Subset selection

Outline

- Best subset selection
- Forward stepwise selection
- Backward stepwise selection
- Selecting the best model

Why improving least squares fitting

Better prediction accuracy

- If the true relationship between the response and the predictors is approximately linear and n>>p, OLS works fine (small bias and small variance).
- However if n is not much larger than p, variance is large and test error will be large.
- If n>p, least squares can’t be used at all => we need to constrain or shrink the estimated coefficients.
- Such shrinkage reduces the variance and improve the model’s generalisation.

Better model interpretability

- Removing irrelevant predictors also makes the model much more interpretable.
- OLS itself is unable to make such predictor (variable) selection.

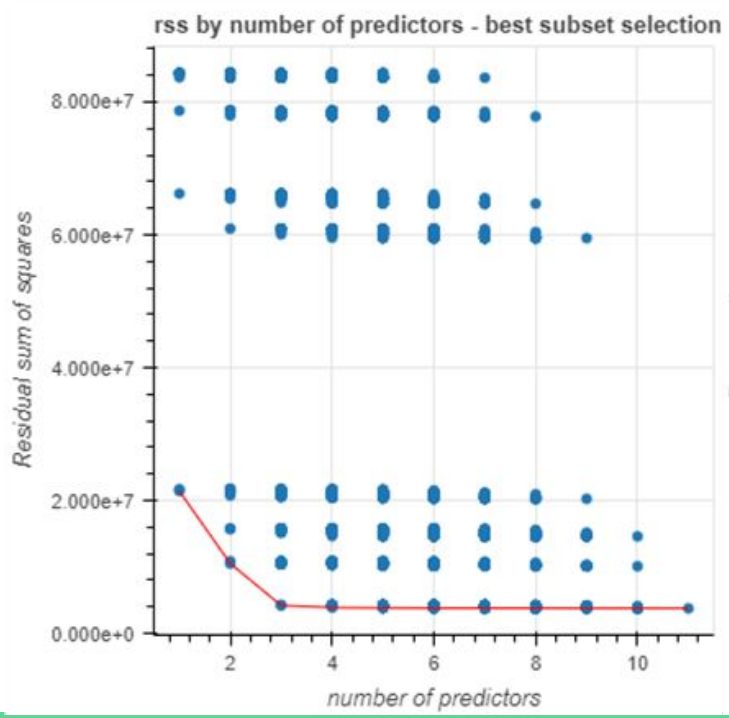
Best subset selection

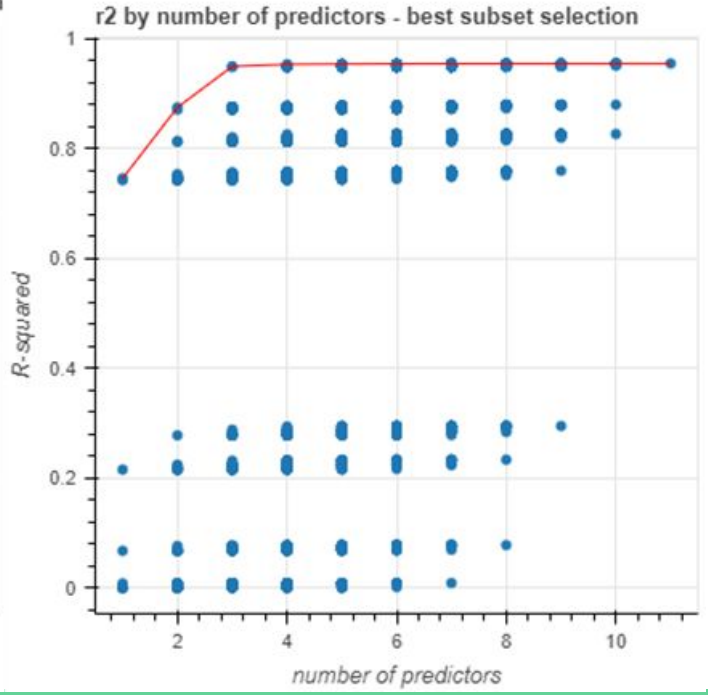
Algorithm 6.1 *Best subset selection*

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Best subset selection results - the Credit dataset

