FALL 2013
Stat 8003: Statistical Methods I
Lecture 9

Jichun Xie

1 Hypothesis Testing

1.1 Likelihood Ratio Test

Lemma 1 (Neyman-Pearson Lemma)**.** *We observe* $Y_1, \ldots, Y_n$ *i.i.d.* $f_Y(y)$.

$$\mathrm{H}_0 : \tau = \tau_0$$
$$\mathrm{H}_1 : \tau = \tau_1$$

The form of the most powerful test of H_0 *versus* H_1 *is given by the rule: "Reject* H_0 *for Large Values of the Likelihood Ratio"*

$$LR = \frac{Lik(\tau_1)}{Lik(\tau_0)}$$

1.2 Generalized Likelihood Ratio Tests (GLRT)

Neyman & Pearson gave us a framework for test statistic construction when our null and alternative are simple and we have parametric distributions. In practice, hypotheses are generally composite.

1. Observe $\mathbf{Y}$ form $f_Y(y, \theta)$. Let Θ_0 and Θ_1 denote subsets of the parameter space. The GLRT rejects H_0 when

$$LR = \frac{\max_{\theta \in \Theta_1} L(\theta)}{\max_{\theta \in \Theta_0} L(\theta)} > k.$$

or a form that is a little easier to work with:

$$\Lambda = \frac{\max_{\theta \in \Theta_0 \cup \Theta_1} L(\theta)}{\max_{\theta \in \Theta_0} L(\theta)} > k^*.$$

The next problem is that we need to find the distribution of Λ in order to set up probability statements and get the error rates. Sometimes we can find an exact distribution of Λ in order to set up probability statements and get the error rates. In other cases we work with an asymptotic approximation.

2. Results which you will prove in STAT 8001 or 8002.

   1. Simple null, *e.g.*:

$$\begin{aligned} &\mathrm{H}_0 : \tau = \tau_0 \\ &\mathrm{H}_1 : \tau \neq \tau_0 \ \text{ or } \ \mathrm{H}_1 : \ \tau = \tau_1, \end{aligned}$$

      where τ is a one-dimensional parameter. Then,

$$2 \log \Lambda \dot{\sim} \chi^2(1) \text{ under } \mathrm{H}_0.$$

   2. Nested null and alternative

$$\begin{aligned} &\mathrm{H}_0 : \boldsymbol{\theta} \in \Theta_{p-r} \text{ (reduced model)} \\ &\mathrm{H}_1 : \boldsymbol{\theta} \in \Theta_p \text{ (full model)}, \end{aligned}$$

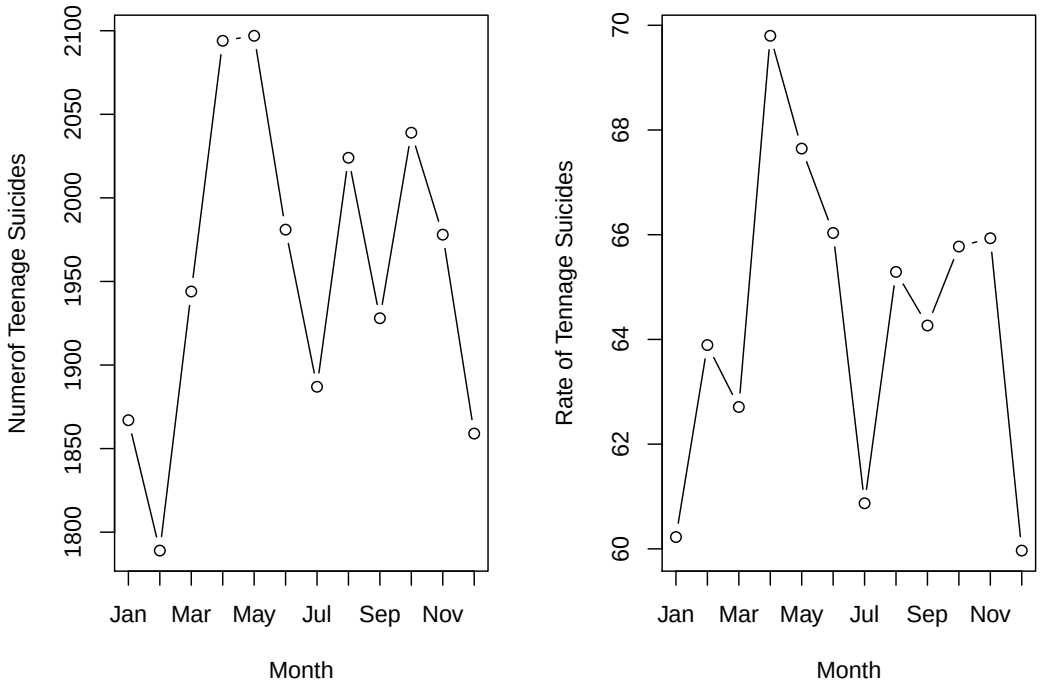
      where Θ_{p-r} is a subset of Θ_p. Then

