FALL 2013
Stat 8003: Statistical Methods I
Lecture 12

Jichun Xie

1 Multiple Linear Regression

1.1 Testing Multiple Contrasts

$$\mathrm{H}_0: \ \mathbf{K}^{\mathrm{T}}\boldsymbol{\beta} = \mathbf{m} \quad vs. \quad \mathbf{K}^{\mathrm{T}}\boldsymbol{\beta} \neq \mathbf{m}.$$

Why are we interested in such hypothesis testings?

Example: In a clinical trial, the patients with obesity are randomized. The bmi of the patients before and after the treatment are measured. Now the researchers are interested in whether the effect of Treatment B is the same as Treatment A. We can use two sample t-test to solve the problem. However, let's see another solution.

Model:

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i, \ i = 1, \ldots, n.$$

If the ith patient is in Group A, then $X_{i1} = 1$, $X_{i2} = 0$; otherwise, $X_{i1} = 0$ and $X_{i2} = 1$. Y_i is the difference of bmi before and after the treatment. There is no intercept term in the model.

Suppose the researchers are not only interested in comparing Treatment A and Treatment B. In addition, they are interested to test whether the effect of Treatment A is equal to 1. How to do the test simultaneously?

Let

$$\mathbf{K} = \begin{pmatrix} 1 & -1 \\ 1 & 0 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \quad \mathbf{m} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Then the hypothesis $\mathrm{H}_0 : \begin{cases} \beta_1 = \beta_2 \\ \beta_1 = 1 \end{cases}$ can be written as $\mathrm{H}_0 :\ \mathbf{K}^{\mathrm{T}}\boldsymbol{\beta} = \mathbf{m}$.

Example 2: Car price example. The investigator is interested in the relationship between the car prise and a bunch of variables, including engine size (X_1), horsepower (X_2), RPM (X_3), passenger capacity (X_4), rear seat room (X_5) and whether the car is imported (X_6).

$$\mathbf{Y} = \beta_0 + \beta_1\mathbf{X}_1 + \beta_2\mathbf{X}_2 + \beta_3\mathbf{X}_3 + \beta_4\mathbf{X}_4 + \beta_5\mathbf{X}_5 + \beta_6\mathbf{X}_6 + \boldsymbol{\epsilon}$$

Here $\mathbf{Y}$ stands for car price; $\mathbf{X}_i$ stands for the covariates.

The researchers would like to know whether the passenger capacity and the rear seat room would affect the car price. The hypothesis would be $\mathrm{H}_0 :\ \beta_4 = \beta_5 = 0$. It can be written as the form $\mathbf{K}^{\mathrm{T}}\boldsymbol{\beta} = \mathbf{m}$. How?

Now suppose $\mathbf{K}$ is an $p \times s$ matrix and $\boldsymbol{\beta}$ is an $p \times 1$ vector. Here p the column number of $\mathbf{X}$. Also suppose $\texttt{Rank}(\mathbf{X}) = p$ and $\texttt{Rank}(\mathbf{K}) = s$.

If we conduct GLR test, in the end, we will reject H_0 if

$$F = \frac{(\mathbf{K}^{\mathrm{T}}\hat{\boldsymbol{\beta}} - \mathbf{m})^{\mathrm{T}}(\mathbf{K}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{K})^{-1}(\mathbf{K}^{\mathrm{T}}\hat{\boldsymbol{\beta}} - \mathbf{m})}{s\hat{\sigma}^2}$$

is large.

Under H_0, F follows F distribution $F_{s,n-p}$. Why?

Then we need to reject H_0 if $F > F^{-1}_{s,n-p}(1-\alpha)$.

Now let's think about some special cases:

1.1.1 Case 1: $\mathrm{H}_0 : \boldsymbol{\beta} = \mathbf{0}$

What does it mean? (Think about Example 1)

In this case, $\mathbf{K} = \mathbf{I}_{p\times p}$ and $\mathbf{m} = \mathbf{0}_{p\times 1}$.

