Part 12: Effects of Predictor Relationships

Text references: §7.1, 10.5 in KNN
Shalizi book, Sec 2.1
S-W notes Lecture 14, 17

As the name implies, multiple regression involves including multiple predictors in the model. Although this can lead to much better predictions of the response, tricky issues and counterintuitive results can arise when predictors have relationships amongst themselves.

We will explore some different issues and tools in this regard here.

Exercise: Suppose that one is conducting a study in which the response variable is blood pressure (Y), while two available predictors are (1) weight at birth (X_1), and (2) current weight (X_2). What are the possible effects of fitting a model with just weight at birth versus a model that includes both weight predictors?

Model 1: $Y = \beta_0 + \beta_1 X_1 + \varepsilon$

Model 2: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ → Same β_1 ?

BP b.w. c.w

No, almost anything is possible

A relationship between BP (Y) and b.w. (X_1) could change once c.w. (X_2) is included in the model because $Cov(X_1, X_2) \neq 0$ (and $cov(X_2, Y) \neq 0$)

From Wasserman, "All of Statistics"

The MLE $\widehat{\beta}$ has to be obtained by maximizing $\mathcal{L}(\beta)$ numerically. There is a fast numerical algorithm called reweighted least squares. The steps are as follows:

Reweighted Least Squares Algorithm

Choose starting values $\widehat{\beta}^0 = (\widehat{\beta}_0^0, \ldots, \widehat{\beta}_k^0)$ and compute p_i^0 using equation (13.32), for $i = 1, \ldots, n$. Set $s = 0$ and iterate the following steps until convergence.

1. Set
$$Z_i = \text{logit}(p_i^s) + \frac{Y_i - p_i^s}{p_i^s(1-p_i^s)}, \quad i = 1, \ldots, n.$$
2. Let W be a diagonal matrix with (i,i) element equal to $p_i^s(1-p_i^s)$.
3. Set
$$\widehat{\beta}^s = (X^T W X)^{-1} X^T W Y.$$
This corresponds to doing a (weighted) linear regression of Z on Y.
4. Set $s = s + 1$ and go back to the first step.

The Fisher information matrix I can also be obtained numerically. The estimate standard error of $\widehat{\beta}_j$ is the (j,j) element of $J = I^{-1}$. Model selection is usually done using the AIC score $\ell_S - |S|$.

13.17 Example. The Coronary Risk-Factor Study (CORIS) data involve 462 males between the ages of 15 and 64 from three rural areas in South Africa, (Rousseauw et al. (1983)). The outcome Y is the presence ($Y = 1$) or absence ($Y = 0$) of coronary heart disease. There are 9 covariates: systolic blood pressure, cumulative tobacco (kg), ldl (low density lipoprotein cholesterol), adiposity, famhist (family history of heart disease), typea (type-A behavior), obesity, alcohol (current alcohol consumption), and age. A logistic regression yields the following estimates and Wald statistics W_j for the coefficients:

Covariate	$\widehat{\beta}_j$	$\widehat{\text{se}}$	W_j	p-value
Intercept	-6.145	1.300	-4.738	0.000
sbp	0.007	0.006	1.138	0.255
tobacco	0.079	0.027	2.991	0.003
ldl	0.174	0.059	2.925	0.003
adiposity	0.019	0.029	0.637	0.524
famhist	0.925	0.227	4.078	0.000
typea	0.040	0.012	3.233	0.001
obesity	-0.063	0.044	-1.427	0.153
alcohol	0.000	0.004	0.027	0.979
age	0.045	0.012	3.754	0.000

Are you surprised by the fact that systolic blood pressure is not significant or by the minus sign for the obesity coefficient? If yes, then you are confusing association and causation. This issue is discussed in Chapter 16. The fact that blood pressure is not significant does not mean that blood pressure is not an important *cause* of heart disease. It means that it is not an important *predictor* of heart disease relative to the other variables in the model. ■

13.8 Bibliographic Remarks

A succinct book on linear regression is Weisberg (1985). A data-mining view of regression is given in Hastie et al. (2001). The Akaike Information Criterion (AIC) is due to Akaike (1973). The Bayesian Information Criterion (BIC) is due to Schwarz (1978). References on logistic regression include Agresti (1990) and Dobson (2001).

