计量经济学

第八讲：非线性回归函数

黄嘉平

工学博士 经济学博士
深圳大学中国经济特区研究中心 讲师

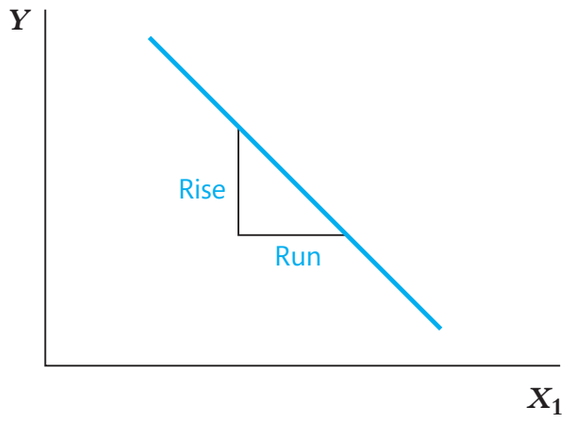
办公室　粤海校区汇文楼2613
E-mail　huangjp@szu.edu.cn
Website　https://huangjp.com

主要内容

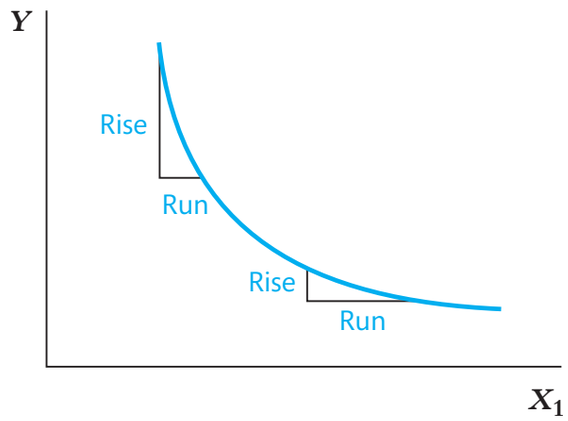
- 非线性回归模型的一般建模方法
- 一元非线性函数
  - 多项式（polynomials）
  - 对数函数（logarithms）
- 自变量的交互作用
  - 两个二值变量的交互作用
  - 连续变量和二值变量的交互作用
  - 两个连续变量的交互作用
- 学生教师比对测试成绩的非线性效应（自学）

线性与非线性回归函数

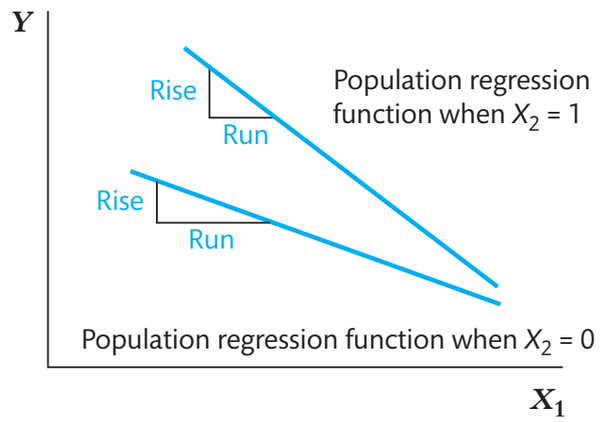
- 线性回归模型中的“线性”包含两重含义：
  1. 指回归函数是回归变量的线性函数，即 Y 是 $X_1, \ldots, X_m$ 的线性函数。 使模型简单
  2. 指回归函数是回归系数的线性函数，即 Y 是 $\beta_0, \beta_1, \ldots, \beta_m$ 的线性函数。 OLS 估计的前提
- 当假设 Y 是 $X_1, \ldots, X_m$ 的线性函数时，X_k 的变化对 Y 的效应为常数，即系数 β_k 。
- 现实中，X_k 的变化对 Y 的效应有可能依赖于 X_k 自身的取值，也有可能依赖于其他回归变量的取值。因此有必要考虑非线性的回归函数。