Then

$$F = \frac{\hat{\boldsymbol{\beta}}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})\hat{\boldsymbol{\beta}}}{(\hat{\mathbf{Y}} - \mathbf{X}\hat{\boldsymbol{\beta}})^{\mathrm{T}}(\hat{\mathbf{Y}} - \mathbf{X}\hat{\boldsymbol{\beta}})} \cdot \frac{n-p}{p} = \frac{\mathbf{Y}^{\mathrm{T}}\mathbf{P}_X\mathbf{Y}}{\mathbf{Y}^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_X)\mathbf{Y}} \cdot \frac{n-p}{p}$$

Reject H_0 if $F > F^{-1}_{p,n-p}(1-\alpha)$.

The numerator and denominator of F have interpretations. We call them

$$\text{Sum Square Regression} = SSR = \mathbf{Y}^{\mathrm{T}}\mathbf{P}_X\mathbf{Y}$$
$$\text{Sum Square Error} = SSE = \mathbf{Y}^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_X)\mathbf{Y}$$

Then, $F = \frac{SSR/p}{SSE/(n-p)}$.

Further,

$$\text{Total Sum Square} = SST = SSR + SSE = \mathbf{Y}^{\mathrm{T}}\mathbf{Y}.$$

We therefore decompose SST into two independent part: SSR and SSE. In order to better display the results, we set up an Analysis of Variance (ANOVA) table for the test.

Table 1: The ANOVA table for $\mathrm{H}_0: \boldsymbol{\beta} = \mathbf{0}$.

Source	SS	d.f.	MS	F-Statistic
Regression	$\mathbf{Y}^{\mathrm{T}}\mathbf{P}_X\mathbf{Y}$	p	$(\mathbf{Y}^{\mathrm{T}}\mathbf{P}_X\mathbf{Y})/p$	$F = \frac{SSR/p}{SSE/(n-p)}$
Error	$\mathbf{Y}^{\mathrm{T}}(\mathbf{I}-\mathbf{P}_X)\mathbf{Y}$	$n-p$	$(\mathbf{Y}^{\mathrm{T}}(\mathbf{I}-\mathbf{P}_X)\mathbf{Y})/(n-p)$	$\sim F_{p,n-p}$ under H_0
Total	$\mathbf{Y}^{\mathrm{T}}\mathbf{Y}$	n		

1.1.2 Case 2: $\boldsymbol{\beta}_1 = 0$

Now consider the linear model with the intercept:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \begin{pmatrix} \mathbf{1}_{n\times 1} & \mathbf{X}_{1,n\times p_1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \boldsymbol{\beta}_1 \end{pmatrix} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}^2)$. Here $\boldsymbol{\beta}_1$ is a $p_1 \times 1$ vector, with $p_1 = p - 1$.

We would like to test $\mathrm{H}_0: \boldsymbol{\beta}_1 = 0$. What does it mean? (Think about Example 2).

If we would like the hypothesis into the form $\mathbf{K}^{\mathrm{T}}\boldsymbol{\beta} = \mathbf{m}$, what would $\mathbf{K}$ and $\mathbf{m}$ be?

In this case,

$$F = \frac{SSR_m/p_1}{SSE/(n-p_1-1)},$$

where

$$SSR_m = \mathbf{Y}^{\mathrm{T}}(\mathbf{P}_X - \mathbf{J}/n)\mathbf{Y}$$
$$SSE = \mathbf{Y}^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_X)\mathbf{Y},$$

where $\mathbf{J}$ is the matrix of all ones. SSR_m is called the Sum Square Regression corrected for Means.

Similarly,

$$SST_m = SSR_m + SSE = \mathbf{Y}^{\mathrm{T}}(\mathbf{I} - \mathbf{J}/n)\mathbf{Y},$$

where SST_m is called Total Sum Square of Errors corrected for Means.

Table 2: The ANOVA table for $\mathrm{H}_0: \boldsymbol{\beta}_1 = \mathbf{0}$.

