Multiple Linear Regression

Wilks, 2011: “Statistical Methods in the Atmospheric Sciences”, Sections 6.2 and 6.4

Multiple Linear Regression (MLR)

- There is still one single predictand, y, but there are multiple predictors (the number of predictors is denoted by K). The relationship between the predictand and the predictors can be approximated by a linear equation:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_K x_K .$$

- These K + 1 regression coefficients are often called the regression parameters. In addition, the parameter b0 is sometimes known as the regression constant.
- To develop the MLR model, we need to determine K+1 parameters, the intercept b_0 and the slopes for the individual predictors (b_1, .. b_k)
- The K + 1 parameters in MLR are found by minimizing the sum of squared residuals. This is achieved by simultaneously solving K + 1 equations using matrix algebra.

Wilks, 2011

Some questions you may have for MLR

- How do we select predictors?
- How many predictors shall we select?
- How do we assess the prediction skill?

*These questions apply to other statistical models with multiple predictors.

Why is Careful Predictor Selection Important?

- Predictor selection is not as simple as adding predictors until an apparently good relationship is achieved.
- Including too many predictor variables in a forecast equation may deteriorate the forecast performance.

Wilks, 2011

An example of overfitting

We will predict total winter snowfall at Ithaca from 1980 to 1986 using the year (yr) and four nonsense predictors: the U.S. federal deficit, the number of personnel in the U.S. Air Force, the sheep population of the U.S., and the average Scholastic Aptitude Test (SAT) scores of college-bound high-school students.

