STAT 139: STATISTICAL SLEUTHING THROUGH LINEAR MODELS

Lecture 22
Nov 18, 2014

Victoria Liublinska

Odds and Ends

- HW 10, Problem 3, part (a): the *p*-values on Display 10.19 have to corrected:

DISPLAY 10.19 Regression output data for Exercise 11

Variable	Estimate	SE	*t*-stat	*p*-value	*p*-value
Constant	3.775	0.3881	9.7321	~~<0.0000~~	**<0.0001**
lsize	0.0809	0.1131	0.7139	~~0.2443~~	**0.4879**
days	0.0774	0.1447	0.5346	~~0.5104~~	**0.6020**

Estimated SD about the regression is 0.8234 on 13 degrees of freedom; $R^2 = 11.41\%$.

Today's overview

- Inferential tools for multiple regression:
  - *t*-tests and CIs for coefficients and their linear combinations;
  - Confidence and prediction intervals for mean response;
  - Extra-sum-of-squares *F*-tests to compare regression models.
- Adjusted R-squared
- Strategies for variable selection

Today's overview

Reading:

- **Required:** Ch. 10 (Ch. 10 R code), Ch. 12 (Ch. 12 R code)
- **Optional Reading:** Gelman and Hill, Chapters 3, 4.
- **Supplementary Theory**: A. Sen and M. Srivastava. "Regression Analysis: Theory, Methods, and Applications", Ch. 3
- Another good reference available online:

Regression Modeling Strategies : With Applications to Linear Models, Logistic Regression, and Survival Analysis

Harrell, Frank, 2001.

Inferential Tools for Multiple Linear Regression

Multiple Regression: Inference

- Least squares estimators of coefficients;
- *t*-tests and confidence intervals for coefficients;
- *t*-tests and confidence intervals for linear combinations of coefficients;
- Confidence and prediction intervals for mean response;
- Extra-sum-of-squares *F*-tests to compare regression models.

Multiple Regression: Inference

- Least squares estimators of coefficients;
- *t*-tests and confidence intervals for coefficients;
- *t*-tests and confidence intervals for linear combinations of coefficients;
- Confidence and prediction intervals for mean response;
- Extra-sum-of-squares *F*-test to compare regression models.

Multiple Linear Regression Model

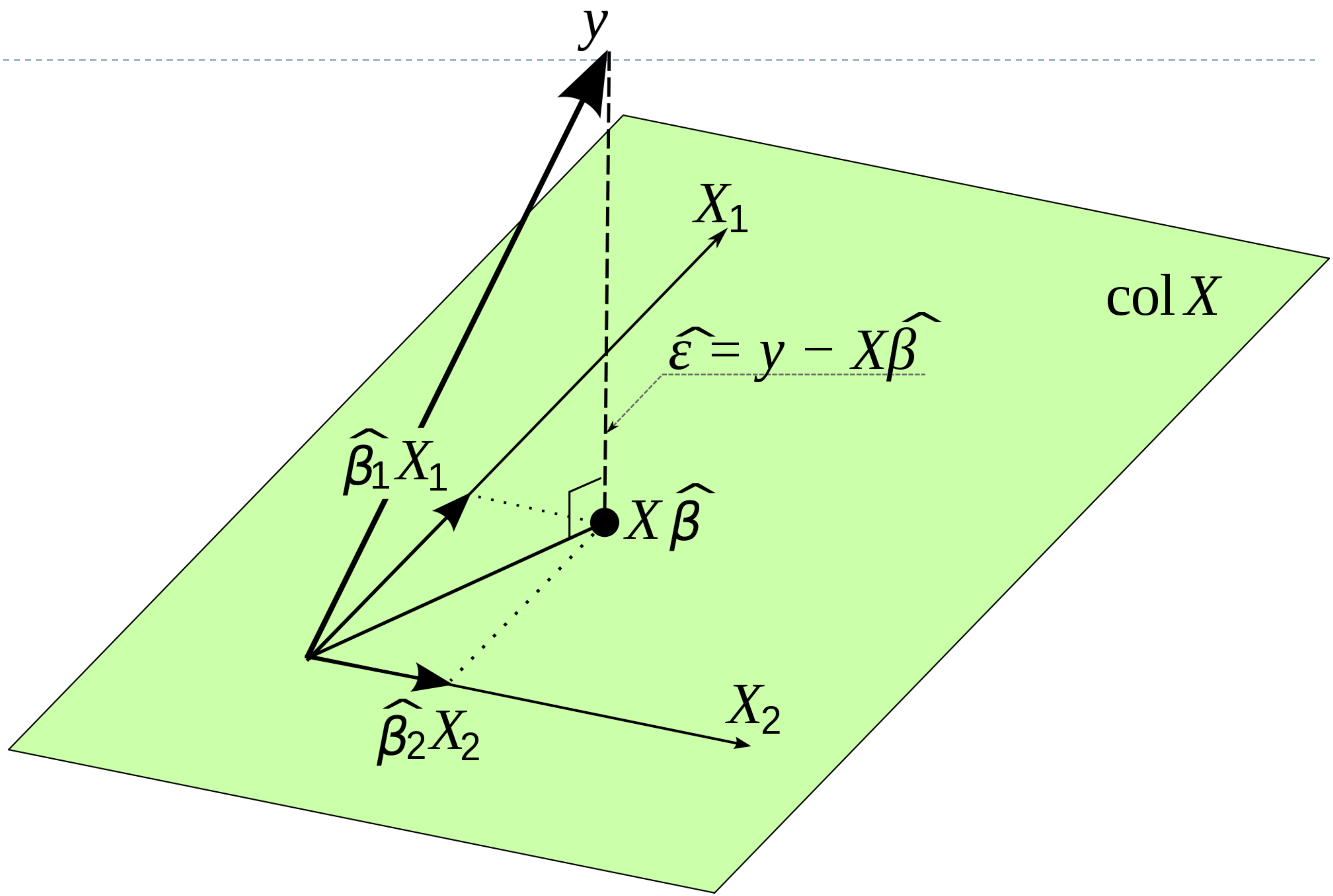
$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_K X_{iK} + \varepsilon_i, \text{ where } \varepsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$$

- *K+2* parameters: $\beta_0, \beta_1, \ldots, \beta_K$, and σ^2

Matrix notation (Ch. 10 Exercises 20 and 21):

$$\boldsymbol{Y}_{n\times 1} = \boldsymbol{X}_{n\times(K+1)} \boldsymbol{\beta}_{(K+1)\times 1} + \boldsymbol{\varepsilon}_{n\times 1},$$
$$\boldsymbol{\varepsilon} \sim N_n\left(0, \sigma^2 I_{n\times n}\right).$$

Geometrically, $\hat{\boldsymbol{Y}}$ can be interpreted as the orthogonal projection of $\boldsymbol{Y}$ onto the subspace generated by the columns of the design matrix $\boldsymbol{X}$.



http://en.wikipedia.org/wiki/Ordinary_least_squares

Sampling Distributions of Multiple Regression Estimators

Coefficients: $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$

$$E(\hat{\beta}_j) = \beta_j$$

$$Var(\hat{\beta}_j) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{[j,j]} \text{ (see HW10)}$$

$$\hat{\boldsymbol{\beta}} \sim N_{K+1}\left(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)$$

Residual variance: $\hat{\sigma}^2 = \dfrac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n-(K+1)} \sim \dfrac{\sigma^2\chi^2_{n-(K+1)}}{n-(K+1)}$

$$\hat{Y}_i = \hat{\mu}(Y_i \mid X_{i1}X_{i2},\ldots,X_{iK}) = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \ldots + \hat{\beta}_K X_{iK}$$

$$E(\hat{\sigma}^2) = \sigma^2$$

Multiple Regression: Inference

- Least squares estimators of coefficients;
- *t*-tests and confidence intervals for coefficients;
- *t*-tests and confidence intervals for linear combinations of coefficients;
- Confidence and prediction intervals for mean response;
- Extra-sum-of-squares *F*-test to compare regression models.

Inference About Individual Coefficients

$$H_0 : \beta_j = \beta_j^0 \ \ (\text{usually } \beta_j^0 = 0)$$

$$H_a : \beta_j \neq \beta_j^0$$

- *t*-test: $$\frac{\hat{\beta}_j - \beta_j^0}{\hat{\sigma}\sqrt{(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{[j,j]}}} \overset{H_0}{\sim} t_{n-(K+1)}$$

- (1-α)100% confidence interval:

$$\hat{\beta}_j \pm t_{n-(K+1),(1-\alpha/2)}\hat{\sigma}\sqrt{(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{[j,j]}}$$

Modeling Car Price: Inference About Individual Coefficients

```
> RegModel <- lm(Price ~ c.Mileage*c.Liter, data = CarData)
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         2.1327e+04  2.844e+02  74.995  < 2e-16 ***
c.Mileage          -1.580e-01  3.472e-02  -4.549 6.23e-06 ***
c.Liter             4.962e+03  2.574e+02  19.278  < 2e-16 ***
c.Mileage:c.Liter  -9.493e-02  3.033e-02  -3.130  0.00181 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8062 on 800 degrees of freedom
```

$$\boxed{\hat{\sigma} = 8062}$$

You can get the vector of standard errors, $\hat{\sigma}\sqrt{(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{[j,j]}}$ using the following code:

```
> Coef.vars <- diag(vcov(RegModel))
> sqrt(Coef.vars)
(Intercept)          c.Mileage          c.Liter   c.Mileage:c.Liter
 284.382             0.0347             257.40        0.0303
```

Modeling Car Price: Inference About Individual Coefficients

```
> RegModel <- lm(Price ~ c.Mileage*c.Liter, data = CarData)
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.1327e+04  2.844e+02  74.995  < 2e-16 ***
c.Mileage         -1.580e-01  3.472e-02  -4.549 6.23e-06 ***
c.Liter            4.962e+03  2.574e+02  19.278  < 2e-16 ***
c.Mileage:c.Liter -9.493e-02  3.033e-02  -3.130  0.00181 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8062 on 800 degrees of freedom
```

$$\frac{\boxed{\hat{\beta}_j} - 0}{\boxed{\hat{\sigma}\sqrt{(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{[j,j]}}}} = t\text{-statistic}$$

A two-sided *p*-value = $2P(t_{800} > |t\text{-statistic}|)$.

Modeling Car Price: Inference About Individual Coefficients

```
> RegModel <- lm(Price ~ c.Mileage*c.Liter, data = CarData)
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.1327e+04  2.844e+02  74.995  < 2e-16 ***
c.Mileage         -1.580e-01  3.472e-02  -4.549 6.23e-06 ***
c.Liter            4.962e+03  2.574e+02  19.278  < 2e-16 ***
c.Mileage:c.Liter -9.493e-02  3.033e-02  -3.130  0.00181 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8062 on 800 degrees of freedom
...
```

- Verify that the *t*-statistic for the intercept is 74.995 and for the slope of *c.Mileage* it is -4.549. How do we calculate the corresponding *p*-values?

- Verify that the 95% CI for the intercept is (20770, 21884), and for the slope of *c.Mileage* it is (-0.226, -0.090).

Modeling Car Price: Inference About Individual Coefficients

```
> RegModel <- lm(Price ~ c.Mileage*c.Liter, data = CarData)
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.1327e+04  2.844e+02  74.995  < 2e-16 ***
c.Mileage         -1.580e-01  3.472e-02  -4.549 6.23e-06 ***
c.Liter            4.962e+03  2.574e+02  19.278  < 2e-16 ***
c.Mileage:c.Liter -9.493e-02  3.033e-02  -3.130  0.00181 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8062 on 800 degrees of freedom
```

Use the following function to calculate all CIs in R:

```
> confint(RegModel, level = 0.95)

                          2.5 %        97.5 %
(Intercept)       20768.9058785  2.188535e+04
c.Mileage            -0.2261215 -8.979740e-02
c.Liter            4456.8969346  5.467411e+03
c.Mileage:c.Liter    -0.1544642 -3.540147e-02
```

Multiple Regression: Inference

- Least squares estimators of coefficients;
- *t*-tests and confidence intervals for coefficients;
- *t*-tests and confidence intervals for linear combinations of coefficients;
- Confidence and prediction intervals for mean response;
- Extra-sum-of-squares *F*-test to compare regression models.

Inference About Linear Combinations of Coefficients

$$H_0: C_0\beta_0 + C_1\beta_1 + \ldots + C_K\beta_K = \boldsymbol{C}^T\boldsymbol{\beta} = \gamma$$

$$H_a: C_0\beta_0 + C_1\beta_1 + \ldots + C_K\beta_K \neq \gamma,$$

where $\boldsymbol{C} = \begin{pmatrix} C_0 \\ C_1 \\ \ldots \\ C_K \end{pmatrix}$ is a column-vector of numbers.

- Let $\hat{\gamma} = \boldsymbol{C}^T\hat{\boldsymbol{\beta}}$
- Then $E(\hat{\gamma}) = \gamma$, $Var(\hat{\gamma}) = \boldsymbol{C}^T Var(\hat{\boldsymbol{\beta}})\boldsymbol{C} = \sigma^2\boldsymbol{C}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{C}$
- Also, $\hat{\gamma} \overset{H_0}{\sim} N\left(\gamma, \sigma^2\boldsymbol{C}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{C}\right)$

Inference About Linear Combinations of Coefficients

$$H_0: C_0\beta_0 + C_1\beta_1 + \ldots + C_K\beta_K = \boldsymbol{C}^{\boldsymbol{T}}\boldsymbol{\beta} = \gamma$$

$$H_a: C_0\beta_0 + C_1\beta_1 + \ldots + C_K\beta_K \neq \gamma,$$

$$\hat{\gamma} \overset{H_0}{\sim} N\left(\gamma, \sigma^2 \boldsymbol{C}^{\boldsymbol{T}}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-\boldsymbol{1}}\boldsymbol{C}\right)$$

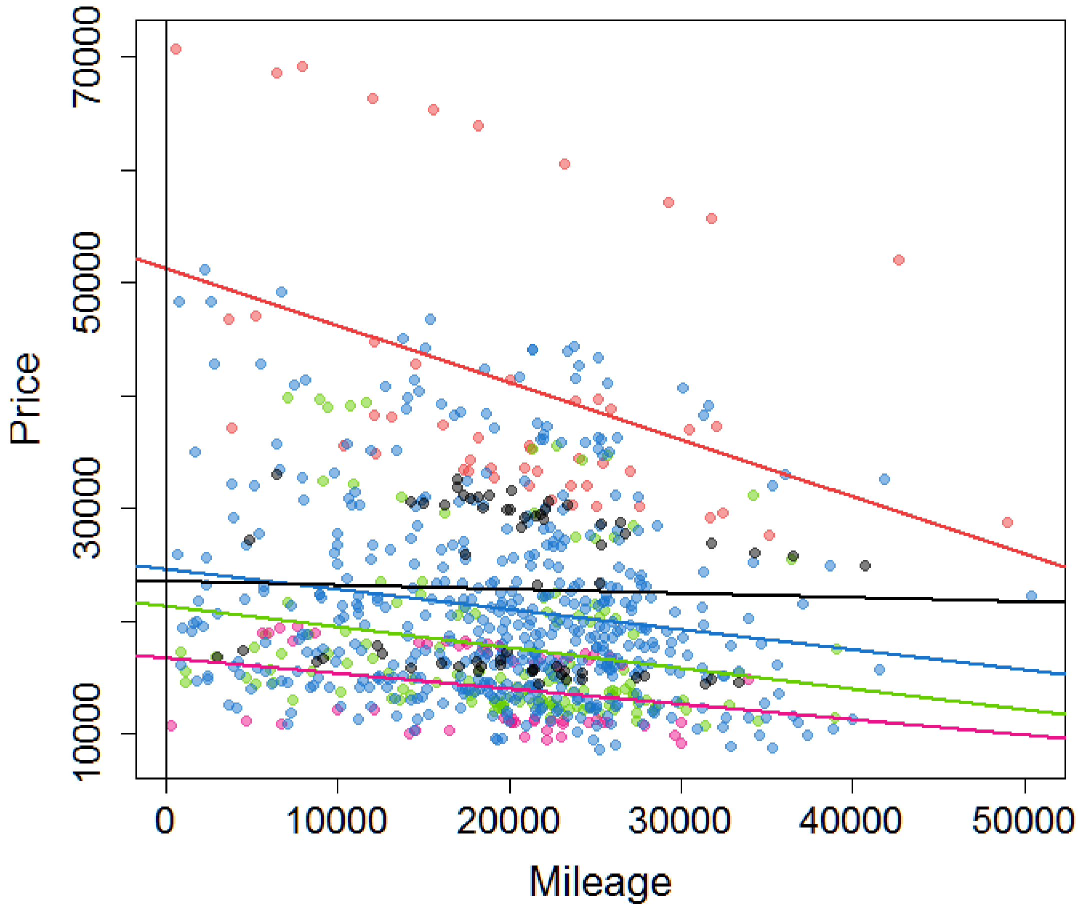
- t-test:

$$\frac{\hat{\gamma} - \gamma}{\hat{\sigma}\sqrt{\boldsymbol{C}^{\boldsymbol{T}}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-\boldsymbol{1}}\boldsymbol{C}}} \overset{H_0}{\sim} t_{n-(K+1)}$$

- See R&S Section 10.4.3 and Display 10.15 for more details.

