**LINEAR REGRESSION IN AN INFERENTIAL SETTING**

**Estimating a Simple Regression**

* Construct a regression model that explains how a district’s sales are affected by the advertising expense incurred in that district.
* Y = SALES ; X = ADVT
* Given a dataset of n=25 points, we want to fit a straight line.
* to the data that is “closest” to the 25 points.
* Here ‘a’ is the intercept and ‘b’ is the slope
* The value that comes from the regression line is called “predicted value”. The value that comes from the data is “observed value”. The difference between predicted value and observed value is called “Residual”.
* The difference between the observed and residual value is called the residual/error denoted as e1.
  + In general, residual value
* The sum of squared errors is denoted as:
* SSE is the measure of discrepancy between the line and n data points
* SSE=0 when all observations lie on the regression line

**Preliminary steps before regression analysis**

1. Univariate Analysis ( Single-Variable Analysis )

a. Investigate distribution of individual variables ( Histograms and Boxplots and Heatmaps )

b. Learn about shape,skew,extreme values,etc.

2. Bivariate Analysis ( Two-variable analysis )

a. Investigate correlation between two variables

b. Learn about type of relationship ( e.g. linear, non-linear, etc )

c. Scatterplots

**R-Square**

*The proportion of the total variation that is explained by the model is called the R-Square*

**Basics**

1. Consider the 25 data points available for the variable SALES (Y). If we are not

provided with the data for the variable ADVT (X), then, the best prediction for is . 2. The variation in the observed sales data around this prediction can be measured by:

3. This quantity is known as sum of squared total or SST (total variation in SALES)

**Understanding R-Squared**

1. Some of the variation in SALES across districts is because of differences in ADVT across districts

2. The rest of the deviation is due to other factors

When all the points fall on the regression line, then the value of R2 = 1. Otherwise R2 < 1

**Multiple Linear Regression Model**

**Introduction:**

1. Simple linear regression model has only one predictor

2. Multiple linear regression models have multiple predictors ( X1, X2, …. Xk ) 3. Main conceptual difference is the interpretation of individual predictors

**Example:**

**Model**

1. In general, we have k predictor variables, the regression equation is:

where is the regression coefficient of

2. Mathematically, it is *the change in Y when*  *is increased by 1 unit, and all other X’s are held*

*constant.*

**Python Code**

| mod3 = smf.ols('SALES~ADVT+INCOME', data=df)  res3 = mod3.fit()  print(res3.summary()) |
| --- |

Use + to add the other columns in the ols function.

After running the model, the equation with the coefficients is as below



We can interpret the coefficients as follows:

1. For every unit increase in ADVT, SALES increases by 5.0691 units on average, provided we keep INCOME constant.

2. For every unit increase in INCOME, SALES increases by 0.8081 units on average, provided we keep ADVT constant.

**Why does the coefficient of ADVT change in multiple linear regression as compared to single linear regression?**

Here, one unit is $1,000. An important point to remember while interpreting the coefficients in the case of multiple linear regression is that they isolate the influence of each variable. Note that the value of the coefficient of ADVT changes from 7.53 in the case of simple linear regression to 5.07 in the case of multiple linear regression. This is because we have controlled for INCOME. As you have seen before, INCOME and ADVT are correlated. Hence, in the case of simple linear regression, when we were allowing INCOME to vary, the influence of the INCOME variable was also being attributed to ADVT.

**Variable Selection**

**Standard Error of Estimate**

The standard error of estimate and the R2 are similar in the case of multiple linear regression as they were in simple linear regression. The adjusted R2 is a new evaluation metric that is useful in the case of multiple linear regression. Its utility can be determined by understanding the problem of variable selection in multiple linear regression.

In the case of multiple linear regression, we need to choose which independent variables to include in our model. But we cannot choose the variables that lead to a model with the highest value of R2. This is because additional variables start explaining noise in the training data, and even though adding variables improves the R2 value, the model does not perform well on unseen data.

**Variable Selection**

To choose between models with different explanatory variables, should we pick the model with the highest R2 ?

***No. R-Square measures the fit of the model to the given data. Sometimes, higher R-Square can hurt the predictability of a model.*** *R-Square can be increased by adding more predictors to the regression model, regardless of whether the predictors have a conceptual relationship to Y. This is because the model starts explaining random noise in the data. Thus, Predictive performance deteriorates even though the R-square is higher.*

*Therefore, R-square should not be considered while selecting the model.*

**Adjusted R-Square:**

1. For variable selection, R-square cannot be used.
2. It is not good to include a lot of predictors.
3. Adjusted R-Square can be used to select the model.

