52571: Regression and Statistical Models

Spring 2024

Asaf Weinstein

1 What is regression?

In Statistics, *regression* refers to the modeling and analysis of the relationship between a response variable Y and a set of explanatory variables $X_1, \dots, X_p$. To use the most basic terms, in a regression problem we are trying to learn how Y changes as $X_1, \dots, X_p$ change. In that sense, regression analysis studies the *dependence* of the response on the explanatory variables. Here are some real-life examples:

Example 1. $Y =$ weight of a person. $X =$ height of a person

Example 2. $Y =$ percent body weight from fat. $X =$ circumference of abdomen.

Example 3. $Y =$ typical (median) house value in a neighborhood of the city of Boston. $X_1 =$ average distance from employment centers, $X_2 =$ number of rooms per house, $X_3 =$ level of air pollution.

In the elementary description above we have intentionally avoided 'causal' terminology: we did not state our aim as studying what change in Y is *caused* by a change in the values of $X_1, \dots, X_p$, because the tools covered in this course are generally designed to learn only about *association* between $X_1, \dots, X_p$ and Y, not causation. This caveat should be kept in mind throughout the course.

In the first two examples $p = 1$, as we have a single explanatory variable. In the third example $p = 3$ as we have three explanatory variables. Now let's think about the general description of a regression problem in the context of the first example above, and imagine that X and Y above were to be measured on each male person in the entire population in the United States. We all know that taller people tend to have higher weight. At the same time, if I consider men with height just about 176 cm, e.g., then of course we do not expect all of them to have exactly the same weight. In other words, weight is not fully *determined* by height, and it is certainly possible to find, for example, a 172 cm tall male whose weight is greater than that of a 176 cm tall male. What exactly could we be referring to, then, when we say that taller people "tend" to have higher weight? We can imagine that, if for some fixed value of x we considered only men of height $X \in (x - \Delta, x + \Delta)$, for small Δ, then for every x the corresponding weights, the corresponding values of the response,

$$\mathscr{Y}_x := \{Y : X \in (x - \Delta, x + \Delta)\},$$

have some *distribution*. This can be regarded as approximating the *conditional distribution* of Y on $X = x$. When we say that taller people *tend* to have higher weight, it is reasonable to expect that the *average* of values in $\mathscr{Y}_x$ increase with x. We can imagine a function that maps each x to the average of the values in $\mathscr{Y}_x$—the *conditional mean* of Y for $X = x$. This function is called the *general regression function*, and is standardly the main object of interest in regression analysis. As illustrated in Figure 9, this line can be thought of as being formed, approximately, by taking small bins on the X axis, and drawing a horizontal line, that spans the bin width, at the average value of Y in every bin separately. The function $x \mapsto \mathrm{mean}\,(\mathscr{Y}_x)$ describes the

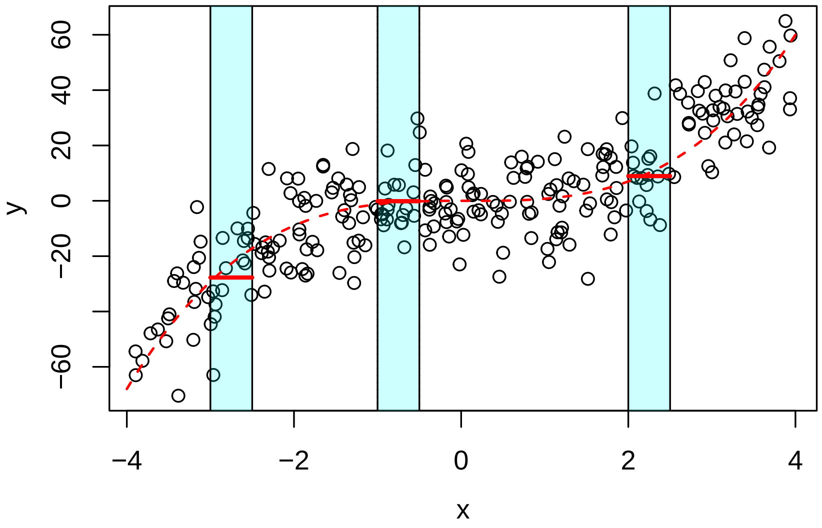


Figure 1: An illustration for general regression. The observed datapoints are represented by the black circles. Light blue rectangles are bins of small width 2Δ on the X-axis, and horizontal red lines mark the average of points (X,Y) with X values in the corresponding bin, i.e. the average of $\mathscr{Y}_x$, for three different values of x. The dashed red line represents a "true" regression line which can be imagined to approximately form if we binned tightly on the entire X range, and consider the (step) function obtained by the horizontal red lines.

population of all males in the USA. In other words, to know this function *exactly* would require access to the weights and heights of all US males. To have access to the records of all subjects in the population is usually impractical. This is where Statistics enters: we will take on the task of *estimating* this function when we only have access to a *sample* from the population.

To estimate a general regression function based on a sample only, could be quite challenging, especially when the number of explanatory variables p is large. In this course we will focus on the case where the population regression line is assumed to be a *linear* function of the explanatory variables, which in the single explanatory case illustrated in Figure 1 the dashed red line would simply be a straight line. Our goal will be to estimate this *linear regression function* and provide statistical inference for it. (Remark: linear regression analysis in fact has a meaning also when the population regression function is not linear, but it is simpler to describe the target in linear regression analysis as we did above, and it is also consistent with what will follow in subsequent chapters).

2 Simple regression and the Least-Squares method

We start with the situation of a single explanatory variable, $p = 1$, in which case we say that we're dealing with *simple* regression. Suppose that we have a dataset of n sample points,

$$(x_i, y_i), \qquad i = 1, ..., n,$$

where x_i records a measurement on some explanatory variable $X \in \mathbb{R}$, and y_i records a measurement of some response variable $Y \in \mathbb{R}$. For illustration, we will use a dataset with $n = 50$ measurements of $X =$car

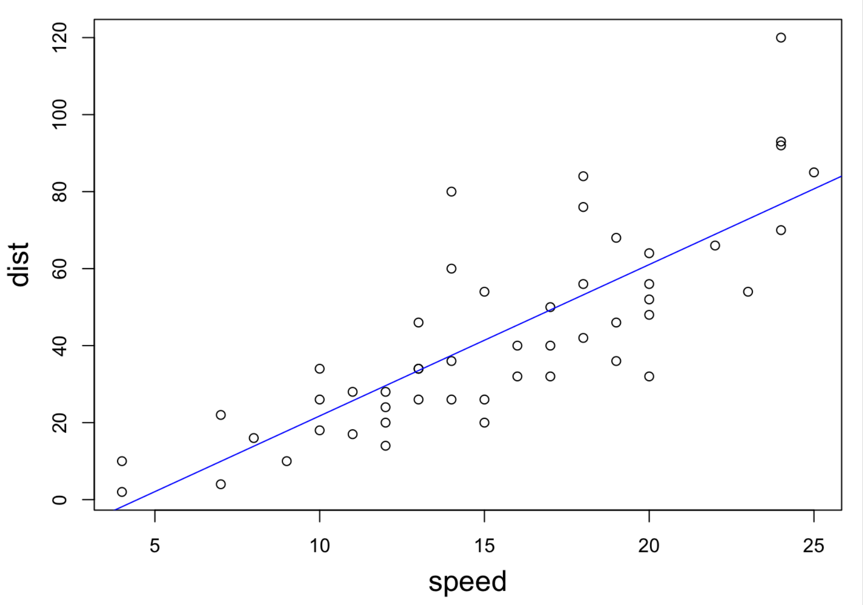


Figure 2: Car data: stopping distance (Y) vs. car speed (X). Blue line is the least squares line.

speed, Y =stopping distance. Figure 2 shows a *scatter plot* of the data, a graph of y_i versus x_i for each of the data points. We want to use this data to learn how X *linearly* explains Y[1]. This means we want to *fit* to (estimate from) the scatter of points a straight line,

$$y = \hat{\beta}_0 + \hat{\beta}_1 x, \tag{1}$$

where the intercept $\hat{\beta}_0$ and slope $\hat{\beta}_1$ are functions of the data $(x_i, y_i),\ i = 1, ..., n$ (hence the "hats" in the notation). For any fitted line (1), we define the *predicted values* (also called *fitted values*) to be

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i. \tag{2}$$

Intuitively, a "good" fit would make the predicted values $\hat{y}_i$ close to the observed values y_i. For any candidate straight line $y = b_0 + b_1 x$, different metrics can be used to measure how well the predicted values agree with the observations. The most standard one is to measure this by the sum of the squared errors between $\hat{y}_i$ and y_i,

$$Q(b_0, b_1) = \sum_{i=1}^{n} \big(y_i - (b_0 + b_1 x_i)\big)^2,$$

and notice that the objective function $Q(b_0, b_1)$ depends on the observations (x_i, y_i). This suggests to obtain the *estimates* $\hat{\beta}_0$ and $\hat{\beta}_1$ by minimizing the objective over (b_0, b_1),

$$(\hat{\beta}_0, \hat{\beta}_1) \ = \ \underset{(b_0, b_1)}{\arg\min}\, Q(b_0, b_1). \tag{3}$$

The criterion entailing the estimation of the slope and intercept by minimizing the sum of squared errors, is called the *least squares* (LS) criterion. The values $(\hat{\beta}_0, \hat{\beta}_1)$ in (3) are the *least squares estimates* of the

[1] for this particular dataset, a straight line indeed seems appropriate for describing the "trend", but what follows does not assume this.

coefficients. The line obtained by substituting $(\hat{\beta}_0, \hat{\beta}_1)$ into (1), appearing in blue in the figure, is called the *least squares line*, or the estimated (linear) *regression line*. From now on, whenever we use the symbols $\hat{\beta}_0, \hat{\beta}_1$ and unless indicated otherwise, we will refer to the least squares coefficients, although we should bear in mind that many other methods (criteria) can be used to fit a line to the data (for example, minimizing the sum of absolute errors instead of the squared errors).

An advantage of using *squared* errors in the objective function, is that it makes $Q\left(b_0, b_1\right)$ differentiable, and so for finding a minimum we just look for a point where the partial derivatives vanish:

$$\begin{aligned}
\frac{\partial Q}{\partial b_0} &= -2\sum_{i=1}^{n}\left(y_i - \left[b_0 + b_1 x_i\right]\right)\left(\frac{\partial}{\partial b_0}\left[b_0 + b_1 s_i\right]\right) = -2\sum_{i=1}^{n}\left(y_i - \left[b_0 + b_1 x_i\right]\right)\\
\frac{\partial Q}{\partial b_1} &= -2\sum_{i=1}^{n}\left(y_i - \left[b_0 + b_1 x_i\right]\right)\left(\frac{\partial}{\partial b_1}\left[b_0 + b_1 x_i\right]\right) = -2\sum_{i=1}^{n}\left(y_i - \left[b_0 + b_1 x_i\right]\right) x_i
\end{aligned}$$

Then, we get

$$\begin{aligned}
\sum_{i=1}^{n}\left(y_i - \left[b_0 + b_1 x_i\right]\right) &= 0\\
\sum_{i=1}^{n}\left(y_i - \left[b_0 + b_1 x_i\right]\right) x_i &= 0
\end{aligned}$$

which can be rewritten as

$$\begin{aligned}
b_0 + \bar{x} b_1 &= \bar{y}\\
\bar{x} b_0 + \left(n^{-1}\sum_{i=1}^{n} x_i^2\right) b_1 &= n^{-1}\sum_{i=1}^{n} x_i y_i
\end{aligned}$$

where $\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i, \bar{y} := \frac{1}{n}\sum_{i=1}^{n} y_i$. From the first equation we have $b_0 = \bar{y} - \bar{x} b_1$. Substituting this into the second equation and solving for b_1 gives

$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

which can also be written as

$$b_1 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}.$$

Exercise. Prove the equivalence between (1.3) and (1.4). Use the fact that for any vector $z = (\boldsymbol{z}_1, \ldots, z_n)$, we have

$$\begin{aligned}
\sum_{i=1}^{n}\left(z_i - \bar{z}\right)^2 &= \sum_{i=1}^{n}\left(z_i^2 - 2\bar{z} z_i + \bar{z}^2\right)\\
&= \sum_{i=1}^{n} z_i^2 - 2\bar{z}\sum_{i=1}^{n} z_i + \sum_{i=1}^{n}\bar{z}^2\\
&= \sum_{i=1}^{n} z_i^2 - 2n\bar{z}^2 + n\bar{z}^2\\
&= \sum_{i=1}^{n} z_i^2 - n\bar{z}^2
\end{aligned}$$

(Remark: if we define a random variable Z by $P\left(Z=z_i\right)=\frac{1}{n}$, then this is just the familiar relationship

$$\underbrace{V(Z)}_{n^{-1}\sum(z_i-\bar{z})^2} = \underbrace{EZ^2}_{n^{-1}\sum z_i^2} - \underbrace{(EZ)^2)}_{\bar{z}^2}.$$

To summarize, the LS solution is given by

$$\hat{\beta}_0+\hat{\beta}_1 x,$$

where

$$\boxed{\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^n\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sum_{i=1}^n\left(x_i-\bar{x}\right)^2} && (4)\\ \hat{\beta}_0 &= \bar{y}-\hat{\beta}_1 \bar{x} && (5) \end{aligned}}$$

Thus, the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ have closed-form expressions as functions of the data. In fact, the LS coefficients are linear in the y_i 's (formally meaning that (1.7) and (1.8), viewed as functions of $\boldsymbol{y}=\left(y_1, \ldots, y_n\right)$ only, are affine).

Remark. That the LS coefficients are linear in y is unrelated to the fact that we are attempting to fit a linear function (a straight line) to the data: e.g., if we used absolute deviations then the fitted line is still a straight line, but the coefficients will not be linear functions in $\boldsymbol{y}$.

Recall that for two vectors $\boldsymbol{u}=\left(u_1, \ldots, u_n\right), \boldsymbol{v}=\left(v_1, \ldots, v_n\right)$, the (empirical) *correlation coefficient* between $\boldsymbol{u}$ and $\boldsymbol{v}$ is defined as

$$r_{u,v}:=\frac{\sum_i\left(u_i-\bar{u}\right)\left(v_i-\bar{v}\right)}{\sqrt{\sum_i\left(u_i-\bar{u}\right)^2}\sqrt{\sum_i\left(v_i-\bar{v}\right)^2}},$$

which can also be written as

$$\frac{\frac{1}{n-1}\sum_i\left(u_i-\bar{u}\right)\left(v_i-\bar{v}\right)}{s_u\cdot s_v}, \qquad s_u^2=\frac{1}{n-1}\sum_i(u_i-\bar{u})^2,\ s_v^2=\frac{1}{n-1}\sum_i(v_i-\bar{v})^2$$

(s_u, s_v are the the *standard errors* of $\boldsymbol{u}$ and $\boldsymbol{v}$, respectively, and the numerator can be interpreted as the empirical *covariance* of $\boldsymbol{u}$ and $\boldsymbol{v}$). Therefore, we can obtain an equivalent expression for the slope of the LS line,

$$\hat{\beta}_1=\frac{\sum_{i=1}^n\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sum_{i=1}^n\left(x_i-\bar{x}\right)^2}=\frac{\sum_{i=1}^n\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sqrt{\sum_{i=1}^n\left(x_i-\bar{x}\right)^2}\sqrt{\sum_{i=1}^n\left(y_i-\bar{y}\right)^2}}\cdot\frac{\sqrt{\sum_{i=1}^n\left(y_i-\bar{y}\right)^2}}{\sqrt{\sum_{i=1}^n\left(x_i-\bar{x}\right)^2}}=r_{\boldsymbol{x},\boldsymbol{y}}\cdot\frac{s_y^2}{s_x^2}. \tag{6}$$

Recall the definition of the predicted values, $\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1 x_i$. We define The i th *residual* is defined to be

$$e_i:=y_i-\hat{y}_i.$$

Proposition 1. *For the least squares fit we have the following properties:*

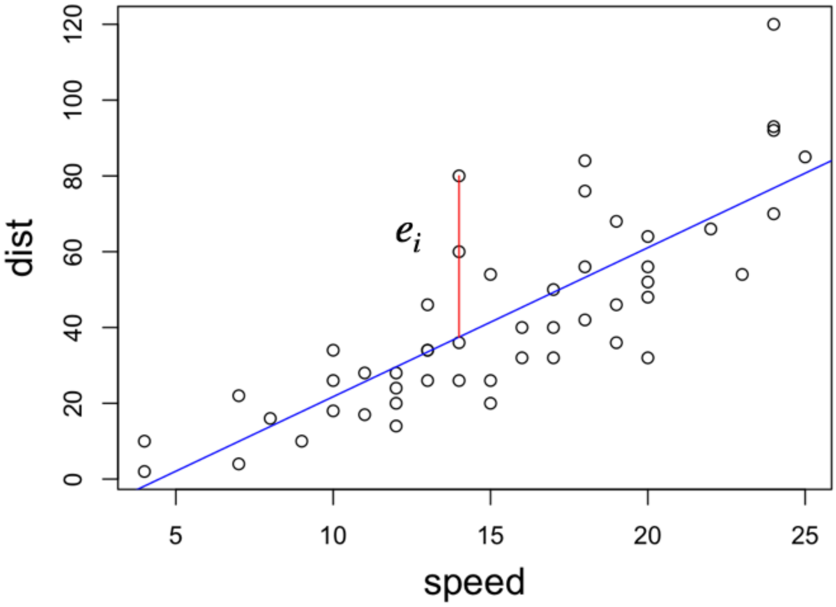


Figure 3: Car data: residuals.

1. $\sum_{i=1}^{n} e_i = 0$
2. $\sum_{i=1}^{n} x_i e_i = 0$
3. $\sum_{i=1}^{n} \hat{y}_i e_i = 0$
4. $n^{-1} \sum_{i=1}^{n} \hat{y}_i = \bar{y}$

Proof. Properties 1 and 2 are true because the LS solution $(\hat{\beta}_0, \hat{\beta}_1)$ satisfies equations (1.3) and (1.4), respectively. Property 3 follows from properties 1 and 2 . For Property 4 , using 1 and the definition of the residuals, we have

$$\frac{1}{n}\sum \hat{y}_i = \frac{1}{n}\sum (y_i - e_i) = \frac{1}{n}\sum y_i = \bar{y}.$$

□

Sums of squares decomposition

$$SST := \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad \text{Sum of Squares Total}$$

$$SSR := \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \qquad \text{Sum of Squares Regression}$$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 := \sum_{i=1}^{n} e_i^2 \qquad \text{Sum of Squares Error}$$

SST measures the total ("marginal") variation in y_i; SSR measures the variation explained by the linear regression, i.e., the variation in the y_i which can be accounted for (linearly) by the x_i; and SSE is the leftover variation, i.e, the variation in the y_i which cannot be explained by the x_i.

Note that we have

$$SSR := \sum_i (\hat{y}_i - \bar{y})^2 = \sum_i (\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})^2 = \sum_i [(\hat{\beta}_1 (x_i - \bar{x})]^2 = \hat{\beta}_1^2 \sum_i (x_i - \bar{x})^2 \tag{7}$$

where we used the fact that $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$.

Proposition 2. *SST decomposes as*

$$SST = SSR + SSE.$$

Proof. We have

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n [(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + 2 \sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

Now,

$$\sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \sum_{i=1}^n e_i (\hat{y}_i - \bar{y}) = \sum_{i=1}^n e_i \hat{y}_i - \bar{y} \sum_{i=1}^n e_i = 0 - 0,$$

using Properties $1 + 3$ of the LS estimates. Together, we get

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n [(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = SSE + SSR$$

□

Definition 1. For simple linear regression, the *coefficient of determination*, also called the "R squared", is

$$R^2 := \frac{SSR}{SST}$$

The R^2 value measures the proportion of the total variance of the y_i's explained (linearly!) by x_i's. Note that, by Proposition 2, we have

$$R^2 = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}.$$

The R^2 value is connected to Pearson's correlation coefficient, as the following proposition asserts.

Proposition 3. *We have*

1. $r_{x,y} = \operatorname{sign}\left(\hat{\beta}_1\right) \sqrt{R^2}$ *(which implies $r_{x,y}^2 = R^2$).*
2. $r_{y,\hat{y}} = \sqrt{R^2}$ *(which implies $r_{\hat{y},y}^2 = R^2$).*

Note: this means that $|r_{x,y}| = |r_{y,\hat{y}}|$, and that $r_{y,\hat{y}} \geq 0$. This relation gives an interpretation for the correlation coefficient between $\boldsymbol{x}$ and $\boldsymbol{y}$: Since $r_{x,y}^2 = R^2$, the square of the Pearson correlation coefficient is the fraction of variation in $\boldsymbol{y}$ that can be linearly explained by $\boldsymbol{x}$ (i.e., by a least-squares regression of $\boldsymbol{y}$ on $\boldsymbol{x}$.

Proof. From (6) we know

$$\hat{\beta}_1 = r_{\boldsymbol{x},\boldsymbol{y}} \cdot \frac{s_y^2}{s_x^2}.$$

Also, Using Equation (7), we get

$$R^2 := \frac{SSR}{SST} = \frac{\hat{\beta}_1^2 \sum_i (x_i - \bar{x})^2}{\sum_i (y_i - \bar{y})^2} = \hat{\beta}_1^2 \frac{s_x^2}{s_y^2}.$$

From the two displays above,

$$r_{x,y} = \hat{\beta}_1 \frac{s_x}{s_y} = \sqrt{\left(\hat{\beta}_1 \frac{s_x}{s_y}\right)^2} \operatorname{sign}\left(\hat{\beta}_1 \frac{s_x}{s_y}\right) = \sqrt{R^2} \cdot \operatorname{sign}(\hat{\beta}_1)$$

For the second part of the claim, we have $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ and $\bar{\hat{y}} := \frac{1}{n} \sum_i \hat{y}_i = \bar{y}$, hence

$$\sum_i (\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y}) = \sum_i (\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})(y_i - \bar{y}) = \sum_i \hat{\beta}_1 (x_i - \bar{x})(y_i - \bar{y}) = \hat{\beta}_1 \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

$$\sum_i (\hat{y}_i - \bar{\hat{y}})^2 = \sum_i [\hat{\beta}_1 (x_i - \bar{x})]^2 = \hat{\beta}_1^2 \sum_i (x_i - \bar{x})^2$$

so that

$$r_{\hat{y},y} = \frac{\sum_i (\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sqrt{\sum_i (\hat{y}_i - \bar{\hat{y}})^2} \sqrt{\sum_i (y_i - \bar{y})^2}} = \frac{\hat{\beta}_1 \sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\hat{\beta}_1^2 \sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}},$$

and the last expression is $r_{x,y}$ by definition. □

3 Multiple linear regression

So far we have dealt with a single explanatory variable. We now generalize the ideas to multiple explanatory variables. Thus, the data is now

$$(x_{i1}, \dots, x_{ip}, y_i)\,, \qquad i = 1, ..., n.$$

The explanatory variable for the ith observation is a p-dimensional vector, $\boldsymbol{x}_i = (x_{i1}, \dots, x_{ip})$, and we assume throughout $p + 1 \leq n$ (we will see soon why the case $p + 1 > n$ is problematic).

In the multiple explanatory variables case it is generally challenging to visualize the data points, because we quickly run out of dimensions if we attempt to generate a scatterplot (the case $p = 2$ is still doable because we can draw a 3-dimensional plot, showing X_1, X_2 on the XY-plane and the response Y on the Z axis). However, the mathematical concepts from the simple regression case can still be extended to the general-p case. Thus, we may still try to predict y by a *linear* function of $x_1, \dots, x_p$, i.e., the analog of (1) will be

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p = \sum_{j=0}^{p} \hat{\beta}_j x_j,$$

where for each observation we prepend $x_{i0} \equiv 1$ to the vector of explanatory variables, that is, we define $\boldsymbol{x}_i = (1, x_{i1}, \dots, x_{ip})$ instead of $\boldsymbol{x}_i = (x_{i1}, \dots, x_{ip})$.

For any method for fitting such a linear function (that is, for calculating $\hat{\beta}_1, \dots, \hat{\beta}_p$ as a function of the observed data), the *fitted values* are

$$\hat{y}_i = \sum_{j=0}^{p} \hat{\beta}_j x_{ij},$$

and the *residuals* are defined exactly as before,

$$e_i := y_i - \hat{y}_i$$

We can generalize also the least squares method, by seeking the set of coefficients $\hat{\beta}_1, \dots, \hat{\beta}_p$ which, as before, minimize the sum of squared errors. Formally, let $\boldsymbol{b} = (b_0, \dots, b_p)^\top \in \mathbb{R}^{p+1}$, and define

$$Q(\boldsymbol{b}) := \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} b_j x_{ij} \right)^2 . \tag{8}$$

This is a differentiable function (now of $p + 1$ variables), and we can take partial derivatives and set them to zero. We have

$$\frac{\partial}{\partial b_r} \sum_{j=0}^{p} x_{ij} b_j = x_{ir},$$

and, therefore,

$$\frac{\partial Q}{\partial b_r} = -2 \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} x_{ij} b_j \right) \left(\frac{\partial}{\partial b_r} \sum_{j=0}^{p} x_{ij} b_j \right) = -2 \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} x_{ij} b_j \right) x_{ir}.$$

Setting this to zero for $r = 0, \dots, p$ yields the so-called *Normal equations*,

$$\boxed{\sum_{j=0}^{p} \left(\sum_{i=1}^{n} x_{ir} x_{ij} \right) b_j = \sum_{i=1}^{n} x_{ir} y_i, \qquad r = 0, \dots, p.} \tag{9}$$

The solution to this set of $p+1$ equations (in $p+1$ variables), assuming it exists and is unique (we will see conditions for this to be the case), gives the least squares coefficients $\hat{\beta}_1, \ldots, \hat{\beta}_p$.

The system of equations (9) can be written equivalently in *matrix* form. Thus, define

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n, \qquad \boldsymbol{X} \ := \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \in \mathbb{R}^{n \times (p+1)}.$$

The $n \times (p+1)$ matrix $\boldsymbol{X}$ is called the *design matrix*, or simply the X-matrix. For $i = 1, \ldots, n$, the i th row of $\boldsymbol{X}$ is the $(p+1)$-dimensional row vector $\boldsymbol{x}_i^\top$, the feature vector for the i th sample point (transposed). For $j = 0, \ldots, p$, the j th column of $\boldsymbol{X}$ is the n-dimensional column vector $\boldsymbol{X}_j := (x_{1j}, \ldots, x_{nj})^\top \in \mathbb{R}^n$, the vector of values of the jth feature for each of the n observations. Note that $\boldsymbol{X}_0 = (1, \ldots, 1)^\top =: \boldsymbol{1}_n$.

With this notation,

$$\sum_{i=1}^{n} x_{ir} x_{ij} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)_{rj}$$

and

$$\sum_{i=1}^{n} x_{ir} y_i = \left(\boldsymbol{X}^\top \boldsymbol{y}\right)_r,$$

so we can write (2.2) as

$$\sum_{j=0}^{p} \left(\boldsymbol{X}^\top \boldsymbol{X}\right)_{rj} b_j = \left(\boldsymbol{X}^\top \boldsymbol{y}\right)_r, \quad r = 0, \ldots, p$$

or, equivalently,

$$\boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{b} = \boldsymbol{X}^\top \boldsymbol{y}$$

Now, $\boldsymbol{X}^\top \boldsymbol{X}$ is a $(p+1) \times (p+1)$ matrix, and it is invertible if and only if the columns of $\boldsymbol{X}$ are linearly independent (prove this!). If $p+1 > n$ the columns of $\boldsymbol{X}$ are necessarily linearly dependent (because column rank $\leq \min(n, p+1) = n < p+1$). If $p+1 \leq n$, the columns of $\boldsymbol{X}$ are linearly independent if no feature vector $\boldsymbol{X}_j$ is a combination of the others, which is a reasonable assumption. If the columns of $\boldsymbol{X}$ are linearly dependent, $\boldsymbol{X}^\top \boldsymbol{X}$ is singular and (2.3) has infinitely many solutions, i.e., the LS coefficients $\hat{\boldsymbol{\beta}}$ are not unique (every solution for (2.3) is then said to be a LS solution). If the columns of $\boldsymbol{X}$ are linearly independent, which we will generally assume from now on, then (2.3) has a unique solution given by

$$\boxed{\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}. \qquad (10)}$$

Geometric interpretation of the LS solution. There is in fact a more direct (and more intuitive) way to derive the LS solution using *geometric* interpretation. We first recall some basic definitions and facts from linear algebra.

