Linear Models

1. The basic model

We now study a *linear statistical model*. That is, we study the models where the observations $\boldsymbol{Y} := (Y_1, \ldots, Y_n)'$ has the following assumed property:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\beta} := (\beta_0, \beta_1, \ldots, \beta_{p-1})$ is a vector of p unknown parameters, and

$$\boldsymbol{X} := \begin{pmatrix} x_{1,0} & \cdots & x_{1,p-1} \\ \vdots & & \vdots \\ x_{n,0} & \cdots & x_{n,p-1} \end{pmatrix}$$

is the socalled "regression matrix," or "design matrix." The elements of the $n \times p$ matrix $\boldsymbol{X}$ are assumed to be known; these are the "descriptive" or "explanatory" variables, and the randomness of the observed values is inherited from the "noise vector," $\boldsymbol{\varepsilon} := (\varepsilon_1, \ldots, \varepsilon_n)'$, which we may think of as being "typically small." Note that we are changing our notation slightly; $\boldsymbol{X}$ is no longer assumed to be a random vector [this is done in order to conform with the historical development of the subject].

Throughout, we assume always that the ε_i's are independent with mean zero and common variance σ^2, where $\sigma > 0$ is possibly [in fact, typically] unknown.

In particular, it follows from this assumption that

$$\mathrm{E}\boldsymbol{\varepsilon} = \boldsymbol{0} \text{ and } \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}. \tag{1}$$

Let us emphasize that our linear model, once written out coordinate-wise, is

$$Y_i = \beta_0 x_{i,0} + \cdots + \beta_{p-1} x_{i,p-1} + \varepsilon_i \qquad (1 \le i \le n).$$

It is intuitively clear that unless $n \ge p$, we cannot hope to effectively use our n observed values in order to estimate the $p+1$ unknowns $\sigma^2, \beta_0, \ldots, \beta_{p-1}$. Therefore, we assume always that

$$n \ge p + 1. \tag{2}$$

This condition guarantees that our linear model is not overspecified.

The best-studied linear models are the *normal models*. Those are linear models for which we assume the more stringent condition that

$$\boldsymbol{\varepsilon} \sim \mathrm{N}_n\left(\mathbf{0}, \sigma^2 \boldsymbol{I}\right). \tag{3}$$

Example 1 (A measurement-error model)**.** Here we study a socalled measurement-error model: Suppose the observations $Y_1, \ldots, Y_n$ satisfy

$$Y_i = \mu + \varepsilon_i \qquad (1 \le i \le n)$$

for an unknown parameter μ. This is a simplest example of a linear model, where $\boldsymbol{\beta} = \mu$ is 1×1, and $\boldsymbol{X} := \mathbf{1}_{n \times 1}$ is a vector of n ones. □

Example 2 (Simple linear regression)**.** In simple linear regression we assume that the observed values have the form

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad (1 \le i \le n),$$

where x_i is the predictive variable the corresponds to observation i, and β_0, β_1 are unknown. Simple linear regression fits into our theory of linear models, once we set the design matrix as

$$\boldsymbol{X} := \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

Example 3 (Polynomial regression)**.** Consider a nonlinear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_{p-1} x_i^{p-1} + \varepsilon_i \qquad (1 \le i \le n),$$

where p is a known integer ≥ 1 [$p-1$ denotes the degree of the polynomial approximation to the observed y's]. Then polynomial regression models are linear models with design matrices of the form

$$\boldsymbol{X} := \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{p-1} \end{pmatrix}.$$

Example 4 (One-way layout with equal observations)**.** In the simplest case of one-way layout, in analysis of variance, our observations are indexed by vectors themselves as follows:

$$Y_{i,j} = \mu_i + \varepsilon_{i,j} \qquad (1 \le i \le I\ ,\ 1 \le j \le J).$$

