

Inference with Multiple Comparisons

Concerns

- Multiple comparisons refers to making several comparisons simultaneously.
- *Comparisonwise error rate* (CWER) is the Type I error rate α for each comparison (i.e. the probability of false rejection for each comparison).
- Note that if each comparison has $\alpha = 0.05$ and suppose multiple comparisons, then the probability of having at least one significant comparison given that all H_0 's are true is > 0.05 .
- A crude analogy (assuming independence): toss a coin once with $P(H) = 0.05$; toss it twice, then $P(\text{at least one } H) = 1 - 0.95 \times 0.95 = 1 - 0.9025 = 0.0975$; toss it many times, then $P(\text{at least one } H)$ becomes much larger than 0.05.
- *Experimentwise error rate* (EWER) is the probability of *at least* one false rejection among multiple comparisons, given that all H_0 's are true.

Inference with Multiple Comparisons

Bonferroni idea

- The problem is that suppose $\text{CWER} = 0.05$, then EWER can be much larger than 0.05 if many comparisons are made. In practice, control CWER , or EWER , or find a compromise.
- Consider two comparisons, each with $\text{CWER} = \alpha$.
- Let A denote the event that Type I error is made on the first comparison.
- Let B denote the event that Type I error is made on the second comparison.
- Then

$$\begin{aligned}\text{EWER} &= P(\text{at least one Type I error is made}) \\ &= P(A \text{ or } B) \\ &= P(A) + P(B) - P(A \text{ and } B) \\ &\leq P(A) + P(B) = 2\alpha\end{aligned}$$

- The inequality is known as the Bonferroni inequality.
- Usually $P(A \text{ and } B)$ is small and thus

$$\text{EWER} \approx 2\alpha.$$

Inference with Multiple Comparisons

Selection bias

- Consider $k > 2$ trt comparisons in the following way.
- Take the largest and the smallest trt means and compare them at α level.
- In testing whether the corresponding population means are equal, the actual Type I error rate is larger than α , because we *selected* the test that has the highest chance of leading to rejection.
- Now we will learn how to make multiple comparisons with these concerns and ideas in mind.

Inference with Multiple Comparisons

Barley root example

In a study of five varieties of barley, the weight of roots is recorded of $n = 7$ plants per variety is recorded. The group means are

$$\bar{y}_{1.} = 16.3, \bar{y}_{2.} = 19.3, \bar{y}_{3.} = 14.7, \bar{y}_{4.} = 20.3, \bar{y}_{5.} = 18.5.$$

The ANOVA table is

Source	df	SS	MS	F	p-value
Trt	4	145.94	36.48	5.09	< 0.01
Error	30	214.74	7.16	—	—
Total	34	360.68	—	—	—

- Case I: general contrasts.
- Case II: all pairwise comparisons.

Inference with Multiple Comparisons

Case I: general contrasts

- Among many approaches, we consider two approaches:
 - Bonferroni method
 - Protected t-test
- The Bonferroni method makes use of the Bonferroni idea to control EWER.
- For example, consider two comparisons in the barley example:

$$H_0 : \mu_1 = (\mu_2 + \mu_3 + \mu_4 + \mu_5)/4$$

$$H_0 : (\mu_2 + \mu_3)/2 = (\mu_4 + \mu_5)/2$$

and we want to control the EWER to be 0.05.

- By the Bonferroni idea, we want $2\alpha \approx 0.05$ and thus $\alpha = 0.025$ for *each* of the two comparisons. That is, for each H_0 , perform a t-test and reject H_0 if the p-value is < 0.025 .
- In general, suppose there are r comparisons to be made (chosen in advance) and we want an EWER to be 0.05, then we want $r\alpha \approx 0.05$ and thus $\alpha = 0.05/r$ for *each* of the r comparisons. That is, for each H_0 , perform a t-test and reject H_0 if the p-value is $< 0.05/r$.

Inference with Multiple Comparisons

Remarks

- The main idea in the Bonferroni method is

$$\text{EWER} = r \times \text{CWER}$$

- Thus EW p-value = $r \times$ CW p-value.
- For example, if the CW p-value is between 0.005 and 0.01, then the EW p-value is between $0.005r$ and $0.01r$. If $r = 5$, then the EW p-value is between 0.025 and 0.05.
- The Bonferroni method is quite conservative, especially when r is large. Some people rarely use it if $r \geq 4$ or 5.

Inference with Multiple Comparisons

Protected t-test

- The main focus is to control CWER, but with protection.
- The procedure is as follows. Perform an overall f-test for

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

and if the f-test is significant at α , then perform a *usual* t-test at α for each comparison.

- Past simulation studies show that the protected t-test performs well, even though it is somewhat liberal.

Inference with Multiple Comparisons

Case II: all pairwise comparisons

- Recall the barley root example and that the f-test is significant.
- Now we want to compare pairwise the group means.

Group	3	1	5	2	4
Mean	14.7	16.3	18.5	19.3	20.3
- Among many approaches, we consider three approaches:
 - Fisher's least significant difference (LSD).
 - Bonferroni test.
 - Q-method (QD).
- We also focus on balanced data with $n_1 = n_2 = \dots = n_k$.