(a) Constant slope



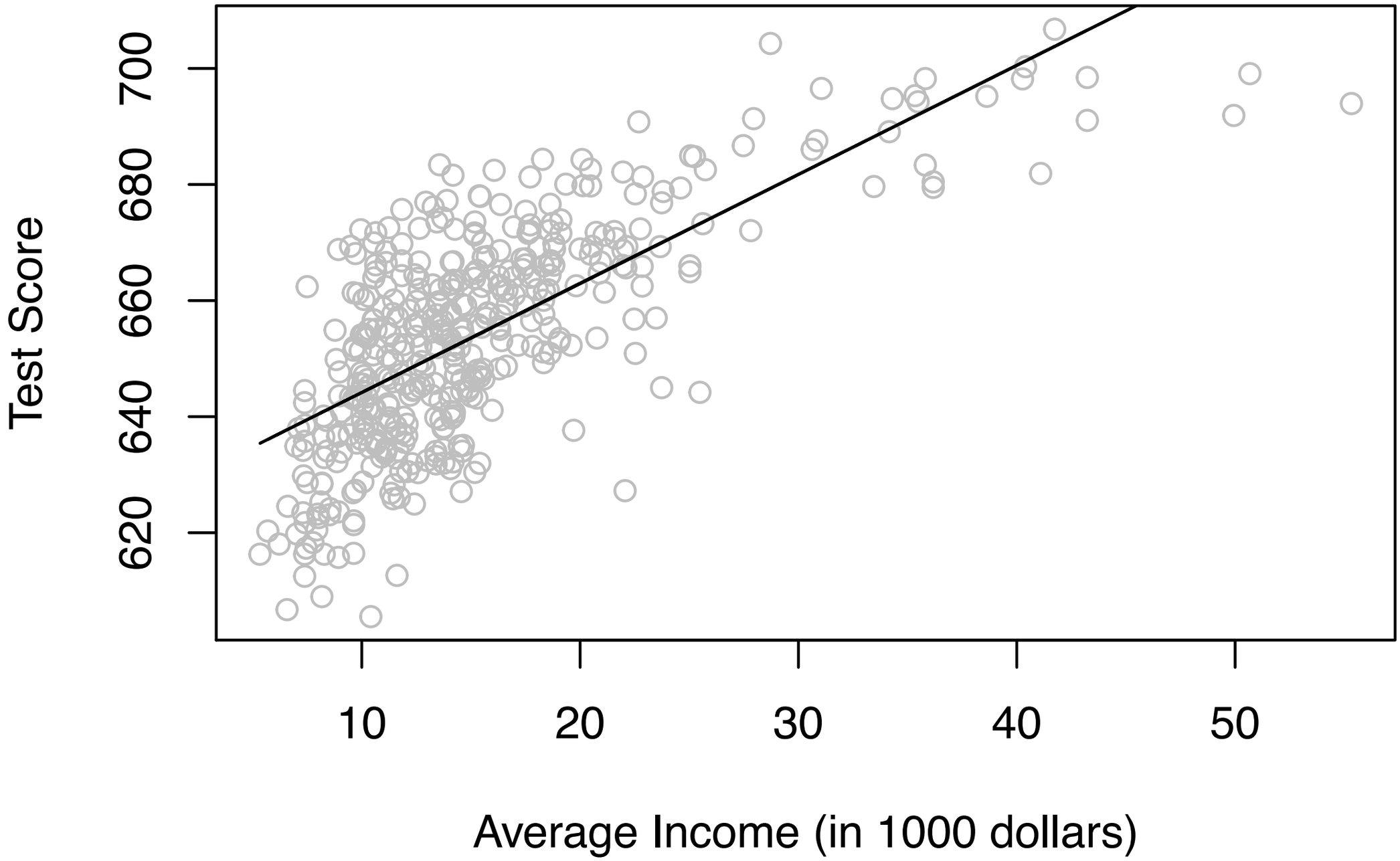
(b) Slope depends on the value of X_1



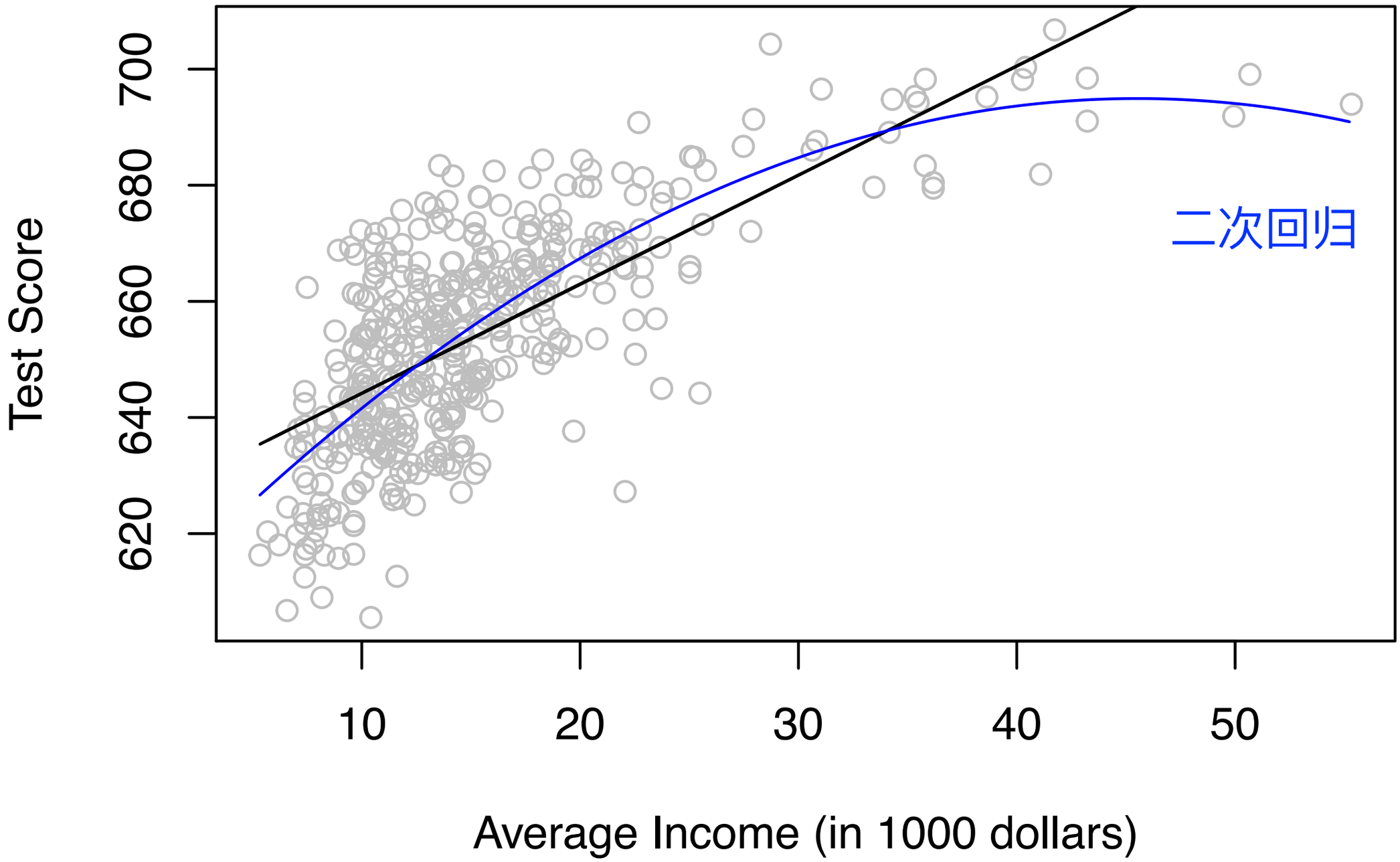
(c) Slope depends on the value of X_2

非线性回归模型的一般建模方法

测试成绩和地区收入


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

测试成绩和地区收入


Test Score
620
640
660
680
700
10
20
30
40
50
Average Income (in 1000 dollars)
二次回归

二次回归模型

Quadratic regression model

- 二次总体回归模型

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

- 线性模型与二次模型的估计结果

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

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

非线性回归模型

Nonlinear regression model

- 非线性回归模型：

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

其中 $f(X_{1i}, X_{2i}, \ldots, X_{mi})$ 为总体非**线性回归函数（nonlinear regression function）**，是自变量 $X_{1i}, X_{2i}, \ldots, X_{mi}$ 的非线性函数。

- 在其他变量保持不变的情况下，当 X_k 变为 $X_k + \Delta X_k$ 时，对应的 Y 的变化为

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

收入变化对测试成绩的非线性效应

- 在测试成绩和收入的模型中，当地区收入从 10 提高到 11 时，测试成绩的变化的预测值为

$$(\widehat{\beta}_0 + \widehat{\beta}_1 \times 11 + \widehat{\beta}_2 \times 11^2) - (\widehat{\beta}_0 + \widehat{\beta}_1 \times 10 + \widehat{\beta}_2 \times 10^2)$$
$$= \widehat{\beta}_1 + 21\widehat{\beta}_2 = 3.85 + 21 \times (0.042) \approx 2.96$$

- 当地区收入从 40 提高到 41 时，测试成绩的变化的预测值为

$$(\widehat{\beta}_0 + \widehat{\beta}_1 \times 41 + \widehat{\beta}_2 \times 41^2) - (\widehat{\beta}_0 + \widehat{\beta}_1 \times 40 + \widehat{\beta}_2 \times 40^2)$$
$$= \widehat{\beta}_1 + 81\widehat{\beta}_2 = 3.85 + 81 \times (0.042) \approx 0.42$$

