# 6. Regression and one-way analysis of variance

Systematic relationship between explanatory variable and response?

*Example:* Father's height - son's height

Scatterplot, see lecture

Data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

$x$'s: explanatory variables

$y$'s: response variables

Denote the sample variance for $x_1, \ldots, x_n$ by $s_x^2$ and the sample variance for $y_1, \ldots, y_n$ by $s_y^2$

The *sample correlation (Pearson's correlation coefficient)* is defined as

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

Often we abbreviate

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

*Exercise:* We can write $R$ as

$$R = \frac{n\sum_i x_i y_i - \sum_i x_i \sum_i y_i}{\sqrt{n\sum_i x_i^2 - (\sum_i x_i)^2}\sqrt{n\sum_i y_i^2 - (\sum_i y_i)^2}}$$

*Recall: Simple linear regression*

Model

$$Y_i = \alpha + \beta(x_i - \bar{x}) + \epsilon_i, \quad i = 1, \ldots, n$$

where the $\epsilon_i$ are independent, identically distributed, mean zero, (unknown) variance $\sigma^2$

The parameters of the model are $\alpha, \beta, \sigma^2$

*Least-squares estimates:* Minimize

$$\sum_{i=1}^{n} (y_i - \alpha - \beta(x_i - \bar{x}))^2$$

gives

$$\hat{\alpha} = \bar{Y}$$

$$\hat{\beta} = \frac{n \sum_i x_i y_i - \sum_i x_i \sum_i y_i}{n \sum_i x_i^2 - (\sum_i x_i)^2}$$

Note:

$$\widehat{\beta} = R\frac{s_y}{s_x}$$

The *least-squares regression line* is

$$y = \widehat{\alpha} + \widehat{\beta}(x - \bar{x})$$

If $x = \bar{x}$ then $y = \bar{y}$, so goes through the point of averages $(\bar{x}, \bar{y})$

If $x$ increases by one $s_x$ then $y$ increases by $Rs_y$

$R^2$ measures how much of the total variability of the data is accounted for by the model.

*Example* Expected life spans for men and women, see lecture.

*Regression fallacy:* In virtually all test-retest situations, the bottom group on the first test will on average show some improvement on the second test - and the top group will on average fall back. This is the regression effect. The regression fallacy consists of thinking that the regression effect must be due to something important, not just the spread around the line.

*Example:* Repeated coin toss, see lecture.

*Recall:* If in addition each $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ then the least-squares estimates are the maximum-likelihood estimates

$\widehat{\alpha}, \widehat{\beta}$ are unbiased, independent,

$$\widehat{\alpha} \sim \mathcal{N}\left(\alpha, \frac{\sigma^2}{n}\right)$$

$$\widehat{\beta} \sim \mathcal{N}\left(\beta, \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}\right)$$

*Variance estimator*

Residual sum of squares (error sum of squares) is

$$SSE = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2$$

with $\widehat{Y}_i = \widehat{\alpha} + \widehat{\beta}(x - \bar{x})$

Then

$$\widehat{\sigma}^2 = \frac{SSE}{n-2}$$

is unbiased for $\sigma^2$

More explanatory variables: **Multiple linear regression**

Data $(x_{i,1}, x_{i,2}, \ldots, x_{i,(p-1)}; y_i)$ for $i = 1, \ldots, n$

(Example: father's height, mother's height, grandfathers' height all as explanatory variables)

*Model*

$$Y_i = \theta_0 + \theta_1 x_{i,1} + \theta_2 x_{i,2} + \cdots + \theta_{p-1} x_{i,p-1} + \epsilon_i,$$

for $i = 1, \ldots, n$

with $\epsilon_i$'s independent, identically distributed, mean zero, (unknown) variance $\sigma^2$

Shorthand: Put

$$\underline{Y} = (Y_1, \ldots, Y_n)^T \quad (n \times 1) - vector$$

$$\underline{\theta} = (\theta_0, \ldots, \theta_{p-1})^T \quad (p \times 1) - vector$$

$$\underline{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T \quad (n \times 1) - vector$$

$$\mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,p-1} \\ 1 & x_{2,1} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,p-1} \end{pmatrix}$$

is an $n \times p$ matrix called the *design matrix*; then we can write the model as

$$\underline{Y} = \mathbf{X}\underline{\theta} + \underline{\epsilon}$$

Least-Squares estimation:

Minimize

$$\begin{aligned} S(\underline{\theta}) &= \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_{i,1} - \theta_2 x_{i,2} - \cdots \\ &\quad -\theta_{p-1} x_{i,p-1})^2 \\ &= \| \underline{Y} - \mathbf{X}\underline{\theta} \|^2, \end{aligned}$$

where $\| \cdot \|$ is the $L_2$-norm

Differentiate:

$$\begin{aligned} \frac{\partial}{\partial \theta_j} S(\underline{\theta}) &= -2 \sum_{i=1}^{n} x_{i,j} (y_i - \theta_0 - \theta_1 x_{i,1} - \theta_2 x_{i,2} - \cdots \\ &\quad -\theta_{p-1} x_{i,p-1}) \end{aligned}$$

with $x_{0,j} = 1$, so

$$\sum_{j=0}^{p-1} \frac{\partial}{\partial \theta_j} S(\underline{\theta}) = -2(\mathbf{X}^T \underline{Y} - \mathbf{X}^T \mathbf{X} \underline{\theta})$$

Thus the minimizing $\underline{\widehat{\theta}}$ satisfies the *normal equations*

$$\mathbf{X}^T\mathbf{X}\underline{\theta} = \mathbf{X}^T\underline{Y}.$$

If $\mathbf{X}^T\mathbf{X}$ is non-singular, then we obtain

$$\underline{\widehat{\theta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\underline{Y}.$$

*Example: Simple linear regression*

Model

$$Y_i = \theta_0 + \theta_1 x_i + \epsilon_i$$

Here $p = 1$ and

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

and

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}$$

so that

$$\begin{aligned}(\mathbf{X}^T\mathbf{X})^{-1} &= \frac{1}{n\sum x_i^2 - (\sum x_i)^2} \times \\ &\quad \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix}\end{aligned}$$

and

$$\mathbf{X}^T\underline{Y} = \begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n x_i Y_i \end{pmatrix}$$

yielding

$$\begin{aligned}\underline{\hat{\theta}} &= \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} \\ &= \frac{1}{n\sum x_i^2 - (\sum x_i)^2} \times \\ &\quad \begin{pmatrix} \sum_{i=1}^n x_i^2 \sum_{i=1}^n Y_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i Y_i \\ -\sum_{i=1}^n x_i \sum_{i=1}^n Y_i + n\sum_{i=1}^n x_i Y_i \end{pmatrix}\end{aligned}$$

and

$$n\sum_{i=1}^{n} x_i^2 - (\sum x_i)^2 \;=\; n\sum_{i=1}^{n}(x_i-\bar{x})^2$$

and

$$\begin{aligned} & -\sum_{i=1}^{n} x_i \sum_{i=1}^{n} Y_i + n\sum_{i=1}^{n} x_i Y_i \\ = & \sum_{i=1}^{n}(x_i-\bar{x})Y_i \\ = & \sum_{i=1}^{n}(x_i-\bar{x})(Y_i-\bar{Y}) \end{aligned}$$

so this reduces to the usual estimators

$$\begin{aligned} \hat{\beta} &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})Y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \\ \hat{\alpha} &= \bar{Y}-\hat{\beta}\bar{x}. \end{aligned}$$

*Vector-valued random elements*

In general, if $\underline{Y} = (Y_1, \ldots, Y_n)^T$ is a vector of random variables, and

$$EY_i = \mu_i, Var(Y_i) = \sigma_i^2 = \sigma_{ii}, Cov(Y_i, Y_j) = \sigma_{ij}$$

then the *mean vector* is

$$E\underline{Y} = (\mu_1, \ldots, \mu_n)^T$$

and the *variance-covariance matrix* is

$$V(\underline{Y}) = \Sigma_{YY} = (\sigma_{ij})_{1 \leq i,j \leq n}$$

If $\underline{W}$ and $\underline{Y}$ are random vectors and if $\mathbf{L}$ is a deterministic matrix, then if $\underline{W} = \mathbf{L}\underline{Y}$

$$\begin{aligned} E\underline{W} &= \mathbf{L}E\underline{Y} \\ V(\underline{W}) &= \mathbf{L}V(\underline{Y})\mathbf{L}^T \end{aligned}$$

