Regression Models for Quantitative & Qualitative Predictors

Polynomial regression models

2 uses

1) When curvilinear response <u>is</u> polynomial
2) " " " " unknown but fit well by a polynomial.

<u>Danger</u> extrapolation in polynomial models may be dangerous.

<u>Model Types</u>

1) One predictor var. - second order

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

where

$$x_i = X_i - \bar{X} \quad .$$

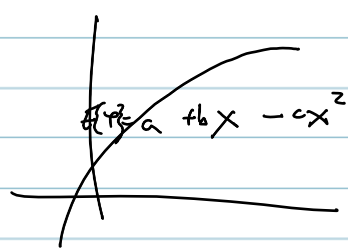
"Centering" vars reduces multicolinearity substantially

<u>Notation</u> (note β indexing)

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

The response function is

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



β_0 - is the intercept as before

β_1 - linear effect coefficient

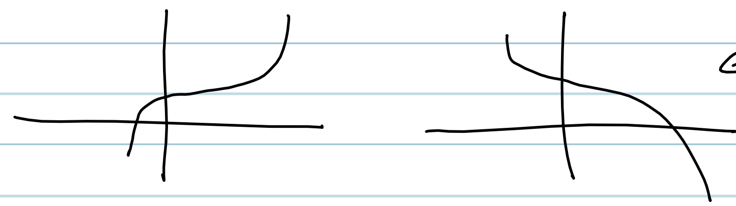
β_2 - quadratic effect coefficient

Third order models

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

where

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



cubic functions

Note: higher orders always improve fit but parameters become highly sensitive to noise and are harder to interpret.

Two predictor vars - second order

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

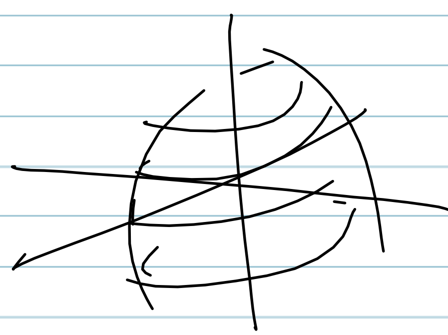
Where

$$x_{i1} = X_{i1} - \bar{X}_1$$
$$x_{i2} = X_{i2} - \bar{X}_2$$

The response function is a conic section

The coefficient β_{12} is called the interaction effect coefficient.

Ex



$$\hat{y} = b_0 + b_1 x + b_{11} x^2$$

$$= b_0 + b_1 (x - \bar{x}) + b_{11} (x - \bar{x})^2$$

$$= b_0 + b_1 x - b_1 \bar{x} + b_{11} (x^2 - 2x\bar{x} + \bar{x}^2)$$

$$= b_0 + b_1 x - b_1 \bar{x} + b_{11} x^2 - 2 b_{11} x \bar{x} + b_{11} \bar{x}^2$$

$$= (b_0 - b_1 \bar{x} + b_{11} \bar{x}^2) + (b_1 - 2 b_{11}) x + b_{11} x^2$$

Implementation of Poly Regression models

Fitting them requires nothing new

e.g.

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ \vdots \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & X_{11}^2 & X_{11}\cdot X_{12} \\ 1 & X_{21} & X_{22} & X_{21}^2 & X_{21}\cdot X_{22} \\ 1 & X_{31} & X_{32} & X_{31}^2 & X_{31}\cdot X_{32} \\ 1 & X_{41} & X_{42} & X_{41}^2 & X_{41}\cdot X_{42} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_{11} \\ \beta_{12} \\ \vdots \\ \vdots \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \vdots \\ \vdots \end{bmatrix}$$

leads to

$$b = (X'X)^{-1} X' Y$$ and tests

as usual.

<u>Model selection</u>: hierarchical approach: it is natural to include vars using a sequential selection process from lower-order to higher order terms.

For instance the model

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

can be fitted with the variables ordered in this way and with partial sums of squares F-tests used to test whether or not the coefficient of the most highest order term is zero. No further terms are considered (why? think about this.)

Regression function in terms of non-centered vars

If we fit

$$\hat{Y} = b_0 + b_1 x + b_{11} x^2 \quad \text{where } x = X - \bar{X}$$

then

$$\hat{Y} = b_0' + b_1' X + b_{11}' X^2$$

where

$$b_{11}' = b_{11}, \quad b_1' = b_1 - 2b_{11}\bar{X}, \quad b_0' = b_0 - b_1\bar{X} + b_{11}\bar{X}^2$$

i.e. the regression func. can be expressed in terms of the original vars

Comments: - Poly. models can be tough; multicolinearity even when centered.

- Tests not as powerful because extra terms eat up degrees of freedom, etc.

Interaction regression models

Terms, interpretation, fitting, etc.

A regression model with p-1 pred. vars contains additive effects if the response func. can be written in the form

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

where f_i, $1 \le i \le p-1$ can be any functions.

