STAT 350 Lecture 8: ANOVA and Regression

Explanation of Variability

(Chapter 11, 13 of WMMY)

Outline

Explanation of Variation

Three Types

For the assess of association between continuous response and certain type of predictor (also called factor), we essentially consider the explanation of variation in the response by the predictor.

1. Factor with two levels: two sample comparison (solved)
2. Factor with more than two levels: analysis of variance (ANOVA)
3. Continuous factor: regression analysis

Example: Middle School Class Data

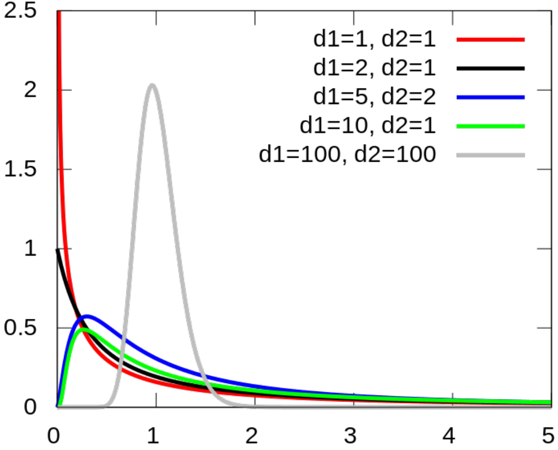
Outline

Two sample comparison

- $\{X_{11}, X_{12}, \cdots, X_{1n} :\overset{\text{IID}}{\sim} \mathrm{N}(\mu_1, \sigma^2)\}$ and $\{X_{21}, X_{22}, \cdots, X_{2n} :\overset{\text{IID}}{\sim} \mathrm{N}(\mu_2, \sigma^2)\}$ are two **independent** samples of the same size n.
- Hypothesis: $H_0 : \mu_1 = \mu_2$.
- Re-formulate: $X_{ij} = \mu_i + \epsilon_{ij}$ with $i = 1, 2$ and $j = 1, 2, \cdots, n$. $\epsilon_{ij} \overset{\text{IID}}{\sim} \mathrm{N}(0, \sigma^2)$.
- T test statistic: $T = \frac{\bar{X}_1 - \bar{X}_2}{s_{\text{pool}}\sqrt{2/n}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{(s_1^2 + s_2^2)/n}} \overset{H_0}{\sim} t(2n-2)$.
- Another test statistic: $F = \frac{n[(\bar{X}_1 - \bar{X})^2 + (\bar{X}_2 - \bar{X})^2]}{[\sum_{j=1}^n (X_{1j} - \bar{X}_1)^2 + \sum_{j=1}^n (X_{2j} - \bar{X}_2)^2]/(2(n-1))} = \frac{n[(\bar{X}_1 - \bar{X})^2 + (\bar{X}_2 - \bar{X})^2]}{(s_1^2 + s_2^2)/2} \overset{H_0}{\sim} F(1, 2(n-1))$

Another distribution: F distribution

$X \sim F(d_1, d_2)$



Analysis of Variance (ANOVA)

- $\{X_{i1}, X_{i2}, \cdots, X_{in} :\overset{\text{IID}}{\sim} \text{N}(\mu_i, \sigma^2)\}$, $i = 1, 2, \cdots, k$ are k **independent** samples of the same size n.
- Hypothesis:
  $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k, H_1$: not all the means are equal.
- Re-formulate: $X_{ij} = \mu_i + \epsilon_{ij}$ with $i = 1, 2, \cdots, k$ and $j = 1, 2, \cdots, n$. $\epsilon_{ij} \overset{\text{IID}}{\sim} \text{N}(0, \sigma^2)$.
- F test statistic:

$$F = \frac{[n \sum_{i=1}^{k} (\bar{X}_i - \bar{X})^2]/(k-1)}{[\sum_{i=1}^{k} \sum_{j=1}^{n} (X_{ij} - \bar{X}_i)^2]/(k(n-1))} \overset{H_0}{\sim} F(k-1, k(n-1))$$

Variance Decomposition

The sum of squares identity:

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij}-\bar{X})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij}-\bar{X}_i)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{X}_i-\bar{X})^2$$
$$= \sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij}-\bar{X}_i)^2 + n\sum_{i=1}^{k}(\bar{X}_i-\bar{X})^2$$

- total sum of squares: SST$=\sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij}-\bar{X})^2$;
- regression sum of squares: SSR$=n\sum_{i=1}^{k}(\bar{X}_i-\bar{X})^2$;
- error sum of squares: SSE$=\sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij}-\bar{X}_i)^2$.

