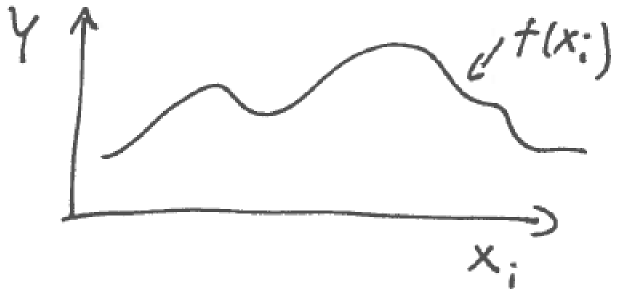
Part III

Regression

Chapter 14

Overview over regression modelling

14.1 General setup



- y: **response variable**, also known as **outcome** or **label**
- $x_1, x_2, x_3, \dots, x_d$: **predictor variables**, also known as **covariates** or **covariables**
- The relationship between the outcomes and the predictor variables is assumed to follow

  $$y = f(x_1, x_2, \dots, x_d) + \varepsilon$$

  where f is the **regression function** (not a density) and ε represents **noise**.

14.2 Objectives

1. **Understand the relationship** between the response y and the predictor variables x_i by **learning the regression function** f from observed data (training data). The estimated regression function is $\hat{f}$.

2. **Prediction of outcomes**

$$\underbrace{\hat{y}}_{\substack{\text{predicted response}\\ \text{using fitted } \hat{f}}} \quad = \hat{f}(x_1, x_2, \dots, x_d)$$

If instead of the fitted function $\hat{f}$ the known regression function f is used we denote this by

$$\underbrace{y^\star}_{\substack{\text{predicted response}\\ \text{using known } f}} \quad = f(x_1, x_2, \dots, x_d)$$

3. **Variable importance**

   - which covariates are most relevant in predicting the outcome?
   - allows to better understand the data and model
     → variable selection (to build simpler model with same predictive capability)

14.3 Regression as a form of supervised learning

Regression modelling is a special case of **supervised learning**.

In supervised learning we make use of labelled data, i.e. $\boldsymbol{x}_i$ has an associated *label* y_i. Thus, the data is consists of pairs $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \dots, (\boldsymbol{x}_n, y_n)$.

The *supervision* part of in supervised learning refers to the fact that the labels are given.

In **regression** typically the label y_i is continuous and called the *response*.

On the other hand, if the label y_i is discrete/categorical then supervised learning is called **classification**.

$$\text{Supervised Learning} \begin{array}{ll} \longrightarrow \text{Discrete } y & \longrightarrow \text{Classification Methods} \\ \\ \longrightarrow \text{Continuous } y & \longrightarrow \text{Regression Methods} \end{array}$$

Another important type of statistical learning is **unsupervised learning** where labels y are inferred from the data $\boldsymbol{x}$ (this is also known as **clustering**). Furthermore, there is also *semi-supervised learning* with labels only partly known.

Note that there are regression models (e.g. logistic regression) with discrete response that are performing classification, so one may argue that "supervised learning"="generalised regression".

14.4 Various regression models used in statistics

In this course we only study linear multiple regression. However, you should be aware that the linear model is in fact just a special cases of some much more general regression approaches.

General regression model:

$$y = f(x_1, \dots, x_d) + \text{"noise"}$$

- The function f is estimated nonparametrically - splines - Gaussian processes
- Generalised Additive Models (GAM): - the function f is assumed to be the sum of individual functions $f_i(x_i)$
- Generalised Linear Models (GLM): - f is a transformed linear predictor $h(\sum b_i x_i)$, noise is assumed from an exponential family
- Linear Model (LM): - linear predictor $\sum b_i x_i$, normal noise

In R the linear model is implemented in the function lm(), and generalised linear models in the function glm(). Generalised additive models are available in the package "mgcv".

In the following we focus on the linear regression model with continuous response.

Chapter 15

Linear Regression

15.1 The linear regression model

In this module we assume that f is a linear function:

$$f(x_1, \dots, x_d) = \beta_0 + \sum_{j=1}^{d} \beta_j x_j = y^\star$$

In vector notation:

$$f(\boldsymbol{x}) = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x} = y^\star$$

with $\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_d \end{pmatrix}$ and $\boldsymbol{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}$

Therefore, the linear regression model is

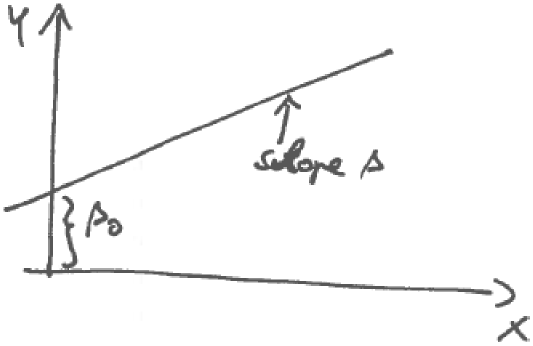
$$\begin{aligned} y &= \beta_0 + \sum_{j=1}^{d} \beta_j x_j + \varepsilon \\ &= \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x} + \varepsilon \\ &= y^\star + \varepsilon \end{aligned}$$

where:

- β_0 is the **intercept**
- $\boldsymbol{\beta} = (\beta_1, \dots, \beta_d)^T$ are the **regression coefficients**
- $\boldsymbol{x} = (x_1, \dots, x_d)^T$ is the predictor vector containing the **predictor variables**

15.2 Interpretation of regression coefficients and intercept

- The regression coefficient β_i corresponds to the slope (first partial derivative) of the regression function in the direction of x_i. In other words, the gradient of $f(\boldsymbol{x})$ are the regression coefficients: $\nabla f(\boldsymbol{x}) = \boldsymbol{\beta}$
- The intercept β_0 is the offset at the origin ($x_1 = x_2 = \ldots = x_d = 0$):



15.3 Different types of linear regression:

- **Simple linear regression**: $y = \beta_0 + \beta x + \varepsilon$ (=single predictor)
- **Multiple linear regression**: $y = \beta_0 + \sum_{j=1}^{d} \beta_j x_j + \varepsilon$ (= multiple predictor variables)
- **Multivariate regression**: multivariate response $\boldsymbol{y}$

15.4 Distributional assumptions and properties

General assumptions:

- We treat y and $x_1, \ldots, x_d$ as the primary observables that can be described by random variables.

- $\beta_0, \boldsymbol{\beta}$ are parameters to be inferred from the observations on y and $x_1, \ldots, x_d$.

- Specifically, will we assume that response and predictors have a mean and a (cov)variance:

  i. Response:
     $\text{E}(y) = \mu_y$
     $\text{Var}(y) = \sigma^2_y$
     The **variance of the response** $\text{Var}(y)$ is also called the **total variation** .

ii. Predictors:
$\text{E}(x_i) = \mu_{x_i}$ (or $\text{E}(\boldsymbol{x}) = \boldsymbol{\mu}_{\boldsymbol{x}}$)
$\text{Var}(x_i) = \sigma^2_{x_i}$ and $\text{Cor}(x_i, x_j) = \rho_{ij}$ (or $\text{Var}(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$)
The **signal variance** $\text{Var}(y^\star) = \text{Var}(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}) = \boldsymbol{\beta}^T \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\beta}$ is also called the **explained variation**.

- We assume that y and $\boldsymbol{x}$ are jointly distributed with correlation $\text{Cor}(y, x_j) = \rho_{y,x_j}$ between each predictor variable x_j and the response y.
- In contrast to y and $\boldsymbol{x}$ the noise variable ε is only indirectly observed via the difference $\varepsilon = y - y^\star$. We denote the mean and variance of the noise by $\text{E}(\varepsilon)$ and $\text{Var}(\varepsilon)$.
  The **noise variance** $\text{Var}(\varepsilon)$ is also called the **unexplained variation** or the **residual variance**. The **residual standard error** is $\text{SD}(\varepsilon)$.

