Understanding Data and Statistical Design (60117)

Chapter 8

Multiple Linear Regression II

Subject Coordinator: Stephen Woodcock
Lecture notes: Scott Alexander

School of Mathematical and Physical Sciences, UTS

Autumn 2024

Chapter outline

Topics:

- categorical predictors
  - dummy variables
- interaction effects
- partial sum squares and F-test
- R examples
  - data set
  - model 1
  - model 2
  - model 3

See Chapter 14 of Draper and Smith (1998).

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 (or factors)**.

But categorical variables can't be used directly in regression models because regression is a numerical procedure and categorical variables are not 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**.

Categorical predictors – dummy variables

2-CATEGORY PREDICTOR

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 could 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

$$\begin{aligned} \mathbb{E}[Y] &= \beta_0 + \sum_{j=1}^{m} \beta_j x_j + \gamma z \\ &= \begin{cases} \beta_0 + \sum_{j=1}^{m} \beta_j x_j, & z = 0 \\ \beta_0 + \gamma + \sum_{j=1}^{m} \beta_j x_j, & z = 1 \end{cases} \end{aligned}$$

which represents two planes separated by a parallel shift of size γ from the category A plane.

Categorical predictors – dummy variables

3-CATEGORY PREDICTOR

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?

We could perhaps define the third value that z takes as something other than $z = 2$, but how would we calculate this estimate and isn't this something the least squares method should take care of anyway?

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

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

which are three planes separated by parallel shifts of size γ_1 and γ_2 from the category A plane.

Categorical predictors – dummy variables

M-CATEGORY PREDICTOR

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

These variables are defined as

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

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**.

Categorical predictors – dummy variables

We then use least squares to fit the model

$$\mathbb{E}[Y] = \beta_0 + \sum_{j=1}^{m} \beta_j x_j + \sum_{j=1}^{M-1} \gamma_j z_j \tag{1a}$$

$$= \begin{cases} \beta_0 + \sum_{j=1}^{m} \beta_j x_j, & (z_1, z_2, \ldots, z_{M-2}, z_{M-1}) = (0, 0, \ldots, 0, 0) \\ \beta_0 + \gamma_1 + \sum_{j=1}^{m} \beta_j x_j, & (z_1, z_2, \ldots, z_{M-2}, z_{M-1}) = (1, 0, \ldots, 0, 0) \\ \vdots & \vdots \\ \beta_0 + \gamma_{M-2} + \sum_{j=1}^{m} \beta_j x_j, & (z_1, z_2, \ldots, z_{M-2}, z_{M-1}) = (0, 0, \ldots, 1, 0) \\ \beta_0 + \gamma_{M-1} + \sum_{j=1}^{m} \beta_j x_j, & (z_1, z_2, \ldots, z_{M-2}, z_{M-1}) = (0, 0, \ldots, 0, 1) \end{cases} \tag{1b}$$

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

Categorical predictors – dummy variables

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

$$\hat{y} = \hat{\beta}_0 + \sum_{j=1}^{m} \hat{\beta}_j x_j + \sum_{j=1}^{M-1} \hat{\gamma}_j z_j. \tag{2}$$

Rather than fitting a single regression model with the M-state categorical variable, one could instead partition the sample data by category and fit M regression models of the form (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.

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**.

Interaction effects

2-CATEGORY PREDICTOR WITH INTERACTION

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.

The model in this situation is

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

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.

Interaction effects

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

Some 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;
- interaction effects involving continuous predictors impart curvature into the model;
- interaction effects do not necessarily have to be two-way – they can be three-way, four-way etc.

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 group of predictors, discrete and continuous alike.

To keep things general we use the β-notation used in previous chapters when describing the **complete F-test**.

Partial sum squares and F-test

PARTITION OF THE PARAMETERS

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

By partition we mean

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

and

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

The partial F-test is used to test the overall significance of the set of predictors associated with the set of coefficients $\{\beta_{q+1}, \ldots, \beta_m\}$.

Note that the partial F-test is not restricted to sets of predictors with sequentially ordered indices – it is just convenient from a notational perspective to present it in this manner.

Partial sum squares and F-test

DECOMPOSITION OF SUM SQUARES

The starting point is the sum square decomposition of the full model (i.e., model with m predictors)

$$SST = SSR + SSE.$$

The sum square regression of the full model can then be decomposed as

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

where

- SSR_q is the sum square regression for the model containing the first q predictors
- SSR_{m-q} is the addition to the sum square regression due to adding the next $m - q$ predictors into the model.

