assignment_4_written

April 10, 2025

1 Assignment 4 Part 1

1.1 Q1

A researcher in digital humanity studying narrative theory is trying to understand how the pace of the story (how fast or how slow the plots move across the story) impacts the success of the narrative. You can find their paper here. He randomly sampled a list of novels texts and measured their semantic speed (story pace) and popularity by the number of downloads (log transformed). Below is a subset of samples he collected:

index	speed	log_popularity
0	1	0
1	0	1
2	2	1
3	2	2
4	1	2

Now, he wants to run a simple linear regression model to test the linear relationship between story speed, denoted as X_i, and its log popularity, denoted as Y_i. To conduct this regression analysis, answer the following questions.

1. Write down the regression model specification with regression intercept β_0 and regression slope β_1.
2. Calculate the means for both variables $\bar{X}$ and $\bar{Y}$
3. Compute the OLS estiamtes of regression slope $\hat{\beta}_1 = \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_i-\bar{X})^2}$ and regression intercept $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$
4. Compute the residuals $e_i = Y_i - \hat{Y} = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$
5. Compute sum of squared errors $SSE = \sum e_i^2$, and what is the degree of freedom of SSE?
6. Compute the mean squared errors $MSE = \frac{SSE}{df(SSE)}$
7. Compute the standard error of OLS estimates of regression slopes $s.e.(\hat{\beta}_1) = \sqrt{MSE\frac{1}{\sum(X_i-\bar{X})^2}}$ and regression intercept $s.e.(\hat{\beta}_0) = \sqrt{MSE(\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i-\bar{X})^2})}$

8. The researcher wants to conduct a two-tailed t test on the relationship between speed and popularity. State the null hypothesis and alternative hypothesis. Think: should we test on regression intercept or regression slope?
9. Compute the T statistic $T = \frac{\hat{\beta}_1}{s.e.(\beta_1)}$, what is the degree of freedom for this T statistic?
10. Write down the Python code to compute the p value.

    Recall the python code for two tailed t test, if T statistic is positive, is

    ```
    from scipy import stats
    p = 2*(1-stats.t.cdf(T, df=df))
    print(p)
    ```

11. The researcher also wants to conduct a F test on the regression model. State the null hypothesis and the alternative hypothesis.
12. Compute $SSTO = \sum(Y_i - \bar{Y})^2$, $SSR = SSTO - SSE$, and $MSR = \frac{SSR}{df(SSR)}$
13. Compute the F statistic $F = \frac{MSE}{MSR}$, what is the degree of freedom of this F statistic?
14. Write down the Python code to compute the p value

    Recall that the python code for F test is

    ```
    from scipy import stats
    p = 1-stats.f.cdf(F, df1=df1, df2=df2)
    print(p)
    ```

15. Compute the coefficient of determination $R^2 = \frac{SSR}{SSTO}$. What does this R^2 mean?

1.2 Q2

Now, you want to use the matrix form to express the regression model and conduct your regression analysis using the matrix form. Using the dataset given in Q1, do the following.

1. Construct the design matrix X, and write down the response vector Y. What is the order (size) of these two matrices?
2. Compute the coefficient vector $b = (X^TX)^{-1}X^TY$, you may need to calculate X^TX first, then $(X^TX)^{-1}$, then X^TY, and finally $(X^TX)^{-1}X^TY$. Recall that $\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$
3. Compute the mean response vector $\hat{Y} = Xb$
4. Compute the residual vector $e = Y - \hat{Y}$
5. Compute the sum of squared errors $SSE = e^Te$, and mean squared errors $MSE = \frac{SSE}{df(SSE)}$
6. Compute the variance covariance matrix of coefficient vector

$$\sigma^2\{b\} = \sigma^2(X^TX)^{-1} = MSE(X^TX)^{-1}$$

7. Compute the standard errors for OLS estimates of regression intercept and regression slopes.

1.3 Assignment 4 part 2

In this part of the assignment, you need to download and open the Jupyter notebook `assignment_4_coding.ipynb` to follow the instructions and complete the assignment.

You also need to download the data file `gutenberg_data.csv` and place it at the same file folder of your jupyter notebook file.

After you have done it, submit both your written assignment and code assignment.