\emptyset, H, T, \Omega\}$$

Take

$$P(\emptyset) = 0\,,\ P(H) = p\,,\ P(T) = 1 - p\,,\ P(\Omega) = 1$$

Note: Fair coin if p = 1/2

Probability Theory

Example: Two coins tossed possibly multiple times and outcome is ordered pair

$$\Omega = \{(H,H),(T,H),(H,T),(T,T)\}$$

$$\mathcal{F} = \{\emptyset,(H,H),(T,H),(H,T),(T,T),\Omega,\{(H,H),(T,H)\},\cdots\}$$

Let

$$A = \{(H,H),(T,H)\}$$

$$B = \{(H,H),(H,T)\}$$

Then

$$P(A) = \frac{1}{2}\ ,\ P(B) = \frac{1}{2}$$

$$P(A \cap B) = \frac{1}{4}\ ,\ P(A \cup B) = \frac{3}{4}$$

Definition: Events A and B are independent if

$$P(A \cap B) = P(A)P(B)$$

Random Variables and Distributions

Definition: A random variable is a function $X : \Omega \to \mathbb{R}$ with the property that $\{\omega \in \Omega \,|\, X(\omega) \leq x\} \in \mathcal{F}$ for each $x \in \mathbb{R}$.

Definition: The value of a random variable X at a point $\omega \in \Omega$

$$x = X(\omega)$$

is a realization of X.

Definition: Associated with every random variable X is a cumulative distribution function $F : \mathbb{R} \to [0,1]$ given by

$$\begin{aligned} F(x) &= P\{\omega \in \Omega \,|\, X(\omega) \leq x\} \\ &= P(X) \leq x \end{aligned}$$

Example: Take $\Omega = \{(H,H),(T,H),(H,T),(T,T)\}$ and define $X(\omega)$ as number of heads

$$\begin{aligned} &X(H,H) = 2 \\ &X(H,T) = X(T,H) = 1 \\ &X(T,T) = 0 \end{aligned} \qquad F(x) = \begin{cases} 0 & \text{if} \quad x < 0 \\ \frac{1}{4} & \text{if} \quad 0 \leq x < 1 \\ \frac{3}{4} & \text{if} \quad 1 \leq x < 2 \\ 1 & \text{if} \quad x \geq 2 \end{cases}$$

Distributions and Densities

Definition: A random variable X is discrete if it takes values in a countable subset $\{x_1, x_2, \cdots\}$, only, of $\mathbb{R}$.

Definition: X is continuous if its distribution function can be expressed as

$$F(x) = \int_{-\infty}^{x} f(s)ds \ , \ x \in \mathbb{R}$$

for some integrable function $f : \mathbb{R} \to [0, \infty)$ where f is called the probability density function (PDF) of X.

Definition: The probability mass function of a discrete random variable X is a function $f : \mathbb{R} \to [0, 1]$ given by $f(x) = P(X = x)$.

PDF Properties:

(i) $f(x) \geq 0$

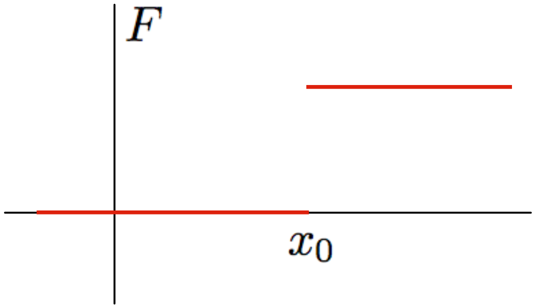
(ii) $\int_{\mathbb{R}} f(x)dx = 1$

(iii) $P(x_1 \leq X \leq x_2) = F(x_2) - F(x_1) = \int_{x_1}^{x_2} f(x)dx$

Density Properties

Example:

$$X(x) = \begin{cases} 0 & , \quad x < x_0 \\ 1 & , \quad x \geq x_0 \end{cases}$$



$$F(x) = \begin{cases} 0 & , \quad x < x_0 \\ 1 & , \quad x \geq x_0 \end{cases} \Rightarrow f(x) = \delta(x - x_0)$$

Example:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \quad , \; -\infty < x < \infty$$

$$F(x) = \int_{-\infty}^{x} f(s)ds = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$$

where

$$\text{erf}(x) = \frac{2}{\pi}\int_0^x e^{-s^2} ds$$

Note: $X \sim N(\mu, \sigma^2)$

Density Properties

Additional Properties:

(i) Mean, first moment or expected value

$$\mu = E(X) = \int_{\mathbb{R}} x f(x) dx$$

(ii) n^{th} moment

$$E(X^n) = \int_{\mathbb{R}} x^n f(x) dx$$

(iii) Second central moment (difference between X and μ)

$$\sigma^2 = \mathrm{var}(X) = E[(X-\mu)^2] = \int_{\mathbb{R}} (x-\mu)^2 f(x) dx$$

Note:

$$\sigma^2 = E(X^2) - \mu^2$$

(iv) The covariance of X and Y is

$$\begin{aligned} \mathrm{cov}(X,Y) &= E[(X - E(X))(Y - E(Y))] \\ &= E(XY) - E(X)E(Y) \end{aligned}$$

Note: X and Y are uncorrelated if cov$(X,Y) = 0$

Multivariate Distributions

Note: Important for longitudinal data

$$X = (X_1, X_2, \cdots, X_n)$$

$$\Rightarrow X : \Omega \to \mathbb{R}^n \ \ \text{Random Vector}$$

Joint CDF: $F : \mathbb{R}^n \to [0,1]$ by

$$\begin{aligned} F(x_1, \cdots, x_n) &= P\{w \in \Omega \,|\, X_j(\omega) \leq x_i \,,\, j = 1, \cdots, n\} \\ &= P(X \leq x) \end{aligned}$$

Joint Density (if it exists): $f : \mathbb{R}^n \to [0, \infty)$

$$F(x) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(s_1, \cdots, s_n) d_{s_1} \cdots d_{s_n}$$

Example: Let $X \sim N(\mu, V)$

$$f(x) = \frac{1}{\sqrt{(2\pi)^n |V|}} \exp\left[-\frac{1}{2}(x-\mu)V^{-1}(x-\mu)^T\right]$$

Multivariate Distributions

Example: Let $X \sim N(\mu, V)$

$$f(x) = \frac{1}{\sqrt{(2\pi)^n |V|}} \exp\left[-\frac{1}{2}(x-\mu)V^{-1}(x-\mu)^T\right]$$

Note:

- $E(X) = \mu$ so $E(X_j) = \mu_j$
- $V = (v_{ij})$ is the covariance matrix since $v_{ij} = \mathrm{cov}(X_i, X_j)$. This is often written

$$V = E((X-\mu)^T(X-\mu))$$

Note:

$$V = \mathrm{cov}(X) = \begin{bmatrix} \mathrm{var}(X_1) & \mathrm{cov}(X_1, X_2) & \cdots & \mathrm{cov}(X_1, X_n) \\ \mathrm{cov}(X_1, X_2) & \mathrm{var}(X_2) & \cdots & \mathrm{cov}(X_2, X_n) \\ \vdots & \vdots & & \vdots \\ \mathrm{cov}(X_1, X_n) & \mathrm{cov}(X_2, X_n) & \cdots & \mathrm{var}(X_n) \end{bmatrix}$$

Multivariate Distributions

Definition: The marginal distribution functions of X and Y are

$$\begin{aligned} F_X(x) &= P(X \leq x) \\ &= \int_{-\infty}^{x} \left(\int_{-\infty}^{\infty} f(u,y)dy \right) du \end{aligned}$$

with a similar definition for $F_Y(y)$.

Definition: Marginal density function of X

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy$$

Definition: X and Y are independent if and only if

$$F(x,y) = F_X(x)F_Y(y)$$

or

$$f(x,y) = f_X(x)f_Y(y)$$

Note: X and Y are independent $\Rightarrow$ $\mathrm{cov}(X,Y) = 0$ $\Leftrightarrow$ X and Y are uncorrelated

Estimators and Estimates

Definition: An estimator is a function or procedure for deriving an estimate from observed data. An estimator is a random variable whereas an estimate is a real number.

Example: Suppose we want to estimate the variance $\mathrm{var}(Y) = \sigma^2$ of a random variable Y using a sample of n independent observations. Consider two statistics.

$$S^2 = \frac{1}{n}\sum_{j=1}^{n}(Y_j - \mu)^2$$

$$T^2 = \frac{1}{n}\sum_{j=1}^{n}(Y_j - \bar{Y})^2 \text{ where } \bar{Y} = \frac{1}{n}[Y_1 + Y_2 + \cdots + Y_n]$$

Now

$$E(S^2) = \frac{1}{n}\sum_{j=1}^{n}E(Y_j - \mu)^2 = \frac{1}{n}\sum_{j=1}^{n}\sigma^2 = \sigma^2 \quad \text{but } \mu \text{ is often unknown}$$

$$E(T^2) = \frac{n-1}{n}\sigma^2 \quad \text{so biased}$$

Unbiased estimator: $\mathcal{T}^2 = \frac{1}{n-1}\sum_{j=1}^{n}(Y_j - \bar{Y})^2$