13.9 Appendix

THE AKAIKE INFORMATION CRITERION (AIC). Consider a set of models $\{M_1, M_2, \ldots\}$. Let $\widehat{f}_j(x)$ denote the estimated probability function obtained by using the maximum likelihood estimator of model M_j. Thus, $\widehat{f}_j(x) = \widehat{f}(x; \widehat{\beta}_j)$ where $\widehat{\beta}_j$ is the MLE of the set of parameters β_j for model M_j. We will use the loss function $D(f, \widehat{f})$ where

$$D(f, g) = \sum_x f(x) \log\left(\frac{f(x)}{g(x)}\right)$$

is the Kullback-Leibler distance between two probability functions. The corresponding risk function is $R(f, \widehat{f}) = \mathbb{E}(D(f, \widehat{f})$. Notice that $D(f, \widehat{f}) = c -$

Chapter 2

The Truth about Linear Regression

We need to say some more about how linear regression, and especially about how it really works and how it can fail. Linear regression is important because

1. it's a fairly straightforward technique which sometimes works tolerably for prediction;
2. it's a simple foundation for some more sophisticated techniques;
3. it's a standard method so people use it to communicate; and
4. it's a standard method so people have come to confuse it with prediction and even with causal inference as such.

We need to go over (1)–(3), and provide prophylaxis against (4).

2.1 Optimal Linear Prediction: Multiple Variables

We have a response variable Y and a p-dimensional vector of predictor variables or features $\vec{X}$. We would like to predict Y using $\vec{X}$. We saw last time that the best predictor we could use, at least in a mean-squared sense, is the conditional expectation,

$$\mu(\vec{x}) = \mathbb{E}\left[Y|\vec{X} = \vec{x}\right] \tag{2.1}$$

Instead of using the optimal predictor $\mu(\vec{x})$, let's try to predict as well as possible while using only a linear function of $\vec{x}$, say $\beta_0 + \beta \cdot \vec{x}$. This is not an assumption about the world, but rather a decision on our part; a choice, not a hypothesis. This decision can be good — $\beta_0 + \vec{x} \cdot \beta$ could be a tolerable approximation to $\mu(\vec{x})$ — even if the linear hypothesis is strictly wrong. It might also be that no linear approximation to μ is much good mathematically, but we might be forced to make it for practical reasons, e.g., speed of computation.

Perhaps the best reason to hope the choice to use a linear model isn't crazy is that we may hope μ is a smooth function. If it is, then we can Taylor expand it about our favorite point, say $\vec{u}$:

$$\mu(\vec{x}) = \mu(\vec{u}) + \sum_{i=1}^{p} \left(\left. \frac{\partial \mu}{\partial x_i} \right|_{\vec{u}} \right) (x_i - u_i) + O(||\vec{x} - \vec{u}||^2) \tag{2.2}$$

or, in the more compact vector-calculus notation,

$$\mu(\vec{x}) = \mu(\vec{u}) + (\vec{x} - \vec{u}) \cdot \nabla \mu(\vec{u}) + O(||\vec{x} - \vec{u}||^2) \tag{2.3}$$

If we only look at points $\vec{x}$ which are close to $\vec{u}$, then the remainder terms $O(||\vec{x} - \vec{u}||^2)$ are small, and a linear approximation is a good one[1]. Here, "close to $\vec{u}$" really means "so close that all the non-linear terms in the Taylor series are comparatively negligible".

