Lecture 5

Matrix approach to linear regression

Aim: to transition to multiple linear regression

1. Elements of multivariate normal distribution
2. Matrix approach
3. Multiple linear regression

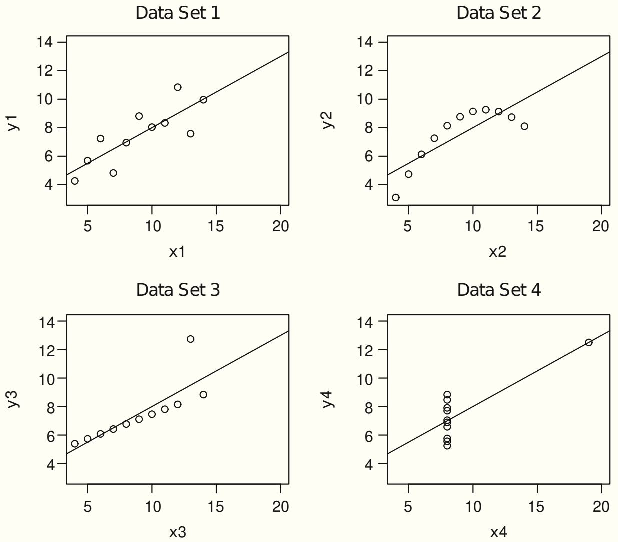
Last week we learned

Anscombe's four data sets

Regression diagnostics

Transformations

Anscombe's four data sets



Regression diagnostics

- Standardised residuals r_i are given by

$$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1-h_{ii})}} \qquad \text{where} \qquad h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{s_{xx}}$$

- The ith case is an outlier if
  - $|r_i| > 2$ for small and medium data sets,
  - $|r_i| > 4$ for large data sets.

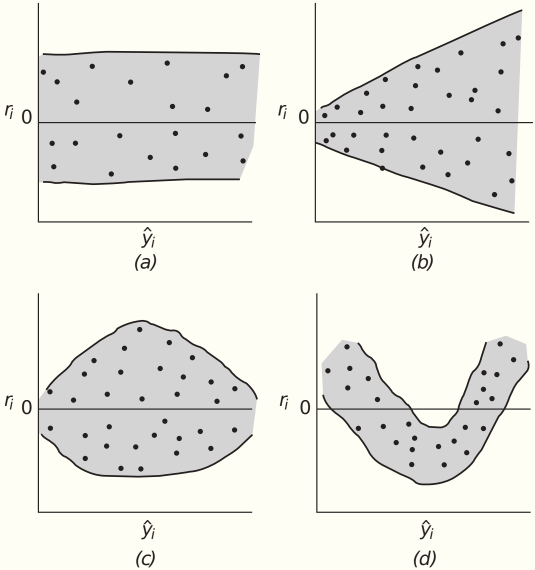
- The ith case is a high leverage case if $h_{ii} > 4/n$. Moreover,
  - it is a bad leverage case if it is also an outlier,
  - it is a good leverage case otherwise.

- Cook's distance is a measure of influence given by

$$D_i = \frac{r_i^2}{2} \cdot \frac{h_{ii}}{1-h_{ii}}$$

  A cut-off for D_i for the simple linear regression is $4/(n-2)$.

Constance of variance plots



Transformations

- Typical transformations are:
  - Square root transformation, i.e. $y_i \to \sqrt{y_i}$ or $x_i \to \sqrt{x_i}$
  - Log transformation, i.e. $y_i \to \log y_i$ or $x_i \to \log x_i$
  - Power transformations, i.e. $y_i \to y_i^\lambda$ or $x_i \to x_i^\lambda$

Lecture 5

Matrix approach to linear regression

Aim: to transition to multiple linear regression

1. Elements of multivariate normal distribution
2. Matrix approach
3. Multiple linear regression

- Let $z_i \sim N(\mu_i, \sigma_{ii}^2)$ for $1 \leq i \leq n$ be normal random variables
- Set $\sigma_{ij}^2 = \mathrm{Cov}(z_i, z_j)$ for $1 \leq i, j \leq n$
- The joint distribution of the z_i's is the **multivariate normal distribution**

$$\boldsymbol{z} \sim N_n(\boldsymbol{\mu}, V)$$

where

$$\boldsymbol{z} = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix} \qquad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} \qquad V = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \cdots & \sigma_{1n}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \cdots & \sigma_{2n}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1}^2 & \sigma_{n2}^2 & \cdots & \sigma_{nn}^2 \end{pmatrix}$$

