# Practical 2: Review of Linear Model

## One-way analysis of variance using the normal linear model.

A cancer researcher obtained measurements of the metabolic rate of glucose utilization for low, intermediate and high grade tumors of a particular type.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Low | 54.7 | 50.5 | 45.9 | 59.7 | 51.5 | 45.5 | 47.4 | 55.6 | 59.4 | 56.9 | |
| Intermediate | 48.6 | 55.8 | 62.6 | 64.7 | 54.3 | 61.3 | 44.6 | 59.8 | 46.0 | 56.5 | 42.4 |
| | 71.4 | 62.6 | 67.9 | 55.2 | 47.3 | 75.9 | | | | | |
| High | 63.8 | 65.4 | 68.1 | 70.0 | 63.9 | 63.6 | 63.6 | 59.5 | 56.3 | 59.1 | 60.6 |

1. Enter these data into a single vector (y):

```
low = c(54.7, 50.5, 45.9 ,59.7 ,51.5, 45.5,
47.4, 55.6, 59.4 ,56.9)

inter = c(48.6, 55.8, 62.6, 64.7 ,54.3, 61.3,
44.6, 59.8 ,46.0, 56.5,42.4 ,71.4, 62.6, 67.9,
55.2, 47.3, 75.9)

high=c(63.8 ,65.4, 68.1, 70.0, 63.9, 63.6, 63.6,
59.5, 56.3 ,59.1, 60.6)
```

   Combine the individual vectors (low, inter and high) into a single vector:

```
y = c(low, inter, high)
```

2. Create a corresponding vector for grade with values 1, 2 & 3 corresponding to low intermediate and high:

```
grade = c(low*0+1, inter*0+2, high*0+3)
```

Can you explain how the above R-code creates the appropriate coding scheme?

3. Plot metabolic rate by grade. Interpret this plot.

```
plot(grade,y,ylab="Metabolic
Rate",xlab="Grade")
```

**MR is higher for High Grade tumors than for both the Low and Intermediate Grades.**

**MR is more variable for the Intermediate Grade.**

4. Prepare a box-plot of metabolic rate by grade. Interpret this plot.

```
boxplot(y~grade)
```

**MR is higher for High Grade tumors than both the Low and Intermediate Grades.**

**MR is more variable for the Intermediate Grade.**

**MR distribution is symmetric for Low and Intermediate, but skewed for High.**

**There are no outliers.**

5. Use the aov function in R to obtain a one-way ANOVA of the metabolic rate with grade being the factor. Interpret the summary output.

```
mr.aov = aov(y ~ factor(grade))
```

```
summary(mr.aov)
```

**There is a statistically significant difference between Grade (p = 0.0105).**

6. What is the statistical model for this ANOVA? What are the assumptions that must be satisfied/checked?

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i \qquad \epsilon_i \sim NID(0, \sigma^2)$$

**where** $x_1$ **= 1 if Grade = 2 & 0 otherwise**
$x_2$ **= 1 if Grade = 3 & 0 otherwise**
$i$ *= 1 to 38*

**Errors are Normally distributed (with mean 0).**
**Errors are Independent.**
**Errors have constant variance.**

7. Plot the model residuals by grade. Interpret this plot.

```
plot(grade, mr.aov$residuals)
abline(h = 0)
```

**The residuals do not have constant variance. There is greater variability in the intermediate group.**

8. Obtain a normal probability plot and interpret.

```
qqnorm(mr.aov$residuals)
qqline(resid(mr.aov))
```

**Normal Probability Plot: Good adherence to the reference line.**

**Minor departures.**

**Conclusion: Normal distribution.**

9. Obtain a histogram for the residuals and interpret.

```
hist(resid(mr.aov))
```

**Histogram: Almost perfectly symmetric. Close to Normal/bell-shape.**

**Conclusion: Normal distribution.**

10. We will now repeat the ANOVA using an explicit linear model and the lm function in R. Create variables x2 and x3 which are indicators of Intermediate and High grade cases.

```
x2 = ifelse(grade==2, 1, 0)
x3 = ifelse(grade==3, 1, 0)
```

Combine the y, grade, x2 and x3 vectors into a matrix:

```
cbind(y, grade, x2, x3)
```

What type of variables are x2 and x3?

11. Fit the linear model $Y_i = \beta_0 + \beta_1 X_{i2} + \beta_2 X_{i3} + e_i$, $e_i \sim NID(0, \sigma^2)$:

```
mr.lm = lm(y ~ x2 + x3)
summary(mr.lm)
```

12. What do the parameters $\beta_0$, $\beta_1$ and $\beta_2$ represent?

**$\beta_0$: mean MR for Low Grade (code 1).**
**$\beta_1$: the difference in mean MR between Intermediate and Low.**
**$\beta_2$: the difference in mean MR between High and Low.**

13. What inferences can be drawn about $\beta_1$ and $\beta_2$ and hence the 3 grades of tumor?

**Accept $H_0$: $\beta_1 = 0$**
**Intermediate Grade is not significantly different to Low Grade (p = 0.11530).**

**Reject $H_0$: $\beta_2 = 0$: High Grade is significantly higher, by 10.372, than Low Grade (p = 0.00282).**

14. Fit the linear model $Y_i = \beta_0 + e_i$, $e_i \sim NID(0, \sigma^2)$. What does this model imply?

```
mr2.lm <- lm(y ~ 1)

summary(mr2.lm)
```

**The model implies that there is no difference between Grades – a common mean.**

**The Null model.**