Statistical inference (MVE155/MSG200)

Multiple regression

Simple linear regression

Describes linear relationship between a (random) response variable Y and a (deterministic) predictor x, i.e.

$$Y = \beta_0 + \beta_1 x + \sigma Z,$$

where $Z \sim N(0,1)$.

Note that the noise $\sigma > 0$ is constant (homoscedastic) and does not depend on the value of x. If σ varies with x, the situation is called heteroscedastic.

Simple linear regression: Data

Data consist of n pairs of independent observations

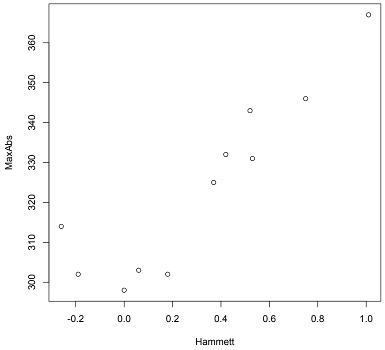
$$(x_1, y_1), ..., (x_n, y_n),$$

where

$$y_i = \beta_0 + \beta_1 x_i + e_i,$$

where E_i's (stochastic variants or e_i's) are iid and $N(0, \sigma)$-distributed.

Can we describe the maximum absorbance rate (y) (in nanomoles) as a linear function of the Hammett constant (x), i.e. $y = \beta_0 + \beta_1 x$, for a particular compound?



Hammett (x)	0.00	0.75	0.06	-0.26	0.18	0.42	-0.19	0.52	1.01	0.37	0.53
Max abs rate (y)	298	346	303	314	302	332	302	343	367	325	331

Parameter estimation: least squares

β_0 and β_1 are estimated by minimizing the sum of squares of the residuals $y_i - \hat{y}_i = y_i - \beta_0 - \beta_1 x_i$, i.e.

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

with respect to β_0 and β_1 (least squares estimates) or by using the maximum likelihood method.

Both methods above lead to the same estimators.

Parameter estimation: maximum likelihood

The parameters β_0, β_1, and σ^2 can be estimated by maximizing the likelihood function

$$
\begin{aligned}
L(\beta_0, \beta_1, \sigma^2) &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \prod_{i=1}^{n} \exp\left(-\frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{2\sigma^2}\right) \\
&= (2\pi)^{-\frac{n}{2}} (\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2\right)
\end{aligned}
$$

or the log likelihood

$$
-\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.
$$

Parameter estimation

The ML estimates for the parameters β_0, β_1, and σ^2 are

$$b_0 = \bar{y} - b_1\bar{x}, \quad b_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}, \quad \hat{\sigma^2} = \frac{ss_E}{n} = \frac{1}{n}\sum_{i=1}^{n} \hat{e}^2,$$

where $\overline{xy} = \frac{1}{n}\sum_i x_i y_i$, $\overline{x^2} = \frac{1}{n}\sum_i x_i^2$, and $\hat{e}_i = y_i - b_0 - b_1 x_i$, $i = 1, \ldots, n$, are the residuals.

The ML estimator $\hat{\sigma}^2 = SS_E/n$ for σ^2 is biased and an unbiased estimator is given by

$$S^2 = \frac{SS_E}{n-2}.$$

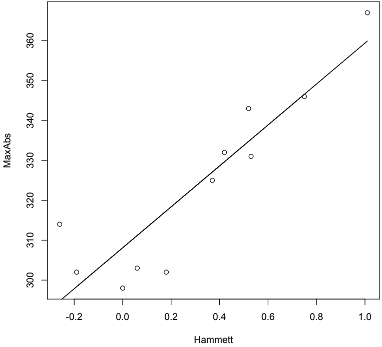
Example (continues)

