Week 9 - Assessed Exercises

November 4, 2023

1 Week 9 - Assessed exercises

In this weeks exercises we will fit a regression model and create a stepwise AIC function.

Submit your answers to the questions below on Moodle.

[1]:

```python
from pandas import Series, DataFrame
import pandas as pd
import numpy as np
import numpy.random as npr
import statsmodels.api as sm
```

In this week's assessed exercises we will look at the prostate data set, which we used in this week's material.

Load in the prostate dataset. Create a Series y which contains the response variable (lpsa).

[ ]:

Create a DataFrame X which contains the explanatory variables (lcavol, lweight, age, lbph, svi, lcp, gleason, and pgg45). Standardise X, such that all variables have mean 0 and standard deviation 1.

[ ]:

Add an intercept column to X. The intercept column should be the first column of X.

[ ]:

Which column of X has the smallest correlation with y?

[ ]:

Use the `OLS` function from statsmodels.api to fit a linear regression with y as your dependent/response variable and the first two columns of X as the explanatory variables, i.e. the intercept column and the lcavol column.

What is the adjusted R-squared to 3 decimal places?

[ ]:

We now want to run a *forward selection AIC regression*. AIC is the Akaike information criterion. It's designed to penalise models with lots of explanatory variables so that we pick models which fit the data well but aren't too complicated. In general, if you have two models fitted to the same data, the model with the lowest AIC is preferable. The AIC is given as part of the model summary with OLS.

The steps to run a forward selection AIC regression are: 1. Begin with a model that contains no variable (other than the intercept). Run a linear regression and record the AIC. For now, this is our *current model*. 2. Find the most significant variable, i.e. the variable that lowers the AIC the most a. Run a linear regression with the *current model* plus one additional variable, and record the decrease in AIC. b. Repeat step 2a for each variable not included in the *current model*. c. Find the variable with the biggest decrease in AIC. d. Update the *current model* to include the variable that decreases the AIC the most. 3. If none of the variables lower the AIC then go to step 4. Otherwise repeat step 2 until adding variables no longer reduces the AIC. 4. Report your final chosen variables

Write a function called `forwardAIC` which performs this algorithm given the DataFrame X and Series y. The function should return the column numbers of the X matrix for the model that gives the lowest AIC.

[ ]:

Which five variables (not including the constant) come back as important?

[ ]:

What's the AIC of this chosen model?

[ ]:

Bonus question (ungraded)

(Included in the non-assessed exercises on Moodle, if you wish to check your answers)

Run the same analysis on the full Diamonds data (given with this notebook) using price as the dependent/response variable. Load the data in and create dummy variables for the categorical variables cut, colour and clarity (using `pd.get_dummies`). You will need to drop one category for each categorical variable (i.e. drop 'Fair' for cut, drop 'D' for color, and drop 'I1' for clarity). Otherwise the model cannot be fully determined.

[ ]:

How many variables (not including the constant) get chosen?

[ ]:

What's the AIC of this chosen model?

[ ]: