1 Introduction

Statistical learning refers to a broad set of tools that can be used to explore and model complex datasets. There are overlaps with computer science and machine learning, where statistical science often places a greater emphasis on identifying and quantifying sources of uncertainty.

[The spam dataset consists of word frequencies and whether or not an e-mail is marked as spam]

[The CU FCQ dataset consists of course questionnaire outcomes such as instructor effectiveness, hours spent on homework, prior student interest in class]

The CU FCQ dataset consists of course questionnaire outcomes such as instructor effectiveness, hours spent on homework, prior student interest in class, etc. We might interpret hours spent on homework as an input variable while instructor effectiveness is an output variable. Our goal is to characterize a relationship between inputs and outputs.

Notation 1. \( X \) will refer to input variables, which also be called predictors, independent variables, features or variables. The output variable will be denoted by \( Y \), and is sometimes called the response or dependent variable.

With multiple inputs, we distinguish between them using adorned \( X \)s, e.g., \( X_1 = \) hours spent on homework and \( X_2 = \) prior interest in class.

Given \( p \) types of predictors \( X_1, \ldots, X_p \), we assume there is some type of relationship

\[
Y = f(X_1, \ldots, X_p) + \varepsilon
\]
where $f$ is a fixed but unknown function of the inputs and $\varepsilon$ is a random, mean zero error term. In this model, $f$ represents the systematic information that $X_1, \ldots, X_p$ provide about $Y$, while $\varepsilon$ is a mean zero error term that represents stochastic (or random) error that cannot be explained using $X_1, \ldots, X_p$.

### 1.1 Typical Goals

The typical major goals are to

- Find relationships between a group of explanatory variables and a response variable that provides good predictive performance
- Interpret this relationship
- Reduce the size of a group of predictors for scientific or computational purposes.

Some possibilities are:

- Classification and regression
- Algorithms for large data analysis
- Recommender systems
- Spam filters
- Text processing
- Disease monitoring

Generally, most of the above reduce to estimating $f$ in

$$Y = f(X_1, \ldots, X_p) + \varepsilon.$$ 

We may wish to do this for many reasons, which can be roughly categorized into prediction and inference.
**Prediction**

In many cases, the input variables \( X_1, \ldots, X_p \) are easily available (or are variables that we can control), whereas \( Y \) is quantity of main interest. If we could estimate \( f \), say using \( \hat{f} \), and we knew \( X_1, \ldots, X_p \), then we could **predict** \( Y \) using

\[
\hat{Y} = \hat{f}(X_1, \ldots, X_p).
\]

For example, we might like to know if doubling homework would lead to an instructor being more effective \((\hat{f}(2X_1, X_2))\).

The accuracy of \( \hat{Y} \) as a predictor for \( Y \) depends on two quantities of **reducible error** and **irreducible error**. The expected squared error between our predictor and the response is

\[
\mathbb{E}(Y - \hat{Y})^2 = \mathbb{E}(f(X_1, \ldots, X_p) + \varepsilon - \hat{f}(X_1, \ldots, X_p))^2
\]

\[
= \mathbb{E}(f + \varepsilon - \hat{f})^2
\]

\[
= \mathbb{E}((f - \hat{f})^2 + \varepsilon^2 - 2\varepsilon(f - \hat{f}))
\]

\[
= \mathbb{E}(f - \hat{f})^2 + \text{Var} \varepsilon.
\]

The first term is **reducible** while the second term is **irreducible**. In other words, by using more data or better learning techniques, we can improve the estimate of \( \hat{f} \approx f \), whereas \( \varepsilon \) is random and cannot be predicted. For prediction, our goal will be to **minimize** this expected squared error \( \mathbb{E}(Y - \hat{Y})^2 \).

**Inference**

*Inference* refers to the act of estimating \( f \) and characterizing/quantifying the unpredictable error \( \varepsilon \) using a probabilistic model. Major inferential questions are:

- Which predictors are associated with the response? (Does frequency of “budget” imply an e-mail is spam?)
- What is the relationship between \( Y \) and each \( X_i \)? (Does frequency of “budget” increase or decrease likelihood an e-mail is spam?)
- Are these relationships linear or nonlinear? (Can we use \( f(X) = \beta_0 + \beta_1 X \) or do we need something more complicated?)
1.2 Some Jargon

Parametric vs. Nonparametric

Learning methods for estimating $f$ can be roughly categorized as parametric or nonparametric (or a mix between the two which is sometimes called semiparametric). For example, a parametric model for $Y = \text{global temperature given } X = \text{CO}_2$ concentration might use

$$f(X) = \beta_0 + \beta_1 X$$

or in other words

$$Y \approx \beta_0 + \beta_1 X.$$ 

The major obstacle is then to estimate the unknown parameters $\beta_0$ and $\beta_1$, for example using least squares. Once we have estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, we can predict $Y$ by

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X.$$ 

A nonparametric model makes no explicit assumption about the form of the relationship between $Y$ and $X_1, \ldots, X_p$, but seeks an $f$ that gets close to a set of training data points without being too wiggly or rough. Typically nonparametric methods are more flexible, but are not as easy to interpret as parametric models, and the modeler must weigh between these two as techniques are applied.

Supervised vs. Unsupervised

Learning techniques can be either supervised or unsupervised. Supervised techniques apply to the cases where we have a response of interest $Y$, given some covariates $X_1, \ldots, X_p$. Most of this class is devoted to supervised learning. Unsupervised learning applies when no response $Y$ is available, but we are interested in learning about (for example) the relationship between covariates. For example, given $X_1, \ldots, X_n$, cluster analysis seeks to determine if the observations fall into distinct groups.

[E.g., given a set of e-mails, determine which is spam, or given a set of Netflix ratings on a shared account, how to tell which rating comes from which user]
Regression vs. Classification

A variable $X$ (or $Y$) is either quantitative, $X \in \mathbb{R}$, or qualitative, $X \in \{s_1, \ldots, s_m\}$. Sometimes qualitative variables are known as categorical. For instance, salary would be a quantitative variable whereas gender would be a categorical variable as it breaks up into classes or categories. Problems with a quantitative response are referred to as regressions while qualitative responses are known as classifications.

A Common Theme: Bias-Variance Tradeoff

A common way to assess the quality of fit of a model is by examining the predictive mean squared error. That is, for a new observation $Y = f(X) + \varepsilon$ and given an estimator $\hat{f}(X)$, we want to minimize

$$\mathbb{E}(Y - \hat{f}(X))^2 = \mathbb{E}(Y - \hat{f})^2.$$

Before we get data, $\hat{f}$ is random (because it depends on unobserved random data $Y_1, \ldots, Y_n$). Assuming $\varepsilon$ and $\hat{f}$ are uncorrelated, and that $f$ is fixed, we have

$$\mathbb{E}(Y - \hat{f})^2 = \mathbb{E}(f + \varepsilon - \hat{f})^2$$

$$= \mathbb{E}(f - \hat{f})^2 + \mathbb{E}\varepsilon^2 + 2\mathbb{E}(\varepsilon(f - \hat{f}))$$

$$= \mathbb{E}(f - \mathbb{E}\hat{f} + \mathbb{E}\hat{f} - \hat{f})^2 + \text{Var}\varepsilon$$

$$= (f - \mathbb{E}\hat{f})^2 + \mathbb{E}(\hat{f} - \mathbb{E}\hat{f})^2 + \text{Var}\varepsilon$$

$$= (f - \mathbb{E}\hat{f})^2 + \text{Var}\hat{f} + \text{Var}\varepsilon$$

$$= (\text{bias of } \hat{f})^2 + \text{variance of } \hat{f} + \text{random error}$$

since $\mathbb{E}(\mathbb{E}\hat{f} - \hat{f}) = 0$.

[Picture of squared-bias/variance tradeoff with sum as MSE]

The predictive mean squared error breaks into the squared bias of $\hat{f}$, the variance of $\hat{f}$ and an irreducible and unpredictable variance of the error term. As the model for $\hat{f}$ becomes more flexible, we reduce the bias but increase the variance, whereas more rigid models exhibit smaller variance at the cost of increased bias. [Picture of MSE on y-axis as function of model complexity on x-axis with squared bias decreasing variance increasing]
2 Linear Regression

Linear regression is a simple but powerful approach for supervised learning. Given a set of predictors (or features or covariates or explanatory variables) $X_1, \ldots, X_p$ and a response (or supervisor) $Y$, we should be able to answer the following questions:

- Is there a relationship between $Y$ and $X_i$? If so, how strong is it?
- How accurately can we quantify a relationship?
- How accurately can we predict $Y$, given a set of covariates?
- Is the relationship linear, and are there interactions between $X_1, \ldots, X_p$?

2.1 Simple Linear Regression

In the simple linear regression case, we have a single quantitative response $Y$ and a single predictor $X$. We assume that, up to some random error, $Y$ and $X$ share a linear relationship,

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

(1)

where $\varepsilon$ is the random error term. Sometimes we say $Y$ is being regressed on $X$. This is our first example of a statistical model. Here, $\beta_0$ and $\beta_1$ are unknown coefficients or parameters which must be estimated. This is the typical model with $f(X) = \beta_0 + \beta_1 X$. This function is known as the population regression line, and $\varepsilon$ is the residual. Of course, $\beta_0$ is the estimated value of $Y$ for $X = 0$ and $\beta_1$ is the increase in $Y$ for a unit increase of $X$.

In practice we have a set of $n$ observation pairs

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

of both covariates $x_i$ and responses $y_i$. If the relationship (1) holds, then we have a set of observation equations

$$y_1 = \beta_0 + \beta_1 x_1 + \varepsilon_1$$

$$\vdots$$

$$y_n = \beta_0 + \beta_1 x_n + \varepsilon_n$$
where the $\varepsilon_i$ are now fixed numbers. Equivalently, a priori these are random variables so

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n$$

(2)

where $\varepsilon_i$ are random.

There are differing levels of assumptions that are typically made in practice regarding (2).

