Quantitatiive & Qualitative Predictors

AU STAT-615

Emil Hvitfeldt

2021-03-24

Polynomial Regression Models

One predictor variable - second order

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$$

where $X_i = X_i - \bar{X}$

We are centering the predictor variable because X and X^2 may be highly correlated and thus $\mathbf{X}^T\mathbf{X}$ will be very difficult to invert. This can lead to computational issues

Notation

Most of the time we use the following notation

$$Y_i = \beta_0 + \beta_1 X_i + \beta_{11} X_i^2 + \varepsilon_i$$

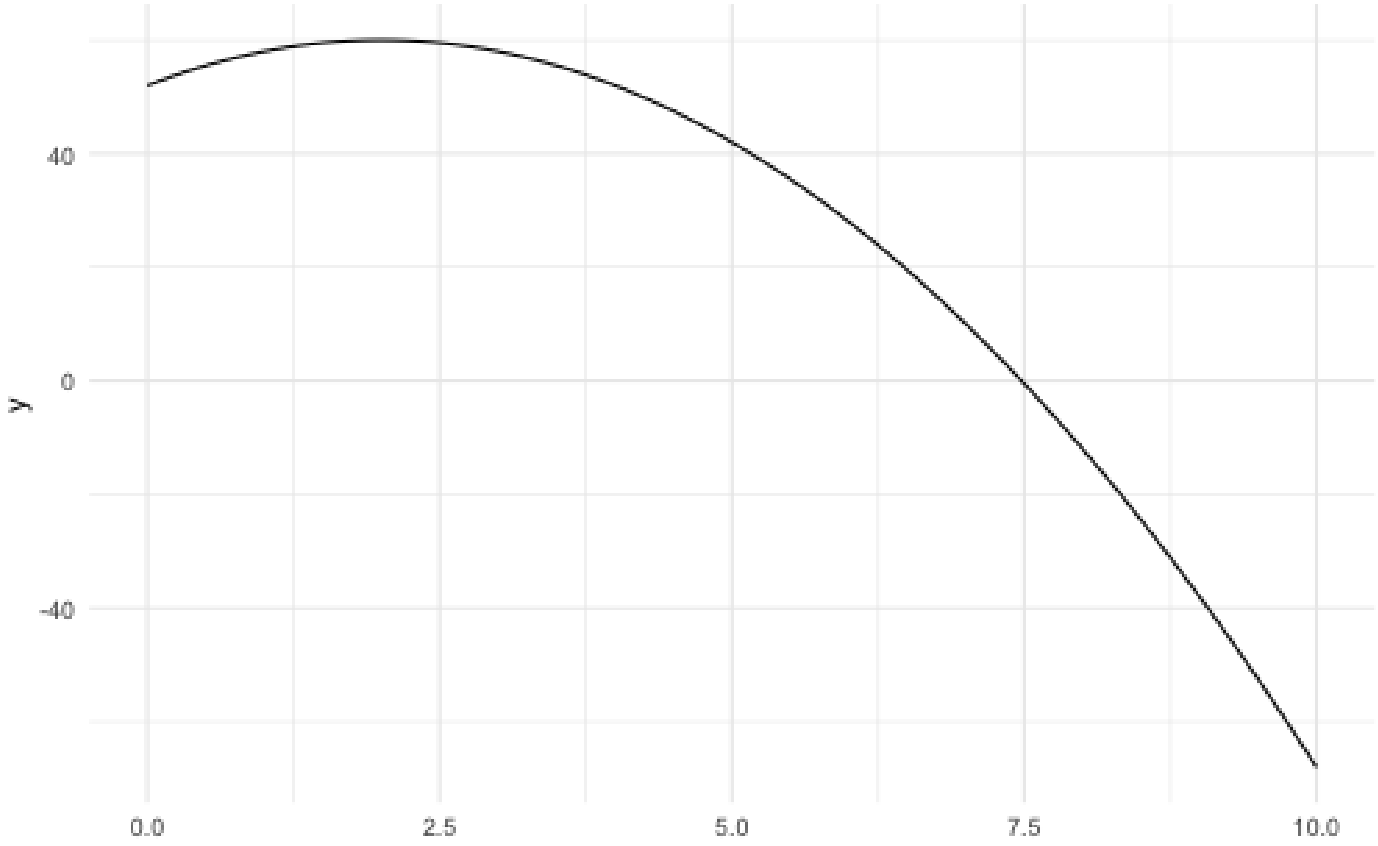
and

$$E\{Y\} = \beta_0 + \beta_1 X_i + \beta_{11} X_i^2$$

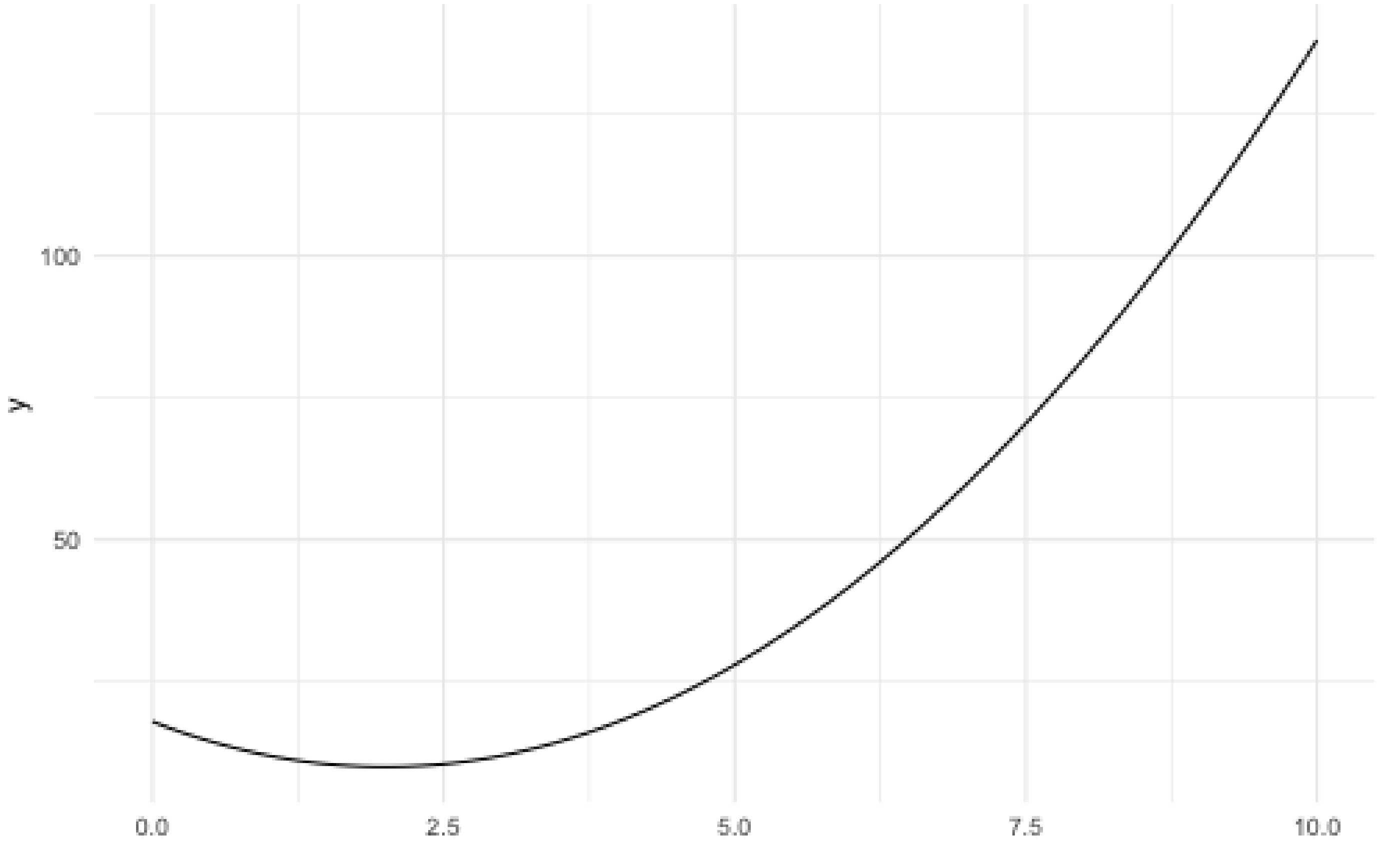
This is done to put emphasize on the exponents

- $\beta_1 \to$ linear effect coefficient
- $\beta_{11} \to$ quadratic effect coefficient