For instance

$$E\{Y\} = \underbrace{\beta_0 + \beta_1 X_1 + \beta_2 X_1^2}_{f_1(X_1)} + \underbrace{\beta_3 X_2}_{f_2(X_2)}$$

has effects X_1 & X_2 which are additive of Y.

The reg. func.

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

~~does~~ is not an additive effects model because it contains an interaction effect.

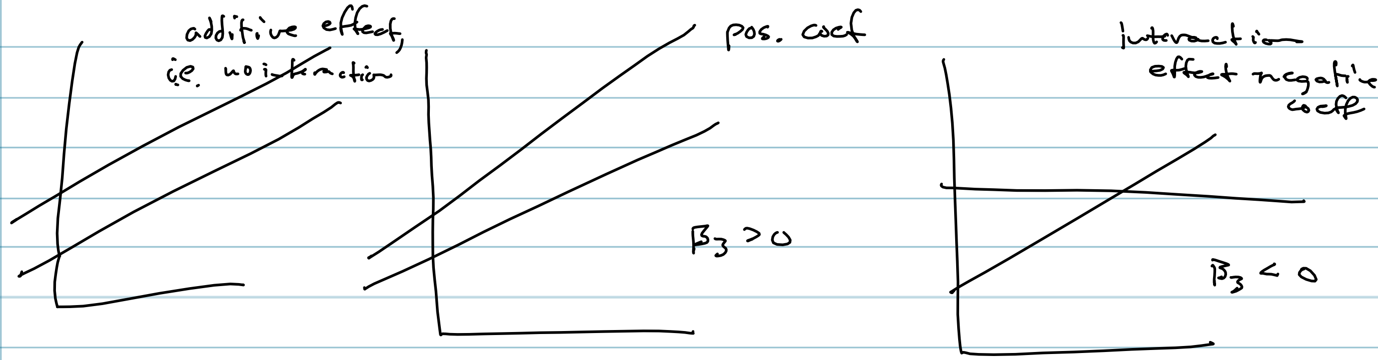
The cross-product term is called an interaction term.

Interpretation of Regression coeffs

Consider
$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i$$
The effects are given by

$$\partial Y_i / \partial X_{i1} = \beta_1 + \beta_3 X_{i2} \Rightarrow \text{the level of the second input var affects slope.}$$

and vice versa.



Note: 1) interaction terms often exhibit high multi-collinearity. Centering predictors individually again helps.

2) the number of potential interaction terms can be quite high $\binom{p}{2}$ for second-order interactions – could need a large amount of data to fit the corresponding model (big n)

Using a priori knowledge is not a bad way to go here. One can plot residuals of the additive affect model against interaction terms to get a sense of which vars matter.

¡muy IMPORTANTE!

Qualitative predictors $\boxed{\text{¥ KEY}}$

Qualitative vars are discrete; gender $\in$ {male, female}, disability status (not disabled, partly disabled, fully dissabled}

One way to identify the classes of a qualitative variables is to use indicator var's that take the values 0 & 1.

For instance if data $X_1, \dots, X_{N_1}$ come from class A and data $X_{N_1+1}, \dots, X_{N_1+N_2}$ come from calss B we can choose class A = 0 & class B = 1

design matrix

$$E\left\{\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix}\right\} = \begin{bmatrix} 1 & X_1 & 0 \\ 1 & X_2 & 0 \\ 1 & X_3 & 0 \\ & \vdots & \\ 1 & X_{N_1} & 0 \\ & \vdots & \\ 1 & \vdots & 1 \\ & & 1 \\ 1 & & 1 \\ & & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

Note: a qualitative var. with c classes can be represented with c-1 indicator variables.

Interpretting regression models with qualitative predictors

If $X_{i1} \in \mathbb{R}$ and $X_{i2} \in \{0,1\}$ and we use the regression model

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

then the response function is

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

For $X_2 = 0$ this reduces to

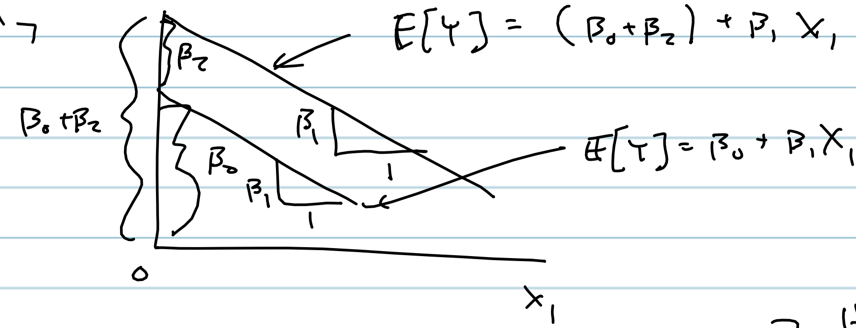
$$E[Y] = \beta_0 + \beta_2 X_1$$

<u>but</u> for $X_2 = 1$ this reduces to

$$E[Y] = (\beta_0 + \beta_2) + \beta_1 X_1$$

so intercept shifts but slope is the same.