RSS and R2 for all possible regression models of Balance on the predictors



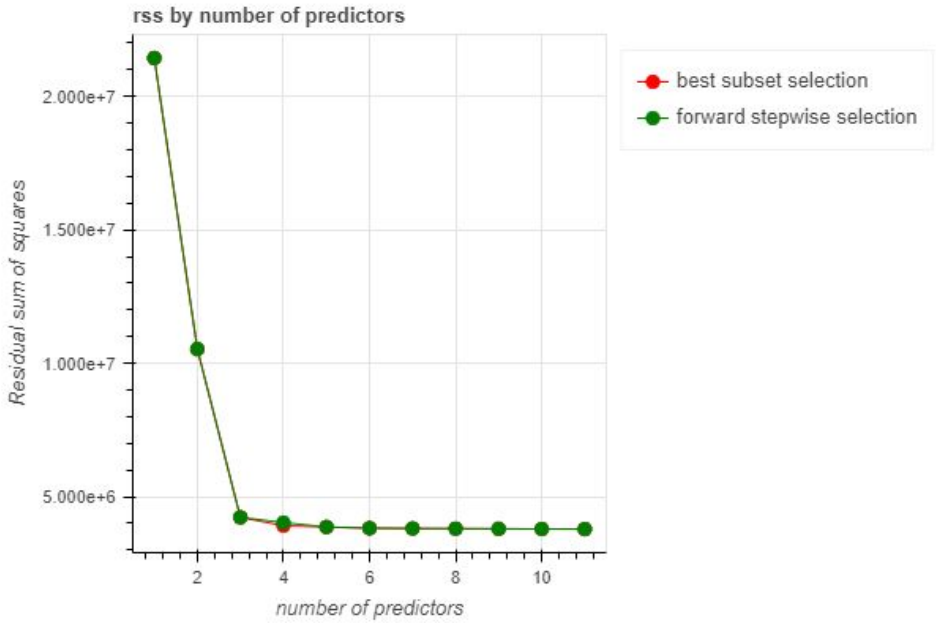


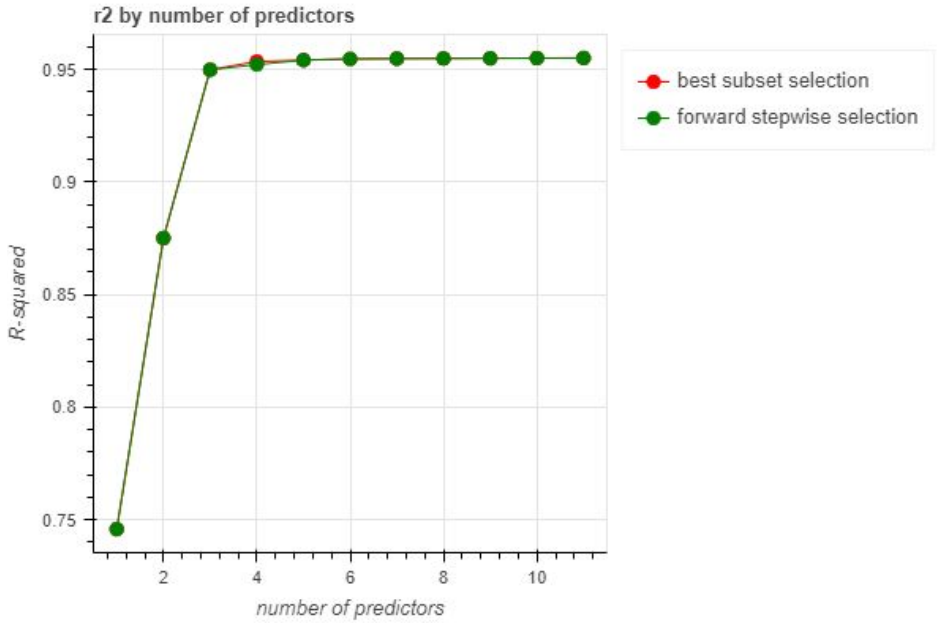
Forward stepwise selection

Algorithm 6.2 *Forward stepwise selection*

1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors.
2. For $k = 0, \ldots, p-1$:
   (a) Consider all $p - k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   (b) Choose the *best* among these $p - k$ models, and call it $\mathcal{M}_{k+1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Forward stepwise selection results - the Credit dataset