A1 1. The relationship (1) holds
   2. $\varepsilon_i$ are iid (independent and identically distributed) $N(0, \sigma^2)$ for all $i = 1, \ldots, n$
   (i.e. $Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$)

A2 1. The relationship (1) holds
   2. $E\varepsilon_i = 0$
   3. $\text{Var}\varepsilon_i = \sigma^2$ (homoskedasticity)
   4. $\varepsilon_i$ and $\varepsilon_j$ are iid

A3 1. The relationship (1) holds
   2. $E\varepsilon_i = 0$
   3. $\text{Var}\varepsilon_i = \sigma^2$ (homoskedasticity)
   4. $\varepsilon_i$ and $\varepsilon_j$ are uncorrelated for $i \neq j$

A4 1. The relationship (1) holds
   2. $E\varepsilon_i = 0$
   3. $\text{Var}\varepsilon_i < \infty$.

[Picture of A1: what would we expect plot to look like with $n \gg 0$?]

What is common to all levels of assumptions is that the posited relationship (1) holds. The differences between the three assumptions are about the assumed behavior of the residuals. A1 is the strongest in that we explicitly assume the residuals are normally distributed, whereas A2 relaxes this assumption to being iid from some unspecified probability distribution. Statistical independence is a stronger statement than correlation, and A4 is the
weakest in that we only put assumptions on the first two moments of the variables. For instance, $\mathbb{E}\varepsilon_1^2$ may be different than $\mathbb{E}\varepsilon_2^2$ in A3, but not for A2 or A1.

The fourth assumption is the weakest, and relaxes the homoskedasticity property (equal variance among all residuals) by allowing for heteroskedasticity (nonconstant variance). We will work with A1. Our goals are usually

1. Estimate parameters $\beta_0, \beta_1$ and $\sigma^2$ (and quantify uncertainty in our estimates)
2. Assess the validity of our assumed model (1)
3. Predict response at a new covariate value $X = x$

### 2.2 OLS

The heuristic for estimating the regression parameters $\beta_0$ and $\beta_1$ is that we should minimize the distance between $y_i$ and its fitted value $\hat{\beta}_0 + \hat{\beta}_1 x_i$.

[Picture of data with candidate lines]

The ordinary least squares (OLS) estimators for $\beta_0$ and $\beta_1$ are found by minimizing

$$RSS = R(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

where $RSS$ refers to the residual sum of squares. Note that other types of distance can be used, but least squares is the most popular, in part due to the simplicity of the solution (and also a connection to maximum likelihood).

Doing this minimization results in the least squares estimators

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where $\bar{x} = n^{-1} \sum_{i=1}^{n} x_i$ and $\bar{y} = n^{-1} \sum_{i=1}^{n} y_i$. Notice that the estimators are linear in the observations. It can be shown that the OLS estimators are the best linear unbiased estimators (BLUEs) under squared loss. Note

$$\mathbb{E}Y_i = \beta_0 + \beta_1 x_i$$
and
\[ EY = \beta_0 + \beta_1 \bar{x}. \]

The OLS estimators are *unbiased* in that
\[
E \hat{\beta}_1 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})E(Y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}
\]
\[
= \sum_{i=1}^{n} (x_i - \bar{x})(\beta_1(x_i - \bar{x})) / \sum_{i=1}^{n} (x_i - \bar{x})^2
\]
\[
= \beta_1 \sum_{i=1}^{n} (x_i - \bar{x})^2 / \sum_{i=1}^{n} (x_i - \bar{x})^2
\]
\[
= \beta_1
\]

and
\[
E \hat{\beta}_0 = EY - \bar{x}E \hat{\beta}_1
\]
\[
= \beta_0 + \beta_1 \bar{x} - \bar{x} \beta_1
\]
\[
= \beta_0.
\]

Lastly, we must estimate \( \sigma^2 \). An unbiased estimate of \( \sigma^2 \) can be found by scaling the residual sum of squares (RSS) by its degrees of freedom,
\[
\hat{\sigma}^2 = \frac{RSS}{n-2} = \frac{1}{n-2} \sum_{i=1}^{n} \varepsilon_i^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2
\]
where the \(-2\) occurs since we are estimating two parameters \( \beta_0 \) and \( \beta_1 \) in the mean function.

Recall the usual unbiased estimator for variance divides by \( n-1 \), which is when only \( \beta_0 \) is being estimated. Sometimes \( \hat{\sigma} \) is known as the *standard error of regression*.

[R example (IntroLinearRegression.R)]

Given estimates \( \beta_0, \beta_1, \) and \( \sigma \), can we assess the uncertainty in these? Note that \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) both involve \( y \), so before the experiment is performed these are random variables. Additionally note that if we work under A3, the \( \varepsilon_i \) are uncorrelated, mean zero and variance
\( \sigma^2 \), so [using \( \sum (x_i - \bar{x})Y = 0 \)]

\[
\text{Var}(\hat{\beta}_1) = \text{Var} \left( \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \varepsilon_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right) \\
= \frac{1}{(\sum_{i=1}^{n}(x_i - \bar{x})^2)^2} \text{Var} \left( \sum_{i=1}^{n}(x_i - \bar{x}) \varepsilon_i \right) \\
= \frac{1}{(\sum_{i=1}^{n}(x_i - \bar{x})^2)^2} \sum_{i=1}^{n}(x_i - \bar{x})^2 \sigma^2 \\
= \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.
\]

Similarly, we get

\[
\text{Var}(\hat{\beta}_0) = \frac{\sigma^2}{n} + \frac{\sigma^2 \bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.
\]

[Note that as \( n \to \infty \) and \( \sum_{i=1}^{n}(x_i - \bar{x})^2 \to \infty \) we get consistency]. Finally, the two estimators are negatively correlated with

\[
\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\sigma^2 \frac{\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.
\]

Note that if A1 held, i.e., the \( \varepsilon_i \) are iid \( N(0, \sigma^2) \), then \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \), as linear combinations of normal random variables, are normal. In particular, under A1,

\[
\hat{\beta}_0 \sim N \left( \beta_0, \frac{\sigma^2}{n} + \frac{\sigma^2 \bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right) \\
\hat{\beta}_1 \sim N \left( \beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right).
\]

Additionally, it can be shown that

\[
\frac{n-2}{\sigma^2} \sigma^2 \sim \chi^2_{n-2}.
\]

**Confidence Intervals**

An aside: If \( X \sim N(\mu, \sigma^2) \), note that

\[
P \left( -1.96 \leq \frac{X - \mu}{\sigma} \leq 1.96 \right) \approx 0.9500042 \approx 95\%.
\]
so,
\[ P(\mu - 1.96\sigma \leq X \leq \mu + 1.96\sigma) \approx 95\% \]
or,
\[ P(X - 1.96\sigma \leq \mu \leq X + 1.96\sigma) \approx 95\%. \]
The middle statement says that \( \mu \pm 1.96\sigma \) contains \( X \) 95% of the time. The last statement says that \( X \pm 1.96\sigma \) will contain \( \mu \) 95% of the time, and is the basis for constructing a confidence interval. Sometimes this is approximated as \( X \pm 2\sigma \). [Note one of these is a fixed interval while the other is random].

Let’s simplify notation just a bit. Call

\[ C = \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \]

so that

\[ \hat{\beta}_0 \sim N(\beta_0, \sigma^2 C). \]

Suppose we knew \( \sigma^2 \). We call the following an exact 95% confidence interval for \( \beta_0 \)

\[ \left( \hat{\beta}_0 - 1.96\sqrt{\text{Var} \hat{\beta}_0}, \hat{\beta}_0 + 1.96\sqrt{\text{Var} \hat{\beta}_0} \right) \]

\[ = \left( \hat{\beta}_0 - 1.96\sigma\sqrt{C}, \hat{\beta}_0 + 1.96\sigma\sqrt{C} \right). \]

This would be the interval that contains \( \beta_0 \) 95% of the time, or, in other words, if we repeated the experiment 1000 times, we would expected this interval to contain \( \beta_0 \) 950 of those times. We don’t know \( \sigma^2 \), however, so we must rely on a plug-in estimator \( \hat{\sigma}^2 \) when constructing such an interval.

Two options:

Option 1 Use

\[ \hat{\beta}_0 \pm 1.96\hat{\sigma}\sqrt{C} \quad \text{or} \quad \hat{\beta}_0 \pm 2\hat{\sigma}\sqrt{C} \]

as an approximate 95% confidence interval for \( \beta_0 \).

Option 2 Be more careful.
Recall that a $t$ random variable with $r$ degrees of freedom can be represented as

$$T = \frac{X}{\sqrt{Y/r}} = d \frac{N(0,1)}{\sqrt{\chi^2_r/r}}$$

where $X \sim N(0,1)$ and $Y \sim \chi^2_r$ are independent random variables and where we write $T \sim t_r$. [Picture of $t_r$ distribution compared to normal] Under A1, $\hat{\beta}_0 \sim N(\beta_0, \sigma^2 C)$ is exact. Thus,

$$\frac{\hat{\beta}_0 - \beta_0}{\sigma \sqrt{C}} \sim N(0,1)$$

and

$$\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma} \sqrt{C}} = d \frac{N(0, \sigma^2)}{\hat{\sigma}} = d \frac{N(0, \sigma^2)}{\sqrt{\hat{\sigma}^2}} = d \frac{N(0, \sigma^2)}{\hat{\sigma}^2} = d \frac{N(0, \sigma^2)}{\sqrt{(n-2)\hat{\sigma}^2}}$$

$$= d \frac{N(0,1)}{\sqrt{\chi^2_{n-2}/(n-2)}} = d t_{n-2}$$

The last step requires a little extra care since it’s not obvious that $\hat{\beta}_0$ and $\hat{\sigma}^2$ are independent.

