**L17 Regression with dummy variables**

1. Dummy variables

   (1) Dummy variables
   Values of a dummy variable specify categories even if the values are numerical. For example a data set contains three variables, $y$: calculus scores, $x$: algebra scores and $z$: schools. Here $z = \begin{cases} 1 & \text{School A} \\ 2 & \text{School B} \\ 3 & \text{School C} \end{cases}$ is a dummy variable with numerical values.

   (2) Problem of interests
   With a dummy variable there are models for each categories, the inferences on the parameters across models might be of interest. For example for $\begin{cases} y_A = \beta_{A0} + \beta_{A1}x + \epsilon \\ y_B = \beta_{B0} + \beta_{B1}x + \epsilon \\ y_C = \beta_{C0} + \beta_{C1}x + \epsilon \end{cases}$
   we want to test $H_0: \beta_{A0} = \beta_{C0}$; $H_0: \beta_{A1} = \beta_{B1}$; $H_0: E(y_A) = E(y_C)$.

   (3) Difficulties
   We only have method on testing $H_0: A\beta = b$ where $\beta$ is a parameter vector from one model. Thus the models must be merged into one and the testing must be carried out in this combined model.

2. Directly use a dummy variable with two numerical values

   (1) Problem
   Data: $y$ calculus score, $x$ algebra score, $z = \begin{cases} 1 & \text{School A} \\ 2 & \text{School B} \end{cases}$.
   Models: $\begin{cases} y_A = \beta_{A0} + \beta_{A1}x + \epsilon \\ y_B = \beta_{B0} + \beta_{B1}x + \epsilon \end{cases}$
   Inferences: Testing $H_0: \beta_{A0} = \beta_{B0}$; $H_0: \beta_{A1} = \beta_{B1}$; $H_0: E(y_A) = E(y_B)$

   (2) A combined model
   $y = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz + \epsilon$ is equivalent to $\begin{cases} y_A = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x + \epsilon \\ y_B = (\beta_0 + 2\beta_2) + (\beta_1 + 2\beta_3)x + \epsilon \end{cases}$
   Thus $H_0: \beta_{A0} = \beta_{B0} \Longleftrightarrow \beta_0 + \beta_2 = \beta_0 + 2\beta_2 \Longleftrightarrow \beta_2 = 0$
   $H_0: \beta_{A1} = \beta_{B1} \longleftrightarrow \beta_1 + \beta_3 = \beta_1 + 2\beta_3 \Longleftrightarrow \beta_3 = 0$
   $H_0: E(y_A) = E(y_B) \Longleftrightarrow \beta_2 = 0$ and $\beta_3 = 0$
   So in the combined model $y = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz + \epsilon$, test
   $H_0: \beta_2 = 0$; $H_0: \beta_3 = 0$; and $H_0: \beta_2 = \beta_3 = 0$.

   (3) Implementation

```
data a; infile "D\Ex.txt";
      input y x z; xz=x*z;
proc reg; model y=x z xz/noprint;
      test z=0; test xz=0; test z=0, xz=0; run;
```

   (4) Presentation

| |
|---|
| $H_0: E(y_A) = E(y_B)$ versus $H_a: E(y_A) \neq E(y_B)$ |
| Test statistic: $F = \frac{MSH}{MSE}$ |
| $p$-value: $P(F(2, n-4) > F_{ob})$ |

3. Indicator variables

(1) A dummy variable
that specifies more than two classes with numerical values can not be directly used. For example with $y$, $x$ and $z$ in (1) of 1,
$y = \beta_0 + \beta_1 x + \beta_2 z + \epsilon \Longrightarrow \begin{cases} y_A = \beta_0 + \beta_2 + \beta_1 x + \epsilon \\ y_B = \beta_0 + 2\beta_2 + \beta_1 x + \epsilon \\ y_C = \beta_0 + 3\beta_2 + \beta_1 x + \epsilon \end{cases}$ . So $\beta_{A0} = \beta_{B0} \Longleftrightarrow \beta_0 + \beta_2 = \beta_0 + 2\beta_2 \Longleftrightarrow \beta_2 = 0 \Longleftrightarrow \beta_{A0} = \beta_{B0} = \beta_{C0}$.

(2) Indicator variables
A dummy variable with $t$ categories defined $t$ indicator variables. For example

| z | IA | IB | IC |
|---|---|---|---|
| A | 1 | 0 | 0 |
| B | 0 | 1 | 0 |
| C | 0 | 0 | 1 |
| B | 0 | 1 | 0 |
| A | 1 | 0 | 0 |

(3) Use indicators
$t-1$ indicator variables can identify $t$ classes. So using $t-1$ indicators one can construct a combined model from $t$ models from $t$ classes. For the data, models and problems in 1, let IA, IB be the indicators. Then
$y = \beta_0 + \beta_1 x + \beta_2 IA + \beta_3 IB + \beta_4 IA\ x + \beta_5 IB\ x + \epsilon$ is equivalent to

$$\begin{cases} y_A = (\beta_0 + \beta_2) + (\beta_1 + \beta_4)x + \epsilon \\ y_B = (\beta_0 + \beta_3) + (\beta_1 + \beta_5)x + \epsilon \\ y_C = \beta_0 + \beta_1 x + \epsilon \end{cases}$$

Therefore $H_0 : \beta_{A0} = \beta_{C0} \Longleftrightarrow \beta_0 + \beta_2 = \beta_0 \Longleftrightarrow \beta_2 = 0$
$H_0 : \beta_{A1} = \beta_{B1} \Longleftrightarrow \beta_1 + \beta_4 = \beta_1 + \beta_5 \Longleftrightarrow \beta_4 = \beta_5$
$H_0 : E(y_A) = E(y_C) \iff \beta_{A0} = \beta_{C0}$ and $\beta_{A_1} = \beta_{C1}$
$\iff \beta_0 + \beta_2 = \beta_0$ and $\beta_1 + \beta_4 = \beta_1 \Longrightarrow \beta_2 = 0$ and $\beta_4 = 0$.