Source	SS	d.f.	MS	F-Statistic
SSR_m	$\mathbf{Y}^{\mathrm{T}}(\mathbf{P}_X - \mathbf{J}/n)\mathbf{Y}$	p_1	$\{\mathbf{Y}^{\mathrm{T}}(\mathbf{P}_X - \mathbf{J}/n)\mathbf{Y}\}/p_1$	$F = \frac{SSR_m/p_1}{SSE/(n-p_1-1)}$
SSE	$\mathbf{Y}^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_X)\mathbf{Y}$	$n - p_1 - 1$	$(\mathbf{Y}^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_X)\mathbf{Y})/(n - p_1 - 1)$	$\sim F_{p_1, n-p_1-1}$ under H_0
SST_m	$\mathbf{Y}^{\mathrm{T}}(\mathbf{I} - \mathbf{J}/n)\mathbf{Y}$	$n - 1$		

Table 2 displays the ANOVA analysis.

Also $SSRm = SSR_1 - SSR_2$, where SSR_1 is the SSR of the full model, and SSR_2 is the SSR of the model with only the intercept. SSR_m can be viewed as the difference of SSR between a larger model and a smaller model. In hypothesis testing, the smaller model is also called the null model or the nested model; the larger model is also called the expanded or the alternative model.

1.1.3 Example: Ret Loss in Flax

The data is from a dew-retting experiment in Ballarat 1942-43, in which flax was laid out under various climactic conditions and for various periods. Retting involves softening the flax stems by soaking in water, thus enabling the separation of the linen fibres from the wooden material by a process called scrutching. The flax variety used was "Liral Crown". Two samples were taken from each trial and the ret loss (Y), as a percentage, was calculated. The other three variables are the mean daily rainfall (in points, X_1), the retting period (in days, X_2) and the mean daily temperature (in degrees Farenheit, X_3). How to test whether the effect of these variables are zero?

Model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon,$$

where $\epsilon \sim \mathrm{N}(0, \sigma^2)$.

Test the effect of all three covariates. Null hypothesis:

$$\mathrm{H}_0: \ \beta_1 = \beta_2 = \beta_3 = 0.$$

Table 3: The ANOVA table

Source	SS	d.f.	MS	F-Statistic
SSR_m	40.64	3	13.55	$F = 7.01 \sim F_{3,50}$ under H_0
SSE	96.63	50	1.93	
SST_m	137.27	53		

Test the effect of rain and temperature. Null hypothesis:

$$\mathrm{H}_0: \ \beta_1 = \beta_3 = 0.$$

Now let $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^{\mathrm{T}}$, and

$$K^{\mathrm{T}} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix}$$

The null hypothesis can be written as $\mathrm{H}_0: \ K^{\mathrm{T}}\boldsymbol{\beta} = 0$. The F-statistic equals to 7.00 and p-value $= 5e-3$.

1.2 Collinearity

If there is some variable $\mathbf{X}_i \approx \sum_{j \neq i} \alpha_j \mathbf{X}_j$. What will happen?

- $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is close to a singular matrix, and therefore can hardly be inverted.
- Even if $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is invertible, it will be highly "unstable".

One way to show the problem of collinearity. Without loss of generality, suppose $X_p \approx \sum_{j=1}^{p-1} \alpha_j X_j$. Let $\mathcal{S} = \{1, \ldots, p-1\}$.

$$\begin{aligned}
(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} &= \left[\begin{pmatrix} \mathbf{X}_{\mathcal{S}}^{\mathrm{T}} \\ \mathbf{X}_p^{\mathrm{T}} \end{pmatrix} \begin{pmatrix} \mathbf{X}_{\mathcal{S}}^{\mathrm{T}} & \mathbf{X}_p^{\mathrm{T}} \end{pmatrix} \right]^{-1} \\
&= \begin{pmatrix} (\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\mathbf{X}_{\mathcal{S}})^{-1} + \frac{1}{k}(\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\mathbf{X}_{\mathcal{S}})^{-1}\mathbf{X}_{\mathcal{S}}\mathbf{X}_p\mathbf{X}_p^{\mathrm{T}}\mathbf{X}_{\mathcal{S}}(\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\mathbf{X}_{\mathcal{S}})^{-1} & -\frac{1}{k}(\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\mathbf{X}_{\mathcal{S}})^{-1}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\mathbf{X}_p \\ -\frac{1}{k}\mathbf{X}_p^{\mathrm{T}}\mathbf{X}_{\mathcal{S}}(\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\mathbf{X})^{-1} & \frac{1}{k} \end{pmatrix},
\end{aligned}$$

where $k = \mathbf{X}_p^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_{\mathbf{X}_{\mathcal{S}}})\mathbf{X}_p$.