1. For a $n \times m$ matrix $\boldsymbol{A}$, the *image* is the linear subspace of $\mathbb{R}^n$ given by

$$\operatorname{Im}(\boldsymbol{A}) := \{\boldsymbol{A}\boldsymbol{v} : \boldsymbol{v} \in \mathbb{R}^m\}$$

Recall that, if $\boldsymbol{A}_j$ denotes the j-th column of $\boldsymbol{A}$, then $\boldsymbol{A}\boldsymbol{v} = \sum_j \boldsymbol{A}_j v_j$, so we have

$$\mathrm{Im}(\boldsymbol{A}) = \mathrm{sp}\left(\boldsymbol{A}_1, \ldots, \boldsymbol{A}_m\right) \subseteq \mathbb{R}^n,$$

i.e., the image is the linear subspace (of $\mathbb{R}^n$) *spanned* by the columns $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_m$.

2. The (standard) *inner product* of vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$ is $\boldsymbol{u}^\top \boldsymbol{v} = \boldsymbol{v}^\top \boldsymbol{u} = \sum_{i=1}^n u_i v_i$.

3. The *norm* of a vector $\boldsymbol{v} \in \mathbb{R}^n$ is $\|\boldsymbol{v}\| = \left(\boldsymbol{v}^\top \boldsymbol{v}\right)^{1/2} = \left(\sum_{i=1}^n v_i^2\right)^{1/2}$. The *Euclidean distance* between two vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$ is $\|\boldsymbol{u} - \boldsymbol{v}\|$. In $\mathbb{R}^2, \mathbb{R}^3$ this coincides with the usual geometric notion of a distance between two points.

4. For $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$, we say that $\boldsymbol{v}$ is *orthogonal* to $\boldsymbol{u}$, and denote $\boldsymbol{u} \perp \boldsymbol{v}$, if $\boldsymbol{u}^\top \boldsymbol{v} = 0$. In $\mathbb{R}^2, \mathbb{R}^3$ this coincides with the usual geometric notion of perpendicularity.

5. The *orthogonal complement* of a subspace $M \subseteq \mathbb{R}^n$ is the subspace

$$M^\perp := \left\{\boldsymbol{v} : \boldsymbol{v}^\top \boldsymbol{u} = 0 \quad \forall \boldsymbol{u} \in M\right\}$$

(Exercise: verify that $M^\perp$ is indeed a linear space.)

6. *Pythagorean theorem*: If $\boldsymbol{u} \perp \boldsymbol{v}$, then $\|\boldsymbol{v} + \boldsymbol{u}\|^2 = \|\boldsymbol{v}\|^2 + \|\boldsymbol{u}\|^2$.

We now derive the LS solution (10) from an alternative, geometric viewpoint. For any $\boldsymbol{b} = (b_0, \ldots, b_p)^\top \in \mathbb{R}^{p+1}$ we can write the objective function (8) in vector form as

$$Q(\boldsymbol{b}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2.$$

Thus, if $\hat{\boldsymbol{\beta}}$ is the minimizer of $Q(\boldsymbol{b})$ over all $\boldsymbol{b} \in \mathbb{R}^{p+1}$, then $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ minimizes the squared norm (equivalently, it minimizes the Euclidean distance) between $\boldsymbol{y}$ and $\boldsymbol{z}$ among all vectors $\boldsymbol{z} \in \mathrm{Im}(\boldsymbol{X})$. Then $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ must be the *projection* of $\boldsymbol{y}$ onto $\mathrm{colsp}(\boldsymbol{X})$ (this claim requires a proof, but it's intuitive geometrically). Now, a necessary and sufficient condition for $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ to be the projection of $\boldsymbol{y}$ onto $\mathrm{colsp}(\boldsymbol{X})$ is

$$\boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{0}, \tag{11}$$

which simply requires that the dot product between the residual $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}$ and each column of $\boldsymbol{X}$ is zero. Rearranging (11), we get

$$\boldsymbol{X}^\top \boldsymbol{y} = \boldsymbol{X}^\top \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

which, assuming again that $\boldsymbol{X}^\top \boldsymbol{X}$ is invertible, yields (10). The vector

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

is called the orthogonal projection of $\boldsymbol{y}$ onto $\mathrm{Im}(\boldsymbol{X})$, and the matrix

$$\boldsymbol{P}_{\boldsymbol{X}} := \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top$$

i.e., the matrix such that $\hat{\boldsymbol{y}} = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}$, is called the projection matrix of $\boldsymbol{X}$ (this is the matrix projecting any vector onto the linear space spanned by the columns of $\boldsymbol{X}$).

Remark. We are assuming, as before, that $\boldsymbol{X}^\top \boldsymbol{X}$ is invertible, although the projection matrix of $\boldsymbol{X}$ is well-defined also when the columns of $\boldsymbol{X}$ are linearly dependent; for example, this can be achieved by choosing a basis for $\mathrm{Im}(\boldsymbol{X})$ and writing (2.5) using the ”thinner” matrix instead of $\boldsymbol{X}$).

Linear algebra interlude: projection matrices and related results.

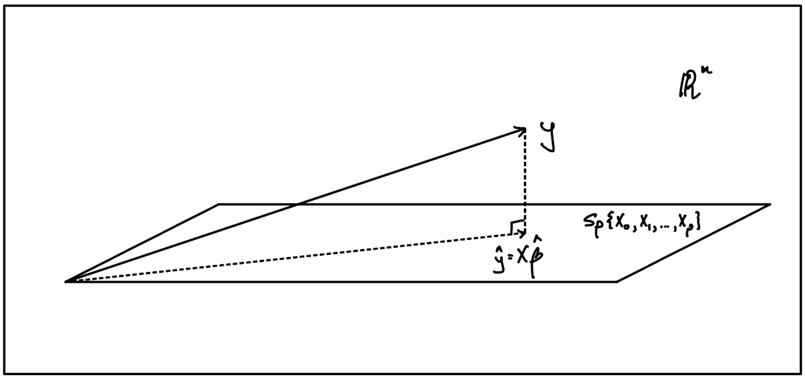


Figure 4: Geometric interpretation of the Least squares estimator

Proposition 4. *Let $\boldsymbol{X}$ be an $n \times m$ matrix and assume that it has linearly independent columns (i.e., full column rank; remember that this implies $m \leq n$). Then the projection matrix $\boldsymbol{P_X}$ has the following properties.*

1. *$\boldsymbol{P}_X$ is symmetric*
2. *$\boldsymbol{P}_X$ is idempotent, $\boldsymbol{P}_X^2 = \boldsymbol{P}_X$*
3. *$\boldsymbol{P}_X \boldsymbol{X} = \boldsymbol{X}$*
4. *$\boldsymbol{X}^\top \left(\boldsymbol{I} - \boldsymbol{P_X}\right) = \boldsymbol{0} \in \mathbb{R}^{m \times n}$*
5. *$\boldsymbol{P}_X v \in \mathrm{Im}(X)$ for all $v \in \mathbb{R}^n$*
6. *If $m = n$ and $\boldsymbol{X}$ is invertible, then $\boldsymbol{P_X} = \boldsymbol{I}$*
7. *$\left(\boldsymbol{I} - \boldsymbol{P_X}\right) \boldsymbol{v} \in \mathrm{Im}(\boldsymbol{X})^\perp$ for all $\boldsymbol{v} \in \mathbb{R}^n$*
8. *If $w \in \mathrm{lm}(\boldsymbol{X})$, then $\boldsymbol{P_X w} = \boldsymbol{w}$*
9. *If $\boldsymbol{w} \in \mathrm{Im}(\boldsymbol{X})^\perp$, then $\boldsymbol{P}_X \boldsymbol{w} = \boldsymbol{0}$*
10. *If $\boldsymbol{Z}$ is another $n \times m$ matrix s.t. $\mathrm{Im}(\boldsymbol{Z}) = \mathrm{Im}(\boldsymbol{X})$, then $\boldsymbol{P_Z} = \boldsymbol{P_X}$. This means that $\boldsymbol{P_X}$ depends on $\boldsymbol{X}$ only through the span of its columns. Hence, for an arbitrary linear space M, we can define the projection matrix $\boldsymbol{P}_M$ onto M (an explicit form for $\boldsymbol{P}_M$ can be obtained by taking any basis of M and stacking its elements as columns in a matrix $\boldsymbol{X}$, then forming $\boldsymbol{P_X} := \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \Big)$*
11. *If L and M are two subspaces with $L \subseteq M$, then $\boldsymbol{P}_M \boldsymbol{P}_L = \boldsymbol{P}_L \boldsymbol{P}_M = \boldsymbol{P}_L$.*

Proof.

1. $\boldsymbol{P_X}^\top = \left[\boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top\right]^\top = \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top$ where we used the fact $\left[\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\right]^\top = \left[\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^\top\right]^{-1} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}$.

2. $\boldsymbol{P}_{\boldsymbol{X}}^2 = \left[\boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\right]\left[\boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\right] = \boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top$

3. $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{X} = \boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{X}$

4. $\boldsymbol{X}^\top(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}}) = \left[(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{X}\right]^\top = \left[\boldsymbol{X}-\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{X}\right]^\top = \left[\boldsymbol{X}-\boldsymbol{X}\right]^\top = \boldsymbol{0}^\top \in \mathbb{R}^{n\times m} = \boldsymbol{0} \in \mathbb{R}^{m\times n}$, where we used fact #3

5. $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{v}$ for all $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{v} = \boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{v} = \boldsymbol{X}\left[\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{v}\right] \in \operatorname{Im}(\boldsymbol{X})$

6. If $\boldsymbol{P}_{\boldsymbol{X}}$ is (square and) invertible, $\left[\boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\right]^{-1} = \left[\boldsymbol{X}^\top\right]^{-1}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)\boldsymbol{X}^{-1} = \boldsymbol{I}_n$.

7. $(\boldsymbol{X}\boldsymbol{u})^\top(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{v} = \boldsymbol{u}^\top\underbrace{\boldsymbol{X}^\top(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})}_{\boldsymbol{0}\in\mathbb{R}^{m\times n}}\boldsymbol{v} = 0$

8. $\boldsymbol{P}_{\boldsymbol{X}}\underbrace{(\boldsymbol{X}\boldsymbol{u})}_{w} = (\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{X})\boldsymbol{u} = \underbrace{X\boldsymbol{u}}_{w}$

9. $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{w} = \boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\underbrace{\boldsymbol{X}^\top\boldsymbol{w}}_{=0} = \boldsymbol{0}$

10. Denote $\boldsymbol{X}_j, \boldsymbol{Z}_j$ for the j th columns of $\boldsymbol{X}, \boldsymbol{Z}$, respectively. Then $\boldsymbol{Z}_j = \boldsymbol{X}\boldsymbol{h}_j$ for some $\boldsymbol{h}_j \in \mathbb{R}^m$ because $\boldsymbol{Z}_j \in \operatorname{Im}(\boldsymbol{Z}) = \operatorname{Im}(\boldsymbol{X})$. Putting $\boldsymbol{H} := [\boldsymbol{h}_1\boldsymbol{h}_2\cdots\boldsymbol{h}_m] \in \mathbb{R}^{m\times m}$, we have $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{H}$. Further, if $\boldsymbol{w} \in \mathbb{R}^m$ is s.t. $\boldsymbol{H}\boldsymbol{w} = \boldsymbol{0}$, then $\mathbf{Z}\boldsymbol{w} = \boldsymbol{X}\boldsymbol{H}\boldsymbol{w} = \boldsymbol{0}$, which implies $\boldsymbol{w} = \boldsymbol{0}$ because $\boldsymbol{Z}$ was assumed to have full column rank. Hence, $\boldsymbol{H}$ is invertible. Then

$$\boldsymbol{P}_{\boldsymbol{Z}} = \boldsymbol{Z}\left(\boldsymbol{Z}^\top\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^\top = \boldsymbol{X}\boldsymbol{H}\left[(\boldsymbol{X}\boldsymbol{H})^\top\boldsymbol{X}\boldsymbol{H}\right]^{-1}(\boldsymbol{X}\boldsymbol{H})^\top =$$
$$= \boldsymbol{X}\boldsymbol{H}\boldsymbol{H}^{-1}\boldsymbol{X}^{-1}\left[(\boldsymbol{X}\boldsymbol{H})^\top\right]^{-1}(\boldsymbol{X}\boldsymbol{H})^\top = \boldsymbol{X}\left[\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\right]\boldsymbol{X}^\top = \boldsymbol{P}_{\boldsymbol{X}}$$

11. Let $\boldsymbol{v}$ be any vector. Then $\boldsymbol{P}_M\underbrace{\boldsymbol{P}_L\boldsymbol{v}}_{\in M} \overset{8}{=} \boldsymbol{P}_L\boldsymbol{v}$, implying $\boldsymbol{P}_M\boldsymbol{P}_L = \boldsymbol{P}_L$. Transposing both sides and using the fact that $\boldsymbol{P}_M, \boldsymbol{P}_L$ are both symmetric, we obtain also $\boldsymbol{P}_L\boldsymbol{P}_M = \boldsymbol{P}_L$.

□

Additionally, we have the following results related to projection matrices.

Proposition 5. *Let M be a subspace of $\mathbb{R}^n$ with $\dim(M) = m \le n$. Then any vector $\boldsymbol{v} \in \mathbb{R}^n$ can be uniquely represented as $\boldsymbol{v} = \boldsymbol{w} + \boldsymbol{z}$ where $\boldsymbol{w} \in M$ and $z \in M^\perp$, the orthogonal complement of M. Moreover, in this representation, $\boldsymbol{w} = \boldsymbol{P}_M\boldsymbol{v}$, the projection of $\boldsymbol{v}$ onto M, and satisfies $\boldsymbol{w} = \underset{\boldsymbol{u}\in M}{\arg\min}\|\boldsymbol{v}-\boldsymbol{u}\|^2$.*

Proof. Taking $\boldsymbol{w} = \boldsymbol{P}_M\boldsymbol{v}, \boldsymbol{z} = \boldsymbol{v} - \boldsymbol{P}_M\boldsymbol{v}$, we have $\boldsymbol{w} \in M, z \in M^\perp$ by the properties above, and $\boldsymbol{v} = \boldsymbol{w} + \boldsymbol{z}$. We show that this representation is unique. Thus, suppose $\boldsymbol{v} = \boldsymbol{w}_1 + \boldsymbol{z}_1, \boldsymbol{v} = \boldsymbol{w}_2 + \boldsymbol{z}_2$ for $\boldsymbol{w}_1, \boldsymbol{w}_2 \in M, \boldsymbol{z}_1, \boldsymbol{z}_2 \in M^\perp$. Then $\boldsymbol{0} = (\boldsymbol{w}_1 + \boldsymbol{z}_1) - (\boldsymbol{w}_2 + \boldsymbol{z}_2) = (\boldsymbol{w}_1 - \boldsymbol{w}_2) + (\boldsymbol{z}_1 - \boldsymbol{z}_2)$, and, furthermore, $(\boldsymbol{w}_1 - \boldsymbol{w}_2) \perp (\boldsymbol{z}_1 - \boldsymbol{z}_2)$ because $(\boldsymbol{w}_1 - \boldsymbol{w}_2) \in M, (\boldsymbol{z}_1 - \boldsymbol{z}_2) \in M^\perp$ (remember that each of $M, M^\perp$ is a subspace so closed under addition/subtraction). Therefore,

$$0 = \|\boldsymbol{0}\|^2 = \|\,(\boldsymbol{w}_1 - \boldsymbol{w}_2) + (\boldsymbol{z}_1 - \boldsymbol{z}_2)\,\|^2 = \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|^2 + \|\boldsymbol{z}_1 - \boldsymbol{z}_2\|^2 \Rightarrow (\boldsymbol{w}_1 - \boldsymbol{w}_2) = \boldsymbol{0}, (\boldsymbol{z}_1 - \boldsymbol{z}_2) = \boldsymbol{0}.$$

It remains to show that $\boldsymbol{w}$ minimizes the Euclidean distance from $\boldsymbol{v}$ to M. Take any $\boldsymbol{u} \in M$. Then

$$\|\boldsymbol{v}-\boldsymbol{u}\|^2 = \|(\boldsymbol{w}+\boldsymbol{z})-\boldsymbol{u}\|^2 = \|\boldsymbol{w}-\boldsymbol{u}+\boldsymbol{z}\|^2 = \|\boldsymbol{w}-\boldsymbol{u}\|^2 + \|\boldsymbol{z}\|^2 \geq \|\boldsymbol{z}\|^2,$$

and on the other hand, for $\boldsymbol{u} = \boldsymbol{w} = \boldsymbol{P}_M \boldsymbol{v}$ we have $\|\boldsymbol{v}-\boldsymbol{u}\|^2 = \|\boldsymbol{z}\|^2$. □

Proposition 6. *We have*

1. $\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}} = P_{Im(\boldsymbol{X})^\perp}$
2. *if* L *and* M *are two subspaces of* $\mathbb{R}^n$ *with* $L \subseteq M$*, then* $\boldsymbol{P}_M - \boldsymbol{P}_L = \boldsymbol{P}_{M \cap L^\perp}$

Proof. (a) this part follows from the uniqueness of the representation in Proposition 1. (b) this also follows from the uniqueness of the representation in Proposition 1, taking the original space to be M (I.e., the space to which v belongs) and the subspace (what's denoted M in Proposition 1) to be L. □

Proposition 7. *Let* $\boldsymbol{Q}$ *be an* $n \times n$ *matrix of rank* $m \leq n$ *which is symmetric and idempotent,* $\boldsymbol{Q}^\top = \boldsymbol{Q}, \boldsymbol{Q}^2 = \boldsymbol{Q}$*. Then* $\boldsymbol{Q} = \boldsymbol{P}_M$ *where* $M := \mathrm{Im}(\boldsymbol{Q})$.

Proof. Exercise. □

Diagonalizability and positive-semidefiniteness of a projection matrix. First, we recall some facts:

1. A matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is diagonalizable (over $\mathbb{R}$) if there is an invertible matrix $\boldsymbol{P} \in \mathbb{R}^{n\times n}$ and a diagonal matrix $\boldsymbol{D}$ such that $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1}$.

   Remark: $\boldsymbol{P}^{-1}$ is the transition matrix from the standard basis of $\mathbb{R}^n$ into the basis in the columns of $\boldsymbol{P}$).

2. If $\boldsymbol{A} \in \mathbb{R}$ is *symmetric*, then there is an *orthogonal* matrix $\boldsymbol{U} \in \mathbb{R}^{n\times n}$ and a diagonal matrix $\boldsymbol{D}$ such that $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^{-1} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\top$.

   Remark: (i) a square matrix $\boldsymbol{U}$ is *orthogonal* if $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}$ (this implies $\boldsymbol{U}^\top = \boldsymbol{U}^{-1}$, because for any two square matrices $\boldsymbol{A}, \boldsymbol{B}$, we have $\boldsymbol{A}\boldsymbol{B} = \boldsymbol{I} \Rightarrow \boldsymbol{B}\boldsymbol{A} = \boldsymbol{I}$; prove this!). (ii) consistent with the above, you should have seen a proof in your linear algebra course that if $\boldsymbol{A}$ is a (real) symmetric matrix, then it is diagonalizable, and all of its eigenvalues are real.

3. A symmetric matrix $\boldsymbol{A}$ is *positive semidefinite* if all of its eigenvalues are nonnegative. It is called *positive definite* if all of its eigenvalues are positive.

   Remark: this is equivalent to saying that the *quadratic form* given by $\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x}$ is *nonnegative* for all vectors $\boldsymbol{x}$ (positive semidefinite) or that $\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x}$ is *positive* (positive definite).

4. Let $\boldsymbol{A}$ be a positive semidefinite matrix. Then:

   (i) there is $\boldsymbol{B}$ square such that $\boldsymbol{B}^2 = \boldsymbol{A}$, and in fact this representation is unique: take $\boldsymbol{B} = \boldsymbol{U}\boldsymbol{D}^{1/2}\boldsymbol{U}^\top$, where $\boldsymbol{D}^{1/2} := \mathrm{diag}(d_1, ..., d_n)$, so $\boldsymbol{B}^2 = \boldsymbol{U}\boldsymbol{D}^{1/2}\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{D}^{1/2}\boldsymbol{U}^\top = \boldsymbol{A}$ (this $\boldsymbol{B}$ is often called the square root of $\boldsymbol{A}$).

   (ii) there is $\boldsymbol{B}$ square such that $\boldsymbol{B}\boldsymbol{B}^\top = \boldsymbol{A}$: take $\boldsymbol{B} = \boldsymbol{U}\boldsymbol{D}^{/12}$, then $\boldsymbol{B}^2 = \boldsymbol{U}\boldsymbol{D}^{/12}(\boldsymbol{U}\boldsymbol{D}^{/12})^\top = \boldsymbol{A}$

Now suppose that $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ is a projection matrix onto a subspace $M \subseteq \mathbb{R}^n$ of dimension $\dim(M) = m$, then $\boldsymbol{Q}$ is positive semidefinite, and there is a representation

$$\boldsymbol{Q} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\top, \qquad \boldsymbol{D} = \mathrm{diag}(\underbrace{1, ..., 1}_{m}, \underbrace{0, ..., 0}_{n-m}),$$

where $\boldsymbol{U}$ is orthogonal, such that the first m columns of $\boldsymbol{U}$ are an orthonormal basis of M, and the last $n-m$ columns of $\boldsymbol{U}$ are an orthonormal basis of $M^{\perp}$.
Remark: consistent with the previous item, note that $\boldsymbol{Q} = \boldsymbol{B}\boldsymbol{B}^{\top}$ for $\boldsymbol{B} = \boldsymbol{U}\boldsymbol{D}^{1/2} = \tilde{\boldsymbol{U}} := [\boldsymbol{U}_1, ..., \boldsymbol{U}_m]$.

4 Statistical modeling

So far we have been working with arbitrary data points $(1, x_{i1}, \ldots, x_{ip}, y_i)\,, i = 1, \ldots, n$. In other words, the $n \times (p+1)$ matrix $\boldsymbol{X}$ and the vector $\boldsymbol{y}$ consisted of any fixed numbers, except that we assumed that $\boldsymbol{X}$ has full column rank (i.e., the columns of $\boldsymbol{X}$ are linearly independent). If we return to the motivation for the regression problem, recall that we are ultimately interested in learning something about the relationship between the covariate vector $(1, X_{i1}, \ldots, X_{ip})$ and the response Y in some *population*, rather than in the particular dataset (sample) we happened to observe. In other words, we want the least squares regression line that we fit on the sample to *estimate* a ‘theoretical’ (or ‘true’) regression line for some target population. This will be possible if we assume that the observations are a *random sample* from the target population. Hence, we will now assume that the data points $(1, x_{i1}, \ldots, x_{ip}, y_i)$ are *i.i.d.* (independent, identically distributed) realizations of a *random vector*

$$(1, X_1, \ldots, X_p, Y) \sim P.$$

Now, we can always write

$$Y_i = \underbrace{\mathbb{E}\left(Y_i \mid X_{i1}, \ldots, X_{ip}\right)}_{f(X_{i1}, \ldots, X_{ip})} + \underbrace{\left(Y_i - \mathbb{E}\left(Y_i \mid X_{i1}, \ldots, X_{ip}\right)\right)}_{\epsilon_i},$$

where $f\left(X_{i1}, \ldots, X_{ip}\right)$, the conditional expectation of Y given $X_1, \ldots, X_p$, is the *systematic* part, and ϵ_i, the deviation of Y from its conditional expectation given the X_{ij}'s, $j = 1, \ldots, p$, is the *error* part. Note that

$$\begin{aligned}&\mathbb{E}\left(\epsilon_i \mid X_{i1}, \ldots, X_{ip}\right) = \mathbb{E}\left[Y_i - \mathbb{E}\left(Y_i \mid X_{i1}, \ldots, X_{ip}\right) \mid X_{i1}, \ldots, X_{ip}\right] = \\ &= \mathbb{E}\left(Y_i \mid X_{i1}, \ldots, X_{ip}\right) - \mathbb{E}\left[\mathbb{E}\left(Y_i \mid X_{i1}, \ldots, X_{ip}\right)\right] = \mathbb{E}\left(Y_i \mid X_{i1}, \ldots, X_{ip}\right) - \mathbb{E}\left(Y_i \mid X_{i1}, \ldots, X_{ip}\right) = 0\end{aligned}$$

i.e., the random variable ϵ has mean zero conditionally on $X_{i1}, \ldots, X_{ip}$ (by the way, note that this implies $\mathbb{E}\epsilon_i = 0$ unconditionally).

Assumption 1 (linearity). In general, the function $f\left(X_{i1}, \ldots, X_{ip}\right)$ can be any function of $\left(X_{i1}, \ldots, X_{ip}\right)$. From now on, we make the assumption that this is a *linear* (in fact, affine) function of $\left(x_{i1}, \ldots, x_{ip}\right)$, i.e.,

$$f\left(1, X_{i1}, \ldots, X_{ip}\right) = \sum_{j=0}^{p} \beta_j X_{ij}.$$

Assumption 2 (homoscedastic (equal variance) and uncorrelated errors). In addition to the linearity assumption on the conditional *mean* of Y, we will make an assumption on the conditional *variance* of the error ϵ_i. Specifically, we assume $V(\epsilon_i | X_{i1}, ..., X_{ip}) = \sigma^2$, i.e., that the errors have equal variance; this is referred to as *homoscedastic* errors. Moreover, we assume that, conditionally on the explanatory variables, the errors are uncorrelated, i.e., $Cov(\epsilon_i, \epsilon_{i'}) = 0$ if $i' \neq i$.

We can summarize all of the above as follows. The general linear model is given by

$$Y_i = \sum_{j=0}^{p} \beta_j X_{ij} + \epsilon_{ij}, \qquad \mathbb{E}[\epsilon_i | \text{all } X_{ij}\text{'s}] = 0, \qquad \text{Cov}(\epsilon_k, \epsilon_l | \text{all } X_{ij}\text{'s}) = \begin{cases} \sigma^2, & k = l \\ 0, & i \neq j \end{cases}$$

Actually, throughout the course we will generally treat the X_{ij}'s as *fixed* (nonrandom). In that case, the above is equivalent to

$$Y_i = \sum_{j=0}^{p} \beta_j x_{ij} + \epsilon_{ij}, \qquad \mathbb{E}[\epsilon_i] = 0, \qquad \text{Cov}(\epsilon_k, \epsilon_l) = \begin{cases} \sigma^2, & k = l \\ 0, & i \neq j \end{cases} \tag{12}$$

Moments of random vectors, algebra of covariance. We are headed toward providing statistical inference for $\boldsymbol{\beta}$ (and σ^2) under the model (12). As this will involve working with random vectors, we begin with some general definitions.

A random vector is a vector $Z = (Z_1, \ldots, Z_n)^\top$ whose components Z_i are random variables with some joint distribution. A random matrix is a matrix

$$\boldsymbol{Z} = \begin{bmatrix} Z_{11} & Z_{12} & \cdots & Z_{1m} \\ Z_{21} & Z_{22} & \cdots & Z_{2m} \\ \vdots & & & \\ Z_{n1} & Z_{n2} & \cdots & Z_{nm} \end{bmatrix}$$

whose components Z_{ij} are random variables with some joint distribution.