Note that SSR_{m-q} can be obtained from the sum square regression or sum square error terms as

$$\begin{aligned} SSR_{m-q} &= SSR - SSR_q \\ &= SSE_q - SSE. \end{aligned}$$

Partial sum squares and F-test

PARTIAL F-TEST

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

$$MSR_{m-q} = \frac{SSR_{m-q}}{m-q} = \frac{SSR - SSR_q}{m-q} = \frac{SSE_q - SSE}{m-q} \tag{4}$$

and the **mean square error**

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

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

Under the condition

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

the RV

$$F^*_{m-q} = \frac{MSR_{m-q}}{MSE} \tag{5}$$

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

Partial sum squares and F-test

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

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

$$f^*_{m-q} = \frac{msr_{m-q}}{mse}$$

is calculated from the data and is an observation of the RV F^*_{m-q} in (5).

Partial sum squares and F-test

Test decision

H_0 is rejected at significance level α if

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

where the quantile $f_{1-\alpha}$ is from $F \sim F(m-q, n-m-1)$ distribution.

Equivalently, H_0 is rejected if the p-value

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

where $F \sim F(m-q, n-m-1)$.

The null hypothesis H_0 is retained if this is not the case.

R examples – data set

Now an example using R.

We are going to build a model to predict performance in an end of year maths exam using a variety of predictors.

Our sample data consists of $n = 39$ observations, with the variables summarised in the table below (data in maths.data.csv on Canvas).

Name	**Type**	**Description**
sch	categorical predictor	school (A, B, C)
sex	categorical predictor	female (F), male (M)
cur	continuous predictor	curriculum coverage during year 2
*m*1	continuous predictor	maths score end of year 1
*m*2	continuous response	maths score end of year 2

We are going to build a variety of models with $m2$ as response.

R examples – data set

Binary dummy variables

The dummy variable *sexF* will be coded as

$$sexF = \begin{cases} 0, & sex = M \\ 1, & sex = F \end{cases} .$$

The dummy variables *schB* and *schC* will be coded as

$$(schB, schC) = \begin{cases} (0,0), & sch = A \\ (1,0), & sch = B \\ (0,1), & sch = C \end{cases} .$$

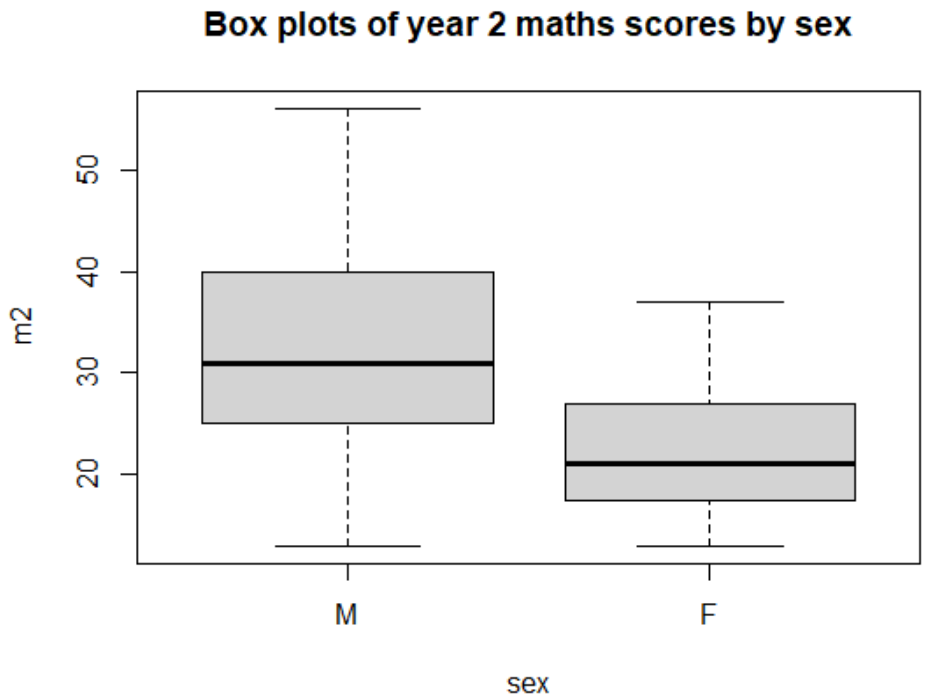
Note that we don't have to create these dummy variables ourselves – R will take care of this automatically.

R examples – model 1

The first model we consider is

$$M2 = \beta_0 + \gamma sexF + \epsilon.$$

Box plots



There appears be a significant difference in the sample distributions of $m2$ for males and females suggesting that *sex* is a significant predictor.

R examples – model 1

Fitted model

R produced the summary information below.

```
Call:
lm(formula = m2 ~ sex, data = maths.data)

Residuals:
    Min      1Q  Median      3Q     Max
-19.826  -5.882  -1.826   5.618  23.174

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   32.826      2.012  16.317  < 2e-16 ***
sexF          -9.889      3.141  -3.148  0.00324 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.648 on 37 degrees of freedom
Multiple R-squared:  0.2113,    Adjusted R-squared:   0.19
F-statistic: 9.912 on 1 and 37 DF,  p-value: 0.003241
```

R examples – model 1

The fitted model is

$$\widehat{m2} = 32.826 - 9.889 sexF$$

or

$$\widehat{m2} = \begin{cases} 32.826, & sex = M \\ 22.937, & sex = F \end{cases} .$$

Note that 32.826 and 22.937 are just the sample means of $m2$ for $sex = M$ and $sex = F$ respectively.

