School of Mathematics and Statistics

Te Kura Mātai Tatauranga

STAT 292 **Assignment 4: Solutions and Feedback** **2020**

Extra comments and feedback are in italics.

1. **Comprehension Test [30 marks]**

   (a) **[24 marks] Two-way ANOVA**
   See the SAS output on pages 3 to 6 of the Assignment 3 Questions.

   Model equation:
   The model equation for the k^{th} response (score in comprehension test) with the first factor (Ethnicity) at level i and the second (Sex) at level j is

   $$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + E_{ijk}.$$

   The factors must be entered into the model in this order, as we are testing Sex after allowing for Ethnicity. Since this is an unbalanced data set, the order of entry matters. We must use Type I sums of squares – as given in the question.

   Assumptions, Diagnostic Graphs and Comments:
   We need to assume the errors are independent and normally distributed, with constant variance.
   Using the raw data and the diagnostic plots (p.4, Ass3 Questions), the plot of Residual versus Predicted Value shows a reasonably level band, supporting the assumption of constant variance. The Q-Q plot, Residual versus Quantile, shows an almost straight line, indicating normality of the residuals.
   We cannot check for independence, as we don't know how the experiment was run.

   ANOVA Table :

Source	df	Sum of Squares	Mean Square	F value	Pr > F
Ethnicity	2	3060.64	1530.32	48.79	<.0001
Sex	1	275.11	275.11	8.77	0.0070
Ethnicity×Sex	2	29.69	14.84	0.47	0.6289
Error	23	721.39	31.36		
Total	28	4086.83			

   Start with the interaction test. Hypotheses:
   H_0: There is no interaction (i.e. all $(\alpha\beta)_{ij} = 0$) versus
   H_A: There is interaction (at least one $(\alpha\beta)_{ij} \neq 0$).

Statistical conclusion:
The null hypothesis is not rejected at either the 5% or the 1% significance level; p-value $= 0.6289 > 0.05$.

Proceed to the main effects tests (since no significant interaction).
H_0: There is no main effect of Ethnicity (i.e. all $\alpha_i = 0$) versus
H_A: There is a main effect of Ethnicity (at least one $\alpha_i \neq 0$).
At either the 5% or the 1% level, H_0 is rejected, as $p < 0.0001 < 0.01$.
We have detected a main effect of Ethnicity – this confirms what was believed before the comprehension test was conducted.

H_0: There is no main effect of Sex (i.e. all $\beta_j = 0$) versus
H_A: There is a main effect of Sex (at least one $\beta_j \neq 0$).
This test is done after allowing for ethnic differences, as was required.
At either the 5% or the 1% level, H_0 is rejected, as $p = 0.0070 < 0.01$. We have detected a main effect of Sex after allowing for ethnic differences.

Interpretation and interaction plot:
Both ethnicity and sex have a main effect on the score in this comprehension test, but there is no interaction between these factors.
Two interaction plots were given (pages 5 and 6, Ass3 Questions). The lines are nearly parallel in both plots, illustrating there was no significant interaction.

On page 5, Ass3 Questions:
• The vertical separation indicates the significant main effect of Sex; this was significant after allowing for ethnicity. *Females are doing better than males.*

• The main effect of Ethnicity is shown by the lines not being horizontal. Recall that ethnicity was thought to be an important factor prior to the comprehension test. *Ethnic Group 1 appears to have the best comprehension.*

On page 6, Ass3 Questions:
• The vertical separation indicates the significant main effect of Ethnicity. Recall that ethnicity was thought to be an important factor prior to the comprehension test. *Ethnic Group 1 appears to have the best comprehension.*

• The non-horizontal lines illustrate the main effect of Sex; this was significant after allowing for ethnicity. *Females are doing better than males.*

The main result is the answer to the point of interest originally posed in the question: **after allowing for the different ethnic groups, there is a difference between the sexes in their comprehension of English**.

(b) **[6 marks] Discrepancy of Sex test:**

In the one-way ANOVA, with the fixed effects model, the hypotheses are:

H$_0$: There is no main effect of Sex (i.e. all $\alpha_i = 0$) versus

H$_A$: There is a main effect of Sex (at least one $\alpha_i \neq 0$).

