# Part II

# Linear Regression Modelling

# 2. Linear regression Modelling

In the previous section, we attempted to explain the variation in an observed response variable by fitting models with one or more factors.

Factors are **discrete** variables taking different levels; in this section we will now utilize **continuous** variables that can similarly explain variation in an observed response.

# 2.1 Simple Linear Regression

We will investigate models relating two quantities $x$ and $y$ through equations of the form

$$y = ax + b$$

where $a$ and $b$ are constants (that is, a straight-line).

Variables $x$ and $y$ will not be treated exchangeably - we will regard $y$ as being a function of $x$.

Such models are deterministic, that is, if we know $x$ (and the values of the constants), we can compute $y$ exactly without error.

A more useful model allows for the possibility that the system is not observed perfectly, that is, we do not observe $(x, y)$ pairs that are always consistent with a simple functional relationship.

## Example (Pharmacokinetic Model)

If a dose of drug is taken at time $x = 0$, the amount (concentration) of drug still in the bloodstream at time $x$ is often well-modelled by a simple equation. Let

- $D$ denote the amount of drug taken at $x = 0$
- $x$ time
- $y^\star$ is the amount (concentration per unit volume) in the bloodstream.

Then

$$y^\star = \frac{D}{V} \exp\{-\lambda x\}$$

where

- $\lambda$ is the elimination rate
- $V$ is the volume of bloodstream.

## Example (Pharmacokinetic Model (continued))

Taking logs of both sides, setting $y = \log y^\star$, then

$$y = -\lambda x + \log(D/V) = -\lambda x + (\log D - \log V)$$

that is, $y = ax + b$ where

- $a = -\lambda$
- $b = (\log D - \log V)$

However, in practice, when we measure concentration, we do so with random error.

# 2.1.1 Probabilistic Models

In a **probabilistic** model, we allow for the possibility that $y$ is observed with random error, that is,

$$y = ax + b + ERROR$$

where *ERROR* is a random term that is present due to imperfect observation of the system due to (i) measurement error or (ii) missing information.

Note that we do not treat $x$ and $y$ exchangeably; $x$ is a fixed observed variable that is measured *without error*, whereas $y$ is an observed variable that is measured *with random error*.

We model the variation in $y$ as a function of $x$. We observe pairs $(x_i, y_i), i = 1, \ldots, n$.

# A Basic Probabilistic Model

Terminology:

- $y$ - *Dependent variable* or *independent variable*
- $x$ - *Independent variable*, or *predictor*, or *covariate*

The model we study takes the form

$$y = \beta_0 + \beta_1 x + \epsilon$$

where $\epsilon$ is a random error term, a random variable with mean zero and finite variance ($E[\epsilon] = 0$, $Var[\epsilon] = \sigma^2$); it represents the error present in the measurement of $y$.

- $\beta_0$ - *Intercept* parameter
- $\beta_1$ - *Slop*e parameter

- $\beta_1 > 0$ - increasing $y$ with increasing $x$
- $\beta_1 < 0$ - decreasing $y$ with increasing $x$
- $\beta_1 = 0$ - no relationship between $x$ and $y$

Note:

$$E[Y|x] = \beta_0 + \beta_1 x$$

where $E[Y|x]$ is the expected value of $Y$ for fixed value of $x$.

Recall the notation

- $Y$ - a random variable with a probability distribution
- $y$ - a fixed value that the variable $Y$ can take.

**Fundamental Problem:** If we believe the straight-line model with error is correct, how do we find the values of parameters $\beta_0$ and $\beta_1$. We only have the observed data $\{(x_i, y_i), i = 1, \ldots, n\}$.

## 2.1.2 Least Squares Fitting

We select the best values of $\beta_0$ and $\beta_1$ by minimizing the *error in fit*. For two data points $(x_1, y_1)$ and $(x_2, y_2)$, the errors in fit are

$$\begin{aligned} e_1 &= y_1 - (\beta_0 + \beta_1 x_1) \\ e_2 &= y_2 - (\beta_0 + \beta_1 x_2) \end{aligned}$$

respectively. But note that, potentially, $e_1 > 0$ and $e_2 < 0$ so there is a possibility that these fitting errors cancel each other out. Therefore we look at **squared** errors (as a large negative error is as bad as a large positive error)

$$\begin{aligned} e_1^2 &= (y_1 - (\beta_0 + \beta_1 x_1))^2 \\ e_2^2 &= (y_2 - (\beta_0 + \beta_1 x_2))^2 \end{aligned}$$

For $n$ data, we obtain $n$ misfit squared errors

$$e_1^2, \ldots, e_n^2$$

We select $\beta_0$ and $\beta_1$ as the values of the parameters that minimize $SSE$, where

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

We wish to make the total misfit squared error as small as possible.

SSE - sum of squared errors - is similar to the SSE for ANOVA. We could write

$$SSE = SSE(\beta_0, \beta_1)$$

to show the dependence of $SSE$ on the parameters.

Minimization of $SSE(\beta_0, \beta_1)$ is achieved **analytically**.