Other Estimators

Commonly Employed Estimators:

- Maximum likelihood
- Bayes estimators
- Particle filter (Sequential Monte Carlo (SMC))
- Markov chain Monte Carlo (MCMC)
- Kalman filter
- Wiener filter

Linear Regression

Consider

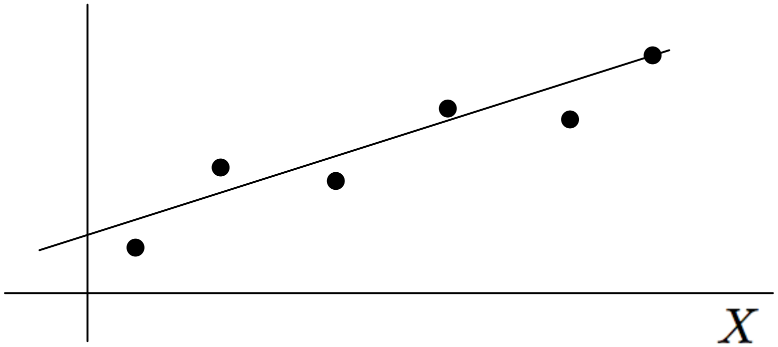
$$y = X\beta + \varepsilon$$

where

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad , \quad X = \begin{bmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & & \vdots \\ X_{n1} & \cdots & X_{np} \end{bmatrix} \quad , \quad \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad , \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Observations   Design Matrix   Unknown Parameters   Errors

Example: $y_j = (\beta_0 + \beta_1 X_j) + \varepsilon_j \ , \ j = 1, \cdots, n$



Linear Regression

Statistical Model:

$$Y = X\beta_0 + \varepsilon$$

Assumptions:

(i) $E(\varepsilon) = 0$

(ii) ε iid (independent and identically distributed)

$$\Rightarrow \quad \mathrm{var}(\varepsilon_j) = \sigma_0^2$$

$$E[(\varepsilon_i - E(\varepsilon_i))(\varepsilon_j - E(\varepsilon_j))] = \mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0 \text{ for } i \neq j$$

(iii*) If $\varepsilon \sim N(0, \sigma_0^2)$ and hence $Y_j \sim N(X\beta_0, \sigma_0^2)$, maximum likelihood estimator (MLE) provides expression for β_0 and σ_0^2.

(iv*) Generalized least squares employed if nonconstant variance.

Goals:

(1) Contruct a 'good' estimator B for β.

(2) Construct an estimator S^2 for σ_0^2.

Least Squares Problem

Minimize

$$T = (y - X\beta)^T (y - X\beta)$$

Note: General result for quadratic forms

$$Q = A^T \Phi A$$

$$\Rightarrow \nabla_\beta Q = 2(\nabla_\beta A^T)\Phi A$$

Thus

$$\nabla_\beta T = 2[\nabla_\beta (y - X\beta)^T][y - X\beta]$$

where

$$\nabla_\beta (y - X\beta)^T = -\nabla_\beta \beta^T X^T = -X^T$$

Least Squares Estimate: $\hat{b} = (X^T X)^{-1} X^T y$

Least Squares Estimator: $B = (X^T X)^{-1} X^T Y$

Note: $B = \arg\min_\beta (Y - X\beta)^T (Y - X\beta)$

Parameter Estimator Properties

Estimator Mean:

$$\begin{aligned} E(B) &= E\left[(X^T X)^{-1} X^T Y\right] \\ &= (X^T X)^{-1} X^T E(Y) \\ &= \beta_0 \end{aligned}$$

Here β_0 denotes the 'true' parameters that generate Y

Estimator Covariance: Let $A = (X^T X)^{-1} X^T$

$$\begin{aligned} \mathrm{cov}(B) &= E[(B - \beta_0)(B - \beta_0)^T] \\ &= E[(\beta_0 + A\varepsilon - \beta_0)(\beta_0 + A\varepsilon - \beta_0)^T] \\ &= AE(\varepsilon\varepsilon^T)A^T \\ &= \sigma_0^2 (X^T X)^{-1} X^T \cdot X(X^T X)^{-T} \\ &= \sigma_0^2 (X^T X)^{-1} \end{aligned}$$

$$\begin{aligned} \text{since } B &= AY \\ &= A(X\beta_0 + \varepsilon) \\ &= \beta_0 + A\varepsilon \end{aligned}$$

