**37252 Sample Exam**

**Question 1. (20 marks)**

The analysis in this question is based on health data collected in 1950 from a sample of US adults, with follow-up data collected in 1962. There are two variables; the dependent variable is the systolic blood pressure (SBP) in 1962 of the respondents, while the independent variable is their SBP in 1950. As an initial analysis a scatterplot has been produced to look at the relationship between the two variables.



1. Based on the scatterplot, comment on the type, direction and strength of any relationship between the two variables.

(3 marks)

Following the scatterplot, a simple linear regression has been undertaken, which has produced the following output:







1. Using the ANOVA Table, test whether the model is significantly better than a model with the intercept only. State clearly the null and alternative hypotheses of the test; and your conclusion based on the appropriate p-value.

(3 marks)

1. Write-down the fitted regression model for the relationship between SBP in1962 and SBP in1950. Interpret the value of the estimate for $β\_{1}$ (the slope) for the regression line.

(2 marks)

1. Find the 90% confidence interval for the intercept parameter $β\_{0}$ and state clearly what a confidence interval means.

(3 marks)

As part of the analysis, SPSS has also produced a normal P-P plot of the residuals and a scatterplot of the (internally) studentized residuals. These plots can be seen on the next page.



1. Discuss whether the two plots of the residuals support the assumptions of the regression model being satisfied.

(3 marks)

1. A colleague suggests transforming the data and fitting a model of the form

$$\frac{y\_{i}}{\sqrt{x\_{i}}}=\frac{β\_{0}}{\sqrt{x\_{i}}}+β\_{1}\sqrt{x\_{i}}+\frac{ε\_{i}}{\sqrt{x\_{i}}}.$$

Explain why this may help improve the fit of the model with respect to the assumptions of the regression model.

(2 marks)

The model suggested in part f) was fitted to the data by transforming the variables. This resulted in the following output for the transformed model.



1. Interpret the fitted model parameters with respect to the original variables of SBP in 1962 and SBP in 1950. Comment on whether the model fits better in terms of how well the parameters are estimated. (*Hint: compare the standard errors for the model parameters from the two fitted models*.)

(4 marks)

(End of Question 1)

**Question 2. (20 marks)**

The analysis in this question uses multiple linear regression to explore the relationship between a country’s Gross National Income per capita (GNI) and two measures of the education system; mean years of schooling and expected years of schooling.

Below is a matrix scatterplot looking at relationships between the three variables.



1. Discuss the strength and direction of the relationship between the dependent variable GNI and the two potential explanatory variables. State why using the log of GNI would be sensible when applying multiple linear regression.

(3 marks)

A multiple regression model is fitted with log10(GNI) as the dependent variable and including both explanatory variables. The output on the following page is created.







1. From the output, find and write-down the value of R-square and explain what this tells us in terms of how well the model fits.

(3 marks)

1. Write-down the fitted regression model.

(1 mark)

1. For the fitted model, interpret the impact of both mean years of schooling and expected years of schooling on a country’s GNI (*original scale*). Comment on the statistical significance of the parameters in relation to your interpretation.

(3 marks)

1. The output contains collinearity statistics. Explain why multi-collinearity is a problem for the interpretation of multiple regression models and comment on whether it is an issue in this particular model. You should refer back to an appropriate part of the matrix scatterplot.

(4 marks)

To extend the analysis, a categorical variable grouping countries into broad regions is included. The regions are; Europe, Americas, Oceania, Middle East, Asia, Africa. Choosing Europe as the reference group, the following SPSS output for the estimated model parameters is created.



1. Interpret the impact of each category of country group (*relative to Europe*) on a country’s GNI. Comment on the statistical significance of the parameters for each category in relation to your interpretation.

(4 marks)

A further extension has included the interaction between mean years of schooling and countries in the Middle East.



1. Test whether the interaction effect is significant stating clearly your hypotheses, test statistic, p-value, and conclusion.

(2 marks)

**Question 3. (20 marks)**



1. Using the percentages in the cross-tabulation, describe the relationship between gender and receiving low pay.
2. Using the percentages (or counts) calculate the odds of males receiving low pay and the odds for females. Hence calculate (and interpret) the odds ratio of receiving low pay for females relative to males.



1. Using the Chi-Square test of association, decide whether there is a significant association between gender and receiving low pay. Make sure you:
* state clearly your null and alternative hypothesis,
* explain the role of the expected counts in the calculation of the test statistic,
* state the value of your test statistic with its associated p-value, and your conclusion with respect to the hypotheses.
1. For the fitted model, show that the estimated odds of low pay from the model are given by

$$\hat{odds}=e^{(\hat{β}\_{0}+\hat{β}\_{1}×Sex)}$$

 and the fitted probabilities are given by

$$\hat{p}=\frac{e^{(\hat{β}\_{0}+\hat{β}\_{1}×Sex)}}{1+e^{(\hat{β}\_{0}+\hat{β}\_{1}×Sex)}}$$

where the dummy variable $Sex=0$ for males, $Sex=1$ for females and $\hat{β}\_{0}$ and $\hat{β}\_{1}$ are the estimates of $β\_{0}$ and $β\_{1}$.



1. From the output, write down the fitted model in terms of the Ln(odds) of receiving low pay. Does the output support a significant relationship between gender and receiving low pay? Make sure you justify your answer with a hypothesis test and associated p-value.
2. Using the relationships given in part d) (or otherwise), show that the binary logistic model gives fitted values for the odds for males and females receiving low pay that match those calculated in part b), and hence that the odds ratio for females relative to males is 2.312.

**Question 4.**



1. Using the appropriate information, discuss whether there is evidence that the overall model fits better than a model with just the intercept. State the value of the test statistic and the associated p-value that supports your answer.



1. Using odds ratios interpret the **linear relationship** between age and low pay. State whether the relationship is statistically significant based on appropriate p-value.
2. Using the odds ratio, interpret the relationship between gender and low pay.



1. Using odds ratios, interpret the relationship between ‘region of residence’ and low pay. State whether the two odds ratios are statistically significant based on appropriate p-values.



1. Explain why the output for the Hosmer and Lemeshow Test supports exploring interactions between the variables in the model.



1. Is the interaction term significant? State the value of the test statistic and the associated p-value that supports your answer.
2. For an individual aged 40 years, work-out the six fitted probabilities associated with combinations of gender and region of residence. (**Hint:** *Write-out the model for the six combinations and then use the formula in part d) of Question 4*.)