# MAT 3375
# Regression Analysis

# Chapter 4
# Extensions of the OLS Model

P. Boily (uOttawa)

Summer – 2023

# Outline

# 4 – Extensions of the OLS Model

We have seen that we can fairly easily extend simple linear regression to multiple linear regression with minimal disruption, simply by using the appropriate matrix notation.

In practice, the MLR assumptions are rarely met; we have also present ways in which we can identify departures from the assumptions, and how we can remedy this situation.

In this chapter, we will discuss more sophisticated extensions of linear regression, extensions that get closer to real-life applications.

# 4.1 – Muticollinearity and Variance Inflation

The multiple linear regression **normal equations** are

$$(\mathbf{X}^\top\mathbf{X})\mathbf{b} = \mathbf{X}^\top\mathbf{Y}.$$

When $\mathbf{X}^\top\mathbf{X}$ is **invertible**, the solution $\mathbf{b} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$ is **unique**.

If one of the variables is a non-trivial linear combination of other variables

$$X_k = \alpha_{j_1}X_{j_1} + \cdots + \alpha_{j_\ell}X_{j_\ell},$$

then $\text{rank}(\mathbf{X}^\top) = \text{rank}(\mathbf{X}^\top\mathbf{X}) < p$ and so $\mathbf{X}^\top\mathbf{X}$ is **singular** (not invertible), and the solution is not **unique** (the system in **under-determined**).

**Example:** consider the design matrix and vector response

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 1 & 2 \\ 1 & 1 & 2 & 3 \\ 1 & 3 & 3 & 6 \end{pmatrix} \quad \text{and} \quad \mathbf{Y} = \begin{pmatrix} 0 \\ 1 \\ 4 \end{pmatrix}.$$

Find the LS model $\mathrm{E}\{Y \mid (X_1, X_2, X_3)\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$.

**Solution:** we compute the constituents of the normal equations

$$\mathbf{X}^\top \mathbf{X} = \begin{pmatrix} 3 & 5 & 6 & 11 \\ 5 & 11 & 12 & 23 \\ 6 & 12 & 14 & 26 \\ 11 & 23 & 26 & 49 \end{pmatrix} \quad \text{and} \quad \mathbf{X}^\top \mathbf{Y} = \begin{pmatrix} 5 \\ 13 \\ 14 \\ 27 \end{pmatrix}.$$

The row echelon form of $[\mathbf{X}^\top\mathbf{X} \mid \mathbf{X}^\top\mathbf{Y}]$ is

$$\begin{pmatrix} 1 & 0 & 0 & 0 & -2 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

meaning that $\boldsymbol{b} = (-2, 1-s, 1-s, s)$ provides a LS solution for all $s \in \mathbb{R}$. Furthermore, we also cannot compute the corresponding variance-covariance matrix $\sigma^2\{\mathbf{b}\} = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$.

In practice, it is quite rate that a predictor is an **exact** linear combination of other predictors; when it is almost so, however, the design matrix may be nearly **singular** (**ill-conditioned**), leading to **uncertainty** in the parameter vector $\mathbf{b}$ that solves the normal equations (and "**wrong coefficient signs**").

In multiple linear regression, the **variance inflation factor for $\beta_k$** is

$$\mathrm{VIF}_k = \frac{1}{1 - R_k^2}, \quad k = 1, \ldots, p,$$

where $R_k^2$ is the coefficient of multiple determination obtained when $X_k$ is regressed on the other $p - 2$ predictor variables in the model.

Note that if $X_k$ is **very nearly** a linear combination of the other predictors, then $R_k^2 \approx 1$, yielding a **large** $\mathrm{VIF}_k$, which influence the least-squares estimates. In practice, $\max_k \mathrm{VIF}_k > 10$ implies that there are likely crucial problems with multicollinearity.

Remedial measures include **centering the data**, **ridge regression**, and **principal component regression**.

**Example:** consider the following dataset

| $X_1$ | $X_2$ | $X_3$ | $X_4$ | $Y$ |
|---|---|---|---|---|
| 1 | 1 | 2.063 | 1 | 2.995 |
| 2 | 1 | 3.184 | 1 | 3.773 |
| 1 | 1 | 2.131 | 2 | 2.846 |
| 2 | 1 | 2.867 | 2 | 3.963 |
| 1 | 2 | 3.104 | 1 | 5.291 |
| 2 | 2 | 3.876 | 1 | 6.070 |
| 1 | 2 | 2.999 | 2 | 5.034 |
| 2 | 2 | 3.865 | 2 | 6.014 |

Compare the linear models

$$\mathrm{E}\{Y \mid (X_1, X_2, X_3)\} \quad \text{and} \quad \mathrm{E}\{Y \mid (X_1, X_2, X_4)\}.$$

**Solution:** the R output for the first of these is

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.08808    0.25637  -0.344   0.7485
X1           1.15062    0.43523   2.644   0.0574 .
X2           2.45248    0.44756   5.480   0.0054 **
X3          -0.27147    0.48792  -0.556   0.6076
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1236 on 4 degrees of freedom
Multiple R-squared:  0.9947,Adjusted R-squared:  0.9907
F-statistic: 249.1 on 3 and 4 DF,  p-value: 5.303e-05
```

The output for the second model is

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.08134    0.22254  -0.365 0.733254
X1           0.91339    0.08411  10.859 0.000408 ***
X2           2.20826    0.08411  26.253 1.25e-05 ***
X4          -0.06812    0.08411  -0.810 0.463473
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.119 on 4 degrees of freedom
Multiple R-squared:  0.9951,Adjusted R-squared:  0.9914
F-statistic: 269.3 on 3 and 4 DF,  p-value: 4.545e-05
```

The estimated parameters $b_0$, $b_1$, and $b_2$ are **quite similar** in both models, but the standard errors are **starkingly different**; the confidence intervals in the second model are **much tighter** for $\beta_1$ and $\beta_2$ than they are in the first model.

Why is this? Note that $\text{VIF}_1 \approx \text{VIF}_2 \approx \text{VIF}_4 \approx 1$ in the second model (the predictors are **linearly independent**), whereas $\text{VIF}_1 \approx \text{VIF}_2 \approx \text{VIF}_3 \approx 25$ in the first model.

This should not come as a surprise, as $X_3$ **is very nearly a linear combination of** $X_1$ **and** $X_2$:

$$\|X_3 - X_1 - X_2\|_2^2 \approx 0.324,$$

whereas $\|X_1\|_2^2 \approx 4.47$, $\|X_2\|_2^2 \approx 4.47$, and $\|X_3\|_2^2 \approx 8.70$.

# 4.2 – Polynomial Regression

In a dataset with a predictor $X$ and a response $Y$, both numerical, if the relationship between $X$ and $Y$ is **not linear**, we may consider transforming the data so that the relationship between $X'$ and $Y'$ is **so**, fitting a **linear LS** model to these new variables, and inverting the results to obtain a relationship between the original $X$ and $Y$.

Another approach is to create a sequence of predictors

$$X_1 = X, \; X_2 = X^2, \quad \ldots, \quad X_k = X^k$$

and to treat the entire situation as a MLR model

$$\mathrm{E}\{Y|(X_1, \ldots, X_k)\} = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k = \beta_0 + \beta_1 X + \cdots + \beta_k X^k.$$