Identifiability assumptions:

In a statistical analysis we would like to be able to separate signal ($y^\star$) from noise (ε). To achieve this we require some **distributional assumptions to ensure identifiability** and avoid confounding:

1) **Assumption 1:** ε and $y^\star$ are are independent. This implies $\text{Var}(y) = \text{Var}(y^\star) + \text{Var}(\varepsilon)$, or equivalently $\text{Var}(\varepsilon) = \text{Var}(y) - \text{Var}(y^\star)$.

   Thus, this assumption implies the **decomposition of variance**, i.e. that the **total variation** $\text{Var}(y)$ equals the sum of the **explained variation**$\text{Var}(y^\star)$ and the **unexplained variation**$\text{Var}(\varepsilon)$.

2) **Assumption 2:** $\text{E}(\varepsilon) = 0$. This allows to identify the intercept β_0 and implies $\text{E}(y) = \text{E}(y^\star)$.

Optional assumptions (often but not always):

- The noise ε is normally distributed
- The response y and and the predictor variables x_i are continuous variables
- The response and predictor variables are jointly normally distributed

Further properties:

- As a result of the independence assumption 1) we can only choose two out of the three variances freely:
  i. in a generative perspective we will choose signal variance $\text{Var}(y^\star)$ (or equivalently the variances $\text{Var}(x_j)$) and the noise variance $\text{Var}(\varepsilon)$, then the variance of the response $\text{Var}(y)$ follows.
  ii. in an observational perspective we will observe the variance of the reponse $\text{Var}(y)$ and the variances $\text{Var}(x_j)$, and then the error variance $\text{Var}(\varepsilon)$ follows.
- As we will see later the regression coefficients β_j depend on the correlations between the response y and and the predictor variables x_j. Thus, the choice of regression coefficients implies a specific correlation pattern, and vice

versa (in fact, we will use this correlation pattern to infer the regression coefficients from data!).

15.5 Regression in data matrix notation

We can also write the regression in terms of actual observed data (rather than in terms of random variables):

Data matrix for the predictors:

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & \dots & x_{1d} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nd} \end{pmatrix}$$

Note the statistics convention: the n rows of $\boldsymbol{X}$ contain the samples, and the d columns contain variables.

Response data vector: $(y_1, \dots, y_n)^T = \boldsymbol{y}$

Then the regression equation is written in data matrix notation:

$$\underbrace{\boldsymbol{y}}_{n\times 1} = \underbrace{\boldsymbol{1}_n \beta_0}_{n\times 1} + \underbrace{\boldsymbol{X}}_{n\times d} \underbrace{\boldsymbol{\beta}}_{d\times 1} + \underbrace{\boldsymbol{\varepsilon}}_{\underbrace{n\times 1}_{\text{residuals}}}$$

where $\boldsymbol{1}_n = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ is a column vector of length n (size $n \times 1$).

Note that here the regression coefficients are now multiplied *after* the data matrix (compare with the original vector notation where the *transpose* of regression coefficients come *before* the vector of the predictors).

The **observed noise** values (i.e. realisations of the random variable ε) are called the **residuals**.

15.6 Centering and vanishing of the intercept β_0

If $\boldsymbol{x}$ and y are centered, i.e. if $\text{E}(\boldsymbol{x}) = \boldsymbol{\mu}_{\boldsymbol{x}} = 0$ and $\text{E}(y) = \mu_y = 0$, then the intercept β_0 disappears:

The regression equation is

$$y = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x} + \varepsilon$$

with $E(\varepsilon)$. Taking the expectation on both sides we get $\mu_y = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{\mu}_{\boldsymbol{x}}$ and therefore

$$\beta_0 = \mu_y - \boldsymbol{\beta}^T \boldsymbol{\mu}_{\boldsymbol{x}}$$

This is zero if the mean of the response μ_y and the mean of predictors $\boldsymbol{\mu}_{\boldsymbol{x}}$ vanish. Conversely, if we assume that the intercept vanishes ($\beta_0 = 0$) this is only possible for general $\boldsymbol{\beta}$ if both $\boldsymbol{\mu}_{\boldsymbol{x}} = 0$ and $\mu_y = 0$.

Thus, in the linear model is always possible to transform y and $\boldsymbol{x}$ (or data $\boldsymbol{y}$ and $\boldsymbol{X}$) so that the intercept vanishes. To simplify equations we will therefore often set $\beta_0 = 0$.

15.7 Objectives in data analysis using linear regression

1. Understand functional relationship: find estimates of the intercept ($\hat{\beta}_0$) and the regression coefficients ($\hat{\beta}_j$), as well as the associated errors.

2. Prediction:

   - Known coefficients β_0 and $\boldsymbol{\beta}$: $y^\star = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}$
   - Estimated coefficients $\hat{\beta}_0$ and $\hat{\beta}$ (note the "hat"!): $\hat{y} = \hat{\beta}_0 + \sum_{j=1}^d \hat{\beta}_j x_j = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$

   For each point prediction find the **corresponding prediction error!**

3. Variable importance: Which predictors x_j are most relevant?
   → test whether $\beta_j = 0$
   → find measures of variable importance

   Remark: as we will see β_j or $\hat{\beta}_j$ itself is **not** a measure of variable importance!

Chapter 16

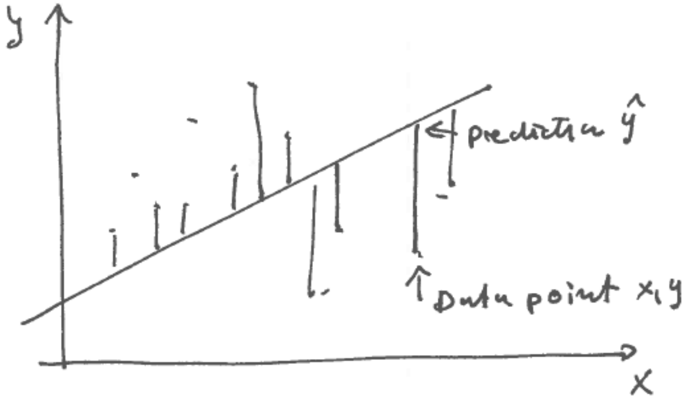
Estimating regression coefficients

In this chapter we discuss various ways to estimate the regression coefficients. First, we discuss estimation by Ordinary Least Squares (OLS) by minimising the residual sum of squares. This yields the famous Gauss estimator. Second, we derive estimates of the regression coefficients using the methods of maximum likelihood assuming normal errors. This also leads to the Gauss estimator. Third, we show that the coefficients in linear regression can written and interpreted in terms of two covariance matrices and that the Gauss estimator of the regression coefficients is a plug-in estimator using the MLEs of these covariance matrices. Furthermore, we show that the (population version) of the Gauss estimator can also be derived by finding the best linear predictor and by conditioning. Finally, we discuss special cases of regression coefficients and their relationship to marginal correlation.

16.1 Ordinary Least Squares (OLS) estimator of regression coefficients

Now we show the classic way (Gauss 1809; Legendre 1805) to **estimate regression coefficients** by the method of **ordinary least squares (OLS)**.

Idea: choose regression coefficients such as to *minimise* the *squared error* between observations and the prediction.



In data matrix notation (note we assume $\beta_0 = 0$ and thus *centered data* $\boldsymbol{X}$ and $\boldsymbol{y}$):

$$\text{RSS}(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

RSS is an abbreviation for "Residual Sum of Squares" which is is a function of $\boldsymbol{\beta}$. Minimising RSS yields the OLS estimate:

$$\widehat{\boldsymbol{\beta}}_{\text{OLS}} = \underset{\boldsymbol{\beta}}{\arg\min}\, \text{RSS}(\boldsymbol{\beta})$$

$$\text{RSS}(\boldsymbol{\beta}) = \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}$$

Gradient:

$$\nabla\text{RSS}(\boldsymbol{\beta}) = -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}$$

$$\nabla\text{RSS}(\widehat{\boldsymbol{\beta}}) = 0 \longrightarrow \boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}$$

$$\Longrightarrow \widehat{\boldsymbol{\beta}}_{\text{OLS}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Note the similarities in the procedure to maximum likelihood (ML) estimation (with minimisation instead of maximisation)! In fact, as we see next this is not by chance as OLS *is* indeed a special case of ML! This also implies that OLS is generally a good method — but only if sample size n is large!

