STAT2170 and STAT6180
Applied Statistics

SGTA Week 11: Two-Way ANOVA 2022

```
## here() starts at C:/Users/hanse/Downloads/Dropbox/Teaching/MQ/Units/STAT2170-6180Applied Statistics/
```

Part 1

Question 1

Growth measurements for children with deficient bone development after being given a hormone treatment are given below. Data taken from Neter et al. (1996) and are contained in the file **bone-growth.dat**.

growth	growth rate difference (cm per month)
gender	male or female
bone.dev	child's bone development (Mild, Moderate and Severe)

The aim is to determine if growth rates are similar for males and females after taking into account the bone development.

- Determine if the data is balanced or unbalanced.
  - Count the number of observations in each combination of the two factor levels. (**Hint**: Use the table command along the columns of the data that specify the levels of both factors.)
- Conduct a preliminary analysis of the data using graphical summaries (plots).
  - Produce an interaction plot using the `interaction.plot` function.
  - Produce a Box plot and comment on each graph.
  - Comment on any structure of the data using the two above preliminary plots.
  - State if the design is balanced or unbalanced.
- Construct a Two-Way ANOVA analysis using either the `lm` command or the `aov` command.
  - First produce a One-Way ANOVA with `growth` explained by `gender`.
  - Add the `bone.dev` factor to the model to create a Two-Way ANOVA. Comment on the changes in Sum of Squares and F-statistic and resulting P-Value for the Two-Way ANOVA compared to the One-Way ANOVA.
  - Reverse the order in your Two-Way ANOVA and comment.
  - Add in a interaction term and inspect the ANOVA table. Write down the model and test the significance of the interaction.
- Choose a model to check the effect of `bone.dev` after controlling for the effect of gender and comment.
- Validate your final Two-Way ANOVA model.

Part 2: Previous exam question

Question 1

An investigation into the respiratory function in developing children and young adults was conducted. A random sample of individuals and relevant measurements were taken. A relationship is suspected between the Forced Expiratory Volume (FEV) measured in litres, and the age in years of an individual. Linear and polynomial regression models are analysed in R below.

FEV	Forced Expiratory Volume measured in litres
Age	Age of patient in years

```
fev.1 = lm(FEV ~ Age, data = fev)
fev.2 = lm(FEV ~ Age + I(Age*Age), data = fev)
fev.3 = lm(FEV ~ Age + I(Age*Age) + I(Age*Age*Age), data = fev)
anova(fev.1, fev.2, fev.3)
```

```
# Analysis of Variance Table
#
# Model 1: FEV ~ Age
# Model 2: FEV ~ Age + I(Age * Age)
# Model 3: FEV ~ Age + I(Age * Age) + I(Age * Age * Age)
#   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
# 1    316 59.991
# 2    315 50.130  1    9.8613 61.9248 5.823e-14 ***
# 3    314 50.003  1    0.1262  0.7925     0.374
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

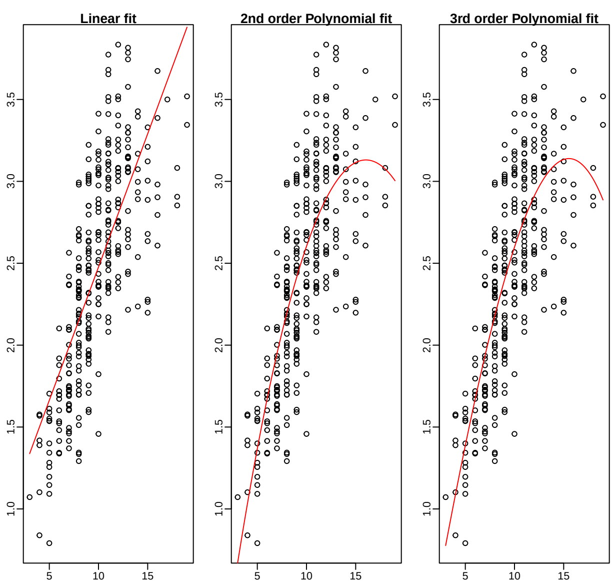


Figure 1: Linear and polynomial fits for the FEV Data

```r
summary(fev.2)
```

```
#
# Call:
# lm(formula = FEV ~ Age + I(Age * Age), data = fev)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -1.14506 -0.27465 -0.00089  0.26202  1.01063
#
# Coefficients:
#               Estimate Std. Error t value Pr(>|t|)
# (Intercept)  -0.591761   0.199190  -2.971   0.0032 **
# Age           0.464187   0.039051  11.887  < 2e-16 ***
# I(Age * Age) -0.014470   0.001838  -7.872 5.68e-14 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 0.3989 on 315 degrees of freedom
# Multiple R-squared:  0.6207,  Adjusted R-squared:  0.6183
# F-statistic: 257.8 on 2 and 315 DF,  p-value: < 2.2e-16
```

