Outline

1. (Cont.) Estimation of $E(Y_h)$

2. Estimation vs. Prediction

3. SLR and power analysis

4. Correlation Coefficient $\hat{\rho}$

Estimation of $E(Y_h)$

- $X_h =$ the level of X for which we want to estimate the mean response.
- X_h could be observed or not, but should be within the range of $\{X_i\}$.
- $\mu_h = E(Y_h) = \beta_0 + \beta_1 X_h =$ the mean response at X_h.
- The estimate of μ_h is

$$\hat{\mu}_h = \hat{\beta}_0 + \hat{\beta}_1 X_h.$$

- $\hat{\mu}_h \sim N(\mu_h, \sqrt{\mathrm{Var}(\hat{\mu}_h)})$. Why?

Estimation of $E(Y_h)$

- The variance of $\hat{\mu}_h$ is

$$\begin{aligned}\text{Var}(\hat{\mu}_h) &= \text{Var}(\hat{\beta}_0 + \hat{\beta}_1 X_h) \\ &= \text{Var}(\hat{\beta}_0) + X_h^2 \text{Var}(\hat{\beta}_1) + 2X_h \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) \\ &= \end{aligned}$$

- The estimated variance of $\hat{\mu}_h$ is

$$\widehat{\text{Var}(\hat{\mu}_h)} = \rule{3cm}{0.4pt}$$

- A useful test statistic is

$$\frac{\hat{\mu}_h - \mu_h}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}} \sim T_{n-2}.$$

- A $(1-\alpha)$ CI for μ_h is

Example: Wetland Species Richness

- The estimated mean number of species at $x_h = 0.10$ is

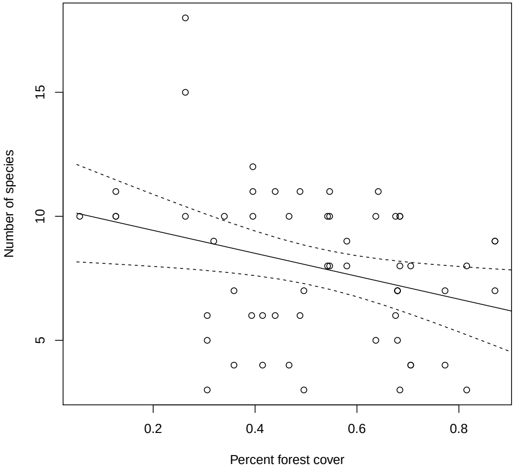
- The estimated variance of $\hat{\mu}_h$ is

$$\widehat{\text{Var}(\hat{\mu}_h)} \quad =$$

- The 95% CI for the mean number of species at $X_h = 0.10$ is

- Interpretation:

Example: Wetland Species Richness



Outline

1. (Cont.) Estimation of $E(Y_h)$
2. Estimation vs. Prediction
3. SLR and power analysis
4. Correlation Coefficient $\hat{\rho}$

Example: Wetland Species Richness

- The fitted regression line is $\hat{y} = 10.357 - 4.622x$.
- The estimated error variance is $\hat{\sigma}^2 = \frac{479.03}{56} = 8.554$.
- Questions of interest:

  1. What is the population mean number of species for a 10% forest cover around the wetland?

  2. What is the number of species for a 10% forest cover around a wetland yet to be sampled?

- In both cases, the estimated/predicted value is:

$$\hat{y} = 10.357 - 4.622 \times 0.10 = 9.895.$$

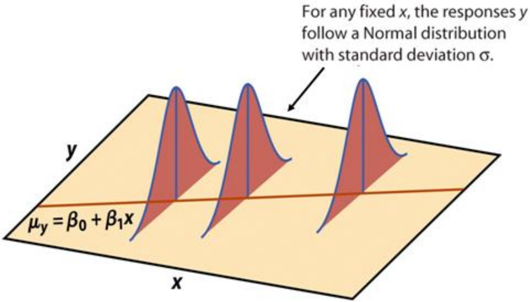
- Q: Which quantity has larger uncertainty?

