Econometrics 1

Lecture 8: Nonlinear Regression Functions

黄嘉平

中国经济特区研究中心 讲师

办公室：文科楼2613

E-mail: huangjp@szu.edu.cn

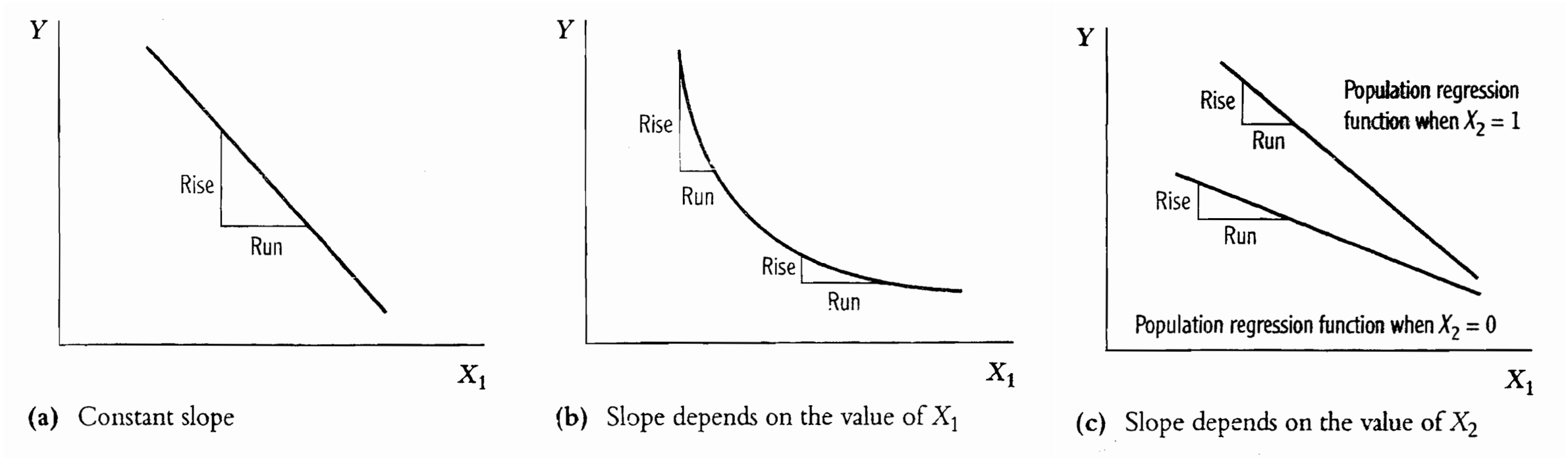
Tel: (0755) 2695 0548

Website: https://huangjp.com

Linearity in coefficients

- The multiple regression models are linear functions of unknown coefficients (or parameters) of the population regression model, but not necessarily linear functions of independent variables.

- Population regression functions with different slopes:

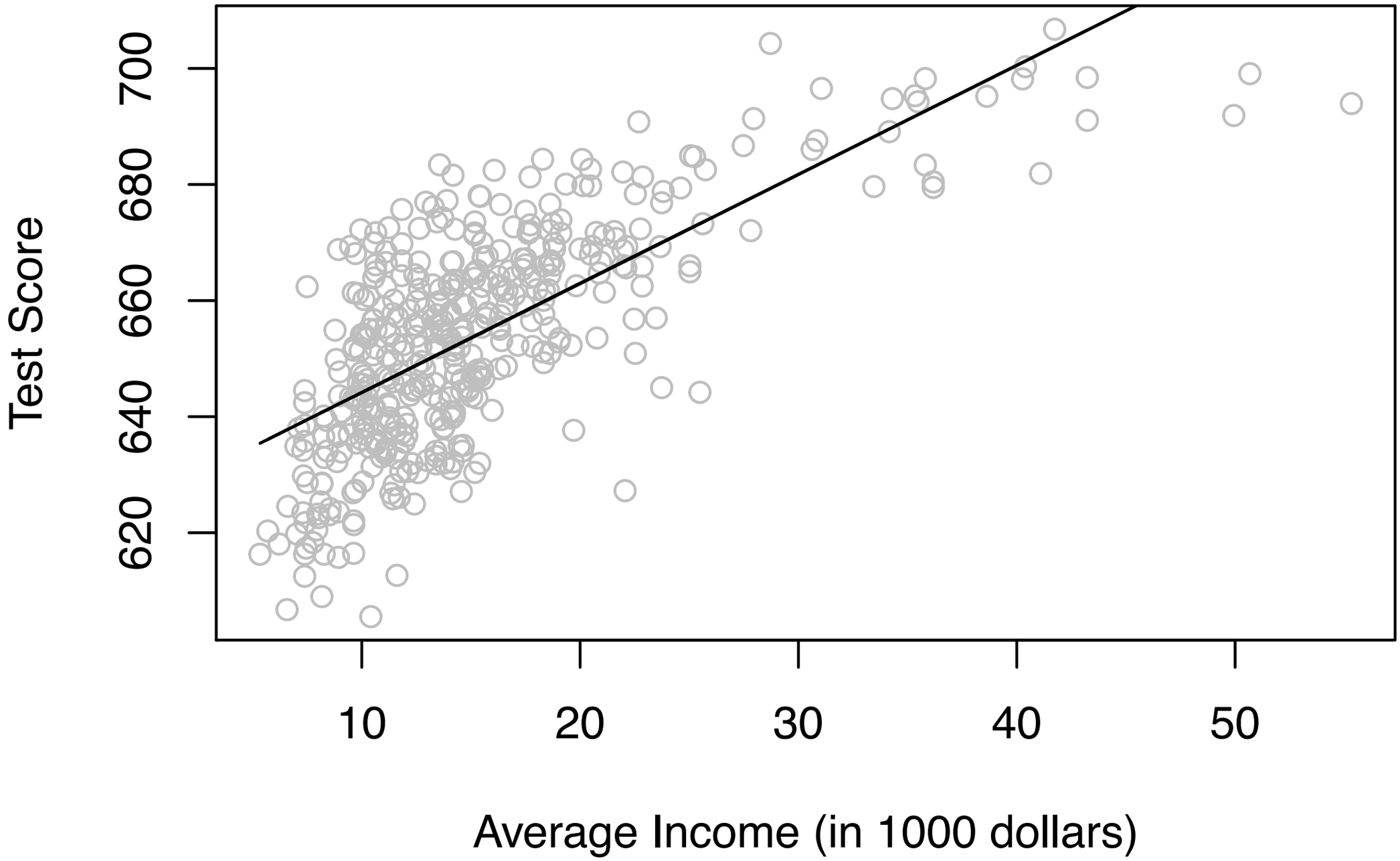


linear function

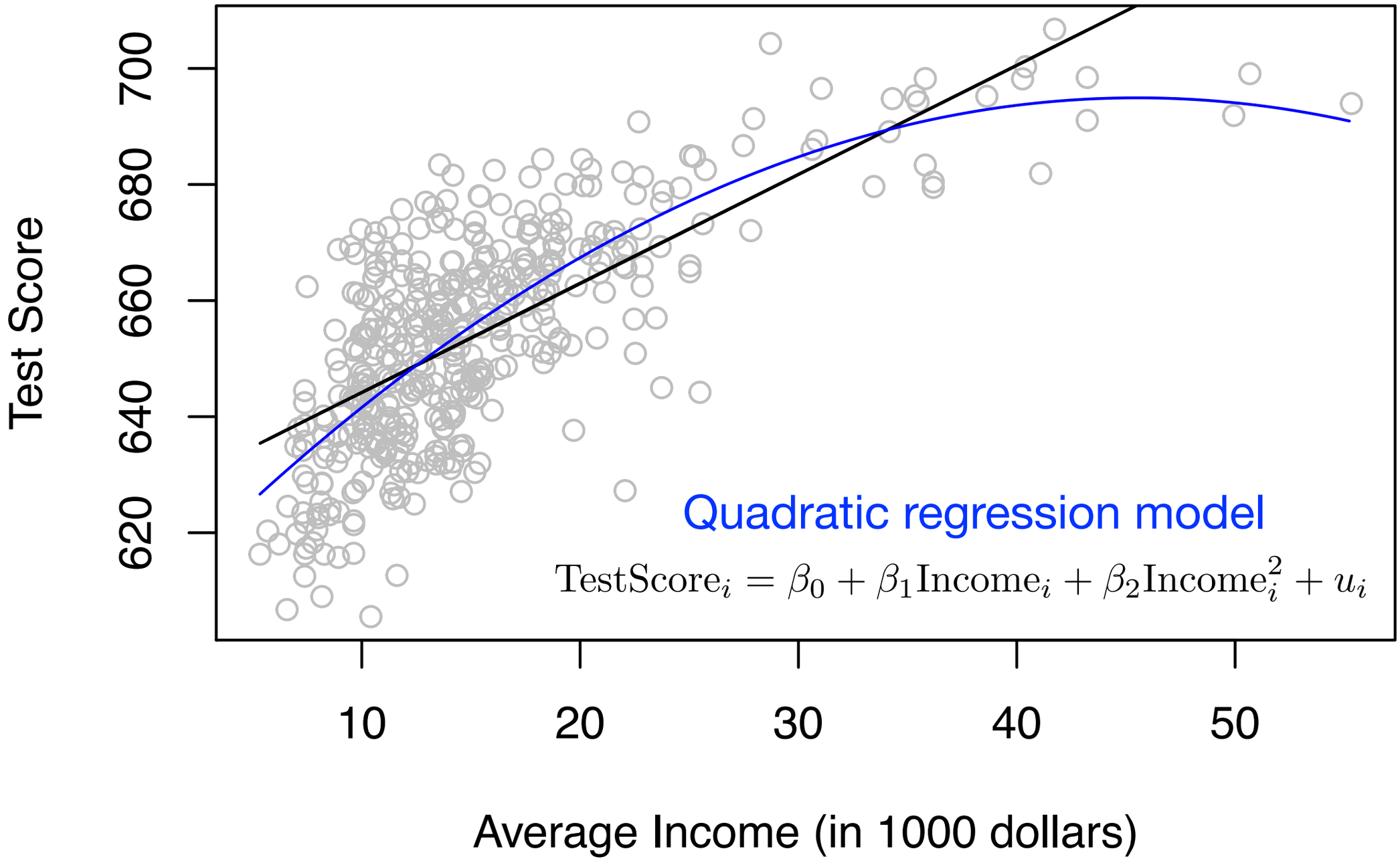
nonlinear functions

A General Strategy for Modeling Nonlinear Regression Functions

Average income vs. test score



Average income vs. test score



The quadratic regression model

- The population regression function

$$\mathrm{E}(\mathrm{TestScore}_i \mid \mathrm{Income}_i) = \beta_0 + \beta_1 \mathrm{Income}_i + \beta_2 \mathrm{Income}_i^2$$

- Regression outputs of the linear and quadratic models

$$\widehat{\mathrm{TestScore}} = \underset{(1.87)}{625.4} + \underset{(0.11)}{1.88}\ \mathrm{Income},\ \ \overline{R}^2 = 0.506$$

$$\widehat{\mathrm{TestScore}} = \underset{(2.90)}{607.3} + \underset{(0.27)}{3.85}\ \mathrm{Income} - \underset{(0.0048)}{0.042}\ \mathrm{Income}^2,\ \ \overline{R}^2 = 0.554$$

General form of nonlinear regression function

- The nonlinear population regression models* are of the form

$$Y_i = f(X_{1i}, X_{2i}, \ldots, X_{ki}) + u_i, \quad i = 1, \ldots, n$$

where $f(X_{1i}, X_{2i}, \ldots, X_{ki})$ is the population **nonlinear regression function**.

* There are other forms of nonlinear regression function, see Appendix 8.1.

The effect on Y of a change in X_k

- When the value X_k is changed to $X_k + \Delta X_k$, the change of Y is

$$\Delta Y = f(X_1, \ldots, X_{k-1}, X_k + \Delta X_k, X_{k+1}, \ldots, X_m) - f(X_1, \ldots, X_{k-1}, X_k, X_{k+1}, \ldots, X_m)$$

- Suppose in our TestScore-Income model, the Income is increased from 10 to 11, then the change of TestScore is

$$(\hat{\beta}_0 + \hat{\beta}_1 \times 11 + \hat{\beta}_2 \times 11^2) - (\hat{\beta}_0 + \hat{\beta}_1 \times 10 + \hat{\beta}_2 \times 10^2)$$
$$= \hat{\beta}_1 + 21\hat{\beta}_2$$

A general approach to modeling nonlinearities using multiple regression

1. Identify a possible nonlinear relationship.

2. Specify a nonlinear function and estimate its parameters by OLS.

3. Determine whether the nonlinear model improves upon a linear model.

   Test the null hypothesis that the population regression function is linear against the alternative that it is nonlinear.

4. Plot the estimated nonlinear regression function.

5. Estimate the effect on Y of a change in X.

Nonlinear Functions of a Single Independent Variable

Nonlinear functions of a single independent variable

- Polynomials

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \cdots + \beta_r X_i^r + u_i$$

- Logarithms

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 \ln(X_i) + u_i \\ \ln(Y_i) &= \beta_0 + \beta_1 X_i + u_i \\ \ln(Y_i) &= \beta_0 + \beta_1 \ln(X_i) + u_i \end{aligned}$$

