Suggested Solution : TMA4255 Applied Statistics
Spring 2024

1a)

H_0: $\mu_1 = \mu_2$, H_1: $\mu_1 > \mu_2$.

Here, $\bar{x} - \bar{y} \sim N(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2)$.

We use the assumption of the same variance in the two groups, estimated by s^2_{pool}. Then

$$T = \frac{\bar{x} - \bar{y}}{s_{\text{pool}}\sqrt{1/7 + 1/5}} \sim t_{10}$$

We reject H_0 if the observed $T > t_{10,0.95}$.

Here, we get $T = \frac{19.64 - 17.62}{2.23\sqrt{(1/7+1/5)}} = 1.55$. The upper 0.05 percentile is $t_{10,0.95} = 1.81$. We do not reject H_0. The mean fluoride level is not significantly larger for cows grazing near the industrial area.

1b)

H_0: $\sigma_1^2 = \sigma_2^2$, H_1: $\sigma_1^2 \neq \sigma_2^2$.

Under H_0, we have that

$$F = \frac{s_1^2}{s_2^2} \sim f_{6,4}$$

We reject H_0 if the observed $F > f_{6,4,0.975} = 9.20$ or $F < f_{6,4,0.025} = 0.16$.

Here, we get $F = 5.17/4.65 = 1.11$. We do not reject H_0.

The approach of using s_{pool} as a common variance in a) is then valid.

1c)

The Wilcoxon test forms the rank sum for each sample. Denote ranks by $r_1, \ldots, r_{12}$ and group label by $g_1, \ldots, g_{12}$, then

$$W_1 = \sum_{j=1}^{12} r_j I(g_j = 1), W_2 = \sum_{j=1}^{12} r_j I(g_j = 2)$$

The hypothesis test is H_0: $\mu_1 = \mu_2$, H_1: $\mu_1 > \mu_2$. We only assume that the two sample distributions are symmetric. Then, under hypothesis H_0, the rank-sum W_2 should not be too small.

Polluted	21.3	18.7	23.0	17.1	16.8	20.9	19.7
Unpolluted	14.2	18.3	17.2	18.4	20.0		
Polluted ranks	11	7	12	3	2	10	8
Unpolluted ranks	1	5	4	6	9		

Table 1: Data and ranks.

Ranking the data (Table 1), we get $W_1 = 2+3+7+8+10+11+12 = 53$ and $W_2 = 25$. This means that $U_2 = 25 - \frac{5\cdot 6}{2} = 10$.

Using the normal approximation of this test statistic for the smaller group (U_2), we have (under H_0) that

$$U_2 = W_2 - \frac{n_2(n_2+1)}{2} \approx N(\frac{7\cdot 5}{2}, \frac{5\cdot 7\cdot 13}{12}) = N(17.5, 6.16^2),$$

Observed $Z = \frac{10-17.5}{6.16} = -1.22$, which is not smaller than the percentile $z_{0.05} = -1.645$. We do not reject H_0.

At significance level $\alpha = 0.05$, an exact test (Table) with 7 (larger group) and 5 (smaller group) has 6 as critical value for U_2 (the smallest sample). The p-value of 10 is 0.134. We do not reject H_0.

2a)

H_0: uniform distribution, H_1: not uniform distribution.

Under H_0, we have that the goodness of fit statistic is χ^2_{k-1}, where k is the number of classes (here 4).

The expected numbers in every bin, according to hypothesis H_0, are $e_i = 54/4 = 13.5$, $i = 1, 2, 3, 4$.

$$X = \sum_{i=1}^{4} \frac{(o_i - e_i)^2}{e_i} = (9 - 13.5)^2/13.5 + 3\cdot(15 - 13.5)^2/13.5 = 2$$

The critical limit is $\chi^2_{0.05,3} = 7.81$. The difference in class numbers can be attributed to random variation. We do not reject H_0.

2b)

H_0: independence, H_1: not independent

Let A be the outcome (bin) of Day 1 and B be the outcome (bin) of Day 2. Under H_0 about independence; $P(A\cap B) = P(A)P(B)$, for all A

and B in the sample space of bins. We can construct $P(A)$ and $P(B)$ from the row and column sums, and this forms the expected numbers which are $e_{A,B} = nP(A)P(B) = N_A N_B/n$. Here, grand total is $n = 53$, while column and row totals are N_A and N_B respectively.

Many of the expected numbers in the table entries are similar, with the 15 repeating several times as N_A or N_B. Table 2 shows both observed and expected values (parentheses).