Expectation value

- Let $\boldsymbol{z}$ be an n-dimensional column vector of random vectors, then

$$\mathbb{E}(\boldsymbol{z}) = \mathbb{E}\begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix} = \begin{pmatrix} \mathbb{E}(z_1) \\ \mathbb{E}(z_2) \\ \vdots \\ \mathbb{E}(z_n) \end{pmatrix}$$

- Let a be a scalar, $\boldsymbol{b}$ – an n-dimensional column vector of constants, and A, B – matrices of constants, then

$$\mathbb{E}(a\boldsymbol{z} + \boldsymbol{b}) = a\,\mathbb{E}(\boldsymbol{z}) + \boldsymbol{b}$$

$$\mathbb{E}(A\boldsymbol{z}) = A\,\mathbb{E}(\boldsymbol{z})$$

$$\mathbb{E}(\boldsymbol{z}^T B) = \mathbb{E}(\boldsymbol{z})^T B$$

Variance-covariance matrix

- Variance-covariance (dispersion) matrix

$$\mathrm{Var}(\boldsymbol{z}) = \begin{pmatrix} \mathrm{Var}(z_1) & \mathrm{Cov}(z_1, z_2) & \dots & \mathrm{Cov}(z_1, z_n) \\ \mathrm{Cov}(z_2, z_1) & \mathrm{Var}(z_2) & \dots & \mathrm{Cov}(z_2, z_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(z_n, z_1) & \mathrm{Cov}(z_n, z_2) & \dots & \mathrm{Var}(z_n) \end{pmatrix}$$

  - The matrix $\mathrm{Var}(\boldsymbol{z})$ is symmetric, $\mathrm{Cov}(z_i, z_j) = \mathrm{Cov}(z_j, z_i)$
  - For mutually uncorrelated random variables it is a diagonal matrix, $\mathrm{Cov}(z_i, z_j) = 0$ for all $i \neq j$
  - It can be written as $\mathrm{Var}(\boldsymbol{z}) = \mathbb{E}\left[(\boldsymbol{z} - \mathbb{E}(\boldsymbol{z}))(\boldsymbol{z} - \mathbb{E}(\boldsymbol{z}))^T\right]$
  - For a transformed variable $\boldsymbol{u} = A\boldsymbol{z}$ we have $\mathrm{Var}(\boldsymbol{u}) = A\mathrm{Var}(\boldsymbol{z})A^T$

Variance-covariance matrix

- $\text{Var}(\boldsymbol{z}) = \mathbb{E}\left[(\boldsymbol{z}-\boldsymbol{\mu})(\boldsymbol{z}-\boldsymbol{\mu})^T\right]$ where $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{z})$. Indeed:

$$\mathbb{E}\left[(\boldsymbol{z}-\boldsymbol{\mu})(\boldsymbol{z}-\boldsymbol{\mu})^T\right] = \mathbb{E}\left[\begin{pmatrix} z_1-\mu_1 \\ z_2-\mu_2 \\ \vdots \\ z_n-\mu_n \end{pmatrix}(z_1-\mu_1,\, z_2-\mu_2,\, \ldots,\, z_n-\mu_n)\right]$$

$$= \begin{pmatrix} \mathbb{E}((z_1-\mu_1)^2) & \mathbb{E}((z_1-\mu_1)(z_2-\mu_2)) & \cdots & \mathbb{E}((z_1-\mu_1)(z_n-\mu_n)) \\ \mathbb{E}((z_2-\mu_2)(z_1-\mu_1)) & \mathbb{E}((z_2-\mu_2)^2) & \cdots & \mathbb{E}((z_2-\mu_2)(z_n-\mu_n)) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}((z_n-\mu_n)(z_1-\mu_1)) & \mathbb{E}((z_n-\mu_n)(z_2-\mu_2)) & \cdots & \mathbb{E}((z_n-\mu_n)^2) \end{pmatrix}$$

$$= \text{Var}(\boldsymbol{z})$$

- Homework: show that $\text{Var}(\boldsymbol{z}) = \mathbb{E}[\boldsymbol{z}\boldsymbol{z}^T] - \boldsymbol{\mu}\boldsymbol{\mu}^T$