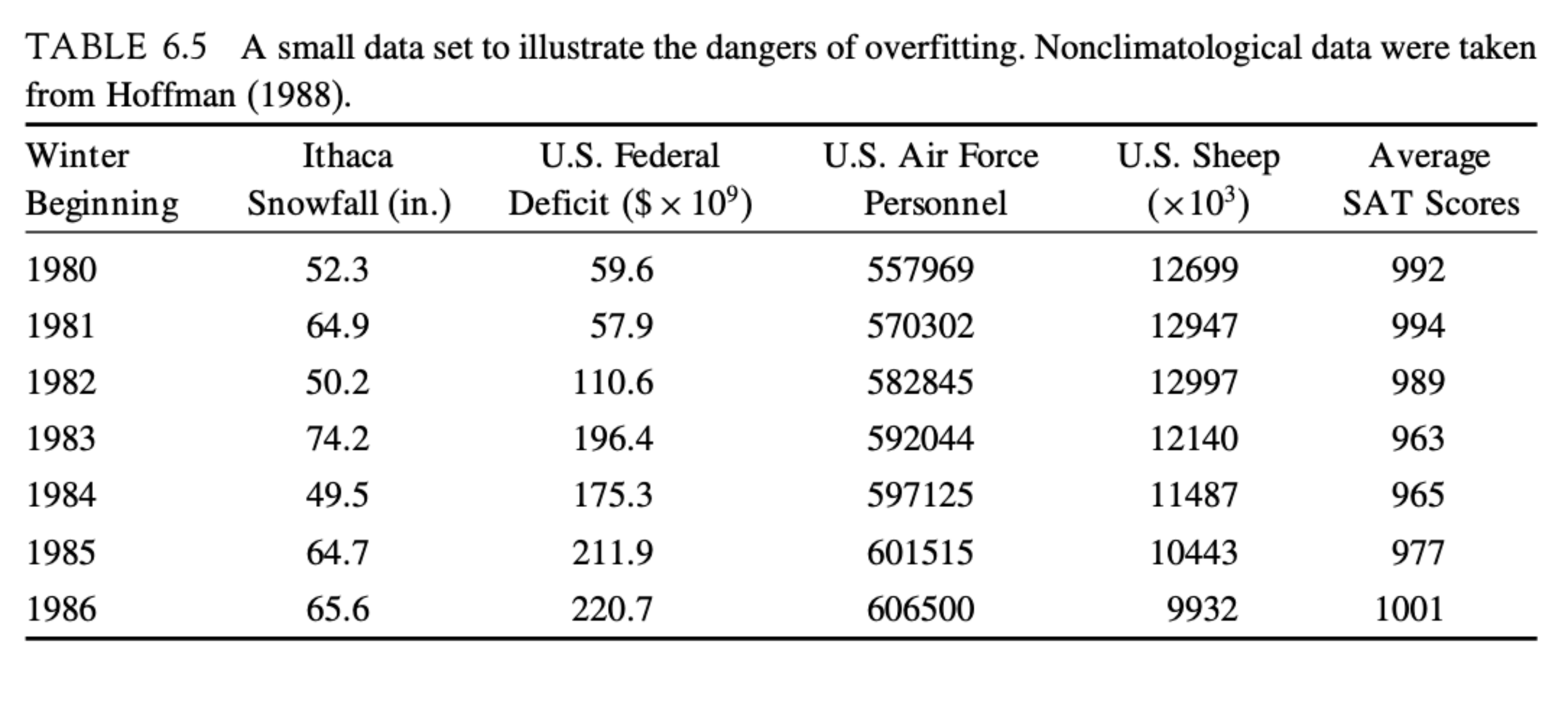

TABLE 6.5 A small data set to illustrate the dangers of overfitting. Nonclimatological data were taken from Hoffman (1988).

Winter Beginning	Ithaca Snowfall (in.)	U.S. Federal Deficit ($\$ \times 10^9$)	U.S. Air Force Personnel	U.S. Sheep ($\times 10^3$)	Average SAT Scores
1980	52.3	59.6	557969	12699	992
1981	64.9	57.9	570302	12947	994
1982	50.2	110.6	582845	12997	989
1983	74.2	196.4	592044	12140	963
1984	49.5	175.3	597125	11487	965
1985	64.7	211.9	601515	10443	977
1986	65.6	220.7	606500	9932	1001

Wilks, 2011

An example of overfitting (cont'd)

The regression will be fit using only the six winters 1980-1985, which is known as training data.

$$\text{Snow} = 1161771 - 601.7(\text{yr}) - 1.733(\text{deficit}) + 0.0567(\text{AF pers.}) \\ -0.3799(\text{sheep}) + 2.882(\text{SAT}).$$

which produces a perfect fit with MSE=0 and R^2=1.

However, the equation performs very poorly outside of the training sample and produce a meaningless forecast for negative snowfall in 1986.

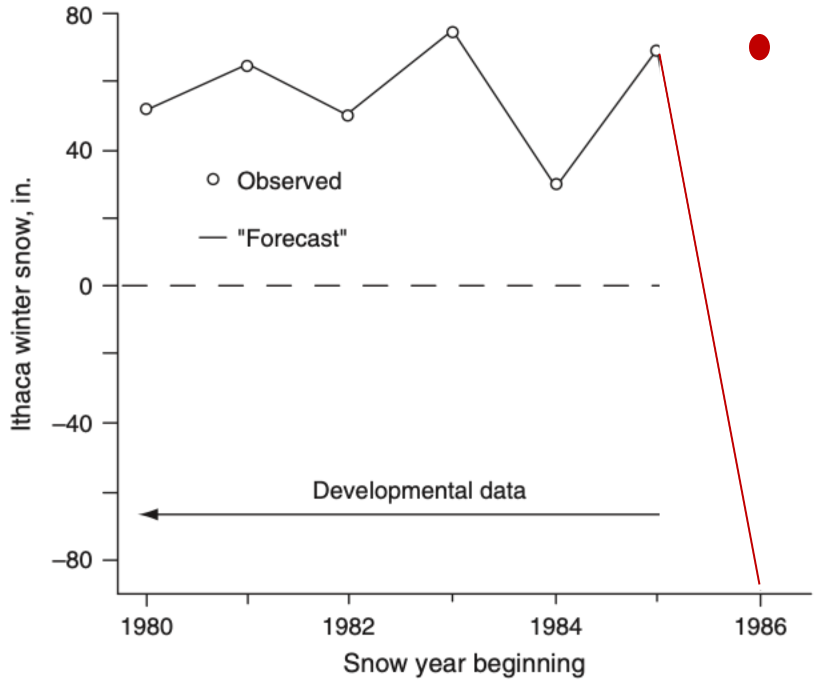

FIGURE 6.14 Forecasting Ithaca winter snowfall using the data in Table 6.5. The number of predictors is one fewer than the number of observations of the predictand in the developmental data, yielding perfect correspondence between the values specified by the regression and the data for this portion of the record. The relationship falls apart completely when used with the 1986 data, which was not used in equation development. The regression equation has been grossly overfit.

