STAT22200 Spring 2016 Chapter 09

Yibi Huang

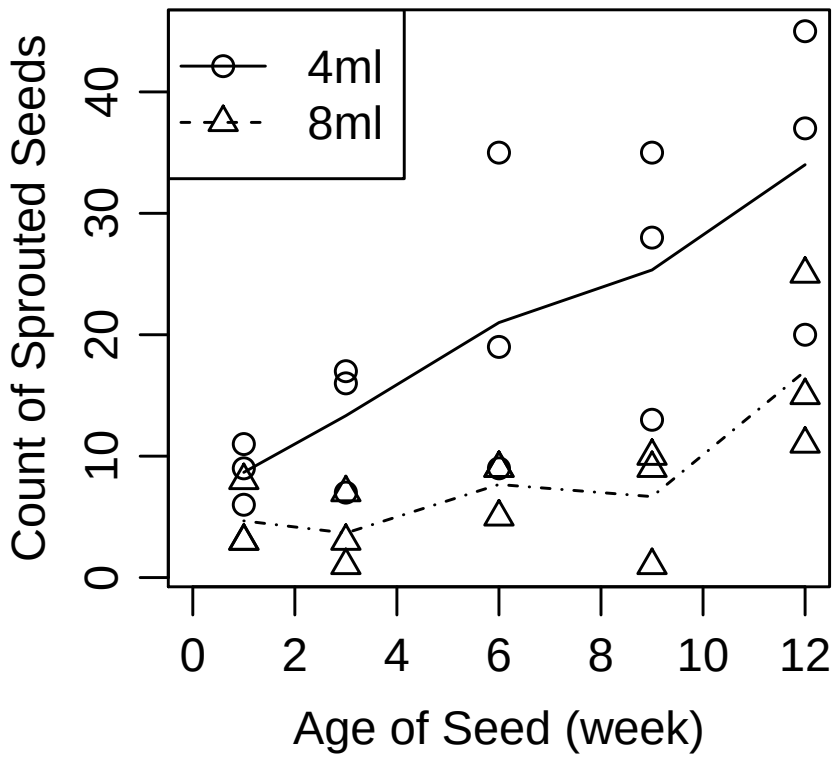
May 17, 2019

9.2.3 Quantitative Factors

9.2.3 Quantitative Factors

Sprouting Barley Example Revisit (p.166, Oehlert)
Recall the sprouting barley study on the conditions barley germinate. The response is the number of seeds germinating in 100 seeds.

water	Age of Seeds (weeks) 1	3	6	9	12
4(ml)	11	7	9	13	20
	9	16	19	35	37
	6	17	35	28	45
8(ml)	8	1	5	1	11
	3	7	9	10	15
	3	3	9	9	25