(4) Implementation
Assume $z = A$, $B$, $C$ for schools $A$, $B$ and $C$.

```
data a; infile "D:\Ex.txt";
    input y x z $;
    if z="A" then do; IA=1; IB=0; end;
    if z="B" then do; IA=0; IB=1; end;
    if z="C" then do; IA=0; IB=0; end;
    IAx=IA*x; IBx=IB&x;
proc reg;
    model y=x IA IB IAx IBx/noprint;
    test IA; test IAx=IBx; test IA=0, IAx=0;
    run;
```

## L18 Polynomial regression and ANOVA model

1. Polynomial regression

   (1) Polynomial regression
   With one response variable $y$ and one predictor $x$, one can have two possible simple linear regression models

   $$y = \beta_0 + \beta_1 x + \epsilon \text{ and } y = \beta x + \epsilon.$$

   But one can also create more predictors $x^2$, $x^3$,..,$x^k$ from $x$ to have

   $$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \epsilon \text{ and } y = \beta_1 x_1 + \cdots + \beta_k x^k + \epsilon$$

   called polynomial regression since the regression function is a polynomial of $x$.

   **Comment:** Polynomial regression model is still a linear regression since the regression function is still a linear function of unknown parameter vector $\beta$.

   (2) Fit data to polynomial regression model
   To fit data to a polynomial regression $x^2$, $x^3, ..., x^k$ must be created from $x$ first.

```
data a;
   infile "C:\Example.txt";
   input y x1;
   x2=x1*x1; x3=x2*x1;
proc reg;
   model y=x1 x2 x3;
   run;
```

**Ex1:** With data $\begin{array}{c|cccccc} \text{x} & 1 & 1 & 0 & 0 & -1 & 1 \\ \text{y} & 1 & 2 & 3 & 4 & 5 & 0 \end{array}$ does the model $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$ shows a significant improvement from the model $y = \beta_0 + \beta_1 x + \epsilon$?
In model $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$ test $H_0 : \beta_2 = 0$.

$H_0 : \beta_2 = 0$ versus $H_a : \beta_2 \neq 0$
Test Statistic: $t = \frac{\widehat{\beta}_2}{S_{\widehat{\beta}_2}}$
$p$-value: $2P(t(n-3) > |t_{ob}|)$
$t = -0.60$, $p$-value: $2P(t(3) > 0.6) = 0.5908$
Fail to reject $H_0$
Data do not show a significant improvement.

2. One-way ANOVA model

   (1) One-way ANOVA model
   One-way ANOVA model specifies $q$ normal populations $N(\mu_i, \sigma^2)$ with possible different

means $\mu_1, ..., \mu_q$ and a common variance $\sigma^2$. These $q$ populations are responses to $q$ treatments by $q$ levels of a factor. The model can be expressed as

$$y = \mu(x) + \epsilon,\ \epsilon \sim N(0,\ \sigma^2)$$

where $\mu(x)$ is an unspecified function of $x$, and $x$ assumes $q$ different values. Thus $\mu(x)$ assumes unknown values $\mu_1, ..., \mu_q$.

(2) Samples in ANOVA model
$y = (y_1, ...., y_n)'$ is observed vector of responses in ANOVA. For this $y$, let $J \in R^{n\times q}$ with the $j$th column being the indicator for the $j$th population, i.e.,

$$J_{i,j} = \begin{cases} 0 & \text{if } y_i \text{ is not from the } j\text{th population} \\ 1 & \text{if } y_i \text{ is from the } j\text{th popualtion} \end{cases}$$

and $\mu = (\mu_1, ..., \mu_q)'$. Then

$$y \sim N\left(J\mu,\ \sigma^2 I_n\right).$$

(3) SSPE
$y \sim N(J\mu,\ \sigma^2 I_n)$, a linear model, has $SSE$ called the SS of Pure Errors (SSPE).
SSP $= y'[I - J(J'J)^{-1}J']y$ with DF $= \text{rank}[I - J(J'J)^{-1}J'] = n - q$, and

$$\max[L(\mu,\ \sigma^2) :\ \mu,\ \sigma^2] = \left(\frac{n}{2\pi e}\right)^{n/2} SSPE^{n/2}.$$

3. Computation for SSPE, ANOVA model SSE

(1) Hand computation
Suppose random sample $y_{i1}, ..., y_{in_i}$ is from $N(\mu_i,\ \sigma^2)$ with size $n_i$, mean $\overline{y}_i$ and $CSS_i = \sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2$. Then $SSE = CSS_1 + \cdots + CSS_q$.

**Ex2:** $y$ in Ex1 are from 3 populations distinguished by the values 1, 0 and $-1$ of $x$.

$$\begin{aligned} SSPE &= CSS_1 + CSS_0 + CSS_{-1} \\ &= [(1-1)^2 + (2-1)^2 + (0-1)^2] + [(3-3.5)^2 + (4-3.5)^2] + (5-5)^2 \\ &= 2 + 0.5 + 0 - 2.5. \end{aligned}$$

(2) SAS
Suppose variable $x$ assumes different values indicating different populations corresponding value of $y$ is from. Then

```
proc anova; class x; model y=x;  run;
```

produces ANOVA table on which SSE=SSPE with DF are displayed.

**Ex3:** For data in Ex1

```
data a;
   input y x @@;
   datalines;
   1 1 2 1 3 0 4 0 5 -1 0 1
   ;
proc anova;
   class x;
   model y=x;
   run;
```

produced SSE=2.5 and DF=6-3=3.