MATH20811 Practical Statistics: Coursework 3 (December 2022)

The marks awarded for this coursework constitute 40% of the total assessment for the module.

Your solution to the coursework should be fairly concise (maximum of about 12 pages) and it should take, on average, about 20 hours to complete.

Please read all the instructions and advice given below carefully.

The submission deadline is 10:00 am on Wednesday 4 January 2023.

Late Submission of Work: Any student's work that is submitted after the given deadline will be classed as late, unless an extension has already been agreed via mitigating circumstances or a DASS extension.

The following rules for the application of penalties for late submission are quoted from the University *Guidance on late submission document* (dated July 2021):

"*Any work submitted at any time within the first 24 hours following the published submission deadline will receive a penalty of 10% of the maximum amount of marks available. Any work submitted at any time between 24 hours and up to 48 hours late will receive a deduction of 20% of the marks available, and so on, at the rate of an additional 10% of available marks deducted per 24 hours, until the assignment is submitted or no marks remain.*"

Your submitted solutions should all be in one document which must be prepared using LaTeX. For each part of the project you should provide explanations as to how you completed what is required, show your workings and also comment on computational results, where applicable.

When you include a plot, be sure to give it a title and label the axes correctly.

When you have written or used `R` code to answer any of the parts, then you should list this `R` code after the particular written answer to which it applies. This may be the `R` code for a function you have written and/or code you have used to produce numerical results, plots and tables. `R` code should also be clearly annotated.

Do not use screenshots of `R` code/output in your report. Instead, to include `R` code use the *verbatim* environment and summarise `R` output in tables using the *table* environment, as demonstrated in the solution of Example Sheet 2.

Your file should be submitted through the Turnitin assessment in the coursework folder entitled "MATH20811 CW3" under Assessment & Feedback on Blackboard and by the above time and date. Work will be marked anonymously on Blackboard so please ensure that your filename is clear but that it does not contain your name and student id number. Similarly, do not include your name and id number in the document itself.

There is a basic LaTeX template file on Blackboard which you may choose to use for typing-up your solutions. The file is called `CW3_submitted_work.tex`.

Turnitin will generate a similarity report for your submitted document and indicate matches to other sources, including billions of internet documents (both live and archived), a subscription repository of periodicals, journals and publications, as well as submissions from other students. Please ensure that the document you upload represents your own work and is written in your own words. The Turnitin report will be available for you to see shortly after the due date.

This coursework should hopefully help to reinforce some of the methodology you have been studying, as well as the skills in `R` you have been developing in the module. Correct interpretation and meaningful discussion of the results (i.e. attempt to put the results into context) are important in order to achieve a high mark for the coursework.

The data for this work is related to auto insurance claims in Sweden over a particular period of time. The data was collected by the Swedish Committee on Analysis of Risk Premium in Motor Insurance. They are in the file `sweden_ins_data.txt` on Blackboard which contains two columns:

`claims` = number of claims - this will be the predictor / independent variable (or covariate) in the regression model. We can denote it by x.

`payment` = total payment for all the claims in thousands of Swedish Kronor for geographical zones in Sweden - this will be the dependent / response variable in the regression model. We denote this variable by y.

There are $n = 63$ observations in the dataset. The simple linear regression model for the data is given by:

$$y_i = \alpha + \beta x_i + \epsilon_i \quad i = 1, \ldots, n$$

where α and β are parameters with unknown values, and the random errors $\epsilon_1, \ldots, \epsilon_n$ are assumed to be independent $N(0, \sigma^2)$ random variables where the value of σ^2 is also unknown.

1. Produce a scatterplot of the data and comment on any evident features, and also the apparent suitability (or not) of the above regression model for the data. [3]

2. Write your own function in `R` to fit, using least squares, a simple linear regression model to a set of data comprising a response variable, y and a single covariate, x. Your function should have only two arguments - a vector of data for y and a vector of data for x.

   The output from your function should be in the form of a `list` object which comprises:

   - a vector containing the parameter estimates $\hat{\alpha}$ and $\hat{\beta}$;
   - a vector containing the fitted values;
   - a vector of the estimated errors (or residuals);
   - a scalar giving the value of the residual degrees of freedom.

   The four items in the output should all be calculated manually (DIY) within the function itself using the input data.

   To complete this part, run you function using the Swedish insurance data. [8]

   Please note that,

   - full marks are obtained in part 2 for writing a function which uses DIY calculations and can successfully output the four items listed in the question;
   - if you are unable to complete a function to do all the calculations then you may write code to calculate them items separately outside of a function, but there will be a mark penalty for this;

- all the items that your DIY regression function should output are available in an object that can be created by running the `lm` function. If this is your only recourse to obtaining the results specified in part 2, then you will only receive 1/8 marks for it. However, you will then have the necessary results available to do the subsequent parts of this coursework.

3. Using the output generated in part 2:

   (i) Report the estimated values of α and β that have been calculated. Superimpose the fitted regression line on to a scatterplot of the data and comment on the results. [3]

   (ii) Using an unbiased estimator, estimate the value of the error variance, σ^2. [2]

   (iii) Use an F-test to test $H_0 : E[Y|x] = \alpha$ vs $H_1 : E[Y|x] = \alpha + \beta x$ at a 5% significance level. Report your conclusions. [2]

   (iv) Calculate a 95% confidence interval for $E[Y \mid x = 80]$. [2]

   In the next parts we will use the residuals to examine the plausibility of the assumptions made about our regression model. It can be shown theoretically that the estimated errors from a fitted model (the residuals) have differing variances which depend on the values of the covariates. Consequently, we will work with the standardised residuals so that they are all on the same scale. To obtain the standardised residuals from the fitted model object created in `R` we can use the `rstandard` function which performs the standardisation in a special way that recognises the unequal variances.

4. Using diagnostic plots of the standardised residuals a) against the fitted values and b) against the predictor, make comments about the validity of the assumptions that were made regarding the model that has been fitted to the data. [4]

5. Manually construct (rather than using an existing `R` function for this purpose) a Normal quantile-quantile plot of the standardised residuals and superimpose a suitable reference line to help gauge Normality. Comment on the form of your plot and say whether you think that Normality was a tenable assumption or not. [3]

6. We now wish to carry out a Kolmogorov-Smirnov (KS) test to assess whether the distribution of the standardised residuals is $N(0, 1)$. Find the value of the KS test statistic and also that standardised residual value where the absolute difference between the empirical and $N(0, 1)$ cdfs is a maximum. [3]

7. Produce a plot containing the empirical cdf of the standardised residuals and the $N(0, 1)$ cdf and indicate on it the point at which the maximum difference between the curves occurs. [3]

8. Write a function in `R` to simulate the sampling distribution of the Kolmogorov-Smirnov test statistic when the $N(0, 1)$ null distribution is true using the same sample size as that of the Swedish insurance data. [3]

Run your function and use the results to plot a histogram of the estimated sampling distribution with a superimposed kernel density estimate of this distribution. [2]

Use your simulated test statistic values to obtain an estimated 5% critical value for your test. Compare your observed value with this and report your conclusions. [2]