OCT21 homework

Answers in black are chatbot-based answer (manually checked) and answers in red are written notes to the answers and codes

Problem 1:

**Components of Simple Linear Regression:**

1. **Predictor (Independent) Variable (X)**: This is the input variable that we use to predict the outcome. It is often observed directly.
2. **Outcome (Dependent) Variable (Y)**: This is the response variable we want to predict or understand.
3. **Slope Coefficient (β1​)**: This describes how much Y is expected to change for a one-unit change in X. It represents the strength and direction of the relationship between X and Y.
4. **Intercept Coefficient (β0)**: This is the value of Y when X=0. It represents the starting point of the regression line on the Y-axis.
5. **Error Term (ϵ)**: This captures the randomness or noise in the relationship between X and Y — factors that affect Y but are not included in the model. It is assumed to be normally distributed with a mean of zero.

The **Simple Linear Regression Model** is given by:

***Y=β0+β1X+ϵ***

In this equation:

* Y is the predicted value (outcome variable).
* β0+β1X represents the deterministic part, which is a linear function of X.
* ϵ is the stochastic part, capturing the variability.

If we assume that the error term follows a normal distribution with mean 0 and standard deviation σ, then Y is normally distributed with mean β0+β1X and standard deviation σ. Each observation of Y given X is a sample from this normal distribution.

Problem2

Done in the code part

Problem 3

The red line in the plot represents the theoretical Simple Linear Regression model, while the green line represents the fitted Simple Linear Regression model.

The difference between the two is: the theoretical Simple Linear Regression model is used in the case to generate the data in the plot. It uses an error term to represent the sampling variance. Usually, this model is highly theoretical and will not reflect anything in reality; however, in this case, since the model iis used to generate the data, the data in the plot fit perfectly well with the model.

The fitted Simple Linear Regression model, on the other hand, is an estimation we made to the real data (or simulated data, as In this case.) Under the model, sum of Squared Errors between the observed and predicted values (under this model) is minimized, making it the best guess of the linear relationship between the two variables (if such relation exists)

Problem 4

print(model.summary()) gives you information about the fitted model. print(model.params) gives you the value of Intercept(β0) and X(β1)

print(model.params.values) gives you the value of Intercept(β0) and X(β1), but only two values withour variable names

Problem 5

Part of the question has been answered in problem 3 (the green part) The reason we use squares is because the difference between observed and predicted Y’s (the residuals) can be either positive or negative, and we are actually making the absolute value of it (“the distance” between the two values) minimal

Problem 6

First copy the 4 expressions

1. 1-((Y-fitted\_model.fittedvalues)\*\*2).sum()/((Y-Y.mean())\*\*2).sum()
2. fitted\_model.rsquared
3. np.corrcoef(Y,fitted\_model.fittedvalues)[0,1]\*\*2
4. np.corrcoef(Y,x)[0,1]\*\*2

note that all four expressions are expressing the same thing ————they all represent r-squared

the yellow part of expression 1 is the variance of all Y’s in the dataset (by definition)

In the blue part, “fitted\_model.fittedvalues” is the predicted value and trys to capture the variance in the yellow part, yet the prediction is still some values (residuals) away from the real observed value, and the residual actually represents “the **unexplained variation** in Y after fitting the model” and thus, the whole of the expression becomes “all variance-unexplained variance=explained variance”

Obviously, the more variance the model captures, the better, that’s why fitted\_model.rsquared can be interpreted as a measure of the accuracy of the model

Expreesion 3 and 4 are basically doing the same——they’re just calculating r first, and then squares it (yet r-squared=r^2 is just a math coincidence)

Problem7

The residuals in the fitted model is not normally distributed, suggesting the relation between the two variables are not linear.

