Econometrics 1 *Applied Econometrics with R*

Lecture 9: Nonlinear Regression

黄嘉平

中国经济特区研究中心 讲师

办公室：文科楼2613

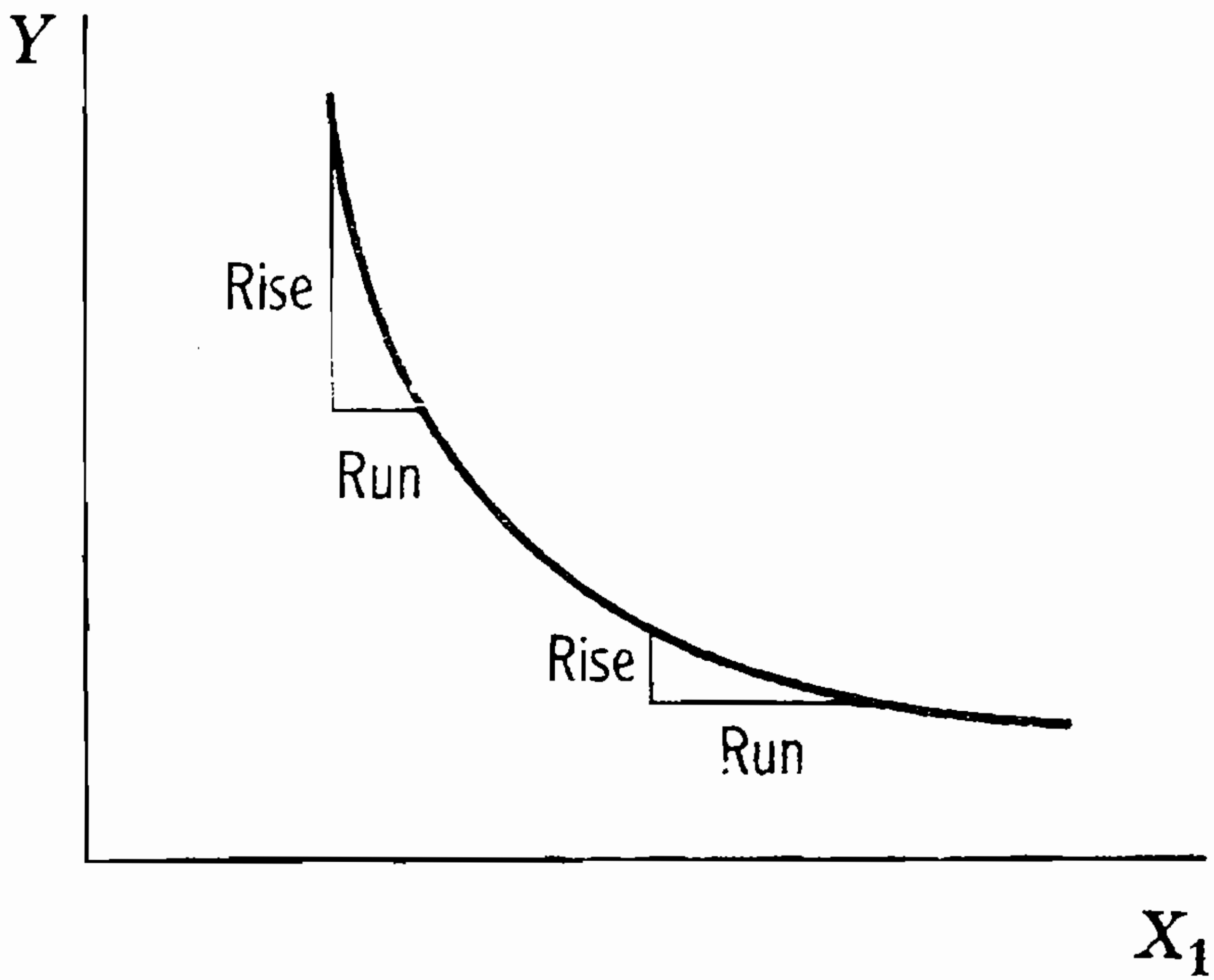
E-mail: huangjp@szu.edu.cn

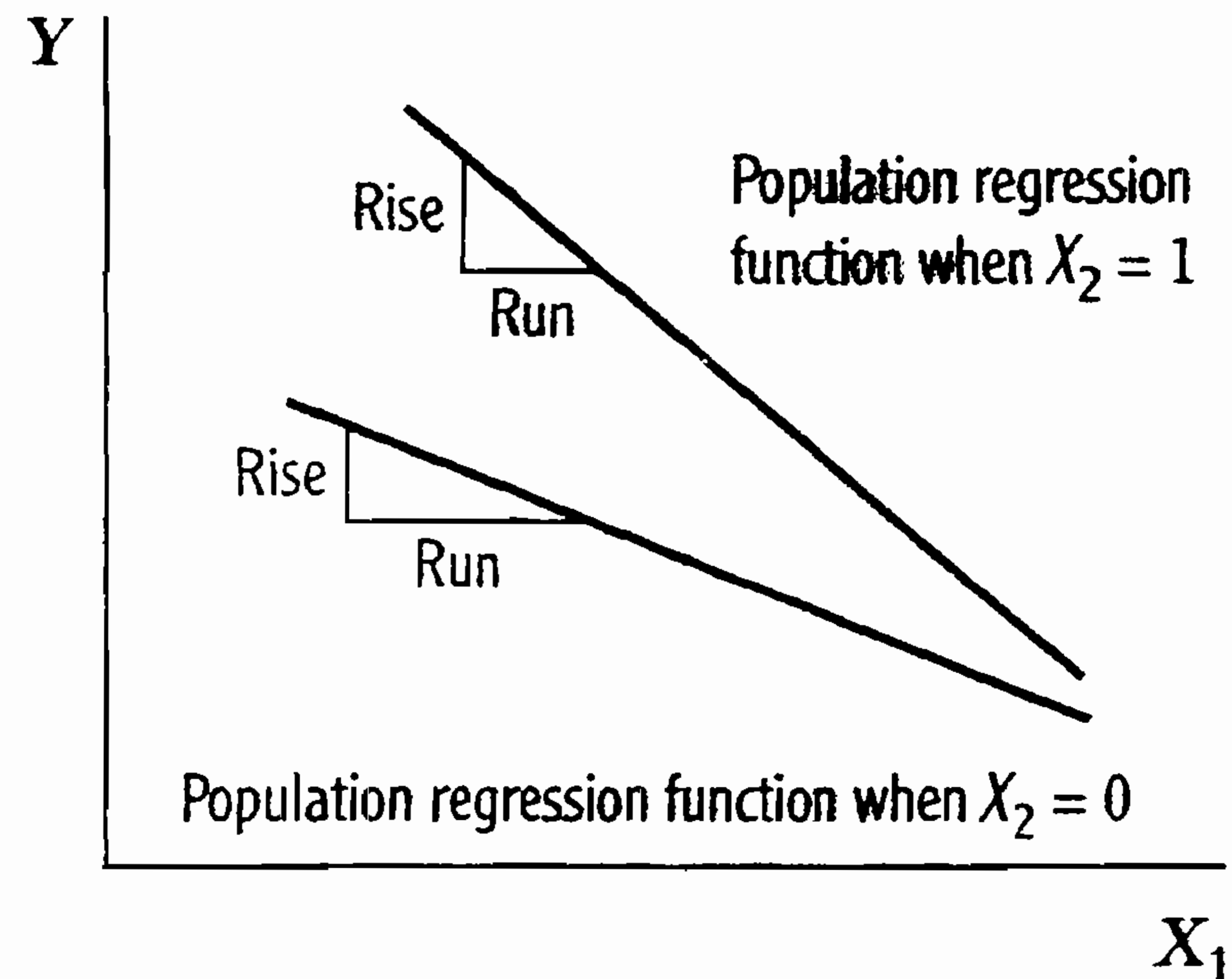
Tel: (0755) 2695 0548

Website: https://huangjp.com

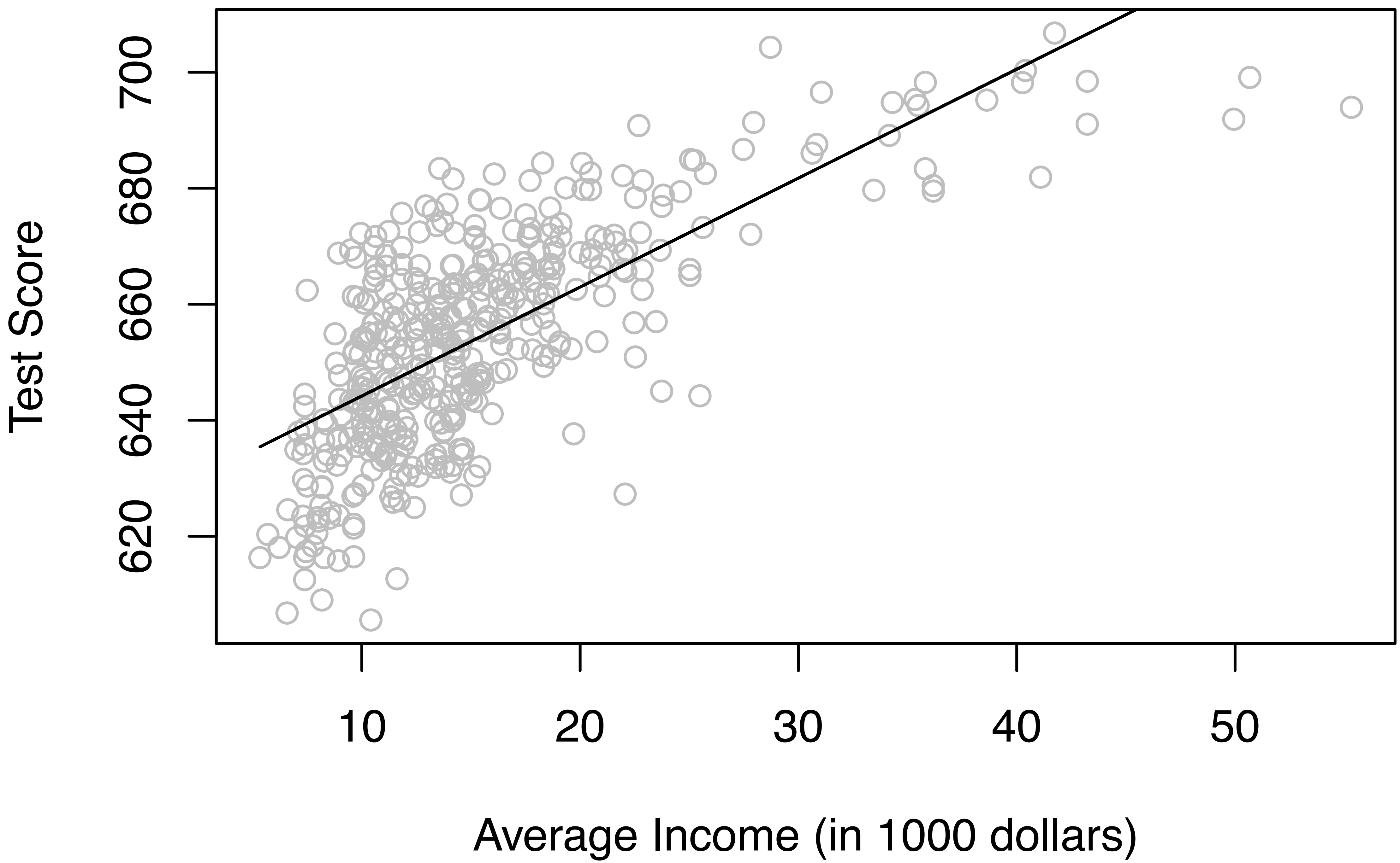
Nonlinear Regression

Two types of nonlinearity

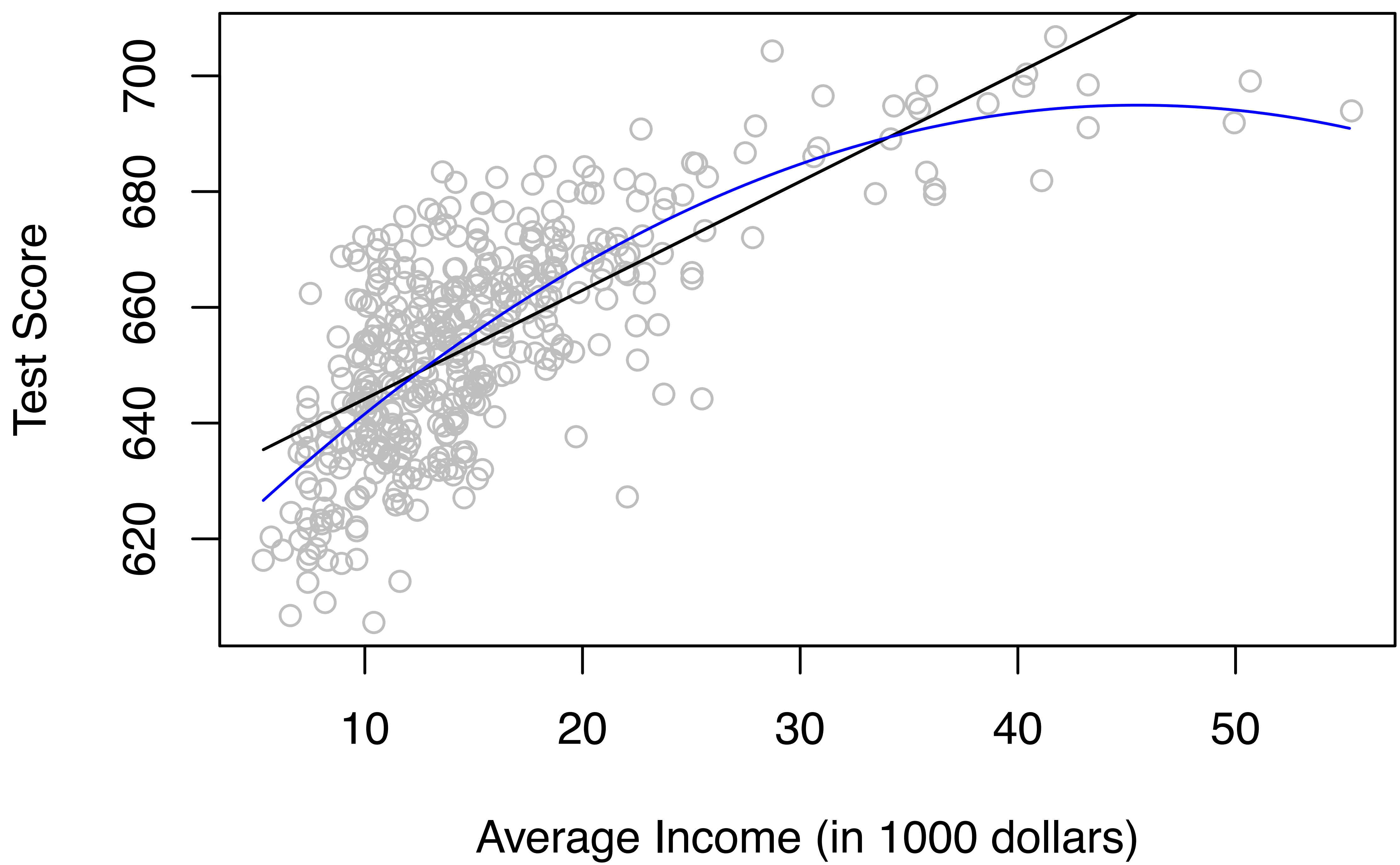




Average income vs. test score



Average income vs. test score



Average income vs. test score

- A quadratic regression model

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

- Implementation in R

```
> lm(testscr ~ avginc + I(avginc^2))
```

- The `I()` command ensures that the term `avginc^2` is an independent variable of the model.

Investigate the `I()` command

- Perform the following commands with `summary()`

```
> lm(testscr ~ avginc)

> lm(testscr ~ avginc + avginc^2)

> lm(testscr ~ avginc + I(avginc^2))

> avginc2 <- avginc^2
> lm(testscr ~ avginc + avginc2)
```

What did you find?

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

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

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

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

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.