Two routes: (i) calculus and (ii) geometric methods. It follows that the best parameters $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are given by

$$\widehat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \qquad \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$$

where

- Sum of Squares $SS_{xx}$:

$$SS_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2$$

- Sum of Squares $SS_{xy}$:

$$SS_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

$\widehat{\beta}_0$ and $\widehat{\beta}_1$ are the **least-squares estimates**

$$y = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

is the **least-squares line of best fit**. The **fitted-values** are

$$\hat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i \qquad i = 1, \dots, n$$

and the **residuals** or **residual errors** are

$$\hat{e}_i = y_i - \hat{y}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \qquad i = 1, \dots, n$$

## 2.1.3 Model Assumptions for Least-Squares

To utilize least-squares for the probabilistic model

$$Y = \beta_0 + \beta_1 x + \epsilon$$

we make the following assumptions

1. The expected error $E[\epsilon]$ is zero so that
$$E[Y] = \beta_0 + \beta_1 x$$
2. The variance of the error, $Var[\epsilon]$, is constant and does not depend on $x$.
3. The probability distribution of $\epsilon$ is a symmetric distribution about zero (a stronger assumption is that $\epsilon$ is Normally distributed).
4. The errors for two different measured responses are independent, i.e. the error $\epsilon_1$ in measuring $y_1$ at $x_1$ is independent of the error $\epsilon_2$ in measuring $y_2$ at $x_2$.

# 2.1.4 Parameter Estimation: Estimating $\sigma^2$

Using the LS procedure, we can construct an estimate of the *error* or *residual error* variance

Recall that

$$Var[\epsilon] = \sigma^2$$

An estimate of $\sigma^2$ is

$$\widehat{\sigma}^2 = \frac{SSE(\widehat{\beta}_0, \widehat{\beta}_1)}{n-2} = s^2$$

say.

Note that the denominator $n-2$ is again a *degrees of freedom* parameter of the form

$$\begin{array}{ccc} \text{TOTAL NUMBER} & - & \text{NUMBER OF PARAMETERS} \\ \text{OF DATA} & & \text{ESTIMATED} \end{array}$$

or $n-p$, where in the simple linear regression, $p=2$ ($\widehat{\beta}_0$ and $\widehat{\beta}_1$). Note also that

$$SSE(\widehat{\beta}_0, \widehat{\beta}_1) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = SS_{yy} - \widehat{\beta}_1 SS_{xy}$$

where

$$SS_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

# Estimation and Testing for Slope

In the model where

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

it is of interest to test the hypothesis

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_a &: \beta_1 \neq 0 \end{aligned}$$

i.e. $H_0$ implies that there is no systematic contribution of $x$ to the variation of $y$.

To test $H_0$ vs $H_a$ we us the test statistic

$$t = \frac{\widehat{\beta}_1}{\text{e.s.e}(\widehat{\beta}_1)} = \frac{\widehat{\beta}_1}{s_{\widehat{\beta}_1}}$$

where e.s.e$(\widehat{\beta}_1)$ is the *Estimated Standard Error* of $\widehat{\beta}_1$, computed as

$$\text{e.s.e}(\widehat{\beta}_1) = \frac{s}{\sqrt{SS_{xx}}}$$

where $s$ is the estimate of $\sigma$ defined previously.

If $H_0$ is true, and $\beta_1 = 0$, then

$$t = \frac{\widehat{\beta}_1}{s/\sqrt{SS_{xx}}} \sim \text{Student}(n-2)$$

so we can carry out a significance test at level $\alpha$ in the usual way (use a $p$-value, or construct the rejection region).

Note: we might also consider a one-sided test, where $H_a : \beta_1 > 0$, say.

- If $H_a : \beta_1 \neq 0$, we use the *two-sided* rejection region, with critical values
$$C_R = \pm t_{n-2}(\alpha/2)$$
- If $H_a : \beta_1 > 0$, we use the *one-sided* rejection region, with critical value
$$C_R = +t_{n-2}(\alpha)$$
- If $H_a : \beta_1 < 0$, we use the *one-sided* rejection region, with critical value
$$C_R = -t_{n-2}(\alpha)$$

Note: To test

$$\begin{aligned} H_0 &: \beta_1 = b \\ H_a &: \beta_1 \neq b \end{aligned}$$

for any $b$, the test statistic is

$$t = \frac{\widehat{\beta}_1 - b}{s/\sqrt{SS_{xx}}}$$

(for example, $b = 1$ may be of interest. If $H_0$ is true

$$t \sim \text{Student}(n-2)$$

# Confidence Interval

A $100(1-\alpha)\%$ confidence interval for $\beta_1$ is

$$\widehat{\beta}_1 \pm t_{n-2}(\alpha/2) \times s_{\widehat{\beta}_1}$$

where

$$\begin{aligned} t_{n-2}(\alpha/2) &: \quad \alpha/2 \text{ prob. point of Student}(n-2) \text{ distn.} \\ s_{\widehat{\beta}_1} &: \quad \text{Estimated standard error of } \widehat{\beta}_1 \end{aligned}$$

Note: we could perform a similar analysis for $\beta_0$, but this is generally of less interest.

The only quantity that needs attention is the estimated standard error of $\widehat{\beta}_0$. It can be shown that

$$\text{e.s.e.}(\widehat{\beta}_0) = s_{\widehat{\beta}_0} = \sqrt{\frac{1}{n}\left(1 + \frac{n\overline{x}^2}{SS_{xx}}\right)}$$

# 2.1.5 The Coefficient of Correlation

To measure the *strength of association* between the two variables $x$ and $y$ we can use the

**Pearson Product Moment Coefficient Of Correlation**

or *correlation coefficient* which measures the strength of the **linear** relationship between $x$ and $y$.