Variance Estimator Properties

Goal: Construct an estimator S^2 for σ_0^2

Residual: $R = Y - XB$

Variance Estimator: $S^2 = \frac{R^T R}{n-p}$

Note:

$$\begin{aligned} R &= Y - X(X^T X)^{-1} X^T Y \\ &= (I_n - H)Y \end{aligned}$$

where

$$H = X(X^T X)^{-1} X^T$$

Properties of H:

$$H^T = H \quad \text{(symmetric)}$$

$$H^2 = H \quad \text{(idempotent)}$$

$$(I_n - H)^2 = (I_n - H)$$

$$(I_n - H)X = X - X(X^T X)^{-1} X^T X = 0$$

Variance Estimator Properties

Note:

$$\begin{aligned} R &= (I_n - H)Y \\ &= (I_n - H)(X\beta_0 + \varepsilon) \\ &= (I_n - H)\varepsilon \end{aligned}$$

so

$$\begin{aligned} R^T R &= [(I_n - H)\varepsilon]^T [(I_n - H)\varepsilon] \\ &= \varepsilon^T (I_n - H)\varepsilon \end{aligned}$$

Thus

$$R^T R = \sum_{i=1}^{n} \sum_{j=1}^{n} g_{ij}\varepsilon_i\varepsilon_j$$

so

$$\begin{aligned} E(R^T R) &= \sum_{i=1}^{n} \sum_{j=1}^{n} g_{ij} E(\varepsilon_i\varepsilon_j) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} g_{ij} \mathrm{cov}(\varepsilon_i, \varepsilon_j) \ \text{ since } E(\varepsilon_i) = E(\varepsilon_j) = 0 \\ &= \sum_{i=1}^{n} g_{ii} \mathrm{var}(\varepsilon_i) = \sigma_0^2 \mathrm{tr}(I_n - H) \end{aligned}$$

Variance Estimator Properties

Note:

$$\mathrm{tr}(A+B) = \mathrm{tr}(B+A)$$

$$\mathrm{tr}(AB) = \mathrm{tr}(BA)$$

Thus

$$\begin{aligned}\mathrm{tr}(I_n - H) &= n - \mathrm{tr}\left[(X^T X)^{-1} X^T X\right] \\ &= n - p\end{aligned}$$

Unbiased Estimator:

$$S^2 = \frac{R^T R}{n-p}$$

Unbiased Estimate:

$$\hat{s}^2 = \frac{1}{n-p} \sum_{j=1}^{n} \left[y_j - (X\hat{b})_j\right]^2$$

Parameter Estimator Properties

Properties of B: Assume ε_j are iid with $E(\varepsilon_j) = 0$ constant variance σ_0

1. $E(B) = \beta_0$
2. $\text{cov}(B) = \sigma_0^2(X^TX)^{-1}$
3. Suppose in addition that $\varepsilon_j \sim N(0, \sigma_0^2)$. Then $B \sim N_p\left(\beta_0, \sigma_0^2(X^TX)^{-1}\right)$.
4. Suppose ε_j are iid with $E(\varepsilon_j) = 0$, constant variance σ_0 and unknown distribution. Then in the limit $n \to \infty$,

   $$B \sim N_p\left(\beta_0, \sigma_0^2(X^TX)^{-1}\right)$$

   Proof: Involves Central Limit Theorem

Central Limit Theorem:

Let $X_1, X_2, \cdots$ be a sequence of iid random variables with mean μ and variance σ. Then

$$\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \xrightarrow{D} N(0,1) \text{ as } n \to \infty$$

where

$$S_n = X_1 + X_2 + \cdots + X_n$$

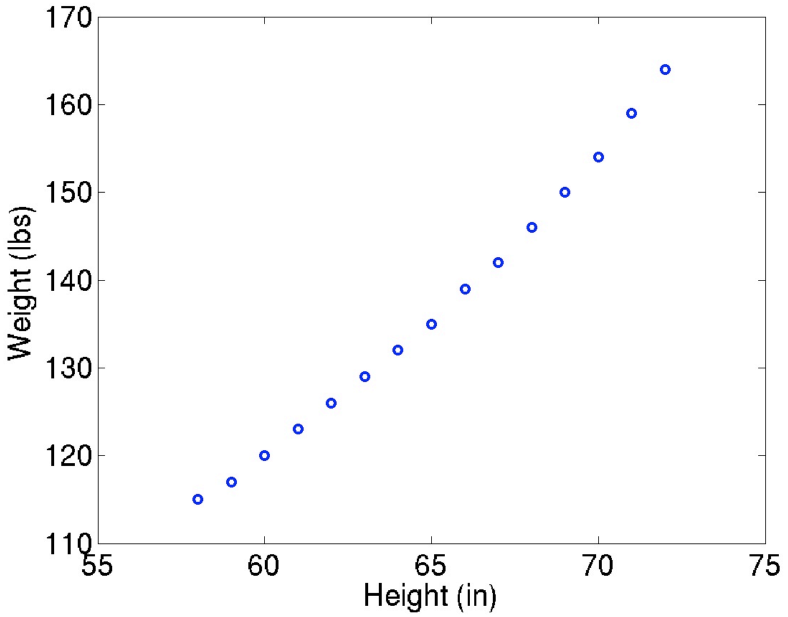
Example

Example: Consider the height-weight data from the *1975 World Almanac and Book of Facts*

