LINEAR REGRESSIONS

SIMON ELLERSGAARD

1. Overview

1.1. Model Philosophy. The most garden variety of regressions - the linear ones - arguably also triumph as the most prevalent in quantitative analysis. This should come as no surprise: first, there is the comparative ease of implementation with countless software packages already containing a complete codification of the process. Secondly, they are readily interpretable: insofar as a connection exists between two variables, we can quickly read off the magnitude and directionality of the relationship between them. Thirdly, there is a substantial body of literature exposing the statistical properties of linear regressions, and the apparent ease with which these notions can be picked up and interpreted by practitioners. Finally, linear regressions are (notwithstanding the name) remarkably generic in the relationships they are able to capture, insofar as the linearity is understood to reside in the parameters only. Polynomial equations, e.g., are linear in the parameters and therefore fall under the spectrum of linear regression.

To set the scene, let $\{y_i\}_{i=1}^n$ be a set of n observations for the *response* variable or *regressand* (the variable we want to predict), and let $\{x_{i1}, x_{i2}, ..., x_{ip-1}\}_{i=1}^n$ be a set of n observations for $p-1$ input *features* or *regressors*. In a linear regression model we postulate that y can be explained in terms of the xs through the following relationship

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip-1} + \varepsilon_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \tag{1}$$

where ε_i is a zero-mean constant-variance stochastic noise term codifying deviations from the linear model, and we have defined the vector quantities $\mathbf{x}_i = (1, x_{i1}, ..., x_{ip})^\top \in \mathbb{R}^p$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_{p-1})^\top \in \mathbb{R}^p$, the latter being a vector of constants to be determined. Ceteris paribus, if a feature x_j is augmented by one unit, the response variable y will increase by β_j units.

To avoid cluttered notation, it will be convenient to codify the collection of observations $i = 1, 2, ..., n$ for (1) as the matrix equation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2}$$

where $\mathbf{y} = (y_1, y_2, ..., y_n)^\top \in \mathbb{R}^n$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)^\top \in \mathbb{R}^n$, and $\mathbf{X}$ is the $n \times p$ matrix

$$\begin{pmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_n^\top \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np-1} \end{pmatrix}.$$

Note that the first feature has been chosen to be a constant column (of ones). This so-called intercept should generally not be ignored unless the features and response have been centered (i.e. normalised) beforehand. To see this, consider summing equation (1) over i and dividing by n. By construction, for centered variables, $n^{-1}\sum_i y_i = n^{-1}\sum_i x_{ij} = n^{-1}\sum_i \varepsilon_i = 0$, so here it must be the case that $\beta_0 = 0$.

Remark 1. While linearity *prima facie* appears overtly restrictive, equation (1) can, in fact, model a wide range of behaviours. Specifically, as the linearity is understood to reside in the parameters, a model such as $y_i = \alpha + \beta \ln(x'_1) + \gamma x'_2 x'_3 + \delta \sin(1/x'_4) + \varepsilon_i$ is perfectly linear, as the change of variables $x_1 \equiv \ln(x'_1)$, $x_2 \equiv x'_2 x'_3$, $x_3 \equiv \sin(1/x'_4)$ clearly shows. This powerful fact is all too often overlooked by newcomers to econometrics.

Date: 2020-07-31.

1.2. Some advice on feature construction.

- Whilst feature normalisation is not a requirement, there is some rationale behind engineering the regression such that features remain roughly on the same scale, lest we run into numerically unstable solutions. E.g. if your input features include both the national GDP and base rate, expressing the former in, say, units of trillions or billions of USD and the latter in units of % p.a. would make sense.
- *Categorical variables*, i.e. features which only take on a a discrete set of states (not necessarily numerical), are often codified erroneously by newcomers to econometrics. Case in point, while a dichotomaous state such as { default, non-default } readily can be identified with the encoding $\{1, 0\}$, and "orderable" polytomaous states such as {buy, neutral, sell} plausibly can be identified with $\{1, 0, -1\}$, it is non-obvious what to do with polytamous variables without any natural ordering into the real line. E.g. if we were to regress S&P500 stock returns onto their sector membership $S = \{$ IT, health care, financials, ..., materials $\}$ how should we proceed? Assuming $|S| = N$ there are $N!$ different orderings of states; each one as meaningless as the next. To circumvent this problem the answer is to deploy "one-hot encoding" in which the N states are associated with $N-1$ binary variables. Thus, the IT sector could be associated with the variable $x_{\text{IT}} \in \{0, 1\}$, where $x_{\text{IT}} = 1$ iff the stock belongs to the IT sector. Similar variables would then be created for the remaining variables (heat care, financials, ...) with the caveat that no variable is needed for the final category (since if all other sectors are classified as zeros, the stock must be default belong to this category).

1.3. Literature.

The literature on linear regressions is plentiful. For ordinary least squares and maximum likelihood estimation an authoritative work to which we owe much inspiration is Heij et al. [4]. Less technical but pedagocially excellent resources are found in James et al. [6] and Rogers and Girolami [9]. For Bayesian linear regression O'Hagan and Forster [2] is an excellent resource.

2. The Ordinary Least Squares Framework

2.1. Assumptions of Ordinary Least Squares (OLS).

Before we begin, it is helpful to stipulate our assumptions. The exact origin (necessity) of these presuppositions will become clearer as we go along.

1. First, we assume that the model coincides with the data generating process, i.e. that our model (2) is a accurate representation of the true governing relationship between regressors and regressand.
2. Secondly, we assume the regressors - codified by $\mathbf{X} \in \mathbb{R}^{n \times p}$ - are fixed (non-stochastic) and that $n \geq p$ and $\text{rank}(\mathbf{X}) = p$ (the matrix has full rank). (We will see how these assumptions can be relaxed in later sections).
3. Thirdly, so-called *strict exogoneity* is enforced wherein errors have conditional mean zero: $\mathbb{E}[\boldsymbol{\varepsilon}] = 0$.
4. Fourthly, the errors are assumed *spherical*, i.e. $\mathbb{V}\text{ar}[\boldsymbol{\varepsilon}] = \sigma^2 \mathbb{I}_n$, where $\mathbb{I}_n$ is the $n \times n$ identity matrix. This encapsulates the notion of *homescedacity* ("similar dispersion"), $\mathbb{E}[\varepsilon_i^2] = \sigma^2$, and the absence of auto-correlation, $\mathbb{E}[\varepsilon_i \varepsilon_j] = 0$, for $i \neq j$. When errors do not have constant variance we say that they are *heteroscedastic* - a matter which can be treated using *weighted least squares*.

Remark 2. It is sometimes claimed that a further assumption is normality of the errors ($\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbb{I}_n)$). While this is a convenient assumption from a statistical perspective, which greatly facilitates hypothesis testing, we emphasise that the validity of the OLS estimator does NOT necessitate normality. The origin of this misconception arguably stems from the fact that the maximum-likelihood estimate for $\boldsymbol{\beta}$ coincides with the OLS estimator under normal errors. Finally, one sometimes sees the stronger claim that the response $\mathbf{y}$ must be normal. This is just plain wrong.

2.2. Analytic Solution.

In the ordinary least squares framework we postulate that

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\text{argmin}} \sum_{i=1}^{n} (\mathbf{x}_i^\top \boldsymbol{\beta} - y_i)^2 \\ &= \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\text{argmin}} ||\mathbf{X}\boldsymbol{\beta} - \mathbf{y}||_2^2 \end{aligned} \tag{3}$$

where $||\cdot||_2^2$ is the L_2-norm, i.e. $||\mathbf{X}\boldsymbol{\beta} - \mathbf{y}||_2^2 \equiv (\mathbf{X}\boldsymbol{\beta} - \mathbf{y})^\top (\mathbf{X}\boldsymbol{\beta} - \mathbf{y})$. Expanding the brackets, we arrive at

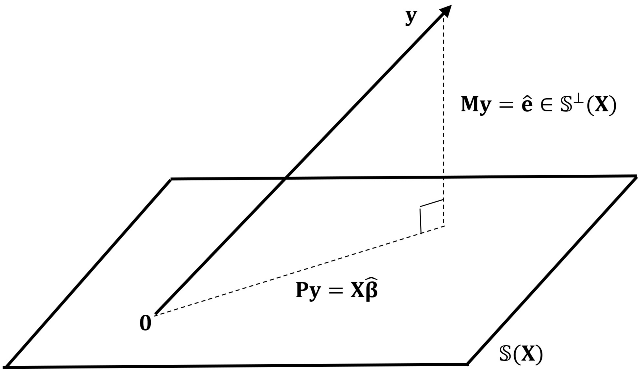

FIGURE 1. The regressand $\mathbf{y} \in \mathbb{R}^n$ can be decomposed as $\mathbf{y} = \mathbf{P}\mathbf{y} + \mathbf{M}\mathbf{y}$, where the matrix $\mathbf{P}$ projects onto the space $\mathbb{S}(\mathbf{X})$ spanned by the columns of $\mathbf{X}$, and the matrix $\mathbf{M}$ projects onto the orthogonal subspace $\mathbb{S}^{\perp}(\mathbf{X})$. Since $\mathbf{P}\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} = \hat{\mathbf{y}}$ and $\mathbf{M}\mathbf{y} = \hat{\boldsymbol{\varepsilon}}$ we therefore have a very simple geometric interpretation of the mechanics of OLS regression.

$$||\mathbf{X}\boldsymbol{\beta} - \mathbf{y}||_2^2 = \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{y} + \mathbf{y}^\top \mathbf{y}. \tag{4}$$

The extremum is found by differentiating the bracketed expression with respect to $\boldsymbol{\beta}$ and setting equating to zero. This gives rise to the so-called *normal equation*, $\mathbf{X}^\top \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^\top \mathbf{y}$, which, assuming the absence of perfect colinearity, can be written as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}. \tag{5}$$

This is the OLS estimator.