For instance, suppose we are interested in the effect of I different fertilizers. We apply these fertilizers to J different blocks, independently, and "measure the effect." Then, $Y_{i,j}$ is the effect of fertilizer i in block j. The preceding model is assuming that, up to sampling error, the effect of fertilizer i is μ_i. This is a linear model. Indeed, we can create a new random vector $\boldsymbol{Y}$ of IJ observations by simply "vectorizing" the $Y_{i,j}$'s:

$$\boldsymbol{Y} := \left(Y_{1,1}\,,\ldots,Y_{1,J}\,,Y_{2,1}\,,\ldots,Y_{2,J}\,,\ldots,Y_{I,1}\,,\ldots,Y_{I,J}\right)'.$$

The vector $\boldsymbol{\beta}$ of unknowns is $\boldsymbol{\beta} := (\mu_1\,,\ldots,\mu_I)'$, and the design matrix is the following $IJ \times I$ matrix:

$$\boldsymbol{X} := \begin{pmatrix} \mathbf{1}_{J\times 1} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{J\times 1} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{J\times 1} \end{pmatrix},$$

where $\mathbf{1}_{J\times 1} := (1\,,\ldots,1)'$ is a J-vector of all ones. It is possible to show that *one-way layout with unequal number of observations* is also a linear model. That is the case where $Y_{i,j} = \mu_i + \varepsilon_{i,j}$, where $1 \le i \le I$ and $1 \le j \le J_i$ [the number of observed values might differ from block to block].

2. The least-squares estimator of $\boldsymbol{\theta} := \boldsymbol{X\beta}$

Let us return to our general linear model

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}.$$

Ultimately, our goal is to first find and then analyze the least-squares estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$. But first, let us find the least-squares estimate for $\boldsymbol{\theta} := \boldsymbol{X\beta}$. In other words, we wish to perform the following optimization problem:

$$\min_{\boldsymbol{\beta}\in\mathbf{R}^p} \|\boldsymbol{Y} - \boldsymbol{X\beta}\| = \min_{\boldsymbol{\theta}\in\mathcal{C}(\boldsymbol{X})} \|\boldsymbol{Y} - \boldsymbol{\theta}\|.$$

Abstractly speaking, the minimizer solves

$$\widehat{\boldsymbol{\theta}} = \boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\boldsymbol{Y}.$$

But is there an optimal $\boldsymbol{\beta}$? As we shall see next, there certainly is a unique $\widehat{\boldsymbol{\beta}}$ when $\boldsymbol{X}$ has full rank.

3. The least-squares estimator of $\boldsymbol{\beta}$

Our *least-squares estimator* $\widehat{\boldsymbol{\beta}}$ of the vector parameter $\boldsymbol{\beta}$ is defined via

$$\min_{\boldsymbol{\beta}\in\mathbf{R}^p} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\| = \left\|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right\|.$$

We aim to solve this minimization problem under natural conditions on the design matrix $\boldsymbol{X}$. But first, let us introduce some notation. The vector

$$\widehat{\boldsymbol{Y}} := \widehat{\boldsymbol{\theta}} := \boldsymbol{X}\widehat{\boldsymbol{\beta}}$$

is called the vector of *fitted values*, and the coordinates of

$$\boldsymbol{e} := \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$$

are the socalled *residuals*.

Now we write our minimization problem in the following form: First find the minimizing $\widehat{\boldsymbol{\beta}} \in \mathbf{R}^p$ that solves

$$\min_{\boldsymbol{z}\in\mathcal{C}(\boldsymbol{X})} \|\boldsymbol{Y} - \boldsymbol{z}\| = \left\|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right\|.$$