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

- Polynomials

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

- Logarithms

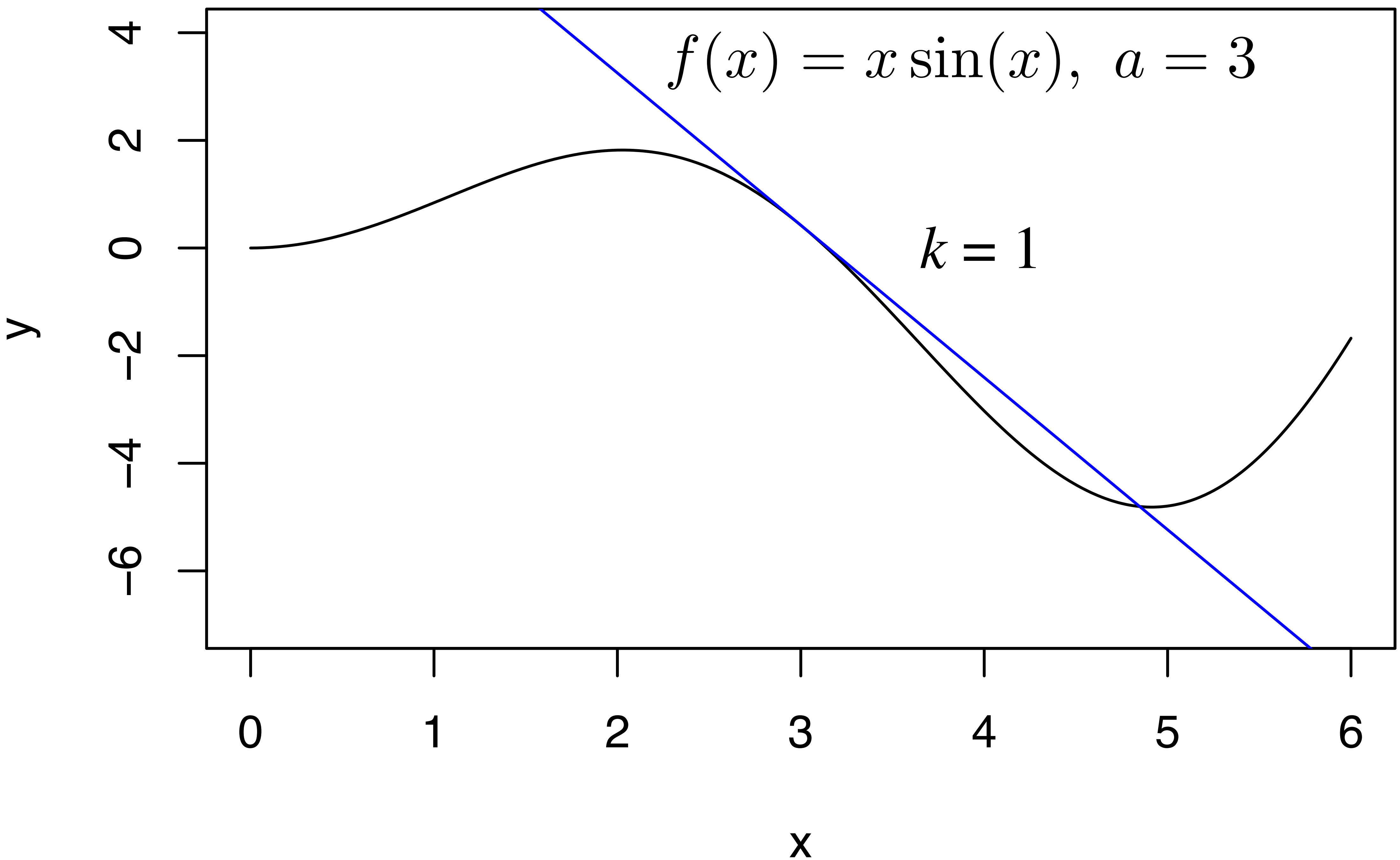
$$\begin{aligned} Y_i &= \beta_0 + \beta_1 \ln(X_i) + u_i \\ \ln(Y_i) &= \beta_0 + \beta_1 X_i + u_i \\ \ln(Y_i) &= \beta_0 + \beta_1 \ln(X_i) + u_i \end{aligned}$$

Polynomials

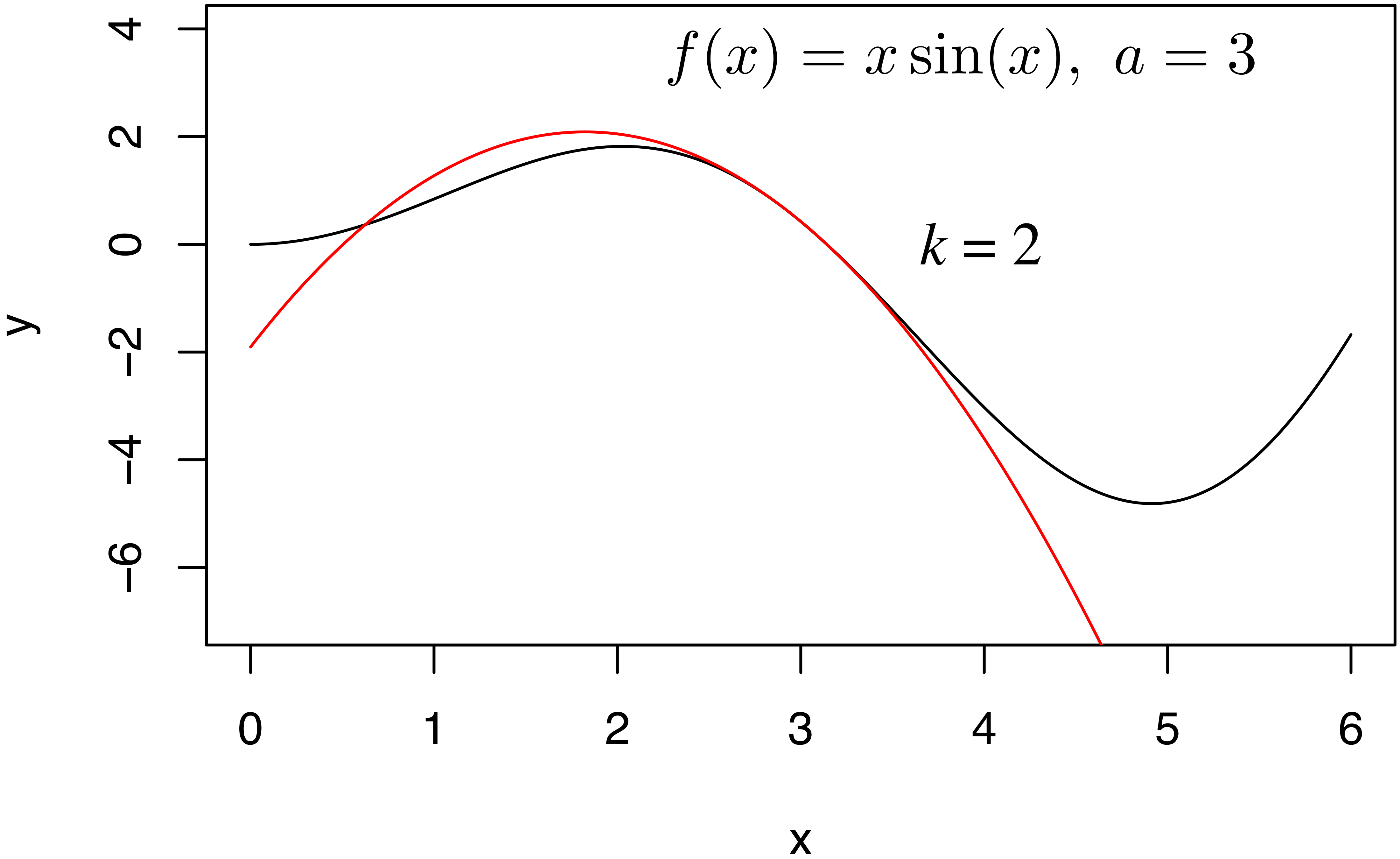
- Why polynomials?
- Taylor series expansion of a smooth function $f(x)$ at point a:

