Polynomial Regression

Polynomial Regression

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1}^2 + \beta_4 X_{i2}^2 + \beta_5 X_{i1} X_{X2} + \epsilon_i$$

- With enough data, can approximate arbitrary nonlinear response functions
- Most useful when the mean response is non-linear but error variance is constant
- In many cases, transformation of Y may make more sense
  - Regression on transformed variables may use fewer degrees of freedom
  - Polynomial regression will **not** correct non-constant variance

- Polynomials and transformations can be used together.
- Polynomials of several predictors can be combined (*response surface methodology*)
- this may or may not involve interactions between variables

Cautions

- Polynomials generally create a multicollinearity problem, which can often be corrected by centering the data $(x_i = X_i - \bar{X})$, or standardization $(\mathrm{x_i} = (X_i - \bar{X})/s)$.

In R, use the scale() function

- Extrapolation beyond the scope of the data is a **very bad idea**

Example: Battery life

A researcher studied the effect of the charge rate and temperature on the life of a new type of power cell in a preliminary small-scale experiment.

X1:Charge rate (3 levels)

X2:Temperature (3 levels)

Y: Number of cycles

- The levels of charge and temperature are planned.
- We want to know:

1. whether a linear or quadratic function is appropriate
2. if there is a significant interaction between charge rate and temperature

Starting with the second-order polynomial regression model with interaction, the researcher aimed to better understand the response function (e.g., the MLR function) within the range of the factor being studied. Despite uncertainty about its nature, this was seen as a necessary step in the research process.

$$Y_i = old\beta_0 + old\beta_1 X_{i1} + old\beta_2 X_{i2} + old\beta_3 X_{i1}^2 + old\beta_4 X_{i2}^2 + old\beta_5 X_{i1} X_{X2} + \epsilon_i$$

Correlation between X_1 and X_1^2 is 0.991, between X_2 and X_2^2 is 0.986

To reduce multicollinearity, the researcher decided to center the variables. For the sake of demonstration, they also scaled the variables using convenient units.

$$\boldsymbol{x_{i1}} = \frac{\boldsymbol{X_{i1}-\bar{X}}}{\mathbf{0.4}} = \frac{\boldsymbol{X_{i1}-1}}{\mathbf{0.4}}, \quad \boldsymbol{x_{i2}} = \frac{\boldsymbol{X_{i2}-\bar{X}}}{\mathbf{10}} = \frac{\boldsymbol{X_{i2}-20}}{\mathbf{10}}, \quad \textbf{hence}$$

- The actual sd[x1}=0.31, and sd{x2}=7.75
- If use scale(x), there will be slightly different
- Choose 0.4 and 10 for convenient.

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1}^2 + \beta_4 x_{i2}^2 + \beta_5 x_{i1} x_{i2} + \epsilon_i$$

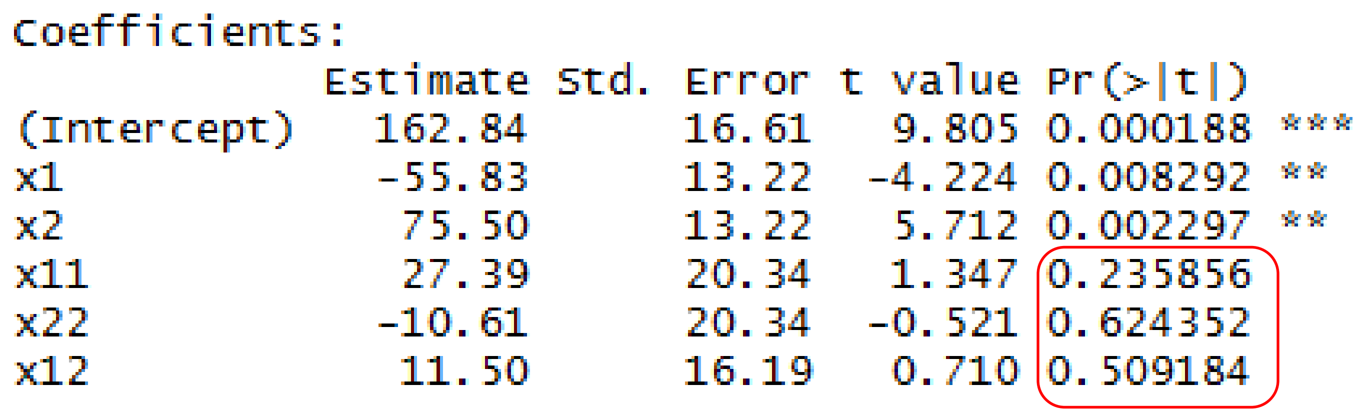
Correlation between x_1 and x_1^2 is now <0.001, between x_2 and x_2^2 is now <0.001