Inference About Linear Combinations of Coefficients: Comparing Slopes of Non-Reference Levels

- Convertible, Coupe, Hatchback, Sedan, Wagon



- `lm(formula = Price ~ Mileage + Type + Mileage:Type, data = CarData)`

Inference About Linear Combinations of Coefficients: Comparing Slopes of Non-Reference Levels

Models for each car type:

$$\mu(\text{Price}_i \mid \text{Mileage}_i, \text{Type}_i = \text{Convertible}) = \beta_0 + \beta_1 \cdot \text{Mileage}_i$$

$$\mu(\text{Price}_i \mid \text{Mileage}_i, \text{Type}_i = \text{Coupe}) = (\beta_0 + \beta_2) + (\beta_1 + \boxed{\beta_6}) \cdot \text{Mileage}_i$$

$$\mu(\text{Price}_i \mid \text{Mileage}_i, \text{Type}_i = \text{Hatchback}) = (\beta_0 + \beta_3) + (\beta_1 + \boxed{\beta_7}) \cdot \text{Mileage}_i$$

$$\mu(\text{Price}_i \mid \text{Mileage}_i, \text{Type}_i = \text{Sedan}) = (\beta_0 + \beta_4) + (\beta_1 + \beta_8) \cdot \text{Mileage}_i$$

$$\mu(\text{Price}_i \mid \text{Mileage}_i, \text{Type}_i = \text{Wagon}) = (\beta_0 + \beta_5) + (\beta_1 + \beta_9) \cdot \text{Mileage}_i$$

$$H_0: \beta_6 = \beta_7,$$
$$H_a: \beta_6 \neq \beta_7.$$

$\Leftrightarrow$

$$\boldsymbol{C} = (0, \quad 0, \quad \ldots, \quad 1, \quad -1, \quad \ldots, \quad 0)$$

$$H_0: C_0\beta_0 + C_1\beta_1 + \ldots + C_6\beta_6 + C_7\beta_7 + \ldots + C_9\beta_9 = \boldsymbol{C}^T\boldsymbol{\beta} = 0$$

$$H_a: C_0\beta_0 + C_1\beta_1 + \ldots + C_6\beta_6 + C_7\beta_7 + \ldots + C_9\beta_9 \neq 0$$

Inference About Linear Combinations of Coefficients: Comparing Slopes of Non-Reference Levels

```
regmodel <- lm(Price ~ Mileage*Type,
                data = CarData)
C = as.matrix(c(0,0,0,0,0,0,1,-1,0,0))
C = t(C)
VAR <- vcov(regmodel)
SE <- sqrt(C %*% VAR %*% t(C))
gamma <- sum(C*coef(regmodel))

t_statistic <- gamma/SE # t=-0.271
2*(1-pt(abs(t_statistic), df=794))# p-value=0.7863
```


EASIER SOLUTION: redefine the reference level

Inference About Linear Combinations of Coefficients: Comparing Slopes of Non-Reference Levels

```
> levels(CarData$Type)
[1] "Convertible“  "Coupe"    "Hatchback"   "Sedan"       "Wagon"
> CarData$Type <- factor(CarData$Type, levels = c("Coupe", "Convertible",
                                       "Hatchback", "Sedan", "Wagon"))
> regmodel <- lm(Price ~ Mileage*Type, data = CarData)
> summary(regmodel)

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              2.138e+04  1.825e+03  11.713   <2e-16 ***
Mileage                 -1.840e-01  8.526e-02  -2.158   0.0312 *
TypeConvertible          2.989e+04  3.274e+03   9.129   <2e-16 ***
TypeHatchback           -4.580e+03  3.517e+03  -1.302   0.1931
TypeSedan                3.201e+03  2.053e+03   1.560   0.1193
TypeWagon                2.267e+03  3.543e+03   0.640   0.5225
Mileage:TypeConvertible -3.203e-01  1.465e-01  -2.186   0.0291 *
Mileage:TypeHatchback    4.625e-02  1.705e-01   0.271   0.7863
Mileage:TypeSedan        5.868e-03  9.587e-02   0.061   0.9512
Mileage:TypeWagon        1.457e-01  1.632e-01   0.893   0.3722
```


EASIER SOLUTION: redefine the reference level

Multiple Regression: Inference

- Least squares estimators of coefficients;
- t-tests and confidence intervals for coefficients;
- t-tests and confidence intervals for linear combinations of coefficients;
- Confidence and prediction intervals for mean response;
- Extra-sum-of-squares *F*-test to compare regression models.

Confidence and Prediction Intervals At X=X_0

$$\hat{\mu}(Y \mid \boldsymbol{X} = \boldsymbol{X}_0) = \hat{\beta}_0 + \hat{\beta}_1 X_{01} + \ldots + \hat{\beta}_K X_{0K}, \text{ where } \boldsymbol{X}_0 = \begin{pmatrix} 1 \\ X_{01} \\ \ldots \\ X_{0K} \end{pmatrix}.$$

- If viewed as a special case of a linear combination, $\boldsymbol{C} = \boldsymbol{X}_0$, we get

$$\hat{\mu}(Y \mid \boldsymbol{X} = \boldsymbol{X}_0) = \boldsymbol{C}^{\boldsymbol{T}} \hat{\boldsymbol{\beta}} = \hat{\gamma} \overset{H_0}{\sim} N\left(\boldsymbol{C}^{\boldsymbol{T}} \boldsymbol{\beta}, \sigma^2 \boldsymbol{C}^{\boldsymbol{T}} (\boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{X})^{-\boldsymbol{1}} \boldsymbol{C}\right)$$

or $\hat{\mu}(Y \mid \boldsymbol{X} = \boldsymbol{X}_0) \overset{H_0}{\sim} N\left(\boldsymbol{X}_{\boldsymbol{0}}^{\boldsymbol{T}} \boldsymbol{\beta}, \sigma^2 \boldsymbol{X}_{\boldsymbol{0}}^{\boldsymbol{T}} (\boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{X})^{-\boldsymbol{1}} \boldsymbol{X}_{\boldsymbol{0}}\right)$

Confidence Interval At $X=X_0$

$$\hat{\mu}(Y \mid \boldsymbol{X} = \boldsymbol{X}_0) \overset{H_0}{\sim} N\left(\boldsymbol{X}_0^T \boldsymbol{\beta}, \sigma^2 \boldsymbol{X}_0^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}_0\right)$$

Point-wise confidence interval:

$$\boldsymbol{X}_0^T \hat{\boldsymbol{\beta}} \pm t_{n-(K+1),(1-\alpha/2)} \hat{\sigma} \sqrt{\boldsymbol{X}_0^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}_0}$$

```
> RegModel <- lm(Price ~ Mileage*Liter,
                 data = CarData)
> NewDataSet <- data.frame(Mileage = c(10000, 20000),
                           Liter = c(4,8))
> predict(RegModel, new = NewDataSet,
                    interval="confidence")
       fit      lwr      upr
1 28555.72 27427.34 29684.10
2 45847.02 43276.85 48417.18
```

Prediction Interval At $X=X_0$

$$\hat{\mu}(Y \mid \boldsymbol{X} = \boldsymbol{X}_0) \overset{H_0}{\sim} N\left(\boldsymbol{X}_0^T \boldsymbol{\beta}, \sigma^2 \boldsymbol{X}_0^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}_0\right)$$

Point-wise prediction interval:

$$\boldsymbol{X}_0^T \hat{\boldsymbol{\beta}} \pm t_{n-(K+1),(1-\alpha/2)} \hat{\sigma} \sqrt{\boxed{1+} \boldsymbol{X}_0^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}_0}$$

```
> RegModel <- lm(Price ~ Mileage*Liter,
                 data = CarData)
> NewDataSet <- data.frame(Mileage = c(10000, 20000),
                           Liter = c(4,8))
> predict(RegModel, new = NewDataSet,
                    interval="prediction")
       fit      lwr      upr
1 28555.72 12689.77 44421.67
2 45847.02 29813.89 61880.14
```

Simultaneous Confidence Intervals At $X=X_0$

Simultaneous confidence interval (Scheffe's method):

$$\boldsymbol{X}_0^T \hat{\boldsymbol{\beta}} \pm \sqrt{(K+1)F_{(K+1),\, n-(K+1),\, (1-\alpha)}}\,\hat{\sigma}\sqrt{\boldsymbol{X}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_0}$$

```
> NewDataSet <- data.frame(Mileage = c(10000,20000),
                           Liter = c(4,8))
> p <- length(RegModel$coef)
> n <- dim(CarData)[1]
> pred.int <- predict(RegModel, new = NewDataSet,
                      interval="confidence")
> pred.se <- (pred.int[,3]-pred.int[,1])/qt(0.975,n-p)
> mu.L = pred.int[,1] - sqrt(p*qf(0.95,p,n-p))*pred.se
> mu.U = pred.int[,2] + sqrt(p*qf(0.95,p,n-p))*pred.se
> cbind(pred=pred.int[,1], mu.L,mu.U)
      pred     mu.L     mu.U
1 28555.72 26780.93 29202.13
2 45847.02 41804.49 47319.38
```

Simultaneous Prediction Intervals At $X = X_0$

Simultaneous prediction interval (Scheffe's method):

$$\boldsymbol{X}_0^T \hat{\boldsymbol{\beta}} \pm \sqrt{(K+1) F_{(K+1),\, n-(K+1),\, (1-\alpha)}}\, \hat{\sigma} \sqrt{1 + \boldsymbol{X}_0^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}_0}$$

```
> NewDataSet <- data.frame(Mileage = c(10000,20000),
                           Liter = c(4,8))
> p <- length(RegModel$coef)
> n <- dim(CarData)[1]
> pred.int <- predict(RegModel, new = NewDataSet,
                               interval="prediction")
> pred.se <- (pred.int[,3]-pred.int[,1])/qt(0.975,n-p)
> mu.L = pred.int[,1] - sqrt(p*qf(0.95,p,n-p))*pred.se
> mu.U = pred.int[,2] + sqrt(p*qf(0.95,p,n-p))*pred.se
> cbind(pred=pred.int[,1], mu.L,mu.U)
      pred      mu.L     mu.U
1 28555.72  3600.688 37644.80
2 45847.02 20629.050 55031.86
```

Multiple Regression: Inference

- Least squares estimators of coefficients;
- t-tests and confidence intervals for coefficients;
- t-tests and confidence intervals for linear combinations of coefficients;
- Confidence and prediction intervals for mean response;
- Extra-sum-of-squares *F*-test to compare regression models.

Extra-Sum-of-Squares *F*-test: Nested (Hierarchical) Models

1. Separate Means:

$$\mu\{Y \mid X\} = \text{separate value for each combination of predictors}$$

2. Regression Model:

$$\mu\{Y \mid X\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

3. "Simpler" Regression Model:

$$\mu\{Y \mid X\} = \beta_0 + \beta_1 X_1$$

4. Equal Means:

$$\mu\{Y \mid X\} = \mu = \beta_0$$

Extra-Sum-of-Squares F-test

- Formal test for several regression coefficients simultaneously:

$$H_0: \mu(Y \mid X_1, X_2, \ldots, X_K) = \beta_0 + \beta_1 X_1 + \ldots + \beta_M X_M$$
$$H_a: \mu(Y \mid X_1, X_2, \ldots, X_K) = \beta_0 + \beta_1 X_1 + \ldots + \beta_M X_M + \ldots + \beta_K X_K$$

- Equivalently,

$$H_0: \beta_{M+1} = \beta_{M+2} = \ldots = \beta_K = 0$$ Reduced

$$H_a: \text{At least one among } \beta_{M+1},\ \beta_{M+2}, \ldots,\ \beta_K \text{ is not } 0$$ Full (M<K)

Extra-Sum-of-Squares F-test

Reduced $H_0 : \mu(Y \mid X_1, X_2, ..., X_K) = \beta_0 + \beta_1 X_1 + ... + \beta_M X_M$

Full (K>M) $H_a : \mu(Y \mid X_1, X_2, ..., X_K) = \beta_0 + \beta_1 X_1 + ... + \beta_M X_M + ... + \beta_K X_K$

$$\boxed{\begin{aligned} R &= \frac{(\mathrm{SSR}_{\mathrm{Reduced}} - \mathrm{SSR}_{\mathrm{Full}})/(\mathrm{d.f.}_{\mathrm{Reduced}} - \mathrm{d.f.}_{\mathrm{Full}})}{\mathrm{SSR}_{\mathrm{Full}} / \mathrm{d.f.}_{\mathrm{Full}}} \\ &= \frac{\mathrm{ESS}/\{\text{number of parameters being tested}\}}{\hat{\sigma}^2 \text{ from Full model}} \end{aligned}}$$

- Exact sampling distribution of R under H_0 is

$$F_{(\mathrm{d.f.}_{\mathrm{Reduced}} - \mathrm{d.f.}_{\mathrm{Full}},\ \mathrm{d.f.}_{\mathrm{Full}})}$$

- Test requires normality assumption to hold!