Now, this implies

$$P \left( t_{n-2}(0.025) \leq \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma} \sqrt{C}} \leq t_{n-2}(0.975) \right) = 0.95$$

where $t_{n-2}(0.025)$ is the 2.5% quantile of a $t_{n-2}$ distribution. [Note that, by symmetry, $t_{n-2}(0.025) = -t_{n-2}(0.975)$]

Option 2 Under A1, an exact 95% confidence interval for $\beta_0$ is

$$\hat{\beta}_0 \pm t_{n-2}(0.025) \hat{\sigma} \sqrt{C}$$

Note that, in general if we wanted a $100(1 - \alpha)%$ confidence interval, we would use

$$\hat{\beta}_0 \pm t_{n-2}(\alpha/2) \hat{\sigma} \sqrt{C}$$

$$\left( \pm t_{n-2}(1 - \alpha/2) \hat{\sigma} \sqrt{C} \right)$$

where $t_{n-2}(\alpha/2)$ is the lower $\alpha/2 \times 100$% quantile of the $t_{n-2}$ distribution. [e.g., if $\alpha = 0.05$, then $100(1 - \alpha)% = 95%$] Table 1 gives some typical values of $t_{n-2}(0.975)$. [Note that small
Table 1: Values of $t_{n-2}(0.975)$ for various $n$.

<table>
<thead>
<tr>
<th>$n - 2$</th>
<th>10</th>
<th>20</th>
<th>50</th>
<th>100</th>
<th>200</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_{n-2}(0.975)$</td>
<td>2.23</td>
<td>2.09</td>
<td>2.01</td>
<td>1.98</td>
<td>1.97</td>
</tr>
</tbody>
</table>

$n$ implies we end up using wider confidence intervals, i.e., we exhibit less confidence if we don’t approximate by using normal cutoffs.

The $\alpha$ level controls the Type I error – that is, the frequency of which the CI does not contain the truth. E.g., $\alpha = 0.05$, then 95% of experiments’ CIs will contain the true parameters. [This is no guarantee that this experiment’s CI does!]

In summary, 95% approximate CIs for our estimated parameters are

\[
\hat{\beta}_0 \pm 1.96SE(\hat{\beta}_0) \\
\hat{\beta}_1 \pm 1.96SE(\hat{\beta}_1)
\]

while exact 95% CIs are

\[
\hat{\beta}_0 \pm t_{n-2}(0.025)SE(\hat{\beta}_0) \\
\hat{\beta}_1 \pm t_{n-2}(0.025)SE(\hat{\beta}_1)
\]

where $SE = \sqrt{\text{Var}}$ are standard errors, that is, estimated standard deviations. In particular,

\[
\text{Var}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \\
\text{Var}(\hat{\beta}_0) = \frac{\hat{\sigma}^2}{n} + \frac{\hat{\sigma}^2 \overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2}.
\]

The standard error attempts to quantify an answer to the question how certain are we about the value of the estimator?

**Hypothesis Testing**

[3 graphs of possible dependence – 1 obvious correlation, 1 little noise close to zero correlation, 1 very noisy but possible linear dependence]
A hypothesis test usually looks like

\[ H_0 : \beta_1 = 0 \]
\[ H_A : \beta_1 \neq 0 \]

where \( H_0 \) is called the null hypothesis and \( H_A \) is the alternative hypothesis. That is, our working hypothesis is that the slope \( \beta_1 \) is zero, i.e., that \( X \) and \( Y \) do not actually share a (linear) relationship. The goal then, based on noisy data, is to assess whether \( H_0 \) is a reasonable hypothesis.

Hypothesis testing follows a series of steps:

1. Set an \( \alpha \) level which controls Type I error, that is, the probability of incorrectly rejecting \( H_0 \) when \( H_0 \) is true. Typically \( \alpha = 0.05 \) or 0.01 (e.g., if \( H_0 \) is true then in 5% of datasets we will accidentally reject \( H_0 \) anyway).

2. Form hypothesis

\[ H_0 : \beta_1 = 0 \]
\[ H_A : \beta_1 \neq 0. \]

3. Develop a theoretical test statistic

\[ \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} \]

and derive its theoretical distribution under \( H_0 \). This statistic here is known as a \( t \) statistic, because under under A1 we have the exact distribution

\[ \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} \sim t_{n-2}. \]

4. Calculate the test statistic under the data. If

\[ \left| \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} \right| > t_{n-2}(1 - \alpha/2). \]

then we reject \( H_0 \) in favor of \( H_A \). Otherwise we fail to reject \( H_0 \).
The heuristic behind step 4 is that if the statistic takes on an unusual value (i.e., falls in the tails of the $t_{n-2}$ distribution) then we might suspect that our null hypothesis is faulty.

Note

$$\left| \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} \right| > t_{n-2}(1 - \alpha/2) \iff 0 \not\in (\hat{\beta}_1 - t_{n-2}(1 - \alpha/2)SE(\hat{\beta}_1), \hat{\beta}_1 + t_{n-2}(1 - \alpha/2)SE(\hat{\beta}_1))$$

so, equivalently, if our confidence interval for $\beta_1$ doesn’t include 0 we would reject $H_0$ in favor of $H_A$. If the confidence interval does include zero, we say that we **fail to reject** $H_0$ (which is different than accepting $H_0$ – the failure is not evidence for $H_0$, we just don’t have evidence that it’s a particularly bad hypothesis).

Note we could just as easily set up a test of the form

$$H_0 : \beta_1 = 5$$

$$H_A : \beta_1 \neq 5$$

(or $\beta_1^*$ instead of 5), in which case our test would reject if

$$\left| \frac{\hat{\beta}_1 - 5}{SE(\hat{\beta}_1)} \right| > t_{n-2}(1 - \alpha/2).$$

The identical argument for

$$H_0 : \beta_0 = \beta_0^*$$

$$H_A : \beta_0 \neq \beta_0^*$$

works, in which case we reject if

$$\left| \frac{\hat{\beta}_0 - \beta_0^*}{SE(\hat{\beta}_0)} \right| > t_{n-2}(1 - \alpha/2).$$

Note that to perform a hypothesis test or create a confidence interval we have to choose the $\alpha$-level (traditionally 5%). An alternative is to report the *p*-value, that is, *the probability of observing something more extreme than our test statistic under $H_0$*. In these cases this p-value is known as a *two sided p-value*.  

Some warnings:
• Interpret a p-value with care: it is the probability of seeing something as-or-more extreme than our observed test statistic under $H_0$. For example, if we set the $\alpha$ level at 5% and $H_0$ is true we would still see “significant” p-values 5% of the time.

• If we cannot reject $H_0$ based on our test statistic at a particular $\alpha$ level, this is NOT evidence for the null, this just means there is insufficient evidence against $H_0$.

Hypothesis testing, confidence intervals and p-values all give (almost) the same information

$$\hat{\beta}_1 \pm t_{n-2}(1 - \alpha/2)SE(\hat{\beta}_1) \text{ does not contain zero}$$

$$\iff$$

$$H_0: \beta_1 = 0 \text{ is rejected at level } \alpha$$

$$\iff$$

$$p\text{-value } p < \alpha.$$

The $t_{n-2}$ quantiles we use throughout this section are exact under assumptions A1, however in reality nothing is normal, but due to the central limit theorem these quantiles are still approximate even for non-normal data when $n \geq 40$ to 50.

**ANOVA**

An analysis of variance (ANOVA) table is often used to compare competing linear models. For example, we might want to compare the two models

$$Y = \beta_0 + \varepsilon \quad \text{(3)}$$

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad \text{(4)}$$

(of course we already know this could boil down to testing $H_0 : \beta_1 = 0$). If (3) were the correct model, then $\hat{\beta}_0 = \bar{y}$ would be the least squares estimate, and the residual sum of squares would be

$$SYY = \sum_{i=1}^{n} (y_i - \hat{\beta}_0)^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$
Table 2: Analysis of variance table for simple regression.

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>$SS_{reg}$</td>
<td>$SS_{reg}/1$</td>
<td>$MS_{reg}/\hat{\sigma}^2$</td>
<td></td>
</tr>
<tr>
<td>Residual</td>
<td>$n-2$</td>
<td>$RSS$</td>
<td>$\hat{\sigma}^2 = RSS/(n-2)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>$n-1$</td>
<td>$SYY$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For the “full” model (4), we have the residual sum of squares as

$$RSS = \sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

using the OLS estimators for $\beta_0$ and $\beta_1$. $SYY$ is an estimate of remaining variability under (3) while $RSS$ is an estimate of remaining variability under (4).

If $\hat{y}_i$ is the fitted value under (4), then

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

and it can be shown that

$$SYY = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = SS_{reg} + RSS$$

where we have defined $SS_{reg}$ as the sum of squares due to regression. Finally recall that

$$\hat{\sigma}^2 = \frac{RSS}{n-2}$$

Then the analysis of variance table (ANOVA table) for simple regression is shown as Table 2.

A few notes:

- If $SS_{reg}$ is large then (4) is a big improvement over (3). To quantify this improvement, we compare comparing mean square regression errors

$$F = \frac{SS_{reg}/1}{RSS/(n-2)} = \frac{MS_{reg}}{\hat{\sigma}^2}$$

where $MS_{reg} = SS_{reg}/1$ is the mean square error of regression. If (4) is true, we expect $MS_{reg} \gg \hat{\sigma}^2$ and $F \gg 1$. 

17
Under A1 (or if the sample size is sufficiently large) and if (3) is assumed to be the true model, then $F$, the $F$-statistic is distributed as a $F(1, n-2)$ random variable, and thus if its p-value is large this is evidence that (4) is a better model.

[Picture of $F$-distribution and p-value]

- The df are degrees of freedom that correspond to the number of free variables minus the number of estimated parameters (e.g. for RSS it is $n-2$ due to estimation of $\beta_0$ and $\beta_1$).
- The regular test of $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$ relies on the $t$-statistic where

$$t^2 = \left( \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \right)^2 = F$$

after some manipulation of terms. Thus, the $F$-test in the ANOVA table is equivalent to the $t$-test in this simple case.

We will see ANOVA as a useful tool to compare more complicated mean functions in the next section.