Inference with Multiple Comparisons

Fisher's LSD

1. Use protected LSD (basically the same as protected t-test).
2. Find the distance $D_L = \bar{Y}_1. - \bar{Y}_2.$ so that this distance leads exactly to a p-value of α :

$$\frac{D_L}{S_p \sqrt{\frac{2}{n}}} = t_{\alpha/2, \text{dfErr}}$$

and thus the LSD is:

$$D_L = t_{\alpha/2, \text{dfErr}} \times S_p \sqrt{\frac{2}{n}}.$$

Inference with Multiple Comparisons

Remarks

- In the barley example, suppose $\alpha = 0.05$. Since $f = 5.19$ on $df = (4,30)$ and the p-value is less than 0.01, proceed to perform all pairwise comparisons.
- Since $n = 7$, $s_p^2 = 7.16$, $df_{Err} = 30$, $t_{0.025,30} = 2.042$, we have

$$D_L = 2.042 \times \sqrt{7.16 \times 2/7} = 2.92$$

and

Group:	3	1	5	2	4
Mean:	14.7	16.3	18.5	19.3	20.3

- That is, two group means that are within 2.92 of each other are connected with a line and are not significantly different.
- Interpretation is not transitive.
- Alternative displays are possible. See the bluebook.

Inference with Multiple Comparisons

Bonferroni tests

- Use the Bonferroni method for all pairwise comparisons.
- In the barley example, the total number of comparisons is 10.
- Bonferroni tests are overly conservative and are not useful.
- But Bonferroni tests may be useful for a subset of comparisons.

Inference with Multiple Comparisons

Q-method

- The Q-method (QD) does not involve a t-test.
- QD is also known as “studentized range”, or “Tukey’s W approach”, or HSD for honestly significant distance.
- The main idea is to control the EWER at α by using the argument of selection bias. If μ_i ’s are all equal, then the probability that the largest group mean is different from the smallest group mean is α .
- Let

$$D_Q = Q_{k,dfErr,\alpha} \times S_p \sqrt{\frac{1}{n}}$$

where $Q_{k,dfErr,\alpha}$ is called the Q-score and we look up the Q-score from Table A15 (Snedecor and Cochran’s book).

Inference with Multiple Comparisons

Remarks

- In the barley example, $k = 5$, $n = 7$, $s_p^2 = 7.16$, $\text{dfErr} = 30$, and at $\alpha = 0.05$, $Q_{5,30,0.05} = 4.11$ and

$$D_Q = 4.11 \times \sqrt{7.16/7} = 4.16$$

and thus

Group:	3	1	5	2	4
Mean:	14.7	16.3	18.5	19.3	20.3

- The QD tends to be conservative.

Inference with Multiple Comparisons

Final remarks

- If the sample sizes are not equal, both LSD and QD methods require adjustment.
- An intermediate approach between LSD and QD is Newman-Keuls (NK).
- Another sequential method commonly used is Duncan's multiple range test (DMRT)
- These methods represent a tradeoff between controlling Type I error and power. From conservative to liberal are Bonferroni, QD, NK, DMRT, and LSD.
- Choice of method depends on objectives and experiences.

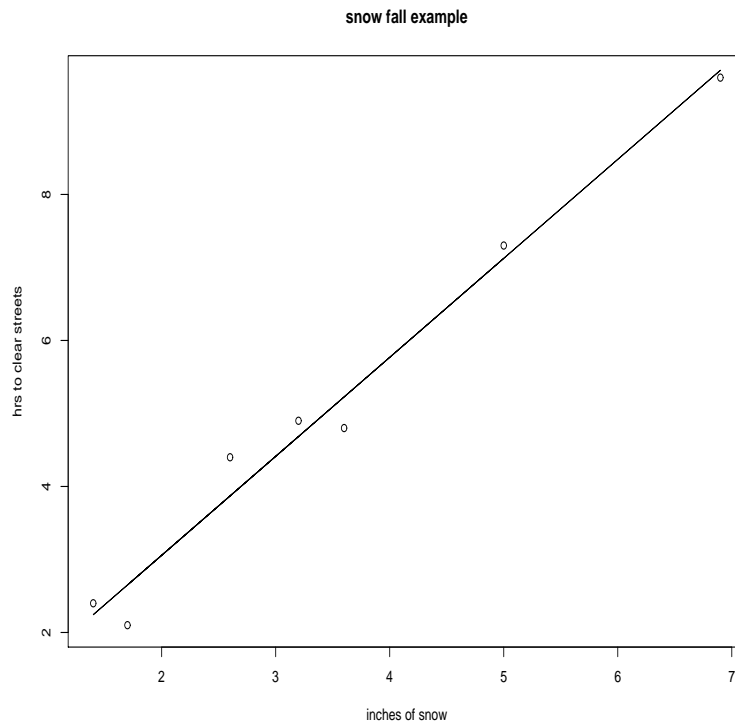
Simple Linear Regression

Snow fall example

From 7 small towns of Wisconsin, the following snow fall data were collected.

inches of snow fall x	3.2	1.4	2.6	6.9	3.6	1.7	5.0
hours to clear streets y	4.9	2.4	4.4	9.6	4.8	2.1	7.3

The question of interest is: What is the relationship between the amount of snow fall (x) and the time it takes to clear the streets (y)?



Simple Linear Regression

Objectives

In the snow fall example, the objectives are to describe the relationship between the amount of snow fall (x) and the time it takes to clear the streets (y), estimate or predict time to clear the streets for a given amount of snowfall.

Model

- The main idea behind simple linear regression is to fit data with a straight line

$$y = b_0 + b_1x$$

Recall equation for a straight line $y = mx + b$.

- Here b_0 is an intercept and b_1 is a slope (rise/run).
- We will discuss the statistical model later.
- The goal is to find b_0, b_1 for the best fitting line.
- The approach is least squares.

Simple Linear Regression

Least squares

- Find b_0, b_1 that minimize the sum of squares

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2$$

where y_i is the observed value and \hat{y}_i is the fitted value $\hat{b}_0 + \hat{b}_1 x_i$.