**Example:** fit the following data

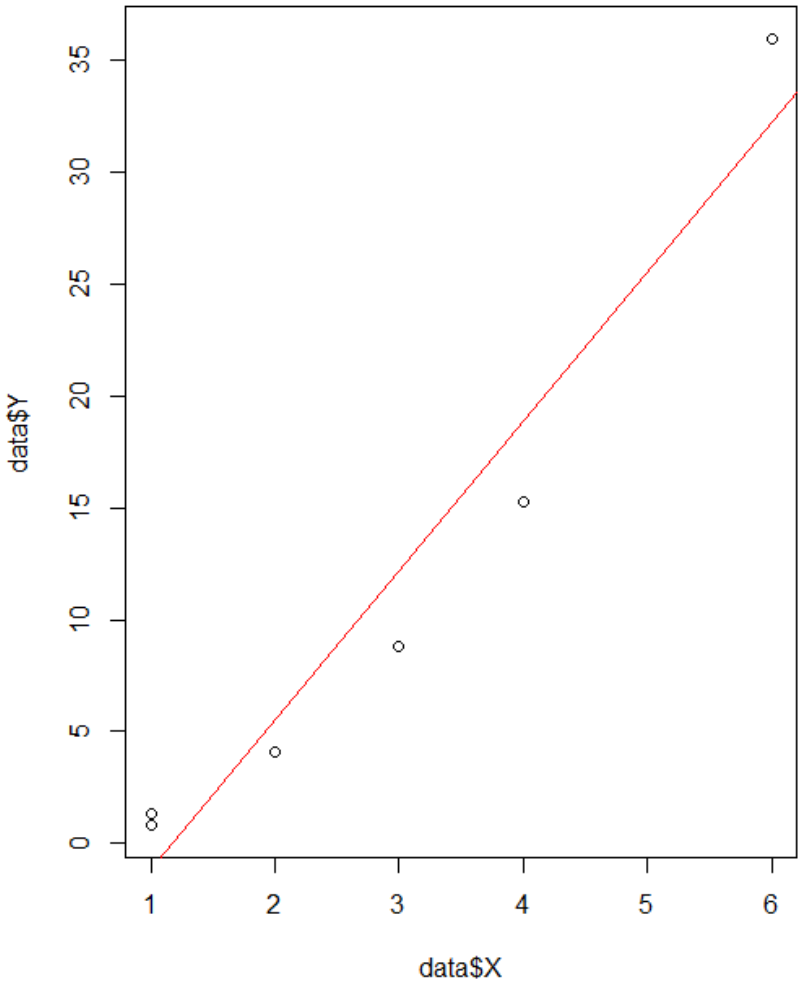
| $X$ | 1 | 1 | 2 | 4 | 3 | 6 |
|---|---|---|---|---|---|---|
| $Y$ | 0.8 | 1.3 | 4.1 | 15.3 | 8.8 | 36 |

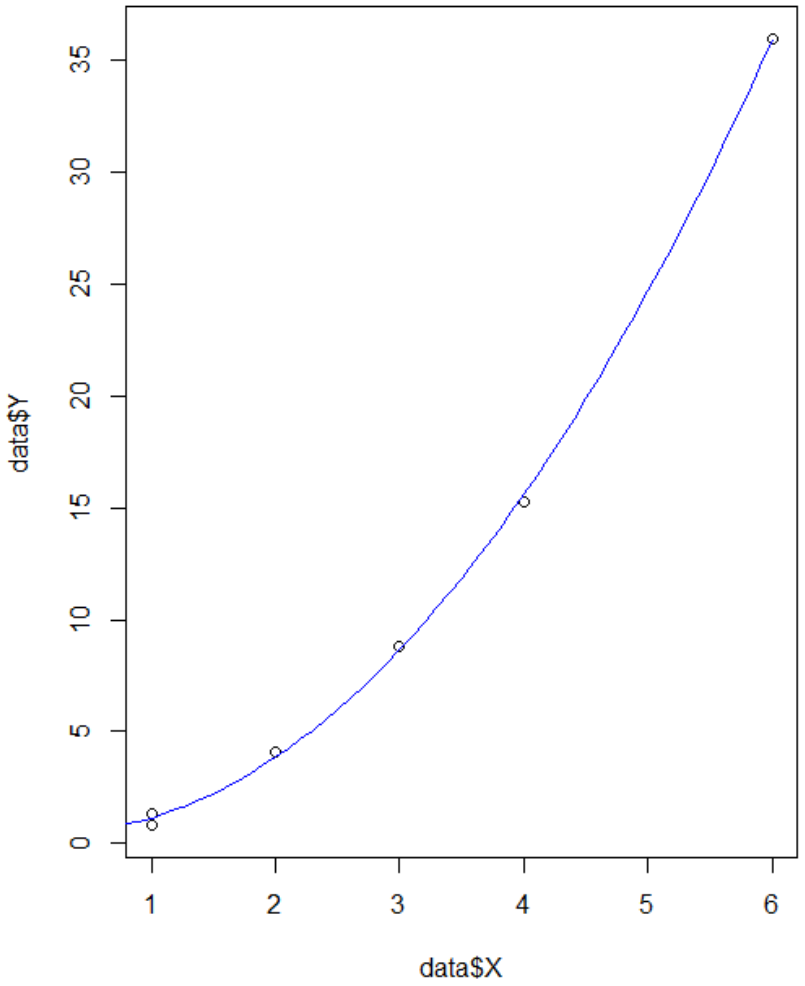
**Solution:** we can fit a linear model to the data

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -7.913      2.734  -2.895  0.04435 *
X              6.693      0.818   8.182  0.00122 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.55 on 4 degrees of freedom
Multiple R-squared:  0.9436, Adjusted R-squared:  0.9295
F-statistic: 66.94 on 1 and 4 DF,  p-value: 0.001215
```

The fit seems decent ($R_a^2 = 0.9295$), but a plot of the data suggests that something is astray: visually, the quadratic fit seems better ($R_a^2 = 0.9994$).

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.56635    0.47768   1.186 0.321128
X           -0.49591    0.34935  -1.420 0.250809
X2           1.06466    0.05046  21.101 0.000233 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3354 on 3 degrees of freedom
Multiple R-squared:  0.9996,Adjusted R-squared:  0.9994
F-statistic:  3973 on 2 and 3 DF,  p-value: 7.331e-06
```


data$Y
data$X
0
5
10
15
20
25
30
35
1
2
3
4
5
6


data$Y
data$X
0
5
10
15
20
25
30
35
1
2
3
4
5
6

One thing we notice is that of the three coefficients, only the quadratic $b_2$ is significant at $\alpha = 0.05$, even though the fit seemed **quite tight**, visually. Part of the problem is that although the relationship between $X$ and $X^2$ is **not linear**, they are still **correlated**, leading to a fairly high VIF term:

$$\text{VIF}_1 = \frac{1}{1 - R_1^2} = \frac{1}{1 - 0.9510685} = 20.43673.$$

This is typical of polynomial regression: the suggested remedial measure is to use **centered predictors** $x_i = X_i - \overline{X}$.

The quadratic fit of the previous example would take the same form:

$$\begin{aligned} \text{E}\{Y\} &= \beta_0 + \beta_1(X - \overline{X}) + \beta_2(X - \overline{X})^2 \\ &= \left\{\beta_0 - \beta_1\overline{X} + \beta_2\overline{X}^2\right\} + \left\{\beta_1 - 2\beta_2\overline{X}\right\} X + \beta_2 X^2 = \beta_0' + \beta_1' X + \beta_2' X^2 \end{aligned}$$