**Coefficient of Determination $r^2$**

Define the coefficient of determination ($r^2$ or $R^2$) as

$$r^2 = 1 - \frac{RSS}{SYY} = \frac{SS_{reg}}{SYY}.$$  

Note $r^2 \in [0, 1]$ is one minus the remaining unexplained variability. If $RSS/SYY$ is close to 1 then we didn’t improve much using our linear model, whereas if $RSS/SYY \approx 0$ then $r^2 \approx 1$ and we explained most of the variability in the data using our model. Indeed, $r^2$ can be interpreted as the percentage of explained variability. For simple regression we have $r^2$ is the squared sample correlation between $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$:

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{SXY}{\sqrt{SXX \cdot SYY}}.$$
Prediction

Given a linear model
\[ Y = \beta_0 + \beta_1 X + \varepsilon \]
there are two quantities we might want to predict with uncertainty at a new covariate value \( X = x_* \):

- The mean value \( \beta_0 + \beta_1 x_* \) (fit)
- A single new observation \( y_* = \beta_0 + \beta_1 x_* + \varepsilon_* \) (prediction).

If we knew \( \beta_0 \) and \( \beta_1 \), our uncertainty for the fit would be zero and our uncertainty for the prediction would be \( \sigma^2 \).

[In the first case we want to quantify our uncertainty about the mean function, whereas in the second case we want to include our uncertainty about what a potential residual might be]

The optimal point predictor for both cases is
\[ \hat{\beta}_0 + \hat{\beta}_1 x_* \].

[Note predictor is linear in the observations!]

The uncertainty (standard error) depends on which case we’re looking at. The predictive standard error is then
\[
SE(\hat{\beta}_0 + \hat{\beta}_1 x_* + \varepsilon_*) = \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{SXX}} \]  (prediction)

[Note where the minimum occurs]

95% predictive confidence intervals are then
\[
\hat{\beta}_0 + \hat{\beta}_1 x_* \pm t_{n-2}(0.975) SE(\hat{\beta}_0 + \hat{\beta}_1 x_*) \]  (fit/confidence interval)
\[
\hat{\beta}_0 + \hat{\beta}_1 x_* \pm t_{n-2}(0.975) SE(\hat{\beta}_0 + \hat{\beta}_1 x_* + \varepsilon_*) \]  (prediction interval)
if A1 holds (we could swap in 1.96 for the $t$-quantile if approximate intervals suffice). [Note the standard error of prediction is greater than just $\hat{\sigma}$, since there is uncertainty in estimating $\beta_0$ and $\beta_1$].

[R example (SimpleLinearRegression.R)]

**Maximum Likelihood**

If A1 holds and $\varepsilon_1, \ldots, \varepsilon_n$ are iid $N(0, \sigma^2)$, then since

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

we have

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

so that it has pdf

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2} \left( \frac{y_i - (\beta_0 + \beta_1 x_i)}{\sigma} \right)^2}.$$  

The joint pdf of $y_1, \ldots, y_n$ is then

$$f(y_1, \ldots, y_n) = \prod_{i=1}^{n} f(y_i) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left( -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 \right).$$

by independence. Minimizing the negative log-likelihood $-\log f(y_1, \ldots, y_n)$ is equivalent to maximizing the likelihood, so the maximum likelihood estimators for $\beta_0$ and $\beta_1$ minimize

$$-\log f(y_1, \ldots, y_n) = C + \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2.$$  

In particular, under A1, the OLS estimators are also ML estimators, and enjoy additional theoretical properties. The MLE of $\sigma^2$ is different, and is biased, but $\hat{\sigma}^2_{MLE} \to \sigma^2$.

**2.3 Diagnostics**

Based on our estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, we have fitted values

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$
and estimated residuals

\[ \hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i), \quad i = 1, \ldots, n. \]

If our assumptions are correct, then the estimated residuals \( \{\hat{\varepsilon}_i\}_i \) should not have any structure.

One important diagnostic plot is the residuals vs. fitted values scatterplot, which can indicate a number of assumption violations:

1. If there are definite trends in the plot then the assumed linear relationship may be violated. [Picture]

2. If the variability changes with fitted values, then homoskedasticity may be violated; this latter case is when the errors exhibit heteroskedasticity. [Picture]

3. A few unusually large residuals may be evidence of outliers. Outliers won’t (necessarily) change \( \hat{\beta}_i \), but they do inflate \( \hat{\sigma}^2 \). [Picture]

The strongest set of assumptions, A1, suggest the residual terms \( \varepsilon_i \) are normally distributed. To assess normality of the estimated residuals \( \{\hat{\varepsilon}_i\}_i \), the quantile-quantile plot (Q-Q plot) is often used. The Q-Q plot plots the theoretical quantiles of a standard normal versus the estimated quantiles of the standardized observed residuals. If the plot falls along the identity line, it may be reasonable to assume the errors arose from a normal distribution. Q-Q plots can be used to assess heavy-tailedness and skewness compared to the normal distribution. [Picture of standard and heavy-tailed Q-Q plot]

Note that the OLS-estimated regression line is forced to go through the pair \((\bar{x}, \bar{y})\), that is,

\[ \hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{y}. \]

Thus, you can imagine if a particular covariate \( x_k \) were unusually far from \( \bar{x} \), it might exert a particularly large leverage on the estimates \( \hat{\beta}_i \), given this constraint. The leverage of the \( k \)th data point \((x_k, y_k)\) is defined by

\[ \frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}. \]
Note that leverage is a quantification of how far $x_k$ is from the average covariate. High leverage points can have a substantial effect on the ordinary least squares fit, but doesn’t necessarily inflate $\hat{\sigma}^2$.

[R example (Diagnostics.R)]

2.4 Pause: Crash Course in Matrix Algebra

A matrix $A$ with $n$ rows and $m$ columns (i.e., an $n \times m$ matrix) is an element of $\mathbb{R}^n \times \mathbb{R}^m$. For example, a $3 \times 2$ matrix has real-valued elements and

$$
A = \begin{pmatrix}
  a_{11} & a_{12} \\
  a_{21} & a_{22} \\
  a_{31} & a_{32}
\end{pmatrix}
$$

where, generally $a_{ij}$ will refer to the $(i, j)$th element, that is, the row $i$ column $j$ element. Sometimes we shorthand this as

$$
A = (a_{ij})_{i=1,j=1}^{3,2} = (a_{ij})
$$

A vector is just a matrix with one column, (known as a column vector), e.g.,

$$
x = \begin{pmatrix}
  x_1 \\
  x_2 \\
  \vdots \\
  x_n
\end{pmatrix}.
$$

The transpose of a matrix is $A^T = (a_{ji})$, switching column and row places. E.g.,

$$
A = \begin{pmatrix}
  2 & \pi \\
  0 & 10 \\
  -3 & 3
\end{pmatrix} \quad \text{has} \quad A^T = \begin{pmatrix}
  2 & 0 & -3 \\
  \pi & 10 & 3
\end{pmatrix}.
$$

Note that $(A^T)^T = A$. A matrix is square if its number of rows is the number of columns. A square matrix is symmetric if $A = A^T$, e.g.,

$$
A = \begin{pmatrix}
  1 & 3 \\
  3 & 10
\end{pmatrix}
$$

22
The elements $\text{diag}(A) = (a_{11}, a_{22}, \ldots, a_{nn})$ define the diagonal of a matrix. The $n$-dimensional identity matrix $I$ is the square $n \times n$ matrix with 1s along the diagonal

$$ I_n = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} $$

An upper triangular matrix is a matrix with zero entries below the main diagonal:

$$ \begin{pmatrix} 2 & 8 & 3 & 0 \\ 0 & 0 & 10 & 7 \\ 0 & 0 & 4 & 7 \\ 0 & 0 & 0 & 5 \end{pmatrix} $$

and a lower triangular matrix has zeros above the diagonal. If we wanted to sum all elements of

$$ A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix}, $$

we can write this as

$$ \sum_{i=1}^{3} \sum_{j=1}^{2} a_{ij} = a_{11} + a_{12} + a_{21} + a_{22} + a_{31} + a_{32} $$

Any two $n \times p$ matrices $A$ and $B$ may be added by taking elementwise sums,

$$ \begin{pmatrix} 2 & \pi \\ 0 & 10 \\ -3 & 3 \end{pmatrix} + \begin{pmatrix} 3 & 1 \\ 4 & 2 \\ 5 & 0 \end{pmatrix} = \begin{pmatrix} 5 & \pi + 1 \\ 4 & 12 \\ 2 & 3 \end{pmatrix}. $$

Multiplication by a constant $bA = b(a_{ij}) = (ba_{ij})$ multiplies elementwise. Some properties of matrix addition: (both $A$ and $B$ must be $n \times m$):

- $A + B = B + A$
- $(A + B)^T = A^T + B^T$
- $(A - B)^T = A^T - B^T$
- $(x + y)^T = x^T + y^T$
• $(x - y)^T = x^T - y^T$

Matrices $A(n \times m)$ and $B(m \times p)$ multiply to form a $n \times p$ matrix $C$ with

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj},$$

that is, take the sum of elementwise products of the $i$th row of $A$ with the $j$th column of $B$. E.g.,

$$
\begin{pmatrix}
2 & 0 \\
0 & 10 \\
-3 & 3
\end{pmatrix}
\begin{pmatrix}
3 & 1 \\
4 & 2 \\
3 & 3
\end{pmatrix}
= 
\begin{pmatrix}
6 & 2 \\
40 & 20 \\
3 & 3
\end{pmatrix}.
$$

Two matrices can multiply if the number of columns of the first equals the number of rows of the second matrix, e.g., $AB$ makes sense, but then $BA$ may not conform, and, even if they do,

$$AB \neq BA.$$

Some properties of matrix multiplication:

• $A(B + C) = AB + AC$
• $A(B - C) = AB - AC$
• $(A + B)C = AC + BC$
• $(A - B)C = AC - BC$
• $(A + B)(C + D) = AC + BC + AD + BD$
• $(AB)C = A(BC)$

Transposes work as

$$(AB)^T = B^T A^T.$$

Note, then, that if $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$, 

$$x^T y = \sum_{i=1}^{n} x_i y_i \in \mathbb{R}.$$
which is sometimes called the dot product of $x$ and $y$. Moreover the squared Euclidean norm of $x$ is defined as

$$\|x\|_2^2 \equiv x^T x = \sum_{i=1}^{n} x_i^2$$

where $\| \cdot \|_2^2$ denotes the squared $L_2$ length of the vector $x$. More properties:

- $\sum_{i=1}^{n} a_i^T x_i = a^T \sum_{i=1}^{n} x_i$
- $\sum_{i=1}^{n} A x_i = A \sum_{i=1}^{n} x_i$
- $\sum_{i=1}^{n} (A x_i)(A x_i)^T = A (\sum_{i=1}^{n} x_i x_i^T) A^T$

If $A$ is square $(n \times n)$ and $x$ and $y$ are vectors,

$$y^T A y = \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j a_{ij}$$

is called a *quadratic form*. Note it is a scalar.