$$2 \log \Lambda \dot{\sim} \chi^2(r).$$

Examples: Seasonal Changes in Teen Suicides. Teenage Suicide Data (available at `http://www.familyfirstaid.org/suicide.html`). In 2001, teen suicide was the 3rd leading cause of death among young adults and adolescents 15 to 24 years of age, following unintentional injuries and homicide. The rate was 9.9/100,000 or .01%. The adolescent suicide rate among youth ages 10-14 was 1.3/100,000 or 272 deaths among 20,910,440 children in this age group. The gender ratio for this age group was 3:1 (males: females). The teen suicide rate among youth aged 15-19 was 7.9/100,000 or 1,611 deaths among 20,271,312 teenagers in this age group. The gender ratio for teenage group was 5:1 (males: females). Among young people 20 to 24 years of age, the youth suicide rate was 12/100,000 or 2,360 deaths among 19,711,423 people in this age group. The gender ratio for this age group was 7:1 (males: females).

For this example we need a little background information on the multinomial distribution. Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ be a vector denoting the number of times that an independent observation falls into the ith category, $i = 1, \ldots, n$ in a series of n trials, where the probability of falling intro the ith category is θ_i; $\sum_{i=1}^n \theta_i = 1$. Then,

$$\mathbb{P}\left(Y_1 = y_1, \ldots, Y_n = y_n\right) = \frac{n!}{y_1! \cdots y_n!} \prod_{i=1}^n \theta_i^{y_i}.$$



Note that in this model, the constraint on the parameters is $\sum_{i=1}^{n} \theta_i = 1$; and the constraint on the counts is $\sum_{i=1}^{n} y_i = n$. Note that this distribution is an extension of the binomial, and the mle for each $\hat{\theta}_i = \frac{Y_i}{n}$.

1. The null hypothesis is that the suicide rate is constant across months. The alternative is that there is seasonal variation in suicide rate.

2. Model: Y_i is the number of suicides in the ith month ($i = 1, \ldots, 12$).

$$\begin{aligned} \mathrm{H}_0 &: \boldsymbol{\theta} = (\theta_{1,0}, \ldots, \theta_{12,0}) \text{ (reduced model)} \\ \mathrm{H}_1 &: \boldsymbol{\theta} = (\theta_1^*, \ldots, \theta_{12}^*) \text{ (full model)} \end{aligned}$$

where under the null hypothesis the $\theta_{i,0}$ are determined by # days in the ith month/365.

The model is

$$\begin{aligned}\Lambda &= \frac{\max L(\hat{\boldsymbol{\theta}}_{mle})}{\max L(\hat{\boldsymbol{\theta}}_{0,mle})} \\ &= \frac{\frac{n!}{y_1 \cdots y_n} \prod_{i=1}^{12} \hat{\theta}_i^{y_i}}{\frac{n!}{y_1 \cdots y_n} \prod_{i=1}^{12} \theta_{i,0}^{y_i}} \\ &= \prod_{i=1}^{12} \left(\frac{\hat{\theta}_i}{\theta_{i,0}} \right)^{y_i} \\ &= \prod_{i=1}^{12} \left(\frac{y_i}{n\theta_{i,0}} \right)^{y_i}.\end{aligned}$$

In this case, large values of Λ do not translate into a test statistic whose distribution is known. However,

$$2 \log \Lambda \dot{\sim} \chi^2(11).$$

There are $12 - 1 = 11$ free parameters in the full model since $\sum_{i=1}^{n} \theta_{i,0} = 1$; and there are 0 free parameters in the restricted model (rate determined by the number of days in the month).

$$2 \log \Lambda = 2 \sum_{i=1}^{12} y_i \left[\log(y_i) - \log(n\theta_{i,0}) \right]$$

This statistic has an "observed-expected" look to it. Why?