	Day 1(1-8)	Day 1(9-16)	Day 1(17-24)	Day 1(25-32)	
Day 2(1-8)	0 (1.36)	0 (2.54)	5 (2.54)	4 (2.54)	9
Day 2(9-16)	4 (2.26)	0 (4.25)	0 (4.25)	11 (4.25)	15
Day 2(17-24)	4 (2.11)	10 (3.96)	0 (3.96)	0 (3.96)	14
Day 2(25-32)	0 (2.26)	5 (4.25)	10 (4.25)	0 (4.25)	15
	8	15	15	15	53

Table 2: Data with observed and expected (parantheses) values.

The test statistic here is then

$$X = \sum_{i=1}^{16} \frac{(o_i - e_i)^2}{e_i}$$

Under H_0, the it is approximately distributed as $\chi^2_{(k-1)(k-1)} = \chi^2_9$.

Here, with the many 0s in observed values, we can quickly see that the test statistic becomes at least $3 \cdot 4.25 + 2 \cdot 3.96$, which is already bigger than the critical limit is $\chi^2_{0.05,9} = 16.9$.

The full sum is

$$X = \sum_{i=1}^{16} \frac{(o_i - e_i)^2}{e_i} = (0 - 1.36)^2/1.36 + \ldots + (0 - 4.25)^2/4.25 = 60.9$$

We reject H_0.

Note that the approximation requires expected numbers in each bin to be reasonably large (a rule-of-thumb is $np > 5$). Here the numbers are somewhat small. Some categories could potentially be merged. Nevertheless, the rejection appears to be very clear.

3a)

$$y_i = \beta_0 + x_{1,i}\beta_1 + x_{2,i}\beta_2 + x_{3,i}\beta_3 + \epsilon_i, \quad i = 1, \ldots, 20$$

The assumption is that the noise terms $\epsilon_i \sim N(0, \sigma^2)$ and independent.

For any effect,

$$T_\beta = \frac{\hat{\beta}}{s_\beta}.$$

This means that the two first missing numbers are: $s_{\beta_1} = 15.306/4.647 = 3.29$, $T_{\beta_2} = 0.0485/0.0784 = 0.62$.

For the p-value, we consider how likely the observed T value (or something more extreme) is. The two-sided test means that p-value$= 2P(t_{16} < -1.631) = 2P(t_{16} > 1.631) \approx 2 \cdot 0.06 = 0.12$. (We see that value of 1.631 is between the quantiles of 0.05 and 0.075 for a t-distribution with 20-4=16 degrees of freedom.)

3b)

The growth of a fish per year is β_1. With $20 - 2 = 18$ degrees of freedom in the t-distribution, we have

$$P(-t_{18,0.05} < \frac{\beta_1 - \hat{\beta}_1}{s_{\beta_1}} < t_{18,0.05}) = 0.90$$

Moving elements around, we have

$$P(\hat{\beta}_1 - s_{\beta_1} t_{18,0.05} < \beta_1 < \hat{\beta}_1 + s_{\beta_1} t_{18,0.05}) = 0.90$$

Here, $t_{18,0.05} = 1.734$, and then $\hat{\beta}_1 \pm s_{\beta_1} t_{18,0.05} = 16.2 \pm 2.807 \cdot 1.734 = (11.3, 21.1)$ mm.

The fitted model of different ages is $\hat{y}(x_1) = 48.8 + 16.2x_1$, which gives $\hat{y}(2) = 81.2$, $\hat{y}(2) = 113.6$ and $\hat{y}(7) = 162.2$. Figure 1 shows the fitted line and the data. The model is overpredicting the data for small ages and then underpredicting for intermediate age before overpredicting again for high age. Residuals would show a clear pattern of negative, positive and then negative.

3c)

R^2 is defined by the explainability of the regression model fit as part of the total variability. Generally, the total variability in the data (y_i) from the mean $(\bar{y})$, can be split in two parts, one explained by the regression model $(\hat{y}_i)$ and one part with the residual variation;

$$SST = SSR + SSE, SST = \sum_{i=1}^{20} (y_i - \bar{y})^2, SSR = \sum_{i=1}^{20} (\hat{y}_i - \bar{y})^2, SSE = \sum_{i=1}^{20} (y_i - \hat{y}_i)^2$$