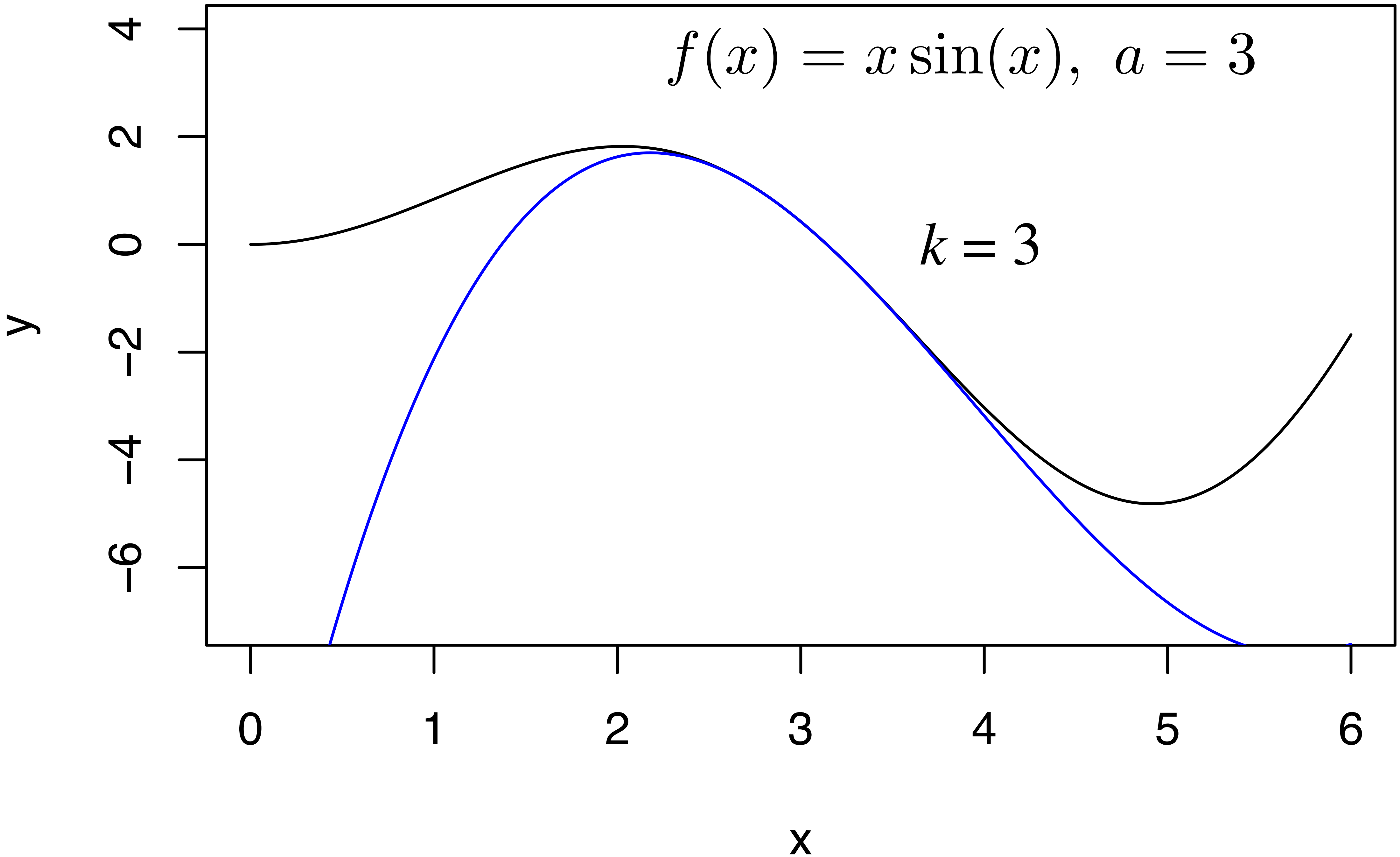
$$
\begin{aligned}
f(x) = & f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 \\
& + \frac{f'''(a)}{3!}(x-a)^3 + \frac{f''''(a)}{4!}(x-a)^4 + \cdots
\end{aligned}
$$

- A function can be approximated by a polynomial.