The above Gauss' estimator is fundamental in statistics so it is worthwile to memorise it!

16.2 Maximum likelihood estimation of regression coefficients

16.2.1 Normal log-likelihood function for regression coefficients and noise variance

We now show how to estimate regression coefficients using the method of maximum likelihood. This is a second method to derive $\hat{\boldsymbol{\beta}}$.

We recall the basic regression equation

$$y = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x} + \varepsilon$$

with independent noise ε and observed data $y_1, \ldots, y_n$ and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$.

Assuming $\text{E}(\varepsilon) = 0$ the intercept is identified as

$$\beta_0 = \mu_y - \boldsymbol{\beta}^T \boldsymbol{\mu}_{\boldsymbol{x}}$$

Combining the two above equations we see that noise variable equals

$$\varepsilon = (y - \mu_y) - \boldsymbol{\beta}^T (\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{x}})$$

Assuming joint (multivariate) normality for the observed data, the response y and predictors $\boldsymbol{x}$, we get as the MLEs for the respective means and (co)variances:

- $\hat{\mu}_y = \hat{\text{E}}(y) = \frac{1}{n} \sum_{i=1}^n y_i$
- $\hat{\sigma}^2_y = \widehat{\text{Var}}(y) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\mu}_y)^2$
- $\hat{\boldsymbol{\mu}}_{\boldsymbol{x}} = \hat{\text{E}}(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^n \boldsymbol{x}_i$
- $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}\boldsymbol{x}} = \widehat{\text{Var}}(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^n (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{x}})(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{x}})^T$
- $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}y} = \widehat{\text{Cov}}(\boldsymbol{x}, y) = \frac{1}{n} \sum_{i=1}^n (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{x}})(y_i - \hat{\mu}_y)$

Note that these are are sufficient statistics and hence summarize perfectly the observed data for $\boldsymbol{x}$ and y under the normal assumption

Consequently, the residuals (indirect observations of the noise variable) for a given choice of regression coefficients $\boldsymbol{\beta}$ and the observed data for $\boldsymbol{x}$ and y are

$$\varepsilon_i = (y_i - \hat{\mu}_y) - \boldsymbol{\beta}^T (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{x}})$$

Assuming that the noise $\varepsilon \sim N(0, \sigma^2_\varepsilon)$ is normally distributed with mean 0 and variance $\text{Var}(\varepsilon) = \sigma^2_\varepsilon$. we can write down the normal log-likelihood function for σ^2_ε and $\boldsymbol{\beta}$:

$$\log L(\boldsymbol{\beta}, \sigma^2_\varepsilon) = -\frac{n}{2} \log \sigma^2_\varepsilon - \frac{1}{2\sigma^2_\varepsilon} \sum_{i=1}^n \left((y_i - \hat{\mu}_y) - \boldsymbol{\beta}^T (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{x}}) \right)^2$$

Maximising this function leads to the MLEs of σ^2_ε and $\boldsymbol{\beta}$!

Note that the residual sum of squares appears in the log-likelihood function (with a minus sign), which implies that ML assuming normal distribution will recover the OLS estimator for the regression coefficients! So OLS is a special case of ML !

16.2.2 Detailed derivation of the MLEs

The gradient with regard to $\boldsymbol{\beta}$ is

$$\begin{aligned}\nabla_{\boldsymbol{\beta}} \log L(\boldsymbol{\beta}, \sigma^2_\varepsilon) &= \frac{1}{\sigma^2_\varepsilon} \sum_{i=1}^{n} \left((\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{x}})(y_i - \hat{\mu}_y) - (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{x}})(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{x}})^T \boldsymbol{\beta} \right) \\ &= \frac{n}{\sigma^2_\varepsilon} \left(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}y} - \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}\boldsymbol{x}} \boldsymbol{\beta} \right)\end{aligned}$$

Setting this equal to zero yields the Gauss estimator

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{x}\boldsymbol{x}} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}y}$$

By plugin we the get the MLE of β_0 as

$$\hat{\beta}_0 = \hat{\mu}_y - \hat{\boldsymbol{\beta}}^T \hat{\boldsymbol{\mu}}_{\boldsymbol{x}}$$

Taking the derivative of $\log L(\hat{\boldsymbol{\beta}}, \sigma^2_\varepsilon)$ with regard to σ^2_ε yields

$$\frac{\partial}{\partial \sigma^2_\varepsilon} \log L(\hat{\boldsymbol{\beta}}, \sigma^2_\varepsilon) = -\frac{n}{2\sigma^2_\varepsilon} + \frac{1}{2\sigma^4_\varepsilon} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

with $\hat{y}_i = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ and the residuals $y_i - \hat{y}_i$ resulting from the fitted linear model. This leads to the MLE of the noise variance

$$\widehat{\sigma^2_\varepsilon} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Note that the MLE $\widehat{\sigma^2_\varepsilon}$ is a biased estimate of σ^2_ε. The unbiased estimate is $\frac{1}{n-d-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$, where d is the dimension of $\boldsymbol{\beta}$ (i.e. the number of predictors).

16.2.3 Asymptotics

The advantage of using maximum likelihood is that we also get the (asympotic) variance associated with each estimator and typically can also assume asymptotic normality.

Specifically, for $\hat{\boldsymbol{\beta}}$ we get via the observed Fisher information at the MLE an asymptotic estimator of its variance

$$\widehat{\text{Var}}(\widehat{\boldsymbol{\beta}}) = \frac{1}{n}\widehat{\sigma^2_\varepsilon}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}\boldsymbol{x}}^{-1}$$

Similarly, for $\hat{\beta}_0$ we have

$$\widehat{\text{Var}}(\widehat{\beta}_0) = \frac{1}{n}\widehat{\sigma^2_\varepsilon}(1 + \hat{\boldsymbol{\mu}}^T\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}\boldsymbol{x}}^{-1}\hat{\boldsymbol{\mu}})$$

For finite sample size n with known $\text{Var}(\varepsilon)$ one can show that the variances are

$$\text{Var}(\widehat{\boldsymbol{\beta}}) = \frac{1}{n}\sigma^2_\varepsilon\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}\boldsymbol{x}}^{-1}$$

and

$$\text{Var}(\widehat{\beta}_0) = \frac{1}{n}\sigma^2_\varepsilon(1 + \hat{\boldsymbol{\mu}}_{\boldsymbol{x}}^T\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}\boldsymbol{x}}^{-1}\hat{\boldsymbol{\mu}}_{\boldsymbol{x}})$$

and that the regression coefficients and the intercept are normally distributed according to

$$\widehat{\boldsymbol{\beta}} \sim N_d(\boldsymbol{\beta}, \text{Var}(\widehat{\boldsymbol{\beta}}))$$

and

$$\widehat{\beta}_0 \sim N(\beta_0, \text{Var}(\widehat{\beta}_0))$$

We may use this to test whether whether $\beta_j = 0$ and $\beta_0 = 0$.

16.3 Covariance plug-in estimator of regression coefficients

16.3.1 Regression coeffients as product of variances

We now try to understand regression coefficients in terms of covariances (thus obtaining a third way to compute and estimate them).

We recall that the Gauss regression coefficients are given by

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

where $\boldsymbol{X}$ is the $n \times d$ data matrix (in statistics convention)

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & \dots & x_{1d} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nd} \end{pmatrix}$$

Note that we assume that the data matrix $\boldsymbol{X}$ is centered (i.e. column sums $\boldsymbol{X}^T \boldsymbol{1}_n = \boldsymbol{0}$ are zero).

Likewise $\boldsymbol{y} = (y_1, \dots, y_n)^T$ is the response data vector (also centered with $\boldsymbol{y}^T \boldsymbol{1}_n = 0$).

