Inference for Unstructured Multisample Studies (Ch. 7.1 and 7.4)

Will Landau

Iowa State University

Apr 25, 2013

Outline

The one-way ANOVA model

Residuals and fitted values

Variance estimation

Standardized residuals

Inference

The one-way ANOVA model

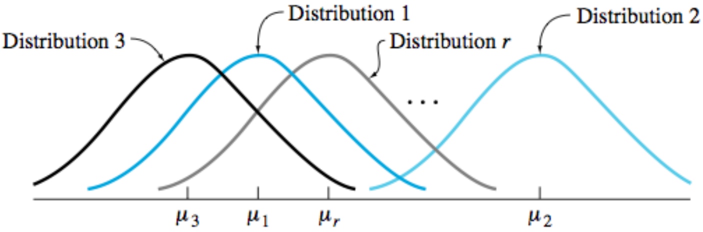
- Suppose we have:
  - Some response variable, Y
  - Some covariate factor, X, with levels $i = 1, 2, \ldots, I$ and n_i observations at level i.
- The **one-way ANOVA model**, sometimes called the one-way normal model, is:

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

where:
  - The ε_{ij}'s are iid $N(0, \sigma^2)$
  - μ_i is the true mean response at level i of the factor.
  - $j = 1, 2, \ldots, n_i$.

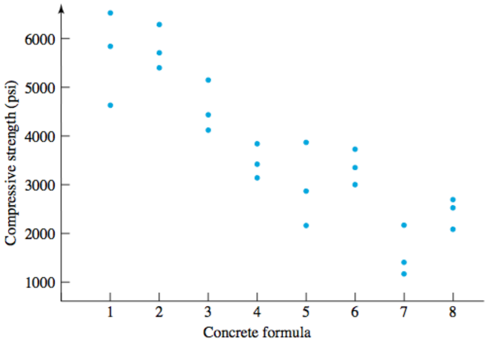
The one-way ANOVA model

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$



Example: concrete

- Compressive strengths of 8 different formulas of concrete:



- But the order of the numbers given to the formulas is meaningless. It wouldn't make sense to do a simple linear regression of strength on formula.

Example: concrete

- Instead of:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

with Y_i as strength and X_i as the formula index, we use:

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

where:

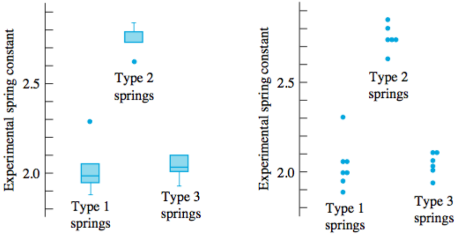
  - i is the formula index, $i = 1, 2, \ldots, 8$
  - j is the index of a specimen within the formula i group.

Example: springs

- Spring constants of three types of steel springs:

Empirical Spring Constants

Type 1 Springs	Type 2 Springs	Type 3 Springs
1.99, 2.06, 1.99	2.85, 2.74, 2.74	2.10, 2.01, 1.93
1.94, 2.05, 1.88	2.63, 2.74, 2.80	2.02, 2.10, 2.05
2.30		



Example: springs

- Doesn't make sense to regress exponential spring constant on spring type.
- Instead, we apply:

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

  where:
  - Y_{ij} is the exponential spring constant of spring type i spring number j.
  - μ_i is the true mean exponential spring constant of type i.
  - i is the formula index, $i = 1, 2, \ldots, 8$
  - j is the index of a specimen within the formula i group.

Outline

The one-way ANOVA model

Residuals and fitted values

Variance estimation

Standardized residuals

Inference

Fitted values

- Similarly to before, $\widehat{y}_{ij}$ is the fitted value corresponding to y_{ij}. It represents an estimate of the true mean response at factor level i and sample unit j.
- We treat all sample units equally, letting;

$$\widehat{y}_{ij} = \overline{y}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$$

  the average of all the responses at factor level i.
- We get $\widehat{y}_{ij} = \overline{y}_{i.}$ by minimizing the loss function:

$$S(\mu_1, \mu_2, \dots, \mu_I) = \sum_{ij} (y_{ij} - \mu_i)^2$$

  over all the choices of $\mu_1, \mu_2, \dots, \mu_I$, selecting $\overline{y}_{i.}$ to estimate μ_i.
- The residuals e_{ij} are then:

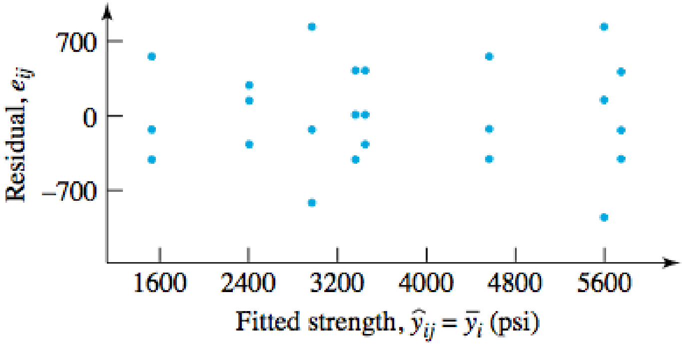
$$e_{ij} = y_{ij} - \overline{y}_{i.}$$

Example: concrete

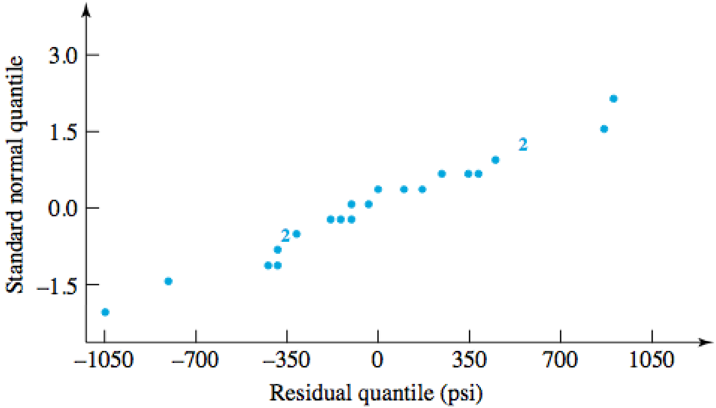
Example Computations of Residuals for the Concrete Strength Study

Specimen	i, Concrete Formula	y_{ij}, Compressive Strength (psi)	$\hat{y}_{ij} = \bar{y}_i$, Fitted Value	e_{ij}, Residual
1	1	5,800	5,635.3	164.7
2	1	4,598	5,635.3	−1,037.3
3	1	6,508	5,635.3	872.7
4	2	5,659	5,753.3	−94.3
5	2	6,225	5,753.3	471.7
⋮	⋮	⋮	⋮	⋮
22	8	2,051	2,390.7	−339.7
23	8	2,631	2,390.7	240.3
24	8	2,490	2,390.7	99.3

Example: concrete



Example: concrete



Outline

Variance estimation

- We can compute a sample variance for each factor level:

$$s_i^2 = \frac{1}{n_i - 1} \sum_j (y_{ij} - \overline{y}_{ij})^2$$

- And we can compute a **pooled sample variance**:

$$s_P^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_I - 1)s_I^2}{(n_1 - 1) + (n_2 - 1) + \cdots + (n_I - 1)}$$

- The pooled sample standard deviation is just $s_P = \sqrt{s_P^2}$

Variance estimation

- If $n = \sum_i n_i$, then:

$$s_P^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + \cdots + (n_I-1)s_I^2}{(n_1-1)+(n_2-1)+\cdots+(n_I-1)}$$
$$= \frac{(n_1-1)\left(\frac{1}{n_1-1}\right)\sum_j (y_{1j}-\overline{y}_1)^2 + \cdots + (n_I-1)\left(\frac{1}{n_I-1}\right)\sum_j (y_{Ij}-\overline{y}_I)^2}{n-I}$$
$$= \frac{1}{n-I}\sum_{ij}(y_{ij}-\overline{y}_i)^2$$
$$= \frac{1}{n-I}\sum_{ij} e_{ij}^2$$