Now we know from Proposition 18 (p. 20) that the vector $\boldsymbol{z}$ that achieves this minimum does so uniquely, and is given by $\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\boldsymbol{Y}$, where we recall $\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})} := \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is projection onto the column space of $\boldsymbol{X}$; this of course is valid provided that $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ exists. Now the matrix $\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}$ plays such an important role that it has its own name: It is called the *hat matrix*, and is denoted as

$$\boldsymbol{H} := \boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'.$$

$\boldsymbol{H}$ is called the *hat matrix* because it maps the observations $\boldsymbol{Y}$ to the fitted values $\widehat{\boldsymbol{Y}}$ [informally, it puts a "hat" over $\boldsymbol{Y}$]. More precisely, the defining feature of $\boldsymbol{H}$ is that

$$\widehat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y},$$

once again provided that $\boldsymbol{X}'\boldsymbol{X}$ is nonsingular. The following gives us a natural method for checking this nonsingularity condition in terms of $\boldsymbol{X}$ directly.

Lemma 5. *$\boldsymbol{X}'\boldsymbol{X}$ is nonsingular if and only if* $\operatorname{rank}(\boldsymbol{X}) = p$.

Proof. Basic linear algebra tells us that the positive semidefinite $\boldsymbol{X}'\boldsymbol{X}$ is nonsingular if and only if it is positive definite; i.e., if and only if it has full rank. Since the rank of $\boldsymbol{X}'\boldsymbol{X}$ is the same as the rank of $\boldsymbol{X}$, and since $n > p$, $\boldsymbol{X}'\boldsymbol{X}$ has full rank if and only if its rank, which is the same as $\operatorname{rank}(\boldsymbol{X})$, is p. □

From now on we always assume [unless we state explicitly otherwise] that $\text{rank}(\boldsymbol{X}) = p$. We have shown that, under this condition, if there is a $\widehat{\boldsymbol{\beta}}$, then certainly $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{H}\boldsymbol{Y} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$. In order to find $\widehat{\boldsymbol{\beta}}$ from this, multiply both sides by $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ to see that $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$.

The quantity $\text{RSS} := \boldsymbol{e}'\boldsymbol{e} = \|\boldsymbol{e}\|^2 := \|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\|^2$ is called the *sum of squared residuals*, also known as the *residual sum of squared errors*, and is given by $\|(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}\|^2 = \|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})^\perp}\boldsymbol{Y}\|^2$. In particular, we have the following:

Proposition 6. *If* $\text{rank}(\boldsymbol{X}) = p$*, then the least-squares estimator of* $\boldsymbol{\beta}$ *is*

$$\widehat{\boldsymbol{\beta}} := (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}, \tag{4}$$

and $\text{RSS} = \|(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}\|^2$.

Let us make some elementary computations with the least squares estimator of $\boldsymbol{\beta}$.

Lemma 7. $\widehat{\boldsymbol{\beta}}$ *is an unbiased estimator of* $\boldsymbol{\beta}$*, and* $\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

Proof. Because $\text{E}\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\text{E}\boldsymbol{Y} = \boldsymbol{\beta}$, it follows that $\widehat{\boldsymbol{\beta}}$ is unbiased. Also, $\text{Var}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\text{Var}(\boldsymbol{Y})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$, because of (1). □

4. Optimality

Theorem 8. *Let* $\boldsymbol{\theta} := \boldsymbol{X}\boldsymbol{\beta}$ *be estimated, via least squares, by* $\widehat{\boldsymbol{\theta}} := \boldsymbol{H}\boldsymbol{Y}$*. Then, for all linear unbiased estimates of* $\boldsymbol{c}'\boldsymbol{\theta}$*, the estimator* $\boldsymbol{c}'\widehat{\boldsymbol{\theta}}$ *uniquely minimizes the variance.*

By a "linear unbiased estimator of $\boldsymbol{c}'\boldsymbol{\theta}$" we mean an estimator of the form $\sum_{j=1}^n a_j Y_j$ whose expectation is $\boldsymbol{c}'\boldsymbol{\theta}$. In this sense, $\boldsymbol{c}'\widehat{\boldsymbol{\theta}}$ is the *best linear unbiased estimator* [or "BLUE"] of $\boldsymbol{c}'\boldsymbol{\theta}$. The preceding can be improved upon as follows, though we will not prove it:

Theorem 9 (Rao)**.** *Under the normal model* (3)*,* $\boldsymbol{c}'\widehat{\boldsymbol{\theta}}$ *is the unique UMVUE of* $\boldsymbol{c}'\boldsymbol{\theta}$*, for every nonrandom vector* $\boldsymbol{c} \in \mathbf{R}^n$.