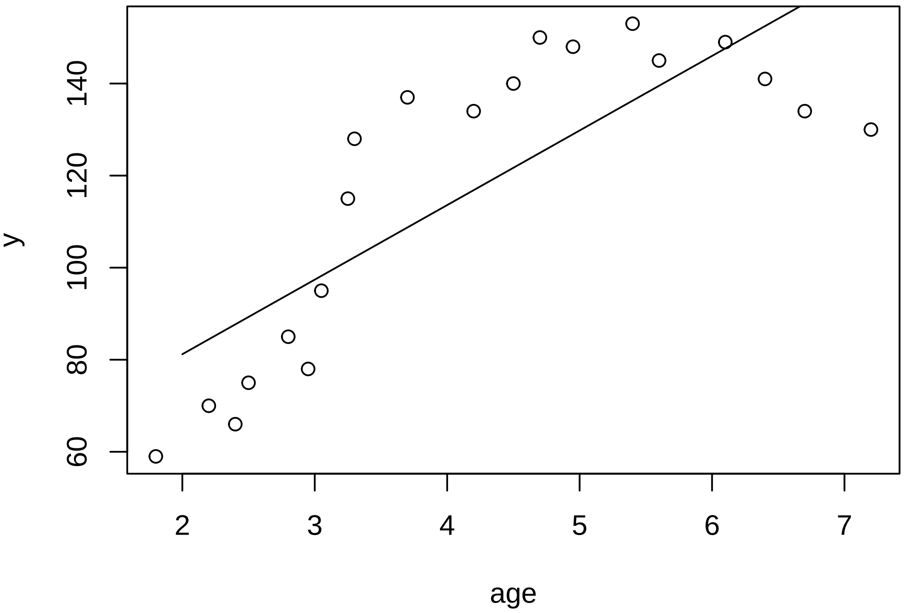

Figure 1: Sketch of fitted line and data. Residuals would show a pattern of negative, positive and then negative.

$R^2 = SSR/SST = 1 - SSE/SST$. This R^2 always increases when one adds more covariates in the model. The adjusted $R^2_{\text{adj}} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$ is a variant of R^2 that compensates for the number of covariates (k) in the model. This does not always increase when more covariates go into the model. Here, R^2 and R^2_{adj} are both much larger for model c) than for a) and b).

The largest length value is where the derivative of the fitted quadratic curve is 0. We denote this by x_1^* We have

$$\frac{d\hat{y}}{dx} = \hat{\beta}_1 + 2\hat{\beta}_2 x_1^* = 0 \leftrightarrow x_1^* = -\frac{\hat{\beta}_1}{2\hat{\beta}_2}$$

In this case we get $x_1^* = 83.715/(2(-7.584)) = 5.52$.

The variance of this is approximated by linearization (derivatives in a Taylor expansion of the function $x_1^*(\hat{\beta}_1, \hat{\beta}_2)$.

$$\mathrm{Var}(x_1^*) = [\frac{dx_1^*}{d\hat{\beta}_1}]^2\mathrm{Var}(\hat{\beta}_1) + [\frac{dx_1^*}{d\hat{\beta}_2}]^2\mathrm{Var}(\hat{\beta}_2) + 2[\frac{dx_1^*}{d\hat{\beta}_1}][\frac{dx_1^*}{d\hat{\beta}_2}]\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2)$$

$$= [1/(2 \cdot 7.58)]^2 \cdot 8.47^2 + [(83.72/(2 \cdot 7.58^2)]^2 \cdot 0.94^2$$

$$+2[1/(2 \cdot 7.58)][(83.72/(2 \cdot 7.58^2)](-7.86) = 0.16^2$$

4a)

$$\hat{A} = \frac{92.2 + 93.7 + 92.7 + 92.9}{4} - \frac{88.1 + 90.1 + 92.1 + 93.1}{4} = 2.025$$

From the D column, we recognize that the generator is $D = ABC$. This means that several effects are confounded: $A = BCD$, $B = ACD$, $C = ABD$, $AB = CD$, $AC = BD$, $AD = BC$.

4b)

We can specify the variance from the three linear combinations giving the interactions, AB, AC, AD. (The other 5 linear combinations of data give the mean and four main effects.)

$$s_{\text{Eff}}^2 = \frac{\hat{AB}^2 + \hat{AC}^2 + \hat{AD}^2}{3}$$

$$= (0.325^2 + 1.825^2 + 0.575^2)/3 = 1.12^2$$

This means that $T_A = \hat{A}/s_{\text{Eff}} = 1.81$, while $t_{0.025,3} = 3.18$. The effect of factor A is not significantly large.