

## 21 Lecture 24: March 27

### Last time

- Diagnosing non-normality, non-constant error variance, and nonlinearity (JF chapter 12)

### Today

- lecture on Friday
- Data transformation
- Collinearity (JF chapter 13, RD 8.3.2)
- Principal component analysis (JF 13.1.1, RD 8.3.4)

### Box-Cox transformation of $Y$

Box and Cox (1964) suggested a power transformation of  $Y$  with the object of normalizing the error distribution, stabilizing the error variance, and straightening the relationship of  $Y$  to the  $X$ s. The general Box-Cox model is

$$Y_i^{(\lambda)} = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i$$

where  $\epsilon_i \stackrel{iid}{\sim} N(0, \sigma_\epsilon^2)$ , and

$$Y_i^{(\lambda)} = \begin{cases} \frac{Y_i^\lambda - 1}{\lambda} & \text{for } \lambda \neq 0 \\ \log_e Y_i & \text{for } \lambda = 0 \end{cases}$$

Note: in statistics,  $\log_e$  is often written as  $\log$ .

For a particular choice of  $\lambda$ , the conditional maximized log-likelihood (see JF 12.5.1 p.324 footnote 55) is

$$\begin{aligned} \log_e L(\beta_0, \beta_1, \dots, \beta_p, \sigma_\epsilon^2 | \lambda) &= -\frac{n}{2} (1 + \log_e 2\pi) \\ &\quad - \frac{n}{2} \log_e \hat{\sigma}_\epsilon^2(\lambda) + (\lambda - 1) \sum_{i=1}^n \log_e Y_i \end{aligned}$$

where  $\hat{\sigma}_\epsilon^2(\lambda) = \sum \hat{\epsilon}_i^2(\lambda)/n$  and where  $\hat{\epsilon}_i(\lambda)$  are the residuals from the least-squares regression of  $Y^{(\lambda)}$  on  $X$ s. The least-squares coefficients from this regression are the maximum-likelihood estimates of  $\beta$ s conditional on the values of  $\lambda$ .

A simple procedure for finding the maximum-likelihood estimator  $\hat{\lambda}$  is to evaluate the maximized  $\log_e L$  (called the profile log-likelihood) for a range of values of  $\lambda$ . To test:  $H_0 : \lambda = 1$ , calculated the likelihood-ratio statistic

$$G_0^2 = -2[\log_e L(\lambda = 1) - \log_e L(\lambda = \hat{\lambda})]$$

which is asymptotically distributed as  $\chi_1^2$  with one degree of freedom under  $H_0$ . A 95% confidence interval for  $\lambda$  includes those values for which

$$\log_e L(\lambda) > \log_e L(\lambda = \hat{\lambda}) - 1.92$$

The number 1.92 comes from  $\frac{1}{2}\chi_{1,0.05}^2 = 0.5 \times 1.96^2$ .

Figure 21.1 shows a plot of the profile log-likelihood against  $\lambda$  for the original SLID regression of composite hourly wages on sex, age, and education. The maximum-likelihood estimate of  $\lambda$  is  $\hat{\lambda} = 0.09$ , and a 95% confidence interval runs from 0.04 to 0.13. Although 0 is outside of the CI (confidence interval), it is essentially the same as log transformation of wages (the correlation between log wages and wages<sup>0.09</sup> is 0.9996).

