UNIVERSITÀ DI ROMA TOR VERGATA

EEBL - Business Statistics

Revision - III

1 Where to study

G James, D Witten, T Hastie, and R Tibshirani and J Friedman. *An Introduction to Statistical Learning with Applications in R*. Springer, Springer Series in Statistics, 2009. Dowloadable at http://www-bcf.usc.edu/ gareth/ISL/

- For assessing model accuracy and predictive performance read sections 2.1.1, 2.1.2, 2.1.3, 2.1.4, and section 2.2. up to page 36,
- Cross-validation: see section 5.1, pages 176-183
- Linear model selection and regularization are dealt with in chapter 6. Study pages 203-215.

T Hastie, R Tibshirani and J Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition. Springer, Springer Series in Statistics, 2009.
Website: http://www-stat.stanford.edu/ElemStatLearn/.

- The linear regression model: chapter 3, sections 3.1, 3.2 up to page 51.
- Model evaluation and selection: chapter 7, sec. 7.1.–7.6. and 7.10. up to page 245.
- Subset selection: sections 3.3.1. and 3.3.2.
- Ridge regression: section 3.4.1.
- The Lasso: section 3.4.2.-3.4.3
- Principal Components Regression: section 3.5.1.

1.1 Exercises

1. The following table summarizes the results of estimating a linear regression model for house prices from a training sample of $N = 1080$ observations:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   4336.851  12084.856   0.359    0.720
sqft            97.862      3.485  28.082  < 2e-16 ***
Age           -694.671    141.292  -4.917 1.02e-06 ***
Pool           815.028   8926.466   0.091    0.927
Bedrooms    -20923.878   4567.168  -4.581 5.16e-06 ***
Fireplace      -97.130   5152.184  -0.019    0.985
Waterfront   63376.186   9389.874   6.749 2.43e-11 ***
DOM            -20.988     24.841  -0.845    0.398
---
Residual standard error: 76390 on 1072 degrees of freedom
Multiple R-squared: 0.6162,Adjusted R-squared: 0.6137
F-statistic: 245.9 on 7 and 1072 DF,  p-value: < 2.2e-16
```

- What is the interpretation of the p-value 0.40 associated to the explanatory variable `DOM` (days on the market)?
- What is the estimate of σ^2, the variance of the disturbance term in the regression model?
- Compute the value of the Bayesian Information Criterion, $BIC = \ln(RSS_p/N) + \frac{p+1}{N} \ln N$, using the information reported in the above summary table.
- What is the `F-statistic` in the last line meant for?
- Use a sentence to illustrate the role and the limitations of `Multiple R-squared` for assessing goodness of fit.

2. What is the meaning of the term "multicollinearity"? What are the consequences of multicollinearity?

3. Illustrate (in words) what happens when irrelevant variables are included as inputs in the regression model or relevant variables are omitted.

4. Consider the linear regression model for a training sample, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{y}$ is a vector of N observations on the output, and $\mathbf{X}$ is an $N \times (p+1)$ matrix of measurements on p inputs; the first column is a vector of 1's. The 'hat' matrix is defined as $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and the least squares residuals are $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$, where $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ are the fitted values. State which of the following statements are true:

   (a) The residuals have zero mean.

   (b) Ordinary least squares is always the best method for estimating $\boldsymbol{\beta}$.

   (c) The variance of the i-th training sample residual e_i is smaller than σ^2.

   (d) The residuals are uncorrelated with each of the p explanatory variables.

   (e) The residuals are uncorrelated with the variables not included in the model.

   (f) The fitted values have the same mean as the original observations y_i.

5. (Exam question) A large part of our course has been dedicated to model selection and evaluation.

   (a) What is meant by "bias-variance trade-off"?

   (b) Discuss the problem of measuring the predictive accuracy of a model, explaining the difference between the training error and the test error.

   (c) Present alternative ways of estimating the test error (out-of-sample predictive performance) (Mallow's C_p, cross-validation, etc.).

   (d) What is the main difference between AIC and BIC?

   (e) Discuss the pros and cons of the subset selection methodology known as forward stepwise.

6. *Exam question*.
   In the linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{y}$ is a vector of N standardised observations on the output, and $\mathbf{X}$ is an $N \times p$ matrix of measurements on p standardised inputs, the ridge estimator of the regression coefficients is
   $$\hat{\boldsymbol{\beta}}_r = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}.$$

   (a) Explain the rationale and the properties of this estimator, stressing the role of the shrinkage parameter λ.

   (b) Why do you need to standardise the inputs and the output?

   (c) What is LASSO, and how does it differ from ridge regression?

7. (Exam question) Illustrate cross-validation as a method for estimating the test error.