利用多元回归建立非线性模型的一般方法

1. 确定一种可能的非线性关系。

2. 确定一个非线性函数并用 OLS 估计其参数。

   OLS 估计仅需要回归函数是回归系数的线性函数，并不要求回归函数是回归变量的线性函数。

3. 确定非线性模型是否改进了线性模型。

   可针对总体回归函数是线性的原假设（非线性项系数均为零）和非线性的备择假设进行检验（参照 8.2 节）。

4. 画出非线性回归函数估计图。

5. 估计 X 的变化对 Y 的效应。

   可针对一个或多个具有代表性的 X 取值计算。

一元非线性函数

单一自变量的非线性函数

- 多项式（polynomials）

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

- 对数（logarithms）

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

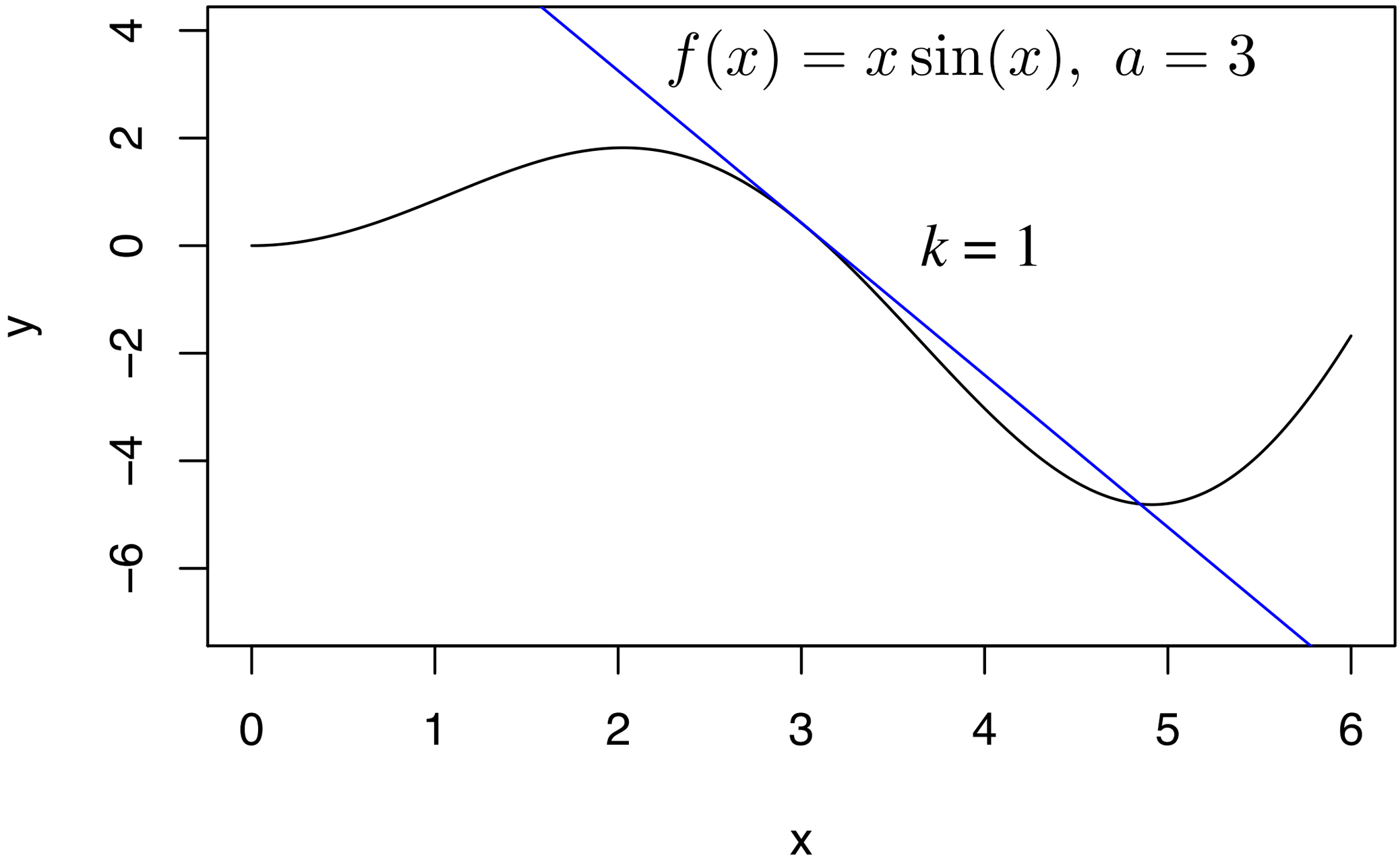
多项式

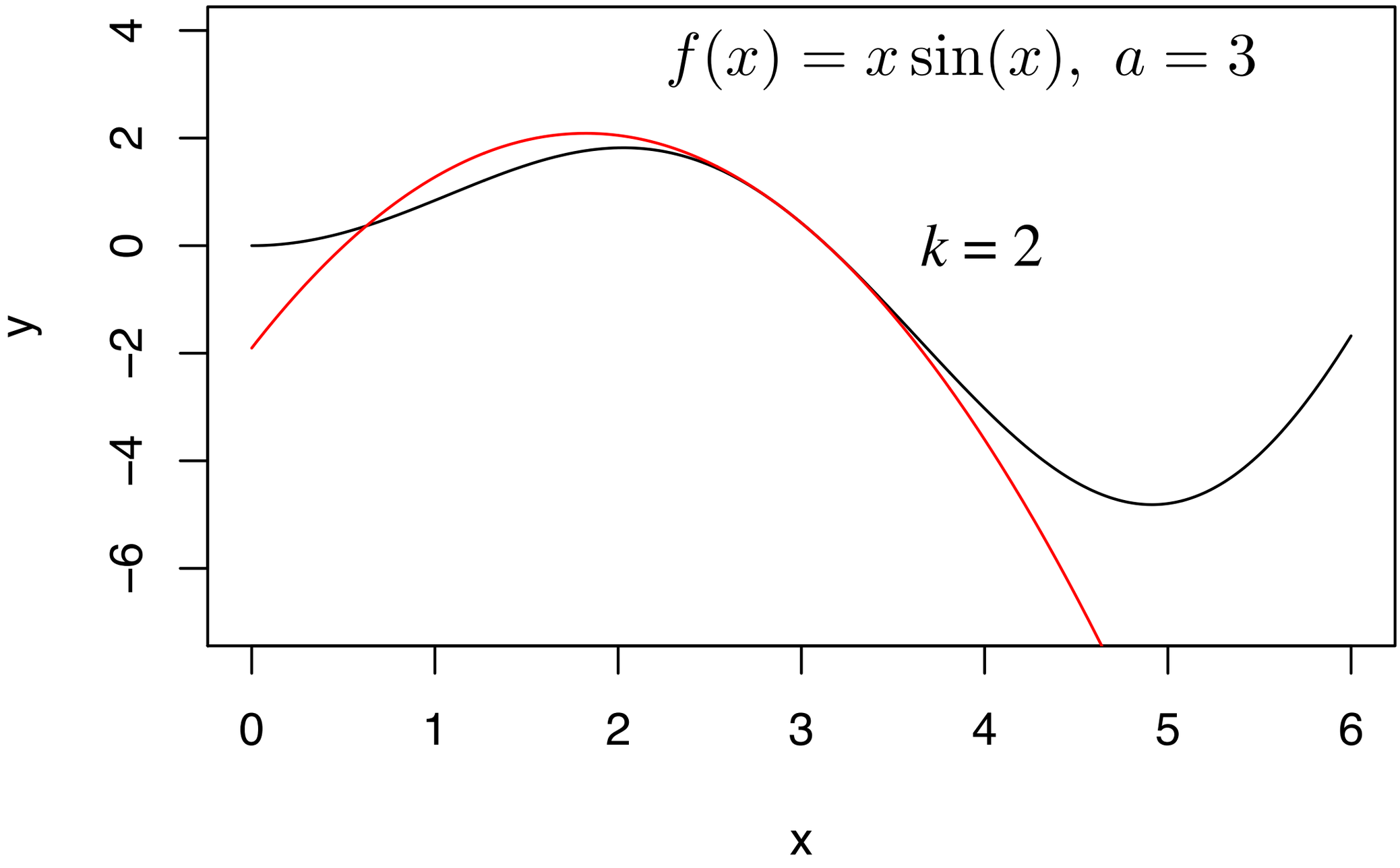
Polynomials

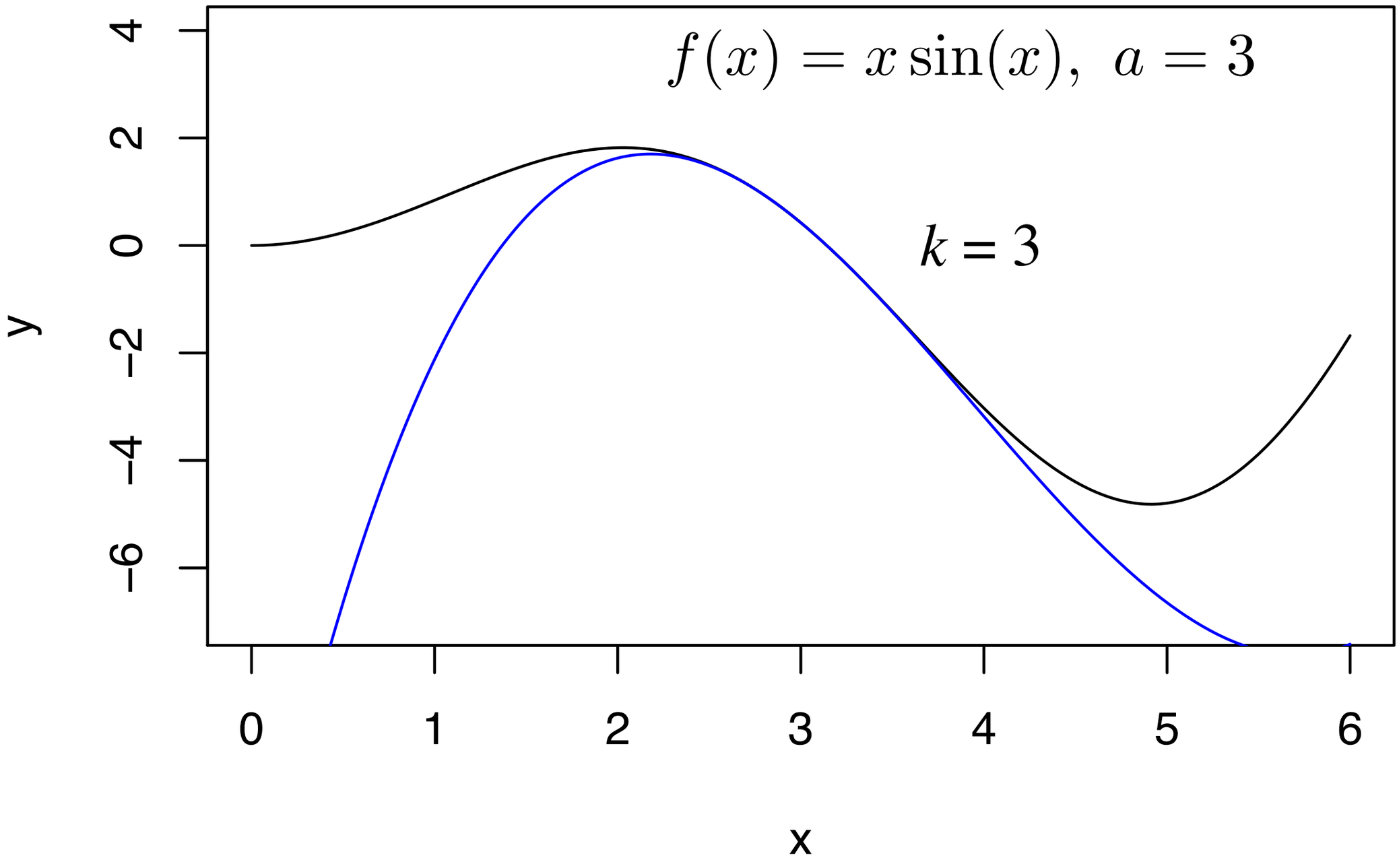
- 为什么考虑多项式?
- 泰勒展开（Taylor expansion）公式

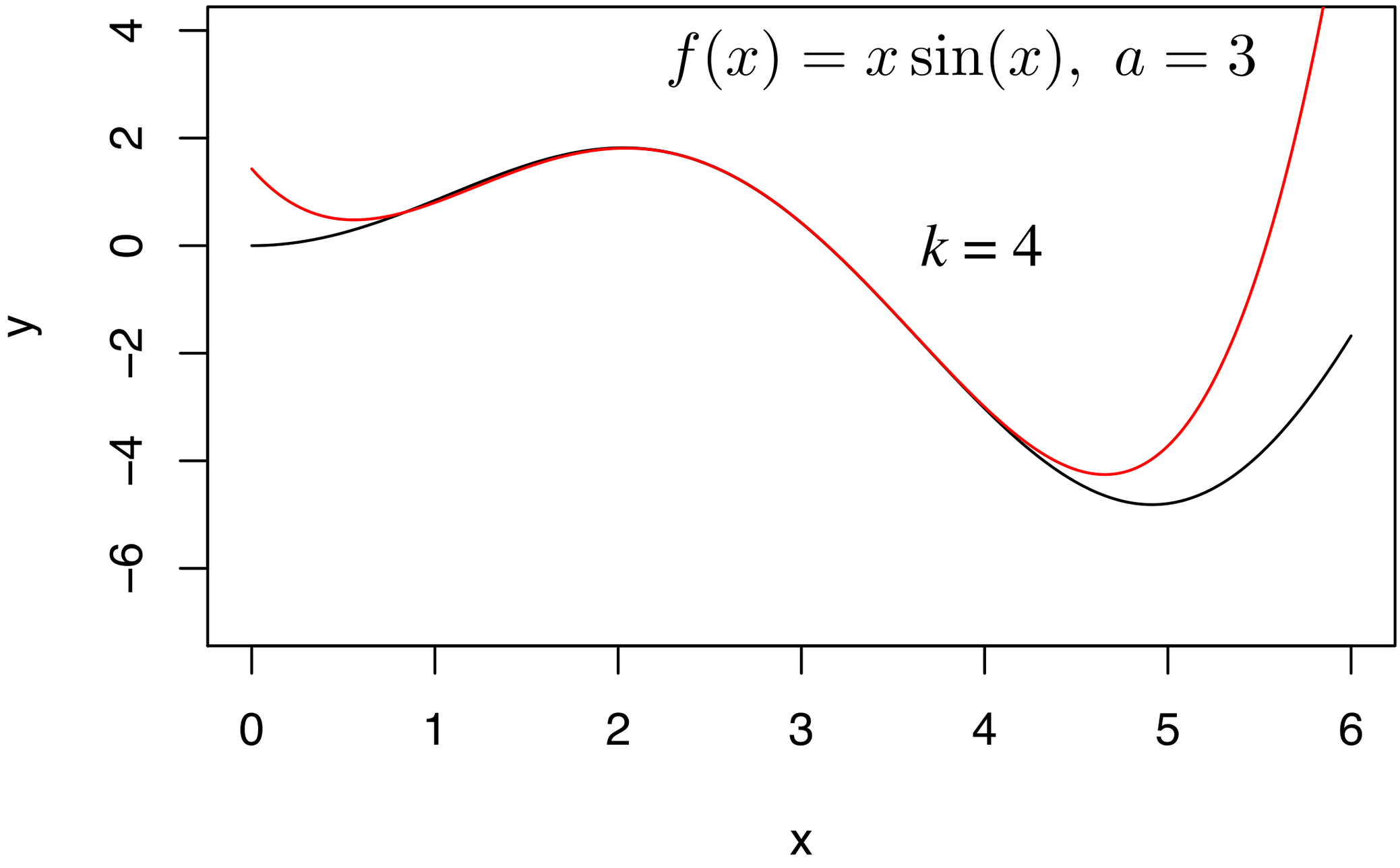
$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 + \frac{f''''(a)}{4!}(x-a)^4 + \cdots$$