Estimation vs. Prediction

Simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \varepsilon_i \sim \text{iid } N(0, \sigma^2), \; i = 1, \ldots, n.$$

- mean response at $X = 0.1$: $\beta_0 + \beta_1 \times 0.1$
- “new" response at $X = 0.1$: $\beta_0 + \beta_1 \times 0.1 + \varepsilon$
- sub-population vs. single observation



Estimation vs. Prediction

Consider a simple model (with covariate **0**)

$$Y_i = \mu + \varepsilon_i, \quad \varepsilon_i \sim \text{iid } N(0, \sigma^2).$$

1. Then, estimate μ by

$$\hat{\mu} = \bar{Y}$$

   - What is $Var(\hat{\mu})$?

2. Also, predict a new observation Y by

$$\hat{Y}_{(\text{new})} = \bar{Y}$$

   - What is the variance of the prediction error?

Outline

1. (Cont.) Estimation of $E(Y_h)$
2. Estimation vs. Prediction
3. SLR and power analysis
4. Correlation Coefficient $\hat{\rho}$

Another view of T-test

- Recall the simple linear regression (SLR) model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \varepsilon_i \sim i.i.d.\ N(0, \sigma^2),$$

for all $i = 1, \ldots, n$.

- Equivalently

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{MVN}(\mathbf{0}, \sigma^2 \boldsymbol{I}).$$

where $\boldsymbol{X}_{n\times 2} = \begin{bmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}$ denote the $n \times 2$ design matrix.

- One-sample test is a special case of SLR.
- Two-sample test is also a special case of SLR.

Another view of T-test

- Let
$$Y_i = \beta_0 + \varepsilon_i, \quad \varepsilon_i \sim i.i.d.\ N(0, \sigma^2),$$
for all $i = 1, \ldots, n$.
- Equivalently
$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{MVN}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}).$$
where $\boldsymbol{X}_{n\times 1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$ denote the $n \times 1$ design matrix, and $\boldsymbol{\beta} = \beta_0$.
- The one-sample test is equivalent to
$$H_0 : \beta_0 = \mu \text{ vs. } H_A : \beta_0 \neq \mu$$

Another view of two-sample test

- Let
  $$Y_i = \beta_0 \mathbb{1}_{i \text{ is in group } 1} + \beta_1 \mathbb{1}_{i \text{ is in group } 2} + \varepsilon_i, \quad \varepsilon_i \sim i.i.d.\ N(0, \sigma^2),$$
  for all $i = 1, \dots, n$.
- Equivalently
  $$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{MVN}(\mathbf{0}, \sigma^2 \boldsymbol{I}).$$
  where $\boldsymbol{X}_{n\times 2} = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix}$ denote the $n \times 2$ design matrix,
  and $\boldsymbol{\beta} = (\beta_0, \beta_1)'$.
- The unpaired two sample test is equivalent to
  $$H_0 : \beta_0 - \beta_1 = 0 \text{ vs. } H_A : \beta_0 - \beta_1 \neq 0$$

Errors in Hypothesis Test

Reality	Our Decision H_0	H_A
H_0	$\surd$	Type I Error
H_A	Type II Error	$\surd$

Example

Consider an SLR $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, where ε_i i.i.d. $N(0, \sigma^2)$. Suppose $n = 20$, $\sigma^2 = 0.01$, and $\bar{X} = 0.6$, $\sum_{i=1}^{n}(X_i - \bar{X})^2 = 10$. Graph the power for the hypothesis

$$H_0 : \beta_0 = 3.35 \quad \text{vs.} \quad H_A : \beta_0 \neq 3.35$$

with type 1 error 0.05.

