The corresponding estimators are: $\hat{\beta} = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})\, Y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$

and $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}$.

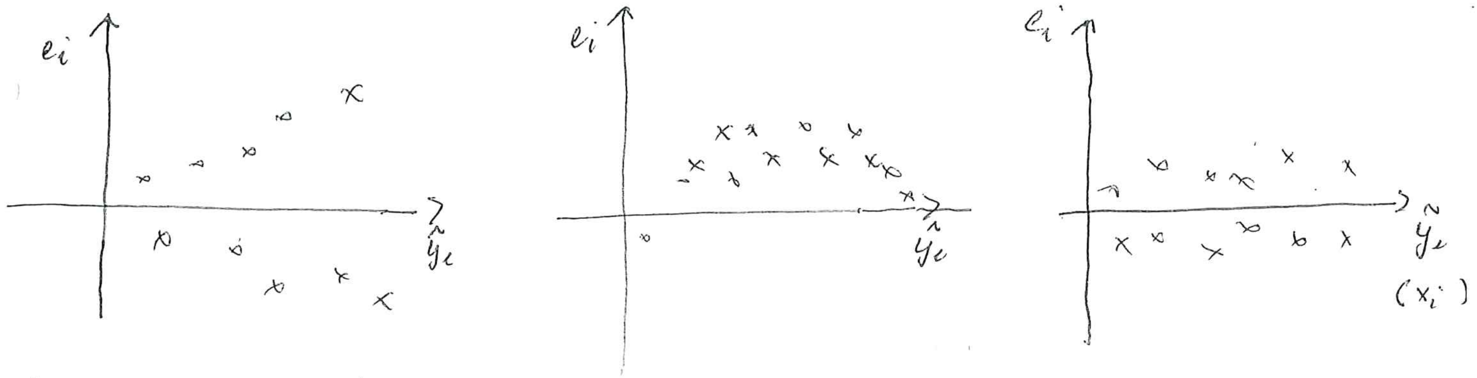
The estimated expected regression line (estimated model) is:

$\hat{y} = a + b x$

Residuals: $e_i = y_i - \hat{y}_i = y_i - a - b x_i$

These should be plotted against $\hat{y}_i$, against x_i and eventually other regression variables. Normal-plot should be used to check for normal distribution

Pattern in the residuals indicates that the model does not fit.



Variance not constant. Transform y

Should consider second order terms

e: $E[Y_i] = \alpha + \beta_1 x_i + \beta_2 x_i^2$

No pattern (ok.)

Transformations for constant variance

Some calculations.

Assume $\quad z = g(y)$

$\Rightarrow \quad z \approx g(\mu) + g'(\mu)(y-\mu) \quad$ where μ is some value. We get for a random variable Z that.

$$Z = g(Y) \approx g(\mu) + g'(\mu)(Y-\mu)$$

and $\quad Var(Z) \approx (g'(\mu))^2 \, Var(Y)$

For this to be constant, we must have

$$(g'(\mu))^2 . \, Var(Y) = k$$

$$\Rightarrow g'(\mu) = \frac{\sqrt{k}}{SD(Y)}$$

Such that if $SD(Y) \propto \mu \Rightarrow g'(\mu) = \frac{\sqrt{k}}{c\mu}$ and $g(Y) = \ln(Y)$ should work.

If $SD(Y) \propto \mu^2 \Rightarrow g'(\mu) = \frac{\sqrt{k}}{c\mu^2}$ and $g(Y) = \frac{1}{Y}$ will help.

If $SD(Y) \propto \sqrt{\mu} \Rightarrow g'(\mu) = \frac{\sqrt{k}}{c\sqrt{\mu}}$ and $g(Y) = \sqrt{Y}$ will help.

The most common transformations for constant variance are thus.

$Y^{\circ} = \ln(Y), \quad \sigma \propto E[Y]$

$Y^{\circ} = \sqrt{Y}, \quad \sigma^2 \propto E[Y]$

$Y^{\circ} = \frac{1}{Y}, \quad \sigma \propto (E[Y])^2.$

Some transformations give directly a linear model.

$$Y_i = \alpha e^{\beta X_i + E_i} \Rightarrow \ln(Y_i) = \ln\alpha + \beta X_i + E_i$$

$$Y_i = \alpha X_i^{\beta} \cdot E_i \Rightarrow \ln(Y_i) = \ln\alpha + \beta \ln X_i + \ln(E_i)$$

$$Y_i = \frac{X_i}{\alpha + (\beta + E_i) X_i} \Rightarrow \frac{1}{Y_i} = \frac{\alpha}{X_i} + \beta + E_i$$

Chapter 12. Multiple Linear Regression

Model: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i \begin{cases} E[\varepsilon_i]=0 \\ Var[\varepsilon_i] \end{cases}$, independent

Response

regression variables.

We get $E[Y_i \mid X_{1i}, X_{2i}, \ldots, X_{ki}] = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$

We observe: $(y_i, x_{1i}, x_{2i}, \ldots, x_{ki})$, $i = 1, 2, \ldots, n$

that according to the model must satisfy

$$y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{21} + \cdots + \beta_k x_{k1} + \varepsilon_1^\circ$$
$$y_2 = \beta_0 + \beta_1 x_{12} + \beta_2 x_{22} + \cdots + \beta_k x_{k2} + \varepsilon_2^\circ$$
$$\vdots$$
$$y_n = \beta_0 + \beta_1 x_{1n} + \beta_2 x_{2n} + \cdots + \beta_k x_{kn} + \varepsilon_n^\circ$$

or,

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & & x_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1^* \\ \varepsilon_2^* \\ \vdots \\ \varepsilon_n^* \end{bmatrix}$$

or $\quad \underline{y} = \underline{X}\,\underline{\beta} + \underline{\varepsilon}^*$

The estimates for $\underline{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$ are the values for $\underline{b} = (b_0, b_1, \ldots, b_k)'$ that minimizes

$$Q = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i} - \cdots - b_k x_{ki})^2 .$$

The estimated model then becomes

$$\hat{y}_i = b_0 + b_1 x_{1i} + \dots + b_k x_{ki}$$

The normal equations are found by derivating Q with respect to $b_0, b_1, \dots, b_k$ and setting the partial derivatives to 0.

We get.

$$\frac{\partial Q}{\partial b_0} = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i} - \dots - b_k x_{ki}) = -2 \sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$$

$$\frac{\partial Q}{\partial b_1} = -2 \sum_{i=1}^{n} x_{1i} (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i} - \dots - b_k x_{ki}) = -2 \sum_{i=1}^{n} x_{1i} (y_i - \tilde{y}_i) = 0$$

$$\vdots$$

$$\frac{\partial Q}{\partial b_k} = -2 \sum_{i=1}^{n} x_{ki} (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i} - \dots - b_k x_{ki}) = -2 \sum_{i=1}^{n} x_{ki} (y_i - \hat{y}_i) = 0$$

Which can be written as

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{12} & \cdots & x_{1n} \\ \vdots & & & \\ x_{k1}, & x_{k2}, & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \\ y_n - \tilde{y}_n \end{bmatrix} = 0 \quad \text{or} \quad \underline{X}'(\underline{y} - \hat{\underline{y}}) = \underline{X}'(\underline{y} - \underline{X}\underline{b}) = 0$$

or $\underline{X}'\underline{X}\,\underline{b} = \underline{X}'\underline{y}$ and $\underline{b}_{LS} = (\underline{X}'\underline{X})^{-1}\underline{X}'\underline{y}$