Car Price: Extra-Sum-of-Squares F-tests

$$R = \frac{(\text{SSR}_{\text{Reduced}} - \text{SSR}_{\text{Full}})/(\text{d.f.}_{\text{Reduced}} - \text{d.f.}_{\text{Full}})}{\text{SSR}_{\text{Full}} / \text{d.f.}_{\text{Full}}}$$

```
> RegModel1 <- lm(Price ~ Mileage*Liter, data = CarData)
> RegModel2 <- lm(Price ~ Mileage, data = CarData)
> summary(RegModel1)
…
Residual standard error: 8062 on 800 degrees of freedom
> summary(RegModel2)
…
Residual standard error: 9789 on 802 degrees of freedom
```

Verify that $R = 191.2$ and d.f. are (2, 800).

Conclusion?

Car Price: Extra-Sum-of-Squares F-tests

In R:

```
> anova(RegModel1, RegModel2)
Analysis of Variance Table

Model 1: Price ~ c.Mileage * c.Liter
Model 2: Price ~ c.Mileage
  Res.Df        RSS Df   Sum of Sq      F    Pr(>F)
1    800 5.2001e+10
2    802 7.6856e+10 -2 -2.4855e+10 191.19   < 2.2e-16 ***
```

Adjusted R-squared Statistic

Occam's Razor (Principle of Parsimony)

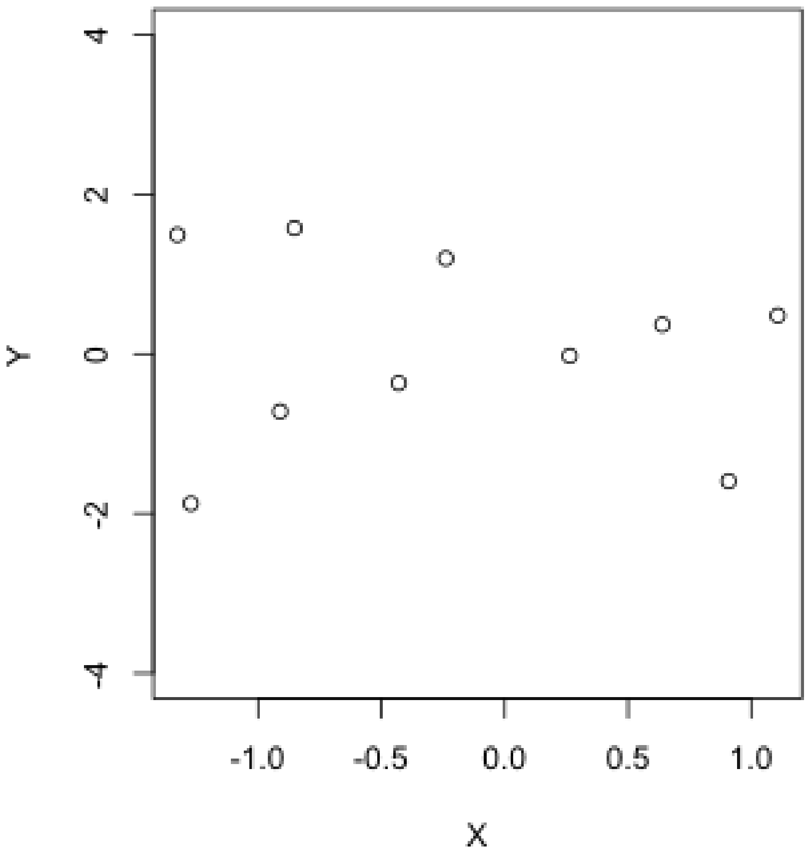
- Observed process can be viewed as a combination of signal and noise.

- Occam's razor: Use the simplest model that explains the **signal,** avoid overfitting!

$$R^2 = \frac{\mathrm{SSR}_{\mathrm{Reduced}} - \mathrm{SSRes}}{\mathrm{SSR}_{\mathrm{Reduced}}} = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}$$

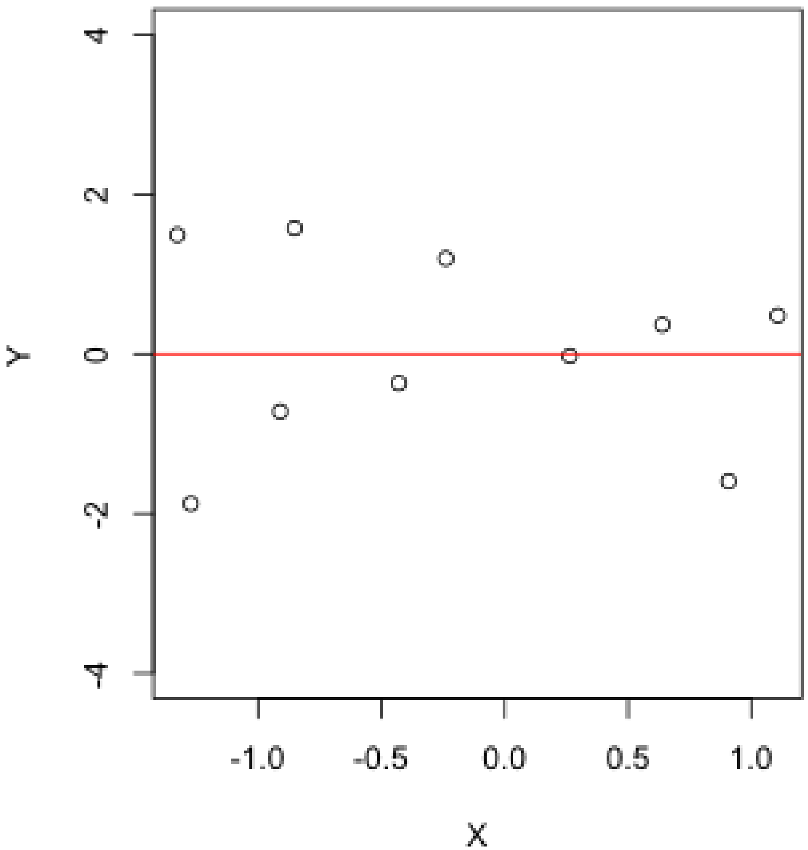
- R^2 evaluates the fit to *signal AND noise;*
  - R^2 increases with every added variable. What is the intuitive reason?

$$E(Y \mid X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + + \beta_K X^K$$



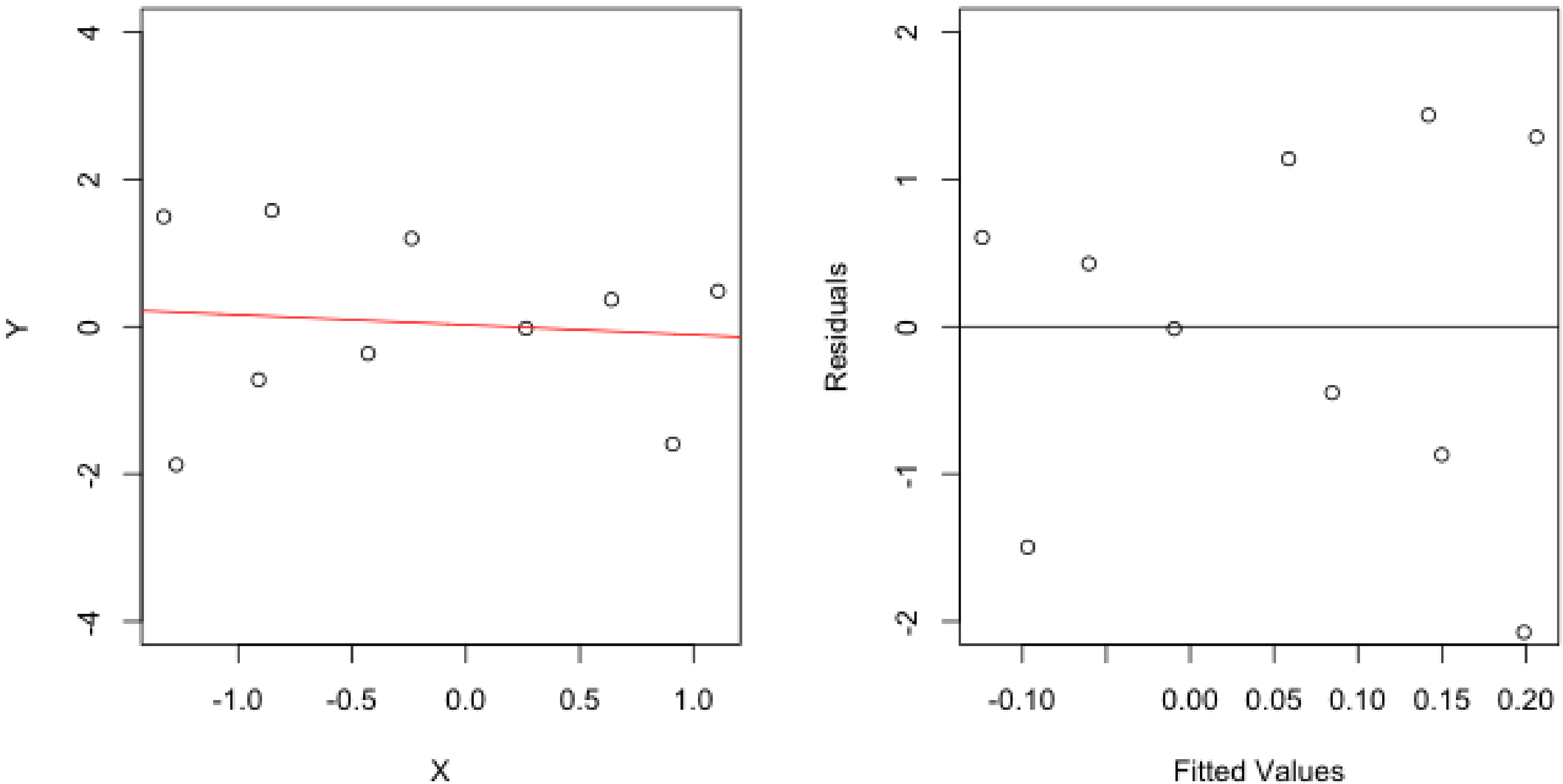
$$E(Y \mid X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + + \beta_K X^K$$

$$R^2 = 0, \; K = 0$$



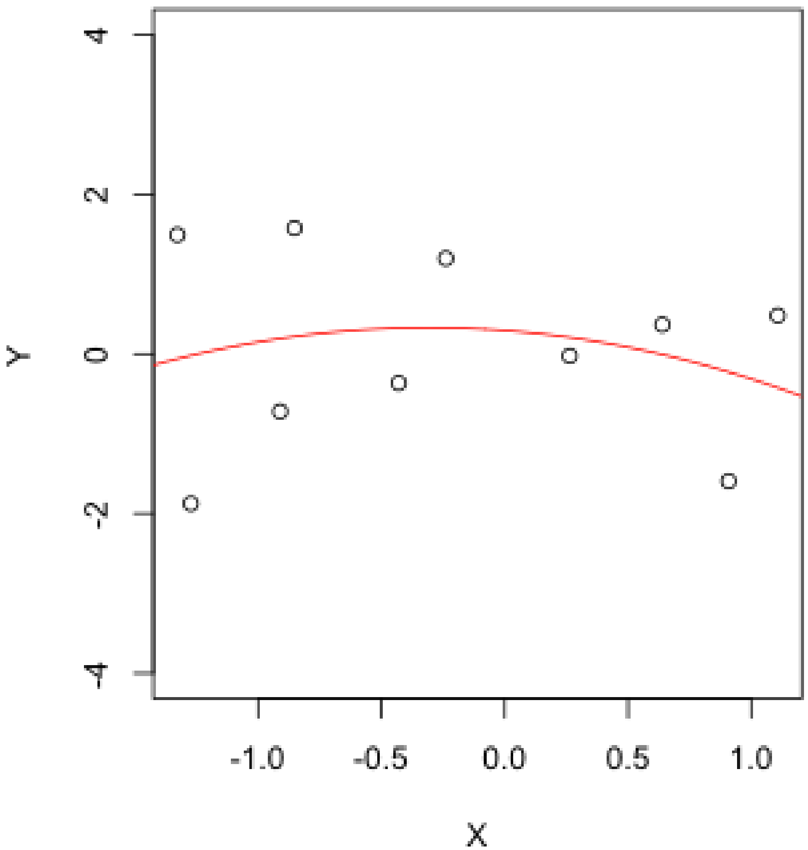
$$E(Y \mid X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + + \beta_K X^K$$

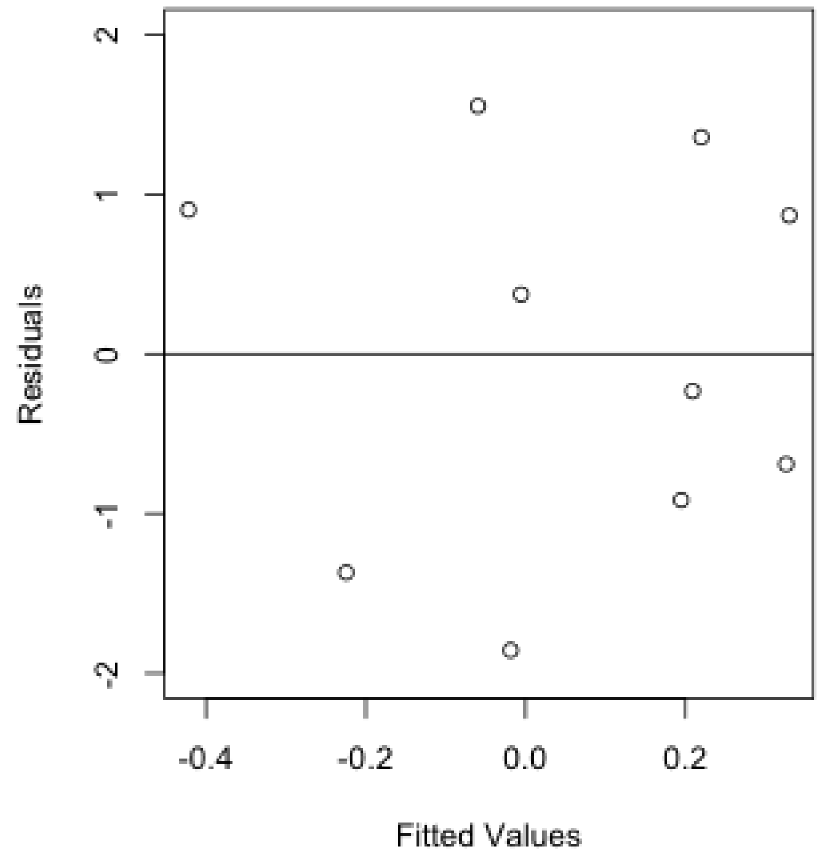
$$R^2 = 0.01,\ K = 1$$



$$E(Y \mid X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + + \beta_K X^K$$

