codecademy

Choosing a Linear Regression Model

Choosing a Linear Model

For multivariate datasets, there are many different linear models that could be used to predict the same outcome variable. Therefore, we need methods for comparing models and choosing the "best" one for the task at hand.

R-Squared

One method for comparing linear regression models is R-squared. R-squared is interpreted as the proportion of variation in an outcome variable which is explained by a particular model. Therefore, we generally prefer models with higher R-squared.

Nested Models

Two models are considered *nested* if one model contains all the same predictors as the other model, plus any number of additional predictors. For example, $model1$ and $model2$ in the example code are nested models because all of the predictors in $model2$ are also in $model1$.

```
import statsmodels.api as sm
model1 = sm.OLS.from_formula('salary ~ age + years_experience', data = data).fit()
model2 = sm.OLS.from_formula('salary ~ age + years_experience + gender + country + age:years_experience', data = data).fit()
```

codecademy

Adjusted R-Squared

If we want to compare nested models, R-squared can be problematic because it will ALWAYS favor the larger (and therefore more complex) model. Adjusted R-squared is an alternative metric that penalizes R-squared for each additional predictor. Therefore, larger nested models will always have larger R-squared but may have smaller adjusted R-squared.

F-test

To compare nested linear models, we can use a hypothesis test called an F-test. The null hypothesis is that the coefficients on all of the additional predictors in the larger model are zero; the alternative hypothesis is that at least one is non-zero. If we reject the null (by calculating a p-value less than our significance threshold), then that suggests that at least one of the additional predictors in the larger model is warranted. The provided code demonstrates how to run an F-test in Python.

```
from statsmodels.stats.anova import
anova_lm
anova_results = anova_lm(model1, model2)
print(anova_results)
```

Log Likelihood

Log-likelihood is a metric that can be used to choose the best linear regression model for predictive purposes; it is related to the probability of the observed data given a particular model. Higher log-likelihood is therefore better; however, log-likelihood is often negative, so “higher” means a smaller negative number.

codecademy

AIC and BIC

For nested linear regression models, log-likelihood is always higher for models with more parameters. Akaike information criterion (AIC) and Bayesian information criterion (BIC) are log-likelihood based criteria that include a penalty for additional predictors (BIC uses a bigger penalty). Therefore, AIC and BIC can be used to compare nested models to find the best model with the smallest predictor set. Lower AIC or BIC is better.

Training and Test Sets

One way to compare two linear regression models is to:

1. Randomly split the data into training and test sets.
2. Fit both models on the training set.
3. Use those fitted models to predict the outcome variable for the test set.
4. Evaluate the predictive accuracy using a metric such as predictive root mean squared error.

This allows us to see how well a model performs with respect to making predictions for new data (that was not used to fit the model).

Predictive Root Mean Squared Error

Predictive root mean square error (PRMSE) can be used to evaluate the predictive accuracy of a linear regression model for new data (that was not used to fit the model). It is calculated as the square root of the of the mean squared difference between the predicted and true values; therefore smaller PRMSE is preferable.

Print Share

codecademy