Polynomial Regression[1]

STA 302 Fall 2020

[1] See last slide for copyright information.

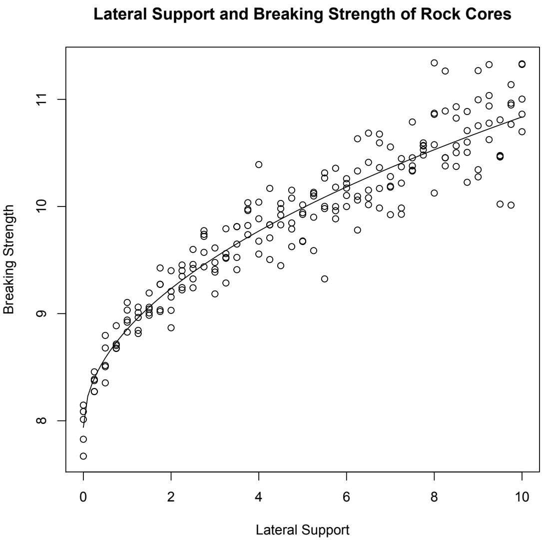
Overview

Fitting curves

- "Linear" regression means linear in the β_j.
- Model can allow for curviness in the x variable(s).
- Suppose you want to fit the curve $y = g(x)$ by least squares.
- Calculate a new x variable using $x^* = g(x)$.
- And do `lm(y ~ xstar)`.
- Implicitly, the model is $y_i = \beta_0 + \beta_1 g(x_i) + \epsilon_i$.
- If you are really sure $\beta_0 = 0$ and $\beta_1 = 1$, try `lm(y ~ 0 + offset(xstar))`.
- $H_0 : \beta_0 = 0, \beta_1 = 1$ is testable.

Fitting a curve: $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1\sqrt{x}$

Transform the explanatory variable



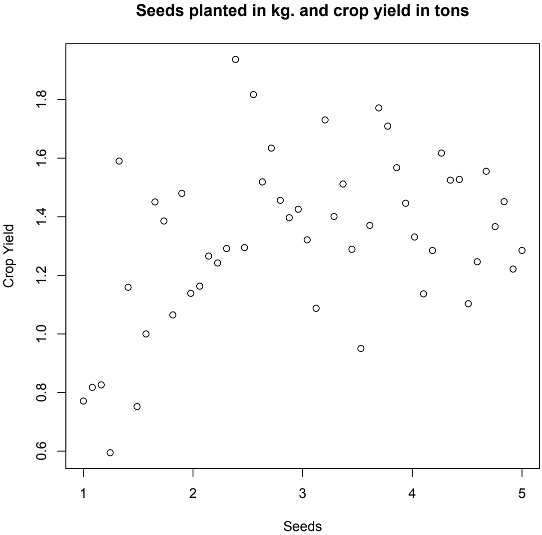
Taylor's Theorem

- What if you don't know $g(x)$, but want to allow for possible curviness?
- Taylor's Theorem says

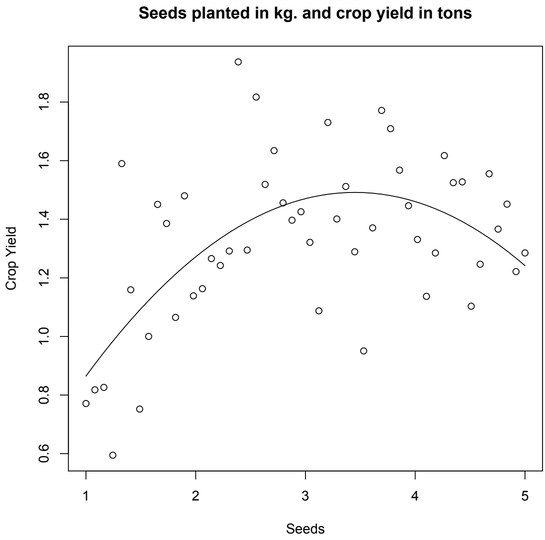
$$
\begin{aligned}
g(x) &= g(x_0) + g'(x_0)(x - x_0) + g''(x_0)\frac{(x - x_0)^2}{2!} \\
&+ g'''(x_0)\frac{(x - x_0)^3}{3!} + \cdots
\end{aligned}
$$

- The first several terms can approximate $g(x)$ quite well.
- The first several terms are a polynomial in x.
- So try something like `lm(y ~ x + xsq + xpow3)`.