Definition 2. The expectation of a random $n \times m$ matrix **Z** is defined as the $n \times m$ matrix $\mathbb{E}\boldsymbol{Z}$ whose (i, j)-th entry is

$$[\mathbb{E}\boldsymbol{Z}]_{ij} = \mathbb{E}Z_{ij}$$

In other words,

$$\mathbb{E}\boldsymbol{Z} = \begin{bmatrix} \mathbb{E}Z_{11} & \mathbb{E}Z_{12} & \cdots & \mathbb{E}Z_{1m} \\ \mathbb{E}Z_{21} & \mathbb{E}Z_{22} & \cdots & \mathbb{E}Z_{2m} \\ \vdots & & & \\ \mathbb{E}Z_{n1} & \mathbb{E}Z_{n2} & \cdots & \mathbb{E}Z_{nm} \end{bmatrix},$$

and, as a special case when $m = 1$, for a random vector $\boldsymbol{Z} = (Z_1, \ldots, Z_n)^\top$,

$$\mathbb{E}\boldsymbol{Z} = \begin{bmatrix} \mathbb{E}Z_1 \\ \mathbb{E}Z_2 \\ \vdots \\ \mathbb{E}Z_n \end{bmatrix}$$

Properties. $\boldsymbol{Z}, \boldsymbol{W}$ random matrices. For any fixed matrices $\boldsymbol{A}, \boldsymbol{B}$ of compatible dimensions, we have:

1. $\mathbb{E}[\boldsymbol{Z} + \boldsymbol{W}] = \mathbb{E}[\boldsymbol{Z}] + \mathbb{E}[\boldsymbol{W}]$

   Proof. $[\mathbb{E}(\boldsymbol{Z} + \boldsymbol{W})]_{ij} \overset{(1)}{=} \mathbb{E}\left([\boldsymbol{Z} + \boldsymbol{W}]_{ij}\right) \overset{(2)}{=} \mathbb{E}\left(\boldsymbol{Z}_{ij} + \boldsymbol{W}_{ij}\right) \overset{(3)}{=} \mathbb{E}\boldsymbol{Z}_{ij} + \mathbb{E}\boldsymbol{W}_{ij} \overset{(4)}{=} [\mathbb{E}\boldsymbol{Z}]_{ij} + [\mathbb{E}\boldsymbol{W}]_{ij}$ where (1) is due to the definition of the expectation of a matrix; (2) is due to the rule of addition of two matrices (the (i, j)-th element of the sum is the sum of the (i, j)-th elements); (3) is due to linearity of expectation for (univariate) random variables (this is the main step of the proof); and (4) is again due to the definition of the expectation of a matrix. □

2. $\mathbb{E}[\boldsymbol{AZB}] = \boldsymbol{A}\mathbb{E}[\boldsymbol{Z}]\boldsymbol{B}$

Proof. First,

$$[\mathbb{E}(\boldsymbol{A}\boldsymbol{Z})]_{ij} = \mathbb{E}\left(\sum_r \boldsymbol{A}_{ir}\boldsymbol{Z}_{rj}\right) = \sum_r \boldsymbol{A}_{ir}\mathbb{E}\boldsymbol{Z}_{rj} = [\boldsymbol{A}\mathbb{E}\boldsymbol{Z}]_{ij} \Rightarrow \mathbb{E}(\boldsymbol{A}\boldsymbol{Z}) = \boldsymbol{A}\mathbb{E}\boldsymbol{Z}$$

A similar argument yields

$$\mathbb{E}(\boldsymbol{Z}\boldsymbol{B}) = (\mathbb{E}\mathbf{Z})\boldsymbol{B}$$

Finally,

$$\mathbb{E}[\boldsymbol{A}\boldsymbol{Z}] = \mathbb{E}[\boldsymbol{A}(\boldsymbol{Z}\boldsymbol{B})] \overset{(1)}{=} \boldsymbol{A}\mathbb{E}[\boldsymbol{Z}\boldsymbol{B}] \overset{(2)}{=} \boldsymbol{A}\mathbb{E}[\boldsymbol{Z}]\boldsymbol{B}$$

□

3. $\mathbb{E}[\boldsymbol{A}\boldsymbol{U} + \boldsymbol{C}] = \boldsymbol{A}\mathbb{E}[\boldsymbol{U}] + \boldsymbol{C}$ (from $1 + 2$)

   Proof. Exercise. □

Reminder. For two random variables Z, W, recall that the covariance of Z and W is

$$\operatorname{Cov}(Z, W) := \mathbb{E}\left(Z - \mu_Z\right)\left(W - \mu_W\right)$$

where $\mu_Z := \mathbb{E}Z, \mu_W := \mathbb{E}W$.

Using linearity of the expectation, we get the identity

$$\operatorname{Cov}(Z, W) = \mathbb{E}[ZW] - \mu_Z\mu_W.$$

In the special case $W = Z$, by the definition we have

$$\operatorname{Cov}(Z, Z) = V(Z) = \mathbb{E}\left(Z - \mu_Z\right)^2$$

As properties of the covariance of two random variables, recall that For any fixed $a \in \mathbb{R}$, we have:

1. $\operatorname{Cov}(W, Z) = \operatorname{Cov}(Z, W)$
2. $\operatorname{Cov}(aZ + R, W) = a\operatorname{Cov}(Z, W) + \operatorname{Cov}(R, W)$

Definition 3. The *covariance matrix* of a random vector $\boldsymbol{Z} \in \mathbb{R}^n$ with a random vector $\boldsymbol{W} \in \mathbb{R}^m$ is denoted $\operatorname{cov}(\boldsymbol{Z}, \boldsymbol{W})$, and defined to be the $n \times m$ matrix whose (i, j)-th entry is

$$[\operatorname{cov}(\boldsymbol{Z}, \boldsymbol{W})]_{ij} := \operatorname{Cov}\left(Z_i, W_j\right)$$

In the special case where $\boldsymbol{W} = \boldsymbol{Z}$, we denote $\operatorname{cov}(\boldsymbol{Z}) := \operatorname{cov}(\boldsymbol{Z}, \boldsymbol{Z})$, and by the above definition,

$$[\operatorname{cov}(\boldsymbol{Z})]_{ij} := [\operatorname{cov}(\mathbf{Z}, \mathbf{Z})]_{ij} = \operatorname{Cov}\left(Z_i, Z_j\right)$$

Equivalently, if we denote $\mu_Z := \mathbb{E}Z, \mu_W := \mathbb{E}\boldsymbol{W}$, (3.7) and (3.8) can be expressed in matrix notation as

$$\boxed{\operatorname{cov}(\mathbf{Z}, \boldsymbol{W}) := \mathbb{E}\left[\left(\boldsymbol{Z} - \mu_{\boldsymbol{Z}}\right)\left(\boldsymbol{W} - \mu_{\boldsymbol{W}}\right)^\top\right], \qquad (13)}$$

which is an $n \times m$ matrix, and

$$\boxed{\operatorname{cov}(\mathbf{Z}) := \operatorname{cov}(\mathbf{Z}, \mathbf{Z}) = \mathbb{E}\left[(\mathbf{Z} - \mu_{\mathbf{Z}})(\mathbf{Z} - \mu_{\mathbf{Z}})^{\top}\right], \qquad (14)}$$

which is an $n \times n$ matrix. Using the identity (3.8) for (univariate) random variables Z, W, and the entry-wise definition of the expectation of a matrix, we also have the multivariate counterparts,

$$\operatorname{cov}(\boldsymbol{Z}, \boldsymbol{W}) = \mathbb{E}\left[\boldsymbol{Z}\boldsymbol{W}^{\top}\right] - \mu_{\boldsymbol{Z}}\mu_{\boldsymbol{W}}^{\top}$$

and

$$\operatorname{cov}(\mathbf{Z}) = \mathbb{E}\left[\mathbf{Z}\mathbf{Z}^{\top}\right] - \mu_{\mathbf{Z}}\mu_{\mathbf{Z}}^{\top}$$

Properties of covariance matrix. $\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{R}$ random vectors; $\boldsymbol{a}$ fixed vector. Then we have the following properties :

1. $\operatorname{cov}(\boldsymbol{Z}, \boldsymbol{W}) = \operatorname{cov}(\boldsymbol{W}, \boldsymbol{Z})^{\top}$
2. $\operatorname{cov}(\boldsymbol{Z} + \boldsymbol{R}, \boldsymbol{W}) = \operatorname{cov}(\boldsymbol{Z}, \boldsymbol{W}) + \operatorname{cov}(\boldsymbol{R}, \boldsymbol{W})$
3. $\operatorname{cov}(\boldsymbol{A}\boldsymbol{Z}, \boldsymbol{B}\boldsymbol{W}) = A \operatorname{cov}(\boldsymbol{Z}, \boldsymbol{W})\boldsymbol{B}^{\top}$
4. $\operatorname{cov}(\boldsymbol{A}\boldsymbol{Z}) = \boldsymbol{A} \operatorname{cov}(\boldsymbol{Z})\boldsymbol{A}^{\top}$ (from 3)
5. $V\left(\boldsymbol{a}^{\top}\boldsymbol{Z}\right) = \boldsymbol{a}^{\top} \operatorname{cov}(\boldsymbol{Z})\boldsymbol{a}$ (from 4)
6. $\operatorname{cov}(\boldsymbol{Z})$ is a nonnegative definite matrix (from 5)

Now return to the linear model. By the definition of the covariance matrix and expectation, (12) is equivalent to

$$\boxed{\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \mathbb{E}\boldsymbol{\epsilon} = \mathbf{0}, \quad \operatorname{cov}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}_n, \qquad (15)}$$

where $\boldsymbol{X}$ is a *fixed* (nonrandom) $n \times p + 1$ matrix, and $\boldsymbol{\beta},\ \sigma^2$ unknown.

The vector $\boldsymbol{\epsilon}$ is called the *errors*. We will sometimes write $\boldsymbol{\epsilon} \sim \left(\mathbf{0}, \sigma^2\boldsymbol{I}_n\right)$ as shorthand for $\mathbb{E}\boldsymbol{\epsilon} = \mathbf{0}, \quad \operatorname{cov}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}_n$ (without any further assumptions on the distribution). With this notation, (16) can be written even more compactly as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \left(\mathbf{0}, \sigma^2\boldsymbol{I}_n\right)$$

5 Inference under the linear model

Recall that the LS estimator is given by

$$\hat{\boldsymbol{\beta}} = \boldsymbol{A}\boldsymbol{Y}, \quad \boldsymbol{A} := \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top.$$

The corresponding vectors of *fitted (predicted) values* and *residuals* are given, respectively, by

$$\hat{\mathbf{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}, \qquad \boldsymbol{e} = \mathbf{Y} - \hat{\mathbf{Y}},$$

and we have $\boldsymbol{e} \perp \hat{\mathbf{Y}}$ (this was a *defining* property of the LS solution). Remember that all of this holds regardless of the assumptions of the linear model, and in fact requires no statistical assumptions at all.

Now, assume the linear model (16). Then the vector $\mathbf{Y}$ becomes a *random* vector, with distribution generally depending on the unknown *parameters* $\boldsymbol{\beta}$ and σ^2. Same goes for $\hat{\boldsymbol{\beta}}$, which now has a meaning as an *esimator* of the unknown parameter $\boldsymbol{\beta}$, the true (unknown) coefficient vector. Let us calculate its mean and covariance matrix. We have

$$\mathbb{E}[\boldsymbol{Y}] = \mathbb{E}[\boldsymbol{X}\boldsymbol{\beta}] + \mathbb{E}[\boldsymbol{\epsilon}] = \boldsymbol{X}\boldsymbol{\beta}$$

and

$$\operatorname{cov}(\boldsymbol{Y}) = \operatorname{cov}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \operatorname{cov}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_n.$$

Also, we can now calculate

$$\mathbb{E}[\hat{\boldsymbol{\beta}}] = \mathbb{E}[\boldsymbol{A}\boldsymbol{Y}] = \boldsymbol{A}\mathbb{E}\boldsymbol{Y} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{\beta},$$

and

$$\begin{aligned}\operatorname{cov}(\hat{\boldsymbol{\beta}}) = \operatorname{cov}(\boldsymbol{A}\boldsymbol{Y}) = \boldsymbol{A}\operatorname{cov}(\boldsymbol{Y})\boldsymbol{A}^\top = \boldsymbol{A}\left(\sigma^2 \boldsymbol{I}\right)\boldsymbol{A}^\top &= \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \left[\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top\right]^\top \\ &= \sigma^2 \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{X} \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} = \sigma^2 \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}.\end{aligned}$$

Hence

$$\boxed{\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}, \qquad \operatorname{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{X} \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} = \sigma^2 \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}. \qquad (16)}$$

Estimating the noise level σ^2. Intuitively, it makes sense to use the residual vector $\boldsymbol{e}$ to estimate σ^2. Define

$$\hat{\sigma}^2 := \frac{1}{n-p-1}\|\boldsymbol{e}\|^2 = \frac{1}{n-p-1}\sum_{i=1}^n e_i^2.$$

Proposition 8. *$\hat{\sigma}^2$ defined above is an unbiased estimator of σ^2.*

Proof. Let $M = \operatorname{Im}(\boldsymbol{X})$. Denote $\boldsymbol{Q} := \boldsymbol{I} - \boldsymbol{P}$ for the projection matrix onto $M^\perp$. Then

$$\boldsymbol{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \boldsymbol{P}\mathbf{Y} = (\boldsymbol{I} - \boldsymbol{P})\mathbf{Y} = \boldsymbol{Q}\boldsymbol{Y},$$

and note that we also have

$$\boldsymbol{Q}\mathbf{Y} = \boldsymbol{Q}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \boldsymbol{Q}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Q}\boldsymbol{\epsilon} = \boldsymbol{Q}\boldsymbol{\epsilon},$$

because $\boldsymbol{QX} = \mathbf{0}$ (every column of $\boldsymbol{X}$ is in $\mathrm{Im}(\boldsymbol{X})$), so we can conclude $\boldsymbol{e} = \boldsymbol{Q\epsilon}$. Therefore,

$$\|\boldsymbol{e}\|^2 = \|\boldsymbol{Q\epsilon}\|^2 = \boldsymbol{\epsilon}^\top \boldsymbol{Q}^\top \boldsymbol{Q\epsilon} = \boldsymbol{\epsilon}^\top \boldsymbol{Q\epsilon} = \sum_i \sum_j Q_{ij}\epsilon_i\epsilon_j,$$

since $\boldsymbol{Q}$ is symmetric and idempotent (projection matrix), and so

$$\mathbb{E}\|\boldsymbol{e}\|^2 = \mathbb{E}\|\boldsymbol{Q\epsilon}\|^2 = \mathbb{E}\left[\boldsymbol{\epsilon}^\top \boldsymbol{Q}^\top \boldsymbol{Q\epsilon}\right] = \mathbb{E}\left[\boldsymbol{\epsilon}^\top \boldsymbol{Q\epsilon}\right] = \mathbb{E}\left[\sum_i \sum_j Q_{ij}\epsilon_i\epsilon_j\right] = \sum_i \sum_j Q_{ij}\mathbb{E}\left[\epsilon_i\epsilon_j\right]. \quad (17)$$

Also, because $\mathbb{E}\epsilon_i = 0$ by assumption, we have

$$\mathbb{E}\left[\epsilon_i\epsilon_j\right] = \mathrm{Cov}\left(\epsilon_i, \epsilon_j\right) = \left\{ \begin{array}{ll} \sigma^2, & i = j \\ 0, & \text{otherwise} \end{array} \right. .$$

Continuing from (17),

$$\mathbb{E}\|\boldsymbol{e}\|^2 = \sum_i \sum_j Q_{ij}\mathbb{E}\left[\epsilon_i\epsilon_j\right] = \sum_i Q_{ii}\mathbb{E}\left[\epsilon_i^2\right] = \sum_i Q_{ii}V\left(\epsilon_i\right) = \sigma^2 \sum_i Q_{ii} = \sigma^2\,\mathrm{tr}(\boldsymbol{Q}). \quad (18)$$

Now, we known that, as a projection matrix, $\boldsymbol{Q}$ is similar to a diagonal matrix $\boldsymbol{D}$ whose diagonal has $\dim(\boldsymbol{Q}) = n - (p+1) = n - p - 1$ entries equal to 1 , and the rest $p + 1$ entries are zero. But the trace is preserved under the similarity relation, meaning that $\mathrm{tr}(Q) = \mathrm{tr}(\boldsymbol{D}) = n - p - 1$. Continuing from (18), we get

$$\mathbb{E}\|\boldsymbol{e}\|^2 = \sigma^2\,\mathrm{tr}(\boldsymbol{Q}) = \sigma^2(n - p - 1)$$

implying

$$\mathbb{E}\left[\frac{1}{n-p-1}\|\boldsymbol{e}\|^2\right] = \sigma^2.$$

□

We can give an alternative, shorter proof using the following general result.

Lemma 1. *For any random vector $\boldsymbol{Z}$ it holds that*

$$\mathbb{E}\|\boldsymbol{Z}\|^2 = \mathrm{tr}\left(\mathbb{E}\left[\boldsymbol{Z}\boldsymbol{Z}^\top\right]\right) = \mathrm{tr}\left(\mathrm{cov}(\boldsymbol{Z}) + \boldsymbol{\mu_Z}\boldsymbol{\mu_Z}^\top\right),$$

where $\boldsymbol{\mu_Z} := \mathbb{E}\boldsymbol{Z}$. As a special case, if $\mu_{\boldsymbol{Z}} = \mathbf{0}$, then $\mathbb{E}\|\boldsymbol{Z}\|^2 = \mathrm{tr}(\mathrm{cov}(\boldsymbol{Z}))$.

Proof of lemma. . We have

$$\mathbb{E}\|\boldsymbol{Z}\|^2 = \mathbb{E}\left(\boldsymbol{Z}^\top\boldsymbol{Z}\right) \overset{(a)}{=} \mathbb{E}\left[\mathrm{tr}\left(\boldsymbol{Z}^\top\boldsymbol{Z}\right)\right] \overset{(b)}{=} \mathbb{E}\left[\mathrm{tr}\left(\boldsymbol{Z}\boldsymbol{Z}^\top\right)\right] \overset{(c)}{=} \mathrm{tr}\left(\mathbb{E}\left[\boldsymbol{Z}\boldsymbol{Z}^\top\right]\right) \overset{(d)}{=} \mathrm{tr}\left(\mathrm{cov}(\boldsymbol{Z}) + \boldsymbol{\mu_Z}\boldsymbol{\mu_Z}^\top\right)$$

where (a) is because $\boldsymbol{Z}^\top\boldsymbol{Z}$ is a scalar; (b) is due to the general identity $\mathrm{tr}(\boldsymbol{AB}) = \mathrm{tr}(\boldsymbol{BA})$; (c) is due to the definition of the expectation of a random matrix, and the linearity of the expectation; and (d) is due to the general identity $\mathrm{cov}(\boldsymbol{Z}) = \mathbb{E}\left[\mathbf{ZZ}^\top\right] - \boldsymbol{\mu}_Z\boldsymbol{\mu}_{\mathbf{Z}}^\top$. □

Alternative proof of Proposition 8. . Taking $\boldsymbol{Z} = \boldsymbol{e}$ in the lemma above, we have

$$\mathbb{E}\|\boldsymbol{e}\|^2 = \operatorname{tr}\left(\operatorname{cov}(\boldsymbol{e}) + \boldsymbol{\mu_e}\boldsymbol{\mu_e}^\top\right) \overset{(a)}{=} \operatorname{tr}(\operatorname{cov}(\boldsymbol{e})) =$$
$$\operatorname{tr}(\operatorname{cov}(\boldsymbol{QY})) = \operatorname{tr}\left(\boldsymbol{Q}\operatorname{cov}(\boldsymbol{Y})\boldsymbol{Q}^\top\right) = \operatorname{tr}\left(\boldsymbol{Q}\left[\sigma^2\boldsymbol{I}\right]\boldsymbol{Q}^\top\right) =$$
$$\sigma^2\operatorname{tr}\left(\boldsymbol{QQ}^\top\right) = \sigma^2\operatorname{tr}(\boldsymbol{Q}) = \sigma^2(n-p-1)$$

where (a) is because $\mathbb{E}\boldsymbol{e} = \mathbb{E}[\boldsymbol{QY}] = \boldsymbol{Q}\mathbb{E}\mathbf{Y} = \boldsymbol{Q}\mathbb{E}(\boldsymbol{X\beta} + \boldsymbol{\epsilon}) = \boldsymbol{Q}[\mathbb{E}\boldsymbol{\epsilon}] = \mathbf{0}$, and the last steps are as in the original proof we gave. □

Linear combinations. If $\boldsymbol{a} = (a_0, \ldots, a_p)^\top \in \mathbb{R}^{p+1}$ is a fixed vector, then

$$\theta := \boldsymbol{a}^\top\boldsymbol{\beta} = \sum_{j=0}^{p} a_j\beta_j \in \mathbb{R} \tag{19}$$

is called a *linear combination* (of $\boldsymbol{\beta}$). Consider estimating a linear combination (19). A natural estimator is

$$\hat{\theta} = \boldsymbol{a}^\top\hat{\boldsymbol{\beta}} = \boldsymbol{a}^\top\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{Y} = \boldsymbol{c}^\top\boldsymbol{Y},$$

where

$$\boldsymbol{c} := \boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{a} \in \mathbb{R}^n.$$

We can calculate its mean and variance under the linear model (3.11),

$$\mathbb{E}[\hat{\theta}] = \mathbb{E}\left[\boldsymbol{c}^\top\boldsymbol{Y}\right] = \boldsymbol{c}^\top\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{c}^\top X\boldsymbol{\beta} = \boldsymbol{a}^\top\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{a}^\top\boldsymbol{\beta} = \theta$$

and

$$V(\hat{\theta}) = V\left(\boldsymbol{c}^\top\boldsymbol{Y}\right) = \boldsymbol{c}^\top\operatorname{cov}(\boldsymbol{Y})\boldsymbol{c} = \boldsymbol{c}^\top\left[\sigma^2\mathbf{I}_n\right]\boldsymbol{c} = \sigma^2\boldsymbol{c}^\top\boldsymbol{c}.$$

Thus $\hat{\theta}$ is a linear (4.2) unbiased (4.4) estimator of θ with the variance in (4.5). Is there are better linear unbiased estimator of θ? First we need to define "better". The mean squared error (MSE) of an estimator $\hat{\theta}$ of θ is

$$\operatorname{MSE}(\hat{\theta}) := \mathbb{E}_\theta\left[(\hat{\theta} - \theta)^2\right].$$

We will say that an estimator $\hat{\theta}$ of θ is *better* than another estimator $\tilde{\theta}$ if

$$\operatorname{MSE}(\hat{\theta}) \leq \operatorname{MSE}(\tilde{\theta}) \quad \forall\theta.$$

For any estimator $\hat{\theta}$, we have

$$\operatorname{MSE}(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right] = \mathbb{E}\left[(\hat{\theta} - \mathbb{E}\hat{\theta}) + (\mathbb{E}\hat{\theta} - \theta)^2\right] =$$
$$= \mathbb{E}\left[(\hat{\theta} - \mathbb{E}\hat{\theta})^2\right] + \mathbb{E}\left[(\mathbb{E}\hat{\theta} - \theta)^2\right] + 2\underbrace{\mathbb{E}[(\hat{\theta} - \mathbb{E}\hat{\theta})(\mathbb{E}\hat{\theta} - \theta)]}_{=0} =$$
$$= \underbrace{\mathbb{E}\left[(\hat{\theta} - \mathbb{E}\hat{\theta})^2\right]}_{V(\hat{\theta})} + \underbrace{\mathbb{E}\left[(\mathbb{E}\hat{\theta} - \theta)^2\right]}_{(\operatorname{bias}(\hat{\theta}))^2}$$

where we used that fact that $\mathbb{E}[(\hat{\theta} - \mathbb{E}\hat{\theta}) = \mathbb{E}\hat{\theta} - \mathbb{E}\hat{\theta} = 0$.

We conclude from the general decomposition (5) that an unbiased estimator has

$$\mathrm{MSE}(\hat{\theta}) = V(\hat{\theta}).$$

Returning to our original question, we will say that $\hat{\theta}$ is a better estimator of θ than another candidate estimator $\tilde{\theta}$ if

$$V(\hat{\theta}) \leq V(\tilde{\theta}) \quad \forall\theta.$$

The following theorem, maybe the most famous result in all of linear regression, says that, under the linear model (16), the LS estimator $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator ("BLUE") of θ.

Theorem 1 (Gauss-Markov)**.** . *Let $\theta := \boldsymbol{a}^\top \boldsymbol{\beta}$ be a linear combination, and assume the linear model (3.11). Denote by $\hat{\theta}$ the LS estimator of θ in (4.2), and consider another linear unbiased estimator $\hat{\theta}$ of θ*

$$\tilde{\theta} = \boldsymbol{d}^\top \boldsymbol{Y}, \quad \mathbb{E}\tilde{\theta} = \theta \quad \forall\theta.$$

Then

$$V(\hat{\theta}) \leq V(\tilde{\theta}) \quad \forall\theta$$

Proof. For c defined in (4.3), write

$$\boldsymbol{d} = \boldsymbol{c} + \boldsymbol{\Delta}, \quad \boldsymbol{\Delta} := \boldsymbol{d} - \boldsymbol{c} \in \mathbb{R}^n.$$

$\tilde{\theta}$ is unbiased, then $\forall\boldsymbol{\beta}$

$$\begin{aligned}\theta = \mathbb{E}\tilde{\theta} = \mathbb{E}\left[\boldsymbol{d}^\top \boldsymbol{Y}\right] = \mathbb{E}\left[(\boldsymbol{c}+\boldsymbol{\Delta})^\top \boldsymbol{Y}\right] = \mathbb{E}\left[(\boldsymbol{c}+\boldsymbol{\Delta})^\top \boldsymbol{Y}\right] &= \mathbb{E}\Big[\Big(\boldsymbol{c}^\top \boldsymbol{Y}\Big] + \mathbb{E}\Big[\Big(\boldsymbol{\Delta}^\top \boldsymbol{Y}\Big] \\ &= \theta + \boldsymbol{\Delta}^\top \mathbb{E}[\boldsymbol{Y}] = \theta + \boldsymbol{\Delta}^\top \boldsymbol{X}\boldsymbol{\beta},\end{aligned}$$

where the second-to-last equality is due to unbiasedness of $\hat{\theta}$. Comparing the two extreme sides of the sequence of the equality, we get

$$\boldsymbol{\Delta}^\top \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0} \quad \forall\boldsymbol{\beta} \quad \Rightarrow \quad \boldsymbol{\Delta}^\top \boldsymbol{X} = \boldsymbol{0},$$

so

$$\boldsymbol{\Delta}^\top \boldsymbol{c} = \underbrace{\boldsymbol{\Delta}^\top \boldsymbol{X}}_{=0} \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{a} = \boldsymbol{0}.$$

We then calculate

$$\begin{aligned}V(\tilde{\theta}) = V\left(\boldsymbol{d}^\top \boldsymbol{Y}\right) &= V\left[(\boldsymbol{c}+\boldsymbol{\Delta})^\top \mathbf{Y}\right] \\ &= \mathrm{cov}\left[(\boldsymbol{c}+\boldsymbol{\Delta})^\top \mathbf{Y}\right] \\ &= (\boldsymbol{c}+\boldsymbol{\Delta})^\top \mathrm{cov}[\boldsymbol{Y}](\boldsymbol{c}+\boldsymbol{\Delta}) \\ &= (\boldsymbol{c}+\boldsymbol{\Delta})^\top \sigma^2 \left[\boldsymbol{I}_n\right] (\boldsymbol{c}+\boldsymbol{\Delta}) \\ &= \sigma^2 (\boldsymbol{c}+\boldsymbol{\Delta})^\top (\boldsymbol{c}+\boldsymbol{\Delta}) \\ &= \sigma^2 \left(\boldsymbol{c}^\top \boldsymbol{c} + \boldsymbol{\Delta}^\top \boldsymbol{\Delta}\right) \\ &\geq \sigma^2 \boldsymbol{c}^\top \boldsymbol{c} \\ &= V(\hat{\theta}).\end{aligned}$$

□

We have considered point estimation of a scalar $\theta = \theta(\boldsymbol{\beta})$, more specifically unbiased estimation of a linear function of $\boldsymbol{\beta}$. We now want to move on to other inferential tasks, for example we'll want to use the LS estimator $\hat{\boldsymbol{\beta}}$ to construct a confidence interval for $\boldsymbol{\beta}$, or to test whether a particular coordinate β_j is equal to zero. For this we will need some further assumptions on the linear model.

Review of multivariate distributions. All the concepts presented here generalize naturally beyond the two dimensional case. If Z_1, Z_2 are two random variables, then $Z = (Z_1, Z_2)^\top$ is a random vector of dimension 2. The joint cumulative distribution function (CDF) of $\mathbf{Z}$ is

$$F_{\boldsymbol{Z}}(z_1, z_2) := P(Z_1 \leq z_1, Z_2 \leq z_2),$$

which is always defined and determines the distribution of $\boldsymbol{Z}$. The variables Z_1 and Z_2 are (statistically) independent if

$$F_{\boldsymbol{Z}}(z_1, z_2) = P(Z_1 \leq z_1) P(Z_2 \leq z_2) \quad \text{for all } z_1, z_2 \in \mathbb{R}.$$

If the derivative

$$f_{\mathbf{Z}}(z_1, z_2) = \frac{\partial^2}{\partial z_1 z_2} F_{\mathbf{Z}}(z_1, z_2)$$

exists (for all except maybe a subset of $\mathbb{R}^2$ of probability zero), we call $f_{\mathbf{Z}}$ the *joint density* of $\boldsymbol{Z}$, and we have the relation

$$F_{\mathbf{Z}}(z_1, z_2) = \int_{-\infty}^{z_1} \int_{-\infty}^{z_2} f_{\mathbf{Z}}(u_1, u_2)\, du_1 du_2.$$

Of course, the derivative is in that case an equivalent characterization of the distribution of Z.