Notes:

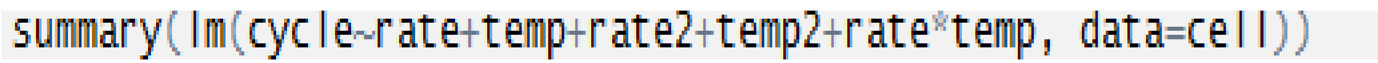
1. $old\beta \neq \beta$
2. For simplicity in coding, sometimes we exchange notations:

$$x_1^2 = x_{11}, \qquad x_2^2 = x_{22}, \qquad x_1 x_2 = x_{12}, \qquad \beta_3 = \beta_{11}, \beta_4 = \beta_{22}, \beta_5 = \beta_{12},$$

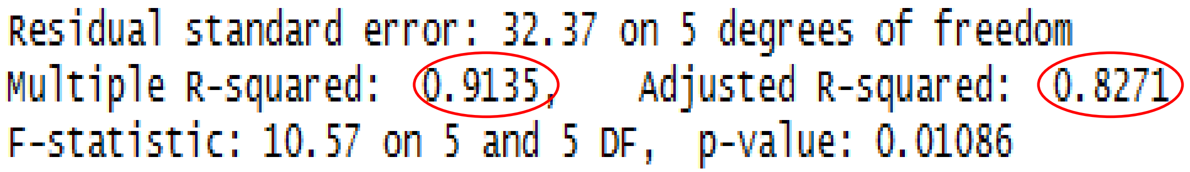
Regression Result Based on the **Scaled** X variables



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   162.84      16.61   9.805 0.000188 ***
x1            -55.83      13.22  -4.224 0.008292 ** 
x2             75.50      13.22   5.712 0.002297 ** 
x11            27.39      20.34   1.347 0.235856    
x22           -10.61      20.34  -0.521 0.624352    
x12            11.50      16.19   0.710 0.509184    
```



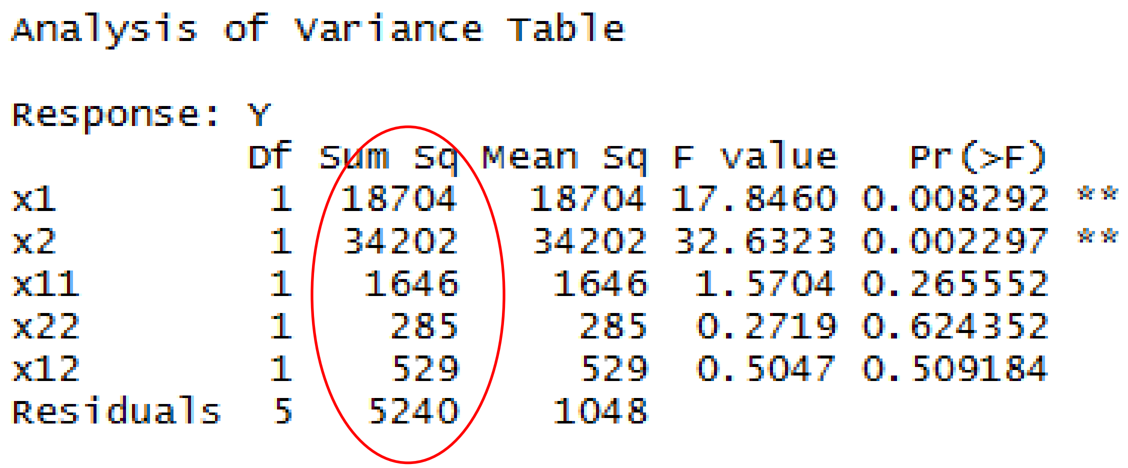
```
summary(lm(cycle~rate+temp+rate2+temp2+rate*temp, data=cell))
```



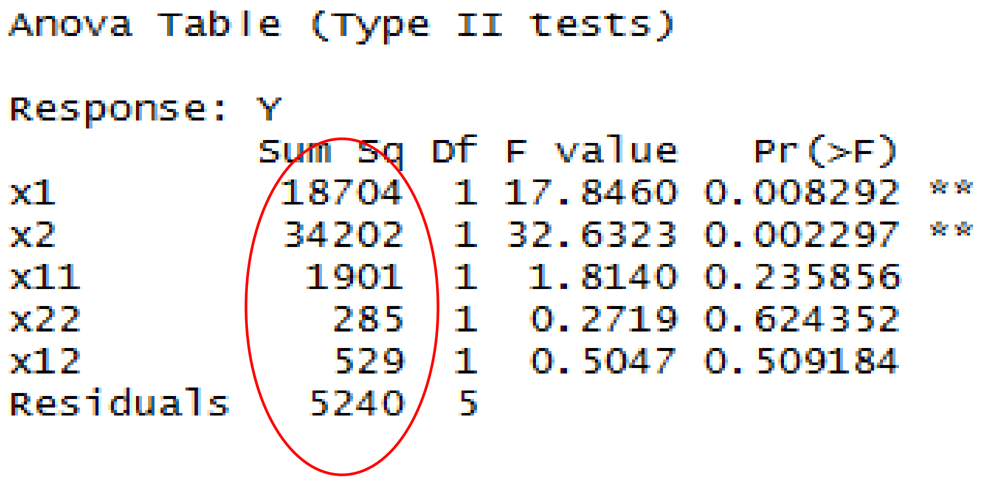
```
Residual standard error: 32.37 on 5 degrees of freedom
Multiple R-squared:  0.9135,	Adjusted R-squared:  0.8271 
F-statistic: 10.57 on 5 and 5 DF,  p-value: 0.01086
```

Type I SS

Type II SS