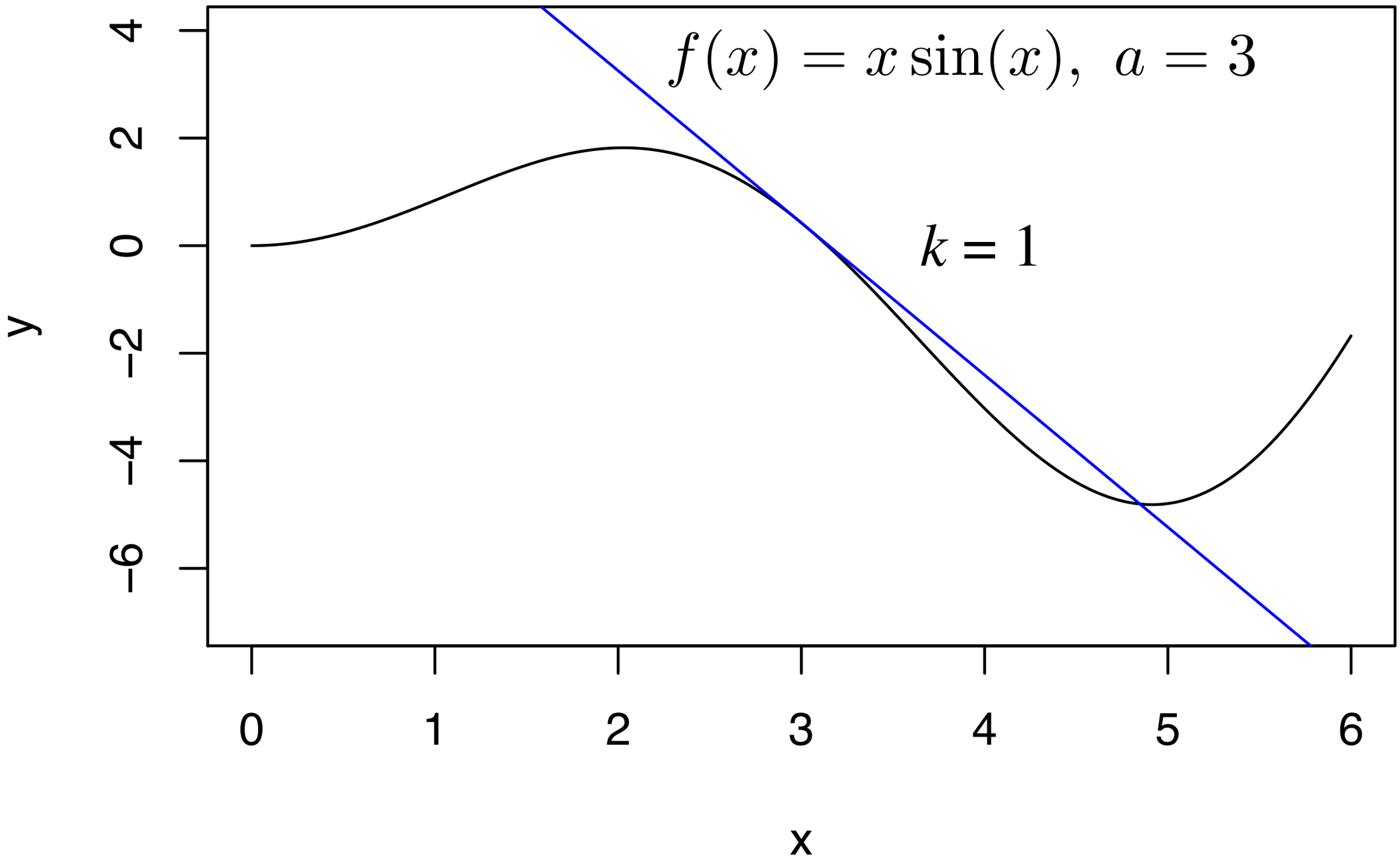
Polynomials

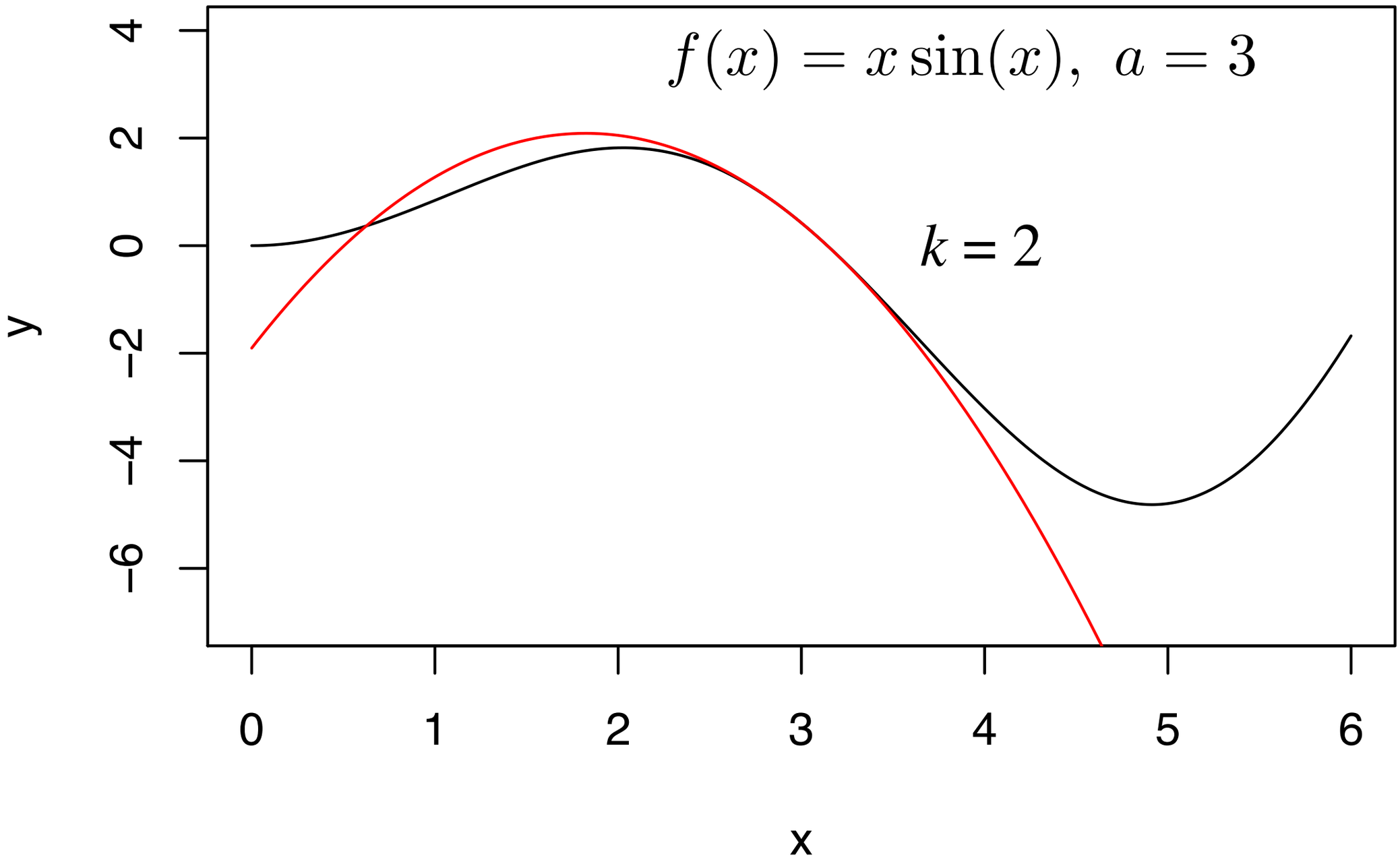
- Why polynomials?

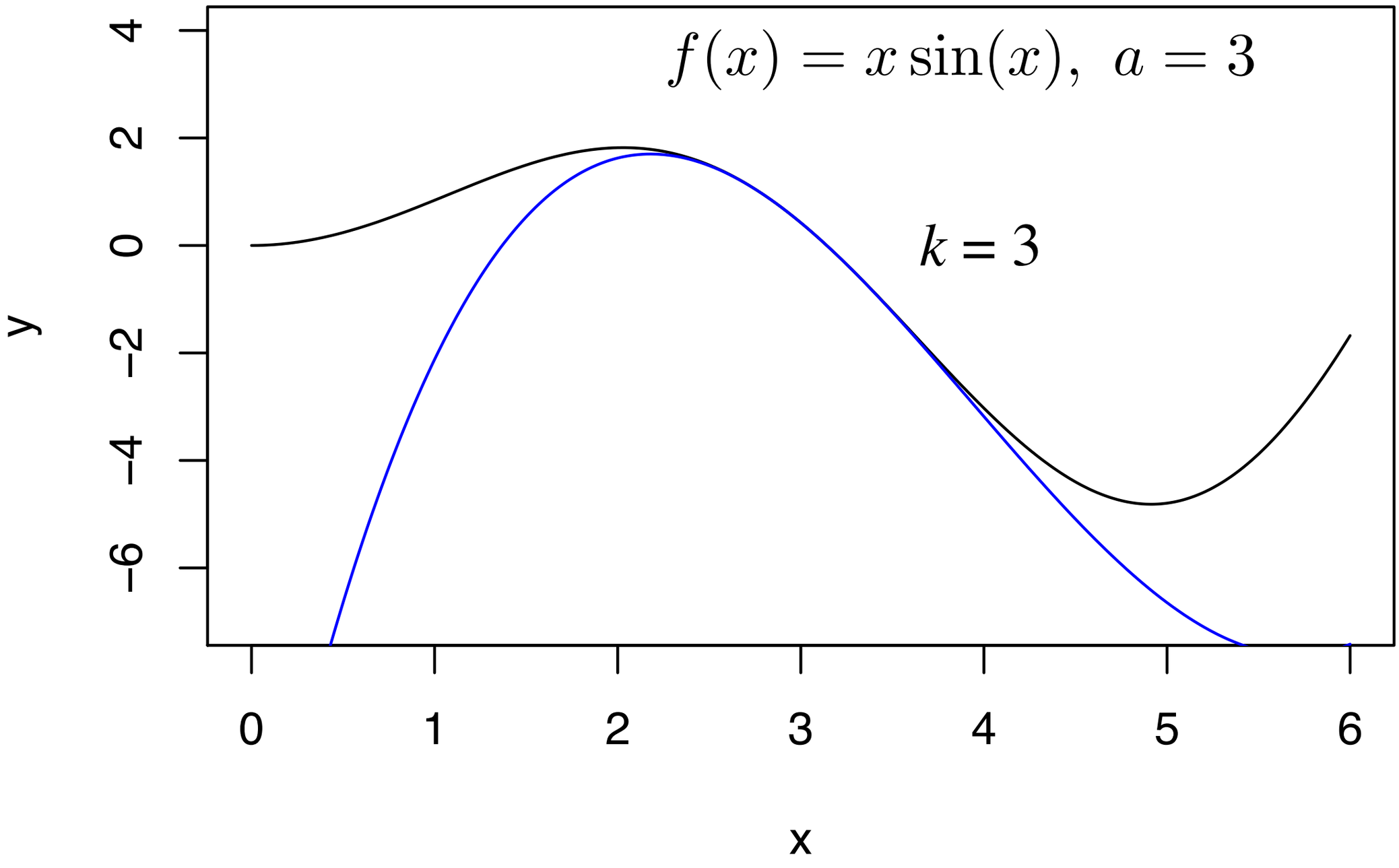
- Taylor series expansion of a smooth function $f(x)$ at point a:

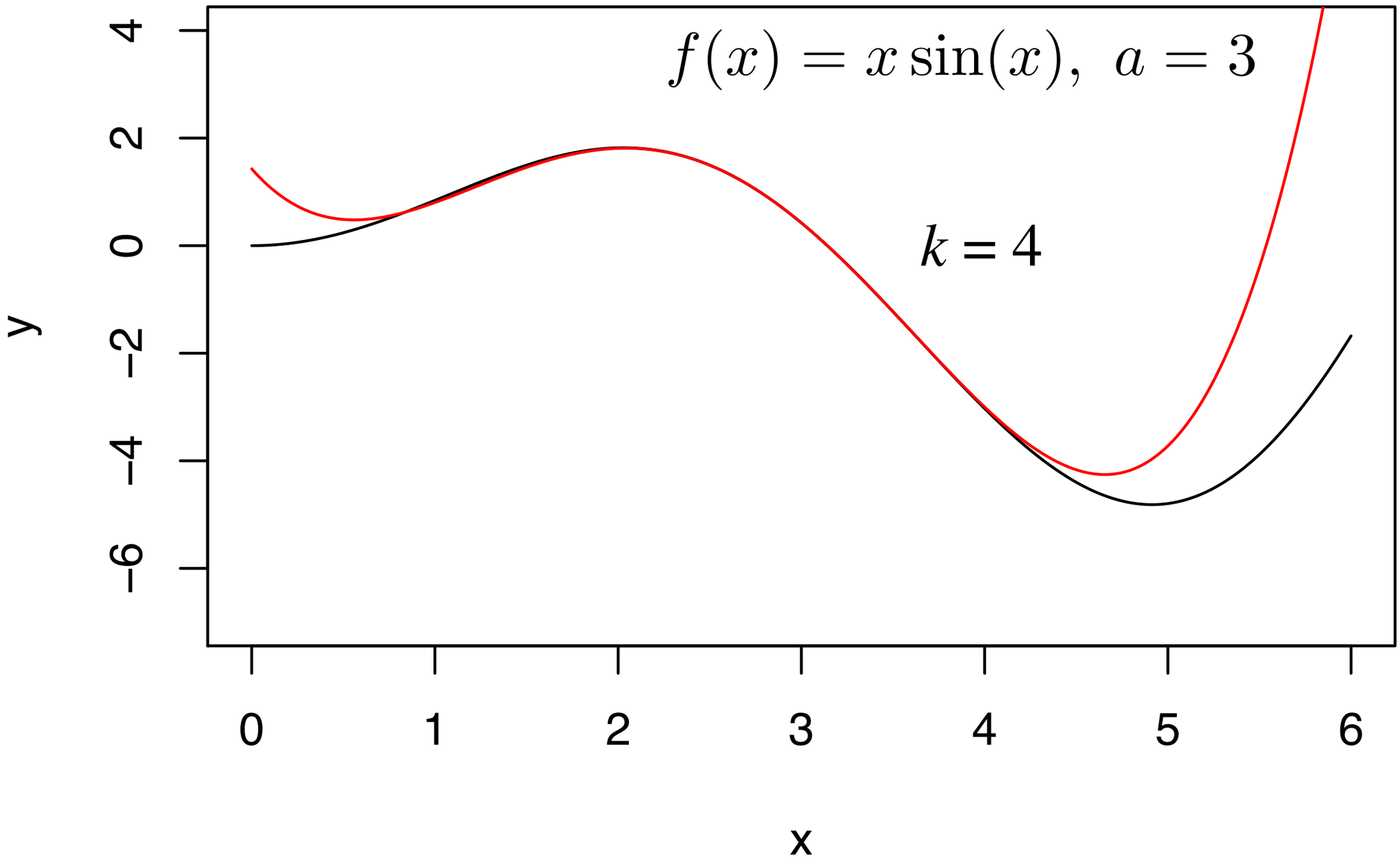
$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 + \frac{f''''(a)}{4!}(x-a)^4 + \cdots$$

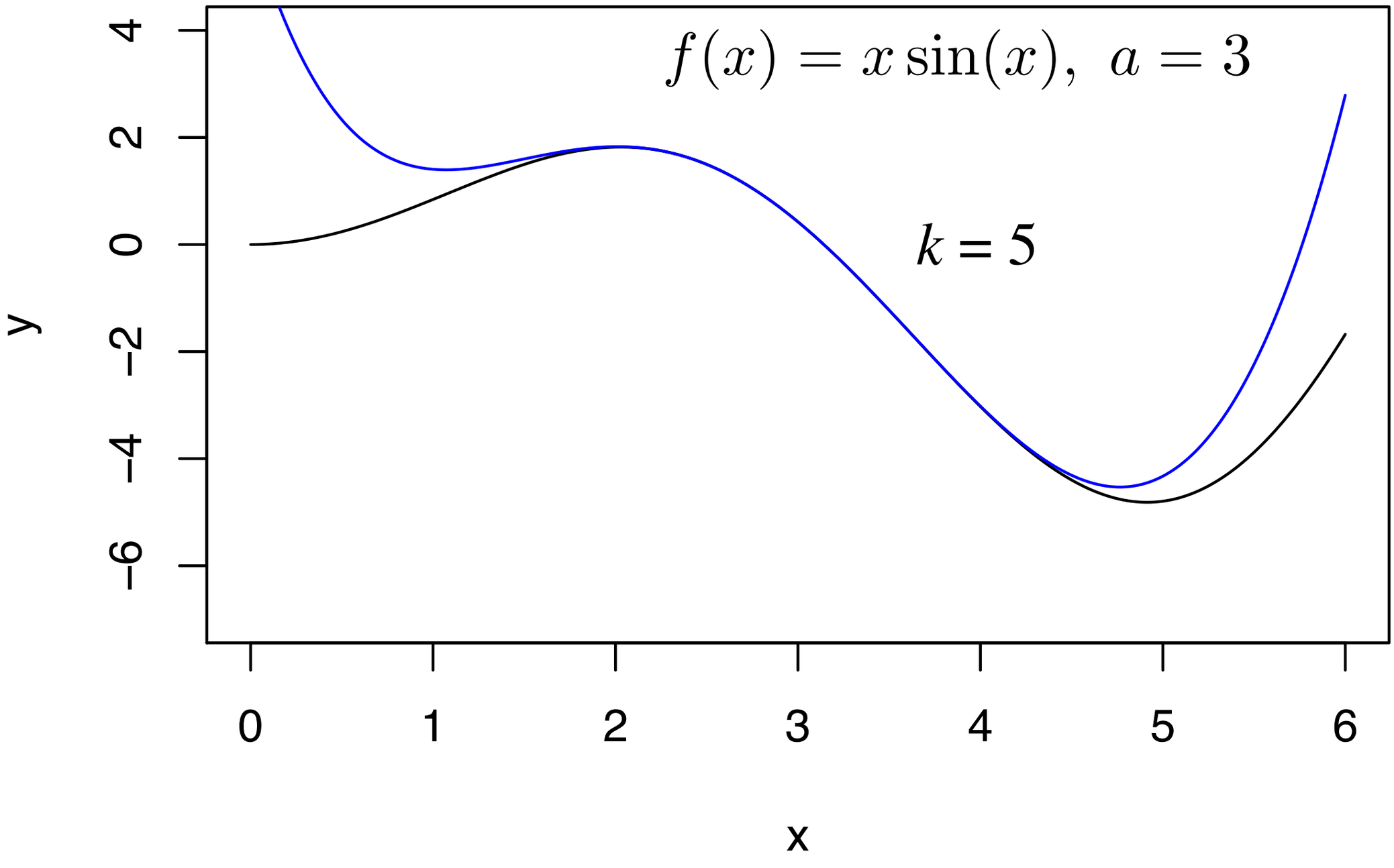
- A function can be approximated by a polynomial.