The coefficient, $r$, is defined by

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$$

where

$$SS_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 \quad SS_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2$$

$$SS_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

Note: $-1 \leq r \leq 1$.

- If $r$ is close to 1, there is a strong linear relationship between $x$ and $y$ where $y$ **increases** with $x$.
- If $r$ is close to -1, there is a strong linear relationship between $x$ and $y$ where $y$ **decreases** with $x$.

Note: In the model

$$y = \beta_0 + \beta_1 x$$

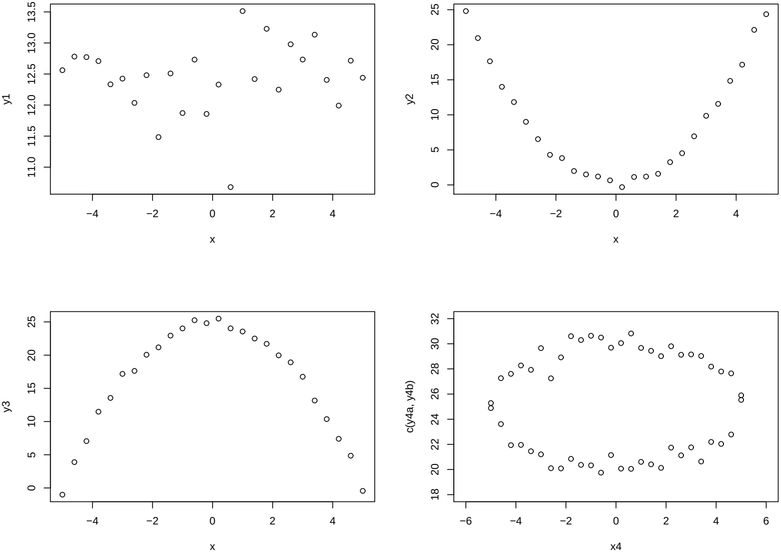
$\beta_1 = 0 \Longrightarrow r \approx 0$, so tests for $\beta_1 = 0$ can also be used to deduce a lack of correlation between the variables.

# Notes

1. A strong linear relationship is not necessarily a **causal** relationship, that is, just because $r \approx 1$ does not mean that $x$ **causes** changes in $y$ (we may have a *spurious* correlation).

2. Just because $r \approx 0$ does not mean that that $x$ and $y$ are unrelated, merely that they are **uncorrelated**. That is, it is possible to construct examples where $x$ and $y$ have a strong functional relationship, but where $r = 0$.

# Examples where $r \approx 0$.



# Testing Correlation

We use $\rho$ to denote the **true** correlation between $X$ and $Y$.

We can test the hypothesis that $\rho = 0$ (that is, that $X$ and $Y$ are uncorrelated using $r$. For testing

$$\begin{aligned} H_0 &: \quad \rho = 0 \\ H_a &: \quad \rho \neq 0 \end{aligned}$$

we can use the test statistic

$$t = \frac{r}{\sqrt{(1 - r^2)/(n - 2)}}$$

If $H_0$ is true, then approximately

$$t \sim \text{Student}(n - 2)$$

Alternately, we could use

$$z = \frac{1}{2} \log \left( \frac{1 + r}{1 - r} \right)$$

and then, if $H_0$ is true, as (approximately)

$$Z \sim N \left( \frac{1}{2} \log \left( \frac{1 + \rho}{1 - \rho} \right), \frac{1}{n - 3} \right)$$

when $\rho = 0$, so that (approximately)

$$\sqrt{n - 3}\, Z \sim N(0, 1)$$

A related quantity is the

**Coefficient of Determination**

or **R$^2$ Statistic**

$$r^2 = \frac{SS_{yy} - SSE}{SS_{yy}} = 1 - \frac{SSE}{SS_{yy}}$$

Note that the *total variation* in $y$ is recorded via

$$SS_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

and the *random variation* is recorded via

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Therefore the **variation explained by the linear regression** is

$$SSR = SS_{yy} - SSE \qquad \text{as} \qquad SS_{yy} = SSR + SSE$$

Thus

$$r^2 = \frac{SSR}{SS_{yy}} = \frac{\text{Variation explained by Regression}}{\text{Total Variation}}$$

$r^2$ is a measure of model adequacy, that is, if $r^2 \approx 1$, then the linear model is a **good fit**.

## Example (Blood Viscosity vs PCV)

We have

- $n = 32$
- $r = 0.879$
- $R^2 = r^2 = (0.879)^2 = 0.772$

Test of $\rho = 0$:

$$t = \frac{r}{\sqrt{(1 - r^2)/(n - 2)}} = 10.087$$

We compare with a Student$(n - 2) \equiv$ Student(30) distribution; the $p$-value is $3.73 \times 10^{-11}$, so there is strong evidence that $\rho \neq 0$.

# 2.1.6 Prediction

After the linear model is fitted, it can be used for **forecasting** or **prediction**. That is, given a new $x$ value we can predict the corresponding $y$.

As before, we see that at any value of $x_p$, the prediction $\hat{y}_p$ is

$$\hat{y}_p = \widehat{\beta}_0 + \widehat{\beta}_1 x_p$$

This is the best predictor of $y$ at this $x$ value.