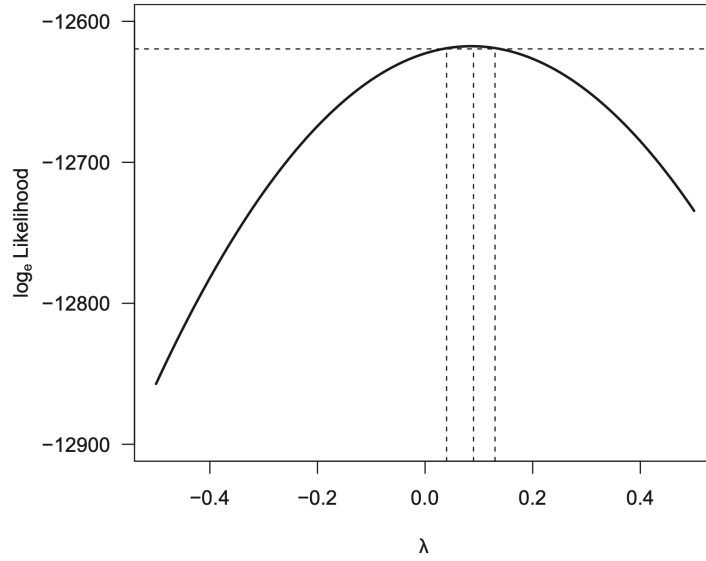


Figure 21.1: Box-Cox transformations for the SLID regression of wages on sex, age, and education. The maximized (profile) log-likelihood is plotted against the transformation parameter  $\lambda$ . The intersection of the line near the top of the graph with the profile log-likelihood curve marks off a 95% confidence interval for  $\lambda$ . The maximum of the log-likelihood corresponds to the MLE of  $\lambda$ . JF Figure 12.14.

### Box-Tidwell transformation of $X$ s

Now, consider the model

$$Y_i = \beta_0 + \beta_1 X_{i1}^{\gamma_1} + \cdots + \beta_p X_{ip}^{\gamma_p} + \epsilon_i$$

where the errors are independently distributed as  $\epsilon_i \stackrel{iid}{\sim} N(0, \sigma_\epsilon^2)$  and all the  $X_{ij}$  are positive.

The parameters of this model ( $\beta_0, \beta_1, \dots, \beta_p, \gamma_1, \dots, \gamma_p$ , and  $\sigma_\epsilon^2$ ) could be estimated by general nonlinear least squares. Box and Tidwell (1962) suggested the following computationally more efficient procedure (also yields a constructed-variable diagnostic):

1. Regress  $Y$  on  $X_1, \dots, X_p$ , obtaining  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ . (“Regress  $A$  on  $B$ s” is the same as “fitting the linear regression model with  $A$  as the response variable and  $B$ s as the explanatory variables”.)
2. Regress  $Y$  on  $X_1, \dots, X_p$  and the constructed variables  $X_1 \log_e X_1, \dots, X_p \log_e X_p$  (again, by fitting the model of  $Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \delta_1 X_1 \log_e X_1 + \dots + \delta_p X_p \log_e X_p + \epsilon_i$ ) to obtain  $\tilde{\beta}_0, \tilde{\beta}_1, \dots, \tilde{\beta}_p, \tilde{\delta}_1, \dots, \tilde{\delta}_p$ . In general  $\hat{\beta}_i \neq \tilde{\beta}_i$ . (The constructed variables result from the first-order Taylor-series approximation to  $X_j^{\gamma_j}$  evaluated at  $\gamma_j = 1$ :  $X_j^{\gamma_j} \approx X_1 + (\gamma_j - 1)X_1 \log_e X_1$ .)
3. The constructed variable  $X_j \log_e X_j$  can be used to assess the need for a transformation of  $X_j$  by testing the null hypothesis  $H_0 : \delta_j = 0$ . Added-variable plots for the constructed variables are useful for assessing leverage and influence on the decision to transform the  $X$ s.
4. A preliminary estimate of the transformation parameter  $\gamma_j$  (not the MLE) is

$$\tilde{\gamma}_j = 1 + \frac{\tilde{\delta}_j}{\hat{\beta}_j}$$

where  $\tilde{\delta}_j$  is from step 2 and  $\hat{\beta}_j$  is from step 1.

## Polynomial regression

A machinery of multiple regression to fit non-linear relationships between predictor(s) and response.

- Linear:  $y = \beta_0 + \beta_1 x + \epsilon$
- Quadratic:  $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$
- Cubic:  $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$
- $k^{th}$  order polynomial:  $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_k x^k + \epsilon$

Question:

Does quadratic model provide a significantly better fit than linear model?

*Solution:* Test  $H_0 : \beta_2 = 0$  vs.  $H_a : \beta_2 \neq 0$ .

Alternatively, compare the corresponding adjusted- $R^2$  values.

## Additional reference

- “A First Course in Linear Model Theory” by Nalini Ravishanker and Kipak K. Dey
- [Lecture notes](#) by Cedric Ginestet

## Collinearity

In linear model

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} &\sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n) \end{aligned}$$

Collinearity (or multicollinearity) exists when there is “near-dependency” between the columns of the design matrix  $\mathbf{X}$ .

- Two or more columns.
- In other words, high correlation between explanatory variables.
- the data/model pair is ill-conditioned when  $\mathbf{X}^T\mathbf{X}$  is nearly singular.

