**Question 1:** Run the OLS related code shared in the “GLM.py” file for the ’us macro quarterly.xlsx’ data: [Points 30]

– Explain the model summary when all the data is fit. Discuss Rsquared, Df Residuals, Df Model, Method, F-statistic, Prob (F-statistic), Log-Likelihood, AIC, BIC, Coef, Std err, t, and P > |t|.

Rsquared: How the independent variable is explained by the changes in the dependent variable. In this case, the independent variable is highly explained by the changes in the dependent variable.

Df Residuals: The degree of freedom of residuals = number of observations – (number of variables+1) - 1 being estimated. In this case, 220 - (6+1) =213.

Df Model: The number of variables. In this case, Df model = 6.

Method: In this case, because the model we used is ordinary least square, the method would be Least Squares.

F-statistic and Prob (F-statistic): Because the Prob (F-statistic) is far less than 0.05, we can reject the null hypothesis that our independent variables is 0.

Log-Likelihood: Log-Likelihood measures the goodness a model fit the data. But it should be used when compare with other models. The higher the Log-Likelihood, the better the model fit the data.

AIC, BIC: Also, methods for selecting models and BIC has a higher penalty on the number of parameters. The smaller the AIC and BIC, the better the model fit the data.

Coef: it is the measurement of how change in that variable influences the change in the independent variable. In this case, x1 has the smallest coefficient and x2 has the highest.

Std err and t: it is the estimate of the SD of the coefficient and lower standard error leads to higher t statistics.

P > |t|: The p value of t statistics. In this case, we assume that alpha is 0.05, then all coefficients are significant.

– Run with the train-test split based on the ordering and compare the results of the test with the above model fitted on the entire data.

From the graph below, the model fits the data relatively well which is compatible with the result from the summary above. But the model is still a little bit off the test prediction part. The result with train-test split looks not as good as the one using entire data.

Chart, line chart

Description automatically generated

Chart, line chart, scatter chart

Description automatically generated

–Run all three variations of ANOVA and discuss results on SSR, df diff, ss dif, df resid, F and Pr(> F).

1. The first model in the comparison only uses an intercept which reasonably has a really high SSR and residual degree of free. Because the Pr(>F) is far smaller than 0.05, we are confident to say that the parameters in the second model are significant.
2. The less complex model with only three parameters for sure has a higher degree of freedom and higher SSR. Because the Pr(>F) is far smaller than 0.05, we are confident to say that the parameters in the second model but not in the first model are significant.
3. Because models compared are the same, they have the same df\_resid, ssr and we are not interested in the df\_diff , ss\_diff and F either.

**Question 2:** Run the code shared in the “GLM.py” file for the ”chip dataset.csv” data: [Points 50]

– Explain the model summary when the entire data is fitted. Discuss R-squared, Df Residuals, Df Model, Method, F-statistic, Prob (Fstatistic), Log-Likelihood, AIC, BIC, Coef, Std err, t, and P > |t|.

Rsquared: How the independent variable is explained by the changes in the dependent variable. In this case, R-squared is 0.790 which means the independent variable is highly explained by the changes in the dependent variable.

Df Residuals: The degree of freedom of residuals = number of observations – (number of variables+1) - 1 being estimated. In this case, 1610 - (5+1) =1604.

Df Model: The number of variables. In this case, Df model = 5.

Method: In this case, because the model we used is ordinary least square, the method would be Least Squares.

F-statistic and Prob (F-statistic): Because the Prob (F-statistic) is far less than 0.05, we can reject the null hypothesis that our independent variables is 0.

Log-Likelihood: Log-Likelihood measures the goodness a model fit the data. But it should be used when compare with other models. The higher the Log-Likelihood, the better the model fit the data.

AIC, BIC: Also, methods for selecting models and BIC has a higher penalty on the number of parameters. The smaller the AIC and BIC, the better the model fit the data.

Coef: it is the measurement of how change in that variable influences the change in the independent variable. In this case, x4 has the smallest coefficient and x1 has the highest.

Std err and t: it is the estimate of the SD of the coefficient and lower standard error leads to higher t statistics.