Let us consider Theorem 8 next.

Proof of Theorem 8. We saw on page 43 that $\widehat{\boldsymbol{\theta}} := \boldsymbol{H}\boldsymbol{Y}$ irrespective of whether or not $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ exists. Therefore,

$$\text{E}\left(\boldsymbol{c}'\widehat{\boldsymbol{\theta}}\right) = \boldsymbol{c}'\text{E}\widehat{\boldsymbol{\theta}} = \boldsymbol{c}'\boldsymbol{\theta}, \quad \text{Var}\left(\boldsymbol{c}'\widehat{\boldsymbol{\theta}}\right) = \boldsymbol{c}'\text{Var}(\boldsymbol{H}\boldsymbol{Y})\boldsymbol{c} = \boldsymbol{c}'\boldsymbol{H}'\text{Var}(\boldsymbol{Y})\boldsymbol{H}\boldsymbol{c}$$
$$= \sigma^2\|\boldsymbol{H}\boldsymbol{c}\|^2.$$

Any other linear estimator has the form $\boldsymbol{a}'\boldsymbol{Y}$, and satisfies

$$\mathrm{E}\left(\boldsymbol{a}'\boldsymbol{Y}\right) = \boldsymbol{a}'\mathrm{E}\boldsymbol{Y} = \boldsymbol{a}'\boldsymbol{\theta}, \quad \mathrm{Var}\left(\boldsymbol{a}'\boldsymbol{Y}\right) = \boldsymbol{a}'\mathrm{Var}(\boldsymbol{Y})\boldsymbol{a} = \sigma^2\|\boldsymbol{a}\|^2.$$

If, in addition, $\boldsymbol{a}'\boldsymbol{Y}$ is unbiased, then it follows that $\boldsymbol{a}'\boldsymbol{\theta} = \boldsymbol{c}'\boldsymbol{\theta}$; i.e., $\boldsymbol{a} - \boldsymbol{c}$ is orthogonal to $\boldsymbol{\theta}$. This should hold no matter what value $\boldsymbol{\beta}$ [and hence $\boldsymbol{\theta}$] takes. Since $\mathcal{C}(\boldsymbol{X})$ is the collection of all possible values of $\boldsymbol{\theta}$, it follows that $\boldsymbol{a} - \boldsymbol{c}$ is orthogonal to $\mathcal{C}(\boldsymbol{X})$. Because $\boldsymbol{H}$ is projection onto $\mathcal{C}(\boldsymbol{X})$, it follows that $\boldsymbol{H}(\boldsymbol{a} - \boldsymbol{c}) = \boldsymbol{a}$; equivalently, $\boldsymbol{H}\boldsymbol{c} = \boldsymbol{H}\boldsymbol{a}$. Therefore, $\mathrm{Var}(\boldsymbol{c}'\widehat{\boldsymbol{\theta}}) = \sigma^2\|\boldsymbol{H}\boldsymbol{a}\|^2$ and

$$\mathrm{Var}\left(\boldsymbol{a}'\boldsymbol{Y}\right) - \mathrm{Var}\left(\boldsymbol{c}'\widehat{\boldsymbol{\theta}}\right) = \sigma^2\left\{\|\boldsymbol{a}\|^2 - \|\boldsymbol{H}\boldsymbol{a}\|^2\right\} = \sigma^2\|(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{a}\|^2 \geq 0,$$

thanks to the Pythagorean property. □

5. Regression and prediction

Now that we have the least-squares estimate for $\boldsymbol{\beta}$, let us use it in order to make prediction.