$$E\{Y\} = 52 + 8x - 2x^2$$



$$E\{Y\} = 18 - 8x + 2x^2$$

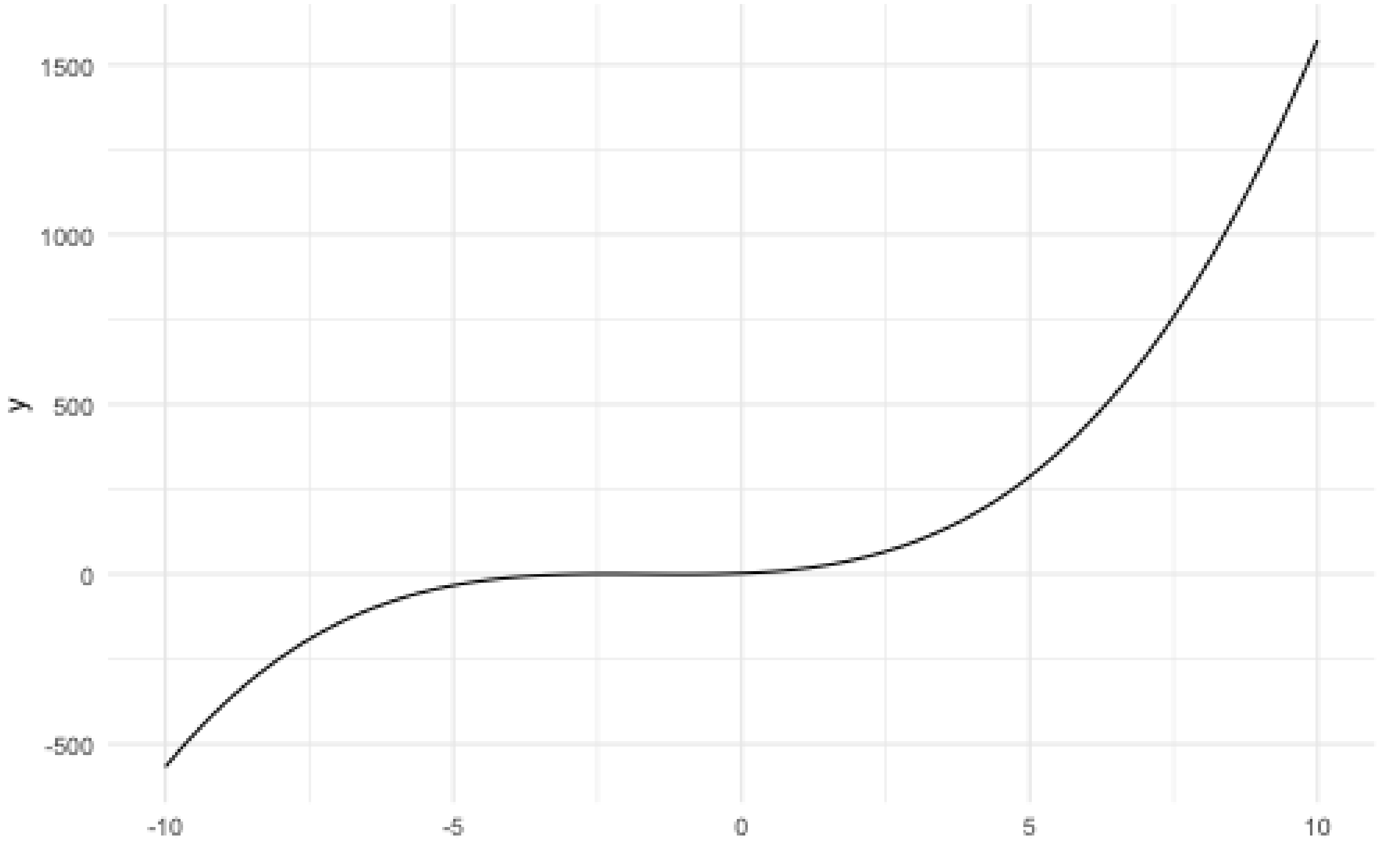


One predictor - third orders

$$Y_i = \beta_0 + \beta_1 X_i + \beta_{11} X_i^2 + \beta_{111} X_i^3 + \varepsilon_i$$

where $X_i = X_i - \bar{X}$

$$E\{Y\} = 2 + 7x + 5x^2 + x^3$$



Polynomial regression model

Models can become more complicated. For instance we can consider

two predictor variables - second order

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_{11} X_{i1}^2 + \beta_{22} X_{i2}^2 + \beta_{12} X_{i1} X_{i2}$$

Interaction Regression models

Let us have $p-1$ predictor variables. A regression model containns additive effects if the response function can be written in the form

$$E\{Y\}f_1(X_1)+f_2(X_2)+\cdots+f_{p-1}(X_{p-1})$$

Interaction Regression models

Example

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2$$

- $f_1(X_1)$
- $f_2(X_2)$

Interaction Regression models

On the other hand, if we have

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

Cannot be expressed in the previous form

This model contains interactionn effects

Interaction Regression models

On the other hand, if we have

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

This is called linear-by-linear or bilinear interaction ter or simply interraction term