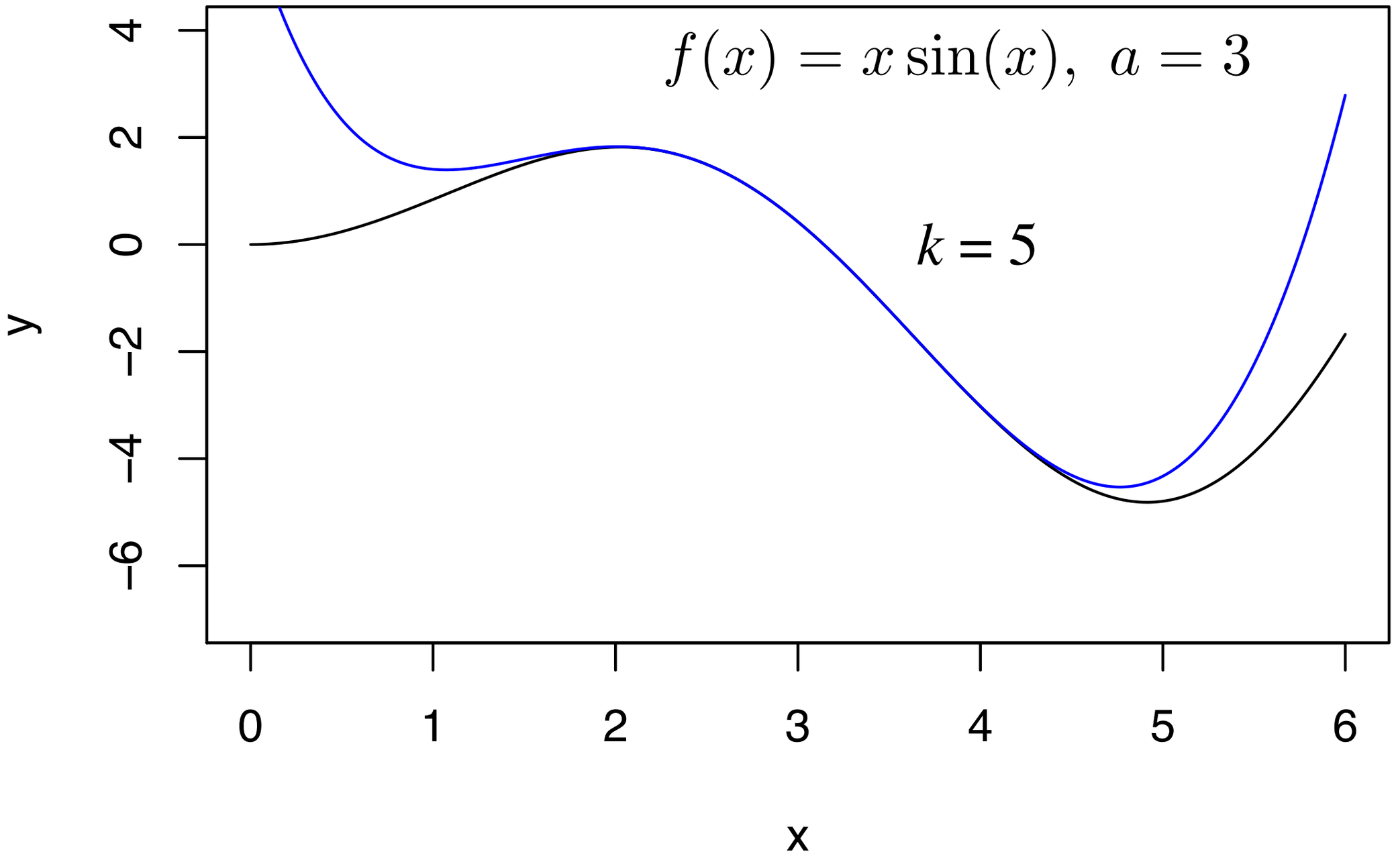
任意的足够平滑的函数 $f(x)$ 在 $x = a$ 附近都可以以有限项的多项式近似。因此，多项式可以用来描述多种非线性关系，哪怕真实的关系不是多项式。