Noting that

$$\hat{\boldsymbol{\Sigma}}_{xx} = \frac{1}{n}(\boldsymbol{X}^T \boldsymbol{X})$$

is the MLE of covariance matrix among $\boldsymbol{x}$ and

$$\hat{\boldsymbol{\Sigma}}_{xy} = \frac{1}{n}(\boldsymbol{X}^T \boldsymbol{y})$$

is the MLE of the covariance between $\boldsymbol{x}$ and y we see that the OLS estimate of the regression coefficients can be expressed as

$$\widehat{\boldsymbol{\beta}} = \left(\hat{\boldsymbol{\Sigma}}_{xx}\right)^{-1} \hat{\boldsymbol{\Sigma}}_{xy}$$

We can write down a population version (with no hats!):

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}_{xx}^{-1} \boldsymbol{\Sigma}_{xy}$$

Thus, OLS regression coefficients can be interpreted as plugin estimator using MLEs of covariances! In fact, we may also use the unbiased estimates since the scale factor ($1/n$ or $1/(n-1)$) cancels out so it does not matter which one you use!

16.3.2 Importance of positive definiteness of estimated covariance matrix

Note that $\hat{\boldsymbol{\Sigma}}_{xx}$ is inverted in $\widehat{\boldsymbol{\beta}} = \left(\hat{\boldsymbol{\Sigma}}_{xx}\right)^{-1} \hat{\boldsymbol{\Sigma}}_{xy}$.

- Hence, the estimate $\hat{\boldsymbol{\Sigma}}_{xx}$ needs to be positive definite!
- But $\hat{\boldsymbol{\Sigma}}_{xx}^{\text{MLE}}$ is only positive definite if $n > d$!

Therefore we can use the ML estimate (empirical estimator) only for large $n > d$, otherwise we need to employ a different (regularised) estimation approach (e.g. Bayes or a penalised ML)!

Remark: writing $\hat{\boldsymbol{\beta}}$ explicitly based on covariance estimates has the advantage that we can construct plug-in estimators of regression coefficients based on regularised covariance estimators that improve over ML for small sample size. This leads to the so-called SCOUT method (=covariance-regularized regression by Witten and Tibshirani, 2008).

16.4 Standardised regression coefficients and their relationship to correlation

We recall the relationship between regression coefficients $\boldsymbol{\beta}$ and the marginal covariance $\boldsymbol{\Sigma}_{xy}$ and the covariances among the predictors $\boldsymbol{\Sigma}_{xx}$:

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$$

We can rewrite the regression coefficients in terms of marginal correlations $\boldsymbol{P}_{xy}$ and correlations $\boldsymbol{P}_{xx}$ among the predictors using the variance-correlation decompositions $\boldsymbol{\Sigma}_{xx} = \boldsymbol{V}_x^{1/2}\boldsymbol{P}_{xx}\boldsymbol{V}_x^{1/2}$ and $\boldsymbol{\Sigma}_{xy} = \boldsymbol{V}_x^{1/2}\boldsymbol{P}_{xy}\sigma_y$:

$$\begin{aligned}\boldsymbol{\beta} &= \underbrace{\boldsymbol{V}_x^{-1/2}}_{\text{(inverse) scale of } x_i} \quad \boldsymbol{P}_{xx}^{-1}\boldsymbol{P}_{xy} \quad \underbrace{\sigma_y}_{\text{scale of } y} \\ &= \boldsymbol{V}_x^{-1/2}\; \boldsymbol{\beta}_{\text{std}}\; \sigma_y\end{aligned}$$

Thus the regression coefficients $\boldsymbol{\beta}$ contain the scale of the variables, and take into account the correlations among the predictors ($\boldsymbol{P}_{xx}$) in addition to the marginal correlations between the response y and the predictors x_i ($\boldsymbol{P}_{xy}$).

This decomposition allows to understand a number special cases for which the regression coefficients simplify further:

a) If the response and the predictors are standardised to have variance one, i.e. $\text{Var}(y) = 1$ and $\text{Var}(x_i) = 1$, then $\boldsymbol{\beta}$ becomes equal to the **standardised regression coefficients**

$$\boldsymbol{\beta}_{\text{std}} = \boldsymbol{P}_{xx}^{-1}\boldsymbol{P}_{xy}$$

Note that standardised regression coefficients do not make use of variances and and thus are scale-independent.

b) If there is no correlation among the predictors , i.e. $\boldsymbol{P}_{xx} = \boldsymbol{I}$ the the regression coefficients reduce to

$$\boldsymbol{\beta} = \boldsymbol{V}_x^{-1}\boldsymbol{\Sigma}_{xy}$$

where $\boldsymbol{V}_x$ is a diagonal matrix containing the variances of the predictors. This is also called **marginal regression**. Note that the inversion of $\boldsymbol{V}_x$ is trival since you only need to invert each diagonal element individually.

c) If both a) and b) apply simultaneously (i.e. there is no correlation among predictors and response and predictors and predictors are standardised) then the regression coefficients simplify even further to

$$\boldsymbol{\beta} = \boldsymbol{P}_{xy}$$

Thus, in this very special case the regression coefficients are identical to the correlations between the response and the predictors!

16.5 Further ways to obtain regression coefficients

16.5.1 Best linear predictor

The **best linear predictor** is a fourth way to arrive at the linear model. This is closely related to OLS and minimising squared residual error.

Without assuming normality the above multiple regression model can be shown to be optimal linear predictor under the minimum mean squared prediction error:

Assumptions:

- y and $\boldsymbol{x}$ are random variables
- we construct a new variable (the linear predictor) $y^{\star\star} = b_0 + \boldsymbol{b}^T \boldsymbol{x}$ to optimally approximate y

Aim:

- choose b_0 and $\boldsymbol{b}$ such to minimize the mean squared prediction error $\text{E}((y - y^{\star\star})^2)$

16.5.1.1 Result:

The mean squared prediction error $MSPE$ in dependence of $(b_0, \boldsymbol{b})$ is

$$\begin{aligned}
\text{E}((y - y^{\star\star})^2) &= \text{Var}(y - y^{\star\star}) + \text{E}(y - y^{\star\star})^2 \\
&= \text{Var}(y - b_0 - \boldsymbol{b}^T\boldsymbol{x}) + (\text{E}(y) - b_0 - \boldsymbol{b}^T\text{E}(\boldsymbol{x}))^2 \\
&= \sigma_y^2 + \text{Var}(\boldsymbol{b}^T\boldsymbol{x}) + 2\,\text{Cov}(y, -\boldsymbol{b}^T\boldsymbol{x}) + (\mu_y - b_0 - \boldsymbol{b}^T\boldsymbol{\mu}_{\boldsymbol{x}})^2 \\
&= \sigma_y^2 + \boldsymbol{b}^T\boldsymbol{\Sigma}_{\boldsymbol{x}\boldsymbol{x}}\boldsymbol{b} - 2\,\boldsymbol{b}^T\boldsymbol{\Sigma}_{\boldsymbol{x}y} + (\mu_y - b_0 - \boldsymbol{b}^T\boldsymbol{\mu}_{\boldsymbol{x}})^2 \\
&= MSPE(b_0, \boldsymbol{b})
\end{aligned}$$

We look for

$$(\beta_0, \boldsymbol{\beta}) = \underset{b_0, \boldsymbol{b}}{\arg\min}\; MSPE(b_0, \boldsymbol{b})$$

In order to find the minimum we compute the gradient with regard to $(b_0, \boldsymbol{b})$

$$\nabla MSPE = \begin{pmatrix} -2(\mu_y - b_0 - \boldsymbol{b}^T\boldsymbol{\mu}_{\boldsymbol{x}}) \\ 2\,\boldsymbol{\Sigma}_{\boldsymbol{x}\boldsymbol{x}}\boldsymbol{b} - 2\,\boldsymbol{\Sigma}_{\boldsymbol{x}y} - 2\boldsymbol{\mu}_{\boldsymbol{x}}(\mu_y - b_0 - \boldsymbol{b}^T\boldsymbol{\mu}_{\boldsymbol{x}}) \end{pmatrix}$$

and setting this equal to zero yields

$$\begin{pmatrix} \beta_0 \\ \boldsymbol{\beta} \end{pmatrix} = \begin{pmatrix} \mu_y - \boldsymbol{\beta}^T\boldsymbol{\mu}_{\boldsymbol{x}} \\ \boldsymbol{\Sigma}_{\boldsymbol{x}\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}y} \end{pmatrix}$$

Thus, the optimal values for b_0 and $\boldsymbol{b}$ in the best linear predictor correspond to the previously derived coefficients β_0 and $\boldsymbol{\beta}$!