Example: Power analysis in SLR

- Since σ is known,

$$\hat{\beta}_0 \sim N\left(\beta_0, \mathrm{Var}(\hat{\beta}_0)\right),$$

where $\mathrm{Var}(\hat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right) \approx 0.03^2$.

- Under the null hypothesis $H_0 : \beta_0 = 3.35$,

____

- Rejection region is $|\hat{\beta}_0 - 3.35| > 1.96 * 0.03 \approx 0.06$.
- Under a particular alternative $H_A : \beta_0 = \mu$,

____

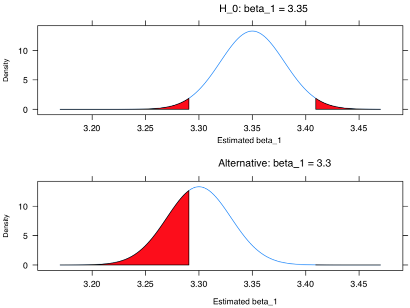
- Power at μ:

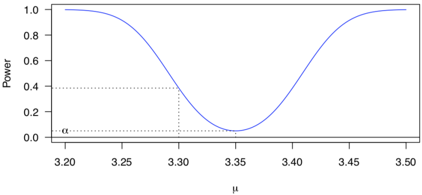
$$\mathrm{Power}(\mu) = \mathbb{P}\left(\underbrace{|\hat{\beta}_0 - 3.35| > 0.06}_{\text{rejection region calculated from } H_0} \;\middle|\; \mu\right)$$

Example: Power analysis in SLR

μ	3.25	3.30	3.35	3.40	3.45
Power (μ)	.0.91	0.37	0.05	0.37	0.91

- The rejection region (top graph) and the power at $\mu = 3.3$ (bottom graph)
- Power curve (right graph): a graph of the power for many alternative hypotheses



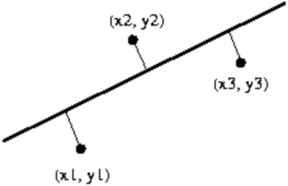


Outline

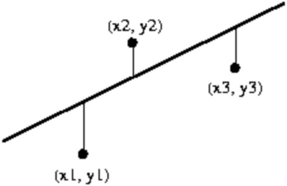
1. (Cont.) Estimation of $E(Y_h)$
2. Estimation vs. Prediction
3. SLR and power analysis
4. Correlation Coefficient $\hat{\rho}$

Brainstorm

Why do we use vertical distance to define the fitted line?



Perpendicular Distances



Vertical Distances

Other choices?

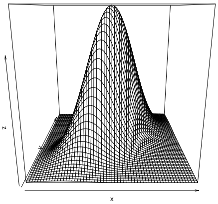
- The sum of the squares of perpendicular distance
- The sum of absolute value of the distance

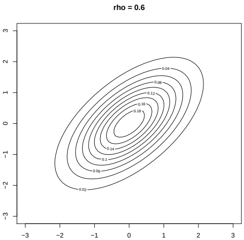
Correlation Model

- In simple linear regression, we model and predict Y given $X = x$.
- If interested in how two variables are related to each other, X and Y are to be treated symmetrically.
- Let X and Y both be random and have a bivariate distribution.
- A useful distribution is a **bivariate normal distribution** with a probability density that is parameterized by
  - μ_Y and σ_Y: the mean and the SD of the marginal distribution of Y
  - μ_X and σ_X: the mean and the SD of the marginal distribution of X
  - ρ_{YX} (or ρ): the **coefficient of correlation** between Y and X

Correlation Model

The probability density surface can be plotted using a 3D or contour plot.