The *inverse* of a square matrix $A$, denoted $A^{-1}$, is the matrix that satisfies

$$A A^{-1} = A^{-1} A = I,$$

if it exists. We have

$$(AB)^{-1} = B^{-1} A^{-1},$$

if all inverses exist.

A square $n \times n$ matrix $A$ is *nonnegative definite* if, for any vector $a$,

$$a^T A a \geq 0.$$

Positive definite matrices (replacing $\geq$ with $>$) *always admit an inverse*. Additionally, $A$ has a “square root” called the *Cholesky factor* or *Cholesky decomposition*

$$A = T^T T$$

where $T$ is an invertible upper triangular matrix. The Cholesky factor $T$ is unique, but there are many other “square root” matrices.

The *determinant* of a matrix is a number and is written $\det(A) = |A|$, and has the following properties:
• \( \det \mathbf{I} = 1 \)

• \( \det (\mathbf{A}^T) = \det \mathbf{A} \)

• \( \det \mathbf{A}^{-1} = 1/(\det \mathbf{A}) \)

• \( \det (\mathbf{AB}) = \det (\mathbf{A}) \det (\mathbf{B}) \) for square matrices

• \( \det (c\mathbf{A}) = c^n \det \mathbf{A} \) where \( \mathbf{A} \) is \( n \times n \).

• The determinant of a triangular or diagonal matrix is the product of its diagonal.

The trace of a matrix is the sum of its diagonals, \( \text{tr} \mathbf{A} = \sum_{i=1}^{n} a_{ii} \). We have

• \( \text{tr} (\mathbf{A} + \mathbf{B}) = \text{tr} \mathbf{A} + \text{tr} \mathbf{B} \)

• \( \text{tr} (\mathbf{AB}) = \text{tr} (\mathbf{BA}) \) (even if \( \mathbf{AB} \neq \mathbf{BA} \))

Two vectors \( \mathbf{a} \) and \( \mathbf{b} \) orthogonal if

\[
\mathbf{a}^T \mathbf{b} = 0.
\]

If \( \|\mathbf{a}\|_2 = 1 \) then \( \mathbf{a} \) is said to be normalized. Any vector can be normalized by

\[
\mathbf{b} = \frac{\mathbf{a}}{\|\mathbf{a}\|_2}.
\]

A square matrix \( \mathbf{C} = (c_1, c_2, \ldots, c_p) \) is said to be orthogonal if its columns and rows are normalized and mutually orthogonal, whence

\[
\mathbf{C}^T \mathbf{C} = \mathbf{C} \mathbf{C}^T = \mathbf{I}.
\]

Thus, \( \mathbf{C}^{-1} = \mathbf{C}^T \). Multiplication by an orthogonal matrix has the effect of rotating the axes, that is, if \( \mathbf{z} = \mathbf{C} \mathbf{x} \) then note

\[
\mathbf{z}^T \mathbf{z} = \mathbf{x}^T \mathbf{C}^T \mathbf{C} \mathbf{x} = \mathbf{x}^T \mathbf{x}
\]

maintains the same length.

An eigenvalue \( \lambda \) of a matrix \( \mathbf{A} \) is a number where

\[
\mathbf{Ax} = \lambda \mathbf{x}
\]
and \( \mathbf{x} \) is called the corresponding \emph{eigenvector}. They can be calculated by finding solutions to \( \det(\mathbf{A} - \lambda \mathbf{I}) = 0 \). Note that, if \( \mathbf{x} \) is an eigenvector, and \( k \) is a number, then

\[
\mathbf{A}(k\mathbf{x}) = k\lambda(\mathbf{x}) = \lambda(k\mathbf{x})
\]

so \( k\mathbf{x} \) is also an eigenvector, and thus we usually scale eigenvectors to be unit length.

If \( \lambda \) is an eigenvalue of \( \mathbf{A} \) and \( \mathbf{x} \) is the corresponding eigenvector, then \( 1\pm\lambda \) is an eigenvalue of \( \mathbf{I} \pm \mathbf{A} \), and \( \mathbf{x} \) is still the corresponding eigenvector.

If \( \mathbf{A} \) is a square matrix with eigenvalues \( \lambda_1,\ldots,\lambda_n \),

- \( \text{tr} \mathbf{A} = \sum_{i=1}^{n} \lambda_i \)
- \( \det \mathbf{A} = \prod_{i=1}^{n} \lambda_i \).

Importantly,

- If \( \mathbf{A} \) is positive definite, then its eigenvalues are \emph{all positive}
- If \( \mathbf{A} \) is nonnegative definite, then its eigenvalues are either positive or zero.
- If \( \mathbf{A} \) is symmetric, its eigenvectors are all mutually orthogonal.

By the last property, if \( \mathbf{A} \) has eigenvectors \( \mathbf{x}_1,\ldots,\mathbf{x}_n \), form

\[
\mathbf{C} = [\mathbf{x}_1 \mathbf{x}_2 \cdots \mathbf{x}_n]
\]

and note \( \mathbf{C} \) is orthogonal. Then

\[
\mathbf{A} = \mathbf{A} \mathbf{I} = \mathbf{A} \mathbf{C} \mathbf{C}^T \\
= [\mathbf{A} \mathbf{x}_1 \mathbf{A} \mathbf{x}_2 \cdots \mathbf{A} \mathbf{x}_n] \mathbf{C}^T \\
= [\lambda_1 \mathbf{x}_1 \lambda_2 \mathbf{x}_2 \cdots \lambda_n \mathbf{x}_n] \mathbf{C}^T \\
= \mathbf{C} \mathbf{D} \mathbf{C}^T
\]
where

\[
D = \begin{pmatrix}
\lambda_1 & 0 & \cdots & 0 \\
0 & \lambda_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \lambda_n
\end{pmatrix}
\]

is diagonal with eigenvalues on the diagonal. This is known as the spectral decomposition or eigendecomposition of \( A \). We can also diagonalize \( A \) by

\[
C^T AC = D.
\]

A square root matrix is \( A^{1/2} = CD^{1/2}C^T \), where \( D^{1/2} = \text{diag}(\lambda_1^{1/2}, \ldots, \lambda_n^{1/2}) \). The square of \( A \) can be calculated

\[
A^2 = CD^2C^T
\]

and the inverse

\[
A^{-1} = CD^{-1}C^T
\]

all of which only involve applying these functions to the eigenvalues in \( D \), which is one reason the eigendecomposition is nice.

The rank of a matrix is the number of linearly independent column vectors (or row vectors). If \( A \) is \( n \times p \), then

- \( \text{rank}(A) \leq \min(n, p) \)
- \( \text{rank}(A^T A) = \text{rank}(AA^T) = \text{rank}(A) \)
- If \( A \) is \( n \times n \) and has full rank \((n)\) then \( A \) is invertible.

We will mostly be only dealing with full rank matrices.

Let \( A \) be an \( n \times p \) matrix of rank \( k \). The singular value decomposition of \( A \) is

\[
A = UDV^T
\]

where \( U \) is \( n \times k \), \( D \) is \( k \times k \) and \( V \) is \( p \times k \). \( D = \text{diag}(\lambda_1, \ldots, \lambda_k) \) contains the positive square roots of the \( \lambda_1^2, \ldots, \lambda_k^2 \) nonzero eigenvalues of \( AA^T \) or \( A^T A \). The \( k \) columns of \( U \) are
normalized eigenvectors of \( \mathbf{A}\mathbf{A}^T \) corresponding to eigenvalues \( \lambda_1^2, \ldots, \lambda_k^2 \). The \( k \) columns of \( \mathbf{V} \) are the normalized eigenvectors of \( \mathbf{A}^T\mathbf{A} \) corresponding to eigenvalues \( \lambda_1^2, \ldots, \lambda_k^2 \). Then,

\[
\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}.
\]

If \( \mathbf{A} \) is positive definite, then the singular value decomposition is the same as the spectral decomposition.

### 2.5 Multivariate normal

Recall the definition of covariance and correlation: if \( X \) and \( Y \) are random variables with mean and standard deviation \( \mu_X, \sigma_X \) and \( \mu_Y, \sigma_Y \), respectively, then

\[
\text{Cov}(X, Y) := \mathbb{E}((X - \mu_X)(Y - \mu_Y))
\]

and

\[
\text{Cor}(X, Y) := \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \quad (= \rho).
\]

In particular, \( \text{Cor}(X, Y) \in [-1, 1] \) is unitless and is a measure of the linear dependence between \( X \) and \( Y \), and \( \text{Cov}(X, Y) \) is in the units of \( X \) times the units of \( Y \). Covariances enjoy the following properties:

- \( \text{Var}X = \text{Cov}(X, X) \)
- \( \text{Cov}(aX + b, Y) = a\text{Cov}(X, Y) \) for any \( a, b \in \mathbb{R} \) [prove]
- \( \text{Cov}(X, Y) = \text{Cov}(Y, X) \)
- \( \text{Cov}(X + Y, Z) = \text{Cov}(X, Z) + \text{Cov}(Y, Z) \).