**What is Adjusted R-Square?**

Adjusted R-Square imposes a penalty on additional predictors that are added, only allowing the predictors that provide sufficient value.

**Usefulness of a predictor variable**

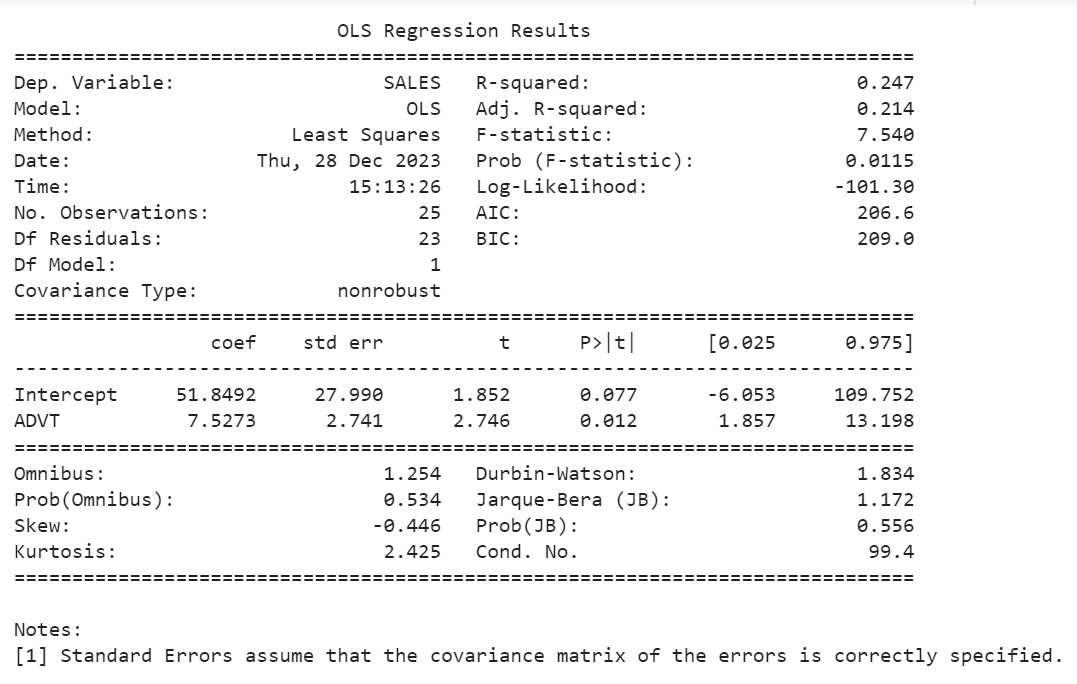
**Testing Usefulness for a predictor variable**

If X is not a useful predictor of Y, then . If it is a useful predictor of Y, then . However, is unknown. Its sample estimator is b. We use the estimator to conduct a hypothesis test to check where or . is also called the *population slope* that remains to be unknown.

**Kool-Karma Example**

* For a simple regression model of SALES and ADVT, the hypothesis are:

* The test statistic is , where b is the estimated coefficient; SE is the standard error and 0 in this case the value assumed under the NULL hypothesis. ( )
* The p-value can be computed from above.