- As it turns out,

$$E(s_P^2) = \sigma^2$$
$$\frac{n-I}{\sigma^2} s_P^2 \sim \chi^2_{n-I}$$

- A $1-\alpha$ confidence interval for σ^2 is of the form:

$$\left(\frac{n-I}{\chi^2_{n-I,\ 1-\alpha/2}} s_P^2,\ \frac{n-I}{\chi^2_{n-I,\ \alpha/2}} s_P^2 \right)$$

Example: concrete

Summary Statistics for the Concrete Strength Study

i, Concrete Formula	n_i, Sample Size	$\bar{y}_i$, Sample Mean (psi)	s_i, Sample Standard Deviation (psi)
1	$n_1 = 3$	$\bar{y}_1 = 5{,}635.3$	$s_1 = 965.6$
2	$n_2 = 3$	$\bar{y}_2 = 5{,}753.3$	$s_2 = 432.3$
3	$n_3 = 3$	$\bar{y}_3 = 4{,}527.3$	$s_3 = 509.9$
4	$n_4 = 3$	$\bar{y}_4 = 3{,}442.3$	$s_4 = 356.4$
5	$n_5 = 3$	$\bar{y}_5 = 2{,}923.7$	$s_5 = 852.9$
6	$n_6 = 3$	$\bar{y}_6 = 3{,}324.7$	$s_6 = 353.5$
7	$n_7 = 3$	$\bar{y}_7 = 1{,}551.3$	$s_7 = 505.5$
8	$n_8 = 3$	$\bar{y}_8 = 2{,}390.7$	$s_8 = 302.5$

Example: concrete

$$
\begin{aligned}
s_P^2 &= \frac{(3-1)(965.6)^2 + (3-1)(432.3)^2 + \cdots + (3-1)(302.5)^2}{(3-1) + \cdots + (3-1)} \\
&= 2\,\frac{965.6^2 + 432.3^2 + \cdots + 302.5^2}{16} \\
&= 338213 \text{ psi}^2 \\
s_P &= \sqrt{338213} = 581.6 psi
\end{aligned}
$$

Example: concrete

- $n = 24$, $I = 8$, $n - I = 16$.
- $\chi^2_{16,\ 0.95} = 26.296$, $\chi^2_{16,\ 0.05} = 7.962$
- Hence, a 90% 2-sided confidence interval for σ^2 is:

$$\left(\frac{16 \cdot 581.6^2}{26.296}, \frac{16 \cdot 581.6^2}{7.962} \right)$$
$$= (205816,\ 679745.9)$$

and you can make a 90% confidence interval for σ by transforming the endpoints of the confidence interval for σ^2:

$$(\sqrt{205816},\ \sqrt{679745.9}) = (453.7,\ 824.5)$$

- We're 90% confident that the true overall standard deviation of compressive strength of the concrete within factor levels is between 453.7 psi and 824.5 psi.

Outline

The one-way ANOVA model

Residuals and fitted values

Variance estimation

Standardized residuals

Inference

Standardized residuals

- Just as before, even though $\varepsilon_{ij} \sim$ iid $\mathrm{N}(0, \sigma^2)$, the e_{ij}'s don't have constant variance.
- The **standardized residuals** for the one-way ANOVA model are of the form:

$$e^*_{ij} = \frac{e_{ij}}{s_P\sqrt{\frac{n_i - 1}{n_i}}}$$

which are about $N(0, 1)$ on average.

Outline

Inference for the one-way ANOVA model

1. $H_0 : \mu_1 = \mu_2 = \cdots = \mu_I$, H_a : not all the μ_i's are equal.
2. α is some sensible value.
3. The test statistic is:

$$K = \frac{MSR}{MSE} = \frac{SSR/(I-1)}{SSE/(n-I)}$$

- Here,
  - n is the number of observations.
  - I is the number of levels of the covariate.
  - $SSR = \sum_{ij}(\widehat{y}_{ij} - \overline{y}_{..})^2 = \sum_{ij}(\overline{y}_{i.} - \overline{y}_{..})^2$
  - $SSE = \sum_{ij}(y_{ij} - \widehat{y}_{ij})^2 = \sum_{ij}(y_{ij} - \overline{y}_{i.})^2$
  - $SST = \sum_{ij}(y_{ij} - \overline{y}_{..})^2$
  - $\overline{y}_{..} = \frac{1}{n}\sum_{ij} y_{ij}$
- Assume H_0 is true, the model is valid, and the ε_{ij}'s are iid $N(0, \sigma^2)$
- Then, $K \sim F_{I-1,\ n-I}$.
- Reject H_0 if $K > F_{I-1,\ n-I,\ 1-\alpha}$