- FACT: The best fitting line has slope and intercept:

$$\begin{aligned}\hat{b}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n} (\sum_{i=1}^n x_i) (\sum_{i=1}^n y_i)}{\sum_{i=1}^n x_i^2 - \frac{1}{n} (\sum_{i=1}^n x_i)^2}, \\ \hat{b}_0 &= \bar{y} - \hat{b}_1 \bar{x}.\end{aligned}$$

Simple Linear Regression

Least squares

In the snow fall example,

$$\sum_{i=1}^n x_i = 24.40, \sum_{i=1}^n x_i^2 = 107.42, \sum_{i=1}^n x_i y_i = 154.07$$
$$\sum_{i=1}^n y_i = 35.50, \sum_{i=1}^n y_i^2 = 222.03.$$

Thus

$$\hat{b}_1 = \frac{154.07 - 24.40 \times 35.50/7}{107.42 - 24.4 \times 24.40/7} = \frac{30.33}{22.37} = 1.356$$
$$\hat{b}_0 = 35.50/7 - 1.356 \times 24.40/7 = 0.345$$

We may also predict y at say $x = 4$,

$$\hat{y} = 0.345 + 1.356 \times 4 = 5.77$$

How to account for uncertainty in the fitted line and variation?

Simple Linear Regression

SLR model

- Model y by random variable Y .
- Regard x as fixed, although x could be random.
- Consider the model of Y conditional on x ($[Y|x]$) such that

$$E(Y|x) = b_0 + b_1x.$$

where b_0, b_1 are fixed unknown parameters (intercept, slope) characterizing the relationship between x and y .

- The formal simple linear regression (SLR) model is:

$$Y_i = b_0 + b_1x_i + e_i$$

where $e_i \sim \text{iid}N(0, \sigma^2)$, $i = 1, \dots, n$.

- Y is called a *dependent variable* or *response variable*.
- x is called an *independent variable* or *covariate*.
- e 's are called *errors*.

Simple Linear Regression

Assumptions

1. The model is correct:

$$E(Y_i) = b_0 + b_1x_i$$

That is, a straightline relationship between y and x .

2. Errors e_i are independent.
3. Errors e_i have homogeneous variance: $Var(e_i) = \sigma^2$.
4. Errors e_i have normal distribution: $e_i \sim N(0, \sigma^2)$.

Remarks

- σ^2 is sometimes written as σ_e^2 .
- Equivalently,
 - 2' Y_i are independent.
 - 3' Y_i have homogeneous variance: $Var(Y_i) = \sigma^2$.
 - 4' Y_i have normal distribution: $Y_i \sim N(b_0 + b_1x_i, \sigma^2)$.

Simple Linear Regression

Model parameters

Estimate model parameters b_0, b_1, σ^2 by estimators $\hat{b}_0, \hat{b}_1, \hat{\sigma}^2$ (MSErr).

Test $H_0 : b_1 = 0$

- Analysis of variance (ANOVA)
- T-test

Simple Linear Regression

ANOVA for testing $H_0 : b_1 = 0$

Partition sum of squares (SS):

$$SSTotal = SSReg + SSErr,$$

where

$$SSTotal = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2$$

$$df = n - 1$$

$$SSReg = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$
$$= \hat{b}_1 \left[\sum_{i=1}^n x_i y_i - \frac{1}{n} \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right) \right]$$

$$df = 1$$

$$SSErr = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = SSTot - SSReg$$

$$df = n - 2$$

where $\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i$ is the fitted value and $r_i = y_i - \hat{y}_i$ is the raw residual.

Simple Linear Regression

ANOVA for testing $H_0 : b_1 = 0$

In the snow fall example,

Source	df	SS	MS	F
Regression	1	41.13	41.13	239.13
Error	5	0.86	0.172	–
Total	6	41.99	–	–

- Estimate σ^2 by $s^2 = \text{MSErr} = 0.172$ on $\text{df} = 5$.
- Fact: Under $H_0 : b_1 = 0$,

$$F = \frac{\text{MSReg}}{\text{MSErr}} \sim F_{1,n-2}$$

- In the snow fall example, the observed

$$f = \frac{41.13}{0.172} = 239.13.$$

Compare with F on $\text{df} = (1,5)$, p-value is less than 0.01. Reject H_0 at 5% and there is strong evidence against $H_0 : b_1 = 0$.

- If $b_1 = 0$, then the model becomes $Y_i = b_0 + e_i$. Hence the test can be viewed as choosing between the model $Y_i = b_0 + e_i$ under H_0 and the model $Y_i = b_0 + b_1x_i + e_i$ under H_A .

Simple Linear Regression

T-test for $H_0 : b_1 = 0$

- Alternatively use \hat{b}_1 directly where

$$\hat{b}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

and thus \hat{b}_1 is a r.v.

- Note that \hat{b}_1 is a weighted sum of normal distributions and hence also has a normal distribution. Thus use t statistic of the form

$$\frac{\hat{b}_1 - \mu_{\hat{b}_1}}{s_{\hat{b}_1}}.$$

- $\mu_{\hat{b}_1} = E(\hat{b}_1) = b_1$ and

$$\begin{aligned} \text{Var}(\hat{b}_1) &= \text{Var}\left(\frac{\sum (x_i - \bar{x})Y_i}{\sum (x_i - \bar{x})^2}\right) = \frac{\sum \text{Var}((x_i - \bar{x})Y_i)}{(\sum (x_i - \bar{x})^2)^2} \\ &= \frac{\sum (x_i - \bar{x})^2 \sigma^2}{(\sum (x_i - \bar{x})^2)^2} = \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \end{aligned}$$

- Hence $s_{\hat{b}_1} = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}}$ and in the snow fall example

$$s_{\hat{b}_1} = \frac{\sqrt{0.172}}{\sqrt{22.37}} = 0.0877.$$

Simple Linear Regression

T-test for $H_0 : b_1 = 0$

- Fact: Under $H_0 : b_1 = 0$,

$$T = \frac{\hat{b}_1 - 0}{s_{\hat{b}_1}} \sim T_{n-2}$$

- In the snow fall example, $s_{\hat{b}_1} = 0.0877$ and thus the observed

$$t = \frac{1.356}{0.0877} = 15.46.$$

Compare with T on $df = 5$, the (two-tailed) p-value is less than 0.01. Reject H_0 at 5% and there is strong evidence against $H_0 : b_1 = 0$.