Variance-covariance matrix

- Let $\boldsymbol{u} = A\boldsymbol{z}$. Then $\mathrm{Var}(\boldsymbol{u}) = A\mathrm{Var}(\boldsymbol{z})A^T$. Indeed:

$$
\begin{aligned}
\mathrm{Var}(\boldsymbol{u}) &= \mathbb{E}\left[(\boldsymbol{u} - \mathbb{E}(\boldsymbol{u}))(\boldsymbol{u} - \mathbb{E}(\boldsymbol{u}))^T\right] \\
&= \mathbb{E}\left[(A\boldsymbol{z} - A\boldsymbol{\mu})(A\boldsymbol{z} - A\boldsymbol{\mu})^T\right] \\
&= \mathbb{E}\left[A(\boldsymbol{z} - \boldsymbol{\mu})(\boldsymbol{z} - \boldsymbol{\mu})^T A^T\right] \\
&= A\mathbb{E}\left[(\boldsymbol{z} - \boldsymbol{\mu})(\boldsymbol{z} - \boldsymbol{\mu})^T\right]A^T \\
&= A\mathrm{Var}(\boldsymbol{z})A^T
\end{aligned}
$$

Lecture 5

Matrix approach to linear regression

Aim: to transition to multiple linear regression

1. Elements of multivariate normal distribution
2. Matrix approach
3. Multiple linear regression

- The SLR model has n equations

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_2 + \varepsilon_2 \\ \vdots & \qquad\quad \vdots \\ y_n &= \beta_0 + \beta_1 x_n + \varepsilon_n \end{aligned}$$

- In the matrix form

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

- The matrix X is known as the design matrix.

Distribution of $\boldsymbol{\varepsilon}$ and $\boldsymbol{y}$

- The SLR model in the matrix form

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

- Standard assumption for the error terms

$$\varepsilon_i \overset{ind}{\sim} N(0, \sigma^2) \quad \Longrightarrow \quad \boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 I_n)$$

- Distribution of $\boldsymbol{y}$

$$y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2) \quad \Longrightarrow \quad \boldsymbol{y} \sim N_n(X\boldsymbol{\beta}, \sigma^2 I_n)$$

where

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad I_n = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{pmatrix}$$

Claim. The least squares estimate of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}$$

Proof. First, note that

$$SS_E = \sum_i e_i e_i = \boldsymbol{e}^T \boldsymbol{e}$$

Hence in vector notation, to minimise SS_E we must solve the equation

$$\begin{pmatrix} \frac{\partial}{\partial \beta_0} \boldsymbol{e}^T \boldsymbol{e} \\ \frac{\partial}{\partial \beta_1} \boldsymbol{e}^T \boldsymbol{e} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \Longrightarrow \quad \frac{\partial \boldsymbol{e}^T \boldsymbol{e}}{\partial \boldsymbol{\beta}} = \boldsymbol{0}$$

Matrix derivatives

- Let $\boldsymbol{z} = (z_1, \ldots, z_r)^T$ and let $f(z_1, \ldots, z_r)$ be a function of $\boldsymbol{z}$. Then

$$\frac{\partial f(z_1, \ldots, z_r)}{\partial \boldsymbol{z}} = \begin{pmatrix} \frac{f(z_1, \ldots, z_r)}{\partial z_1} \\ \vdots \\ \frac{\partial f(z_1, \ldots, z_r)}{\partial z_n} \end{pmatrix}$$

- For any vector $\boldsymbol{a} = (a_1, \ldots, a_r)^T$ we have

$$\frac{\partial \boldsymbol{a}^T \boldsymbol{z}}{\partial \boldsymbol{z}} = \frac{\partial \boldsymbol{z}^T \boldsymbol{a}}{\partial \boldsymbol{z}} = \frac{\partial (a_1 z_1 + \ldots + a_r z_r)}{\partial \boldsymbol{z}} = \begin{pmatrix} a_1 \\ \vdots \\ a_r \end{pmatrix} = \boldsymbol{a}$$

- If M is a square $r \times r$ matrix then

$$\frac{\partial \boldsymbol{z}^T M \boldsymbol{z}}{\partial \boldsymbol{z}} = (M + M^T)\boldsymbol{z}$$

- Using

$$\frac{\partial \boldsymbol{a}^T \boldsymbol{z}}{\partial \boldsymbol{z}} = \frac{\partial \boldsymbol{z}^T \boldsymbol{a}}{\partial \boldsymbol{z}} = \boldsymbol{a} \qquad \frac{\partial \boldsymbol{z}^T M \boldsymbol{z}}{\partial \boldsymbol{z}} = (M + M^T)\,\boldsymbol{z}$$