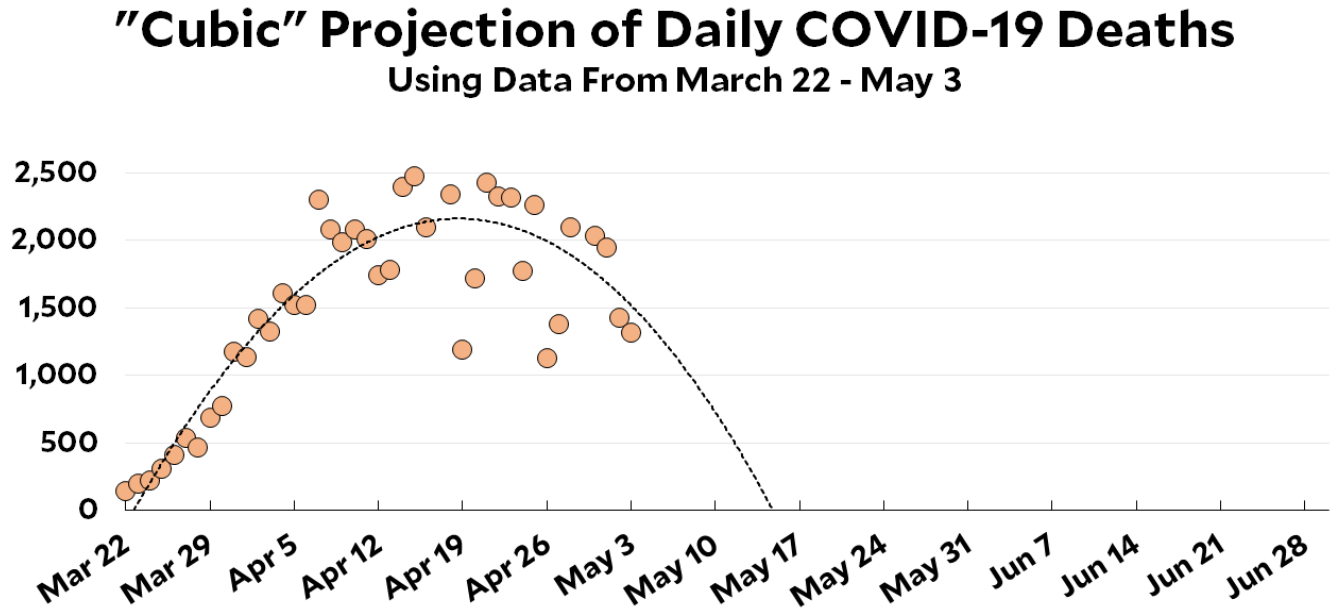
but now **all** coefficients are significant at $\alpha = 0.05$:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.70814    0.20935   36.82 4.41e-05 ***
Xm           5.53718    0.09472   58.46 1.10e-05 ***
X2m          1.06466    0.05046   21.10 0.000233 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3354 on 3 degrees of freedom
Multiple R-squared:  0.9996,Adjusted R-squared:  0.9994
F-statistic:  3973 on 2 and 3 DF,  p-value: 7.331e-06
```

Not surprisingly, the centered $\mathrm{VIF}_1$ is much lower at $1.502374$.

The rest of the ordinary least square machinery easily carries over.

Graphically and/or mathematically, then, polynomial regression can prove quite powerful and convenient to use. But convenience is not always a sufficient reason to use a regression model...



**Example:** we fit a response variable against a centered cubic regression with predictor $x = X - \overline{X}$ by adding one variable at a time to obtain the simple linear model

$$\mathrm{E}\{Y \mid x\} = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3.$$

The regression's sum of squares summary is shown below:

| source of variation | df | SS |
|---|---|---|
| $x$ | 1 | 485.71303 |
| $x^2 \vert x$ | 1 | 9.11434 |
| $x^3 \vert x, x^2$ | 1 | 6.33018 |
| Error | 23 | 285.50912 |

Using $\alpha = 0.05$, test for $H_0 : \beta_2 = \beta_3 = 0$ against $H_1 : \beta_2 \neq 0$ or $\beta_3 \neq 0$.

**Solution:** if $H_0$ holds, the statistic

$$F^* = \frac{\text{SSR}(R)/(p-q)}{\text{SSR}(F)/(n-p)} = \frac{\text{SSR}(x^2, x^3|x)/(p-q)}{\text{SSE}(x, x^2, x^3)/(n-p)}$$

follows a $F(p-q, n-p)$ distribution, where $q = 2$ is the number of parameters in the **reduced model** and $n - p = n - 4 = 23$ is the df of the error, so that $n = 27$.

With $\alpha = 0.05$, the critical value is $F(0.95; 2, 23) = 3.422$. Since

$$F^* = \frac{\left[\text{SSR}(x^2|x) + \text{SSR}(x^3|x, x^2)\right]/2}{\text{SSE}(x, x^2, x^3)/23} = \frac{(9.114 + 6.332)/2}{285.509/23} = 0.622,$$

then $F^* < F(0.95; 2, 23)$ and we **cannot** reject $H_0$ at $\alpha = 0.05$.

# 4.3 – Interaction Effects

We have seen that we can extend simple linear regression in $X$ to include higher power terms (after centering the data to minimize the effects of multicolinearity).

There is nothing to stop us from doing so with any number of predictors $X_1, \ldots, X_p$, leading to an **additive model**

$$E\{Y\} = f_1(X_1) + \cdots + f_p(X_p),$$

where the $f_i$ are **polynomial functions** in 1 variable (this could be modified to any linear function of the regression coefficients $\beta_{i,j}$).

Assume that $p = 2$ for simplicity's sake.

We can refine the model with an **interaction term** $f_3(X_1, X_2) = \beta_3 X_1 X_2$. In keeping with the **hierarchical principle**, we might consider the model

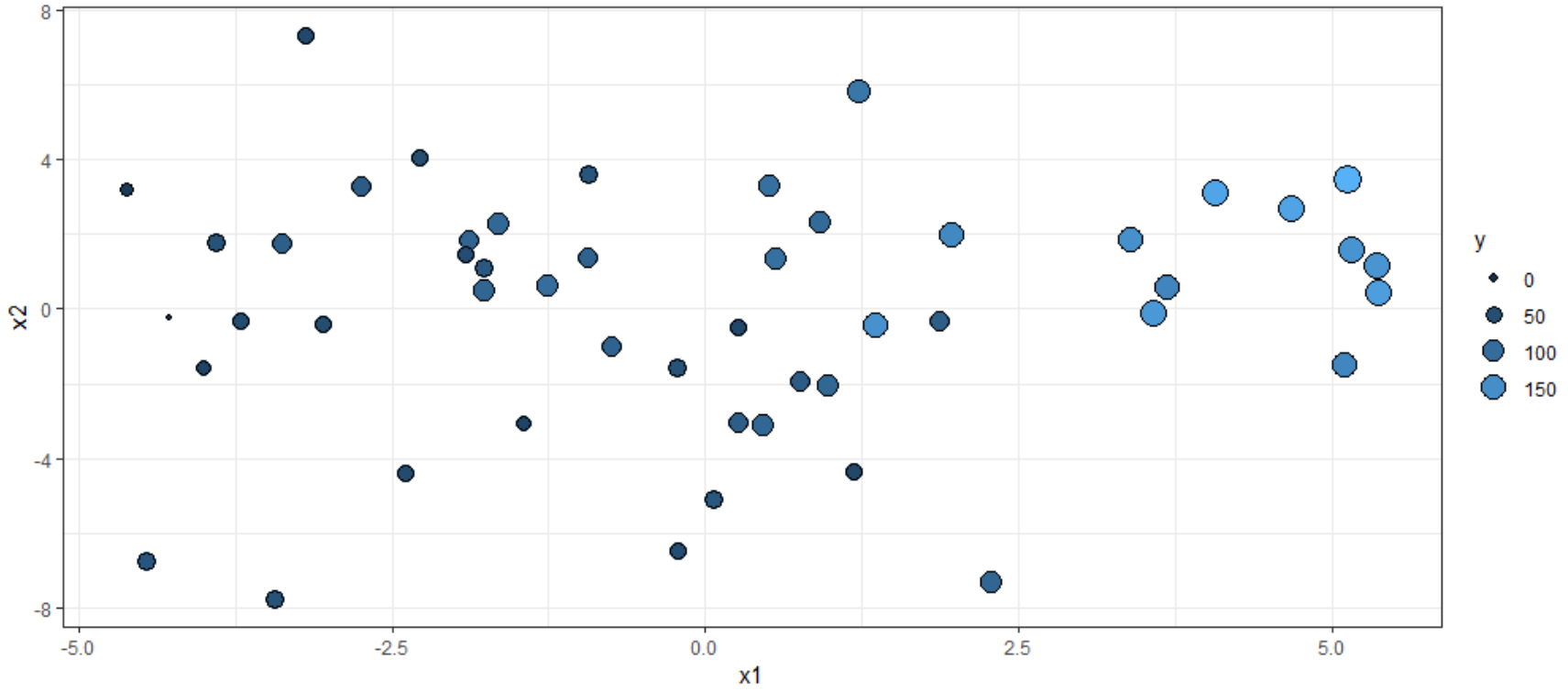
$$\begin{aligned} \mathrm{E}\{Y\} &= f_1(X_1) + f_2(X_2) + f_3(X_1, X_2) \\ &= \beta_0 + \beta_{1,1} X_1 + \beta_{2,1} X_2 + \beta_{1,2} X_1^2 + \beta_3 X_1 X_2 + \beta_{2,2} X_2^2, \end{aligned}$$

although there could also be good reasons to consider something like

$$\mathrm{E}\{Y\} = \beta_0 + \beta_1 X + \beta_2 X_2 + \beta_3 X_1 X_2.$$

In the latter case, if we assume that $\beta_1 \beta_2 > 0$, then if $\beta_1 \beta_3 > 0$, we have a **reinforcement interaction**; if $\beta_1 \beta_3 < 0$, we have an **interference interaction**.

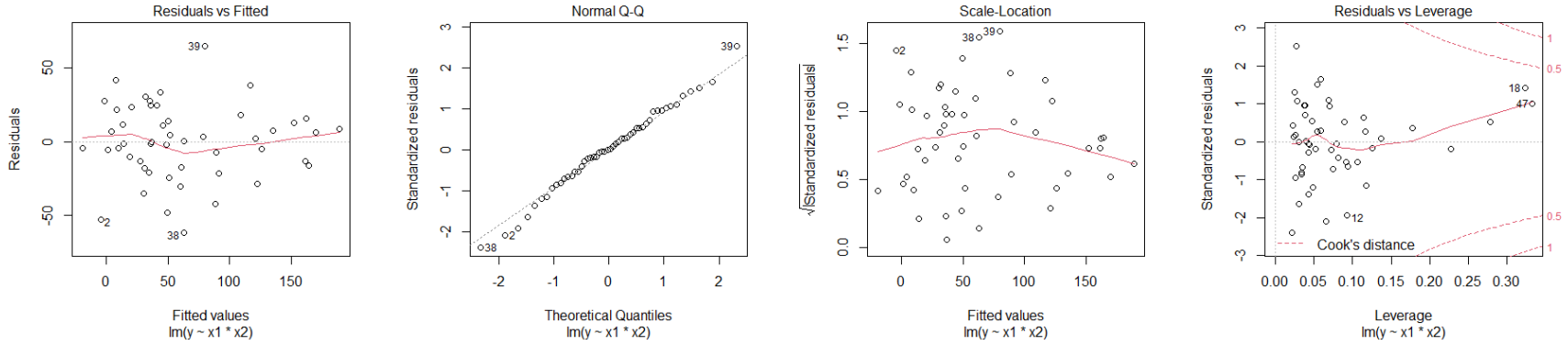
**Example:** we consider a dataset with $n = 50$ observations (2 centered predictors $X_1, X_2$ and a response $Y$, see below).



We compute the fit for the reduced and the full interaction models. The former exhibits reinforcement interaction ($\beta_1\beta_3 > 0$).

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  61.7494     3.7043  16.669  < 2e-16 ***
x1           15.6463     1.3017  12.020 8.55e-16 ***
x2            5.1396     1.2010   4.279 9.40e-05 ***
x1:x2         1.6886     0.4379   3.856 0.000356 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26.06 on 46 degrees of freedom
Multiple R-squared:  0.8166,Adjusted R-squared:  0.8047
F-statistic: 68.28 on 3 and 46 DF,  p-value: < 2.2e-16
```

