The University of Sydney
School of Mathematics and Statistics

# Assignment 1

MATH1905: Statistics (Advanced) Semester 2, 2017

Web Page: `http://sydney.edu.au/science/maths/MATH1905`
Lecturer: Michael Stewart

This assignment is due by **5pm Monday 21st August** 2017, via Turnitin. A PDF copy of your answers must be uploaded in the Learning Management System (Blackboard) at `https://elearning.sydney.edu.au`. Please submit only a PDF document. It should include your name and SID; your tutorial time, day, room and Tutor's name. It is your responsibility to preview each page of your assignment after uploading to ensure each page is included in correct order and is legible (not sideways or upside down) before confirming your submission, and then to check your submission receipt. The School of Mathematics and Statistics encourages some collaboration between students when working on problems, but students must write up and submit their own version of the solutions.

This assignment is worth 5% of your final assessment for this course. Your answers should be well written, neat, thoughtful, mathematically concise, and a pleasure to read. Please cite any resources used and show all working. Present your arguments clearly using words of explanation and diagrams where relevant. After all, mathematics is about communicating your ideas. This is a worthwhile skill which takes time and effort to master. The marker will give you feedback and allocate an overall letter grade and mark to your assignment using the following criteria:

| Mark | Grade | Criterion |
|---|---|---|
| 10 | A+ | Outstanding and scholarly work, answering all parts of all questions correctly, with clear accurate explanations and all relevant diagrams and working. There are at most only minor or trivial errors or omissions. |
| 9 | A | Very good work, making excellent progress on both questions, but with one or two substantial errors, misunderstandings or omissions throughout the assignment. |
| 7 | B | Good work, making good progress on 1 question and moderate progress on the other, but making more than two distinct substantial errors, misunderstandings or omissions throughout the assignment. |
| 6 | C | A reasonable attempt, making moderate progress on both questions. |
| 4 | D | Some attempt, with moderate progress made on only 1 question. |
| 2 | E | Some attempt, with minimal progress made on only 1 question. |
| 0 | F | No credit awarded. |

You may use R to check your answers, but please provide a complete handwritten solution (including residual plots in question 2), scan it to PDF and then submit it through Turnitin. **Please do not include any R output/graphics in your assignment solution**.

This assignment explores **multiple least-squares regression**. Elementary calculus and linear algebra are needed to answer some of these questions. Seek assistance from a lecturer or tutor if you need help.

**1.** Suppose that on each of $n$ individuals we have 3 measurements giving ordered triples $(w_1, x_1, y_1)$, $\ldots$,$(w_n, x_n, y_n)$ and that it is desired to find constants $a$, $b$ and $c$ such that we may express each $y_i$ as

$$y_i = aw_i + bx_i + c + \varepsilon_i$$

where the $\varepsilon_i$'s resemble "random errors". It is proposed to choose $a$, $b$ and $c$ using the method of least squares, that is to minimise the function

$$S_1(a, b, c) = \sum_{i=1}^{n} [y_i - (aw_i + bx_i + c)]^2$$

with respect to $a$, $b$ and $c$.

We shall perform the minimisation in two steps:

$$\min_{a,b,c} S_1(a, b, c) = \min_{a,b} \left[ \min_{c} S_1(a, b, c) \right] ,$$

that is find $\hat{c}(a, b) = \arg\min_c S_1(a, b, c)$, the "best" $c$ when $a$ and $b$ are held fixed; then minimise $S_2(a, b) = S_1(a, b, \hat{c}(a, b))$ over $a$ and $b$ simultaneously.

Write $\bar{w}$, $\bar{x}$ and $\bar{y}$ for the averages of the $w_i$'s, $x_i$'s and $y_i$'s respectively.

(a) To perform the inner minimisation it suffices to solve the equation

$$\frac{\partial S_1(a, b, c)}{\partial c} = 0 \, .$$

Show that the solution to this equation is $\hat{c}(a, b) = \bar{y} - a\bar{w} - b\bar{x}$ and thus that

$$S_2(a, b) = \sum_{i=1}^{n} [(y_i - \bar{y}) - a(w_i - \bar{w}) - b(x_i - \bar{x})]^2 \tag{1}$$

(b) To perform the outer minimisation it suffices to solve the pair of equations

$$\frac{\partial S_2(a, b)}{\partial a} = 0 \tag{2}$$

$$\frac{\partial S_2(a, b)}{\partial b} = 0 \tag{3}$$

so long as a unique solution exists. Let $S_{xy}$, $S_{xx}$ and $S_{yy}$ have their usual definitions and define also $S_{wx} = \sum_{i=1}^{n}(w_i - \bar{w})(x_i - \bar{x})$, $S_{wy} = \sum_{i=1}^{n}(w_i - \bar{w})(y_i - \bar{y})$ and $S_{ww} = \sum_{i=1}^{n}(w_i - \bar{w})^2$.

