Regression

Linear and Multi-Linear Regression

Regression is an approach to modeling the relationship between a dependent variable and one or more independent variables. For one independent variable, the approach is called linear regression and for more than one independent variables, the process is called multiple linear regression. The relation between independent and dependent variables is depicted as a function called regression function or regression equation.

A sample of the multiple linear regression equation is as follows:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p$$

where $\hat{Y}$ is the predicted or expected value of the dependent variable.

X1 through Xp are p distinct independent or predictor variables, b0 is the value of Y when all of the independent variables (X1 through Xp) are equal to zero, and b1 through bp is the estimated regression coefficients. Each regression coefficient represents the change in Y relative to a one unit change in the respective independent variable. Example, change in Y relative to a one unit change in X1, holding all other independent variables constant (i.e., when the remaining independent variables are held at the same value or are fixed).

Identifying & Controlling for Confounding With Multiple Linear Regression

Linear regression equation relating the risk factor (the independent variable) to the dependent variable as follows:

$$\hat{Y} = b_0 + b_1 X_1$$

Where b1 is the estimated regression coefficient that quantifies the association between the risk factor and the outcome.

To assess whether a third variable is a confounder, we can denote the potential confounder X2, and then estimate a multiple linear regression equation as follows:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$$

In the multiple linear regression equation, b1 is the estimated regression coefficient that quantifies the association between the risk factor X1 and the outcome, adjusted for X2 (b2 is the estimated regression coefficient that quantifies the association between the potential confounder and the outcome).

To determine the value of b0, b1, and b2, that gives the minimum error for the given dataset. Least Squares method is used.

A-Stack Solutions Use Cases

Steam Demand Prediction

We will use multiple linear regression to predict the Steam demand (i.e., the dependent variable) by using 4 independent/ input variables:
- Temperature
- Humidity
- Rain
- Wind

Validation of several assumptions is met before applying linear regression models. Identifying if a linear relationship exists between the dependent variable and the independent variables is the starting point.

Linearity Check

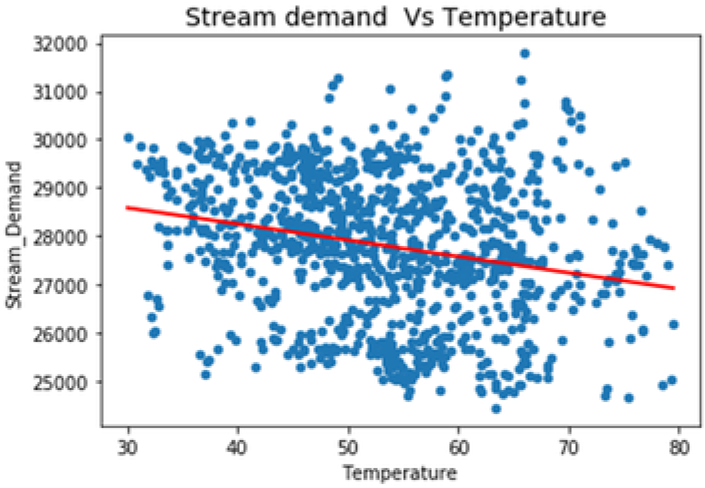
To check that a linear relationship exists between the dependent variable and the independent variables. We check the linear relationship exists between the following :
- The Steam_Demand (dependent variable) and the Temperature(independent variable)
- The Steam_Demand (dependent variable) and the Humidity (independent variable)
- The Steam_Demand (dependent variable) and the Rain(independent variable)
- The Steam_Demand (dependent variable) and the WindSpeed (independent variable.

To perform a quick linearity check, scatter diagrams are used.

Linear relationship existence can be seen in the following diagrams:

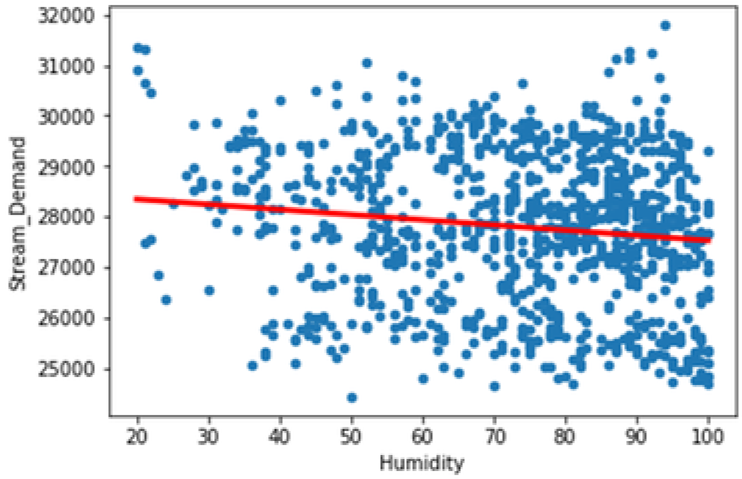
Steam Demand Vs Temperature

In this case, when Temperature goes up, the Steam_Demand goes down.