Perfect collinearity leads to rank-deficiency in  $\mathbf{X}$  such that  $\mathbf{X}^T\mathbf{X}$  is singular. In the case of perfect collinearity, two or more columns are linear-dependent.

An example of perfect collinearity

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \epsilon_i$$

Consider the case, where

- $Y_i$  represents the amount of sales.
- $X_{i1}, X_{i2}, \dots, X_{i4}$  are categorical that represent the quarter in which the sample is collected:  $X_{ij} = \mathbf{1}(\text{sample } i \text{ collected in quarter } j)$ .
- $X_{i5}$  represents expense spent in advertising.

The dummy variable trap  $X_{i4} = 1 - X_{i1} - X_{i2} - X_{i3}$ . Recall that we need  $m - 1$  dummy variables for  $m$  categories.

An example of high correlation between predictors

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

Consider the case, where

- $Y_i$  represents the salary of individual  $i$ .
- $X_{i1}$  represents the age of individual  $i$ .
- $X_{i2}$  represents the experience of individual  $i$ .

How to interpret  $\beta_1$ ?

We expect high correlation between age and experience.

Problems caused by multicollinearity

1. large standard errors of the regression coefficients
  - small associated t-statistics
  - conclusion that truly useful explanatory variables are insignificant in explaining the regression

2. the sign of regression coefficients may be the opposite of what a mechanistic understanding of the problem would suggest
3. deleting a column of the predictor matrix will cause large changes in the coefficient estimates for other variables

However, multicollinearity does **not** greatly affect the **predicted values**.

### Signs and detections of multicollinearity

Some signs for multicollinearity:

1. Simple correlation between a pair of predictors exceeds 0.9 or  $R^2$ .
2. High value of the multiple correlation coefficient with some high partial correlations between the explanatory variables.
3. Large  $F$ -statistics with some small  $t$ -statistics for individual regression coefficients

Some approaches for detecting multicollinearity:

1. Pairwise correlations among the explanatory variables
2. Variance inflation factor
3. Condition number

### Variance inflation factor

For a multiple linear regression with  $k$  explanatory variables. We can regress  $X_j$  on the  $(k - 1)$  other explanatory variables and denote  $R_j$  as the coefficient of determination.

Then the variance inflation factor (VIF) is defined as

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

- $\text{VIF}_j \in [1, +\infty)$
- A suggested threshold is 10
- May use the averaged  $\overline{\text{VIF}} = \sum_{j=1}^k \text{VIF}_j / k$ .

### Condition index and condition number

We first scale the design matrix  $\mathbf{X}$  into column-equilibrated predictor matrix  $\mathbf{X}_E$  such that  $\{X_E\}_{ij} = X_{ij} / \sqrt{\mathbf{X}_j^T \mathbf{X}_j}$ .

Let  $\mathbf{X}_E = \mathbf{U}\mathbf{D}\mathbf{V}^T$  be the singular-value decomposition (SVD) of the  $n \times p$  matrix  $\mathbf{X}_E$  where  $\mathbf{U}^T \mathbf{U} = \mathbf{V}^T \mathbf{V} = \mathbf{I}_p$  and  $\mathbf{D} = \text{diag}(d_1, d_2, \dots, d_p)$  is a diagonal matrix with  $d_j \geq 0$ .

The  $j^{th}$  condition index is defined as

$$\eta(\mathbf{X}_E) = d_{\max}/d_j, \quad j = 1, 2, \dots, p$$

The condition number is defined as

$$C = d_{\max}/d_{\min}$$

$$C \geq 1, \quad d_{\max} = \max_{1 \leq j \leq p} d_j \quad \text{and} \quad d_{\min} = \min_{1 \leq j \leq p} d_j$$

Some properties of the condition number

- Large condition number indicates evidence of multicollinearity
- Typical cutoff values, 10, 15 to 30.

Some problems with the condition number

- practitioners have different opinions of whether  $\mathbf{X}$  should be centered around their means for SVD.
  - centering may remove nonessential ill conditioning, e.g.  $Cor(X, X^2)$
  - centering may mask the role of the constant term in any underlying near-dependencies
- the degree of multicollinearity with dummy variables may be influenced by the choice of reference category
- condition number is affected by the scale of the  $\mathbf{X}$  measurements
  - By scaling down any column of  $\mathbf{X}$ , the condition number can be made arbitrarily large
  - Known as *artificial ill-conditioning*
  - The condition number of the scaled matrix  $\mathbf{X}_E$  is also referred to as the *scaled condition number*

Recall that  $\mathbf{X}_E = \mathbf{U}\mathbf{D}\mathbf{V}^T$  is the singular-value decomposition (SVD) of  $\mathbf{X}_E$ , where  $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}_p$  and  $\mathbf{D} = \text{diag}(d_1, d_2, \dots, d_p)$  is a diagonal matrix with  $d_j \geq 0$ .