- Note that $t^2 = (15.46)^2 = 239.13 = f$. Again this relation holds only for F on $df = (1, \text{something})$.
- In general, under $H_0 : b_1 = b_1^*$,

$$T = \frac{\hat{b}_1 - b_1^*}{s_{\hat{b}_1}} \sim T_{n-2}$$

- A $(1 - \alpha)$ CI for b_1 is

$$\hat{b}_1 \pm t_{\alpha/2, n-2} s_{\hat{b}_1}.$$

Simple Linear Regression

T-test for $H_0 : b_0 = b_0^*$

- For inference of b_0 , use

$$\hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{x}$$

- Fact: \hat{b}_0 has a normal distribution with $E(\hat{b}_0) = b_0$ and

$$Var(\hat{b}_0) = \sigma^2 \left(\frac{1}{n} + \frac{(\bar{x})^2}{\sum (x_i - \bar{x})^2} \right)$$

- Thus

$$s_{\hat{b}_0} = s \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

- For the snow fall example,

$$s_{\hat{b}_0} = \sqrt{0.172} \times \sqrt{\frac{1}{7} + \frac{3.486^2}{22.37}} = 0.344$$

Simple Linear Regression

T-test for $H_0 : b_0 = b_0^*$

- Fact: Under $H_0 : b_0 = b_0^*$,

$$T = \frac{\hat{b}_0 - b_0^*}{s_{\hat{b}_0}} \sim T_{n-2}$$

- In the snow fall example, suppose $H_0 : b_0 = 0$ and since $s_{\hat{b}_0} = 0.344$, the observed

$$t = \frac{0.345 - 0}{0.344} = 1.00.$$

Compare with T on $df = 5$, the p-value $2 \times P(T_5 \geq 1.00) > 0.10$. Do not reject H_0 and there is no evidence against $H_0 : b_0 = 0$.

- Suppose $H_0 : b_0 = 0$. If $b_0 = 0$, then the model becomes $Y_i = b_1x_i + e_i$. Hence the test can be viewed as choosing between the model $Y_i = b_1x_i + e_i$ under H_0 and the model $Y_i = b_0 + b_1x_i + e_i$ under H_A .
- A $(1 - \alpha)$ CI for b_0 is

$$\hat{b}_0 \pm t_{\alpha/2, n-2} s_{\hat{b}_0}.$$

Simple Linear Regression

Estimation vs prediction

- Consider a simpler model

$$Y_i = \mu_i + e_i$$

where $e_i \sim \text{iid } N(0, \sigma^2)$.

- Then

$$\hat{Y}_{\text{est}} = \bar{Y}$$

estimates μ with

$$\text{Var}(\hat{Y}_{\text{est}}) = \text{Var}(\bar{Y}) = \frac{\sigma^2}{n}$$

- Also

$$\hat{Y}_{\text{pred}} = \bar{Y}$$

predicts a future observation with

$$\text{Var}(\hat{Y}_{\text{pred}}) = \text{Var}(\bar{Y} + e) = \frac{\sigma^2}{n} + \sigma^2$$

Simple Linear Regression

Inference of the fitted line

- Estimate (predict) Y at a given x^* of interest by

$$\hat{Y} = \hat{b}_0 + \hat{b}_1 x^*$$

- In the snow fall example, suppose $x^* = 6$, then the estimated (predicted) y is

$$\hat{y} = 0.345 + 1.356 \times 6 = 8.48.$$

- But the standard error depends on the objective.
- Case 1: use \hat{Y} to estimate the true value $b_0 + b_1 x^*$ for a given x^* .
- Case 2: use \hat{Y} to predict a future obs for a given x^* .

Simple Linear Regression

Case 1: estimation

- If \hat{Y} is an estimator of the true value $b_0 + b_1x^*$, then denote $\hat{Y} = \hat{b}_0 + \hat{b}_1x^*$ by \hat{Y}_{est} .
- Then we have

$$E(\hat{Y}_{\text{est}}) = E(\hat{b}_0 + \hat{b}_1x^*) = b_0 + b_1x^*$$
$$Var(\hat{Y}_{\text{est}}) = Var(\hat{b}_0 + \hat{b}_1x^*) = \sigma^2 \left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right)$$

- Hence

$$s_{\hat{Y}_{\text{est}}} = s \times \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

- A $(1 - \alpha)$ CI for $b_0 + b_1x^*$ is

$$\hat{y}_{\text{est}} \pm t_{\alpha/2, n-2} s_{\hat{Y}_{\text{est}}}$$

Simple Linear Regression

Remarks

- In the snow fall example, $\bar{x} = 3.486$. For $x^* = 6$,

$$s_{\hat{Y}_{\text{est}}} = \sqrt{0.172} \times \sqrt{\frac{1}{7} + \frac{(6 - 3.486)^2}{22.37}} = 0.271.$$

- A 95% CI for $b_0 + b_1x^*$ is

$$8.48 \pm 2.571 \times 0.271$$

which is $[7.78, 9.18]$ or 8.48 ± 0.70 .

- Note that for $x^* = \bar{x} = 3.486$,

$$s_{\hat{Y}_{\text{est}}} = \sqrt{0.172 \times \frac{1}{7}} = 0.157.$$

- In general, $s_{\hat{Y}_{\text{est}}}$ is larger when x^* is far away from \bar{x} and smallest when $x^* = \bar{x}$.
- Caution against extrapolation.

