Adapted from STAT1201, semester one 2011

Q1. Embryonic stem cells have the capacity to produce neural progenitor cells, providing a potential means of repopulating cells lost due to spinal cord injury. However cell survival using existing methods is often low. Aiming to improve this, researchers investigated growing cells inside fibrin scaffolds and looked at the effect of the growth factor neurotropin-3 (NT3) on the outcomes. Mouse embryonic stem cells were cultured in 8 wells seeded with fibrin scaffolds to which various concentrations of the growth factor (ng/mL) were added. Fluorescence-activated cell sorting was used after three days to count the number of cells and the percentage of living cells in each well.

a) The least squares line fitted using MATLAB was Alive = 55.04 + 0.639 NT3 with a standard error for the slope of 0.1805 %/ng/mL. Give bounds on the *P*-value for the test of linear association between the percentage of cells alive and NT3 concentration. What do you conclude?

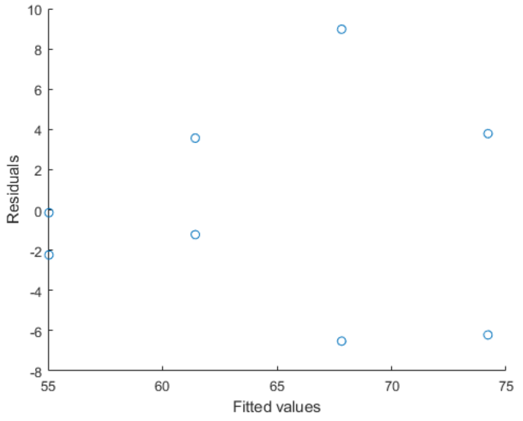
b) What is the sum of the residuals from the least-squares line?

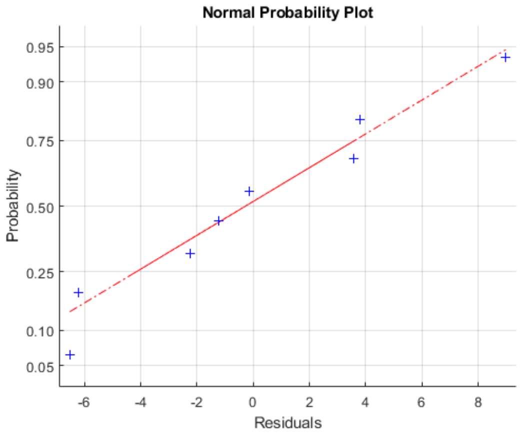
c) The estimated covariance matrix for the estimators of intercept and slope is:

$$\begin{bmatrix} 11.397 & -0.4884 \\ -0.4884 & 0.0326 \end{bmatrix}$$

Give a 95% confidence interval for the mean percentage of cells alive for an NT3 concentration of 25 ng/mL.

d) The following figures show a scatter plot of the residuals against fitted values from the regression along with a Normal probability plot of the residuals. State the assumptions of the linear regression model and comment on their validity based on these plots.





Adapted from STAT1201, semester two 2011

Q2. *Staphylococcus aureus* is amongst the most important pathogenic bacteria responsible for bloodstream nosocomial infections and for biofilm formation on indwelling medical devices. Researchers evaluated the effect of a coral associated actinomycete (CAA) on the growth of both *S. aureus* (SA) and methicillin resistant *S. aureus* (MRSA) biofilms. Various concentrations of CAA extract (μg/mL) were applied to 20 biofilm preparations where 10 had SA and 10 had MRSA. The initial thickness (μm) and final thickness (μm) of the biofilm were recorded.

a) If the CAA concentrations were randomly assigned to the preparations then there should be no association between initial biofilm thickness and the CAA concentration. The researchers found a correlation between the 20 pairs of r = -0.3508. Carry out a test for an association. Show your working and give bounds on the *P*-value. What do you conclude?

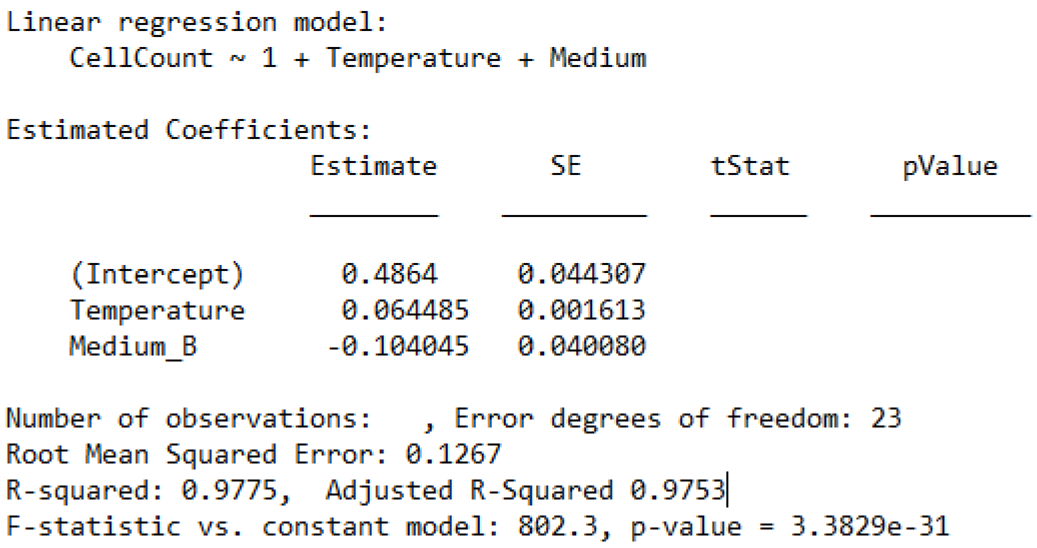
b) The least squares line for the relationship between biofilm thickness reduction and CAA concentration was *Reduction* = 5.002+ 0.02956 *CAA*. What are the units of the slope value?

c) Based on the linear model in (b), what is the estimated reduction in biofilm thickness for a CAA concentration of 75 μg/mL?

d) The standard error for the slope in (b) was 0.01910. Show the appropriate test statistic and give bounds on the *P*-value for the test of linear association between the biofilm thickness reduction and CAA concentration. What do you conclude?

Adapted from STAT1201, semester two 2012

Q3. A biologist compared the effect of temperature for each of two media, *A* and *B*, on the growth of human amniotic cells in a tissue culture. The following linear regression analysis was obtained for the cell counts (×10^6) in relation to the temperature (°C) and the medium. Note that MATLAB has introduced the dummy variable Medium_B that is 1 for a culture with *B* and 0 for a culture with *A*.



```
Linear regression model:
    CellCount ~ 1 + Temperature + Medium

Estimated Coefficients:
                   Estimate       SE         tStat       pValue
                   ________    _________    ______    __________

    (Intercept)      0.4864     0.044307
    Temperature    0.064485     0.001613
    Medium_B      -0.104045     0.040080

Number of observations:    , Error degrees of freedom: 23
Root Mean Squared Error: 0.1267
R-squared: 0.9775,  Adjusted R-Squared 0.9753
F-statistic vs. constant model: 802.3, p-value = 3.3829e-31
```

a) How many experimental trials were used in this analysis?

b) Temperatures ranged from 4°C to 38°C. Sketch the linear relationship between cell count and temperature for the two different media. Clearly indicate the intercept and slope for each line.

c) Carry out a *t* test for whether this model gives evidence of an association between cell count and temperature. Show your working and state your conclusion.

d) Construct a 90% confidence interval for the estimated coefficient of `Medium_B`.

e) What is the estimated mean cell count for a tissue culture with medium *A* at 30°C?

f) The $(X^TX)^{-1}$ matrix from the regression fit was

$$\begin{bmatrix} 0.1223 & 2.143e-05 & 5.334e-04 \\ 2.143e-05 & 1.6207e-04 & 1.186e-05 \\ 5.334e-04 & 1.186e-05 & 0.1001 \end{bmatrix}$$

Determine a 95% confidence interval for the mean cell count for a tissue culture with medium *A* at 30°C?

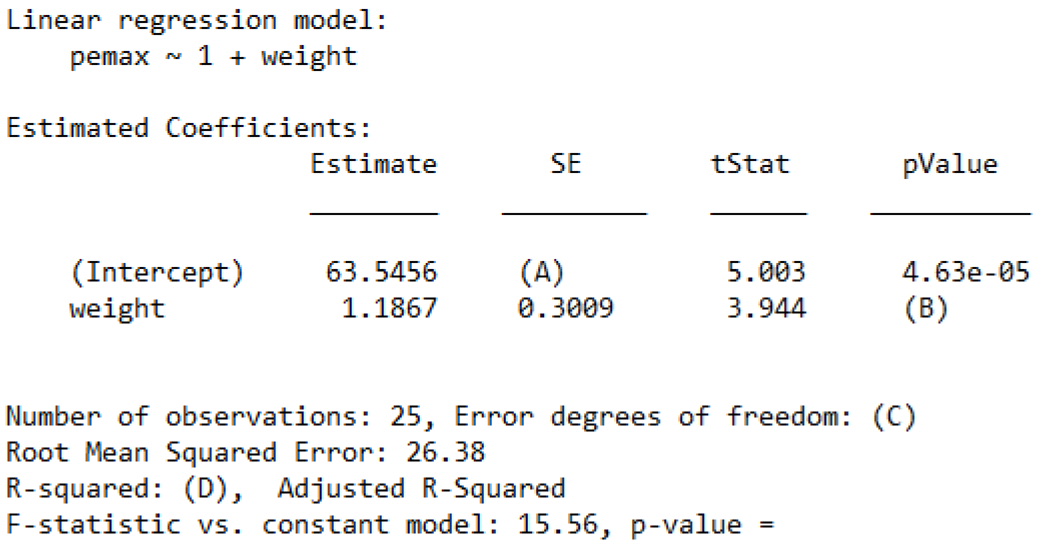
Adapted from STAT1201, semester two 2013

Q4. We are interested in analysing the data from a study conducted by David Altman in 1991. The aim of this study was to investigate lung function for cystic fibrosis patients (7-23 years old). The data contains a number of variables and in this question we are interested in the relationship between `weight` and `pemax`:

- `weight` = Weight (kg)
- `pemax` = Maximum expiratory pressure

a) Write down the linear model and its assumptions to study the relationship between the response variable `pemax` and the explanatory variable `weight`.

b) The edited MATLAB output for the linear model is presented below. Compute the values of (A) and (C) and give a range for (B).



