# Advanced Quantitative Methods in the Social Sciences and Multivariate Statistical Methods in Education

# Assignment

This assignment consists of three exercises. You must answer all three exercises. Each exercise carries equal weight in terms of assessment. Each exercise includes information about a research study, an SPSS dataset or output, and a series of questions for you to answer. The word limit is 4000 words including all tables and figures and any references (we will ignore any words over this word limit).

You will be assessed according to the criteria in the M-level student handbook. You will therefore need to demonstrate Knowledge and Understanding (ability to conduct the taught statistical methods successfully), Approach/Analysis (ability to explain the use of particular statistical methods and make appropriate interpretations of findings), Organisation and Structure (ability to structure your arguments and explanations coherently and logically), and Style and Presentation (ability to report findings according to academic conventions). You will not be assessed on Use of Sources as everything you need to complete the assignment can be found in the course materials (lecture slides and worksheets) and this requires no citation. Similarly, we do not expect you to do any additional research on the substantive topics addressed by the research studies and no marks will be given for any such additional research.

Tips:

* There is no need to begin your answers by excessively repeating the exercise information or typing out the questions in your answers;
* Structure your answer according to the exercise and question numbers and letters;
* Write in a clear concise academic style avoiding repetition and redundancy. Write in an engaging fashion, adding value to the reported statistics by carefully interpreting them and by adding substantive insights and giving intuition for the patterns you find;
* Do not paste SPSS output other than graphs directly into your assignment. Rather,

be selective in which results you choose to show and present these results as tables or inline following academic conventions;

* Where you present calculations, show your workings;
* Do not include references (see above).

# Question 1 ANOVA and Factor analysis,

# (weeks 1-3)

A researcher is interested in the stereotypes that British people hold about migrants to the UK. In particular, she wants to know if stereotypes apply to all migrants in general, or whether groups from different countries or regions are perceived differently by members of the host country. To investigate, she devises a 9-item scale that asks participants to rate society’s perceptions of a particular group of migrants in terms of 9 different attributes (see **Stereotypes scale.pdf** for details). Each item is scored on a Likert scale from 1 (not at all) to 5 (extremely) and the responses are summed to create a single composite stereotype scale with a possible minimum of 9 and a maximum of 45. Higher scores on the scale indicate more *positive* perceptions of the migrant group.

In her experiment she recruits 95 participants at a bus station in a medium-sized British city. Only participants who confirmed they were British citizens were eligible to participate. Each participant was asked to complete the stereotypes scale *four times*, each time considering a different migrant group: people from Eastern Europe; East Asia; South Asia; and Africa.

The original data, containing participants’ total stereotypes scores for each of the four migrant groups, can be found in the file **stereotypes\_ex1.sav**.

1. Produce some descriptive statistics and/or charts to explore the data in **stereotypes\_ex1.sav** and comment on what these suggest about the research questions of interest. What kind of ANOVA test could be used to analyse these data?
2. Conduct the ANOVA test selected in part a). Are there significant differences in stereotypes about different groups of migrants? If so, how do perceptions of the different groups compare? (REPEATED MEASURED ANOVA)
3. Suppose that instead of asking a single group of 95 participants to rate perceptions of all four groups of migrants, the researcher had recruited a larger sample of 380 participants, but asked each participant to complete the scale for only *one* migrant group (i.e. randomly allocating each participant to rate only one of the four possible groups). Discuss the implications of the differences between the two research designs. The data in file **stereotypes\_ex2.sav** contain the same scores as in the original experiment, but now formatted as if they were collected from four independent groups of participants. Analyse these data and contrast the results with those from part b) to help illustrate your answer. (THINK ABOUT DIFFERENT WAYS TO COLLECT) TWO DIFFERENT ANOVAS
4. The researcher reports measures of the reliability of the scales in the original experiment for the four migrant groups as follows: Cronbach’s alpha for Eastern Europeans = 0.82; for East Asians = 0.57; for South Asians = 0.92; and for Africans = 0.72. Comment on these statistics. MAYBE A BAR CHART, for the spread of the data. DON’T HAVE TO DO THE FACTOR ANALYSIS, IT IS PROVIDED. NINE ITEMS, DISCUSSION

Another researcher is given access to the underlying item data that was summed to create stereotype scores for the East European group and uses these data to conduct a factor analysis. The results can be found in the SPSS Viewer document **item\_analysis.spv**. What are the implications of this analysis for the original study? What recommendations for future research would you make on the basis of these results?

.

# Question 2 Linear regression

# (WEEKS 4-6, SUMMARIES, BAR CHARTS)

# The data considered here are from a study of gender differences in salaries among full-time permanent academics at the University of Rummidge in the academic year 2006-2007. Four-hundred-and-ninety-seven academics (18% of those approached) agreed for their anonymised University administrative records to be shared with the researchers.