See R handout for the results of the example. $2 \log \Lambda$ is 47.66 based on the data; and the $\alpha = 0.05$ level cut-off is 19.67. So we should reject the H_0 under the level of 0.05. That is to say, there is a significant difference among the teenage suicide rate per month.

Lastly Collaborators often like to report p-values in their work. We can think of a p-value as the probability of the observed result, or a more extreme result, given the null hypothesis is true. Note: A p-value is a random variable (it's a function of the data) whereas the level of the test is fixed by the design. For the suicide data the p-value is $p = 1.7 \times 10^{-6}$ or $p < .0001$.

1.3 Other likelihood-based hypothesis tests

For the hypotheses:

$$\begin{aligned}\mathrm{H}_0 &: \theta = \theta_0 \\ \mathrm{H}_1 &: \theta \neq \theta_0.\end{aligned}$$

1.3.1 Wald Test

Recall that we discussed in the previous lecture the variance of the MLE estimators. Suppose $Y_1, \ldots, Y_n \sim f(y;\theta)$, *i.i.d.*. We define the log likelihood function by

$$l(\theta) = \sum_{i=1}^{n} \log f(y_i;\theta).$$

The MLE $\hat{\theta}_{mle}$ maximized the (log) likelihood:

$$\hat{\theta}_{mle} = \arg\max_{\theta} l(\theta).$$

Suppose the Fisher information

$$I(\theta) = \mathtt{E}\left\{ (l'(\theta))^2 \middle| \theta \right\} = -\mathtt{E}\left\{ l''(\theta) | \theta \right\}$$

exists. Then under some regularity conditions,

$$\mathtt{Var}(\hat{\theta}_{mle}) = (I(\theta))^{-1}.$$

Now define the observed Fisher information

$$i(\hat{\theta}) = I(\hat{\theta}).$$

Then we can estimate the variance of MLE by

$$\widehat{\mathtt{Var}}(\hat{\theta}_{mle}) = i^{-1}(\hat{\theta}_{mle}).$$

To test $\mathrm{H}_0 : \theta = \theta_0$, we construct

$$W = \frac{|\hat{\theta}_{mle} - \theta_0|}{\sqrt{i(\hat{\theta}_{mle})^{-1}}},$$

Under the null hypothesis, W has a standard normal distribution. Can do one-sided tests as well.

Example. For Poisson distribution,

$$W = \frac{\sqrt{n}|\bar{Y} - \tau_0|}{\sqrt{\bar{Y}}}.$$

More generally if we have an estimate of $\hat{\theta}$ of θ that is asymptotically normal, you sometimes see a 'Wald-type' test

$$W = \frac{|\hat{\theta} - \theta_0|}{\sqrt{\mathtt{Var}(\hat{\theta})}},$$

where $\mathtt{Var}(\hat{\theta})$ is the variance of $\hat{\theta}$. We generally use a consistent estimate of $\mathtt{Var}(\hat{\theta})$ and justify the asymptotic distribution of W using Slutsky's theorem.

1.3.2 Score Test

The score is

$$U(\theta) = \frac{\mathrm{d}l(\theta)}{\mathrm{d}\theta},$$

where under H_0,

$$\mathtt{E}[U(\theta_0)] = 0 \text{ and } \mathtt{Var}[U(\theta_0)] = I(\theta_0),$$

where $I(\theta_0)$ is the information evaluated at θ_0.

Now $U(\theta)$ is the sum of *i.i.d.* random variables (since the log of the likelihood is the sum of the log likelihood for the individual observations). Thus by the central limit theorem

$$\frac{U(\theta_0)}{\sqrt{I(\theta_0)}} \xrightarrow{D} N(0,1).$$

The beauty of the score is that it does not require finding the mle, and thus it sometimes taks a simple form.