Significance

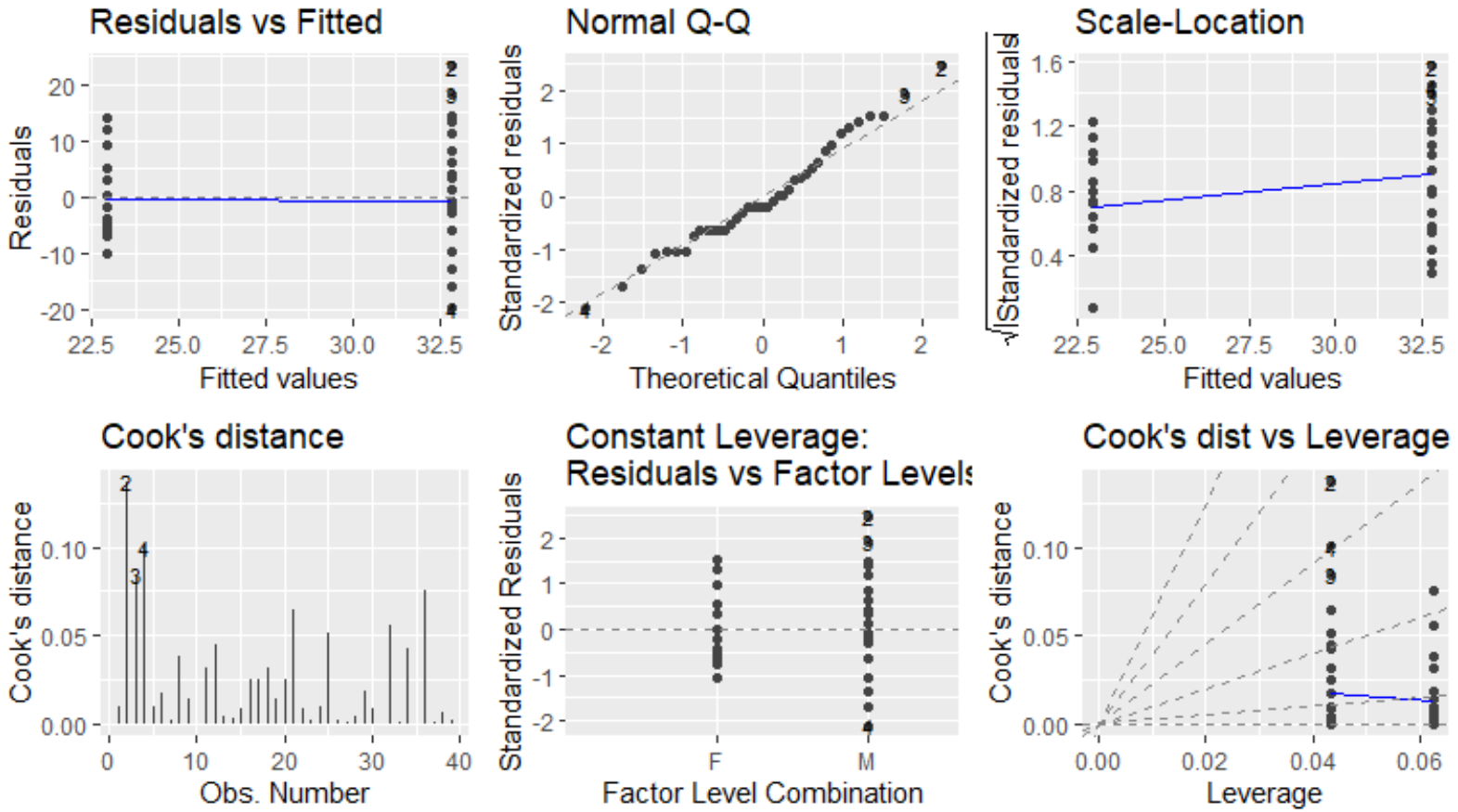
At $\alpha = 0.05$ we see that the model is significant (reject T-test hypothesis $\gamma = 0$ with $p = 0.00324$).

Estimated parameter interpretations

- $\hat{\beta}_0 = 32.826$ is predicted year 2 maths score for males.
- $\hat{\gamma} = -9.889$ is predicted difference in year 2 maths score for females compared to males.