```r
summary(fev.3)
```

```
#
# Call:
# lm(formula = FEV ~ Age + I(Age * Age) + I(Age * Age * Age), data = fev)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -1.14051 -0.27603  0.00088  0.27385  1.00542
#
# Coefficients:
#                      Estimate Std. Error t value Pr(>|t|)
# (Intercept)        -0.2196691  0.4630499  -0.474   0.6355
# Age                 0.3423119  0.1423709   2.404   0.0168 *
# I(Age * Age)       -0.0023132  0.0137800  -0.168   0.8668
# I(Age * Age * Age) -0.0003736  0.0004197  -0.890   0.3740
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 0.3991 on 314 degrees of freedom
# Multiple R-squared:  0.6217,  Adjusted R-squared:  0.6181
# F-statistic:   172 on 3 and 314 DF,  p-value: < 2.2e-16
```

```r
fev.31 = update(fev.3, FEV ~ . - I(Age*Age))
summary(fev.31)
```

```
#
# Call:
# lm(formula = FEV ~ Age + I(Age * Age * Age), data = fev)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -1.13930 -0.27585  0.00172  0.27215  1.00480
#
# Coefficients:
#                     Estimate Std. Error t value Pr(>|t|)
# (Intercept)       -1.460e-01  1.480e-01  -0.987    0.325
# Age                3.187e-01  2.109e-02  15.107  < 2e-16 ***
# I(Age * Age * Age) -4.434e-04  5.592e-05  -7.930 3.85e-14 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 0.3984 on 315 degrees of freedom
# Multiple R-squared:  0.6217,  Adjusted R-squared:  0.6193
# F-statistic: 258.8 on 2 and 315 DF,  p-value: < 2.2e-16
```