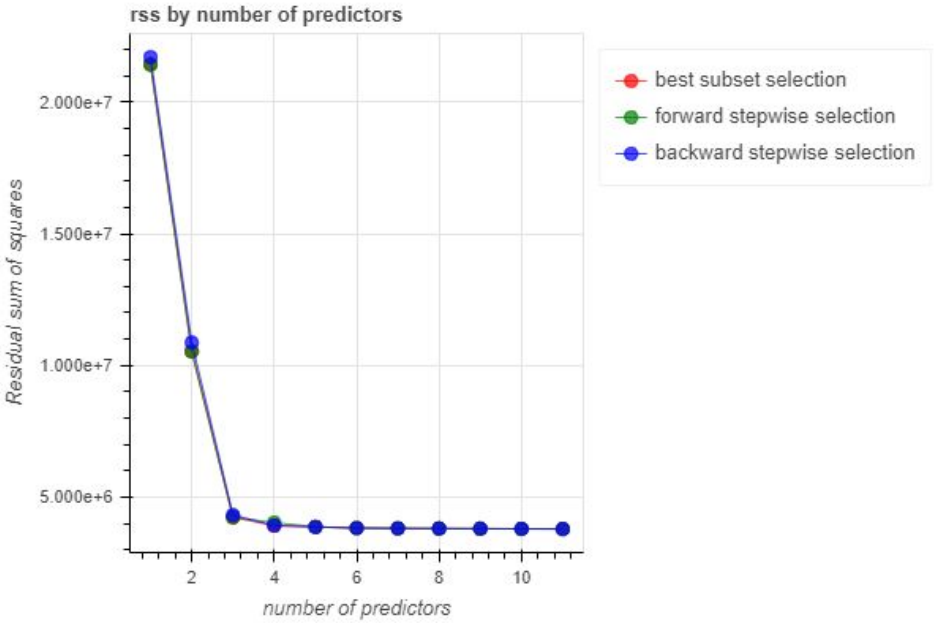


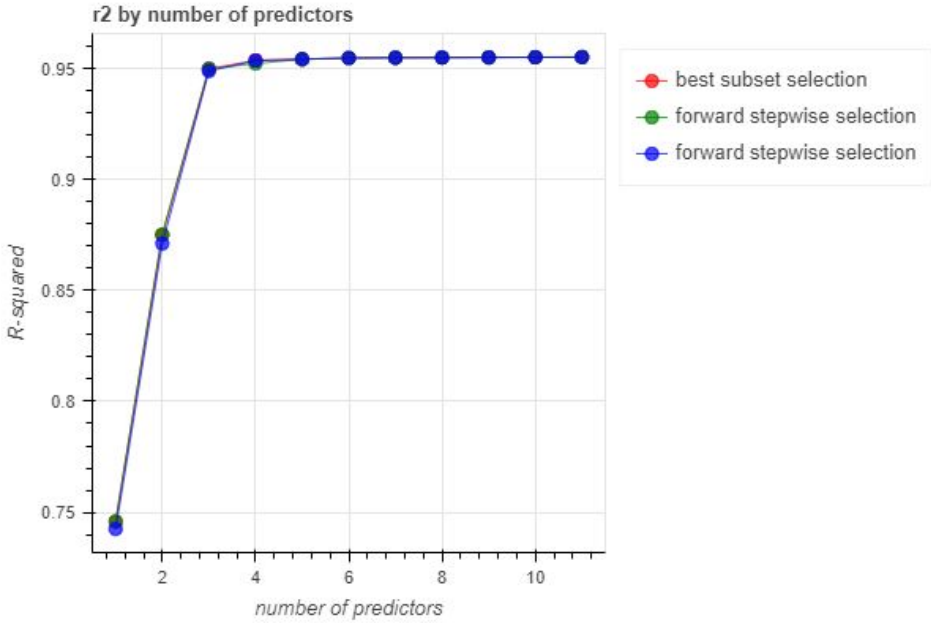
Backward stepwise selection

Algorithm 6.3 *Backward stepwise selection*

1. Let $\mathcal{M}_p$ denote the *full* model, which contains all p predictors.
2. For $k = p, p-1, \ldots, 1$:
   (a) Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
   (b) Choose the *best* among these k models, and call it $\mathcal{M}_{k-1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Backward stepwise selection - results from Credit dataset





Hybrid stepwise selection

Similar to forward stepwise selection, except that after adding a new variable to the model, we remove any existing variables that no longer (statistically significantly) contribute to explaining the response.

Observations

- Best subset selection is computational demanding because we have to fit 2^p models
- Stepwise selection methods have the computational advantage over best subset selection because they only have to fit 1 + p(p+1)/2 models.
- Forward and backward selection do not guarantee the best possible model out of 2^p models. Hybrid is getting closer to best subset selection while preserving the computational advantage of forward stepwise selection.
- Backward selection can only be used when n > p.
- RSE can be a better metric compared to RSS or R2 in selecting the best training model. Why?

Selecting the best model

Because RSS and R-squared are associated with the training error, they are not suitable to select the best model.

We need to choose the best model on test error, which can be estimated either

- Directly using cross-validation, or
- Indirectly by making adjustment to the training error (to account for the bias due to overfitting)

Selecting the best model - the indirect approach

$$C_p = \frac{1}{n}\left(\mathrm{RSS} + 2d\hat{\sigma}^2\right)$$

$$\mathrm{AIC} = \frac{1}{n\hat{\sigma}^2}\left(\mathrm{RSS} + 2d\hat{\sigma}^2\right)$$

$$\mathrm{BIC} = \frac{1}{n\hat{\sigma}^2}\left(\mathrm{RSS} + \log(n)d\hat{\sigma}^2\right)$$

$$\text{Adjusted } R^2 = 1 - \frac{\mathrm{RSS}/(n-d-1)}{\mathrm{TSS}/(n-1)}$$

Selecting the best model - using cross-validation

- Making fewer assumptions (and way simpler to understand) compared to Cp, AIC, BIC, adjusted R-squared
- Can be used for a wider range of model selection tasks
- Always select the simplest model among the models that have approximately equal cross-validation error => the *one-standard-error rule*.

Exercises

Implement the hybrid stepwise selection method in Python (or a language of your choice) combined with cross-validation to select the best subset of regression predictors for predicting Balance in the Credit dataset.