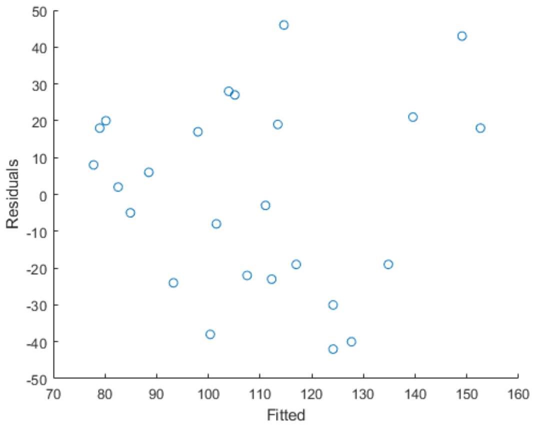
```
Linear regression model:
    pemax ~ 1 + weight

Estimated Coefficients:
                   Estimate       SE        tStat       pValue  
                   ________    _________    ______    __________

    (Intercept)     63.5456     (A)          5.003     4.63e-05
    weight           1.1867     0.3009       3.944     (B)


Number of observations: 25, Error degrees of freedom: (C)
Root Mean Squared Error: 26.38
R-squared: (D),  Adjusted R-Squared
F-statistic vs. constant model: 15.56, p-value =
```

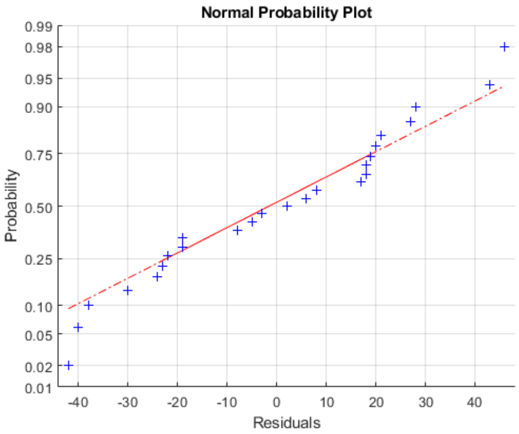
c) Based on the linear model, what is the estimated mean of `pemax` for a `weight` of 50kg?

d) The Pearson correlation coefficient *r* between the variables `pemax` and `weight` is equal to 0.635. Calculate the coefficient of determination R^2 between `pemax` and `weight`.

e) Conduct a hypothesis test to determine whether there is any evidence of a linear association between `pemax` and `weight`. Show your working and state your conclusion.

f) The following figures were produced to check the assumptions of the model. Interpret these two figures.





Adapted from STAT1201, semester two 2014

Q5. A research group measured the inorganic phosphorous (Pi) content (ppm) of soils, and performed an analysis to see how this was related to the available phosphorous (Pp) (ppm at 20 C) for plants. They were interested in the relationship between the inorganic phosphorous (Pi) and the available phosphorus (Pp) for plant growth. They undertook an analysis and the output is provided below.

```
Linear regression model:
    Pp ~ 1 + Pi

Estimated Coefficients:
                   Estimate       SE       tStat        pValue
                   ________    ________    ______    ____________

    (Intercept)     62.5694     4.4519     14.055      4.85e-10
    Pi               1.2291     0.3058      (A)        (B)


Number of observations: (C), Error degrees of freedom: 15
Root Mean Squared Error: 11.92
R-squared: 0.5185,  Adjusted R-Squared: 0.4864
F-statistic vs. constant model: 16.15, p-value = 0.001115
```

a) How many soil samples were measured in this study?

b) Define the linear model (using symbols) for investigating the relationship between the inorganic phosphorous and the available phosphorus for plant growth. Include the assumption on the distributions.

c) Give the slope and intercept of the regression line. Use this to estimate the mean availability of phosphorous at 17 ppm of inorganic phosphorous.

d) Determine the values of (A) and (B) from the incomplete output.

e) Is there any evidence that the slope in the linear relationship between inorganic phosphorous and available phosphorus is different to *one*? State the null and alternative hypothesis, compute the test statistic and *P*-value from the output. What do you conclude?

f) Determine the residual for the observation Pi=12.6 ppm and Pp=51 ppm.

g) Suggest two diagnostic tools to check the model assumptions?

h) The matrix below is the estimated covariance matrix for the estimator of intercept and slope. Fill in the missing entries from the above output.

$$\begin{bmatrix} (D) & (E) \\ 0.2763 & (F) \end{bmatrix}$$