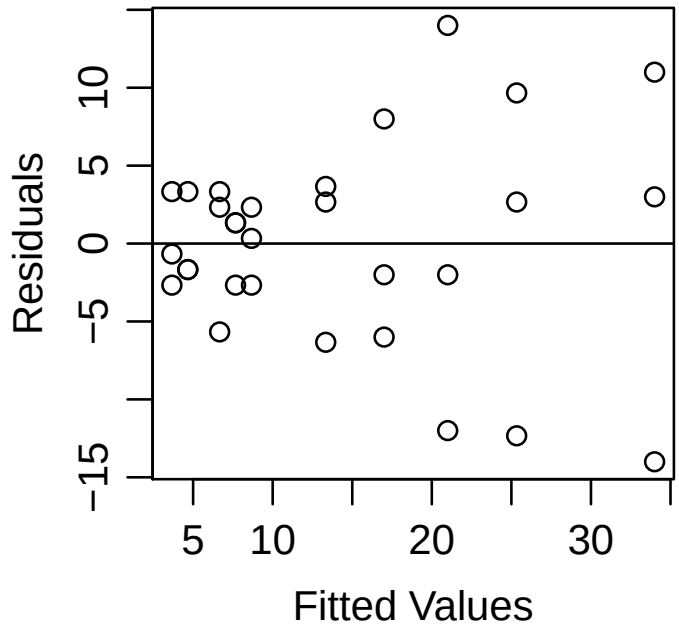


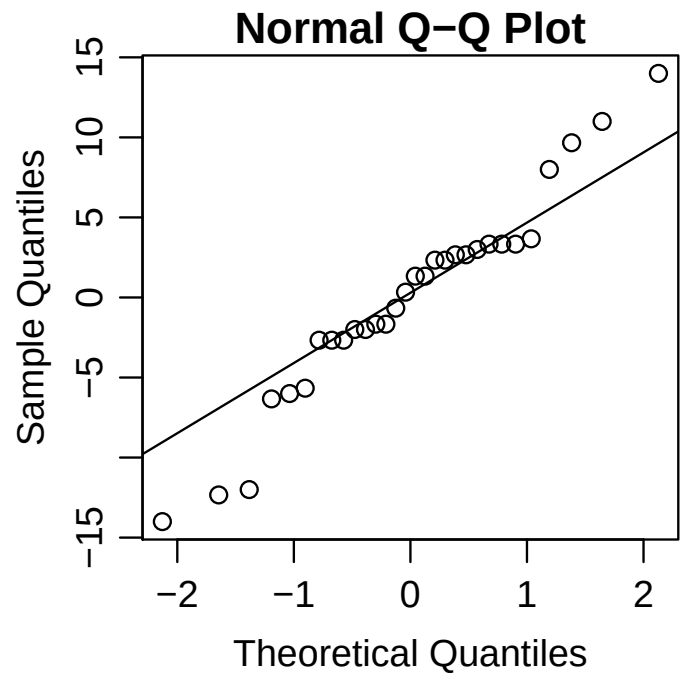
Sprouting Barley Example Revisit (2)

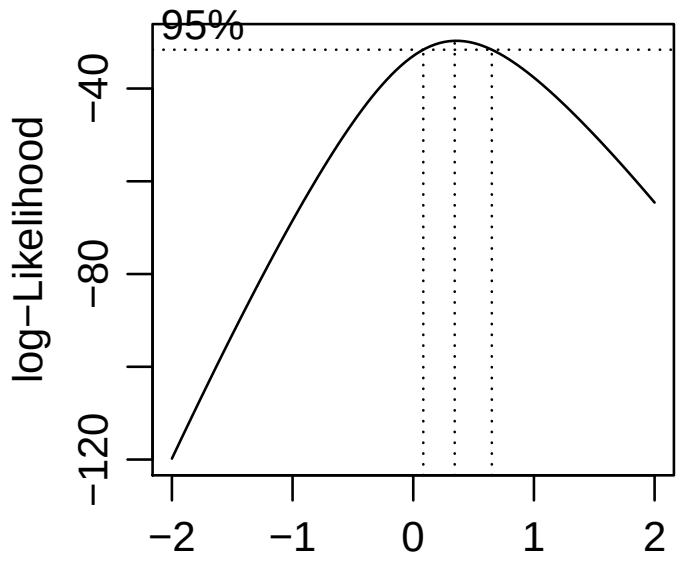
```
> barley = read.table("SproutingBarley.txt",header=T)
> lmfull = lm(y ~ as.factor(week)*as.factor(water),data=barley)
> anova(lmfull)

Response: y
                                 Df  Sum Sq Mean Sq F value   Pr(>F)
as.factor(week)                   4 1321.13  330.28  5.5293 0.003645 **
as.factor(water)                  1 1178.13 1178.13 19.7232 0.000251 ***
as.factor(week):as.factor(water)  4  208.87   52.22  0.8742 0.496726
Residuals                        20 1194.67   59.73
```

Before making conclusions, let's check model assumptions.

Sprouting Barley Example Revisit (3)



