# Multiple Linear Regression

## Worksheet (with answers)

Part of the resource: <https://www.ncrm.ac.uk/resources/online/all/?id=20848>

1. **Using the GHS3.SAV dataset and doing any necessary exploratory analysis first, run a regression predicting number of alcohol units consumed per week using hours of work and age.**

What are the values of the regression coefficients:

1. the constant? (intercept/cutting point)
2. The coefficient for hours of work?
3. The coefficient for age?

Are these all significantly different from zero?

How do you interpret the regression coefficients in this regression?

**2. After recoding the Sex variable appropriately, add it to the model.**

Does this variable improve the model?

 What criteria did you use to draw this conclusion?

1. **Remove Sex and add another variable of your choice.**

Does this variable improve the model?

 What criteria did you use to draw this conclusion?

**4. Create an index of deprivation using the edlev2 and gincome. (it will be necessary to apply filters and recodes to make this work). *You will (probably) need to watch the lecture 2 podcast before you can answer this question.***

 What filters did you apply?

 What recodes did you do?

 What is the correlation of the deprivation index with number of hours worked?

***Answers:***

**1. Using the GHS3.SAV dataset and doing any necessary exploratory analysis first, run a regression predicting number of alcohol units consumed per week using hours of work and age.**

*I did some exploratory analysis and found that one individual had a consumption of 438 units per week! This may be an error and is certainly an extreme outlier. I elected to filter this person out before running the model. I also found one case that has a coding of -8 for work hours. This coding is unusual in the dataset and derives from the person been recorded as working but unpaid. -8 should be coded as missing in the dataset so I therefore set -8 to missing.*

What are the regression coefficients?

What is the constant? (intercept/cutting point) *6.572*

The coefficient for hours of work? *0.245*

The coefficient for age? *-0.084*

Are these all significantly different from zero? *Age is borderline (p=0.083) however I will still leave it in the model, the others are significantly different from 0 (p<0.05).*

How do you interpret the regression coefficients in this regression?

*If the number of hours is held constant then we would predict that an individual will drink 0.084 units less for each year older s/he is. If whether age is held constant then we expect them to drink 0.244 units for each hour worked. The model is:*

*Estimated units = 6.572+ 0.245\*hours worked -0.084(age).*

*Using the standardised coefficients I determine that hours worked is over 3 times as important as age in accounting for variability in units drunk.*

**2. After recoding the Sex variable appropriately, add it to the model.**

*I produced a new variable called “male” which was coded as 1 if the individual is male and 0 if they are female – this is a true binary.*

Does this variable improve the model? *yes*

What criteria did you use to draw this conclusion? *The adjusted R-squared value has increased from 0.046 to 0.084. All p-values are now significant.* 2

**3. Remove Sex and add a another variable of your choice.**

*I added “How long a person has been with current employer”*

Does this variable improve the model? *no the model is worse than that in q2*

What criteria did you use to draw this conclusion? *The adjusted R-squared value has decreased to 0.045. “How long a person has been with current employer” is non significant p=0.687.*

1. **Create an index of deprivation using the edlev2 and gincome. (It will be necessary to apply filters and recodes to make this work).**

What filters did you apply?

*I removed the foreign other.*

What recodes did you do?

*I recoded Edlev2 to so that none became category 6 (there are no instances of the actual category 6 in the data set and we have filtered out the category 7. This new variable has a median of 4 and an inter-quartile range of 4 (I get this by opening the frequencies dialog box and clicking on the stats button). I will need these as I will have to calculate the standardised scores for this variable using a compute as the automatic procedure in SPSS now assumes the variable is continuous.*

*The Index is created by subtracting the standardised income from the standardised educational level variable (so as the variables are working in the same direction). A positive score on the index indicates deprivation a negative score the lack of deprivation.*

What is the correlation of the deprivation index with number of hours worked?

*Spearman’s correlation is -0.448. I have used spearman’s because the combination of the number of hours worked being a borderline case and the fact that the index has been constructed using one ordinal and one interval variable means that I am wary of using a parametric measure such as Pearson’s.*

National Centre for Research Methods (NCRM)
Social Sciences
Murray Building (Bldg 58)
University of Southampton
Southampton SO17 1BJ
United Kingdom

**Web** www.ncrm.ac.uk
**Email** info@ncrm.ac.uk
**Tel** +44 23 8059 4539
**Twitter** @NCRMUK