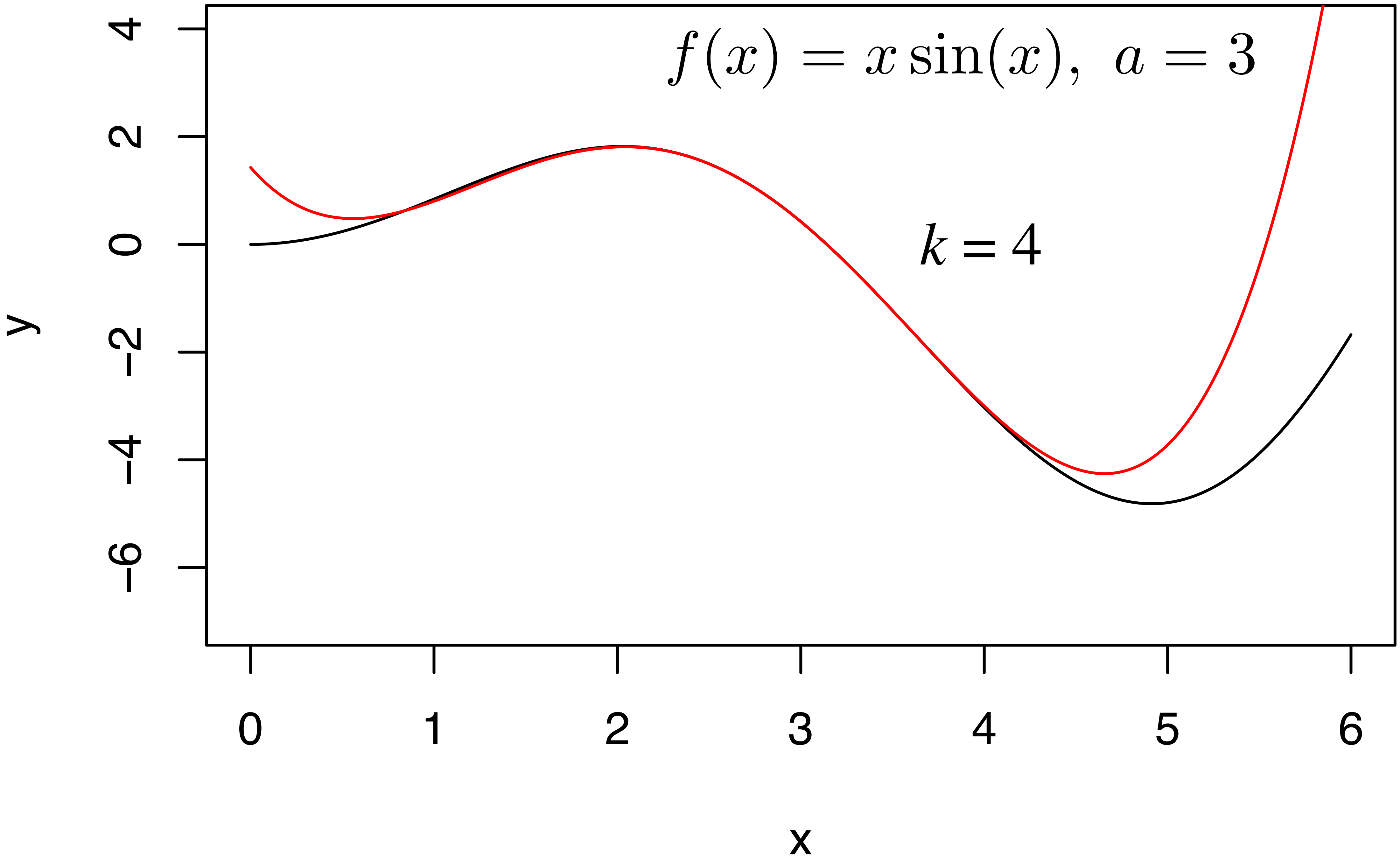


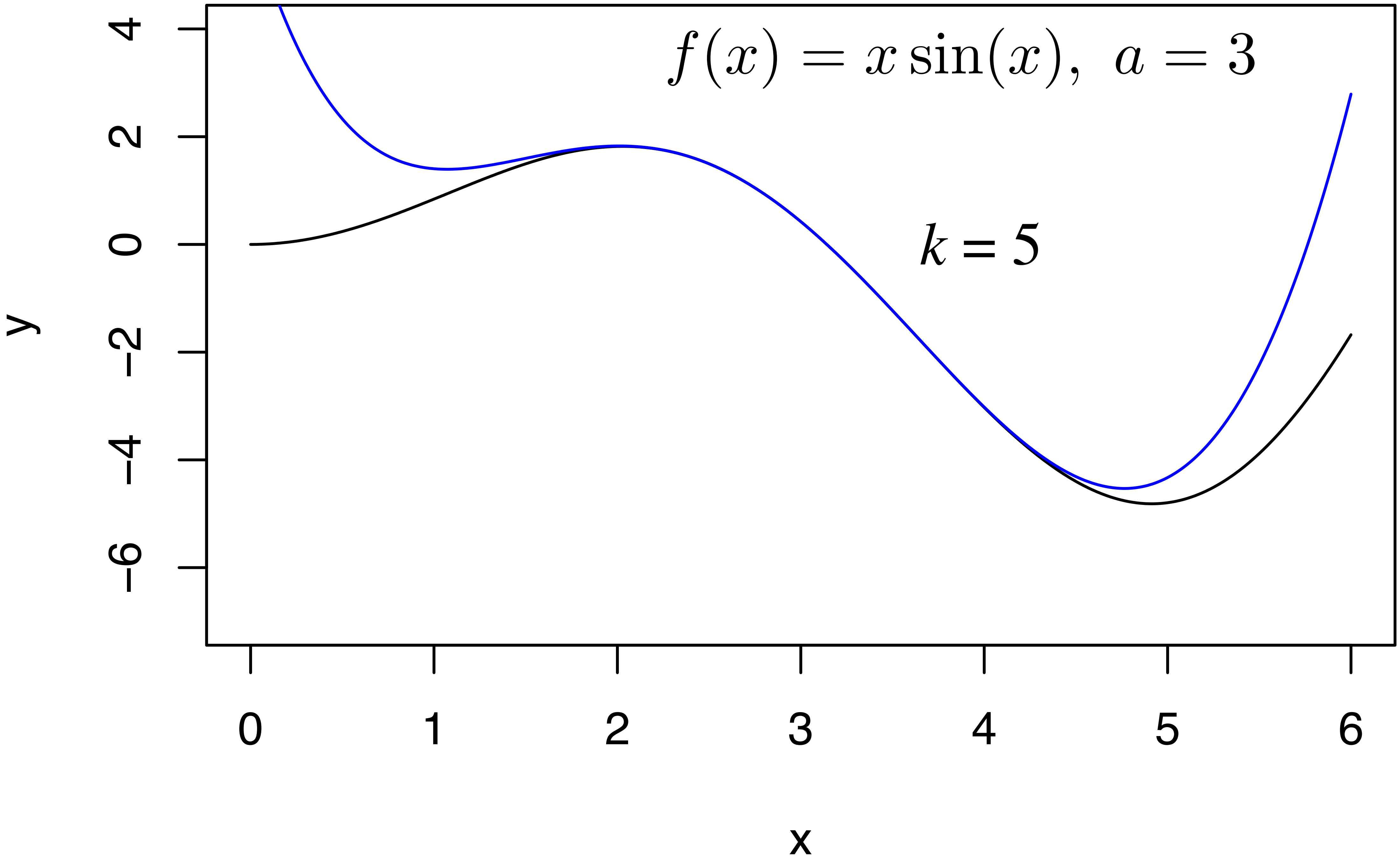
$f(x) = x\sin(x),\ a = 3$
$k = 1$
y
x
4
2
0
−2
−4
−6
0
1
2
3
4
5
6