Height (in)	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72
Weight (lbs)	115	117	120	123	126	129	132	135	139	142	146	150	154	159	164

Consider the model

$$y_j = \beta_0 + \beta_1(x_j/12) + \beta_2(x_j/12)^3 + \varepsilon_j$$



Example

Here

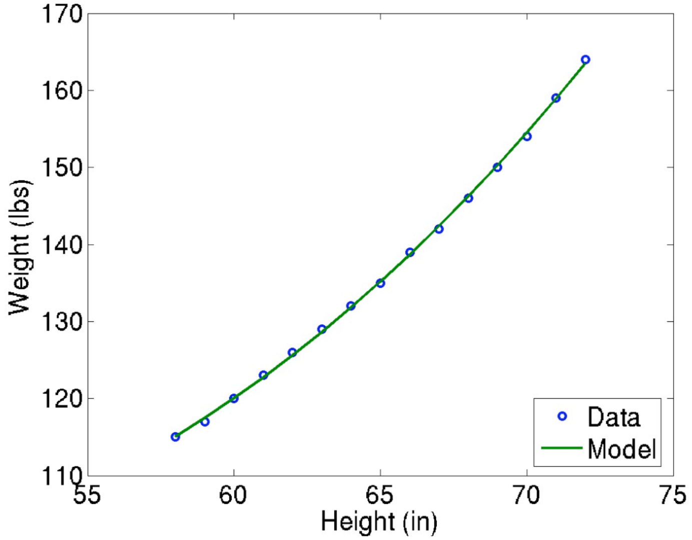
$$X = \begin{bmatrix} 1 & 4.83 & 112.91 \\ 1 & 4.92 & 118.85 \\ 1 & 5.00 & 125.00 \\ 1 & 5.08 & 131.35 \\ 1 & 5.17 & 137.92 \\ 1 & 5.25 & 144.70 \\ 1 & 5.33 & 151.70 \\ 1 & 5.42 & 158.93 \\ 1 & 5.50 & 166.38 \\ 1 & 5.58 & 174.05 \\ 1 & 5.67 & 181.96 \\ 1 & 5.75 & 190.11 \\ 1 & 5.83 & 198.50 \\ 1 & 5.92 & 207.12 \\ 1 & 6.00 & 216.00 \end{bmatrix}$$

Least Square Estimate:

$$(X^T X)\hat{b} = X^T y$$

$$\Rightarrow \begin{bmatrix} 15 & 81 & 2415 \\ 81 & 442 & 13255 \\ 2415 & 13255 & 404171 \end{bmatrix} \begin{bmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \hat{b}_2 \end{bmatrix} = \begin{bmatrix} 2051 \\ 11190 \\ 337421 \end{bmatrix}$$

$$\Rightarrow \begin{aligned} \hat{b}_0 &= 146.59 \\ \hat{b}_1 &= -23.77 \\ \hat{b}_2 &= 0.74 \end{aligned}$$



Note: $\text{cond}(X^T X) = 8.4 \times 10^8$

Note: $\text{eig}(H) = \text{eig}(X(X^T X)^{-1} X^T) = 0, 1$

Example

Variance Estimate:

$$\hat{s}^2 = \frac{\hat{r}^T \hat{r}}{n-p} = \frac{1}{n-p} \sum_{j=1}^{n} \left[y_j - (X\hat{b})_j \right]^2 = 0.131$$

