Interpreting Linear Regression Through statsmodels .summary()

Tim McAleer Follow
Dec 5, 2020 · 7 min read

Images taken from https://www.statsmodels.org/
All coding done using Python and Python's statsmodels library.

```
In [8]: mod = smf.ols(formula='Lottery ~ Literacy + Wealth + Region', data=df)

In [9]: res = mod.fit()

In [10]: print(res.summary())
                            OLS Regression Results
==============================================================================
Dep. Variable:                Lottery   R-squared:                       0.338
Model:                            OLS   Adj. R-squared:                  0.287
Method:                 Least Squares   F-statistic:                     6.636
Date:                Sat, 28 Nov 2020   Prob (F-statistic):           1.07e-05
Time:                        14:39:43   Log-Likelihood:                -375.30
No. Observations:                  85   AIC:                             764.6
Df Residuals:                      78   BIC:                             781.7
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      38.6517      9.456      4.087      0.000      19.826      57.478
Region[T.E]   -15.4278      9.727     -1.586      0.117     -34.793       3.938
Region[T.N]   -10.0170      9.260     -1.082      0.283     -28.453       8.419
Region[T.S]    -4.5483      7.279     -0.625      0.534     -19.039       9.943
Region[T.W]   -10.0913      7.196     -1.402      0.165     -24.418       4.235
Literacy       -0.1858      0.210     -0.886      0.378      -0.603       0.232
Wealth          0.4515      0.103      4.390      0.000       0.247       0.656
==============================================================================
Omnibus:                        3.049   Durbin-Watson:                   1.785
Prob(Omnibus):                  0.218   Jarque-Bera (JB):                2.694
Skew:                          -0.340   Prob(JB):                        0.260
Kurtosis:                       2.454   Cond. No.                         371.
==============================================================================
```

Let's Figure Out What This Means!

Don't be intimidated by the big words and the numbers! This blog is here to translate all that information into plain English. Our goal is to provide a general overview of all statistics. Further research is highly recommended for in depth analysis for each component.
Let's start at the beginning.

```
In [8]: mod = smf.ols(formula='Lottery ~ Literacy + Wealth + Region', data=df)

In [9]: res = mod.fit()

In [10]: print(res.summary())
```

Coding our summary.

The earlier line of code we're missing here is `import statsmodels.formula.api as smf` So what we're doing here is using the supplied ols() or Ordinary Least Squares function from the statsmodels library. OLS is a common technique used in analyzing linear regression. In brief, it compares the difference between individual points in your data set and the predicted best fit line to measure the amount of error produced. The smf.ols() function requires two inputs, the formula for producing the best fit line, and the dataset.

The formula is provided as a string, in the following form: 'dependent variable ~ list of independent variables separated by the + symbol' In plain terms, the dependent variable is the factor you are trying to predict, and on the other side of the formula are the variables you are using to predict. The data set in this case is named 'df' and is being used to determine per capita wager in the Royal Lottery of 1830's France using a few characteristics. For the purpose of this lesson, the data is irrelevant but is available https://cran.r-project.org/web/packages/HistData/HistData.pdf for your interest.

Our first line of code creates a model, so we name it 'mod' and the second uses the model to create a best fit line, hence the linear regression. We name it 'res' because it analyzes the residuals of our model. Then we print our summary.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                Lottery   R-squared:                       0.338
Model:                            OLS   Adj. R-squared:                  0.287
Method:                 Least Squares   F-statistic:                     6.636
Date:                Sat, 28 Nov 2020   Prob (F-statistic):           1.07e-05
Time:                        14:39:43   Log-Likelihood:                -375.30
No. Observations:                  85   AIC:                             764.6
```