Properties for bivariate normal (homework):

- Marginal distribution: $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$.
- Conditional distribution: $Y|X = x \sim N(\alpha + \beta x, \sigma_{Y|x}^2)$ where

$$\alpha = \mu_Y - \mu_X \rho \frac{\sigma_Y}{\sigma_X}, \quad \beta = \rho \frac{\sigma_Y}{\sigma_X}, \quad \sigma_{Y|x}^2 = \sigma_Y^2(1 - \rho^2).$$

Population Correlation Coefficient

- The **population correlation coefficient** (also called Pearson correlation coefficient) between X and Y is

$$\rho = Cor(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}.$$

- ρ is a measure of linear relationship between X and Y, $-1 \leq \rho \leq 1$.
- $\rho = 1$ indicates ________________ correlation.
- $0 < \rho < 1$ indicates ________________ correlation.
- $\rho = 0$ indicates ________________ relationship.
- $-1 < \rho < 0$ indicates ________________ correlation.
- $\rho = -1$ indicates ________________ correlation.

Example 1

Correlation coefficients

Example 2

Correlation coefficients

Example 3

Correlation coefficients

Pearson's Sample Correlation Coefficient

- Based on the data $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, the **sample correlation coefficient**

$$\hat{\rho} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

  estimates ρ.
- Note the symmetry between X and Y in $\hat{\rho}$.
- Sample covariance

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

- $\hat{\rho} = \frac{s_{xy}}{s_x s_y}$ estimates the Pearson correlation coefficient.

Independence

Let (X, Y) be a bivariate random variable in $\mathbb{R}^2$.

- Independence:

$$\mathbb{P}(X \le x, Y \le y) = \mathbb{P}(X = x)\mathbb{P}(Y \le y), \quad \text{for all } x, y \in \mathbb{R}.$$

- Uncorrelated:

$$\mathrm{Cov}(X, Y) = 0.$$

- Independence $\longrightarrow$ uncorrelated, but not vice versa.
- Pearson correlation coefficient is a measure of the strength of linear dependence between two random variables.
- If $Y = aX + b$, then $\rho(X, Y) = 1$ when $a > 0$, and $\rho(X, Y) = -1$ when $a < 0$.

Linear independence

Correlation coefficients

Statistical Inference on ρ

- Assume X and Y are from a bivariate normal distribution.
- Define Fisher's transformation

$$\lambda(\rho) = \frac{1}{2} ln\left(\frac{1+\rho}{1-\rho}\right) = \text{arctanh}(\rho),$$

- Fisher R.A. (1915) shows that

$$\lambda(\hat{\rho}) = \frac{1}{2}\ln\left(\frac{1+\hat{\rho}}{1-\hat{\rho}}\right) \approx N\left(\lambda(\rho), \frac{1}{n-3}\right).$$

- An approximate $(1-\alpha)$ CI for $\lambda(\rho)$ is

$$\lambda(\hat{\rho}) \pm z_{\alpha/2}\sqrt{\frac{1}{n-3}} = \left[\hat{\lambda}_1, \hat{\lambda}_2\right].$$

- An approximate $(1-\alpha)$ CI for ρ is

$$\left(\frac{e^{2\hat{\lambda}_1}-1}{e^{2\hat{\lambda}_1}+1} \equiv\right) tanh(\hat{\lambda}_1) \leq \rho \leq tanh(\hat{\lambda}_2) \left(\equiv \frac{e^{2\hat{\lambda}_2}-1}{e^{2\hat{\lambda}_2}+1}\right).$$

Example: Wetland Species Richness

- From the summary statistics, we have

$$\hat{\rho} = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_$$

- Find the Fisher's transformation

$$\lambda(\hat{\rho}) = \frac{1}{2} \log \left\{ \frac{1 + (-0.307)}{1 - (-0.307)} \right\} = \_\_\_\_\_\_$$

- An approximate 95% CI for $\lambda(\rho)$ is

$$\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_$$

- An approximate 95% CI for ρ is

$$\frac{e^{2(-0.582)} - 1}{e^{2(-0.582)} + 1} \leq \rho \leq \frac{e^{2(-0.0529)} - 1}{e^{2(-0.0529)} + 1}$$

which is ________________.

Remarks on $\hat{\rho}$

Correlation $\neq$ Causation