R examples – model 1

Diagnostic plots



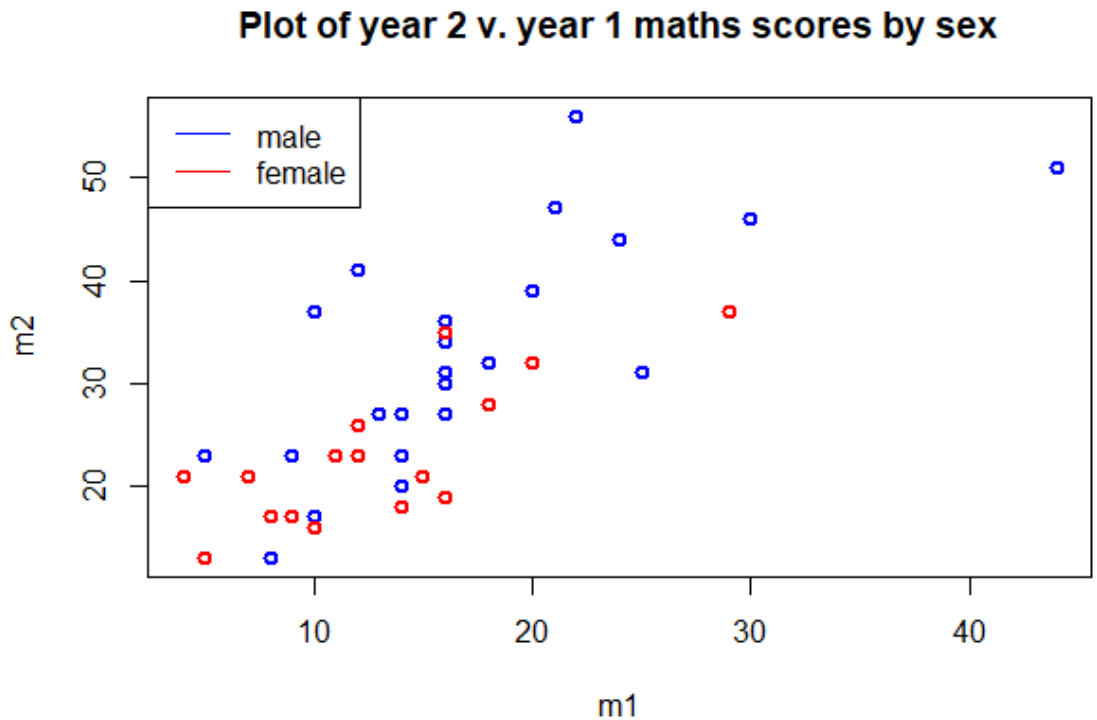
Normality (Shapiro-Wilk test $p = 0.6151$) and independence ($DW = 1.669$, $p = 0.2907$) assumptions look OK but there is strong evidence of non-equal variance in the two groups.

R examples – model 2

The second model we consider is

$$M2 = \beta_0 + \gamma sexF + \beta_m m1 + \delta sexF \times m1 + \epsilon.$$

Scatter plot



It appears that *m*1 is significant but *sex* appears insignificant (no strong evidence of different intercepts) and the interaction term appears insignificant (no strong evidence of different slopes).

R examples – model 2

Fitted model

R produced the summary information below.

```
Call:
lm(formula = m2 ~ sex * m1, data = maths.data)

Residuals:
    Min      1Q  Median      3Q     Max
-11.190  -4.114  -0.793   3.269  18.505

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.58660    3.31108   5.009 1.56e-05 ***
sexF        -4.75849    5.19321  -0.916    0.366
m1           0.95040    0.17496   5.432 4.33e-06 ***
sexF:m1     -0.08754    0.33107  -0.264    0.793
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.825 on 35 degrees of freedom
Multiple R-squared:  0.6266,    Adjusted R-squared:  0.5946
F-statistic: 19.58 on 3 and 35 DF,  p-value: 1.262e-07
```

The fitted model is

$$\widehat{m2} = 16.5866 - 4.75849 sexF + 0.9504 m1 - 0.08754 sexF \times m1$$

or

$$\widehat{m2} = \begin{cases} 16.5866 + 0.95040 m1, & sex = M \\ 11.8281 + 0.86286 m1, & sex = F \end{cases}.$$

R examples – model 2

Significance

At $\alpha = 0.05$ we see that

- the model is significant (reject F-test hypothesis $\gamma = \beta_m = \delta = 0$ with $p = 1.262 \times 10^{-7}$)
- the predictor *sex* is insignificant (retain T-test hypothesis $\gamma = 0$ with $p = 0.366$)
- the predictor $m1$ is significant (reject T-test hypothesis $\beta_m = 0$ with $p = 4.33 \times 10^{-6}$)
- the interaction between *sex* and $m1$ is insignificant (retain T-test hypothesis $\delta = 0$ with $p = 0.793$).

After controlling for $m1$, the categorical predictor *sex* and the interaction term are not required.