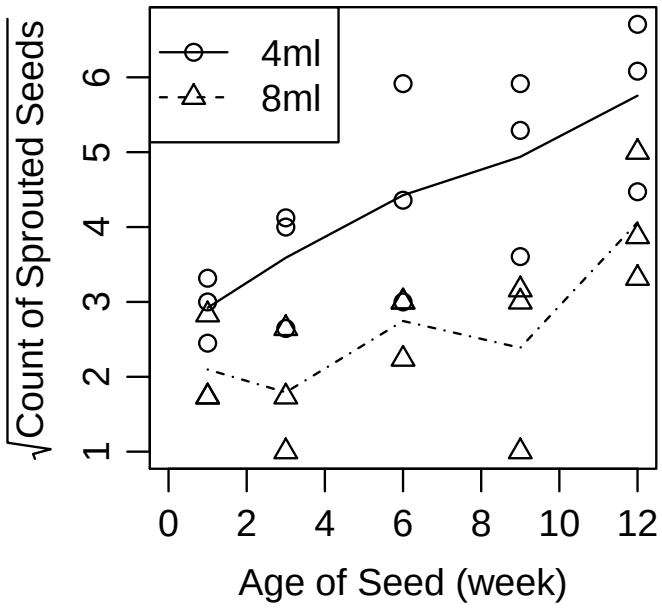


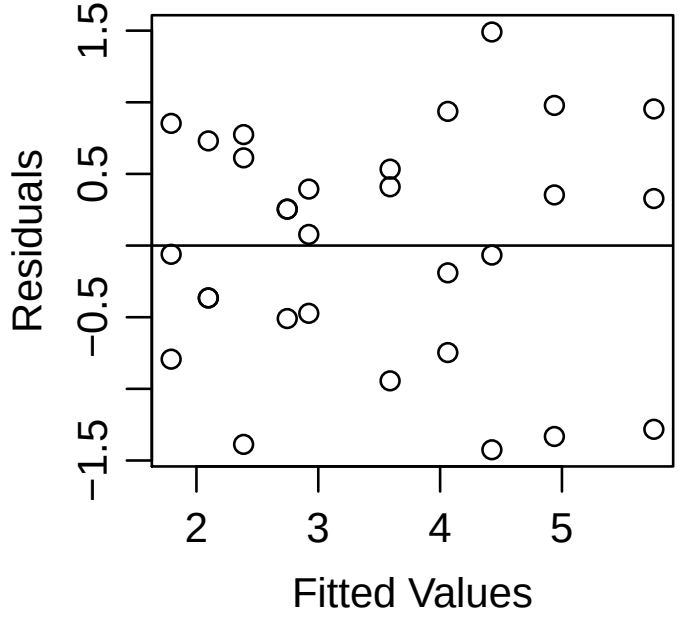


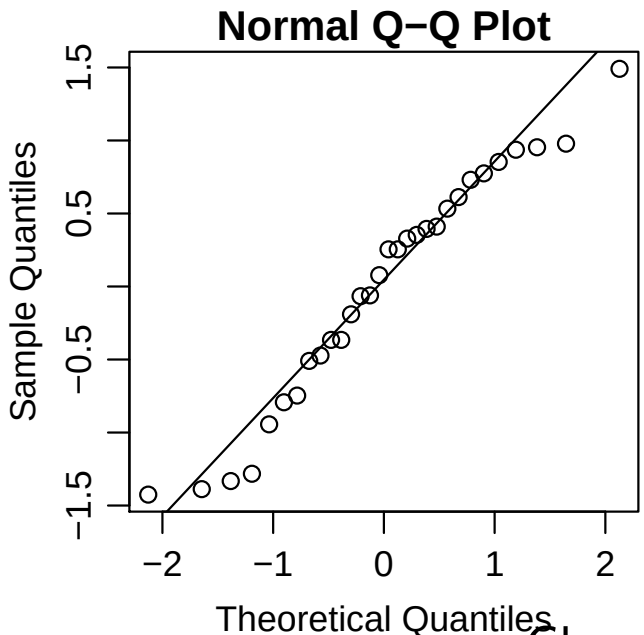
Spot any problem?

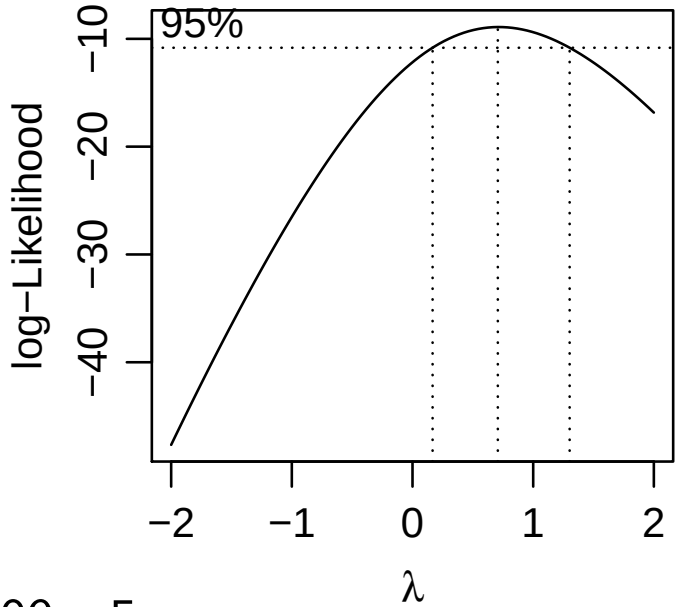
Remedy?

