Regression Models for Quantitative & Qualitative Predictors

Polynomial regression models

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

Danger: extrapolation in polynomial models may be dangerous.

Model Types
1) One predictor var. - second order

\[ Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i \]

where
\[ x_i = x_i - \bar{x} \]

"Centering" var. reduces multicollinearity substantially.

Notation (with \( i \) index)

\[ Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i \]

The response function is:

\[ E \{ Y \} = \beta_0 + \beta_1 x + \beta_2 x^2 \]

\( \beta_0 \) - is the intercept as before
\( \beta_1 \) - linear effect coefficient
\( \beta_2 \) - quadratic effect coefficient
Third order model

\[ y_i = \beta_0 + \beta_1 x_i + \beta_{11} x_i^2 + \beta_{12} x_i^3 + \varepsilon_i \]

where

\[ x_i = x_i - \bar{x} \]

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 \( \beta_{12} \) is called the interaction effect coefficient.
\[\hat{Y} = b_0 + b_1 X + b_{11} X^2\]
\[= b_0 + b_1 (X - \bar{X}) + b_{11} (X - \bar{X})^2\]
\[= b_0 + b_1 X - b_1 \bar{X} + b_{11} X^2 - 2b_{11} \bar{X}X + b_{11} \bar{X}^2\]
\[= \left(b_0 - b_1 \bar{X} + b_{11} \bar{X}^2\right) + \left(b_1 - 2b_{11}\right)X + b_{11} X^2\]
Implementation of Poly Regression Models

Fitting the regression requires solving

given

\[
\begin{bmatrix}
Y_1 \\
Y_2 \\
\vdots \\
Y_n
\end{bmatrix}
= 
\begin{bmatrix}
1 & x_{11} & x_{12} & x_{13}^2 & x_{14} \\
1 & x_{21} & x_{22} & x_{23}^2 & x_{24} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
1 & x_{n1} & x_{n2} & x_{n3}^2 & x_{n4}
\end{bmatrix}
\begin{bmatrix}
\beta_0 \\
\beta_1 \\
\vdots \\
\beta_n
\end{bmatrix}
+ 
\begin{bmatrix}
\epsilon_1 \\
\epsilon_2 \\
\vdots \\
\epsilon_n
\end{bmatrix}
\]

leads to

\[ b = (X'X)^{-1}X'Y \]

as usual.

Model selection: hierarchical approach: it is natural to include vers using a sequential selection process from lower-order to higher-order terms.

For instance, the model

\[ Y_i = \beta_0 + \beta_1 x_i + \beta_{11} x_i^2 + \beta_{111} x_i^3 + \epsilon_i \]

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

Regression function in terms of centered vers

If we fit

\[ \bar{y} = b_0 + b_1 x + b_{11} x^2 \]

then

\[ \bar{y} = b_0' + b_1' x + b_{11}' x^2 \]

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 function can be expressed in terms of
The original work

Counter: - Poly. models can be highly multicollinearity even when ordered.
- Tests not as powerful because extra terms cut up degrees of freedom etc.

Interaction regress. 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 for

\[
E\{Y\} = \beta_0 + f_1(x_1) + f_2(x_2) + \ldots + f_{p-1}(x_{p-1})
\]

where \( f_i, 1 \leq i \leq p-1 \) can be any functions.

For n. ex.

\[
E\{Y\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + f_1(x) + f_2(x_2)
\]

has effects \( x_1 \) \& \( x_2 \) which are additive of \( Y \).

The reg. func.

\[
E\{Y\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1 x_2
\]

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

The cross-product term is called an interaction term.
Interpretation of Regression Coefficients

Consider

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

The effects are given by

$$\frac{\partial Y}{\partial X_1} = \beta_1 + \beta_3 X_2 \Rightarrow \text{the level of the second input var. affects slope.}$$

and vice versa

**Additive Effect:**
- **Positive Coeff:**
- **Interaction Effect Negative:**

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 ($2^n$ for second order interactions) could need a large amount of data to fit the corresponding model (big data).

**A priori knowledge is not a bad way to go here.** One can plot residuals of the additive effect model against interaction terms to get a sense of which vars matter.
Qualitative predictors

Qualitative vars are discrete: gender ∈ {male, female}, disability status ∈ \{not disabled, partly disabled, fully disabled\}, etc.

One way to identify the classes of a qualitative variable is to use indicator vars that take the values 0 or 1.

For instance, if data \(X_1, \ldots, X_N\) come from class A and data \(X_{N+1}, \ldots, X_{2N}\) come from class B, we can choose class A = 0 and class B = 1

design matrix

\[
\begin{bmatrix}
1 & x_1 & 0 \\
1 & x_2 & 0 \\
\vdots & \vdots & \vdots \\
1 & x_N & 0 \\
1 & x_{N+1} & 0 \\
\end{bmatrix}
\begin{bmatrix}
B_0 \\
B_1 \\
B_2 \\
\vdots \\
B_{N-1} \\
\end{bmatrix}
\]

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

Interpreting regression models with qualitative predictors

If \(X_1 \in \mathbb{R}\) and \(X_2 \in \{0, 1\}\)

and we use the regression model

\[Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon\]
then the response function is

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

For \(X_2 = 0\) this reduces to

\[E[Y] = \beta_0 + \beta_2 X_2\]

But 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:

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

So a formal test of \( H_0 : \beta_2 = 0 \) 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 (\( \beta_1 \)) leading to better estimates and greater confidence.

**More than two classes:**

<table>
<thead>
<tr>
<th>Model</th>
<th>Model</th>
<th>( X_1 )</th>
<th>( X_2 )</th>
<th>( X_3 )</th>
<th>( X_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>( X_{i1} )</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>M2</td>
<td>( X_{i2} )</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>M3</td>
<td>( X_{i3} )</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>M4</td>
<td>( X_{i4} )</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Different models that can be selected through testing include:

- \( M_4: E[Y] = \beta_0 + \beta_1 X_{i1} \)
- \( M_3: E[Y] = (\beta_0 + \beta_4) + \beta_1 X_{i1} \)

One inference question might be the difference between \( \beta_4 \) and \( \beta_2 \) (this measures the difference between two regression functions). This question can be answered by remembering that \( b \sim N(\beta, \sigma^2(X'X)^{-1}) \) and that any linear function \( ba \) is also normally
distributed so choosing $\Delta \approx [0.01000 -1.0000]$ for instance allows us to derive the sampling distribution (normal) of the difference between two regression coefficients $\beta_1$, accordingly to construct hypothesis tests, etc.

Time series data

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

$$y_t = \beta_0 + \beta_1 x_t + \epsilon_t \quad t = 1, \ldots, n$$

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

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \epsilon_t$$

where

$$x_{t1} = \begin{cases} 1 & \text{regime 1} \\ 0 & \text{regime 2} \end{cases}$$

Replacing quantitative variables with indicator variables of ranges

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

$$X^1 = \mathbb{I}(0 \leq X < a)$$
$$X^2 = \mathbb{I}(a \leq X \leq b)$$

and use either the indicator variables alone or in combination with the original data (modulo the obvious collinearity problems) to learn different regression functions for different regimes of the data.
Interactions between Quantitative & Qualitative Predictors

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

and

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

then the response function is

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

Meaning of regression coefficients

If $X_2 = 0$

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

If $X_2 = 1$

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

So the indicator effects both the slope and the intercept of the relationship.

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

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