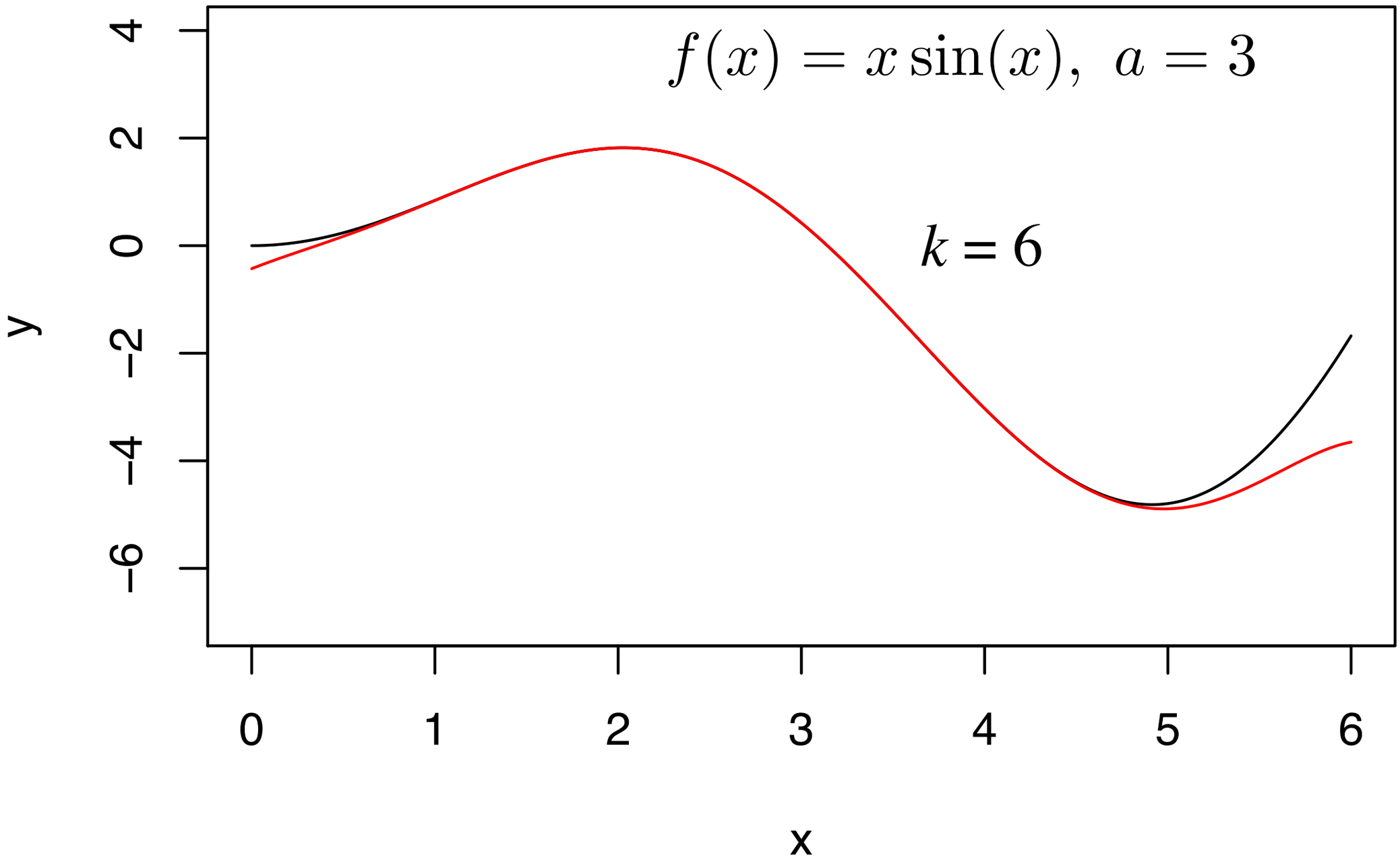

$f(x) = x\sin(x),\ a = 3$
$k = 1$
y
4
2
0
−2
−4
−6
0
1
2
3
4
5
6
x

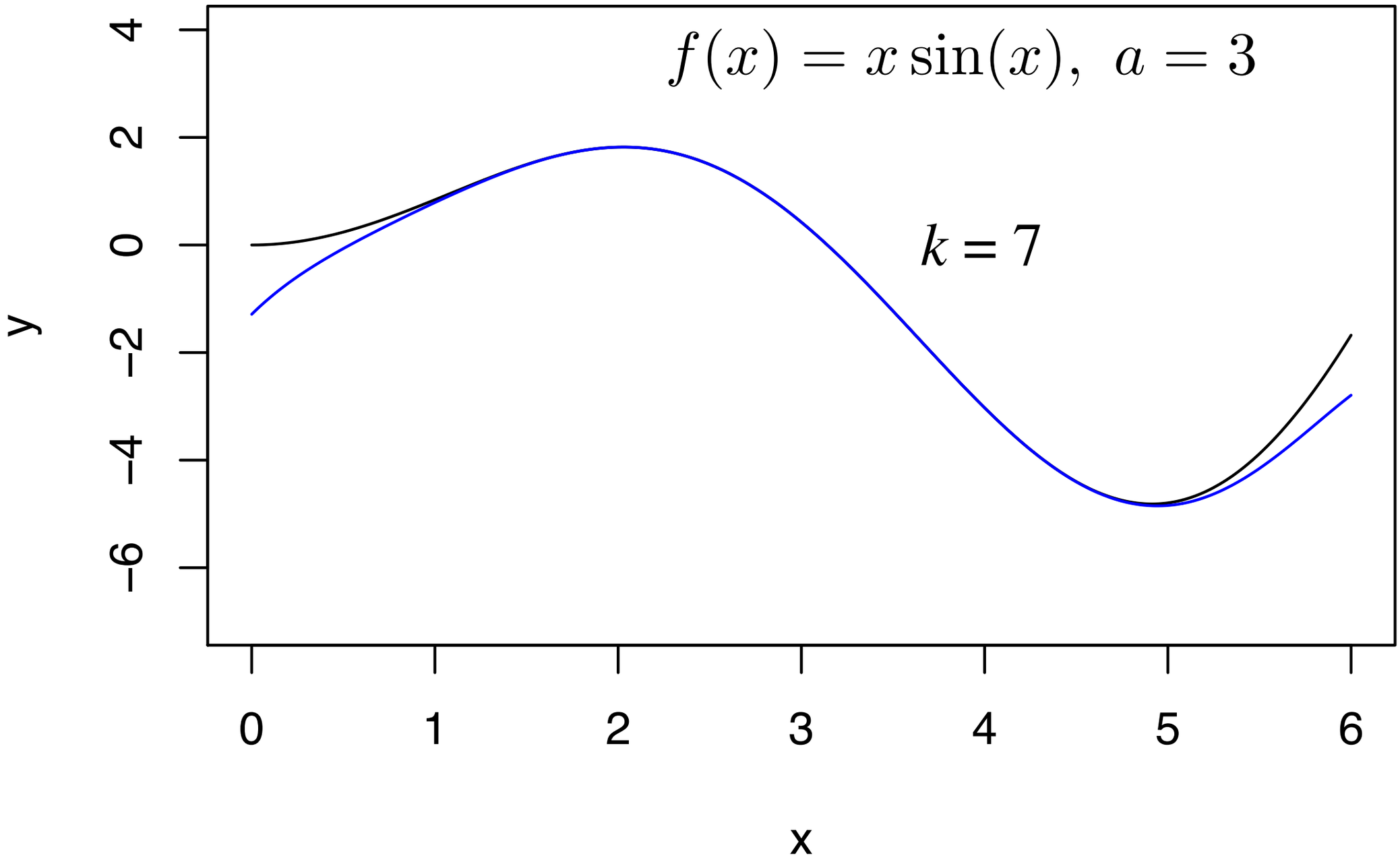

$f(x) = x\sin(x),\ a = 3$
$k = 2$
y
4
2
0
−2
−4
−6
0
1
2
3
4
5
6
x

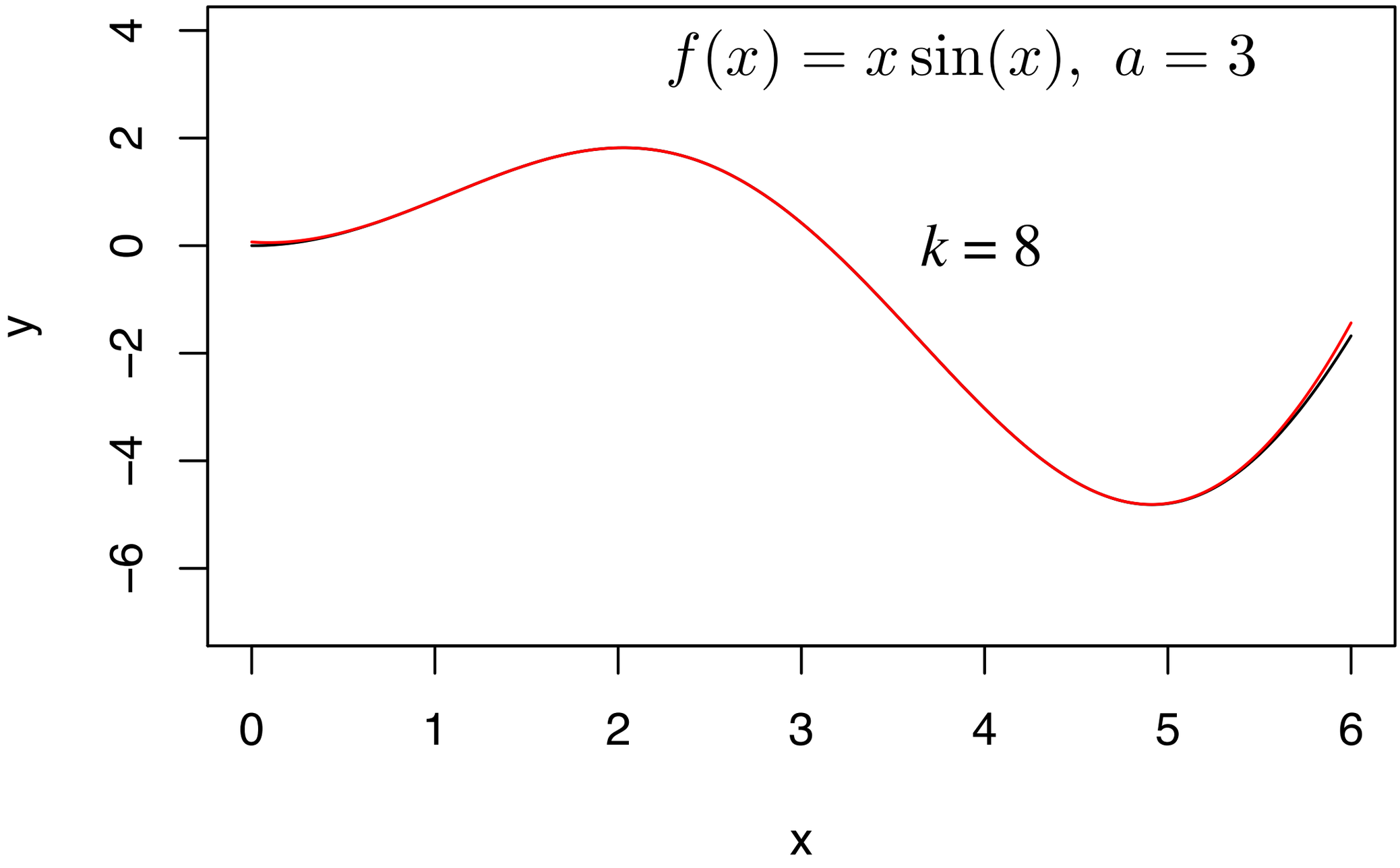

$f(x) = x\sin(x),\ a = 3$
$k = 3$
y
4
2
0
−2
−4
−6
0
1
2
3
4
5
6
x


$f(x) = x\sin(x),\ a = 3$
$k = 4$
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

关于回归中的多项式

- 回归模型

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

中 X_i^k 项的次数越高，该项对 Y_i 的影响越小。

- 回归函数中最高应包含到几次项需要在尝试中确定（参照 8.2 节），但是在实践中，往往加入二次项就足以反映变量间的非线性关系，且在经济理论中很难找到包含三次项以上的非线性关系。

- 虽然加入更高次项时的拟合效果一般会更好，但边际效应往往不高，且容易陷入过度拟合的状态。

- 因此，通常只需考虑二次回归函数。

练习

- 针对下列模型分别进行 OLS 回归

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

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

- 考察三次项 Income_i^3 的作用，即加入三次项对回归结果有怎样的影响。你认为有必要加入三次项吗?