How are covariances defined for vectors of random variables? Suppose \( \mathbf{X} = (X_1, \ldots, X_n)' \) and \( \mathbf{Y} = (Y_1, \ldots, Y_m)' \) are two random vectors, where ’ denotes the transpose. Define \( \text{Cov}(\mathbf{X}, \mathbf{Y}) \) to be the \( n \times m \) matrix with \((i, j)\)th entry \( \text{Cov}(X_i, Y_j) \), that is,

\[
\text{Cov}(\mathbf{X}, \mathbf{Y}) = \begin{pmatrix}
\text{Cov}(X_1, Y_1) & \text{Cov}(X_1, Y_2) & \cdots & \text{Cov}(X_1, Y_m) \\
\text{Cov}(X_2, Y_1) & \text{Cov}(X_2, Y_2) & \cdots & \text{Cov}(X_2, Y_m) \\
\vdots & \vdots & \ddots & \vdots \\
\text{Cov}(X_n, Y_1) & \text{Cov}(X_n, Y_2) & \cdots & \text{Cov}(X_n, Y_m)
\end{pmatrix}
\]

The following properties also hold for covariances of random vectors:
• \( \text{Var} \mathbf{X} = \text{Cov}(\mathbf{X}, \mathbf{X}) \) [This is how we define the variance of a random vector]

• \( \text{Cov}(A \mathbf{X} + \mathbf{\mu}, B \mathbf{Y} + \mathbf{\nu}) = A \text{Cov}(\mathbf{X}, \mathbf{Y}) B' \) for any \( k \times n \) matrix \( A \), \( j \times m \) matrix \( B \), \( \mathbf{\mu} \in \mathbb{R}^k \) and \( \mathbf{\nu} \in \mathbb{R}^j \)

• \( \text{Cov}(\mathbf{X}, \mathbf{Y}) = \text{Cov}(\mathbf{Y}, \mathbf{X})' \).

If \( \mathbf{X} \) is a random vector such that \( \text{Var} \mathbf{X}_i = \sigma_i^2 \), then the covariance matrix \( \text{Cov}(\mathbf{X}, \mathbf{X}) = (\text{Cov}(\mathbf{X}_i, \mathbf{X}_j))_{i,j=1}^n \) can be transformed into a correlation matrix

\[
\text{Cor}(\mathbf{X}, \mathbf{X}) = \left( \frac{\text{Cov}(\mathbf{X}_i, \mathbf{X}_j)}{\sigma_i \sigma_j} \right)_{i,j=1}^n.
\]

The correlation matrix is usually easier to interpret since it consists of all pairwise correlations between the component random variables. Note the diagonal of \( \text{Cor}(\mathbf{X}, \mathbf{X}) \) is 1s.

**Definition 2.** We say the random vector \((\mathbf{X}_1, \mathbf{X}_2)\) has a bivariate normal distribution if

\[
f(x_1, x_2) = \frac{1}{\sqrt{(2\pi)^2 \text{det} \begin{pmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & \sigma_2^2 \end{pmatrix}}} \exp \left( -\frac{1}{2} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}' \begin{pmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & \sigma_2^2 \end{pmatrix}^{-1} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix} \right)
\]

for \((x_1, x_2) \in \mathbb{R}^2 \).

Here, parameters have the following interpretations,

• \( \mathbb{E} \mathbf{X}_1 = \mu_1 \)

• \( \mathbb{E} \mathbf{X}_1 = \mu_2 \)

• \( \text{Var} \mathbf{X}_1 = \sigma_1^2 \)

• \( \text{Var} \mathbf{X}_2 = \sigma_2^2 \)

• \( \text{Cov}(\mathbf{X}_1, \mathbf{X}_2) = \rho \sigma_1 \sigma_2 \)

• \( \text{Cor}(\mathbf{X}_1, \mathbf{X}_2) = \rho \).

Note that the marginals are also normally distributed, e.g., \( \mathbf{X}_1 \sim N(\mu_1, \sigma_1^2) \). [Picture of bivariate density with projected contours]
Definition 3. We say the random vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)'$ has a multivariate normal distribution if

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} \exp \left( -\frac{1}{2} (\mathbf{x} - \mu)' \Sigma^{-1} (\mathbf{x} - \mu) \right)$$

for $\mathbf{x} \in \mathbb{R}^n$. We often write $\mathbf{X} \sim \text{MVN}({\mu}, {\Sigma})$, $\mathbf{X} \sim N_n({\mu}, {\Sigma})$ or $\mathbf{X} \sim N(\mu, \Sigma)$.

Here, $\mathbb{E}\mathbf{X} = \mu$ is the vector of means (i.e., $\mathbb{E} X_i = \mu_i$), and $\text{Cov}(\mathbf{X}, \mathbf{X}) = \Sigma$ is the covariance matrix (sometimes called variance-covariance matrix). The $(i, j)$th entry of $\Sigma$ is $\text{Cov}(X_i, X_j)$. [Imagine this as two parameters, $\mu$ tell us the average behavior of the vector, and $\Sigma$ tells us how all elements relate to each other]

The covariance matrix is symmetric since $\Sigma_{ij} = \text{Cov}(X_i, X_j) = \text{Cov}(X_j, X_i) = \Sigma_{ji}$. The diagonal of $\Sigma$ contains the variances of each component. Additionally, we have

- Each $X_i$ is marginally normally distributed: $X_i \sim N(\mu_i, \sigma_i^2)$
- If $A$ is a $k \times n$ real matrix and $\mathbf{b} \in \mathbb{R}^k$, then $A\mathbf{X} + \mathbf{b} \sim \text{MVN}(A\mu + \mathbf{b}, A\Sigma A')$
- If $\Sigma$ is a covariance matrix then it is nonnegative definite
- If $\Sigma$ is nonnegative definite, then it is a covariance matrix.

[R example (MultivariateNormal.R)]

### 2.6 Multiple Regression

[R example (IntroMultipleRegression.R)]

Sometimes it makes sense to regress $Y$ on multiple covariates $X_1, X_2, \ldots, X_p$. In this case the multiple linear regression model assumes

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon$$

with the only difference here being the introduction of extra covariates. If we have $n$ observations, $y_1, \ldots, y_n$, where the $i$th has $p$ corresponding covariates $x_{i1}, \ldots, x_{ip}$, then

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i$$
where we might make any of the analogous assumptions A1-A4 for \( \{ \varepsilon_i \} \). [Example: temperature on elevation and latitude, temperature changes with elev but also latitude, and neither can explain the other]

Note the interpretation of \( \beta_i \) is different here: for a unit increase in \( X_i \), \( \beta_i \) is the average increase in \( Y \) with all other covariates held fixed.

The model for \( y_1, \ldots, y_n \) can be written in matrix notation. Define

\[
\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \mathbf{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix},
\]

Note \( \mathbf{\beta} \) is \((p+1) \times 1\) and \( \mathbf{X} \) is \(n \times (p+1)\). The first column of 1s in \( \mathbf{X} \) is for the \( \beta_0 \) term. We can then write the model for all observations \( \mathbf{y} \) succinctly as

\[
\mathbf{y} = \mathbf{X}\mathbf{\beta} + \mathbf{\varepsilon}.
\]

Note under A1-A3,

\[
\mathbb{E}\mathbf{\varepsilon} = \mathbf{0} \quad \text{and} \quad \text{Var}\mathbf{\varepsilon} = \text{Cov}(\mathbf{\varepsilon}, \mathbf{\varepsilon}) = \sigma^2 \mathbf{I}_n.
\]

Sometimes it’s convenient to write

\[
\mathbf{x}_i = \begin{pmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix}, \quad \text{so that} \quad y_i = \mathbf{x}_i'\mathbf{\beta} + \varepsilon_i \quad \text{and} \quad \mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix}.
\]

**OLS**

The OLS estimator for \( \mathbf{\beta} \) is found by minimizing

\[
\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}))^2 = (\mathbf{y} - \mathbf{X}\mathbf{\beta})'(\mathbf{y} - \mathbf{X}\mathbf{\beta}),
\]

which results in

\[
\hat{\mathbf{\beta}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.
\]

[Check dimensions on board] Note this is *not* the best way to compute \( \hat{\mathbf{\beta}} \); in practice most programs use a QR decomposition that is more robust against rounding errors.
The OLS estimators are unbiased:

\[ \mathbb{E} \hat{\beta} = \mathbb{E}(X'X)^{-1}X'Y \]
\[ = (X'X)^{-1}X'EY \]
\[ = (X'X)^{-1}X'X\beta \quad \text{[why?] \[= \beta.} \]

Moreover,

\[ \text{Var(} \hat{\beta} \text{)} = \text{Var}(X'X)^{-1}X'Y \]
\[ = (X'X)^{-1}X'\text{Var}(Y, Y)X(X'X)^{-1} \]
\[ = \sigma^2(X'X)^{-1}X'X(X'X)^{-1} \quad \text{[why?] \[= \sigma^2(X'X)^{-1}.} \]

[note this is a variance-covariance matrix]

Standard errors for the OLS estimator \( \hat{\beta}_{OLS} \) are found by taking the square root of the diagonal of

\[ \text{Var}(\hat{\beta}_{OLS}) = \hat{\sigma}^2(X'X)^{-1} \]

where the residual variance is estimated by

\[ \hat{\sigma}^2 = \frac{(y - X\hat{\beta})'(y - X\hat{\beta})}{n - (p + 1)}. \]

This is an unbiased estimator for \( \sigma^2 \) (but is not the MLE). We can also write the residual sum of squares in a few formats:

\[ \text{RSS} = (y - X\hat{\beta})'(y - X\hat{\beta}) \]
\[ = y'y - \hat{\beta}'(X'X)\hat{\beta} \]
\[ = y'y - \hat{\beta}'X'y. \]

As previously, under A1,

\[ \frac{n - (p + 1)}{\hat{\sigma}^2} \hat{\sigma}^2 \sim \chi^2_{n-(p+1)}. \]
Hypothesis Testing and Confidence Intervals

The most basic hypothesis test we would want to check is

\[ H_0 : \beta_1 = \cdots = \beta_p = 0 \]
\[ H_A : \text{At least one } \beta_j \neq 0, \]

that is, are any of the covariates useful as linear predictors? The test statistic is the F-statistic,

\[ F = \frac{(TSS - RSS)/p}{RSS/(n - (p + 1))} \]

where

\[ SYY = TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2. \]

Under \( H_0 \),

\[ \mathbb{E}(RSS/(n - (p + 1))) = \sigma^2 \]

and

\[ \mathbb{E}((TSS - RSS)/p) = \sigma^2, \]

so we would expect \( F \approx 1 \). Thus, our test would reject \( H_0 \) if \( F \) were too large. In fact, under \( H_0 \),

\[ F \sim F_{p,n-(p+1)}, \]

so we would reject in favor of \( H_A \) if \( F > F_{p,n-(p+1)}(1 - \alpha) \) (this is a one-sided test). Equivalently, we can calculate the p-value of \( F \),

\[ 1 - \text{quantile}_{F_{p,n-(p+1)}}(F), \]

and just report this.