F Test

$$\text{SST} = \sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2 + n\sum_{i=1}^{k}(\bar{X}_i - \bar{X})^2 = \text{SSE} + \text{SSR}$$

Source of Variation	Sum of Squares	Degree of Freedom	Mean Squares	F Statistic
Regression	SSR	k-1	MSR=SSR/(k-1)	F=$\frac{\text{MSR}}{\text{MSE}}$
Error	SSE	k(n-1)	MSE=SSE/(k(n-1))	
Total	SST	kn-1		

- Hypothesis:
  $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k, H_1$: not all the means are equal.
- F test statistic:

$$F = \frac{\text{MSR}}{\text{MSE}} \overset{H_0}{\sim} F(k-1, k(n-1))$$

Example: Drug and Pain

A drug company tested three formulations of a pain relief medicine for migraine headache sufferers. For the experiment 27 volunteers were selected and 9 were randomly assigned to one of three drug formulations. The subjects were instructed to take the drug during their next migraine headache episode and to report their pain on a scale of 1 to 10 (10 being most pain).

```
Drug A 4 5 4 3 2 4 3 4 4
Drug B 6 8 4 5 4 6 5 8 6
Drug C 6 7 6 6 7 5 6 5 5
```

```
> pain = c(4, 5, 4, 3, 2, 4, 3, 4, 4, 6, 8, 4, 5, 4,
           6, 5, 8, 6, 6, 7, 6, 6, 7, 5, 6, 5, 5)
> drug = c(rep("A",9), rep("B",9), rep("C",9))
> migraine = data.frame(pain,drug)
```

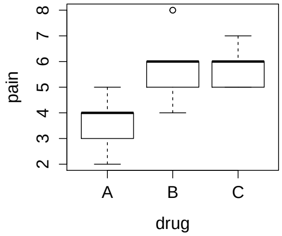
Example: Drug and Pain

	xij	xij-xbar	xij-xibar	xibar-xbar
[1,]	4	1.23456790	0.11111111	2.0864198
[2,]	5	0.01234568	1.77777778	2.0864198
[3,]	4	1.23456790	0.11111111	2.0864198
[4,]	3	4.45679012	0.44444444	2.0864198
[5,]	2	9.67901235	2.77777778	2.0864198
[6,]	4	1.23456790	0.11111111	2.0864198
[7,]	3	4.45679012	0.44444444	2.0864198
[8,]	4	1.23456790	0.11111111	2.0864198
[9,]	4	1.23456790	0.11111111	2.0864198
[10,]	6	0.79012346	0.04938272	0.4444444
[11,]	8	8.34567901	4.93827160	0.4444444
[12,]	4	1.23456790	3.16049383	0.4444444
[13,]	5	0.01234568	0.60493827	0.4444444
[14,]	4	1.23456790	3.16049383	0.4444444
[15,]	6	0.79012346	0.04938272	0.4444444
[16,]	5	0.01234568	0.60493827	0.4444444
[17,]	8	8.34567901	4.93827160	0.4444444
[18,]	6	0.79012346	0.04938272	0.4444444
[19,]	6	0.79012346	0.01234568	0.6049383
[20,]	7	3.56790123	1.23456790	0.6049383
[21,]	6	0.79012346	0.01234568	0.6049383
[22,]	6	0.79012346	0.01234568	0.6049383
[23,]	7	3.56790123	1.23456790	0.6049383
[24,]	5	0.01234568	0.79012346	0.6049383
[25,]	6	0.79012346	0.01234568	0.6049383
[26,]	5	0.01234568	0.79012346	0.6049383
[27,]	5	0.01234568	0.79012346	0.6049383
sum		56.66667	28.44444	28.22222