The multivariate Normal distribution

Definition 4. We say that a random vector $\boldsymbol{W} = (W_1, \ldots, W_k)^\top$ has a multivariate normal distribution if there exists a representation

$$\boldsymbol{W} \overset{d}{=} \boldsymbol{\mu} + \boldsymbol{A}\mathbf{Z} \tag{20}$$

where $\boldsymbol{\mu} \in \mathbb{R}^k$ and $\boldsymbol{A} \in \mathbb{R}^{k \times l}$ are constant (nonrandom) and where $\boldsymbol{Z} = (Z_1, \ldots, Z_l)^\top$ is a random vector whose components Z_i are i.i.d. $\mathcal{N}(0,1)$ random variables ("$\overset{d}{=}$" means "equal in distribution").

Properties of the multivariate Normal distribution.

1. If $\boldsymbol{W}$ has a multivariate normal distribution, then

$$\begin{gathered}\mathbb{E}[\boldsymbol{W}] = \mathbb{E}[\boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{Z}] = \mathbb{E}[\boldsymbol{\mu}] + \mathbb{E}[\boldsymbol{A}\boldsymbol{Z}] = \boldsymbol{\mu} + \boldsymbol{A}\mathbb{E}[\boldsymbol{Z}] = \boldsymbol{\mu} \\ \operatorname{cov}(\boldsymbol{W}) = \operatorname{cov}(\boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{Z}) = \operatorname{cov}(\boldsymbol{A}\boldsymbol{Z}) = \boldsymbol{A}\operatorname{cov}(\boldsymbol{Z})\boldsymbol{A}^\top = \boldsymbol{A}\boldsymbol{A}^\top\end{gathered}$$

2. In (20) suppose that $l = k$, and if $\boldsymbol{A}_{k\times k}$ has linearly independent columns, and denote $\boldsymbol{V} := \boldsymbol{A}\boldsymbol{A}^\top$. Then

$$f_W(\boldsymbol{w}) = (2\pi)^{-m/2} |\boldsymbol{V}|^{-1/2} \exp\left[-(\boldsymbol{w} - \boldsymbol{\mu})^\top \boldsymbol{V}^{-1} (\boldsymbol{w} - \boldsymbol{\mu})/2\right], \quad \boldsymbol{w} \in \mathbb{R}^m$$

It follows that the distribution of $\boldsymbol{W}$ in (4.12) depends on $(\boldsymbol{\mu}, \boldsymbol{A})$ only through $(\boldsymbol{\mu}, \boldsymbol{V})$, and we denote

$$\boldsymbol{W} \sim \mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{V})$$

for the multivariate distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{V}$.

3. It is a consequence of 2 that if $\boldsymbol{W}^{(1)} = \boldsymbol{\mu} + \boldsymbol{A}^{(1)}\mathbf{Z}^{(1)}$ and $\boldsymbol{W}^{(2)} = \boldsymbol{\mu} + \boldsymbol{A}^{(2)}\mathbf{Z}^{(2)}$, and if $\boldsymbol{A}^{(1)}\boldsymbol{A}^{(1)\top} = \boldsymbol{A}^{(2)}\boldsymbol{A}^{(2)\top}$, then $\boldsymbol{W}^{(1)} \overset{d}{=} \boldsymbol{W}^{(2)} \sim \mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{V})$.

4. From the previous properties, if $\boldsymbol{c} \in \mathbb{R}^k$ is a constant vector, then

$$\boldsymbol{c}^\top \boldsymbol{W} \sim \mathcal{N}\left(\boldsymbol{c}^\top \boldsymbol{\mu}, \boldsymbol{c}^\top \boldsymbol{V}\boldsymbol{c}\right)$$

In words, a linear combination of a multivariate normal vector has a univariate normal distribution. In particular, if we take $c = (\underbrace{0, \ldots, 0}_{j-1}, 1, \underbrace{0, \ldots, 0}_{k-j})^\top$, then

$$W_j = \boldsymbol{c}^\top \boldsymbol{W} \sim \mathcal{N}\left(\mu_j, \boldsymbol{V}_{jj}\right)$$

5. If for a random vector $\boldsymbol{W}$ it holds that $\boldsymbol{c}^\top \boldsymbol{W} \sim \mathcal{N}\left(\boldsymbol{c}^\top \boldsymbol{\mu}, \boldsymbol{c}^\top \boldsymbol{V}\boldsymbol{c}\right) \forall \boldsymbol{c} \in \mathbb{R}^m$, where $\boldsymbol{\mu}$ and $\boldsymbol{V}$ denote the mean and covariance of $\boldsymbol{W}$, then $\boldsymbol{W}$ has a multivariate normal distribution. Combined with property 4 , this says

$$\boldsymbol{c}^\top \boldsymbol{W} \sim \mathcal{N}\left(\boldsymbol{c}^\top \boldsymbol{\mu}, \boldsymbol{c}^\top \boldsymbol{V}\boldsymbol{c}\right) \quad \forall \boldsymbol{c} \in \mathbb{R}^m \quad \iff \quad \boldsymbol{W} \sim \mathcal{N}_m(\boldsymbol{\mu}, \boldsymbol{V}).$$

Thus, Property 4 is in fact a *defining property* of the multivariate normal distribution.

5. If $\boldsymbol{C} \in \mathbb{R}^{m \times k}$ constant matrix then $\boldsymbol{C}\boldsymbol{W} \sim \mathscr{N}_m\left(\boldsymbol{C}\boldsymbol{\mu}, \boldsymbol{C}\boldsymbol{V}\boldsymbol{C}^T\right)$.

6. If $\boldsymbol{W}^{(j)} \sim \mathcal{N}_k\left(\boldsymbol{\mu}^{(j)}, \boldsymbol{V}^{(j)}\right), j = 1, \ldots, p$, independent, and if d_j are scalar constants, then

$$\sum_{j=1}^{p} d_j \boldsymbol{W}^{(j)} \sim \mathcal{N}_k\left(\sum_{j=1}^{p} d_j \boldsymbol{\mu}^{(j)}, \sum_{j=1}^{p} d_j^2 \boldsymbol{V}^{(j)}\right)$$

7. Let $\boldsymbol{W} \sim \mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{V})$ and $\mathscr{I}_1, \mathscr{I}_2 \subseteq \{1, \ldots, k\}$ disjoint subsets of indices. If $\operatorname{Cov}(W_i, W_j) = 0 \quad \forall i \in \mathscr{J}_1, j \in \mathscr{J}_2$, then the vectors

$$\boldsymbol{W}^{(1)} = (W_l : l \in \mathscr{J}_2) \in \mathbb{R}^{|\mathcal{F}_2|}, \quad \boldsymbol{W}^{(2)} = (W_k : k \in \mathscr{J}_1) \in \mathbb{R}^{|\mathscr{I}_1|}$$

are statistically independent.

Distributions related to the normal.

Definition 5 (Chi-square distribution)**.** If $Z_1, Z_2, \ldots, Z_k \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$, then the distribution of

$$Q = \sum_{j=1}^{k} Z_j^2$$

is called the Chi-square distribution with k degrees of freedom, and we denote $Q \sim \chi_k^2$ (in R: `pchisq()`, `qchisg()`, `rchisq()`).

Definition 6 (t-distribution)**.** . If $Z \sim \mathcal{N}(0, 1), V \sim \chi_k^2$, are independent random variables, then the distribution of

$$T = \frac{Z}{\sqrt{V/k}}$$

is called the t-distribution with k degrees of freedom, and we denote $T \sim t_k$ (in R: `pt ()`, `qt( )`, `rt( )`).

Definition 7 (F distribution)**.** If $V_1 \sim \chi^2_{k_1}, V_2 \sim \chi^2_{k_2}$, are independent random variables, the distribution of

$$F = \frac{V_1/k_1}{V_2/k_2}$$

is called the F*-distribution with* k_1 *and* k_2 (numerator and denominator, respectively) *degrees of freedom*, and we denote $F \sim \mathrm{F}_{k_1,k_2}$.

Proposition. If $Q \sim \chi^2_k$, then $\mathbb{E}Q = k$.

Proof. For $Z_i \sim \mathcal{N}(0,1)$, iid for $i = 1, \ldots, k$, we can write $Q \overset{d}{=} \sum_{i=1}^k Z_i^2$, where " $\overset{d}{=}$ " means "equal in distribution". Then $\mathbb{E}Q \overset{d}{=} \mathbb{E}\sum_{i=1}^k Z_i^2 = \sum_{i=1}^k \mathbb{E}Z_i^2 = \sum_{i=1}^k V\left(Z_i\right) = k$. □

Proposition. Let $\boldsymbol{Z} \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{I})$, and $\boldsymbol{P}$ be a square symmetric $\left(\boldsymbol{P}^\top = \boldsymbol{P}\right)$ and idempotent $\left(\boldsymbol{P}^2 = \boldsymbol{P}\right)$ matrix with $\operatorname{rank}(\boldsymbol{P}) = r$. Then $\|\boldsymbol{P}Z\|^2 \sim \chi^2_r$.

Proof. From a previous lemma, since $\mathbb{E}[\boldsymbol{P}\boldsymbol{Z}] = \boldsymbol{P}\mathbb{E}[\boldsymbol{Z}] = \boldsymbol{0}$, we have $\mathbb{E}\|\boldsymbol{P}\boldsymbol{Z}\|^2 = \operatorname{tr}(\operatorname{cov}[\boldsymbol{P}\boldsymbol{Z}]) = \operatorname{tr}\left(\boldsymbol{P}\boldsymbol{I}\boldsymbol{P}^\top\right) = \operatorname{tr}(\boldsymbol{P}) = r$, where the last equality is because $\boldsymbol{P}$ is similar to a diagonal matrix with r nonzero elements on its diagonal. □

Inference under the normal linear model. Recall:

$$\text{the (general) linear model:} \qquad \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \mathbb{E}\boldsymbol{\epsilon} = \boldsymbol{0}, \operatorname{cov}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_n$$

We will now make the additional assumption that the error term $\boldsymbol{\epsilon}$ has a *multivariate normal* distribution. In other words, we will assume The normal linear model:

$$\text{the } \textit{normal} \text{ linear model:} \qquad \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}_n\left(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n\right)$$

The additional normality assumption will enable us to address inferential tasks beyond point estimation, e.g., to construct a confidence interval for a linear combination of $\hat{\boldsymbol{\beta}}$. Indeed, if we assume $\boldsymbol{\epsilon}$ has a multivariate normal distribution, then we can derive exact distributions of $\hat{\boldsymbol{\beta}}, \hat{\sigma}^2$, and their joint.

Distribution of $\hat{\boldsymbol{\beta}}$. Recall that, for $\boldsymbol{A} := \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \in \mathbb{R}^{(p+1)\times n}$, we have

$$\begin{aligned}
\hat{\boldsymbol{\beta}} = \boldsymbol{A}\boldsymbol{Y} &= \boldsymbol{A}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\
&= \boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{A}\boldsymbol{\epsilon} \\
&= \boldsymbol{\beta} + \boldsymbol{A}\boldsymbol{\epsilon} \\
&\overset{d}{=} \boldsymbol{\beta} + (\sigma \boldsymbol{A})\mathbf{Z}
\end{aligned}$$

where $\boldsymbol{Z} \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{I})$. Hence, by definition, $\hat{\boldsymbol{\beta}}$ has a multivariate normal distribution. We have already calculated the moments of $\hat{\boldsymbol{\beta}}$,

$$\mathbb{E}\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}, \quad \operatorname{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1},$$

so in conclusion we have

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}_{p+1}\left(\boldsymbol{\beta}, \sigma^2 \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\right)$$

Distribution of $\hat{\sigma}^2$. Recall that $\boldsymbol{e} = \boldsymbol{Q}\boldsymbol{\epsilon}$, where $\boldsymbol{Q}$ is the $n \times n$ projection matrix onto the orthogonal complement of $\operatorname{Im}(\boldsymbol{X})$. By a previous result, $\|\boldsymbol{e}\|^2 \sim \sigma^2 \chi^2_{n-p-1}$. This gives

$$\frac{n-p-1}{\sigma^2}\hat{\sigma}^2 \sim \chi^2_{n-p-1} \quad \Longleftrightarrow \quad \frac{\hat{\sigma}^2}{\sigma^2} \sim \frac{\chi^2_{n-p-1}}{n-p-1} \quad \sim t_{n-p-1}.$$

Joint distribution of $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$. For $\boldsymbol{A} := \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \in \mathbb{R}^{(p+1)\times n}$, first note that

$$\hat{\boldsymbol{\beta}} = \boldsymbol{A}\boldsymbol{Y} = \boldsymbol{A}\left(\boldsymbol{P}_M \boldsymbol{Y} + \boldsymbol{P}_{M^\perp}\boldsymbol{Y}\right) = \boldsymbol{P}\boldsymbol{P}\boldsymbol{P}_M \boldsymbol{Y} + \boldsymbol{A}\boldsymbol{P}_{\boldsymbol{M}^\perp}\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{P}_{\boldsymbol{M}}\boldsymbol{Y}$$

Then,

$$\begin{aligned}
\operatorname{cov}(\hat{\boldsymbol{\beta}}, \boldsymbol{e}) = \operatorname{cov}\left(\boldsymbol{A}\boldsymbol{P}_M \boldsymbol{Y}, \left(\boldsymbol{I}_n - \boldsymbol{P}_M\right)\boldsymbol{Y}\right) &= \boldsymbol{A}\boldsymbol{P}_M \operatorname{cov}(\boldsymbol{Y})\left(\boldsymbol{I}_n - \boldsymbol{P}_M\right)^\top \\
&= \sigma^2 \boldsymbol{A}\boldsymbol{P}_M\left(\boldsymbol{I}_n - \boldsymbol{P}_M\right) = \boldsymbol{0}
\end{aligned} \tag{21}$$

Moreover,

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \boldsymbol{e} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A} \\ \boldsymbol{I}_n - \boldsymbol{P}_M \end{bmatrix} \boldsymbol{Y} \overset{d}{=} \begin{bmatrix} \boldsymbol{A} \\ \boldsymbol{I}_n - \boldsymbol{P}_M \end{bmatrix} (\boldsymbol{X}\boldsymbol{\beta} + \sigma \boldsymbol{Z}) \tag{22}$$

i.e., $\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \boldsymbol{e} \end{bmatrix}$ has a multivariate normal distribution. Together, (21) and (22) imply that $\hat{\beta}$ and $\hat{\sigma}^2$ are statistically independent (because uncorrelated=independent under joint normality).

Statistical inference for a single coefficient.

I. Confidence interval. By the derivations above, we have

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 V_{jj}}} \sim \mathcal{N}(0,1), \quad \boldsymbol{V} := \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}$$

Since σ^2 is unknown, we replace it naturally by its estimator $\hat{\sigma}^2$:

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 V_{jj}}} = \frac{\left(\hat{\beta}_j - \beta_j\right)/\sqrt{\sigma^2 V_{jj}}}{\sqrt{\hat{\sigma}^2 V_{jj}}/\sqrt{\sigma^2 V_{jj}}} = \frac{\left(\hat{\beta}_j - \beta_j\right)/\sqrt{\sigma^2 V_{jj}}}{\sqrt{\hat{\sigma}^2/\sigma^2}} \sim \frac{\mathcal{N}(0,1)}{\sqrt{\chi^2_{n-p-1}/(n-p-1)}}$$

By definition, then,

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 V_{jj}}} \sim t_{n-p-1}. \tag{23}$$

Hence,

$$\begin{aligned} 1-\alpha &= \mathbb{P}\left(t_{n-p-1;\alpha/2} \le \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 V_{jj}}} \le t_{n-p-1;1-\alpha/2}\right) \\ &= \mathbb{P}\left(\hat{\beta}_j - \sqrt{\hat{\sigma}^2 V_{jj}} \cdot t_{n-p-1;\alpha/2} \le \beta_j \le \hat{\beta}_j + \sqrt{\hat{\sigma}^2 V_{jj}} \cdot t_{n-p-1;1-\alpha/2}\right) \end{aligned}$$

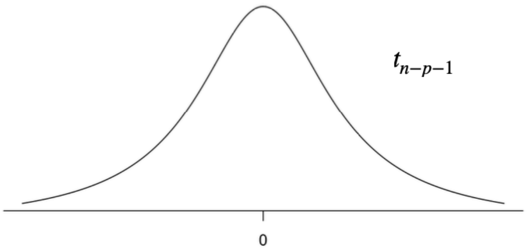


Figure 5: t distribution

In other words,

$$\boxed{CI = \left(\hat{\beta}_j - \sqrt{\hat{\sigma}^2 V_{jj}} \cdot t_{n-p-1;\alpha/2}, \hat{\beta}_j + \sqrt{\hat{\sigma}^2 V_{jj}} \cdot t_{n-p-1;1-\alpha/2}\right) \tag{24}}$$

is a confidence interval (CI) of level $1-\alpha$ for β_j.

More generally, let $\theta = \boldsymbol{a}^\top \boldsymbol{\beta}$ be a linear combination. Then, for the BLUE estimator $\hat{\theta} = \boldsymbol{a}^\top \hat{\boldsymbol{\beta}}$, we have

$$\mathbb{E}[\hat{\theta}] = \boldsymbol{a}^\top \boldsymbol{\beta} = \theta, \qquad V(\hat{\theta}) = \operatorname{cov}\left(\boldsymbol{a}^\top \hat{\boldsymbol{\beta}}\right) = \sigma^2 \boldsymbol{a}^\top \boldsymbol{V} \boldsymbol{a}.$$

Therefore,

$$\frac{\hat{\theta} - \theta}{\sqrt{\sigma^2 a^\top V a}} \sim \mathcal{N}(0,1), \quad \boldsymbol{V} := \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \tag{25}$$

Repeating the argument from before,

$$\frac{\hat{\theta} - \theta}{\sqrt{\hat{\sigma}^2 a^\top V \boldsymbol{a}}} = \frac{(\hat{\theta} - \theta)/\sqrt{\sigma^2 \boldsymbol{a}^\top V \boldsymbol{a}}}{\sqrt{\hat{\sigma}^2/\sigma^2}} = \frac{(\hat{\theta} - \theta)/\sqrt{\sigma^2 a^\top V \boldsymbol{a}}}{\sqrt{\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2}/(n-p-1)}},$$

so

$$\frac{\hat{\theta} - \theta}{\sqrt{\hat{\sigma}^2 \boldsymbol{a}^\top \boldsymbol{V} \boldsymbol{a}}} \sim t_{n-p-1}. \tag{26}$$

In conclusion,

$$\boxed{CI = \hat{\theta} \pm t_{n-p-1;1-\alpha/2} \cdot \sqrt{\hat{\sigma}^2 \boldsymbol{a}^\top \boldsymbol{V} \boldsymbol{a}} \tag{27}}$$

is a Cl of level $1 - \alpha$ for θ.

Note: for $\boldsymbol{a} = \boldsymbol{e}_j = (0, \dots, 0, 1, 0, \dots, 0)^\top$ we get $\theta = \beta_j$ and (27) coincides with (24).

II. Hypothesis testing. Based on (23), we can also derive a statistical test for

$$H_0 : \beta_j = 0 \quad \text{vs.} \quad H_1 : \beta_j \neq 0.$$

Indeed,

$$T_j := \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 V_{jj}}} \overset{H_0}{\sim} t_{n-p-1} \Longrightarrow \mathbb{P}_{H_0}\left(t_{n-p-1;\alpha/2} \leq \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 V_{jj}}} \leq t_{n-p-1;1-\alpha/2}\right) = 1 - \alpha,$$

i.e., a level- α test is to reject H_0 if

$$|T_j| := |\frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 V_{jj}}}| \geq t_{n-p-1;1-\alpha/2}. \tag{28}$$

Extending this to a general linear combination, let $\theta := \boldsymbol{a}^\top \boldsymbol{\beta}$, and suppose we want to test

$$H_0 : \theta = 0 \quad \text{vs.} \quad H_1 : \theta \neq 0.$$

Then, by (26), under the null we have

$$\frac{\hat{\theta}}{\sqrt{\hat{\sigma}^2 \boldsymbol{a}^\top \boldsymbol{V} \boldsymbol{a}}} \overset{H_0}{\sim} t_{n-p-1}.$$

By a similar calculation to the one above, the test that rejects H_0 if

$$|\frac{\hat{\theta}}{\sqrt{\hat{\sigma}^2 \boldsymbol{a}^\top \boldsymbol{V} \boldsymbol{a}}}| \geq t_{n-p-1;1-\alpha/2} \tag{29}$$

is a level-α test.

Note that any CI for θ naturally defines a test of

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta \neq 0.$$

for *any* value θ_0. Indeed, let CI_α be a $(1-\alpha)$-level CI for θ, and consider the test that rejects H_0 whenever $\theta_0 \notin CI_\alpha$. Then, since a CI covers the true parameter θ for *all* θ, it holds in particular for $\theta = \theta_0$, and we have

$$P_{H_0}(\text{type I error}) = P_{\theta_0}(\theta_0 \notin CI_\alpha) \leq 1 - \alpha.$$

We emphasize that this holds for *any* valid CI for θ. If we choose the particular CI (27), then by definition the corresponding test rejects when

$$\theta_0 \notin \left(\hat{\theta} - t_{n-p-1;1-\alpha/2} \cdot \sqrt{\hat{\sigma}^2 \boldsymbol{a}^\top \boldsymbol{V} \boldsymbol{a}},\ \hat{\theta} + t_{n-p-1;1-\alpha/2} \cdot \sqrt{\hat{\sigma}^2 \boldsymbol{a}^\top \boldsymbol{V} \boldsymbol{a}}\right),$$

which is equivalent to rejecting H_0 if

$$|\frac{\hat{\theta} - \theta_0}{\sqrt{\hat{\sigma}^2 \boldsymbol{a}^\top \boldsymbol{V} \boldsymbol{a}}}| \geq t_{n-p-1;1-\alpha/2}. \tag{30}$$

In the special case $\theta_0 = 0$, this coincides with (29). (Of course, since this holds for any linear combination θ, it holds in particular for $\theta = \beta_j$, and we get as a special case the duality between (24) and (28).

III. Prediction interval. Assume the normal linear model, $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}_n\left(\boldsymbol{0}, \sigma^2 \boldsymbol{I}\right)$. Now suppose we had a new independent observation $(\boldsymbol{X}_*, Y_*)$ from the same model, where $\boldsymbol{X}_* = (1, X_{*1}, ..., X_{*p})^\top \in \mathbb{R}^{p+1}$ is the feature (explanatory variables) vector and Y_* is the response for the new observation,

$$Y_* = \boldsymbol{X}_*^\top \boldsymbol{\beta} + \epsilon_*, \quad \epsilon_* \sim \mathcal{N}\left(0, \sigma^2\right).$$

We now want to construct a *prediction interval* for Y_*, i.e., a random interval $PI = PI\left(\boldsymbol{X}_*; \boldsymbol{X}, \boldsymbol{Y}\right)$ which is a function of the observed sample $(\boldsymbol{X}, \boldsymbol{Y})$, s.t.

$$\mathbb{P}\left(Y_* \in PI\right) = 1 - \alpha.$$

A natural *point predictor* for Y_* is

$$\hat{Y}_* = X_*^\top \hat{\boldsymbol{\beta}}$$

Now,

$$\begin{aligned}
\mathbb{E}\left[\hat{Y}_* - Y_*\right] &= \mathbb{E}\hat{Y}_* - \mathbb{E}Y_* = \boldsymbol{X}_*^\top \mathbb{E}\hat{\boldsymbol{\beta}} - \boldsymbol{X}_*^\top \boldsymbol{\beta} = 0 \\
V\left[\hat{Y}_* - Y_*\right] &= V(\hat{Y}_*) + V(Y_*) - 2\operatorname{Cov}(\hat{Y}_*, Y_*) = \sigma^2 \boldsymbol{X}_*^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}_* + \sigma^2 - 2\operatorname{Cov}(\boldsymbol{X}_*^\top \hat{\boldsymbol{\beta}}, \boldsymbol{X}_*^\top \boldsymbol{\beta} + \epsilon_*) = \\
&= \sigma^2[1 + \boldsymbol{X}_*^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}_*]
\end{aligned}$$

Moreover,

$$\begin{pmatrix} Y^* \\ \hat{Y}^* \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}^{*T}\boldsymbol{\beta} + \epsilon^* \\ \boldsymbol{x}^{*T}\hat{\boldsymbol{\beta}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}^{*T}\boldsymbol{\beta} + \epsilon^* \\ \boldsymbol{x}^{*T}\boldsymbol{A}\boldsymbol{Y} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}^{*T}\boldsymbol{\beta} + \epsilon^* \\ \boldsymbol{x}^{*T}\boldsymbol{A}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}^{*T}\boldsymbol{\beta} \\ \boldsymbol{x}^{*T}\boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta} \end{pmatrix} + \begin{pmatrix} 1 & \boldsymbol{0} \\ 0 & x^{*T}\boldsymbol{A} \end{pmatrix} \begin{pmatrix} \epsilon^* \\ \boldsymbol{\epsilon} \end{pmatrix}$$

and, since ϵ^* and $\boldsymbol{\epsilon}$ are *independent* normals, we conclude that $\binom{Y^*}{\hat{Y}^*}$ has a multivariate (2-dim) normal distribution. This implies

$$Y^* - \hat{Y}^* \sim \mathcal{N}\big(0, \sigma^2[1 + \boldsymbol{X}_*^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}_*]\big),$$

because $Y^* - \hat{Y}^*$ is a linear transformation of $(\hat{Y}_*, Y_*)^\top$. With the usual argument replacing σ^2 with $\hat{\sigma}^2$, a $(1-\alpha)$-level prediction inverval for Y^* is then given by

$$\boxed{PI = \hat{Y}_* \pm t_{n-p-1;1-\alpha/2} \cdot \hat{\sigma}\sqrt{1 + \boldsymbol{X}_*^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}_*}. \qquad (31)}$$

Least squares regression: demonstration with R. (in separate file)

6 Categorical explanatory variables

So far, we have treated the explanatory variables $X_j, j = 1, ..., p$, as continuous (numeric) variables. In many cases we would like to incorporate into the regression a *categorical variable* (in `R`: ”factor”), i.e., a variable whose values indicate membership in one of several categories. For example: sex (”male”, ”female”); blood type (”A”, ”B”, ”AB”, ”O”); grape variety for wine (”cabernet sauvignon”, ”merlot”, ”pinot noir”, ”syrah”, ”tempranillo”). We can do this via coding with binary variables.

Suppose that one of the explanatory variables, say X_p, takes on values in $\{0, 1\}$. Recall that the mean of the response, under the linear model, is

$$\mathbb{E}Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j = \begin{cases} \beta_0 + \sum_{j=1}^{p-1}, & X_p = 0 \\ \beta_0 + \beta_p + \sum_{j=1}^{p-1}, & X_p = 1 \end{cases} .$$

That is, the effect of including a binary variable in the regression model is a shift of the intercept (each of $X_p = 0, X_p = 1$ has its own intercept). Specifically, β_0 is the value of the intercept for the category encoded $X_p = 0$, and $\beta_0 + \beta_p$ is the value of the intercept for the category encoded $X_p = 1$, so that β_p is the difference in the intercept values.

Dummy variables. If we have a categorical variable with only 2 categories, we can use a binary variable to represent it. If we have a categorical variable with more than 2 categories, we can encode it with a collection of corresponding binary variables, commonly called *dummy variables*.

Example. Consider a categorical variable indicating grape variety for wine,“cabernet sauvignon”, “merlot”, “pinot noir”, “syrah”, “tempranillo”, 5 categories in total. We represent this with a collection of 4 dummy (binary) variables, $X_{p(k)}, k = 1, \ldots, 4$, corresponding to any 4 of the 5original categories; the remaining, left out category, is called the ”reference” (or ”baseline”) category. In this way, only one of the dummy variables equals 1, and all the rest are zero, except when encoding the baseline category, in which case all of the dummies are zero.

	$X_{p(1)}$	$X_{p(2)}$	$X_{p(3)}$	$X_{p(4)}$
Cabernet Sauvignon	0	0	0	0
Merlot	1	0	0	0
Pinot Noir	0	1	0	0
Syrah	0	0	1	0
Tempranillo	0	0	0	1

In the example above, ”cabernet sauvignon” was chosen as the baseline category, but this choice is arbitrary (R chooses this automatically, generally according to alphabetical order, which also determines which category is left out as baseline).