Example. For the Poisson recall that

$$f_Y(y) = \theta^y \exp(-\theta)/y!$$
$$U(\theta) = -n + \frac{\sum_{i=1}^n Y_i}{\theta}$$

$$\frac{\mathrm{d}^2 \log L(\theta, \mathbf{y})}{\mathrm{d}\theta^2} = \frac{-\sum_{i=1}^n Y_i}{\theta^2}$$
$$I(\theta_0) = \frac{n}{\theta}\Big|_{\theta=\theta_0}$$

$$U = \frac{-n + \frac{\sum_{i=1}^n Y_i}{\theta_0}}{\sqrt{\frac{n}{\theta_0}}}$$
$$= \sqrt{\frac{n}{\theta_0}}(\bar{Y} - \theta_0)$$

Questions:

- How to compare different methods?
- What will happen if the null hypothesis is true and the sample size is large?
- What will happen if the alternative hypothesis is true and the sample size is large?

1.4 Sample Size and Power Analysis for Normal Distribution

Example: We are interested in determining whether the mean volume of fluid delivered to patients during a particular type of neurosurgery is at least 50ml greater than 1500ml. If so then the surgeons need to think about modifying the procedures to reduce the large fluid volumes. Our null hypothesis is that the mean volume, μ, is 1500 ml. We cannot fule out $\mu < 1500$. We believe σ^2 is roughly 10,000 ml^2.

Questions of Interest:

1. What power will we have to detect a difference in the mean of 50 ml if we use a sample size of 10 patients?

2. How many patients will we need to detect a difference of 50 ml with at least 80% power?

3. Setup:

$$\begin{aligned} \mathrm{H}_0 &: \mu = \mu_0 \\ \mathrm{H}_1 &: \mu \neq \mu_0, \end{aligned}$$

where μ is the mean volume of fluid. The type I error rate is $\alpha = 0.05$.

4. Let $Y_1, \ldots, Y_n$ represent the sample of volumes. The test-statistic will be the T statistic (not T-distribution! Here, $Y_1, \ldots, Y_n$ can be non-Normal.) For the purpose of sample size/power calculations we use:

$$Z = \frac{\bar{Y} - \mu_0}{\sqrt{\hat{\sigma}^2/n}}.$$

Under H_0, asymptotically $Z \sim \mathrm{N}(0, 1)$.

5. Under the null

$$\mathbb{P}\left(|Z| > \left|\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right| \mid \mathrm{H}_0 \text{ is true}\right) = \alpha.$$

6. Under the alternative we want

$$\mathbb{P}\left(|Z| > \left|\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right| \mid \mathrm{H}_1 \text{ is true}\right) = 1 - \beta.$$

$$\mathbb{P}\left(\frac{\bar{Y} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} > \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right) + \mathbb{P}\left(\frac{\bar{Y} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} < \Phi^{-1}\left(\frac{\alpha}{2}\right)\right) = 1 - \beta.$$

7. Let μ_1 be the mean under the alternative. WLOG, suppose $\mu_1 > \mu_0$.

$$\mathbb{P}\left(\frac{\bar{Y} - \mu_1}{\sqrt{\frac{\sigma^2}{n}}} > \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) + \frac{\mu_0 - \mu_1}{\sqrt{\frac{\sigma^2}{n}}}\right) + \mathbb{P}\left(\frac{\bar{Y} - \mu_1}{\sqrt{\frac{\sigma^2}{n}}} < \Phi^{-1}\left(\frac{\alpha}{2}\right) + \frac{\mu_0 - \mu_1}{\sqrt{\frac{\sigma^2}{n}}}\right) = 1 - \beta.$$

8. But since $\mu_1 > \mu_0$, the second term is small, and we usually ignore it. Why?

9. What is the distribution fo the first term in (7) when H_1 is true?

$$1 - \Phi\left(\Phi^{-1}\left(1 - \frac{\alpha}{2}\right) + \frac{\mu_0 - \mu_1}{\sqrt{\frac{\sigma^2}{n}}}\right) \approx 1 - \beta.$$

