Dealing with curvature

- Suppose we have paired numerical data
  \((X_1, Y_1), \ldots, (X_n, Y_n)\)
  and want to explain \(Y\) as a function of \(X\). (We work assuming that \(X\) is given, and focus on the distribution of \(Y \mid X\).)

- We saw how to fit a straight line, but the fit may be poor. Without resorting to testing, we might see this on a scatterplot... What can we do?

- Here are two of the main options:
  1. Transform the variables. In regression, we can apply different transformations to \(X\) and \(Y\). Popular transformations include the logarithm and the power functions.
  2. Augment (or expand) the model. In other words, enrich the model (add degrees of freedom) to better fit the data.

Polynomial regression

- A polynomial model is of the form:
  \[\mathbb{E}(Y \mid X = x) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p\]

- Fitting the model by least squares regression amounts to
  \[
  \text{minimize } \text{SSE}(b_0, b_1, \ldots, b_p) = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i - \cdots - b_p X_i^p)^2
  \]
  over \(b_0, b_1, \ldots, b_p \in \mathbb{R}\)

- Theory. The minimizer is unique if and only if the number of unique \(X_i\)'s greater than or equal to \(p + 1\) (the number of parameters).

- Assume the minimizer is unique and denote it by \((\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)\).
Under the standard assumptions on the measurement errors, this corresponds to maximum likelihood estimation.

We estimate \( \sigma^2 \) by
\[
\hat{\sigma}^2 = \frac{SSE(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)}{n - p - 1}.
\]

**Some issues**

Choosing the degree \( p \) is non-trivial.

The larger \( p \) is:
- the richer the model is (better fit to the data)
- the more variable the model is (more coefficients need to be estimated)

In any case, the model becomes quickly unstable as \( p \) increases.

**Splines**

Splines are a popular alternative to polynomials.

A spline of degree \( q \) with knots \( \xi_1, \ldots, \xi_K \in \mathbb{R} \) is a function such that
- It is a polynomial of degree \( q \) on each interval of the form \([\xi_k, \xi_{k+1}]\)
- It has continuous derivatives up to order \( q - 1 \).

The space of such functions is a linear space of dimension
\[
(K + 1)(q + 1) - Kq = K + q + 1
\]

In other words, it takes \( K + q + 1 \) parameters to determine such a function.

**Bases for splines**

The following is a basis for the space of splines of degree \( q \) and knots \( \xi_1, \ldots, \xi_K \):
\[
f_l(x) = x^l, \quad l = 0, \ldots, q
\]

and
\[
g_k(x) = \max(x - \xi_k, 0)^q
\]

This means that for any function \( f(x) \) in that space there are \( a_0, \ldots, a_q \in \mathbb{R} \) and \( b_1, \ldots, b_K \in \mathbb{R} \) such that
\[
f(x) = \sum_{l=0}^{q} a_l f_l(x) + \sum_{k=1}^{K} b_k g_k(x)
\]

In practice another basis (the B-splines basis) is preferred.
Fitting a spline model

□ Suppose that \( h_0, \ldots, h_{K+q} \) forms a basis for the space of splines of degree \( q \) with knots \( \xi_1, \ldots, \xi_K \in \mathbb{R} \).

□ A fit by least squares amounts to solving the following optimization problem

\[
\text{minimize} \quad \text{SSE}(c_0, \ldots, c_{K+q}) = \sum_{i=1}^{n} \left( Y_i - \sum_{j=0}^{K+q} c_j h_j(X_i) \right)^2 \\
\text{over} \quad c_0, \ldots, c_{K+1} \in \mathbb{R}
\]

An example of nonparametric regression

□ Consider a model of the form

\[
\mathbb{E}(Y \mid X = x) = f(x)
\]

where we only assume that \( f \) is twice differentiable.

□ Note that this model is nonparametric = 'not parametric'.

Smoothing splines

□ Consider the following optimization problem

\[
\text{minimize} \quad \sum_{i=1}^{n} (Y_i - g(X_i))^2 + \lambda \int g''(x)^2 \, dx \\
\text{over twice differentiable functions } g
\]

□ The parameter \( \lambda \) drives the degrees of freedom in the fit: the smaller \( \lambda \), the more flexible the model is.

\( \triangleright \) \( \lambda = 0 \) corresponds to \( n \) degrees of freedom if the \( X_i \)'s are all distinct.

\( \triangleright \) \( \lambda = \infty \) corresponds to 2 degrees of freedom (simple linear regression).

□ It turns out that there is always an optimizer among splines of degree 3 (cubic splines) with knots \( X_1, \ldots, X_n \).

□ The tuning parameter \( \lambda \) can be chosen based on the data. [More on that later.]
Other bases for model expansion

- Examples:
  - Trigonometric polynomials (i.e. Fourier basis).
  - Wavelets (and related families).

- In general, suppose we want to fit the model

\[ \mathbb{E}(Y \mid X = x) = \sum_{j=0}^{p} \beta_j f_j(x), \]

where \( f_0, \ldots, f_p \) are known functions.

- Then the model coefficients \( \beta = (\beta_0, \ldots, \beta_p) \) can be fitted by least-squares

\[
\text{minimize} \quad \text{SSE}(b_0, b_1, \ldots, b_p) = \sum_{i=1}^{n} \left( Y_i - \sum_{j=0}^{p} b_j f_j(x) \right)^2 \\
\text{over} \quad b_0, b_1, \ldots, b_p \in \mathbb{R}
\]