	Df	Sum Sq	Mean Sq	F value
drug	?	?	?	?
error	?	28.44	?	
---				
total		56.66		

Example: Drug and Pain

```
> plot(pain ~ drug, data=migraine)
> results = aov(pain ~ drug,
                 data=migraine)
> summary(results)
```




```
            Df Sum Sq Mean Sq F value   Pr(>F)
drug         2  28.22  14.111   11.91 0.000256 ***
Residuals   24  28.44   1.185
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Example: Drug and Pain

Conduct a hypothesis test for
$H_0 : \mu_A = \mu_B = \mu_C, H_1$: not all the means are equal

1. What is the F test statistic?
2. What is the p value of test?
3. What is the conclusion?

Example: Drug and Pain

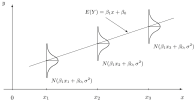
Conduct a hypothesis test for
$H_0 : \mu_A = \mu_B = \mu_C, H_1$: not all the means are equal

1. What is the F test statistic?
2. What is the p value of test?
3. What is the conclusion?

- Rejection region: $\{F > F_\alpha^*(2, 24)\}$
- p-value: $Fcdf(F_{\text{obs}}, 9^9, 2, 24) = 0.00026$

Outline

Scatter Plot Again



Model and assumptions

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

- y_i is the value of the response variable in the i-th trial
- α and β are parameters
- x_i is a known constant, the value of the predictor variable in the $i-$th trial
- $\varepsilon_1, \ldots, \varepsilon_n$ are independent $\mathrm{N}(0, \sigma^2)$ RVs.

Properties

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

- The response y_i is the sum of two components
  - Deterministic term $\alpha + \beta x_i$
  - Random term ϵ_i
- The expected response is $E(y_i) = \alpha + \beta x_i$.
- The variance of the response is $V(y_i) = \sigma^2$.
- $y_i \sim N(\alpha + \beta x_i, \sigma^2)$.

Sampling distribution of $\hat{\beta}$

- Recall the least square estimator:

$$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}, \quad \hat{\beta} = r\frac{s_y}{s_x}.$$

- If we write $\hat{\beta}$ as

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2},$$

  then it's clear that $\hat{\beta}$ is a *linear function* of $Y_1, \ldots, Y_n$.
- Therefore, the sampling distribution of $\hat{\beta}$ is Normal; indeed

$$\hat{\beta} \sim \mathrm{N}(\beta, \sigma^2 / \textstyle\sum_{i=1}^{n}(x_i - \overline{x})^2).$$

- But we don't know σ, so when we standardize $\hat{\beta}$ using an estimate of σ, we expect that a Student-t distribution will emerge.

Tests and CIs for β

- Key quantities are the estimate $\hat{\beta}$ and the standard error $\mathrm{SE}_{\hat{\beta}}$.
- The standard error comes from Minitab output, or can be calculated as

$$\mathrm{SE}_{\hat{\beta}} = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}} = \frac{s_y}{s_x}\sqrt{\frac{1-r^2}{n-2}} = \frac{\hat{\beta}}{r}\sqrt{\frac{1-r^2}{n-2}}.$$

  with $\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.
- Sampling distribution:

$$\frac{\hat{\beta} - \beta}{\mathrm{SE}_{\hat{\beta}}} \sim t(n-2).$$

Tests and CIs for β – cont.

- $100(1-\alpha)\%$ CI for β: $\hat{\beta} \pm t^*_{\alpha/2}(n-2) \times \mathrm{SE}_{\hat{\beta}}$.
- Hypothesis test about β:
  - Null hypothesis – $H_0 : \beta = 0$
  - Alternative hypothesis – one of

$$H_1 : \beta > 0, \quad H_1 : \beta < 0, \quad \text{or} \quad H_1 : \beta \neq 0.$$

  - Test statistic:

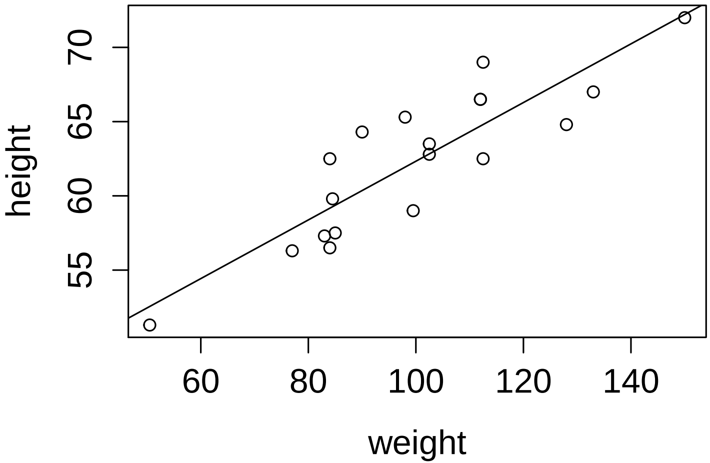
$$t = \frac{\hat{\beta}}{\mathrm{SE}_{\hat{\beta}}} \quad (\sim \mathrm{t}(n-2) \text{ when } H_0 \text{ is true}).$$

  - Compute critical region or p-value as we did before in t-tests, but now use $df = n - 2$.

Middle School Class Data Example

	name	sex	age	height	weight
1	Alice	F	13	56.5	84.0
2	Becka	F	13	65.3	98.0
3	Gail	F	14	64.3	90.0
4	Karen	F	12	56.3	77.0
5	Kathy	F	12	59.8	84.5
6	Mary	F	15	66.5	112.0
...	...	...	...	...	...
12	Guido	M	15	67.0	133.0
13	James	M	12	57.3	83.0
14	Jeffrey	M	13	62.5	84.0
15	John	M	12	59.0	99.5
16	Philip	M	16	72.0	150.0
17	Robert	M	12	64.8	128.0
18	Thomas	M	11	57.5	85.0
19	William	M	15	66.5	112.0

Example: Middle School Class Data



Example: Middle School Class Data

