Understanding P-Values in Regression Models

The Challenge of Model Tuning

- Regression models often involve multiple predictors, leading to numerous possible model configurations.
- It's impractical to test every single arrangement of variables using out-of-sample metrics alone. With only five variables there are 64 different model choices.
- We need an efficient way to identify which variables are likely to make a meaningful contribution to the model.

Introduction to P-values

A **p-value** is our new tool. The idea is that we can fit a linear model with k predictors and we can get a p-value for each predictor.

- We make a hypothesis that $\beta_i = 0$, meaning X_i is not a significant predictor. If $\beta_i = 0$, we can just leave X_i out and we have the same model.
- A P-value quantifies the evidence for that hypothesis. Large p-values means there is a lot of evidence that $\beta_i = 0$.
- Smaller P-values indicate that the hypothesis is wrong, and β_i is definitely not 0.
- We often put a cut-off, if the p-value is above 0.05 it is large

Real Example with P-values

- Consider a regression model predicting house prices with three predictors: square footage (*sqft*), number of bedrooms (*bedrooms*), and distance to city center (*distance*).

Variable	P-value
sqft	0.001
bedrooms	0.045
distance	0.200

- The variable *distance* has a P-value of 0.200, greater than the common alpha level of 0.05.
- Removing *distance* may simplify the model without sacrificing predictive power.

Fitting a Model with Statsmodels Formula API: Code Example

Unfortunately, the scikit learn regression formulas do not give p-values. We will use a different package with a different syntax. Luckily it's similar to the Bambi syntax.

Fitting a Model with Statsmodels Formula API: Code Example

```
# Importing statsmodels formula API
import statsmodels.formula.api as smf

# Defining our data
data = df

# Fitting the model using formula
model = smf.ols(formula='y~x1+x2+x3',data=data).
    fit()

#Summary statistics
model.summary()
```

Reading the Summary Table

- The `coef` column shows the estimated coefficients, indicating the effect size of each predictor.
- The `P>|t|` column shows the P-values for each predictor.

Statistical vs Practical Significance

Making a hard cut-off for determining when a p-value if large or small is convenient but sometimes we need to take a closer look

- A variable that has a p-value that is small is considered to be **statistically significant**. This means that the math says it should be included in the model.
- Variables with large p-values are not statistically significant, however, we should not always completely write them off
- **Practical Significance** is using your brain in close calls or in special cases
- It's possible to have statistical significance without practical significance and vice versa

Practical vs Statistical Significance: Real Example

Consider a regression model predicting customer spending based on age (*age*), monthly visits (*visits*), and customer loyalty (*loyalty*).

Variable	P-value	Coefficient
age	0.005	0.02
visits	0.001	50
loyalty	0.07	200

- *age* has a coefficient of 0.02, meaning each additional year of age increases spending by only 2 cents - a practically insignificant amount.
- *visits* has a coefficient of 50, meaning each additional visit increases spending by \$50 - a practically significant amount.
- *loyalty* has a P-value of 0.07, indicating it's not statistically significant at the 0.05 level. However, its coefficient of 200 suggests that loyalty could still be practically significant.

Interactions Between Variables and P-values

- Interaction between variables can significantly affect P-values.
- Two variables may individually have high P-values but can become significant when one is removed.
- This complexity increases with the number of variables, making interpretation challenging.
- Avoid blindly removing all variables with high P-values at once.
- Safe approach: Check out-of-sample predictive performance each time a variable is removed to ensure the model improves.

Real Example: Interactions Affecting P-values

- Consider a model with three predictors: *age*, *income*, and *education*.

Variable	Initial P-value	P-value after Removing *income*
age	0.25	0.045
income	0.2	–
education	0.001	0.0009

- Initially, *age* and *income* have high P-values.
- After removing *income*, the P-value for *age* drops below 0.05, making it significant.
- Out-of-sample performance improved upon removing *income*.

Special situations

- Never remove the intercept, even if it has a p-value greater than 0.05.
- If there are multiple dummy variables created from a single factor, it may be possible for one dummy variable to be significant and the other is not. It's okay to drop the insignificant dummy variable and keep the other.

Model Tuning with P-values

Follow these steps to use p-values in your model tuning

- Fit the model with all the variables you think are significant
- Choose a cut-off, perhaps 0.05
- If there are p-values above 0.05, remove the largest one or perhaps a few that are large
- Refit the model and check p-values again
- Repeat the previous two steps
- Check out of sample predictive accuracy along the way to be safe
- Consider the practical significance of variables you may be removing
- Stop when all the variables are significant, either statistically or practically

Summary of P-values in Regression Models

- P-values can help you choose which variables to include in your model.
- Use them alongside other criteria, such as practical significance, to make your final decision.
- Always exercise caution and use P-values as one of many tools in your statistical toolbox.