```
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value   Pr(>F)   
x1         1  18704   18704 17.8460 0.008292 **
x2         1  34202   34202 32.6323 0.002297 **
x11        1   1646    1646  1.5704 0.265552   
x22        1    285     285  0.2719 0.624352   
x12        1    529     529  0.5047 0.509184   
Residuals  5   5240    1048                    
```



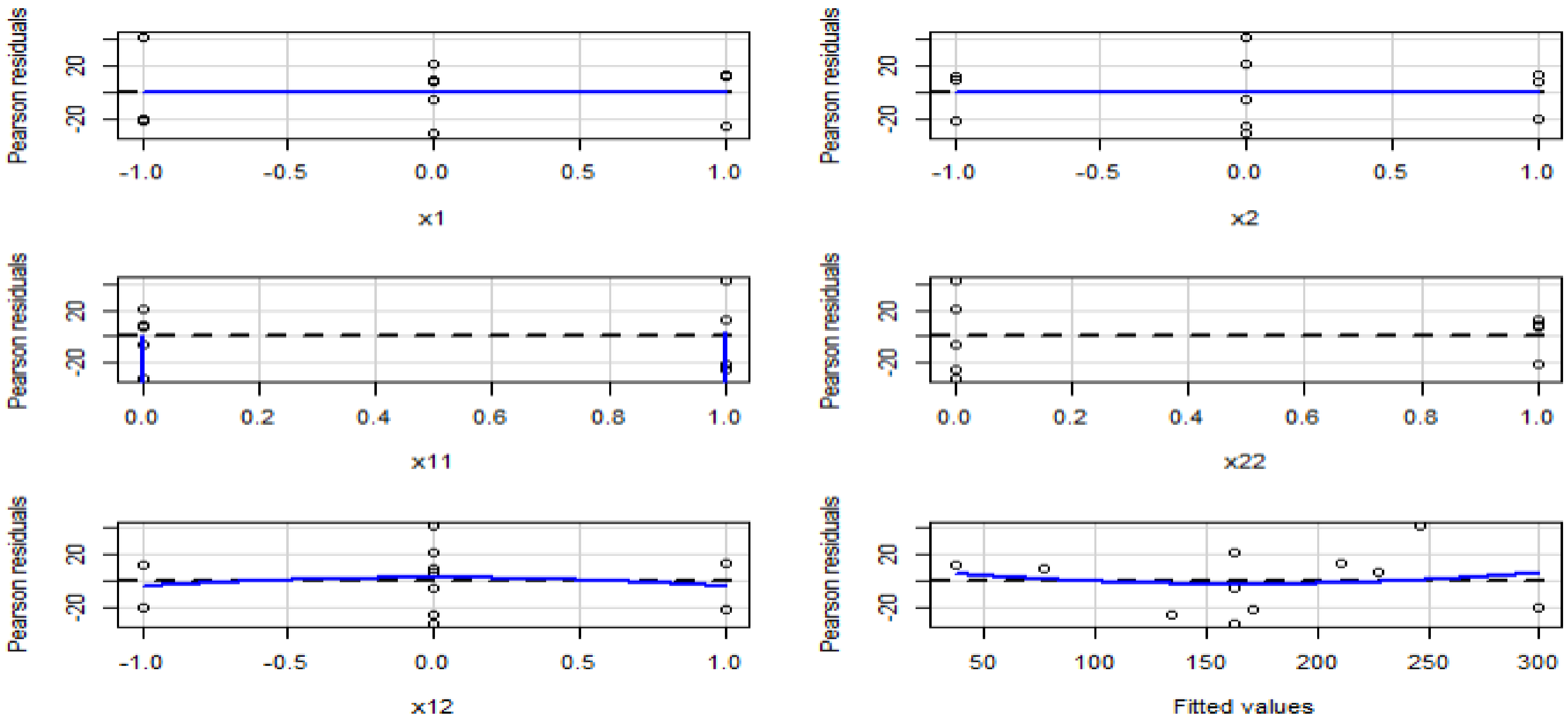
```
Anova Table (Type II tests)

