Date:

→ When we run to many test type I error increases

Type 1 error

rejecting the null hypothesis when its actually true

→ The scenario of testing many pairs of groups is called multiple comparisons

→ The Bonferroni correction suggests that a more strigent significance lvl is more appropriate for these tests

$$\alpha^2 = \alpha / k$$

where k is the no. of comparison being considered

→ If there are k groups, then usually all possible pairs are compared and

$$k = \frac{k(k-1)}{2}$$

Date:

Difference in two means: after ANOVA

$$SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \approx \sqrt{\frac{MSE}{n_1} + \frac{MSE}{n_2}}$$

$$T_{df_E} = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{MSE}{n_1} + \frac{MSE}{n_2}}}$$

Residuals

Diff b/w observed (y_i) and predicted ($\hat{y}_i$)

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

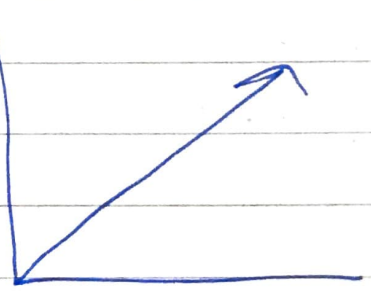
Quantifying the relationship

→ Correlation describes the strength of the linear association b/w two variable

→ It takes values b/w −1 (perfectly neg) and +1 (perfectly pos)

→ A value of 0 indicates no linear association

Fitting a line by least Sq regression



→ Minimize the sum of mag (abs values) of residuals

$$|e_1| + |e_2| + \dots = |e_i|$$

Date:

→ Minimize the sum of sq of residuals least squares

$$e_1^2 + e_2^2 + \cdots + e_n^2$$

The Least Sq line

$$\hat{y} = \beta_0 + \beta_1 x$$

$\hat{y}$ → pred y

β_0 → intercept

β_1 → slope

x → explanatory variable

Intercept

- Parameter : β_0
- Point estimate : b_0

Slope

- Parameter : β_1
- Point estimate : b_1

Conditions

1) Linearity

The relation b/w explanatory and the response variable should be linear

2) Nearly Normal residuals

The residuals should be nearly normal.

3) Constant Variability

The variability of points around

Date:

the least sq line should be roughly const. This implies that the variability of residuals around the 0 line should be roughly const. also called homoseedasticity

Slope

The slope of a regression can be calculated as

$$b_1 = \frac{S_y}{S_x} R$$

Intercept

The intercept is where the regression line intersects the y-axis.

$$b_0 = \bar{y} - b_1 \bar{x}$$

→ when $x = 0$, y is expected to equal the intercept

→ For each unit in x, y is expected to increase / decrease on average by the slope

Date:

Prediction

→ using linear model to pred the valve of response variable for a given valve of the explanatory variables is called prediction, simply plugging in the x valve in the linear model eqn

Extrapolation

→ Applying a model estimate to valves outside of the realm of the original data is called extrapolation

R^2

The strength of the fit of a linear model is commonly evaluted using R^2

→ R^2 is calculated by the sq of correlation coefficient

least sq regression Line Formula

$$\hat{\beta} = \frac{SS_{xy}}{SS_{xx}}$$

$$\hat{\beta_0} = \bar{y} - \hat{\beta_1} \bar{x}$$

$$SS_{xx} = \sum x^2 - \frac{1}{n} \left(\sum x \right)^2$$

Date:

$$SS_{xy} = \sum xy - \frac{1}{n}\left(\sum x\right)\left(\sum y\right)$$

$$\bar{y} = \hat{\beta}_1 x + \hat{\beta}_0$$

Example

verify that it fits the data better than the line $\hat{y} = \frac{1}{2}x - 1$

x	y
2	0
2	1
6	2
8	3
10	3

	x	y	x^2	xy	y^2
	2	0	4	0	0
	2	1	4	2	1
	6	2	36	12	4
	8	3	64	24	9
	10	3	100	30	9
Total	28	9	208	68	23

$$SS_{xx} = \sum x^2 - \frac{1}{n}\left(\sum x\right)^2 = 208 - \frac{1}{5}(28)^2$$

$$= 51.2$$

Date:

$SS_{xy} = \sum xy - \frac{1}{n}(\sum x)(\sum y)$

$= 68 - \frac{1}{5}(28)(9)$

$= 17.6$

$SS_{yy} = \sum y^2 - \frac{1}{n}(\sum y)^2$

$= 23 - \frac{1}{5}(9)^2$

$= 23 - \frac{1}{5}(81)$

$= \frac{115 - 81}{5} = 6.8$

$\bar{x} = \sum x / n$

$= 28/5$

$= 5.6$

$\bar{y} = \sum y / n$

$= 9/5$

$= 1.8$

$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{17.6}{51.2} = 0.34375$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$= 1.8 - 0.34375(5.6)$

$= -0.125$

$\hat{y} = 0.34375x - 0.125$

$r = \frac{SS_{xy}}{\sqrt{SS_{xx} \cdot SS_{yy}}} = \frac{17.6}{\sqrt{51.2 \times 6.8}} = \frac{17.6}{348.16} = 0.9432$

Date:

SSE = $SS_{yy} - \hat{\beta}_1 . SS_{xy}$

= 6.8 − 0.34375 × 17.6

= 6.8 − 6.05

= 0.75

Slope = $\sqrt{SS_{yy}/n} * R$ / $\sqrt{SS_{xx}/n-1}$

= $\sqrt{\frac{6.8}{5-1}} * 0.9432$ / $\sqrt{\frac{51.2}{5-1}}$

= $\sqrt{\frac{6.8}{4}} * 0.9432$ / $\sqrt{\frac{51.2}{4}}$

= $\frac{1.30}{3.58} * 0.9432$

= 0.36 * 0.9432

= 0.34

Date:

Types of Outliers

→ Outliers are points that lie away from the cloud of points

→ Outliers that lie horizontally away from the center of the cloud are called high leverage points

→ High leverage points that actually influence the slope of the regression line are called influential points

ANOVA Example

u_1	u_2	u_3
Group 1	Group 2	Group 3
2	3	12
8	4	15
6	11	2
10	13	5
20	17	8
$\sum \bar{u_1} = 9.2$	$\sum \bar{u_2} = 9.6$	$\sum \bar{u_3} = 8.4$

$$\bar{\bar{X}} = \frac{\sum \bar{u_1} + \sum \bar{u_2} + \sum \bar{u_3}}{3} = \frac{9.2 + 9.6 + 8.4}{3} = 9.07$$

Date:

$df_g = k-1$
$= 3-1$
$= 2$

$df_t = n-1$
$= 15-1$
$= 14$

$df_E = 14-2$
$= 12$

$SSG = [(5 \times (9.2-9.07)^2) + (5 \times (9.6-9.07)^2 + (5 \times (8.4-9.07)^2)]$

$= [5 \times (0.13)^2 + 5 \times (0.53) + 5 \times (-0.67)^2]$

$= (5 \times 0.0169) + (5 \times (0.2809)) + (5 \times 0.4489)$

$= 0.0845 + 1.4045 + 2.2445$

$= 3.7335$

$SST = (2-9.07)^2 + (3-9.07)^2 + (12-9.07)^2 + (8-9.07)^2$
$+ (4-9.07)^2 + (15-9.07)^2 + (6-9.07)^2 +$
$(11-9.07)^2 + (2-9.07)^2 + (10-9.07)^2 + (13-9.07)^2$
$+ (5-(9.07)^2 + (20-9.07)^2 + (17-9.07)^2$
$+ (8-9.07)^2$

Date:

$= 49.9849 + 36.8449 + 8.5849 + 1.1449$
$+ 25.7049 + 35.1649 + 9.4249 + 3.7249 +$
$49.9849 + 0.8649 + 15.4449 + 16.5649$
$+ 119.4649 + 62.8849 + 1.1449$

$= 372.9078 + 62.8849 + 1.1449$
$= 436.9076$

$SSE = 436.9076 - 3.7335$
$= 433.1741$

$MSG = SSG / df_G$
$= 3.7335 / 2$
$= 1.86675$

$MSE = SSE / df_E$
$= 433.1741 / 12$
$= 36.10$

$F = \frac{MSG}{MSE} = \frac{1.86675}{36.10} = 0.05$