If $\mathbf{X}_p \approx \sum_{j=1}^{p-1} \alpha_j \mathbf{X}_j$. Then $(\mathbf{I} - \mathbf{P}_{\mathbf{X}_{\mathcal{S}}})\mathbf{X}_p \approx 0$. $\Rightarrow k \approx 0$. $\Rightarrow 1/k \approx \infty$. Note that $\mathtt{Var}(\hat{\beta}_p) = 1/k$. This means that , if X_p can be almost linearly represented by other covariates, $\mathtt{Var}(\hat{\beta}_p)$ will be very large. In other words, $\hat{\beta}_p$ is very unstable.

Another way to show the problem of collinearity:

$$\mathbf{Y} = \mathbf{X}_1\beta_1 + \ldots + \mathbf{X}_{p-1}\beta_{p-1} + \mathbf{X}_p\beta_p + \mathbf{e},$$

where $\mathbf{e} \sim N(0, \sigma^2\mathbf{I})$. If $\mathbf{X}_p \approx \sum_{j=1}^{p-1} \alpha_j X_j$, then

$$\begin{aligned}
\mathbf{Y} &\approx \mathbf{X}_1\beta_1 + \ldots + \mathbf{X}_{p-1}\beta_{p-1} + (\sum_{j=1}^{p-1} \alpha_j \mathbf{X}_j)\beta_p + \mathbf{e} \\
&= \mathbf{X}_1(\beta_1 + \alpha_1\beta_p) + \ldots + X_{p-1}(\beta_{p-1} + \alpha_{p-1}\beta_p) + \mathbf{e}
\end{aligned}$$

It will reduce to an almost non-full rank model. The solution of $\hat{\boldsymbol{\beta}}$ is close to non-unique (unstable).

Therefore, when there are lots of covariates, it increases the probability of collinearity. It is better to reduce the number of covariates. In practice, we can use scatter plot matrix to examine whether there is any collinearity.

1.2.1 Example: Mass and Physical Measurements for Male Subjects

For his MS305 data project, Michael Larner measured the weight and various physical measurements for 22 male subjects aged 16 - 30. Subjects were randomly chosen volunteers, all in reasonable good health. Subjects were requested to slightly tense each muscle being measured to ensure measurement consistency. Apart from Mass, all measurements are in cm.

Variable	Description
Mass	Weight in kg
Fore	Maximum circumference of forearm
Bicep	Maximum circumference of bicep
Chest	Distance around chest directly under the armpits
Neck	Distance around neck, approximately halfway up
Waist	Distance around waist, approximately trouser line
Thigh	Circumference of thigh, measured halfway between the knee and the top of the leg
Calf	Maximum circumference of calf
Height	Height from top to toe
Shoulders	Distance around shoulders, measured around the peak of the shoulder blades

```
> summary(fit)

Call:
lm(formula = Mass ~ Fore + Bicep + Chest + Neck + Shoulder +
    Waist + Height + Calf + Thigh + Head, data = mass)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5523 -0.9965  0.0461  1.0499  4.1719

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -69.51714   29.03739  -2.394 0.035605 *
```

