MATH08051: Statistics

Worksheet 4 2017–2018

Bring your notes to workshops! Hand in attempts to the assessed questions by 14.10 Monday 19th March (i.e. before the lecture). These should be placed inside the collection cabinet situated outside room 5312 (The Maths Hub). **Please clearly mark your name and workshop number/time on your solution.** Make sure your name is on every page submitted and you securely staple/secure multiple pages together.

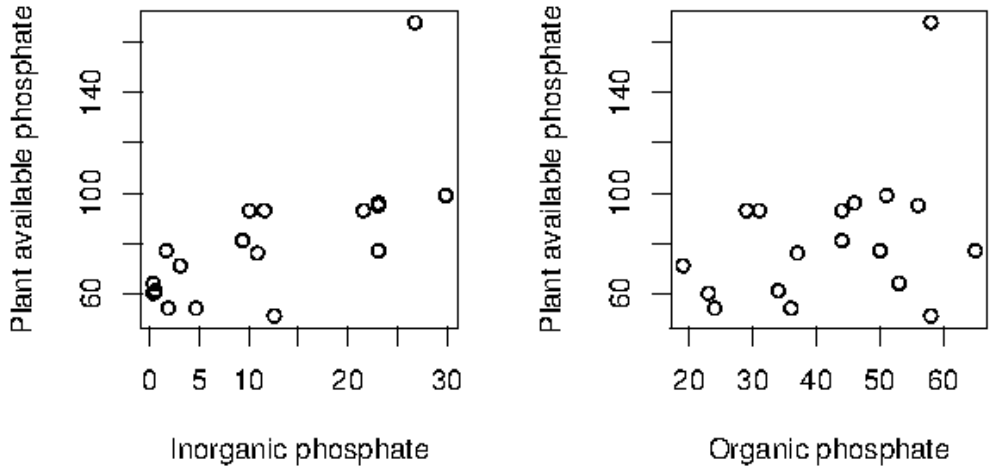
Question 1 will contribute to your final mark - the number of marks for each question are given in [ ] brackets at the end of each question (or part question). The best four marks, out of the five worksheets, will be used for assessment. In total these will account for 15% of the final mark for the course.

If you choose to use R using a Jupyter notebook, a Jupyter version of this worksheet is available on Learn.

Assessed Questions

1. We wish to fit a linear regression analysis in which the response is estimated plant-available phosphorus (PAphos) in 18 Iowa soils at 20°C in parts per million, and the two explanatory variables are inorganic phosphorus (inorg) and organic phosphorus (org). The R commands for reading in the data and plotting the data are given below:

```
> PAphos <- c(64,60,71,61,54,77,81,93,93,51,76,96,77,93,95,54,168,99)
> inorg <- c(0.4,0.4,3.1,0.6,4.7,1.7,9.4,10.1,11.6,12.6,10.9,23.1,
>            23.1,21.6,23.1,1.9,26.8,29.9)
> org <-  c(53,23,19,34,24,65,44,31,29,58,37,46,50,44,56,36,58,51)
> par(mfrow=c(1,2))
> plot(inorg,PAphos,xlab="Inorganic phosphate", ylab="Plant available phosphate")
> plot(org,PAphos,xlab="Organic phosphate", ylab="Plant available phosphate")
```



(a) Fit the three different normal linear models in R corresponding to:

(i) The plant-available phosphorous linearly regressed on inorganic phosphorous;

(ii) The plant-available phosphorous linearly regressed on organic phosphorous;

(iii) The plant-available phosphorous linearly regressed on both inorganic and organic phosphorous.

You should provide both your `R` commands and the associated `R` output. This is most easily done by simply cutting and pasting the `R` commands/output into an editor. [3]

(b) Which of the three models would you use for further analyses? Justify your answer. [2]

(c) For the favoured model state the fitted regression model for the expected response. [1]

(d) For model (i) above calculate a 95% interval for the slope of the regression line. [2]

(e) State the underlying assumptions made in these analyses. [2]

Additional Questions for Workshop

2. Consider the straight line model $\mathbb{E}(Y) = \alpha + \beta x$. Only one of the following assumptions is required when estimating α and β, using a simple linear regression of Y on x. Which one?

   (a) The observations on x are independent.

   (b) The observations on x are independent of those on Y.

   (c) The observations on Y are independent.