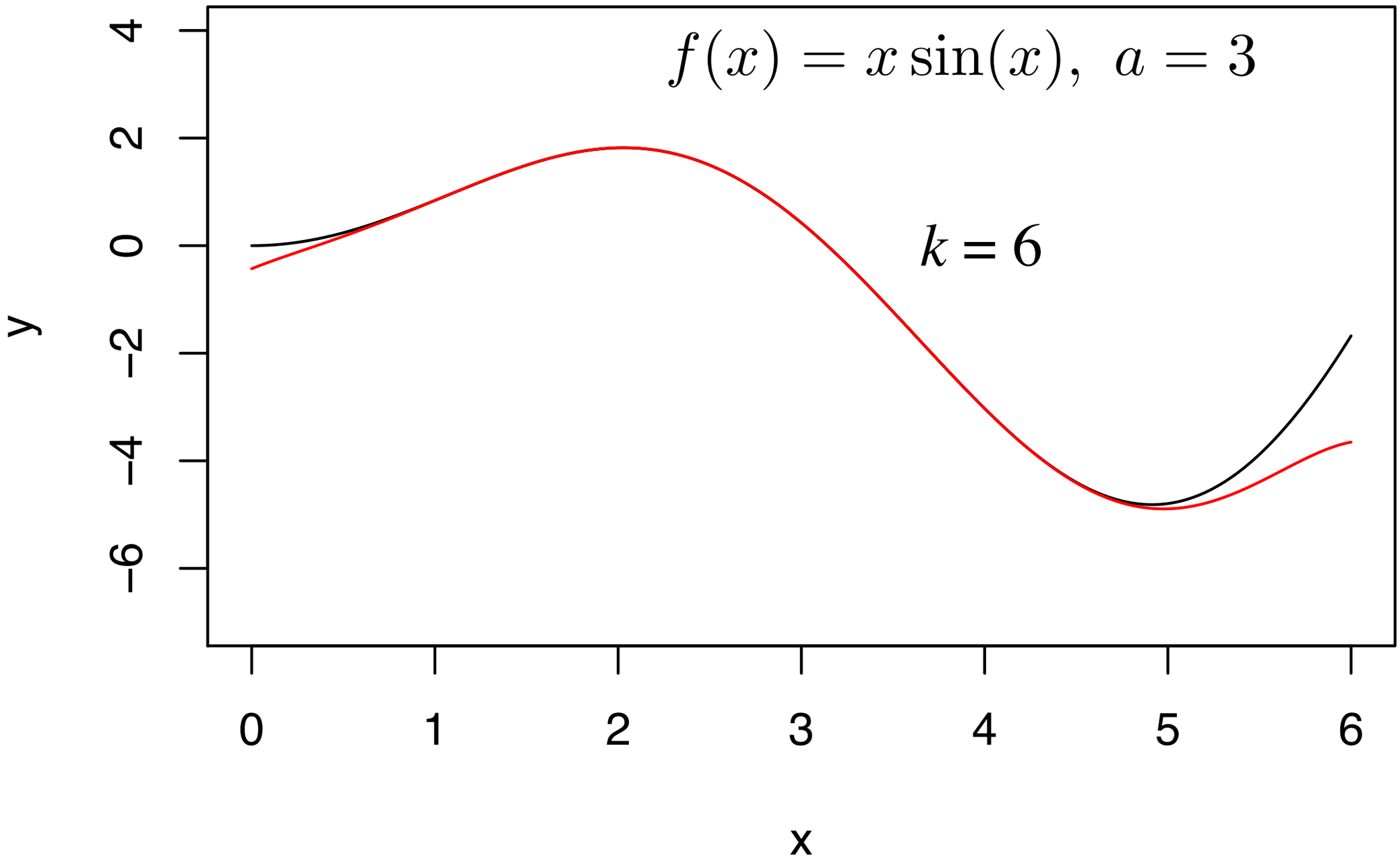

$f(x) = x\sin(x),\ a = 3$
$k = 1$
y
4
2
0
−2
−4
−6
0
1
2
3
4
5
6
x

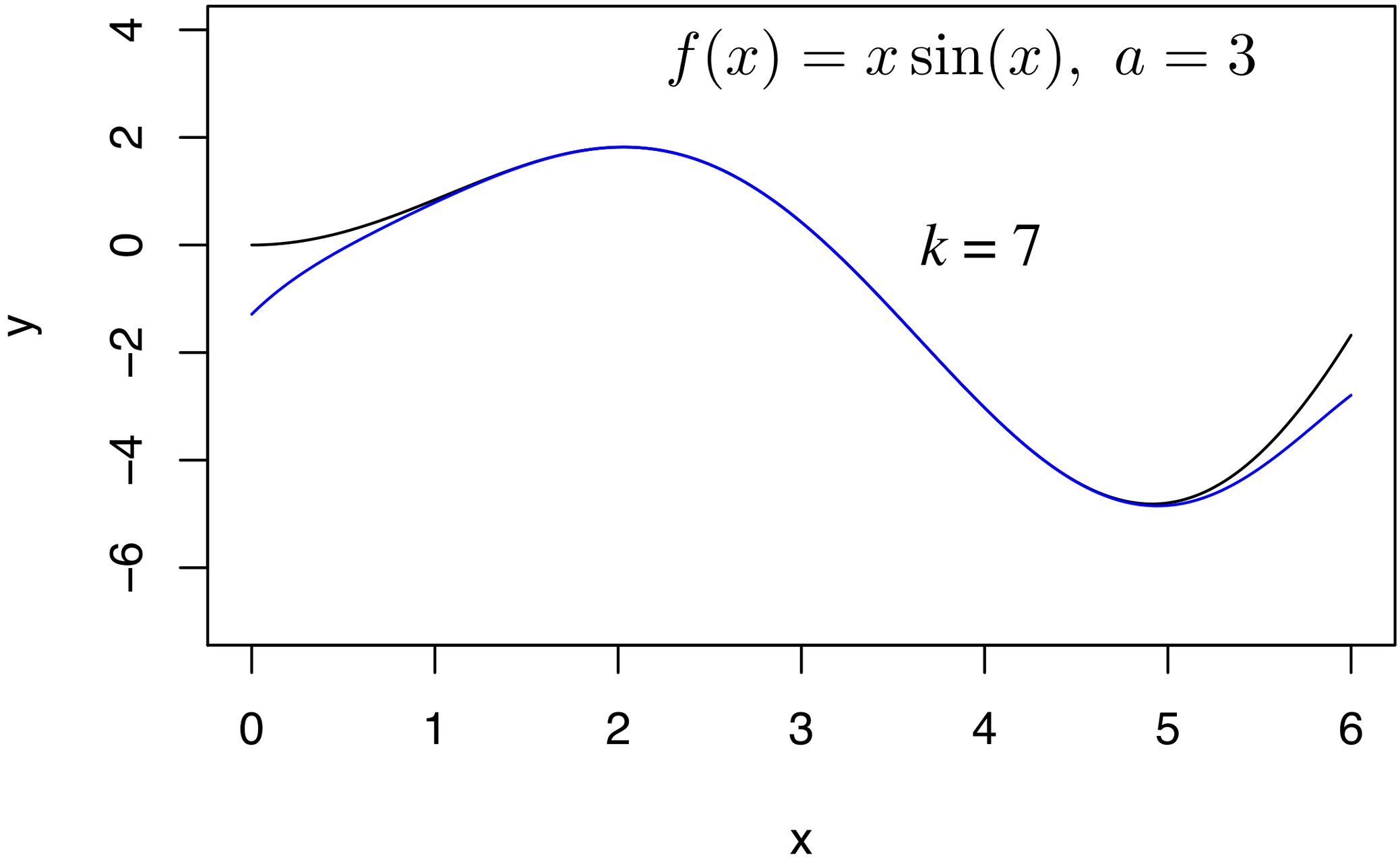

$f(x) = x\sin(x),\ a = 3$
$k = 2$
y
4
2
0
−2
−4
−6
0
1
2
3
4
5
6
x

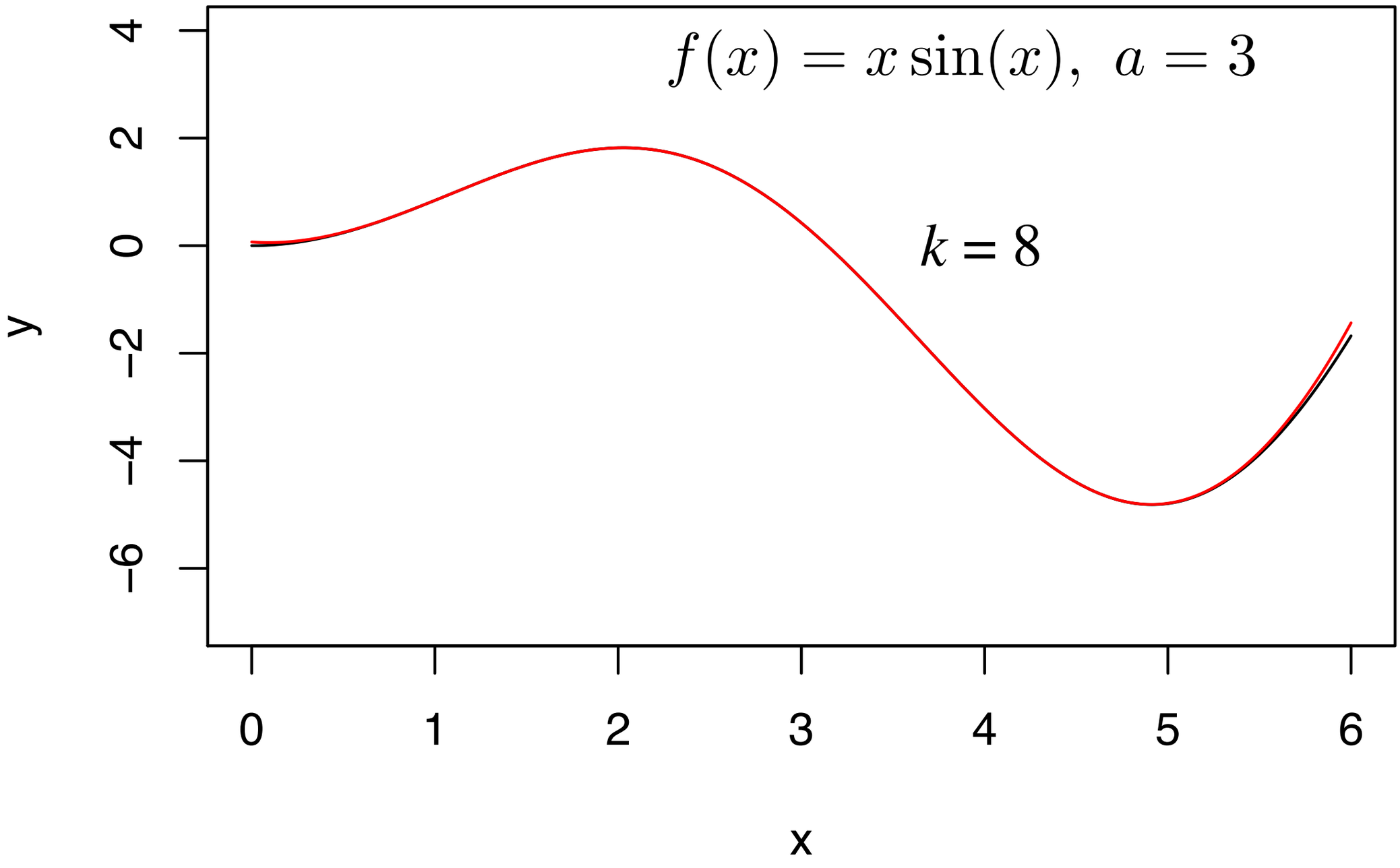

$f(x) = x\sin(x),\ a = 3$
$k = 3$
y
4
2
0
−2
−4
−6
0
1
2
3
4
5
6
x


$f(x) = x\sin(x),\ a = 3$
$k = 4$
y
4
2
0
−2
−4
−6
0
1
2
3
4
5
6
x


$f(x) = x\sin(x),\ a = 3$
$k = 5$
y
4
2
0
−2
−4
−6
0
1
2
3
4
5
6
x


$f(x) = x\sin(x),\ a = 3$
$k = 6$
y
4
2
0
−2
−4
−6
0
1
2
3
4
5
6
x


$f(x) = x\sin(x),\ a = 3$
$k = 7$
y
4
2
0
−2
−4
−6
0
1
2
3
4
5
6
x


$f(x) = x\sin(x),\ a = 3$
$k = 8$
y
4
2
0
−2
−4
−6
0
1
2
3
4
5
6
x

Practice

- Fit the following regression models

$$\text{TestScore}_i = \beta_0 + \beta_1 \text{Income}_i + \beta_2 \text{Income}_i^2 + u_i$$

$$\text{TestScore}_i = \beta_0 + \beta_1 \text{Income}_i + \beta_2 \text{Income}_i^2 + \beta_3 \text{Income}_i^3 + u_i$$

- Do you think the term Income^3 is helpful in explaining test score or not? Why?

  In gretl: add the quadratic and cubic terms, and run OLS.

A sample script

```
square avginc
series cub_avginc = avginc^3
ols testscr const avginc sq_avginc –robust
ols testscr const avginc sq_avginc cub_avginc –robust
```

The quadratic model

```
              coefficient   std. error      z      p-value
  ---------------------------------------------------------
  const       607.302       2.90175      209.3    0.0000   ***