Try a Quadratic



Try a Quadratic

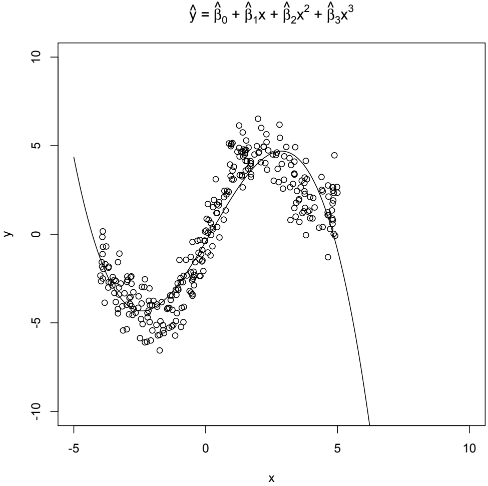


Diminishing Returns

```
> Seedsq = Seeds^2
> mod = lm(Yield ~ Seeds + Seedsq); summary(mod)
Call:
lm(formula = Yield ~ Seeds + Seedsq)

Residuals:
     Min       1Q   Median       3Q      Max
-0.54050 -0.12593 -0.04656  0.15393  0.56948

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.2487     0.2217   1.122 0.267621
Seeds         0.7202     0.1620   4.445 5.34e-05 ***
Seedsq       -0.1043     0.0266  -3.922 0.000284 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.2333 on 47 degrees of freedom
Multiple R-squared:  0.3623,Adjusted R-squared:  0.3352
F-statistic: 13.35 on 2 and 47 DF,  p-value: 2.559e-05
```

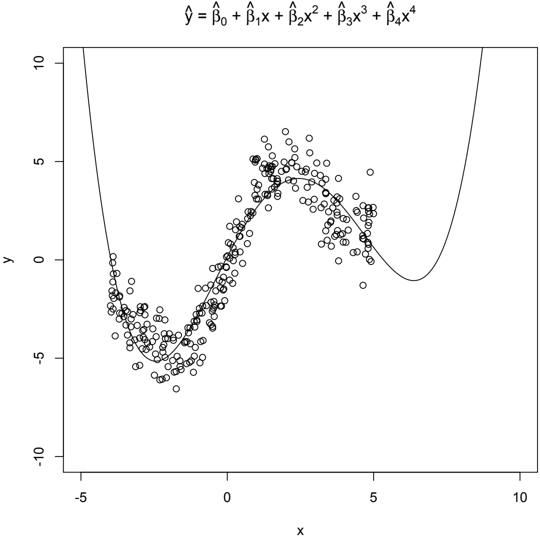
Warning

Polynomial regression can give very bad predictions outside the range of the data.



Even worse

An extra term, which is statistically significant



Response Surface Methodology

- Have an experiment with different dosages of two drugs, or different amounts of water and fertilizer.
- What is the optimal combination?
- Include quadratic terms *and* an interaction:

$$\begin{aligned} y = \beta_0 \quad & + \quad \beta_1 x_1 + \beta_2 x_1^2 \\ & + \quad \beta_3 x_2 + \beta_4 x_2^2 \\ & + \quad \beta_5 x_1 x_2 \qquad\qquad + \epsilon \end{aligned}$$

- The result is a curvy surface that could have a maximum or minimum.
- Estimate the β_j parameters, differentiate $E(y)$ with respect to x_1 and x_2, set the derivatives to zero, and solve.

Solution

$$\widehat{x}_1 = \frac{2\,\widehat{\beta}_1\widehat{\beta}_4 - \widehat{\beta}_3\widehat{\beta}_5}{\widehat{\beta}_5^2 - 4\,\widehat{\beta}_2\widehat{\beta}_4}$$

$$\widehat{x}_2 = \frac{2\,\widehat{\beta}_2\widehat{\beta}_3 - \widehat{\beta}_1\widehat{\beta}_5}{\widehat{\beta}_5^2 - 4\,\widehat{\beta}_2\widehat{\beta}_4}$$

- $\widehat{x}_1$ and $\widehat{x}_2$ really are estimates – estimates of non-linear functions of the β_j
- There's more to it – for example checking that it's really a maximum.
- Is the answer in the range of the data?

Copyright Information