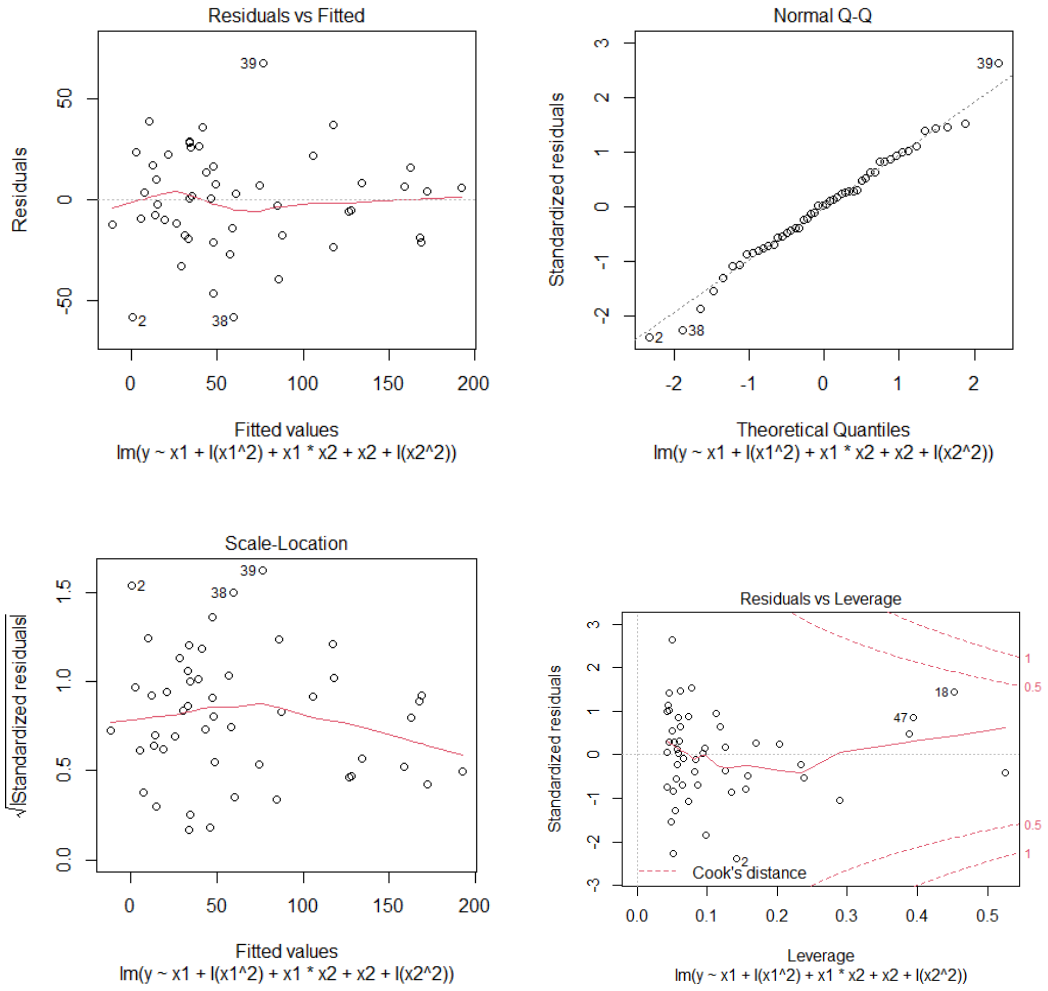
The summary indicates that the reduced interaction linear model is **appropriate**, which is **supported** by the diagnostic plots:




But what about the full model?