Of course there are lots of linear functions so we need to pick one, and we may as well do that by minimizing mean-squared error again:

$$MSE(\beta) = \mathbb{E}\left[\left(Y - \beta_0 - \vec{X} \cdot \beta \right)^2 \right] \tag{2.4}$$

Going through the optimization is parallel to the one-dimensional case (see last chapter), with the conclusion that the optimal β is

$$\beta = \mathbf{v}^{-1} \mathrm{Cov}\left[\vec{X}, Y\right] \tag{2.5}$$

where $\mathbf{v}$ is the covariance matrix of $\vec{X}$, i.e., $v_{ij} = \mathrm{Cov}\left[X_i, X_j\right]$, and $\mathrm{Cov}\left[\vec{X}, Y\right]$ is the vector of covariances between the predictor variables and Y, i.e. $\mathrm{Cov}\left[\vec{X}, Y\right]_i = \mathrm{Cov}\left[X_i, Y\right]$. We also get

$$\beta_0 = \mathbb{E}[Y] - \beta \cdot \mathbb{E}\left[\vec{X}\right] \tag{2.6}$$

just as in the one-dimensional case (Exercise 1).

Multiple regression would be a lot simpler if we could just do a simple regression for each predictor variable, and add them up; but really, this is what multiple regression *does*, just in a disguised form. If the input variables are uncorrelated, $\mathbf{v}$ is diagonal ($v_{ij} = 0$ unless $i = j$), and so is $\mathbf{v}^{-1}$. Then doing multiple regression breaks up into a sum of separate simple regressions across each input variable. When the input variables are correlated and $\mathbf{v}$ is not diagonal, we can think of the multiplication by $\mathbf{v}^{-1}$ as **de-correlating** $\vec{X}$ — applying a linear transformation to come up with a new set of inputs which are uncorrelated with each other.[2]

[1] If you are not familiar with the big-O notation like $O(||\vec{x} - \vec{u}||^2)$, now would be a good time to read Appendix C.

[2] If $\vec{Z}$ is a random vector with covariance matrix I, then $\mathbf{w}\vec{Z}$ is a random vector with covariance matrix $\mathbf{w}^T\mathbf{w}$. Conversely, if we start with a random vector $\vec{X}$ with covariance matrix $\mathbf{v}$, the latter has a "square root" $\mathbf{v}^{1/2}$ (i.e., $\mathbf{v}^{1/2}\mathbf{v}^{1/2} = \mathbf{v}$), and $\mathbf{v}^{-1/2}\vec{X}$ will be a random vector with covariance matrix $\mathbf{I}$. When we write our predictions as $\vec{x}\mathbf{v}^{-1}\mathrm{Cov}\left[\vec{X}, Y\right]$, we should think of this as $\left(\vec{x}\mathbf{v}^{-1/2}\right)\left(\mathbf{v}^{-1/2}\mathrm{Cov}\left[\vec{X}, Y\right]\right)$. We use one power of $\mathbf{v}^{-1/2}$ to transform the input features into uncorrelated variables before taking their correlations with the response, and the other power to decorrelate $\vec{X}$.

Review: [Cf. Part 3, doc cam pp. 12-13]

What's $\vec{\beta}$ and $\hat{\vec{\beta}}$?
(in multiple linear regression)

Ref. Shalizi book p.53

Model: $\boxed{Y = f(\vec{X}) + \varepsilon}$

$\vec{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix}$ — random vector; $X_1, \dots, X_p$ — random variables; ε — random

Minimize true MSE, $MSE(f) = E\left[(Y - f(\vec{X}))^2\right]$

then the optimal predictor $\boxed{f(\vec{x}) = E(Y | \vec{X} = \vec{x})}$

"regression function"

Now assume linear predictor $f(\vec{x}) = \vec{\beta}^T \vec{x}$ (2) $\quad \vec{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$