Then

$$\begin{aligned} \mathbf{X}_E^T \mathbf{X}_E &= \mathbf{V} \mathbf{D} \mathbf{U}^T \mathbf{U} \mathbf{D} \mathbf{V}^T \\ &= \mathbf{V} \mathbf{D}^2 \mathbf{V}^T \end{aligned}$$

is the spectral decomposition of the Gramian matrix  $\mathbf{X}_E^T \mathbf{X}_E$  with  $\{d_j^2\}$  being the eigenvalues and  $\mathbf{V}$  being the corresponding eigen vector matrix. This relationship links the condition numbers to the eigen values of the Gramian matrix.

Condition Index	Proportions of variance			
	$Var(\hat{\beta}_1)$	$Var(\hat{\beta}_2)$	...	$Var(\hat{\beta}_3)$
$\eta_1$	$\pi_{11}$	$\pi_{12}$	...	$\pi_{1p}$
$\eta_2$	$\pi_{21}$	$\pi_{22}$	...	$\pi_{2p}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
$\eta_p$	$\pi_{p1}$	$\pi_{p2}$	...	$\pi_{pp}$

Table 1: Table of condition index and proportions of variance

### Variance decomposition method

The variance-covariance matrix of the coefficient

$$\begin{aligned} Cov(\hat{\beta}) &= \sigma^2(\mathbf{X}_E^T \mathbf{X}_E)^{-1} \\ &= \sigma^2 \mathbf{V} \mathbf{D}^{-2} \mathbf{V}^T \end{aligned}$$

Its  $j^{th}$  diagonal element is the estimated variance of the  $j^{th}$  coefficient,  $\hat{\beta}_j$ . Then

$$Var(\hat{\beta}_j) = \sigma^2 \sum_{h=1}^p \frac{v_{jh}^2}{d_h^2}$$

- Let  $q_{jh} = \frac{v_{jh}^2}{d_h^2}$  and  $q_j = \sum_{h=1}^p q_{jh}$ .
- The variance decomposition proportion is  $\pi_{jh} = q_{jh}/q_j$ .
- $\pi_{jh}$  denotes the proportion of the variance of the  $j^{th}$  regression coefficient associated with the  $h^{th}$  component of its decomposition.
- The variance decomposition proportion matrix is  $\mathbf{\Pi} = \{\pi_{jh}\}$ .

In practice, it is suggested to combine condition index and proportions of variance for multicollinearity diagnostic. Identify multicollinearity if

- Two or more elements in the  $j^{th}$  row of matrix  $\mathbf{\Pi}$  are relatively large
- And its associated condition index  $\eta_j$  is large too

### Principal Components

The method of principal components, introduced by Karl Pearson (1901) and Harold Hotelling (1933), provides a useful representation of the correlational structure of a set of variables. Some advantages of the principal component analysis include

- more unified
- linear transformation of the original predictors into a new set of orthogonal predictors
- the new orthogonal predictors are called principal components

Principal components regression is an approach that inspects the sample data  $(\mathbf{Y}, \mathbf{X})$  for directions of variability and uses this information to reduce the dimensionality of the estimation problem. The procedure is based on the observation that every linear regression model can be restated in terms of a set of orthogonal predictor variables, which are constructed as linear combinations of the original variables. The new orthogonal variables are called the principal components of the original variables.

Let  $\mathbf{X}^T \mathbf{X} = \mathbf{Q} \mathbf{\Delta} \mathbf{Q}^T$  denote the spectral decomposition of  $\mathbf{X}^T \mathbf{X}$ , where  $\mathbf{\Delta} = \text{diag}\{\lambda_1, \dots, \lambda_p\}$  is a diagonal matrix consisting of the (real) eigenvalues of  $\mathbf{X}^T \mathbf{X}$ , with  $\lambda_1 \geq \dots \geq \lambda_p$  and  $\mathbf{Q} = (\mathbf{q}_1, \dots, \mathbf{q}_p)$  denotes the matrix whose columns are the orthogonal eigenvectors of  $\mathbf{X}^T \mathbf{X}$  corresponding to the ordered eigenvalues. Consider the transformation

$$\mathbf{Y} = \mathbf{X} \mathbf{Q} \mathbf{Q}^T \boldsymbol{\beta} + \boldsymbol{\epsilon} = \mathbf{Z} \boldsymbol{\theta} + \boldsymbol{\epsilon},$$

where  $\mathbf{Z} = \mathbf{X} \mathbf{Q}$ , and  $\boldsymbol{\theta} = \mathbf{Q}^T \boldsymbol{\beta}$ .