We can also compute the standard error of this prediction; if the value of the random error variance $\sigma^2$ is known, then

$$\text{s.e.}(\hat{y}_p) = \sigma \sqrt{\frac{1}{n} + \frac{(x_p - \overline{x})^2}{SS_{xx}}}$$

If $\sigma$ is unknown, we estimate $\sigma$ by $\widehat{\sigma} = s$ as defined previously

$$s^2 = \frac{SSE(\widehat{\beta}_0, \widehat{\beta}_1)}{n-2}$$

so that

$$\text{e.s.e.}(\hat{y}_p) = s \sqrt{\frac{1}{n} + \frac{(x_p - \overline{x})^2}{SS_{xx}}}$$

Note: This prediction is the expected value of $y$ at $x = x_p$. That is, we have worked out

$$Var[\widehat{Y}_p] = Var[\widehat{\beta}_0 + \widehat{\beta}_1 x_p]$$

to compute the s.e. for $\widehat{Y}_p$.

But we can actually predict an **error corrupted** version of $\widehat{Y}_p$, $\widehat{Y}_p^\star$ say, where

$$\widehat{Y}_p^\star = \widehat{Y}_p + \epsilon_p$$

where $\epsilon_p$ is a new random error.

But

$$Var[\widehat{Y}_p^\star] = Var[\widehat{Y}_p + \epsilon_p] = Var[\widehat{Y}_p] + Var[\epsilon_p] = Var[\widehat{Y}_p] + \sigma^2$$

that is, there is an **extra** piece of variation due to $\epsilon_p$.

Thus

$$\text{e.s.e.}(\hat{y}_p^\star) = s\sqrt{1 + \frac{1}{n} + \frac{(x_p - \overline{x})^2}{SS_{xx}}} > \text{e.s.e.}(\hat{y}_p)$$

# Prediction Intervals

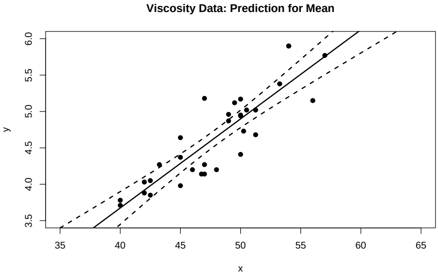
A $100(1-\alpha)\%$ prediction interval for the **mean** value at $x = x_p$ is

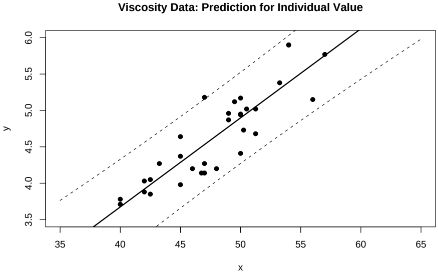
$$\hat{y}_p \pm t_{n-2}(\alpha/2)s\sqrt{\frac{1}{n} + \frac{(x_p - \overline{x})^2}{SS_{xx}}}$$

whereas for an individual new value (predicted with error) at $x = x_p$ is

$$\hat{y}_p \pm t_{n-2}(\alpha/2)s\sqrt{1 + \frac{1}{n} + \frac{(x_p - \overline{x})^2}{SS_{xx}}}$$

# Prediction Intervals





# ANOVA-F test in Regression

An ANOVA-F test can be constructed to test overall (*global*) fit of the linear regression model.

The decomposition of sums of squares for regression takes the form

$$SS = SSR + SSE$$

where

- $SS = SS_{yy}$: overall or total sum of squares
- $SSR$: sum of squares due to <u>R</u>egression
- $SSE$: sum of squares due to <u>E</u>rror

$$
\begin{aligned}
SS &= \sum_{i=1}^{n} (y_i - \overline{y})^2 \\
SSR &= \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 \\
SSE &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2
\end{aligned}
$$

where

$$\hat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i \qquad i = 1, \ldots, n$$

Degrees of Freedom

- TOTAL: $n - 1$
- REGRESSION: 1
- ERROR: $n - 2$

(error d.f. is $n - p$, here $p = 2$).

# The ANOVA Table

| SOURCE | DF | SS | MS | F |
|---|---|---|---|---|
| REGRESSION | 1 | $SSR$ | $MSR = \dfrac{SSR}{1}$ | $F = \dfrac{MSR}{MSE}$ |
| ERROR | $n-2$ | $SSE$ | $MSE = \dfrac{SSE}{(n-2)}$ | |
| TOTAL | $n-1$ | $SS$ | | |

The test of the hypothesis

$$\begin{aligned} H_0 &: \quad E[Y] = \beta_0 \\ H_a &: \quad E[Y] = \beta_0 + \beta_1 x \end{aligned}$$

can be completed by using the test statistic

$$F = \frac{MSR}{MSE}$$

If $H_0$ is true

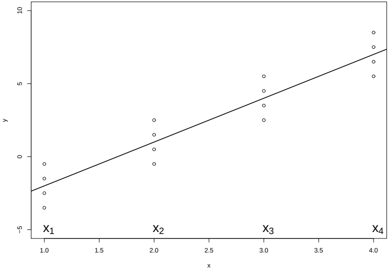
$$F \sim \text{Fisher-F}(1, n-2)$$

This is just like the ANOVA in the one-way layout (CRD) with $n$ groups, but where

$$\mu_i = \beta_0 + \beta_1 x_i$$

That is, the group means are **structured**, that is, we have a formula relating the $\mu_i$ quantities.