Interpretation

The regressionn model for two quanntitative predictor variables with linear effects on Y and interacting effect on X_1 and X_2 on Y represented by a cross product is as follows

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i$$

Interpretation

Note: The regressionn coefficients β_1 and β_2 no longer indicate the change in the mean response with a unit increase of the predictor variable with the other predictor variable held constant at any given level

It can be shown that the change in the mean response with a unit increase in X_1 when X_2 is held constant is

$$\beta_1 + \beta_2 X_2$$

Qualitative Predictors

Example of qualitative predictors

$$X_2 = \begin{cases} 1 & \text{If stock company} \\ 0 & \text{Otherwise} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{If mutual company} \\ 0 & \text{Otherwise} \end{cases}$$

Qualitative Predictors

In order to define the qualitative variables, we used indicator functions and generate the indicator variables or dummy variable

Qualitative Predictors

Let

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

Where Y indicates the speed with which a particular insurance innovation is adopted

X_1 is the size of the firm annd X_2 and X_3 indicate the type of firm

Qualitative Predictors

Let us assume that we have $n = 4$ observations

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & 1 & 0 \\ 1 & X_{21} & 1 & 0 \\ 1 & X_{31} & 0 & 1 \\ 1 & X_{41} & 0 & 1 \end{bmatrix}$$

Qualitative Predictors

Note that

$$\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = 1 \cdot \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} + 1 \cdot \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

Which implies that the columns are linearly dependent

Thus we cannot invert $\mathbf{X}^T\mathbf{X}$, so we cannot have unique solutions for the estimator.

Solution: Drop one of the indicator variables

Qualitative Predictors

Note: A qualitative variable with c classes will be represented by $c - 1$ indicator variables, each taking on the values 0 and 1.

Interpretationn of Regression coefficents

Suppose that we drop the indicator variable X_3 from the model

Then we have

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

Case 1: Mutual firms

$$E\{Y\} = \beta_0 + \beta_1 X_1$$

Case 2: Stock firms

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 \cdot 1$$

Interpretationn of Regression coefficents

The critical question here is why we do not simply fit seperate regressions for stock firms and mutual forms and instead we adopted the approach of fitting one regression with an indicator variable

Reason 1

Since the model assumes equal slopes and same constant error term variancee for each type of firm, the common slope β_1 can best bbe estimated by pooling the two types of firms

Interpretationn of Regression coefficents

The critical question here is why we do not simply fit seperate regressions for stock firms and mutual forms and instead we adopted the approach of fitting one regression with an indicator variable

Reason 2

Other inferences such as for β_0 and β_2 can be made more precisely by working with one regression model containing an indicator variable since more degrees of freedom will be associated with MSE

We want a small MSE, so we nneed to devide by more degrees of freedom

Qualitative Predictors

If a qualitative variable has more than two classes, we requiree additionnal indicator variables in the regressionn model

$$X_2 = \begin{cases} 1 & \text{If } M_1 \\ 0 & \text{Otherwise} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{If } M_2 \\ 0 & \text{Otherwise} \end{cases}$$

$$X_4 = \begin{cases} 1 & \text{If } M_3 \\ 0 & \text{Otherwise} \end{cases}$$

And we can work in the same way we did previously

Alternatives to indicator variables

Consider the following table

Class	X_1
Frequent user	3
Occasional	2
Non user	1

Alternatives to indicator variables

The allocated codes that define the metric may not be reasonable as a quantitative variable

The mean response would change by the same amount when going from a non user to an occasional user as when going from a occasional user to a frequent user

Indicator variables

Indicator variables can be used even if the predictor variable is quantitative

For example

If we have data regarding ages of people, then we can arrrange the groups such as

- under 21
- 21-34
- 35-49
- 50-65
- over 65

Indicator variables

$$X_2 = \begin{cases} 1 & \text{If stock company} \\ -1 & \text{If mutual company} \end{cases}$$

here a meaningful test will be $H_0 : \beta_2 = 0$ vs $H_\alpha : \beta_2 \neq 0$

since the two sides would be equal to each other when $\beta_2 = 0$

Inteeractionn between qualitative and quantitative predictors

For example

$$X_{i1} = \text{size of firm}$$
$$X_{i2} = \begin{cases} 1 & \text{If stock company} \\ 0 & \text{otherwise} \end{cases}$$

We can have

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i \\ E\{Y\} &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 \end{aligned}$$

Inteeractionn between qualitative and quantitative predictors

For mutual firm ($X_2 = 0$)

$$E\{Y\} = \beta_0 + \beta_1 X_1$$

For stock firm ($X_2 = 0$)

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 + \beta_3 X_1 = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1$$