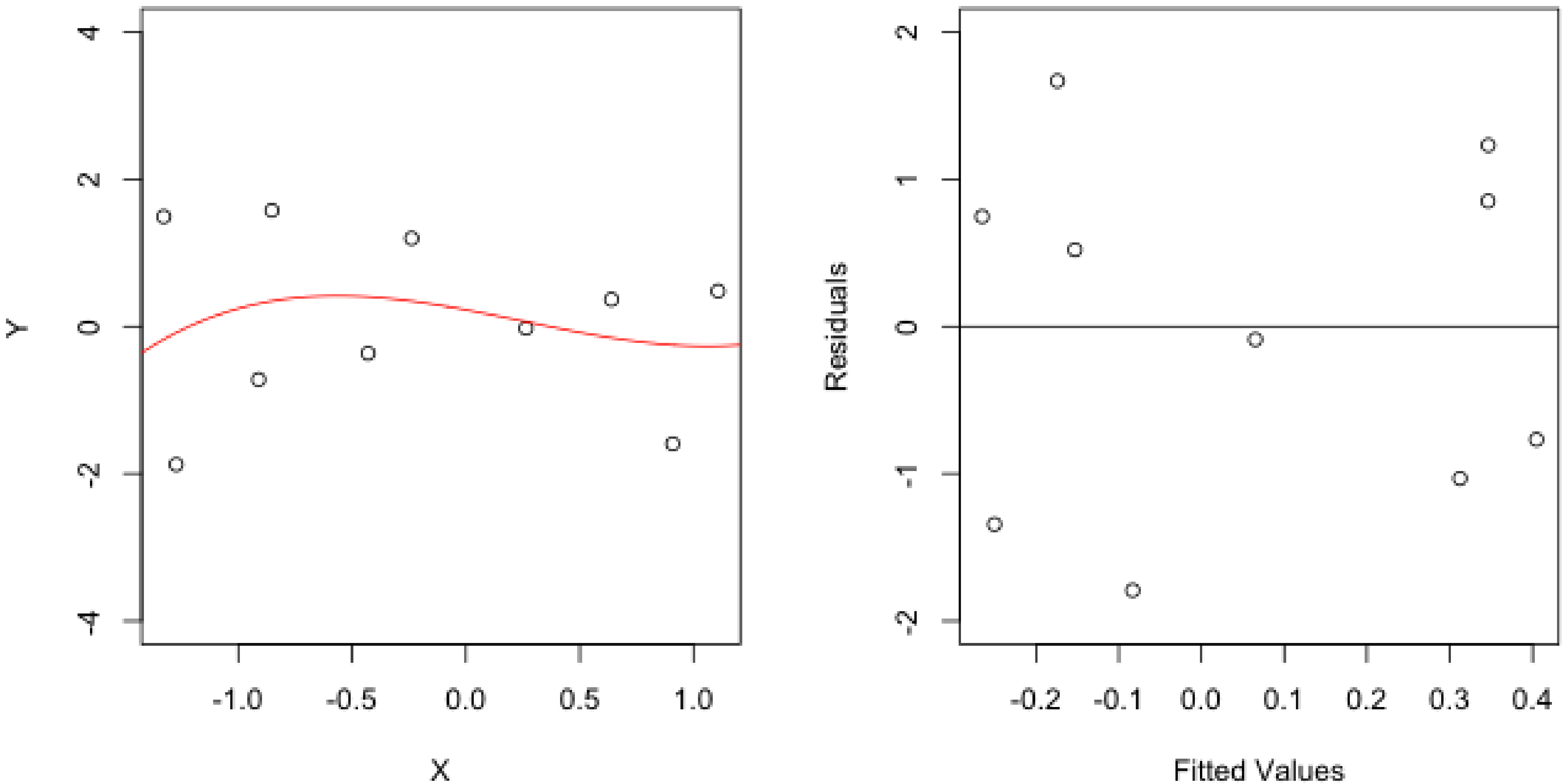
$$R^2 = 0.041, \ K = 2$$





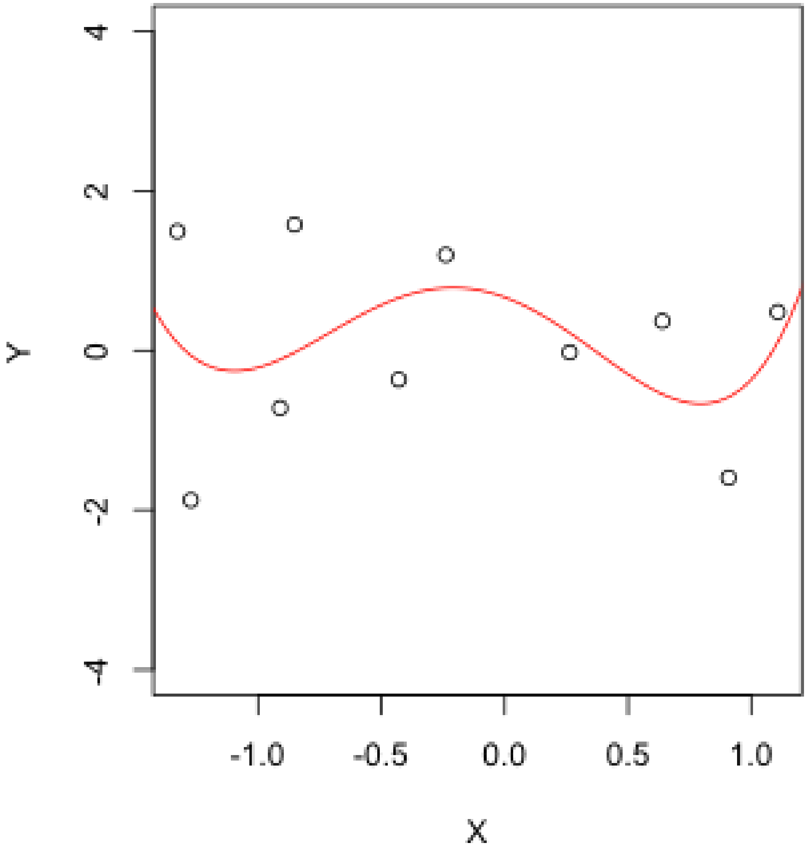
$$E(Y \mid X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + + \beta_K X^K$$

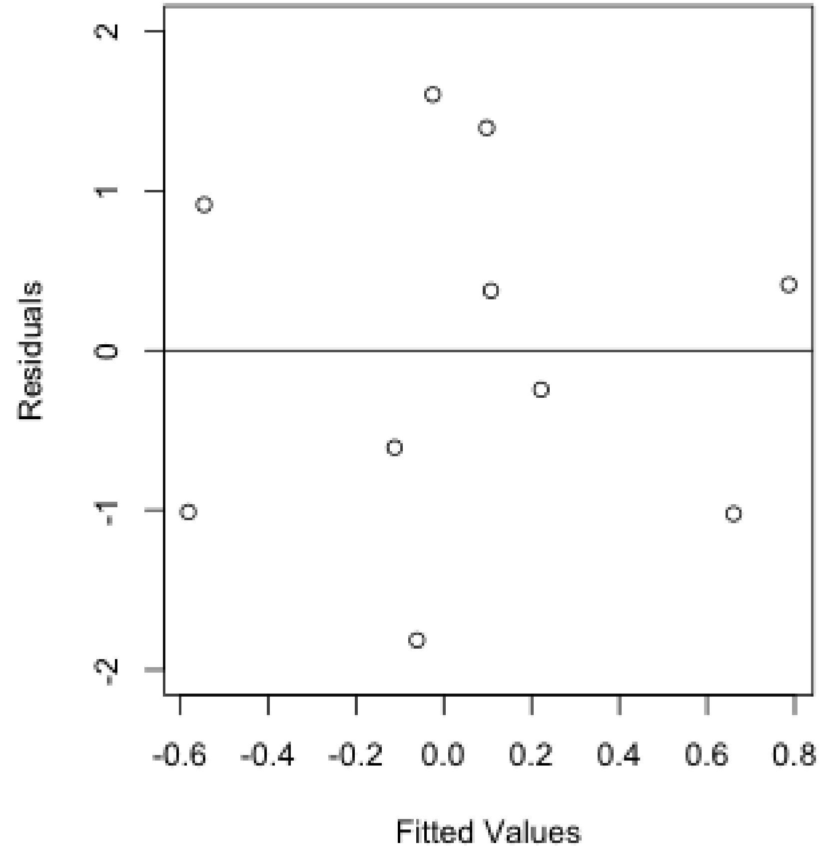
$$R^2 = 0.051, \; K = 3$$



$$E(Y \mid X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + + \beta_K X^K$$

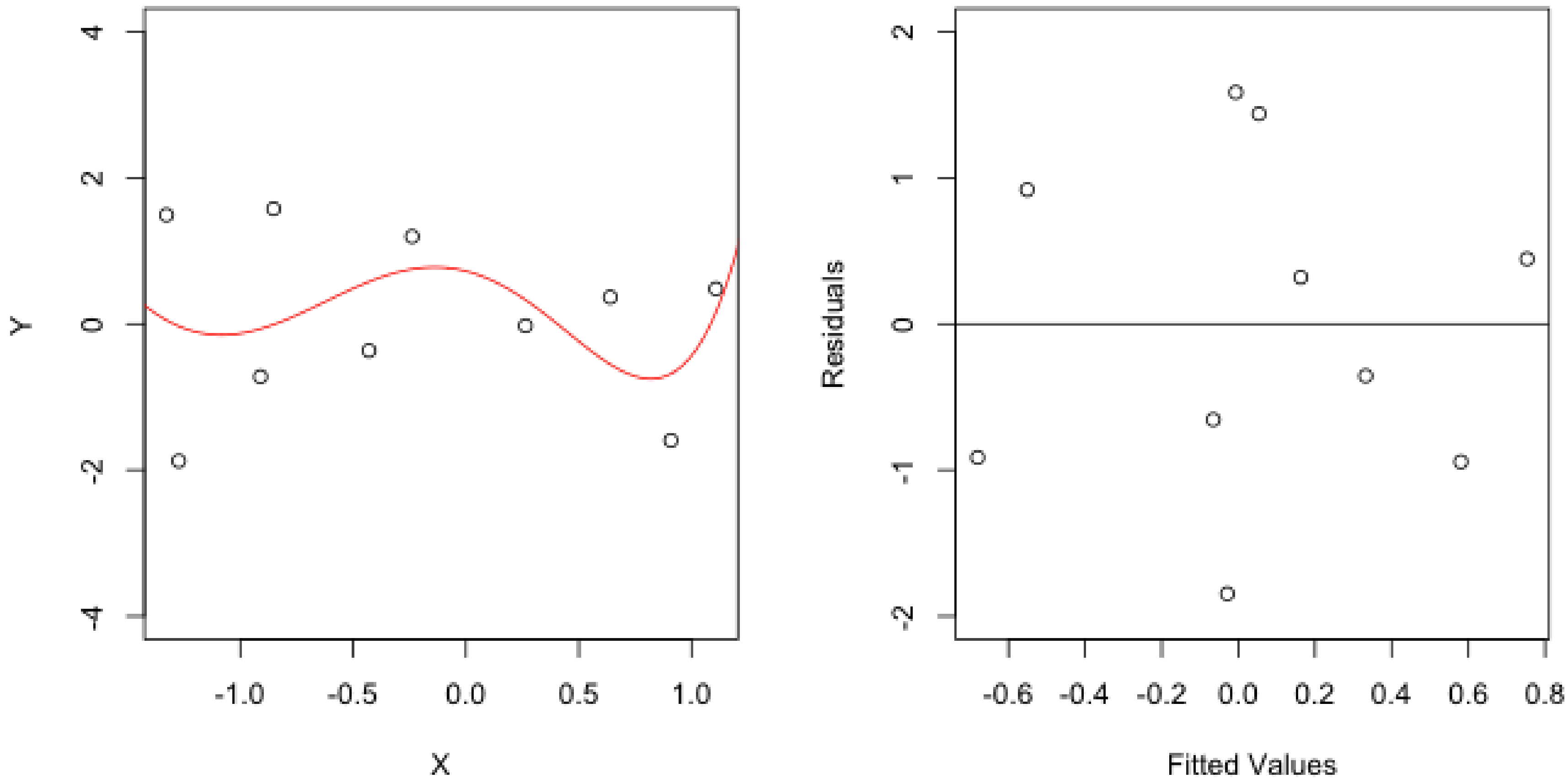
$$R^2 = 0.132,\ K = 4$$





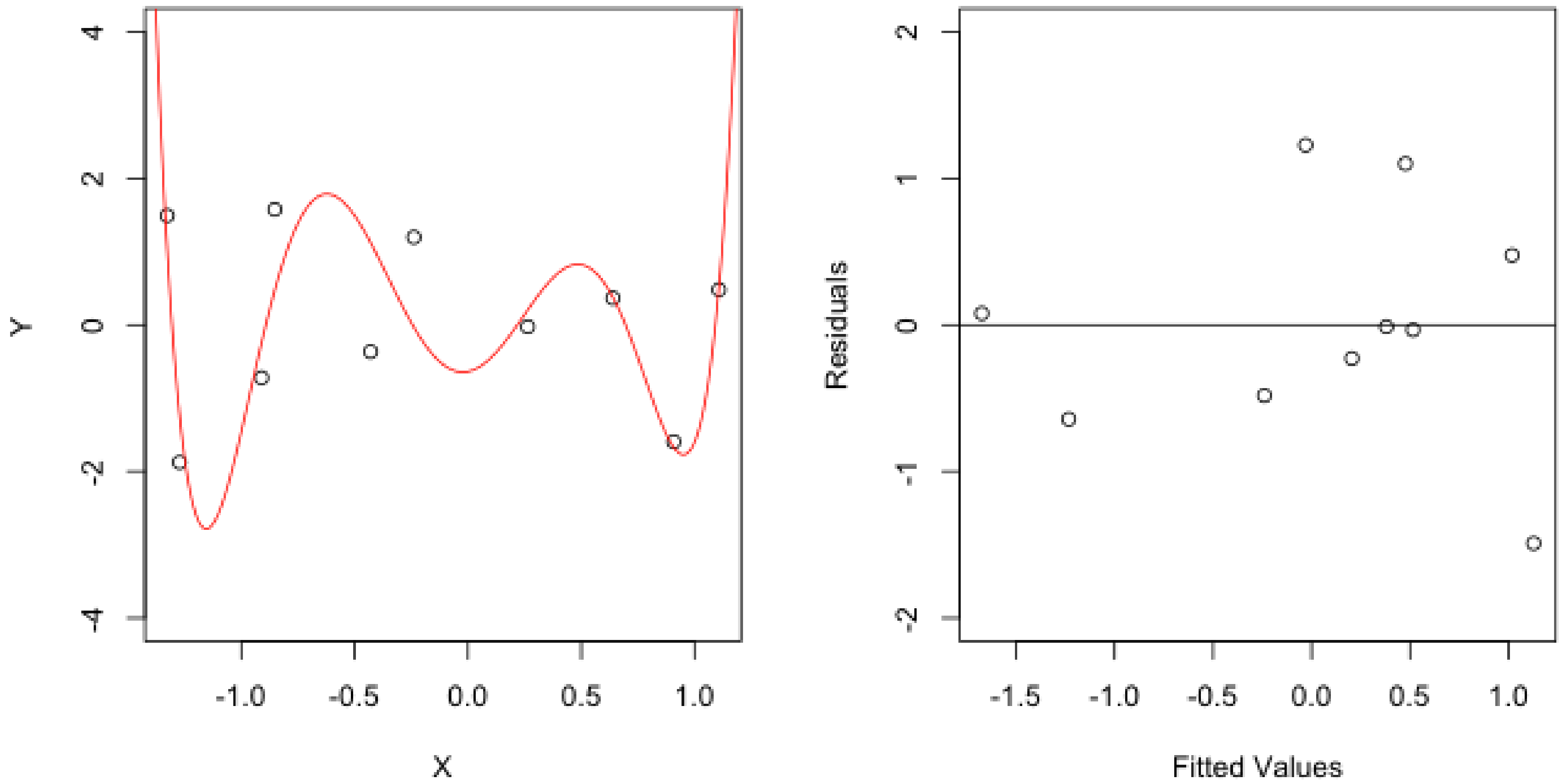
$$E(Y \mid X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + + \beta_K X^K$$

