## Chapter 8 Regression models for Quantitative and Qualitative Predictors

We consider in greater detail standard modeling techniques for quantitative predictors, for qualitative predictors, and for regression models containing both quantitative and qualitative predictors.

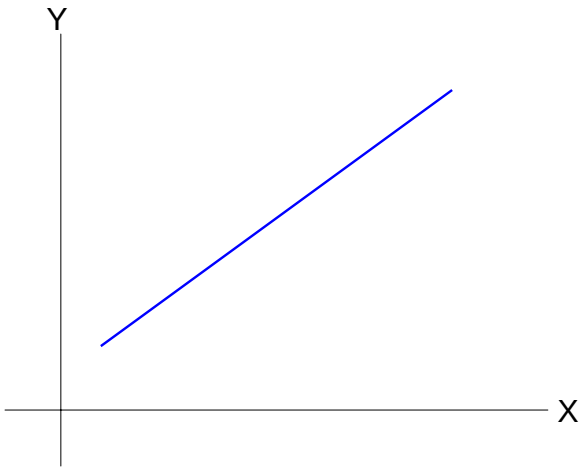
- Interaction and polynomial terms for quantitative predictors
- Indicator variables for qualitative predictors

### 8.1 Polynomial regression models

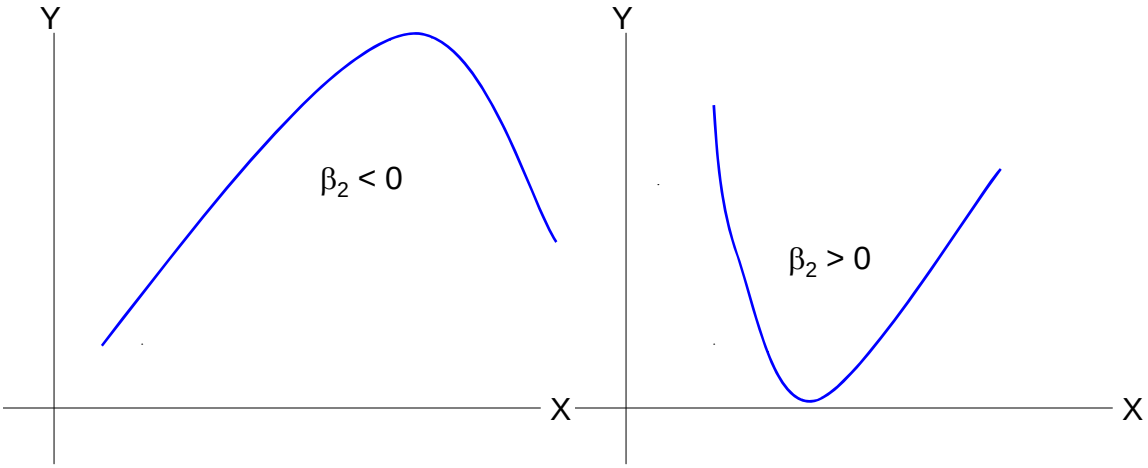
Uses:

1) When the true curvilinear response function is indeed a polynomial function
2) When the true curvilinear response function is unknown (or complex) but a polynomial function is a good approximation to the true function

**First-order model:** $E(Y_i) = \beta_0 + \beta_1 X_i$




**Second-order model**: $E(Y_i) = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$



$X_i^2$ is called the second-order or "quadratic" term of the model. It allows for curvature in the relationship between X and Y.

The sign of $\beta_2$ determines if the curve opens upwards or downwards.

Since $X_i^2$ is a transformation of $X_i$, these two model terms are often highly correlated. Thus, problems with multicollinearity may occur. To partially avoid this, the independent variable is transformed to be **deviations from its mea**n, $x_i=X_i-\overline{X}_1$. The second order model becomes, $E(Y_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$.

Second order models can become more complicated with additional independent variables.

1) Third order model with 1 independent variable:

$$E(Y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2 + \beta_3 x_{i1}^3$$

2) Second order model with 2 independent variables:

$$E(Y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \beta_4 x_{i1}^2 + \beta_5 x_{i2}^2$$

3) Second order model with 3 independent variables:

$$E(Y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i1} x_{i2} + \beta_5 x_{i1} x_{i3} + \beta_6 x_{i2} x_{i3} + \beta_7 x_{i1}^2 + \beta_8 x_{i2}^2 + \beta_9 x_{i3}^2$$

Notice the last two models above contain "**interaction**" terms. Along with the "squared" terms, these are considered to be second order model terms also.

There is a hierarchical approach to fitting the regression model. For example, if $\beta_3 \neq 0$ in 1), then $x_{i1}^2$ and $x_{i1}$ are kept in the model. See Page 299 of KNN for a discussion.