2.3. **Interpretation.** Formally, OLS estimation can be viewed as a projection of the response vector $\mathbf{y}$ onto the linear subspace $\mathbb{S}(\mathbf{X})$ spanned by the feature vectors (i.e. the columns of $\mathbf{X}$). To see this, note that the OLS prediction, $\hat{\mathbf{y}} \equiv \mathbf{X}\hat{\boldsymbol{\beta}}$, can be written as $\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}$. Now the matrix defined by $\mathbf{P} \equiv \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top$ is trivially shown to be an orthogonal projection matrix as it satisfies the conditions of *idempotence*, $\mathbf{P}^2 = \mathbf{P}$, and *symmetry*, $\mathbf{P} = \mathbf{P}^\top$. Indeed it has the meaningful property that $\mathbf{P}\mathbf{X} = \mathbf{X}$. As for the vector of residuals, $\hat{\boldsymbol{\varepsilon}} \equiv \mathbf{y} - \hat{\mathbf{y}}$, this can be written as $\hat{\boldsymbol{\varepsilon}} = \mathbf{M}\mathbf{y}$ where we have defined $\mathbf{M} \equiv \mathbb{I}_n - \mathbf{P}$. $\mathbf{M}$ is likewise an orthogonal projection matrix onto the space $\mathbb{S}^{\perp}(\mathbf{X})$ perpendicular to $\mathbb{S}(\mathbf{X})$. In particular, we have that $\mathbf{P}\mathbf{M} = \mathbf{M}\mathbf{P} = \mathbf{0}$ so $\mathbf{M}$ is the null space of $\mathbf{P}$, and, of course, $\mathbf{M}\mathbf{X} = \mathbf{0}$. See figure 1.

2.4. **Moments.** We will now demonstrate the important results

$$\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}, \quad \text{and} \quad \mathbb{V}\text{ar}[\hat{\boldsymbol{\beta}}] = \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}. \tag{6}$$

The expectation of the $\hat{\beta}$ is easily derivable:

$$\begin{aligned}
\mathbb{E}[\hat{\boldsymbol{\beta}}] &= \mathbb{E}[(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}] \\
&= \mathbb{E}[(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top (\mathbf{X}\boldsymbol{\beta} + \varepsilon)] \\
&= \mathbb{E}[(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} + (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \boldsymbol{\varepsilon}] \\
&= \boldsymbol{\beta} + (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbb{E}[\varepsilon] = \boldsymbol{\beta},
\end{aligned}$$

where the last line makes use of the strict exogeneity assumption. This shows that the OLS estimator is *unbiased.*

To obtain the covariance matrix we only need to work slightly harder. First, we note that $\mathbb{V}\text{ar}[\mathbf{y}] = \sigma^2 \mathbb{I}_n$, which follows from the assumption of spherical errors. From (2) we therefore have $\mathbb{E}[\mathbf{y}\mathbf{y}^\top] = \sigma^2 \mathbb{I}_n + \mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}^\top \mathbf{X}^\top$. Now,

$$\begin{aligned}\mathbb{V}\text{ar}[\hat{\boldsymbol{\beta}}] &= \mathbb{E}[\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^\top] - \mathbb{E}[\hat{\boldsymbol{\beta}}]\mathbb{E}[\hat{\boldsymbol{\beta}}^\top]\\ &= \mathbb{E}[\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^\top] - \boldsymbol{\beta}\boldsymbol{\beta}^\top\\ &= \mathbb{E}[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}\mathbf{y}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}] - \boldsymbol{\beta}\boldsymbol{\beta}^\top\\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbb{E}[\mathbf{y}\mathbf{y}^\top]\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} - \boldsymbol{\beta}\boldsymbol{\beta}^\top\\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\sigma^2\mathbb{I}_n + \mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}^\top\mathbf{X}^\top)\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} - \boldsymbol{\beta}\boldsymbol{\beta}^\top\\ &= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} + (\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\mathbf{X})\boldsymbol{\beta}\boldsymbol{\beta}^\top(\mathbf{X}^\top\mathbf{X})(\mathbf{X}^\top\mathbf{X})^{-1} - \boldsymbol{\beta}\boldsymbol{\beta}^\top\\ &= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} + \boldsymbol{\beta}\boldsymbol{\beta}^\top - \boldsymbol{\beta}\boldsymbol{\beta}^\top = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}.\end{aligned}$$

2.5. **The Gauss-Markov Theorem.** It will be noted that the OLS estimator (5) is linear in the response variable, i.e. it is of the form $\hat{\boldsymbol{\beta}} = \mathbf{C}\mathbf{y}$ where $\mathbf{C} \in \mathbb{R}^{p\times n}$ is a constant matrix. We will now justify the OLS estimator by demonstrating that it is the *best linear unbiased estimator* (acronymically, the B.L.U.E.), which is to say that it has the lowest possible variance of all unbiased estimators. (As we shall see later, there are *biased* estimators which manage to achieve a variance lower than that of the OLS estimator). To prove the Gauss-Markov theorem, let $\tilde{\boldsymbol{\beta}}$ be any other estimator. Without loss of generality, $\tilde{\boldsymbol{\beta}}$ may be written as $\tilde{\boldsymbol{\beta}} = [(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top + \mathbf{D}]\mathbf{y}$ where we have defined the arbitrary constant matrix $\mathbf{D} \in \mathbb{R}^{p\times n}$. Perhaps unsurprisingly, to guarantee that $\tilde{\boldsymbol{\beta}}$ is also unbiased, we must enforce a constraint on $\mathbf{D}$ viz. $\mathbf{D}\mathbf{X} = \mathbf{0}$. To see this note that

$$\begin{aligned}\mathbb{E}[\tilde{\boldsymbol{\beta}}] &= \mathbb{E}[[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top + \mathbf{D}]\mathbf{y}]\\ &= \mathbb{E}[[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top + \mathbf{D}](\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})]\\ &= \mathbb{E}[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon} + \mathbf{D}\mathbf{X}\boldsymbol{\beta} + \mathbf{D}\boldsymbol{\varepsilon}]\\ &= \boldsymbol{\beta} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbb{E}[\boldsymbol{\varepsilon}] + \mathbf{D}\mathbf{X}\boldsymbol{\beta} + \mathbf{D}\mathbb{E}[\boldsymbol{\varepsilon}]\\ &= \boldsymbol{\beta} + \mathbf{D}\mathbf{X}\boldsymbol{\beta}.\end{aligned}$$

where the last line uses the assumption of strict exogeneity. Evidently, $\mathbb{E}[\tilde{\boldsymbol{\beta}}] = \boldsymbol{\beta}$ if and only if $\mathbf{D}\mathbf{X} = \mathbf{0}$ as purported. We are now in a position to prove that the OLS estimator $\hat{\boldsymbol{\beta}}$ has a smaller variance than $\tilde{\boldsymbol{\beta}}$ (formally: that the matrix $\Omega \equiv \mathbb{V}\text{ar}[\tilde{\boldsymbol{\beta}}] - \mathbb{V}\text{ar}[\hat{\boldsymbol{\beta}}]$ is a positive semi-definite matrix):

$$\begin{aligned}\mathbb{V}\text{ar}[\tilde{\boldsymbol{\beta}}] &= \mathbb{V}\text{ar}[[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top + \mathbf{D}]\mathbf{y}]\\ &= [(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top + \mathbf{D}]\mathbb{V}\text{ar}[\mathbf{y}][(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top + \mathbf{D}]^\top\\ &= [(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top + \mathbf{D}]\mathbb{V}\text{ar}[\boldsymbol{\varepsilon}][(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top + \mathbf{D}]^\top\\ &= \sigma^2[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top + \mathbf{D}][(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top + \mathbf{D}]^\top\\ &= \sigma^2[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{D}^\top + \mathbf{D}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} + \mathbf{D}\mathbf{D}^\top]\\ &= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} + \mathbf{D}\mathbf{D}^\top\\ &= \mathbb{V}\text{ar}[\hat{\boldsymbol{\beta}}] + \sigma^2\mathbf{D}\mathbf{D}^\top.\end{aligned}$$

Thus, $\Omega = \sigma^2\mathbf{D}\mathbf{D}^\top$ and clearly, $\forall\mathbf{x} \in \mathbb{R}^p$, $\mathbf{x}^\top\mathbf{D}\mathbf{D}^\top\mathbf{x} = (\mathbf{D}^\top\mathbf{x})^\top\mathbf{D}^\top\mathbf{x} = ||\mathbf{D}^\top\mathbf{x}||_2^2 \geq 0$ so Ω is positive semi-definite.