10. The equation in (9) gives us the power. What about the sample size?

$$\begin{aligned}
\Phi^{-1}\left(1 - \frac{\alpha}{2}\right) + \frac{\mu_0 - \mu_1}{\sqrt{\frac{\sigma^2}{n}}} &= \Phi^{-1}(\beta) \\
\frac{\mu_0 - \mu_1}{\sqrt{\frac{\sigma^2}{n}}} &= \Phi^{-1}(\beta) - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \\
\frac{(\mu_0 - \mu_1)^2}{[\Phi^{-1}(\beta) - \Phi^{-1}(1 - \frac{\alpha}{2})]^2} &= \frac{\sigma^2}{n} \\
n &= \frac{\sigma^2[\Phi^{-1}(\beta) - \Phi^{-1}(1 - \frac{\alpha}{2})]^2}{(\mu_0 - \mu_1)^2}
\end{aligned}$$

11. Back to the example: $\sigma^2 = 100^2(\mu_0 - \mu_1) = -50$, $\Phi^{-1}(\beta) = -0.84$, $\Phi^{-1}(1 - \frac{\alpha}{2}) = 1.96$,

$$n = 31.36.$$

Conclude that we need at least 32 patients to have at least 80% power to detect an increase of 50 ml.

12. Suppose that we were wrong and the difference is actually a decrease in the volume by 50 ml. What would the power be?

13. What if we use 10 patients? What is the power?

$$1 - \beta = 1 - \Phi\left(1.96 + \frac{-50}{\sqrt{\frac{100^2}{10}}}\right) = 0.36.$$

14. Suppose we carry out the experiment with 10 patients and don't reject the null. What would we conclude from the experiment?

2 Linear Regression

2.1 Introduction

In lots of the real problems, the data can be formulated to a response variable Y and some covariates X.

Y	X
dependent variable	independent variable
response variable	explanatory variable
outcome variable	covariate
	predictor

Linear regression can handle the case where Y is continuous. In next semester, we are going to discuss logistic regression and Poisson regression. These two models can handle the case where Y is categorical or count.

Purpose of "Regression": quantify the magnitude of the association/relationship between Y and X.

Example. In a phase-II clinical trial, the researchers are interested in the effect of a new cholesterol lowering drug. The patients are randomized to different drug-dose group and the cholesterol levels before and after taking the drug are measured for each patient. The researchers want to know:

1. Is the new drug has any effect of lowering cholesterols? If there is some effect, is the effect strong or weak? (Estimation)

2. Given a new dose X, what would Y be? (Prediction)

Suppose

$$X_i = \text{dose of the drug the } i\text{-th patent took}$$
$$Y_i = \text{the difference of the cholesterol level of the } i\text{-th patient}$$

We can view (X_i, Y_i) are n *i.i.d.* realization of the random variable X and Y.

How to study the relationship between Y and X?

Useful plots:

1. Scatterplot: for continuous Y and X (See Figure 1)

   Plot Y_i *vs.* X_i for all i.

   Features:

- Overal shape: linear/nonlinear
- Spread
- Isolated points

2. Boxplot: for continuous Y but categorical X.

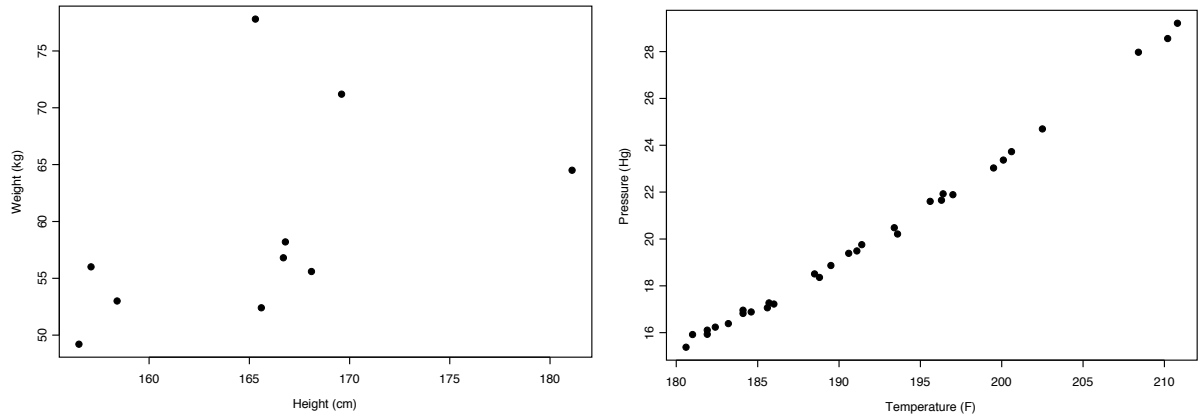


Figure 1: Scatterplot examples

Sample Conditional Mean and Population Conditional Mean:

- $\text{Avg}(Y \mid X)$, "$\mid$" means given.
- $\texttt{E}(Y \mid X)$, conditional expectation.

