Regression Analysis
Chapter 5: Complex Regressors

Shaobo Jin

Department of Mathematics

Qualitative Regressor

If a regressor is qualitative or factors (e.g., Country = "Sweden", "Norway", "Finland", "Denmark"), it cannot be added to the model directly. They can be included in the model using dummy variables.

- Suppose that a factor Country has d levels. We can define

$$U_j = \begin{cases} 1, & \text{if the observation belongs to group } j, \\ 0, & \text{otherwise.} \end{cases}$$

- If we want to regress y on the factor Country, then the model is

$$\mathrm{E}\left(Y \mid \text{Country}\right) = \beta_0 + \beta_1 U_1 + \beta_2 U_2 + \beta_3 U_3 + \beta_4 U_4.$$

Example

If country = "Sweden", "Norway", "Finland", "Denmark", then

$$
\begin{aligned}
\mathrm{E}\left(Y \mid \text{Sweden}\right) &= \beta_0 + \beta_1, \\
\mathrm{E}\left(Y \mid \text{Norway}\right) &= \beta_0 + \beta_2, \\
\mathrm{E}\left(Y \mid \text{Finland}\right) &= \beta_0 + \beta_3, \\
\mathrm{E}\left(Y \mid \text{Denmark}\right) &= \beta_0 + \beta_4.
\end{aligned}
$$

Trap of Dummy Variables

Suppose that we have 2 observations for each country. Then, the design matrix of the above example is

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

$\boldsymbol{X}$ is not full column rank and hence $\boldsymbol{X}^T\boldsymbol{X}$ is singular!

Trap of Dummy Variables

In general, suppose that a factor has d levels and we create dummy variables U_j, $j = 1, ..., d$.

- The dummy variables satisfy

$$\sum_{j=1}^{d} U_j = 1.$$

- The model satisfy

$$\begin{aligned} \mathrm{E}\left(Y \mid \boldsymbol{u}\right) &= \beta_0 + \sum_{j=1}^{d} \beta_j U_j \\ &= \left(\beta_0 + c\right) + \sum_{j=1}^{d} \left(\beta_j - c\right) U_j, \end{aligned}$$

  for any constant c.

This means that the model is not identified.

Reference Level

We need to drop one dummy variable from the model and treat the corresponding label as the reference level.

- If we treat "Sweden" as the reference level, then we need to drop U_1 and the model is

$$\mathrm{E}\left(Y \mid \text{Country}\right) = \beta_0 + \beta_2 U_2 + \beta_3 U_3 + \beta_4 U_4.$$

- Our model is equivalent to

$$\begin{aligned}
\mathrm{E}\left(Y \mid \text{Sweden}\right) &= \beta_0, \\
\mathrm{E}\left(Y \mid \text{Norway}\right) &= \beta_0 + \beta_2, \\
\mathrm{E}\left(Y \mid \text{Finland}\right) &= \beta_0 + \beta_3, \\
\mathrm{E}\left(Y \mid \text{Denmark}\right) &= \beta_0 + \beta_4.
\end{aligned}$$

Interpretation of Coefficients

$$
\begin{aligned}
\mathrm{E}(Y \mid \text{Sweden}) &= \beta_0, \\
\mathrm{E}(Y \mid \text{Norway}) &= \beta_0 + \beta_2, \\
\mathrm{E}(Y \mid \text{Finland}) &= \beta_0 + \beta_3, \\
\mathrm{E}(Y \mid \text{Denmark}) &= \beta_0 + \beta_4.
\end{aligned}
$$

- β_0 is the average of Sweden and the other $\beta_j' s$ are the difference relative to Sweden.
- $\beta_2 - \beta_3$ is the expected difference between Norway and Finland.

Using Dummy Variables in R

An experiment was conducted to measure and compare the effectiveness of various feed supplements on the growth rate of chickens.