For $\epsilon_1, \ldots, \epsilon_n$ independent $\mathcal{N}(0, \sigma^2)$, we abbreviate

$$\underline{\epsilon} \sim \mathcal{MVN}(\underline{0}, \sigma^2 \mathbf{I}_n)$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix and $\mathcal{MVN}$ stands for the multivariate normal distribution

If, in the multiple linear regression model, we assume in addition that the errors $\underline{\epsilon} \sim \mathcal{MVN}(\underline{0}, \sigma^2 \mathbf{I}_n)$, then the likelihood is

$$L(\underline{\theta}, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \times \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \underline{x}_i^T \underline{\theta})^2 \right\}$$

where $\underline{x}_i^T = (1, x_{i,1}, \ldots, x_{i,p-1})$

Taking logarithms yields that the maximum likelihood estimators for $\underline{\theta}$ are obtained by maximizing

$$-\sum_{i=1}^{n} (y_i - \underline{x}_i^T \underline{\theta})^2 = -(\underline{y} - \mathbf{X}\underline{\theta})^T (\underline{y} - \mathbf{X}\underline{\theta})$$

and so again the same as the least-squares estimators

*Distribution of* $\underline{\widehat{\theta}}$

*Assumption:* $\mathbf{X}^T\mathbf{X}$ is non-singular

$$\underline{Y} = \mathbf{X}\underline{\theta} + \underline{\epsilon}$$

with $\underline{\epsilon} \sim \mathcal{MVN}(\underline{0}, \sigma^2\mathbf{I}_n)$, then

$$\underline{Y} \sim \mathcal{MVN}(\mathbf{X}\underline{\theta}, \sigma^2\mathbf{I}_n)$$

and

$$\underline{\widehat{\theta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\underline{Y}.$$

*Claim:* $\underline{\widehat{\theta}}$ is unbiased for $\underline{\theta}$

*Proof:*

$$\begin{aligned} E\underline{\widehat{\theta}} &= E(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\underline{Y} \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E\underline{Y} \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\underline{\theta} = \underline{\theta}. \end{aligned}$$

*Claim:* $V(\underline{\hat{\theta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$

*Proof:* Put $B = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, then $\underline{\hat{\theta}} = B\underline{Y}$ and

$$\begin{aligned} V(\underline{\hat{\theta}}) &= BV(\underline{Y})B^T \\ &= \sigma^2 BB^T \\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}. \end{aligned}$$

Also $\underline{\hat{\theta}}$ is a vector of linear combinations of normally distributed random variables, so: multivariate normal

Conclude:

$$\underline{\hat{\theta}} \sim \mathcal{MVN}(\underline{\theta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$$

*Example: Simple linear regression*

Model $Y_i = \theta_0 + \theta_1 x_i + \epsilon_i$, so

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{n\sum x_i^2 - (\sum x_i)^2} \times \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix}$$

and if $\sum_{i=1}^n x_i = 0$ then

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{n\sum x_i^2} \times \begin{pmatrix} \sum_{i=1}^n x_i^2 & 0 \\ 0 & n \end{pmatrix}$$

and the estimators $\hat{\alpha}$ and $\hat{\beta}$ are independent, unbiased,

$$\hat{\alpha} \sim \mathcal{N}\left(\alpha, \frac{\sigma^2}{n}\right)$$

$$\hat{\beta} \sim \mathcal{N}\left(\beta, \frac{\sigma^2}{\sum_i x_i^2}\right)$$

as obtained before.