gretl 命令及结果

事先生成二次项和三次项

```
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

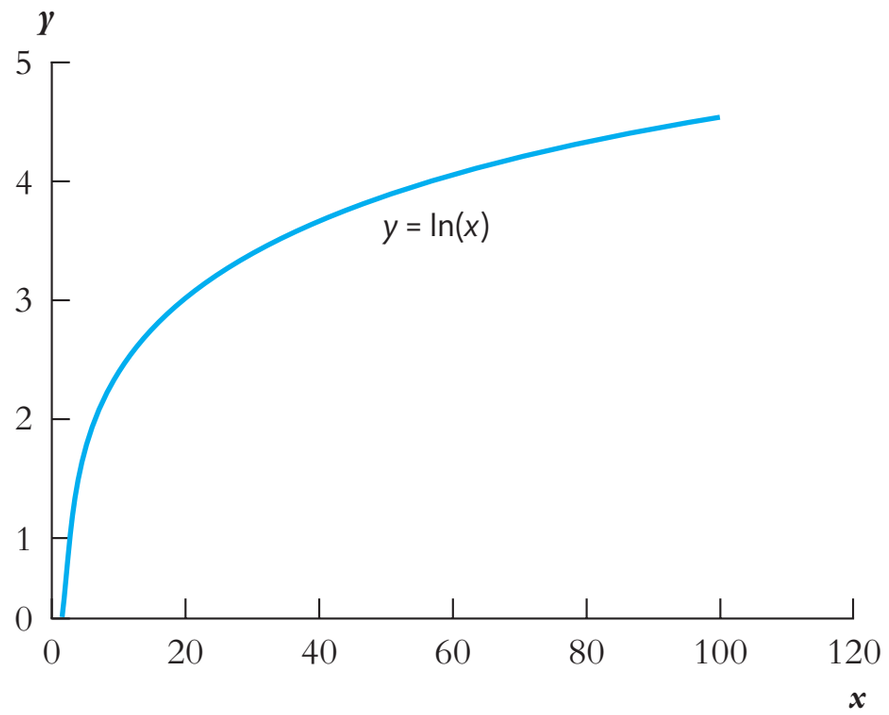
```
              coefficient   std. error      z      p-value
  ---------------------------------------------------------
  const       607.302       2.90175      209.3    0.0000   ***
  avginc        3.85099     0.268094      14.36   8.66e-47 ***
  sq_avginc    -0.0423085   0.00478034    -8.851  8.71e-19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

