Interpreting Significance and P-values

Introduction

In this lesson, you'll learn more about some of the ideas mentioned in earlier lessons around interpreting the significance of results. You'll learn that associating a confidence level to our model's output is very important during decision making and an essential part of data science.

Objectives

You will be able to:

- Evaluate a linear regression model by using statistical performance metrics pertaining to overall model and specific parameters

Let's get started

The ideas of hypothesis testing can be applied to regression, as well as statistical inference. For now, You'll learn about how it works in regression, and we'll spend time on formal hypothesis testing in a variety of contexts later on.

When performing hypothesis tests, a set of hypotheses are developed including a so-called **null hypothesis** and an **alternative hypothesis**. In regression, you will generally try to reject the null hypothesis, because that means that there is some relationship between the dependent and independent variable and your model at hand makes sense. Rejecting or not rejecting a null hypothesis always goes hand in hand with a so-called **confidence level**, often defined by the Greek letter alpha, α.

Hypothesis Testing in Regression

During regression, you try to measure the model parameters (coefficients). The null and alternative hypotheses are also set up in those terms. Think about the simple regression model you created using the advertising dataset. For a simple dataset like this, you can set up the hypotheses as follows:

> **Null Hypothesis (H_0)**: There is no relationship between the amount spent on TV advertisement and sales figures
>
> **Alternative Hypothesis (H_a):** There is "some" relation between the amount spent on TV advertisement and sales figures

If we reject the null hypothesis, it means that we believe that there is some sort of relationship between TV advertisement and sales. If we fail to reject the null hypothesis, it means that we believe that there is no *significant* relationship between TV advertisement and sales, and that our slope parameter m in $y = mx + c$ is not *significantly* different from zero.

You see that we used the word *significant* and *significantly*. What does this mean, and when is something *significant*?

P-value as a level of statistical significance

When performing statistical analyses, rejecting or not rejecting a null hypothesis always goes along with an associated **significance level** or **p-value**.

> The p-value represents a **probability of observing your results (or something more extreme) given that the null hypothesis is true**

Applied to a regression model, p-values associated with coefficients estimates indicate the probability of observing the associated coefficient given that the null-hypothesis is true. As a result, very small p-values indicate that coefficients are **statistically significant**. A very commonly used cut-off value for the p-value is 0.05. If your p-value is smaller than 0.05, you would say:

> The parameter is statistically significant at α level 0.05.

Just like for statistical significance, rejecting the null hypothesis at an alpha level of 0.05 is the equivalent for having a 95% confidence interval around the coefficient that does not include zero. In short

> The p-value represents the probability that the coefficient is actually zero.

Let's import the code from the previous lesson and let's have a look at the p-value.

In [1]:

```python
import pandas as pd
import matplotlib.pyplot as plt
import statsmodels.api as sm
import statsmodels.formula.api as smf
import scipy.stats as stats
plt.style.use('fivethirtyeight')

data = pd.read_csv('Advertising.csv', index_col=0)
f = 'sales~TV'
f2 = 'sales~radio'
model = smf.ols(formula=f, data=data).fit()
```

Now, we can check for the p-value associated with our coefficient for TV:

In [2]:

```python
model.pvalues
```

Out[2]:

```
Intercept    1.406300e-35
TV           1.467390e-42
dtype: float64
```

You'll notice that you also get a p-value for the intercept. Except if the intercept is actually 0, you should get a pretty low p-value here anytime. The tiny p-value for TV indicates that our coefficient for TV is definitely significant, and we can reject the null hypothesis. Next, there are two final things from the entire model summary we'd like to highlight:

In [3]:

```python
model.summary()
```

Out[3]:

OLS Regression Results

Dep. Variable:	sales	R-squared:	0.612
Model:	OLS	Adj. R-squared:	0.610
Method:	Least Squares	F-statistic:	312.1
Date:	Fri, 10 Apr 2020	Prob (F-statistic):	1.47e-42
Time:	17:22:16	Log-Likelihood:	-519.05
No. Observations:	200	AIC:	1042.
Df Residuals:	198	BIC:	1049.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	7.0326	0.458	15.360	0.000	6.130	7.935
TV	0.0475	0.003	17.668	0.000	0.042	0.053

Omnibus:	0.531	Durbin-Watson:	1.935
Prob(Omnibus):	0.767	Jarque-Bera (JB):	0.669
Skew:	-0.089	Prob(JB):	0.716
Kurtosis:	2.779	Cond. No.	338.

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

1. Have a look at the probability for the F-statistic in the upper right plane. This is a p-value for the **overall model fit**. You can see how this probability value is the exact same as the p-value for TV. You'll see that this is always the case if there is only one

independent variable. Here, TV drives the entire model, so the model p-value is the same as the TV coefficient p-value.

2. In the plane with the coefficient estimates, p-values are in the column `P > |t|` and rounded to 3 digits (hence 0.000 for Intercept and TV). You can find the confidence intervals in the two last columns: The confidence interval for intercept is [6,13, 7.935] meaning that there is a 95% chance that the actual coefficient value is in that range. For TV, there is a 95% chance that the actual coefficient value is in the interval [0.042, 0.053]. Note that 0 is in none of these intervals as expected given the very low p-value.

Note that the alpha level of 0.05 is just a convention. You'll come across alpha levels of 0.1 and 0.001. Note that the confidence intervals change too as alpha levels change (to 90% and 99% confidence intervals, yet the standard output of statsmodels is a 95% confidence interval).

Try running the above and check for p-values for `radio` and see how would interpret the outcome.

Additional Resources

We encourage you to visit the following links to see more examples and explanations around hypothesis testing and p values in regression!

Hypothesis Test for Regression Slope

Regression Continued

Summary

In this lesson, you learned how to apply the ideas of hypothesis testing in regression analysis to associate significance and confidence level with your model. You used this with your previous regression model to check the association. In the next lab, you'll combine all the ideas around simple linear regression with OLS on a slightly more complex dataset with more features.