Data: $n = 11$, $\bar{x} = 0.3082$, $\bar{y} = 323.9$, $\overline{xy} = 107.0$, $\overline{x^2} = 0.2352$
giving

$$b_1 = 51.2,\ \ b_0 = \bar{y} - b_1\bar{x} = 308.1,\ \ \hat{\sigma} = 10.2.$$

and

$$y = 308.1 + 51.2x.$$



ML estimators

B_0 and B_1 (stochastic versions of b_0 and b_1) are unbiased estimators for β_0 and β_1, respectively. Also,

$$B_0 \sim N\left(\beta_0, \sqrt{\frac{\sigma^2 \sum x_i^2}{n(n-1)s_x^2}}\right), \quad B_1 \sim N\left(\beta_1, \sqrt{\frac{\sigma^2}{(n-1)s_x^2}}\right)$$

Therefore,

$$\frac{B_0 - \beta_0}{S_{B_0}} \sim t_{n-2} \quad \text{and} \quad \frac{B_1 - \beta_1}{S_{B_1}} \sim t_{n-2}$$

(where $S_{B_0}^2 = S^2 \sum x_i^2 / n(n-1)s_x^2$ and $S_{B_1}^2 = S^2/(n-1)s_x^2$).

Furthermore, there is a weak correlation between the estimators,

$$\text{Cov}(B_0, B_1) = -\frac{\sigma^2 \bar{x}}{(n-1)s_x^2}.$$

Confidence intervals

$100(1-\alpha)\%$ confidence intervals for β_0 and β_1 become

$$I_{\beta_0} = b_0 \pm t_{n-2}(\alpha/2) \cdot s_{b_0} \quad \text{and} \quad I_{\beta_1} = b_1 \pm t_{n-2}(\alpha/2) \cdot s_{b_1}.$$

and the null hypotheses $H_0 : \beta_1 = \beta^*$ and $H_0 : \beta_0 = \beta^*$ can be tested by using the test statistics

$$T = \frac{B_1 - \beta^*}{S_{B_1}} \quad \text{and} \quad T = \frac{B_0 - \beta^*}{S_{B_0}},$$

respectively, which are both t_{n-2}-distributed under H_0. Typically, one tests

- $H_0 : \beta_1 = 0$, no linear relationship between the response y and predictor x.
- $H_0 : \beta_0 = 0$, the intercept is zero.

Prediction intervals

Given the parameter estimates b_0 and b_1, we can predict the value of a new x-value, x_p, (within the interval $(\min\{x_1, ..., x_n\}, \max\{x_1, ..., x_n\})$) using

$$y_p = b_0 + b_1 x_p + \hat{\sigma} z_p,$$

where $Z_p \sim N(0, 1)$ is independent of the sample $(x_1, y_1), ..., (x_n, y_n)$.

The expected value of Y_p is

$$\mu_p = \beta_0 + \beta_1 x_p$$

and its estimator $\hat{\mu}_p = B_0 + B_1 x_p$.

Prediction intervals

The variance of $\hat{\mu}_p$ is

$$\begin{aligned} \mathrm{Var}(B_0 + B_1 x_p) &= \mathrm{Var}(B_0) + x_p^2 \mathrm{Var}(B_1) + 2x_p \mathrm{Cov}(B_0, B_1) \\ &= \frac{\sigma^2}{n} + \frac{\sigma^2}{n-1}\left(\frac{x_p - \bar{x}}{s_x}\right)^2 \end{aligned}$$

and the variance of Y_p

$$\mathrm{Var}(Y_p) = \mathrm{Var}(\hat{\mu}_p + \sigma Z_p) = \sigma^2 \left(1 + \frac{1}{n} + \frac{1}{n-1}\left(\frac{x_p - \bar{x}}{s_x}\right)^2\right)$$

