The main body of the report should include descriptive and/or test statistics, *p*-values where appropriate, your conclusions and full interpretation of the results.

**Task 1**

The data for this task are stored in the SPSS file **Armenia birth weights.sav**. This data file contains information for a sample of 1,148 births in Armenia between 2006 and 2010. The data were obtained from a retrospective survey of mothers aged 15-49 years which was undertaken in 2010. They give the birth weight in grammes of the most recently born child for each mother.

The data also include a set of other variables about either the child whose birth weight has been measured, or the mother of that child: the age of the mother at the time of the birth, whether the mother lived in an urban or a rural district, the highest educational level the mother had attained, the number of years of education the mother had received, a categorical variable denoting the wealth stratum into which the mother’s household fell, the birth order of the child (i.e. whether the child was the mother’s first, second, third, fourth, etc. child), and whether or not the child was a twin,

The table below gives specific details for each variable.

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Description** | **Coding** |
| **Age** | Age of the mother in years at the time of the birth | Continuous variable |
| **Urban** | Whether the mother lived in an urban or a rural district | 0 = ‘Rural’, 1 = ‘Urban’ |
| **Education** | Highest educational level attained by the mother | 1 = ‘Primary’, 2 = ‘Secondary’ 3 = ‘Higher’ |
| **EducYears** | Number of years of education received by the mother | Continuous variable |
| **Wealth** | Wealth stratum of the mother’s household | 1 = ‘Poorest’, 2 = ‘Poorer’, 3 = ‘Middle, 4 = ‘Richer’, 5 = ‘Richest’ |
| **Order** | Birth order of the child | Continuous variable |
| **Twin** | Whether or not the child was a twin | 0 = ‘Not a twin’, 1 = ‘Twin’ |
| **BirthWt** | Birth weight of child in grammes | Continuous variable |

1. (a) Calculate the sample proportion of births to women living in ‘Urban’ districts.

(b) Give a 95% confidence interval for the population proportion of births to women in ‘Urban’ districts, and interpret this interval.

(c) Give a 99% confidence interval for the population proportion of births to women in ‘Urban’ districts. Is the 99% confidence interval wider or narrower than the 95% confidence interval? Why?

2. Recode the variable **Age** into another variable which has seven categories: ‘15-19 years’, ‘20-24 years’, ‘25-29 years’, ‘30-34 years’, ‘35-39 years’, ‘40-44 years’ and ‘45-49 years’. Make sure that your recoded variable is *clearly labelled* and named **Age\_cat**. Produce a frequency tabulation of your recoded variable and state what proportion of births occurred to women aged 20-29 years.

3. Produce suitable descriptive statistics and graphs to compare the birth weights for women living in ‘Urban’ and ‘Rural’ areas. Comment on any differences/similarities between the distributions of birth weights of the two groups. [Note: Do *not* carry out any formal hypothesis tests.]

4. Using the variables **Wealth** and **Urban**, carry out a suitable hypothesis test to determine whether there is an association between household wealth and whether the household was in an ‘Urban’ or a ‘Rural’ area. State your hypotheses clearly and the value of the test statistic, and give your conclusions based on the result of the test. Comment on any patterns you find.

5. A paediatrician claims that the mean birth weight of children born in ‘Rural’ districts in Armenia between 2006 and 2010 was 3,100 grammes. Carry out an appropriate test at the 5% significance level to find out whether this claim is valid. State your hypotheses clearly and the value of the test statistic, and give your conclusions based on the result of the test.

6. Test at the 5% level of significance whether the mean birth weight was different for births in ‘Urban’ and ‘Rural’ areas. Make sure you state the null and alternative hypotheses and the value of the test statistic, and give your conclusions based on the result of the test.

7. Use suitable graphs and calculate simple statistics to show the association between the birth weight and the mother’s years of education. Comment on the association between these two variables.

**Task 2**

1. Fit a simple linear regression model in which the birth weight (**BirthWt**) is the dependent variable and the mother’s years of education (**EducYears**) is the independent variable.

a) Write down the fitted regression equation and interpret the estimated regression coefficients.

b) Carry out a formal hypothesis test to assess whether mother’s years of education is a significant explanatory variable. Make sure you specify your hypotheses, give the test statistic value, and state your conclusion. Why is a significance test like this needed?

c) Given your regression equation, what would you predict the birth weight to be of a child whose mother had eight years of education?

d) Give a 99% confidence interval for the population regression coefficient β, the ‘true’ effect of mother’s years of education on birth weight, and interpret this.

e) Check whether the regression assumptions hold (at least approximately) by producing some suitable residual plots. Discuss your results.

2. Run a regression model with birth weight (**BirthWt**)as the dependent variable. Call this Model 1. Use as independent varlables the mother’s age in years (**Age**), the mother’s education in years (**EducYears**), whether the birth was a twin (**Twin**), whether the mother lived in an ‘Urban’ or a ‘Rural’ area (**Urban**), and the wealth of the household (**Wealth**).

Then run a separate regression model, with the same variables as above but adding birth order (**Order**) as an extra independent variable. Call this Model 2.

(a) Present the results of both regression models in *one single table* that (i) includes all the important information for interpreting the results, (ii) is clear, and (iii) makes it possible to compare regression coefficients for both models easily. This table should be constructed ‘by hand’, i.e. do NOT cut and paste from SPSS output.

(b) Using the results from Model 2 only, interpret the regression results by discussing the estimated associations of *all* the explanatory variables with birth weight.

(c) Using Model 2 only, write down the fitted regression equation for mothers living in ‘Urban’ areas.

(d) Using your equation in (c), predict the expected birth weight for (i) the first child of an ‘Urban’ woman in the ‘Middle’ wealth stratum who was aged 23 years and who had 8 years of education and (ii) the fourth child of a ‘Rural’ woman in the ‘Poorest’ wealth stratum who was aged 35 years and who had 4 years of education. Assume that neither of the births was a twin.

(e) Which do you prefer between Model 1 and Model 2? Why?

(f) Someone suggests that it is not so much years of education that matters for birth weight, as the highest educational level the mother has attained. Test this suggesting by running Model 2 again, replacing **EducYears** by highest educational level. To do this you will need to create two ‘dummy’ variables from the variable **Education**: one takes the value 1 if the mother only has ‘Primary’ education and 0 otherwise, and a second takes the value 1 if the mother has ‘Higher’ education and 0 otherwise. Call these **Primary** and **Higher** respectively. Then replace the variable **EducYears** in Model 2 by the two variables **Primary** and **Higher**. Comment on whether the suggestion is correct.

(g) Does removing **Age** from Model 2 make the model any worse? Explain your answer.