Response: Y
          Sum Sq Df F value   Pr(>F)   
x1         18704  1 17.8460 0.008292 **
x2         34202  1 32.6323 0.002297 **
x11         1901  1  1.8140 0.235856   
x22          285  1  0.2719 0.624352   
x12          529  1  0.5047 0.509184   
Residuals   5240  5                    
```

The Type I and II ANOVA tables may be similar due to a potential lack of multicollinearity issues among the variables.

Fitting of Model

- $\hat{Y} = 162.84 - 55.83x_1 + 75.50x_2 + 27.39x_1^2 - 10.61x_2^2 + 11.5x_1x_2$

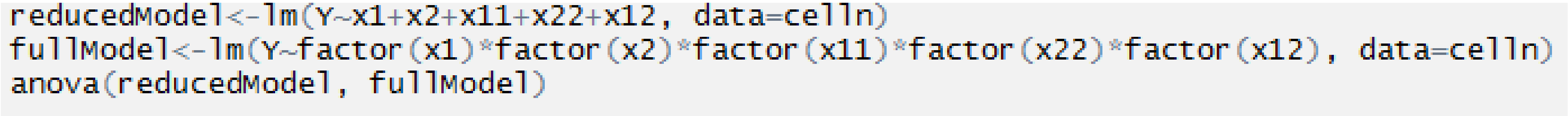
Residual Plots



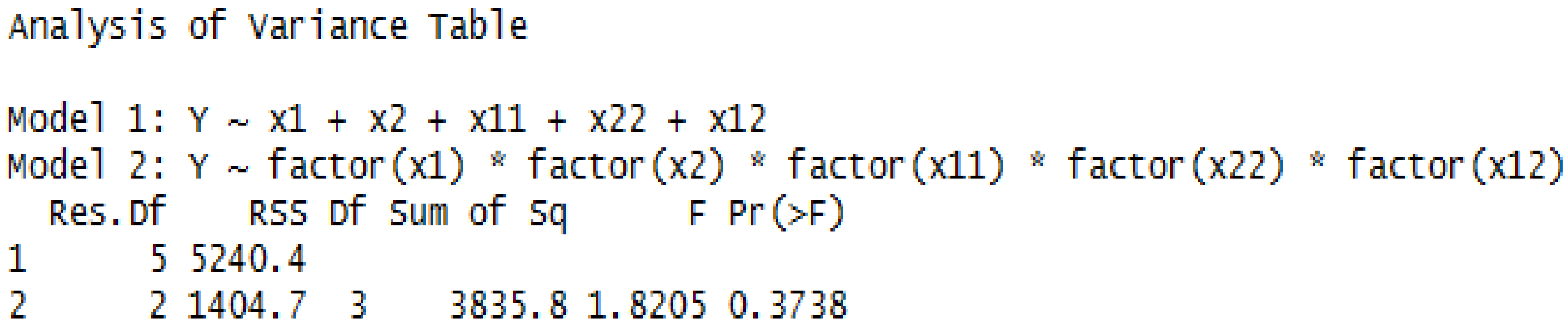
Lack of Fit test

- Ho: $Y_{ij} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 \ + \epsilon_{ij}$ (Reduced model)

Ha: $Y_{ij} = \mu_j + \epsilon_{ij}$ (Full model)