Sprouting Barley Example Revisit (4)









Sprouting Barley Example Revisit (4)

```
> lmfull2 = lm(sqrt(y) ~ as.factor(week)*as.factor(water), data=barley)
> anova(lmfull2)
Response: sqrt(y)
                                 Df  Sum Sq Mean Sq F value    Pr(>F)
as.factor(week)                   4 21.8949  5.4737  5.9406  0.002555 **
as.factor(water)                  1 21.8930 21.8930 23.7606 9.177e-05 ***
as.factor(week):as.factor(water)  4  2.2485  0.5621  0.6101  0.660139
Residuals                        20 18.4280  0.9214
```

- Now what is your conclusion?
- The main effect "seed age" being significant just means that seeds of different ages (1, 3, 6, 9, and 12 weeks) have *different* germination rates. It doesn't even tell us whether sprouting rate *increases* with the age of seeds.
- ANOVA models treat all factors as *categorical.*

$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk}$$

- Are there models that take quantitative levels of factors into account?

9.2.3 Quantitative Factors

Recall in Section 3.10, when treatments in a CRD are quantitative, we can fit a polynomial model

$$y_{ij} = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k + \varepsilon_{ij}.$$

In a 2-way $a \times b$ factorial design, of which factor A is quantitative with a numerical levels $x_1, \ldots, x_a$, and factor B is categorial, we may consider a polynomial model

$$y_{ijk} = \theta_0 + \theta_1 x_i + \theta_2 x_i^2 + \cdots + \theta_m x_i^m + \beta_j + \varepsilon_{ijk}.$$

If

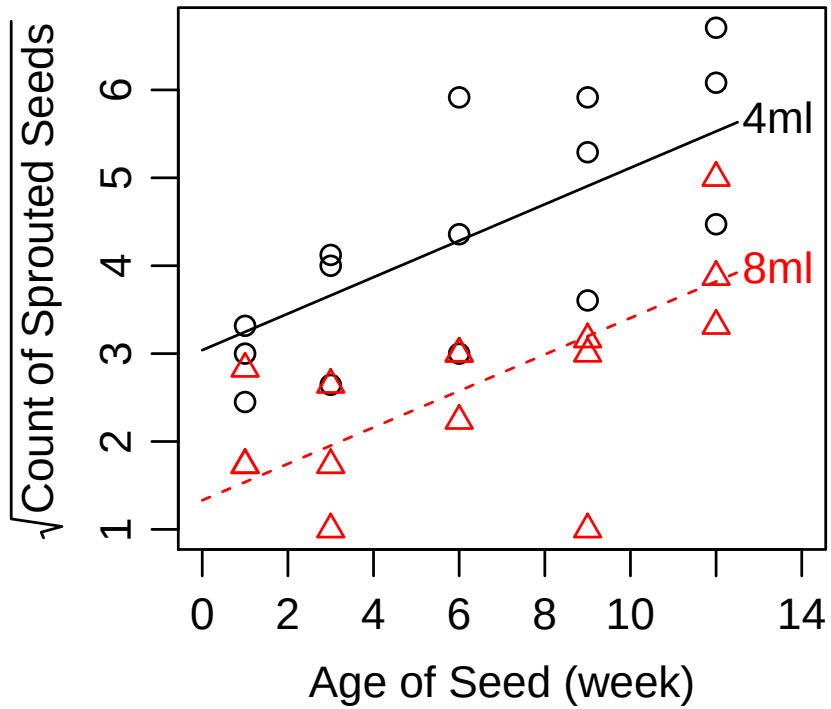
$$\mu + \alpha_i = \theta_0 + \theta_1 x_i + \theta_2 x_i^2 + \cdots + \theta_m x_i^m \quad \text{for all } i = 1, \ldots, a,$$