R examples – model 2

Collinearity

```
     sex       m1   sex:m1
5.462509 1.497526 5.067559
```

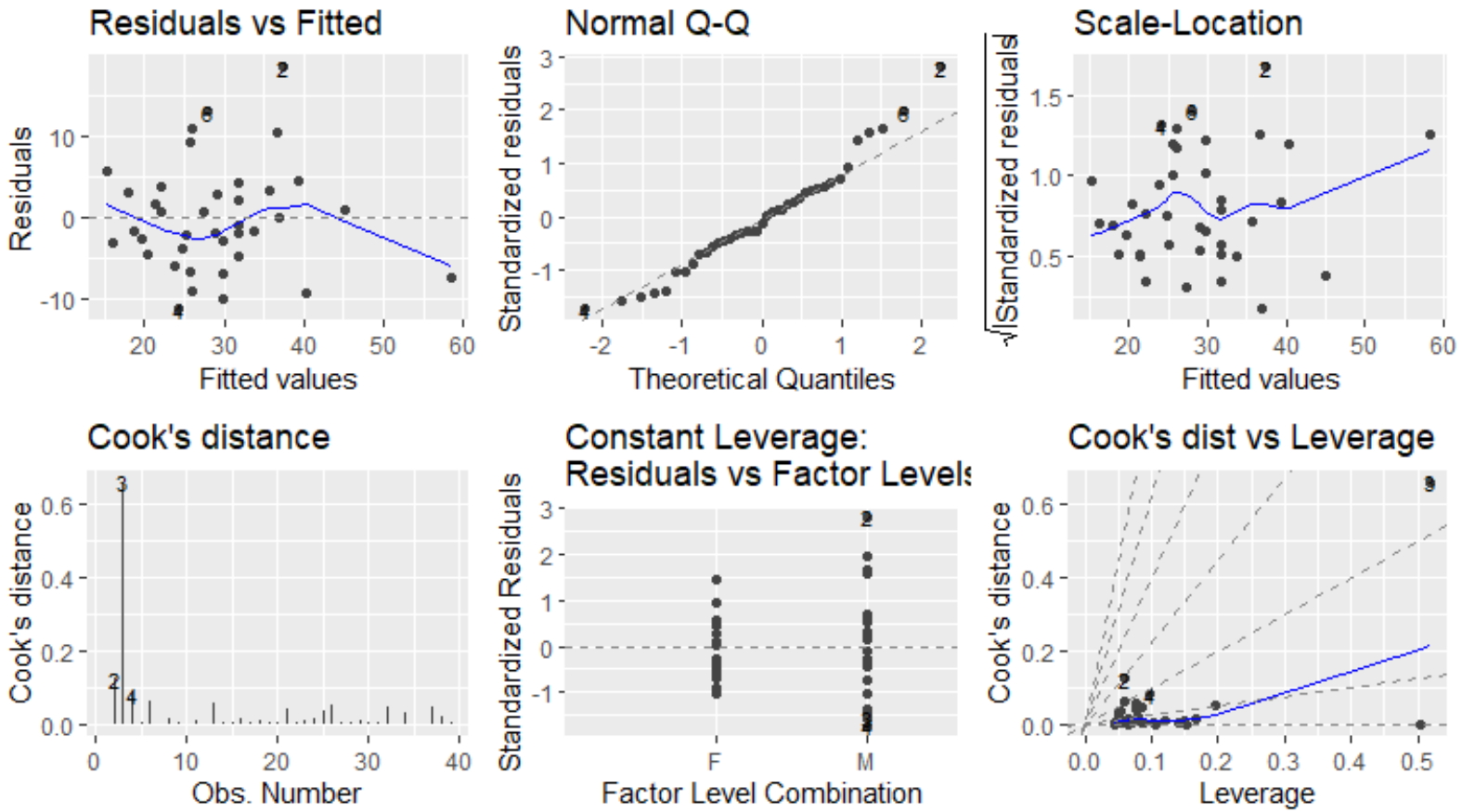
There is some evidence of collinearity ($VIF > 5$) which shouldn't surprise given the interaction term present.

Estimated parameter interpretations

- $\hat{\beta}_0 = 16.5866$ is predicted year 2 maths score for males who score 0 in year 1.
- $\hat{\gamma} = -4.75849$ is predicted difference in year 2 maths score for females compared to males who score 0 in year 1.
- $\hat{\beta}_m = 0.9504$ is predicted change in year 2 maths score for males for each additional point in year 1.
- $\hat{\delta} = -0.08754$ is predicted difference in the change in year 2 maths score for females compared to males for each additional point in year 1.