Simple Linear Regression

Case 2: prediction

- If \hat{Y} is a predictor of a new/future observation, then denote $\hat{Y} = \hat{b}_0 + \hat{b}_1 x^*$ by \hat{Y}_{pred} .
- Then we have

$$\begin{aligned} E(\hat{Y}_{\text{pred}}) &= E(\hat{b}_0 + \hat{b}_1 x^*) = b_0 + b_1 x^* \\ \text{Var}(\hat{Y}_{\text{pred}}) &= \text{Var}(\hat{b}_0 + \hat{b}_1 x^* + e) \\ &= \sigma^2 \left(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right) \end{aligned}$$

- Hence

$$s_{\hat{Y}_{\text{pred}}} = s \times \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

- A $(1 - \alpha)$ prediction interval (PI) is

$$\hat{y}_{\text{pred}} \pm t_{\alpha/2, n-2} s_{\hat{Y}_{\text{pred}}}$$

Simple Linear Regression

Remarks

- In the snow fall example, for $x^* = 6$,

$$s_{\hat{Y}_{\text{pred}}} = \sqrt{0.172} \times \sqrt{1 + \frac{1}{7} + \frac{(6 - 3.486)^2}{22.37}} = 0.495.$$

- A 95% PI is

$$8.48 \pm 2.571 \times 0.495$$

which is $[7.21, 9.75]$ or 8.48 ± 1.27 .

- How about predicting Y at $x^* = 14$? Again caution against extrapolation.

Simple Linear Regression

Model fitting

- A useful quantity for assessing the overall regression fit is the *coefficient of determination*:

$$R^2 = \frac{\text{SS Regression}}{\text{SS Total}}$$

- R^2 represents the proportion of the total SS that is explained by the regression model.
- In the snow fall example,

$$R^2 = \frac{41.13}{41.99} = 0.98$$

which is very high.

Simple Linear Regression

Model diagnostics

- Recall the four model assumptions:

1. The model is correct:

$$E(Y_i) = b_0 + b_1x_i$$

That is, a straightline relationship between y and x .

2. Errors e_i are independent.
 3. Errors e_i have homogeneous variance: $Var(e_i) = \sigma^2$.
 4. Errors e_i have normal distribution: $e_i \sim N(0, \sigma^2)$.
- Check model assumptions by examining the residuals

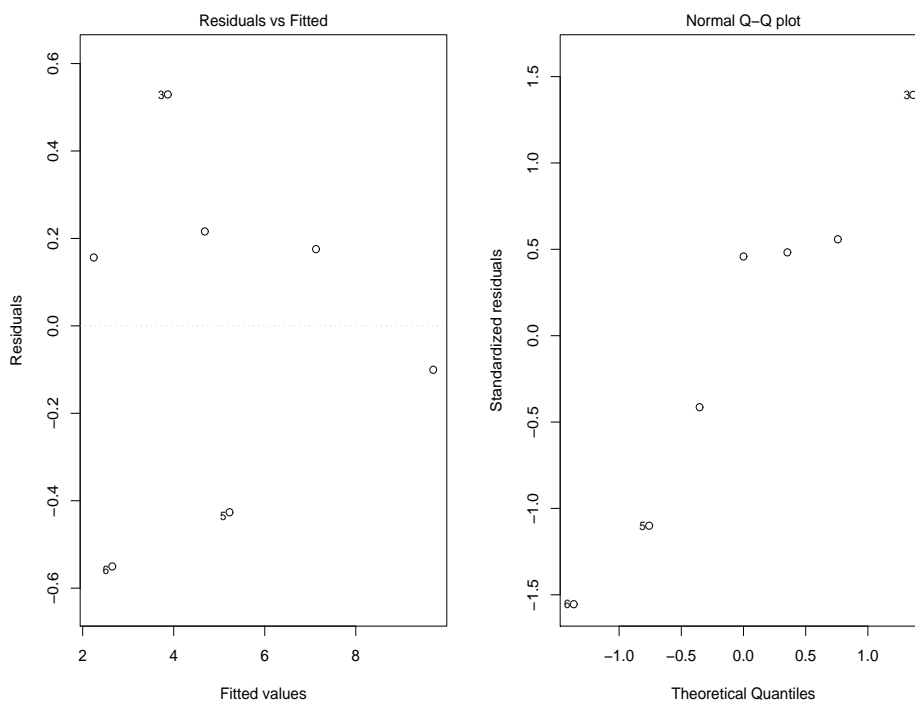
$$r_i = y_i - \hat{y}_i$$

- Note that $\sum_{i=1}^n r_i = 0$ and $\sum_{i=1}^n r_i^2 = \text{SSError}$.
- Residual plot: r_i versus \hat{y}_i .
- The assumptions are probably OK if the residual plot is a random scatter. Otherwise various patterns may indicate problems such as wrong model, or nonhomogeneous variance, or outliers.
- It may be hard to interpret when n is small.

Simple Linear Regression

Model diagnostics

In the snow fall example, the residual plot and the normal scores plot are shown below.