$$R^2 = 0.135, \; K = 5$$



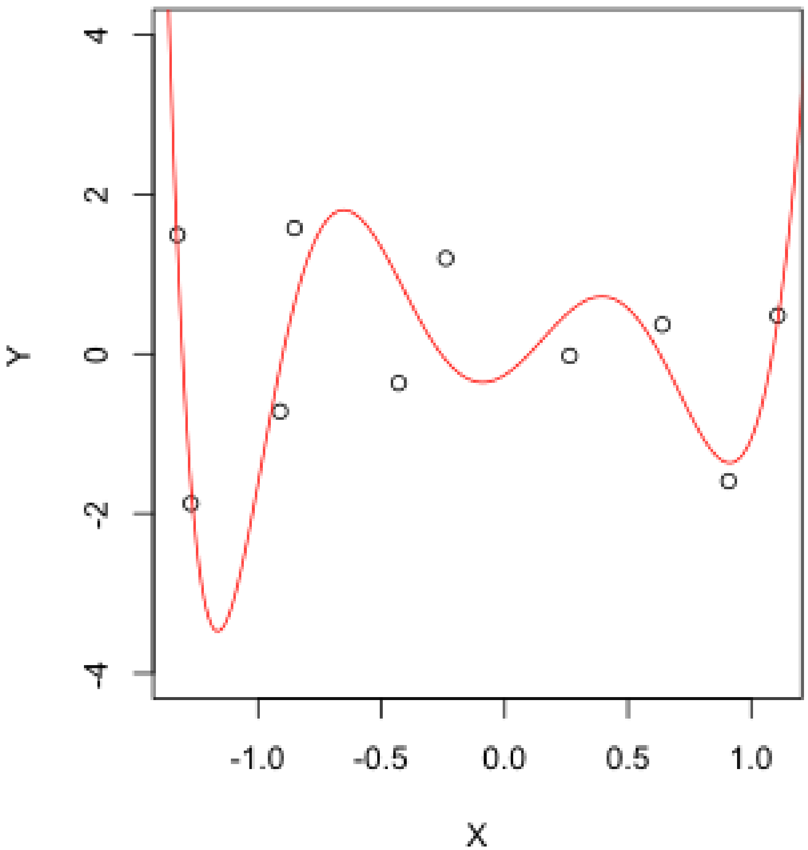
$$E(Y \mid X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + +\beta_K X^K$$

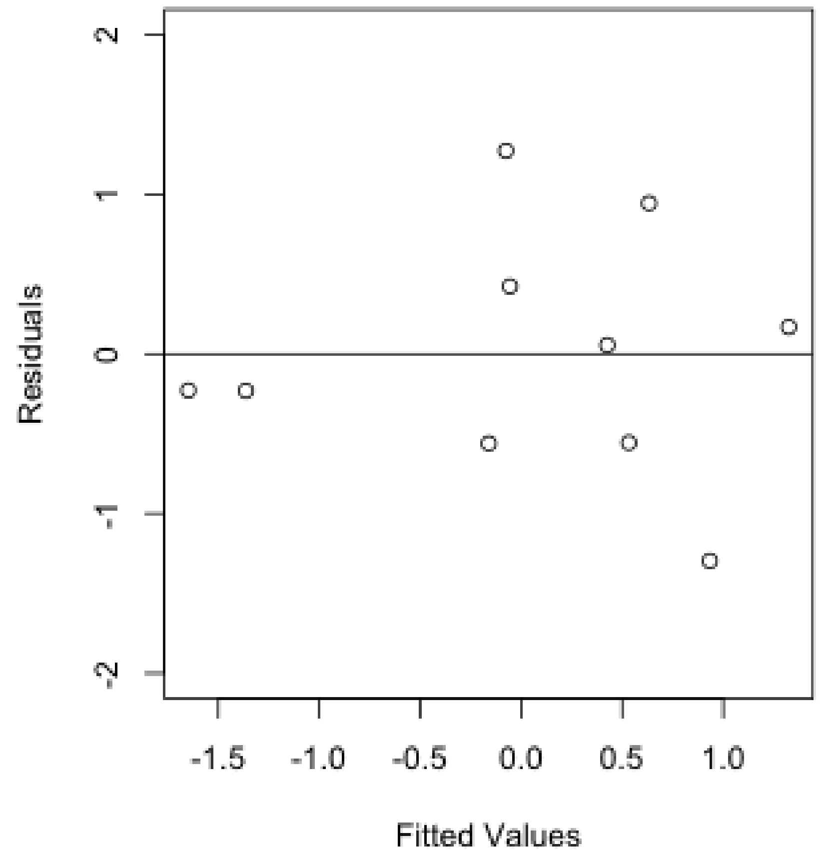
$$R^2 = 0.555, \;\; K = 6$$



$$E(Y \mid X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + + \beta_K X^K$$

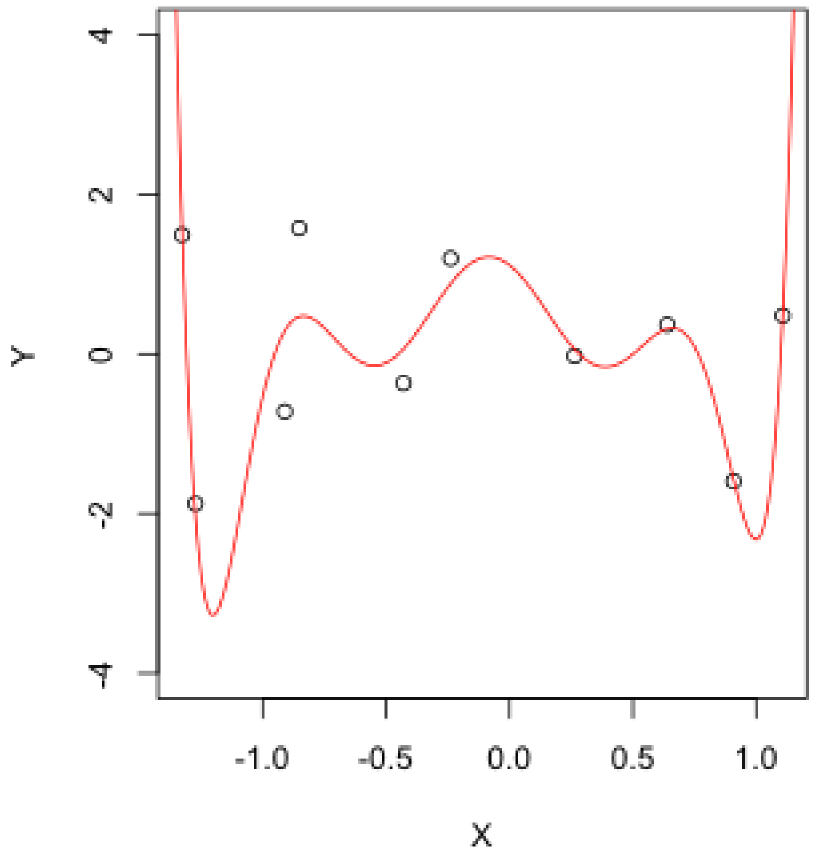
$$R^2 = 0.611, \ K = 7$$

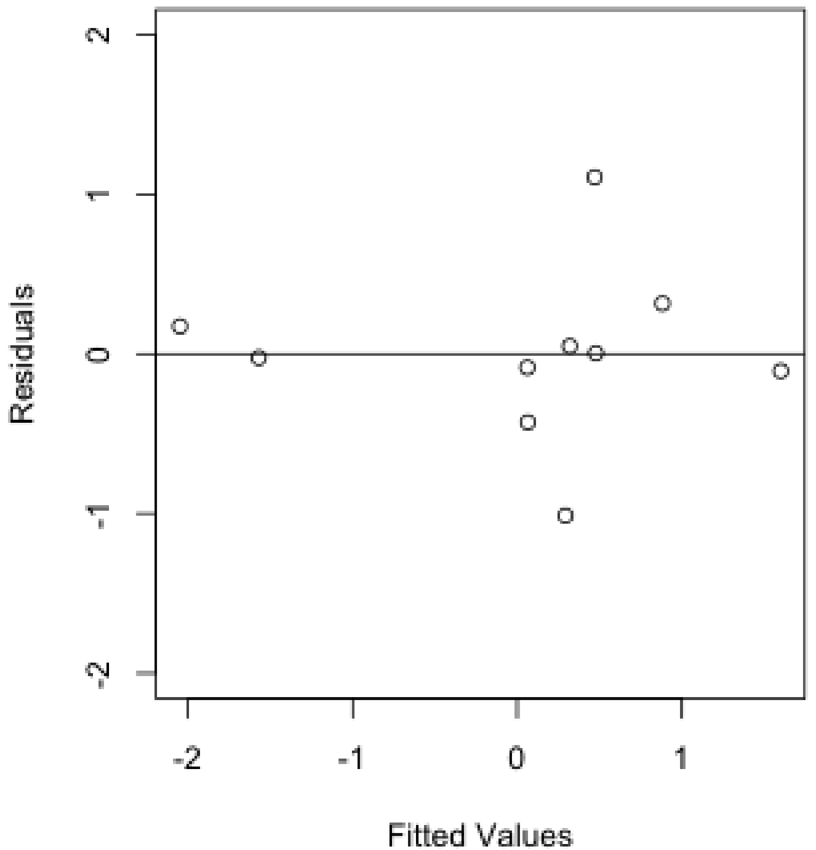




$$E(Y \mid X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + + \beta_K X^K$$

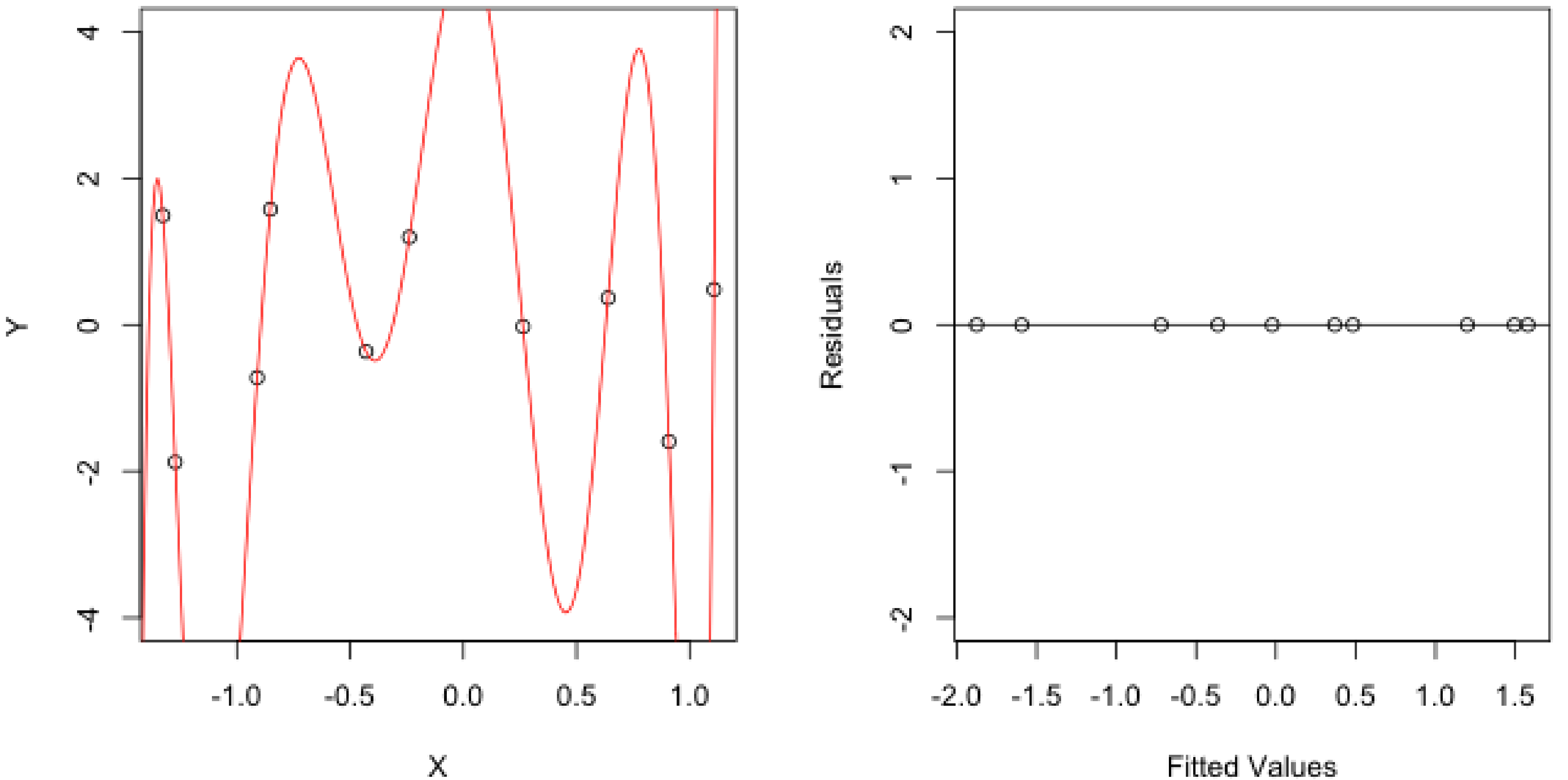
$$R^2 = 0.804, \;\; K = 8$$





$$E(Y \mid X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + + \beta_K X^K$$

$$R^2 = 1,\ K = 9$$



Adjusted R^2 (Adjusted R-squared)

$$\text{Adjusted } R^2 = 1 - \frac{\text{SSRes}/(n - (K+1))}{\text{SSR}_{\text{Reduced}}/(n-1)} = 1 - \frac{\hat{\sigma}^2_{\text{Reg}}}{s_Y^2}$$

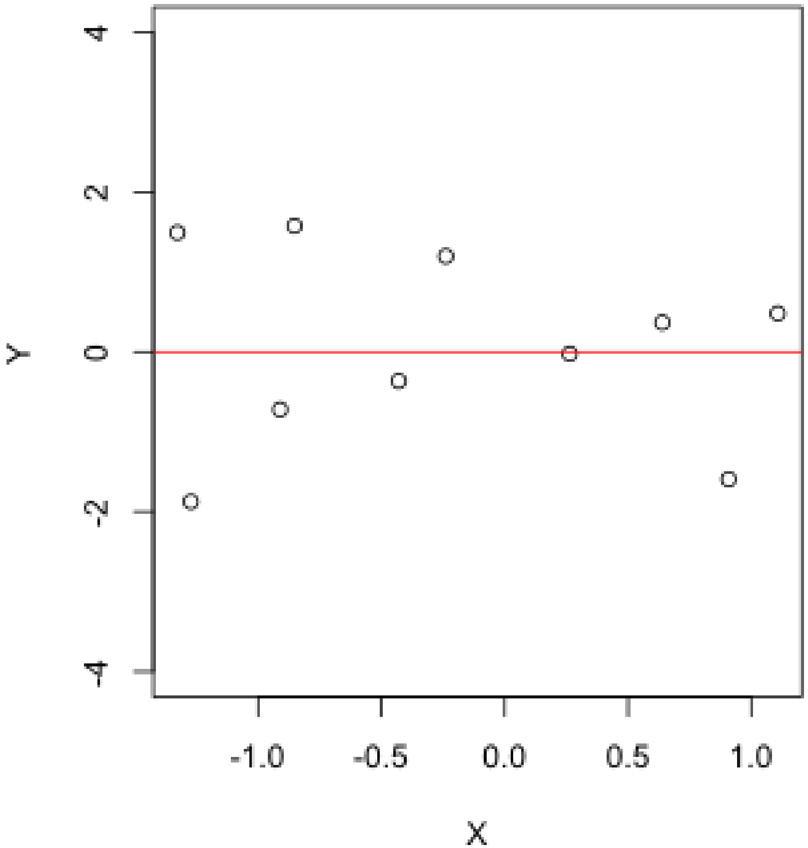
$$= 1 - (1 - R^2)\frac{n-1}{n-(K+1)}$$

K is the number of predictors and n is the sample size.