Recall that our model is $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. In applications, Y_i is the ith observation for the y variable, and the linear model is really saying that given an explanatory variable $\boldsymbol{x} = (x_0, \ldots, x_{p-1})'$,

$$y = \beta_0 x_0 + \cdots + \beta_{p-1}x_{p-1} + \text{“noise.”}$$

Therefore, our prediction, for a given $\boldsymbol{x}$, is

$$\text{[predicted value] } y = \widehat{\beta}_0 x_0 + \cdots + \widehat{\beta}_{p-1}x_{p-1}, \tag{5}$$

where $\widehat{\beta} = (\widehat{\beta}_0, \ldots, \widehat{\beta}_{p-1})'$ is the least-squares estimate of $\boldsymbol{\beta}$. We may view the right-hand side of (5) as a function of $\boldsymbol{x}$, and call (5) the equation for the “regression line.”

6. Estimation of σ^2

We wish to also estimate σ^2. The estimator of interest to us turns out to be the following:

$$S^2 := \frac{1}{n-p}\mathrm{RSS}, \tag{6}$$

as the next lemma suggests.

Lemma 10. *S^2 is an unbiased estimator of σ^2.*

Proof. Recall that RSS $= \boldsymbol{e}'\boldsymbol{e} = \|\boldsymbol{Y} - \boldsymbol{H}\boldsymbol{Y}\|^2$. We can write the RSS as $\|(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}\|^2 = \boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{H})'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} = \boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$. In other words, the RSS is a random quadratic form for the matrix $\boldsymbol{A} := \boldsymbol{I} - \boldsymbol{H}$, and hence

$$\begin{aligned}\mathrm{E}(\mathrm{RSS}) &= \mathrm{tr}((\boldsymbol{I} - \boldsymbol{H})\mathrm{Var}(\boldsymbol{Y})) + (\mathrm{E}\boldsymbol{Y})'(\boldsymbol{I} - \boldsymbol{H})(\mathrm{E}\boldsymbol{Y}) \\ &= \sigma^2\mathrm{tr}(\boldsymbol{I} - \boldsymbol{H}) + (\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X}\boldsymbol{\beta}.\end{aligned}$$

Because $\boldsymbol{X}\boldsymbol{\beta} \in \mathcal{C}(\boldsymbol{X})$, $\boldsymbol{I} - \boldsymbol{H}$ projects onto the orthogonal subspace of where it is, therefore $(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0}$. And the trace of the projection matrix $(\boldsymbol{I} - \boldsymbol{H})$ is its rank, which is $n - \mathrm{tr}(\boldsymbol{H}) = n - p$, since $\boldsymbol{X}$ has full rank p. It follows that $\mathrm{E}(\mathrm{RSS}) = \sigma^2(n - p)$, and therefore $\mathrm{E}(S^2) = \sigma^2$. □

7. The normal model

In the important special case of the normal model,

$$\boldsymbol{Y} \sim \mathrm{N}_n\left(\boldsymbol{X}\boldsymbol{\beta}\,, \sigma^2\boldsymbol{I}\right).$$

Therefore, $\widehat{\boldsymbol{\beta}} \sim \mathrm{N}_p(\boldsymbol{\beta}\,, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$. And the vector $(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ of *residual errors* is also a multivariate normal:

$$(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} \sim \mathrm{N}_{n-p}\left((\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X}\boldsymbol{\beta}\,, \sigma^2(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H})'\right) = \mathrm{N}_{n-p}\left(\boldsymbol{0}\,, \sigma^2(\boldsymbol{I} - \boldsymbol{H})\right).$$

Therefore, in the normal model,

$$S^2 = \frac{1}{n-p}\,\|(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}\|^2 \sim \sigma^2\frac{\chi^2_{n-p}}{n-p}.$$