2.2 Relationship in Statistical Terms

Correlation $\rho(X, Y)$ or ρ_{XY}: measures the linear association between two r.v. Y and X.

$$\rho_{XY} = \frac{\texttt{Cov}(X, Y)}{\sqrt{\texttt{Var}(X) \cdot \texttt{Var}(Y)}},$$

where

$$\texttt{Cov}(X, Y) = \mathbb{E}\left\{(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\right\}, \quad \texttt{Var}(Y) = \mathbb{E}\left\{(Y - \mathbb{E}(Y))^2\right\} = \sigma_Y^2$$

Note that $-1 \leq \rho(X, Y) \leq 1$. Why?

Consider $Z_1 = \frac{Y}{\sigma_Y} + \frac{X}{\sigma_X}$ and $Z_2 = \frac{Y}{\sigma_Y} - \frac{X}{\sigma_X}$.

$$\begin{aligned}
\texttt{Var}\left(Z_1\right) &= \texttt{Var}\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right) \\
&= \texttt{Var}\left(\frac{X}{\sigma_X}\right) + \texttt{Var}\left(\frac{Y}{\sigma_Y}\right) + 2\texttt{Cov}\left(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right) \\
&= 1 + 1 + 2\rho(X, Y).
\end{aligned}$$

Since $\texttt{Var}(Z_1) \geq 0$, $1 + 1 + 2\rho(X, Y) \geq 0$, and then $\rho(X, Y) \geq -1$.

Similarly by calculating $\texttt{Var}(Z_2)$ and use the fact $\texttt{Var}(Z_2) \geq 0$, we can show $\rho(X, Y) \leq 1$.

Here are some other facts about $\rho_{X,Y}$:

- When $\rho_{XY} = 0$,
$$\texttt{Var}\left(\frac{Y}{\sigma_Y} + \frac{X}{\sigma_X}\right) = \texttt{Var}\left(\frac{Y}{\sigma_Y} - \frac{X}{\sigma_X}\right).$$
- When $\rho_{XY} > 0$,
$$\texttt{Var}\left(\frac{Y}{\sigma_Y} + \frac{X}{\sigma_X}\right) > \texttt{Var}\left(\frac{Y}{\sigma_Y} - \frac{X}{\sigma_X}\right).$$
- When $\rho_{XY} = 1$,
$$\texttt{Var}\left(\frac{Y}{\sigma_Y} - \frac{X}{\sigma_X}\right) = 0.$$

Some extra notes about $\rho_{X,Y}$:

- If X and Y are independent ($X \perp\!\!\!\perp Y$), then $\rho_{X,Y} = 0$.
- However $\rho_{X,Y} = 0$ doesn't necessarily lead to $X \perp\!\!\!\perp Y$.
- However, if X and Y follow normal distribution, then $\rho_{X,Y} = 0$ leads to $X \perp\!\!\!\perp Y$.
- ρ_{XY} is a measure of linear association, it is NOT a measure of causality.

Sample correlation: After obtaining the data (X_i, Y_i), $i = 1, \ldots, n$. We can define the sample

correlation (Pearson's correlation) as:

$$\begin{aligned} r_{XY} &= \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \cdot \frac{1}{n}\sum i = 1^n (X_i - \bar{X})^2}} \\ &= \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \cdot \sum_{i=1}^{n}(X_i - \bar{X})^2}} \\ &= \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}, \end{aligned}$$

where $S_{xx} = \sum_{i=1}^{n}(X_i - \bar{X})^2$, $S_{yy} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ and $S_{xy} = \sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})$.

2.3 Basic Regression Model

$$\begin{aligned} Y &= \text{Systematic component} + \text{Random component} \\ &= \text{Fit} + \text{Error or "Residuals"} \\ &= f(X;\beta) + \epsilon \end{aligned}$$

1. Fit $f(X;\beta)$: estimate the parameter β. The logic is $f(X;\beta)$ should be close to Y so that the error term ϵ is small.