2.6. **On Estimating the Variance of the Residuals.** So far, we have completely disregarded the estimation of the disturbance variance, σ^2. Yet, if we are to conduct any tests of the statistical significance of the OLS estimator this evidently needs to be rectified. To this end it is opportune to look at $\mathbb{E}[\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}]$ where $\hat{\boldsymbol{\varepsilon}}$ is the empirical residual: $\hat{\boldsymbol{\varepsilon}} \equiv \mathbf{y} - \hat{\mathbf{y}} = \mathbf{M}\mathbf{y}$ cf. section 2.3. It turns out that it is convenient to rewrite $\mathbb{E}[\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}]$ in terms of the trace operator, specifically $\mathbb{E}[\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}] = \mathbb{E}[\text{tr}(\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}})] = \mathbb{E}[\text{tr}(\hat{\boldsymbol{\varepsilon}}\hat{\boldsymbol{\varepsilon}}^\top)] = \text{tr}(\mathbb{E}[\hat{\boldsymbol{\varepsilon}}\hat{\boldsymbol{\varepsilon}}^\top])$, where we have used the cyclicality of the trace operator and the linearity of the trace and expectation operators. Combining this expression with $\hat{\boldsymbol{\varepsilon}} = \mathbf{M}\mathbf{y} = \mathbf{M}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \mathbf{M}\boldsymbol{\varepsilon}$ we therefore have $\mathbb{E}[\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}] = \text{tr}(\mathbb{E}[\mathbf{M}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\mathbf{M}^\top]) = \text{tr}(\mathbf{M}\mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top]\mathbf{M}^\top) = \sigma^2\text{tr}(\boldsymbol{M}^2) = \sigma^2\text{tr}(\boldsymbol{M})$, where we have used the spherical error property, as well as the symmetry and idempotence of the projection matrix. Using the fact that $\text{tr}(\mathbf{M}) = \text{tr}(\mathbb{I}_n - \mathbf{P}) = \text{tr}(\mathbb{I}_n - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top) = n - \text{tr}(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top) = n - \text{tr}((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}) = n - \text{tr}(\mathbb{I}_p) =$

$n-p$ we therefore have that $\mathbb{E}[\hat{\varepsilon}^\top\hat{\varepsilon}] = (n-p)\sigma^2$. Thus, for an unbiased estimate of σ^2 one would compute the sample variance

$$s^2 = \frac{\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}}{n-p}. \tag{7}$$

Combining this with the result in (6) we therefore have that the standard error of the j^{th} component of $\hat{\boldsymbol{\beta}}$ is given by

$$\text{SE}(\hat{\beta}_j) = s\sqrt{[(\mathbf{X}^\top\mathbf{X})^{-1}]_{jj}}. \tag{8}$$

Remark 3. Note that as the number of features increases for fixed n so does the sample variance increase, becoming infinite when $p = n$. This should make sense intuitively: when $p = n$ we obtain a perfect fit in our regression, but obviously we haven't really explained anything. Rather, as it would quickly become apparent if we were to introduce more data, we would have been extremely likely to have over-fit our initial observations.

2.7. **Goodness of fit. The R^2 measure.** Ultimately, we desire a singular measure which codifies the goodness of fit of the regression. To this end, we introduce the *coefficient of determination*, R^2, which measures the proportion of total variation which is explained by the model. Specifically, we compute the ratio

$$R^2 \equiv \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}, \tag{9}$$

where ESS is the *explained sum of squares*, defined as $ESS \equiv \sum_i(\hat{y}_i - \bar{y})^2$, TSS is the *total sum of squares*, defined as $TSS \equiv \sum_i(y_i - \bar{y})^2$, and RSS is the *residual sum of squares*, defined as $RSS = \sum_i \hat{\varepsilon}_i^2 = \sum_i(y_i - \hat{y}_i)$ ($\bar{y}$ denotes the mean of y). To justify this measure, we will show that - insofar as the regression includes the constant intercept term ($\mathbf{X}$ contains a column of ones) - the following decomposition holds true:

$$TSS = ESS + RSS \tag{10}$$

To this end, let us introduce the matrix $\mathbf{N} \equiv \mathbb{I}_n - n^{-1}\boldsymbol{\iota}\boldsymbol{\iota}^\top$ where $\boldsymbol{\iota} = (1, 1, ..., 1)^\top \in \mathbb{R}^n$. Note that $\mathbf{N}$ is symmetric and idempotent.[1] Clearly, $\forall \mathbf{x} \in \mathbb{R}^n$: $\sum_i(x_i - \bar{x})^2 = (\mathbf{x} - \bar{x}\boldsymbol{\iota})^\top(\mathbf{x} - \bar{x}\boldsymbol{\iota}) = (\mathbf{N}\mathbf{x})^\top(\mathbf{N}\mathbf{x}) = \mathbf{x}^\top\mathbf{N}^\top\mathbf{N}\mathbf{x} = \mathbf{x}^\top\mathbf{N}\mathbf{x}$, so we can write $TSS = \mathbf{y}^\top\mathbf{N}\mathbf{y}$. Inserting $\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\varepsilon}}$ we get $TSS = (\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\varepsilon}})^\top\mathbf{N}(\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\varepsilon}}) = \hat{\boldsymbol{\beta}}^\top\mathbf{X}^\top\mathbf{N}\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}^\top\mathbf{X}^\top\mathbf{N}\hat{\boldsymbol{\varepsilon}} + \hat{\boldsymbol{\varepsilon}}^\top\mathbf{N}\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\varepsilon}}^\top\mathbf{N}\hat{\boldsymbol{\varepsilon}}$. Now take $\hat{\boldsymbol{\varepsilon}} = \mathbf{M}\mathbf{y}$ and left multiply by $\mathbf{X}^\top$ to get $\mathbf{X}^\top\hat{\boldsymbol{\varepsilon}} = \mathbf{X}^\top\mathbf{M}\mathbf{y}$. Since $\mathbf{M}\mathbf{X} = \mathbf{0}$ we must have that $\mathbf{X}^\top\hat{\boldsymbol{\varepsilon}} = \mathbf{0}$. In particular, using the assumption of a constant column in $\mathbf{X}$, it must be the case that $\boldsymbol{\iota}^\top\hat{\boldsymbol{\varepsilon}} = \mathbf{0}$ whence $\mathbf{N}\hat{\boldsymbol{\varepsilon}} = \hat{\boldsymbol{\varepsilon}}$. This allows us to simplify the expression for the total sum of squares as follows:

$$TSS = \hat{\boldsymbol{\beta}}^\top\mathbf{X}^\top\mathbf{N}\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}} = \hat{\mathbf{y}}^\top\mathbf{N}\hat{\mathbf{y}} + \hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}},$$

where we readily identify $ESS = \hat{\mathbf{y}}^\top\mathbf{N}\hat{\mathbf{y}}$ and $RSS = \hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}$. This establishes (10) and justifies R^2 as a ratio of variation explained by the model over the total variation exhibited. Clearly, $R^2 \in [0, 1]$ as $0 \leq ESS \leq TSS$.

A rather interesting corollary is the equivalence between R^2 and the (squared) correlation between the observed values y_i and the fitted values $\hat{y}_i$, i.e.

$$R^2 = \rho^2_{y,\hat{y}}. \tag{11}$$

Proving this is straightforward. Per definition, $\rho^2_{y,\hat{y}} \equiv \mathbb{C}\text{ov}[y, \hat{y}]^2/(\mathbb{V}\text{ar}[y]\mathbb{V}\text{ar}[\hat{y}])$. Meanwhile $y = \hat{y} + \hat{\varepsilon}$ with $\mathbb{C}\text{ov}[\hat{y}, \hat{\varepsilon}] = 0$, so $\rho^2_{y,\hat{y}} = \mathbb{C}\text{ov}[\hat{y} + \hat{\varepsilon}, \hat{y}]^2/(\mathbb{V}\text{ar}[y]\mathbb{V}\text{ar}[\hat{y}]) = \mathbb{C}\text{ov}[\hat{y}, \hat{y}]^2/\ (\mathbb{V}\text{ar}[y]\mathbb{V}\text{ar}[\hat{y}]) = \mathbb{V}\text{ar}[\hat{y}]/\mathbb{V}\text{ar}[y]$. Obviously, $\mathbb{V}\text{ar}[\hat{y}] = n \cdot ESS$ and $\mathbb{V}\text{ar}[y] = n \cdot TSS$, which completes the proof.

Remark 4. What happens when there is no constant intercept term in the regression model? For starters (10) is no longer valid and RSS may exceed TSS so computing R^2 as $1 - RSS/TSS$ could result in a negative number. Imagine, for example, that we fit a data generating process with negative slope and positive intercept in the first quadrant. A zero intercept OLS regression would fit this data with

[1] It is, in fact, a special example of the $\mathbf{M}$ projection matrix when $\mathbf{X} = \boldsymbol{\iota}$ cf. section 2.3

a positively slope straight line, and it is clear that we would have fared better by simply postulating a constant value for y. To avoid negative R^2s one may therefore see the alternative definition ESS/TSS being used. However, this is not the complete picture. In some software packages one can sometimes witness a considerable boost in R^2 if only one enforces the constraint $\beta_0 = 0$. Nonetheless, this situation is not a legitimate increase in performance, but rather is an artefact of the software package tacitly setting $\bar{y} = 0$ in the definition of R^2 in zero-intercept models. Comparing the R^2 outputs with and without the intercept term is thus like comparing chalk and cheese.

Remark 5. One of the key shortcomings of R^2 is that the measure incentivises overfitting: adding explanatory variables to the model never decreases R^2. To remedy this, one may consider discounting performance by the number of features included in what is known as the adjusted R^2:

$$R^2_{\text{adj}} \equiv 1 - \frac{n-1}{n-p-1}(1 - R^2). \tag{12}$$

We note that $R^2_{\text{adj}} < R^2$ and that as $p \to n-1$, $R^2_{\text{adj}} \to -\infty$.

2.8. **A Brief Note on Model Uncertainty.** Suppose the true data generating process is of the form $\mathbf{y} = \mathbf{X}_1\beta_1 + \boldsymbol{\varepsilon}$, but we believe the right model is of the more generic form $\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \boldsymbol{\varepsilon}$. What happens? Well, we are in effect specifying the right model, with the caveat that $\beta_2 = 0$. The OLS machinery developed above rightfully shows that $\mathbb{E}[\hat{\beta}_1] = \beta_1$ and $\mathbb{E}[\hat{\beta}_2] = \beta_2 = 0$. However, the catch here is that by introducing auxiliary regressors we inflate the uncertainty around our estimators. Specifically, let $\hat{\beta}_1^{(1)}$ be the OLS estimator associated with specifying the data generating process correctly, and let $\hat{\beta}_1^{(2)}$ be the OLS estimator for β_1 when we've introduced the redundant feature $\mathbf{X}_2$, then it can be shown that $\mathbb{V}\text{ar}[\hat{\beta}_1^{(1)}] \leq \mathbb{V}\text{ar}[\hat{\beta}_1^{(2)}]$.