The pure quadratic terms are **not significant**, which suggests that the reduced model is **likely** (although **not necessarily**) a better choice.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 58.25684    5.94511   9.799 1.24e-12 ***
x1          15.36026    1.38371  11.101 2.42e-14 ***
I(x1^2)      0.41459    0.46486   0.892 0.377316
x2           4.91100    1.31831   3.725 0.000553 ***
I(x2^2)      0.01042    0.26562   0.039 0.968891
x1:x2        1.56368    0.46519   3.361 0.001613 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26.4 on 44 degrees of freedom
Multiple R-squared:  0.8199,Adjusted R-squared:  0.7994
F-statistic: 40.06 on 5 and 44 DF,  p-value: 2.654e-15
```


Residuals vs Fitted
Residuals
Fitted values
lm(y ~ x1 + I(x1^2) + x1 * x2 + x2 + I(x2^2))
Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(y ~ x1 + I(x1^2) + x1 * x2 + x2 + I(x2^2))
Scale-Location
√|Standardized residuals|
Fitted values
lm(y ~ x1 + I(x1^2) + x1 * x2 + x2 + I(x2^2))
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage
lm(y ~ x1 + I(x1^2) + x1 * x2 + x2 + I(x2^2))

# 4.4 – ANOVA/ANCOVA Models for Categorical Variables

We can also include categorical variables within the OLS framework. Suppose there are $K$ treatments (levels) for predictor $X$.

1. In the **dummy variable** encoding, we set

$$X_j = \begin{cases} 1 & \text{treatment } j \\ 0 & \text{else} \end{cases}$$

for $j = 1, \ldots, K-1$. The ANOVA/OLS model is then

$$Y_i = \beta_0 + \sum_{j=1}^{K-1} \beta_j X_{i,j} + \varepsilon_i \quad \text{and} \quad E\{Y\} = \begin{cases} \beta_0 & \text{treatment } K \\ \beta_0 + \beta_j & \text{treatment } j \end{cases}$$

2. In the **treatment effect** encoding, we set

$$X_j = \begin{cases} 1 & \text{treatment } j \\ -1 & \text{treatment } K \\ 0 & \text{else} \end{cases}$$

for $j = 1, \ldots, K-1$. The ANOVA/OLS model is as in the dummy encoding case and

$$E\{Y\} = \begin{cases} \beta_0 - (\beta_1 + \cdots + \beta_{K-1}) & \text{treatment } K \\ \beta_0 + \beta_j & \text{treatment } j \end{cases}$$

Specific examples will illustrate the main principles.

# 4.5 – Weighted Least Squares

We have seen that the OLS regression model $\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ requires **constant variance**. When that assumption is not met – but in a "monotonic" manner, such as $\sigma^2\{\varepsilon_i\} = \sigma^2 x_i$, say – various data transformations on the predictors $X$ may be appropriate.

What do we do when the linearity assumption is valid, but the variance $\sigma_i$ does not change in a **systematic** manner?

One way to approach the problem is $via$ **weighted least squares** (WLS), which does not require all observations to be **treated equally** (i.e., to be given the **same weight**).

Let $w_i \geq 0$ be the weight of observation $i$ and write $Z_i = \sqrt{w_i}Y_i$.

Define the **weight matrix** as $\mathbf{W} = \text{diag}(w_1, \ldots, w_n)$.

The **WLS problem** is to find the coefficient vector $\boldsymbol{\beta}$ which **minimizes** the weighted sum of squared errors

$$\begin{aligned}
\text{SSE}_w &= Q_w(\boldsymbol{\beta}) = \|\mathbf{Z} - \hat{\mathbf{Z}}\|_2^2 \\
&= \|\sqrt{\mathbf{W}}\mathbf{Y} - \sqrt{\mathbf{W}}\hat{\mathbf{Y}}\|_2^2 = \|\sqrt{\mathbf{W}}\mathbf{Y} - \sqrt{\mathbf{W}}\mathbf{X}\boldsymbol{\beta}\|_2^2 \\
&= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{W}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\
&= \mathbf{Y}^\top\mathbf{W}\mathbf{Y} - \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{W}\mathbf{Y} - \mathbf{Y}^\top\mathbf{W}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{W}\mathbf{X}\boldsymbol{\beta}.
\end{aligned}$$

But $\nabla_{\boldsymbol{\beta}} Q_w(\boldsymbol{\beta}) = -2\mathbf{X}^\top\mathbf{W}\mathbf{Y} + 2\mathbf{X}^\top\mathbf{W}\mathbf{X}\boldsymbol{\beta}$, so that the WLS estimator $\mathbf{b}$ of $\boldsymbol{\beta}$ is

$$\nabla_{\boldsymbol{\beta}} Q_w(\boldsymbol{\beta}) = \mathbf{0} \implies \boldsymbol{b} = (\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{W}\mathbf{Y}.$$

The entire OLS machinery can then be used in the WLS context simply by replacing $\mathbf{Y}$ by $\sqrt{\mathbf{W}}\mathbf{Y}$ and $\mathbf{X}$ by $\sqrt{\mathbf{W}}\mathbf{X}$ throughout.

**Example:** consider a dataset with $n = 11$ observations

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $x$ | 0.82 | 1.09 | 1.22 | 1.24 | 1.29 | 1.30 | 1.36 | 1.38 | 1.39 | 1.40 | 1.55 |
| $y$ | 1.47 | 1.33 | 1.32 | 1.30 | 1.35 | 1.34 | 1.38 | 1.52 | 1.40 | 1.44 | 1.58 |

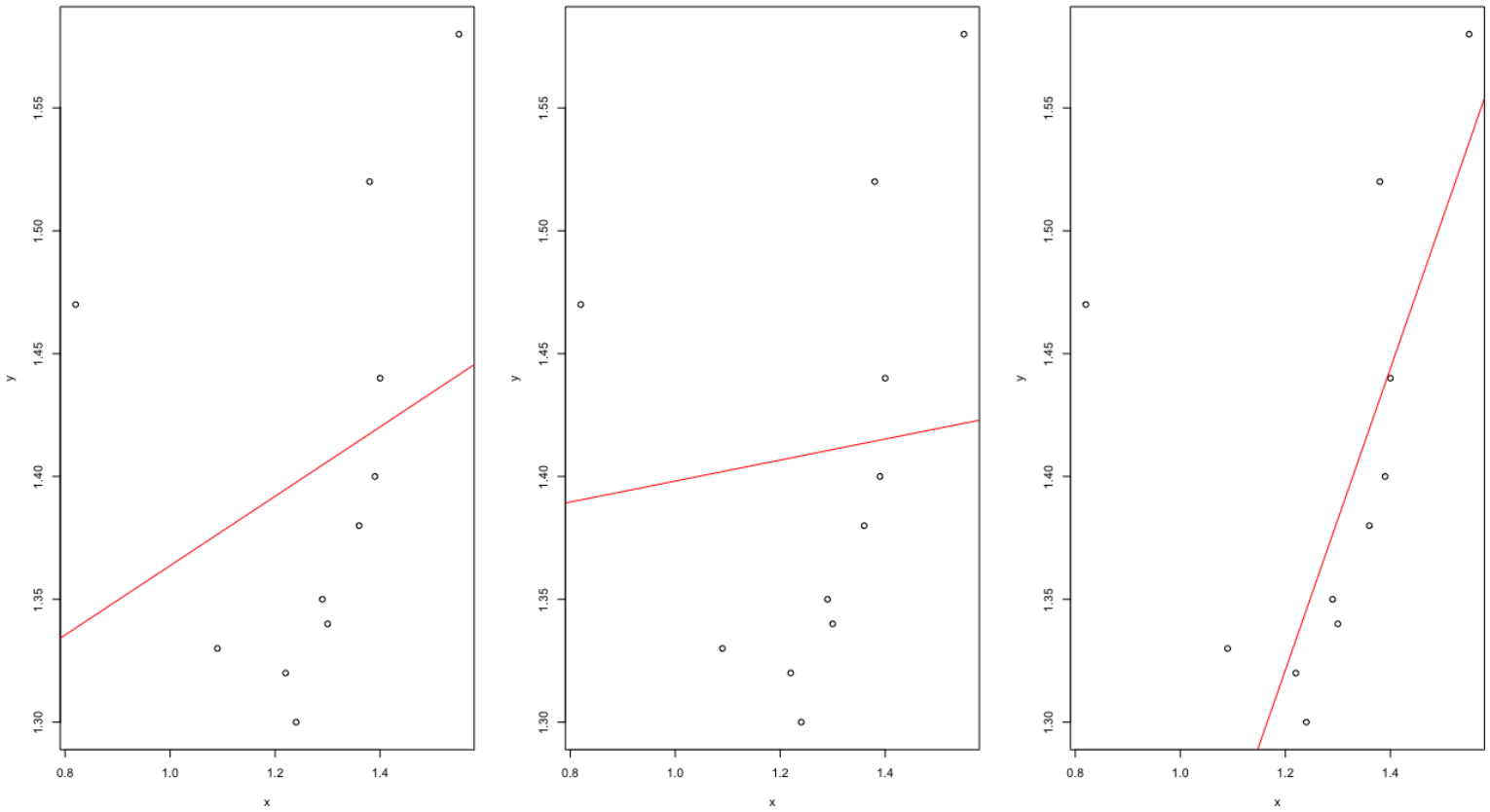
The OLS model is

$$\hat{y} = 1.223 + 0.1412x \quad \text{(left)};$$

the WLS model with $w_1 = 2$ and $w_i = 1$, $i = 2, \ldots, 11$ is

$$\hat{y} = 1.3553 + 0.0428x \quad \text{(middle)};$$

the OLS/WLS without the first observation is $\hat{y} = 0.5848 + 0.6136x$ (right).


y
x
1.30
1.35
1.40
1.45
1.50
1.55
0.8
1.0
1.2
1.4

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.09759 -0.06036 -0.03454  0.06123  0.13864

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.2225     0.1920   6.366  0.00013 ***
x             0.1412     0.1489   0.948  0.36782
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09047 on 9 degrees of freedom
Multiple R-squared:  0.09081, Adjusted R-squared:  -0.01021
F-statistic: 0.899 on 1 and 9 DF,  p-value: 0.3678
```

```
Weighted Residuals:
     Min       1Q   Median       3Q      Max
-0.10841 -0.07148 -0.03354  0.06517  0.15833

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.3553     0.1624   8.344 1.58e-05 ***
x             0.0428     0.1292   0.331    0.748
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09669 on 9 degrees of freedom
Multiple R-squared:  0.01204, Adjusted R-squared:  -0.09773
F-statistic: 0.1097 on 1 and 9 DF,  p-value: 0.748
```