Wilks, 2011

General Statements on Overfitting

- More generally, any K = n − 1 predictors will produce a perfect regression fit to any predictand for which there are n observations.
  - For the case of n = 2, a straight line can be fit using any K = 1 predictor (simple linear regression), since a line can be found that will pass through any two points in the plane, and only an intercept and a slope are necessary to define a line.
  - Degree of freedom is equal to the sample size!
- The problem of overfitting is not limited to cases where nonsense predictors are used in a forecast equation and will be a problem when too many meaningful predictors are included as well.

Wilks, 2011

General Guidance

- Begin development of a regression equation by choosing only physically reasonable or meaningful potential predictors.
- A tentative regression equation needs to be tested on a sample of data not involved in its development.
  - A very large difference in performance between the dependent and independent samples would lead to the suspicion that the equation had been overfit.
- A reasonably large developmental sample is needed to develop a stable forecast equation.
  - Stability means that the fitted coefficients are also applicable to independent (i.e., future) data, so that the coefficients would be substantially unchanged if based on a different sample of the same kind of data.
- Avoid inclusion of predictors with strong mutual correlation
- The problem of selecting a good set of predictors from a pool of potential predictors is called screening regression.

Wilks, 2011

How do we select predictors: Forward selection

- Forward selection or stepwise regression
  1. All the possible M simple linear regressions between the available predictors and the predictand are computed, and that predictor whose linear regression is best among all candidate predictors is chosen as x1.
  2. Given the particular x1 chosen on the previous step, that predictor variable yielding the best regression yˆ = b0 + b1 x1 + b2 x2 is chosen as x2.
  3. At a subsequent step K, that member of the potential predictor pool not yet in the regression is chosen as xk that produces the best regression in conjunction with the K − 1 predictors chosen on previous steps.
- In general, when these regression equations are recomputed the regression coefficients for the intercept and for the previously chosen predictors will change, because the predictors usually are correlated to a greater or lesser degree.

Wilks, 2011

How do we select predictors: backward elimination

1. The initial point is a regression containing all M potential predictors, $y\hat{} = b_0 + b_1 x_1 + b_2 x_2 + \cdot \cdot \cdot + b_M x_M$
   - Backward elimination will not be computationally feasible if $M \geq n$.
2. At each step of the backward elimination procedure, the least important predictor variable is removed from the regression equation
   - The regression coefficients for the remaining variables require recomputation if (as is usually the case) the predictors are mutually correlated.

Wilks, 2011

Stopping Rules

- Both forward selection and backward elimination require a stopping criterion or stopping rule.
  - Without such a rule, forward selection would continue until all M candidate predictor variables were included in the regression equation, and backward elimination would continue until all predictors had been eliminated.
- The stopping criterion is somewhat subjective.
  - We might stop adding predictors in a forward selection when none of the remaining predictors would increase the R2 by a specified amount.
  - The stopping criterion can also be at the point where the MSE does not decline appreciably with the addition of more predictors.
- For forecasting purposes, the exact stopping point is not usually critical as long as it is approximately right.

Wilks, 2011

How do we assess the prediction skill?
Cross Validation

- Leave-one-out cross validation (sample size = n):
  - An observation is omitted, and the forecast model is developed using the remaining n-1 data values. The model is then used to predict the omitted value.
  - This is repeated for every observation, yielding a time series of the predictand.
  - Then one can calculate the MSE or R^2 between the observed predictand (y) and prediction y^.
- Cross validation can also be carried out for any number m of withheld data points and developmental data sets of size n−m
- Cross validation requires some special care when the data are serially correlated. A solution is to leave out blocks of an odd number of consecutive observations, L, and the cross-validation prediction is made only for that middle value.

Wilks, 2011

References

- Wilks, 2011: “Statistical Methods in the Atmospheric Sciences”, Sections 6.2 and 6.4