A test for an individual parameter \( \beta_j \)

\[ H_0 : \beta_j = 0 \]
\[ H_A : \beta_j \neq 0 \]

follows from the test statistic

\[ \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \sim t_{n-(p+1)} \]

(and from which individual confidence intervals can be derived).
Prediction

Suppose we want to predict $Y_\star$ for new set of covariates $x_\star = (1, x_{s1}, x_{s2}, \ldots, x_{sp})'$. There are two we might want to predict with uncertainty at $x_\star$:

- The mean value $x_\star' \beta$ (fit)
- A single new observation $y_\star = x_\star' \beta + \varepsilon_\star$ (prediction).

[In the first case we want to quantify our uncertainty about the mean function, whereas in the second case we want to include our uncertainty about what the residual will be]

The natural point predictor for both cases is

$$\hat{y}_\star = x_\star' \hat{\beta} = x_\star' (X' X)^{-1} X y.$$

[Note predictor is linear in the observations $y$!]

Our uncertainty depends on which case we’re looking at. Recall we define $SE(\cdot) = \sqrt{\text{Var}(\cdot)}$. Note

$$\text{Var}(x_\star^T \hat{\beta}) = x_\star^T \text{Var}(\hat{\beta}) x.$$

The standard error of prediction is then

$$\sqrt{\text{Var}(x_\star' \hat{\beta})} = \hat{\sigma}^2 x_\star' (X' X)^{-1} x_\star \quad \text{(fit/confidence)}$$

$$\sqrt{\text{Var}(x_\star' \hat{\beta} + \varepsilon_\star)} = \hat{\sigma}^2(1 + x_\star' (X' X)^{-1} x_\star) \quad \text{(prediction)}$$

and 95% predictive confidence intervals are then

$$\hat{y}_\star \pm t_{n-(p+1)}(\alpha/2)SE(x_\star' \hat{\beta}) \quad \text{(fit/confidence)}$$

$$\hat{y}_\star \pm t_{n-(p+1)}(\alpha/2)SE(x_\star' \hat{\beta} + \varepsilon_\star) \quad \text{(prediction)}$$

under A1. [Note the standard error of prediction is greater than just $\hat{\sigma}$, since there is uncertainty in estimating $\beta$ – follows since $X^T X$ is nonnegative definite]. If the residuals $\varepsilon$ are correlated, then we can improve the point and interval estimates by taking account of this extra structure, stay tuned.
Maximum Likelihood

Under A1, the OLS estimators $\hat{\beta}$ are the same as the MLEs, since

$$- \log f(y_1, \ldots, y_n) = - \log f(y) = C + \frac{1}{2\sigma^2} (y - X\beta)'(y - X\beta),$$

and minimizing the negative log-likelihood is equivalent to performing OLS.

Diagnostics

Similar diagnostics can be used for simple and multiple regression. For instance, $r^2$ still is a measure of fraction of variance explained, and values close to 1 indicate a good model fit. Beware, though, that $r^2$ will always increase as more covariates are added, whether they are useful or not. Root mean squared error (RMSE or RSE) is sometimes a useful summary for model comparisons

$$RMSE = \sqrt{\frac{RSS}{n - (p + 1)}} (= \hat{\sigma}),$$

and is in the units of the response $Y$. Note that the dependence on $p$ implies that RMSE can increase with the addition of covariates.

[R example (MultipleRegression.R)]

2.7 Model Selection (Variable Selection)

A fundamental problem in multiple regression is in choosing a set of relevant covariates. This problem is known as variable selection. To start, if we reject

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

then we know at least one covariate is useful, but this test doesn’t indicate which one.

Why Choose Variables?

Why not just throw all covariates into the dataset? If

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

but we have available covariates $X_1, \ldots, X_p, X_{p+1}, \ldots, X_{p+k}$ and use all of them, then
• At risk of overfitting (that is, “finding” trends that do not exist)
• $\hat{\beta}$ is still unbiased
• Var($\hat{\beta}$) is inflated.

What if we don’t use enough covariates? If $X = [X_1, X_2]$ contains the true set of covariates for a dataset, but we only use $X_1$, then

$$E\hat{\beta} = E(X'_1X_1)^{-1}X'_1Y$$

$$= (X'_1X_1)^{-1}X'_1EY$$

$$= (X'_1X_1)^{-1}X'_1(X_1\beta_1 + X_2\beta_2)$$

$$= \beta_1 + (X'_1X_1)^{-1}X'_1X_2\beta_2$$

$$\neq \beta_1.$$ 

So,

• $\hat{\beta}$ is biased.

• At risk of underfitting.

The Wrong Way

What about the proposed approach: for $i = 1, \ldots, p$, perform the hypothesis test

$$H_0 : \beta_i = 0$$

$$H_A : \beta_i > 0$$

and keep all $X_i$ where the test resulted in a rejection?

Pause: suppose we know our test statistic $T$ has a cdf $F$. If $t$ is the observed statistic, then the p-value is

$$P(T \geq t) = 1 - P(T < t)$$

$$= 1 - F(t)$$

$$\sim 1 - U(0, 1)$$

$$\sim U(0, 1)$$

37
Assessing Quality of Model Fit

Given two competing models (e.g., one with / one without a particular covariate), how do we decide which is better? That is, how do we quantify the goodness-of-fit? Recall that

\[ RSS = (y - X\hat{\beta})'(y - X\hat{\beta}) = \sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}))^2 \]

and

\[ r^2 = 1 - \frac{RSS}{SYY} \]

We could use the model that has a better (lower RSS or higher \( r^2 \)) value, except that these will always decrease / increase respectively with the addition of new covariates. Thus, we need an approach that rewards model fit while penalizing complexity.

If we entertain a model with \( d \) covariates \( X_1, \ldots, X_d \) fit to \( n \) observations, then we define Mallow’s \( C_p \) as

\[ C_p = \frac{1}{n}(RSS + 2\hat{\sigma}^2). \]

The second term typically increases as \( d \) increases, and we favor models with the smallest \( C_p \) value. [Note the book gives a different version of \( C_p \) which some people use in practice].

If a model is fit by maximum likelihood, then Akaike’s information criterion (AIC) or the Bayesian information criterion (BIC) may be used. [Recall that OLS is equivalent to ML if errors are normal]. Up to a constant, AIC is

\[ AIC = \frac{1}{n\sigma^2}(RSS + 2d\hat{\sigma}^2) \]

so, for least squares models, \( C_p \) and AIC are proportional (i.e., they agree on the best model). Up to a constant, BIC is

\[ BIC = \frac{1}{n}(RSS + \log(n)d\hat{\sigma}^2). \]

Note that since \( \log n > 2 \) for any \( n > 7 \), BIC tends to place more penalty on models with many variables \( d \). Under BIC, the best model is that with the smallest BIC value.
Finally, the \textit{adjusted} $r^2$ statistic is defined by

\[
\text{Adjusted } r^2 = 1 - \frac{RSS/(n - (d + 1))}{TSS/(n - 1)}
\]

The best model is that that \textit{maximizes} the adjusted $r^2$. Intuition: once all relevant variables are in the model, RSS is relatively stable, and increasing $d$ should be penalized (note that if RSS constant then the adjusted $r^2$ is decreasing in $d$).

\textbf{Variable Selection Methods}

Warning: Our model selection technique may depend on our objective. We may care about:

- \textit{Prediction accuracy}: in which case we don’t (really) care about which variables are used, so long as we can predict new $Y$’s with good skill.

- \textit{Model interpretability}: Including numerous variables with lots of interactions usually results in a model that is \textit{difficult to interpret}.

Ideally, we would entertain \textit{all} possible models. To do this we would fit the model

\[ Y = \beta_0 + \varepsilon \]

and the models

\[ Y = \beta_0 + \beta_1 X_i + \varepsilon \]

for $i = 1, \ldots, p$ and the models

\[ Y = \beta_0 + \beta_1 X_i + \beta_2 X_j + \varepsilon \]

for all $i \neq j = 1, \ldots, p$, etc., up to the full model

\[ Y = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \varepsilon. \]

For each of these competing models, we calculate any goodness-of-fit criterion (AIC, BIC, $C_p$, adjusted $r^2$), and choose the single model with the best value (e.g., lowest BIC). Although appealing, note that for $p$ covariates there are $2^p$ possible models, so this method is often computationally prohibited. This method is called \textit{best subset selection}.

The \textit{forward stepwise selection} algorithm works as follows:
• Fit the null model $M_0$, that is, the model with no covariates

• For $k = 1, \ldots, p$
  – Consider all models that add one covariate to $M_{k-1}$
  – Pick the “best” of these and call it $M_k$

• Pick the “best” model amongst $M_0, M_1, \ldots, M_p$.

Here, picking the “best” model amounts to choosing one of the criterions AIC, BIC, $C_p$, adjusted $r^2$ to quantify quality. Forward selection tends to computationally much easier than best subset selection, as it is a guided search over model space.

Alternatively, we could start with all covariates and then remove them in a stepwise fashion. This is known as backward elimination or backward stepwise selection. The algorithm is:

• Fit the full model $M_p$ with all covariates

• For $k = p, \ldots, 1$
  – Consider all models that delete one covariate from $M_k$
  – Pick the “best” model and call it $M_{k-1}$

• Pick the “best” model amongst $M_0, M_1, \ldots, M_p$.

Neither forward or backward selection is guaranteed to find the best model that best subset selection does, and indeed these techniques often disagree in practice. Also note that backward elimination will not work for high-dimensional problems, $p > n$. One could also consider a hybrid algorithm that adds and removes variables simultaneously, since sometimes adding a variable makes another’s coefficient insignificant.

[R example (VariableSelection.R)]

2.8 Potential Issues

In multiple regression there are numerous potential issues to consider.
Categorical (Qualitative) Predictors

So far we've dealt with quantitative predictors, e.g., \( X \in \mathbb{D} \subseteq \mathbb{R} \). Categorical or qualitative covariates only take on finitely many values (also sometimes called factors).

