Regression and Linear Models (37252)

Lecture 6 - Multiple Linear Regression III

Lecturer: Joanna Wang
Notes adopted from Scott Alexander

School of Mathematical and Physical Sciences, UTS

Autumn 2024

Acknowledgement

These notes are based, in part, on earlier versions prepared by Dr Ed Lidums and Prof. James Brown.

6. Lecture outline

Topics:

- categorical predictors
  - dummy variables
- interaction effects
- examples
- partial sum squares and F-test
  - data set
  - two categories
  - three categories
  - two category with interaction

See Chapter 14 of Draper and Smith (1998).

6. Categorical predictors

So far we have built regression models using **continuous predictors**.

Now we will allow for discrete effects in our models, effects often represented by **categorical predictors**.

But categorical variables can't be used directly in regression models because regression is a numerical procedure and categorical variables don't have to be numeric.

A numerical proxy is required, a discrete variable taking values corresponding to defined states of the categorical predictor.

Suitable for the purposes of regression is the **dummy variable**.

6. Categorical predictors – dummy variables

Consider developing a multiple regression model including a categorical predictor defined for two categories A and B.

There are many (actually infinite) alternatives as to how one would define such a variable, but the most common is to specify a variable z described by

$$z = \begin{cases} 0, & \text{category } A \\ 1, & \text{category } B. \end{cases}$$

Using z and the other m independent variables we look to fit the model

$$\mathbb{E}[Y|x_1, \ldots, x_m, z] = \beta_0 + \sum_{j=1}^{m} \beta_j x_j + \gamma z$$

$$= \begin{cases} \beta_0 + \sum_{j=1}^{m} \beta_j x_j, & z = 0 \\ \beta_0 + \gamma + \sum_{j=1}^{m} \beta_j x_j, & z = 1 \end{cases}$$

which represents two planes separated by a parallel shift of size γ.

6. Categorical predictors – dummy variables

Now instead of two categories, suppose there are three.

A naive approach would be to redefine z as

$$z = \begin{cases} 0, & \text{category } A \\ 1, & \text{category } B \\ 2, & \text{category } C \end{cases}$$

and refit the model previously described.

This might be OK if γ is the common difference between the three planes associated with the three categories, but what if it isn't?

6. Categorical predictors – dummy variables

The correct approach is to use two dummy variables, z_1 and z_2, defined as

$$(z_1, z_2) = \begin{cases} (0,0), & \text{category } A \\ (1,0), & \text{category } B \\ (0,1), & \text{category } C \end{cases}$$

and then fit the model

$$\mathbb{E}[Y|x_1, \ldots, x_m, z_1, z_2] = \beta_0 + \sum_{j=1}^{m} \beta_j x_j + \gamma_1 z_1 + \gamma_2 z_2$$

$$= \begin{cases} \beta_0 + \sum_{j=1}^{m} \beta_j x_j, & (z_1, z_2) = (0,0) \\ \beta_0 + \gamma_1 + \sum_{j=1}^{m} \beta_j x_j, & (z_1, z_2) = (1,0) \\ \beta_0 + \gamma_2 + \sum_{j=1}^{m} \beta_j x_j, & (z_1, z_2) = (0,1) \end{cases}$$

which are three planes separated by parallel shifts of size γ_1 and γ_2.

6. Categorical predictors – dummy variables

More generally, if there are M categories then $M-1$ dummy variables $z_1, \ldots, z_{M-1}$ are required.

These variables are defined to take values according to

$$(z_1, z_2, \ldots, z_{M-2}, z_{M-1}) = \begin{cases} (0,0,\ldots,0,0), & \text{category } A \\ (1,0,\ldots,0,0), & \text{category } B \\ \vdots & \vdots \\ (0,0,\ldots,1,0), & \text{category } M-1 \\ (0,0,\ldots,0,1), & \text{category } M \end{cases}.$$

The category associated with

$$(z_1, z_2, \ldots, z_{M-2}, z_{M-1}) = (0,0,\ldots,0,0)$$

is called the **reference category**.

6. Categorical predictors – dummy variables

We then use least squares to fit the model

$$\begin{aligned}
&\mathbb{E}[Y|x_1,\ldots,x_m,z_1,\ldots,z_{M-1}] \\
&\quad = \beta_0 + \sum_{j=1}^{m}\beta_j x_j + \sum_{j=1}^{M-1}\gamma_j z_j \qquad (1a) \\
&\quad = \begin{cases}
\beta_0 + \sum_{j=1}^{m}\beta_j x_j, & (z_1,z_2,\ldots,z_{M-2},z_{M-1}) = (0,0,\ldots,0,0) \\
\beta_0 + \gamma_1 + \sum_{j=1}^{m}\beta_j x_j, & (z_1,z_2,\ldots,z_{M-2},z_{M-1}) = (1,0,\ldots,0,0) \\
\quad\vdots & \quad\vdots \\
\beta_0 + \gamma_{M-2} + \sum_{j=1}^{m}\beta_j x_j, & (z_1,z_2,\ldots,z_{M-2},z_{M-1}) = (0,0,\ldots,1,0) \\
\beta_0 + \gamma_{M-1} + \sum_{j=1}^{m}\beta_j x_j, & (z_1,z_2,\ldots,z_{M-2},z_{M-1}) = (0,0,\ldots,0,1)
\end{cases} \qquad (1b)
\end{aligned}$$

which are M planes separated by parallel shifts of size $\gamma_1,\gamma_2,\ldots,\gamma_{M-1}$.

6. Categorical predictors – dummy variables

Using the estimated parameters determined by least squares gives the fitted model

$$\hat{Y}|x_1, \ldots, x_m, z_1, \ldots, z_{M-1} = \hat{\beta}_0 + \sum_{j=1}^{m} \hat{\beta}_j x_j + \sum_{j=1}^{M-1} \hat{\gamma}_j z_j. \qquad (2)$$

We see that the model in (1a) is equivalent to the M simpler models in (1b).

However, if we fit the model (2) and compare it to the M simpler models fitted to the sample data partitioned by category, we will see that we get slightly different estimates of the parameters.

This is due to the non-linear properties of the residuals.

6. Interaction effects

The $M - 1$ dummy variables introduced in the last section are designed to capture categorical effects independently of the effect of the other m predictors in the model.

However, it is possible that the effects of these other predictors differ according to category.

To capture these effects in the model requires **interaction terms**.

To keep the notation simple, we will illustrate interaction in a model with one continuous predictor x and one categorical predictor defined on two states (A and B), represented in the model by the dummy variable z.

6. Interaction effects

The model in this situation is

$$\begin{aligned}\mathbb{E}[Y|x,z] &= \beta_0 + \beta_1 x + \gamma z + \delta xz \\ &= \begin{cases} \beta_0 + \beta_1 x, & z = 0 \text{ (category } A\text{)} \\ \beta_0 + \gamma + (\beta_1 + \delta)x & z = 1 \text{ (category } B\text{)} \end{cases},\end{aligned}$$

where the effect of the interaction term z is to change the slope of line.

This simple situation can be generalised to m continuous predictors and a categorial predictor defined on M states, represented in the model by $M - 1$ dummy variables.

6. Interaction effects

The terms involving the numerical and categorical predictors are known as the **main effects** and the interaction terms as the **interaction effects**.

Two points worth noting:

- if an interaction effect is deemed statistically-significant then the main effects variable involved in the interaction **MUST** be included in the model;
- interaction effects do not necessarily have to involve both numerical and categorical predictors – they can involve combinations of any type of predictor.

6. Partial sum squares and F-test

Recall a predictor taking three (or more) categories requires two (or more) dummy variables and that these dummies must be treated as a group. As the T-test can no longer be used to test significance, we require an alternative – the **partial F-test** .

This test is quite general, and can be used on any groups of predictors, discrete and continuous alike. For this reason we use the β-notation used previously when describing the F-test.

Simple partition of the set parameters

Consider the partition of parameters $\{\beta_0, \ldots, \beta_q\}$ and $\{\beta_{q+1}, \ldots, \beta_m\}$.

By partition we mean

$$\{\beta_0, \ldots, \beta_q\} \bigcap \{\beta_{q+1}, \ldots, \beta_m\} = \emptyset$$

and

$$\{\beta_0, \ldots, \beta_q\} \bigcup \{\beta_{q+1}, \ldots, \beta_m\} = \{\beta_0, \ldots, \beta_m\}.$$

6. Partial sum squares and F-test

The starting point is the full model **decomposition of sum squares** (i.e., decomposition from F-test on the model of m predictors)

$$SST = SSR + SSE$$

but we go further and decompose SSR as

$$SSR = SSR_q + SSR_{m-q} \tag{3}$$

where SSR_q is the total sum squares of the regression model on the first q predictors.