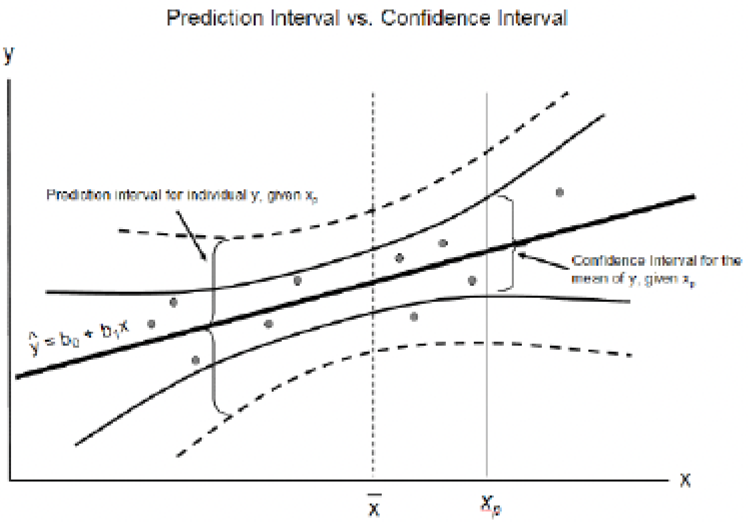
leading to the $100(1-\alpha)\%$ confidence interval for μ_p

$$I_{\mu_p} = b_0 + b_1 x_p \pm t_{n-2}(\alpha/2)\, s \sqrt{\frac{1}{n} + \frac{1}{n-1}\left(\frac{x_p - \bar{x}}{s_x}\right)^2}$$

and the $100(1-\alpha)\%$ prediction interval for y_p

$$I_{Y_p} = b_0 + b_1 x_p \pm t_{n-2}(\alpha/2)\, s \sqrt{1 + \frac{1}{n} + \frac{1}{n-1}\left(\frac{x_p - \bar{x}}{s_x}\right)^2}$$

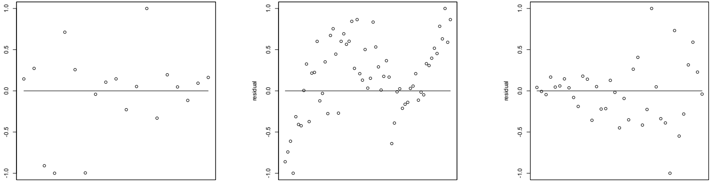
Prediction intervals



Residuals

The random variables (residuals) $\hat{E}_i$ are normally distributed with zero means and weakly correlated with each other.

Under the simple regression model, the scatter plot of the residuals $\hat{e}_i$ versus x_i should be randomly scattered around the x-axis (left, our previous example). The residual plot will reveal if the simple linear model is not good (middle) or if the noise variance is not constant (right).



The normality can be checked by plotting a normal QQ-plot between ordered residuals and standard normal quantiles.

Connection between b_1 and sample correlation coefficient

The sample correlation coefficient is

$$r = \frac{s_{xy}}{s_x s_y},$$

where

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \quad s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2,$$

and

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

Note that

$$b_1 = \frac{r s_y}{s_x}.$$

Coefficient of determination

As in ANOVA, we can describe the observations by using sums of squares. First, we can write

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

and then, by taking squares and summing over all observations, we obtain

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

or equivalently,

$$ss_T = ss_R + ss_E,$$

where

- $ss_T = (n-1)s_y^2$ is the total sum of squares
- $ss_R = (n-1)b_1^2 s_x^2 = (n-1)r^2 s_y^2$ is the regression sum of squares
- ss_E is the residual (error) sum of squares.

Coefficient of determination

Therefore,

$$\frac{ss_R}{ss_T} = \frac{(n-1)r^2 s_y^2}{(n-1)s_y^2} = r^2 \quad \text{and} \quad \frac{ss_E}{ss_T} = 1 - r^2,$$

and r^2 is called the coefficient of determination.

Also,

$$ss_E = ss_T(1-r^2) = (n-1)s_y^2(1-r^2)$$

giving an unbiased estimator for σ^2, namely

$$s^2 = \frac{ss_E}{n-2} = \frac{n-1}{n-2} s_y^2 (1-r^2).$$

Multiple linear regression

We can have any number (less than n) of predictors in a regression model. If we have $p-1$, $p \geq 2$, predictors, our data consist of

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

where $e_1, \ldots, e_n$ are independently generated from the distribution $N(0, \sigma)$.