If we minimize $MSE(\vec{\beta}) = E\left[(Y - \vec{\beta}^T \vec{X})^2\right]$

then the solution is the "optimal" linear predictor of Y

$$\begin{cases} f(\vec{x}) = \vec{\beta}^T \vec{x} \\ \vec{\beta} = [Var(\vec{X})]^{-1} Cov(\vec{X}, Y) \end{cases}$$

$\vec{\beta}$ is a fixed vector (unknown) — Shalizi 2.4, 2.5

Compare this with the L.S. estimator

$$\hat{\vec{\beta}} = (X^T X)^{-1} (X^T \vec{Y}) \qquad X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

approx $\approx (\hat{Var}(\vec{X}))^{-1} \cdot \hat{Cov}(\vec{X}, Y)$

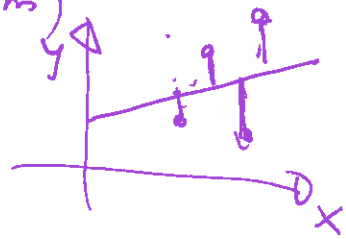
Recall: $\hat{\vec{\beta}} = (X^T X)^{-1}(X^T \vec{Y})$ ← estimator; ← random vector

minimizes the empirical MSE or $\frac{RSS}{n-q}$

$$\widehat{MSE}(\vec{\beta}) = \hat{E}\left[\left(Y - \vec{\beta}^T \vec{X}\right)^2\right]$$

Computed from data i.e. n observations

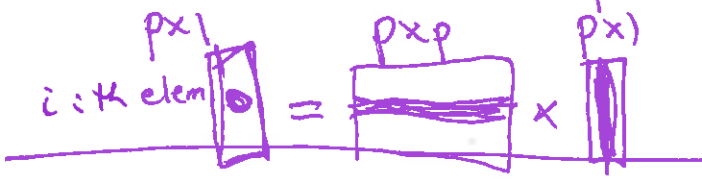
Y: r.v.; $\vec{\beta}$: fixed vector; $\vec{X}$: (fixed or random) vector



$$= \frac{1}{n-q}\sum_{i=1}^{n}\left(Y_i - \vec{\beta}^T \vec{X}_i\right)^2 = \frac{1}{n-q}\left\|\vec{Y} - X\vec{\beta}\right\|^2$$

$$= \frac{RSS}{n-q}$$

$$\vec{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$$



$$\vec{\beta} = \underbrace{\left[\mathrm{var}(\vec{X})\right]^{-1}}_{\dim: p\times p}\ \underbrace{\mathrm{cov}(\vec{X}, Y)}_{p\times 1\text{-vector}}$$

[Shalizi 2.5]

$$\left[\mathrm{var}(\vec{X})\right]_{i,j} = \mathrm{cov}(X_i, X_j)$$

predictors X_i, X_j

$$\left[\mathrm{cov}(\vec{X}, Y)\right]_i = \mathrm{cov}(X_i, Y)$$

$$\beta_i = \sum_{k=1}^{p}\left[\mathrm{var}(\vec{X})^{-1}\right]_{ik}\left[\mathrm{cov}(\vec{X}, Y)\right]_k \qquad i = 1, \dots, p$$

If the predictors are uncorrelated

then $\text{var}(\vec{X}) = \begin{bmatrix} \text{var}(X_1) & & 0 \\ & \ddots & \\ 0 & & \text{var}(X_p) \end{bmatrix}$

$$\left(\text{var}(\vec{X})\right)^{-1} = \begin{bmatrix} \frac{1}{\text{var}(X_1)} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\text{var}(X_p)} \end{bmatrix}$$

Spec. case (SLR)

$$\beta_i = \frac{\text{cov}(X_i, Y)}{\text{var}(X_i)}$$

This is just the same β as in simple linear regression with one predictor only

Otherwise, + more generally, β_i depends on $\text{cov}(X_i, X_k)$ and $\text{cov}(X_k, Y)$ for all $k \neq i$ — Important!