```
Call:
lm(formula = height ~ weight, data = classdata)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.2328 -1.8602 -0.2124  1.7970  4.1982 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 42.57014    2.67989  15.885 1.24e-11 ***
weight       0.19761    0.02616   7.555 7.89e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.527 on 17 degrees of freedom
Multiple R-squared:  0.7705,Adjusted R-squared:  0.757 
F-statistic: 57.08 on 1 and 17 DF,  p-value: 7.887e-07
```

Example: Middle School Class Data

1. Give a 90% CI for the slope.
2. Is the slope positive? Perform a test at the 1% level.

Example: Middle School Class Data

1. Give a 90% CI for the slope.
2. Is the slope positive? Perform a test at the 1% level.

1. $\overline{x} = 100.03, \overline{y} = 62.34, s_x = \sqrt{518.6520}, s_y = \sqrt{26.2869}, s_{xy}^2 = 102.4934$
2. $r = \frac{s_{xy}^2}{s_x s_y} = 0.8777813$
3. $\hat{\beta} = r\frac{s_y}{s_x} = 0.19761$
4. $\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x} = 42.57014$
5. $\mathrm{SE}_{\hat{\beta}} = \frac{\hat{\beta}}{r}\sqrt{\frac{1-r^2}{n-2}} = \frac{0.19761}{0.8777813}\sqrt{\frac{1-0.8777813^2}{17}} = 0.02615709$

- 90% CI for β: $\hat{\beta} \pm t^*_{\alpha/2}(n-2)\mathrm{SE}_{\hat{\beta}} = 0.19761 \pm 1.74 \times 0.02615709 = [0.152, 0.243]$
- $H_0 : \beta = 0, H_1 : \beta > 0$;
  $t_{obs} = \hat{\beta}/\mathrm{SE}_{\hat{\beta}} = 0.19761/0.02615709 = 7.555$;
  p-value=tcdf(7.555,9^9,17)=3.943e-7