R examples – model 2

Diagnostic plots



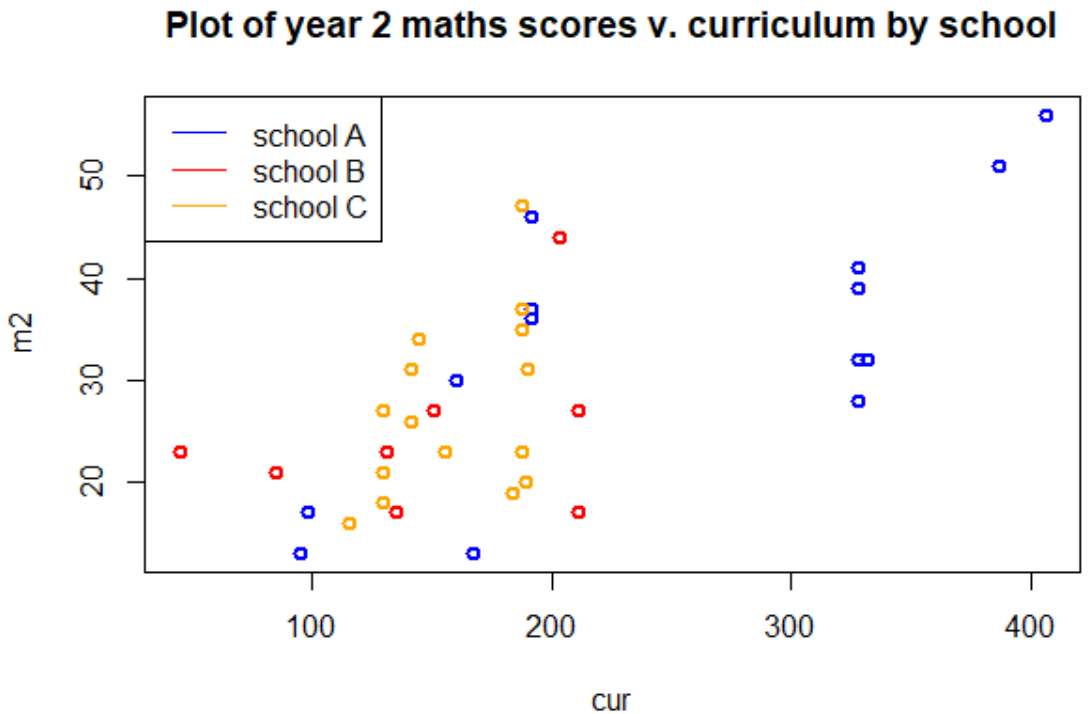
Normality (Shapiro-Wilk $p = 0.2755$) and independence ($DW = 1.867$, $p = 0.6731$) assumptions look OK, but perhaps slight evidence of non-constant variance.

R examples – model 3

The third model we consider is

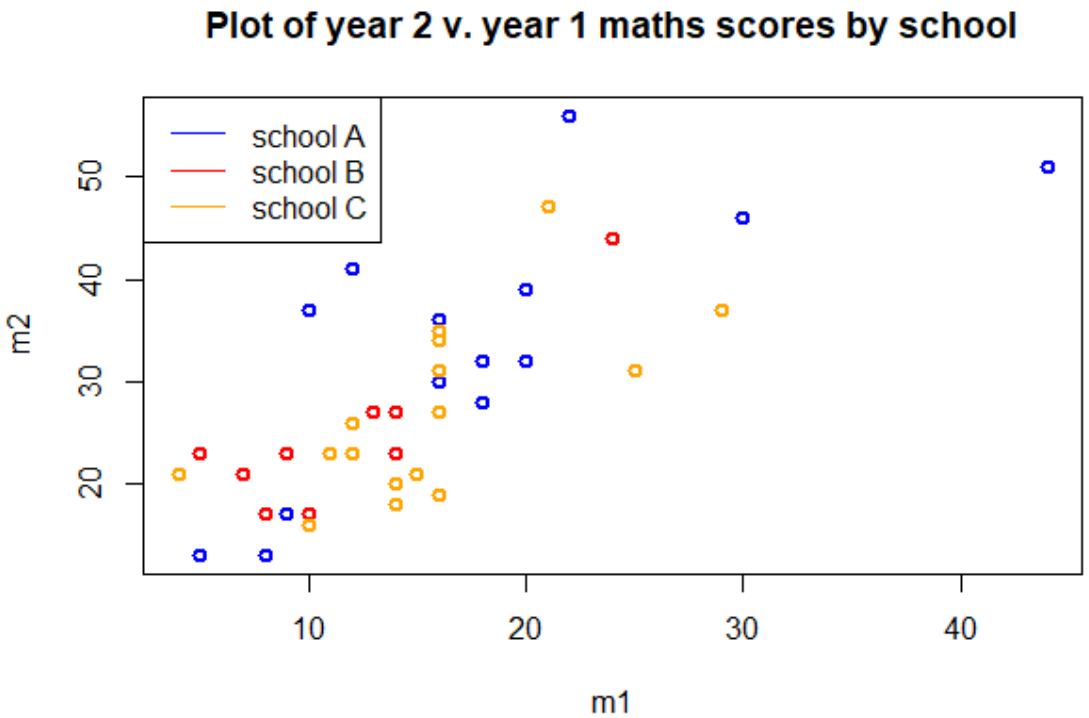
$$M2 = \beta_0 + \gamma_B schB + \gamma_C schC + \beta_c cur$$
$$+\beta_m m1 + \delta_B schB \times m1 + \delta_C schC \times m1 + \epsilon.$$

Scatter plots



It appears that *cur* is significant but *sex* appears insignificant (no strong evidence of different intercepts).