16.5.1.2 Irreducible Error

The minimum achieved MSPE (=**irreducible error**) is

$$MSPE(\beta_0, \boldsymbol{\beta}) = \sigma_y^2 - \boldsymbol{\beta}^T \boldsymbol{\Sigma}_{\boldsymbol{x}\boldsymbol{x}} \boldsymbol{\beta} = \sigma_y^2 - \boldsymbol{\Sigma}_{y\boldsymbol{x}} \boldsymbol{\Sigma}_{\boldsymbol{x}\boldsymbol{x}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{x}y}$$

With the abbreviation $\Omega^2 = \boldsymbol{P}_{y\boldsymbol{x}} \boldsymbol{P}_{\boldsymbol{x}\boldsymbol{x}}^{-1} \boldsymbol{P}_{\boldsymbol{x}y} = \sigma_y^{-2} \boldsymbol{\Sigma}_{y\boldsymbol{x}} \boldsymbol{\Sigma}_{\boldsymbol{x}\boldsymbol{x}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{x}y}$ we can simplify this to

$$MSPE(\beta_0, \boldsymbol{\beta}) = \sigma_y^2 (1 - \Omega^2) = \text{Var}(\varepsilon)$$

Writing $b_0 = \beta_0 + \Delta_0$ and $\boldsymbol{b} = \boldsymbol{\beta} + \boldsymbol{\Delta}$ it is easy to see that the mean squared predictive error is a quadratic function around the minimum:

$$MSPE(\beta_0 + \Delta_0, \boldsymbol{\beta} + \boldsymbol{\Delta}) = \text{Var}(\varepsilon) + \Delta_0^2 + \boldsymbol{\Delta}^T \boldsymbol{\Sigma}_{\boldsymbol{x}\boldsymbol{x}} \boldsymbol{\Delta}$$

Note that usually $y^\star = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}$ does not perfectly approximate y so there *will* be an irreducible error (= noise variance)

$$\text{Var}(\varepsilon) = \sigma_y^2 (1 - \Omega^2) > 0$$

which implies $\Omega^2 < 1$.

The quantity Ω^2 has a further interpretation of the population version of as the squared multiple correlation coefficient between the response and the predictors and plays a vital role in decomposition of variance, as discussed later.

16.5.2 Regression by conditioning

Conditioning is a fifth way to arrive at the linear model. This is also the most general way and can be used to derive many other regression models (not just the simple linear model).

16.5.2.1 General idea:

- two random variables y (response, scalar) and $\boldsymbol{x}$ (predictor variables, vector)
- we assume that y and $\boldsymbol{x}$ have a joint distribution $F_{y,\boldsymbol{x}}$
- compute *conditional* random variable $y|\boldsymbol{x}$ and the corresponding distribution $F_{y|\boldsymbol{x}}$

16.5.2.2 Multivariate normal assumption

Now we assume that y and $\boldsymbol{x}$ are (jointly) multivariate normal. Then the conditional distribution $F_{y|\boldsymbol{x}}$ is a univariate normal with the following moments (you can verify this by looking up the general conditional multivariate normal distribution):

a) Conditional expectation:

$$\text{E}(y|\boldsymbol{x}) = y^\star = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}$$

with coefficients $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{\boldsymbol{x}\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}y}$ and intercept $\beta_0 = \mu_y - \boldsymbol{\beta}^T \boldsymbol{\mu}_{\boldsymbol{x}}$.

Note that as $y^\star$ depends on $\boldsymbol{x}$ it is a random variable itself with mean

$$\text{E}(y^\star) = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{\mu}_{\boldsymbol{x}} = \mu_y$$

and variance

$$\begin{aligned}\text{Var}(y^\star) &= \text{Var}(\text{E}(y|\boldsymbol{x}))\\ &= \boldsymbol{\beta}^T \boldsymbol{\Sigma}_{\boldsymbol{x}\boldsymbol{x}} \boldsymbol{\beta} = \boldsymbol{\Sigma}_{y\boldsymbol{x}} \boldsymbol{\Sigma}_{\boldsymbol{x}\boldsymbol{x}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{x}y}\\ &= \sigma_y^2 \boldsymbol{P}_{y\boldsymbol{x}} \boldsymbol{P}_{\boldsymbol{x}\boldsymbol{x}}^{-1} \boldsymbol{P}_{\boldsymbol{x}y}\\ &= \sigma_y^2 \Omega^2\end{aligned}$$

b) Conditional variance:

$$\begin{aligned}\text{Var}(y|\boldsymbol{x}) &= \sigma_y^2 - \boldsymbol{\beta}^T \boldsymbol{\Sigma}_{\boldsymbol{x}\boldsymbol{x}} \boldsymbol{\beta}\\ &= \sigma_y^2 - \boldsymbol{\Sigma}_{y\boldsymbol{x}} \boldsymbol{\Sigma}_{\boldsymbol{x}\boldsymbol{x}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{x}y}\\ &= \sigma_y^2 (1 - \Omega^2)\end{aligned}$$

Note this is a constant so $\text{E}(\text{Var}(y|\boldsymbol{x})) = \sigma_y^2(1 - \Omega^2)$ as well.

Chapter 17

Squared multiple correlation and variance decomposition in linear regression

In this chapter we first introduce the (squared) multiple correlation and the multiple and adjusted R^2 coefficients as estimators. Subsequently we discuss variance decomposition.

17.1 Squared multiple correlation Ω^2 and the R^2 coefficient

In the previous chapter we encountered the following quantity:

$$\Omega^2 = \boldsymbol{P}_{yx}\boldsymbol{P}_{xx}^{-1}\boldsymbol{P}_{xy} = \sigma_y^{-2}\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$$

With $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$ and $\beta_0 = \mu_y - \boldsymbol{\beta}^T\boldsymbol{\mu}_x$ it is straightforward to verify the following:

- the cross-covariance between y and $y^\star$ is

$$\begin{aligned}\text{Cov}(y, y^\star) &= \boldsymbol{\Sigma}_{yx}\boldsymbol{\beta} = \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}\\ &= \sigma_y^2\boldsymbol{P}_{yx}\boldsymbol{P}_{xx}^{-1}\boldsymbol{P}_{xy} = \sigma_y^2\Omega^2\end{aligned}$$

- the (signal) variance of $y^\star$ is

$$\begin{aligned}\text{Var}(y^\star) &= \boldsymbol{\beta}^T\boldsymbol{\Sigma}_{xx}\boldsymbol{\beta} = \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}\\ &= \sigma_y^2\boldsymbol{P}_{yx}\boldsymbol{P}_{xx}^{-1}\boldsymbol{P}_{xy} = \sigma_y^2\Omega^2\end{aligned}$$

hence the correlation $\text{Cor}(y, y^\star) = \frac{\text{Cov}(y,y^\star)}{\text{SD}(y)\text{SD}(y^\star)} = \Omega$ with $\Omega \geq 0$.

This helps to understand the Ω and Ω^2 coefficients:

- Ω is the linear correlation between the response (y) and prediction $y^\star$.
- Ω^2 is called the **squared multiple correlation** between the scalar y and the vector $\boldsymbol{x}$.
- Note that if we only have one predictor (if x is a scalar) then $\boldsymbol{P}_{xx} = 1$ and $\boldsymbol{P}_{yx} = \rho_{yx}$ so the multiple squared correlation coefficient reduces to squared correlation $\Omega^2 = \rho_{yx}^2$ between two scalar random variables y and x.

17.1.1 Estimation of Ω^2 and the multiple R^2 coefficient

The multiple squared correlation coefficient Ω^2 can be estimated by plug-in of empirical estimates for the corresponding correlation matrices:

$$R^2 = \hat{\boldsymbol{P}}_{yx}\hat{\boldsymbol{P}}_{xx}^{-1}\hat{\boldsymbol{P}}_{xy} = \hat{\sigma}_y^{-2}\hat{\boldsymbol{\Sigma}}_{yx}\hat{\boldsymbol{\Sigma}}_{xx}^{-1}\hat{\boldsymbol{\Sigma}}_{xy}$$

This estimator of Ω^2 is called the **multiple R^2 coefficient**.