we compute

$$\begin{aligned}
\frac{\partial \boldsymbol{e}^T \boldsymbol{e}}{\partial \boldsymbol{\beta}} &= \frac{\partial}{\partial \boldsymbol{\beta}} (\boldsymbol{y} - X\boldsymbol{\beta})^T (\boldsymbol{y} - X\boldsymbol{\beta}) \\
&= \frac{\partial}{\partial \boldsymbol{\beta}} \left(\boldsymbol{y}^T \boldsymbol{y} - \boldsymbol{\beta}^T X^T \boldsymbol{y} - \boldsymbol{y}^T X \boldsymbol{\beta} + \boldsymbol{\beta}^T (X^T X)\,\boldsymbol{\beta} \right) \\
&= 0 - X^T \boldsymbol{y} - (\boldsymbol{y}^T X)^T + (X^T X + (X^T X)^T\,)\,\boldsymbol{\beta} \\
&= -2X^T \boldsymbol{y} + 2\,(X^T X)\,\boldsymbol{\beta}
\end{aligned}$$

- Next, equate the derivative to zero

$$\left. \frac{\partial \boldsymbol{e}^T \boldsymbol{e}}{\partial \boldsymbol{\beta}} \right|_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}} = \mathbf{0} \quad \Longrightarrow \quad (X^T X)\,\hat{\boldsymbol{\beta}} = X^T \boldsymbol{y} \quad \Longrightarrow \quad \hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}$$

The SLR model in the matrix form

- We found that $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y}$. This must agree with our earlier results.

- The design matrix is

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

Therefore (Homework: verify the steps on this and next slide)

$$X^T \mathbf{y} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}$$

$$X^T X = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}$$

$$\det(X^T X) = n \sum_i x_i^2 - \left(\sum_i x_i \right)^2 = n \sum_i \left(x_i^2 - x_i \bar{x} \right) = n\, s_{xx}$$

- The inverse of $X^T X$ is

$$(X^TX)^{-1} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}^{-1} = \frac{1}{n s_{xx}} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix} = \frac{1}{s_{xx}} \begin{pmatrix} \frac{1}{n}\sum x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}$$

- Therefore

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (X^TX)^{-1}X^T\boldsymbol{y} \\
&= \frac{1}{s_{xx}} \begin{pmatrix} \frac{1}{n}\sum x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix} = \frac{1}{s_{xx}} \begin{pmatrix} \frac{1}{n}\sum x_i^2 \sum y_i - \bar{x}\sum x_i y_i \\ \sum x_i y_i - \bar{x}\sum y_i \end{pmatrix} \\
&= \frac{1}{s_{xx}} \begin{pmatrix} \frac{1}{n}\sum x_i^2 \sum y_i - \bar{x}^2 \sum y_i - \bar{x}(\sum x_i y_i - \bar{x}\sum y_i) \\ s_{xy} \end{pmatrix} \\
&= \frac{1}{s_{xx}} \begin{pmatrix} s_{xx}\bar{y} - \bar{x}s_{xy} \\ s_{xy} \end{pmatrix} = \begin{pmatrix} \bar{y} - \hat{\beta}_1\bar{x} \\ \hat{\beta}_1 \end{pmatrix}
\end{aligned}$$

which is the same result we found earlier.

Lecture 5

Matrix approach to linear regression

Aim: to transition to multiple linear regression

Multiple linear regression

- A linear regression model that has more than one predictor variable, X_1, X_2, . . . , is called a multiple linear regression model.
- This model generalizes the simple linear regression in two ways:
  - it allows the mean function $\mathbb{E}(Y)$ to depend on more than one predictor,
  - it can take shapes more complicated than straight lines.
- A multiple linear regression model can regress:
  - continuous data, e.g. weight in kg, height in cm, income in £,
  - ordinal categorical data, e.g.

    level: very low, low, medium, high, very high

    likeness: dislike, dislike somewhat, neutral, like somewhat, like
  - non-ordinal categorical data, e.g. colour or gender, blood type.

- Suppose that a multiple linear regression model has p predictors $X_1, \ldots, X_p$. This means that

$$\mathbb{E}(Y|X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- When $X_1 = x_1, \ldots, X_p = x_p$ we write

$$\mathbb{E}(Y|X = \boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

- Parameters $\beta_0, \ldots, \beta_p$ are called partial regression coefficients – each β_i represents the expected change in the response per unit change in x_i when all x_j with $j \neq i$ are held *constant*:

  – when $p = 1$, we obtain the simple linear regression model;

  – when $p = 2$, the mean function $\mathbb{E}(Y|X)$ is a plane in 3 dimensions;

  – when $p \geq 3$, the mean function $\mathbb{E}(Y|X)$ is a hyperplane, the generalization of a p-dimensional plane in a $(p+1)$-dimensional space.