```
reducedModel<-lm(Y~x1+x2+x11+x22+x12, data=celln)
fullModel<-lm(Y~factor(x1)*factor(x2)*factor(x11)*factor(x22)*factor(x12), data=celln)
anova(reducedModel, fullModel)
```



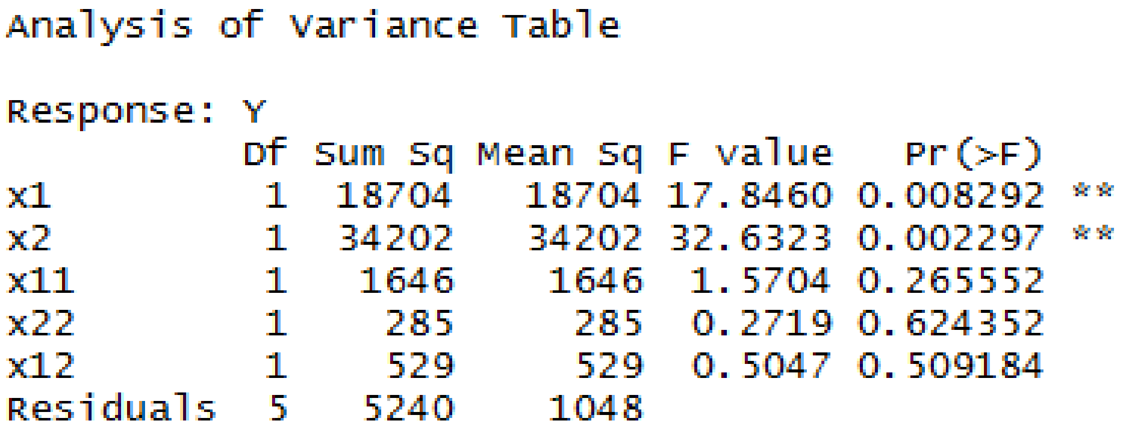
```
Analysis of Variance Table

Model 1: Y ~ x1 + x2 + x11 + x22 + x12
Model 2: Y ~ factor(x1) * factor(x2) * factor(x11) * factor(x22) * factor(x12)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1      5 5240.4
2      2 1404.7  3    3835.8 1.8205 0.3738
```

$$F_s = \frac{SSLF}{c-p} \Big/ \frac{SSPE}{n-c} = \frac{3835.77}{3} \Big/ \frac{1404.7}{2} = 1.82 \quad \sim F(3,2)$$

Do not reject the Ho, conclude that there isn't a lack of fit issue.

The GLT (Partial F test): Now consider whether a first-order model would be sufficient

- $Ho: \beta_3 = \beta_4 = \beta_5 = 0$; *(Reduced model)*
- Ha: *not all* βs *in Ho equal zero (Full model)*