```
Df Residuals:                       78   BIC:                             781.7
Df Model:                            6
Covariance Type:            nonrobust
================================================================================
```

Details and statistics

The top of our summary starts by giving us a few details we already know. Our **Dependent Variable** is 'Lottery,' we've using OLS known as Ordinary Least Squares, and the **Date** and **Time** we've created the **Model**. Next, it details our **Number of Observations** in the dataset. **Df Residuals** is another name for our Degrees of Freedom in our mode. This is calculated in the form of 'n-k-1' or 'number of observations-number of predicting variables-1.' **Df Model** numbers our predicting variables. If you're wondering why we only entered 3 predicting variables into the formula but both Df Residuals and Model are saying there are 6, we'll get into this later. Our **Covariance Type** is listed as nonrobust. Covariance is a measure of how two variables are linked in a positive or negative manner, and a robust covariance is one that is calculated in a way to minimize or eliminate variables, which is not the case here.

R-squared is possibly the most important measurement produced by this summary. R-squared is the measurement of how much of the independent variable is explained by changes in our dependent variables. In percentage terms, 0.338 would mean our model explains 33.8% of the change in our 'Lottery' variable. **Adjusted R-squared** is important for analyzing multiple dependent variables' efficacy on the model. Linear regression has the quality that your model's R-squared value will never go down with additional variables, only equal or higher. Therefore, your model could look more accurate with multiple variables even if they are poorly contributing. The adjusted R-squared penalizes the R-squared formula based on the number of variables, therefore a lower adjusted score may be telling you some variables are not contributing to your model's R-squared properly.

The **F-statistic** in linear regression is comparing your produced linear model for your variables against a model that replaces your variables' effect to 0, to find out if your group of variables are *statistically significant*. To interpret this number correctly, using a chosen alpha value and an F-table is necessary. **Prob (F-Statistic)** uses this number to tell you the accuracy of the null hypothesis, or whether it is accurate that your variables' effect is 0. In this case, it is telling us 0.00107% chance of this. **Log-likelihood** is a numerical signifier of the likelihood that your produced model produced the given data. It is used to compare

coefficient values for each variable in the process of creating the model. **AIC** and **BIC** are both used to compare the efficacy of models in the process of linear regression, using a penalty system for measuring multiple variables. These numbers are used for feature selection of variables.

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	38.6517	9.456	4.087	0.000	19.826	57.478
Region[T.E]	-15.4278	9.727	-1.586	0.117	-34.793	3.938
Region[T.N]	-10.0170	9.260	-1.082	0.283	-28.453	8.419
Region[T.S]	-4.5483	7.279	-0.625	0.534	-19.039	9.943
Region[T.W]	-10.0913	7.196	-1.402	0.165	-24.418	4.235
Literacy	-0.1858	0.210	-0.886	0.378	-0.603	0.232
Wealth	0.4515	0.103	4.390	0.000	0.247	0.656

Omnibus:	3.049	Durbin-Watson:	1.785
Prob(Omnibus):	0.218	Jarque-Bera (JB):	2.694
Skew:	-0.340	Prob(JB):	0.260
Kurtosis:	2.454	Cond. No.	371.

Onto our coefficients!

Now we see the work of our model! Let's break it down.

The Intercept is the result of our model if all variables were tuned to 0. In the classic 'y = mx+b' linear formula, it is our b, a constant added to explain a starting value for our line.

Beneath the intercept are our variables. Remember our formula? 'Lottery ~ Region + Literacy + Wealth' Here we see our dependent variables represented. But why are there four different versions of Region when we only input one? Simply put, the formula expects continuous values in the form of numbers. By inputting region with data points as strings, the formula separates each string into categories and analyzes the category separately. Formatting your data ahead of time can help you organize and analyze this properly.

Our first informative column is the coefficient. For our intercept, it is the value of the intercept. For each variable, it is the measurement of how change in that variable affects the independent variable. It is the 'm' in 'y = mx + b' One unit of change in the dependent variable will affect the variable's coefficient's worth of change in the independent variable. If the coefficient is negative, they have an *inverse* relationship. As one rises, the other falls.

Our std error is an estimate of the standard deviation of the coefficient, a measurement of the amount of variation in the coefficient throughout its data points. The t is related and is a measurement of the precision with which the coefficient was measured. A low std error compared to a high coefficient produces a high t statistic, which signifies a high significance for your coefficient.

P>|t| is one of the most important statistics in the summary. It uses the t statistic to produce the *p value*, a measurement of how likely your coefficient is measured through our model by chance. The p value of 0.378 for Wealth is saying there is a 37.8% chance the Wealth variable has no affect on the dependent variable, Lottery, and our results are produced by chance. Proper model analysis will compare the p value to a previously established *alpha value*, or a threshold with which we can apply significance to our coefficient. A common alpha is 0.05, which few of our variables pass in this instance.

[0.025 and 0.975] are both measurements of values of our coefficients within 95% of our data, or within two standard deviations. Outside of these values can generally be considered outliers.

Omnibus describes the normalcy of the distribution of our residuals using skew and kurtosis as measurements. A 0 would indicate perfect normalcy. **Prob(Omnibus)** is a statistical test measuring the probability the residuals are normally distributed. A 1 would indicate perfectly normal distribution. **Skew** is a measurement of symmetry in our data, with 0 being perfect symmetry. **Kurtosis** measures the peakiness of our data, or its concentration around 0 in a normal curve. Higher kurtosis implies fewer outliers.

Durbin-Watson is a measurement of homoscedasticity, or an even distribution of errors throughout our data. Heteroscedasticity would imply an uneven distribution, for example as the data point grows higher the relative error grows higher. Ideal homoscedasticity will lie between 1 and 2. **Jarque-Bera (JB)** and **Prob(JB)** are alternate methods of measuring the same value as Omnibus and Prob(Omnibus) using skewness and kurtosis. We use these values to confirm each other. **Condition number** is a measurement of the sensitivity of our model as compared to the size of changes in the data it is analyzing. Multicollinearity is strongly implied by a high condition number. Multicollinearity a term to describe two or more independent variables that are strongly related to each other and are falsely affecting our predicted variable by redundancy.

Our definitions barely scratch the surface of any one of these topics. Independent research is strongly encouraged for an understanding of these terms and how they relate to one another. Hopefully this blog has given you enough of an understanding to begin to interpret your model and ways in which it can be improved!