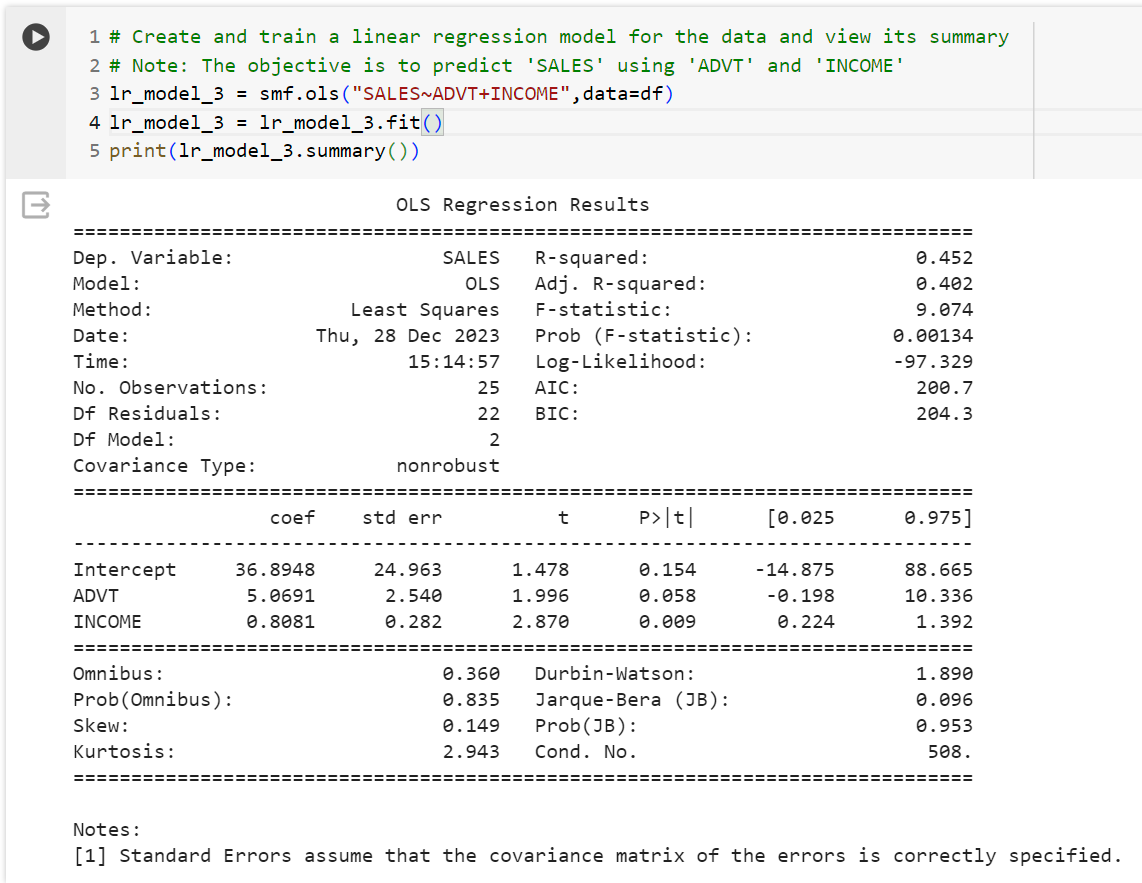
* Above is the output of the single regression analysis.
* The T-statistic is equal to 2.746 and p-value is equal to 0.012
* T-statistic can be computed as
* 95% confidence intervals are

**Conclusion**

* Since p-value is less than 5% significance level, we reject the null hypothesis and conclude that
* Consequently, the coefficient of ADVT is statistically significant at the 5% level of significance.
* Bottom line: ADVT is useful in explaining the variability in sales across different districts.

**F-Test**

**Multiple Predictors**

****

* In this, the p-value of ADVT is equal to 0.058. Therefore, we are no longer able to conclude that ADVT is a useful predictor because the p-value is above the 5% significance level.
* The p-value is INCOME is 0.009. Therefore, at 5% significance, it is a useful predictor.
* Bottom Line: In the presence of INCOME, ADVT is not statistically significant at 5% significance level. Therefore, ADVT does not have additional predictive power.
* On the other hand, at a 5% significance level, INCOME does have additional predictive power for predicting sales as its p-value is 0.009 compared to a regression model that just uses INCOME.

**Statistical Significance v/s Practical Significance**

*Statistical Significance* only implies whether a parameter is non-zero. It does not imply anything about the practical importance or the magnitude of the parameter. *Practical significance* is more specific.

For example, we may consider a predictor practically only significant if every dollar spent on advertising, the return on sales is at least $1.5 (or $3). On the other hand, statistical significance is telling you that the return is not zero but it can be $0.5 which is statistically significant but from a business point of view, it is not practically significant.

**Diagnostic Plots**

*Used to diagnose the accuracy/error in the regression model*

1. *Actual vs Fitted values* - scatter plot of actual values in X-axis and fitted/predicted values in Y-axis
2. *Residuals vs Fitted values* - Residuals on the Y-axis and fitted values on the X-axis
3. Histograms of the residuals: Should be “bell-shaped”
4. Q-Q Plot - should be in the central diagonal