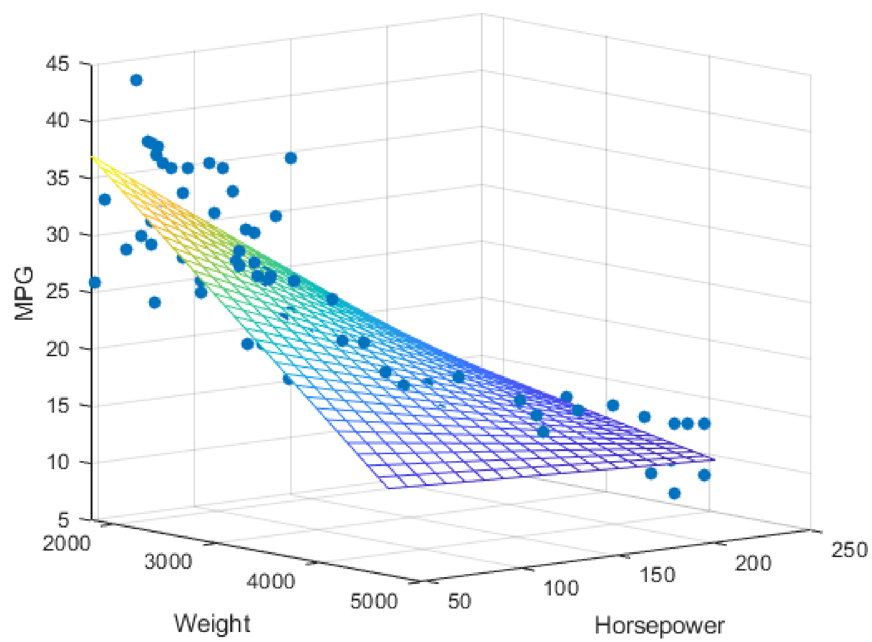
# Slides 17: Multiple regression

- Design matrix
- Least squares estimates
- Matrix formulation of
  the simple linear regression
- t-values
- Quadratic regression
- Adjusted coefficient
  of determination
- Collinearity problem



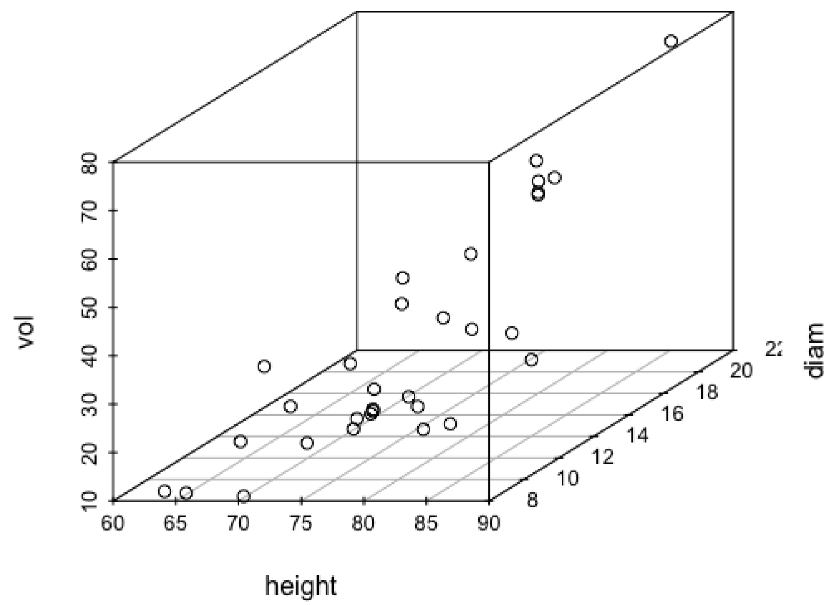
## Example

Trees: $n = 31$

$x_1 =$ diameter

$x_2 =$ height

$y =$ volume




Linear model

$$y_1 = \beta_0 + \beta_1 x_{1,1} + \beta_2 x_{1,2} + e_1,$$

$$\cdots$$

$$y_n = \beta_0 + \beta_1 x_{n,1} + \beta_2 x_{n,2} + e_n,$$

where $e_1, \ldots, e_n$ are independent realisations of the random noise

$$\epsilon \sim \mathrm{N}(0, \sigma).$$

With $p-1$ predictors, the corresponding data set consists of $n$ vectors $(x_{i,1},\ldots,x_{i,p-1},y_i)$ with $n>p$ and

$$y_1 = \beta_0 + \beta_1 x_{1,1} + \ldots + \beta_{p-1} x_{1,p-1} + e_1,$$
$$\ldots$$
$$y_n = \beta_0 + \beta_1 x_{n,1} + \ldots + \beta_{p-1} x_{n,p-1} + e_n.$$