1. weight: a numeric variable giving the chick weight.
2. feed: a factor giving the feed type, i.e., casein, horsebean, linseed, meatmeal, soybean, sunflower.

```
LR <- lm(weight ~ factor(feed), data = chickwts)
```

Using Dummy Variables in R

```
summary(LR)

##
## Call:
## lm(formula = weight ~ factor(feed), data = chickwts)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -123.909  -34.413    1.571   38.170  103.091
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             323.583     15.834  20.436  < 2e-16 ***
## factor(feed)horsebean -163.383     23.485  -6.957 2.07e-09 ***
## factor(feed)linseed   -104.833     22.393  -4.682 1.49e-05 ***
## factor(feed)meatmeal   -46.674     22.896  -2.039 0.045567 *
## factor(feed)soybean    -77.155     21.578  -3.576 0.000665 ***
## factor(feed)sunflower    5.333     22.393   0.238 0.812495
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.85 on 65 degrees of freedom
## Multiple R-squared:  0.5417,Adjusted R-squared:  0.5064
```

Test Coefficients

What if we want to test two levels have the same mean? For example, we want to test $\beta_2 - \beta_3 = 0$. It is the same as testing a linear combination

$$\boldsymbol{a}^T \boldsymbol{\beta} = 0.$$

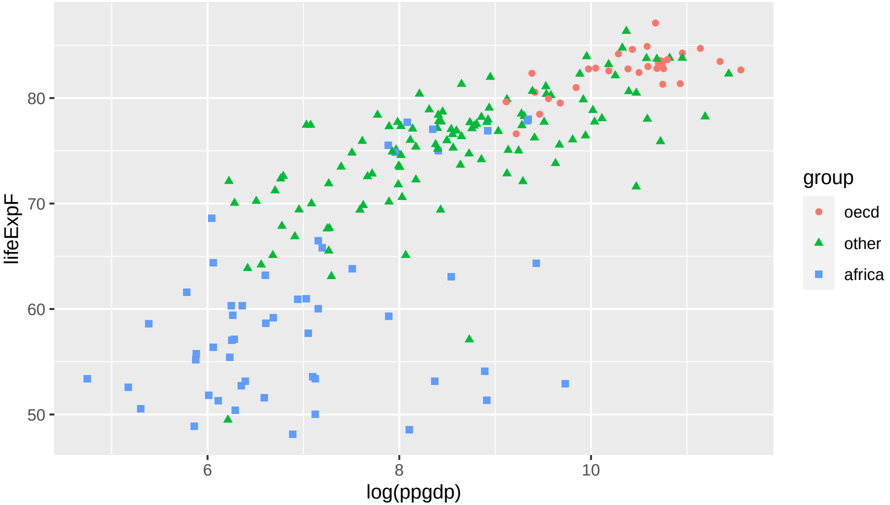
Under the normality assumption

$$\hat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \quad \sigma^2 \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}\right),$$
$$\text{and } \boldsymbol{a}^T \hat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{a}^T \boldsymbol{\beta}, \quad \sigma^2 \boldsymbol{a}^T \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{a}\right).$$

Hence,

$$\frac{\left(\boldsymbol{a}^T \hat{\boldsymbol{\beta}} - \boldsymbol{a}^T \boldsymbol{\beta}\right) / \sqrt{\boldsymbol{a}^T \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{a}}}{\sqrt{\hat{\boldsymbol{e}}^T \hat{\boldsymbol{e}} / (n-p)}} \sim t(n-p).$$

Add One Continuous Regressor



Add One Continuous Regressor

```
LR <- lm(lifeExpF ~ group + log(ppgdp), data = UN11)
summary(LR)

##
## Call:
## lm(formula = lifeExpF ~ group + log(ppgdp), data = UN11)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.6348  -2.1741   0.2441   2.3537  14.6539
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   49.529      3.400  14.569  < 2e-16 ***
## groupother    -1.535      1.174  -1.308    0.193
## groupafrica  -12.170      1.557  -7.814 3.35e-13 ***
## log(ppgdp)     3.177      0.316  10.056  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.109 on 195 degrees of freedom
## Multiple R-squared:  0.7492,Adjusted R-squared:  0.7453
## F-statistic: 194.1 on 3 and 195 DF,  p-value: < 2.2e-16
```

Interaction

The model that we fitted can be viewed as

$$\mathrm{E}\left(Y \mid X = x, \boldsymbol{u} = \boldsymbol{u}\right) = \beta_0 + \beta_1 x + \sum_{j=2}^{d} \beta_j U_j.$$

How x affects y is the same across different groups. We only have main effects in the model.