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   **
```

对数函数

Logarithms



- 对数函数是指数函数的反函数，即 $x = \ln(\exp(x))$
- 对数函数可以用来近似比例变化（proportional change），即当 Δx 很小时，

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

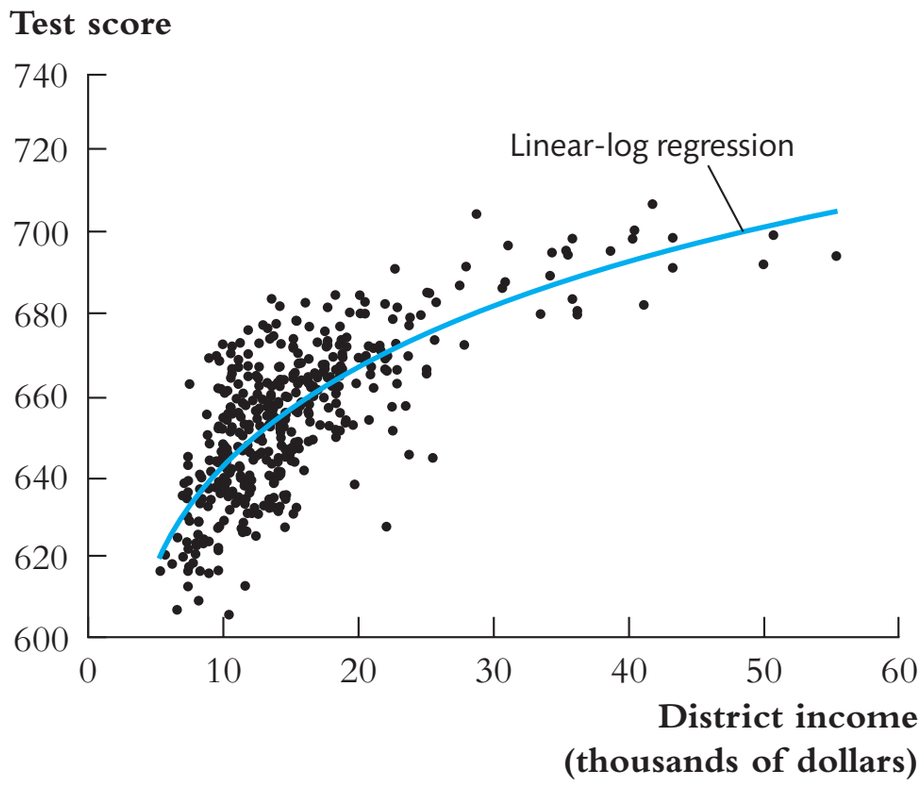
- 比例变化也可以用百分比的形式表述，即 $\frac{\Delta x}{x} \times 100$ 为 x 的百分比变化。在经济学中，收入或价格等变量更适合以百分比变化来表示，因此习惯上用其对数函数进行回归。

回归中的对数（一）：线性对数模型

The linear-log model

- 线性对数模型的定义是

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

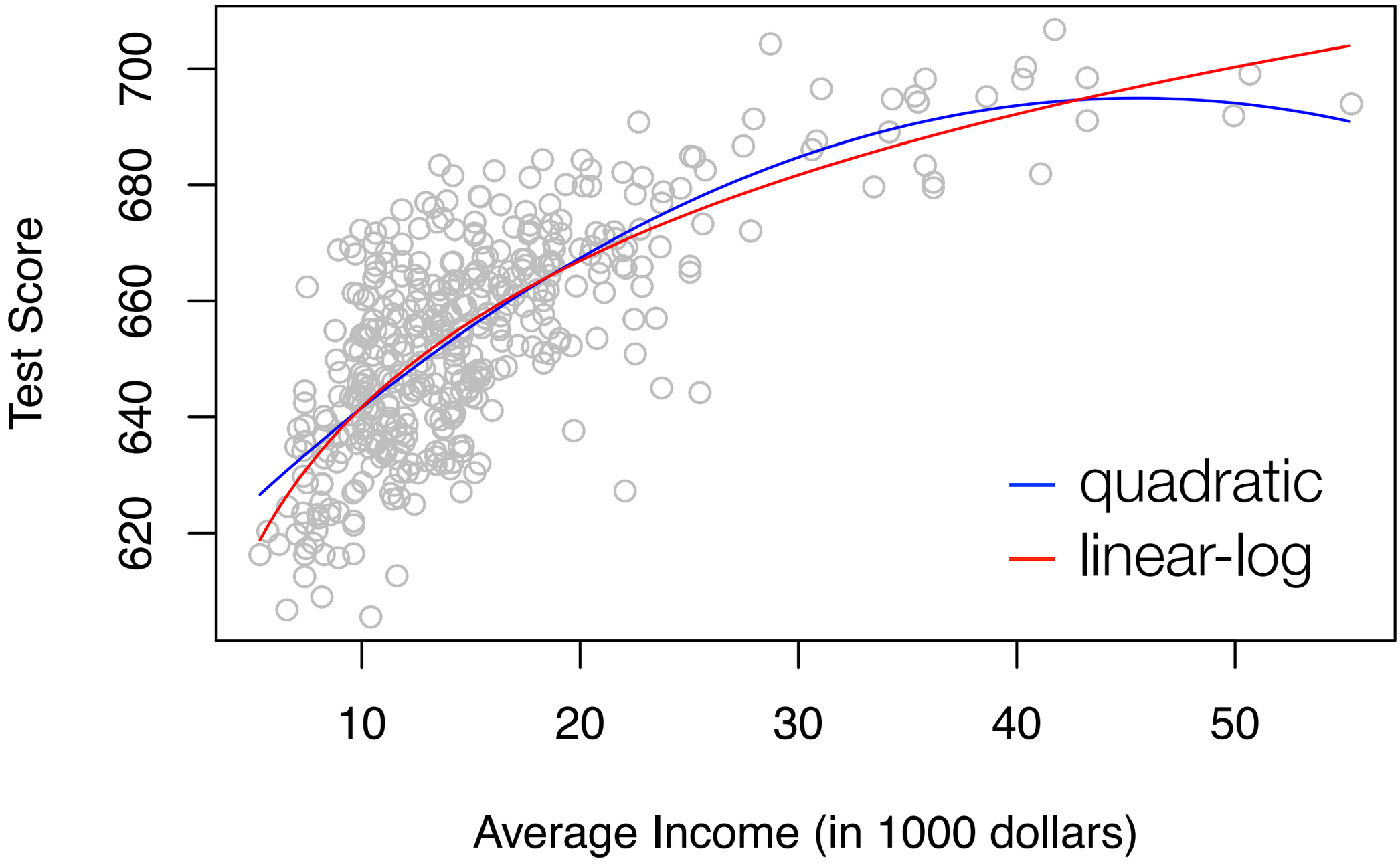


- 如果变量间的真实关系是

$$\exp(Y_i) = \beta_0^{\#} X_i^{\beta_1} v_i$$

则令 $\beta_1 = \ln(\beta_0^{\#})$，$u_i = \ln(v_i)$ 可推出线性对数模型。

- 线性对数模型中系数 β_1 代表 X_i 的比例变化引起的 Y_i 的变化。若换算成百分比就是当 X_i 变化1%时，Y_i 的变化为 $0.01\beta_1$。


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

回归中的对数（二）：对数线性模型

The log-linear model

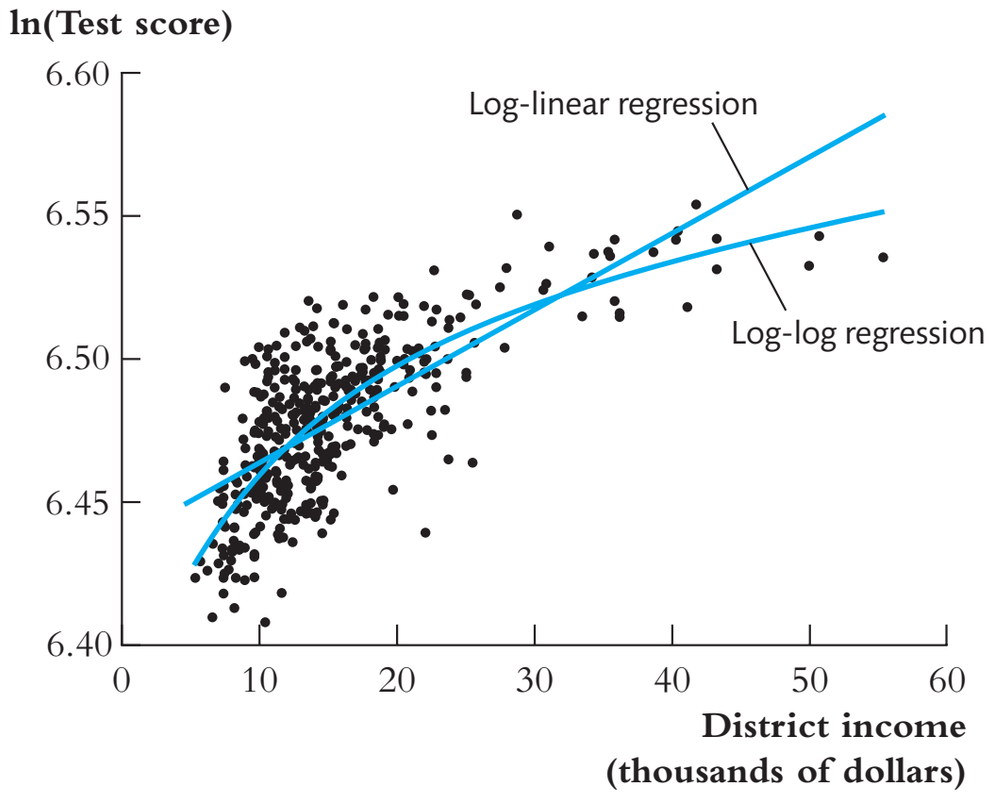
- 对数线性模型的定义是

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

- 如果变量间的真实关系是

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

则可推出对数线性模型。



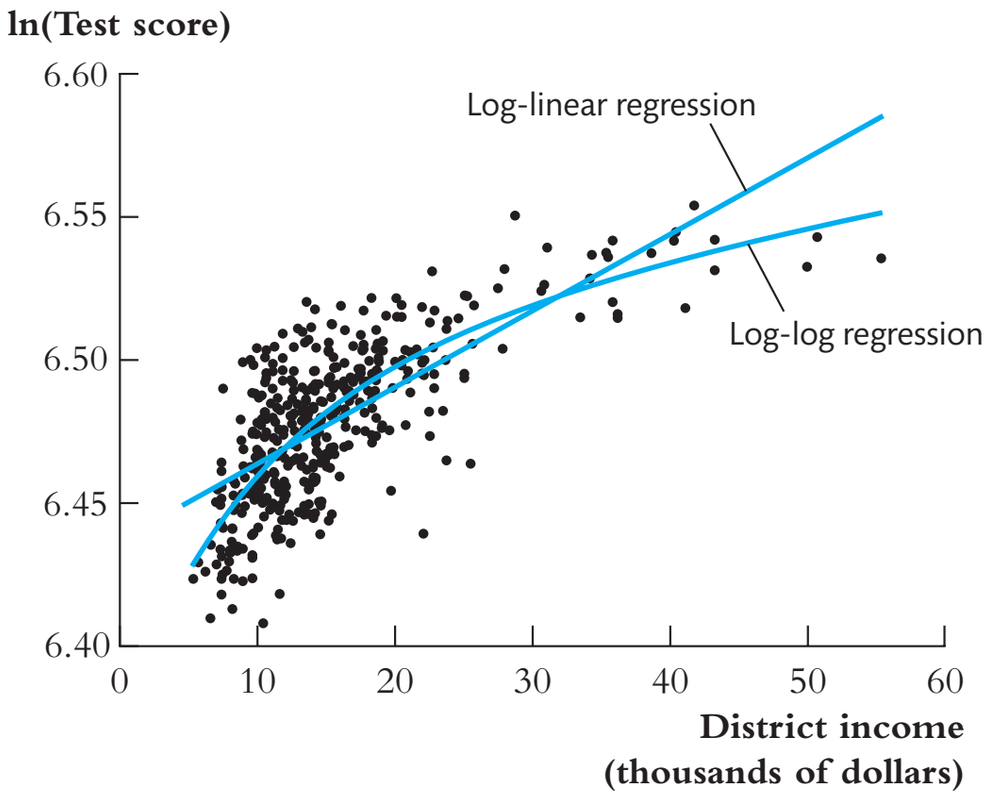
- 对数线性模型中系数 β_1 代表 X_i 的变化引起的 Y_i 的比例变化。若换算成百分比就是当 X_i 变化1个单位时，Y_i 的变化为 $100\beta_1\%$。

回归中的对数（三）：双对数模型

The log-log model

- 双对数模型的定义是

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



- 如果变量间的真实关系是

$$Y_i = \beta_0^{\#} X_i^{\beta_1} v_i$$

则令 $\beta_1 = \ln(\beta_0^{\#})$，$u_i = \ln(v_i)$ 可推出双对数模型。例如 Cobb-Douglas 生产函数 $Y = AK^{\alpha}L^{\beta}$ 可用双对数模型进行回归。

- 对数线性模型中系数 β_1 代表 X_i 的比例变化引起的 Y_i 的比例变化。在经济学中称为 Y_i 关于 X_i 的弹性。

例如，需求的价格弹性 = $\dfrac{\Delta Q/Q}{\Delta P/P}$

练习

- 在 8.2.2 小节中给出了含有对数函数的三种回归模型

$$\underset{(3.8)}{\widehat{TestScore}} = \underset{(1.40)}{557.8 + 36.42\ln(Income)}, \overline{R}^2 = 0.561. \tag{8.18}$$

$$\underset{(0.006)}{\widehat{\ln(TestScore)}} = \underset{(0.0021)}{6.336 + 0.0554\ln(Income)}, \overline{R}^2 = 0.557. \tag{8.23}$$

$$\underset{(0.003)}{\widehat{\ln(TestScore)}} = \underset{(0.00018)}{6.439 + 0.00284\,Income}, \overline{R}^2 = 0.497. \tag{8.24}$$

在 gretl 中尝试复制以上结果。

包含对数项时的注意事项

- 当因变量不同时（例如一个模型的因变量是 Y，另一个模型的因变量是 $\ln(Y)$），不能直接比较两个回归的 R^2 或调整 R^2。

  这是因为 R^2 的计算基于 SSR，而因变量单位不同时 SSR 的单位也不同，因此直接比较没有意义。

- 当因变量为 $\ln(Y)$ 时，可以利用回归估计预测值 $\widehat{\ln(Y)}$，但是不能简单的通过取指数 $\exp\left(\widehat{\ln(Y)}\right)$ 作为 Y 的估计值。

  例如在对数线性模型中，$Y_i = \exp(\beta_0 + \beta_1 X_i + u_i) = e^{\beta_0+\beta_1 X_i} e^{u_i}$，因此 $E(Y_i \mid X_i) = e^{\beta_0+\beta_1 X_i} E(e^{u_i} \mid X_i)$。但是，即使 $E(u_i \mid X_i) = 0$，$E(e^{u_i} \mid X_i) \neq 1$。是否能够正确预测 Y_i 取决于是否能够正确估计 $E(e^{u_i} \mid X_i)$ 的取值。

非线性最小二乘估计（详见附录 8.1）

Nonlinear least squares estimation

- 当回归函数不是回归系数的线性函数时，无法用 OLS 进行估计。例如：

$$Y_i = \beta_0[1 - e^{-\beta_1(X_i - \beta_2)}] + u_i$$

- 这种情况下可以用非线性最小二乘估计。非线性最小二乘估计量没有通用公式，需要用计算机寻找最优解。

- 在 gretl 中用 `nls` 命令或从菜单中选择

  `> Model > Nonlinear Least Squares`

  实现。

关于二次函数的错误应用

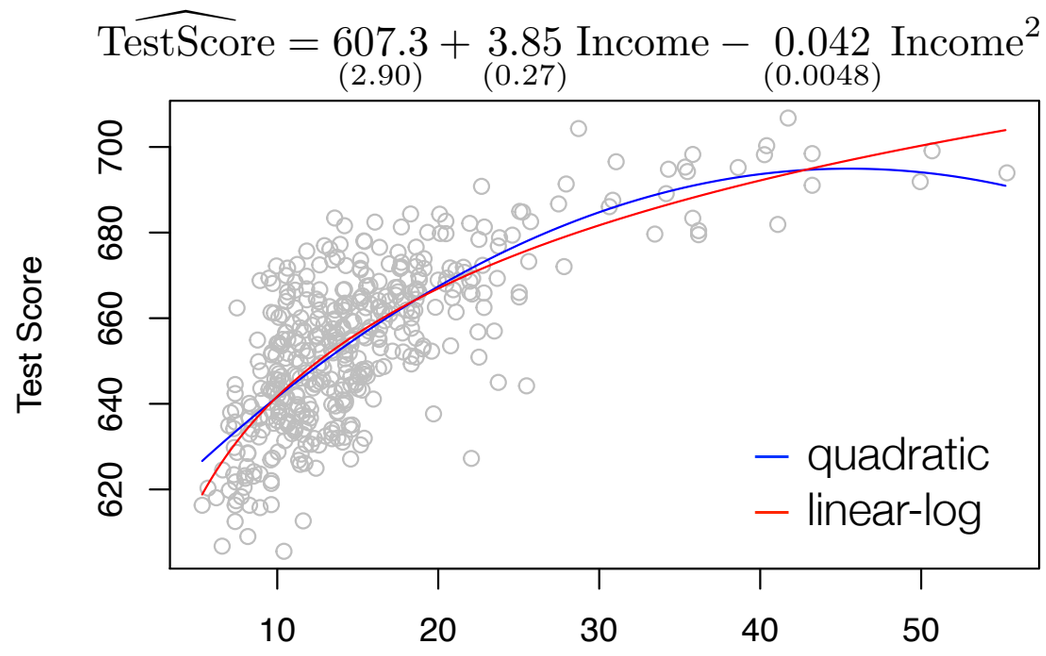
检验 U 型或倒 U 型关系

- 在实证研究中常有人针对 X 的 U 型或倒 U 型函数关系进行检验，例如库兹涅茨曲线（收入不平等程度随着经济发展先增后减）。

- 很多研究试图通过二次函数回归时二次项系数非零的统计学显著性证明 U 型或倒 U 型关系的存在，如周云波 (2009)。这种做法是错误的。

  - 当数据本不呈现 U 型或倒 U 型关系时，二次回归结果也可能显著，如右图。

  - 理论上二次关系成立，但数据中拐点还未出现时，以二次回归结果预测拐点的出现节点是不严谨的，而且也无法得出拐点的置信区间。



- 更加严谨有效的方法见 Lind & Mehlum (2010)。

自变量的交互作用

自变量的交互作用

Interactions between independent variables

- 有时一个自变量的变化对因变量的效应会随着其他自变量的取值而产生变化，例如
  - 某种药物不同剂量的效果可能因患者的年龄、体重、性别等特征的不同而不同
  - 宏观金融政策（如降低贷款利率等）的效果可能因各地区的社会、经济发展状态的不同而不同

- 在回归分析中，这种自变量间的影响被称为**交互作用（interaction）**，以自变量的乘积的形式体现，称为**交互项（interaction term）**。

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

交互项

交互作用的分类

- 按照自变量的种类，可以将交互作用分成三类：
  - 两个二值变量的交互作用

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

  - 连续变量和二值变量的交互作用

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

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

  - 两个连续变量的交互作用

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

两个二值变量的交互作用

- 当回归模型仅包含虚拟变量时，因变量的预测值可以解释为不同分类的组内均值。这同样适用于包含交互项的模型。

- 针对 $Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 (D_{1i} \times D_{2i}) + u_i$ 中自变量的不同取值计算条件期望，可得

$$
\begin{aligned}
E(Y_i \mid D_{1i} = 0,\, D_{2i} = 0) &= \beta_0 \\
E(Y_i \mid D_{1i} = 1,\, D_{2i} = 0) &= \beta_0 + \beta_1 \\
E(Y_i \mid D_{1i} = 0,\, D_{2i} = 1) &= \beta_0 + \beta_2 \\
E(Y_i \mid D_{1i} = 1,\, D_{2i} = 1) &= \beta_0 + \beta_1 + \beta_2 + \beta_3
\end{aligned}
$$

在 STAR 数据中的应用

- 令 HiSTR_i 为二值变量：当 $\mathrm{str}_i \geq 20$ 时取值为 1，否则为 0。
- 令 HiEL_i 为二值变量：当 $\mathrm{el\_pct}_i \geq 10$ 时取值为 1，否则为 0。
- 对以下模型进行 OLS 回归

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

- gretl 命令：