Inference for the one-way ANOVA model

4. The moment of truth: construct the ANOVA table:

Source	SS	df	MS	F
Covariate	SSR	$I-1$	$SSR/(I-1)$	MSR/MSE
Error	SSE	$n-I$	$SSE/(n-I)$	

Example: concrete

1. $H_0 : \mu_1 = \mu_2 = \cdots = \mu_8$, H_a : not all the μ_i's are equal.
2. $\alpha = 0.05$
3. The test statistic is:

$$K = \frac{MSR}{MSE} = \frac{SSR/(I-1)}{SSE/(n-I)} = \frac{SSR/7}{SSE/16}$$

- Assume H_0 is true, the model is valid, and the ε_{ij}'s are iid $N(0, \sigma^2)$
- Then, $K \sim F_{I-1,\ n-I}$.
- Reject H_0 if $K > F_{I-1,\ n-I,\ 1-\alpha} = F_{7,16,0.95} = 2.66$

Example: concrete

4. The moment of truth: we start by calculating SST, s_P^2, and SSE:

$$\begin{aligned}&(5{,}800-3{,}693.6)^2+(4{,}598-3{,}693.6)^2+(6{,}508-3{,}693.6)^2\\&+\cdots+(2{,}631-3{,}693.6)^2+(2{,}490-3{,}693.6)^2\\=\ &52{,}772{,}190\ (\text{psi})^2\end{aligned}$$

$$s_P^2 = 338{,}213.1\ (\text{psi})^2 \text{ and } n-r=16, \text{ so}$$

$$SSE=(n-r)s_P^2=5{,}411{,}410\ (\text{psi})^2$$

Lastly, we calculate SSR:

$$\sum_{i=1}^{r} n_i(\bar{y}_i-\bar{y})^2 = 47{,}360{,}780$$

Example: concrete

ANOVA Table (for testing $H_0: \mu_1 = \mu_2 = \cdots = \mu_8$)

Source	*SS*	*df*	*MS*	*F*
Treatments	47,360,780	7	6,765,826	20.0
Error	5,411,410	16	338,213	
Total	52,772,190	23		

5. With $K = 20.0 > 2.66$, we reject H_0 and conclude H_a.
6. There is enough evidence to conclude that the compressive strength of the concrete varies with formula.

Example: railroad rails

- The following data are taken from the paper Zero- Force Travel-Time Parameters for Ultrasonic Head-Waves in Railroad Rail by Bray and Leon- Salamanca (Materials Evaluation, 1985).
- Given are measurements in nanoseconds of the travel time (in excess of 36.1 μs) of a certain type of mechanical wave induced by mechanical stress in railroad rails.

Rail	Travel Time (nanoseconds above 36.1 μs)
1	55, 53, 54
2	26, 37, 32
3	78, 91, 85
4	92, 100, 96
5	49, 51, 50
6	80, 85, 83

Example: railroad rails

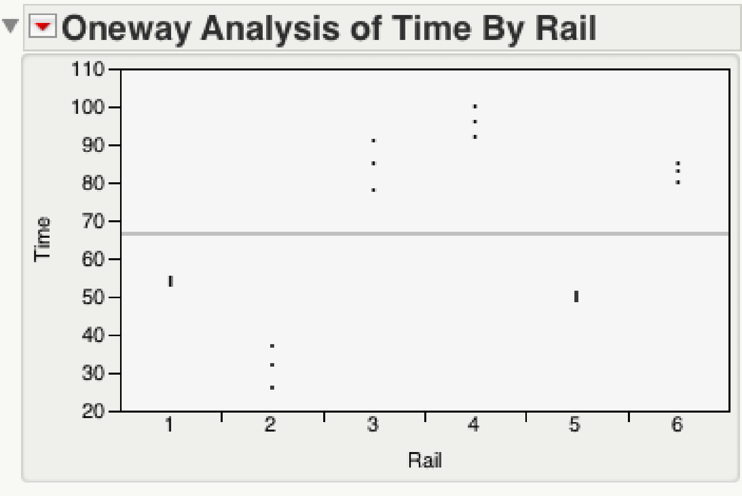
- We apply the model:

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

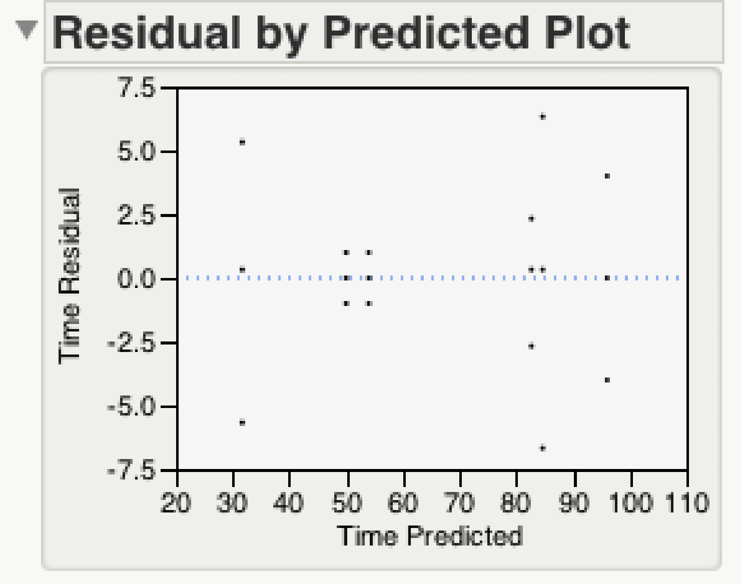
  where:

  - Y_{ij} is the observed travel time (ns) of the wave in excess of 26.1 μs for Rail i wave j.
  - μ_i is the true mean travel time (ns) in excess of 26.1 μs of waves through Rail i.

Example: railroad rails



Example: railroad rails



Example: railroad rails

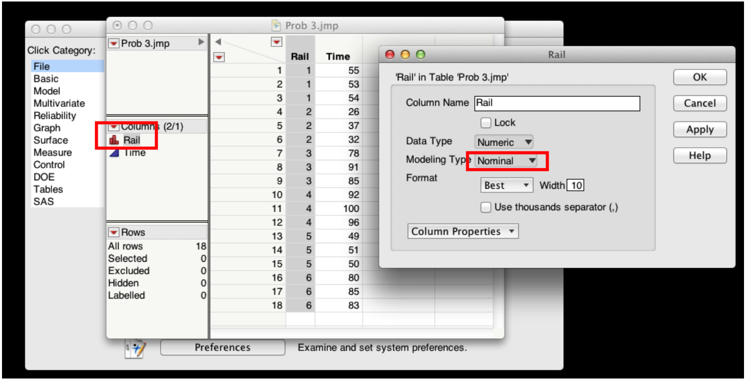
1. $H_0 : \mu_1 = \mu_2 = \cdots = \mu_6$, H_a : not all the μ_i's are equal.
2. $\alpha = 0.05$
3. The test statistic is:

$$K = \frac{MSR}{MSE} = \frac{SSR/(I-1)}{SSE/(n-I)} = \frac{SSR/(6-1)}{SSE/(18-6)} = \frac{SSR/5}{SSE/12}$$

- Assume H_0 is true, the model is valid, and the ε_{ij}'s are iid $N(0, \sigma^2)$
- Then, $K \sim F_{I-1,\ n-I}$.
- Reject H_0 if $K > F_{I-1,\ n-I,\ 1-\alpha} = F_{5,12,0.95} = 3.11$

Example: railroad rails

4. The moment of truth: load the data into JMP and fit travel time on rail, and *make sure the rail variable is a factor*.



Example: railroad rails

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	9310.5000	1862.10	115.1814
Error	12	194.0000	16.17	**Prob > F**
C. Total	17	9504.5000		<.0001*

5. With $K = 115.18 > 3.11$, we reject H_0 and conclude H_a.
6. There is enough evidence to conclude that the true mean excess travel time of waves along the rails depends on the rail.