If the effect of x on y is different for different groups, we need to add an interaction.

$$\mathrm{E}\left(Y \mid X = x, \boldsymbol{u} = \boldsymbol{u}\right) = \beta_0 + \beta_1 x + \sum_{j=2}^{d} \beta_j U_j + \sum_{j=2}^{d} \gamma_j x U_j.$$

One Factor and One Continuous Regressor

```
G.Other <- subset(UN11, UN11$group == "other")
G.Africa <- subset(UN11, UN11$group == "africa")
G.Oecd <- subset(UN11, UN11$group == "oecd")
LR <- lm(lifeExpF ~ log(ppgdp) * group, data = UN11)
LR1 <- lm(lifeExpF ~ log(ppgdp), data = G.Other)
LR2 <- lm(lifeExpF ~ log(ppgdp), data = G.Africa)
LR3 <- lm(lifeExpF ~ log(ppgdp), data = G.Oecd)
```

One Factor and One Continuous Regressor

```
## Group Other
coef(LR1)

## (Intercept)  log(ppgdp)
##   48.040558    3.171973

c(coef(LR)["(Intercept)"] + coef(LR)["groupother"],
  coef(LR)["log(ppgdp)"] + coef(LR)["log(ppgdp):groupother"])

## (Intercept)  log(ppgdp)
##   48.040558    3.171973
```

One Factor and One Continuous Regressor

```r
## Group OECD
coef(LR3)

## (Intercept)  log(ppgdp)
##   59.213661    2.242535

coef(LR)

##            (Intercept)             log(ppgdp)            groupother
##             59.2136614              2.2425354           -11.1731029
##            groupafrica  log(ppgdp):groupother log(ppgdp):groupafrica
##            -22.9848394              0.9294372             1.0949810
```

Interactions

If we have two continuous regressors x_1 and x_2, the model can be

$$
\begin{aligned}
\mathrm{E}(Y \mid x_1, x_2) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2, \\
\mathrm{E}(Y \mid x_1, x_2) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2.
\end{aligned}
$$

If we have two factors x_1 and x_2 with d_1 and d_2 levels each, the model can be

$$
\begin{aligned}
\mathrm{E}(Y \mid x_1, x_2) &= \beta_0 + \sum_{j=2}^{d_1} \beta_j u_j + \sum_{k=2}^{d_2} \gamma_k v_k, \\
\mathrm{E}(Y \mid x_1, x_2) &= \beta_0 + \sum_{j=2}^{d_1} \beta_j u_j + \sum_{k=2}^{d_2} \gamma_k v_k + \sum_{j=2}^{d_1} \sum_{k=2}^{d_2} w_{jk} u_j v_k.
\end{aligned}
$$

Polynomial Regression with One Regressor

The simple linear regression $\mathrm{E}(Y \mid X = x) = \beta_0 + \beta_1 x$ can be generalized to a quadratic regression as

$$\mathrm{E}(Y \mid X = x) = \beta_0 + \beta_1 x + \beta_2 x^2,$$

or a cubic regression as

$$\mathrm{E}(Y \mid X = x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3,$$

or a polynomial regression as

$$\mathrm{E}(Y \mid X = x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_d x^d.$$

These models are not linear in x, but still linear in the parameters.

Polynomial Regression with Multiple Regressors

If we have more than one regressor, we can add power terms and products of regressors into the model.

$$
\begin{aligned}
\mathrm{E}\left(Y \mid X_1 = x_1, X_2 = x_2\right) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2, \\
\mathrm{E}\left(Y \mid X_1 = x_1, X_2 = x_2\right) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2, \\
\mathrm{E}\left(Y \mid X_1 = x_1, X_2 = x_2\right) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2, \\
\mathrm{E}\left(Y \mid X_1 = x_1, X_2 = x_2\right) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 \\
&\quad + \beta_6 x^3.
\end{aligned}
$$

Always Keep in Mind

1. We often want to keep marginality of our model: when you include higher-order terms into the model, always include all lower-order and often include same-order terms.
2. If the base is large and the exponent is also large, then the power can be too large.
3. The power terms can be highly correlated. Always check the correlations when you have higher-order terms.

```
x <- seq(0, 2, length.out = 100)
round(cor(cbind(x, x ^ 2, x ^ 3, x ^ 4)), 4)

##         x
## x 1.0000 0.9676 0.9155 0.8648
##   0.9676 1.0000 0.9860 0.9582
##   0.9155 0.9860 1.0000 0.9921
##   0.8648 0.9582 0.9921 1.0000
```