Consider four replicates at $x$ values $(x_1, x_2, x_3, x_4)$ in a regression;



Then for group $i$, $\mu_i = \beta_0 + \beta_1 x_i$, $i = 1, 2, 3, 4$.

# Checking the Local Fit

A plot of the *residuals*

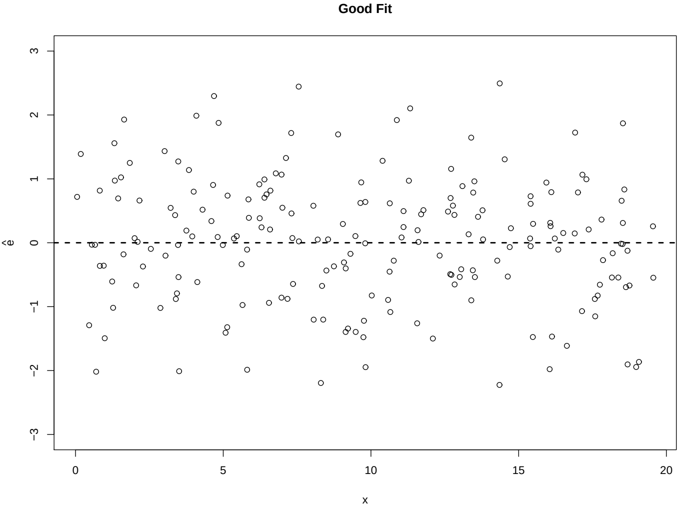
$$\hat{e}_i = y_i - \hat{y}_i$$

can reveal model inadequacies. We should observe that in plots of

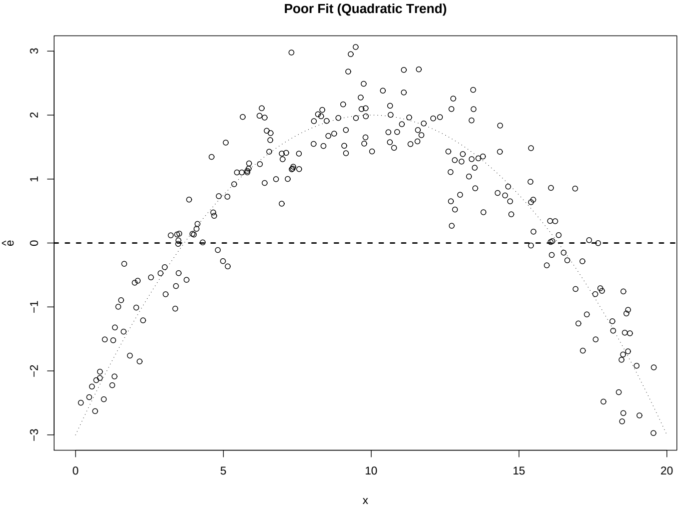
- $x$ vs $\hat{e}$
- $y$ vs $\hat{e}$
- $\hat{y}$ vs $\hat{e}$

there is no discernible pattern

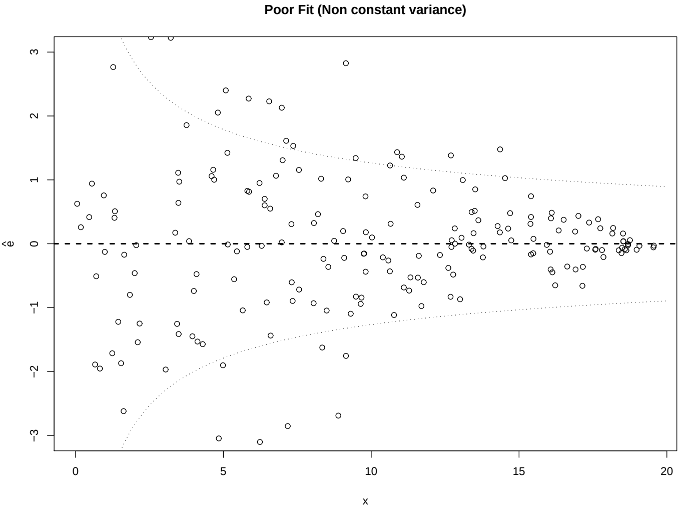
# Checking the Local Fit: Good Fit



# Checking the Local Fit: Poor Fit



# Checking the Local Fit: Poor Fit



# $R^2$ and adjusted $R^2$

SPSS reports both the $R^2$ statistic

$$R^2 = 1 - \frac{SSE}{SS}$$

and the **adjusted** $R^2$ statistic

$$R^2 = 1 - \frac{SSE/EDF}{SS/TDF}$$

where

- $EDF =$ error degrees of freedom $= n - 2$
- $TDF =$ total degrees of freedom $= n - 1$

# 2.1.7 Polynomial Regression

In many practical situations, the simple straight line

$$y = \beta_0 + \beta_1 x$$

is not appropriate. Instead, a model including powers of $x$

$$x^2, x^3, \dots, x^k$$

should be considered. For example

$$y = \beta_0 + \sum_{j=1}^{k} \beta_j x^j = \beta_0 + \beta_1 x + \cdots + \beta_k x^k$$

The **Polynomial Regression Model**

$$Y = \beta_0 + \beta_1 x + \cdots + \beta_k x^k + \epsilon$$

where $\epsilon$ is a random error term as before can be used to model data.