In general, if X_p (say) is a categorical variable with K categories (called ”levels”), we use $K - 1$ dummy variables in the regression to encode it:

$$\mathbb{E}[Y_i] = \beta_0 + \beta_{p(1)} X_{ip(1)} + \beta_{p(2)} X_{ip(2)} + \cdots + \beta_{p(K-1)} X_{ip(K-1)} + \sum_{j=1}^{p-1} \beta_j X_{ij}$$

This results in a linear mode with $p + K - 2$ variables + intercept, i.e., $p + k - 1$ variables in total. The ”effective” intercept of level k, for $k = 1, \ldots, K - 1$, is $\beta_0 + \beta_{p(k)}$, so that $\beta_{p(k)}$ is the difference in intercepts for the kt th category.

Of course, we can include more than one categorical variable in a regression model. In that case, each of the categorical variables will have a baseline level, and the overall intercept will correspond to the combination of all baseline levels. Here is an example analyzed with R.

```
> # load dataset
> salaries <- carData::Salaries
> head(salaries)

          rank  discipline  yrs.since.phd     yrs.service   sex salary
  1       Prof  B                      19  18 Male 139750
  2       Prof  B                      20  16 Male 173200
  3   AsstProf  B                       4          3 Male          79750
  4       Prof  B                      45                39  Male 115000
  5       Prof  B                      40  41 Male 141500
  6  AssocProf  6                  6 Male            97000

> # extract variable names and types
> names(salaries)

[1] "rank" "discipline" "yrs.since.phd" "yrs.service" "sex" "salary"
```

We can obtain variable type with the function `str()`. Note the variables of type ”Factor”:

```
> str(salaries)
 'data.frame':  397 obs.  of 6 variables:
$ rank :  Factor w/ 3 levels "AsstProf", "AssocProf",..:  3 3 1 3 3 2 3 3
3 3 ...
$ discipline :  Factor w/ 2 levels "A","B":  2 2 2 2 2 2 2 2 2 2 ...
$ yrs.since.phd:  int 19 20 4 45 40 6 30 45 21 18 ...
$ yrs.service :  int 18 16 3 39 41 6 23 45 20 18 ...
$ sex :  Factor w/ 2 levels "Female","Male":  2 2 2 2 2 2 2 2 2 1 ...
$ salary :  int 139750 173200 79750 115000 141500 97000 175000 147765 119250
129000 ...
>
```

To view which level (“Male” or “Female”) is assigned the value 1 (by default):

```
> # view dummy coding for the factor 'sex' and 'rank'
> contrasts(salaries$sex)
          Male
    Female     0
    Male       1
```

Same for “rank”:

```
  > contrasts(salaries$rank)
               AssocProf  Prof
    AsstProf           0     0
    As.socRrof         1     0
    Prof               0     1
```

Frequency table (how many men, how many women in entire dataset):

```
> # frequency table for sex
> table(salaries$sex)
     Female    Male
         39     358
```

Linear regression of ”salary” on ”sex” and ”yrs.service”:

```
> ## regress y = salary on sex + yrs.service
> fm <- lm(salary šex + yrs.service, data = salaries)
> summary(fm)
Call:   lm(formula = salary šex + yrs.service, data = salaries)
Residuals:

       Min       1Q   Median      3Q      Max
    -81757   -20614    -3376   16779   101707

 Coefficients:  Estimate Std.  Error t value Pr(> |t|)
 (Intercept)   92356.9  4740.2  19.484   < 2e − 16   ***
 sexMale        9071.8  4861.6   1.866    0.0628   .
 yrs.service     747.6   111.4   6.711  6.74e − 11   ***

Signif.  codes:  0 ‘***’ 0.001 ’**’ 0.01 ‘*’ 0.05 ’.’  0.1 ’, 1

Residual standard error:  28490 on 394 degrees of freedom
Multiple R-squared:  0.1198, Adjusted R-squared:  0.1154
F-statistic:  26.82 on 2 and 394DF, p-value:  1.201e − 11
```

The fitted regression line is

$$\hat{Y}_i = 92356.9 + 9071.8 \cdot \text{ sexMale }_i + 747.6 \cdot \text{ yrs.service }_i$$

Hence, 92356.9 is the ’effective’ intercept for a female. The estimated mean salary for a male with 3 years of service, for example, is

$$92356.9 + 9071.8 + 747.6 \cdot 3 = 103671.5.$$

We can view the corresponding X matrix with `model.matrix()`:

```
> # obtain X matrix
> head( model.matrix(fm) )

(Intercept) sexMale yrs.service
 1   1   1   18
 2   1   1   16
 3   1   1    3
 4   1   1   39
 5   1   1   41
 6   1   1    6
```

Add another categorical variable:

```
> regress y = salary on sex + rank + yrs.service
> # frequency table for rank*sex
> table(salaries$rank, salaries$sex)

             Female Male
  AsstProf       11   56
  AssogProf      10   54
  Prof           18  248

> fm1 <- lm(salary   sex + rank + yrs.service, data = salaries)

> summary(fm1)$coef
                 Estimate  Std. Error    t value      Pr(>| t |)
  (Intercept)   76612.810   4426.0007  17.309715  2.847735e − 50
  sexMale        5468.708   4035.3366   1.355205  1.761327e − 01
  rankAssocProf. 14702.856  4266.5563   3.446071  6.303299e − 04
  rankProf      48980.224   3991.8299  12.270118  1.635066e − 29
  yrs.service    -171.792    115.2707  -1.490335  1.369404e − 01

#The general intercept 76612.810 is the 'effective' intercept
for female assistant professor
#The 'effective' intercept for a female professor
will be 76612.810 + 48980.224
```

The fitted regression line is

$$\begin{aligned}\hat{Y}_i =& 76612.810 + 5468.708 \cdot \texttt{sexMale}_i + \\ & + 14702.856\texttt{rankAssocProf}_i + 48980.224\texttt{rankProf}_i - 171.792 \cdot \texttt{yrs.service}_i\end{aligned}$$

Notice that the general intercept β_0, and the coefficient of the dummy variable "sexMale" and of the continuous variable "yrs.service", are different from the values in the previously fitted model (without "rank"); in fact, the coefficient of "yrs.service" even changes sign! 76612.810 is the 'effective' intercept for a female assistant professor. The estimated mean salary for a male professor, with 3 years of service, for example, is

$$76612.810 + 5468.708 + 48980.224 - 171.792 \cdot 3 = 130546.366.$$

7 Interactions

An interaction term (variable) for two explanatory variables, say X_p and X_j, is a new explanatory variable given by $X_{\text{new}} = X_p X_j$. Including interaction variables allows the effect of one variable to depend on the value of another variable.

Example 1. Suppose we start with a regression model with 3 explanatory variables,

$$\mathbb{E}\left[Y_i\right] = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3}.$$

In this model, the interpretation of the third coefficient (for example) is as follows: β_3 is the increase in the mean value of Y per unit increase in X_3, if we hold the values of X_1, X_2 fixed. I.e., this is the increase in the mean response value "conditional" on $X_1 = x_1, X_2 = x_2$, but it holds regardless of the specific values x_1, x_2.

Now let's add an interaction between X_2 and X_3, i.e., add the new variable $X_{i4} = X_{i2}X_{i3}$:

$$\mathbb{E}\left[Y_i\right] = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4}.$$

If X_2 and X_3 are both binary, the interaction has an effect on the effective intercept only:

$$\text{intercept} \;=\; \begin{cases} \beta_0, & \text{if } X_{i2} = 0, X_{i3} = 0 \\ \beta_0 + \beta_2, & \text{if } X_{i2} = 1, X_{i3} = 0 \\ \beta_0 + \beta_3, & \text{if } X_{i2} = 0, X_{i3} = 1 \\ \beta_0 + \beta_2 + \beta_3 + \beta_4, & \text{if } X_{i2} = 1, X_{i3} = 1 \end{cases}$$

If X_2 is binary and X_3 is continuous:

$$\text{slope of } X_{i2} = \begin{cases} \beta_2, & \text{if } X_{i3} = 0 \\ \beta_2 + \beta_4, & \text{if } X_{i3} = 1 \end{cases}$$

Finally, if X_2 and X_3 are both continuous:

$$\begin{array}{ll} \text{slope of } X_{i2} = \beta_2 + \beta_4 u, & \text{if } X_{i3} = u \\ \text{slope of } X_{i3} = \beta_3 + \beta_4 v, & \text{if } X_{i2} = v \end{array}$$

Example 2. $Y =$ salary. $X_1 =$ yrs.service. $X_2 =$ sex. Regress $Y =$ salary on $X_1 =$ yrs.service:

$$Y_i = \beta_0 + \beta_1 \times \text{ yrs.service } + \epsilon_i.$$

```
> # regress salary on yrs.service
> fm0 <- lm(salary yrs.service, data=salaries)
> fm0$coefficients

 (Intercept)  yrs.service
 99974.6529   779.5691
```

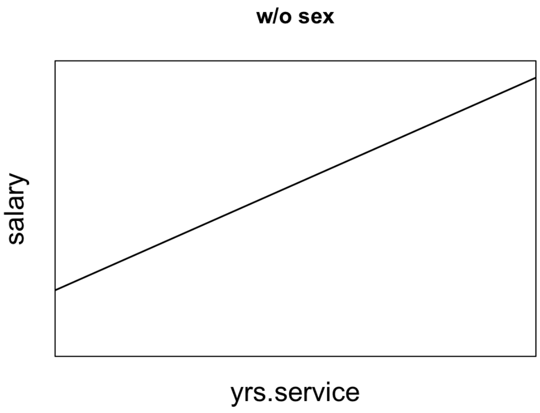


Figure 6: Simple regression of salary on years of service for Example 2

Regress Y = salary on X_1 = yrs.service and X_2 = sex, no interaction:

$$Y_i = \beta_0 + \beta_1 \times \text{yrs.service} + \beta_2 \times (\text{sex} = \text{Male}) + \epsilon_i$$

```
> # regress salary on yrs.service + sex, no interaction
> fm1 <- lm(salary ỹrs.service + sex, data=salaries)
> fm1$coefficients
 (Intercept)  yrs.service  sexMale
 92356.9467   747.6121     9071.8000
```

Hence, estimated mean salary is:

For Males: $92356.9467 + 9071.8000 + 747.6121\cdot$ yrs.service
For Females: $92356.9467 + 747.6121\cdot$ yrs.service

I.e., two parallel lines, because no interaction. Important: the slope (=the coefficient of ”yrs.service”) is not the same as in the previous that excludes ”sex”.

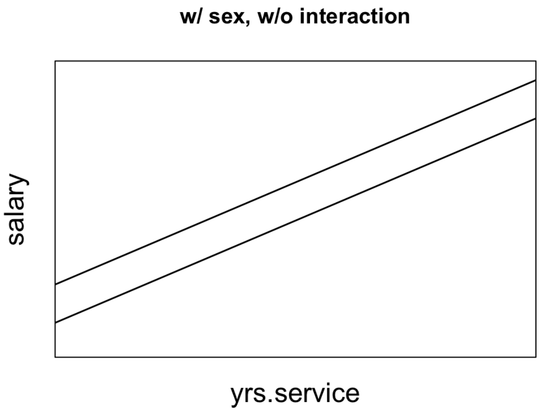


Figure 7: Regression of “yrs.service” on “sex”, no interaction

Regress Y = salary on X_1 = yrs.service and X_2 = sex, with interaction:

$$Y_i = \beta_0 + \beta_1 \times \text{yrs.service} + \beta_2 \times (\text{sex = Male}) + \epsilon_i + \beta_3 \times \text{yrs.service} \times (\text{sex=Male}) + \epsilon_i.$$

```
> # regress salary on yrs.service + sex, with interaction
> fm2 <- lm(salary ỹrs.service + sex + yrs.service:sex, data=salaries)
> fm2$coefficients
  (Intercept)  yrs.service     sexMale  yrs.service:sexMale
   82068.5087    1637.2997  20128.6258            -931.7363
```

Hence, estimated mean salary is:

For Males: $82068.5087 + 20128.6258 + (1637.2997 - 931.7363)\cdot$ yrs.service
For Females: $82068.5087 + 1637.2997\cdot$ yrs.service

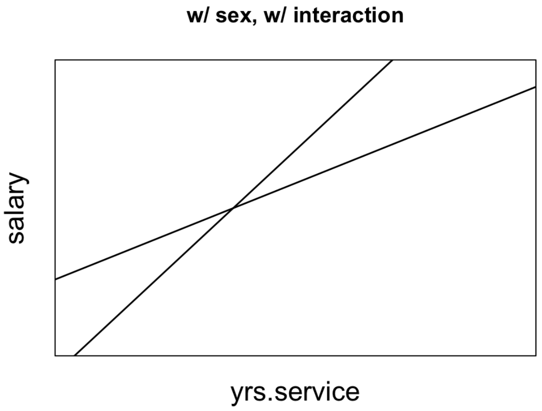


Figure 8: Regression of “yrs.service” on “sex”, with interaction

How to evaluate the significance of interaction term (i.e., whether or not the effect of ”yrs.service” differs significantly between males and females)? Look at the p-value for the interaction term. In the output below this is ≈ 0.083, so, e.g., it is significant at $\alpha = 0.1$ level.

```
> fm2 <- lm(salary   yrs.service + sex + yrs.service:sex,data=salaries)
> summary(fm2)
```

```
Call:
lm(formula = salary   yrs.service + sex + yrs.service:sex, data = salaries)
Residuals:
     Min       1Q   Median      3Q    Max
    −80381   −20258      −3727  16353  102536

Coefficients:
                         Estimate  Std.  Error  t value   Pr(> |t|)
     (Intercept)          82068.5        7568.7   10.843  < 2e − 16  ∗∗∗
     yrs.service           1637.3         523.0    3.130  0.00188   **
     sexMale              20128.6        7991.1    2.519  0.01217   *
     yrs.service:sexMale   -931.7         535.2   -1.741  0.08251   .
---
Signif.  codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’  0.1 ‘ ’ 1

Residual standard error:  28420 on 393 degrees of freedom
Multiple R-squared:  0.1266, Adjusted R-squared:  0.1199
F-statistic:  18.98 on 3 and 393 DF, p-value:  1.622e-11
```

8 Residual analysis: checking model assumptions

We move on to discuss some practical aspects of regression analysis. Our first topic will be residual analysis: the general goal is to check if the modeling assumptions seem adequate (correct). To remind, the general linear model is given by

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \mathbb{E}[\boldsymbol{\epsilon}] = \boldsymbol{0}, \mathrm{cov}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_n.$$

We can break this down into three separate assumptions:

1. *Linearity*: $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0} \Longleftrightarrow \mathbb{E}[\boldsymbol{Y}] = \boldsymbol{X}\boldsymbol{\beta}$ (make sure you can prove this equivalence).
2. *Homoscedastic (equal-variance) errors*: $\mathrm{Var}\,(\epsilon_i) \equiv \sigma^2$ is the same for all $i = 1, \ldots, n$
3. *Uncorrelated errors*: $\mathrm{Cov}\,(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$.

In the normal model, also:

4. *Normality of errors*: $\boldsymbol{\epsilon}$ is multivariate normal $\Rightarrow \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma^2\right)$ (combined with 2,3).

We will first present some basic tools for detecting "substantial" departures from assumptions 1-4. If such departures are detected, we will discuss some possible fixes.
The main assumptions are linearity (Assumption 1) and homoscedasticity (Assumption 2); we will want to check these first. Let us start intuitively with the case of simple linear regression (i.e., single predictor, $p = 1$). In that case we can look at a scatterplot of the data and try to visually detect violations of the linearity and homoscedasticity assumptions.

The plot below shows 4 different simulated datasets. The fitted (LS) regression line is shown in blue in all four panels. In the top left panel the data was generated from a linear model with equal error variances. We see that there is roughly the number of points lying above and below the LS line throughout the range of the x values; this indicates good agreement with the linearity assumption. Furthermore, the spread of the points around the LS line is roughly constant throughout the range of the x values; this indicates good agreement with the homoscedasticity (equal-variances) assumption.

In the top right there is strong violation of the linearity assumption: the points show a clear trend in their location about the LS line: for x values with small absolute value, almost all points are below the LS line, and the trend is opposite for x values with large absolute value. However, there is still good agreement with the homoscedasticity assumption: the dashed black line represents a nonparametric regression fit; you can think of this as a smoothed version of an estimator that bins the points on the x axis, and takes the average of the y-values in each bin separately. The points do seem to have a constant spread (throughout the range of the x values) about this dotted line.

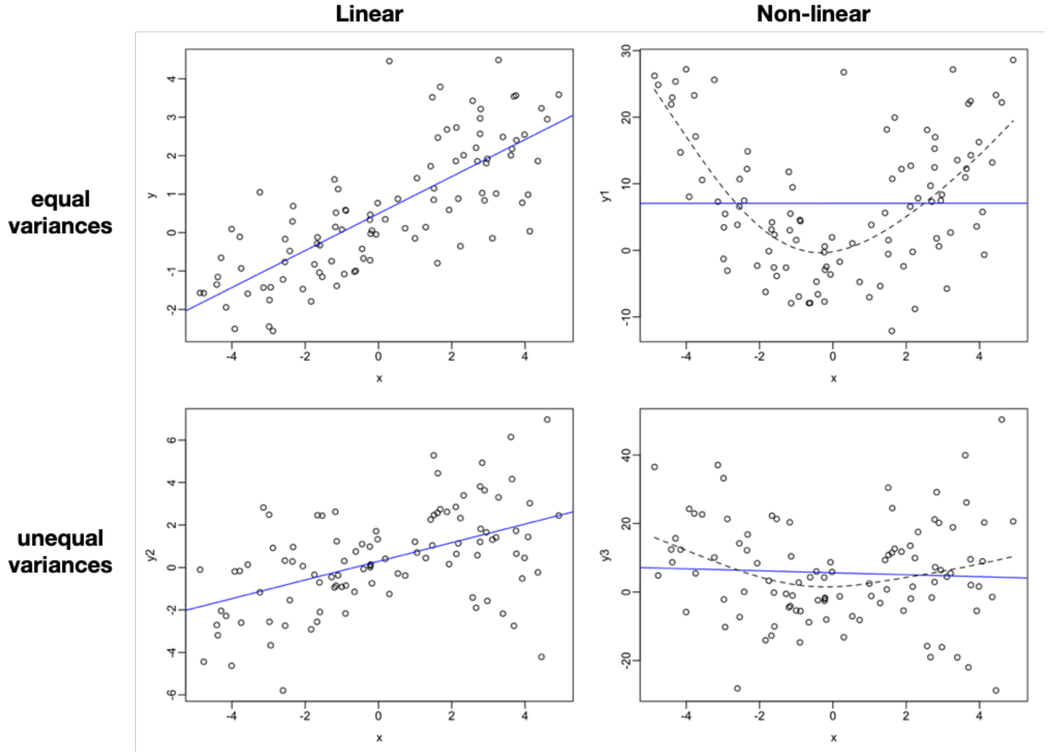


Figure 9: Violations of linearity and/or homoscedasticity

In the bottom right panel the data is simulated from a linear model, but the errors have unequal variances: this is shown clearly in the "fanning out" of the points as we move away from $x = 0$ in either direction. Lastly, in the bottom right panel both linearity and homoscedasticity are violated.

It is harder to check the Normality assumption just by looking at the scatterplot; we will need some device to help us check normality. But also if we stick to the linearity and homoscedasticity assumptions, the situation is more complicated when dealing with multiple regression (general number p of predictors), because a scatterplot is basically irrelevant for $p > 2$ (we can't really visualize a scatterplot for the case of more than two predictors, because this will require a plot in more than 3 dimensions). The workaround wil be to consider residuals plots instead of scatterplots. In fact, residual plots are usually more convenient to inspect than scatterplots also in the simple regression case.

Recall that the residuals are

$$e_i = Y_i - \hat{Y}_i$$

which in vector form can be written

$$\boldsymbol{e} = \boldsymbol{Q}, \quad \boldsymbol{Q} := \boldsymbol{P}_{\mathrm{Im}(X)^\perp}.$$

Under the general linear model, we have

$$\mathbb{E}(\boldsymbol{e}) = \mathbb{E}(\boldsymbol{Q}Y) = \boldsymbol{Q}\mathbb{E}(\boldsymbol{Y}) = \boldsymbol{Q}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0}$$

(by the linearity assumption). So, under the general linear model, the residuals are zeromean, $\mathbb{E}\left(e_i\right)=0$, regardless of the value of $\boldsymbol{x}_i \in \mathbb{R}^{p+1}$, in particular, regardless of the $\mathbb{E}\left[Y_i\right]=\sum \beta_j x_{ij}$ (which is a function of $\boldsymbol{x}_i$). When the sample size n is large, and under reasonable assumptions, the fitted value $\hat{Y}_i=\boldsymbol{x}_i^{\top} \hat{\boldsymbol{\beta}}$ will be a consistent (in a proper sense) estimator of its mean $\mathbb{E}\left[\hat{Y}_i\right]=\mathbb{E}\left[Y_i\right]=\sum \beta_j x_{ij}$, i.e. for very large n we can roughly treat $\hat{Y}_i \approx \mathbb{E}\left[Y_i\right]=\sum \beta_j x_{ij}$ (this is of course imprecise). This means we roughly expect the residuals to be approximately zero-mean, $\mathbb{E}\left(e_i\right)=0$ regardless of the value of $\hat{Y}_i$.

As for the variances, note that

$$\operatorname{cov}(\boldsymbol{e})=\operatorname{cov}(\boldsymbol{Q})=\boldsymbol{Q} \operatorname{cov}(\boldsymbol{Y}) \boldsymbol{Q}^{\top}=\sigma^2 \boldsymbol{Q}$$

so even under the linear model the residuals are generally not uncorrelated (hence not independent), and not even homoscedastic (i.e., $\operatorname{Var}\left(\epsilon_i\right)=Q_{ii}$ depends on i); compare this with ϵ_i, which have equal variances and are independent. Still, when the sample size n is large we argued that $\hat{Y}_i \approx \mathbb{E}\left[Y_i\right]=\sum \beta_j x_{ij}$, so its variance is approximately zero. In that case

$$\operatorname{Var}\left(e_i\right)=\operatorname{Var}\left(Y_i-\hat{Y}_i\right) \approx \operatorname{Var}\left(Y_i\right)=\operatorname{Var}\left(\epsilon_i\right) \equiv \sigma^2$$

so we can also expect the spread of the e_i 's about the x axis (the horizontal linear at $y=0$) to be roughly constant regardless of the value of $\hat{Y}_i$.

The discussion above can be made even more intuitive by saying that, when n is large, we expect “$e_i \approx \epsilon_i$” and “$\hat{Y}_i \approx \sum \beta_j x_{ij}$ ” in some sense, but a plot of ϵ_i vs. $\sum \beta_j x_{ij}$ should be throughout centered about zero, and have constant variance. In other words, we expect no pattern at all (”shapeless cloud of points”). The figure below shows the corresponding plots for e_i vs. $\hat{Y}_i$ for the 4 scenarios in the previous plot.

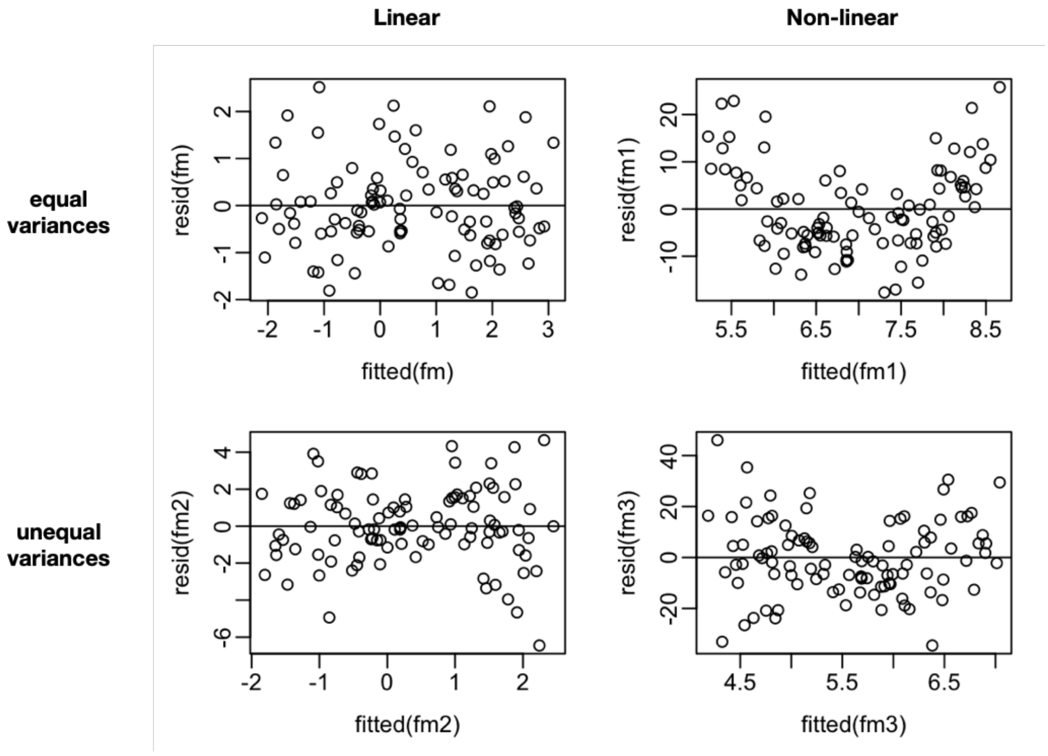


Figure 10: Residuals e_i vs. fitted values $\hat{Y}_i$

Checking normality. To check the normality assumption, we need a better tool because this is a more subtle assumption and harder to see visually just by staring at the residual plots. A preliminary check can be done using simply a histogram of the residuals e_i. Recall that, to form a histogram for a sample $W_1, \ldots, W_n$, we fix a grid $w_1, \ldots, w_K$, and let $n_k =$ number of observations falling in the bin $(w_{k-1}, w_k]$, and $h_k = w_k - w_{k-1}$ be the width of the k th bin, $k = 1, \ldots, K$. Then we plot a rectangle whose base is the bin $(w_{k-1}, w_k]$ and whose height is

$$f_k := \frac{1}{n} \cdot \frac{n_k}{h_k}.$$

The resulting histogram is an estimate for the density of W_i. The following command in R forms a histogram with bins of equal width: hist(r, breaks = 'scott', freq = FALSE).

Thus, under the normal model, the histogram of the residuals e_i should approximate a $\mathcal{N}\left(0, \sigma^2\right)$ density:

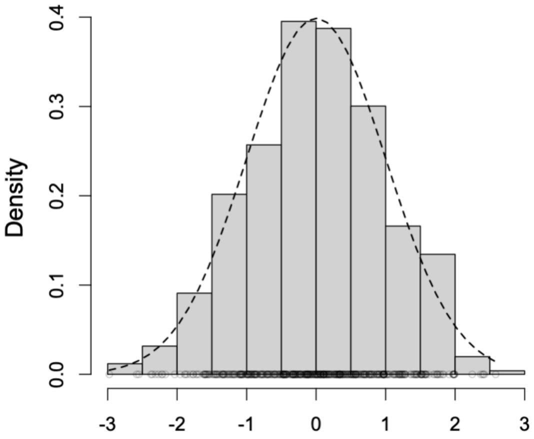


Figure 11: Histogram of residuals e_i

We could overlay a density curve for $\mathcal{N}\left(0, \sigma^2\right)$ and compare, but a more convenient and precise way is to use a Quantile-Quantile plot (Q-Q plot).

To introduce the Q-Q plot, consider first CDF's F and G corresponding to any two distributions, and consider the plot

$$y = F^{-1}(p) \quad \text{vs.} \quad x = G^{-1}(p), \quad p \in (0,1)$$

If the two distributions are the same, i.e., $F(t) = G(t)$ for all t, then of course we expect the plot to form the line $y = x$. In a Q-Q plot, G is taken to be a theoretical reference distribution, and F is taken to be an empirical CDF computed from the sample. Note that, if F is an empirical CDF, then F^{-1} gives the order statistics $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. Of course, in this case we never expect the points to align exactly on a straight line, but we do expect this approximately if the sample is iid draws from G.