Example: income vs education and age

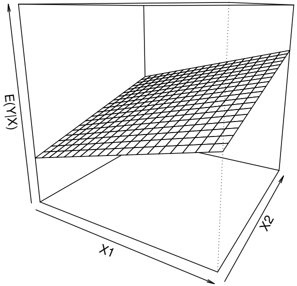
- Consider the relation between the income and education of a person:

  – On an average, higher level of education provides higher income.

  – Most people have higher income when they are older than when they are young, regardless of education.

- A MLR model for income Y vs. education X_1 and age X_2

$$\mathbb{E}(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

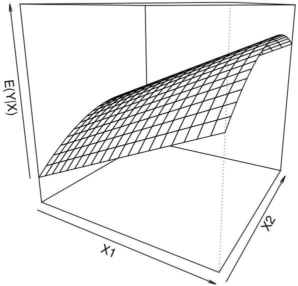


Example: income vs education and age

- Often the income tends to rise less rapidly in the later earning years than in early years.

- To accommodate such possibility, we might include a quadratic term to our model:

$$\mathbb{E}(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{22} X_2^2$$

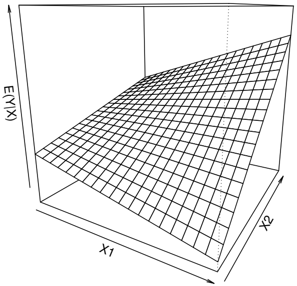


Example: a model with interactions

- A MLR model with interactions

$$\mathbb{E}(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2$$

where X_1X_2 represents an interaction between X_1 and X_2 – the effect of a change in X_1 on Y depends on the level of X_2, and vice versa



- If we let $X_3 = X_1X_2$, we recover a MLR model with three variables.

Example: a polynomial model

- A polynomial regression model

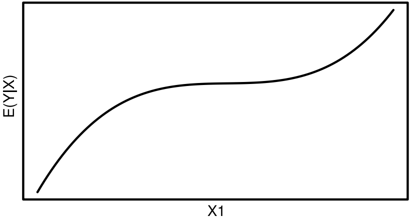
$$\mathbb{E}(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1^3$$

  – The response surface is curvilinear.

  – If we let $X_2 = X_1^2$, $X_3 = X_1^3$, we recover a MLR model with three variables:

$$\mathbb{E}(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

  – It is often advised to avoid higher order term when building the models; the reasons will be explained further in the course.



The error term

- As in the case of a simple linear regression, $\mathbb{E}(Y|X = \boldsymbol{x})$ does not adequately describe the data which show some randomness.
- To deal with this problem we introduce a random error $\varepsilon \sim N(0, \sigma^2)$

$$Y = \mathbb{E}(Y|X = \boldsymbol{x}) + \varepsilon = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon$$

- MLR models in R:

```
> lm( y ~ x1 + x2 + ... + xp )  #  b0 + b1 x1 + ...
```

– interaction term

```
> lm( y ~ x1*x2 )  # b0 + b1 x1 + b2 x2 + b12 x1 x2
> lm( y ~ x1 + x2 + x1:x2 )  # the same as above
```

– polynomial model

```
> lm( y ~ x1 + I(x1^2) + I(x1^3) ) # b0 + b1 x1 + ...
```

Quiz

- Which of these regression models are linear?

(a) $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2 + \beta_3 x_3^3 + \varepsilon$

(b) $y = \beta_0 + \beta_1 x_1 + \beta_2^2 x_2 + \beta_3^3 x_3^3 + \varepsilon$

(c) $y = \alpha + \beta \sin(x) + \varepsilon$

(d) $y = \alpha + \log(\beta x) + \varepsilon$

(e) $y = \beta_0 + \beta_1 x_1^{-1} + \beta_2 x_2 + \beta_{123} x_1 x_2 x_3 + \varepsilon$

(f) $y = \exp(\beta_0) + \beta_1 \exp(x_1) + \exp(\beta_2 x_2) + \varepsilon$

(g) $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

The principle of parsimony

- Given a set of equally good explanations for a given phenomenon, then the correct explanation is the simplest explanation.

- In statistical modelling, the principle of parsimony means that:
  - models should have as few parameters as possible
  - linear models should be preferred to non-linear models
  - experiments relying on few assumptions should be preferred to those relying on many
  - models should be pared down until they are minimal adequate
  - simple explanations should be preferred to complex explanations

- In general, a variable is retained in the model only if it causes a significant increase in deviance when it is removed from the current model:
  - a model should be as simple as possible, but no simpler.

Next week

Multiple Linear Regression