**MCQ4 – Model Answer**

1. With respect to the well-established relationship between atmospheric CO2 and global temperature which variable is which? *CO2 is the explanatory and temperature is the response. Or, alternatively, CO2 is the independent and temperature is the dependent (this was not an option!). Personally I prefer the terms explanatory and response as they have more intuitive meanings, but independent and dependent are just as valid terms.*
2. What is the term for using the equation describing the relationship between two variables to predict unsampled values outside of the range of sampled values? *This is extrapolation. If we wanted values inside the sampled range this would be interpolation. Regression is how we derive the equation in the first place and correlation does not really allow us to predict anything.*
3. In determining the intercept during linear regression which point can we use that we know will always fall upon the least squares regression line? *The point whose coordinates are the mean of x and the mean of y. This is an extension of the assumption that both x and y are normally distributed and so as two univariate data sets their middle value would be accurately captured by their means. In a bivariate setting we are just combining these two middle values together to get the middle of the bivariate data. The other answers are all wrong as we cannot guarantee that any other value of x or y falls exactly on the regression line. In other words, regression (like many other types of statistics) only works by assuming the data are normally distributed.*
4. What range of values does Pearson's correlation coefficient (r) span? *Negative one to one. I.e., a perfect negative relationship between x and y through to a perfect positive relationship between x and y. Probabilities and r-squared values are examples of a zero to one scale. Negative infinity to infinity is the full range of any normal distribution and zero to infinity is the range of some other probability distributions we will encounter later in the course.*
5. What relationship do the variables "Number of pirates (approximate)" and "Average global temperature" have according to pastafarians? *A strong negative relationship. We found this in lecture by performing a correlation test on their “data”. I.e., as pirates have decreased in number global temperature has risen. We should naturally be suspicious of this data (amongst other reasons) as it is an example of “temporal autocorrelation”. I.e., a lot of data sets when plotted against time have some form of trend and thus combining any two such data sets together will tend to lead to finding a strong correlation even if this doesn’t exist in reality (e.g., go explore* [*http://www.tylervigen.com/spurious-correlations*](http://www.tylervigen.com/spurious-correlations)*). As geologists we deal with time a lot (e.g., the dinosaur increases and decreases data from Lecture 1 shows an overall trend of increase) and so this is something we should always be wary of.*
6. In Anscombe's Quartet II what relationship do x and y have that makes correlation inappropriate? *A clear curvilinear (polynomial) relationship. The other answers refer to different quartets, but the key point here is that although software such as R will give you an answer if you ask for an estimate of r if you have plotted your data and found a non-linear relationship you should know that this estimate is not meaningful (i.e., Lesson #1 and Lesson #2).*
7. In Anscombe's Quartet IV what effect does removing the outlier have on the slope estimate? *It becomes NA as it is not possible to calculate. This is because the answer would logically be a vertical line (slope of either infinity or negative infinity) and so R doesn’t know what to do with this data. Again, the major issue here is a poorly designed study as x needs to be varied to understand it’s effect on y.*
8. What would an estimate for r of -0.97 indicate? *That x and y are nearly perfectly negatively correlated (i.e., it is very close to minus one). It could be a calculation error, but we cannot know this without further information. I.e., if it were less than -1 or more than 1 we would know straight away that it was a calculation error without needing additional information.*
9. For your own ammonite data what range of values did you find for the slope of your regression at a confidence level of 100%? *Negative infinity to infinity. Again, this reveals that although we are doing bivariate statistics here, univariate assumptions of normally distributed data are “lurking” underneath. I.e., at 100% we are effectively asking for the full range of possible values, which under a normal distribution is the limits: negative infinity to infinity.*
10. How does the coefficient of determination relate to Pearson's correlation coefficient? *The coefficient of determination (r-squared) is the square of Pearson's correlation coefficient (r), but only if the regression line is straight. However, it is better to use the r-squared output directly from the model (rather than squaring the correlation coefficient) in practice as this approach will always be appropriate even if looking at a curvilinear regression as we did with Quartet II. More broadly this is another example of terrible terminology in statistics as r-squared is (not always) the square of r.*