```
series hi_str = (str >= 20)
series hi_el = (el_pct >= 10)
series inter_hs_he = hi_str * hi_el
ols testscr const hi_str hi_el inter_hs_he --robust
```

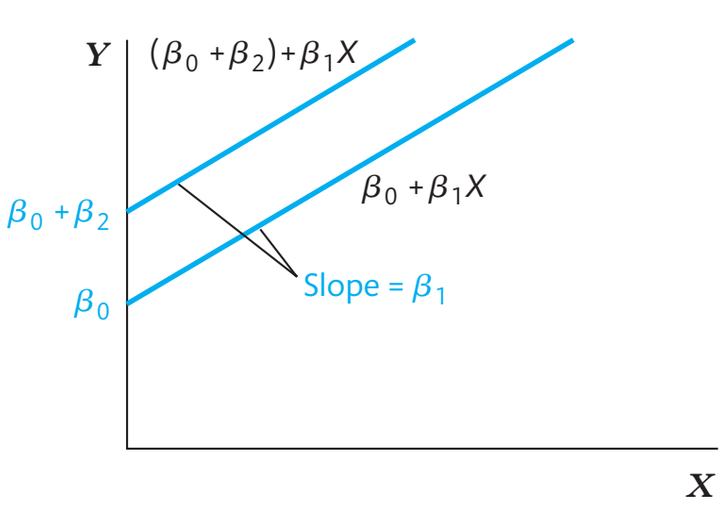
事先生成交互项

在 STAR 数据中的应用

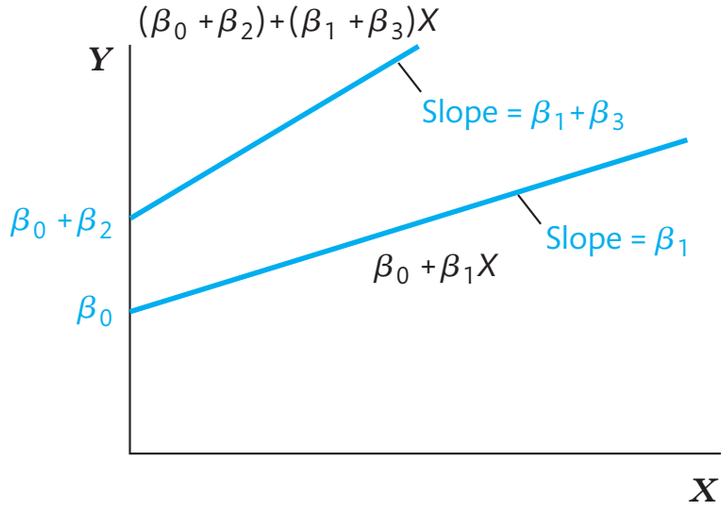
```
Model 5: OLS, using observations 1-420
Dependent variable: testscr
Heteroskedasticity-robust standard errors, variant HC1

                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
```

连续变量和二值变量的交互作用

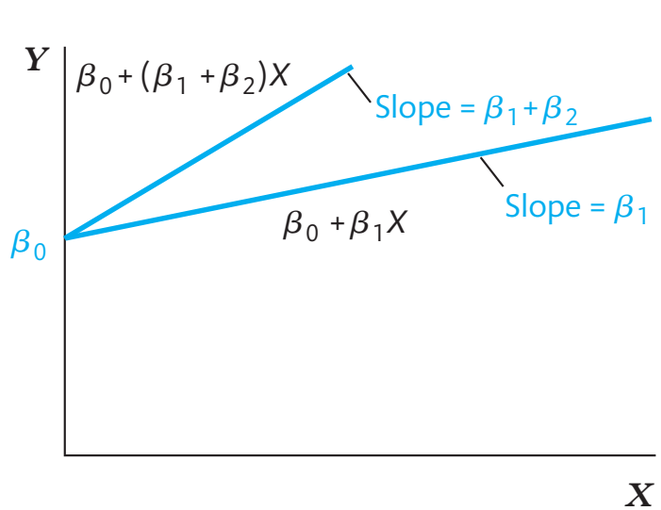
- 当回归模型包含连续变量和二值变量、以及两者的交互项时，回归函数的图形可以帮助我们立即回归系数的含义。



(a) Different intercepts, same slope



(b) Different intercepts, different slopes



(c) Same intercept, different slopes

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

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

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

两个连续变量的交互作用

- 在以下回归模型中，

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

交互项的存在使 X_1 的单位变化效应依赖于 X_2，反之亦然。即

  - 当 X_1 变化一个单位时，对应的 Y 的变化为 $\beta_1 + \beta_3 X_2$；
  - 当 X_2 变化一个单位时，对应的 Y 的变化为 $\beta_2 + \beta_3 X_1$。

- 当 X_1 和 X_2 同时增加一个单位时，对应的 Y 的变化为

$$\underbrace{(\beta_1 + \beta_3 X_2)}_{\text{固定 } X_2 \text{ 时 } X_1 \text{ 的变化效应}} + \underbrace{(\beta_2 + \beta_3 X_1)}_{\text{固定 } X_1 \text{ 时 } X_2 \text{ 的变化效应}} + \underbrace{\beta_3}_{X_1 \text{ 和 } X_2 \text{ 同时变化的附加效应}}$$

课后练习（不需提交）

- 阅读 8.4 节《学生教师比对测试成绩的非线性效应》，并尝试复制表 8.3 中的回归结果。

拓展阅读

- Kuznets, S., Economic Growth and Income Inequality, *American Economic Review*, 45(1):1-28, 1955.

- Acemoglu, D. and Robinson, J., The Political Economy of the Kuznets Curve, *Review of Development Economics*, 6(2):183-203, 2002.

- 周云波, 城市化、城乡差距以及全国居民总体收入差距的变动——收入差距倒 U 型假说的实证检验, 经济学（季刊）, 8(4), 2009.

- Lind, J. and Mehlum, H., With or Without U? The Appropriate Test for a U-Shaped Relationship, *Oxford Bulletin of Economics and Statistics*, 71(1), 2010.