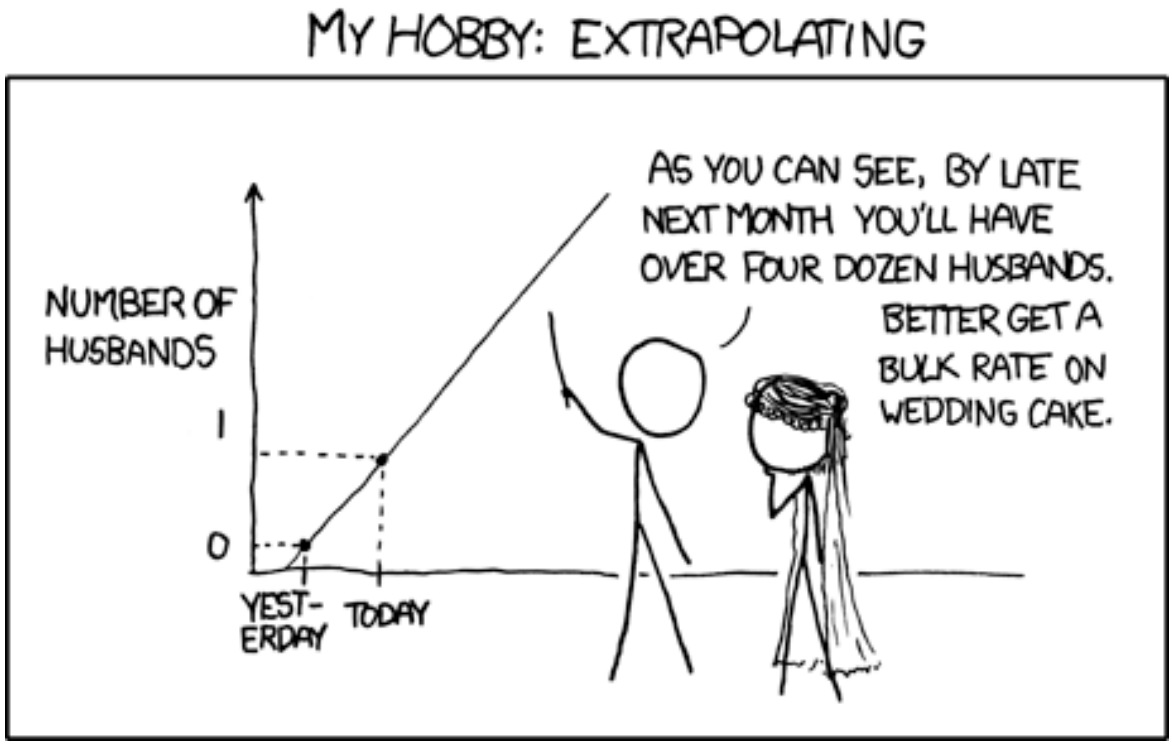
Problem sheet 3

Useful regressions



A couple is planning a road trip for their honeymoon. They want to drive all over Ireland in a camping car. They are planning a 1000km round trip and they wonder how much they should allocate for the fuel. The couple thinks there must be a way to estimate the amount of money needed, based on the distance they are going to travel. What they want to figure out is "If we drive for 1000km, how much will we pay for gas?"

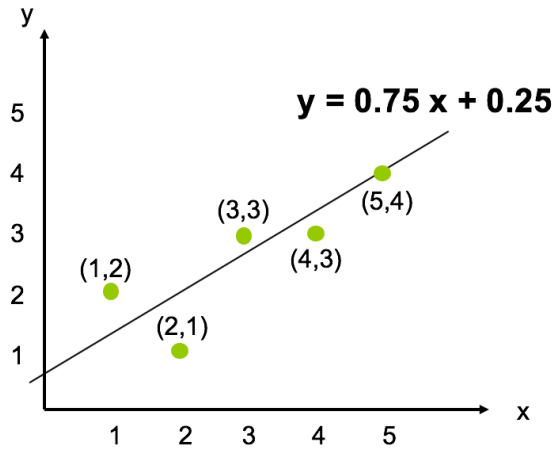
Fortunately, they have been laboriously tracking their car's efficiency for the last year and have made the following graph which seems to show a connection between km driven and fuel paid. They now want to use the data they have been collecting so far and use it to **predict** how much they are going to spend. The idea is that you can make estimated guesses about the future based on data from the past—the data points they have been laboriously logging. To do that they need a mathematical **model** that describes the relationship between km driven and money spent to fill the tank.

Concretely By using linear regression on their log, they can fit a line and find its equation. From this equation they can then predict how much they will pay for gas for their 1000km round trip. Tada!

Questions

Four questions, each worth two marks with two marks for attendance.

Let's try to compute some goodness of fit but I have used some simpler data than our newly wed example to make the computation faster on paper. Below is a graph representing five observations and a regression line.



Q1. Using the equation of the line, compute the goodness of fit using the standard error of the estimate. You can do this by filling in the following table.

x	Actual y	Estimated $\hat{y}$	$\hat{y}$ - y	$(\hat{y} - y)^2$
1	2	1	-1	1
2	1	1.75	0.75	0.5625
3	3	2.5	0.5	0.25
4	3	3.25	0.25	0.0625
5	4	4	0	0

square root $(\sum (\hat{y} - y)^2 / (N-2))$ = square root (1.875 / (5-2)) = square root (0.625)

Q2. Using the equation of the line, compute the goodness of fit using the R square. The mean of y is (2+1+3+3+4)/5= 2.6. You can do this by filling in the following table.

x	Actual y	y – mean(y)	$(y - \text{mean}(y))^2$	Estimated $\hat{y}$	$\hat{y}$ - mean(y)	$(\hat{y} - \text{mean}(y))^2$
1	2	-0.6	0.36	1	-1.6	2.56
2	1	-1.6	2.56	1.75	-0.85	0.7225
3	3	0.4	0.16	2.5	-0.1	0.01
4	3	0.4	0.16	3.25	0.65	0.4225
5	4	1.4	1.96	4	1.4	1.96

$\sum (\hat{y} - \text{mean}(y))^2 = 5.675$
$\sum (y - \text{mean}(y))^2 = 5.2$
$R^2 = 5.675 / 5.2 = 1.09$

Note that having a R > 1 is not usual and is due to the fact that I did not use what is called the "least square regression" method to find the best line. I just randomly fitted a line. In this case we could have used another more general method for computing R square which is given by a slightly different formula (see below).

Rsquared 1 (seen in class):

$$\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$:

Rsquared 2 (more generic equation):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \in [0,1]$$

If we use the second equation, we have:

x	Actual y	y – mean(y)	(y-mean(y)) ²	Estimated ŷ	y - ŷ	(y - ŷ) ²
1	2	-0.6	0.36	1	1	1
2	1	-1.6	2.56	1.75	-0.75	0.5625
3	3	0.4	0.16	2.5	0.5	0.25
4	3	0.4	0.16	3.25	-0.25	0.0625
5	4	1.4	1.96	4	0	0

∑ (y - ŷ)² = 1.875
∑ (y-mean(y)) ² = 5.2
R² = 1 – (1.875/ 5.2) = 1 - 0.36 = 0.64 (64% fit!)

Useful

Here are two reminders to compute the goodness of fit.