Two immediate problems:

1. How to choose $k$
2. How to carry out inference
    - estimation
    - testing
    - prediction

We begin by addressing 2. The estimation of parameters can be again carried out using **Least Squares** provided that the model assumptions listed before are valid. Consider $k = 2$.

We choose $\underset{\sim}{\beta} = (\beta_0, \beta_1, \beta_2)^{\mathsf{T}}$ to minimize the **sum of squared errors**

$$SSE(\underset{\sim}{\beta}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i - \beta_2 x_i^2)^2$$

that is the fitted values for parameters $\underset{\sim}{\beta}$ are

$$\hat{y}_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$$

$\underset{\sim}{\widehat{\beta}}$ can be found to minimize $SSE$ using calculus techniques (differentiating with respect to the elements of $\underset{\sim}{\beta}$) to give the minimum SSE

$$SSE(\underset{\sim}{\beta}) = \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i - \widehat{\beta}_2 x_i^2)^2$$

We can also compute the estimated **standard errors**

$$s_{\widehat{\beta}_0}, s_{\widehat{\beta}_1}, s_{\widehat{\beta}_2}$$

which allow tests of parameters to be carried out, and confidence intervals calculated.

We can also compute prediction intervals.

The best estimate of the residual error variance $\sigma^2$ is

$$\widehat{\sigma}^2 = \frac{SSE(\underset{\sim}{\widehat{\beta}})}{n-3}$$

$p$ is the number of parameters estimated equal to three, so we divide by $n-3$.

We can also compute

- Residuals
  - can be used to assess the fit of the model.
  - the residuals should be *patternless* if the model fit is good.
- $R^2$, Adjusted $R^2$ statistics
  - used to assess the global fit of the model.
  - used to compare the quality of fit with other models.

### Example (Hooker Pressure Data)

For the Hooker pressure data, a **quadratic** polynomial ($k = 2$) might be suitable.

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2$$

We need to estimate $\beta_0$, $\beta_1$ and $\beta_2$ for these data to see if the model fits better than the straight line model we fitted previously. This can be achieved using SPSS.

It transpires that the quadratic model produces a set of residuals that are patternless, which the straight line model when fitted does not.

See Handout for full details.

Note: It is common to use the **Standardized Residuals**

$$\widehat{z}_i = \frac{\widehat{e}_i}{\widehat{\sigma}} = \frac{y_i - \hat{y}_i}{\widehat{\sigma}}$$

where $\widehat{\sigma}^2$ is the estimate of $\sigma^2$ defined previously, as

$$\mathrm{Var}[\widehat{z}_i] \approx 1$$

if the model fit is good, whereas

$$\mathrm{Var}[\widehat{e}_i] \approx \sigma^2$$

which clearly depends on $\sigma$. This makes it hard to compare $\widehat{e}_i$ across different models when inspecting residuals.

Note: Although the model based on

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

is **not** linear in $x$, it **is** linear in the parameters. Because of this, we still term this a *linear model*. It is this fact that makes the least-squares solutions easy to find.

This model is no more difficult to fit than the model

$$y = \beta_0 + \beta_1 \frac{x}{1+x} + \beta_2 (1 - e^{-x})$$

say - it is still a *linear in the parameters model*. It is in the general class of models

$$y = \beta_0 + \beta_1 g_1(x) + \beta_2 g_2(x)$$

where $g_1(x)$ and $g_2(x)$ are general functions of $x$.

In fact, any model of the form

$$y = \sum_{j=0}^{k} \beta_j g_j(x) + \epsilon \tag{1}$$

can be fitted routinely using least-squares; if we know $x$, then we can compute

$$g_0(x), g_1(x), \ldots, g_k(x)$$

and plug those values into the formula (1).

## Example (Harmonic Regression)

Let

$$\begin{aligned} g_0(x) &= 1 \\ g_1(x) &= \begin{cases} \cos(\lambda_j x) & j \text{ odd} \\ \sin(\lambda_j x) & j \text{ even} \end{cases} \end{aligned}$$

where $k$ is an even number, $k = 2p$ say.

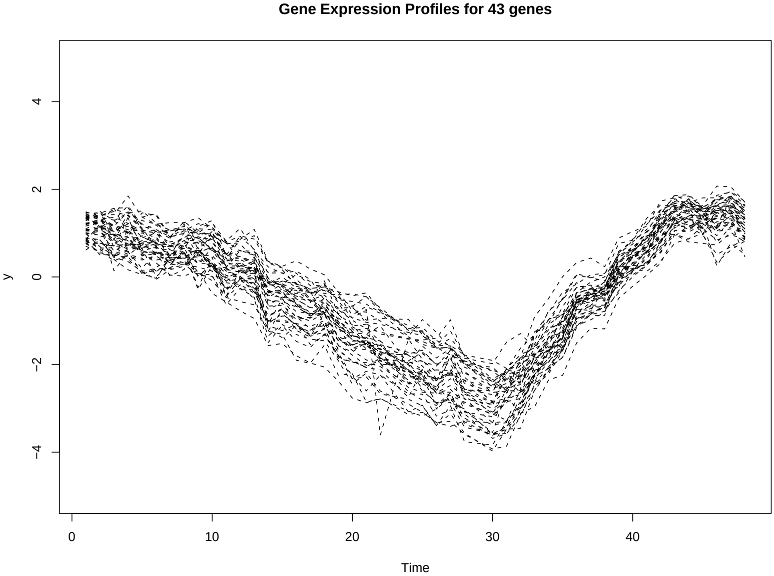
$\lambda_j, j = 1, 2, \ldots, p$ are constants

$$\lambda_1 < \lambda_2 < \cdots < \lambda_p$$

For fixed $x$, $\cos(\lambda_j x)$ and $\sin(\lambda_j x)$ are also fixed, known values.