Parameter Covariance Estimate:

$$\widehat{\mathrm{cov}(B)} = \hat{s}^2 (X^T X)^{-1}$$

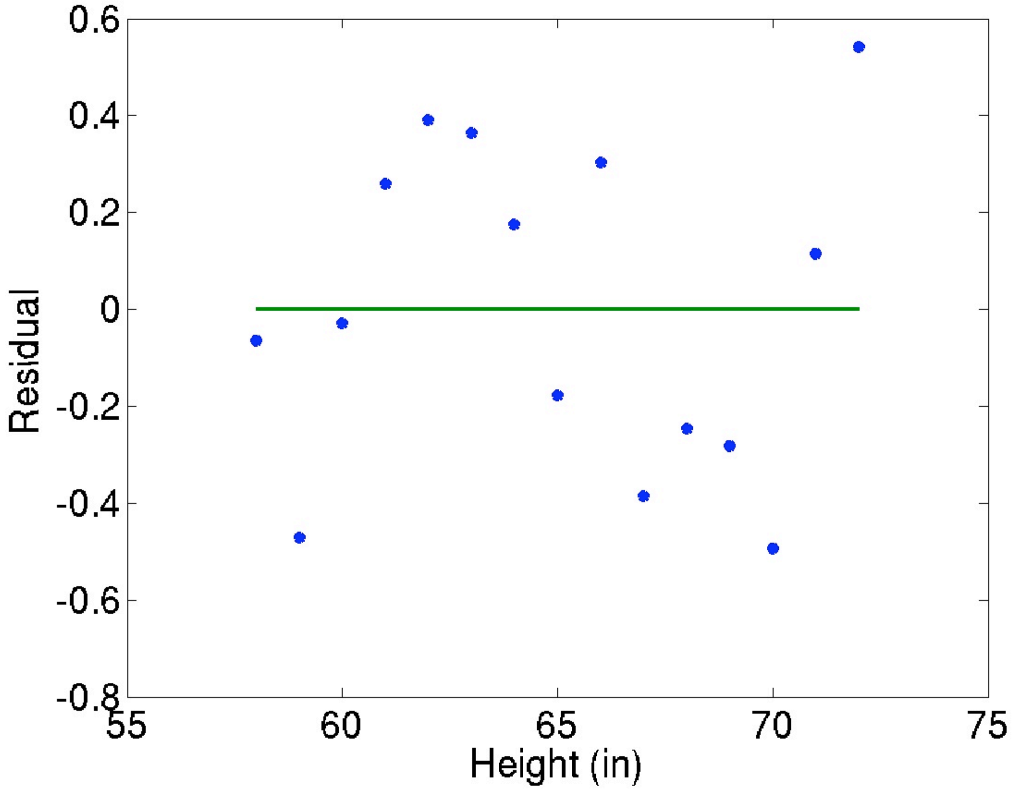
Note: This yields variances and standard deviations for parameter estimates

$$\hat{b}_0 = 146.59 \pm 15.9$$
$$\hat{b}_1 = -23.77 \pm 4.4$$
$$\hat{b}_2 = 0.74 \pm 0.05$$

Goal: Can we additionally compute confidence intervals? Yes, but we need a little more statistics.

Example

Hypothesis: One way to check the hypothesis of iid is to plot the residuals



Random Variables Related to the Normal

Chi-Square Random Variables:

If $X \sim N(0,1)$, then the random variable $Y = X^2$ is said to have a χ^2-distribution with 1 dof. Furthermore, if $Y_i, i = 1, \cdots, n$ are independent χ^2 random variables with 1 dof, then their sum

$$\chi^2(n) = \sum_{i=1}^{n} Y_i$$

is a χ^2 random variable with n dof.

T Random Variables:

If $Z \sim N(0,1)$ and $\chi^2(n)$ is a chi-square random variable with n dof that is independent from Z, then

$$T = \frac{Z}{\sqrt{\chi^2(n)/n}}$$

has a (Student's) t-distribution with n dof.

Variance Estimator Properties

Assumption: Assume that $\varepsilon_j \sim N(0, \sigma_0^2)$ are iid (use asymptotic results if not normal).

Property 1: $(n-p)S^2/\sigma_0^2$ has a χ^2-distribution with $n-p$ dof.

Property 2: If δ_k denotes the k^{th} diagonal element of $(X^TX)^{-1}$, then

$$T_k = \frac{B_k - \beta_{0k}}{S\sqrt{\delta_k}}$$

has a t-distribution with $n-p$ dof.

Verification of 1: Recall that

$$\begin{aligned}
(n-p)S^2/\sigma_0^2 &= \frac{1}{\sigma_0^2} R^T R \\
&= \frac{1}{\sigma_0^2} \varepsilon^T (I_n - H)\varepsilon \\
&= \frac{1}{\sigma_0^2} \langle \varepsilon, (I_n - H)\varepsilon \rangle \\
&= \frac{1}{\sigma_0^2} \langle \varepsilon, Q\Lambda Q^T \varepsilon \rangle \quad , \; Q \text{ orthogonal} \\
&= \frac{1}{\sigma_0^2} \langle Q^T\varepsilon, \Lambda Q^T \varepsilon \rangle
\end{aligned}$$