```
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value   Pr(>F)   
x1         1  18704   18704 17.8460 0.008292 **
x2         1  34202   34202 32.6323 0.002297 **
x11        1   1646    1646  1.5704 0.265552   
x22        1    285     285  0.2719 0.624352   
x12        1    529     529  0.5047 0.509184   
Residuals  5   5240    1048                    
```

$$Fs = \frac{\frac{SSR(x_1^2, x_2^2, x_1x_2 \mid x_1, x_2)}{3}}{\frac{SSE(x_1, x_2, x_1^2, x_2^2, x_1x_2)}{dfE}} = \frac{\frac{1646+285+519}{3}}{1048} = 0.78$$

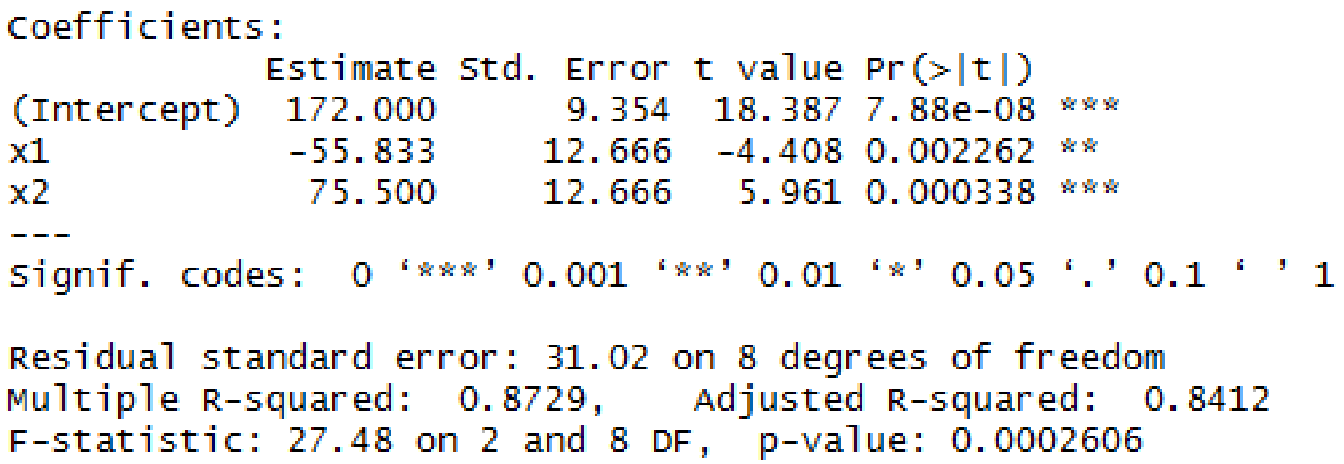
At $\alpha = 0.05$, *the critical value* $F(0.95; 3,5) = 5.41$, *since* $Fs < 5.44$, *we do not reject the Ho*.

We conclude that the first-order model is adequate for the range of the charge rates and temperatures considered.

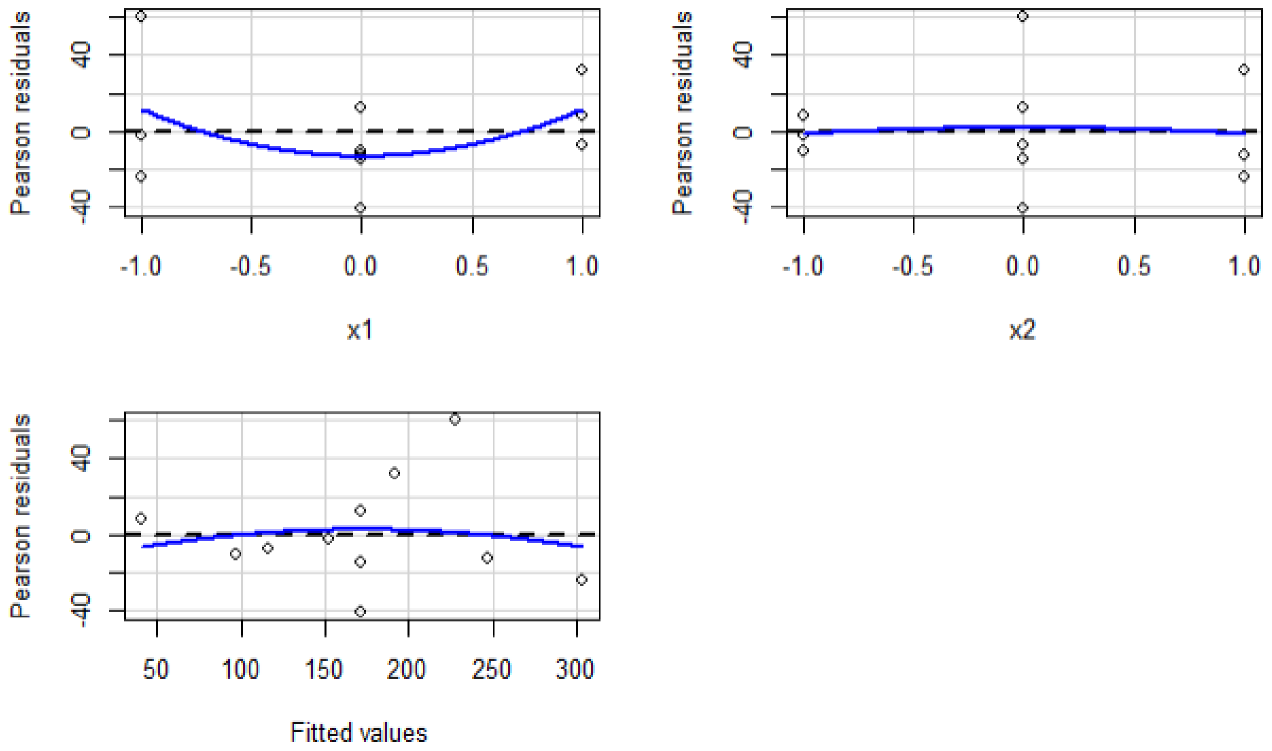
Fit the first-order Model

$$Y_{ij} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon_{ij}$$

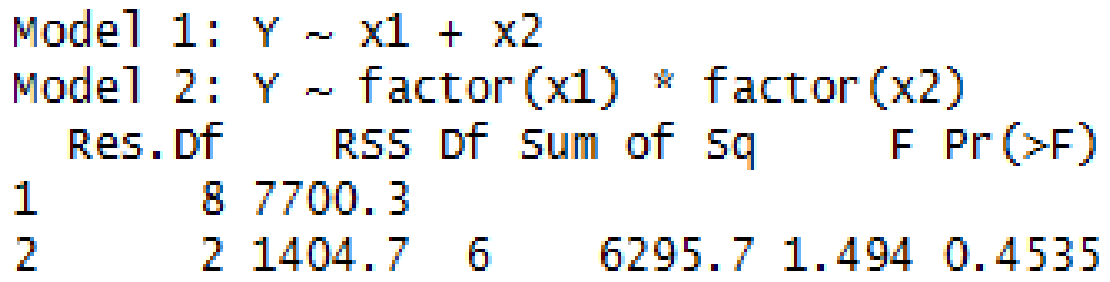
$$\hat{Y} = 172 - 55.83x_1 + 75.5x_2$$



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  172.000      9.354  18.387 7.88e-08 ***
x1           -55.833     12.666  -4.408 0.002262 ** 
x2            75.500     12.666   5.961 0.000338 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31.02 on 8 degrees of freedom
Multiple R-squared:  0.8729,	Adjusted R-squared:  0.8412 
F-statistic: 27.48 on 2 and 8 DF,  p-value: 0.0002606
```