NOV4 homework

Problem 8

The H0 will be:” β1=0”

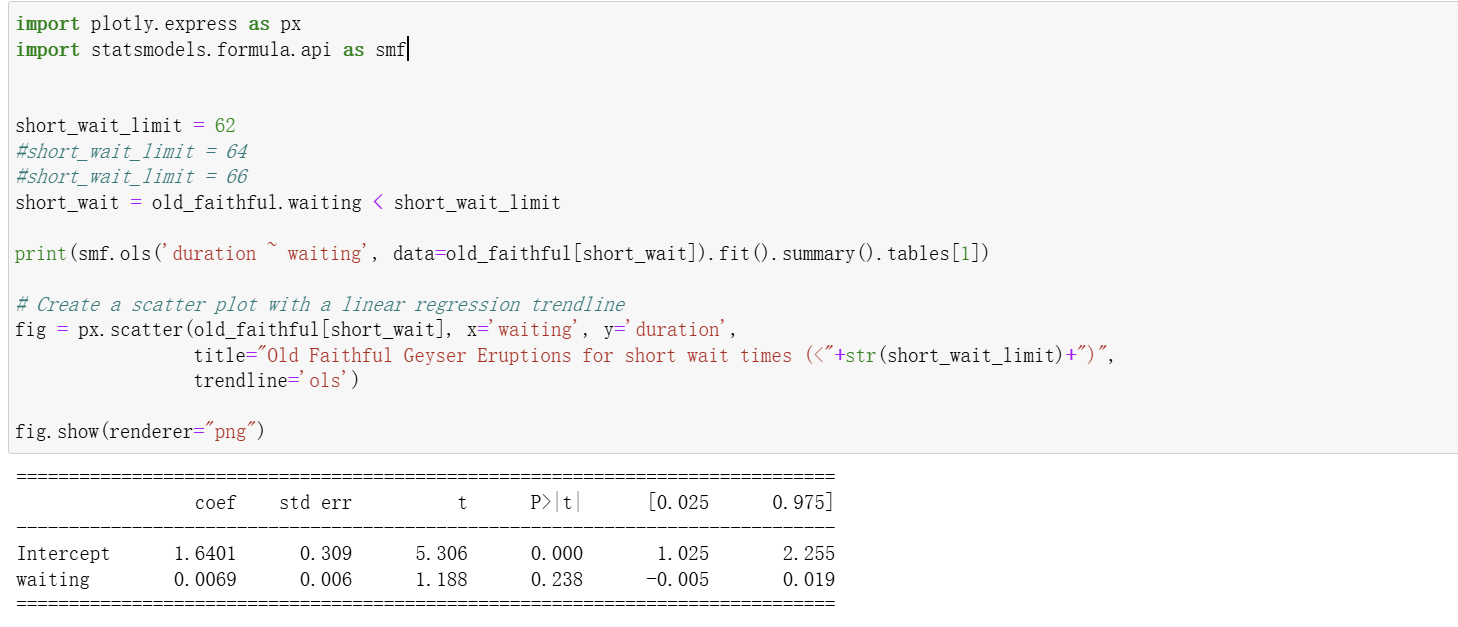


Given the P-value<0.001 in the data given by fitted\_model.summary(),

There is very strong evidence against H0

Problem 9

Wait<62



Wait<64

Wait<66



As seen in the screenshot, There is very strong evidence against H0 when wait time<62 or 64, and there are weaker yet still strong evidence against H0 when wait time<66

Problem 11

1. smf.ols('duration ~ waiting', data=old\_faithful)
2. smf.ols('duration ~ waiting', data=old\_faithful[short\_wait])
3. smf.ols('duration ~ waiting', data=old\_faithful[long\_wait])

these are all traditional kind of regression where B1 represent the slope of the best-of -fit line. While when we are using an indicator variable, B1 is the difference between “short” and “long” group and is just an constant. It is useful in Hypothesis testing, but it’s not slope of a graph, and is highly dependent on the indicator we assign. The value can vary a lot when we change the indicator, yet it is just not 0.