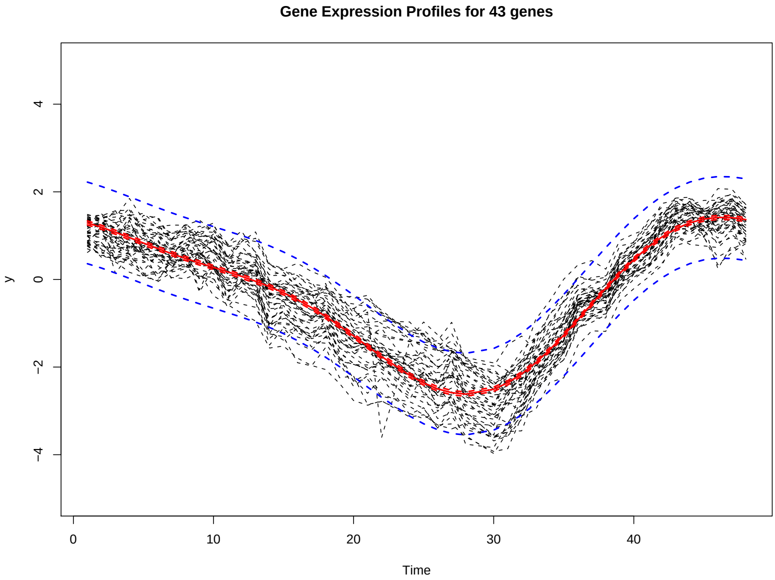
# Gene Expression Data Example

## Harmonic Regression Fit with $p = 2$.



# Gene Expression Data Example

## Harmonic Regression Fit with $p = 2$.



Why are things so straightforward ?

- because the system of equations based on the derivatives

$$\frac{\partial}{\partial \beta_j} \left\{ SSE(\underset{\sim}{\beta}) \right\} = 0 \qquad j = 0, 1, \ldots, k$$

can always be solved routinely, so we can always find $\widehat{\underset{\sim}{\beta}}$.

In the general model (1), simple formulae for

- $\widehat{\underset{\sim}{\beta}}$
- $s.e.(\widehat{\underset{\sim}{\beta}})$
- $\widehat{\sigma}^2$

can be found using a matrix formulation.

**See handout on website - NOT EXAMINABLE !**

Note: One-way ANOVA can be formulated in the form of model (1). Recall

- $k$ independent groups
- means $\mu_1, \ldots, \mu_k$
- $y_{ij}$ - $j$th observation in the $i$th group

Let

$$\begin{aligned} \beta_0 &= \mu_k \\ \beta_t &= \mu_t - \mu_k \qquad t = 1, 2, \ldots, k-1. \end{aligned}$$

Define new data $x_{ij}(t)$ where

$$x_{ij}(t) = \begin{cases} 1 & \text{if } t = i \\ 0 & \text{if } t \neq i \end{cases}$$

Then, using the linear regression formulation

$$y_{ij} = \beta_0 + \sum_{t=1}^{k-1} \beta_t x_{ij}(t) + \epsilon_{ij}.$$

For any $i, j$, $x_{ij}(t)$ is non-zero for only one value of $t$, when $t = i$.

We term this a regression on a *factor predictor*; it is clear that $\beta_0, \beta_1, \ldots, \beta_{k-1}$ can be estimated using least-squares.

This clarifies the link between

ANOVA and Linear Modelling

- they are essentially the SAME MODEL formulation.

This link extends to **ALL ANOVA** models; recall that we used the **General Linear Model** option in SPSS to fit two-way ANOVA.

# 2.2 Multiple Linear Regression

Multiple linear regression models model the variation in response $y$ as a function of **more than one** independent variable.

Suppose we have variables

$$X_1, X_2, \ldots, X_k$$

recording different features of the experimental units. We wish to model response $Y$ as a function of $X_1, X_2, \ldots, X_k$.

# 2.2.1 Multiple Linear Regression Models

Consider the model for datum $i$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$$

where $x_{ij}$ is the measured value of *covariate* $j$ on experimental unit $i$. That is

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \epsilon_i$$

where the first two terms on the right hand side are the *systematic* or *deterministic* components, and the final term $\epsilon_i$ is the *random* component.

## Example ($k = 2$)

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

A three parameter model.

Note: We can also include *interaction* terms

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12}(x_{i1} \,.\, x_{i2}) + \epsilon_i$$

where

- The first two terms in $x_{i1}$ and $x_{i2}$ are **main effects**
- The third term in $(x_{i1} \,.\, x_{i2})$ is an **interaction**

This is a four parameter model.

# Multiple Linear Regression Examples

**SEE HANDOUT**

- Multiple regression: Viscosity Example
- Factor Regression:
- Interaction
- Residuals
- SPSS Instructions

**Subgroup analysis**, with a factor predictor and a continuous covariate, is a form of interaction modelling; the factor predictor *interacts* with the covariate to modify the slope across the subgroups, for example.