Variance Estimator Properties

Note: The eigenvalues of H are either 0 or 1 and $\mathrm{tr}(I_n - H) = \mathrm{rank}(I_n - H) = n - p$. Thus

$$\Lambda = \begin{bmatrix} I_{n-p} & 0 \\ 0 & 0 \end{bmatrix}$$

Note: Q^T orthogonal implies that $q = Q^T \varepsilon \sim N(0, \sigma_0)$ since $\varepsilon \sim N(0, \sigma_0)$. Thus

$$\frac{\langle q, \Lambda q \rangle}{\sigma_0^2} = \sum_{i=1}^{n-p} \frac{q_i^2}{\sigma_0^2}$$

is the sum of squares of $n - p$ independent $N(0, 1)$ random variables.

Result: $(n - p)S^2/\sigma_0^2$ is $\chi^2(n - p)$.

Variance Estimator Properties

Verification of 2: Recall that $B_k - \beta_{0k} \sim N(0, \sigma_0^2 \delta_k)$ so $\frac{B_k - \beta_{0k}}{\sigma_0 \sqrt{\delta_k}} \sim N(0,1)$. Now

$$T_k = \frac{B_k - \beta_{0k}}{S\sqrt{\delta_k}} = \frac{(B_k - \beta_{0k})/\sigma_0\sqrt{\delta_k}}{S/\sigma_0} = \frac{Z}{\sqrt{\chi^2(n-p)/(n-p)}}$$

Result: T_k has a t-distribution with $n-p$ dof.

Confidence Interval: The $(1-\alpha) \times 100\%$ confidence interval for β_{0k} is

$$(B_k - t_{n-p,1-\alpha/2} S\sqrt{\delta_k}, B_k + t_{n-p,1-\alpha/2} S\sqrt{\delta_k})$$

since

$$P\left[-t_{n-p,1-\alpha/2} \leq T_k \leq t_{n-p,1-\alpha/2}\right] = 1-\alpha$$

$$\Rightarrow P\left[-t_{n-p,1-\alpha/2} \leq \frac{B_k - \beta_{0k}}{S\sqrt{\delta_k}} \leq t_{n-p,1-\alpha/2}\right] = 1-\alpha$$

$$\Rightarrow P\left[B_k - t_{n-p,1-\alpha/2} S\sqrt{\delta_k} \leq \beta_{0k} \leq B_k + t_{n-p,1-\alpha/2} S\sqrt{\delta_k}\right] = 1-\alpha$$

Example

Previous Example:

For $\alpha = .05$, $t_{n-p,1-\alpha/2} = 2.2$. The 95% confidence intervals are thus

$$\hat{b}_0 \in [111.6, 181.5]$$

$$\hat{b}_1 \in [-33.5, -14.1]$$

$$\hat{b}_2 \in [.63, .85]$$

Note: This is consistent with the results on Slide 23 with $2\sigma = 94.45\%$.

Summary of Linear Theory

Statistical Model:

$$Y = X\beta_0 + \varepsilon \quad , \; \beta_0 \in \mathbb{R}^p$$

$$y = X\beta_0 + \varepsilon \quad \text{(Realization)}$$

Assumptions: $E(\varepsilon_j) = 0$, ε_j iid with $\mathrm{var}(\varepsilon_j) = \sigma_0^2$

Least Squares Estimator and Estimate:

Note that $E(B) = \beta_0$

$$B = (X^T X)^{-1} X^T Y \quad , \quad \hat{b} = (X^T X)^{-1} X^T y$$

Variance Estimator and Estimate:

$R = Y - XB$, $\hat{r} = y - XB$

$$S^2 = \frac{R^T R}{n-p} \quad , \quad \hat{s}^2 = \frac{\hat{r}^T \hat{r}}{n-p}$$

Covariance Estimator and Estimate:

$$\mathrm{cov}(B) = \sigma_0^2 (X^T X)^{-1} \quad , \quad \widehat{\mathrm{cov}(B)} = \hat{s}^2 (X^T X)^{-1}$$

Summary of Linear Theory

Statistical Properties: (if $\varepsilon_k \sim N(0, \sigma_0^2)$ or in the limit $n \to \infty$)

- $B \sim N_p\left(\beta_0, \sigma_0^2 (X^T X)^{-1}\right)$
- $\dfrac{n-p}{\sigma_0^2} S^2 \sim \chi^2(n-p)$
- The $(1-\alpha) \times 100\%$ confidence interval for $\hat{b}_k$ is

$$(\hat{b}_k - t_{n-p,1-\alpha/2} SE_k, \hat{b}_k + t_{n-p,1-\alpha/2} SE_k)$$

where $SE_k = \sqrt{\widehat{\mathrm{cov}(B)}_{kk}}\ ,\ k = 1, \cdots, p$

Hypothesis Testing

Statistical Testing:

- An objective of statistics is to make inferences about unknown population parameters and models based on information in sample data.
- Inferences may be estimates of parameters or tests of hypotheses regarding their values.