It is very convenient to use the matrix notation

$$\mathbf{y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{e},$$

where $\mathbf{y} = (y_1,\ldots,y_n)^T$, $\boldsymbol{\beta} = (\beta_0,\ldots,\beta_{p-1})^T$, $\boldsymbol{e} = (e_1,\ldots,e_n)^T$ are column vectors, and

$$\mathbb{X} = \begin{pmatrix} 1 & x_{1,1} & \ldots & x_{1,p-1} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n,1} & \ldots & x_{n,p-1} \end{pmatrix}$$

is the so called design matrix assumed to have rank $p$.

The machinery developed for the simple linear regression model works well for the multiple regression. The least squares estimates

$$\mathbf{b} = (b_0, \ldots, b_{p-1})^T$$

give the predicted responses $\hat{\mathbf{y}} = \mathbb{X}\mathbf{b}$ that minimise the sum of squares

$$S(\mathbf{b}) = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 = (y_1 - \hat{y}_1)^2 + \ldots + (y_n - \hat{y}_n)^2$$

The LS estimates must satisfy the normal equations

$$\mathbb{X}^T\mathbb{X}\mathbf{b} = \mathbb{X}^T\mathbf{y}$$

Solving this system of linear equations we get

$$\mathbf{b} = \mathbb{M}\mathbb{X}^T\mathbf{y}, \quad \mathbb{M} = (\mathbb{X}^T\mathbb{X})^{-1}$$

**Question**. What are the dimensions of the matrix $\mathbb{M}$?

In particular, in the simple linear regression case with $p = 2$, we have

$$\mathbb{X}^T = \begin{pmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_n \end{pmatrix}$$

as the transposed design matrix, so that

$$\mathbb{X}^T\mathbb{X} = \begin{pmatrix} n & x_1 + \ldots + x_n \\ x_1 + \ldots + x_n & x_1^2 + \ldots + x_n^2 \end{pmatrix} = n \begin{pmatrix} 1 & \bar{x} \\ \bar{x} & \overline{x^2} \end{pmatrix}$$

Taking the inverse matrix

$$\mathbb{M} = (\mathbb{X}^T\mathbb{X})^{-1} = \frac{1}{n(\overline{x^2} - (\bar{x})^2)} \begin{pmatrix} \overline{x^2} & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}$$

we get LS estimates for the simple linear regression in the matrix form

$$\mathbf{b} = \mathbb{M}\mathbb{X}^T\mathbf{y} = \frac{1}{\overline{x^2} - (\bar{x})^2} \begin{pmatrix} \overline{x^2} & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \begin{pmatrix} \bar{y} \\ \overline{xy} \end{pmatrix}$$

With $\mathbf{b} = \mathbb{M}\mathbb{X}^T\mathbf{y}$, the predicted responses are computed as

$$\hat{\mathbf{y}} = \mathbb{X}\mathbf{b} = \mathbb{P}\mathbf{y}, \qquad \mathbb{P} = \mathbb{X}\mathbb{M}\mathbb{X}^T$$

Check that $\mathbb{P}$ is a projection matrix such that $\mathbb{P}^2 = \mathbb{P}$.

For the random vector $\mathbf{B}$ behind the LS estimates $\mathbf{b}$, we find that

$$\mathrm{E}(\mathbf{B}) = \boldsymbol{\beta}.$$

Furthermore, the covariance matrix, the $p \times p$ matrix with elements $\mathrm{Cov}(B_i, B_j)$, is given by

$$\mathrm{E}\{(\mathbf{B} - \boldsymbol{\beta})(\mathbf{B} - \boldsymbol{\beta})^T\} = \sigma^2\mathbb{M}.$$

The vector of residuals

$$\hat{\boldsymbol{e}} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbb{I} - \mathbb{P})\mathbf{y}$$

has a zero mean vector and a covariance matrix $\sigma^2(\mathbb{I} - \mathbb{P})$.

$$\boxed{s^2 = \frac{SS_{\mathrm{E}}}{n-p}, \quad \text{where } SS_{\mathrm{E}} = S(\mathbf{b}) = \|\hat{\boldsymbol{e}}\|^2 \text{ is an unbiased estimate of } \sigma^2}$$

