Selección de variables y regularización

Francesc Carmona

8 de mayo de 2019

Los ejercicios con (∗) son opcionales, con (∗∗) además son difíciles.

Ejercicios del libro de Faraway

1. (Ejercicio 1 cap. 10 pág. 159)

   Use the `prostate` data with `lpsa` as the response and the other variables as predictors. Implement the following variable selection methods to determine the "best" model:

   (a) Backward elimination

   (b) AIC

   (c) Adjusted R^2

   (d) Mallows C_p

2. (∗) (Ejercicio 2 cap. 10 pág. 159)

   Using the `teengamb` dataset with `gamble` as the response and the other variables as predictors, repeat the work of the first question.

3. (Ejercicio 3 cap. 10 pág. 159)

   Using the `divusa` dataset with `divorce` as the response and the other variables as predictors, repeat the work of the first question.

4. (Ejercicio 4 cap. 10 pág. 160)

   Using the `trees` data, fit a model with `log(Volume)` as the response and a second-order polynomial (including the interaction term) in `Girth` and `Height`. Determine whether the model may be reasonably simplified.

5. (∗) (Ejercicio 5 cap. 10 pág. 160)

   Fit a linear model to the `stackloss` data with `stack.loss` as the predictor and the other variables as predictors. Simplify the model if possible. Check the model for outliers and influential points. Now return to the full model, determine whether there are any outliers or influential points, eliminate them and then repeat the variable selection procedures.

6. (Ejercicio 6 cap. 10 pág. 160)

   Use the `seatpos` data with `hipcenter` as the response.

   (a) Fit a model with all eight predictors. Comment on the effect of leg length on the response.

   (b) Compute a 95% prediction interval for the mean value of the predictors.

   (c) Use AIC to select a model. Now interpret the effect of leg length and compute the prediction interval. Compare the conclusions from the two models.

7. (∗) (Ejercicio 8 cap. 10 pág. 160)

   Use the `odor` data with `odor` as the response.

   (a) Fit a second-order response surface model which contains all quadratic and cross-product terms of the three predictors. Explicitly include all nine terms in your model statement. You will need to "protect" these terms using the `I()` function. Show the regression summary.

   (b) Use the backward elimination method with a cut-off of 5% to select a smaller model.

   (c) Use the `step` method to select a model using AIC. Using this selected model determine the optimal values of the predictors to minimize odor.

   (d) Use the `polym` function to fit a second-order orthogonal polynomial regression model to the response using the three predictors. Confirm that the fit is identical to (a).

   (e) Apply the `step` procedure to this last model. What terms are considered for elimination? What model was selected? Compare the outcome with (c).

8. (Ejercicio 1 cap. 11 pág. 180)

   Using the `seatpos` data, perform a PCR analysis with `hipcenter` as the response and `HtShoes`, `Ht`, `Seated`, `Arm`, `Thigh` and `Leg` as predictors. Select an appropriate number of components and give an interpretation to those you choose. Add `Age` and `Weight` as predictors and repeat the analysis. Use both models to predict the response for predictors taking these values:

```
    Age    Weight   HtShoes        Ht    Seated
 64.800   263.700   181.080   178.560    91.440
    Arm     Thigh       Leg
 35.640    40.950    38.790
```

9. (Ejercicio 2 cap. 11 pág. 181)

   Fit a PLS model to the `seatpos` data with `hipcenter` as the response and all other variables as predictors. Take care to select an appropriate number of components. Use the model to predict the response at the values of the predictors specified in the first question.

10. (Ejercicio 3 cap. 11 pág. 181)

    Fit a ridge regression model to the `seatpos` data with `hipcenter` as the response and all other variables as predictors. Take care to select an appropriate amount of shrinkage. Use the model to predict the response at the values of the predictors specified in the first question.

11. (Ejercicio 4 cap. 11 pág. 181)

    Take the `fat` data, and use the percentage of body fat, `siri`, as the response and the other variables, except `brozek` and `density` as potential predictors. Remove every tenth observation from the data for use as a test sample. Use the remaining data as a training sample building the following models:

    (a) Linear regression with all predictors

    (b) Linear regression with variables selected using AIC

    (c) Principal component regression

    (d) Partial least squares

    (e) Ridge regression

    Use the models you find to predict the response in the test sample. Make a report on the performances of the models.

12. (Ejercicio 5 cap. 11 pág. 181)

    Some near infrared spectra on 60 samples of gasoline and corresponding octane numbers can be found by `data(gasoline, package="pls")`. Compute the mean value for each frequency and predict the response for the best model using the five different methods from Question 4.

13. ($*$) (Ejercicio 6 cap. 11 pág. 181)

    The dataset `kanga` contains data on the skulls of historical kangaroo specimens.

    (a) Compute a PCA on the 18 skull measurements. You will need to exclude observations with missing values. What percentage of variation is explained by the first principal component?

    (b) Provide the loadings for the first principal component. What variables are prominent?

    (c) Repeat the PCA but with the variables all scaled to the same standard deviation. How do the percentage of variation explained and the first principal component differ from those found in the previous PCA?

    (d) Give an interpretation of the second principal component.

    (e) Compute the Mahalanobis distances and plot appropriately to check for outliers.

    (f) Make a scatterplot of the first and second principal components using a different plotting symbol depending on the sex of the specimen. Do you think these two components would be effective in determining the sex of a skull?