- Adjusted R^2 adjusts both the numerator and the denominator by their respective degrees of freedom.
- Adjusted R^2 penalizes an excess of variables.
- Not a proportion, does not have to be between 0 and 1.

$$E(Y \mid X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + + \beta_K X^K$$

$$R^2 = 0, K = 0, \text{ Adjusted } R^2 = 0$$

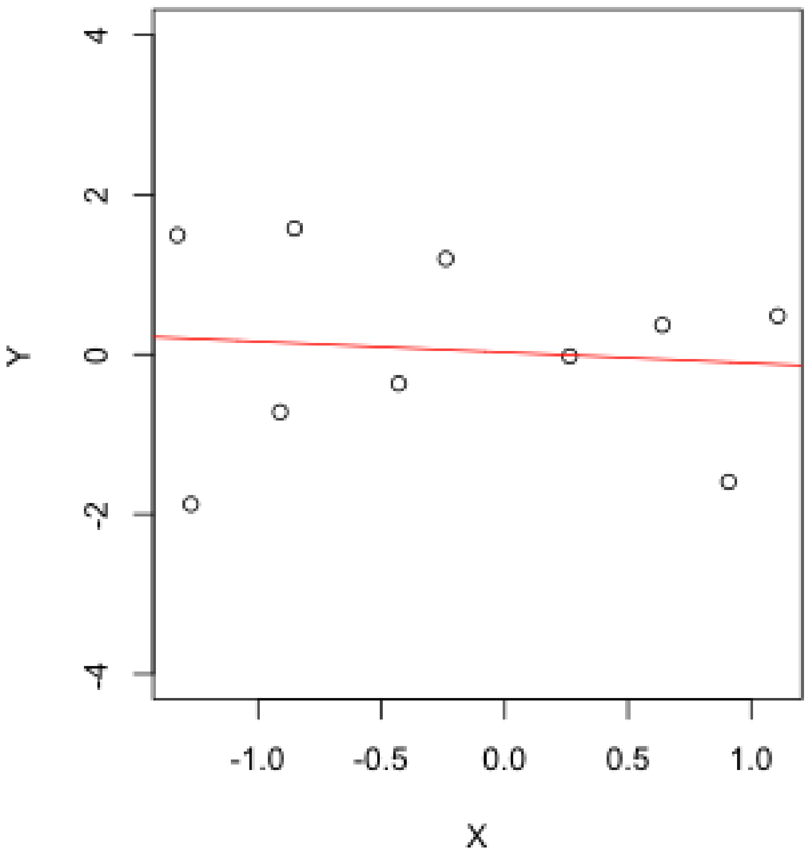


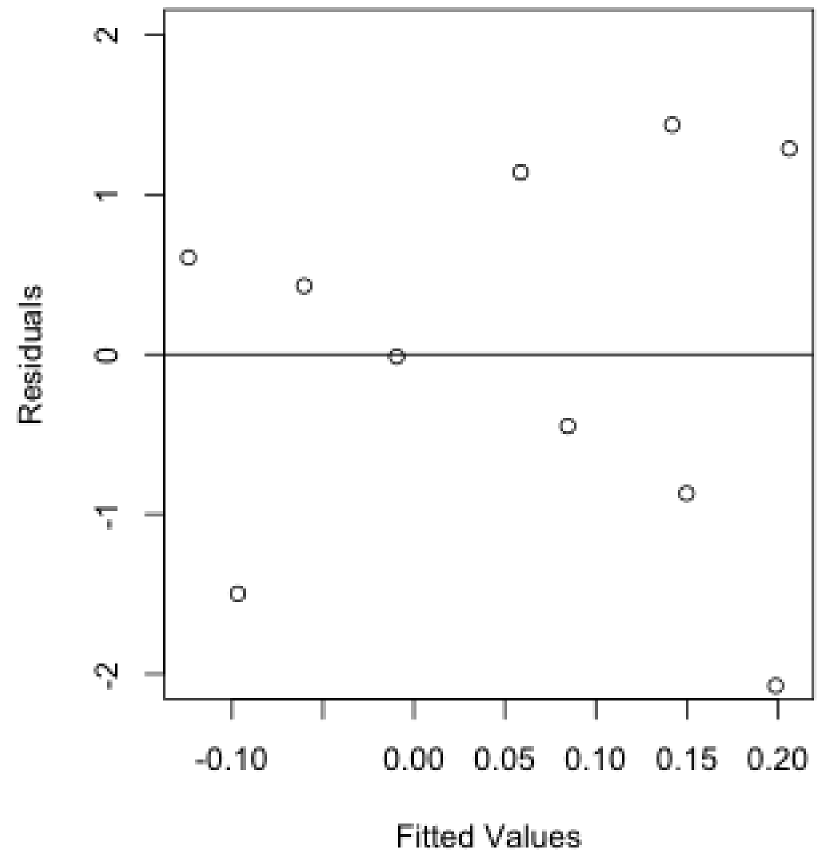
$$n = 10$$

$$\text{Adjusted } R^2 = 1 - \frac{n-1}{n-(K+1)} = 0$$

$$E(Y \mid X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + + \beta_K X^K$$

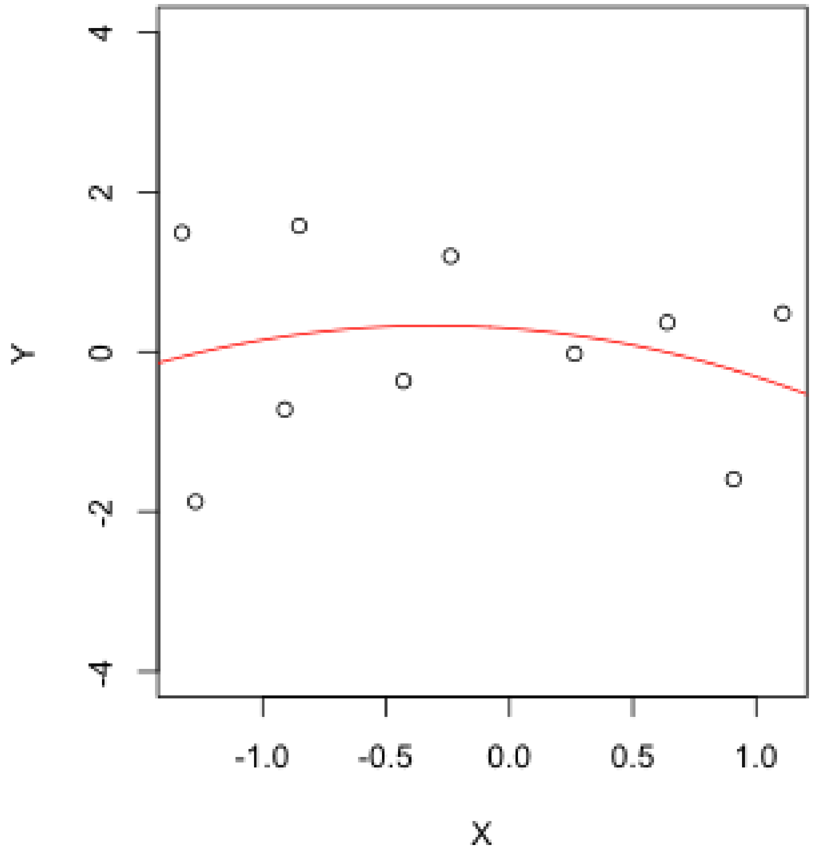
$$R^2 = 0.01, K = 1, \text{ Adjusted } R^2 = -0.114$$

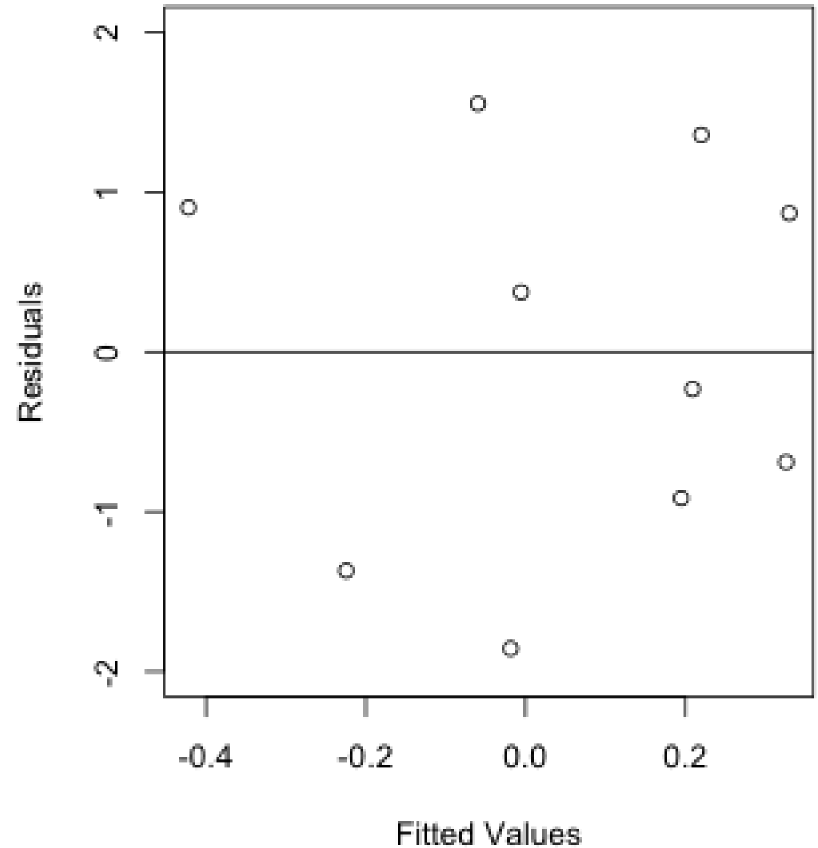




$$E(Y \mid X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + + \beta_K X^K$$

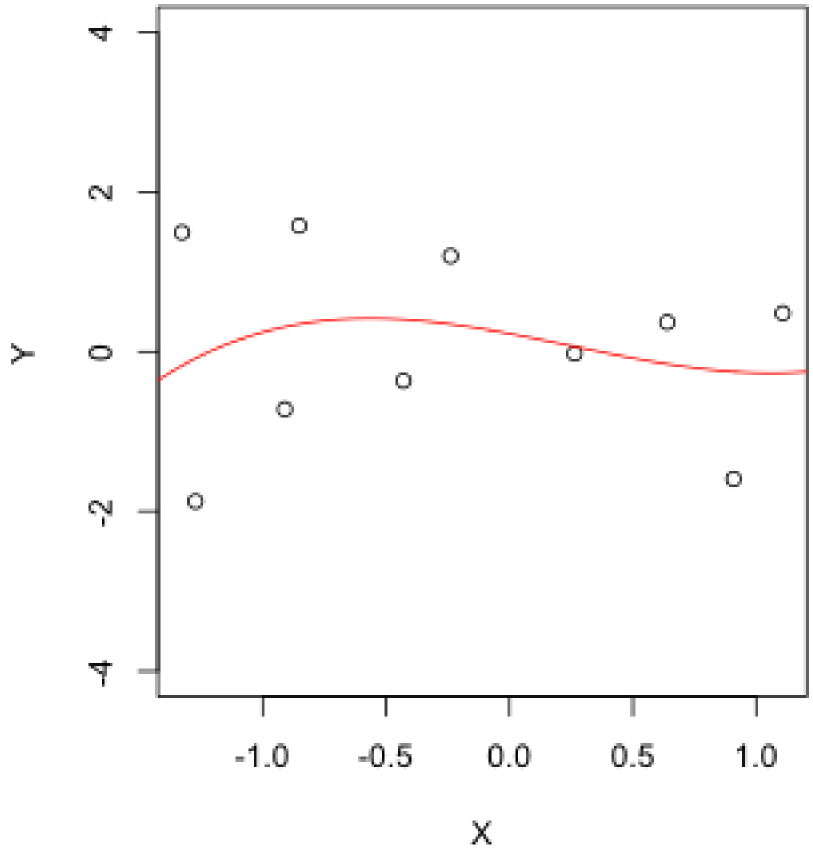
$$R^2 = 0.041, K = 2, \text{ Adjusted } R^2 = -0.131$$

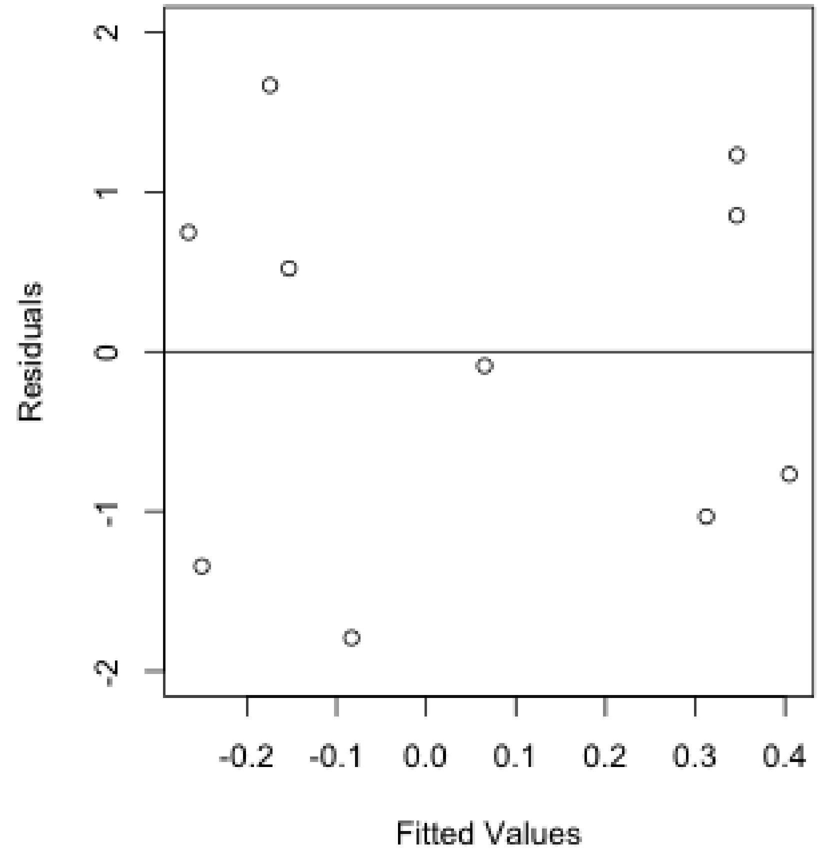




$$E(Y \mid X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + + \beta_K X^K$$