If the same scale factor $1/n$ or $1/(n-1)$ is used in estimating the variance σ_y^2 and the covariances $\boldsymbol{\Sigma}_{xx}$ and $\boldsymbol{\Sigma}_{yx}$ then this factor will cancel out.

Above we have seen that Ω^2 is directly linked with the noise variance via

$$\text{Var}(\varepsilon) = \sigma_y^2(1 - \Omega^2)\,.$$

so we can express the squared multiple correlation as

$$\Omega^2 = 1 - \text{Var}(\varepsilon)/\sigma_y^2$$

The **maximum likelihood estimate** of the noise variance $\text{Var}(\varepsilon)$ (also called **residual variance**) can be computed from the residual sum of squares $RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ as follows:

$$\widehat{\text{Var}}(\varepsilon)_{ML} = \frac{RSS}{n}$$

whereas the **unbiased estimate** is obtained by

$$\widehat{\text{Var}}(\varepsilon)_{UB} = \frac{RSS}{n-d-1} = \frac{RSS}{df}$$

where the **degree of freedom** is $df = n - d - 1$ and d is the number of predictors.

Similarly, we can find the maximum likelihood estimate v_y^{ML} for σ_y^2 (with factor $1/n$) as well as an unbiased estimate v_y^{UB} (with scale factor $1/(n-1)$)

The **multiple R^2 coefficient** can then be written as

$$R^2 = 1 - \widehat{\text{Var}}(\varepsilon)_{ML}/v_y^{ML}$$

Note we use MLEs.

In contrast, the so-called **adjusted multiple R^2 coefficient** is given by

$$R^2_{\text{adj}} = 1 - \widehat{\text{Var}}(\varepsilon)_{UB}/v_y^{UB}$$

where the unbiased variances are used.

Both R^2 and R^2_{adj} are estimates of Ω^2 and are related by

$$1 - R^2 = (1 - R^2_{\text{adj}}) \frac{df}{n-1}$$

17.1.2 R commands

In R the command `lm()` fits the linear regression model.

In addition to the regression cofficients (and derived quantities) the R function `lm()` also lists

- the multiple R-squared R^2,
- the adjusted R-squared R^2_{adj},
- the degrees of freedom df and
- the residual standard error $\sqrt{\widehat{\text{Var}}(\varepsilon)_{UB}}$ (computed from the unbiased variance estimate).

See also Worksheet R3 which provides R code to reproduce the exact output of the native `lm()` R function.

17.2 Variance decomposition in regression

The squared multiple correlation coefficient is useful also because it plays an important role in the decomposition of the total variance:

- total variance: $\text{Var}(y) = \sigma_y^2$
- unexplained variance (irreducible error): $\sigma_y^2(1 - \Omega^2) = \text{Var}(\varepsilon)$
- the explained variance is the complement: $\sigma_y^2\Omega^2 = \text{Var}(y^\star)$

In summary:

$$\text{Var}(y) = \text{Var}(y^\star) + \text{Var}(\varepsilon)$$

becomes

$$\underbrace{\sigma_y^2}_{\text{total variance}} = \underbrace{\sigma_y^2\Omega^2}_{\text{explained variance}} + \underbrace{\sigma_y^2(1-\Omega^2)}_{\text{unexplained variance}}$$

The unexplained variance measures the fit after introducing predictors into the model (smaller means better fit). The total variance measures the fit of the model without any predictors. The explained variance is the difference between total and unexplained variance, it indicates the increase in model fit due to the predictors.

17.2.1 Law of total variance and variance decomposition

The **law of total variance** is

$$\underbrace{\text{Var}(y)}_{\text{total variance}} = \underbrace{\text{Var}(\text{E}(y|\boldsymbol{x}))}_{\text{explained variance}} + \underbrace{\text{E}(\text{Var}(y|\boldsymbol{x}))}_{\text{unexplained variance}}$$

provides a very general decomposition in explained and unexplained parts of the variance that is valid regardless of the form of the distributions $F_{y,\boldsymbol{x}}$ and $F_{y|\boldsymbol{x}}$.

In regression it conncects variance decomposition and conditioning. If you plug-in the conditional expections for the multivariate normal model (cf. previous chapter) we recover

$$\underbrace{\sigma_y^2}_{\text{total variance}} = \underbrace{\sigma_y^2\Omega^2}_{\text{explained variance}} + \underbrace{\sigma_y^2(1-\Omega^2)}_{\text{unexplained variance}}$$

17.2.2 Related quantities

Using the above three quantities (total variance, explained variance, and unexplained variance) we can construct a number of scores:

1) **coefficient of determination, squared multiple correlation**:

$$\frac{\text{explained var}}{\text{total var}} = \frac{\sigma_y^2\Omega^2}{\sigma_y^2} = \Omega^2$$

(range 0 to 1, with 1 indicating perfect fit)

2) **coefficient of non-determination, coefficient of alienation**:

$$\frac{\text{unexplained var}}{\text{total var}} = \frac{\sigma_y^2(1-\Omega^2)}{\sigma_y^2} = 1-\Omega^2$$

(range 0 to 1, with 0 indicating perfect fit)

3) F **score**, t^2 **score**:

$$\frac{\text{explained var}}{\text{unexplained var}} = \frac{\sigma_y^2 \Omega^2}{\sigma_y^2(1-\Omega^2)} = \frac{\Omega^2}{1-\Omega^2} = \mathcal{F} = \frac{\tau^2}{n}$$

(range 0 to ∞, with ∞ indicating perfect fit)

Note that the $\mathcal{F}$ and τ^2 scores are population versions of the F and t^2 statistics.

Also note that $\Omega^2 = \frac{\tau^2}{\tau^2+n} = \frac{\mathcal{F}}{\mathcal{F}+1}$ links squared correlation with squared t-scores and F-scores.

17.3 Sample version of variance decomposition

If Ω^2 and σ_y^2 are replaced by their MLEs this can be written in a sample version as follows using data points y_i, predictions $\hat{y}_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$

$$\underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{\text{total sum of squares (TSS)}} = \underbrace{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{\text{explained sum of squares (ESS)}} + \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)^2}_{\text{residual sum of squares (RSS)}}$$

Note that TSS, ESS and RSS all scale with n. Using data vector notation the sample-based variance decomposition can be written in form of the Pythagorean theorem:

$$\underbrace{||\boldsymbol{y} - \bar{y}\mathbf{1}\,||^2}_{\text{total sum of squares (TSS)}} = \underbrace{||\hat{\boldsymbol{y}} - \bar{y}\mathbf{1}||^2}_{\text{explained sum of squares (ESS)}} + \underbrace{||\boldsymbol{y} - \hat{\boldsymbol{y}}||^2}_{\text{residual sum of squares (RSS)}}$$

17.3.1 Geometric interpretation of regression as orthogonal projection:

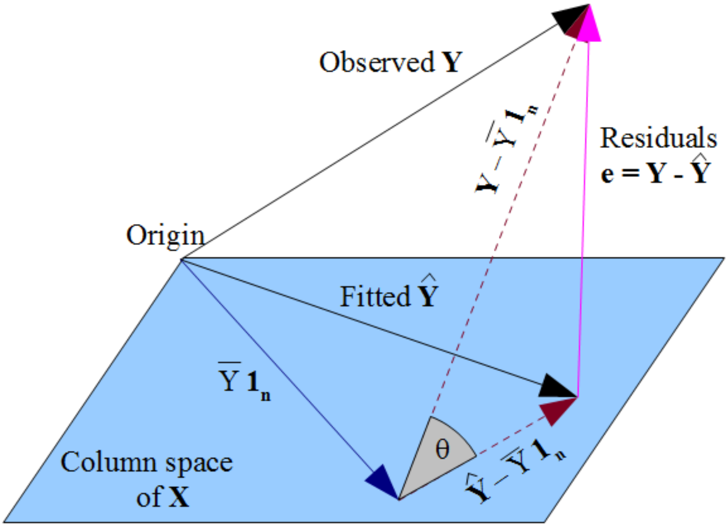
The above equation can be further simplified to

$$||\boldsymbol{y}||^2 = ||\hat{\boldsymbol{y}}||^2 + \underbrace{||\boldsymbol{y} - \hat{\boldsymbol{y}}||^2}_{\text{RSS}}$$

Geometrically speaking, this implies $\hat{\boldsymbol{y}}$ is an orthogonal projection of $\boldsymbol{y}$, since the residuals $\boldsymbol{y} - \hat{\boldsymbol{y}}$ and the predictions $\hat{\boldsymbol{y}}$ are orthogonal (by construction!).