Residual plots should always be examined to evaluate the assumptions of the model. When a squared or higher order term is in the model corresponding to $x_1$, only the plot of $e_i$ vs. $x_{i1}$ needs to be examined for the "linearity" assumption. This is because $e_i$ vs. $x_{i1}^2$, $e_i$ vs. $x_{i1}^3$, … give the same information as $e_i$ vs. $x_1$

See p.300 of KNN for an example of a second order model with two independent variables.

**Case Example:** (Table 8.1, R codes and outputs)
A researcher studied the effects of the charge rate (X1) and temperature (X2) on the life of a new type of power cell (Y) in a preliminary small-scale study. Because of the balanced nature of X1 and X2 levels studied, the researcher not only centered the variables around their respective means but also scaled them in convenient units, as follows:

$$x_{i1} = \frac{X_{i1} - \overline{X}_1}{0.4} = \frac{X_{i1} - 1.0}{0.4}$$

$$x_{i2} = \frac{X_{i2} - \overline{X}_2}{10} = \frac{X_{i2} - 20}{10}$$

(a) Consider the following model:

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2$$

Find the corresponding estimated regression model.

(b) Perform a lack of fit test of the regression function in (a).

$H_0$: $E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2$

$H_a$: $E(Y) \neq \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2$

(c) Perform a partial F test to determine if a first order model should be used instead of a second order model. Use $\alpha$=0.05.

$H_0$: $E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

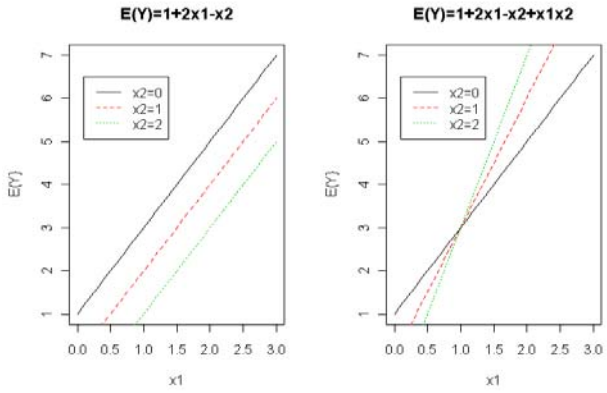
$H_a$: $E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2$

(d) Fit a first-order regression model and estimate the **original** regression coefficients.

**R codes and outputs** (Chapter08_Table1.R and its outputs)

## 8.2 Interaction regression models

The effect of one independent variable on the dependent variable depends on another independent variable.



**Example**: Suppose there are two independent variables

Consider the first order model: $E(Y_i)=\beta_0+\beta_1X_{i1}+\beta_2X_{i2}$. The effect of $X_1$ on E(Y) is measured by $\beta_1$.

Consider the model $E(Y_i)=\beta_0+\beta_1X_{i1}+\beta_2X_{i2}+\beta_3X_{i1}X_{i2}$. The effect of $X_1$ on E(Y) is measured only by $\beta_1$ and $\beta_3X_{i2}$. Since $X_2$ is an independent variable, the effect of $X_1$ on E(Y) is dependent on $X_2$. Similarly, the effect of $X_2$ on E(Y) is dependent on $X_1$. Thus, we say there is an "interaction" between $X_1$ and $X_2$.

For a model containing an interaction term, the regression function is no longer a plane.

Conditional effects plot: **Figure 8.7** on page 307.
Response surfaces and contour plots**: Figure 8.8** on page 310.

**Example: Grandfather clocks**

**Implementation of interaction regression models:**

Sometimes problems with multicollinearity can occur when interaction terms are in the model. Similar to polynomial models, a transformation of $x_{ij}=X_{ij}-\overline{X}_j$ can be done to partially remedy the problem.

**Example:** We wish to test formally in the body fat example of Table 7.1 whether interaction terms between the three predictor variables should be included in the regression model.

Since we know some X variables are highly correlated. We first center the variables around their respective means and the centered variables are denoted by $x_1$, $x_2$ an $x_3$.