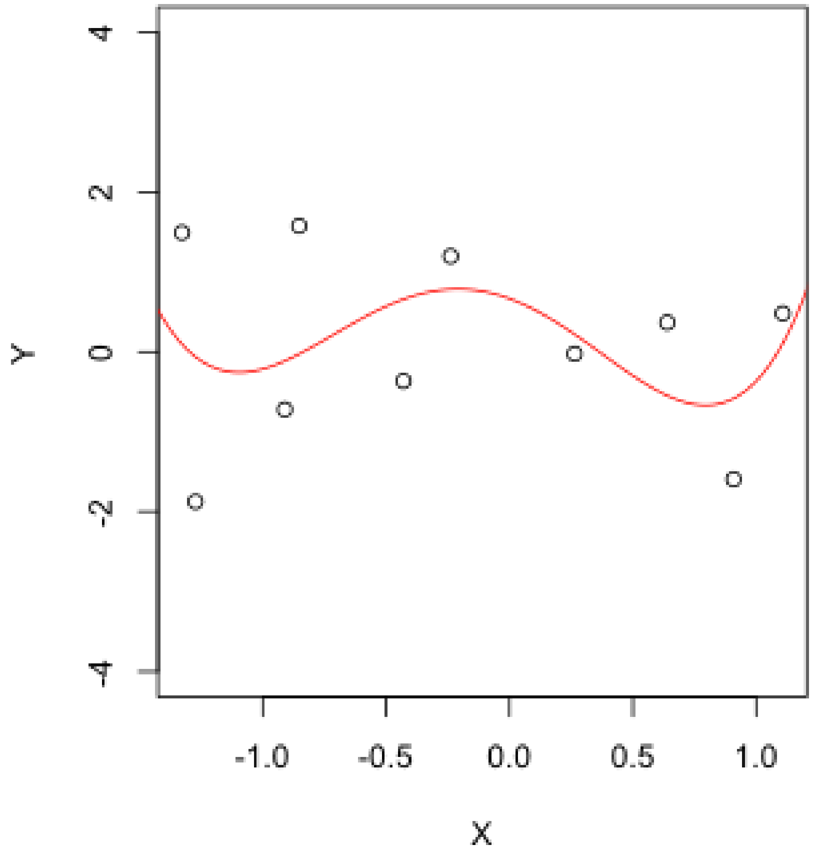
$$R^2 = 0.051, K = 3, \text{ Adjusted } R^2 = -0.424$$

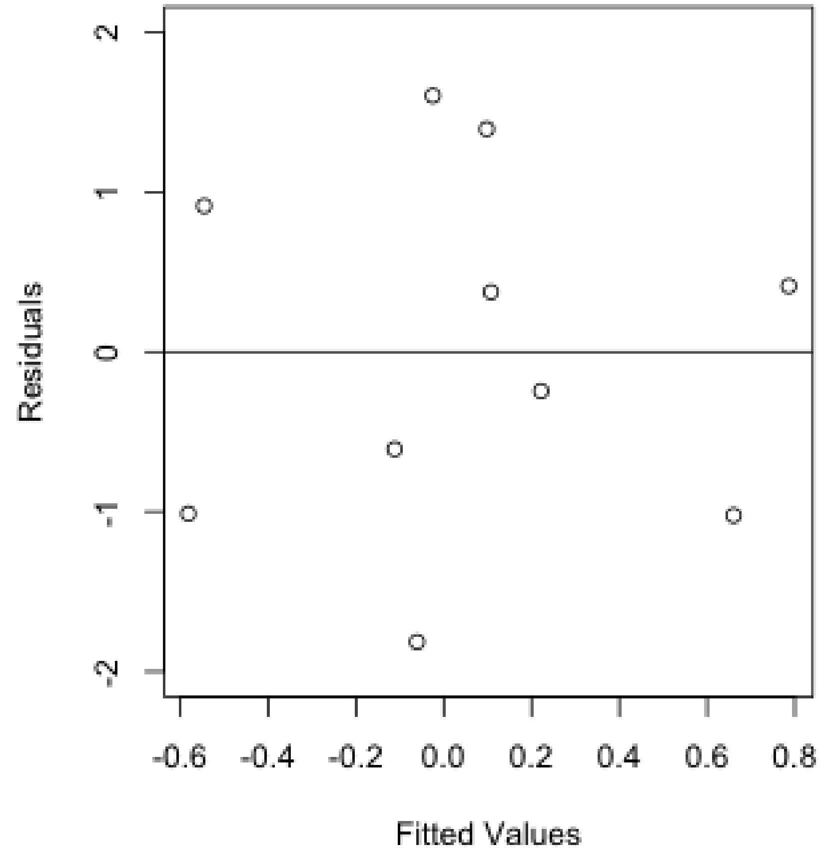




$$E(Y \mid X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + + \beta_K X^K$$

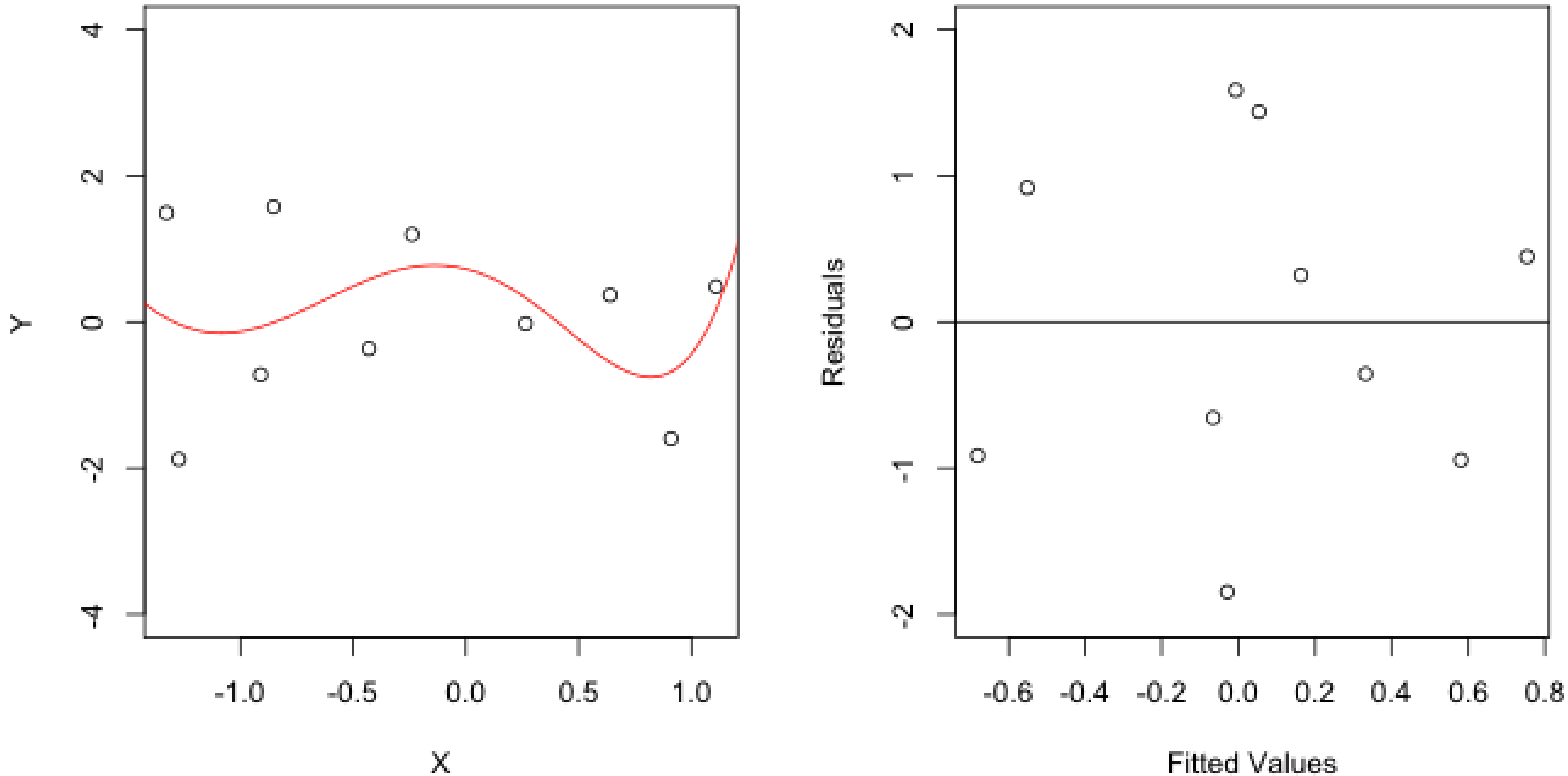
$$R^2 = 0.132, K = 4, \ \text{Adjusted } R^2 = -0.562$$





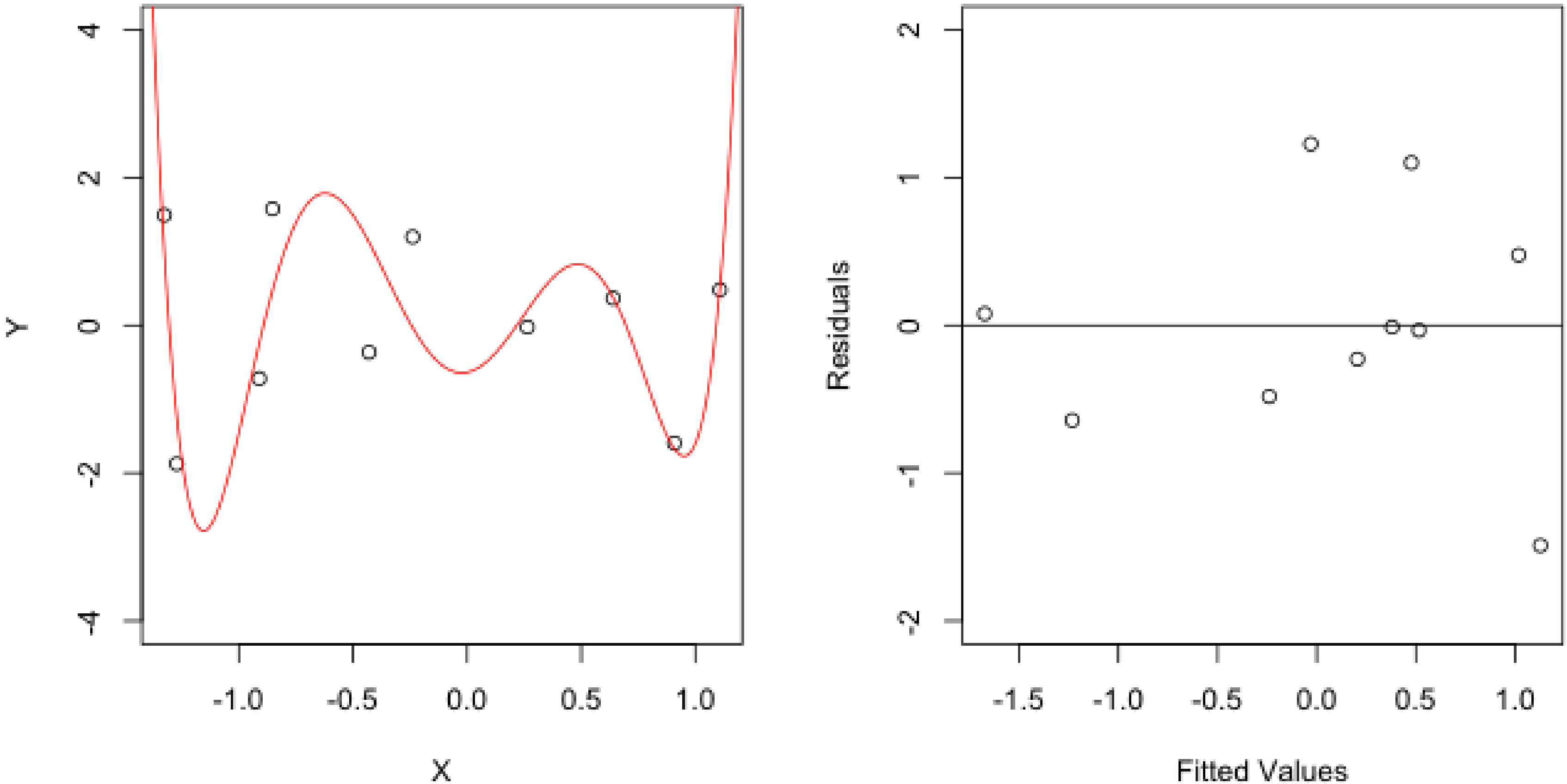
$$E(Y \mid X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + + \beta_K X^K$$

$$R^2 = 0.135, K = 5, \text{ Adjusted } R^2 = -0.946$$



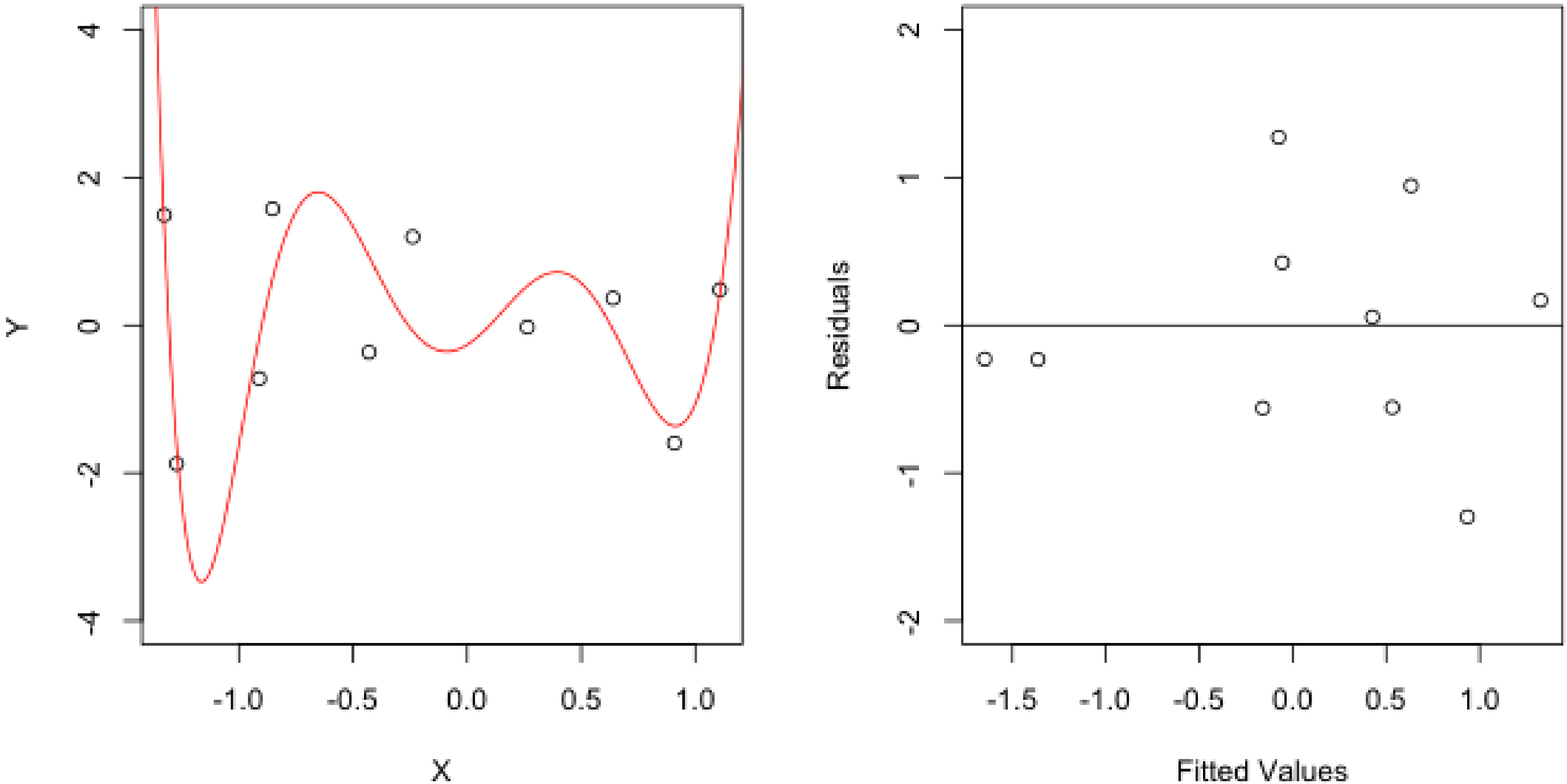
$$E(Y \mid X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + + \beta_K X^K$$

