The University of Sydney
School of Mathematics and Statistics

# Assignment 1

MATH1905: Statistics (Advanced) — Semester 2, 2016

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

This assignment is due by **11:59pm Monday 15th August 2016**, 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 (scan or convert other formats). 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. |

This assignment explores **multiple least-squares regression**. Elementary calculus and linear algebra are needed to answer these two questions; seek assistance from a lecturer or tutor if you need help. The two parts marked with an **asterisk*** are quite challenging, don't feel bad if you find them difficult!

Suppose that on each of $n$ individuals we have 3 measurements giving ordered triples $(w_1, x_1, y_1), \ldots, (w_n, x_n, y_n)$.

**1.** Suppose it is desired to find constants $a$ and $b$ such that we may express each $y_i$ via

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

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

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

with respect to $a$ and $b$. It suffices to solve the pair of equations

$$\frac{\partial S_1}{\partial a} = 0 \qquad (1)$$

$$\frac{\partial S_1}{\partial b} = 0 \qquad (2)$$

so long as a unique solution exists. Write $\Sigma_{wx} = \sum_{i=1}^{n} w_i x_i$, $\Sigma_{xx} = \sum_{i=1}^{n} x_i^2$, etc..

(a) Determine both partial derivatives and write the equations (1) and (2) in matrix form, i.e.

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

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

(b) Write an inequality involving $\Sigma_{wx}$, $\Sigma_{xx}$, etc. which holds if and only if the determinant of **M** is positive.

(c) Assuming the determinant is positive, by inverting **M** solve the equations and express the least squares solutions $a$ and $b$ in terms of $\Sigma_{wx}$, $\Sigma_{xx}$, etc..

(d) Show that in the special case where $w_1 = w_2 = \cdots = w_n = 1$, the expressions for the solutions in the previous part reduce to

(i) $b = S_{xy}/S_{xx}$ and

*(ii) $a = \bar{y} - b\bar{x}$

where as usual $S_{xy} = \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})$, $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (**hint:** recall the computing formulae for $S_{xy}$ and $S_{xx}$).

**2.** Suppose now that we wish to include an intercept, that is choose constants $a$, $b$ and $c$ so that with

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

the resultant $\varepsilon_i$'s resemble "random errors". This may be desirable if the "no intercept" version in question 1 did not give a good fit. It is proposed to again use the method of least squares, that is to choose $a$, $b$ and $c$ minimising

$$S_2(a, b, c) = \sum_{i=1}^{n} \left[y_i - (aw_i + bx_i + c)\right]^2 \, .$$

(a) Use calculus to show that the value of $c$ which minimises $S_2(a, b, c)$ when $a$ and $b$ are held fixed is $c = \bar{y} - a\bar{w} - b\bar{x}$ and thus that this problem reduces to minimising

$$S_3(a, b) = \sum_{i=1}^{n} \left\{(y_i - \bar{y}) - [a(w_i - \bar{w}) + b(x_i - \bar{x})]\right\}^2$$

over $a$ and $b$ which is mathematically equivalent to the problem in question 1.

*(b) The solution to this last minimisation problem is the same as that for question 1 after replacing $\Sigma_{wx}$, $\Sigma_{xx}$, etc. with $S_{wx}$, $S_{xx}$, etc. respectively ($S_{wx}$, $S_{ww}$ etc. are defined in the same way as $S_{xy}$, $S_{xx}$). A unique solution exists if and only if the inequality analogous to that derived in question 1 part (b) above holds.

Derive **three equivalent conditions**:

- one only involving $s_w$, the standard deviation of the $w_i$'s,
- one only involving $s_x$, the standard deviation of the $x_i$'s and
- one only involving $r_{wx}$, the correlation between them

which all hold if and only if a unique solution exists.