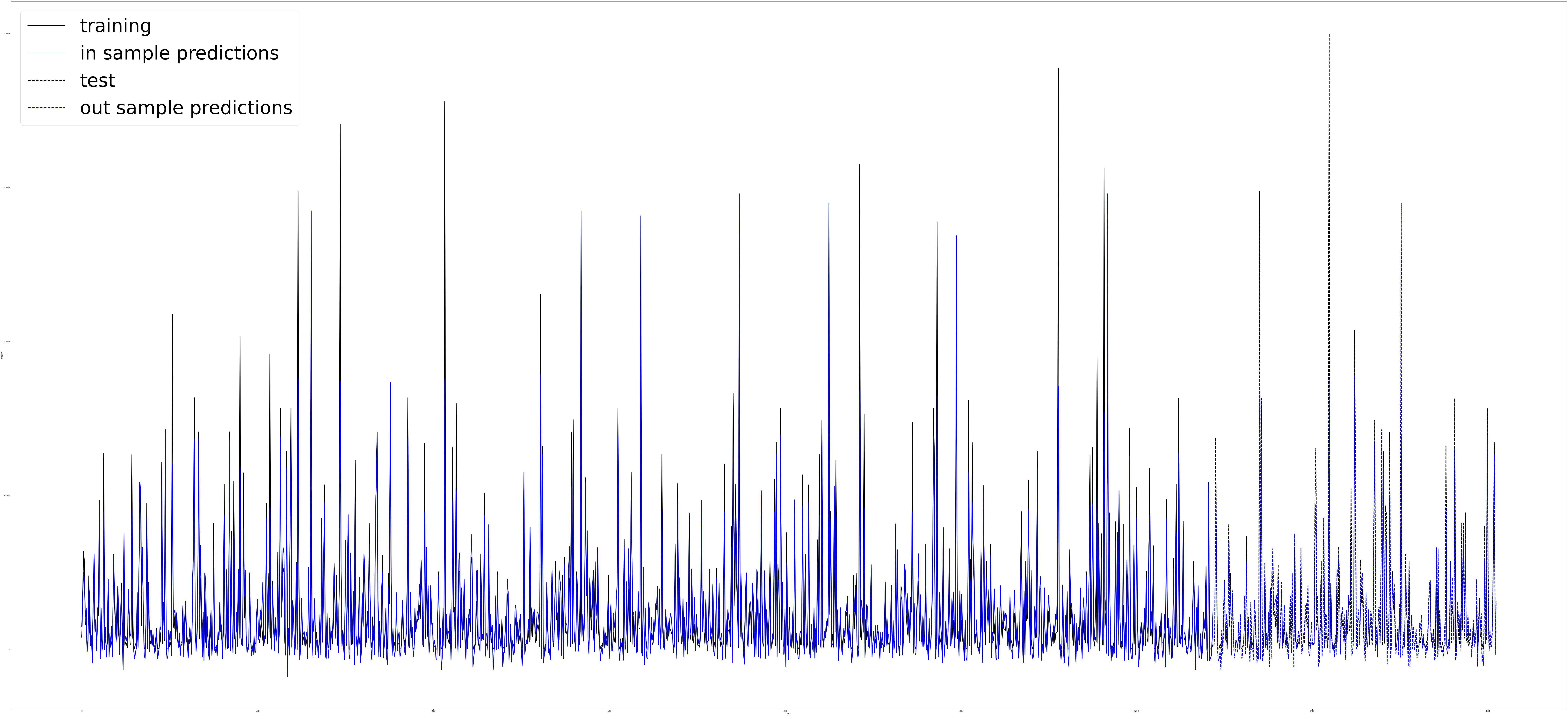
P > |t|: The p value of t statistics. In this case, we assume that alpha is 0.05, then all other parameters except x2 are significant.

– Run with the train-test split when the order is not important and compare the results of the test with the above model fitted on the entire data.

From the plots below, the model fits the data relatively good especially when we focus on the testing part. It is hard to tell the model using the entire data fits the data better the one using train-test split.

Chart

Description automatically generated



Chart, histogram

Description automatically generated

– Run all three variations of ANOVA and discuss results on SSR, df diff, ss dif, df resid, F and Pr(> F).

1. The first model in the comparison only uses an intercept which reasonably has a really high SSR and residual degree of free. Because the Pr(>F) is far smaller than 0.05, we are confident to say that the parameters in the second model are significant.
2. The less complex model with only three parameters for sure has a higher degree of freedom and higher SSR. Because the Pr(>F) is far smaller than 0.05, we are confident to say that the parameters in the second model but not in the first model are significant.
3. Because models compared are the same, they have the same df\_resid, ssr and we are not interested in the df\_diff , ss\_diff and F either.

– Question 3: Run the code: GLM using Poisson Regression on the ’Smokers Age.xlsx’  
data. And discuss the summary results [Points 20].

1. The R-square here is 1 and the log-likelihood is low which indicates that this is a good model fit the data.
2. The P-value of PersonYears is larger than 0.05 which means that this parameter is not significant.