Suggested Solution : TMA4255 Applied Statistics Spring 2021

1a)

H_0: $\mu = 3.5$, H_1: $\mu < 3.5$.

Data are assumed to be independent and identically distributed $X_i \sim N(\mu, \sigma^2)$, $i = 1, \ldots, 11$.

We estimate σ^2 by $s^2 = \frac{1}{11-1}\sum_{i=1}^{11}(x_i - \bar{x})^2 = 0.518^2$. We have $\bar{x} = 3.4$.

$$t = \frac{\bar{x} - 3.5}{s/\sqrt{11}} = \frac{-0.1}{0.156} = -0.64.$$

Here, $t_{10,\alpha} = -1.81$, which is much smaller than t, so we do not reject H_0. The p-value is $P(t_{10} < -0.64) = 0.27$.

1b)

The sign test test if the median of the data is 3.5. There are no parametric assumptions and none about symmetry of the distribution.

$Y = \sum_{i=1}^{11} I(x_i < 3.5) = 8$.

The binomial distribution, under H_0, says that the number of samples below 3.5 should be binomial distributed with parameters $p = 1/2$ and $n = 11$ trials.

$$\text{p-value} = P(Y \geq 8) = \sum_{k=8}^{11} \frac{11!}{(11-k)!k!}\frac{1}{2^{11}} = 0.113$$

The Wilcoxon sign-rank test has no assumptions about parametric distribution, but it assumes symmetry around the mean/median.

To do this test one must sort the distances in samples away from the hypothesis mean 3.5 The ranks of the samples above 3.5 are : 4, 8 and 10.5 with a sum of $W_+ = 22.5$. The rank sum of samples below are then $W_- = 66 - 22.5 = 43.5$.

The normal assumption for rank-sums says that $W_+ \sim N(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24})$, under H_0. If the observed value is significantly small under this distribution, we reject the hypothesis of the same mean in the new brand.

Here, with $n = 11$, $Z = \frac{22.5-33}{\sqrt{126.5}} = -0.93$. We have p-value $= P(Z < -0.93) = 0.18$. This is not very small, and we cannot reject H_0. In fact the p-value is larger than for the sign-test because we have some large value (4.1 and 4.4) in the data.

2a)

The hypotheses are: $H_0 : p_1 = p_2$, $H_1 : p_1 < p_2$.

Using normal approximation;

$$Z = \hat{p}_1 - \hat{p}_2 \sim N(p_1 - p_2, \frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2})$$

Assuming $p_1 - p_2 = 0$ in the mean, and using a pooled estimate of p in the variance, we have

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}}$$

We reject H_0 if $Z < z_\alpha = -1.64$.

From the data we have $\hat{p}_1 = 3/10 = 0.3$ and $\hat{p}_2 = 6/11 = 0.545$, and $\hat{p}_1 - \hat{p}_2 = -0.245$. $\hat{p} = (3+6)/(10+11) = 0.43$.

$Z = \frac{-0.245}{\sqrt{0.0468}} = -1.13$. This means we do not reject H_0.

If we do not pool the variance estimate we get: $v = \sqrt{0.3 \cdot 0.7/n_1 + 0.545 \cdot 0.455/n_2} = 0.208$, and overall a similar conclusion: $Z = \frac{-0.245}{v} = -1.18$.

2b)

Power is the probability of rejecting a hypothesis H_0 when it is not true: $P(\text{reject}H_0|H_1) = P(Z < z_\alpha | p_1 < p_2)$. The power will depend on the difference $p_1 - p_2 < 0$. If this difference is very close to 0, the power is close to the significance level of the test. If this distance is much less than 0, the power goes towards 1. If the numbers of samples n_1 and n_2 increase, the power gets larger as it becomes easier to detect the difference.

For this case we have $p_1 - p_2 = -0.2$. Assuming the same variance estimate in the denominator, we define $v = \sqrt{0.3 \cdot 0.7/n_1 + 0.545 \cdot 0.455/n_2} = 0.208$

$$P(\text{reject}H_0|H_1) = P(Z < z_\alpha | p_1 < p_2) = P(Z < z_\alpha - \frac{-0.2}{v}) = P(Z < -0.681) = 0.248$$

.

To achieve a power of 0.5 we need

$$P(\text{reject}H_0|H_1) = P(Z < z_\alpha - \frac{-0.2}{v(n_1, n_2)}) = 0.5$$

. Then we must have $v(n_1, n_2) = 0.2/1.64 = 0.12$. That means that

$$v(n_1, n_2) = \sqrt{0.3 \cdot 0.7/n_1 + 0.545 \cdot 0.455/n_2} = 0.12.$$

Here, we can increase both n_1 and n_2, or one of them more than the other. By trial and error: $n_1 = 30$ and $n_2 = 33$ gives $v(n_1, n_2) = 0.1214$, $n_1 = 20$ and $n_2 = 60$ gives $v(n_1, n_2) = 0.1210$, $n_1 = 65$ and $n_2 = 22$ gives $v(n_1, n_2) = 0.1204$.

3a)

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$. The alternative H_1 is that at least one of the group means is different from the rest.