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

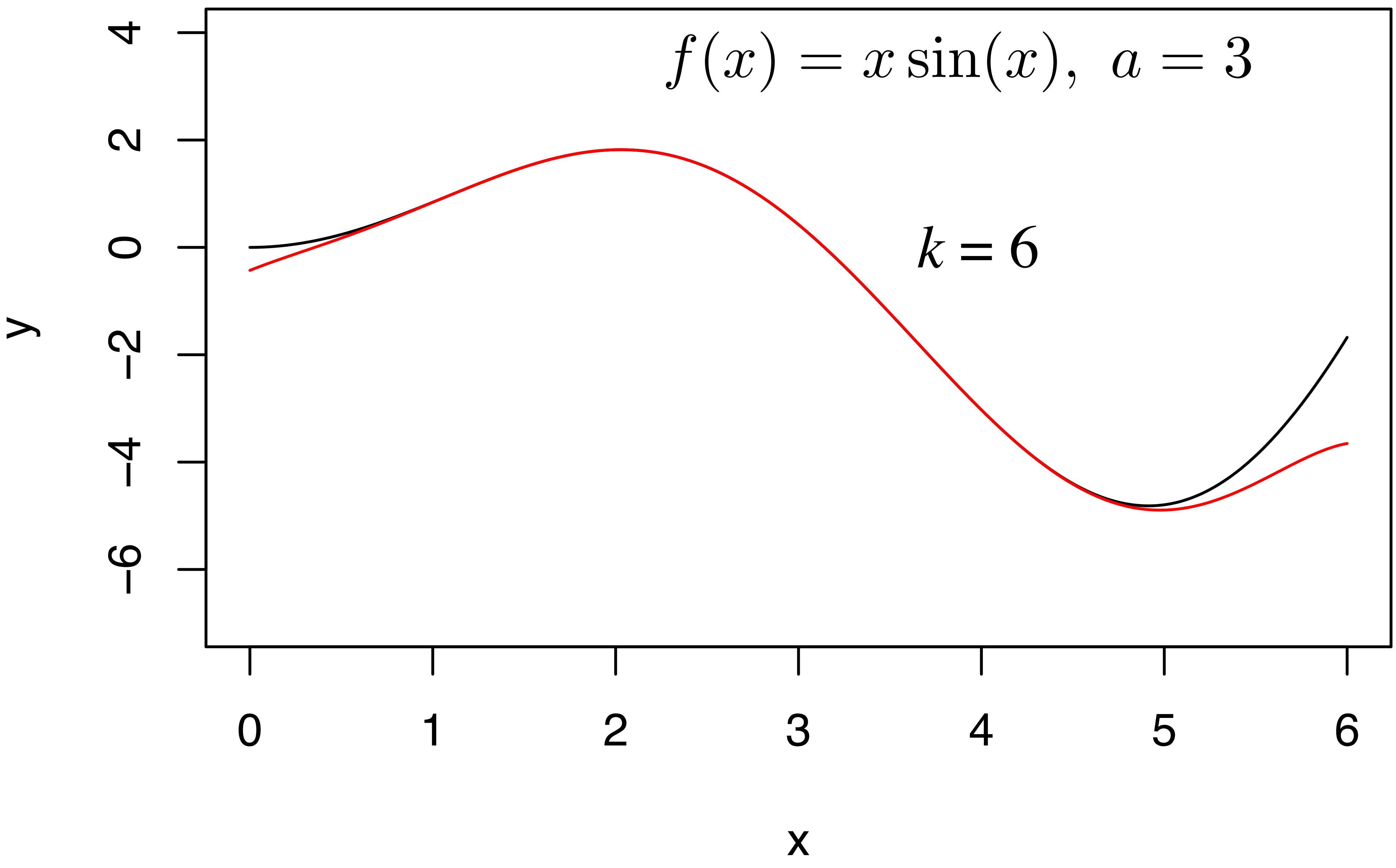


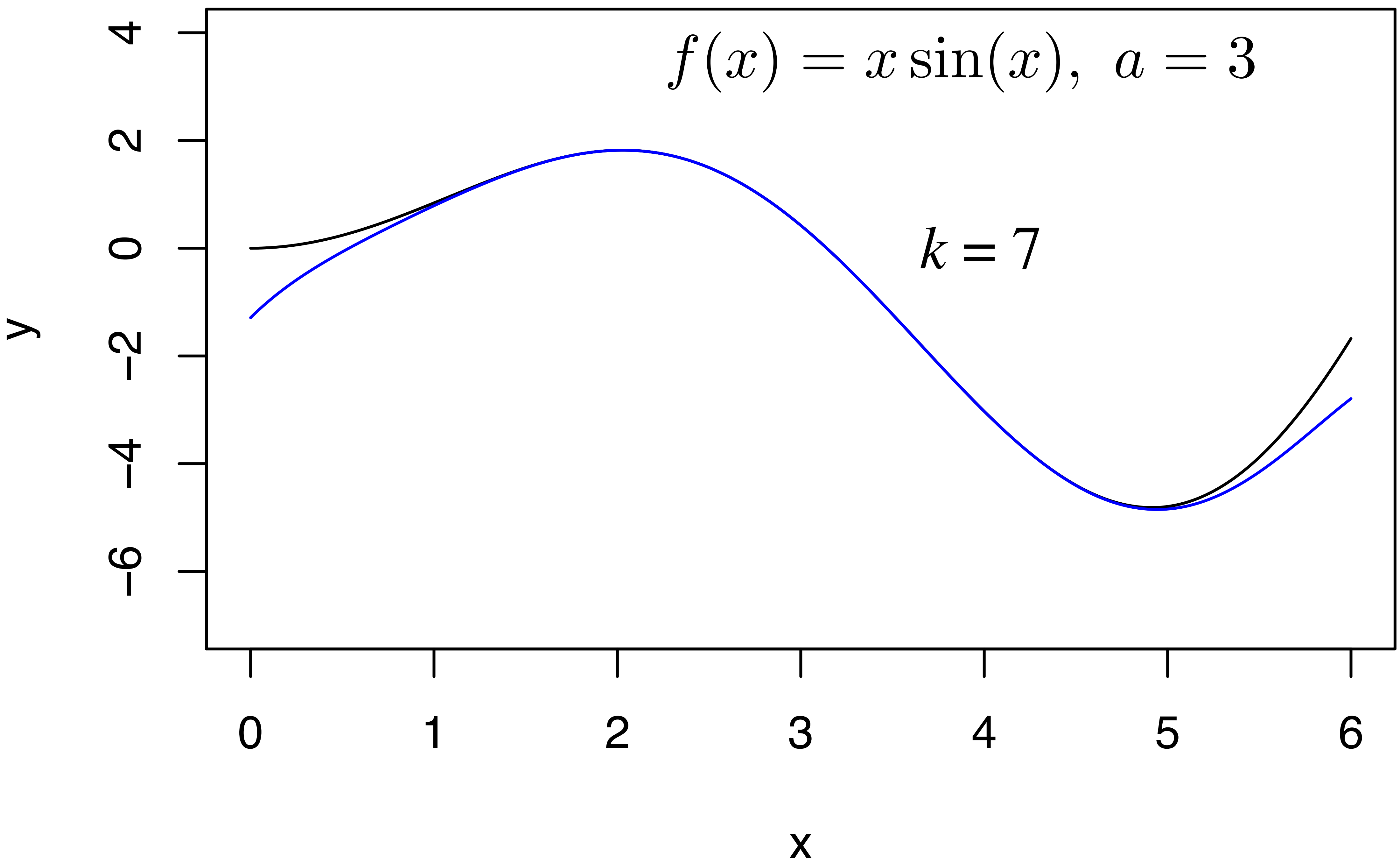
$f(x) = x\sin(x),\ a = 3$
$k = 3$
y
x
4
2
0
−2
−4
−6
0
1
2
3
4
5
6


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

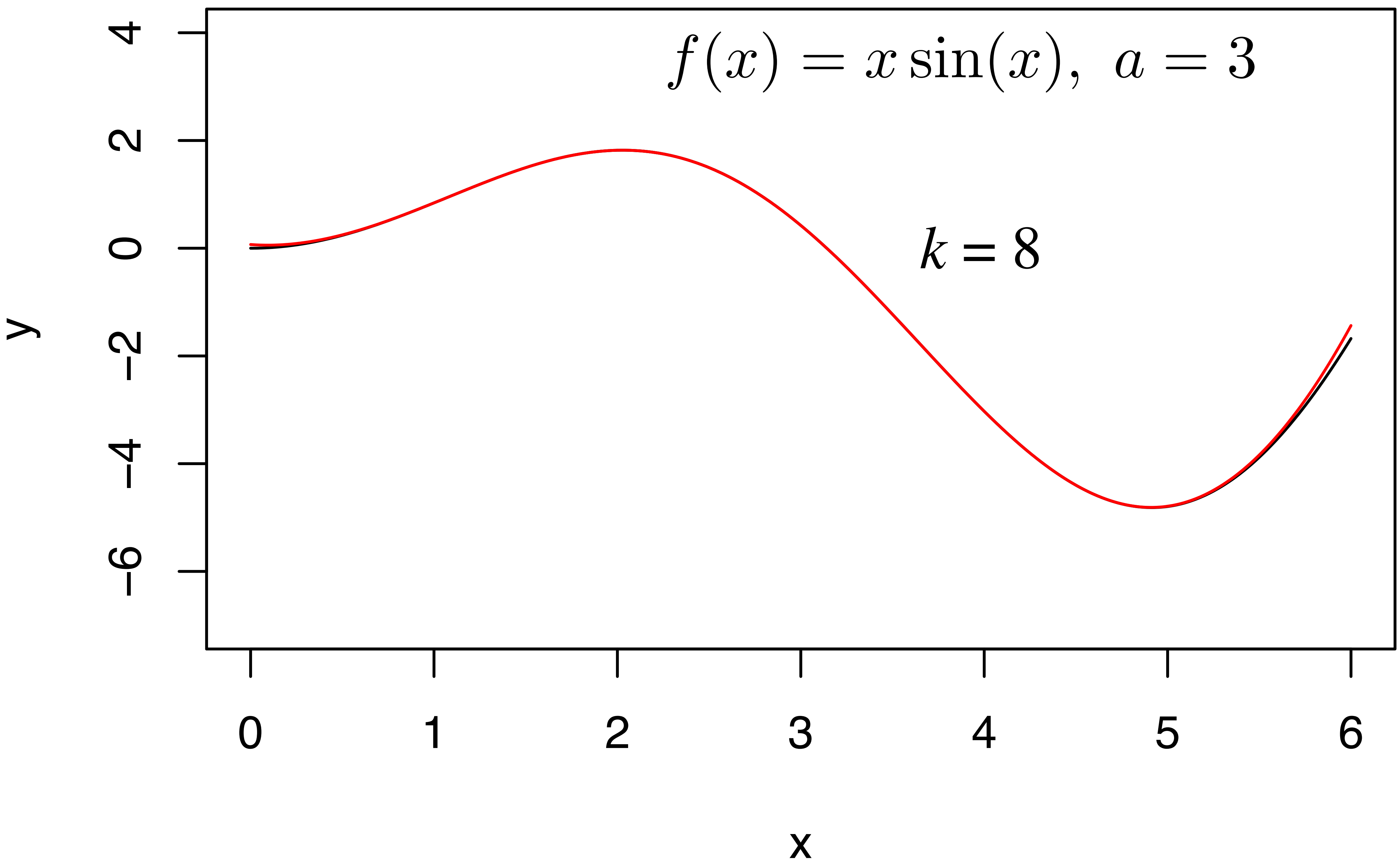


$f(x) = x\sin(x),\ a = 3$
$k = 5$
4
2
0
−2
−4
−6
y
0
1
2
3
4
5
6
x


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



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


$f(x) = x\sin(x),\ a = 3$
$k = 8$
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?

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 …

Heteroskedasticity-robust standard errors

- Regression with the `lm` command in R is under the homoskedasiticity assumption (see Section 5.4).

$$\widehat{\text{TestScore}} = \underset{(5.83)}{600.1} + \underset{(0.86)}{5.02}\,\text{Income} + \underset{(0.037)}{0.096}\,\text{Income}^2 + \underset{(0.00047)}{0.00069}\,\text{Income}^3$$

- The heteroskedasticity-robust standard errors given in the book are

$$\widehat{\text{TestScore}} = \underset{(5.1)}{600.1} + \underset{(0.71)}{5.02}\,\text{Income} + \underset{(0.029)}{0.096}\,\text{Income}^2 + \underset{(0.00035)}{0.00069}\,\text{Income}^3$$

Heteroskedasticity-robust standard errors in R

- Use the `coeftest` command in `lmtest` package in combination with the `vcov` command in `sandwich` package:

```
> library(sandwich)
> library(lmtest)
> fm <- lm(testscr ~ avginc + I(avginc^2) + I(avginc^3))
> coeftest(fm, vcov = vcovHC, type = “HC0”)
```

- Alternative settings of `type`:

  `“const”` — homoskedastic case
  `“HC0”` — the model used in the book
  `“HC3”` — default