6. Partial sum squares and F-test

The **mean square** regression for the partial F-test is

$$MSR_{m-q} = \frac{SSR_{m-q}}{m-q} \tag{4}$$

and the **mean square error**

$$MSE = \frac{SSE}{n-m-1}$$

where the denominators are the relevant **degrees of freedom**.

If the usual assumptions are satisfied and if

$$\beta_{q+1} = \cdots = \beta_m = 0$$

then the RV

$$F_{m-q} = \frac{MSR_{m-q}}{MSE} \tag{5}$$

follows an F(m-q,$n-m-1$) distribution.

6. Partial sum squares and F-test

The partial F-test procedure is a minor modification of the F-test procedure.

Partial F-test hypotheses

$$H_0: \ \beta_{q+1} = \cdots = \beta_m = 0$$
$$H_A: \ \text{at least one } \beta_j \neq 0.$$

Test statistic

The test statistic f^*_{m-q} is the value that the RV F_{m-q}, given in (5), takes for each particular test.

6. Partial sum squares and F-test

Rejection of null hypothesis

H_0 is **rejected** in favour of H_A at **significance level** α if

$$f^*_{m-q} > f_{1-\alpha}$$

where $f_{1-\alpha}$ is the **quantile** satisfying

$$\text{Prob}(F_{m-q} > f_{1-\alpha}) = \alpha.$$

Equivalently, H_0 is rejected if the **p-value**

$$p = \text{Prob}(F_{m-q} > f^*_{m-q}) < \alpha.$$

The null hypothesis H_0 is **retained** in any other case.

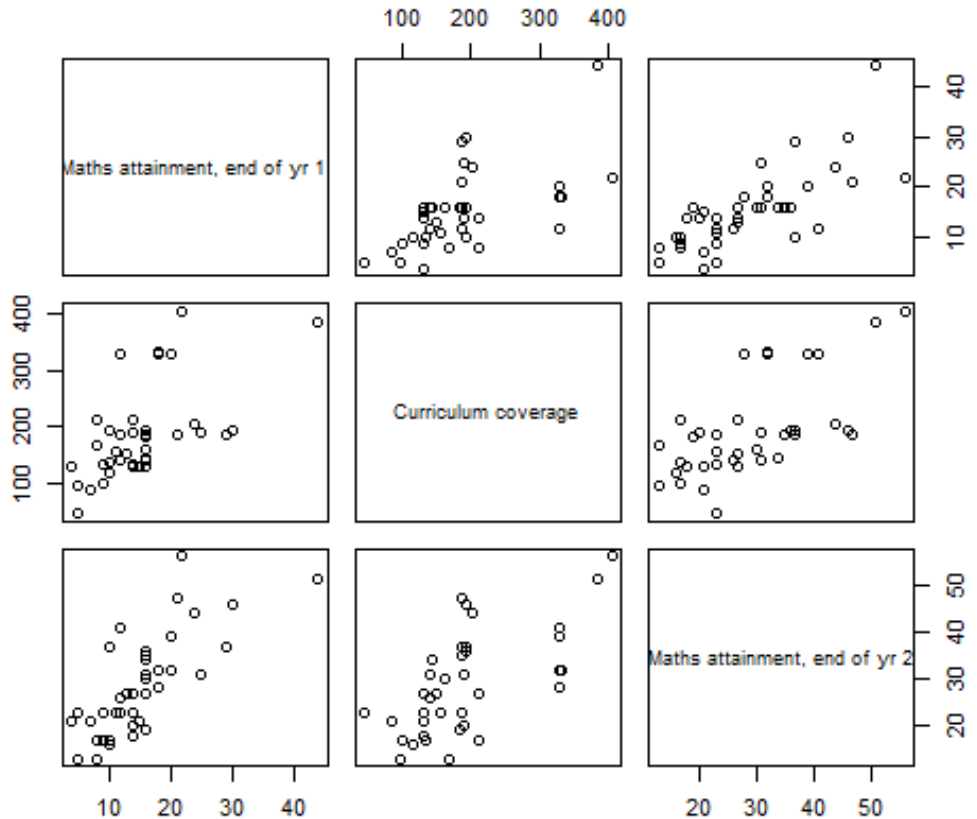
6. Examples – data set

In what follows we use the data set mathsdata.csv, available on Canvas, which includes the variables

- *school* – categorical variable taking values 1, 2 and 3
- *sex* – categorical variable taking values 0 (male) and 1 (female)
- *eg* – categorical variable taking values 0 (white) and 1 (African Carribean)
- *mathatt*1 – numerical variable representing maths attainment at end of year 1
- *curric* – numerical variable representing curriculum coverage during year 2
- *mathatt*2 – numerical variable representing maths attainment at end of year 2.