The elements of  $\boldsymbol{\theta}$  are known as the regression parameters of the principal components. The matrix  $\mathbf{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_p\}$  is called the matrix of principal components of  $\mathbf{X}^T \mathbf{X}$ .  $\mathbf{z}_j = \mathbf{X} \mathbf{q}_j$  is the  $j$ th principal component of  $\mathbf{X}^T \mathbf{X}$  and  $\mathbf{z}_j^T \mathbf{z}_j = \lambda_j$ , the  $j$ th largest eigenvalue of  $\mathbf{X}^T \mathbf{X}$ .

Principal components regression consists of deleting one or more of the variables  $\mathbf{z}_j$  (which correspond to small values of  $\lambda_j$ ), and using OLS estimation on the resulting reduced regression model.

#### Derivation under standardized predictors, JF 13.1.1

Consider the vectors of standardized predictors,  $\mathbf{x}_1^*, \mathbf{x}_2^*, \dots, \mathbf{x}_p^*$  (obtained by subtracting the mean and divided by standard deviation of the original predictor vectors). Because the principal components are linear combinations of the original predictors, we write the first principal component as

$$\begin{aligned} \mathbf{w}_1 &= A_{11} \mathbf{x}_1^* + A_{21} \mathbf{x}_2^* + \dots + A_{p1} \mathbf{x}_p^* \\ &= \mathbf{X}^* \mathbf{a}_1 \end{aligned}$$

The variance of the first component becomes

$$\begin{aligned} S_{w_1}^2 &= \frac{1}{n-1} \mathbf{w}_1^T \mathbf{w}_1 \\ &= \frac{1}{n-1} \mathbf{a}_1^T \mathbf{X}^{*T} \mathbf{X}^* \mathbf{a}_1 \\ &= \mathbf{a}_1^T \mathbf{R}_{XX} \mathbf{a}_1 \end{aligned}$$

where  $\mathbf{R}_{XX} = \frac{1}{n-1} \mathbf{X}^{*T} \mathbf{X}^*$ . We want to maximize  $S_{w_1}^2$  under the normalizing constraint  $\mathbf{a}_1^T \mathbf{a}_1 = 1$  (otherwise  $S_{w_1}^2$  can be arbitrarily large by inflating  $\mathbf{a}_1$ ). Consider

$$F_1 \equiv \mathbf{a}_1^T \mathbf{R}_{XX} \mathbf{a}_1 - L_1 (\mathbf{a}_1^T \mathbf{a}_1 - 1)$$



where  $L_1$  is a Lagrange multiplier. By differentiating this equation with respect to  $\mathbf{a}_1$  and  $L_1$ ,

$$\begin{aligned}\frac{\partial F_1}{\partial \mathbf{a}_1} &= 2\mathbf{R}_{XX}\mathbf{a}_1 - 2L_1\mathbf{a}_1 \\ \frac{\partial F_1}{\partial L_1} &= -(\mathbf{a}_1^T \mathbf{a}_1 - 1)\end{aligned}$$

Setting the partial derivatives to 0 produces

$$\begin{aligned}(\mathbf{R}_{XX} - L_1\mathbf{I}_p)\mathbf{a}_1 &= \mathbf{0} \\ \mathbf{a}_1^T \mathbf{a}_1 &= 1\end{aligned}$$

From the first equation, we see that  $L_1$  is an eigenvalue of  $\mathbf{R}_{XX}$  such that  $\mathbf{R}_{XX}\mathbf{a}_1 = L_1\mathbf{a}_1$  such that

$$S_{w_1}^2 = \mathbf{a}_1^T \mathbf{R}_{XX} \mathbf{a}_1 = L_1 \mathbf{a}_1^T \mathbf{a}_1 = L_1$$

To maximize  $S_{w_1}^2$ , we only need to pick the largest eigenvalue of  $\mathbf{R}_{XX}$ .