Conversely, suppose now that the true data generating process is bivarite, $\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \boldsymbol{\varepsilon}$, but we believe the right model to be of the restricted form $\mathbf{y} = \mathbf{X}_1\beta_1 + \boldsymbol{\varepsilon}$. This has the immediate effect that we end up biasing the OLS estimator. Specifically, since $\hat{\beta}_1^{(1)} \equiv (\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\mathbf{X}_1^\top\mathbf{y} = \beta_1 + (\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\mathbf{X}_1^\top\mathbf{X}_2\beta_2 + (\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\mathbf{X}_1^\top\boldsymbol{\varepsilon}$, so $\mathbb{E}[\hat{\beta}_1^{(1)}] = \beta_1 + (\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\mathbf{X}_1^\top\mathbf{X}_2\beta_2$. The caveat here is that variance correspondingly has dropped, which is to say that $\mathbb{V}\text{ar}[\hat{\beta}_1^{(1)}] \leq \mathbb{V}\text{ar}[\hat{\beta}_1^{(2)}]$.

In practice, we obviously do not have epistemic access to the true governing model when we conduct our regressions. One would be well advised to keep this extra layer of model uncertainty in mind (and their respective implications) when interpreting results.

2.9. **On the Significance of the Coefficients. The t-test.** Generally the OLS estimator will tend to yield non-zero coefficients regardless of which features we include in the model. To avoid spuriously attributing meaning to this, we need to test whether we can reject the null hypothesis that each individual coefficient is in fact zero. To this end, let us assume that $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbb{I}_n)$ (see remark 2). From section 2.4 we immediately deduce that

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}). \tag{13}$$

whence $\forall i$

$$\frac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{[(\mathbf{X}^\top\mathbf{X})^{-1}]_{ii}}} \sim N(0,1). \tag{14}$$

By setting $\beta_i = 0$ this test-statistic allows us to assess the null hypothesis that β_i is nil. This argument is *essentially* correct, but does not account for the fact the population parameter σ must in turn be estimated through equation (8). Factoring this in, the appropriate test statistic becomes

$$t_i \equiv \frac{\hat{\beta}_i - \beta_i}{s\sqrt{[(\mathbf{X}^\top\mathbf{X})^{-1}]_{ii}}} \sim t(n-p), \tag{15}$$

where $t(n-p)$ is the student t-distribution with $n-p$ degrees of freedom. The proof runs along the following lines: first recall that if $\boldsymbol{z} = (z_1, z_2, ..., z_n)^\top$ is a vector of i.i.d. $N(0,1)$ random variables then by definition $\boldsymbol{z}^\top\boldsymbol{z} \sim \chi^2(n)$ - the chi-squared distribution with n degrees of freedom. In particular, if $\boldsymbol{A}$ is a symmetric idempotent matrix of rank r then $\boldsymbol{z}^\top\boldsymbol{A}\boldsymbol{z} \sim \chi^2(r)$. Finally, if $z \sim N(0,1)$ and $x \sim \chi^2(r)$ where $z \perp x$ then per definition $z/\sqrt{x/r} \sim t(r)$, the student-t distribution with r degrees of freedom.

Now from (7) we may rewrite the test-statistic as

$$t_i = \frac{(\hat{\beta}_i - \beta_i)/(\sigma\sqrt{[(\mathbf{X}^\top\mathbf{X})^{-1}]_{ii}})}{\sqrt{\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}/(\sigma^2(n-p))}}.$$

We just argued in (13) that the numerator of this expression must be $N(0,1)$. As for the denominator recall from section 2.6 that $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{M}\boldsymbol{\varepsilon}$, where $\boldsymbol{M}$ is a rank $n-p$ orthogonal projection matrix projection onto the subspace perpendicular to $\mathbf{X}$. It follows that $\frac{\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}}{\sigma^2} = \frac{\boldsymbol{\varepsilon}^\top}{\sigma}\boldsymbol{M}\frac{\boldsymbol{\varepsilon}}{\sigma} \sim \chi^2(n-p)$. Finally, since $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon}$ and $(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{M} = 0$ it can be shown that the numerator and denominator must be independent. This establishes the result.

Thus, for a given OLS coefficient $\hat{\beta}_i$ we reject the null-hypothesis of it being zero, at a particular level of significance $\alpha \in (0,1)$, just in case the p-value $\mathbb{P}(|t| > |t_i|) < 1-\alpha$, where t_i is calculated from (15) with $\beta_i = 0$ and t is a random variable following a $t(n-k)$ distribution. As a rule of thumb, when the degrees of freedom exceeds 30, the normal distribution may be used in lieu of t-distribution (the distributions can be shown to be equivalent in the infinite limit).

2.10. **Confidence Regions.** Whilst the previous subsection focused on the significance of individual components of $\hat{\boldsymbol{\beta}}$, we will here consider the vectorized estimator in its entirety. For starters let us consider the *confidence region* for $\boldsymbol{\beta}$ - an abstract extension of the confidence interval, which can be thought of as a p-dimensional ellipsoid, the volume of which is determined by an ad-hoc level of significance $1-\alpha$, centered around the estimator $\hat{\boldsymbol{\beta}}$. Such regions are sometimes portrayed as encapsulating the true population parameter with the probability $1-\alpha$, but the reality is more subtle: if we were to repeat the sampling of n observations many times, carefully constructing confidence regions for each such set, then $(1-\alpha)\%$ of those regions would contain the true population parameter.

To construct the confidence region, first recall that if $\mathbf{x} = (x_1, x_2, ..., x_n)^\top \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then $\boldsymbol{z} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu}) \sim N(\mathbf{0}, \mathbb{I}_n)$ and, by definition, $\boldsymbol{z}^\top\boldsymbol{z} = (\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \sim \chi^2(n)$. Applying this to our distribution for $\hat{\boldsymbol{\beta}}$ (13) we deduce that

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \frac{\mathbf{X}^\top\mathbf{X}}{\sigma^2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \chi^2(p).$$

Again, the population variance σ^2 really ought to be replaced with the sample variance $s^2 = \hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}/(n-p)$ where $\frac{\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}}{\sigma^2} \sim \chi^2(n-p)$. Inserting this, and recalling that for the generic independent random variables $v \sim \chi^2(\nu)$ and $w \sim \chi^2(\omega)$ the ratio $(v/\nu)/(w/\omega)$ follows the so-called F-distribution with ν and ω degrees of freedom, $F(\nu, \omega)$, we readily find that

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \frac{\mathbf{X}^\top\mathbf{X}}{ps^2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim F(p, n-p). \tag{16}$$

Now for a particular level of significance $1-\alpha$ this may be recast as the formula

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \mathbf{X}^\top\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \leq ps^2 F_{1-\alpha}(p, n-p), \tag{17}$$

which codifies the equation of an ellipsoid in $\mathbb{R}^p$. To see this more vividly note that $\mathbf{X}^\top\mathbf{X}$ is a symmetric matrix of rank p so it admits the eigendecomposition $\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$ where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_0, \lambda_2, ..., \lambda_{p-1})$ is the diagonal matrix of eigenvalues, and $\boldsymbol{Q} = (\boldsymbol{u}_0 \;\; \boldsymbol{u}_1 \;\; ... \;\; \boldsymbol{u}_{p-1})$ is an orthogonal matrix containing the associated (unit) eigenvectors. Using the change of coordinates $\boldsymbol{\beta} \mapsto \boldsymbol{\beta}_* : \boldsymbol{\beta}_* = \boldsymbol{Q}^\top(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$ ("a translation followed by an orthogonal transformation")[2] we find that $\boldsymbol{\beta}_*^\top\boldsymbol{\Lambda}\boldsymbol{\beta}_* \leq ps^2F_{1-\alpha}(p, n-p)$, which in component form amounts to

$$\frac{\beta_{*0}^2}{\alpha_0^2} + \frac{\beta_{*1}^2}{\alpha_1^2} + \cdots + \frac{\beta_{*p-1}^2}{\alpha_{p-1}^2} \leq 1, \tag{18}$$

where we have defined $\alpha_i \equiv \sqrt{ps^2F_{1-\alpha}(p, n-p)/\lambda_i}$. This clearly highlights the ellipsoidal nature of the confidence region: each of the principal semi-axes pointing in the direction of one of the eigenvectors $\boldsymbol{u}_i$ of $\mathbf{X}^\top\mathbf{X}$ with α_i codifying the half-length thereof.