  avginc        3.85099     0.268094      14.36   8.66e-47 ***
  sq_avginc    -0.0423085   0.00478034    -8.851  8.71e-19 ***
```

The cubic model

```
                coefficient    std. error      z      p-value
  -----------------------------------------------------------
  const         600.079        5.10206      117.6    0.0000   ***
  avginc          5.01868      0.707350       7.095  1.29e-12 ***
  sq_avginc      -0.0958052    0.0289537     -3.309   0.0009   ***
  cub_avginc      0.000685484  0.000347065    1.975   0.0483   **
```

Determine the degree of polynomial

1. Pick a maximum value of r (start with 2, 3, or 4) and estimate the polynomial regression for that r.

2. Test the hypothesis $\beta_r = 0$. If it is rejected, than X^r belongs in the regression, so use the polynomial of degree r.

3. If the hypothesis cannot be rejected in step 2, eliminate X^r from the regression and estimate a polynomial regression of degree r-1. Test whether the coefficient is zero. If rejected, than use the polynomial of degree r-1.

4. If not rejected, try r-2 …

Logarithms

- Definition $x = \ln(\exp(x))$
- Logarithms and percentages

$$\ln(x + \Delta x) - \ln(x) \approx \frac{\Delta x}{x}$$

when $\Delta x/x$ is small. For example,

$$\ln(101) - \ln(100) = 0.00995$$

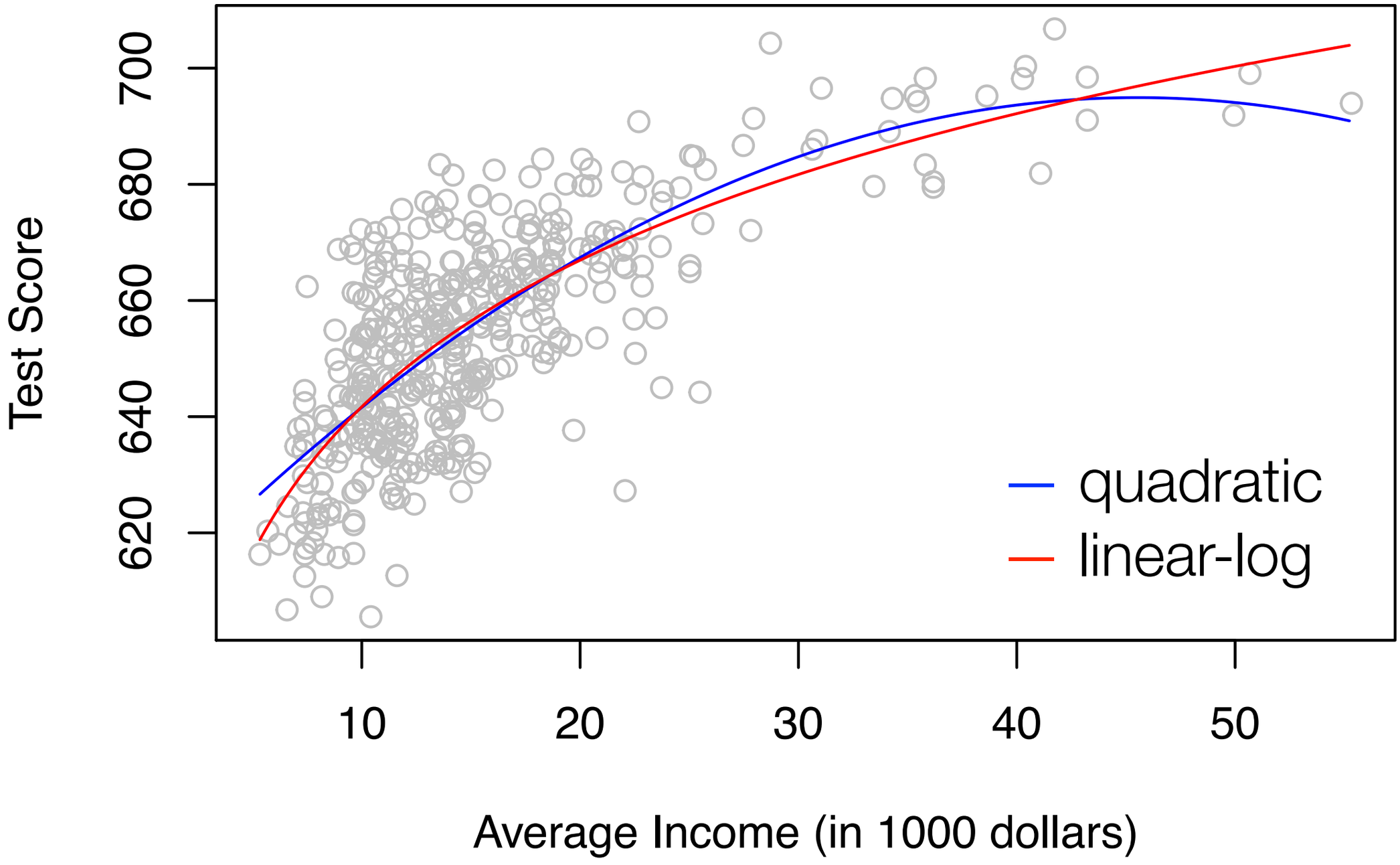
- $\Delta x/x$ is the percentage change in x divided by 100.
- Usually, changes in *price* and *wages* are expressed in logarithms.