(i) Determine both partial derivatives and write the equations (2) and (3) in matrix form, i.e.

$$\mathbf{M}\begin{pmatrix} a \\ b \end{pmatrix} = \mathbf{v}$$

for a 2-by-2 matrix $\mathbf{M}$ and a column vector $\mathbf{v}$, writing $\mathbf{M}$ and $\mathbf{v}$ in terms of $S_{wx}$, $S_{xx}$, etc. .

(ii) The pair of equations (2) and (3) have a unique solution if and only if the determinant of $\mathbf{M}$ is non-zero. Assume that the $w_i$'s are not all equal and that the $x_i$'s are (also) not all equal. Determine a further necessary and sufficient condition for there to **not** be a unique solution to equations (2) and (3), express it in terms of a well-known bivariate summary statistic and give a geometric interpretation of it.

(iii) Assuming the determinant is non-zero, by inverting $\mathbf{M}$ solve the equations and express the least-squares solutions $a$ and $b$ in terms of $S_{wx}$, $S_{xx}$, etc. .

**2.** Consider the data below which reports flow as a function of depth:

| Depth | 0.340 | 0.290 | 0.280 | 0.420 | 0.290 | 0.410 | 0.760 | 0.730 | 0.460 | 0.400 |
|---|---|---|---|---|---|---|---|---|---|---|
| Flow | 0.636 | 0.319 | 0.734 | 1.327 | 0.487 | 0.924 | 7.350 | 5.890 | 1.979 | 1.124 |

Examine the R commands and output below and answer the questions which accompany them.

(a)

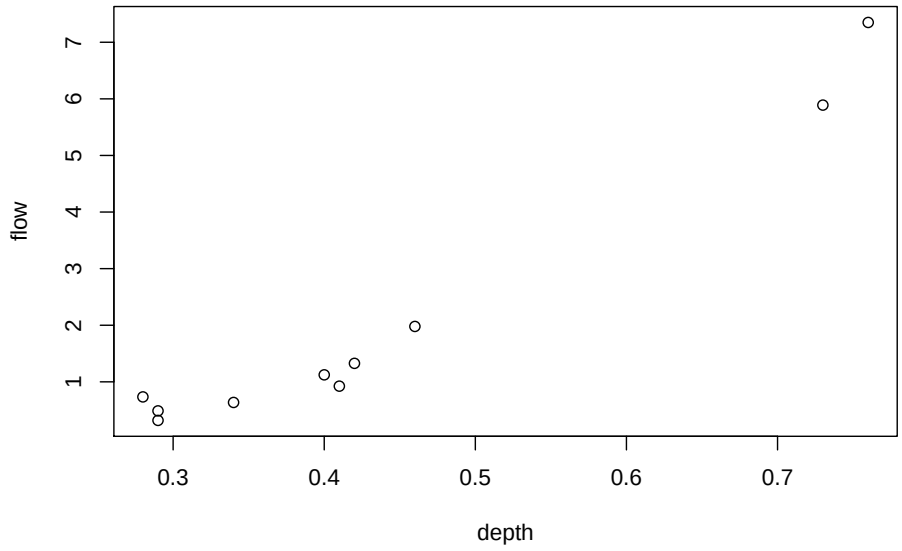
```
depth
```

```
[1] 0.34 0.29 0.28 0.42 0.29 0.41 0.76 0.73 0.46 0.40
```

```
flow
```

```
[1] 0.636 0.319 0.734 1.327 0.487 0.924 7.350 5.890 1.979 1.124
```

```
plot(depth,flow)
```




```
x=depth
y=flow
Sxy=sum((x-mean(x))*(y-mean(y)))
Sxx=sum((x-mean(x))*(x-mean(x)))
Syy=sum((y-mean(y))*(y-mean(y)))
Sxy
```

```
[1] 3.74006
```

```
Sxx
```

```
[1] 0.27036
```

```
Syy
```

```
[1] 54.65201
```

Is there a strong linear relationship between flow and depth? Explain.

(b)

```
b=Sxy/Sxx
a=mean(y)-b*mean(x)
lin.res=y-(a+b*x)
lin.res
```

```
[1] -0.08530433  0.28937713  0.84271342 -0.50099467  0.45737713 -0.76565838
[7]  0.81857139 -0.22641974 -0.40233984 -0.42732209
```

Are the flow values well explained as a linear function of depth, plus "random errors"? Explain, by neatly sketching an appropriate residual plot and describing clearly what it reveals about the linear least-squares fit.

(c) A *least-squares parabola* can be fitted to a set of points $(x_1, y_1), \ldots, (x_n, y_n)$ by defining $w_i = x_i^2$ and then using the procedure analysed in question 1. The fit is performed using the R commands below.

```
w=x^2
quad.fit=lm(y~w+x)
quad.fit$res
```

```
          1           2           3           4           5           6
-0.07465133 -0.19333726  0.24720442  0.05427986 -0.02533726 -0.26198689
          7           8           9          10
 0.32765780 -0.40614505  0.31227639  0.02003933
```

Are the flow values well explained as a *quadratic* function of depth, plus "random errors"? Explain, by neatly sketching an appropriate residual plot and describing clearly what it reveals about the quadratic least-squares fit.