The data in the following table were gathered for an environmental impact study that examined the relationship between the depth of a stream and the rate of its flow (Ryan et al 1976).

| Depth $x$ | .34 | .29 | .28 | .42 | .29 | .41 | .76 | .73 | .46 | .40 |
|---|---|---|---|---|---|---|---|---|---|---|
| Flow rate $y$ | .64 | .32 | .73 | 1.33 | .49 | .92 | 7.35 | 5.89 | 1.98 | 1.12 |

A bowed shape of the plot of the residuals versus depth suggests that the relation between $x$ and $y$ is not linear. The multiple regression framework can by applied to the quadratic model

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

with $x_1 = x$ and $x_2 = x^2$.

| Coefficient | Estimate | Standard Error | $t$ value |
|---|---|---|---|
| $\beta_0$ | 1.68 | 1.06 | 1.52 |
| $\beta_1$ | $-10.86$ | 4.52 | $-2.40$ |
| $\beta_2$ | 23.54 | 4.27 | 5.51 |

The residuals show no sign of systematic misfit. The test statistic $t = 5.51$ of the utility test of $H_0 : \beta_2 = 0$ shows that the quadratic term in the model is statistically significant.

Define in terms of the diagonal elements $m_{jj}$ of matrix $\mathbb{M}$

$$m_j = m_{j+1,j+1}, \quad j = 0, 1, \ldots, p-1$$

Then the standard error of $b_j$ is computed as

$$s_{b_j} = s\sqrt{m_j}, \quad j = 0, 1, \ldots, p-1.$$

$$\boxed{\text{Exact sampling distributions } \frac{B_j - \beta_j}{S_{B_j}} \sim t_{n-p}, \quad j = 0, 1, \ldots, p-1.}$$

To check the underlying normality assumption inspect the normal probability plot for the standardised residuals $\frac{\hat{e}_i}{s\sqrt{1-p_{ii}}}$, where $p_{ii}$ are the diagonal elements of $\mathbb{P}$.

$$\boxed{\text{Exact } 100(1-\alpha)\% \text{ confidence intervals } I_{\beta_j} = b_j \pm t_{n-p}(\tfrac{\alpha}{2}) \cdot s_{b_j}}$$

For a utility test of $H_0 : \beta_j = 0$, use the t-value

$$b_j / s_{b_j}$$

having $t_{n-p}$-distribution under $H_0 : \beta_j = 0$.

Coefficient of multiple determination can be computed similarly to the simple linear regression model as

$$R^2 = 1 - \frac{SS_{\mathrm{E}}}{SS_{\mathrm{T}}},$$

where $SS_{\mathrm{T}} = (n-1)s_y^2$. The problem with $R^2$ is that it increases even if irrelevant variables are added to the model.

To punish for irrelevant variables it is better to use the adjusted coefficient of multiple determination

$$\boxed{R_a^2 = 1 - \tfrac{n-1}{n-p} \cdot \tfrac{SS_{\mathrm{E}}}{SS_{\mathrm{T}}}}$$

Observe that the adjustment factor $\frac{n-1}{n-p}$ gets larger for the larger number of predictors $p$ in the model, and that

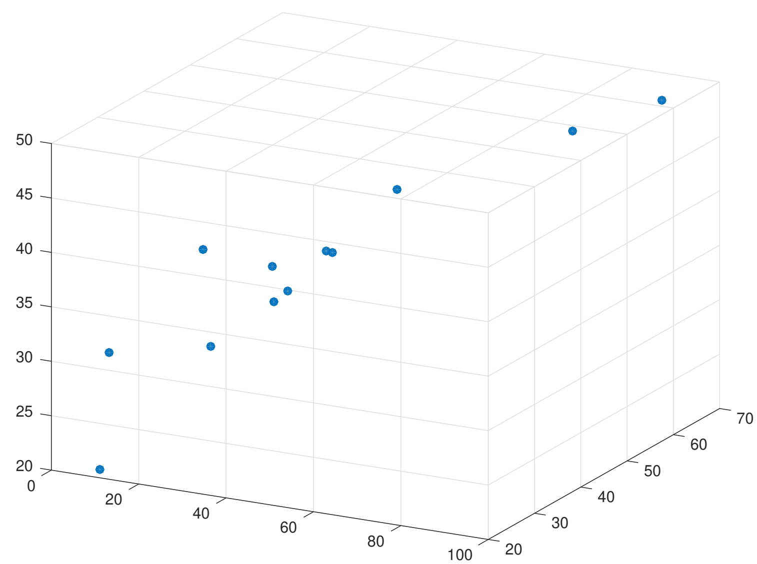
$$1 - R_a^2 = \frac{s^2}{s_y^2}$$

is the proportion of the noise variance of the total variance of responses.

