Chapter 6
What if n is not $>> p$?

When $p > n$, the fitted model should do better than i) interpolating the data or ii) discarding all of the predictors and using the location model of Section 1.3.5 for inference. If $p > n$, forward selection, lasso, relaxed lasso, elastic net, and relaxed elastic net can be useful for several regression models. Ridge regression, partial least squares, and principal components regression can also be computed for multiple linear regression. Sections 4.3, 5.9, and 10.7 give prediction intervals.

One of the **biggest errors in regression** is to use the response variable to build the regression model using all n cases, and then do inference as if the built model was selected without using the response, e.g., selected before gathering data. Using the response variable to build the model is called *data snooping*, then inference is generally no longer valid, and the model built from data snooping tends to fit the data too well. In particular, do not use data snooping and then use variable selection or cross validation. See Hastie et al (2009, p. 245) and Olive (2017a, pp. 85-89).

Building a regression model from data is one of the most challenging regression problems. The "final full model" will have response variable $Y = t(Z)$, a constant x_1, and predictor variables $x_2 = t_2(w_2, ..., w_r), ..., x_p = t_p(w_2, ..., w_r)$ where the initial data consists of $Z, w_2, ..., w_r$. Choosing $t, t_2, ..., t_p$ so that the final full model is a useful regression approximation to the data can be difficult.

As a rule of thumb, if strong nonlinearities are apparent in the predictors $w_2, ..., w_p$, it is often useful to remove the nonlinearities by transforming the predictors using power transformations. When p is large, a scatterplot matrix of $w_2, ..., w_p$ can not be made, but the log rule of Section 1.2 can be useful. Plots from Chapter 7, such as the DD plot, can also be useful. A scatterplot matrix of the w_i is an array of scatterplots of w_i versus w_j. A scatterplot is a plot of w_i versus w_j.

In the literature, it is sometimes stated that predictor transformations that are made without looking at the response are "free." The reasoning

is that the conditional distribution of $Y|(x_2 = a_2, ..., x_p = a_p)$ is the same as the conditional distribution of $Y|[t_2(x_2) = t_2(a_2), ..., t_p(x_p) = t_p(a_p)]$: there is simply a change of labelling. Certainly if $Y|x = 9 \sim N(0, 1)$, then $Y|\sqrt{x} = 3 \sim N(0, 1)$. To see that the above rule of thumb does not always work, suppose that $Y = \beta_1 + \beta_2 x_2 + \cdots + \beta_p x_p + e$ where the x_i are iid lognormal(0,1) random variables. Then $w_i = \log(x_i) \sim N(0, 1)$ for $i = 2, ..., p$ and the scatterplot matrix of the w_i will be linear while the scatterplot matrix of the x_i will show strong nonlinearities if the sample size is large. However, there is an MLR relationship between Y and the x_i while the relationship between Y and the w_i is nonlinear: $Y = \beta_1 + \beta_2 e^{w_2} + \cdots + \beta_p e^{w_p} + e \neq \boldsymbol{\beta}^T \boldsymbol{w} + e$. Given Y and the w_i with no information of the relationship, it would be difficult to find the exponential transformation and to estimate the β_i. The moral is that predictor transformations, especially the log transformation, can and often do greatly simplify the MLR analysis, but predictor transformations can turn a simple MLR analysis into a very complex nonlinear analysis.

Recall the 1D regression model from Definition 1.2 with

$$Y \perp\!\!\!\perp \boldsymbol{x} | SP \quad \text{or} \quad \mathrm{Y} \perp\!\!\!\perp \boldsymbol{x} | \mathrm{h}(\boldsymbol{x}),$$

where the real valued function $h : \mathbb{R}^p \to \mathbb{R}$. An important special case is a model with a linear predictor $h(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{\beta}$.

For the 1D regression model, let the ith case be $(Y_i, \boldsymbol{x}_i)$ for $i = 1, ..., n$ where the n cases are independent. Variable selection is the search for a subset of predictor variables that can be deleted with little loss of information if n/p is large, and so that the model with the remaining predictors is useful for prediction even if n/p is not large. The *model for variable selection* given by Equation (4.1) can be useful even if n/p is not large:

$$\boldsymbol{x}^T \boldsymbol{\beta} = \boldsymbol{x}_S^T \boldsymbol{\beta}_S + \boldsymbol{x}_E^T \boldsymbol{\beta}_E = \boldsymbol{x}_S^T \boldsymbol{\beta}_S \tag{6.1}$$