$$R^2 = 0.555, K = 6, \text{ Adjusted } R^2 = -0.335$$



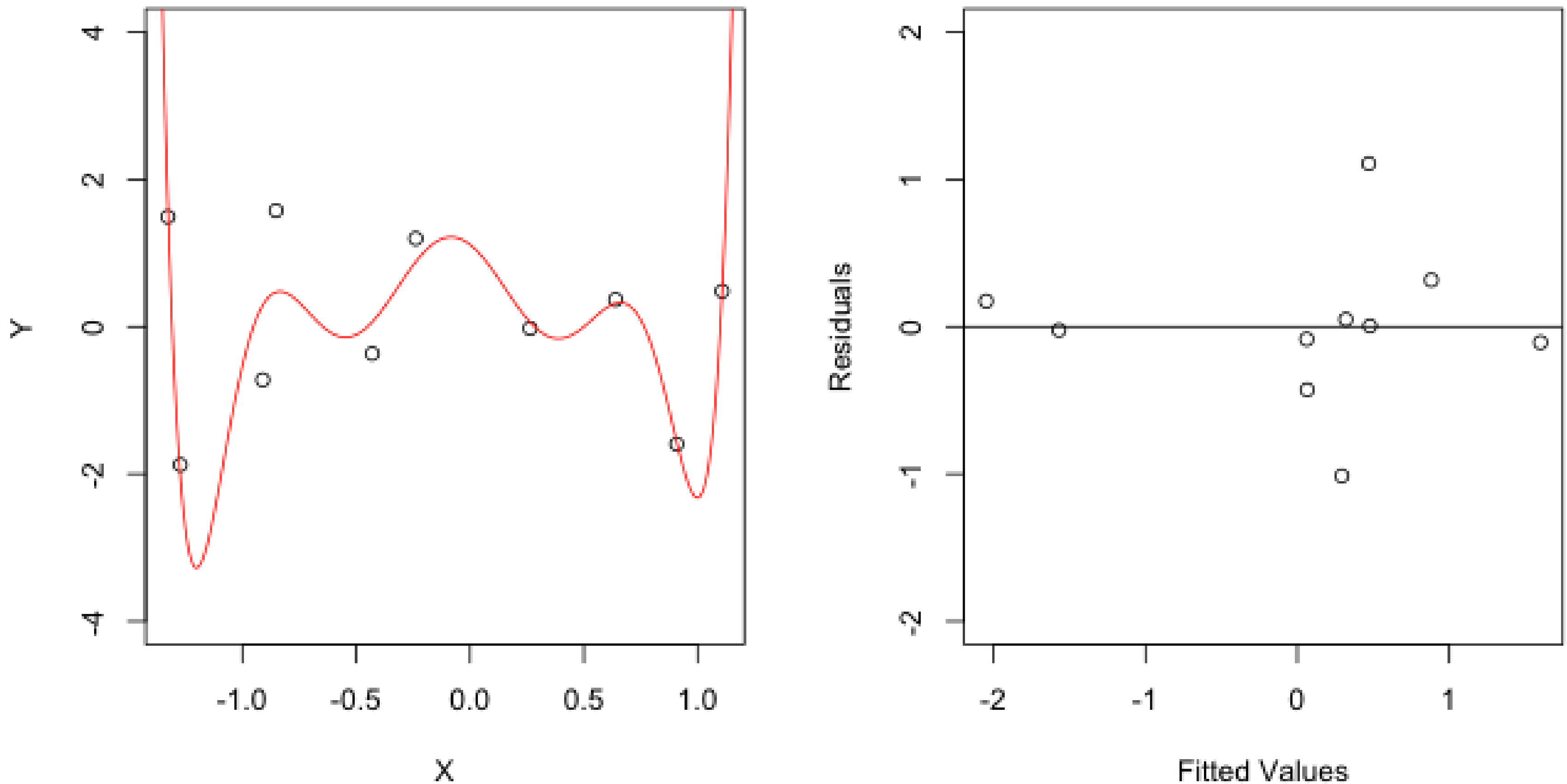
$$E(Y \mid X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + + \beta_K X^K$$

$$R^2 = 0.611, K = 7, \ \text{Adjusted } R^2 = -0.751$$



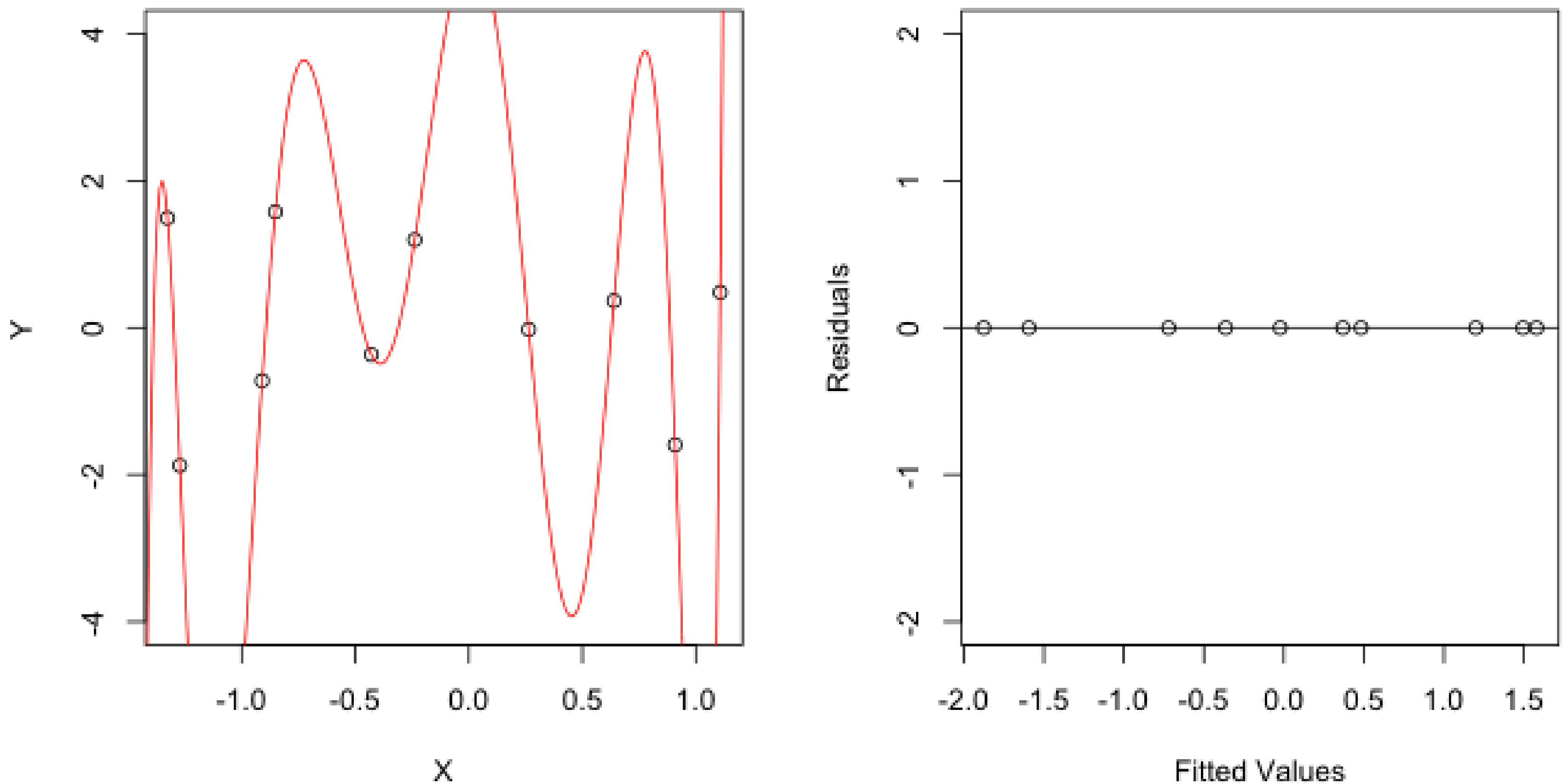
$$E(Y \mid X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + + \beta_K X^K$$

$$R^2 = 0.804, K = 8, \ \text{Adjusted } R^2 = -0.765$$



$$E(Y \mid X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + + \beta_K X^K$$

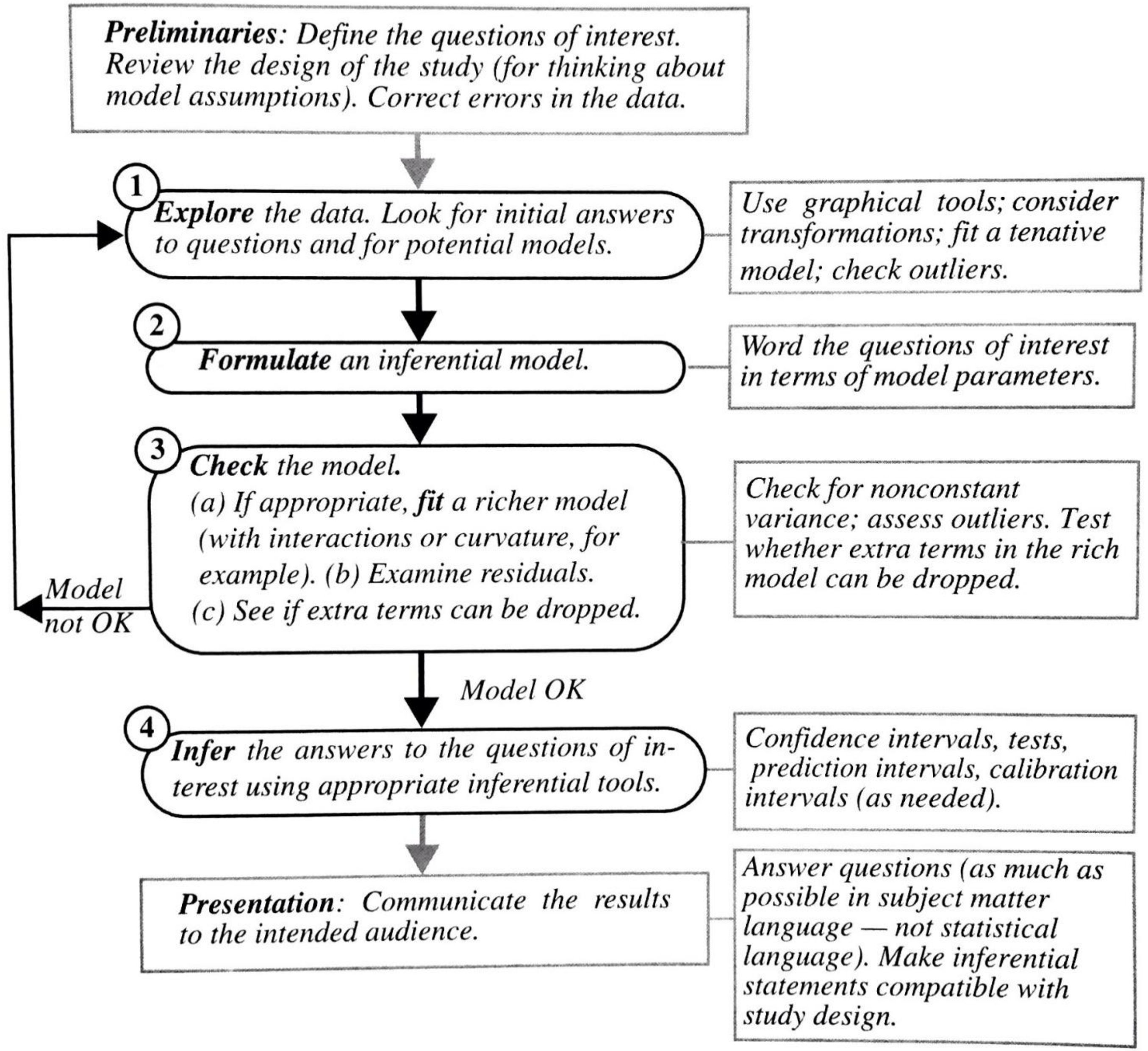
$$R^2 = 1, K = 9, \text{ Adjusted } R^2 = NA$$



Best model : $E(Y \mid X) = \beta_0$, $K = 0$ and adjusted $R^2 = 0$

Strategies for Variable Selection

Display 9.9 A strategy for data analysis using statistical models



Strategies for Variable Selection

1. Identify key objectives.
2. Screen variables and indentify the ones that are sensitive to the objectives, exclude redundancies.
3. Exploratory analysis: graphical displays and correlation coefficients.
   - Apply transformations, if necessary.
4. Fit a rich model and perform model check: residual plot, QQplot, consider outliers.
5. Simplify the model without loosing too much of the initial explanatory qualities.
   - Possibly, perform automatic variable selection.
   - Perform cross-validation – set aside a portion of the data set to check the model.
6. Finalize the model and proceed with analysis.

Strategies for Variable Selection

Depend on our objective:

1. Adjusting for auxiliary explanatory variables prior to inclusion of the main variable of interest (sometimes, for the purpose of causal inference):
   - Ok to use variable selection if causal inference is not required (e.g., sex discrimination case in Ch. 12);
   - Otherwise, need more advanced techniques (e.g., subclassification or matching).

2. Looking for prediction or best set of predictors, risk-factors.
   - No interpretation needed;
   - Ok to use variable selection techniques.