R examples – model 3



It appears that *m*1 is significant but *sex* appears insignificant (no strong evidence of different intercepts) and the interaction term involving *sch* and *m*1 appears insignificant (no strong evidence of different slopes).

R examples – model 3

Fitted model

R produced the summary information below.

```
Call:
lm(formula = m2 ~ cur + sch * m1, data = maths.data)

Residuals:
     Min       1Q   Median       3Q      Max
-10.4004  -5.2165  -0.0468   4.6672  14.4615

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.27848    4.88861   2.307  0.02767 *
cur          0.03816    0.02064   1.849  0.07370 .
schB        -3.74220    7.03315  -0.532  0.59835
schC        -3.27991    6.49180  -0.505  0.61685
m1           0.71903    0.23642   3.041  0.00467 **
schB:m1      0.28502    0.48530   0.587  0.56112
schC:m1      0.10893    0.37601   0.290  0.77392
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.077 on 32 degrees of freedom
Multiple R-squared:  0.633,     Adjusted R-squared:  0.5642
F-statistic: 9.198 on 6 and 32 DF,  p-value: 7.112e-06
```

R examples – model 3

The fitted model is

$$\widehat{m2} = 11.27848 - 3.7422 schB - 3.72991 schC + 0.03816 cur + 0.71903 m1 + 0.28502 schB \times m1 + 0.10893 schC \times m1.$$

or

$$\widehat{m2} = \begin{cases} 11.2785 + 0.03816 cur + 0.71903 m1, & school = A \\ 7.53628 + 0.03816 cur + 1.00405 m1, & school = B \\ 7.54857 + 0.03816 cur + 0.82796 m1, & school = C \end{cases}.$$

R examples – model 3

Significance

At $\alpha = 0.05$ we see that

- the model is significant (reject F-test hypothesis $\gamma_B = \gamma_C = \beta_c = \beta_m = \delta_B = \delta_C = 0$ with $p = 7.112 \times 10^{-6}$)
- the predictor *cur* is insignificant (retain T-test hypothesis $\beta_c = 0$ with $p = 0.0737$)
- the predictor $m1$ is significant (reject T-test hypothesis $\beta_m = 0$ with $p = 0.00467$).

To determine the significance of the categorical predictor *sch* and the interaction between *sch* and $m1$ we need to perform partial F-tests.

We will perform the partial F-test for the interaction between *sch* and $m1$.

To do so we need to know the contribution the interaction term makes to the sum square regression given the presence of *sch*, *cur* and $m1$ in the model.

R examples – model 3

We can find this contribution with the ANOVA table below.

```
Analysis of Variance Table

Response: m2
          Df  Sum Sq Mean Sq F value    Pr(>F)
cur        1 1885.03 1885.03 37.6365  7.35e-07 ***
sch        2   10.86    5.43  0.1085 0.8975389
m1         1  849.58  849.58 16.9627 0.0002507 ***
sch:m1     2   18.71    9.36  0.1868 0.8305137
Residuals 32 1602.73   50.09
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA table produced with the "anova" function in R is the so-called "Type I" or "sequential" sum squares, i.e.

- 1885.03 – SSR for model with *cur* as predictor
- 10.86 – addition to SSR by adding *sch* to model with *cur* as predictor
- 849.58 – addition to SSR by adding *m*1 to model with *cur* and *sch* as predictors
- 18.71 – addition to SSR by adding *sch* and *m*1 interaction to model with *cur*, *sch* and *m*1 as predictors.

R examples – model 3

The **order of the terms appearing in Type I sum squares matters**, so we can only use the p-value in this output to test the interaction term.

(Note that for models with categorical variables only and equal observations for each variable combination the order does not matter, but that is not the case here.)

At $\alpha = 0.05$ we see the interaction term is insignificant (retain partial F-test hypothesis $\delta_B = \delta_C = 0$ with $p = 0.8305137$).

The next step would be to remove the interaction term from the model and re-do the analysis (omitted).

R examples – model 3

Collinearity