Doctors want predictions on heart catheter length depending on child's height and weight.

The data consist of $n = 12$ observations for the distance to pulmonary artery coming from 12 operations performed earlier:

| Height (in) | Weight (lb) | Length (cm) |
|---|---|---|
| 42.8 | 40.0 | 37.0 |
| 63.5 | 93.5 | 49.5 |
| 37.5 | 35.5 | 34.5 |
| 39.5 | 30.0 | 36.0 |
| 45.5 | 52.0 | 43.0 |
| 38.5 | 17.0 | 28.0 |
| 43.0 | 38.5 | 37.0 |
| 22.5 | 8.5 | 20.0 |
| 37.0 | 33.0 | 33.5 |
| 23.5 | 9.5 | 30.5 |
| 33.0 | 21.0 | 38.5 |
| 58.0 | 79.0 | 47.0 |



We start with two simple linear regressions

$$\text{H-model: } L = \beta_0 + \beta_1 H + \epsilon, \qquad \text{W-model: } L = \beta_0 + \beta_1 W + \epsilon.$$

The analysis of these two models is summarised as follows

| Estimate | H-model | $t$ value | W-model | $t$ value |
|---|---|---|---|---|
| $b_0(s_{b_0})$ | 12.1(4.3) | 2.8 | 25.6(2.0) | 12.8 |
| $b_1(s_{b_1})$ | 0.60(0.10) | 6.0 | 0.28(0.04) | 7.0 |
| $s$ | 4.0 | | 3.8 | |
| $r^2$ | 0.78 | | 0.80 | |

The plots of standardised residuals do not contradict the normality assumptions.

**Question 1**. How can we use the four t-values in the table?

**Question 2**. Which of the two simple linear regression models is more preferable?

These two simple regression models should be compared to the multiple regression model

$$L = \beta_0 + \beta_1 H + \beta_2 W + \epsilon,$$

which gives

$$\begin{array}{lll} b_0 = 21, & s_{b_0} = 8.8, & b_0/s_{b_0} = 2.39, \\ b_1 = 0.20, & s_{b_1} = 0.36, & b_1/s_{b_1} = 0.56, \\ b_2 = 0.19, & s_{b_2} = 0.17, & b_2/s_{b_2} = 1.12, \\ s = 3.9, & R^2 = 0.81 & \end{array}$$

In contrast to the simple models, we can not reject neither $H_1 : \beta_1 = 0$ nor $H_2 : \beta_2 = 0$. This paradox is explained by different meaning of the slope parameters in the simple and multiple regression models.

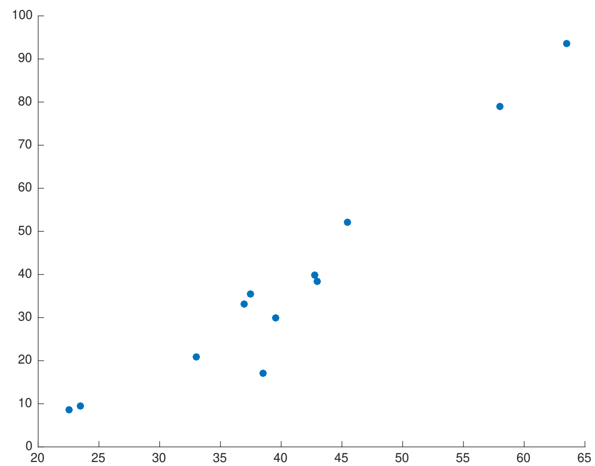
In the multiple model $\beta_1$ is the expected change in $L$ when $H$ increased by one unit and $W$ held constant.

**Question**. Is this model better than H-model and W-model?

The values of $R_a^2$ for three models show that the W-model is the best

| H-model | W-model | (H,W)-model |
|---|---|---|
| 0.76 | 0.78 | 0.77 |

Adding height variable to the weight does not improve the model, since the height and weight have a strong linear relationship.




The fitted plane has a well resolved slope along the line about which the $(H, W)$ points fall and poorly resolved slopes along the $H$ and $W$ axes.