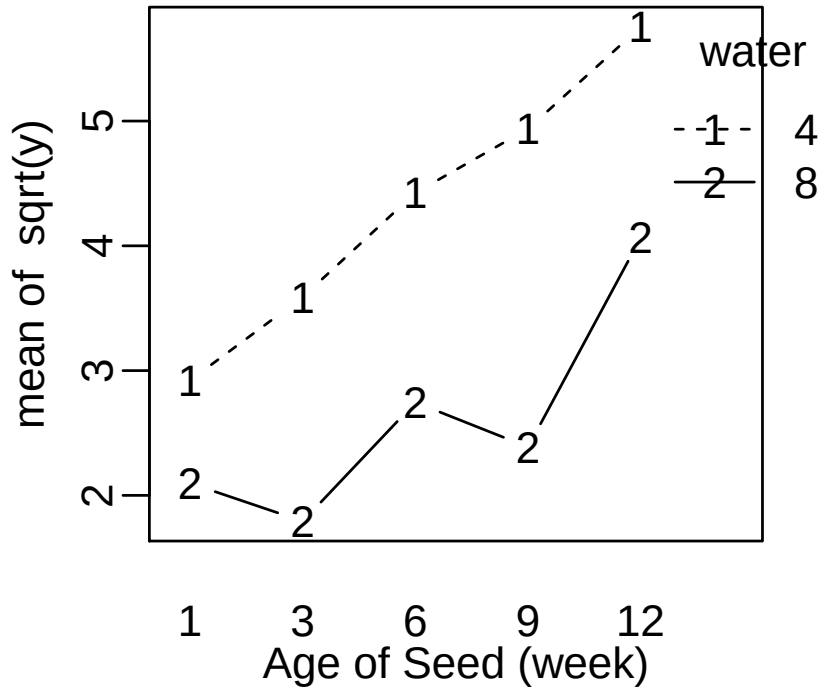
then the polynomial model is equivalent to the additive model

$$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}.$$

- The order of the polynomial m cannot exceed $a - 1$. (Why?)
- As long as $m \leq a - 1$, the polynomial model is *nested* in the additive model $y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$. (Why?)

A Polynomial Model for the Sprouting Barley Data (1)





As both `week` is quantitative, the ANOVA table shows no significant interaction, and the square-root transformed response is roughly linearly with `week` from the plot, we thus fit the polynomial model

$$\sqrt{y_{ijk}} = \theta_0 + \theta_1 \text{week}_i + \beta_j + \varepsilon_{ijk}.$$

Compare the ANOVA tables of the 3 models.

Full Model: $\sqrt{y_{ijk}} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk}$

```
                                Df  Sum Sq Mean Sq F value    Pr(>F)
as.factor(week)                  4 21.8949  5.4737  5.9406  0.002555 **
as.factor(water)                 1 21.8930 21.8930 23.7606 9.177e-05 ***
as.factor(week):as.factor(water)  4  2.2485  0.5621  0.6101  0.660139
Residuals                        20 18.4280  0.9214
```

Additive Model: $\sqrt{y_{ijk}} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$

```
                Df Sum Sq Mean Sq F value    Pr(>F)
as.factor(week)  4 21.895  5.4737  6.3536  0.001236 **
as.factor(water) 1 21.893 21.8930 25.4121 3.746e-05 ***
Residuals       24 20.677  0.8615
```

Polynomial Model: $\sqrt{y_{ijk}} = \theta_0 + \theta_1 \text{week}_i + \beta_j + \varepsilon_{ijk}$

```
                Df Sum Sq Mean Sq F value    Pr(>F)
week             1 20.346 20.3464  24.718 3.286e-05 ***
as.factor(water) 1 21.893 21.8930  26.597 1.997e-05 ***
Residuals       27 22.225  0.8231
```

Compare the SSE and MSE of the 3 models. Which one has the minimal SSE? Which one has the minimal MSE?

Should We Consider a Higher Order Polynomial Model?