Thus, to check normality of the errors, we will basically want to plot

$$e_{(i)} \text{ vs. } \quad \Phi^{-1}\left(\frac{i}{n}\right), \qquad i = 1, \ldots, n$$

where $e_{(1)} \leq e_{(2)} \leq \cdots \leq e_{(n)}$ are the order statistics of the residuals, and Φ^{-1} is the inverse CDF of a normal distribution with proper mean and variance. However, if F, G in (11.2) are both normal distributions (each with its own-but possibly different-mean and variance), it is a simple exercise to show that the graph

(11.2) will still be a straight line, though not necessarily with slope 1 and intercept 0 . In other words, in (11.3) we can take Φ^{-1} to be a standard normal $\mathcal{N}(0,1)$, and compare to a general straight line instead of the identity line $y = x$. Recall that, under the normal model and assuming n is large, we expect e_i to be approximately $\mathscr{N}\left(0,\sigma^2\right)$ and approximately independent. In this case, we expect (11.3) to lie approximately on a straight line. The basic command in R is

```
qqnorm(resid(fm1))  # generate Normal Q-Q plot
qqline(resid(fm1))  # add reference line
```

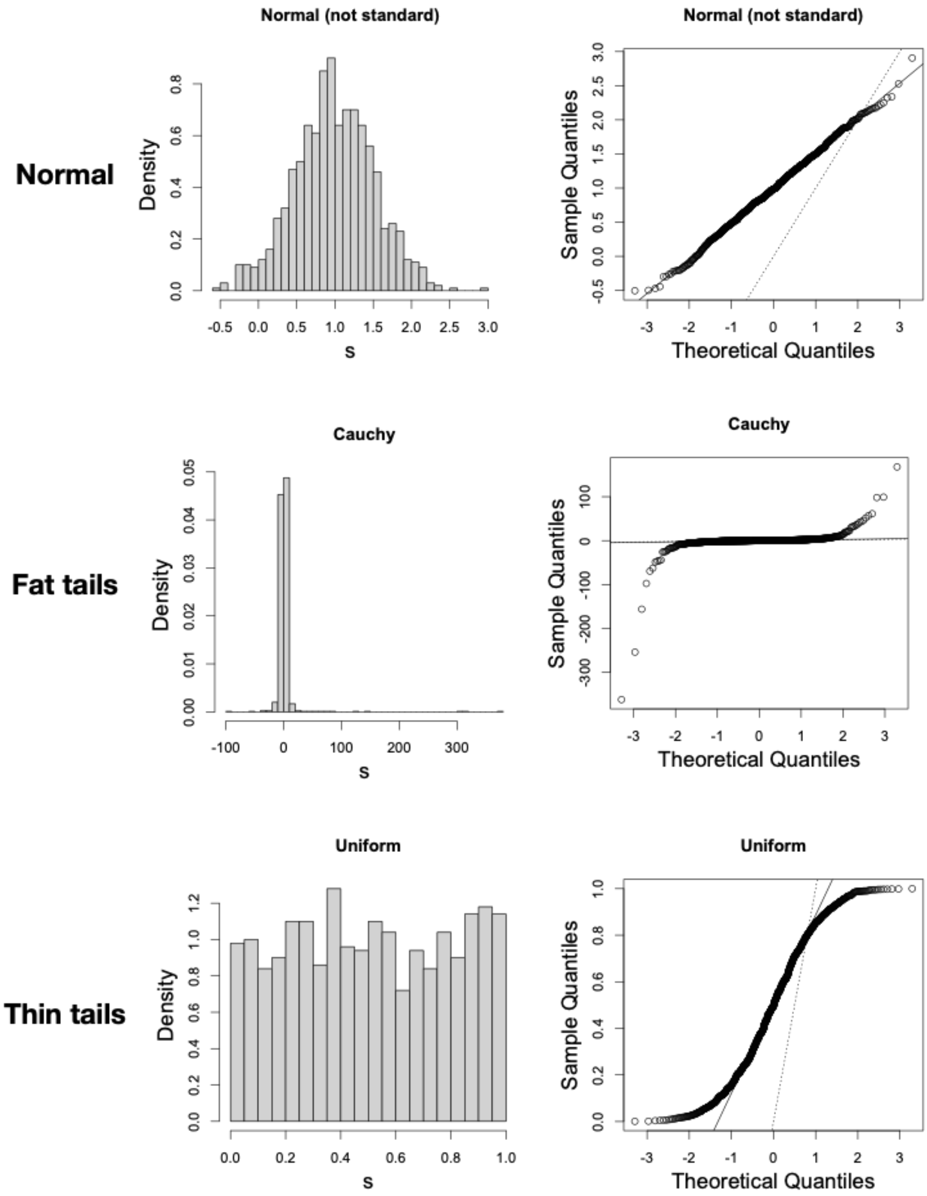


Figure 12: QQ-plot

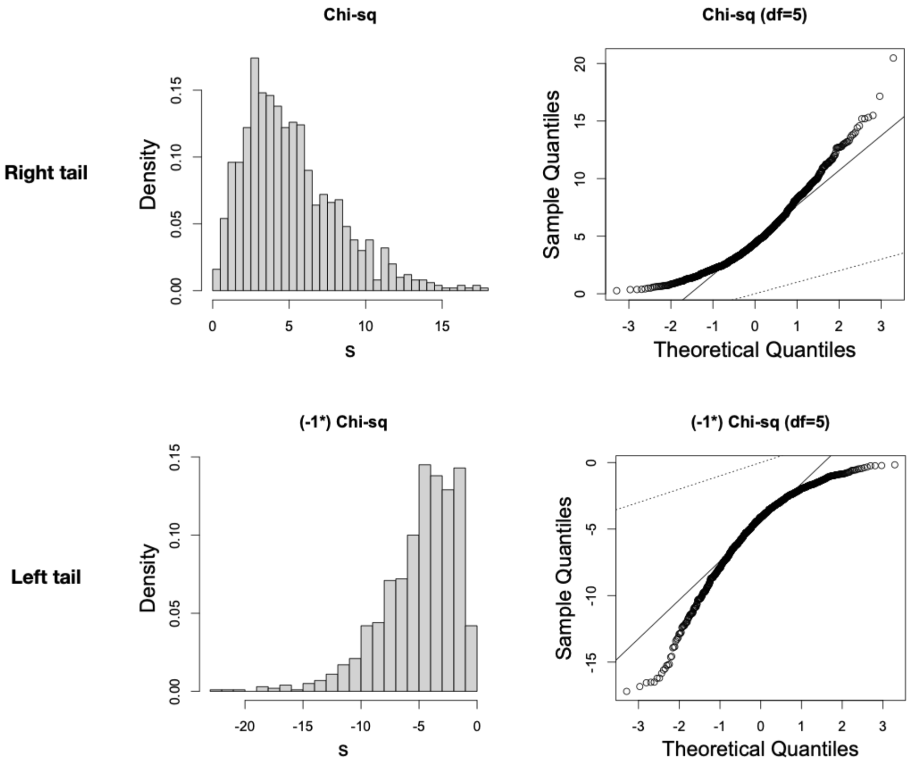


Figure 13: QQ-plot (continued)

Outliers. An outlier is a sample point $(\boldsymbol{X}_i, Y_i)$ that shows clear disagreement with the fitted model for the dataset. To identify outliers, one may compute $\hat{Y}_{(i)} =$ fitted value for X_i when the i th observation is removed from the dataset. If the difference $Y_i - \hat{Y}_{(i)}$ is large, the i th observation is suspect as an outlier. Note: this does not necessarily imply that the residual from the fit to the entire dataset (including the i th point) is small!

A formal measure for identifying candidate outliers is the leverage of a point, defined as

$$\text{leverage } := \operatorname{Cov}\left(\hat{Y}_i, Y_i\right) = [\operatorname{cov}(\hat{\boldsymbol{Y}}, \boldsymbol{Y})]_{ii} = [\boldsymbol{P}_X]_{ii} = \left[\boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top\right]_{ii}$$

Note that this depends on the matrix $\boldsymbol{X}$ only, i.e., only on the explanatory variables. Observations with high leverage have the potential of being outliers.

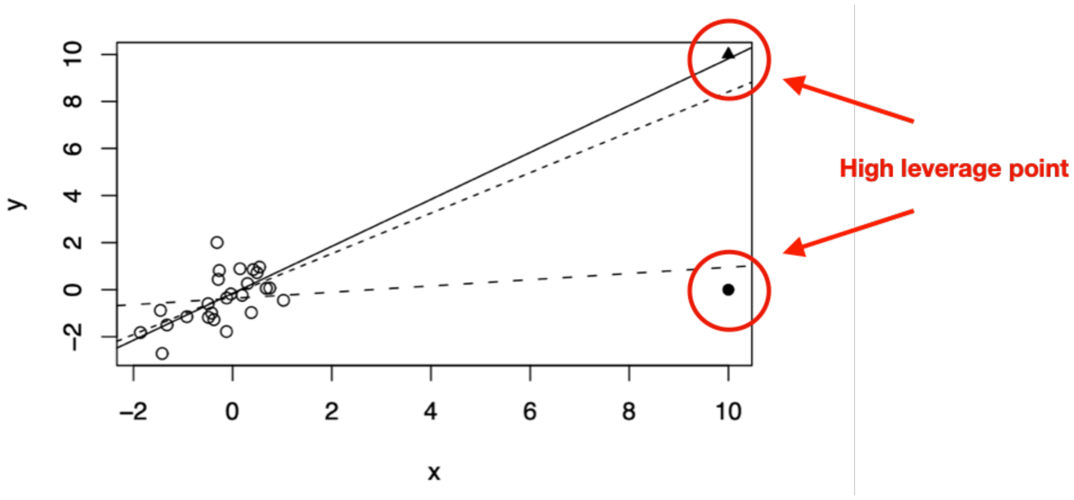
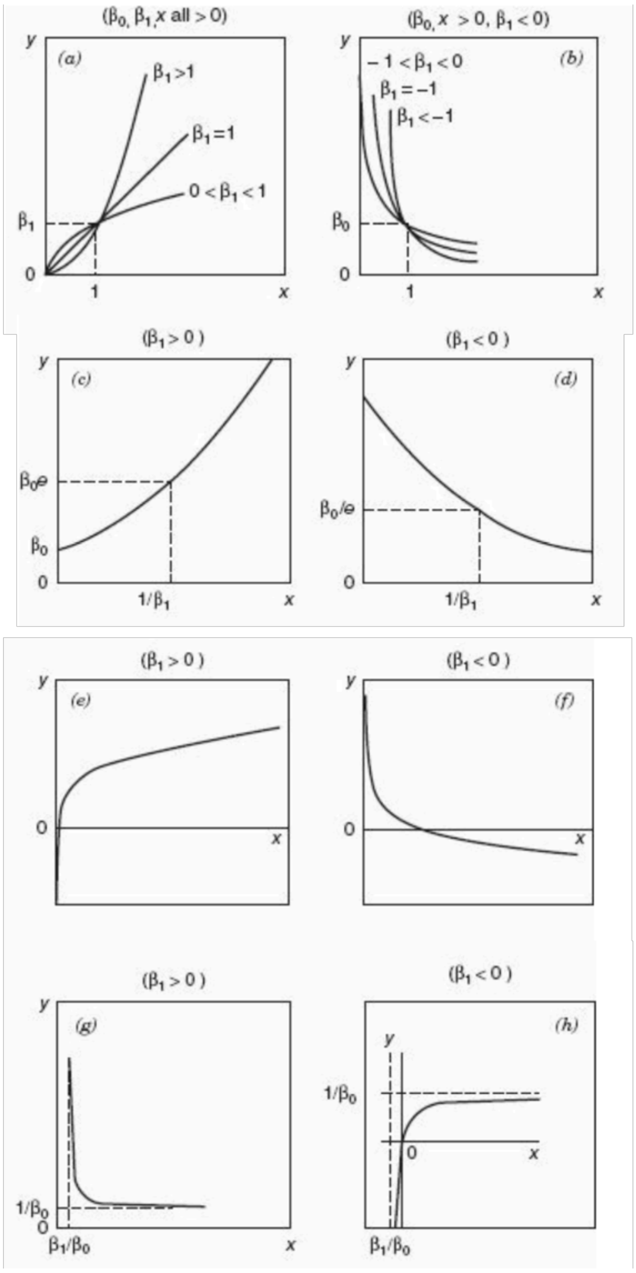


Figure 14: Illustration of leverage

Strategies for fixing violations of the linear model. We will now see a few workarounds the can help to correct the situation when there are clear violations of the basic model assumptions. Violations of the linearity assumption. The linearity assumption says that the expectation of Y_i is linear in $\boldsymbol{X}_i, \mathbb{E}Y_i = \boldsymbol{X}_i^\top \boldsymbol{\beta}$. There are situations where the linear model is inadequate for regressing the original Y_i s on the original X_i s, but will become appropriate one applying a transformation to Y_i or X_i or both.

Examples. Here are some examples that can be used as guidelines for suggesting a ”correcting” transformation.



$$y = \beta_0 x^{\beta_1} \longrightarrow \tilde{y} = \log(y), \tilde{x} = \log(x)$$
$$\tilde{y} = \log\left(\beta_0\right) + \beta_1 \tilde{x}$$

$$y = \beta_0 e^{\beta_1 x} \longrightarrow \tilde{y} = \ln(y)$$
$$\tilde{y} = \ln\left(\beta_0\right) + \beta_1 x$$

$$y = \beta_0 + \beta_1 \log(x) \quad \longrightarrow \quad \tilde{x} = \log(x)$$
$$\tilde{y} = \beta_0 + \beta_1 \tilde{x}$$

$$y = \frac{x}{\beta_0 x - \beta_1} \quad \longrightarrow \quad \tilde{y} = \frac{1}{y}, \tilde{x} = \frac{1}{x}$$
$$\tilde{y} = \beta_0 - \beta_1 \tilde{x}$$

Figure 15: Transformations

Violations of the equal-variance (homoscedasticity) assumption. There are cases where the assumption of equal variances for the errors ϵ_i is not appropriate. For example, this would be the case when the distribution of the outcome variable Y is such that the variance $\mathrm{Var}[Y]$ is functionally related to $\mathbb{E}[Y]$.

Examples.

- $Y \sim \mathrm{Pois}(\lambda) : \mathbb{E}[Y]=\lambda, \quad \mathrm{Var}(Y)=\lambda$

- $Y \sim \mathrm{Binom}(n,p)/n : \mathbb{E}[Y]=p, \quad \mathrm{Var}(Y)=p(1-p)/n$

Hence, if, say, $Y_i \sim \mathrm{Pois}\left(\lambda_i\right)$ with the mean $\lambda_i=\mathbb{E}\left[Y_i\right]$ being a function of X_i, then $\mathrm{Var}\left[\epsilon_i\right]=\lambda_i$ will not be the same for all $i=1,\ldots,n$.

In such cases, it may be possible to apply a variance stabilizing transformation to Y in order to recover again a situation with equal-variance errors.

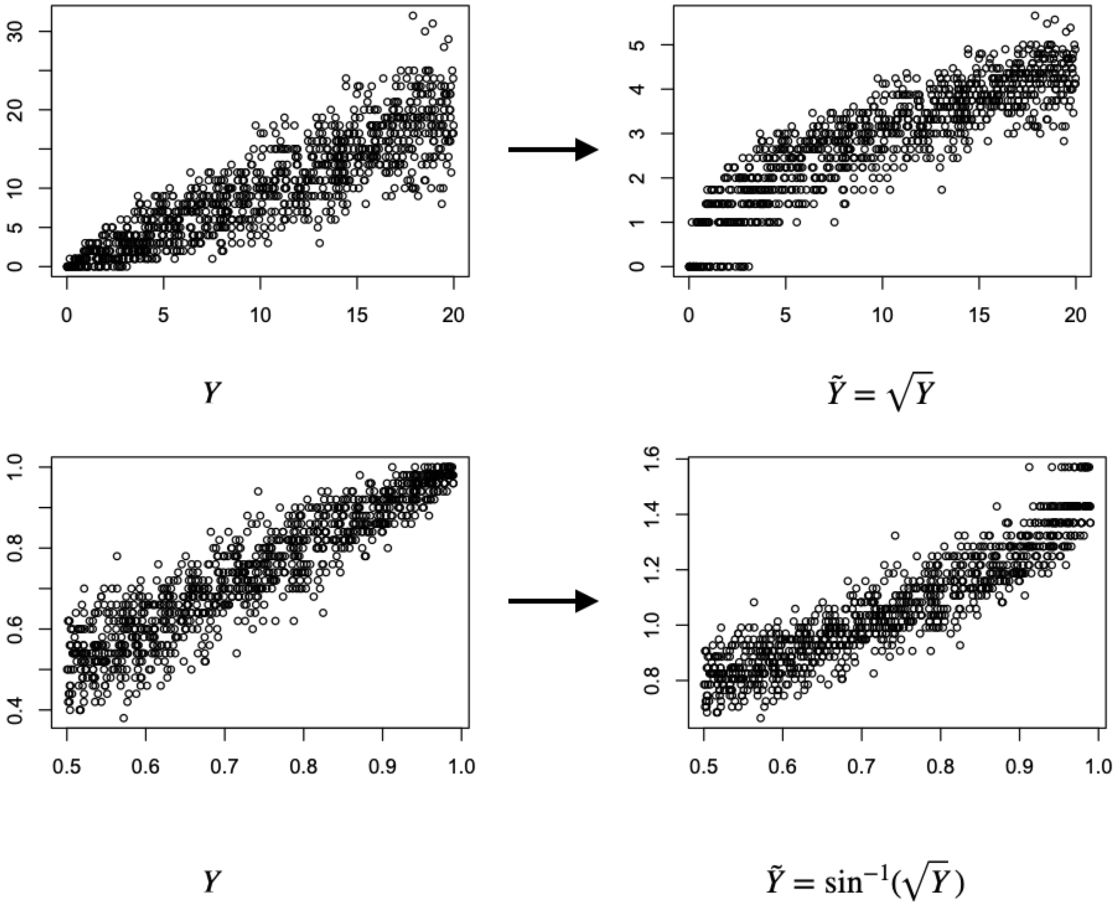


Figure 16: Variance-stabilizing transformations

How to obtain the correct transformation for variance stabilizing? Suppose that we know that

$$\sigma^2=\phi\left(\mu_Y\right)$$

for some known function $\phi(\cdot)$. For example, in the binomial case,

$$\mu_Y=np, \quad \sigma_Y=[np(1-p)]^{1/2}=\left[\mu_Y\left(1-\mu_Y/n\right)\right]^{1/2}.$$

If we now define $Z = f(Y)$, and use the so-called delta method to write

$$z = f(y) \approx f(\mu_Y) + f'(\mu_Y)(y - \mu_Y) \Longrightarrow \sigma_Z^2 \approx [f'(\mu_Y)]^2 \sigma_Y^2 = [f'(\mu_Y)]^2 \phi^2(\mu_Y),$$

then to stabilize the variance would mean approximately to have

$$f'(\mu_Y)\sigma_Y = f'(\mu_Y)\phi(\mu_Y)$$

implying

$$f(y) = \sigma_Z \int \frac{1}{\phi(y)} dy$$

Thus, for example, in the Binomial case,

$$f(y) = \sigma_Z \int [y(1 - y/n)]^{-1/2} dy = \sigma_Z \cdot 2n^{1/2} \arcsin\left[(y/n)^{1/2}\right]$$

Generalized least squares. A variance-stabilizing transformation may help in correcting the situation back to the case with (approximately) equal-variance errors, but the transformation applied might, at the same time, impact linearity: if the original means $\mathbb{E}[Y_i]$ are linear in $\boldsymbol{X}_i$, then, by the same delta method argument, after transformation the means are $\mathbb{E}[Y_i] \approx f(\mu_Y)$ and we generally lose linearity.

There is another method to deal with violations of the equal-variance assumption by working directly with the original data, rather than transforming. Thus, assume an *Extended linear model*:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \mathbb{E}[\boldsymbol{\epsilon}] = \boldsymbol{0}, \operatorname{cov}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{V} \tag{32}$$

where $\boldsymbol{V}$ is a known $n \times n$ positive-definite covariance matrix. Note that in the special case $\boldsymbol{V} = \boldsymbol{I}_n$ we are back to the standard linear model. It is easy to verify that the usual LS estimator,

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$$

is still an unbiased estimator of $\boldsymbol{\beta}$, and, hence, $\hat{\boldsymbol{\theta}} = \boldsymbol{a}^\top \hat{\boldsymbol{\beta}}$ is unbiased for $\theta = \boldsymbol{a}^\top \boldsymbol{\beta}$. In the special case $\boldsymbol{V} = \boldsymbol{I}_n$, we further have by the Gauss-Markov theorem that $\hat{\theta} = \boldsymbol{a}^\top \hat{\boldsymbol{\beta}}$ is BLUE, i.e., it has minimum variance among all linear unbiased estimators of $\theta = \boldsymbol{a}^\top \boldsymbol{\beta}$. This is no longer true in the case of a general $\boldsymbol{V}$; however, by reducing the model back to the familiar case $\boldsymbol{V} = \boldsymbol{I}_n$, we can obtain a BLUE for the more extended model (11.4): first, we find an invertible $n \times n$ matrix $\boldsymbol{A}$ s.t.

$$\boldsymbol{V} = \boldsymbol{A}\boldsymbol{A}^\top$$

(this is always possible when $\boldsymbol{V}$ is positive-definite, e.g. we find such $\boldsymbol{A}$ if we orthogonally diagonalize $\boldsymbol{V}$). Now define

$$\tilde{\boldsymbol{Y}} = \boldsymbol{A}^{-1}\boldsymbol{Y}, \quad \tilde{\boldsymbol{X}} = \boldsymbol{A}^{-1}\boldsymbol{X}, \quad \tilde{\boldsymbol{\epsilon}} = \boldsymbol{A}^{-1}\boldsymbol{\epsilon}.$$

Then we have

$$\tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{\epsilon}}, \quad \tilde{\boldsymbol{\epsilon}} \sim \left(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n\right), \tag{33}$$

i.e., the usual linear model holds for the transformed variables. The Gauss-Markov theorem then says that the estimator

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{\mathrm{GLS}} = \left(\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}\right)^{-1} \tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{Y}} &= \left[\boldsymbol{X}^\top \left(\boldsymbol{A}^\top\right)^{-1} \boldsymbol{A}^{-1}\boldsymbol{X}\right]^{-1} \boldsymbol{X}^\top \left(\boldsymbol{A}^\top\right)^{-1} \boldsymbol{A}^{-1}\boldsymbol{Y} \\ &= \left[\boldsymbol{X}^\top \left(\boldsymbol{A}\boldsymbol{A}^\top\right)^{-1} \boldsymbol{X}\right]^{-1} \boldsymbol{X}^\top \left(\boldsymbol{A}\boldsymbol{A}^\top\right)^{-1} \boldsymbol{Y} \\ &= \left(\boldsymbol{X}^\top \boldsymbol{V}^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{V}^{-1} \boldsymbol{Y}\end{aligned}$$

is BLUE for $\boldsymbol{\beta}$ under the original model (32), but this means that this estimator is also BLUE for the transformed model (33) because this is the same $\boldsymbol{\beta}$ in both models. The estimator $\hat{\boldsymbol{\beta}}^{\mathrm{GLS}}$ above is called the generalized least squares (GLS) estimator.

Special case: if $\boldsymbol{V} = \boldsymbol{W} = \operatorname{diag}(w_1, \ldots, w_n)$ is diagonal, meaning that the errors are uncorrelated but do not have equal variances, the GLS estimator is called the weighted least squares (WLS) estimator.

9 Multicollinearity

Recall that, throughout, we have had the assumption that the columns of the $n \times (p+1)$ matrix $\boldsymbol{X}$ are linearly independent (implying necessarily that $p+1 \leq n$). If the columns of $\boldsymbol{X}$ were linearly dependent, then $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$ is not defined because $\boldsymbol{X}^\top \boldsymbol{X}$ is not invertible, and there is indeed no unique LS estimator (comment: a solution to the LS criterion, i.e., a minimizer of the sum of squared residuals, always exists, but is unique only when the columns of $\boldsymbol{X}$ are linearly independent). In fact, even the true parameter vector $\boldsymbol{\beta}$ is, in a sense, not well-defined because it is non-identifiable $\iff$ there exist several choices of $\boldsymbol{\beta}$ yielding the same value for $\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{X}\boldsymbol{\beta}$).

While we assume that the columns of $\boldsymbol{X}$ are never exactly linearly dependent, i.e.,

$$\boldsymbol{X}\boldsymbol{c} = \sum_{j=0}^{p} c_j \boldsymbol{X}^{(j)} \neq 0$$

for all $\boldsymbol{c} \in \mathbb{R}^p$, they may still be nearly linearly dependent, i.e.,

$$\boldsymbol{X}\boldsymbol{c} = \sum_{j=0}^{p} c_j \boldsymbol{X}^{(j)} \approx 0$$

for some $\boldsymbol{c} \in \mathbb{R}^p$. In other words, there is redundancy in the explanatory variables in the sense that there is an explanatory variable that's nearly a linear combination of the others. If this is the case, we say that we have a high degree of multicollinearity in $\boldsymbol{X}$ matrix. While multicollinearity generally does not affect prediction accuracy (recall that $\hat{\boldsymbol{Y}} = \boldsymbol{P}_{\mathrm{Im}(\boldsymbol{X})}\boldsymbol{Y}$ does not depend on $\boldsymbol{X}$ itself, only on the span of its columns), it does affect the variance of the coefficients of the explanatory variables. Specifically, if there's substantial multicollinearity, the LS estimator $\hat{\boldsymbol{\beta}}$ will be highly sensitive to small changes in $\boldsymbol{Y}$, which will result in large variances for the estimators $\hat{\beta}_j$. We first show this formally by giving an alternative representation for the LS solution $\hat{\beta}_j$.

For this end, we consider first simple regression with no intercept. We can think of this as a special case of the general setting, where $\boldsymbol{X} = (x_1, \dots, x_n)^\top$ is a $n \times 1$ matrix (i.e., a column vector) that records the values of a single explanatory variable, and, importantly, without the intercept column $\mathbf{1} = (1, \dots, 1)^\top$. The least squares solution is given by the usual formula,

$$\hat{c} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{Y} = \frac{\boldsymbol{X}^\top \boldsymbol{Y}}{\boldsymbol{X}^\top \boldsymbol{X}} = \frac{\boldsymbol{X}^\top \boldsymbol{Y}}{\|\boldsymbol{X}\|^2} \in \mathbb{R},$$

where we used the symbol $\hat{c}$ instead of $\hat{\boldsymbol{\beta}}$ to distinguish from the usual, vector estimator which uses a matrix $\boldsymbol{X}$ that includes a column of 1 's. remembering that now $\boldsymbol{X}^\top \boldsymbol{X} \in \mathbb{R}$ is a scalar, and so is $\hat{\beta}$ (note that it appears in regular font, not boldface as usual).

Now, come back to the usual multiple regression setting, where $\boldsymbol{Y} \in \mathbb{R}^n$, and

$$\boldsymbol{X} = \begin{bmatrix} | & | & & | \\ \boldsymbol{X}^{(0)} & \boldsymbol{X}^{(1)} & \cdots & \boldsymbol{X}^{(0)} \\ | & | & & | \end{bmatrix} \in \mathbb{R}^{n \times (p+1)}$$

and $\boldsymbol{X}^{(0)} = (1, \dots, 1)^\top \in \mathbb{R}^n$. Consider first the multiple regression of the j-th explanatory variable on all other p explanatory variables. That is, write the usual LS solution when replacing $\boldsymbol{Y}$ with the vector $\boldsymbol{X}^{(j)}$, and the $n \times (p+1)$ matrix $\boldsymbol{X}$ with the $n \times p$ matrix

$$\boldsymbol{X}^{(-j)} := \begin{bmatrix} | & | & \cdots & | & | & \cdots & | \\ \boldsymbol{X}^{(0)} & \boldsymbol{X}^{(1)} & \cdots & \boldsymbol{X}^{(j-1)} & \boldsymbol{X}^{(j+1)} & \cdots & \boldsymbol{X}^{(p)} \\ | & | & \cdots & | & | & \cdots & | \end{bmatrix} \in \mathbb{R}^{n \times p}.$$

Next, denote by $\boldsymbol{z}^{(j)}$ the residuals from this regression, i.e., $\boldsymbol{z}^{(j)} = \left(\boldsymbol{I}_n - \boldsymbol{P}_{\mathrm{Im}\left(\boldsymbol{X}^{(-j)}\right)}\right)\boldsymbol{X}^{(j)}$. Finally, consider the simple regression, without intercept, of $\boldsymbol{Y}$ on $\boldsymbol{z}^{(j)}$, and denote the resulting LS coefficient by $\hat{c}^{(j)}$. Then it can be shown that

$$\hat{\beta}_j = \hat{c}^{(j)}.$$

To summarize in words, the LS estimator $\hat{\beta}_j$ for the multiple regression of $\boldsymbol{Y}$ on $\boldsymbol{X}$, is given by the simple regression, without intercept, of $\boldsymbol{z}^{(j)}$ on $\boldsymbol{X}^{(-j)}$. Since this amounts to substituting $\boldsymbol{z}^{(j)}$ for $\boldsymbol{X}$ in (11.5), we get

$$\hat{\beta}_j = \hat{c}^{(j)} = \frac{\boldsymbol{z}^{(j)\top}\boldsymbol{Y}}{\|\boldsymbol{z}^{(j)}\|^2}$$

The variance of this estimator is

$$\mathrm{Var}\left(\hat{\beta}_j\right) = \mathrm{Var}\left(\frac{\boldsymbol{z}^{(j)\top}\boldsymbol{Y}}{\|\boldsymbol{z}^{(j)}\|^2}\right) = \mathrm{cov}\left(\frac{\boldsymbol{z}^{(j)\top}\boldsymbol{Y}}{\|\boldsymbol{z}^{(j)}\|^2}\right) = \sigma^2 \frac{\boldsymbol{z}^{(j)\top}\boldsymbol{z}^{(j)}}{\left(\boldsymbol{z}^{(j)\top}\boldsymbol{z}^{(j)}\right)^2} = \sigma^2 \frac{1}{\|\boldsymbol{z}^{(j)}\|^2}$$

(Remark: note that this means $\left[\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\right]_{jj} = \frac{1}{\|\boldsymbol{z}^{(j)}\|^2}\Bigg)$. Now, if the columns of $\boldsymbol{X}$ are nearly linearly dependent, this means that the squared norm of $\boldsymbol{z}^{(j)} = \left(\boldsymbol{I}_n - \boldsymbol{P}_{\mathrm{Im}\left(\boldsymbol{X}^{(-j)}\right)}\right)\boldsymbol{X}^{(j)}$ will be small, remembering that the residuals in the LS method are the residuals from the best (in terms of squared error) approximation of the response as a linear combination of the explanatory variables. This implies that $\mathrm{Var}\left(\hat{\beta}_j\right)$ will be large.