```r
fev.32 = update(fev.31, FEV ~ . - 1)
summary(fev.32)
```

```
#
# Call:
# lm(formula = FEV ~ Age + I(Age * Age * Age) - 1, data = fev)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -1.12930 -0.29488 -0.00938  0.26609  1.01959
#
# Coefficients:
#                     Estimate Std. Error t value Pr(>|t|)
# Age                2.984e-01  4.738e-03   62.98   <2e-16 ***
# I(Age * Age * Age) -3.966e-04  2.961e-05  -13.39   <2e-16 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 0.3984 on 316 degrees of freedom
# Multiple R-squared:  0.9754,  Adjusted R-squared:  0.9753
# F-statistic:  6276 on 2 and 316 DF,  p-value: < 2.2e-16
```

a) Referring to the plots on linear and polynomial fits for the FEV data, explain why the simple linear regression model `fev.1` is inadequate for this FEV dataset.

b) Notice that in model `fev.3`, the intercept and quadratic terms are not statistically significant. One may consider removing the intercept and quadratic terms from model `fev.3` according to the backward selection technique and obtain `fev.32` as the final best model. Why is it **not** appropriate to use the backward selection technique to remove the intercept and quadratic terms in model `fev.3`?

c) Given the technique described above is invalid, and only using the output on the previous page, `fev.2` seems better than `fev.3`? Why?

The fitted polynomial model for `fev.2` is given by

$$\widehat{FEV} = -0.591761 + 0.464187(Age) - 0.014470(Age^2) \tag{1}$$

d) According to (1), how much does FEV change when Age increases by 1 year? Explain.

e) Using your fitted polynomial in (1), predict the FEV for a 10 year old child.

f) Using your fitted polynomial in (1), can you predict the FEV for a 100 year old adult? If not, why?

g) Both models, `fev.2` and `fev.3`, require validation. Some diagnostic plots are shown below. Using only these plots, explain whether the models are valid for analysis.

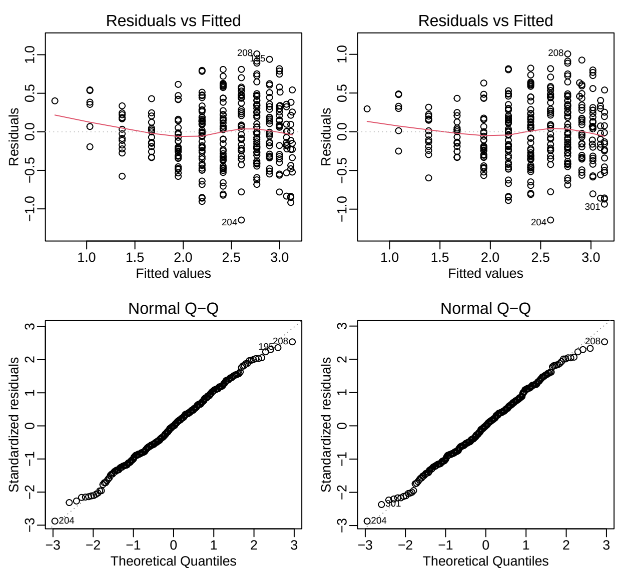


Figure 2: Diagnostic plots for the polynomial regression models. Left plots: `fev.2`; Right plots: `fev.3`

Part 3: Bibliography Management in RMarkdown

In `RMarkdown`, the `bibliography` field in `YAML` allows one to automatically collate & print out the references of in-text citations at the end of the html/pdf document. For an example, for **Neter et al. (1996)** in the beginning of this `pdf`, one can automatically include its corresponding bibliography entry `Applied linear statistical models` in the end of this document. Now let's see how.

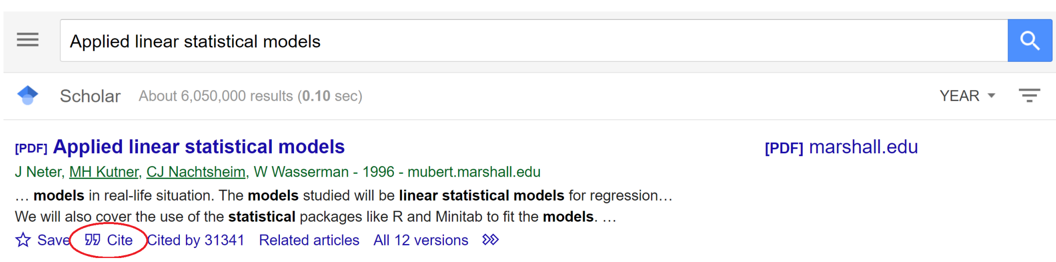
a) Create a `RMarkdown` file with default output format `PDF`. Save this `RMarkdown` file on your desktop.

b) Now we link this `RMarkdown` file to a bibliography document with extension `.bib`. Please

- add a line of `bibliography: Reference.bib` below the `output: pdf_document` at the start of this `RMarkdown` file.
- create a `Text Document` named `Reference.txt` on your desktop.
- change the extension of `Reference.txt` from `.txt` to `.bib`.

c) Now we add the reference of Neter et al. (1996) to the previous `.bib` bibliography document.

- Open scholar.google.com on your web browser
- In the search box, enter `Applied linear statistical models` and then press `Enter` on your keyboard
- Under the first entry of the search result, find a quotation mark and a link `Cite` (see figure below). Click on this `Cite`.
- In the Pop-up window, you may see various format of the reference; you may also find a link `BibTeX` at the bottom of this window (where `BibTeX` is a file format that describes lists of references). Click on this `BibTex`.
- In the Pop-up window, you may see some text starting with `@article{neter1996applied,` (where `neter1996applied` is the label for the reference). Copy all the texts in this window and paste them to your `Reference.bib` file on your desktop.



d) Now we add another reference to the previous `.bib` bibliography document.

- Open scholar.google.com on your web browser
- In the search box, enter `Package 'mpcmp'` and then press `Enter` on your keyboard
- Under the first entry of the search result, find a quotation mark and a link `Cite` (see figure below). Click on this `Cite`.
- In the Pop-up window, you may see various format of the reference; you may also find a link `BibTeX` at the bottom of this window (where `BibTeX` is a file format that describes lists of references). Click on this `BibTex`.
- In the Pop-up window, you may see some text starting with `@article{fung2020package,` (where `fung2020package` is the label for the reference). Copy all the texts in this window and paste them (below the entry of `Applied linear statistical models`) to your `Reference.bib` file on your desktop.

e) Now we cite **Neter et al. (1996)** in our `RMarkdown` document

- Enter `@neter1996applied` in the `RMarkdown` document
- Knit the `RMarkdown` document and generate a `PDF` file
- Find the reference corresponding to `@neter1996applied` at the end of this `PDF` file.

f) Now we play with the references in our `RMarkdown` document

- Enter `[@neter1996applied]`, which put the citation in parenthesis
- Enter `[@neter1996applied; @fung2020package]`, which cites multiple entries
- Knit the `RMarkdown` document and check the `PDF` file you generate

References

Neter, John, Michael H Kutner, Christopher J Nachtsheim, and William Wasserman. 1996. *Applied Linear Statistical Models*. Vol. 4. Irwin Chicago.