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

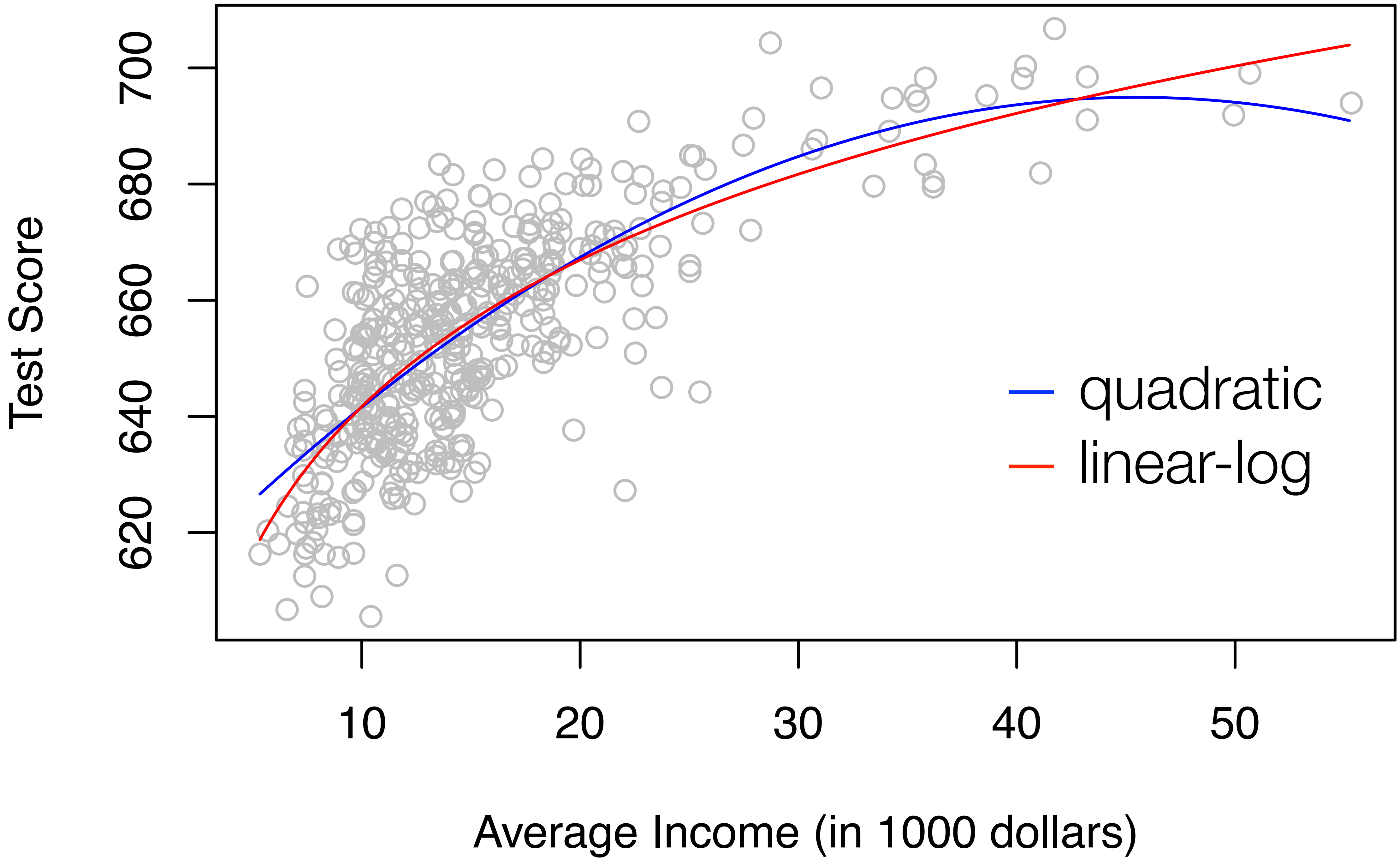
- The *linear-log* model

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

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

- Practice
  - Fit the model $\text{TestScore}_i = \beta_0 + \beta_1 \ln(\text{Income}_i) + u_i$
  - Plot your estimated regression line with sample data.

```
> fm <- lm(testscr ~ log(avginc))
> newx <- seq(min(avginc), max(avginc), 0.1)
> newy <- fm$coefficients[1] + fm$coefficients[2] * log(newx)
> plot(avginc, testscr, col = “gray")
> lines(newx, newy, col = “red")
```



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

- Try the log-log model in the regression of `testscr` on `avginc`.

- Plot your regression function on the scatter plot.
  x axis: average income
  y axis: log of test score

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, than 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