At the 5% level, H$_0$ is not rejected since $p = 0.3292 > 0.05$. So there is no statistically significant evidence of a difference between the sexes.

The failure to detect a difference between the sexes occurs because of a failure to first allow for ethnic differences, which were (thought to be) important.

When Ethnicity was allowed for, as in part (a), the MSE (estimated residual unexplained variation in the data) was 31.36. However, when Ethnicity was omitted from the model, the MSE was 146.025, a substantial increase in the variability **not** being explained by the model. This made the test for Sex non-significant, through a loss of power to detect any difference.

2. **Invertebrates in Mussel Clumps [20 marks]**

See the SAS output on pages 8 to 10 of the Assignment 3 Questions.

(a) **[3 marks]** Yes, the scatterplot (p.8, Ass3 Questions) seems to show both linearity and constant variance.

(i) Linearity: There is no obvious curvature, so linearity is supported.

(ii) Constant variance: The data points are similar vertical distances from an imagined fitted line (slanting upwards through the middle of the data), with no obvious funnelling, supporting constant variance.

(b) **[17 marks] Simple Linear Regression**

Model equation:

$$Y = \beta_0 + \beta_1 x + E$$

where Y = number of Species, x = log(Area), β_0 = intercept and β_1 = slope of the regression line.

Hypotheses:

H$_0$: $\beta_1 = 0$ (i.e. zero slope) versus

H$_A$: $\beta_1 \neq 0$ (non-zero slope).

Assumptions and Comments on Diagnostic Graphs:

We assume the error term has a $N(0, \sigma^2)$ distribution, the errors are independent of each other and of x.

See page 10 of the Assignment 3 Questions for the diagnostic graphs.

The second diagnostic graph, RStudent versus Predicted Value, shows a fairly level band across the page, supporting the assumption of constant variance.

The Q-Q plot of Residual versus Quantile is a straight line (other than a couple of points), supporting the normality assumption.

The Cook's distances are well below 1, so no outliers or points of high leverage have been signalled.

We cannot check for independence – it depends on how the data were collected.

Note: if there are outliers or points of high leverage, the regression procedure is still valid, as no actual assumptions have been violated, but the results may not be very useful as they give too much weight to those special points.

Statistical conclusion:

The statistical analyses are on page 9 of the Assignment 3 Questions.

The t test statistic is 10.86, and the p-value is < 0.0001, so we reject H_0 and conclude that the slope **is** significantly different from zero.

Interpretation:

We note from the scatterplot that the slope is positive. There is a significant increase in the number of species as the area of the clump increases. The increase is linear if log(Area) is used as a predictor.

3. **Coarse Woody Debris in Lakes [30 marks]**

   (a) **[4 marks] Comments on the Scatterplots**

   The scatterplots are on page 12 of the Assignment 3 Questions.

   There is a negative association between CWD and L10CABIN, and between RIP.DENS and L10CABIN. The association between CWD and RIP.DENS is positive, which would be expected.

   (b) **[22 marks] Presentation of the three regression analyses, plus comments on the diagnostic graphs**

   See the SAS output on pages 13 to 16 of the Assignment 3 Questions.

   i. **First model**
   Model equation: $Y = \beta_0 + \beta_1 X_1 + E$
   Hypotheses: H_0: $\beta_1 = 0$ versus H_A: $\beta_1 \neq 0$.
   Conclusion: Reject H_0 at the 5% level of significance, as $p = 0.0002 < 0.05$.
   Interpretation: Used alone, riparian tree density is a useful predictor of coarse woody debris.

   ii. **Second model**
   Model equation: $Y = \beta_0 + \beta_2 X_2 + E$
   Hypotheses: H_0: $\beta_2 = 0$ versus H_A: $\beta_2 \neq 0$.
   Conclusion: Reject H_0 at the 5% level of significance, as $p = 0.0002 < 0.05$.

Interpretation: Used alone, log(1 + number of cabins) is a useful predictor of coarse woody debris.

iii. **Third model**
Model equation: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + E$
Two tests are done, as follows.