where $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$, $\boldsymbol{x}_S$ is an $a_S \times 1$ vector, and $\boldsymbol{x}_E$ is a $(p - a_S) \times 1$ vector. Given that $\boldsymbol{x}_S$ is in the model, $\boldsymbol{\beta}_E = \boldsymbol{0}$ and E denotes the subset of terms that can be eliminated given that the subset S is in the model. Let $\boldsymbol{x}_I$ be the vector of a terms from a candidate subset indexed by I, and let $\boldsymbol{x}_O$ be the vector of the remaining predictors (out of the candidate submodel). Suppose that S is a subset of I and that model (6.1) holds. Then

$$\boldsymbol{x}^T \boldsymbol{\beta} = \boldsymbol{x}_S^T \boldsymbol{\beta}_S = \boldsymbol{x}_S^T \boldsymbol{\beta}_S + \boldsymbol{x}_{I/S}^T \boldsymbol{\beta}_{(I/S)} + \boldsymbol{x}_O^T \boldsymbol{0} = \boldsymbol{x}_I^T \boldsymbol{\beta}_I$$

where $\boldsymbol{x}_{I/S}$ denotes the predictors in I that are not in S. Since this is true regardless of the values of the predictors, $\boldsymbol{\beta}_O = \boldsymbol{0}$ if $S \subseteq I$.

6.1 Sparse Models

When $n/p \to 0$ as $n \to \infty$, consistent estimators generally cannot be found unless the model has a simplifying structure. A sparse model is one such structure. For Equation (6.1), a population regression model is *sparse* if a_S is small. We want $n \geq 10a_S$.

For multiple linear regression with $p > n$, results from Hastie et al. (2015, pp. 20, 296, ch. 6, ch. 11) and Luo and Chen (2013) suggest that lasso, relaxed lasso, and forward selection with EBIC can perform well for sparse models. Least angle regression, elastic net, and relaxed elastic net can also be useful.

Suppose the selected model is I_d, and $\boldsymbol{\beta}_{I_d}$ is $a_d \times 1$. For multiple linear regression, forward selection with C_p and AIC often gives useful results if $n \geq 5p$ and if the final model I has $n \geq 10a_d$. For $p < n < 5p$, forward selection with C_p and AIC tends to pick the full model (which overfits since $n < 5p$) too often, especially if $\hat{\sigma}^2 = MSE$. The Hurvich and Tsai (1989) AIC_C criterion can be useful for MLR and time series if $n \geq \max(2p, 10a_d)$. If $n \geq 5p$, AIC and BIC are useful for many regression models, and forward selection with EBIC can be used for some models if n/p is small. See Section 4.1 and Chen and Chen (2008).

6.2 Data Splitting

Data splitting is useful for many regression models when the n cases are independent, including multiple linear regression, multivariate linear regression where there are $m \geq 2$ response variables, generalized linear models (GLMs), the Cox (1972) proportional hazards regression model, and parametric survival regression models.

Consider a regression model with response variable Y and a $p \times 1$ vector of predictors $\boldsymbol{x}$. This model is the full model. Suppose the n cases are independent. To perform data splitting, randomly divide the data into two sets H and V where H has n_H of the cases and V has the remaining $n_V = n - n_H$ cases $i_1, ..., i_{n_V}$. Find a model I, possibly with data snooping or model selection, using the data in the training set H. Use the model I as the full model to perform inference using the data in the validation set V. That is, regress Y_V on $\boldsymbol{X}_{V,I}$ and perform the usual inference for the model using the $j = 1, ..., n_V$ cases in the validation set V. If $\boldsymbol{\beta}_I$ uses a predictors, we want $n_V \geq 10a$ and we want $P(S \subseteq I) \to 1$ as $n \to \infty$ or for $(Y_V, \boldsymbol{X}_{V,I})$ to follow the regression model.

In the literature, often $n_H \approx \lceil n/2 \rceil$. For model selection, use the training data set to fit the model selection method, e.g. forward selection or lasso, to get the a predictors. On the test set, use the standard regression inference from regressing the response on the predictors found from the training set. This method can be inefficient if $n \geq 10p$, but is useful for a sparse model

if $n \leq 5p$, if the probability that the model underfits goes to zero, and if $n \geq 20a$.

The method is simple, use one half set to get the predictors, then fit the regression model, such as a GLM or OLS, to the validation half set $(\boldsymbol{Y}_V, \boldsymbol{X}_{V,I})$. The regression model needs to hold for $(\boldsymbol{Y}_V, \boldsymbol{X}_{V,I})$ and we want $n_V \geq 10a$ if I uses a predictors. The regression model can hold if $S \subseteq I$ and the model is sparse. Let $\boldsymbol{x} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_p)^T$ where $\boldsymbol{x}_1$ is a constant. If $(Y, \boldsymbol{x}_2, ..., \boldsymbol{x}_p)^T$ follows a multivariate normal distribution, then $(Y, \boldsymbol{x}_I)$ follows a multiple linear regression model for every I. Hence the full model need not be sparse, although the selected model may be suboptimal.