Simple Linear Regression

Key R Commands

```
> x = c(3.2, 1.4, 2.6, 6.9, 3.6, 1.7, 5.0)
> y = c(4.9, 2.4, 4.4, 9.6, 4.8, 2.1, 7.3)
> snow.lm = lm(y~x) #SLR
> summary(snow.lm)
```

Call:

```
lm(formula = y ~ x)
```

Residuals:

1	2	3	4	5	6	7
0.2159	0.1564	0.5294	-0.1005	-0.4264	-0.5504	0.1755

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.34552	0.34688	0.996	0.365
x	1.35579	0.08855	15.311	2.16e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4188 on 5 degrees of freedom

Multiple R-Squared: 0.9791, Adjusted R-squared: 0.9749

F-statistic: 234.4 on 1 and 5 DF, p-value: 2.156e-05

```
> anova(snow.lm) #ANOVA
```

Analysis of Variance Table

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x	1	41.117	41.117	234.43	2.156e-05 ***
Residuals	5	0.877	0.175		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

>

```
> #scatter plot
```

```
> plot(x,y,xlab="inches of snow", ylab="hrs to clear streets", main="snow fall example")
```

```
> lines(x,fitted(snow.lm))
```

```
> predict(snow.lm, data.frame(x=6), se.fit=T, interval="confidence")
$fit
      fit      lwr      upr
[1,] 8.480278 7.778058 9.182498

$se.fit
[1] 0.2731755

$df
[1] 5

$residual.scale
[1] 0.4188009

> predict(snow.lm, data.frame(x=6), se.fit=T, interval="prediction")
$fit
      fit      lwr      upr
[1,] 8.480278 7.194939 9.765618

$se.fit
[1] 0.2731755

$df
[1] 5

$residual.scale
[1] 0.4188009

# model diagnostics
> par(mfrow=c(1,2))
> plot(snow.lm, which=1)
> plot(snow.lm, which=2)
```

Correlation

An overview

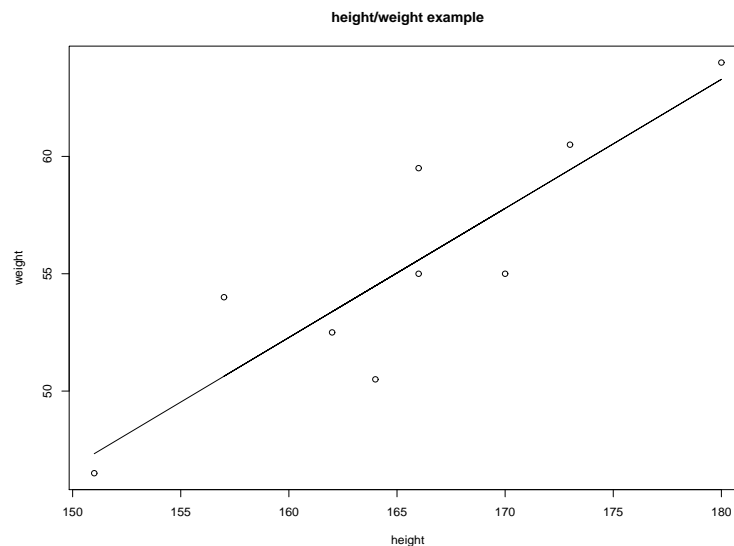
In simple linear regression, we predict Y given x . Now we are interested in how two variables are related to each other and hence X and Y are treated symmetrically.

Height/weight example

Relation of height (X) and weight (Y) of adult women

$X(\text{cm})$: 166 162 170 164 157 173 180 166 151

$Y(\text{kg})$: 59.5 52.5 55.0 50.5 54.0 60.5 64.0 55.0 46.5



Correlation

Simple linear regression

```
> ht = c(166, 162, 170, 164, 157, 173, 180, 166, 151)
> wt = c(59.5, 52.5, 55.0, 50.5, 54.0, 60.5, 64.0, 55.0, 46.5)
> htwt.lm = lm(wt~ht) #SLR
> summary(htwt.lm)
```

```
Call:
lm(formula = wt ~ ht)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-3.9832 -0.8829 -0.5834  1.0658  3.9166
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -35.7355    18.9906  -1.882  0.10190
ht           0.5501     0.1146   4.798  0.00197 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 2.771 on 7 degrees of freedom
Multiple R-Squared:  0.7668,    Adjusted R-squared:  0.7335
F-statistic: 23.02 on 1 and 7 DF,  p-value: 0.001970
```

```
> anova(htwt.lm) #ANOVA
Analysis of Variance Table
```

```
Response: wt
      Df Sum Sq Mean Sq F value    Pr(>F)
ht     1  176.801  176.801   23.023 0.001970 **
Residuals  7   53.755    7.679
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> cor(ht, wt)
[1] 0.8756973
> cor(wt, ht)
[1] 0.8756973
>
```

Correlation

Model

- X and Y are both random and have a bivariate distribution.
- The most useful distribution is a bivariate normal distribution.
- Probability density surface can be plotted using a 3D or contour plot.
- $Y|X = x$ is normal and so is $X|Y = y$.

Correlation

Population correlation coefficient

- $\rho = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}$ is the population correlation coefficient between X and Y .
- ρ is a measure of linear relationship between X and Y .
- $-1 \leq \rho \leq 1$.
- $\rho = 1$ indicates perfect positive correlation.
- $0 < \rho < 1$ indicates modest positive correlation.
- $\rho = 0$ indicates no linear relationship.
- $-1 < \rho < 0$ indicates modest negative correlation.
- $\rho = -1$ indicates perfect negative correlation.

Correlation

Sample correlation coefficient

- Based on data $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, the sample correlation coefficient

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

estimates ρ .

- Note the symmetry between x and y in r .
- Working formula is

$$r = \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n} (\sum_{i=1}^n x_i) (\sum_{i=1}^n y_i)}{\sqrt{\sum_{i=1}^n x_i^2 - \frac{1}{n} (\sum_{i=1}^n x_i)^2} \sqrt{\sum_{i=1}^n y_i^2 - \frac{1}{n} (\sum_{i=1}^n y_i)^2}}$$

- For the height-weight data, the observed r is

$$r = \frac{82630 - 82308.61}{\sqrt{584.22} \sqrt{230.56}} = 0.876.$$

- Note that $\hat{b}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = r \times \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$

Correlation

Statistical inference

- Assume X and Y are from a bivariate normal distribution.
- Test $H_0 : \rho = 0$ versus $H_a : \rho \neq 0$.
- Use $T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim T_{n-2}$ under H_0 .
- For the height-weight data, the observed

$$t = \frac{0.876 \times \sqrt{7}}{\sqrt{1 - 0.876^2}} = 4.80$$

on $df = 7$ with a p-value < 0.01 .