[2] Note that $\boldsymbol{Q}$ cannot necessarily be interpreted as just a rotation matrix. The determinant of $\boldsymbol{Q}$ is ±1 (since $\boldsymbol{Q}^\top\boldsymbol{Q} = \mathbb{I}_p \to \det(\boldsymbol{Q})^2 = 1$) which encapsulates rotations as well as

2.11. **General Linear Restrictions. The F-test.** Building on our work in the previous subsections, suppose now we wish to investigate the *joint* significance of the coefficients ($H_0 : \beta_1 = \beta_2 = ... = \beta_{p-1} = 0$), or more broadly test our regression against any other set of linear restriction on the coefficients (such as $H_0 : \beta_i = \beta_j = 0$ or $H_0 : \beta_i = -2\beta_j$ etc. for particular i, j). Such restrictions generally take on the form H_0: $\boldsymbol{R\beta} = \boldsymbol{r}$ where $\boldsymbol{R} \in \mathbb{R}^{g\times p}$ and $\boldsymbol{r} \in \mathbb{R}^g$ (for example, for the joint significance of the coefficients, we would have $\boldsymbol{R} \in \mathbb{R}^{(p-1)\times p}$ with $R_{ij} = \delta_{i,j-1}$ (the Kronecker delta) and $\boldsymbol{r} = \mathbf{0} \in \mathbb{R}^{p-1}$). To test the hypothesis, note that (13) implies that $\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r} \sim N(\boldsymbol{R\beta} - \boldsymbol{r}, \sigma^2\boldsymbol{R}(\mathbf{X}^\top\mathbf{X})^{-1}\boldsymbol{R}^\top)$. If the null hypothesis is true, $\boldsymbol{R\beta} - \boldsymbol{r} = \mathbf{0}$, hence following the reasoning that led to (16) we deduce that

$$(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})^\top \frac{[\boldsymbol{R}(\mathbf{X}^\top\mathbf{X})^{-1}\boldsymbol{R}^\top]^{-1}}{gs^2}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}) \sim F(g, n-p). \tag{19}$$

Rejecting the null hypothesis of the linear restriction on the OLS estimator is then a matter of getting a value for this expression to exceed some pre-specified threshold (significance level) in the F-distribution. The only slight caveat here is that (19) fundamentally is difficult to interpret. Fortunately, it transpires that the expression can be re-written as a more intuitive ratio of "sum of squared restricted OLS errors to sum of squared unrestricted OLS errors":

$$\frac{(\hat{\boldsymbol{\varepsilon}}_R^\top\hat{\boldsymbol{\varepsilon}}_R - \hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}})/g}{\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}/(n-p)} \sim F(g, n-p), \tag{20}$$

where $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\varepsilon}}_R = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_R$ are the errors resulting from the OLS estimator and the restricted OLS estimator respectively. Proving this is algebraically cumbersome, but shall will sketch the general idea, leaving the reader to fill in the details (see Hallam [1] for details). First, note that the restricted OLS estimator $\hat{\boldsymbol{\beta}}_R$ arises from solving the minimization problem

$$\mathscr{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = ||\mathbf{X}\boldsymbol{\beta} - \mathbf{y}||_2^2 - \boldsymbol{\lambda}^\top(\boldsymbol{r} - \boldsymbol{R\beta}),$$

where $\boldsymbol{\lambda} \in \mathbb{R}^g$ is a Lagrange multiplier. Differentiating partially with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ and equating to zero, the two first order conditions are readily shown to be i. $-2\mathbf{X}^\top\mathbf{y} + 2\mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}}_R + \boldsymbol{R}^\top\boldsymbol{\lambda} = \mathbf{0}$ and ii. $\boldsymbol{R}\hat{\boldsymbol{\beta}}_R - \boldsymbol{r} = \mathbf{0}$. Pre-multiplying i. by $\boldsymbol{R}(\mathbf{X}^\top\mathbf{X})^{-1}$ and rearranging we find $\boldsymbol{\lambda} = -2[\boldsymbol{R}(\mathbf{X}^\top\mathbf{X})^{-1}\boldsymbol{R}^\top]^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_R - \boldsymbol{R}(\mathbf{X}^\top\mathbf{X})^\top\mathbf{X}^\top\mathbf{y})$. Substituting this back into i. alongside the conditions $\boldsymbol{R}\hat{\boldsymbol{\beta}}_R = \boldsymbol{r}$ and $(\mathbf{X}^\top\mathbf{X})^\top\mathbf{X}^\top\mathbf{y} = \hat{\boldsymbol{\beta}}$ we arrive at the following relationship between $\hat{\boldsymbol{\beta}}_R$ and $\hat{\boldsymbol{\beta}}$: $\hat{\boldsymbol{\beta}}_R = \hat{\boldsymbol{\beta}} + (\mathbf{X}^\top\mathbf{X})^{-1}\boldsymbol{R}[\boldsymbol{R}(\mathbf{X}^\top\mathbf{X})^{-1}\boldsymbol{R}^\top]^{-1}(\boldsymbol{r} - \boldsymbol{R}\hat{\boldsymbol{\beta}})$. Pre-multiplying both sides of this equation by $\mathbf{X}$ and subtracting $\mathbf{y}$ we therefore have $\hat{\boldsymbol{\varepsilon}}_R = \hat{\boldsymbol{\varepsilon}} + \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\boldsymbol{R}[\boldsymbol{R}(\mathbf{X}^\top\mathbf{X})^{-1}\boldsymbol{R}^\top]^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})$. Taking the inner product of this expression with itself and using the fact that $\mathbf{X}\hat{\boldsymbol{\varepsilon}} = \mathbf{0}$ we find after a bit of manipulation that $\hat{\boldsymbol{\beta}}_R^\top\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^\top\hat{\boldsymbol{\beta}} + (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})^\top[\boldsymbol{R}(\mathbf{X}^\top\mathbf{X})^{-1}\boldsymbol{R}^\top]^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})$. Plugging this into the numerator of (19) and using (7) for the denominator we obtain the desired expression (20).

2.12. **Predictions.** Suppose we've fit our regression based on n training points. Now consider an auxiliary point $\mathbf{x}_*$. What's our prediction for $y_* = y(\mathbf{x}_*)$ and what is our confidence interval therefore? Importantly, how does this differ from the confidence interval for the *average* value of y, at the same point $\mathbf{x}_*$ in feature space? Surprisingly, perhaps, there is a difference here! To see this, note that the new point by assumption must obey the true formula $y_* = \mathbf{x}_*^\top\boldsymbol{\beta} + \varepsilon_*$. Since $\mathbb{E}[\varepsilon_*] = 0$ our point prediction will be $\hat{y}_* = \mathbf{x}_*^\top\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the OLS estimator from our n training points. As for the variance of our prediction, note that we derive uncertainty both from $\hat{\boldsymbol{\beta}}$ (variance $\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$) and ε_* (variance σ^2). These sources of error are independent so we have $\mathbb{V}\mathrm{ar}[\hat{y}_*] = \sigma^2\mathbf{x}_*^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_* + \sigma^2$ whence the $1-\alpha$ confidence interval for our prediction (the so-called *prediction interval*) becomes

$$\hat{y}_* \pm t_{1-\alpha/2}(n-p)s\sqrt{\mathbf{x}_*^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_* + 1}, \tag{21}$$

where s (squared) is the oft-mentioned sample variance, and $t_{1-\alpha/2}(n-p)$ is the t-distribution with $n-p$ degrees of freedom at probability level $1-\alpha/2$.

On the other hand, suppose we are interested in the mean response at $\mathbf{x}_*$. Obviously, the point estimate remains $\bar{y}_* = \mathbf{x}_*^\top\hat{\boldsymbol{\beta}}$, but our source of uncertainty now solely lies with the OLS coefficient (by construction, there is no error term here). Hence, the confidence interval for the mean response amounts to

$$\bar{y}_* \pm t_{1-\alpha/2}(n-p)s\sqrt{\mathbf{x}_*^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_*}. \tag{22}$$

For an illustration of prediction vs. confidence intervals in the context of fraternal twin IQ the reader is referred to Leininger [10].

2.13. **Example: Simple Linear Regression.** Suppose $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ for $i = 1, 2, ..., n$ where $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$ - the case of *simple linear regression* with normally distributed errors. For didactic reasons let's rehash some of the results above under this simplified set-up. For starters, the matrix version (2) would obviously have

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \text{and} \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

From this, using the the standard identity for the 2 × 2 matrix inverse, we readily find that

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix}^{-1} = \frac{1}{n\sum_i(x_i-\bar{x})^2}\begin{pmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & n \end{pmatrix}, \tag{23}$$

where $\bar{x} = n^{-1}\sum_i x_i$ is the sample mean. Hence the OLS estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ becomes

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \frac{1}{n\sum_i(x_i-\bar{x})^2}\begin{pmatrix} \sum_i x_i^2 \sum_i y_i - \sum_i x_i \sum_i x_i y_i \\ n\sum_i x_i y_i - \sum_i x_i \sum_i y_i \end{pmatrix},$$

which we can simplify further by making the following observations: first, $n\sum_i x_i y_i - \sum_i x_i \sum_i y_i = n\sum_i(x_i - \bar{x})(y_i - \bar{y})$ where $\bar{y} = n^{-1}\sum_i y_i$. Secondly, by adding and subtracting the term $\bar{x}\sum_i x_i \sum_i y_i$ it can be shown that $\sum_i x_i^2 \sum_i y_i - \sum_i x_i \sum_i x_i y_i = n\sum_i(x_i-\bar{x})^2\bar{y} - n\sum_i(x_i-\bar{x})(y_i-\bar{y})\bar{x}$. The upshot of this is

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}, \tag{24}$$

where

$$\hat{\beta}_1 = \frac{\sum_i(x_i-\bar{x})(y_i-\bar{y})}{\sum_i(x_i-\bar{x})^2}. \tag{25}$$

Note that $\hat{\beta}_1$ is nothing but the sample covariance divided by the sample variance, i.e. $\hat{\beta}_1 = s_{x,y}/s_x^2$, or identically $\hat{\beta}_1 = \rho_{x,y}s_y/s_x$, where $\rho_{x,y}$ is the sample correlation. This makes for a useful mnemonic. It is also convenient, as it gives us the coefficient of determination for free. Specifically, we have that

$$R^2 = \rho_{x,y}^2. \tag{26}$$

To see this, recall that $R^2 = ESS/TSS$ where $ESS = \sum_i(\hat{y}_i - \bar{y})^2$ and $TSS = \sum_i(y_i - \bar{y})^2$. Substituting $\hat{y}_i = \hat{\beta}_0 + \hat{b}_1 x_i$ into ESS followed by our expression for $\hat{\beta}_0$ we get $ESS = \sum_i(\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})^2 = \sum_i(\bar{y} - \hat{\beta}_1\bar{x} + \hat{\beta}_1 x_i - \bar{y})^2 = \hat{\beta}_1^2\sum_i(x_i - \bar{x})^2$. Hence, using our expression for $\hat{\beta}_1$ we arrive at

$$ESS = \frac{[\sum_i(x_i-\bar{x})(y_i-\bar{y})]^2}{\sum_i(x_i-\bar{x})^2}.$$

ESS/TSS therefore amounts to

$$R^2 = \frac{[\sum_i(x_i-\bar{x})(y_i-\bar{y})]^2}{\sum_i(x_i-\bar{x})^2\sum_i(y_i-\bar{y})^2} = \rho_{x,y}^2.$$

Emphatically, (26) is *not* the same relationship as was found in (11). While the latter broadly applies to linear regressions, the former is only true for simple linear regressions. This is worthwhile keeping in mind!

