Name:
Date:

Regression Examples

- The regression line is just a line of best fit through the middle of the data. Since it is a straight line, it has the form
$$\hat{y} = a + b\,\hat{x}$$
where a is the y-intercept and b is the gradient of the line.
- We use $\hat{y}$ and $\hat{x}$ to say that these are *predicted* data points. The *actual* data points (x_i, y_i) may not lie on the line at all, but they should be close to the line.
- The *residuals* are the distance bewteen a real y_i and a predicted $\hat{y}$ - they tell us how close the line is to the real data. The residual for the i^{th} data point is given by
$$r_i = y_i - \hat{y}$$
- When asked to interpret b, you should always refer to what happends when $\hat{x}$ increases by 1 unit; when asked to interpret a you should explain what happens when $\hat{x} = 0$. You should always use the language given in the question - don't just say x and y.
- It is always true that
$$\sum r_i = 0$$
meaning that the sum of the residuals is zero.

Example (Calculating a regression line) The following data give the fastest running times (x) for the olympics women's 100m sprint over the past 10 years (y):

x	1996	2000	2004	2008	2012
y	11.1	10.94	10.93	10.78	10.75

Calculate the equation of the least squares regression line of y on x

Example (Interpreting the gradient) 1. For the regression line you have calculated above, give an interpretation of value of b within the context of the question.

2. It is calculated that, for men, the regression line has equation $\hat{y} = 49.84 - 0.02\hat{x}$. Make a comparison between the value of b for men and for women.

3. Interpret your comparison within the context of the question.

Example (Extrapolation and Interpolation) 1. For the women's regression line you have calculated above, find an estimate for the winning running time if the competition were held in 2002

2. Also find an estimate for the winning time in the year 2400

3. Which of the two estimates above is more reliable than the other? Explain your answer.

4. If the linear trend continues, in what year should we expect to see a running time of 10 seconds for women? Comment on the validity of your answer.

Example (Residuals) The following data shows the number of people who die by falling down the stairs (x) against sales of the iPhone in millions (y):

x	1927	1935	1960	1991
y	1.39	11.63	20.73	39.99

1. Calculate the equation of the least squares regression line
2. Interpret your values for a and b in the context of the question
3. Assuming that 1980 people fall down the stairs in the next year, find an estimate for the number of iPhone sales
4. Calculate the residuals r_3 and r_4
5. By considering the average of these residuals, improve the estimate you gave in part 3.
6. Explain why the average of all the r_i has not been used to improve the estimate in part 3.

Example (Exam style question) In an experiment, it is found that the computing speed y GHz is related to the temperature x °C of the processor of a computer by the regression formula

$$\hat{y} = 4.2 - 0.023\hat{x}$$

1. Interpret the regression line in the context of the data
2. Estimate the computing speed of a computer processor at 27°C
3. Given that the greatest temperature a computer was tested at was 25°C, comment of the validity of the estimate obatined above.
4. The table below represents **some** of the data collected.

x	5	10	15	20	25
y	4.1	3.9	3.86	3.72	3.6

   Calculate the mean of the residuals from this data.
5. Use this to improve the estimate obtained in question 2.
6. Given that there was only one other computer which was tested, and that this computer was tested at 0°C, find the real processing speed of a computer at freezing.