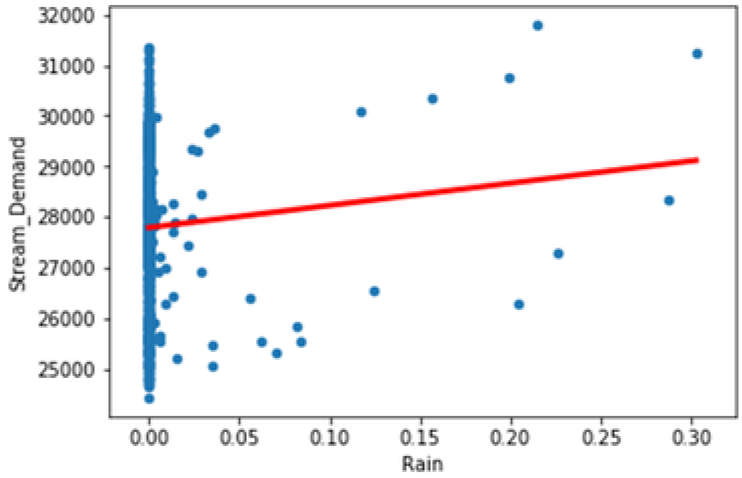
Steam Demand Vs Humidity

In this case, when Humidity goes up , the Steam_Demand goes down.



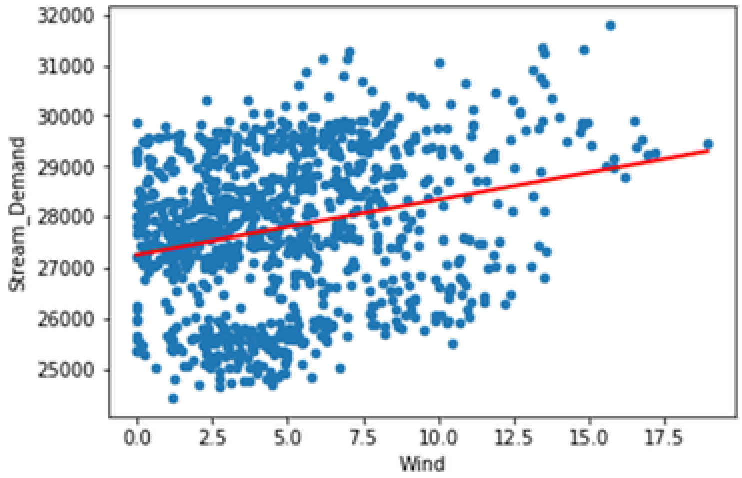
Steam Demand Vs Rain

In this case, when RainFall goes up , the Steam_Demand also goes up.



Steam Demand Vs Windspeed

In this case, when Windspeed goes up , the Stream_Demand goes up too.



Independent Variable Coefficient Calculation

Once the Linear relation has been established between dependent and independent variables, we need to quantify the dependency. This results in a regression equation of following form.

Steam_Demand = (Intercept) + (Temperature)*X1 + (Humidity)*X2 + (Rain)*X3 + (Wind)*X4

Using the regression libraries and the historical data for identified variables we can get values for Intercept and Coefficients X1-..4.

intercept : 29562.847075500533

coefficients : [-3.97750965 -3.05173578 4.20839003 1.14234164]

and form Equation as:-

Steam_Demand = (29562.847075500533) + (Temperature)*-3.97750965 + (Humidity)*-3.05173578 + (Rain)*4.20839003 + (Wind)*1.14234164

Using the regression equation and values for the dependent variable the demand can be predicted. eg