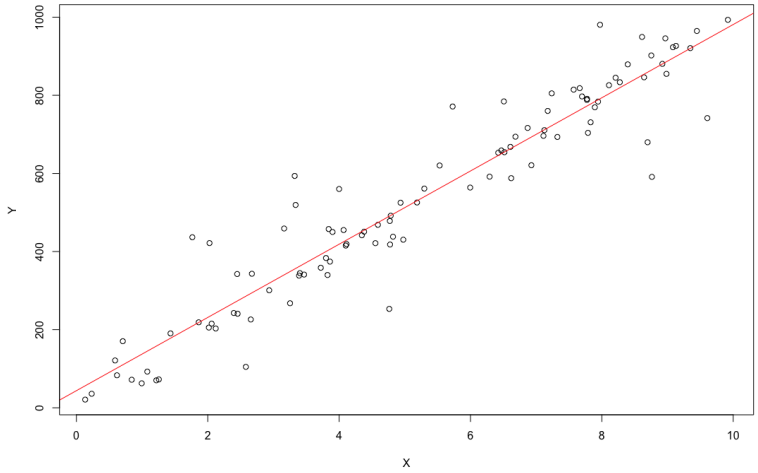
```
Residuals:
     Min       1Q   Median       3Q      Max
-0.04568 -0.03852 -0.01341  0.02205  0.08841

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.5848     0.1916   3.052   0.0158 *
x             0.6136     0.1444   4.250   0.0028 **
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05402 on 8 degrees of freedom
Multiple R-squared:  0.693, Adjusted R-squared:  0.6546
F-statistic: 18.06 on 1 and 8 DF,  p-value: 0.002801
```

How can WLS be used to deal with a error variance which is not constant?

We consider the underlying model

$$\mathbf{Y} \sim \mathcal{N}\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\{\boldsymbol{\varepsilon}\}\right), \quad \text{where} \quad \sigma^2\{\varepsilon_i\} = \sigma_i^2 \not\equiv \sigma^2,$$

such as may be found in the image below:



The procedure goes as in the OLS case, with some slight modifications:

1. if the $\sigma_i^2$ are known, we use the weights $w_i = \frac{1}{\sigma_i^2} \geq 0$;

2. if the $\sigma_i^2$ are unknown:
   (a) we use OLS and find the residuals $e_i$ ($e_i^2$ is an estimate of $\sigma_i^2$ when there are no $Y$-**outliers**; $|e_i|$ is an estimate of $\sigma_i$ when there are some);
   (b) depending on the choice made above, regress either $e_i^2$ or $|e_i|$ on $X_1, \ldots, X_{p-1}$ to obtain fitted values $\hat{v}_i$ or $\hat{s}_i$, which are point estimate of $\sigma_i^2$ or $\sigma_i$, respectively;
   (c) depending on the choice made above, use WLS with $w_i = \frac{1}{\hat{v}_i}$ or $w_i = \frac{1}{\hat{s}_i^2}$ and compute $\mathrm{SSE}_w$ and $\mathrm{MSE}_w = \frac{\mathrm{SSE}_w}{n-p}$.
   If $\mathrm{MSE}_w \approx 1$, the scaling is **appropriate**; otherwise, repeat steps (a) to (c), starting with the current **WLS residuals**.

**Example:** the number of defective items $Y$ produced by a machine is known to be linearly related to the speed setting $X$ of the machine:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad \varepsilon_i \text{ indép.}$$

An analyst regresses the squared residuals $e_i^2 = (\hat{Y}_i - Y_i)^2$ on the speed setting $X_i$ and obtains the following $n = 12$ fitted values:

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\hat{v}_i$ | 68.7 | 317.4 | 193 | 317.4 | 68.7 | 193 | 193 | 317.4 | 68.7 | 317.4 | 68.7 | 193 |

Then, using weighted LS with $w_i = \frac{1}{\hat{v}_i}$, she obtains residuals $e_i^w = \hat{Y}_i^w - Y_i$:

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $e_i$ | −3.6 | 5.6 | −13.5 | −16.4 | −9.6 | 7.5 | −10.5 | 26.6 | 14.4 | −17.4 | −1.6 | 18.5 |

Is the use of these weights appropriate?

**Solution :** we have

$$\text{SSE}_w = \sum_{i=1}^{12} w_i e_i^2 = \sum_{i=1}^{12} \frac{1}{\hat{v}_i} e_i^2 = 12.2953,$$

a sum of squares with $n - p = 12 - 2 = 10$ degrees of freedom, so that

$$\text{MSE}_w = \frac{\text{SSE}_w}{n-p} = \frac{12.2953}{10} = 1.22953.$$

Since $\text{MSE}_w \approx 1$, we have evidence that the weights are **appropriate** and that the initial $\hat{v}_i$ provide **reasonable approximations** of $\sigma_i^2$ for $i = 1, \ldots, 12$.