Logarithms 1: the linear-log model

- The *linear-log* model

$$Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i$$

- In this model, a 1% change in X is associated with a change in Y of $0.01\beta_1$.


Test Score
620
640
660
680
700
10
20
30
40
50
Average Income (in 1000 dollars)
quadratic
linear-log

Logarithms 2: the log-linear model

- The *log-linear* model

$$\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i$$

- In this model, a one-unit change in X is associated with a $100 \times \beta_1\%$ change in Y.

Logarithms 3: the log-log model

- The *log-log* model

$$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i$$

- In this model, a 1% change in X is associated with a $\beta_1\%$ change in Y.

- Here, β_1 is the *elasticity* of Y with respect to X.

Practice

- Reproduce Equations (8.18), (8.23), and (8.24) in gretl.

  Hint: transform variables first.

Comparing different models

- The log-linear and log-log models can be compared using the R^2 or adjusted R^2.

- It does not make sense to compare the log-log model with the linear-log model using R^2, since the dependent variables are different. (Recall the definition of R^2)

- You should use economic theory and experts' knowledge to judge which model is better.

Interactions between independent variables

Interactions between independent variables

- Sometimes the effect on the dependent variable of one independent variable could depend on another independent variable.

- Example: Student-teacher ratio and percentage of English learners.

If the students who are still learning English benefit more from small group instruction, then the effect on test scores of a change in the student-teacher ratio would depend on the percentage of English learners.

Interactions between two binary variables

- Binary variable (dummy variable) $D_i \in \{0, 1\}$
- Regression with two binary variables

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + u_i$$

- E.g., Y: earnings, D1: college degree, D2: gender.
- Model with interaction

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 (D_{1i} \times D_{2i}) + u_i$$

Construct a binary var. from a continuous var.

- Let $HiSTR_i$ be a binary variable that equals 1 if the student-teacher ratio is 20 or more and equals 0 otherwise.

- Let $HiEL_i$ be a binary variable that equals 1 if the percentage of English learners is 10% or more and equals 0 otherwise.

- Fit the model

$$\text{TestScore} = \beta_0 + \beta_1 \text{HiSTR}_i + \beta_2 \text{HiEL}_i + \beta_3 (\text{HiSTR}_i \times \text{HiEL}_i) + u_i$$

A sample script

```
series hi_str = (str >= 20) ? 1 : 0
series hi_el = (el_pct >= 10) ? 1 : 0
series inter_hs_he = hi_str * hi_el
ols testscr const hi_str hi_el inter_hs_he –robust
```

```
Model 5: OLS, using observations 1-420
Dependent variable: testscr
Heteroskedasticity-robust standard errors, variant HC1

              coefficient   std. error      z       p-value
  ---------------------------------------------------------
  const          664.143       1.38809     478.5      0.0000   ***
  hi_str          -1.90784     1.93221      -0.9874   0.3235
  hi_el          -18.1629      2.34595      -7.742    9.77e-15 ***
  inter_hs_he     -3.49434     3.12123      -1.120    0.2629

Mean dependent var   654.1565   S.D. dependent var   19.05335
Sum squared resid    107152.8   S.E. of regression   16.04926
R-squared            0.295555   Adjusted R-squared   0.290475
F(3, 416)            60.19527   P-value(F)           2.44e-32
Log-likelihood      -1759.723   Akaike criterion     3527.446
Schwarz criterion    3543.607   Hannan-Quinn         3533.834
```

This result corresponds to Equation (8.30)