*Variance estimation*

Use *residual sum of squares RSS, SSE*

$$Q^2 = (\underline{Y} - \mathbf{X}\underline{\hat{\theta}})^T(\underline{Y} - \mathbf{X}\underline{\hat{\theta}})$$

The vector $\underline{Y} - \mathbf{X}\underline{\hat{\theta}}$ is also called *vector of residuals*

Chapter 2 Theorem 7:

$$\frac{Q^2}{\sigma^2} \sim \chi^2_{n-p},$$

independent of $\underline{\hat{\theta}}$

so the estimator

$$s^2 = \hat{\sigma^2} = \frac{Q^2}{n-p}$$

is unbiased

*Tests:* To test hypotheses about single parameters $\theta_i$, use from Chapter 2 that

$$\frac{\widehat{\theta}_i - \theta_i}{\widehat{\sigma}\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{ii}}} \sim t_{n-p}$$

*Example:* see lecture.

*Prediction*

Task: For $p$-vector $\underline{a} = (1, a_1, \ldots, a_{p-1})$ predict response $y$, and give prediction interval

Case 1: $\underline{a}$ is among $\underline{x}_i, i = 1, \ldots, n$: have information about $\epsilon$

predict $\tilde{Y} = \underline{a\hat{\theta}}$

then

$$E\tilde{Y} = \underline{a\theta}$$

and

$$V(\tilde{Y}) = \underline{a}V(\underline{\hat{\theta}})\underline{a}^T = \sigma^2 \underline{a}(\mathbf{X}^T\mathbf{X})^{-1}\underline{a}^T$$

so

$$\tilde{Y} \sim \mathcal{N}(\underline{a\theta}, \sigma^2 \underline{a}(\mathbf{X}^T\mathbf{X})^{-1}\underline{a}^T)$$

Estimate $\sigma^2$ so use $t_{n-p}$-distribution

two-sided $100(1-\alpha)\%$ confidence interval for $\tilde{y}$ is

$$\underline{a\theta} \pm t_{n-p}(\alpha/2)\hat{\sigma}\left(\underline{a}(\mathbf{X}^T\mathbf{X})^{-1}\underline{a}^T\right)^{\frac{1}{2}}$$

Case 2: $\underline{a}$ is a new vector yet unobserved

$Y_0 = \underline{a\theta} + \epsilon_0 \sim \mathcal{N}(\underline{a\theta}, \sigma^2)$

$Y_0 - \underline{a}\hat{\theta}$ is again normal, mean zero, variance

$$\begin{aligned} V(Y_0 - \underline{a}\hat{\theta}) &= V(Y_0) + 2Cov(\epsilon_0, \underline{a\hat{\theta}}) \\ &\quad +\sigma^2\underline{a}(\mathbf{X}^T\mathbf{X})^{-1}\underline{a}^T \\ &= \sigma^2(1 + \underline{a}(\mathbf{X}^T\mathbf{X})^{-1}\underline{a}^T) \end{aligned}$$

Using the independence of $Y_0, \hat{\underline{\theta}}$ and $\hat{\sigma}$, we have

$$\frac{Y_0 - \underline{a}\hat{\theta}}{\hat{\sigma}\sqrt{(1 + \underline{a}(\mathbf{X}^T\mathbf{X})^{-1}\underline{a}^T)}} \sim t_{n-p}$$

*Example: Simple linear regression*

Model

$$Y_i = \alpha + \beta(x_i - \bar{x}) + \epsilon_i$$

$\underline{a} = (1, x_0)$ then

$$\underline{a}(\mathbf{X}^T\mathbf{X})^{-1}\underline{a}^T = \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

so

$$\frac{Y_0 - (\hat{\alpha} + \hat{\beta}(x_0 - \bar{x})}{\hat{\sigma}\sqrt{(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \sim t_{n-2}$$

*Example* see lecture.

*Checking the model assumptions*

Residuals $\widehat{\epsilon}_i = y_i - \underline{x_i}\underline{\widehat{\theta}}$

1. Constant variance? (*Homoscedasticity?*)
Residual plot: $\widehat{\epsilon}_i$'s against $y_i$'s: if too heterogeneous, transform the data

2. Linear trend? Again: residual plot

3. Normality? Normal Q-Q plot; if not normal, then use a transformation (typically a power transformation)

4. Independence? Plot $\widehat{\epsilon}_{i+1}$ against $\widehat{\epsilon}_i$: correlation? If so: time series analysis