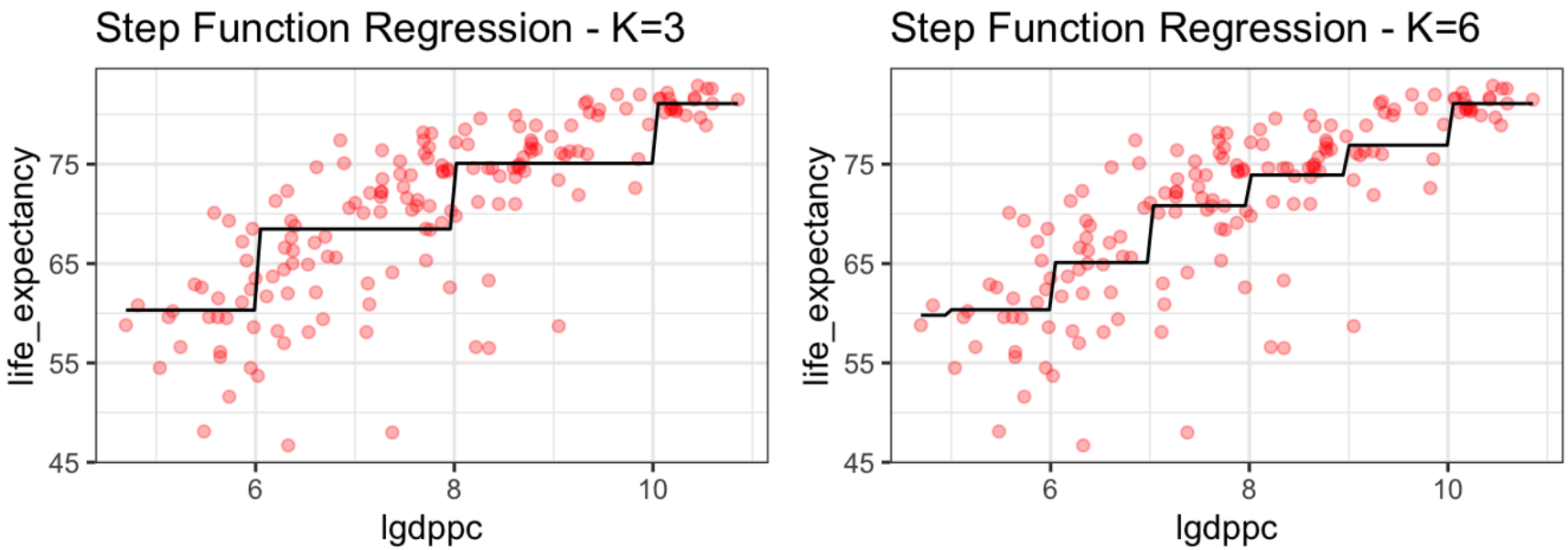
# 4.6 – Other Extensions

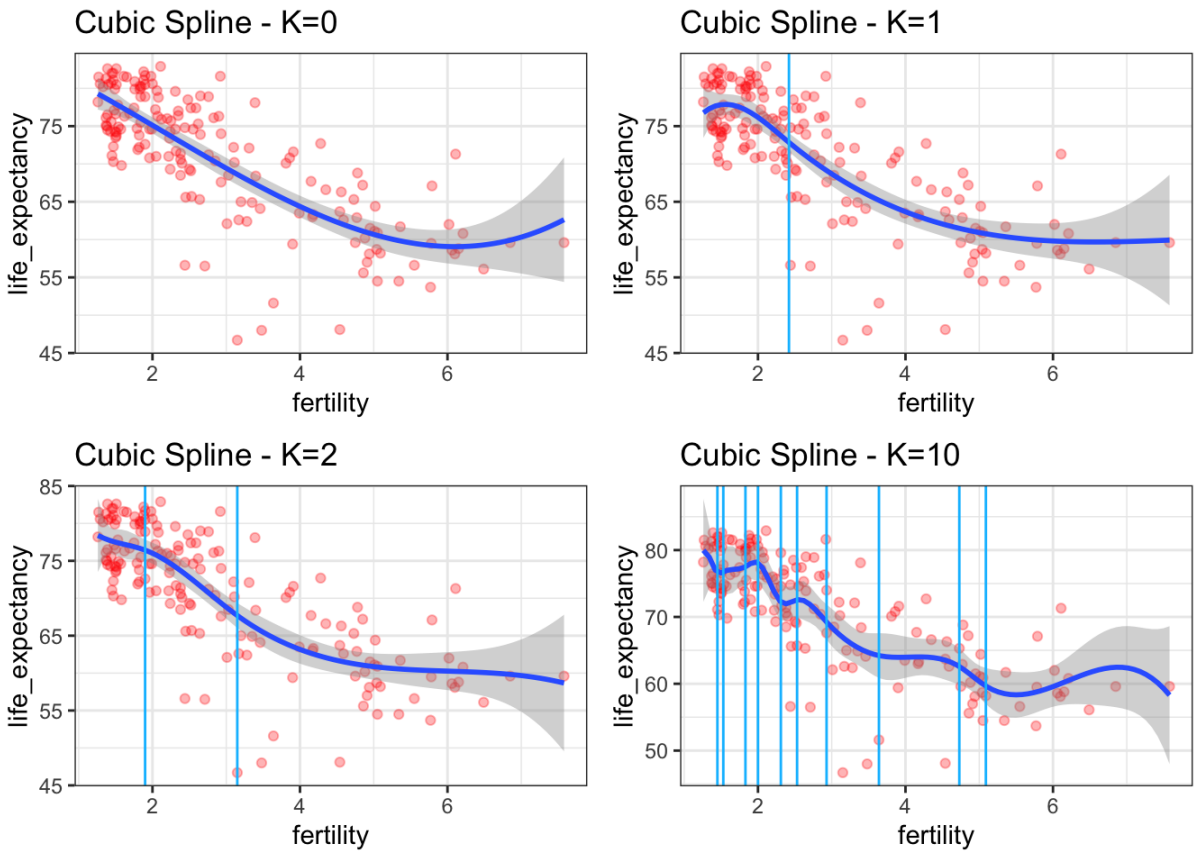
The LS assumptions are **convenient** from a mathematical perspective, but they are not always met in practice. One way out of this conundrum is to use **remedial measures** to transform the data into **compliant inputs**.

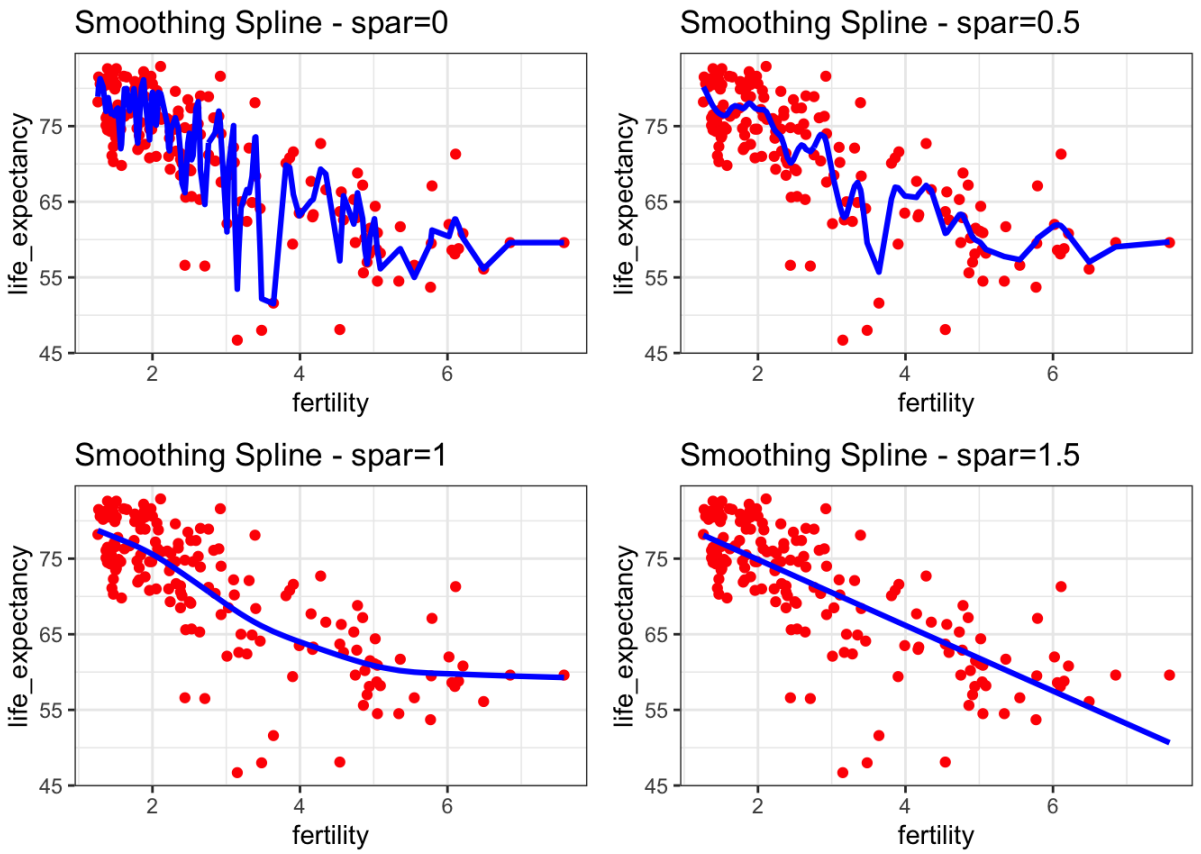
Another approach is to **extend/expand the assumptions** and to work out the corresponding mathematical formalism:

- **generalized linear models (GLM)** implement responses with **non-normal** conditional distributions (see chapter 7);
- **classifiers**, such as logistic regression, decision trees, support vector machines, naïve Bayes methods, neural networks, etc., extend regression to **categorical responses** (not in this course's scope, save for LogReg);