We can describe the models using the notation previously introduced for ANOVA; consider the single binary factor predictor and single covariate case;

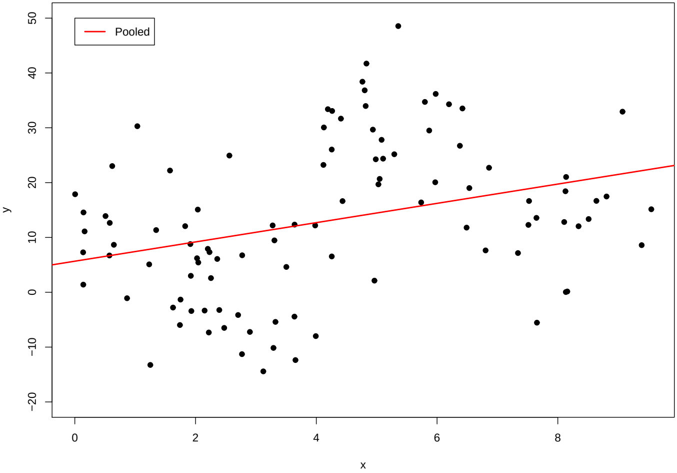
| | | |
|---|---|---|
| MODEL 0 | Single horizontal straight line | $1$ |
| MODEL 1 | Two parallel horizontal straight lines | $X_2$ |
| MODEL 2 | Single straight line, non-zero slope | $X_1$ |
| MODEL 3 | Two parallel straight lines, non-zero slope | $X_1 + X_2$ |
| MODEL 4 | Two non-parallel straight lines | $X_1 + X_2 + X_1.X_2$ |

Note: Always be on the lookout for *lurking* subgroups (subgroups determined by the levels of an unnoticed factor predictor)

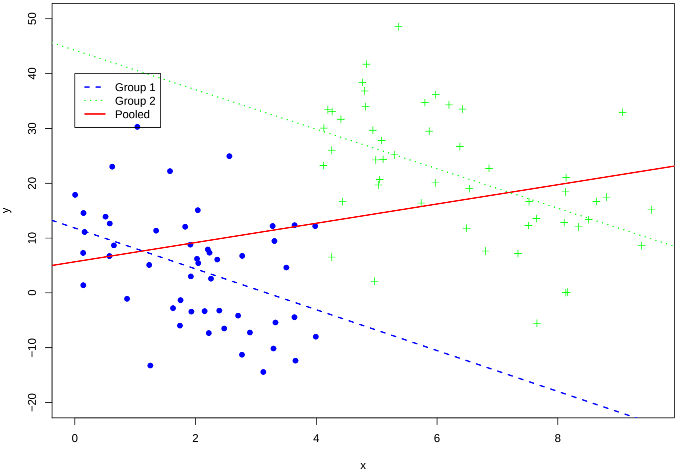
Inferences can change radically when the lurking factor is included in the model

- positive association can be converted into negative association with the continuous covariate.

For example, for factor predictor $X_2$ taking two levels and continuous covariate $X_1$. When the pooled data are examined, a **positive association** between $Y$ and $X_1$ is revealed.



When the pooled data are separated into subgroups, a **negative association** between $Y$ and $X_1$ in each subgroup is revealed.



$X_2 = 1$ (●) and $X_2 = 2$ (+).

i.e. increasing $X_1$ decreases response in subgroup 1, and decreases response in subgroup 2, but appears to increase response overall.

This is known as **Simpson's Paradox in Regression**. It illustrates that pooling data over subgroups must be carried out with care !

- you must fit the factor predictor in the model if you suspect subgroup differences exist.

In the example, the problem arises due to **dependence** between $X_1$ and $X_2$; all the group with $X_2 = 0$ have **low** values of $X_1$, whereas all the group with $X_2 = 1$ have **high** values of $X_1$

Dependence between covariates and factor predictors makes model fitting and results interpretation complicated.

Recap: we can build general models

$$y_i = \beta_0 + \sum_{j=1}^{k} x_{ij} + \epsilon_i$$

to explain the variation of $y$ in terms of covariates and factor predictors $x_1, \dots, x_k$.

- Simple Linear Regression
- Polynomial Regression
- Multiple Regression
- Factor Predictor Regression
- Interaction Models

We can fit each of these models easily using least-squares to obtain

- estimates $\underset{\sim}{\widehat{\beta}} = (\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_k)^\mathsf{T}$
- standard errors
- goodness of fit measures $R^2$ and Adjusted $R^2$
- residuals for model checking
- predictions

# Interpreting $\widehat{\beta}_j$

$\widehat{\beta}_j$ can be interpreted as the amount of increase in response $y$ when $x_j$ increases by one unit when the other predictors

$$x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k$$

are held fixed.

We can test the hypothesis

$$\begin{aligned} H_0 &: \beta_j = 0 \\ H_0 &: \beta_j \neq 0 \end{aligned}$$

using the usual hypothesis testing approach.

Test statistic:

$$t_j = \frac{\widehat{\beta}_j}{s_{\widehat{\beta}_j}} = \frac{\text{ESTIMATE}}{\text{STANDARD ERROR}}$$

If $H_0$ is **true**,

$$t_j \sim Student(n - k - 1)$$

as we are estimating $k + 1$ parameters overall.

Note: In multiple regression, when testing each of

$$\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_k$$

we should strictly use a **multiple testing correction** (as in post-hoc tests in ANOVA)