1 Remarks on the Lectures

1.1 Prerequisites and Background
i. MA 214: Applied Statistics (ANOVA, t-test, etc...)
ii. MA 242: Linear algebra (Real vector spaces).
iii. MA 442: Honors Linear algebra (abstract vector spaces).
iv. MA 581: Probability (CLT, etc...).

1.2 The R Language
i. Free statistical platform (cf. SPSS, SAS, STATA).
ii. A scripting language, or low-level.
iii. An object-oriented language.

1.3 On Taking Notes
i. Research has shown that students taking notes during lectures retain more information than students, who do not. This is mainly due to the fact that we memorize better, when we engage with the material through several sensory channels, simultaneously, such as by reading, listening and writing.

ii. Taking notes also preclude you from day-dreaming, and ensures that you stay focused on the task at hand. You are more likely to pay attention to what is being explained.

iii. Finally, even though a full transcript of the lectures will be available after each class, there is material and observations, which will not be contained in the lecture notes. Therefore, you will benefit from having your record of the classes.
2 Linear Models

Why are linear models so ubiquitous in every corner of data analysis?

i. They allow to understand how one variable is dependent over another.

ii. They permit to evaluate which variables are important.

iii. These are not black-box models. They are straightforward to compute and easy to communicate to other researchers in applied settings.

iv. Most relationships between real-life variables are, in fact, linear, or at least partly linear. Moreover, once this linear relationship has been accounted for, there generally does not remain a large amount of unexplained variance.

Deterministic Part

In simple regression, we are given two sequences of data points, and we wish to understand how they are related to each other. Each pair of observations will be here referred to as a case.

\[(y_i, x_i), \quad i = 1, \ldots, n.\]

In regression, as opposed to correlation, one of these variables is treated as the outcome variable or dependent variable, generally denoted by the \(y_i\)'s. We will then use the other variables for predicting that outcome. As a result, the other variables are referred to as predictors, or independent variables, and are denoted by \(x_i\)'s.

The univariate linear regression model consists of a mean function, expressed as a conditional expectation,

\[E[Y|X = x_i] = \beta_0 + \beta_1 x_i,\]

and a variance function, expressed as a conditional variance operator,

\[\text{Var}[Y|X = x_i] = \sigma^2.\]

The (unknown) parameters in this model are \((\beta_0, \beta_1, \sigma^2)\), where:

i. \(\beta_0\) is the y-intercept of \(E[Y|X]\), when \(X = 0\). Thus, we have

\[E[Y|X = 0] = \beta_0.\]

ii. \(\beta_1\) is the rate of change of \(E[Y|X]\), such that

\[E[Y|X = x + 1] - E[Y|X = x] = \beta_1.\]

iii. \(\sigma^2\) is the variance of \(Y\), given \(X\). It is strictly positive,

\[\sigma^2 > 0.\]
Stochastic Part

Naturally, except in the most trivial circumstances, we do not have a situation where $E[Y|X = x_i] = y_i$, for every $i = 1, \ldots, n$. This would be only true, if we were considering a completely deterministic model, which would not include any noise. The observables or observed data, denoted by $y_i$’s, therefore differ from the expected value of $Y$, given $X$. We account for this difference with the following equation,

$$y_i = E[Y|X = x_i] + e_i, \quad i = 1, \ldots, n,$$

where the $e_i$’s are the statistical error terms. In effect, we are here literally adding noise to the model. Thus, the $e_i$’s are collectively referred to as additive noise.

The $e_i$’s are defined as the difference between the observables and the conditional expected values –that is, $e_i = y_i - E[Y|X = x_i]$. Geometrically, the errors correspond to the vertical distances between each $y_i$ and each conditional expectation or mean function, $E[Y|X = x_i]$. Note that the error terms are not observable, since they depend on the unknown parameters ($\beta_0, \beta_1$).

