Olzhas Shortanbaiuly

MATH 541 Data Analysis and Statistical Learning

Project 1

1-2. Estimating covariance and correlation between each feature X_i and target feature Y:

Covariance values are calculated by pandas.DataFrame.cov() library function and selecting the last column (meaning the correlation between each feature X_i and target feature Y), correlation values are obtained by pandas.DataFrame.corr() library function similarly to covariance. Abs() is used to compute absolute values of correlations.

	covariance	correlation	|correlation|
X1	11.648579	0.017686	0.017686
X2	113.686372	0.169065	0.169065
X3	3.872471	0.006639	0.006639
X4	-194.082744	-0.334646	0.334646
X5	280.593934	0.475479	0.475479
X6	14.507238	0.024879	0.024879
X7	-95.825753	-0.134272	0.134272
X8	-553.472653	-0.791981	0.791981
X9	-193.956879	-0.301703	0.301703
X10	285.889569	0.423752	0.423752
X11	77.566158	0.130866	0.130866
X12	-521.112217	-0.700715	0.700715
X13	-89.453521	-0.140852	0.140852

3. **Question 1: Based on latter correlation values, list of all variables that are relevant to the prediction?**
   As the computed values are Pearson's correlation coefficient telling about the existence of a linear relationship between each predictor X_i and target Y, we can't claim that certain variables are not relevant to the prediction of Y.
   Indeed, if the assumption of predicting a linear model relating all predictors X_i's and Y, the variables relevant to the prediction are those having correlations closer to 1. So, the variables with lowest correlations can be dropped in this case. Then, X_2, X_4, X_5, X_7, X_8, X_9, X_{10}, X_{11}, X_{12}, X_{13} can be selected as variables that are relevant for predicting a linear regression model.

4. The correlation matrix is obtained as follows:
   It was computed using pandas.DataFrame.corr() library function.

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13
X1	1.000000	-0.126214	0.895314	-0.011600	-0.094530	-0.037804	-0.140954	0.114312	0.019232	-0.107263	-0.067862	0.114528	-0.103715
X2	-0.126214	1.000000	-0.136522	-0.143213	0.115770	-0.115540	0.094172	-0.023571	-0.156356	0.094479	0.897628	-0.011275	0.086927
X3	0.895314	-0.136522	1.000000	-0.006078	-0.135922	-0.026325	-0.133774	0.119132	0.049778	-0.141082	-0.070532	0.108259	-0.085854
X4	-0.011600	-0.143213	-0.006078	1.000000	0.088465	0.130147	-0.024134	0.052491	0.887312	0.060582	-0.151375	0.032886	-0.050325
X5	-0.094530	0.115770	-0.135922	0.088465	1.000000	0.022658	-0.096994	-0.010652	0.037033	0.904310	0.078530	0.022285	-0.118563
X6	-0.037804	-0.115540	-0.026325	0.130147	0.022658	1.000000	-0.012402	-0.085202	0.114448	0.025233	-0.123994	-0.080808	-0.042785
X7	-0.140954	0.094172	-0.133774	-0.024134	-0.096994	-0.012402	1.000000	-0.007859	0.002428	-0.112295	0.060963	-0.056745	0.922130
X8	0.114312	-0.023571	0.119132	0.052491	-0.010652	-0.085202	-0.007859	1.000000	0.030411	0.004538	0.030912	0.916386	-0.000268
X9	0.019232	-0.156356	0.049778	0.887312	0.037033	0.114448	0.002428	0.030411	1.000000	0.015870	-0.167556	0.003816	-0.010559
X10	-0.107263	0.094479	-0.141082	0.060582	0.904310	0.025233	-0.112295	0.004538	0.015870	1.000000	0.060065	0.055109	-0.134995
X11	-0.067862	0.897628	-0.070532	-0.151375	0.078530	-0.123994	0.060963	0.030912	-0.167556	0.060065	1.000000	0.024213	0.069625
X12	0.114528	-0.011275	0.108259	0.032886	0.022285	-0.080808	-0.056745	0.916386	0.003816	0.055109	0.024213	1.000000	-0.050502
X13	-0.103715	0.086927	-0.085854	-0.050325	-0.118563	-0.042785	0.922130	-0.000268	-0.010559	-0.134995	0.069625	-0.050502	1.000000

5. **Question 2: What conclusion can you make about the structure of your predictors after analysis of correlation matrix?**
   As can be observed from the correlation matrix, certain predictors are highly correlated with each other, meaning that one can be interchanged with the other. This happens for correlation values close to 1 or -1: X_1 and X_3, X_2 and X_{11}, X_4 and X_9, X_5 and X_{10}, X_7 and X_{13}, X_8 and X_{12}.
   As such highly correlated predictors may result in collinearity issues resulting in increased model variance, dropping one of the two variables in each pair can be advised.