The variability within groups is $SSE = \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i\cdot})^2 = 4900 + 5100 + 4600 + 5000 = 19600$

The variability between groups is $SSA = n \sum_{i=1}^{k} (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$, and with $\bar{y}_{\cdot\cdot} = (128 + 136 + 157 + 141)/4 = 140.5$, $SSA = 4939$.

We reject H_0 is $F = (SSA/3)/(SSE/40) > f_{0.05,3,40} = 2.84$.

Here, $F = 3.36$ so H_0 is rejected. There is significant difference in the group means.

3b)

Define linear contrast $\mu_w = 0.5(\mu_1 + \mu_2) - 0.5(\mu_3 + \mu_4)$.

H_0: $\mu_w = 0$, H_1: $\mu_w \neq 0$.

We have $W = 0.5(\bar{Y}_{1\cdot} + \bar{Y}_{2\cdot}) - 0.5(\bar{Y}_{3\cdot} + \bar{Y}_{4\cdot})$ as an estimate for the contrast, and its distribution is defined by

$$W \sim N(\mu_w, \frac{1}{4}(4\sigma^2/n) = N(\mu_w, \sigma^2/n)$$

Since we do not know σ^2, this is estimated by $s^2 = SSE/40 = 22.1^2$. We then get a T distribution, and under H_0: $\mu_w = 0$, we have

$$T = \frac{W}{s/\sqrt{n}} \sim t_{40}$$

We reject H_0 if $|T| > t_{0.025,40} = -2.02$.

With the above data, we have $t = \frac{-17}{22.1}\sqrt{11} = -2.55$. We then reject H_0.

4a)

The method of least squares minimize the sum of square errors from the fitted model to the data:

$$SSE(\beta_0, \beta_1, \beta_2) = \sum_{i=1}^{14} (y_i - \beta_0 - \beta_1 x_i + \beta_2 x_i^2)^2$$

$$(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) = \text{argmin} SSE(\beta_0, \beta_1, \beta_2)$$

The actual minimum is found by setting the derivatives equal to 0 and solving the linear equations.

In Figure 1 it appears as if a quadratic function would systematically go under the curve for intermediate small and intermediate large covariate values. This means that the residual terms have a structure and might not be independent nor identically distributed.

4b)

The regression parameter estimates would in this case with 14 data and 3 parameters be distributed according to $\frac{\hat{\beta}_i - \beta_i}{s_i} \sim t_{11}$, where s_i is the standard deviation of regression parameter $i = 0, 1, 2$.

A confidence interval starts by

$$P(t_{0.025,11} < \frac{\hat{\beta}_1 - \beta_1}{s_1} < t_{0.975,11}) = 0.95,$$

and here $t_{0.025,11} = -2.20$, and $t_{0.975,11} = 2.20$. By moving around and getting only β_i in the middle, we have

$$P(\hat{\beta}_1 - s_1 2.20 < \beta_1 < \hat{\beta}_1 + s_1 2.20) = 0.95$$

The interval is then $(-15569, -7839)$.

The width of the confidence interval for β_2 is $23906 - 7581 = 16325$. Then $s_2 = 16325/(2 \cdot 2.20) = 3710$.

The t-value is $T = 15743/3701 = 4.25$.

4c)

Setting $SSE = SSE(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$, we have

$$S = \sqrt{SSE/(n-3)}$$

as an estimate of the error standard deviation σ.

$$R^2 = \frac{SSR}{SSE} = 1 - \frac{SSE}{SST},$$

where the sum of squares are $SST = \sum_{i=1}^{14}(y_i - \bar{y})^2$ and $SSR = \sum_{i=1}^{14}(\hat{y}_i - \bar{y})^2$ with $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2$ is the fitted regression model. The $R^2 \in (0, 100)$ %

should be large for models where the regression terms really matter for the prediction, compared with just using the mean value, without any covariates.

The prediction residual sum of squares (PRESS) is defined as the $\sum_{i=1}^{14}(y_i - \hat{y}_{i,-i})^2$, where the prediction $\hat{y}_{i,-i}$ is based on all data except (x_i, y_i). This PRESS should be small, otherwise one is likely overfitting to data and not necessarily predicting well for a hold-out test data set.

The PRESS is 40568 for the quadratic model while it is only 22823 for the piecewise linear model. Moreover, the S is smaller for the piecewise model, indicating less variability around the fitted line using all data as well. The R^2 is also larger for the piecewise linear model and it should be relatively comparable since they have the same number of parameters (see also adjusted and predictive R^2).

4d)

In a forward selection one add one covariate at a time into the model. Next, a check is conducted to see if another covariate should be added to the model, and this continues for step 3. The covariate that explains the most of the variability in the regression data is added, at each step. Here, x_i is added first, then $(x_i - c)I(x_i > c)$ and finally x_i^2 at step 3.

The coefficient for poro (x_i) changes because the model fitting is done with more parameters in the model and the estimates from the least squares method would then also depend on the non-zero value from the other parameters.

The forward selection goes through all three steps, and suggest a model with x_i, $(x_i - c)I(x_i > c)$ and x_i^2. Then again, at the last step, the poro variable is not very significant (when the other two are involved). A model with $(x_i - c)I(x_i > c)$ and x_i^2 could hence also be a good fit, even though this was not evaluated in the stepwise forward selection.