Finally, we note that $\boldsymbol{t}'\widehat{\boldsymbol{\beta}} + \boldsymbol{s}'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$ is a matrix times $\boldsymbol{Y}$ for all $\boldsymbol{t} \in \mathbf{R}^p$ and $\boldsymbol{s} \in \mathbf{R}^{n-p}$. Therefore, $(\widehat{\boldsymbol{\beta}}\,, (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y})$ is also a multivariate normal. But

$$\mathrm{Cov}\left(\widehat{\boldsymbol{\beta}}\,,\ (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}\right) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathrm{Var}(\boldsymbol{Y})(\boldsymbol{I} - \boldsymbol{H})' = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{I} - \boldsymbol{H}) = \boldsymbol{0},$$

since the columns of $\boldsymbol{X}$ are obviously orthogonal to every element in $\mathcal{C}(\boldsymbol{X})^\perp$ and $\boldsymbol{I} - \boldsymbol{H} = \boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})^\perp}$. This shows that $\widehat{\boldsymbol{\beta}}$ and $(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$ are independent, and hence $\widehat{\boldsymbol{\beta}}$ and S^2 are also independent. Thus, we summarize our findings.

Theorem 11. *The least-squares estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is given by (4); it is always unbiased. Moreover, S^2 is an unbiased estimator of σ^2. Under the normal model, S and $\widehat{\boldsymbol{\beta}}$ are independent, and*

$$\widehat{\boldsymbol{\beta}} \sim \mathrm{N}_p\left(\boldsymbol{\beta}\,, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\right), \quad S^2 \sim \sigma^2\frac{\chi^2_{n-p}}{n-p}.$$

Recall that $\boldsymbol{Y}$ has the nondegenerate multivariate normal distribution $N_n(\boldsymbol{X\beta}\,,\sigma^2\boldsymbol{I})$. Therefore, its pdf is

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{y}-\boldsymbol{X\beta}\|^2\right).$$

This shows readily the following.

Lemma 12. *In the normal model, $\widehat{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$ and $\left(\frac{n-p}{n}\right)S^2$ is the MLE for σ^2.*

Proof. Clearly, maximizing the likelihood function, over all $\boldsymbol{\beta}$, is the same as minimizing $\|\boldsymbol{Y}-\boldsymbol{X\beta}\|^2$. Therefore, MLE = least squares for $\boldsymbol{\beta}$. As for σ^2, we write the log likelihood function:

$$L(\sigma) = -\frac{n}{2}\ln(2\pi) - n\ln\sigma - \frac{1}{2\sigma^2}\|\boldsymbol{Y}-\boldsymbol{X\beta}\|^2.$$

Then,

$$L'(\sigma) = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\|\boldsymbol{Y}-\boldsymbol{X\beta}\|^2.$$

Set $L'(\sigma)=0$ and solve to see that the MLE of σ^2 is $\frac{1}{n}\|\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 = \left(\frac{n-p}{n}\right)S^2$, thanks to the MLE principle and the already-proven fact that the MLE of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}}$. □

8. Some examples

1. A measurement-error model. Recall the measurement-error model

$$Y_i = \mu + \varepsilon_i \qquad (1 \le i \le n).$$

We have seen that this is a linear model with $p=1$, $\boldsymbol{\beta}=\mu$, and $\boldsymbol{X} := \mathbf{1}_{n\times 1}$. Since $\boldsymbol{X}'\boldsymbol{X} = n$ and $\boldsymbol{X}'\boldsymbol{Y} = \sum_{i=1}^n Y_i$, we have

$$\widehat{\boldsymbol{\beta}} = \widehat{\mu} := \bar{Y},$$

and

$$\frac{1}{n-1}S^2 = \frac{1}{n-1}\left\|\boldsymbol{Y}-\bar{Y}\mathbf{1}_{n\times 1}\right\|^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i-\bar{Y})^2 := S^2.$$

These are unbiased estimators for μ and σ^2, respectively. Under the normal model, $\bar{Y}$ and S^2 are independent, $\bar{Y}\sim \mathrm{N}(\mu\,,\sigma^2)$ and $(n-1)S^2 \sim \sigma^2\chi^2_{n-1}$. These are some of the highlights of Math. 5080.