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

Interaction between a continuous and a binary variable

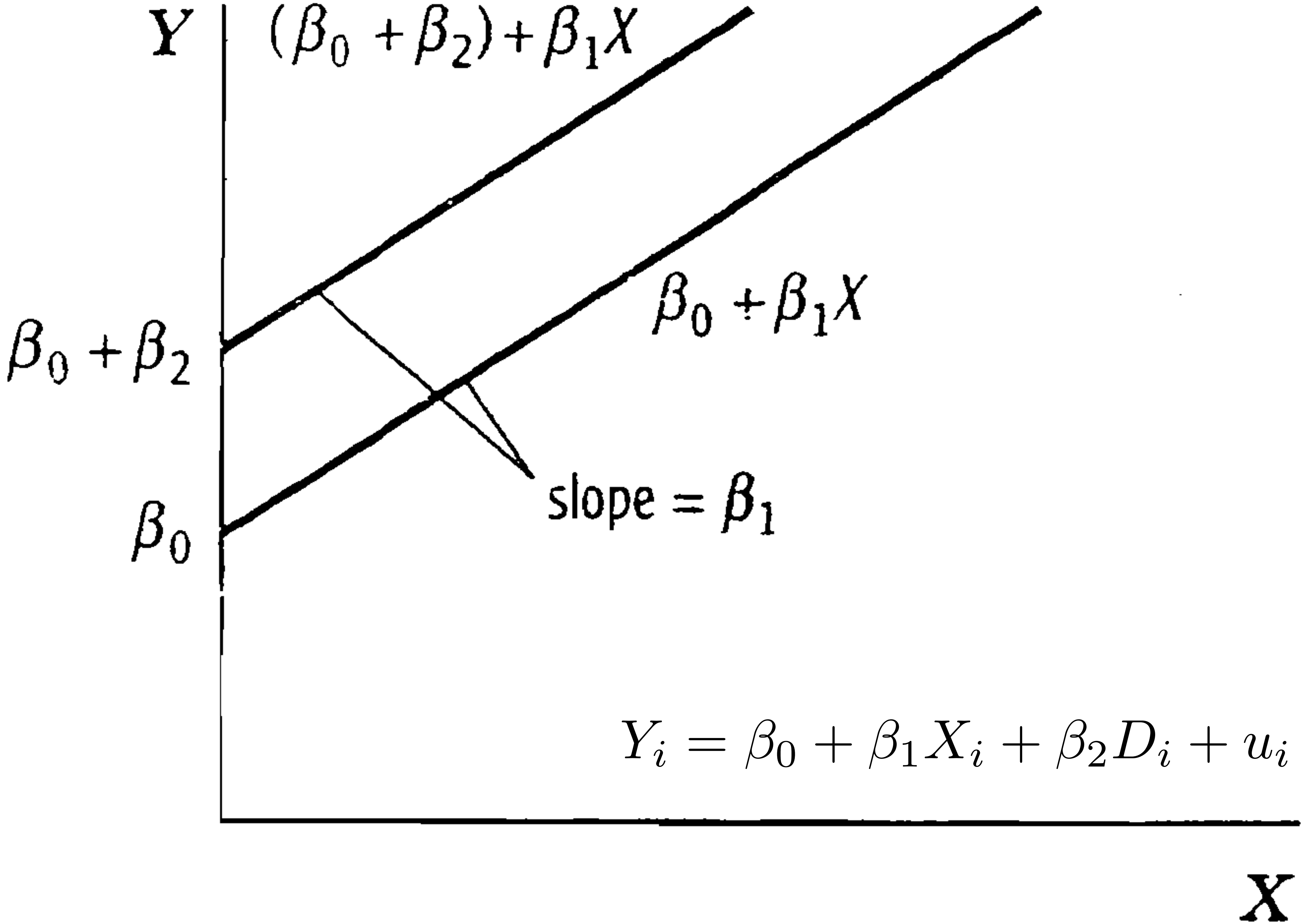
- Model without interaction

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

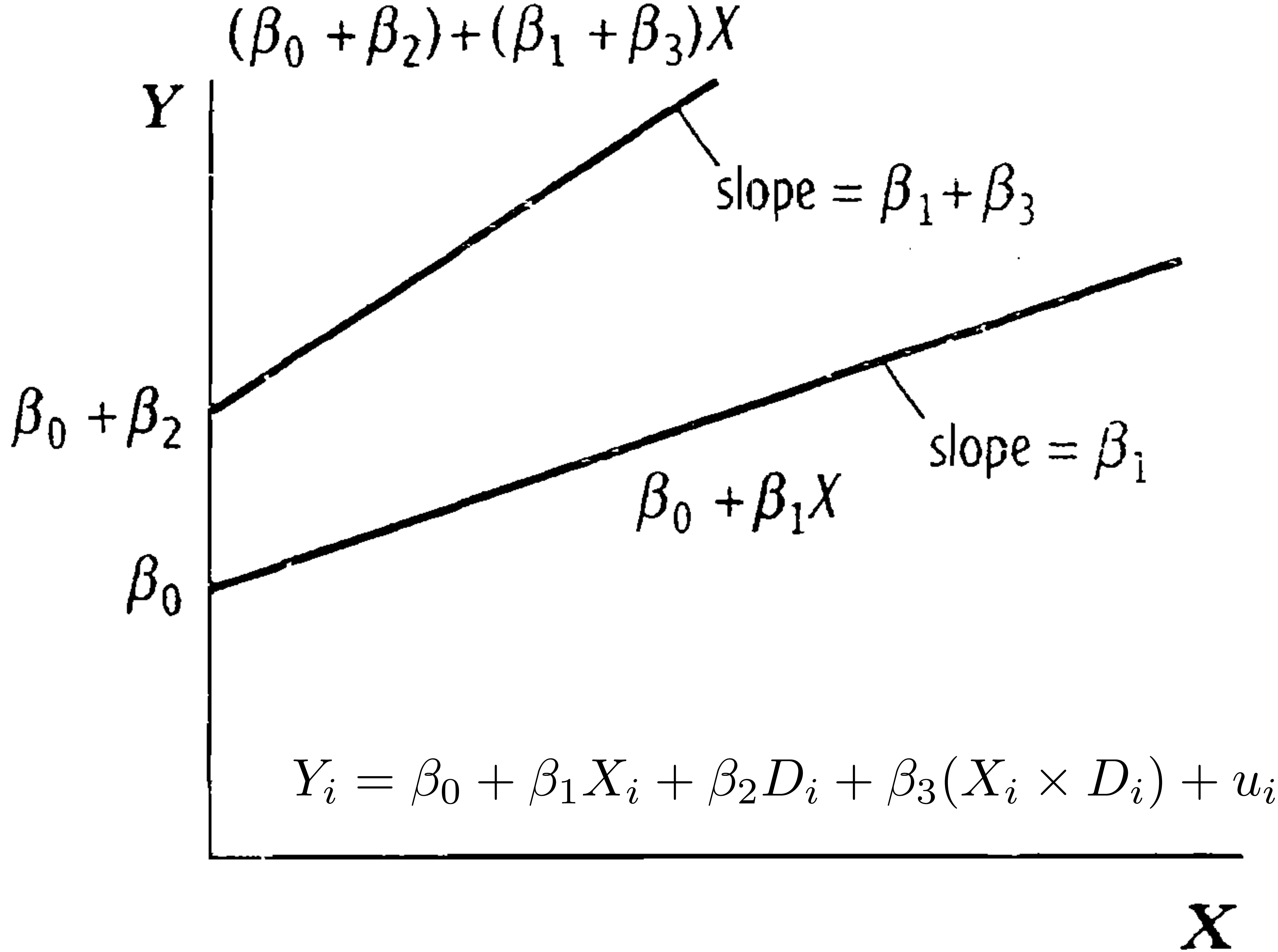
- Models with interaction

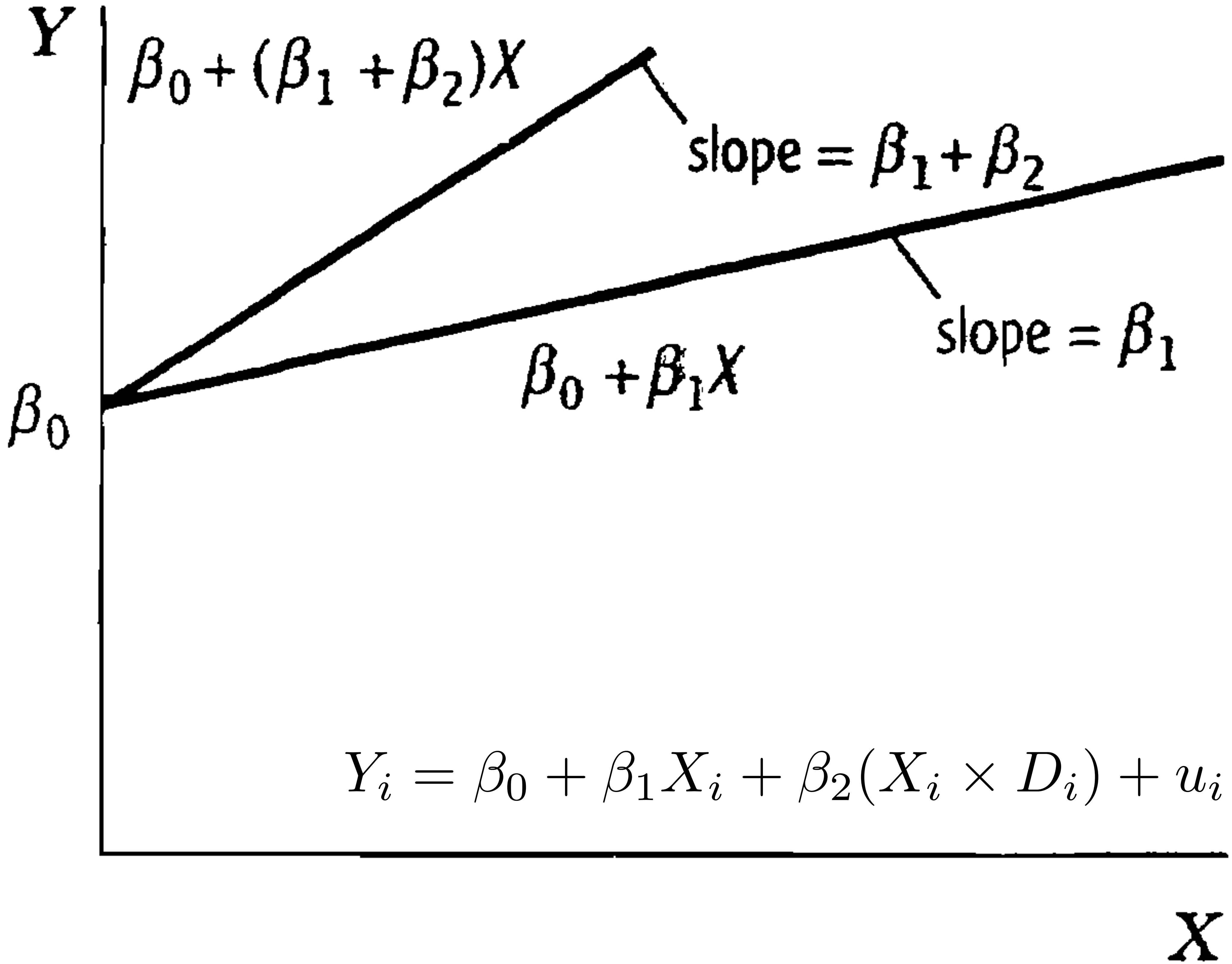
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (X_i \times D_i) + u_i$$

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



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


$(\beta_0 + \beta_2) + (\beta_1 + \beta_3)X$
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



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

Binary variables in R

- If you want a variable (column of data) to be treated as a binary (dummy) variable, it must be defined as a `factor` type of data.

- Example:
```
> factor(c(0, 1, 1))
```

- Try to figure out what the following codes do:

```
> scoredata <- read.csv("caschool.csv")
> d <- factor(el_pct > 10)
> nld <- lm(testscr ~ str + d + str:d)
```

Interaction of variables in `lm()`

- A interaction term between `x1` and `x2` is specified by

  `x1:x2`

- `x1*x2` is equivalent to `x1 + x2 + x1:x2`

- Practice with

```
> lm(testscr ~ str + d + str:d)

> lm(testscr ~ str * d)
```

The `I()` command

- The `^` operator used with `lm()` command has the following meaning

```
  (x1 + x2 + x3)^2
= (x1 + x2 + x3) * (x1 + x2 + x3)
= x1 + x2 + x3 + x1:x2 + x1:x3 + x2:x3
```

  for more details, see `help(formula)`.

- Therefore, if you want to evaluate the quadratic term of `x` (which is an arithmetic operation), you need to use `I(x^2)`.

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.

Take-home exercise (not an assignment)

- Try other possible models that contain nonlinear terms of student-teacher ratio and percentage of English learners, as well as the interactions between them, and predict test scores.

- Read Section 8.4 and Chapter 9.

References

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

2. Kleiber, C. and Zeileis, A., *Applied Econometrics with R*, Springer, 2008.