The regression model is:

$$E(Y_i) = \beta_0 + \beta_1x_{i1} + \beta_2x_{i2} + \beta_3x_{i3} + \beta_4x_{i1}x_{i2} + \beta_5x_{i1}x_{i3} + \beta_6x_{i2}x_{i3} + \varepsilon_i$$

We wish to test:

$H_0$: $\beta_4 = \beta_5 = \beta_6=0$
$H_a$: not all β's in $H_0$ equal zero

Do a partial F-test (Compute the test-statistic …):

**R output:**

```
> Data <- read.table('C:/ /Ch07TA01.txt', header=FALSE)
> colnames(Data) <- c("X1", "X2", "X3", "Y")
>  # X1 = triceps skinfold thickness;
```

```
>  # X2 = thigh circumference;
>  # X3 = midarm circumference;
>  # Y = body fat
>
> # Subtract mean from each of the X's
> # note the transformed variables have the same names
> for(i in 1:3)
+ {
+  Data[,i] <- Data[,i] - mean(Data[,i])
+ }
>
> anova(lm(Y~X1+X2+X3+I(X1*X2)+I(X1*X3)+I(X2*X3),
data=Data))
Analysis of Variance Table

Response: Y
           Df Sum Sq Mean Sq F value    Pr(>F)
X1          1 352.27  352.27 52.2238 6.682e-06 ***
X2          1  33.17   33.17  4.9173   0.04503 *
X3          1  11.55   11.55  1.7117   0.21343
I(X1 * X2)  1   1.50    1.50  0.2217   0.64552
I(X1 * X3)  1   2.70    2.70  0.4009   0.53760
I(X2 * X3)  1   6.51    6.51  0.9658   0.34366
Residuals  13  87.69    6.75
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1
>
> Full <- lm(Y~X1+X2+X3+I(X1*X2)+I(X1*X3)+I(X2*X3),
data=Data)
> Reduced <- lm(Y~X1+X2+X3, data=Data)
> anova(Reduced, Full)
Analysis of Variance Table

Model 1: Y ~ X1 + X2 + X3
Model 2: Y ~ X1 + X2 + X3 + I(X1 * X2) + I(X1 * X3) + I(X2
* X3)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     16 98.405
2     13 87.690  3    10.715 0.5295 0.6699
```

## 8.3 Qualitative Predictors

Qualitative predictor variables can be used in regression.
Examples: gender (male, female)
disability status (not disabled, partly disabled, fully disabled)

One way of quantitatively identifying the classes of a qualitative variable is to use **indicator** variables (also called dummy variables or binary variables) that take on the values 0 and 1.

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

**Example:** An economist wished to relate the speed with which a particular insurance innovation is adopted (Y) to the size of the insurance firm ($X_1$, quantitative variable) and the type of firm.

Y is measured by the number of months elapsed between the time the first firm adopted the innovation and the time given the given firm adopted the innovation.

The size of firm is measured by the amount of total assets of the firm.

Type of firm is composed of two classes – stock companies and mutual companies. We can define the indicator variable ($X_2$)

$$X_2 = \begin{cases} 1 & \text{if stock company} \\ 0 & \text{if mutual company} \end{cases}$$

The response function for this regression model is

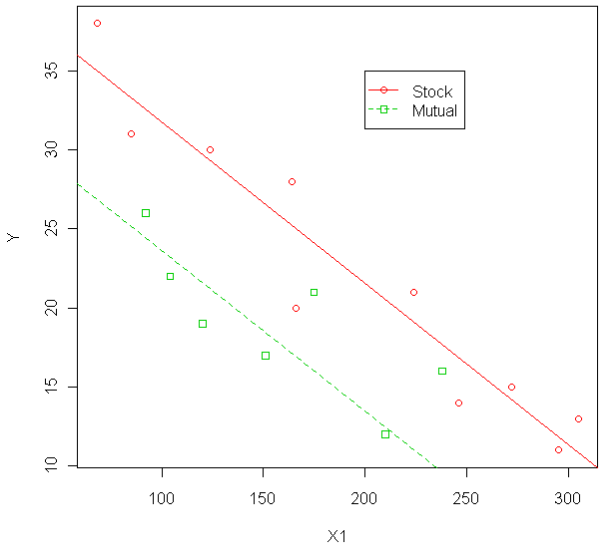
$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

### Interpretation of regression coefficients

Mutual firms: $E\{Y\} = \beta_0 + \beta_1 X_1$

Stock firms: $E\{Y\} = (\beta_0 + \beta_2) + \beta_1 X_1$

(Figure 8.11)
Thus these functions have same slope $\beta_1$.
The parameter, $\beta_2$, measures the differential effect of type of firm.
In general, $\beta_2$ shows how much higher (lower) the mean response line is for the class coded 1 than the line for the class coded 0, for any given level of $X_1$.




Questions (R output from Chapter8_Table2_Insurance.R):

1) Write out the fitted response function:

$$\hat{Y} = 33.87 - 0.10174X_1 + 8.0554X_2$$

2) Get a 95% confidence interval for $\beta_2$.

The 95% confidence interval for $\beta_2$ is $4.98 \leq \beta_2 \leq 11.13$. Thus with 95% confidence, we conclude that stock companies tend to adopt the innovation somewhere between 5 and 11 months later, on the average, than mutual companies, for any given size of firm.