- Remark:

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{\hat{b}_1}{s.e.(\hat{b}_1)}.$$

Thus testing $H_0 : \rho = 0$ is the same as testing $H_0 : b_1 = 0$.

- Remark: CI uses Fisher transformation and is more involved. See the bluebook.

Correlation

More relation to regression

For the height-weight data, ANOVA for regression is

Source	df	SS	MS	F
regression	1	176.80	176.80	23.02
error	7	53.76	7.68	–
total	8	230.56	–	–

- $f = 23.02 = (4.8)^2 = t$.
- $R^2 = \frac{176.80}{230.56} = 0.767 = (0.876)^2 = r^2$.

Final remarks

- ρ measures linear relationship. Y can be perfectly related to X , but not linear (e.g., $Y = X^2$).
- It is easy to find spurious correlation. No causality established here.

Categorical Data

An overview

- Case 1: Binomial, 1 sample
- Case 2: Multinomial, 1 sample
- Case 3: Binomial, 2 samples
- Case 4: Binomial, multiple samples

Case 1: Binomial, 1 sample

Example

- For $Y \sim B(100, p)$, test $H_0 : p = 0.6$ versus $H_0 : p \neq 0.6$
- Suppose we observe 72 heads, by normal approximation, we have

$$Y_{NA} \sim N(60, 24)$$

and

$$\hat{p}_{NA} \sim N(0.6, 0.24)$$

under H_0 .

- Thus

$$Z = \frac{Y_{NA} - 60}{\sqrt{24}} \sim N(0, 1)$$

and the observed

$$z = \frac{72 - 60}{\sqrt{24}} = 2.45$$

with p-value = $0.007 \times 2 = 0.014$.

Case 1: Binomial, 1 sample

New approach

- Draw the following contingency tables

	H	T	
observed data	72	28	100
	H	T	
expected values	60	40	100

- Compute

$$X^2 = \sum_{\text{all possibilities}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

- FACT: If H_0 is true, then X^2 is approximately χ^2 on 1 df.
- Thus the observed

$$x^2 = \frac{(72 - 60)^2}{60} + \frac{(28 - 40)^2}{40} = 6$$

and compared with χ^2 on 1 df, p-value = $P(\chi_1^2 \geq 6)$ [one-sided]. From Table B, the p-value is between 0.01 and 0.025.

- $x^2 = 6 = (2.45)^2 = z^2$ holds for χ_1^2 .
- Same condition as Z -test for a good approximation: $np \geq 5$ and $n(1 - p) \geq 5$.

Case 2: Multinomial, 1 sample

Example

A specially constructed die is such that three sides are labeled 1, the other three sides are labeled 2, 3, and 4. Roll the die 240 times for testing $H_0 : p_1 = 1/2, p_2 = p_3 = p_4 = 1/6$ versus $H_a : \text{not } H_0$, where $p_i = \text{probability that the die comes up with } i$. Note that $p_1 + p_2 + p_3 + p_4 = 1$.

	1	2	3	4		1	2	3	4		
obs	108	27	39	66	240	exp	120	40	40	40	240

$$X^2 = \sum_{\text{all possibilities}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

Thus the observed

$$\begin{aligned} x^2 &= \frac{(108 - 120)^2}{120} + \frac{(27 - 40)^2}{40} + \frac{(39 - 40)^2}{40} + \frac{(66 - 40)^2}{40} \\ &= 1.2 + 4.225 + 0.025 + 16.9 = 22.35 \end{aligned}$$

and compared with χ^2 on 3 df, p-value = $P(\chi_3^2 \geq 22.35) < 0.01$.

Case 2: Multinomial, 1 sample

Remarks

- In general, $df = \# \text{ of cells} - 1$.
- Formally, the model is a multinomial distribution (a generalization of binomial) with 3 assumptions:
 1. n independent trials.
 2. Each trial has k mutually exclusive outcomes.
 3. Constant probability for each outcome in each trial $p_i =$ prob. of the i -th outcome.

Then $Y_i = \#$ of i -th outcome in n trial, $i = 1, \dots, k$, follows a multinomial distribution.

- Conditions for χ^2 test:
 1. All expected values > 1 .
 2. At least 80% of the expected values ≥ 5 .

Case 3: Binomial, 2 samples

Example

- Compare two treatments A and B. For A, there are 71 successes among 105 trials. For B, there are 45 successes among 87 trials. Test $H_0 : p_1 = p_2$ versus $H_0 : p_1 \neq p_2$.

- We compute

$$\hat{p}_1 = 71/105 = 0.676$$

for A and

$$\hat{p}_2 = 45/87 = 0.517$$

for B.

- Then use the fact that

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p(1-p)(1/n_1 + 1/n_2)}} \sim N(0, 1)$$

under H_0 .

- Here the pooled

$$p = \frac{71 + 45}{105 + 87} = \frac{116}{192} = 0.604$$

- Thus the observed $z = 2.24$ with a p-value of 0.025.