6. Examples – data set

Correlation information involving the numerical variables *mathatt*1, *curric* and *mathatt*2 is displayed below.



6. Examples – data set

```
> round(mathsdata_cor_res$r,3)
                              Maths attainment, end of yr 1 Curriculum coverage
Maths attainment, end of yr 1                         1.000                0.596
Curriculum coverage                                   0.596                1.000
Maths attainment, end of yr 2                         0.744                0.657
                              Maths attainment, end of yr 2
Maths attainment, end of yr 1                         0.744
Curriculum coverage                                   0.657
Maths attainment, end of yr 2                         1.000

> round(mathsdata_cor_res$P,3)
                              Maths attainment, end of yr 1 Curriculum coverage
Maths attainment, end of yr 1                            NA                   0
Curriculum coverage                                       0                  NA
Maths attainment, end of yr 2                             0                   0
                              Maths attainment, end of yr 2
Maths attainment, end of yr 1                             0
Curriculum coverage                                       0
Maths attainment, end of yr 2                            NA
```

6. Examples – two categories

The first regression model we build uses the independent categorical variable *sex* and dependent variable *mathatt*2.

As *sex* is defined as a zero/one variable it is already in the form required of a dummy variable and can be used directly in the estimated regression model

$$\widehat{mathatt2} = \hat{\beta} + \hat{\gamma} sex.$$

The Coefficients Table from R is displayed below.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   32.826      2.012  16.317  < 2e-16 ***
sex           -9.889      3.141  -3.148  0.00324 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
```

6. Examples – two categories

Extracting the parameter estimates gives the fitted model

$$\begin{aligned}\widehat{mathatt2} &= \hat{\beta} + \hat{\gamma} sex \\ &= 32.826 - 9.889 sex \\ &= \begin{cases} 32.826, & sex = 0 \text{ (male)} \\ 22.937, & sex = 1 \text{ (female)} \end{cases}\end{aligned}$$

where the estimates have the interpretations

- $\hat{\beta} = 32.826$ as the predicted year 2 maths attainment score for males (the reference category)
- $\hat{\gamma} = -9.889$ as the predicted difference in year 2 maths attainment score for females compared to males.

With very low p-values, both parameters must be deemed statistically significant which means that there is a statistically significant difference in predicted year 2 maths attainment scores between males and females.

6. Examples – two categories

For such a simple example we do not require regression and could have obtained the same result via other statistical means, as the output below confirms.

```
> leveneTest(mathsdata$mathatt2, as.factor(mathsdata$sex))
Levene's Test for Homogeneity of Variance (center = median)
      Df F value  Pr(>F)
group  1  3.2785 0.07832 .
      37
---

> t.test(mathatt2 ~ sex, data = mathsdata, var.equal=T)

Two Sample t-test

data:  mathatt2 by sex
t = 3.1483, df = 37, p-value = 0.003241
alternative hypothesis: true difference in means between group 0 and group 1 is not
equal to 0
95 percent confidence interval:
  3.524516 16.252658
sample estimates:
mean in group 0 mean in group 1
       32.82609        22.93750
```

6. Examples – three categories

The second regression model we build uses the independent categorical variable *school* and dependent variable *mathatt*2.

Unlike the previous example, *school* is not in the optimal form for capturing the categorical effect in a regression model (refer earlier discussion).

We define the dummy variables *sc*2 and *sc*3 as

$$(sc2, sc3) = \begin{cases} (0,0), & \text{school 1} \\ (1,0), & \text{school 2} \\ (0,1), & \text{school 3} \end{cases}$$

and estimate the model

$$\widehat{mathatt2} = \hat{\beta} + \hat{\gamma}_1 sc2 + \hat{\gamma}_2 sc3.$$

6. Examples – three categories

From the Coefficients table