This also valid for the centered versions of the vectors, i.e. $\hat{\boldsymbol{y}} - \bar{y}\mathbf{1}_n$ is an orthogonal projection of $\boldsymbol{y} - \bar{y}\mathbf{1}_n$ (see Figure).

Also note that the angle θ between the two centered vectors is directly related to the (estimated) multiple correlation, with $R = \cos(\theta) = \frac{||\hat{\boldsymbol{y}}-\bar{y}\mathbf{1}_n||}{||\boldsymbol{y}-\bar{y}\mathbf{1}_n||}$, or $R^2 = \cos(\theta)^2 = \frac{||\hat{\boldsymbol{y}}-\bar{y}\mathbf{1}_n||^2}{||\boldsymbol{y}-\bar{y}\mathbf{1}_n||^2} = \frac{\text{ESS}}{\text{TSS}}$.


Observed **Y**
$\mathbf{Y} - \overline{\mathrm{Y}}\,\mathbf{1}_n$
Residuals
$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$
Origin
Fitted $\hat{\mathbf{Y}}$
$\overline{\mathrm{Y}}\,\mathbf{1}_n$
θ
$\hat{\mathbf{Y}} - \overline{\mathrm{Y}}\,\mathbf{1}_n$
Column space
of **X**

Source of Figure: Stack Exchange

Chapter 18

Prediction and variable selection

In this chapter we discuss how to compute (lower bounds) of the prediction error and how to select variables relevant for prediction

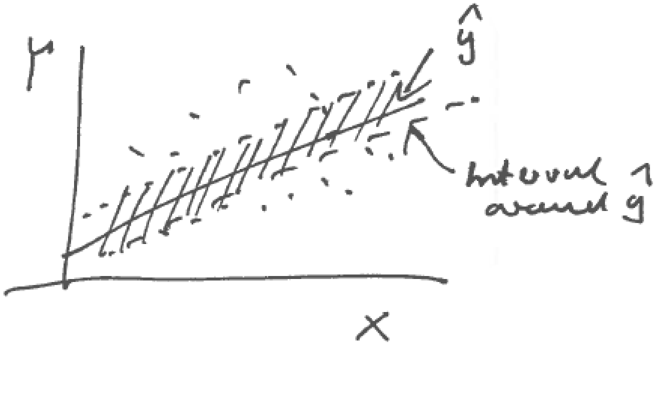
18.1 Prediction and prediction intervals

Learning the regression function from (training) data is only the first step in application of regression models.

The next step is to actually make **prediction** of future outcomes y^{test} given test data $\boldsymbol{x}^{\text{test}}$:

$$y^{\text{test}} = \hat{y}(\boldsymbol{x}^{\text{test}}) = \hat{f}_{\hat{\beta}_0, \hat{\boldsymbol{\beta}}}(\boldsymbol{x}^{\text{test}})$$

Note that y^{test} is a point estimator. Is it possible also to construct a corresponding interval estimate?



The answer is yes, and leads back to the conditioning approach:

$$y^\star = \text{E}(y|\boldsymbol{x}) = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}$$

$$\text{Var}(\varepsilon) = \text{Var}(y|\boldsymbol{x}) = \sigma_y^2(1 - \Omega^2)$$

We know that the mean squared prediction error for $y^\star$ is $\text{E}((y - y^\star)^2) = \text{Var}(\varepsilon)$ and that this is the minimal irreducible error. Hence, we may use $\text{Var}(\varepsilon)$ as the *minimum* variability for the prediction.

The corresponding prediction interval is

$$\left[y^\star(\boldsymbol{x}^{\text{test}}) \pm c \times \text{SD}(\varepsilon)\right]$$

where c is some suitable constant (e.g. 1.96 for symmetric 95% normal intervals).

However, please note that the prediction interval constructed in this fashion will be an *underestimate*. The reason is that this assumes that we employ $y^\star = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}$ but in reality we actually use $\hat{y} = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ for prediction — note the estimated coefficients! We recall from an earlier chapter (best linear predictor) that this leads to increase of MSPE compared with using the optimal β_0 and $\boldsymbol{\beta}$.

Thus, for better prediction intervals we would need to consider the mean squared prediction error of $\hat{y}$ that can be written as $\text{E}((y - \hat{y})^2) = \text{Var}(\varepsilon) + \delta$ where δ is an **additional error term due to using an estimated rather than the true regression function**. δ typically declines with $1/n$ but can be substantial for small n (in particular as it usually depends on the number of predictors d).

For more details on this we refer to later modules on regression.

18.2 Variable importance and prediction

Another key question in regression modelling is to find out predictor variables $x_1, x_2, \ldots, x_d$ are actually important for predicting the outcome y.

→ We need to study variable importance measures (VIM).

18.2.1 How to quantify variable importance?

A variable x_i is **important** if it **improves prediction** of the response y.

Recall variance decomposition:

$$\text{Var}(y) = \sigma_y^2 = \underbrace{\sigma_y^2 \Omega^2}_{\text{explained variance}} + \underbrace{\sigma_y^2(1 - \Omega^2)}_{\text{unexplained/residual variance} = \text{Var}(\varepsilon)}$$

- Ω^2 squared multiple correlation $\in [0,1]$
- Ω^2 large $\to 1$ predictor variables explain most of σ_y^2
- Ω^2 small $\to 0$ linear model fails and predictors do not explain variability
- $\Rightarrow$ If a predictor helps to $\begin{array}{c}\text{increase explained variance}\\ \text{decrease unexplained variance}\end{array}$ then it is important!
- $\Omega^2 = \boldsymbol{P}_{yx}\boldsymbol{P}_{xx}^{-1}\boldsymbol{P}_{xy} \hat{=}$ a function of the X!

VIM: which predictors contribute most to Ω^2

18.2.2 Some candidates for VIMs

1. The regression coefficients $\boldsymbol{\beta}$

   - $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy} = \boldsymbol{V}_x^{-1/2}\boldsymbol{P}_{xx}^{-1}\boldsymbol{P}_{xy}\sigma_y$
   - Not a good VIM since $\boldsymbol{\beta}$ contains the scale!
   - Large $\hat{\beta}_i$ does not indicate that x_i is important.
   - Small $\hat{\beta}_i$ does not indicate that x_i is not important.

2. Standardised regression coefficients $\boldsymbol{\beta}_{\text{std}}$

   - $\boldsymbol{\beta}_{\text{std}} = \boldsymbol{P}_{xx}^{-1}\boldsymbol{P}_{xy}$
   - implies $\text{Var}(y) = 1$, $\text{Var}(x_i) = 1$
   - These do not contain the scale (so better than $\hat{\beta}$)
   - But still unclear how this relates to decomposition of variance

3. Squared marginal correlations ρ_{y,x_i}^2

   Consider case of uncorrelated predictors: $\boldsymbol{P}_{xx} = \boldsymbol{I}$ (no correlation among x_i)

   $$\Rightarrow \Omega^2 = \boldsymbol{P}_{yx}\boldsymbol{P}_{xy} = \sum_{i=1}^{d} \rho_{y,x_i}^2$$

   $\rho_{y,x_i}^2 = \text{Cor}(y, x_i)$ is the marginal correlation between y and x_i, and Ω^2 is (for uncorrelated predictors) the sum of squared marginal correlations.

   - If $\boldsymbol{P}_{xx} = \boldsymbol{I}$, then *ranking* predictors by ρ_{y,x_i}^2 is optimal!
   - The predictor with largest marginal correlation reduces the unexplained variance most!
   - good news: even if there is weak correlation among predictors the marginal correlations are still good as VIM (but then they will not perfectly add up to Ω^2)
   - Advantage: very simple but often also very effective.
   - Caution! If there is strong correlation in $\boldsymbol{P}_{xx}$, then there is **colinearity** (in this case it is oftern best to remove one of the strongly correlated variables, or to merge the correlated variables).