The Lack of Fit test



```
Model 1: Y ~ x1 + x2
Model 2: Y ~ factor(x1) * factor(x2)
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1      8 7700.3                          
2      2 1404.7  6    6295.7 1.494 0.4535
```

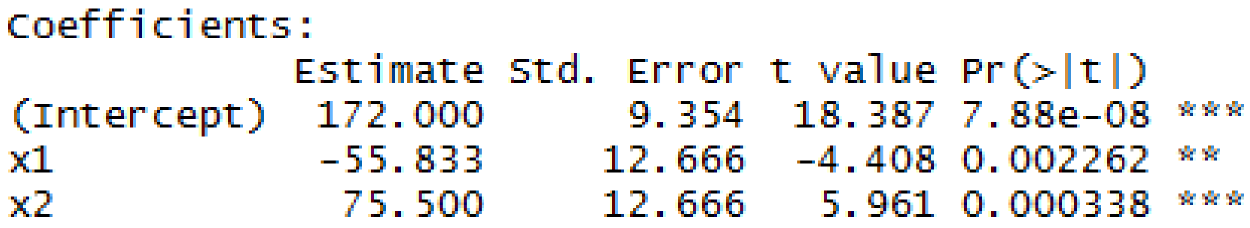
$$F_s = \frac{SSLF}{c-p} \Big/ \frac{SSPE}{n-c} = \frac{6295.7}{6} \Big/ \frac{1404.7}{2} = 1.49$$

Do not reject the Ho, conclude that there isn’t a lack of fit.

Simultaneously estimation the regression coefficients (the linear impacts, the betas, etc.). Use the Bonferroni method with 90% confidence level.