Graphically



So a formal test of $H_0: \beta_2 = 0$, $H_a: \beta_2 \neq 0$ } How? F-test, t-test

effectively asks if the class of the qualitative variable has an effect on the regression relationship, in particular in terms of a constant offset in the relationship.

Question? Why not estimate 2-different models? Estimating a single model pools the data when estimating the shared slope (β_1) leading to better estimates and greater confidence.

More than two classes:

Consider:

Model	X_1	X_2	X_3	X_4
M1	X_{i1}	1	0	0
M2	X_{i1}	0	1	0
M3	X_{i1}	0	0	1
M4	X_{i1}	0	0	0

Different models that can be selected through testing include

$$M4 : E[Y] = \beta_0 + \beta_1 X_{i1}$$
$$M3 : E[Y] = (\beta_0 + \beta_4) + \beta_1 X_{i1}$$
$$\vdots$$

One inference question might be ~~the~~ the difference between β_4 & β_3 (this measures the difference b/twn two regression funcs). This question can be answered by remembering that $b \sim N(\beta, \sigma^2(X'X)^{-1})$ and that any linear function $b^T a$ is also normally distributed so choosing $\vec{a} = [0\,0\,1\,0\,0\,0\,-1\,0\,0\,0]^T$ for instance allows us to derive the sampling distribution (normal) of the difference btwn any two regression coefficients and, accordingly, to construct hypothesis tests, etc.

Time series data

Often linear regression models are used to do forecasting, etc. For instance

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t \qquad t = 1, \ldots, n$$

If two different "regimes" (different economic environments, different patient states, etc.) might result in different forecasts, then indicator vars and hypothesis tests can be employed to test this. I.e.

$$Y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \varepsilon_t$$

where

$$X_{t2} = \begin{cases} 1 & \text{regime 1} \\ 0 & \text{regime 2} \end{cases}$$

Replacing Quantitative variables with Indicator vars of ranges

If a sufficient amount of data is available, sometimes it makes sense to split the data $X \in \mathbb{R}$ into

$$X_1' = \mathbb{I}(0 \le X \le a)$$
$$X_2' = \mathbb{I}(a \le X \le b)$$
$$\vdots$$

and use either the indicator var's alone or in combination with the original data (modulo the obvious colinearity problems) to learn different regression functions ~~for~~ different ranges of the data.

Interactions between Quantitative & Qualitative Predictors

if $X_{i1} \in \mathbb{R}$ and $X_{i2} \in \{0,1\}$

and

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

then the response func. is

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

Meaning of regression coefficients

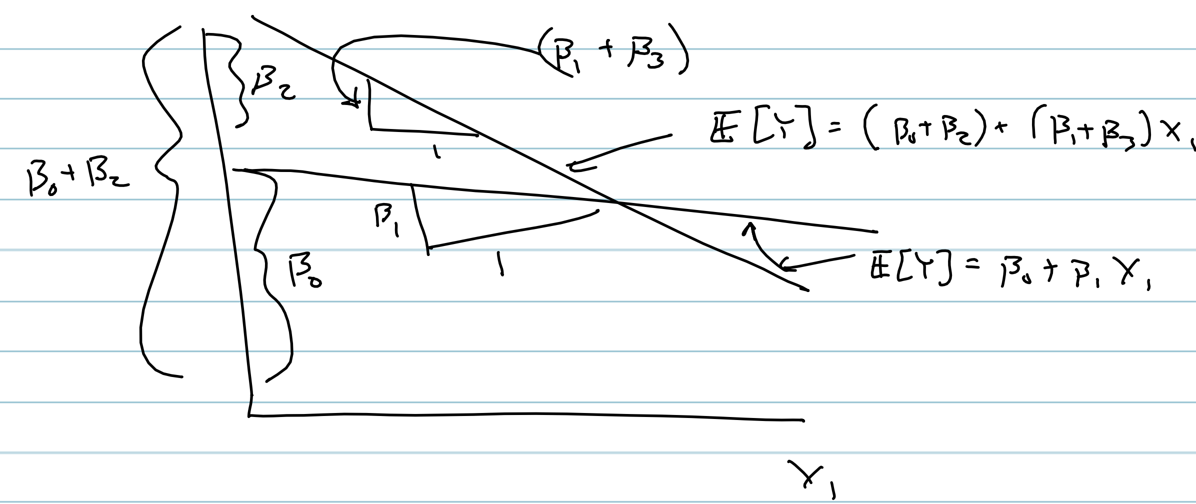
If $X_2 = 0$ then

$$E[Y] = \beta_0 + \beta_1 X_1$$

If $X_2 = 1$ then

$$E[Y] = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1$$

So the indicator effects both the slope <u>and</u> the intercept of the relationship



So testing whether $H_0: \beta_3 = 0$ asks whether the slope is the same btwn two models, $H_0: \beta_2 = 0$ tests if intercepts are same, simultaneous tests (Bonferroni, joint Gaussian tests) test whether or not the two regression models are the same.