Multiple linear regression

We can write

$$\mathbf{y} = (y_1, ..., y_n)^T, \; \beta = (\beta_0, ..., \beta_{p-1})^T, \; \mathbf{e} = (e_1, ..., e_n)^T$$

and give the multiple regression model in the form

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

where

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

is called a design matrix.

Multiple linear regression: estimates

The least squares estimates $\mathbf{b} = (b_0, ..., b_{p-1}^T)$ are

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

which (the stochastic variants) are unbiased estimators for β. The covariance matrix is given by

$$\mathbb{E}((\mathbf{B} - \beta)(\mathbf{B} - \beta)^T) = \sigma^2 (\mathbb{X}^T \mathbb{X})^{-1}.$$

Note that the diagonal elements of this matrix give the variances of the parameter estimators.

The predicted responses become

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

where $\mathbb{P} = \mathbb{X}(\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T$.

Multiple linear regression: residuals

The residuals are defined as in the single predictor case, i.e.

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

The residuals have zero means and the covariance matrix $\sigma^2(\mathbb{I} - \mathbb{P})$.

An unbiased estimate for σ^2 is given by

$$s^2 = \frac{\hat{e}_1^2 + \ldots + \hat{e}_n^2}{n-p} = \frac{ss_E}{n-p}.$$

Multiple linear regression: Hypothesis testing

As in the single predictor case, the parameter estimators B_i, $i = 0, ..., p-1$ are normally distributed and

$$\frac{B_j - \beta_j}{S_{B_j}} \sim t_{n-p}.$$

Often, one tests the null hypotheses $H_0 : \beta_i = 0$, against $H_1 : \beta_i \neq 0$, $i = 0, ..., p-1$.

Multiple linear regression: Coefficient of multiple determination

The coefficient of multiple determination can be computed as in the simple regression model with one predictor, i.e.

$$R^2 = 1 - \frac{ss_E}{ss_T},$$

where $ss_T = (n-1)s_y^2$.

Since R^2 is increasing when new predictors are added (whether they have a relationship with the response variable of not), the coefficient should be adjusted so that it does not overestimate the contribution of the predictors. The adjusted coefficient is defined as

$$R_a^2 = 1 - \frac{n-1}{n-p} \cdot \frac{ss_E}{ss_T} = 1 - \frac{s^2}{s_y^2}.$$

which approaches to R^2 when p decreases.

Remark

We can use multiple regression even in the case of a more complex model in terms of one variable, for example

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 \quad \text{or} \quad y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3.$$

In the first case, we can set

$$x_1 = x \quad \text{and} \quad x_2 = x^2$$

and in the second case,

$$x_1 = x, \quad x_2 = x^2, \quad \text{and} \quad x_3 = x^3.$$

Case study: catheter length

Heart catherization is sometimes performed on children with congenital heart defects by using a Teflon tube (catheter). The length of the catheter, y, is determined by the child's height h and/or the child's weight w. In the study, $n = 12$.Three regression models are compared:

- Model 1: $y = \beta_0 + \beta_1 h + \sigma z$
- Model 2: $y = \beta_0 + \beta_1 w + \sigma z$
- Model 3: $y = \beta_0 + \beta_1 h + \beta_2 w + \sigma z$

Case study: catheter length

The null hypotheses that are tested below are $H_0 : \beta_i = 0$, $i = 0, 1, 2$. In the table below, * means that the test result is significant at 5% level.

Estimates	Model 1 (height)	t-value	Model 2 (weight)	t-value	Model 3 (both)	t-value
$b_0(s_{b_0})$	12.1(4.3)	2.8^*	25.6(2.0)	12.8^*	21(8.8)	2.39^*
$b_1(s_{b_1})$	0.6(0.10)	6.0^*	0.28(0.04)	7.0^*	0.20(0.36)	0.56
$b_2(s_{b_2})$	-	-	-	-	0.19(0.17)	1.12
s	4.0		3.8		3.9	
R^2	0.78		0.80		0.81	
R_a^2	0.76		0.78		0.77	

Which model is the best one?