```
(Intercept)       33.643      2.754  12.218 2.28e-14 ***
factor(school)2   -8.976      4.402  -2.039   0.0488 *
factor(school)3   -6.830      3.771  -1.811   0.0784 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
```

we extract the parameter estimates and write the fitted model as

$$\begin{aligned}\widehat{mathatt}2 &= \hat{\beta} + \hat{\gamma}_1 sc2 + \hat{\gamma}_2 sc3 \\ &= 33.643 - 8.976 sc2 - 6.830 sc3 \\ &= \begin{cases} 33.643, & (sc2, sc3) = (0,0) \text{ (school 1)} \\ 24.667, & (sc2, sc3) = (1,0) \text{ (school 2)} \\ 26.813, & (sc2, sc3) = (0,1) \text{ (school 3)} \end{cases} .\end{aligned}$$

6. Examples – three categories

The parameter estimates have the interpretations

- $\hat{\beta} = 33.643$ as the predicted year 2 maths attainment score for school 1 (the reference category)
- $\hat{\gamma}_1 = -8.976$ as the predicted difference in year 2 maths attainment score for school 2 compared to school 1
- $\hat{\gamma}_2 = -6.830$ as the predicted difference in year 2 maths attainment score for school 3 compared to school 1

of which their associated p-values allow $\hat{\beta}$ and $\hat{\gamma}_1$ to be deemed statistically significant (at 0.05 level) but not $\hat{\gamma}_2$.

From the estimates $\hat{\gamma}_1$ and $\hat{\gamma}_2$ we might infer 2.146 as the predicted difference in year 2 maths attainment score for school 3 compared to school 2.

The easiest way to test this formally is to fit a modified regression model with either school 2 or 3 as the reference category.

6. Examples – three categories

To determine whether there is a statistically significant "school" effect we can refer to the ANOVA results copied below.

```
> mod2 <- lm(mathatt2 ~ factor(school), data = mathsdata)
> anova(mod2)
Analysis of Variance Table

Response: mathatt2
               Df Sum Sq Mean Sq F value  Pr(>F)
factor(school)  2  545.3  272.64  2.5682 0.09065 .
Residuals      36 3821.7  106.16
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

With an insignificant p-value (at 0.05 level) associated with the F-test we conclude there is no evidence that school makes a difference (defined with school 1 as reference category) to mean year 2 maths attainment scores.

Note: ANOVA involving categorical variables is sometimes referred to as ***analysis of covariance (ANCOVA)****.*

6. Examples – two category with interaction

For the final example we model the dependent variable *mathatt*2 using the main effect predictors *mathatt*1 and *sex*.

We also include the interaction term *mathatt*1 $\times$ *sex*, which is in appropriate form as the zero/one variable *sex* can be used directly as a dummy variable in the model we seek to fit

$$\widehat{mathatt2} = \hat{\beta}_0 + \hat{\beta}_1 mathatt1 + \hat{\gamma} sex + \delta mathatt1 \times sex.$$

R output is reproduced below.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  16.58660    3.31108   5.009 1.56e-05 ***
mathatt1      0.95040    0.17496   5.432 4.33e-06 ***
sex          -4.75849    5.19321  -0.916    0.366
mathatt1:sex -0.08754    0.33107  -0.264    0.793
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
```

6. Examples – two category with interaction

From the coefficients table we extract the parameter estimates and write the fitted model as

$$\begin{aligned}\widehat{mathatt2} &= \hat{\beta}_0 + \hat{\beta}_1 mathatt1 + \hat{\gamma} sex + \delta mathatt1 \times sex \\ &= 16.587 + 0.950 mathatt1 - 4.758 sex - 0.088 mathatt1 \times sex \\ &= \begin{cases} 16.587 + 0.950 mathatt1, & sex = 0 \text{ (male)} \\ 11.829 + 0.862 mathatt1, & sex = 1 \text{ (female)} \end{cases}.\end{aligned}$$

Of the parameter estimates, only $\hat{\beta}_0$ and $\hat{\beta}_1$ can be deemed statistically significant.

The interaction effect and the main effect variable associated with this term are deemed statistically insignificant.

From here we would refit the model with the interaction term excluded and see if the model improves.

References I

Draper, N. R. and Smith, H. (1998). *Applied regression analysis*. Wiley-Interscience, Somerset, US.