$$B = t\left(1-\frac{\alpha}{2g}, df\right) = t(0.975; 8)\text{=2.306}$$

Where $g = 2$, and $df = n - p = 11 - 3 = 8$



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  172.000      9.354  18.387 7.88e-08 ***
x1           -55.833     12.666  -4.408 0.002262 ** 
x2            75.500     12.666   5.961 0.000338 ***
```

The confidence interval, $b\_k \pm Bs\{b\_k \}$, represents the estimate of the impact of the transformed X on Y. However, our objective is to determine the impact of the original X on Y.

The original variable X1 and X2 are transformed to x_1 and x_2 through the following transformation function.

$$x_{i1} = \frac{X_{i1} - \bar{X}}{0.4} = \frac{X_{i1} - 1}{0.4}, \qquad x_{i2} = \frac{X_{i2} - \bar{X}}{10} = \frac{X_{i2} - 20}{10}$$

We need to back-transform the β to the $old\beta$.

The back transformation process on the betas

$$Y_{ij} = old\beta_0 + old\beta_1 X_{i1} + old\beta_2 X_{i2} + \epsilon_{ij}$$

$$Y_{ij} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon_{ij} \quad \boldsymbol{where}\ \boldsymbol{x_{i1} = \frac{X_{i1}-\bar{X}}{0.4} = \frac{X_{i1}-1}{0.4},} \qquad \boldsymbol{x_{i2} = \frac{X_{i2}-\bar{X}}{10} = \frac{X_{i2}-20}{10}}$$

$$= \beta_0 + \beta_1 \frac{X_{i1}-1}{0.4} + \beta_2 \frac{X_{i2}-20}{10} + \epsilon_{ij} = (\beta_0 - \frac{\beta_1}{0.4} - 2\beta_2) + \frac{\beta_1}{0.4} X_{i1} + \frac{\beta_2}{10} X_{i2}$$

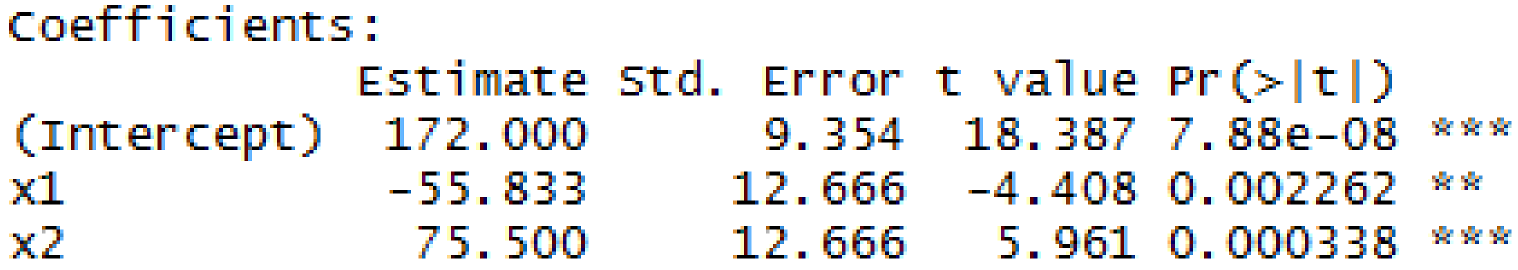
Hence

$$\boxed{\begin{array}{ll} old\beta_1 = \dfrac{\beta_1}{0.4}, & \sigma(old\beta_1) = \dfrac{\sigma(\beta_1)}{0.4} \\ old\beta_2 = \dfrac{\beta_2}{10}, & \sigma(old\beta_2) = \dfrac{\sigma(\beta_2)}{10} \end{array}} \Rightarrow \boxed{\begin{array}{ll} oldb_1 = \dfrac{b_1}{0.4}, & s\{oldb_1\} = \dfrac{s\{b_1\}}{0.4} \\ oldb_2 = \dfrac{b_2}{10}, & s\{oldb_2\} = \dfrac{s\{b_2\}}{10} \end{array}}$$

Estimate the Regression coefficients:

$$oldb_1 = \frac{b_1}{0.4}\,, \qquad s\{oldb_1\} = \frac{s\{b_1\}}{0.4}$$

$$oldb_2 = \frac{b_2}{10}\,, \qquad s\{oldb_2\} = \frac{s\{b_2\}}{10}$$



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  172.000      9.354  18.387 7.88e-08 ***
x1           -55.833     12.666  -4.408 0.002262 ** 
x2            75.500     12.666   5.961 0.000338 ***
```

The Bonferroni Confidence interval for $old\beta$:

$$B = t\left(1 - \frac{\alpha}{2g}, df\right) = t(0.975; 8)\text{=2.306}$$

$$\frac{b_1}{0.4} \pm \frac{Bs\{b_1\}}{0.4} = -\frac{55.833}{0.4} \pm 2.306\left(\frac{12.666}{0.4}\right) = (-212.6, -66.5)$$

$$\frac{b_2}{10} \pm \frac{Bs\{b_2\}}{10} = \frac{75.5}{0.4} \pm 2.306\left(\frac{12.666}{10}\right) = (4.6, \quad 10.5)$$

Based on the analysis, we can conclude that an increase in charge rate by 1-unit results in a decrease in battery life ranging from at least 66.5 cycles to at most 212.6 cycles. Additionally, an increase in temperature by 1-unit results in an increase in battery life ranging from at least 4.6 cycles to at most 10.5 cycles.

Summary on the battery case

- Linear terms are significant
- Quadratic terms are not significant
- Interaction is not significant
- General linear test shows that quadratic and interaction terms can be omitted
- Type I and Type II SS are almost identical

Caution: the coefficient of estimate of high order item is 0 doesn't necessarily establish that a linear response function is appropriate. Examination of residuals would disclose this lack of fit and should always accompany formal testing of polynomial regression coefficients.