Of course other sample sizes than half sets could be used. For example if $n = 1000p$, use $n = 10p$ for the training set and $n = 990p$ for the validation set.

Remark 6.1. i) One use of data splitting is to try to transform the $p \geq n$ problem into an $n \geq 10k$ problem. This method can work if the model is sparse. For multiple linear regression, this method can work if $\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I})$, since then all subsets I satisfy the MLR model: $Y_i = \boldsymbol{x}_{I,i}^T \boldsymbol{\beta}_I + \boldsymbol{e}_{I,i}$. See Remark 1.5. If $\boldsymbol{\beta}_I$ is $k \times 1$, we want $n \geq 10k$ and $V(e_{I,i}) = \sigma_I^2$ to be small. For binary logistic regression, the discriminant function model of Definition 10.7 can be useful if $\boldsymbol{x}_I | Y = j \sim N_k(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$ for $j = 0, 1$. Of course, the models may not be sparse, and the multivariate normal assumptions for MLR and binary logistic regression rarely hold.

ii) Data splitting can be tricky for lasso, ridge regression, and elastic net if the sample sizes of the training and validation sets differ. Roughly set $\lambda_{1,n_1}/(2n_1) = \lambda_{2,n_2}/(2n_2)$. Data splitting is much easier for variable selection methods such as forward selection, relaxed lasso, and relaxed elastic net. Find the variables $x_1^*, ..., x_k^*$ indexed by I from the training set, and use model I as the full model for the validation set.

iii) Another use of data splitting is that data snooping can be used on the training set: use the model as the full model for the validation set.

6.3 Summary

1) Using the response variable to build a model is known as data snooping, and invalidates inference if data snooping is used on the entire data set of n cases.

2) Suppose $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E = \boldsymbol{x}_S^T\boldsymbol{\beta}_S$ where $\boldsymbol{\beta}_S$ is an $a_S \times 1$ vector. A regression model is sparse if a_S is small. We want $n \geq 10a_S$.

3) Assume the cases are independent. To perform data splitting, randomly divide the data into two half sets H and V where H has n_H of the cases and V has the remaining $n_V = n - n_H$ cases $i_1, ..., i_{n_V}$. Build the model, possibly with data snooping, or perform variable selection to Find a model I, possibly with data snooping or model selection, using the data in the training set H.

Use the model I as the full model to perform inference using the data in the validation set V.

6.4 Complements

Suppose model I_k contains k predictors including a constant. For multiple linear regression, the forward selection algorithm in Chapter 4 adds a predictor x_{k+1}^* that minimizes the residual sum of squares, while the Pati et al. (1993) "orthogonal matching pursuit algorithm" uses predictors (scaled to have unit norm: $\boldsymbol{x}_i^T \boldsymbol{x}_i = 1$ for the nontrivial predictors), and adds the scaled predictor x_{k+1}^* that maximizes $|\boldsymbol{x}_{k+1}^{*T} \boldsymbol{r}_k|$ where the maximization is over variables not yet selected and the $\boldsymbol{r}_k$ are the OLS residuals from regressing Y on $\boldsymbol{X}_{I_k}^*$. Fan and Li (2001) and Candes and Tao (2007) gave competitors to lasso. Some fast methods seem similar to the first PLS component. A useful reference for data splitting is Rinaldo et al (2019).

Fan and Li (2002) give a method of variable selection for the Cox (1972) proportional hazards regression model. Using AIC is also useful if p is fixed.

For a time series $Y_1, ..., Y_n$, we could use $Y_1, ..., Y_m$ as one set and $Y_{m+1}, ..., Y_n$ as the other set. Three set inference may be needed. Use $Y_1, ..., Y_m$ as the first set (trianing data), $Y_{m+1}, ..., Y_{m+k}$ as a burn in set, and $Y_{m+k+1}, ..., Y_n$ as the third set for inference.

When the entire data set is used to build a model with the response variable, the inference tends to be invalid, and cross validation should not be used to check the model. See Hastie et al. (2009, p. 245). In order for the inference and cross validation to be useful, the response variable and the predictors for the regression should be chosen before looking at the response variable. Predictor transformations can be done as long as the response variable is not used to choose the transformation. You can do model building on the test set, and then inference for the chosen (built) model as the full model with the validation set, provided this model follows the regression model used for inference (e.g. multiple linear regression or a GLM). This process is difficult to simulate.

6.5 Problems