Case 3: Binomial, 2 samples

New approach

observed	A	B	
success	71	45	116
failure	34	42	76
	105	87	192

Under H_0 , the probability of success is the pooled $p = \frac{116}{192}$.

expected	A	B	
success	63.44	52.56	116
failure	41.56	34.44	76
	105	87	192

Under H_0 ,

the expected # of successes for A is $105 \times \frac{116}{192} = 63.44$,

the expected # of failures for A is $105 \times \frac{76}{192} = 41.56$,

the expected # of successes for B is $87 \times \frac{116}{192} = 52.56$,

the expected # of failures for B is $87 \times \frac{76}{192} = 34.44$.

$$X^2 = \sum_{\text{all possibilities}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

Case 3: Binomial, 2 samples

New approach

- FACT: If H_0 is true, then X^2 is approximately χ^2 on 1 df.
- Thus the observed

$$\begin{aligned}x^2 &= \frac{(71 - 63.44)^2}{63.44} + \frac{(34 - 41.56)^2}{41.56} + \frac{(45 - 52.56)^2}{52.56} + \frac{(42 - 34.44)^2}{34.44} \\ &= 5.023\end{aligned}$$

compared with χ^2 on 1 df, p-value = $P(\chi_1^2 \geq 5.023)$ [one-sided]. From Table B, the p-value is between 0.025 and 0.05.

- $x^2 = 5.023 = (2.24)^2 = z^2$.
- df = 1 because given the marginals and the total, there is only 1 piece of independent information.

Case 4: Binomial, multiple samples

Contingency tables

Compare 4 species (1–4) of pine for disease resistance in a study:

observed	1	2	3	4	
disease	22	10	15	20	67
no disease	29	28	29	17	103
	51	38	44	37	170

Under $H_0 : p_1 = p_2 = p_3 = p_4$,

expected	1	2	3	4	
disease	20.10	14.98	17.34	14.58	67
no disease	30.90	23.02	26.66	22.42	103
	51	38	44	37	170

Case 4: Binomial, multiple samples

Contingency tables

Let p_i denote the disease rate for the i th species of pine. Then $\hat{p}_1 = 0.431$, $\hat{p}_2 = 0.263$, $\hat{p}_3 = 0.341$, $\hat{p}_4 = 0.541$.

Under $H_0 : p_1 = p_2 = p_3 = p_4$,

expected # of diseased pines for species 1 = $51 \times \frac{67}{170} = 20.10$,

expected # of healthy pines for species 1 = $51 \times \frac{103}{170} = 30.90$,

expected # of diseased pines for species 2 = $38 \times \frac{67}{170} = 14.98$,

expected # of healthy pines for species 2 = $38 \times \frac{103}{170} = 23.02$,

etc...

Let

$$X^2 = \sum_{\text{all possibilities}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

Case 4: Binomial, multiple samples

Fact and remarks

- FACT: If H_0 is true, then X^2 is approximately χ^2 on 3 df.
- Thus the observed

$$x^2 = \frac{(22 - 20.10)^2}{20.10} + \frac{(29 - 30.90)^2}{30.90} + \dots = 6.876$$

compared with χ^2 on 3 df, p-value = $P(\chi_3^2 \geq 6.876)$ [one-sided]. From Table B, the p-value is between 0.05 and 0.10.

- This idea can be extended to general $r \times c$ case, where r is the # of rows and c is the # of columns. Then df = $(r - 1) \times (c - 1)$.
- This is an overall test. It may be important to look at individual pieces.
- Conditions for the χ^2 test are again:
 1. All expected values > 1 .
 2. At least 80% of the expected values ≥ 5 .

Categorical Data

Key R commands

```
> # case 1 binomial 1 sample
> prop.test(72, 100, p=0.6, correct=F)
```

1-sample proportions test without continuity correction

```
data: 72 out of 100, null probability 0.6
X-squared = 6, df = 1, p-value = 0.01431
alternative hypothesis: true p is not equal to 0.6
95 percent confidence interval:
 0.6251197 0.7986031
sample estimates:
  p
0.72
```

```
> chisq.test(c(72,28), p=c(0.6,0.4), correct=F)
```

Chi-squared test for given probabilities

```
data: c(72, 28)
X-squared = 6, df = 1, p-value = 0.01431
```

```
> # case 2 multinomial 1 sample
> chisq.test(c(108,27,39,66), p=c(1/2,1/6,1/6,1/6), correct=F)
```

Chi-squared test for given probabilities

```
data: c(108, 27, 39, 66)
X-squared = 22.35, df = 3, p-value = 5.516e-05
```

```
> # case 3 binomial 2 samples
> prop.test(c(71,45), c(105,87), correct=F)
```

2-sample test for equality of proportions without continuity correction

```
data: c(71, 45) out of c(105, 87)
X-squared = 5.0264, df = 1, p-value = 0.02496
```

```

alternative hypothesis: two.sided
95 percent confidence interval:
 0.02097751 0.29692068
sample estimates:
 prop 1    prop 2
0.6761905 0.5172414

> matrix(c(71,34,45,42),2,2)
  [,1] [,2]
[1,]  71  45
[2,]  34  42
> chisq.test(matrix(c(71,34,45,42),2,2), correct=F)

```

Pearson's Chi-squared test

```

data: matrix(c(71, 34, 45, 42), 2, 2)
X-squared = 5.0264, df = 1, p-value = 0.02496

```

```

> # case 4 binomial multiple samples
> prop.test(c(22,10,15,20), c(51,38,44,37), correct=F)

```

4-sample test for equality of proportions without continuity correction

```

data: c(22, 10, 15, 20) out of c(51, 38, 44, 37)
X-squared = 6.8694, df = 3, p-value = 0.07618
alternative hypothesis: two.sided
sample estimates:

```

```

 prop 1    prop 2    prop 3    prop 4
0.4313725 0.2631579 0.3409091 0.5405405

```

```

> matrix(c(22,29,10,28,15,29,20,17),2,4)
  [,1] [,2] [,3] [,4]
[1,]  22  10  15  20
[2,]  29  28  29  17
> chisq.test(matrix(c(22,29,10,28,15,29,20,17),2,4), correct=F)

```

Pearson's Chi-squared test

```

data: matrix(c(22, 29, 10, 28, 15, 29, 20, 17), 2, 4)
X-squared = 6.8694, df = 3, p-value = 0.07618

```