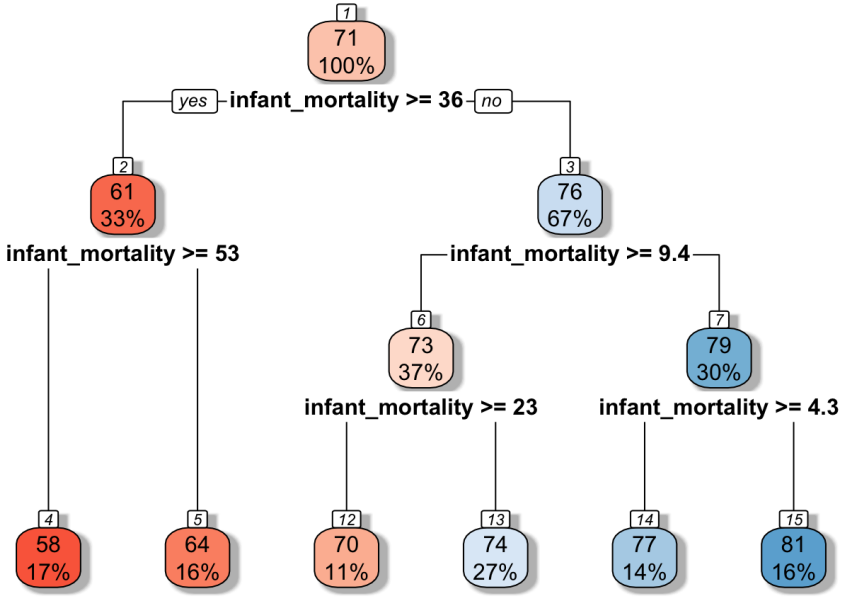
- **non-linear methods**, such as splines, generalized additive models (GAM), nearest neighbour methods, kernel smoothing methods, etc., are used for responses that are **not linear combinations of the predictors**;

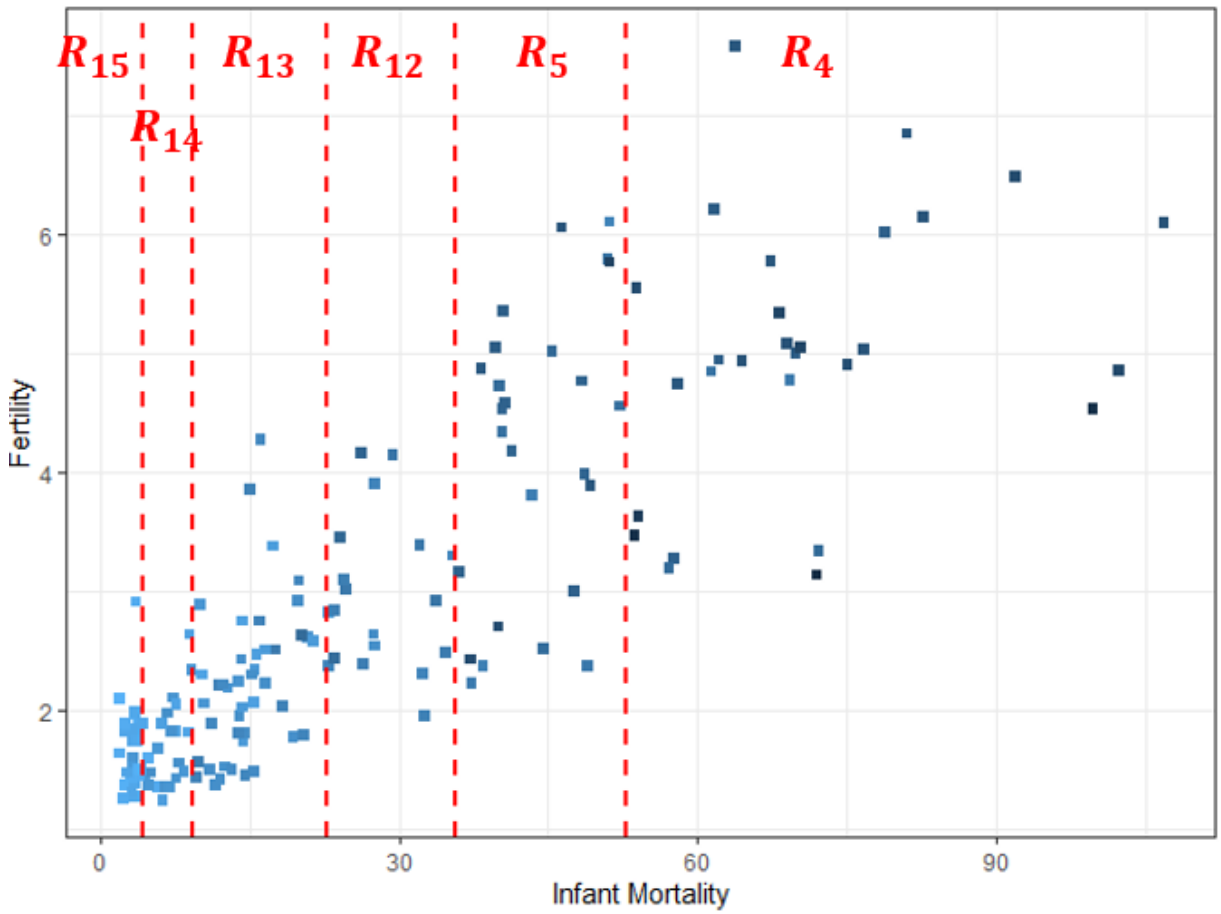



Cubic Spline - K=0
Cubic Spline - K=1
Cubic Spline - K=2
Cubic Spline - K=10
life_expectancy
fertility


Smoothing Spline - spar=0
Smoothing Spline - spar=0.5
Smoothing Spline - spar=1
Smoothing Spline - spar=1.5
life_expectancy
fertility
45
55
65
75
2
4
6

- **tree-based methods** and **ensemble learning methods**, such as bagging, random forests, and boosting, are used to simplify the modeling of **predictor interactions**;




$R_{15}$
$R_{14}$
$R_{13}$
$R_{12}$
$R_5$
$R_4$
Fertility
6
4
2
0
30
60
90
Infant Mortality

- **regularization methods**, such as ridge regression, the LASSO, and elastic nets, facilitate the process of **model selection** and **feature selection**.

On the last topic, assume that the training set consists of $n$ **centered**, **scaled** observations $\mathbf{x}_i$, together with target observations $y_i$.

Let $b_{\text{LS},j}$ be the $j$th LS coefficient, and set a **threshold** $\lambda > 0$, whose value is dataset-dependent.

We have seen that $\mathbf{b}_{\text{LS}}$ is the exact solution to the LS problem

$$\mathbf{b}_{\text{LS}} = \arg\min_{\boldsymbol{\beta}}\{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2\} = \arg\min_{\boldsymbol{\beta}}\{\text{SSE}\}.$$

In general, **no restrictions** are assumed on the values of the coefficients $b_{\mathrm{LS},j}$; large magnitudes imply that corresponding features **play an important role** in predicting the response.

**Ridge regression** (RR) is a method to **regularize** the LS coefficients.

Effectively, it shrinks the LS coefficients by **penalizing** solutions with large magnitudes – if the magnitude of a specific coefficient is **large**, then it must have **great** relevance in predicting the response.

This leads to a modified LS problem:

$$\mathbf{b}_{\mathrm{RR}} = \arg\min_{\boldsymbol{\beta}} \{ \underbrace{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2}_{\mathrm{SSE}} + \underbrace{\lambda n \|\boldsymbol{\beta}\|_2^2}_{\mathrm{penalty}} \}.$$

This quantity is small when $\mathrm{SSE}$ is **small** (i.e., the model is a good fit to the data) and when the **shrinkage penalty** is small (i.e., when each $\beta_j$ is small); RR solutions are typically obtained $via$ numerical methods.

The hyperparameter $\lambda$ controls the **relative impact** of both components. There are other variants, such as **best subset regression** (BS) and the **LASSO**, which both tend to yield $\beta_j = 0$ for some $j$:

$$\mathbf{b}_{\mathrm{BS}} = \arg\min_{\boldsymbol{\beta}} \{ \underbrace{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2}_{\mathrm{SSE}} + \underbrace{\lambda n \|\boldsymbol{\beta}\|_0}_{\mathrm{penalty}} \}, \quad \|\boldsymbol{\beta}\|_0 = \sum_{j=1}^{p} \mathrm{sgn}(|\beta_j|)$$

$$\mathbf{b}_{\mathrm{L}} = \arg\min_{\boldsymbol{\beta}} \{ \underbrace{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2}_{\mathrm{SSE}} + \underbrace{\lambda n \|\boldsymbol{\beta}\|_1}_{\mathrm{penalty}} \}, \quad \|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|.$$