Assumptions on the Error Terms

We make three crucial assumptions, which control the distribution of the error terms,

i. Firstly, we assume that the expectation of all error terms are centered on zero, such that

$$E[E_i|X = x_i] = 0, \quad i = 1, \ldots, n.$$

ii. Secondly, we assume that the error terms are uncorrelated, such that knowledge of one error term is not informative about another error term.

iii. Thirdly, we also assume that the variances of the error terms are constant for every $i = 1, \ldots, n$. This assumption is referred to as homoscedasticity.
W2: Ordinary Least Squares (OLS)

Historically, OLS is the original technique developed by Legendre and Gauss, in the early 1800s, while analyzing astronomical data. This is a method for estimating the unknown parameters in a regression model. Here, we proceed by minimizing a particular statistical criterion. Thus, we choose \((\beta_0, \beta_1)\) that minimize the following residual sum of squares (RSS),

\[
\text{RSS}(\beta_0, \beta_1) := \sum_{i=1}^{n} \left( y_i - (\beta_0 + \beta_1 x_i) \right)^2 .
\]

This may be re-written using the argmin operator, as follows,

\[
\left( \hat{\beta}_0, \hat{\beta}_1 \right) := \text{argmin}_{\beta_0, \beta_1 \in \mathbb{R}^2} \text{RSS}(\beta_0, \beta_1).
\]

Importantly, these minimizers can be obtained in closed-form.

Finally, it remains to estimate the variance parameter, \(\sigma^2\). This is done by normalizing the RSS, estimated at the estimated parameters \(\hat{\beta}_0\) and \(\hat{\beta}_1\),

\[
\hat{\sigma}^2 := \frac{1}{n - 2} \text{RSS}(\hat{\beta}_0, \hat{\beta}_1),
\]

where the denominator is \(n - 2\), because we have already estimated two parameters. That is, from the \(n\) data points that we have started with, we have used two degrees of freedom, and we therefore need to correct for this. We are here incurring a penalty for using some of the data points, which translate into a larger variance estimate. This estimated quantity, \(\hat{\sigma}^2\), is referred to as the residual mean square.

W3: Comparing Models

An alternative model to the one above is a model that only contains a single y-intercept. That is,

\[
\text{E}[Y|X = x_i] = \beta_0, \quad i = 1, \ldots, n,
\]

where \(\beta_0\) is estimated as the mean of the \(y_i\)’s,

\[
\hat{\beta}_0 = \frac{1}{n} \sum_{i=1}^{n} y_i.
\]

This is naturally a simpler model than the one containing both \(\beta_0\) and \(\beta_1\). Geometrically, such a model can be represented as a straight line. This basic model will be used as point of reference, against which we can evaluate the goodness-of-fit of the model including a slope coefficient for the \(x_i\)’s.

Model comparison proceeds by taking the ratio of the sums of squares of each model, such that the coefficient of determination is defined as follows,

\[
R^2 := 1 - \frac{\text{RSS}}{SYY},
\]

where

\[
SYY := \sum_{i=1}^{n} (y_i - \hat{\beta}_0)^2.
\]

It can be shown that this quantity is, in fact, the squared correlation between the predicted values, \(\hat{y}_i\)’s, and the observed values, \(y_i\)’s.
W4: Multiple Regression

The simple linear regression setting can be extended to the case of \( p \) independent variables, such that we may now have the following array of data points,

\[
(y_i, x_{i1}, \ldots, x_{ip}), \quad i = 1, \ldots, n,
\]

with the first predictor is a dummy variable, such that \( x_{i1} = 1 \), for every \( i \). In this case, multiple regression, also referred to as the general linear model (GLM), is expressed as follows,

\[
y_i = \sum_{j=1}^{p} x_{ij} \beta_j + e_i, \quad i = 1, \ldots, n,
\]

which may then be reformulated, using linear algebra,

\[
y_i = x_i^T \beta + e_i, \quad i = 1, \ldots, n,
\]

where \( \beta \) and the \( x_i \)'s are \((p \times 1)\) vectors. Alternatively, one could also re-express this as a single equation, alongside the assumption on the error terms,

\[
y = X\beta + e, \quad \text{Var}[e] = \sigma^2 I,
\]

where \( y \) and \( e \) are \((n \times 1)\) vectors, \( X \) is an \((n \times p)\) matrix, and \( \beta \) is a \((p \times 1)\) vector.

The generality of the GLM comes from the fact that it subsumes most standard statistical models, such as the \( t \)-test, or the analysis of variance (ANOVA). For multiple regression, we make the four main assumptions:

1. **Linearity** in the parameters;
2. Error terms are uncorrelated;
3. Error terms have identical variance;
4. \( X \) has full rank.