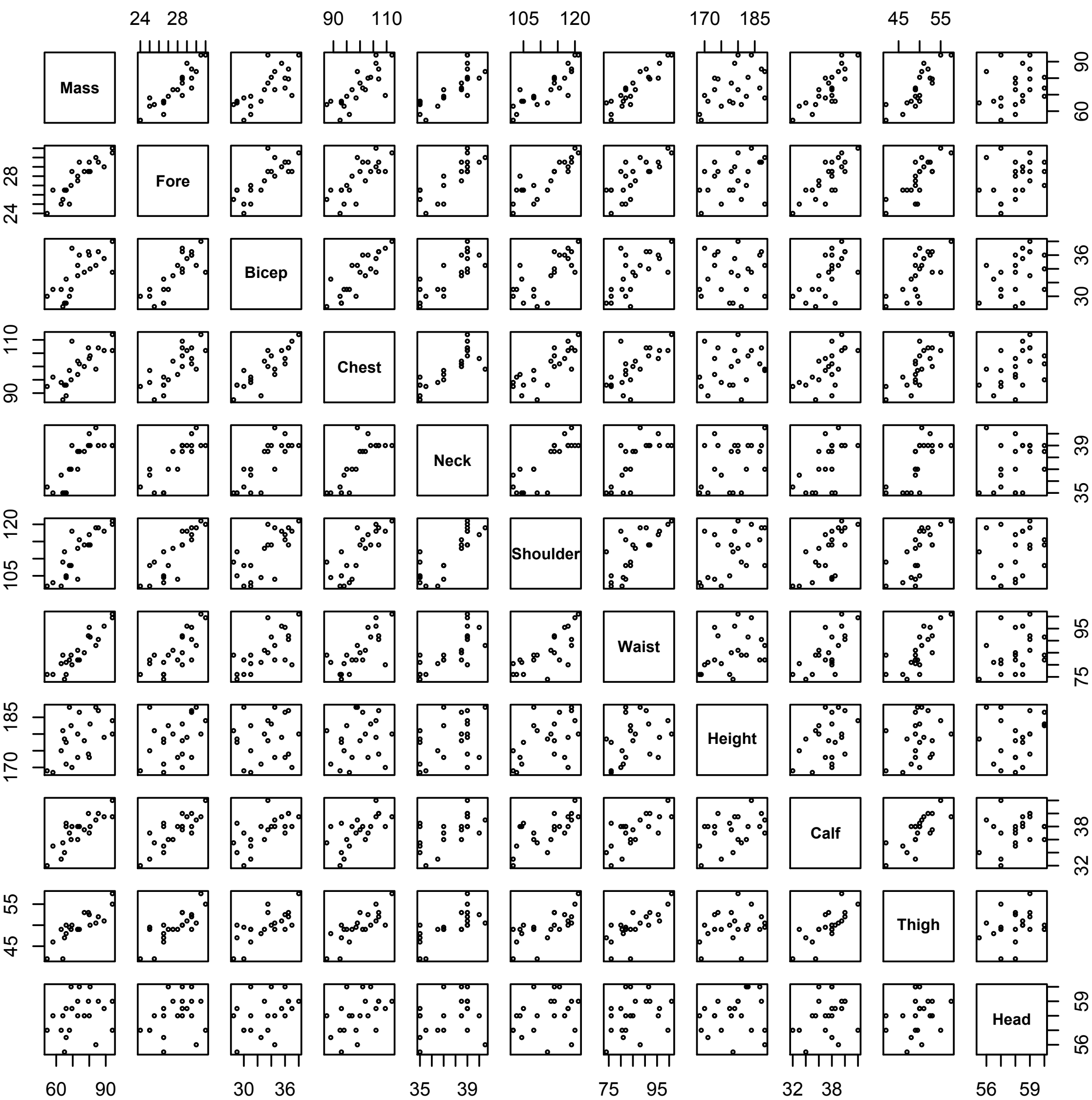

Figure 1: Scatter Plot of Man Mass

```
Fore          1.78182    0.85473   2.085 0.061204 .
Bicep         0.15509    0.48530   0.320 0.755275
Chest         0.18914    0.22583   0.838 0.420132
Neck         -0.48184    0.72067  -0.669 0.517537
Shoulder     -0.02931    0.23943  -0.122 0.904769
Waist         0.66144    0.11648   5.679 0.000143 ***
Height        0.31785    0.13037   2.438 0.032935 *
Calf          0.44589    0.41251   1.081 0.302865
Thigh         0.29721    0.30510   0.974 0.350917
Head         -0.91956    0.52009  -1.768 0.104735
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.287 on 11 degrees of freedom
Multiple R-squared:  0.9772,  Adjusted R-squared:  0.9565
F-statistic: 47.17 on 10 and 11 DF,  p-value: 1.408e-07
```

1.2.2 Singular Value Decomposition

Suppose $\mathbf{X}$ is a centered input matrix with dimension $n \times p$. The SVD (Singular Value Decompostion) of $\mathbf{X}$ has the form

$$\mathbf{X} = \mathbf{U}^{\mathrm{T}}\mathbf{D}\mathbf{V},$$

Here $\mathbf{U}$ and $\mathbf{V}$ are $n \times p$ and $p \times p$ orthogonal matrices. The columns of $\mathbf{U}$ spans the column space of $\mathbf{X}$ and the columns of $\mathbf{V}$ spans the row space. D is a $p \times p$ diagonal matrix, with diagonal entries $d_1 \geq d_2 \geq \ldots \geq d_p \geq 0$ called the singular values of X. If one or more values $d_j = 0$, X is singular.