For hypothesis testing purposes it follows from equations (8) and (23) that

$$SE^2(\hat{\beta}_0) = s^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2} \right), \tag{27a}$$

$$SE^2(\hat{\beta}_1) = \frac{s^2}{\sum_i (x_i - \bar{x})^2}, \tag{27b}$$

where $s^2 = \sum_i (y_i - \hat{y}_i)^2/(n-2)$. In particular, using (15) we may test the hypothesis $H_0 : \hat{\beta}_i = 0$, $H_1 : \hat{\beta}_i \neq 0$ by computing the statistic $\hat{\beta}_i / SE(\hat{\beta}_i) \sim t(n-2)$. Again, to represent these uncertainties in the estimators graphically, one could consider plotting the confidence region in (β_0, β_1)-space, or (what is more common) plot the $1-\alpha$ confidence interval for the mean response around the regression line. Following (22), it is readily shown that said interval is

$$(\hat{\beta}_0 + \hat{\beta}_1 x_*) \pm t_{1-\alpha/2}(n-p) s \sqrt{\frac{1}{n} + \frac{(x_* - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}}. \tag{28}$$

The square root function is readily shown to be convex in x_* with the minimum attained at $x_* = \bar{x}$. The confidence region for the mean response thus forms a concave "lens" around the regression line.

3. Linear Regressions in a Maximum-Likelihood Framework

3.1. **Philosophy.** In section 2 we derived our estimator $\hat{\boldsymbol{\beta}}$ for the linear model by requiring that the sum of squared errors is minimized. It transpires that an equivalent way of arriving at $\hat{\boldsymbol{\beta}}$ is to find the parameters of the probability density function of the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ that maximize the likelihood of obtaining the observed data points $(\mathbf{y}, \mathbf{X})$. This so-called method of *maximum-likelihood estimation* operates along the following general principles: let $p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})$ be the joint probability density codifying a stochastic model (not necessarily linear) coupling n observable data points $\mathbf{X}$ and $\mathbf{y}$ and let $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^k$ be a k-vector of parameters that go into specifying the distribution. Then for fixed $(\mathbf{y}, \mathbf{X})$ we may define the *likelihood function* as $L_n(\boldsymbol{\theta}) = p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})$, and in particular we may define the *maximum likelihood estimator* (MLE) as the parametric choice

$$\hat{\boldsymbol{\theta}}_{ML} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\text{argmax}}\, L_n(\boldsymbol{\theta}), \tag{29}$$

i.e. $\hat{\boldsymbol{\theta}}_{ML}$ is in effect the parameter-specification which makes the observable data "most probable". For computational convenience, one often first applies the natural logarithm to $L_n(\boldsymbol{\theta})$ and then proceeds to maximize the *log-likelihood function* $\ell(\boldsymbol{\theta}) = \ln(L_n(\boldsymbol{\theta}))$. Evidently, as ln is a monotonic function, the maximum of ℓ occurs at a value θ which is also the maximum for L_n. If we let the observations $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$ be i.i.d. then the joint density function may be decomposed as $p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = \prod_{i=1}^n p_{\boldsymbol{\theta}}(y_i|\mathbf{x}_i)$ where $p_{\boldsymbol{\theta}}(y_i|\mathbf{x}_i)$ is the pdf of the ith observation. Writing this in log-likelihood terms we get

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^n \ln(p_{\boldsymbol{\theta}}(y_i|\mathbf{x}_i)) \equiv \sum_{i=1}^n \ell_i(\boldsymbol{\theta}). \tag{30}$$

Remark 6. Non-analytical optimizations to the log-likelihood function can be handled through the multi-dimensional Newton-Raphson method. Specifically, we would solve

$$\hat{\boldsymbol{\theta}}_{t+1} = \hat{\boldsymbol{\theta}}_t - [H(\ell(\boldsymbol{\theta}))]^{-1} \nabla \ell(\boldsymbol{\theta}),$$

where $\nabla \ell(\boldsymbol{\theta}) = \sum_{i=1}^n \nabla \ell_i(\boldsymbol{\theta}) = \sum_{i=1}^n \partial \ell_i(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ is the gradient vector and $H(\ell(\boldsymbol{\theta})) = \sum_{i=1}^n H(\ell_i(\boldsymbol{\theta})) = \sum_{i=1}^n \partial^2 \ell_i / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top$ is the Hessian matrix. Rather conveniently it can be shown that $H(\ell(\boldsymbol{\theta})) \approx -\sum_{i=1}^n \nabla \ell_i(\boldsymbol{\theta}) \nabla^\top \ell_i(\boldsymbol{\theta})$ so one really only needs to compute the first derivative in this algorithm (see Casella and Berger [5]).

Under technical conditions it can be shown that the maximum likelihood estimator has a number of desirable limiting properties. In particular, the maximum likelihood estimator is

(1) *Consistent.* If the true data generating process has parameter $\boldsymbol{\theta}_0$ then as the number of observations n goes to infinity $\hat{\theta}_{ML}$ converges in probability to $\boldsymbol{\theta}_0$: $\hat{\theta}_{ML} \xrightarrow{p} \boldsymbol{\theta}_0$.

(2) *Functionally invariant.* Let $\hat{\boldsymbol{\theta}}_{ML}$ be the maximum likelihood estimator of $\boldsymbol{\theta}$, and let $g(\boldsymbol{\theta})$ be any transformation of $\boldsymbol{\theta}$, then the maximum likelihood estimator of $\boldsymbol{\theta}' = g(\boldsymbol{\theta})$ is $\hat{\boldsymbol{\theta}}'_{ML} = g(\hat{\boldsymbol{\theta}}_{ML})$.

(3) *Efficient* (it achieves the so-called Cramér-Rao bound, having the smallest covariance of all consistent estimators). Specifically as $n \to \infty$ the MLE converges in distribution to the normal distribution,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}) \xrightarrow{d} N(\mathbf{0}, \mathcal{I}_0^{-1}), \tag{31}$$

where

$$\mathcal{I}_0 = \lim_{n\to\infty} n^{-1}\mathcal{I}_n(\boldsymbol{\theta}_0) = \lim_{n\to\infty} -n^{-1}\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$$

is the asymptotic information matrix evaluated at $\boldsymbol{\theta}_0$.

3.2. **Applications to the Linear Model.** Let us suppose we have collected n observations from the linear data-generating process $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with normally distributed spherical errors $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbb{I})$. Clearly, the parameters of the model are $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$ and the relevant distribution $\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2 \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbb{I})$; in particular, since the observations are i.i.d. the likelihood function is a product over normal densities i.e.

$$L_n(\boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^n p_{\boldsymbol{\beta},\sigma^2}(y_i|\mathbf{x}_i) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_i - \mathbf{x}_i^\top\boldsymbol{\beta})^2}{2\sigma^2}\right\}.$$

Thus, the log-likelihood function is

$$\begin{aligned}\ell(\boldsymbol{\beta}, \sigma^2) &= \sum_{i=1}^n \left(-\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(y_i - \mathbf{x}_i^\top\boldsymbol{\beta})^2\right) \\ &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).\end{aligned}$$

The first order conditions are readily shown to be

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma^2}\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0}, \tag{32a}$$

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0, \tag{32b}$$

the solutions to which are

$$\hat{\boldsymbol{\beta}}_{ML} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}, \tag{33a}$$

$$\hat{\sigma}^2_{ML} = \frac{1}{n}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \frac{n-p}{n}s^2. \tag{33b}$$

We see that $\hat{\boldsymbol{\beta}}_{ML}$ is identical to the OLS estimator (5). Interestingly, $\hat{\sigma}^2_{ML}$ is a biased estimator of the variance of the error term (with the bias disappearing as $n \to \infty$).

To extract the distributive properties of $\hat{\boldsymbol{\beta}}_{ML}$ and s^2_{ML} we first compute the Hessian:

$$\begin{pmatrix} \frac{\partial^2 \ell}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top} & \frac{\partial^2 \ell}{\partial\boldsymbol{\beta}\partial\sigma^2} \\ \frac{\partial^2 \ell}{\partial\sigma^2\partial\boldsymbol{\beta}^\top} & \frac{\partial^2 \ell}{\partial\sigma^2\partial\sigma^2} \end{pmatrix} = \begin{pmatrix} -\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X} & -\frac{1}{\sigma^4}\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ -\frac{1}{\sigma^4}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top\mathbf{X} & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \end{pmatrix}.$$

Noting that $\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ and $\mathbb{E}[\varepsilon] = 0$ and $\mathbb{E}[\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}] = n\sigma^2$ it follows that the asymptotic information matrix is

$$\mathcal{I}_0 = \begin{pmatrix} \frac{1}{n\sigma^2}\mathbf{X}^\top\mathbf{X} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}.$$

From (31) we have $\hat{\boldsymbol{\theta}}_{ML} \approx N(\boldsymbol{\theta}_0, \mathcal{I}_n^{-1}(\hat{\boldsymbol{\theta}}_{ML}))$, whence the approximate distributions of our estimators are found to be

$$\hat{\boldsymbol{\beta}}_{ML} \approx N(\boldsymbol{\beta}_0, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}), \tag{34a}$$

$$\hat{\sigma}^2_{ML} \approx N\left(\sigma^2, \frac{2\sigma^4}{n}\right). \tag{34b}$$

4. Linear Regressions in a Bayesian Framework

4.1. **Philosophy.** Hitherto our treatment of the linear model has been exclusively confined to the frequentists' paradigm. It will probably come as no surprise that we may recast the regression process in Bayesian terms. In particular, we may start out with a prior assumption about the distribution of the constituent parameters of the model and then use the observable data to form a rational update with regards to the distribution of said parameters.