W5: Parameter Estimation

Once we have obtained specific estimates of the parameters of interest, \((\hat{\beta}_0, \hat{\beta}_1, \sigma^2)\), it is reasonable to ask whether these estimates are somewhat stable. In particular, we may wish to ask certain questions about the relationship between the estimates, \((\hat{\beta}_0, \hat{\beta}_1)\), and the true population parameters, \((\beta_0, \beta_1)\). This is only meaningful if we treat our original set of data points as a random sample, from an unknown population distribution, such that

\[
(Y_i, x_i), \quad i = 1, \ldots, n.
\]

(This is one point, upon which the two schools of thought in statistics—the frequentist and Bayesian viewpoints—differ substantially.)

To do so, we will need to make another crucial assumption about the distribution of the error terms, such that

\[
e_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n;
\]

which means that all the \( e_i \)'s are independently and identically distributed as normal variates, centered at zero with variance \( \sigma^2 \). Naturally, by an appeal to the central limit theorem (CLT), it turns out that this condition holds asymptotically, as the number of data points goes to infinity, \( n \to \infty \).
When this condition is satisfied, we can use this extra knowledge to compute the **confidence intervals** of each parameter, and evaluate the probability of these parameters to be non-zero, thereby providing a hypothesis test for each of them. For instance, considering the \( y \)-intercept, \( \beta_0 \), we may compute the following \( t \)-statistic,

\[
t = \frac{\hat{\beta}_0 - \beta_0^*}{se(\hat{\beta}_0)},
\]

where under the **null hypothesis**, it is assumed that \( \beta_0^* = 0 \), and where the denominator is the **standard error** of \( \hat{\beta}_0 \).

**W6: Bootstrap**

Another important assumption made when conducting simple or multiple regression is that the error terms are **normally distributed**. However, in many practical situations, such an assumption is difficult to verify or is likely to be untenable. When this occurs, one can resort to the **bootstrap estimation** of the parameters of interest. Such an estimation is conducted by sampling with replacement from the empirical distribution function (EDF) of the \( y_i \)'s. This gives a set of \( m \) bootstrap samples,

\[
y_i^{(t)} \sim \text{EDF}(y_1, \ldots, y_n), \quad i = 1, \ldots, n; \quad t = 1, \ldots, m.
\]

**W8: Weighted Least Squares (WLS)**

In OLS, we assume that the mean and variance functions correspond to the following conditional moments,

\[
E[Y|X = x_i] = \beta_0 + \beta_1 x_i, \quad \text{Var}[Y|X = x_i] = \sigma^2.
\]

However, there may be situations, when we know for sure that the variance function is not identical for every \( i \). We refer to such situations as **heteroscedasticity**. This may happen when the data points are **funnel-shaped**, for instance. We can then re-weight the individual variances as follows,

\[
\text{Var}[Y|X = x_i] = \frac{\sigma^2}{w_i}, \quad i = 1, \ldots, n,
\]

where the weights \( w_i \)'s are **known positive** numbers. This setup naturally extends to multiple regression, such that we obtain the following modified general linear model,

\[
y = X\beta + e, \quad \text{Var}[e] = \sigma^2 W^{-1},
\]

where \( W \) is a diagonal matrix of order \((n \times n)\), with the \( w_i \)'s in the diagonal.

**W9: Polynomial Regression**

Another variant on the standard setting of multiple regression is the use of **exponentiated independent variables**. In this case, predictors may be elevated to different orders, such that the mean function of our regression model now becomes,

\[
E[Y|X] = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2 + \ldots + \beta_d x_d^d.
\]

Is this model still **linear**? By linear models, in this course, and in statistics in general, we are referring to **linearity in the parameters**. The mean function that we are estimating is a function of both the \( x_i \)'s and the vector of parameters \( \beta = [\beta_0, \beta_1]^T \). Therefore, given any \( x_i = [x_{i1}, x_{i2}]^T \), where \( x_{i1} = 1 \), we have

\[
E[Y|X = x_i] = x_i^T \beta =: f(\beta; x_i).
\]
Therefore, the mean function is linear in $\beta$, and this still holds under any non-linear transformation of the vector of IV, $x_i$. That is,

$$E[Y|X = x_i] = \phi(x_i)^T \beta,$$

where $\phi$ is some non-linear map of the $x_i$’s.

**W10-15: Transformations, Outliers, Residuals**

In the remainder of the course, we will also touch on other issues, such as data transformations to correct for non-normality or reduce the variance of the parameters of interest. We will investigate how to detect outliers, and learn how to evaluate the influence of individual cases. Most importantly, we will also consider how to assess the fit of a model using its residuals. Finally, we will briefly cover non-linear models, such as logistic regression, as an example of the generalized linear model.