By SVD, write the least-square fitted vector as

$$\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{U}\mathbf{U}^{\mathrm{T}}\mathbf{y},$$

after some simplication. Note that $\mathbf{U}^{\mathrm{T}}\mathbf{y}$ are the coordinates of $\mathbf{y}$ with respect to the orthonormal basis $\mathbf{U}$.

After centering, the sample covariance matrix is given by $\mathbf{S} = \mathbf{X}^{\mathrm{T}}\mathbf{X}/n$. The eigenvalue decomposition is

$$\mathbf{X}^{\mathrm{T}}\mathbf{X} = \mathbf{V}^{\mathrm{T}}\mathbf{D}^2\mathbf{V}.$$

The eigenvactors v_j (columns of $\mathbf{V}$) are also called the *principal components* directions of $\mathbf{X}$ (Figure 2). The first principal component direction v_1 has the property that $\mathbf{z}_1 = \mathbf{X}v_1$ has the largest sample variance amongst all normalized linear combinations of the columns of $\mathbf{X}$. This sample variance is easily seen to be

$$\mathtt{Var}(\mathbf{z}_1) = \mathtt{Var}(\mathbf{X}v_1) = \frac{d_1^2}{n},$$

and in fact $\mathbf{z}_1 = \mathbf{X}v_1 = \mathbf{u}_1 d_1$. The derived variable $\mathbf{z}_1$ is called the first principal component of $\mathbf{X}$, and hence $\mathbf{u}_1$ is the normalized first principal component. Subsequent principal components $\mathbf{z}_j$ have maximum variance d_j^2/n, subject to being orthogonal to the earlier ones. Conversely the last principal component has minimum variance. Hence the small singular values d_j correspond to directions in the column space of $\mathbf{X}$ having small variance.

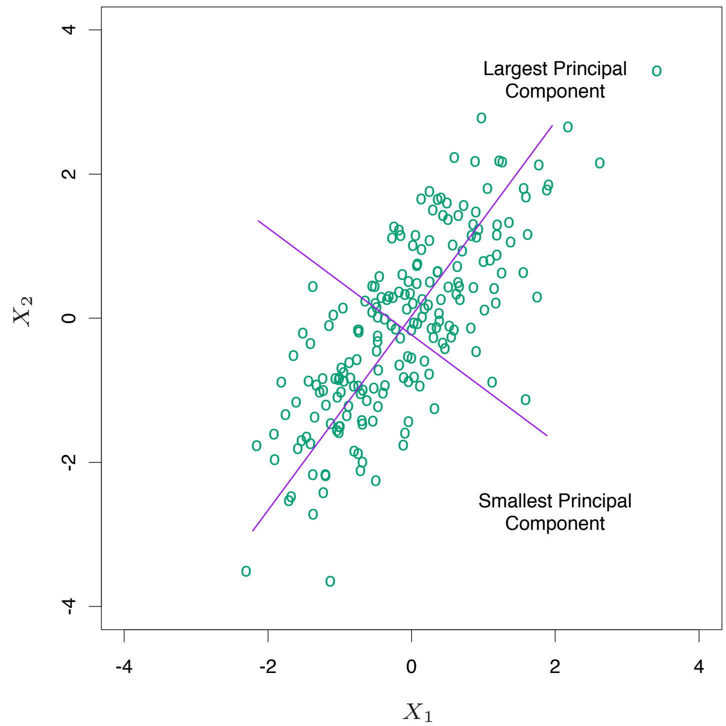

FIGURE 3.9. *Principal components of some input data points. The largest principal component is the direction that maximizes the variance of the projected data, and the smallest principal component minimizes that variance. Ridge regression projects* $\mathbf{y}$ *onto these components, and then shrinks the coefficients of the low–variance components more than the high-variance components.*