The general logic proceeds as follows: let $(\mathbf{y}, \mathbf{X})$ be a collection of n data points, connected through a stochastic (not necessarily linear) model with parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^k$. Our *prior* on the parameters is that $\boldsymbol{\theta} \sim p(\boldsymbol{\theta}|\boldsymbol{\xi})$, where $\boldsymbol{\xi}$ is a vector of *hyper-parameters.* We note that the prior is inherently subjective: there are no ontologically superior priors, but there may be computationally opportune ones (*conjugate priors* that lead to posteriors of the same distributive family) and certainly *uninformative priors* (maximum entropy priors) wherein we effectively acknowledge that there is no a priori reason to prefer one parameter specification over another.

Turning to the observations, our model will obviously codify a particular density function for any arbitrary choice of $\boldsymbol{\theta}$, $p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})$, which the reader should recognize as the *likelihood function* $L_n(\boldsymbol{\theta})$ from section 3.1. In particular, the *marginal likelihood* (the evidence) may then be construed as the density of the observed data after we have marginalized over the possible parameters: $p(\mathbf{y}|\mathbf{X}, \boldsymbol{\xi}) = \int_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\xi})d\boldsymbol{\theta}$.

Putting this together, it follows from Bayes' theorem that a rational *posterior* distribution for the parameters based on the observable data is

$$p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \boldsymbol{\xi}) = \frac{p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\xi})}{p(\mathbf{y}|\mathbf{X}, \boldsymbol{\xi})} = \frac{p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\xi})}{\int_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\xi})d\boldsymbol{\theta}}. \tag{35}$$

Note that the denominator really only serves as a normalization constant (we must require that the posterior is well defined probability density function in the sense that it sums to unity: $\int_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \boldsymbol{\xi})d\boldsymbol{\theta} = 1$). Therefore, it is not uncommon to see (35) expressed simply as

$$p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \boldsymbol{\xi}) \propto p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\xi}). \tag{36}$$

Obviously, the posterior does not yield a point estimate for $\boldsymbol{\theta}$ but we can meaningfully compute quantities such as the *maximum a posteriori* viz. the mode of the posterior distribution. Furthermore, we may proceed to calculate *credible regions* (intervals) for the parameter $\boldsymbol{\theta}$, which is to say for a given level of probability $1 - \alpha$ we may find bounds $\boldsymbol{\theta}_l$ and $\boldsymbol{\theta}_u$ such that $\int_{\boldsymbol{\theta}_l\leq\boldsymbol{\theta}\leq\boldsymbol{\theta}_u} p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \boldsymbol{\xi})d\boldsymbol{\theta} = 1 - \alpha$.[3] Credible regions distinguish themselves from confidence regions (see section 2.10) in the sense that the tacitly depend on our prior distribution (for the frequentist, there is a true population parameters will be covered by her confidence regions through repeat sampling $(1 - \alpha)\%$ of the time). Finally, concerning *predictions* for a novel observation point $\mathbf{x}_*$ note we can construct a density function $p(y_*|\mathbf{x}_*, \mathbf{y}, \mathbf{X})$ for $\mathbf{y}_*$ by marginalizing over the posterior:

$$p(y_*|\mathbf{x}_*, \mathbf{y}, \mathbf{X}, \boldsymbol{\xi}) = \int_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} p(y_*|\mathbf{x}_*, \boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \boldsymbol{\xi})d\boldsymbol{\theta}, \tag{37}$$

or identically $p(y_*|\mathbf{x}_*, \mathbf{y}, \mathbf{X}, \boldsymbol{\xi}) = \mathbb{E}_{p(\boldsymbol{\theta}|\mathbf{y},\mathbf{X},\boldsymbol{\xi})}[p(y_*|\mathbf{x}_*, \boldsymbol{\theta})]$.

4.2. **Applications to the Linear Model.** In deploying the Bayesian paradigm to the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbb{I}_n)$ we shall make the simplifying assumption that σ^2 is known a priori. This is by no means a necessary assumption (see e.g. O'Hagan and Forster [2] for a complete account), yet there is little pedagogical value gained in making matters more complicated at this stage.

Starting with the likelihood function we have already established that $p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2) = N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbb{I}_n)$. Furthermore, it is known that the prior that is conjugate to the normal likelihood is in turn normally distributed.[4] For ease of computation we will therefore set $p(\boldsymbol{\beta}|\boldsymbol{\xi}) = N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, where $\boldsymbol{\mu}_0 \in \mathbb{R}^p, \boldsymbol{\Sigma}_0 \in \mathbb{R}^{p\times p}$ are subjective hyper-parameters chosen based on the context of the problem. In calculating the posterior after n observations we shall actively make use of the fact that we expect $p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \sigma^2, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) = N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$ where $\boldsymbol{\mu}_n \in \mathbb{R}^p, \boldsymbol{\Sigma}_n \in \mathbb{R}^{p\times p}$ are to be determined. In particular, since

[3] Credence regions are in themselves not uniquely determined. To fix them, auxiliary constraints need to be enforced. For example, for the interval one would typically set the two tails to have equal probability mass.

[4] See e.g. `https://en.wikipedia.org/wiki/Conjugate_prior`.

$$\begin{aligned} N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n) &= \frac{1}{\sqrt{(2\pi)^p|\boldsymbol{\Sigma}_n|}} \exp\left\{-\tfrac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu}_n)^\top\boldsymbol{\Sigma}_n^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_n)\right\} \\ &\propto \exp\left\{-\tfrac{1}{2}\left(\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_n^{-1}\boldsymbol{\beta} - 2\boldsymbol{\mu}_n^\top\boldsymbol{\Sigma}_n^{-1}\boldsymbol{\beta}\right)\right\} \end{aligned} \tag{38}$$

it makes sense to write Bayes's formula $p(\boldsymbol{\beta}|\mathbf{y},\mathbf{X},\sigma^2,\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0) \propto p(\mathbf{y}|\mathbf{X},\boldsymbol{\beta},\sigma^2)p(\boldsymbol{\beta}|\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)$ in powers of $\boldsymbol{\beta}$ and read off $\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n$ directly. To this end,

$$\begin{aligned} p(\boldsymbol{\beta}|\mathbf{y},\mathbf{X},\sigma^2,\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0) \propto& \exp\left\{-\tfrac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top(\sigma^2\mathbb{I}_n)^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\} \times \\ & \exp\left\{-\tfrac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu}_0)^\top\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_0)\right\} \\ =& \exp\left\{-\tfrac{1}{2}\left(\sigma^{-2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + (\boldsymbol{\beta}-\boldsymbol{\mu}_0)^\top\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_0)\right)\right\} \\ \propto& \exp\left\{-\tfrac{1}{2}\left(\sigma^{-2}\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} - 2\sigma^{-2}\mathbf{y}^\top\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta} - 2\boldsymbol{\mu}_0^\top\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}\right)\right\} \\ =& \exp\left\{-\tfrac{1}{2}\left(\boldsymbol{\beta}^\top[\sigma^{-2}\mathbf{X}^\top\mathbf{X} + \boldsymbol{\Sigma}_0^{-1}]\boldsymbol{\beta} - 2[\sigma^{-2}\mathbf{y}^\top\mathbf{X} + \boldsymbol{\mu}_0^\top\boldsymbol{\Sigma}_0^{-1}]\boldsymbol{\beta}\right)\right\}. \end{aligned}$$

Comparing with (38) we readily see from the second order terms that

$$\boldsymbol{\Sigma}_n = [\sigma^{-2}\mathbf{X}^\top\mathbf{X} + \boldsymbol{\Sigma}_0^{-1}]^{-1}. \tag{39}$$

As for the first order terms, since $\sigma^{-2}\mathbf{y}^\top\mathbf{X} + \boldsymbol{\mu}_0^\top\boldsymbol{\Sigma}_0^{-1} = [\sigma^{-2}\mathbf{X}^\top\mathbf{y} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0]^\top\mathbb{I}_n = [\sigma^{-2}\mathbf{X}^\top\mathbf{y} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0]^\top \cdot (\boldsymbol{\Sigma}_n\boldsymbol{\Sigma}_n^{-1}) = (\boldsymbol{\Sigma}_n[\sigma^{-2}\mathbf{X}^\top\mathbf{y} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0])^\top\boldsymbol{\Sigma}_n^{-1}$ it follows, again by comparison, that

$$\boldsymbol{\mu}_n = \boldsymbol{\Sigma}_n\left(\sigma^{-2}\mathbf{X}^\top\mathbf{y} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right). \tag{40}$$

Thus, the posterior distribution for the parameter vector $\boldsymbol{\beta}$ is $N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$ where $\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n$ are given by (40) and (39). As the Bayesian approach yields a density rather than a point estimate for the coefficients, one would need to determine what is a sensible reduction of complexity for analytic and graphing purposes. To this end, one could consider working with the maximum a posteriori (the mode) which for the normal distribution is identical to the mean: $\boldsymbol{\beta}_{\text{mode}} = \boldsymbol{\mu}_n$. Note that this is *not* the OLS estimator we encountered before. In fact in order for $\boldsymbol{\beta}_{\text{mode}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ we would require $\boldsymbol{\Sigma}_0^{-1} = \mathbf{0}$ which informally corresponds to the large variance case $\boldsymbol{\Sigma}_0 = \lim_{\sigma_0\to\infty}\sigma_0^2\mathbb{I}$ in which the prior has been flattened out. As we shall see later, the Bayesian maximum a posteriori is closer in spirit to constrained (ridge) regression.