The variables in the IBM SPSS data document **salary.sav** that we will use are the following:

# salary: annual salary (in British pounds)

* **gender**: gender (1: male; 2: female)
* **yearsphd**: time since PhD (in years), a measure of academic age and experience
* **market**: marketability of academic’s discipline, defined as the ratio of the UK national average academic salary paid in the academic’s discipline to the national average across all disciplines. This variable is meant to capture the notion that some disciplines (e.g., finance, engineering) have to pay higher salaries than others (e.g., sociology, history) to prevent their academics from seeking more lucrative employment opportunities outside of academia.
* **ethgroup**: ethnicity (1: white; 2: black; 3 Asian; 4: other)

Answer the following:

# Report and interpret univariate and bivariate descriptive statistics and distributional plots for the variables in the data. SUMMARISE VARIABLES, WALK YOU THROUGH AN ANALYSIS.

# Fit a linear regression of academic salary on a dummy variable for being a female academic. Present the results of this and subsequent models in a single table as you might see in a journal article. Fully interpret the results of this first model. Do you consider the gender difference in mean salaries to be large? Can we interpret this gender difference as evidence of gender discrimination? To what extent do you think you can generalize your results to the University of Rummidge as a whole? How about to other UK universities? WHY THE DIFFERENCES IN SALARIES

# Add ‘time since PhD’ to the model. Fully interpret your results. Explain why the gender gap has reduced and back up your explanation with appropriate descriptive statistics.

# Add marketability to the model. Explain why we should be somewhat careful when interpreting the estimated intercept and coefficient of marketability in this model. Explain any further change you see in the gender gap. Which of the two continuous explanatory variables appears to be the more important predictor: time since PhD or marketability?

# Add ethnicity to the model as a series of dummy variables, choosing white academics as the reference category. To what extent do the effects of gender, time since PhD and marketability change when we adjust for ethnicity? Can you explain why? Having adjusted for the other covariates, does ethnicity significantly predict salaries? Interpret the estimated coefficients of the ethnicity dummy variables. Report the predicted salary differences between the three minority ethnic groups (i.e., Asian vs. black; Other vs. Black; Other vs. Asian) and indicate whether each is statistically significant or not. Report the predicted salary for a female Asian academic working in a discipline with marketability score 1.20 who finished their PhD 15 years ago. Show your calculation.

# FOURTH VARIABLE, ethnicity, dummy variable

# Question 3 Linear regression

# (WEEKS 4,5,6, some of 7,8 a bit more open ended, try different things)

The topic in this question is predictors of a child’s birth weight, and in particular the effect of maternal smoking during pregnancy on birth weight. Birth weight is commonly used as an indicator of child’s health at birth and heavier birth weight has been shown to predict better health, social and economic outcomes later on in life.

The dataset **birthweight.sav** contains the following variables. The variable **bweight** is the outcome variable of interest and the variable **pregsmk** is the initial predictor of interest, though you may choose to focus on other predictors in addition to **pregsmk** in your answer.

|  |  |
| --- | --- |
| **Variable** | **Label** |
| id | Child ID |
| bweight | Child's birth weight in grammes |
| pregsmk | Mother smoked in pregnancy (0 = no; 1 = yes) |
| nbirths | Number of previous births of mother |
| female | Child is female |
| gestation | Gestation in days (duration of pregnancy) |
| motherage | Mother's age at child's birth |
| mumed | Mother's highest education qualification |
| nhooddep | Neighbourhood deprivation score (10=most deprived) |
| healthprob | Mother had health problem in pregnancy |
| planned | Pregnancy was planned |
| bfeed | Mother breastfed baby |
| mumheight | Mother's height in cm |
| mumweight | Mother's weight before pregnancy in kg |

Your answer should be written up in the form of a results section (with additional introduction and discussion) from an academic article, with sections for: an introduction; descriptive statistics; inferential statistics; and some discussion/conclusions. Your work should include consideration of the following issues:

1. What is the estimated effect of smoking during pregnancy on birth weight from a simple (single predictor) linear regression? How do you interpret the unstandardized coefficient on the smoking in pregnancy variable? Why might this not give a good estimate of the *causal* effect of smoking in pregnancy on birth weight?
2. Of the other variables available to you, which do you think might be strongest predictors of birth weight? Which might it be most important to control for when trying to isolate the effect of smoking in pregnancy? Do your regression results support your initial hypotheses?
3. Some research has suggested that the effect of smoking is more harmful for second- or higher order-births and less harmful for first-born children. Is this hypothesis supported in your data? **ABOUT A PARTICULAR VARIABLE, FOCUS ON INTERACTION. Add an interaction, requires one**
4. Previous evidence indicates that both younger, and older, mothers tend to have babies of lower birthweight compared to mothers of mid-childbearing age. Is there any evidence to support such a non-linear relationship in this dataset? **Model H and H SQUARED, quadratic relation**
5. Are there any variables not available here that you would have liked to have included in your regression models? Please explain your reasons. **DISCUSSION PART,**

*Please note: you are not expected to be an expert on the determinants of birth weight! It doesn’t matter if your hypotheses turn out to be incorrect. Marks will be allocated for how well your hypotheses, descriptive statistics, regression results and interpretation join together to make a coherent argument. Note also that you do not necessarily need to make use of every variable in the dataset or explore a large number of different relationships. Answers that provide a lot of statistics without demonstrating understanding of what they mean will likely score lower than more focused analyses with strong discussion and interpretation.*

***Check Students piece of work example.***

**Don’t copy and paste spss output**

**Structure part a, part b, part c.**

**Make sure to use the slides in the worksheets.**

**Try to do a parallel interpretation of the worksheet.**