Temperature=64.50	Humidity=99.0	Rain=0.0	Wind=4.30

Predict the Stream Demand: 27186.438414230535

OLS Regression Performance Results

Dep. Variable:	Stream_Demand	R-squared:	0.142
Model:	OLS	**Adj. R-squared:**	0.139
Method:	Least Squares	**F-statistic:**	41.18
Date:	Tue, 26 Feb 2019	**Prob (F-statistic):**	5.66E-32
Time:	01:18:49	**Log-Likelihood:**	-8644.4
No. Observations:	1000	**AIC:**	1.730E+04
Df Residuals:	995	**BIC:**	1.732E+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.956E+04	307.219	96.227	0.000	2.9E+04	3.02E+04
Temperature	-39.7751	4.299	-9.253	0.000	-48.210	-31.340
Humidity	-3.0517	2.520	-1.211	0.226	-7.996	1.893
Rain	4208.3900	2161.447	1.947	0.052	-33.127	8449.907
Wind	114.2342	13.670	8.357	0.000	87.410	141.059

Omnibus:	73.304	Durbin-Watson:	0.153
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	30.278
Skew:	-0.194	**Prob(JB):**	2.66E-07
Kurtosis:	2.241	**Cond. No.**	4.58E+03

Sample TQL Query

Regression Model Performance Matrix

- **Mean Absolute Error (MAE)**
- **Root mean squared error (RMSE)**
- **R-Squared**
- **Adjusted R-squared**
- **F-Static**

Regression metrics are used to measure accuracy for continuous variables.

Mean Absolute Error (MAE)

Mean Absolute Error (MAE): MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It’s the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight.

$$MAE = \frac{1}{n}\sum_{j=1}^{n} | e_j |$$

Legend: – actual values

– the mean of the actual values

– predicted values

= – error

– size of the data set

If the absolute value is not taken (the signs of the errors are not removed), the average error becomes the Mean Bias Error (MBE) and is usually intended to measure average model bias. MBE can convey useful information, but should be interpreted cautiously because positive and negative errors will cancel out.

Root mean squared error (RMSE)

Root mean squared error (RMSE): RMSE is a quadratic scoring rule that also measures the average magnitude of the error. It’s the square root of the average of squared differences between prediction and actual observation.It represents the sample standard deviation of the differences between predicted values and observed values (called residuals).

$$RMSE = \sqrt{\frac{\sum_{j=1}^{n} e_j^2}{n}}$$

Legend: – actual values

– the mean of the actual values

– predicted values

= – error

– size of the data set

(or)

$$RMSE = \sqrt{MSE}$$

here MSE is

$$MSE = \frac{1}{n}\sum_{j=1}^{n} e_j^2$$

Legend: – actual values

– the mean of the actual values

– predicted values

= – error

– size of the data set

Mean squared error (MSE) is a single value that provides information about the goodness of fit of the regression line. The smaller the MSE value, the better the fit, as smaller values imply smaller magnitudes of error.

MSE is more popular than MAE because MSE "punishes" larger errors. But, RMSE is even more popular than MSE because RMSE is interpretable in the “steam demand” units.

RMSE helps to keep what attributes are included & what needs to be excluded

feature_cols = ['Temperature', 'Humidity','Wind']

RMSE : 1364.508031065501

feature_cols = ['Temperature', 'Humidity',]

RMSE : 1432.5796807515342

RMSE increased when we remove feature Wind.

R-Squared

R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.R-squared is a statistical measure that represents the extent to which the predictor variables (X) explain the variation of the response variable (Y).

Mathematically, R_Squared is given by:

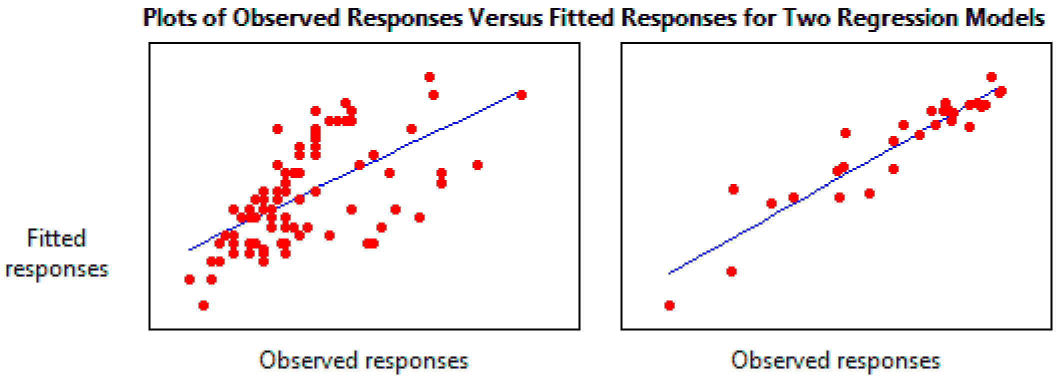
The numerator is MSE (average of the squares of the residuals) and the denominator is the variance in Y values. Higher the MSE, smaller the R_squared and poorer is the model.