2. Error ϵ: described by some probability distributions.

Consider fitting a function $f(X;\beta)$ such that $\mathbb{E}(Y \mid X) = f(X;\beta)$. If $f(X;\beta) = X\beta$, then the model is called a linear model.

Simple linear regression. Suppose the data are (X_i, Y_i), $i = 1, \ldots, n$. Consider the model

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \epsilon_i \\ &= \text{fit}(X;\beta) + \text{Error} \end{aligned}$$

Here the systematic component $f(X;\beta) = \beta_0 + \beta_1 X_i$ is linear. The random component is ϵ_i.

Figure 2 shows an example of simple linear regression.

Remarks:

1. The model is called "simple" because there is only one covariate X.

2. There is not too much difference between the fix design and the random design. The key quantity is $\mathbb{E}(Y \mid X) = \beta_0 + \beta_1 X$.

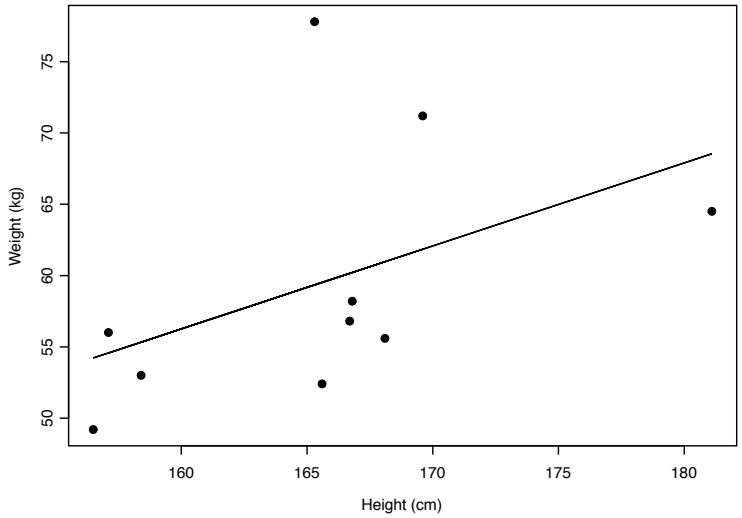


Figure 2: Simple linear regression: weight *vs.* height

Here are some assumptions about SLR:

1. Zero mean of the error: $\mathbb{E}(\epsilon_i) = 0$.
2. Constant variance of the error: $\texttt{Var}(\epsilon_i) = \sigma^2$.
3. ϵ_i's are mutually independent: $\epsilon_i \perp\!\!\!\perp \epsilon_j$, $i \neq j$.
4. ϵ_i's are normal r.v. (needed for testing).

Assumption 1 and 2 are equivalent to the following

1*. $\mathbb{E}(Y_i \mid X_i) = \beta_0 + \beta_1 X_i$.

2*. $\texttt{Var}(Y_i \mid X_i) = \sigma^2$.

If assumptions 1, 2, 3 and 4 holds, $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

How to interpretate the model?

Example. Suppose X = Years of experience, Y = annual salary of an employee (in \$1000). Consider two employees with x, $x + 1$ years of experience.

The expected salary for employee 1 is

$$\mathbb{E}(Y \mid X = x) = \beta_0 + \beta_1 x. \tag{1}$$

And the expected salary for employee 2 is

$$\mathbb{E}(Y \mid X = x + 1) = \beta_0 + \beta_1 (x + 1). \tag{2}$$

Then eq(2) - eq(1) = β_1. Therefore, β_1 can be interpreted as

- *difference* in *average* salary (Y) between employees with $x+1$ years of experience and x years of experience

- *difference (change)* in *average* Y with 1 unit *increase (change)* in X.

Interpretation about β_0: $\beta_0 = \mathbb{E}(Y \mid X = 0)$ is the expected salary for an employee with no working experience.

In other examples, β_0 can be less meaningful, *e.g.*. relationship between babies' weight (Y) and height (X).

Next we discuss the estimation of the unknown parameter β_0, β_1 and σ^2. Previously, we discussed two general types of estimators, including method of moments (MOM) and maximum likelihood estimators (MLE). For simple linear regression, we can use the above methods too. But due to the special properties of the model, we first start with another method with more intuitive explanations.