Qualitative predictor with **more than two classes** (Read on page 318).

**Example:** Suppose we want to regress tool wear (Y) on tool speed (X1) and tool model, where the latter has four classes (M1, M2, M3, and M4). We need to have three indicator variables:

$$X_2 = \begin{cases} 1 & \text{if tool model M1} \\ 0 & \text{otherwise} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{if tool model M2} \\ 0 & \text{otherwise} \end{cases}$$

$$X_4 = \begin{cases} 1 & \text{if tool model M3} \\ 0 & \text{otherwise} \end{cases}$$

A first-order regression model is:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i$$

The data input for X variables would be like this:

| Tool Model | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
|---|---|---|---|---|
| M1 | * | 1 | 0 | 0 |
| M2 | * | 0 | 1 | 0 |
| M3 | * | 0 | 0 | 1 |
| M4 | * | 0 | 0 | 0 |

**8.4 Some considerations in using indicator variables**

- Used allocated codes: a single variable with a set of arbitrary numbers
- Quantitative variables can also be represented by indicator variables
- Other codings for indicator variables: -1 and 1

**8.5 Modeling interaction between quantitative and qualitative predictors**

For the insurance example, the response function is

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

Mutual firms: $E\{Y\} = \beta_0 + \beta_1 X_1$

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

Thus, $\beta_2$ measures how much greater (smaller) is the Y intercept of the response function for the class coded 1 than that for class coded 0. Similarly, $\beta_3$ measures how much greater (smaller) is the slope of the response function for the class coded 1 than that for the class coded 0.

Tests can be performed to check

(1) If the interaction term can be dropped from the model ($\beta_3 = 0$)
(2) If the two regression functions are identical ($\beta_2 = \beta_3 = 0$)

**8.6 More complex models (Read on page 327)**

**8.7 Comparison of two or more regression functions**

Encounter regression for two or more populations and wish to study their similarities and differences.

**Example:** A company operates two production lines for making soap bars. For each line, the relation between the speed of the line ($X_1$) and the amount of scrap (Y) for the day was studied. A formal test is desired to determine whether or not the two regression lines are identical.

**Solution:** Introduce an indicator variable

$$X_2 = \begin{cases} 1 & \text{Prodcution line 1} \\ 0 & \text{Prodcution line 2} \end{cases}$$

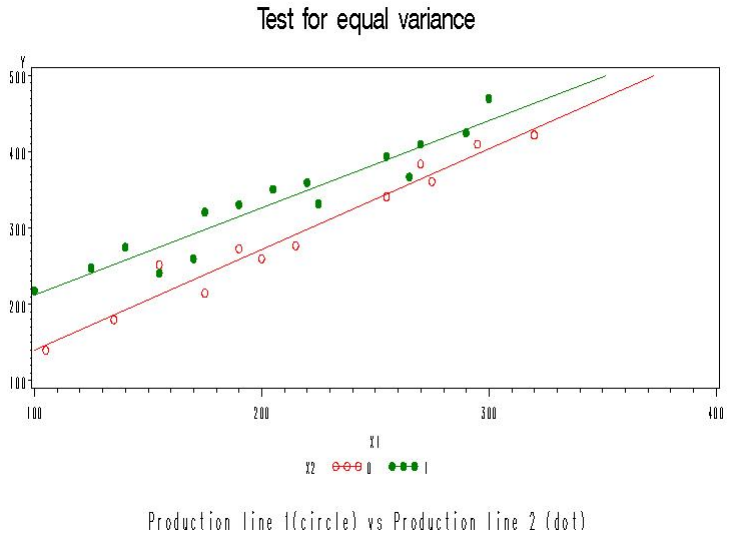
Set up the model $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$

Test if both $\beta_2$ and $\beta_3$ are equal to 0.

Assumption: The error term variances in the regression models for the different populations are equal. If not, transformations may be needed.

**Diagnosis:**

1) Scatter plots




2) Normality test using residuals

3) Test of equality of variances of the error terms for the two population lines using Brown-Forsythe test

**Tests:**

1) Identity of the regression functions for the two production lines (Test if both $\beta_2$ and $\beta_3$ are equal to 0).

2) If the answer to 1) is NO, we would like to examine if the slopes of the two regression lines are the same (Test if $\beta_3$ is equal to 0).

If we have $\beta_2 \neq 0$ and $\beta_3 = 0$, then we have the following conclusion:

1) A given increase in line speed leads to the same amount of increase in expected scrap in each of the two production lines.

2) The expected amount of scrap for any given line speed differs by a constant amount for the two production lines.