Basic checks for multicollinearity.

1. Look at the Pearson correlations (pairwise correlations) for all pairs of explanatory variables. High absolute values are a sign of redundancy.

2. For each $j = 0, 1, \ldots, p$, look at the R^2 value in the regression of the j-the explanatory variable $\boldsymbol{X}^{(j)} = (X_{1j}, \ldots, X_{nj})^\top$ on the remaining p explanatory variables (=columns of $\boldsymbol{X}$). If we denote this by
$$R_j^2 := \frac{SSR_j}{SST_j}$$
then large values of R_j^2 means that $\boldsymbol{X}^{(j)}$ can be approximated with high accuracy as a linear combination of the others, ie., the residuals of the simple regression of $\mathbf{Y}$ on $\boldsymbol{z}^{(j)}$ (=the residuals from regressing $\boldsymbol{X}^{(j)}$ on $\boldsymbol{X}^{(-j)}$) are small. Two standard metrics for measuring R_j^2 are the Tolerance and the Variance Inflation Factor,
$$\mathrm{Tol}_j := 1 - R_j^2 \quad \text{and} \quad \mathrm{VIF}_j := \frac{1}{1 - R_j^2},$$
Respectively. Hence, small values of Tol_j, or large values of VIF_j, are indication for a problem (redundancy).

   The Variance Inflation Factor gets its name from the fact that
$$V = \sigma^2 F, \quad F = \sum_{i=1}^{n} \left(X_{ij} - \bar{X}_{j}.\right)^2$$
is the variance of $\hat{\beta}_j$ in a simple regression (with intercept) of $\boldsymbol{Y}$ *on all predictors except for* $\boldsymbol{X}^{(j)}$, and
$$V_* = \sigma^2 F_B, \quad F_* = \sum_{i=1}^{n} \left(X_{ij}^* - \bar{X}_j^*\right)^2$$
is the variance of $\hat{\beta}_j$ in the usual multiple regression of $\boldsymbol{Y}$ on $\boldsymbol{X}$, then
$$\mathrm{VIF}_j = \frac{V_*}{V}$$
As a rough guideline, R_j^2 values exceeding 0.85 , i.e. Tol_j falling below 0.15 or VIF_j exceeding 6.6, can be considered extreme (indicating substantial redundancy).

3. Condition number and condition index. The condition number of a matrix $\boldsymbol{X}$ (whose columns are not exactly linearly dependent) is defined by
$$\gamma(\boldsymbol{X}) := \frac{\max_{\|c\|=1} \|\boldsymbol{X}\boldsymbol{c}\|}{\min_{\|\boldsymbol{c}\|=1} \|\boldsymbol{X}\boldsymbol{c}\|}$$
Large values of $\gamma(\boldsymbol{X})$ indicate higher degree of redundancy (the restriction $\|\boldsymbol{c}\| = 1$ keeps the numerator and denominator calibrated); indeed, in that case the denominator is approximately zero. The smallest possible value for $\gamma(\boldsymbol{X})$ is 1, which obtains when the column of $\boldsymbol{X}$ are orthogonal with the same norm, i.e., $\boldsymbol{X}^\top \boldsymbol{X} = \lambda \boldsymbol{I}_{p+1}$. As a rough guideline, we can consider values of $\gamma(\boldsymbol{X})$ between 5-10 as low degree of multicollinearity, and between $30 - 100$ as high degree of multicollinearity. It

can be shown, by considering the diagonal representation of the positive-definite matrix $\boldsymbol{X}^\top \boldsymbol{X}$, that the numerator in (11.6) is equal to the largest eigenvalue of $\boldsymbol{X}^\top \boldsymbol{X}$, and the denominator in (11.6) to the smallest eigenvalue, so that

$$\gamma(\boldsymbol{X}) := \left(\frac{\lambda_{\max}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)}{\lambda_{\min}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)} \right)^{1/2}$$

More generally, we we define the condition index corresponding to the j-th eigenvalue λ_j to be

$$\alpha_j := \left(\frac{\lambda_{\max}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)}{\lambda_j\left(\boldsymbol{X}^\top \boldsymbol{X}\right)} \right)^{1/2} .$$

Since $\lambda_j = \|\boldsymbol{X}\boldsymbol{u}_j\|$ where $\boldsymbol{u}_j$ is a unit vector in the direction of the j-th eigenvalue (equivalently, the j th column of a matrix $\boldsymbol{U}$ holding an orthonormal diagonalizing basis), small values of α_j indicate ”directions” with substantial redundancy, and we can identify explanatory variables $\boldsymbol{X}^{(j)}$ exhibiting redundancy by looking at the entries of the corresponding $\boldsymbol{u}_j$ that have the largest coefficients.

For example, if in a case with 3 explanatory variables (+intercept) the eigenvalues of $\boldsymbol{X}^\top \boldsymbol{X}$ are

$$400.0565138 \quad 200.0000000 \quad 199.7850065 \quad 0.1584797$$

with corresponding condition indices

$$1.000000 \quad 1.414313 \quad 1.415074 \quad 50.242803,$$

then λ_{max} is very large compared to λ_{min}, indicating near linear dependence in the linear combination corresponding to the eigenvector of the smallest eigenvalue. The eigenvectors are

```
> eig$vectors
            [,1] [,2]       [,3]         [,4]
[1,]  0.00000000    1 0.00000000  0.000000000
[2,] -0.70670271    0 0.02461521  0.707082292
[3,] -0.70674878    0 0.02180557 -0.707128479
[4,]  0.03282445    0 0.99945916 -0.001986711
```

The last eigenvector (corresponding to $\lambda_{\min}$) is the problematic ”direction”, and we see that it is approximately equal to $.707\boldsymbol{X}^{(1)} - .707\boldsymbol{X}^{(2)}$, equivalently to $\boldsymbol{X}^{(1)} - \boldsymbol{X}^{(2)}$, indicating that $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ are nearly linearly dependent.

4. Proportion of variance table. Recall that

$$\operatorname{cov}\left(\hat{\boldsymbol{\beta}}_j\right) = \sigma^2 \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}$$

Now, if $\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ is a spectral decomposition of $\boldsymbol{X}^\top \boldsymbol{X}$, then $\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} = \boldsymbol{U}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^\top$ is a spectral decomposition of $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$, and we have

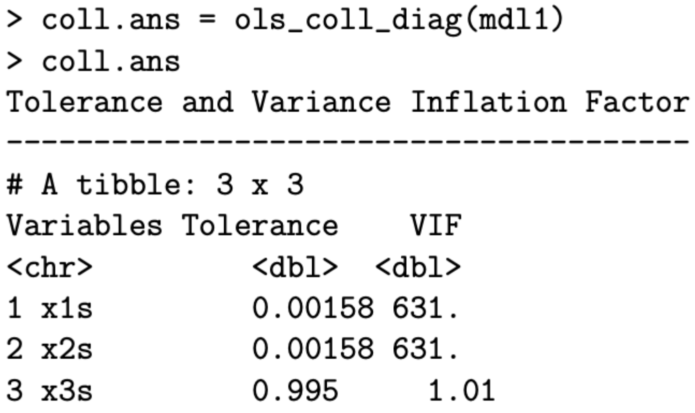
$$\operatorname{Var}\left(\hat{\beta}_j\right) = [\operatorname{Cov}(\hat{\beta})]_{jj} = \sum_r \sum_s U_{jr} \left(\Lambda^{-1}\right)_{rs} \left(U^T\right)_{sj} = \sum_r \lambda_r^{-1} U_{jr}^2$$

The quantity

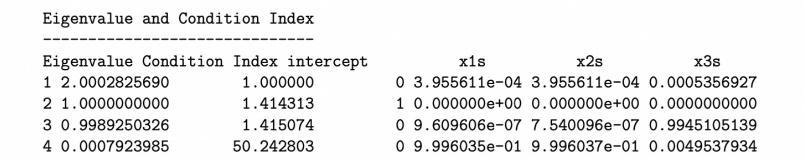
$$\Pi_{rj} = \frac{\lambda_r^{-1} U_{jr}^2}{\sum_s \lambda_s^{-1} U_{js}^2}$$

is the *proportion of* $\mathrm{Var}\left(\hat{\beta}_j\right)$ contributed by the "direction" (=linear combination of the original explanatory variables $\boldsymbol{X}^{(j)}\Big)$ corresponding to λ_r, i.e., that represented by the eigenvector $\boldsymbol{u}_r$. For a small value λ_r, we can identify "problematic" combinations by finding j 's with large value of Π_{rj}.

In R, Tol and VIF, and variance proportions, can be calculated automatically using the function `ols_coll_diag` in the R package `olsr`. If I understand correctly, the package first normalizes all explanatory variables so that the diagonal entries of $\boldsymbol{X}^\top \boldsymbol{X}$ are all 1's, then diagonalizes the resulting matrix. Example:



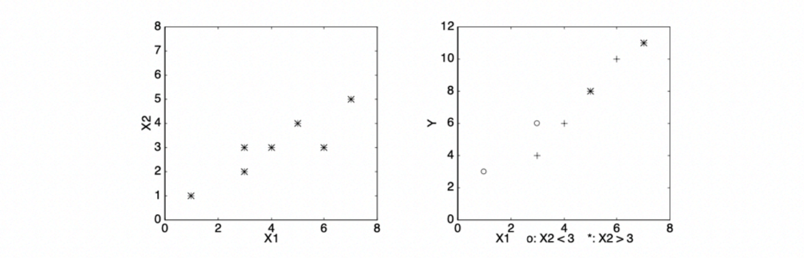
```
> coll.ans = ols_coll_diag(mdl1)
> coll.ans
Tolerance and Variance Inflation Factor
---------------------------------------
# A tibble: 3 x 3
Variables Tolerance    VIF
<chr>         <dbl>  <dbl>
1 x1s         0.00158 631.
2 x2s         0.00158 631.
3 x3s         0.995     1.01
```



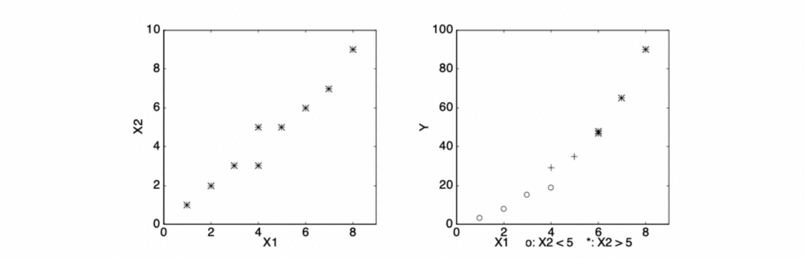
```
Eigenvalue and Condition Index
------------------------------
Eigenvalue Condition Index intercept          x1s          x2s          x3s
1 2.0002825690        1.000000        0 3.955611e-04 3.955611e-04 0.0005356927
2 1.0000000000        1.414313        1 0.000000e+00 0.000000e+00 0.0000000000
3 0.9989250326        1.415074        0 9.609606e-07 7.540096e-07 0.9945105139
4 0.0007923985       50.242803        0 9.996035e-01 9.996037e-01 0.0049537934
```

We see that almost 100% of the variance in $\hat{\beta}_1$ and $\hat{\beta}_2$ come from the term with the low eigenvalue, thus indicating a multicollinearity problem.

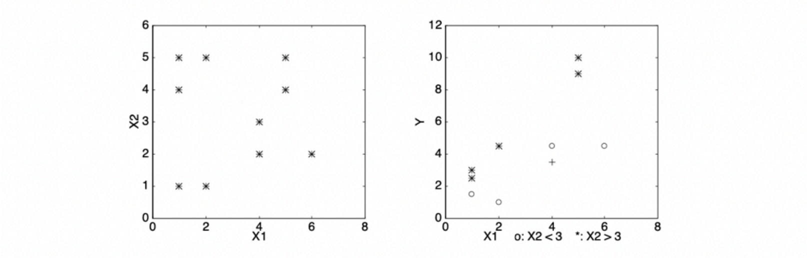
Visual checks for multicollinearity and interactions. We give some illustrating examples for how multicollinearity and interaction each look visually in graphs (credit to Prof. Sam Oman).



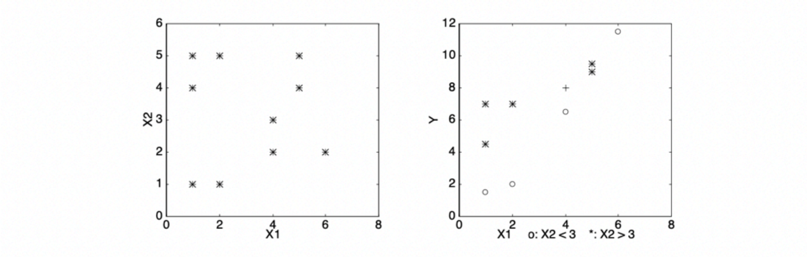
Set A: multicollinearity, no interaction



Set B: multicollinearity, interaction



Set C: no multicollinearity, interaction



Set D: no multicollinearity, no interaction

Note: Sets C, D have the same values.

10 Variable selection

For a given set of measured explanatory variables, $\boldsymbol{X}_i = (X_{i1}, \dots, X_{ip})$, we have so far considered fitting the full model,

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{P}\boldsymbol{Y}, \quad \hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{Y}, \quad \boldsymbol{P} = \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top$$

Instead, we may consider fitting a sub-model consisting of only a subset of explanatory variables. Thus, for a subset $\Omega \subseteq \{1, \dots, p\}$ of size $r := |\Omega|$ (in particular, $r = 0$ for $\Omega = \varnothing$), denote by

$$\boldsymbol{X}_\Omega = \left[\boldsymbol{X}^{(j)} : j \in \Omega \cup \{0\}\right]$$

the $n \times (r+1)$ matrix obtained from $\boldsymbol{X}$ by keeping only the columns $\boldsymbol{X}^{(j)}$ with $j \in \Omega$, and consider fitting on this sub-model,

$$\hat{\boldsymbol{Y}}_\Omega = \boldsymbol{X}_\Omega\hat{\boldsymbol{\beta}}_\Omega = \boldsymbol{P}_\Omega\boldsymbol{Y}, \quad \hat{\boldsymbol{\beta}}_\Omega = \left(\boldsymbol{X}_\Omega^\top \boldsymbol{X}_\Omega\right)^{-1} \boldsymbol{X}_\Omega^\top \boldsymbol{Y}, \quad \boldsymbol{P}_\Omega = \boldsymbol{X}_\Omega\left(\boldsymbol{X}_\Omega^\top \boldsymbol{X}_\Omega\right)^{-1} \boldsymbol{X}_\Omega^\top.$$

There are generally two different reasons why we may be interested in comparing the full model to a submodel:

1. *Prediction*: compare sub-models in terms of the prediction performance on a new and independent sample from the same (true) model
2. *Inference*: test whether $\beta_j = 0$ for all $j \in \{1, \dots, p\}\backslash\Omega$.

Selecting a submodel from a prediction prespective. Assume the full model,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \mathbb{E}[\boldsymbol{\epsilon}] = \boldsymbol{0}, \quad \text{cov}[\boldsymbol{\epsilon}] = \sigma^2 \boldsymbol{I}_n \tag{34}$$

and let $\boldsymbol{Y}^* = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}^*$ be a fresh, independent realization from (34), i.e., $\boldsymbol{\epsilon}^*$ and $\boldsymbol{\epsilon}$ are iid. The goal is to predict $\boldsymbol{Y}^*$ using the observed vector $\mathbf{Y}$. The natural predictor is the LS predictor from the full model, $\hat{\mathbf{Y}} = \boldsymbol{P}\mathbf{Y}$ for $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$ (which is also the quantity we based a *prediction interval* on), but more generally we may consider the LS predictor corresponding to a submodel, $\hat{\boldsymbol{Y}}_\Omega = \boldsymbol{P}_\Omega\boldsymbol{Y}$. For any Ω define the mean squared prediction error of $\hat{\mathbf{Y}}_\Omega$ to be

$$MSPE(\Omega) := \mathbb{E}\|\boldsymbol{Y}^* - \hat{\boldsymbol{Y}}_\Omega\|^2.$$

We have

$$\begin{aligned}
MSPE(\Omega) &:= \mathbb{E}\|\boldsymbol{Y}^* - \hat{\boldsymbol{Y}}_\Omega\|^2 \\
&= \mathbb{E}\|\left(\boldsymbol{Y}^* - \boldsymbol{Y}\right) + \left(\boldsymbol{Y} - \hat{\boldsymbol{Y}}_\Omega\right)\|^2 \\
&= \mathbb{E}\|\boldsymbol{Y}^* - \boldsymbol{Y}\|^2 + \mathbb{E}\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}_\Omega\|^2 + 2\mathbb{E}\left[(\boldsymbol{Y}^* - \boldsymbol{Y})^\top\left(\boldsymbol{Y} - \hat{\boldsymbol{Y}}_\Omega\right)\right] \\
&\overset{(a)}{=} \mathbb{E}\|\boldsymbol{Y}^* - \boldsymbol{Y}\|^2 + \mathbb{E}\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}_\Omega\|^2 + 2\,\text{tr}\left[\text{cov}\left(\boldsymbol{Y} - \hat{\boldsymbol{Y}}_\Omega, \boldsymbol{Y}^* - \boldsymbol{Y}\right)\right] \\
&= \mathbb{E}\|\boldsymbol{Y}^* - \boldsymbol{Y}\|^2 + \mathbb{E}\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}_\Omega\|^2 + 2\,\text{tr}\left[\sigma^2\boldsymbol{P}_{\boldsymbol{\Omega}} - \sigma^2\boldsymbol{I}_n\right] \\
&= 2\sigma^2 n + \mathbb{E}\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}_\Omega\|^2 + 2\sigma^2 r - 2\sigma^2 n \\
&= \mathbb{E}\,\underbrace{\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}_\Omega\|^2}_{:=SSE(\Omega)} + 2\sigma^2 r,
\end{aligned}$$

Above, (a) uses the general fact that $\boldsymbol{Z}^\top \boldsymbol{W} = \operatorname{tr}(\operatorname{cov}[\boldsymbol{Z}, \boldsymbol{W}])$ if $\mathbb{E}[\boldsymbol{Z}] = \mathbf{0}$ or $\mathbb{E}[\boldsymbol{W}] = \mathbf{0}$ (or both), applied to the vectors $\boldsymbol{Z} = \boldsymbol{Y}^* - \boldsymbol{Y}$ (that has expectation zero), and $\boldsymbol{W} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}_\Omega$ (we've seen this result before for the special case $\boldsymbol{W} = \boldsymbol{Z}$; prove the more general statement by adjusting the proof we saw in class for the special case).

What the calculation above says is that the *in-sample* predictor error, $SSE(\Omega) = \|\boldsymbol{Y} - \hat{\boldsymbol{Y}}_\Omega\|^2$, for any partial model (including the full model $\Omega = \{1, \ldots, p\}$), is an under-estimate of the *out-of-sample* prediction error,

$$\underbrace{\mathbb{E}\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}_\Omega\|^2}_{\mathbb{E}[SSE(\Omega)]} = \underbrace{\mathbb{E}\|\boldsymbol{Y}^* - \hat{\boldsymbol{Y}}_\Omega\|^2}_{MSPE(\Omega)} - 2\sigma^2 r,$$

and the larger the size of Ω, the more it is biased downward. Intuitively, this is because $\hat{\boldsymbol{Y}}_\Omega$ is fitted to (trained on) $\mathbf{Y}$, hence the error on the left hand side is indeed expected to be smaller than on the right hand side. Note that, of course, SSE for the full model is smaller than $SSE(\Omega)$, but $MSPE(\Omega) = \mathbb{E}\left[SSE(\Omega) - 2\sigma^2 r\right]$ might actually be larger for the full model!

Goodness of fit measures. The calculation above motivates an adjustment to the usual goodness-of-fit measure,

$$R^2(\Omega) = 1 - \frac{SSE(\Omega)}{SST}$$

(for $\Omega = \{1, \ldots, p\}$ we get the usual quantity R^2 for the full model). Indeed, $R^2(\Omega)$ calculated for any sub-model Ω is decreasing in $SSE(\Omega)$, hence, if we compare any two nested models $\Omega \subseteq \Omega'$, we'll always get $R^2(\Omega) \leq R^2(\Omega')$, but we already saw that this does not necessarily imply that $MSPE(\Omega') \leq MSPE(\Omega)$. This calls for an adjustment to $SSE(\Omega)$ that penalizes for the size (=number of variables) of the sub-model Ω, consistent with (13.4).

A simple correction of the R^2 statistic is the so-called Adjusted- R^2 statistic,

$$\text{adjusted-} R^2(\Omega) := 1 - \frac{SSE(\Omega)/(n-r-1)}{SST/(n-1)} = 1 - \left(\frac{n-1}{n-r-1}\right)\left(1 - R^2(\Omega)\right).$$

Comparing any two nested models $\Omega \subseteq \Omega'$, the adjusted- R^2 can either increase or decrease, because $1 - R^2(\Omega') \leq 1 - R^2(\Omega)$, but the factor $(n-1)/(n-r-1)$ is larger for Ω'. In practice, it is often the case that SSE decreases sharply as the first few "important" variables are entered into the model, then levels off as we move on to variables of lesser importance.

There are alternative measures of goodness-of-fit to the adjusted- R^2, all of which employ penalty for the "complexity" of the model - the number of variables included. One common alternative is Mallows's C_p statistic,

$$C_p = \frac{SSE_\Omega}{SSE/(n-p-1)} - (n - 2(r+1)).$$

Small values of the C_p statistic indicate better fit. i.e., a model consisting of a certain subset of variables is considered better than another model consisting of another subset, if the C_p statistic for the former is smaller than that for the latter.

Two other "classical" measures for goodness-of-fit are AIC (the Akaike Information Criterion) and BIC (the Bayesian Information Criterion); we'll skip the description here, but you can find many resources online. We move on to discussing model selection for inference, in which we treat the problem of variable selection as the problem of identifying nonzero coefficients β_j. Before we proceed, we mention that this "inference" goal is of interest when one wants to know which explanatory variables are "relevant" to the response; for example, in genetics the explanatory variables $X^{(j)}$ might represent different genetic markets, the response some measure for disease expression, and we may want to know which markers are associated with the disease. Importantly, this is a distinct goal from prediction: even if $X^{(j)}$ has a nonzero coefficient β_j, if only prediction is of interest, it might still be worthwhile to leave $X^{(j)}$ out of the model if β_j is small enough in magnitude.

The F-test for a submodel. For some $\Omega \subseteq \{1, ..., p\}$ of size $|\Omega| = r$, and under the normal linear model (note that normality was not assumed in the preceding discussion), suppose that we want to formally test the null hypothesis

$$H_0 : \beta_j = 0 \text{ for all } j \in \{1, ..., p\} \setminus \Omega. \tag{35}$$

i.e., that the coefficients $\beta_j, j \in \{1, \ldots, p\} \backslash \Omega$ are all zero. (note: by convention, we always keep an intercept term, so $\Omega \subseteq \{1, \ldots, p\}$, not $\Omega \subseteq \{0, 1, \ldots, p\}$). Previously in the course, we considered the special case where $r = p - 1$ and Ω^c was a singleton, i.e., when we tested $H_0 : \beta_j = 0$ for a single index j. Recall that in this case the t-test rejects for large absolute values of the t-statistic,

$$T_j^0 = \left[\widehat{\operatorname{Var}\left(\hat{\beta}_j\right)}\right]^{-1/2} \hat{\beta}_j$$

where

$$\widehat{\operatorname{Var}(\hat{\beta}_j)} = \hat{\sigma}^2 \left[\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\right]_{jj} = \frac{\hat{\sigma}^2}{\sigma^2} \operatorname{Var}\left(\hat{\beta}_j\right)$$

This is equivalent to rejecting for large values of

$$(T_j^0)^2 = [\widehat{\operatorname{Var}(\hat{\beta}_j)}]^{-1} \hat{\beta}_j^2 \tag{36}$$

Up to the "hat" on the covariance, and under the null, note that T_j^0 is a standardized version of $\hat{\beta}_j$; with the "hat", this is called sometimes a "studentized" version of $\hat{\beta}_j$. If we write

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_{(1)} \\ \boldsymbol{\beta}_{(2)} \end{bmatrix}$$

with $\boldsymbol{\beta}_{(1)} = (\beta_0, \beta_1, \ldots, \beta_r)^\top$ and $\boldsymbol{\beta}_{(2)} = (\beta_{r+1}, \ldots, \beta_{p+1})^\top$, then, without loss of generality, we can write (35) as

$$H_0 : \boldsymbol{\beta}_{(2)} = \mathbf{0}$$

i.e., that the last $p - r$ coordinates of $\boldsymbol{\beta}$ are zero. As an extension of (36) for the single variable case, consider the statistic

$$W := \hat{\boldsymbol{\beta}}_{(2)}^\top \left[\widehat{\operatorname{cov}\left(\hat{\boldsymbol{\beta}}_{(2)}\right)}\right]^{-1} \hat{\boldsymbol{\beta}}_{(2)}$$

where

$$\widehat{\operatorname{cov}}\left(\hat{\boldsymbol{\beta}}_{(2)}\right) = \hat{\sigma}^2 \left(X^\top \boldsymbol{X}\right)^{-1} = \frac{\hat{\sigma}^2}{\sigma^2} \cdot \operatorname{cov}\left(\hat{\boldsymbol{\beta}}_{(2)}\right) \tag{37}$$

and $\hat{\sigma}^2 = \|\boldsymbol{e}\|^2/(n - p - 1)$. A statistic of the form (37), the generalized version of the squared studentized statistic, is usually called a *Wald statistic*. We now derive its distribution under H_0. We will need the following results in the development.

Definition 8. If $V_1 \sim \chi^2_{d_1}$ and $V_2 \sim \chi^2_{d_2}$ are independent, then the distribution of

$$F := \frac{V_1/d_1}{V_2/d_2}$$

is called the F distribution with d_1 and d_2 degrees of freedom (numerator and denominator, respectively), and we denote $F \sim \mathcal{F}_{d_1, d_2}$.