For example :

if R-square is 0.7, this shows that 70% of the variation in the response variable is explained by the predictor variables. Therefore, the higher the R squared, the more significant is the predictor variable.

Graphical Representation of R-squared

Plotting fitted values by observed values graphically illustrates different R-squared values for regression models.



The regression model on the left accounts for 38.0% of the variance while the one on the right accounts for 87.4%. The more variance that is accounted for by the regression model the closer the data points will fall to the fitted regression line. Theoretically, if a model could explain 100%

of the variance, the fitted values would always equal the observed values and, therefore, all the data points would fall on the fitted regression line.

Generally, R-square is used when there is only one predictor variable, but what if there are multiple predictor variables?

The problem with R-squared is that it will either remain the same or increase with the addition of more variables, even if they do not have any association with the output variables.

Dep. Variable:	Steam_Demand	**R-squared:**	0.070
Model:	OLS	**Adj. R-squared:**	0.069
Method:	Least Squares	**F-statistic:**	37.76
Date:	Tue, 26 Feb 2019	**Prob (F-statistic):**	1.56e-16
Time:	05:06:43	**Log-Likelihood:**	-8684.5
No. Observations:	1000	**AIC:**	1.737e+04
Df Residuals:	997	**BIC:**	1.739e+04
Df Model:	2		
Covariance Type:	nonrobust		

We can say that 70% of the variation in the Steam demand rate is reduced by taking into account **Temperature and Humidity** . Or, we can say that 70% of the variation in SteamDemand is 'due to' or is 'explained by (reduced)' **Temperature and Humidity** .

only 30% of the variation in SteamDemand is left to explain after taking into account **Temperature and Humidity** in a linear way

Adjusted R-squared

Adjusted R-square restricts us from adding variables which do not improve the performance of your regression model. So, if we are building a regression model using multiple predictor variables, it is always recommended that you measure the Adjusted R-squared in order to detect the effectiveness of the model.

where n is the total number of observations and k is the number of predictors. Adjusted R² will always be less than or equal to R²

The adjusted R-squared adjusts for the number of terms in the model. Importantly, its value increases only when the new term improves the model fit more than expected by chance alone. The adjusted R-squared value actually decreases when the term doesn't improve the model fit by a sufficient amount.

Dep. Variable:	Steam_Demand	**R-squared:**	0.070
Model:	OLS	**Adj. R-squared:**	0.069
Method:	Least Squares	**F-statistic:**	37.76
Date:	Tue, 26 Feb 2019	**Prob (F-statistic):**	1.56e-16
Time:	05:06:43	**Log-Likelihood:**	-8684.5
No. Observations:	1000	**AIC:**	1.737e+04
Df Residuals:	997	**BIC:**	1.739e+04
Df Model:	2		
Covariance Type:	nonrobust		

In the above OLS results how the adjusted R-squared decreased (69%). On the other hand, R-squared blithely increases (70%) with each and every additional independent variables **Temperature and Humidity**.

F-Statistic

The F-statistic is a statistical measure used to judge whether at least one independent variable has a non-zero coefficient. A high F-statistic value leads to a statistically accepted p-value (i.e., p < 0.05).

P-Values

The p-value for each term tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that you can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable.

Conversely, a larger (insignificant) p-value suggests that changes in the predictor are not associated with changes in the response.

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	3.031e+04	295.447	102.603	0.000	2.97e+04	3.09e+04
Temperature	-33.3209	4.403	-7.567	0.000	-41.962	-24.680
Humidity	-10.0606	2.393	-4.205	0.000	-14.756	-5.366

In the output above , we can see that the predictor variables of **Temperature** and **Humidity** are significant because both of their p-values are 0.000.

However, the p-value for any predictor variables Rain is greater than the common alpha level of 0.05, which indicates that it is not statistically significant.