Since the polynomial model is nested in the full effects model, we can perform a goodness-of-fit test

$$F = \frac{(SSE_{reduced} - SSE_{full})/(df_{reduced} - df_{full})}{SSE_{full}/df_{full}}$$

```
> lmadd1 = lm(sqrt(y) ~ week + as.factor(water), data=barley)
> lmfull2 = lm(sqrt(y) ~ as.factor(week)*as.factor(water), data=barley)
> anova(lmadd1,lmfull2)
Analysis of Variance Table

Model 1: sqrt(y) ~ week + as.factor(water)
Model 2: sqrt(y) ~ as.factor(week) * as.factor(water)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     27 22.225
2     20 18.428  7    3.7969 0.5887 0.7574
```

The large P-value shows that the linear model fits the data as good as the more general effects model, so the linear model is adequate. There is no need to consider higher order polynomials.

Final Model for the Sprouting Barley Data

From the analysis above, a simplest model with a adequate fit is

$$\sqrt{y_{ijk}} = \theta_0 + \theta_1 \text{week}_i + \beta_j + \varepsilon_{ijk}.$$

```
> lmadd1 = lm(sqrt(y) ~ week + as.factor(water), data=barley)
> lmadd1$coef
      (Intercept)                week  as.factor(water)8
        3.0395813           0.2074459          -1.7085268
```

The estimated coefficients are

$$\widehat{\theta}_0 \approx 3.0396, \quad \widehat{\theta}_1 \approx 0.2074, \quad \widehat{\beta}_{8ml} \approx -1.7085$$

and $\widehat{\beta}_{4ml} = 0$ because R use the constraint $\beta_{4ml} = 0$.
The fitted model is

$$\sqrt{\text{predicted count of sprouted barley in 100 barley seeds}}$$

$$= \begin{cases} 3.0396 + 0.2074(\text{age of seeds in weeks}) & \text{if water} = 4\text{ml} \\ 3.0396 - 1.7085 + 0.2074(\text{age of seeds in weeks}) & \text{if water} = 8\text{ml} \end{cases}$$

Final Model for the Sprouting Barley Data

Note that the response y is the count of sprouted barley in 100 seeds, $y/100$ is the germanation rate of barley seeds. The fitted model in terms of germination rate is

$$\widehat{\sqrt{y/100}} = \sqrt{\text{predicted germination rate of barley seeds}}$$
$$= \begin{cases} 0.30396 + 0.02074\,\text{week} & \text{if water} = 4\text{ml} \\ 0.30396 - 0.17085 + 0.02074\,\text{week} & \text{if water} = 8\text{ml} \end{cases}$$

Interpretation: The square root of the predicted germination rate of barley seeds

- increases by 0.02074 for every additional week after harvest, regardless of whether it's 4 ml or 8 ml of water used in germination
- decreases by 0.17085 if the amount of water is increased from 4 ml to 8 ml, regardless of the age of seeds.

If Both Factors are Quantitative ...

In a 2-way $a \times b$ factorial design, if both factors are quantitative, say factor A has a levels at $x_1, \ldots, x_a$, and factor B has b levels at $z_1, \ldots, z_b$, we may consider a polynomial model

$$y_{ijk} = \theta_0 + \theta_1 x_i + \cdots + \theta_m x_i^m + \phi_1 z_j + \cdots + \phi_r z_j^r + \varepsilon_{ijk}.$$

- the order m of the polynomial for factor A must $\leq a - 1$,
- the order r of the polynomial for factor B must $\leq b - 1$
- this is an additive model
- As long as the orders of polynomial $m \leq a - 1$ and $r \leq b - 1$, the polynomial model is *nested* in the additive model $y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$. (Why?)

  (Of course, it is also nested in the main-effect-interaction model $y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk}$.)

How to Include Interaction in a Polynomial Model?

Observe the polynomial model is additive.

$$y_{ijk} = \theta_0 + \theta_1 x_i + \theta_2 x_i^2 + \cdots + \theta_m x_i^m + \beta_j + \varepsilon_{ijk}$$

Factor B only affect the intercept $\theta_0 + \beta_j$ of the polynomial, but not other coefficients.

To allow interaction, one should allow the coefficients of the polynomial vary with levels of B. The model would be like

$$y_{ijk} = \theta_{0j} + \theta_{1j} x_i + \theta_{2j} x_i^2 + \cdots + \theta_{mj} x_i^m + \varepsilon_{ijk}.$$