Proposition. If $\boldsymbol{Z} \sim \mathcal{N}_m(\mathbf{0}, \boldsymbol{V})$, then $\boldsymbol{Z}^\top \boldsymbol{V}^{-1} \boldsymbol{Z} \sim \chi_m^2$ (assume $\boldsymbol{V}$ has full rank m)

Proof. Since $\boldsymbol{V}$ is of full rank, it is positive definite, and we can write $\boldsymbol{V} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\top$ with $\boldsymbol{U}_{m\times m}$ orthogonal and $\boldsymbol{D}_{m\times m}$ diagonal with all diagonal entries positive. Then $\boldsymbol{V} = \boldsymbol{A}\boldsymbol{A}^\top$ for $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}^{1/2}$, and $\boldsymbol{V}^{-1} = \left(\boldsymbol{A}^{-1}\right)^\top \boldsymbol{A}^{-1}$. Therefore,

$$\boldsymbol{Z}^\top \boldsymbol{V}^{-1} \boldsymbol{Z} = \boldsymbol{Z}^\top \left(\boldsymbol{A}^{-1}\right)^\top \boldsymbol{A}^{-1} \boldsymbol{Z} = \left\|\boldsymbol{A}^{-1}\boldsymbol{Z}\right\|^2. \tag{38}$$

But $\mathbb{E}\left[\boldsymbol{A}^{-1}\boldsymbol{Z}\right] = \boldsymbol{A}^{-1}\mathbb{E}[\boldsymbol{Z}] = \mathbf{0}$, and

$$\begin{aligned}\operatorname{cov}\left(\boldsymbol{A}^{-1}\boldsymbol{Z}\right) = \boldsymbol{A}^{-1}\operatorname{cov}(\boldsymbol{Z})\left(\boldsymbol{A}^{-1}\right)^\top = \boldsymbol{A}^{-1}\boldsymbol{V}\left(\boldsymbol{A}^{-1}\right)^\top &= \boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{A}^\top\left(\boldsymbol{A}^{-1}\right)^\top \\ &= \boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{A}^\top\left(\boldsymbol{A}^\top\right)^{-1} = \boldsymbol{I},\end{aligned}$$

so that

$$\boldsymbol{A}^{-1}\boldsymbol{Z} \sim \mathscr{N}_m(\mathbf{0}, \boldsymbol{I}). \tag{39}$$

The result follows from (38)+(39), and the definition of the χ_m^2 distribution. □

Recall that, under the normal linear model,

$$\hat{\boldsymbol{\beta}} \sim \mathscr{N}_{p+1}(\boldsymbol{\beta}, \operatorname{cov}(\hat{\boldsymbol{\beta}}))$$

In particular,

$$\hat{\boldsymbol{\beta}}_{(2)} \sim \mathscr{N}_{p-r}\left(\boldsymbol{\beta}_{(2)}, \operatorname{cov}\left(\hat{\boldsymbol{\beta}}_{(2)}\right)\right)$$

By (37), we have

$$\left[\widehat{\operatorname{cov}\left(\hat{\boldsymbol{\beta}}_{(2)}\right)}\right]^{-1} = \left(\hat{\sigma}^2/\sigma^2\right)^{-1}\left[\operatorname{cov}\left(\hat{\boldsymbol{\beta}}_{(2)}\right)\right]^{-1}.$$

Therefore, we can write

$$W = \hat{\boldsymbol{\beta}}_{(2)}^\top \left[\widehat{\operatorname{cov}\left(\hat{\boldsymbol{\beta}}_{(2)}\right)}\right]^{-1} \hat{\boldsymbol{\beta}}_{(2)} = \frac{\hat{\boldsymbol{\beta}}_{(2)}^\top \left[\operatorname{cov}\left(\hat{\boldsymbol{\beta}}_{(2)}\right)\right]^{-1} \hat{\boldsymbol{\beta}}_{(2)}}{\hat{\boldsymbol{\sigma}}^2/\sigma^2} = \frac{V_1}{V_2/(n-p-1)},$$

where

$$V_1 = \hat{\boldsymbol{\beta}}_{(2)}^\top \left[\operatorname{cov}\left(\hat{\boldsymbol{\beta}}_{(2)}\right)\right]^{-1} \hat{\boldsymbol{\beta}}_{(2)} \quad \text{and} \quad V_2 = (n-p-1)\hat{\sigma}^2/\sigma^2.$$

Now, under $H_0 : \boldsymbol{\beta}_{(2)} = \mathbf{0}$, we have

$$\hat{\boldsymbol{\beta}}_{(2)} \sim \mathcal{N}_{p-r}\left(\mathbf{0}, \operatorname{cov}\left(\hat{\boldsymbol{\beta}}_{(2)}\right)\right).$$

By the proposition above, $V_1 \overset{H_0}{\sim} \chi_{p-r}^2$. also, we saw previously that

$$V_2 \sim \chi_{n-p-1}^2$$

and that $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are independent, hence so are V_1 (which is a function of $\hat{\boldsymbol{\beta}}$) and V_2 (which is a function of $\hat{\sigma}^2$). Putting everything together, if we define

$$F = W^0/(p-r) = \frac{V_1/(p-r)}{V_2/(n-p-1)}, \tag{40}$$

then

$$F \overset{H_0}{\sim} \mathcal{F}_{p-r,n-p-1} \ ,$$

and a level-α test rejects if $F \geq f_{1-\alpha;\ p-r,n-p-1}$, ie, if the test statistic F exceeds the $1-\alpha$ quantile of the $\mathcal{F}_{p-r,n-p-1}$ distribution.

Equivalent forms for the F statistic.

Proposition 9. *The F statistic* (40) *has the following equivalent forms:*

$$F = \frac{\left\|\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_{(1)}\right\|^2 /(p-r)}{\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2/(n-p-1)}$$

$$F = \frac{(SSE_1 - SSE)/(p-r)}{SSE/(n-p-1)}$$

where: $\hat{\boldsymbol{Y}}$ $=$ *vector of predicted values for the full model* $\hat{\boldsymbol{Y}}_{(1)}$ $=$ *vector of predicted values for the partial model* SSE $=$ *sum-of-squares error for the full model* SSE_1 $=$ *sum-of-squares error for the partial model*

A proof of the equivalence to the statistic (40) will be posted in a separate file to the course website. Here we only show the equivalence of the two forms in the proposition above. Thus, we only need to prove that

$$\|\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_{(1)}\|^2 = SSE_1 - SSE.$$

To see that this indeed holds, write

$$\underbrace{\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}_{(1)}\|^2}_{SSE_1} = \left\|(\boldsymbol{Y} - \hat{\boldsymbol{Y}}) + \left(\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_{(1)}\right)\right\|^2 = \underbrace{\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2}_{SSE} + \|\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_{(1)})\|^2$$

where the last equality is because $\boldsymbol{Y} - \hat{\boldsymbol{Y}} \in \operatorname{Im}^{\perp}(\boldsymbol{X})$ is orthogonal to $\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_{(1)} \in \operatorname{Im}(\boldsymbol{X})$: indeed, if we denote by $\boldsymbol{X}_{(1)}$ the matrix obtained from $\boldsymbol{X}$ by keeping only the first $r+1$ columns, then it is clear that $\hat{\boldsymbol{Y}}_{(1)} \in \operatorname{Im}\left(\boldsymbol{X}_{(1)}\right) \subseteq \operatorname{Im}(\boldsymbol{X})$; in addition, we obviously have $\hat{\boldsymbol{Y}} \in \operatorname{Im}(\boldsymbol{X})$. I.e., both $\hat{\boldsymbol{Y}}, \hat{\boldsymbol{Y}}_{(1)} \in \operatorname{Im}(\boldsymbol{X})$, so $\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_{(1)} \in \operatorname{Im}(\boldsymbol{X})$ as well (a linear space is closed). The figure below illustrates the identity geometrically (Pythagoras in the highlighted triangle).

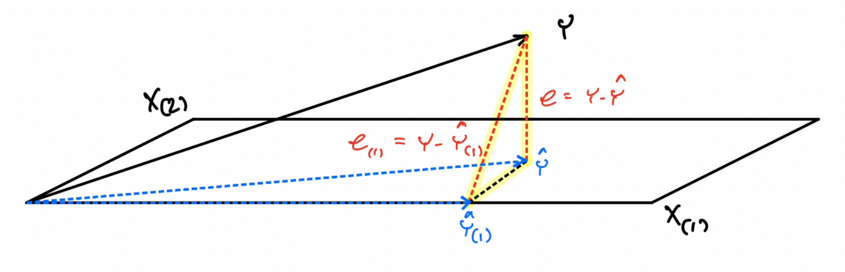


Figure 17: Sum-of-squares decomposition for the F-statistic

Application of the F-test. We present a few examples for the application of the F-test.

Global testing. The ”global test” is the test that all coefficients, except β_0, are zero:

$$H_0 : \beta_1 = \ldots = \beta_p = 0$$

In this case $r = 0$, and we have $\hat{\boldsymbol{Y}}_{(1)} = \bar{Y}\mathbf{1}_n$ and $SSE_{(1)} = SST$, hence

$$F = \frac{(SST - SSE)/p}{SSE/(n-p-1)} = \frac{SST - SSE}{SSE} = \frac{1-(SSE/SST)}{SSE/SST}\frac{n-p-1}{p} = \frac{R^2}{1-R^2}\frac{n-p-1}{p}$$

Example. We consider a dataset analyzed in Daniel (1999) to compare the effectiveness of three treatments for severe depression.

```
# F test

df = read.table(’/Users/asaf/Dropbox/Teaching/52571_Regression/
52571_2022-23/datasets/depression.txt’,sep=’\t’, header = TRUE)

> df <- df[,c(’y’,’age’,’TRT’)]
> head(df)
   y age TRT
1 56  21   A
2 41  23   B
3 40  30   B
4 28  19   C
5 55  28   A
6 25  23   C
>
> str(df)  #’TRT’ is of type "character"
’data.frame’: 36 obs. of  3 variables:
 $ y  : int  56 41 40 28 55 25 46 71 48 63 ...
 $ age: int  21 23 30 19 28 23 33 67 42 33 ...
 $ TRT: chr  "A" "B" "B" "C" ...
>
> df$TRT <- as.factor(df$TRT)  #change variable ’TRT’ to type "factor"
> str(df) # ’y’ and ’age’ are of type "integer", ’TRT’ is of type "factor"
’data.frame’: 36 obs. of  3 variables:
 $ y  : int  56 41 40 28 55 25 46 71 48 63 ...
 $ age: int  21 23 30 19 28 23 33 67 42 33 ...
 $ TRT: Factor w/ 3 levels "A","B","C": 1 2 2 3 1 3 2 3 2 1 ...
>
> fm <- lm(y˜age*TRT, data = df)
> summary(fm)

Call:
lm(formula = y ˜ age * TRT, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4366 -2.7637  0.1887  2.9075  6.5634
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  47.51559    3.82523  12.422 2.34e-13 ***
age           0.33051    0.08149   4.056 0.000328 ***
TRTB        -18.59739    5.41573  -3.434 0.001759 **
TRTC        -41.30421    5.08453  -8.124 4.56e-09 ***
age:TRTB      0.19318    0.11660   1.657 0.108001
age:TRTC      0.70288    0.10896   6.451 3.98e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.925 on 30 degrees of freedom
Multiple R-squared:  0.9143,Adjusted R-squared:  0.9001
F-statistic: 64.04 on 5 and 30 DF,  p-value: 4.264e-15

>
> n <- 36
> p <- 5
> rsquared <- summary(fm)$r.squared
> sse <- summary(fm)$sigma^2*(n-p-1)
> sst <- sse * (1/(1-rsquared))
>
> # H_0: beta_1=...=beta_p=0 (global test)
> fstat <- ((sst-sse)/p) / (sse/(n-p-1))
> # alternative calculation
> fstat.alt <- rsquared/(1-rsquared) * (n-p-1)/p
> fstat==fstat.alt #verify equality
[1] TRUE
>
> # H_0: beta_4=beta_5=0 (no interaction)
> fm.additive <- lm(y~age+TRT, data=df)
> summary(fm.additive)

Call:
lm(formula = y ~ age + TRT, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-12.5732  -3.3922   0.9829   3.9613   9.5062

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  32.54335    3.58105   9.088 2.23e-10 ***
age           0.66446    0.06978   9.522 7.42e-11 ***
TRTB         -9.80758    2.46471  -3.979 0.000371 ***
TRTC        -10.25276    2.46542  -4.159 0.000224 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.035 on 32 degrees of freedom
Multiple R-squared:  0.784,Adjusted R-squared:  0.7637
F-statistic: 38.71 on 3 and 32 DF,  p-value: 9.287e-11

>
> sse.additive <- summary(fm.additive)$sigma^2 * 32
> fstat <- (sse.additive/(32-30))/(sse/30)
> 1-pf(fstat,df1=32-30, df2=30) # compute pvalue: effectively zero
[1] 6.280687e-09
>
```

F-test for the general linear hypothesis. The F-test we've seen above tests $H_0 : \boldsymbol{\beta}_{(2)} = \mathbf{0}$. We are now interested in testing a more general hypothesis,

$$H_0 : \boldsymbol{G}\boldsymbol{\beta} = \mathbf{0}$$

For some fixed $m \times (p+1)$ matrix $\boldsymbol{G}$. Just as an example, the null hypothesis

$$H_0 : \beta_1 = 2\beta_2 \text{ and } \beta_3 = \beta_4 + \beta_1$$

can be written in the form (13.14) as follows:

$$H_0 : \underbrace{\begin{bmatrix} 0 & 1 & -2 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 & -1 & 0 \end{bmatrix}}_{\boldsymbol{G}} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix}}_{\boldsymbol{\beta}} = \mathbf{0}.$$

To obtain a test statistic, let $\boldsymbol{\theta} := \boldsymbol{G}\boldsymbol{\beta} \in \mathbb{R}^m$, and denote the corresponding LS estimator by $\hat{\boldsymbol{\theta}} := \boldsymbol{G}\hat{\boldsymbol{\beta}}$. Now, if we consider the (Wald) statistic

$$F = \hat{\boldsymbol{\theta}}^\top [\widehat{\text{cov}}(\hat{\boldsymbol{\theta}})]^{-1} \hat{\boldsymbol{\theta}}/m$$

then by similar argument to the ones invoked for the development of the original F statistic, we get

$$F \overset{H_0}{\sim} \mathcal{F}_{m,n-p-1}.$$

Automatic variable selection procedures. We return to the general problem of variable selection, not necessarily from the hypothesis testing perspective. In other words, we now want to use the data to select a "promising" set of explanatory variables, e.g. for good prediction performance. There are many automatic variable selection procedures, and we list (versions of) only a few of the most popular ones.

Forward Selection:

0. Start with the minimum model (either the intercept-only model or the model with just the forced-in variables)

1. At each cycle, add the variable that will produce the greatest SSE reduction. Once a variable is added it stays in the model

2. Stop if the variable to be added will not be statistically significant at a specified significance level or will not yield an improvement in a specified model goodness criterion - possible goodness criteria: adjusted-$R^2, C_p, \text{AIC}, \text{BIC}$.

Backward Selection:
0. Start with the model including all the candidate variables.

1. At each cycle, drop the variable whose removal will produce the smallest increase in the SSE.

2. Once a variable has been dropped it stays out of the model

3. Stop if all variables in the current model are statistically significant at a specified level or if the removal will lead to a less favorable value of a specified model goodness criterion (adjusted- R^2, etc).

Stepwise Selection:
0. Start with the minimum model.

1. At each cycle, do the following:

- For each variable not in the model, check the SSE reduction yielded by adding it.
- For each variable in the model, check the SSE increase yielded by dropping it.
- Perform the move that leads to the greatest improvement in the specified model goodness criterion.

Variables dropped previously can be added back, and variables added previously can be dropped.

2. The algorithm stops when there is no one-variable move (addition/deletion) from the current model that produces an improvement in the specified model goodness criterion.

Best Subsets Selection:
0. Specify the following

- `pmax` = maximum number of variables allowed in the model
- M= number of models that you want to examine (Top 5, Top 10, etc)

1. Consider all subsets of the candidate variable pool with pmax or fewer variables. Rank the models according to a selected model goodness criterion.

2. For the top models, report the variables included and the goodness criterion. Typically, there is a collection of models with a similar goodness criterion value, and often the analyst will want to choose the smallest model in this group

There are a number of `R` functions for implementing these algorithms. For example, for forward, backward, and stepwise variable selection, you can use `stepAIC` in the `MASS` package. For best-subset selection, one can use the function `bestglm` in the `MASS` package. The function `bestglm` has the drawback that it does not allow forcing variables into the model, but we can work around this by first adjusting all variables to X_1 and X_2: suppose we have a dataset with variables $X1, X2, X3, X4, X5$ and Y, and we want to select variables for a regression model with response Y with $X1$ and $X2$ forced into the model. We will create a new dataset with variables $X3_{new}$, $X4_{new}$, $X5_{new}$, and Y_{new}, given, respectively, by the residuals of each of the variables $X3, X4, X5$, and Y in a regression on $X1$ and $X2$. We then apply `bestglm` to the new dataset.

11 Analysis of Variance

We turn to another topic, called *analysis of variance* (ANOVA), which is a collection of statistical tools for analyzing the differences between the means of different groups (populations). The connection to our course is that the model we will use for ANOVA is going to be a linear regression model, consisting of only categorical variables.

Let us start with an example. Suppose that we are trying to compare the expected yield (measurement of the amount of a crop grown) for different varieties of tomatoes. For each variety $i = 1, ..., I$, we observe n_i samples $Y_{ij}, j = 1, ..., n_i$, of the yield of tomatoes of that variety, regarded as iid random variables with mean yield μ_i. Based on the $n := n_1 + n_2 + ... + n_I$ observations, we would like to answer questions such as: are the (population) means for all I varieties equal? is there a difference between the mean yield for variety $i = 3$ and $i = 1$? which pairs of varieties, $k, l \in \{1, ..., I\}$ have different means? and so on.

The name "analysis of variance" might sound a bit confusing, because, as described above, we are trying to compare means, not variances. However, the name comes from the fact that the analysis itself will be based on decomposing the (empirical) variance of all sample points $Y_{ij}, i = 1, ..., I, j = 1, ..., n_i$ into two parts: the part attributed to the variability of the individual observation Y_{ij} *within* a particular variety of tomatoes, and the part attributed to the variability of the mean μ_i *between* different varieties. We could be more formal about this: consider the experiment in which I draw a single variety among the I varieties at random, i.e., let J be a random variable distributed uniformly on $1, ..., I$; conditionally on $J = i$, I draw a random variable Y distributed as Y_{ij}, i.e., a random sample of the yield from the ith variety. Define also $M := \mu_J$, the random variable which takes on the value μ_i if $J = i$. Then (Y, M) have a joint distribution, and by the law of total variance,

$$V(Y) = \underbrace{E[V(Y|M)]}_{\text{variance “within” variety}} + \underbrace{V(E[Y|M])}_{\text{variance “between” varieties}}$$

In the example considered the different populations were defined by a single categorical variable, the variety. We could also consider populations defined by two categorical variables, say the variety of the tomatoes and the plant density. The first is called a *one-way* layout, and the corresponding analysis is called a *one-way* ANOVA. The second is a *two-way* layout, and the corresponding analysis is called a *two-way* ANOVA. More generally, if the populations are defined by K categorical variables, we have a case of k*-way* ANOVA.

11.1 One-way ANOVA

We now formalize things and present the one-way layout in terms of the general linear model. For each category $i = 1, ..., I$, we observe n_i iid normal observations of some response variable:

$$Y_{ij} = \mu_i + \epsilon_{ij}, \qquad \epsilon_{ij} \overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2\right), \qquad i = 1, \ldots, I, \quad j = 1, \ldots, n_i. \tag{41}$$

Now, for each $k = 1, ..., I$, define the dummy variables

$$X_{ijk} = \begin{cases} 1, & \text{if observation } ij \text{ is in category } k \\ 0, & \text{otherwise} \end{cases}, \quad i = 1, \ldots, I, \quad j = 1, \ldots, n_i.$$

and note that this is defined now for all I categories (not $I - 1$ categories as we are used to when forming dummy variables). Then (41) can be written

$$Y_{ij} = \mu_1 X_{ij1} + \mu_2 X_{ij2} + \ldots + \mu_I X_{ijI} + \epsilon_{ij}, \qquad \epsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma^2\right), \tag{42}$$

or in vector form

$$\mathbf{Y} = \boldsymbol{X}\boldsymbol{\mu} + \boldsymbol{\epsilon}, \tag{43}$$

with

$$\boldsymbol{X} = [X^{(1)}\ X^{(2)} \cdots\ X^{(I)}] \in \mathbb{R}^{n\times I}$$

for $n := \sum_{i=1}^{I} n_i$, and

$$X^{(k)} := (X_{11k}, ..., X_{1n_1k}, ..., X_{21k}, ..., X_{1n_2k}, ..., X_{I1k}, ..., X_{1n_Ik})^\top \in \mathbb{R}^n, \qquad k = 1, ..., I.$$

Note that (42) has no intercept (but it's still a perfectly legitimate linear model). The first question we want to address is testing if all means are equal. In terms of the model (42), this is testing

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_I. \tag{44}$$

An alternative way to write (41) is as usual, with $I - 1$ dummy variables and an intercept:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \ldots + \beta_{I-1} X_{ij(I-1)} + \epsilon_{ij}, \qquad \epsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma^2\right), \tag{45}$$

and we have used $\beta_0, ..., \beta_{I-1}$ instead of $\mu_1, ..., \mu_I$ to distinguish between the two (equivalent!) *parametrizations* (42) and (45) of the model. The model (45) can be written in the vector form (43) with the corresponding X matrix,

$$\boldsymbol{X} = [\mathbf{1}\ X^{(1)}\ X^{(2)} \cdots\ X^{(I-1)}] \in \mathbb{R}^{n\times I}.$$

Now, in terms of (45), the hypothesis (44) is equivalent to

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{I-1} = 0.$$

This is exactly to test what we called the global null hypothesis under the model (42), so the (global) F-test can be used.

If we're only interested in a p-value, we could proceed with the global F-test in the usual parametrization (45). But in the special case of a one-way ANOVA the F *statistic* itself takes on a special, intuitive form. We now derive this. First, we recall the following form of the F statistic,

$$F = \frac{\left\|\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_{(1)}\right\|^2 /(p-r)}{\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2/(n-p-1)},$$

where $\boldsymbol{Y}_{(1)}$ denotes the fitted values under the *restricted* model, i.e. under H_0. Note that we may use the form above of the F statistic under either parametrizations (42) or (45) (of course, each of these two model will have a different $\boldsymbol{X}^{(1)}$ matrix. It will be convenient to use (42) for our calculation. Thus, we want to test (44). Under the null hypothesis (44), we have

$$Y_{ij} \overset{\text{iid}}{\sim} \mathcal{N}\left(\mu, \sigma^2\right), \quad i = 1, \ldots, I, j = 1, \ldots, n_i,$$

in other words under H_0 we have

$$\mathbf{Y} = \mathbf{1}\mu + \boldsymbol{\epsilon},$$

and so $\mathbf{Y}_{(1)} = \mathbf{1}\hat{\mu}$ with

$$\hat{\mu} = \bar{Y}_{\cdot\cdot} := \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{n_i} Y_{ij}.$$

Under the full model (42), we have

$$Y_{ij} \sim \mathcal{N}(\mu_i, \sigma^2),$$

and $\hat{Y}_{ij} = \hat{\mu}_i$ with

$$\hat{\mu}_i = \bar{Y}_{i\cdot} := \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}.$$

Together, we get

$$\|\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_{(1)}\|^2 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(\hat{Y}_{ij} - \widehat{Y_{(1)}}_{ij}\right)^2 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}\right)^2 = \sum_{i=1}^{I} n_i\left(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}\right)^2 := SSB$$

$$\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(Y_{ij} - \hat{Y}_{ij}\right)^2 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_{i\cdot}\right)^2 := SSE$$

and

$$F = \frac{\|\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_{(1)}\|^2/(I-1)}{\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2/(n-I)} = \frac{SSB/(I-1)}{\mathrm{SSE}\,/(n-I)} = \frac{MSB}{MSE}$$

where

$$MSB := \frac{SSB}{I-1}, \qquad MSE := \frac{SSE}{n-I}.$$

Sum-of-squares decomposition. Recall that, in general, we have $SST = SSR + SSE$. This decomposition also takes on a special form in the one-way ANOVA situation,

$$\underbrace{\sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_{\cdot\cdot}\right)^2}_{SST} = \underbrace{\sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(\hat{Y}_{ij} - \bar{Y}_{\cdot\cdot}\right)^2}_{SSR} + \underbrace{\sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(Y_{ij} - \hat{Y}_{ij}\right)^2}_{SSE}$$

$$= \underbrace{\sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}\right)^2}_{SSB} + \underbrace{\sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_{i\cdot}\right)^2}_{SSE}$$

Inference for the difference between a pair of means. Under the one-way model (41), i.e., $Y_{ij} \sim \mathcal{N}(\mu_i)$, we would now like to provide inference for the difference between any pair of means, say

$$\Delta_{12} := \mu_2 - \mu_1.$$

As we asserted earlier, the LS estimate of μ_i is $\hat{\mu}_i = \bar{Y}_{i\cdot}$, therefore

$$\hat{\mu}_1 = \bar{Y}_{1\cdot}, \quad \hat{\mu}_2 = \bar{Y}_{2\cdot}.$$

Hence,

$$\hat{\Delta}_{12} = \hat{\mu}_2 - \hat{\mu}_1 = \bar{Y}_{2\cdot} - \bar{Y}_1 \sim \mathcal{N}\left(\Delta_{12}, \sigma^2\left(n_1^{-1} + n_2^{-1}\right)\right)$$

and the corresponding t statistics is

$$T_{12}^0 := \frac{\hat{\Delta}_{12}}{\sqrt{\hat{\sigma}^2\left(n_1^{-1} + n_2^{-1}\right)}} = \frac{\hat{\Delta}_{12}/\sqrt{\sigma^2\left(n_1^{-1} + n_2^{-1}\right)}}{\sqrt{\hat{\sigma}^2/\sigma^2}} \overset{H_0}{\sim} t_{n-I}, \quad \hat{\sigma}^2 = \mathrm{SSE}\,/(n-I).$$

Then, for example, a test for

$$H_0 : \mu_1 = \mu_2$$

rejects whenever $|T_{12}^0| > t_{n-I-1;1-\alpha/2}$.

Inference for a general contrast. In the one-way ANOVA context, it is common to consider special kinds of linear combinations. Thus, for a fixed vector $\boldsymbol{c} = (c_1, \ldots, c_I)^\top$ such that $\sum_{i=1}^I c_i \mu_i = 0$, the linear combination

$$\psi(\boldsymbol{c}) := \sum_i c_i \mu_i = \boldsymbol{c}^\top \boldsymbol{\mu}$$

is called a *contrast*. Generalizing the previous part, suppose we want to test that a certain contrast is zero,

$$H_0 : \psi(\boldsymbol{c}) = 0$$

The LS estimate is

$$\hat{\psi}(\boldsymbol{c}) = \sum_{i=1}^I c_i \hat{\mu}_i = \sum_{i=1}^I c_i \bar{Y}_i,$$

and

$$T_{\boldsymbol{c}}^0 := \frac{\hat{\psi}(\boldsymbol{c})}{\sqrt{\hat{\sigma}^2 V(c)}} \overset{H_0}{\sim} t_{n-I}, \quad V(\boldsymbol{c}) := \frac{1}{\sigma^2} \operatorname{Var}[\hat{\psi}(\boldsymbol{c})] = \sum_{i=1}^I \frac{c_i^2}{n_i}.$$

Note: this is really no more than an application in the ANOVA context of the theory we saw earlier in the course for inference for linear combinations.

Simultaneous inference contrasts. Recall that with the global F-statistic we can test the global hypothesis that all μ_i are equal. Equivalently, it test the hypothesis that $\mu_i - \mu_{i'} = 0$ for all $1 \le i < i' \le I$. If the global F-test rejects, then we can declare with confidence $1 - \alpha$ that there is at least one pair for which $\mu_i - \mu_{i'} \neq 0$, but it doesn't tell us which pairs of indices i, i' have unequal means. In most cases, we *will* want to know which specific pairs differ in their means. The t test can do that if we specify i, i' in advance, but now the question is different: we want to go over all $\binom{I}{2}$ pairs $i < i'$, and test whether $H_0^{\{i,i'\}} : \mu_i = \mu_{i'}$ is true. This kind of problem is called a *multiple comparisons problem*.

The Bonferroni correction. Suppose we used a test of level α' for each of the $\binom{I}{2}$ hypotheses $H_0^{\{i,i'\}}$. Then

$$E[\# \text{ false rejections }] = E\left[\sum_{\mathcal{H}} 1\left(H_0\left(\{i,i'\}\right) \text{ rejected }\right)\right] =$$
$$= \sum_{\mathcal{H}} \mathbb{E}\left[1\left(H_0\left(\{i,i'\}\right) \text{ rejected }\right)\right] = \sum_{\mathcal{H}} \mathbb{P}\left(H_0\left(\{i,i'\}\right) \text{ rejected }\right) \le \alpha' \cdot \binom{I}{2}.$$

Thus, if we take $\alpha' = \alpha / \binom{I}{2}$, we will get that

$$P(\text{at least one false rejection}) = E[1(\text{at least one false rejection})] \le E[\#\text{false rejections}] \le \alpha.$$

Hence taking $\alpha' = \alpha / \binom{I}{2}$, widely known as the *Bonferroni correction*, will ensure that the probability of at least one false rejection is $\le \alpha$.