Figure 2: Principal components

When there are some variables with colinearity, we can consider using first several principal components as covariates to fit a regression model. Let's reconsider the man mass example. We can first computing the principal components of the covariates (Figure 3). And then we can fit linear regression using the first two principal components. We can include more principal components, depending on the problems. Here, the third pc is not significant. Please check the R code for more details.

```
> summary(fit2)

Call:
```

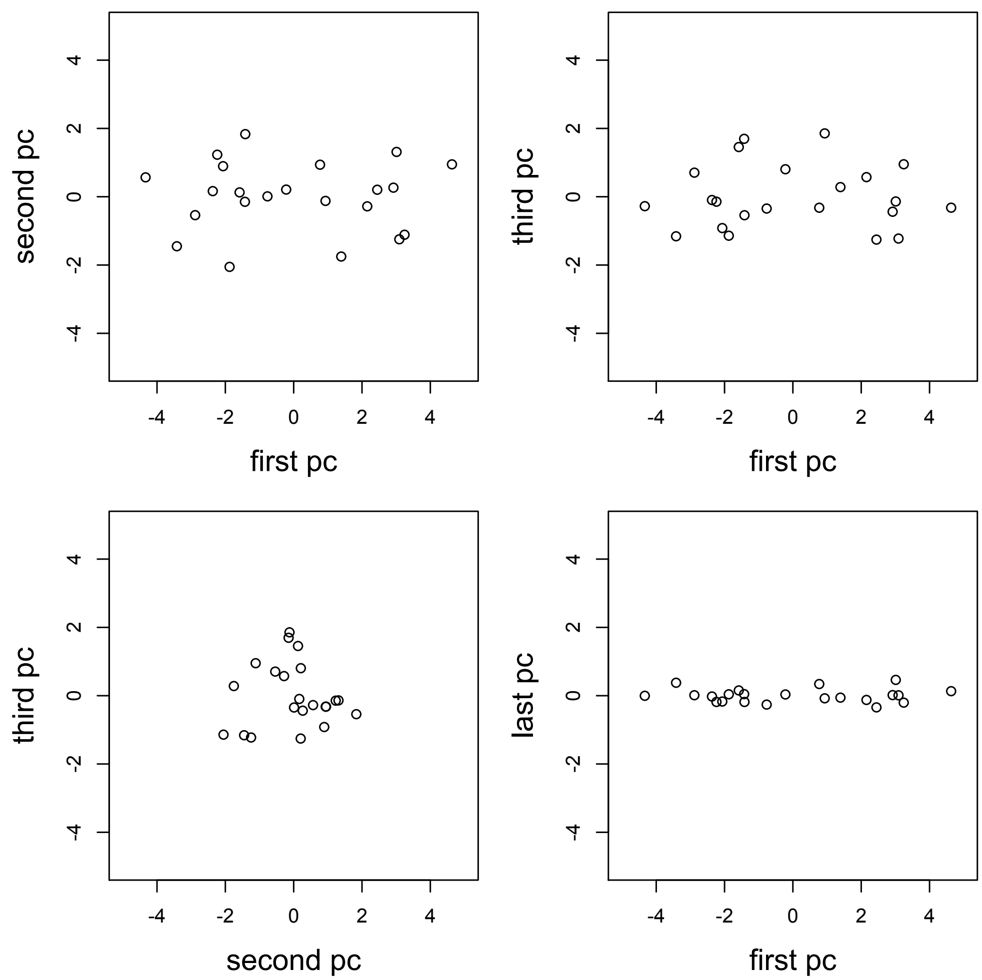

Figure 3: Principal Components of Man Mass Covariates

```
lm(formula = Y ~ Z[, 1:2] - 1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.57508 -0.12126  0.00742  0.12394  0.53477

Coefficients:
          Estimate Std. Error t value Pr(>|t|)
Z[, 1:2]1 -0.36785    0.02393 -15.371 1.53e-12 ***
Z[, 1:2]2 -0.19104    0.05945  -3.213  0.00436 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.2807 on 20 degrees of freedom
Multiple R-squared:  0.925, Adjusted R-squared:  0.9175
F-statistic: 123.3 on 2 and 20 DF,  p-value: 5.644e-12
```