Often, ranking predictors by their squared marginal correlations is done as a prefiltering step (independence screening).

18.3 Regression t-scores.

18.3.1 Wald statistic for regression coefficients

So far, we discussed three obvious candidates for for variable importance measures (regression coefficients, standardised regression coefficients, marginal correlations).

In this section we consider a further quantity, the **regression t−score**:

Recall that ML estimation of the regression coefficients yields

- a point estimate $\hat{\boldsymbol{\beta}}$
- the (asymptotic) variance $\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})$
- the asymptotic normal distribution $\hat{\boldsymbol{\beta}} \overset{a}{\sim} N_d(\boldsymbol{\beta}, \widehat{\text{Var}}(\hat{\boldsymbol{\beta}}))$

Corresponding to each predictor x_i we can construct from the above a t-score

$$t_i = \frac{\hat{\beta}_i}{\widehat{\text{SD}}(\hat{\beta}_i)}$$

where the standard deviations are computed by $\widehat{\text{SD}}(\hat{\beta}_i) = \text{Diag}(\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}))_i$. This corresponds to the **Wald statistic** to test that the underlying true regression coefficient is zero ($\beta_i = 0$).

Correspondingly, under the null hypthesis that $\beta_i = 0$ asymptotically for large n the regression t-score is standard normal distributed:

$$t_i \overset{a}{\sim} N(0, 1)$$

This allows to compute (symmetric) p-values $p = 2\Phi(-|t_i|)$ where Φ is the standard normal distribution function.

For finite n, assuming normality of the observation and using the unbiased estimate for variance when computing t_i, the exact distribution of i_i is given by the Student-t distribution:

$$t_i \sim t_{n-d-1}$$

Regression t-scores can thus be used to test whether a regression coefficient is zero. A large magnitude of the t_i score indicates that the hypothesis $\beta_i = 0$ can be rejected. Thus, a small p-value (say smaller than 0.05) signals that the regression coefficient is non-zero and hence that the corresponding predictor variable should be included in the model.

This allows rank predictor variables by $|t_i|$ or the corresponding p-values with regard to their relevance in the linear model. Typically, in order to simplify a model, predictors with the largest p-values (and thus smallest absolute t-scores) may be removed from a model. However, note that having a p-value say larger than 0.05 by itself is not sufficient to declare a regression coefficient to be zero (because in classical statistical testing you can only reject the null hypothesis, but not accept it!).

Note that by construction the regression t-scores do not depend on the scale, so when the original data are rescaled this will not affect the corresponding regression t-scores. Furthermore, if $\widehat{\text{SD}}(\hat{\beta}_i)$ is small, then the regression t-score t_i can still be large even if $\hat{\beta}_i$ is small!

18.3.2 Computing

When you perform regression analysis in R (or another statistical software package) the computer will return the following:

$\hat{\beta}_i$ Estimated repression coefficient	$\widehat{\text{SD}}(\hat{\beta}_i)$ Error of $\hat{\beta}_i$	$t_i = \frac{\hat{\beta}_i}{\widehat{\text{SD}}(\hat{\beta}_i)}$ t-score computed from first two columns	p-values for t_i based on t-distribution	Indicator of Significance * 0.9 ** 0.95 *** 0.99

In the `lm()` function in R the standard deviation is the square root of the unbiased estimate of the variance (but note that it itself is not unbiased!).

18.3.3 Connection with partial correlation

The deeper reason why ranking predictors by regression t-scores and associated p-values is useful is their link with **partial correlation**.

In particular, the (squared) regression t-score can be 1:1 transformed into the (estimated) (squared) partial correlation

$$\hat{\rho}^2_{y,x_i|x_{j\neq i}} = \frac{t_i^2}{t_i^2 + df}$$

with $df = n - d - 1$, and it can be shown that the p-values for testing that $\beta_i = 0$ are exactly the same as the p-values for testing that the partial correlation $\rho_{y,x_i|x_{j\neq i}}$ vanishes!

Therefore, ranking the predictors x_i by regression t-scores leads to exactly the same ranking and p-values as partial correlation!

18.3.4 Squared Wald statistic and the F statistic

In the above we looked at individual regression coefficients. However, we can also construct a Wald test using the complete vector $\hat{\boldsymbol{\beta}}$. The squared Wald statistic to test that $\boldsymbol{\beta} = 0$ is given by

$$
\begin{aligned}
t^2 &= \hat{\boldsymbol{\beta}}^T \widehat{\text{Var}}(\hat{\boldsymbol{\beta}}^{-1}) \hat{\boldsymbol{\beta}} \\
&= \left(\hat{\boldsymbol{\Sigma}}_{yx} \hat{\boldsymbol{\Sigma}}_{xx}^{-1}\right) \left(\frac{n}{\widehat{\sigma^2_\varepsilon}} \hat{\boldsymbol{\Sigma}}_{xx}\right) \left(\hat{\boldsymbol{\Sigma}}_{xx}^{-1} \hat{\boldsymbol{\Sigma}}_{xy}\right) \\
&= \frac{n}{\widehat{\sigma^2_\varepsilon}} \hat{\boldsymbol{\Sigma}}_{yx} \hat{\boldsymbol{\Sigma}}_{xx}^{-1} \hat{\boldsymbol{\Sigma}}_{xy} \\
&= \frac{n}{\widehat{\sigma^2_\varepsilon}} \hat{\sigma}^2_y R^2
\end{aligned}
$$

With $\widehat{\sigma^2_\varepsilon} / \hat{\sigma}^2_y = 1 - R^2$ we finally get the related F statistic

$$
\frac{t^2}{n} = \frac{R^2}{1 - R^2} = F
$$

which is a function of R^2. If $R^2 = 0$ then $F = 0$. If R^2 is large (< 1) then F is large as well ($< \infty$) and the null hypothesis $\boldsymbol{\beta} = 0$ can be rejected, which implies that at least one regression coefficient is non-zero. Note that the squared Wald statistic t^2 is asymptotically χ^2_d distributed which is useful to find critical values and to compute p-values.

18.4 Further approaches for variable selection

In addition to ranking by marginal and partial correlation, there are many other approaches for variable selection in regression!

a) Search-based methods:

- search through subsets of linear models for d variables, ranging from full model (including all predictors) to the empty model (includes no predictor) and everything inbetween.
- Problem: exhaustive search not possible even for relatively small d as space of models is very large!
- Therefore heuristic approaches such as forward selection (adding predictors), backward selection (removing predictors), or monte-carlo random search are employed.
- Problem: maximum likelihood cannot be used for choosing among the models - since ML will always pick the best model. Therefore, penalised ML criteria such as AIC or Bayesian criteria are often employed instead.

b) Integrative estimation and variable selection:

- there are methods that fit the regression model and perform variable selection *simultaneously*.
- the most well-known approach of this type is "lasso" regression (Tibshirani 1996)
- This applies a (generalised) linear model with ML plus L1 penalty.
- Alternative: Bayesian variable selection and estimation procedures

c) Entropy-based variable selection:

As seen above, two of the most popular approaches in linear models are based on correlation, either marginal correlation or partial correlation (via regression t-scores).

Correlation measures can be generalised to non-linear settings. One very popular measure is the **mutual information** which is computed using the KL divergence. In case of two variables x and y with joint normal distribution and correlation ρ the mutual information is a function of the correlation:

$$\text{MI}(x, y) = \frac{1}{2}\log(1 - \rho^2)$$

In regression he mutual information between the response y and predictor x_i is $\text{MI}(y, x_i)$, and this widely used for feature selection, in particular in machine learning.

d) FDR based variable selection in regression:

Feature selection controling the false discovery rate (FDR) among the selected features are becoming more popular, in particular a number of procedures using so-called "knockoffs", see https://web.stanford.edu/group/candes/knockoffs/ .

e) Variable importance using Shapley values:

Borrowing a concept from game theory Shapley values have recently become popular in machine learning to evaluate the variable importance of predictors in nonlinear models. Their relationship to other statistical methods for measuring variable importance is the focus of current research.