It may also be of interest to plot a credible interval for the regression line. Here one could consider sampling a large number of $\boldsymbol{\beta}$s from $N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$ and computing the corresponding mean responses $\mathbf{y}(\boldsymbol{\beta})$. A $(1-\alpha)\%$ credible interval may thereby be constructed by computing the $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ quantiles for the appropriate range of independent variables.

Finally, regarding predictive density $p(y_*|\mathbf{x}_*,\mathbf{y},\mathbf{X},\sigma^2,\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)$ we note from (37) that this is tantamount to computing the expectation $\mathbb{E}_{p(\boldsymbol{\beta}|\mathbf{y},\mathbf{X},\sigma^2,\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)}[p(y_*|\mathbf{x}_*,\boldsymbol{\beta},\sigma^2)]$ where $p(y_*|\mathbf{x}_*,\boldsymbol{\beta},\sigma^2) = N(\mathbf{x}_*^\top\boldsymbol{\beta},\sigma^2)$ and $p(\boldsymbol{\beta}|\mathbf{y},\mathbf{X},\sigma^2,\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0) = N(\boldsymbol{\mu}_n,\boldsymbol{\Sigma}_n)$. A moments reflection will reveal that we are dealing with a random variable $Y = \mathbf{x}_*^\top\boldsymbol{\beta} + \sigma Z$ where $\boldsymbol{\beta} \sim N(\boldsymbol{\mu}_n,\boldsymbol{\Sigma}_n)$ and $Z \sim N(0,1)$ are independent. Such a random variable will in turn be normal with $\mathbb{E}[Y] = \mathbf{x}_*^\top\mathbb{E}[\boldsymbol{\beta}] + \sigma^2\mathbb{E}[Z] = \mathbf{x}_*^\top\boldsymbol{\mu}_n$ and variance $\mathbb{V}\text{ar}[Y] = \mathbf{x}_*^\top\mathbb{V}\text{ar}[\boldsymbol{\beta}]\mathbf{x}_* + \sigma^2\mathbb{V}\text{ar}[Z] = \mathbf{x}_*^\top\boldsymbol{\Sigma}_n\mathbf{x}_* + \sigma^2$, i.e.

$$p(y_*|\mathbf{x}_*,\mathbf{y},\mathbf{X},\sigma^2,\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0) = N(\mathbf{x}_*^\top\boldsymbol{\mu}_n, \mathbf{x}_*^\top\boldsymbol{\Sigma}_n\mathbf{x}_* + \sigma^2). \tag{41}$$

Setting $\boldsymbol{\Sigma}_0^{-1} = \mathbf{0}$ we note that we recover the prediction interval obtained in (21). In particular, our maximum a posteriori for the prediction is the mean response $\mathbf{x}_*^\top\boldsymbol{\mu}_n$.

4.3. **Numerical Bayes: Markov Chain Monte Carlo.** Finding an explicit expression for the posterior may at times prove onerous if not downright impossible. Under these circumstances, it is desirable to have a numerical procedure in place that allows us to sample from the posterior distribution without ever having to calculate it explicitly. The *Markov Chain Monte Carlo* (MCMC) method known as the *Metropolis-Hastings* algorithm accomplishes just that. Specifically, starting from an arbitrary point $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta} \subseteq \mathbb{R}^k$ in parameter-space, it allows us to generate a sequence of samples $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, ..., \boldsymbol{\theta}_N, ...$ which eventually will converge to draws from the desired posterior distribution $p(\boldsymbol{\theta}|\mathbf{y},\mathbf{X},\boldsymbol{\xi})$. In generating consecutive variates in this sequence one will first need to specify a *proposal density* $q(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t)$, which need not have any connection to the posterior. By far the most common choice here is the random walk $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \boldsymbol{\varepsilon}_t$ where $\boldsymbol{\varepsilon} \sim N(\mathbf{0},\boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \sigma_2^2, ..., \sigma_k^2)$, although this really is down to the researcher's discretion.

Once we have a candidate solution $\boldsymbol{\theta}_{t+1}$ we then need to decide whether we accept or reject it. To this end we generate a uniform random variate $u \sim \mathcal{U}[0,1]$. If $u \leq \alpha(\boldsymbol{\theta}_{t+1}, \boldsymbol{\theta}_t)$ where

$$\alpha(\boldsymbol{\theta}_{t+1}, \boldsymbol{\theta}_t) \equiv \min \left\{ \frac{p(\boldsymbol{\theta}_{t+1}|\mathbf{y}, \mathbf{X}, \boldsymbol{\xi}) q(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1})}{p(\boldsymbol{\theta}_t|\mathbf{y}, \mathbf{X}, \boldsymbol{\xi}) q(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t)}, 1 \right\}. \tag{42}$$

we accept $\boldsymbol{\theta}_{t+1}$ and proceed with the next sample in the chain, otherwise we discard it and draw another candidate solution $\boldsymbol{\theta}_{t+1}$ from the proposal density. As suggested, after an initial *burn-in* phase, one should witness a certain stabilization in the sequence of generated $\boldsymbol{\theta}$. This means that we have started to sample from the posterior distribution as desired - and we're done! In summary, for Metropolis-Hastings we

(1) Posit on an initial point $\boldsymbol{\theta}_0$ in the chain, alongside a proposal density $q(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t)$.
(2) For $t = 0, 1, 2, ..., N$ we:
    (a) Draw the candidate solution $\boldsymbol{\theta}_{t+1}$ from $q(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t)$.
    (b) Accept said candidate with probability $\alpha(\boldsymbol{\theta}_{t+1}, \boldsymbol{\theta}_t)$. If accepted, we proceed to the next link in the chain. Otherwise, we continue drawing from the proposal density.

If all members of $\boldsymbol{\theta}$ have converged when $t = N$ we are successfully sampling from the posterior. Otherwise, the chain needs to be longer. Indeed, one might want to consider a more opportune starting point and proposal density.

The ingenious part of this algorithm is that while (42) does contain the posterior, it only does so through a ratio of posteriors. Recalling (36) this means that we avoid having to deal with the troublesome normalization constant. Specifically, we have

$$\frac{p(\boldsymbol{\theta}_{t+1}|\mathbf{y}, \mathbf{X}, \boldsymbol{\xi})}{p(\boldsymbol{\theta}_t|\mathbf{y}, \mathbf{X}, \boldsymbol{\xi})} = \frac{p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}_{t+1}) p(\boldsymbol{\theta}_{t+1}|\boldsymbol{\xi})}{p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}_t) p(\boldsymbol{\theta}_t|\boldsymbol{\xi})}.$$

Furthermore, working under the assumption of a random walk for the proposal density, we note that moving from $\boldsymbol{\theta}_t$ to $\boldsymbol{\theta}_{t+1}$ is just as likely as moving from $\boldsymbol{\theta}_{t+1}$ to $\boldsymbol{\theta}_t$, i.e. $q(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t) = q(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1})$. Jointly, this means that (42) can be written as

$$\alpha(\boldsymbol{\theta}_{t+1}, \boldsymbol{\theta}_t) \equiv \min \left\{ \frac{p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}_{t+1}) p(\boldsymbol{\theta}_{t+1}|\boldsymbol{\xi})}{p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}_t) p(\boldsymbol{\theta}_t|\boldsymbol{\xi})}, 1 \right\}, \tag{43}$$

which is computationally straight-forward.

Remark 7. For excellent technical exegeses of the Metropolis-Hastings algorithm we refer the reader to Chib and Greenberg [8] and Johannes and Polson [7]. The latter source being of particular interest to financial practitioners who seek to calibrate econometric models (stochastic volatility under the real-world probability measure etc.). There is also a very interesting overview of convergence theory. For basic to intermediate introductions see Lambert [3] and Rogers and Girolami [9].

References

[1] Hallam A. Restricted least squares, hypothesis testing, and prediction in the classical linear regression model. Accessed: 2020-04-21.

[2] O'hagan A. and Forster J. *Kendall's Advanced Theory of Statistics, Volume 2B, Bayesian Inference.* Wiley, 2004.

[3] Lambert B. *A Student's Guide to Bayesian Statistics.* Sage, 2018.

[4] Heij C., de Boer P., Franses P.H., Kloek T., and van Dijk H.K. *Econometric Methods with Applications in Business and Economics.* Oxford University Press, 2004.

[5] Casella G. and Berger R.L. *Statistical Inference.* Cengage Learning, 2 edition, 2001.

[6] James G., Witten D., Hastie T., and Tibshirani R. *An Introduction to Statistical Learning with Applications in R.* Springer, 2013.

[7] Johannes M. and Polson N. *MCMC Methods for Continuous-Time Financial Econometrics.* Elsevier, 2018.

[8] Chib S. and Greenberg E. Understanding the metropolis-hastings algorithm. *The American Statistician*, 49(4):327–335, 11 1995.

[9] Rogers S. and Girolami M. *A First Course in Machine Learning.* CRC Press, 2017.

[10] Leininger T. Unit 6: Simple linear regression lecture 3: Confidence and prediction intervals for slr. Accessed: 2020-04-22.