6. Scikit-learn/Python code for least square estimation of weights β_i in the model:

   from sklearn.linear_model import LinearRegression
   linreg = LinearRegression().fit(X_train, Y_train)
   linreg.coef_ (for β_i with $i > 1$)

   linreg.intercept_ (for β_0)

7. Obtaining values of least square estimates b_i's: (from b_0 to b_{13})

   9.994665255521712, 3.98329913e-03, 4.56300321e-03, 9.97178214e-01, -1.99695855e+00, 2.99724507e+00, -1.61843712e-03, 7.63246123e-05, -3.99864305e+00, -9.55522292e-04, 1.86009997e-03, 4.96900126e-01, -1.75246022e-03, -4.99506191e-01

8. Estimating variance of noise:

   For this purpose, mean_squared_error was computed based on Y_test and $\hat{Y}$.
   Obtained value: **0.027109956382164388**

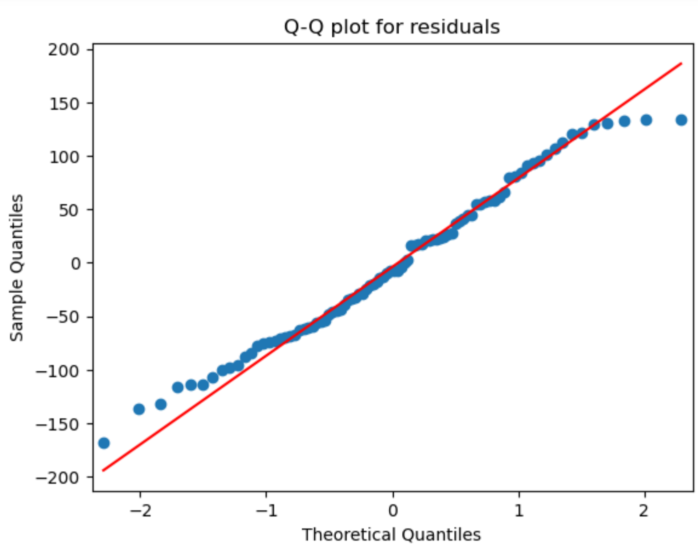
9. Calculating t-value (on a training set) of every non-target variable
   Using the values that were obtained in Steps 7-8, the following table of t values are obtained:

	t-value
X1	1.710762
X2	1.941018
X3	374.654420
X4	-773.182615
X5	1104.187879
X6	-1.386149
X7	0.030706
X8	-1615.223462
X9	-0.409769
X10	0.780795
X11	188.135305
X12	-0.754050
X13	-179.130956

10. To test the relevance of each predictor, the t-test is conducted with a 95% confidence level, and the absolute value of each t-value is compared to a Z-value of 1.96 (for a 95% confidence level). Variables with t-values higher than 1.96 are then discarded.
    Relevant for prediction: X_1, X_2, X_4, X_6, X_7, X_8, X_9, X_{10}, X_{12}, X_{13}

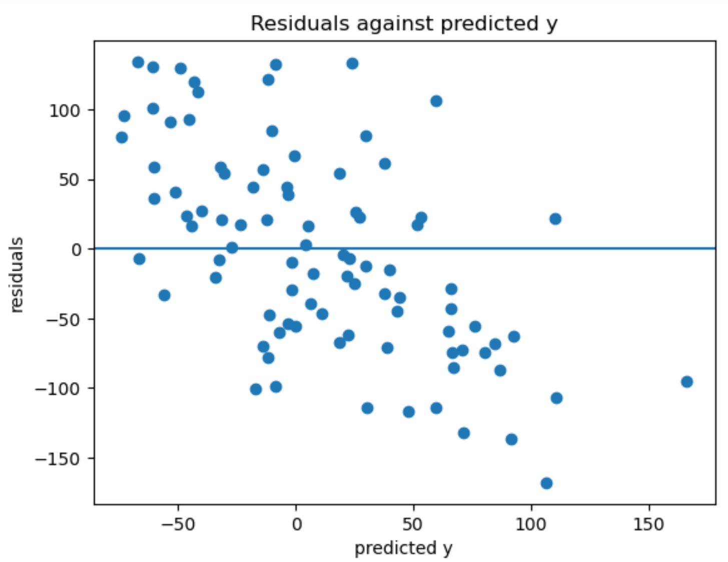
11. **Question 3: List variables that can be discarded.**
    Variables that can be discarded: X_3, X_5, X_{11}.

12. Calculate residuals and draw a Q-Q plot.
    Residuals are calculated through $\hat{Y} - Y$. sm.qqplot() function is used for the Q-Q plot.



13. **Question 4: Is an error normally distributed, yes or no (based on the Q-Q plot)?**
Yes, even though a plot is a little light-tailed, the majority of errors lie along the line, so it can concluded that the error is normally distributed.

14. Plotting "Residuals against $\hat{Y}$"



There is no cluster of points and no trend between residuals and predicted y values. Residuals are evenly distributed along both sides of the horizontal line $\hat{Y} - Y = 0$.

15. **Question 5: If your error is not normal, what would you assume about the real distribution of an error (based on the latter plot)?**
If the error is not normal, in the residuals versus predicted values plot, a constant spread of the residuals to the horizontal $\hat{Y} - Y = 0$ line across all of the predicted y values is required. Residuals should be evenly distributed between two sides of the horizontal line. Such trends mean the model satisfies the linearity and normality of variances assumption.

16. Calculating R^2: using sklearn.metrics.r2_score(), the following value is obtained: **0.9999892057969771**

17. **Question 6: Give the final verdict: did the linear regression model solve the prediction problem**
The linear regression model solved the prediction problem due to several reasons. First of all, R^2 is almost close to 1, meaning that the model explains all variability in response variables around its mean value and the linear regression model fits the data well. Other than that, the Q-Q plot showed that errors are normally distributed. Also, homogeneity of variance and linearity assumptions of the linear regression model are proven to be not violated by the "Residuals against $\hat{Y}$" plot.