Linear Regression: P-Value Basics

I'm deciding which variables to take into account for my linear regression model. Tell me what is the p-value, how to calculate it and how to use it

What is the p-value?

The p-value is a statistical measure that helps you determine whether there is evidence to reject the null hypothesis in a hypothesis test. In the context of linear regression, it helps you understand the significance of each predictor variable.

- **Null Hypothesis (H0):** The coefficient of the predictor variable is equal to zero (i.e., the variable has no effect).
- **Alternative Hypothesis (H1):** The coefficient of the predictor variable is not equal to zero (i.e., the variable has an effect).

A low p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis, suggesting that the predictor variable is significantly related to the dependent variable.

How to Calculate the p-value

1. **Fit the Linear Regression Model:**
   - Use your data to fit a linear regression model.
2. **Obtain the t-Statistic for Each Coefficient:**
   - For each predictor variable, the t-statistic is calculated as the estimated coefficient divided by its standard error.

$$t = \frac{\beta_i}{SE(\beta_i)}$$

   where β_i is the estimated coefficient for predictor i, and $SE(\beta_i)$ is its standard error.
3. **Calculate the p-value:**
   - The p-value is derived from the t-distribution, and it represents the probability of observing a t-statistic at least as extreme as the one observed, under the null hypothesis.
   - For a two-tailed test:

$$p\text{-value} = 2 \times (1 - \text{CDF}(|t|))$$

   where CDF is the cumulative distribution function of the t-distribution.

How to Use the p-value

1. **Determine Significance:**

- Compare the p-value of each predictor variable to your chosen significance level (commonly 0.05).
- If $p \leq 0.05$, reject the null hypothesis for that predictor, indicating it is significant.
- If $p > 0.05$, fail to reject the null hypothesis, indicating it is not significant.

2. **Model Refinement:**
   - Remove predictor variables with high p-values (insignificant variables) to simplify the model.
   - Keep predictor variables with low p-values (significant variables) as they contribute to the model.

Example with Python

Here's an example using Python and the `statsmodels` library to fit a linear regression model and obtain p-values for the predictors:

```python
import statsmodels.api as sm
import pandas as pd

# Sample data
data = {
    'X1': [1, 2, 3, 4, 5],
    'X2': [5, 4, 3, 2, 1],
    'Y': [1.2, 2.3, 3.3, 4.4, 5.5]
}

df = pd.DataFrame(data)

# Add a constant term for the intercept
X = sm.add_constant(df[['X1', 'X2']])
y = df['Y']

# Fit the regression model
model = sm.OLS(y, X).fit()

# Print the summary to get p-values
print(model.summary())
```

Interpreting the Output

The output of `model.summary()` will include:

- **Coefficients:** Estimated coefficients for each predictor.
- **Standard Errors:** Standard errors of the estimated coefficients.
- **t-values:** t-statistics for each coefficient.
- **p-values:** p-values for each coefficient.

You can use the p-values to decide which predictors to keep in your model based on their statistical significance.

Summary

- **p-value:** Measures the probability of obtaining a result at least as extreme as the one observed, under the null hypothesis.
- **Calculation:** Derived from the t-statistic of each predictor's coefficient.

- **Usage:** Helps determine the significance of predictors in the regression model, guiding model refinement by retaining significant variables and discarding insignificant ones.