3Y03 - PROBABILITY AND STATISTICS FOR ENGINEERING

WS19 Lecture 33

Yesterday ANOVA for Linear Regression
↳ Analysis of Variance Approach

Work with $SS_T = SS_E + SS_R$ ← Variablity accounted for by regression line

SS_T: Total variability in Y

SS_E: Variability unexplained by regression line (explained by error)

e.g. • F-test for $H_0: \beta_1 = 0$; $H_1: \beta_1 \neq 0$; test stat. $F_0 = \frac{MS_R}{MS_E} := \frac{SS_R/\boxed{1}}{SS_E/\boxed{(n-2)}}$

numerator & denominator degrees of freedom

reject H_0 at level α if $f_0 > f_{\alpha, 1, n-2}$

• Coefficient of determination $R^2 = \frac{SS_R}{SS_T}$ } proportion of total variability of Y explained by regression line

---

13-2 Single Factor Experiments & (One-way) ANOVA

↓

Testing a response/factor at different "levels" (called treatments)

e.g. Treatment	Response
Different drug dosages (low, med., high)	Improvement in illness
" concrete mixing techniques	Compressive strength
" study techniques	

(cramming v. "directed" studying) Test score

" BMI levels (in groups) Baseball batting performance

" LCD screen luminescence levels screen performance

" caffeine levels consumed (high, very high, outrageously high) Ability to stay awake in class

---

\# Treatments = a

At ith treatment ($i = 1, \ldots, a$) get a random sample $\{Y_{i1}, \ldots, Y_{in_i}\}$

Y_{ij} – jth observation at ith treatment

n_i observations at ith treatment

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

μ_i: mean of ith treatment

ε_{ij}: random error dependent on treatment $\sim N(0, \sigma_i^2)$

$\mu_i = \mu + \tau_i$

μ: overall mean

τ_i: "ith treatment effect"

We assume a <u>completely randomized experimental design</u> i.e. all observations Y_{ij} made in random order & uniform environments

This lets us assume errors ε_{ij} have same underlying variance σ^2

FOR NOW assume all treatments have <u>same</u> # of observations $n_i = n$ for every $i = 1, \ldots, a$.
(only for simplicity — we'll fix this later).

Model simplifies to: $Y_{ij} = \overbrace{\mu + \tau_i}^{\mu_i} + \varepsilon_{ij}$

where $\varepsilon_{ij} \sim N(0, \sigma^2)$

$i = 1, \ldots, a$
$j = 1, \ldots, n$

$= \mu_i + \varepsilon_{ij}$

<u>Goal</u> Compare means $\mu_1, \ldots, \mu_a$
(to see effect of each treatment on factor)

<u>Notice</u> When $a = 2$, this is comparison of 2 means of Normal pop., variance unknown (10-2)

Recall $H_0 : \mu_1 = \mu_2$ $\rightsquigarrow$ t-test

(from 10-2) $H_1 : \mu_1 \neq \mu_2$ (say) (whether $\sigma_1^2 = \sigma_2^2$ or $\sigma_1^2 \neq \sigma_2^2$)

A (bad) idea

pair up: $H_0 : \mu_i = \mu_k$ for each $i \neq k$ pair

$H_1 : \mu_i \neq \mu_k$

$\rightarrow$ There are $\binom{a}{2}$ many such tests

$\rightarrow$ At significance level α for each test

$\alpha =$ P(Type I error in a single test)

So EER = Experiment-wise Error Rate

= P(at least one Type I Error across all $\binom{a}{2}$ tests)

= 1 − P(no Type I Errors)

$= 1 - (1-\alpha)^{\binom{a}{2}}$

$\binom{6}{2} \to 15$

If $\alpha = 0.05$ and $a = 6$, $EER = 1 - (0.95)^{15} = \underline{\underline{0.54}}$

OUCH

ANOVA (an f-test) will compare all means simultaneously (i.e. $\underbrace{\mu_1 = \mu_2 = \ldots = \mu_a}_{H_0}$).

Model: $Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$

We estimate Y_{ij} with estimators $\hat{\mu}, \hat{\tau}_i, \hat{\varepsilon}_{ij}$ for $\mu, \tau_i, \varepsilon_{ij}$.

<u>Notation</u> $\bullet$ = "sum over whichever subscript usually goes here"

$$Y_{i\bullet} = \sum_{j=1}^{n} Y_{ij} \quad = \text{ total under } i\text{th treatment}$$

$$\bar{Y}_{i\bullet} = \frac{1}{n} Y_{i\bullet} \quad = \text{ sample mean under } i\text{th treatment}$$

$$Y_{\bullet\bullet} = \sum_{i=1}^{a} \sum_{j=1}^{n} Y_{ij} = \sum_{i=1}^{a} Y_{i\bullet} \quad = \text{ overall total}$$

$$N = na \quad = \text{ total \# observations across all treatments}$$

$$\bar{Y}_{\cdot\cdot} = \frac{1}{N} Y_{\cdot\cdot} = \text{sample overall mean}$$

[$$S_i^2 = \frac{1}{n-1} \left(\sum_{j=1}^{n} (Y_{ij} - \bar{Y}_{i\cdot} \right)^2 = \text{sample variance under ith treatment}$$
etc.]

Then $\hat{\mu} = \bar{Y}_{\cdot\cdot}$

$\hat{\tau}_i = \bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}$ — difference between ith treatment mean and overall mean

$\hat{\varepsilon}_{ij} = e_{ij} = Y_{ij} - \bar{Y}_{i\cdot}$ (error <u>within</u> ith treatment)