Hypotheses: H_0: $\beta_1 = 0$ versus H_A: $\beta_1 \neq 0$.
Conclusion: Reject H_0 at the 5% level of significance, as $p = 0.0367 < 0.05$.
Interpretation: Assuming log(1 + number of cabins) is included in the model anyway, riparian tree density still provides useful extra information for predicting the amount of coarse woody debris.

Hypotheses: H_0: $\beta_2 = 0$ versus H_A: $\beta_2 \neq 0$.
Conclusion: Reject H_0 at the 5% level of significance, as $p = 0.0267 < 0.05$.
Interpretation: Assuming riparian tree density is included in the model anyway, log(1 + number of cabins) still provides useful extra information for predicting the amount of coarse woody debris.

The diagnostic graphs support the assumption of constant variance (there is a level band in the plot of studentized residuals versus predicted values), and the assumption of normality of residuals also seems justified (there is a straightish line in the Q-Q plot of residuals). There are no outliers or points of high leverage, as no Cook's distance goes above 1. Independence can't be checked, so must be assumed.

(c) [**4 marks**] **Which test answers the question of interest and what is the answer?**

The main question was whether, after allowing for riparian tree density, information about human habitation (measured by L10CABIN) also has an effect on CWD. This is answered in the final test presented above – we compare two models, a reduced model including riparian tree density only, and a complete model with riparian tree density plus L10CABIN.

The result is that (at a 5% significance level) it **is** useful to include L10CABIN as an extra predictor, since $p = 0.0267 < 0.05$. Since the estimate of β_2 was negative (-56.26), the extra impact of human habitation on CWD is negative.

This means that although riparian tree density alone provides some explanatory power for CWD, extra explanation is obtained if we know about the number of cabins, which are a surrogate measure for human habitation.

4. **ANCOVA Question [20 marks]**

(a) **[3 marks] Confirmatory model**

The confirmatory modelling uses analysis of covariance (ANCOVA). Using the notation from the question, the model equation is

$$Y = \alpha_i + \beta_i x + E$$

with the factor Method at level i. Note there are only two Methods, A and B.

(b) **[17 marks] Analysis and interpretation**

Results from the fitted model for Y explained by terms x, Method and $x \times$Method (entered in that order) gave the output on pages 18 to 20 of the Assignment 3 Questions.

The model assumptions are that the error term E has a $N(0, \sigma^2)$ distribution and that the errors are independent of each other and of x.

The plot of RStudent versus Predicted Value (p.19, Ass3 Questions) shows a reasonably level band, with no funnelling. This suggests constant variance. Possibly the variance is lower at the intermediate values of Y. The Residual versus Quantile plot (Q-Q plot, same page) shows a fairly straight line, indicating normality. We cannot check independence without knowing more details of how the experiment was run.

The analysis starts with the interaction test:
H_0: there is no interaction (parallel lines, both slopes are equal, $\beta_A = \beta_B = \beta$),
H_A: there is interaction (slopes are not equal, $\beta_A \neq \beta_B$).
With $p = 0.2938 > 0.05$, H_0 is not rejected. We have not detected any significant interaction between Method and x in their effect on Y. So we can proceed to the main effects tests.

Main effect of Method:
H_0: there is no main effect of Method after allowing for x, ($\alpha_A = \alpha_B$),
H_A: there is a main effect of Method after allowing for x ($\alpha_A \neq \alpha_B$).
Using Type I sums of squares, $p = 0.0199 < 0.05$, we reject H_0. There is a significant main effect of Method, after allowing for the effect of $x =$ true age of tooth.

Main effect of true age:
H_0: there is no main effect of $x =$ true age of tooth ($\beta = 0$),
H_A: there is a main effect of x (non-zero slope, $\beta \neq 0$).
Using Type I sums of squares, $p < 0.0001$, we reject H_0. There is a significant main effect of x. *Note: it would be extremely unusual if the estimated age of a tooth was not influenced by the true age of that tooth!*

Interpretation: Referring to the interaction plot (p.20, Ass3 Questions), the lack of interaction is illustrated by the two fitted lines being nearly parallel. The main effect of Method is seen in the vertical separation of the lines (age estimated with Method A is slightly above the estimate from Method B). The main effect of $x =$ true age is seen by the positive slope of the lines. The lines should definitely have a positive slope, since Y is being used to estimate the true age x.