Interaction between a continuous and a binary variable

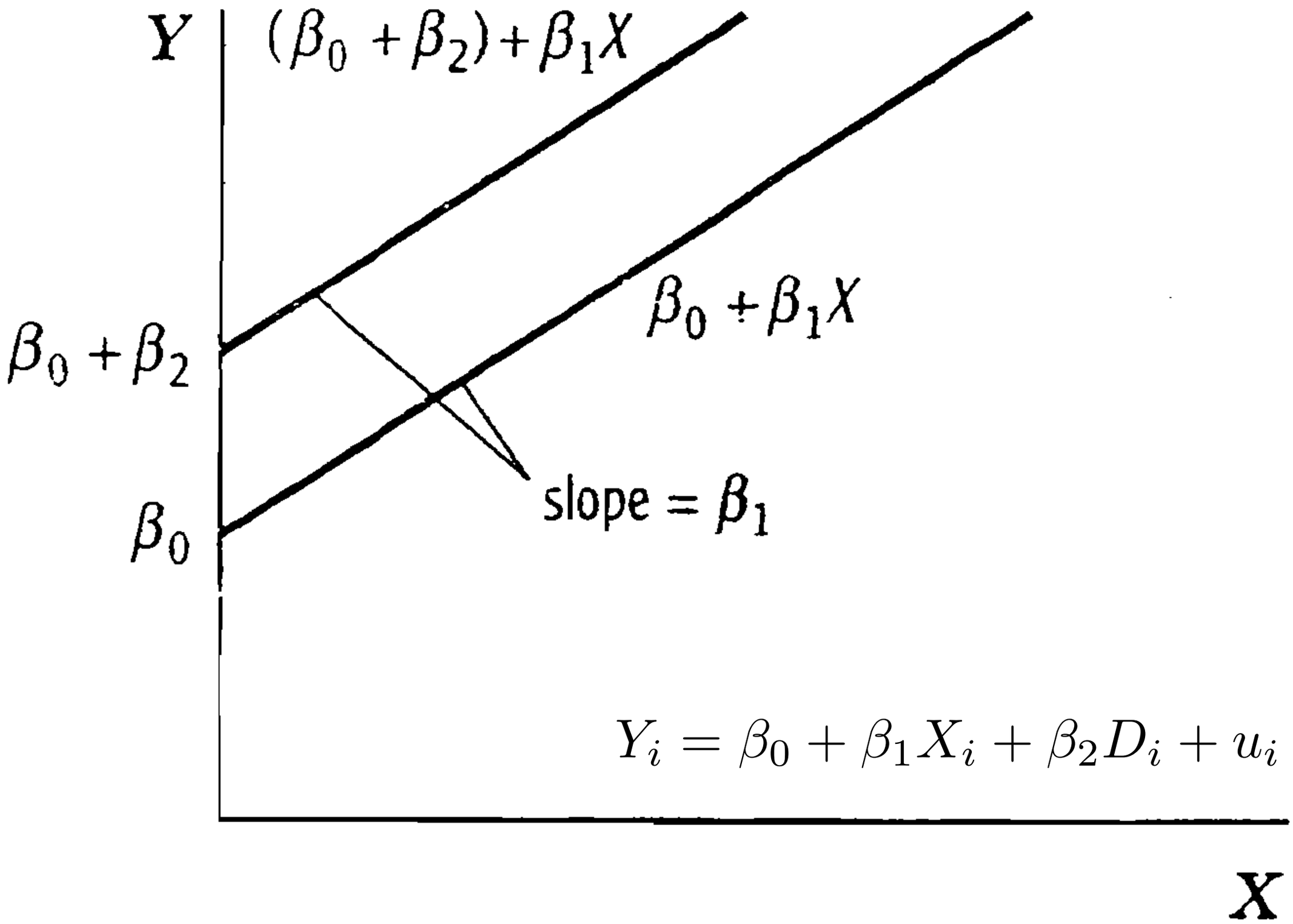
- Model without interaction

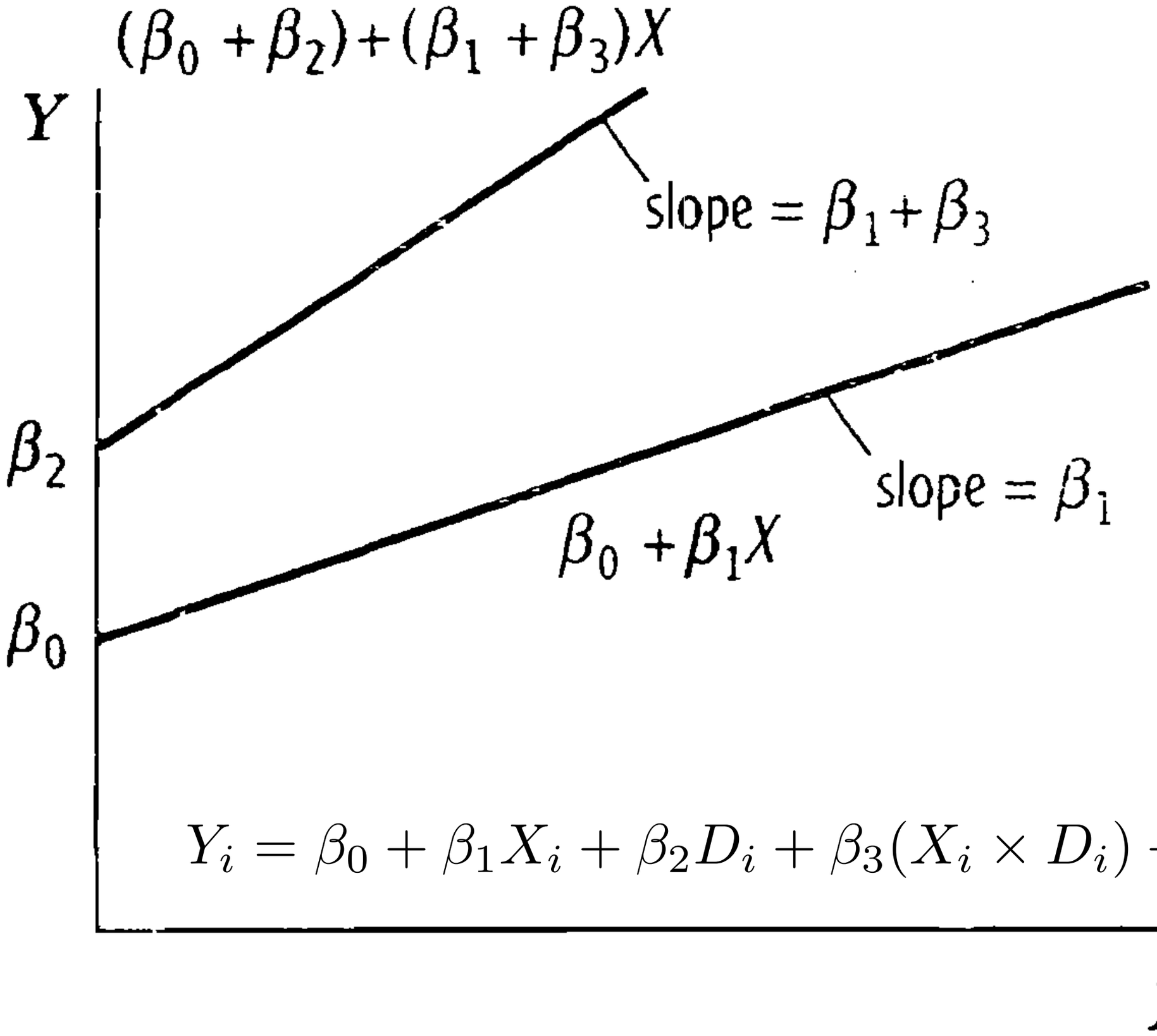
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + u_i$$

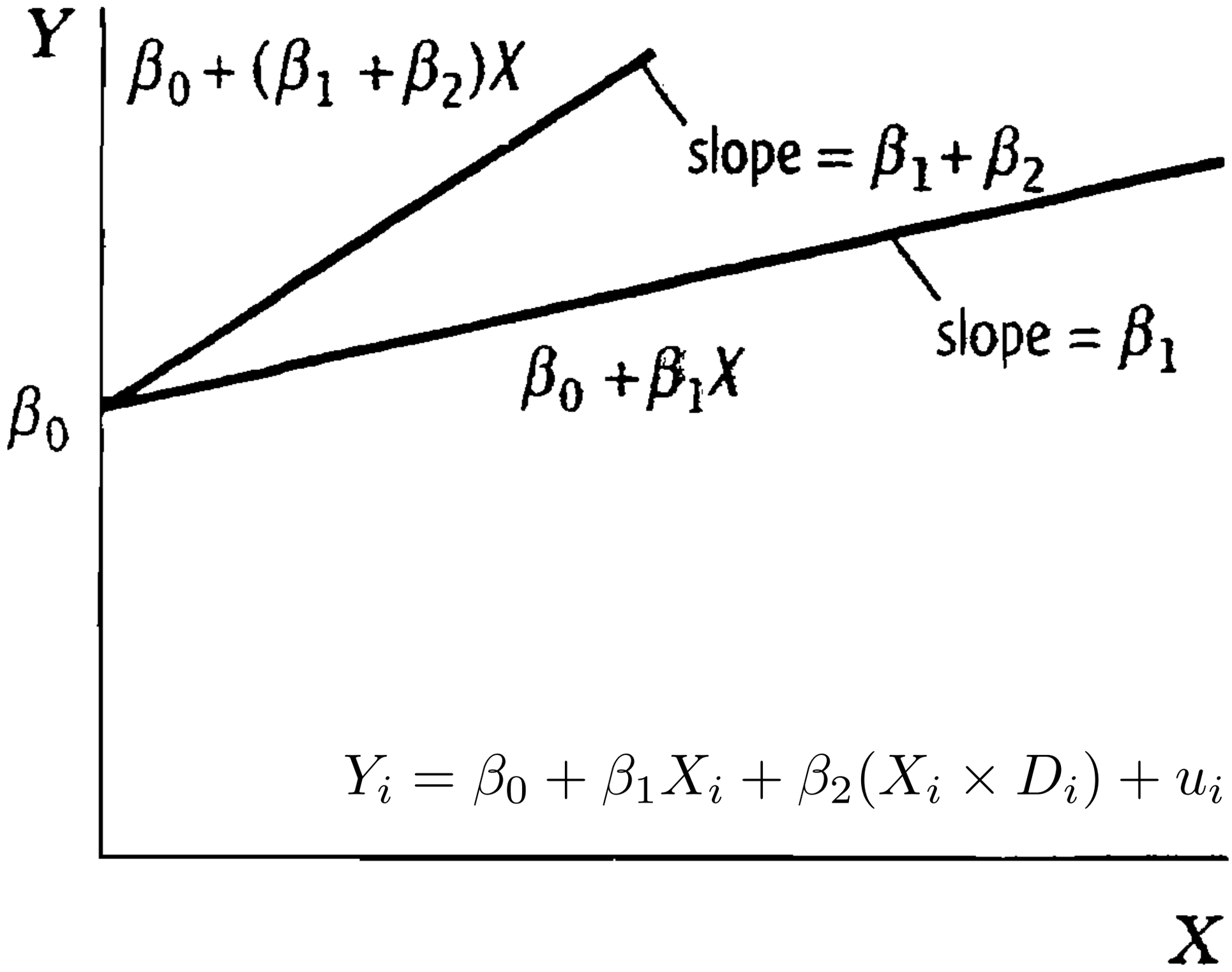
- Models with interaction

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (X_i \times D_i) + u_i$$

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 (X_i \times D_i) + u_i$$


Y
$(\beta_0 + \beta_2) + \beta_1 X$
$\beta_0 + \beta_1 X$
$\beta_0 + \beta_2$
slope = β_1
β_0
$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + u_i$
X


$(\beta_0 + \beta_2) + (\beta_1 + \beta_3)X$
Y
slope = $\beta_1 + \beta_3$
$\beta_0 + \beta_2$
slope = β_1
$\beta_0 + \beta_1 X$
β_0
$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (X_i \times D_i) + u_i$
X



$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 (X_i \times D_i) + u_i$$

Interaction between two continuous variables

- The model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} \times X_{2i}) + u_i$$

- Take test scores as Y, student-teacher ratio as X_1, and percentage of English learners as X_2. Fit this regression model.

- Read pages 324-328 about the interpretation of the model.

Practice

- Reproduce Equations (8.34) and (8.37) in gretl.
- Read Section 8.4 and reproduce Table 8.3 (only the model fitting part) in gretl. Try to use `loop` in combination with `list` to simplify your scripts.

An example:

```
list Z1 = 0 14 15
list Z2 = const str el_pct
loop i=1..2
     ols testscr Z$i
endloop
```

References

1. Stock, J. H. and Watson, M. M., *Introduction to Econometrics*, 3rd Edition, Pearson, 2012.