Hypothesis Testing:

- Largely originated with Ronald Fisher, Jerzy Neyman, Karl Pearson and Egon Pearson
- Fisher: Agricultural statistician: emphasized rigorous experiments and designs
- Neyman: Emphasized mathematical rigor
- Early Paper: R. Fisher, ``Mathematics of a Lady Tasting Tea,'' 1956

  -- Question: Could lady determine means of tea preparation based on taste?

  -- Null Hypothesis: Lady had no such ability

  -- Fisher asserted that no alternative hypothesis was required

Hypothesis Testing

Elements of Test:

- Let X be a random variable with density $f(x;\theta)$
- Assume θ can take values in two possible sets Θ_0 or Θ_1
- Based on experiment, we want to decide if $\theta \in \Theta_0$ or $\theta \in \Theta_1$
- Null Hypothesis H_0: statement that $\theta \in \Theta_0$
- Alternative Hypothesis H_1: statement that $\theta \in \Theta_1$

Strategy:

- Take random sample of size n of X
- Decide in favor of H_0 if $x = (x_1, \cdots, x_n) \in S_0 \subset \mathbb{R}^n(X)$
- Decide in favor of H_1 if $x \in S_1 = \overline{S}_0$
- The partition (S_0, S_1) of the sample space $\mathbb{R}^n(X)$ is termed a test of the hypothesis H_0 against H_1

Hypothesis Testing

Elements of Test:

1. Null hypothesis H_0
2. Alternative hypothesis H_1
3. Test statistics
4. Rejection or critical region S_1

Definitions:

- Test Statistic: Function of sample measurement upon which decision is made.
- Rejection Region: Value of test statistic for which null hypothesis is rejected.

Definitions:

- A Type I error is made if H_0 is rejected when H_0 is true. The probability of a Type I error is denoted α.
- A Type II error is made if H_0 is accepted when H_1 is true. The probability of a Type II error is denoted β.

Hypothesis Testing

Example: Adam is running for Student body president and thinks he will gain more than 50% of the votes and hence win. His committee is very pragmatic and wants to test the hypothesis that he will receive less than 50% of the vote. Here we take

$H_0 : p = 0.5$

$H_1 : p < 0.5$

Sample Size: $n = 15$ in poll

Test Statistic X: Number of sampled voters favoring Adam

Case i: Calculate α if $S_1 = \{x \leq 2\}$

$$\begin{aligned} \alpha &= P(\text{Type I error}) \\ &= P(\text{rejecting } H_0 \text{ when } H_0 \text{ is true}) \\ &= P(\text{value of test statistic is in } S_1 \text{ when } H_0 \text{ is true}) \\ &= P(X \leq 2 \text{ when } p = 0.5) \end{aligned}$$

Note: X is a binomial random variable with $n = 15$ and $p = 0.5$. Thus

$$\alpha = \sum_{x=0}^{2} \binom{15}{x} (0.5)^{15} = 0.004$$

Hypothesis Testing

Example: Is this test equally protect us from erroneously concluding that Adam is the winner when, in fact, he will lose? Suppose that he will really win 30% of the vote so that p = 0.3. What is the probability of a Type II error?

Here

$$\begin{aligned}\beta &= P(\text{Type II error}) \\ &= P(\text{accepting } H_0 \text{ when } H_1 \text{ is true}) \\ &= P(\text{value of test statistic is not in } S_1 \text{ when } H_1 \text{ is true})\end{aligned}$$

Consider value of β when $p = 0.3$ (a value in H_1):

$$\begin{aligned}\beta &= P(X > 2 \text{ when } p = 0.3) \\ &= \sum_{x=3}^{15} \binom{15}{x} (0.3)^x (0.7)^{15-x} \\ &= 0.873\end{aligned}$$

Note: The test using this rejection region protects Adam from Type I errors but not Type II errors.

Hypothesis Testing

One Solution: Use a larger critical or rejection region.

e.g., Consider $S_1 = \{x \leq 5\}$. In this case,

$$\alpha = 0.151 \text{ when } p = 0.5$$

$$\beta = 0.278 \text{ when } p = 0.3$$

Conclusion: This provides a better balance between Type I and Type II errors.

Question: How can we reduce both errors?