```
             GVIF Df GVIF^(1/(2*Df))
cur      2.326270  1        1.525212
sch     37.475463  2        2.474211
m1       2.543256  1        1.594759
sch:m1  35.818001  2        2.446388
```

We won't go into the details, but the statistics in the second column are the ones we should compare to our usual threshold values – here we don't have strong evidence of collinearity.

Estimated parameter interpretations

- $\hat{\beta}_0 = 11.27848$ is predicted year 2 maths score for school A students with 0 curriculum coverage who score 0 in year 1.
- $\hat{\gamma}_B = -3.7422$ is predicted difference in year 2 maths score for school B students compared to school A students with equal curriculum coverage who score 0 in year 1.
- $\hat{\gamma}_C = -3.72991$ is predicted difference in year 2 maths score for school C students compared to school A students with equal curriculum coverage who score 0 in year 1.

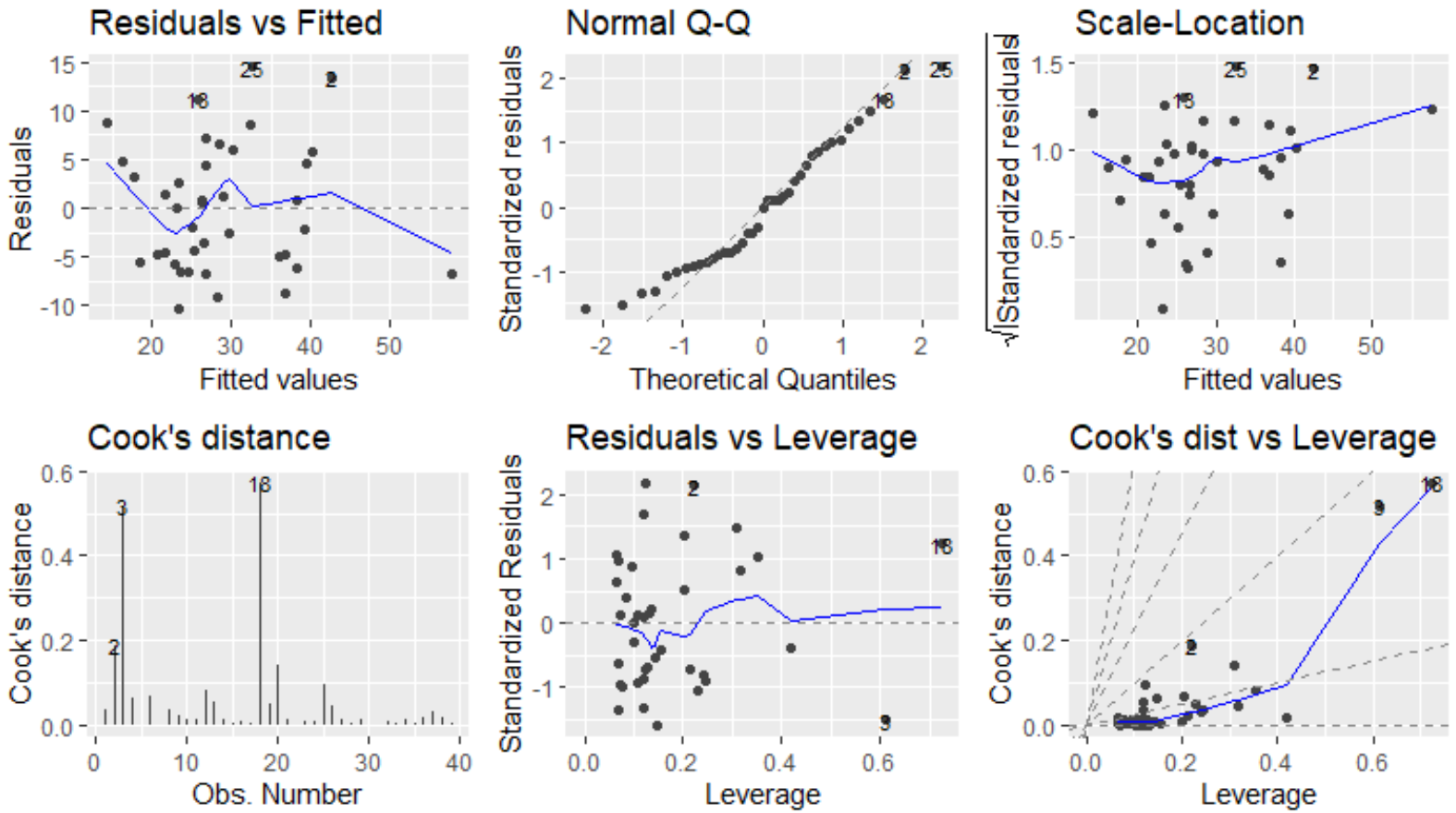
R examples – model 3

Estimated parameter interpretations (cont.)

- $\hat{\beta}_c = 0.03816$ is predicted change in year 2 maths score for students from all schools with equal year 1 maths score for each additional unit of curriculum coverage.
- $\hat{\beta}_m = 0.71906$ is predicted change in year 2 maths score for students from school A with equal curriculum coverage for each additional point in year 1.
- $\hat{\delta}_B = 0.28502$ is predicted difference in the change in year 2 maths score for students from school B compared to students from school A with equal curriculum coverage for each additional point in year 1.
- $\hat{\delta}_C = 0.10893$ is predicted difference in the change in year 2 maths score for students from school C compared to students from school A with equal curriculum coverage for each additional point in year 1.

R examples – model 3

Diagnostic plots



Normality (Shapiro-Wilk $p = 0.1242$), independence ($DW = 1.9439$, $p = 0.618$) and constant variance assumptions look OK.

References I

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