   (d) The observations on Y are normally distributed.

3. Consider the model $\mathbb{E}(Y) = \alpha + \beta x$, where observations $Y \sim N(\alpha + \beta x, \sigma^2)$, and $s^2 = \hat{\sigma}^2$ is estimated in the standard way. Which of the following expressions provides the 95% confidence interval for the mean response $\hat{\mathbb{E}}(Y_0) = \hat{\alpha} + \hat{\beta} x_0$?

   (a)
   $$\hat{\alpha} + \hat{\beta} x_0 \pm z_{0.025} \sqrt{s^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}} \right)}.$$

   (b)
   $$\hat{\alpha} + \hat{\beta} x_0 \pm z_{0.025} \sqrt{s^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}} \right)}.$$

   (c)
   $$\hat{\alpha} + \hat{\beta} x_0 \pm t_{n-2;0.025} \sqrt{s^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}} \right)}.$$

   (d)
   $$\hat{\alpha} + \hat{\beta} x_0 \pm t_{n-2;0.025} \sqrt{s^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}} \right)}.$$

   where $z_{0.025}$ is the upper 2.5% quantile of the $N(0, 1)$ distribution, and $t_{n-2;0.025}$ is the upper 2.5% quantile of the t distribution with $n - 2$ degrees of freedom.

4. The following table gives the death rates per 100,000 from typhoid fever in the United States for six years in the period 1900–1920.

Year	1900	1904	1908	1912	1916	1920
Rate	31.1	23.9	19.6	13.2	8.8	5.0

The following analysis was conducted in R, regressing the death rate on year.

```
> year<-c(1900,1904,1908,1912,1916,1920)
> rate<-c(31.1,23.9,19.6,13.2,8.8,5.0)
> typhoidreg<-lm(rate~year)
> summary(typhoidreg)

Call:
lm(formula = rate ~ year)

Residuals:
       1        2        3        4        5        6
 1.15238 -0.84190  0.06381 -1.13048 -0.32476  1.08095

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 2502.66190  122.31060   20.46 3.37e-05 ***
year          -1.30143    0.06404  -20.32 3.46e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.072 on 4 degrees of freedom
Multiple R-Squared: 0.9904,     Adjusted R-squared: 0.988
F-statistic:   413 on 1 and 4 degrees of freedom,  p-value: 3.461e-05
```

(a) State the fitted regression model.

(b) Estimate the year in which typhoid fever would have been eradicated in the United States, if the linear trend had continued. How much trust would you put in this estimate?

(c) It is more plausible that the death rate is proportional to $e^{\beta t}$, where t denotes the year and β is an unknown parameter. How could simple linear regression be used to estimate β?

5. Consider the following R session.

```
> sdm.reg1<-lm(w~t+u)
> summary(sdm.reg1)

Call:
lm(formula = w ~ t + u)

Residuals:
    Min      1Q  Median      3Q     Max
-25.254 -10.366  -3.759  11.878  24.743

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) -20.9939     4.3774  -4.796 0.000436 ***
t             9.3295     1.0127   9.213 8.62e-07 ***
u             7.8727     0.8764   8.983 1.13e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.87 on 12 degrees of freedom
Multiple R-Squared: 0.9383,     Adjusted R-squared: 0.928
F-statistic: 91.27 on 2 and 12 degrees of freedom,    p-value: 5.51e-08
```

(a) Describe the model being fitted to the data, mentioning any assumptions made. Does the model fit the data well?

(b) An analysis of the same data but including an additional variable V produced the following output:

```
> sdm.reg<-lm(w~t+u+v)
> summary(sdm.reg)

Call:
lm(formula = w ~ t + u + v)

Residuals:
    Min      1Q  Median      3Q     Max
-25.315  -8.786   6.164   8.828  16.819

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -12.166      6.045  -2.012  0.06932 .
t             21.769      6.511   3.343  0.00655 **
u             15.215      3.886   3.915  0.00241 **
v            -11.557      5.989  -1.930  0.07984 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.23 on 11 degrees of freedom
Multiple R-Squared: 0.9539,     Adjusted R-squared: 0.9413
F-statistic: 75.89 on 3 and 11 degrees of freedom,   p-value: 1.235e-07
```

Is this model more appropriate than that in (a) at $\alpha = 0.05$? Justify your answer.