2. Simple linear regression. Recall that simple linear regression is our linear model in the special case that $p = 2$, $\boldsymbol{\beta} = (\beta_0\,, \beta_1)'$, and

$$\boldsymbol{X} := \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

We have

$$\boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}, \quad \boldsymbol{X}'\boldsymbol{Y} = \begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n x_i Y_i \end{pmatrix},$$

and

$$\begin{aligned} (\boldsymbol{X}'\boldsymbol{X})^{-1} &= \frac{1}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix} \\ &= \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}. \end{aligned}$$

Therefore,

$$\begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{pmatrix} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y},$$

which leads to

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \text{ and } \widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{x}.$$

We have derived these formulas by direct computation already. In this way we find that the fitted values are

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\beta}_0 + \widehat{\beta}_1 x_1 \\ \vdots \\ \widehat{\beta}_0 + \widehat{\beta}_1 x_n \end{pmatrix}.$$

Also,

$$S^2 = \frac{1}{n-2}\left\|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\right\|^2 = \frac{1}{n-2}\sum_{i=1}^n \left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right)^2,$$

and this is independent of $(\widehat{\beta}_0\,, \widehat{\beta}_1)$ under the normal model.

Recall that our linear model is, at the moment, the simple regression model,

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

Perhaps the most important first question that we can ask in this context is $\beta_1 = 0$; that is, we wish to know whether the x variables are [linearly] independent of the Y's. Let us try to find a confidence interval for β_1 in

order to answer this question. From now on, we work under the normal model. Recall that under the normal model,

$$\widehat{\boldsymbol{\beta}} \sim \mathrm{N}_2\left(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\right) \Rightarrow \widehat{\beta}_1 \sim \mathrm{N}\left(\beta_1, \sigma^2\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\right]_{2,2}\right) = \mathrm{N}\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right).$$

Equivalently,

$$Z := \frac{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}{\sigma}\left(\widehat{\beta}_1 - \beta_1\right) \sim \mathrm{N}(0, 1).$$

Now, S^2/σ^2 is independent of Z and is distributed as $\chi^2_{n-2}/(n-2)$. Therefore,

$$\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\left(\frac{\widehat{\beta}_1 - \beta_1}{S}\right) = \frac{Z}{\sqrt{S^2/\sigma^2}} \sim t_{n-2}.$$

Therefore, a $(1-\alpha)\times 100\%$ confidence interval for β_1 is

$$\widehat{\beta}_1 \pm \frac{S}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} t^{\alpha/2}_{n-2},$$

where t^r_v is the point whose right area, under the t_v pdf, is r. If zero is not in this confidence interval, then our statistical prediction is that β_1 is not zero [at the confidence level α].

3. A remark on polynomial regression. Recall that, in polynomial regression, we have $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{X}$ is the following design matrix:

$$\boldsymbol{X} := \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{p-1} \end{pmatrix}.$$

If $p = 2$, then this is simple linear regression. Next consider the "quadratic regression" model where $p = 3$. That is,

$$\boldsymbol{X} := \begin{pmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix} \quad \Rightarrow \quad \boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} n & \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i^3 \\ \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i^3 & \sum_{i=1}^n x_i^4 \end{pmatrix}.$$

Because $n \geq 4$, $\boldsymbol{X}$ is nonsingular if and only if $\boldsymbol{x}$ is not a vector of constants [a natural condition]. But you see that already it is painful to invert $\boldsymbol{X}'\boldsymbol{X}$. This example shows the importance of using a computer in linear models: Even fairly simple models are hard to work with, using only direct calculations.