For example, if \( x_i \) is gender, then it has only two levels. We need to choose a convention to code it numerically using an indicator or dummy variable that takes on two possible values, e.g.,

\[
x_i = \begin{cases} 
1 & \text{if } i\text{th person is female} \\
0 & \text{if } i\text{th person is male} 
\end{cases}
\]

What happens to a simple model in this case?

\[
y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} 
\beta_0 + \beta_1 + \varepsilon_i & \text{if } i\text{th person is female} \\
\beta_0 + \varepsilon_i & \text{if } i\text{th person is male} 
\end{cases}
\]

Now \( \beta_0 \) is the average outcome for males, while \( \beta_0 + \beta_1 \) is the average outcome for females. Another interpretation is that \( \beta_1 \) is the average difference between females and males. The coefficient can be interpreted as a shift in the mean for \( x_i = 1 \). If we coded this differently, say

\[
x_i = \begin{cases} 
1 & \text{if } i\text{th person is female} \\
-1 & \text{if } i\text{th person is male} 
\end{cases}
\]

then

\[
y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} 
\beta_0 + \beta_1 + \varepsilon_i & \text{if } i\text{th person is female} \\
\beta_0 - \beta_1 + \varepsilon_i & \text{if } i\text{th person is male} 
\end{cases}
\]

Now \( \beta_0 \) is the overall average (ignoring gender), and \( \beta_1 \) is the amount females are above (and males are below) the average. Note that predictions will always be the same, regardless of the coding of \( x_i \), but the interpretation will change.

If there are more than two levels, then we use additional dummy variables. For example, if \( Y \) is life expectancy measured in each country of the world, and countries are grouped into \textit{Africa}, \textit{OECD} and \textit{Other} [OECD is the Organization for Economic Cooperation and Development, an international think tank charged with promoting policies that will improve global social and economic well-being] then to regress on group we might set

\[
x_{i1} = \begin{cases} 
1 & \text{if } i\text{th country is in OECD} \\
0 & \text{if } i\text{th country is not in OECD} 
\end{cases}
\]
\[ x_{i2} = \begin{cases} 
1 & \text{if } i\text{th country is Other} \\
0 & \text{if } i\text{th country is not Other} 
\end{cases} \]

Then

\[ y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i = \begin{cases} 
\beta_0 + \varepsilon_i & \text{if } i\text{th country is in Africa} \\
\beta_0 + \beta_1 + \varepsilon_i & \text{if } i\text{th country is in OECD} \\
\beta_0 + \beta_2 + \varepsilon_i & \text{if } i\text{th country is neither in Africa or OECD} 
\end{cases} \]

Thus, \( \beta_0 \) is the average life expectancy in Africa, \( \beta_1 \) is the additional expectancy for a person in an OECD country and \( \beta_2 \) is the additional expectancy for a person in a different (non-African or OECD) country. For instance, we found

\[ \hat{\beta}_0 = 59.8 \quad \hat{\beta}_1 = 22.7 \quad \hat{\beta}_2 = 15.6 \]

so that life expectancy in Africa is 59.8, life expectancy in OECD countries is 59.8 + 22.7 = 82.5 and everywhere else is 59.8 + 15.6 = 75.4 years. In this case, \( \beta_0 \) is known as the \textit{baseline}. Switching coding will keep predictions the same, but will change interpretations of coefficients.

Bad idea: set

\[ x_i = \begin{cases} 
0 & \text{if } i\text{th country is Africa} \\
1 & \text{if } i\text{th country is OECD (switch to Other)} \\
2 & \text{if } i\text{th country is Other (switch to OECD)} 
\end{cases} \]

[Brief mention of the trade-off between violating assumptions and the fragility of coefficients through this transformation]

**Beyond Additivity and Linearity**

Two restrictive assumptions we have made so far are \textit{additivity} and \textit{linearity}. The additive assumption implies the effect of one predictor on \( Y \) is independent of the values of the other predictors. The linearity assumption implies that the change in \( Y \) for a one-unit change in \( X \) does \textit{not} depend on the value of \( X \).

The usual generalization to remove the additive assumption is to consider \textit{interactions}. A simple model with an interaction is

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon. \]
For instance, if $Y$ is salary, $X_1$ is years since degree and $X_2$ is 1 for male (0 for female), then $\beta_1$ is the average raise per year for females, while $\beta_1 + \beta_3$ is the average raise per year for males, and $\beta_0$ is the average starting salary for females while $\beta_0 + \beta_2$ is the average starting salary for males. The hierarchical principle states that if an interaction is to be included, then the main effects of $X_1$ and $X_2$ alone should also be included, even if their coefficients are not significantly different than zero.

[R example (Salary.R)]

To overcome linearity, polynomial regression is usually the first step. For example, to generalize

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

we could consider

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

which fits a quadratic function to the data, but note the model is still a linear model in that it is linear in the coefficients. This is equivalent to using two predictors $X_1 = X$ and $X_2 = X^2$. We will spend a lot of time later on more methods for nonlinear modeling. Variable selection techniques can still be used for each

[R example (LifeExpectancy.R)]

Collinearity

Collinearity refers to when two (or more) covariates are closely related to one another. If two covariates are nearly linearly related, then it can be difficult to tease out the effect of one versus the other. Collinearity results in a less stable $(X'X)^{-1}$ and causes standard errors to grow. How to detect collinearity:

- Pairwise scatterplot of covariates
- Correlation matrix of covariates
It is possible for three or more covariates to be linearly related, in which case we can examine a variance inflation factor. In particular,

$$VIF(\hat{\beta}_j) = \frac{1}{1 - r^2_{\hat{X}_j|X_{-j}}}$$

where $r^2_{\hat{X}_j|X_{-j}}$ is the $r^2$ from regressing $X_j$ on the remaining predictors. Rule of thumb: VIFs greater than 5 – 10 are problematic.

To overcome collinearity:

- Use only one covariate from the set of collinear ones
- Orthogonalize:

  $$x^*_2 = x_2 - x_1 (x'_1 x_1)^{-1} x'_1 x_2$$

  where now

  $$x'_1 x^*_2 = 0.$$  

  (Problem: $x^*_2$ may be difficult to interpret)

- Combine via PCA or other methods

  [R example (MultipleRegression2.R)]

**Transformations**

Recall assumptions A1:

1. $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$
2. $\text{Var}\varepsilon_i = \sigma^2$ for all $i$.
3. $\varepsilon_i$ are normally distributed
4. $\varepsilon_i$ are independent (uncorrelated)
If these are not met, then we can either transform the response, predictors or both. Transformation of $Y$ or $X$ can help ensure 1, whereas only transformations of $Y$ can help ensure 2-3. Data transformation cannot help meeting assumption 4.

We’ve seen polynomial regression as a possible way to transform covariates, and will focus on response transformations. If $Y > 0$ is strictly positive then a log transformation $\log Y$ is common (and often helps ensure assumption 2). Other common possibilities are square-root or reciprocal ($\sqrt{Y}$ or $1/Y$). Note that if we use

$$\log Y = \beta_0 + \beta_1 X + \varepsilon$$

then

$$Y = e^{\beta_0}e^{\beta_1 X}e^\varepsilon,$$

or, in other words, the errors are *multiplicative* on the original scale.

Another common family of transformation is the *Box-Cox* family,

$$g_\lambda(Y) = \begin{cases} 
\frac{Y^{\lambda-1}}{\lambda} & \lambda \neq 0 \\
\log Y & \lambda = 0.
\end{cases}$$

where we would then model, e.g.,

$$g_\lambda(Y) = \beta_0 + \beta_1 X + \varepsilon.$$

The best value of $\lambda$ is usually chosen by maximizing the profile-likelihood.

[R example (Transformations.R)]

**Diagnostics Revisited**

A fundamental quantity in regression (and beyond) is the *hat matrix*,

$$H = X(X'X)^{-1}X'.$$

First note

$$H^2 = HH = H$$

so $H$ is idempotent. The *fitted values* are

$$\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'y = Hy,$$
so the hat matrix converts observations into fitted values. The estimated residuals are

\[ \hat{\epsilon}_i = y_i - \hat{y}_i \]

which we can write as

\[ \hat{\epsilon} = y - \hat{y} = y - Hy = (I - H)y. \]

Thus,

\[
\text{Var} \hat{\epsilon} = \text{Cov}(\hat{\epsilon}, \hat{\epsilon}) = (I - H)\text{Cov}(\epsilon, \epsilon)(I - H)' = \sigma^2(I - H).
\]

[We need to use $H^2 = H$ for this] Thus, the standard error for $\hat{\epsilon}_i$ is

\[ SE(\hat{\epsilon}_i) = \hat{\sigma} \sqrt{1 - H_{ii}} \]

where $H_{ii}$ is the $i$th diagonal entry of $H$. Under A1, the studentized residuals

\[ \hat{\epsilon}^*_i = \frac{\hat{\epsilon}_i - 0}{SE(\hat{\epsilon}_i)} \approx N(0, 1). \]

This gives us a method to look for outliers, where we might suspect a particular $y_i$ is an outlier if $\hat{\epsilon}^*_i$ is smaller than $-2$ or larger than $2$.

Now note

\[ \hat{y}_i = H_{i1}y_1 + \cdots + H_{ii}y_i + \cdots + H_{in}y_n, \]

so $H_{ii}$ is the influence of the observation $y_i$ on its own fitted value. We define $H_{ii}$ to be the leverage of $y_i$, a measure of its potential for being influential on the final fit, but note that $H_{ii}$ only depends on $X$, and not the actual value of $y_i$. To measure the the actual influence of a data point $y_i$, we use Cook’s distance (Cook’s D) as

\[ D_i = \frac{H_{ii}}{p(1 - H_{ii})} (\hat{\epsilon}^*_i)^2. \]

High values of leverage imply potential influence on the model fit while high values of Cook’s D imply actual influence on the model fit.

[R example (Diagnostics2.R)]
Degrees of Freedom

If we fit the model

\[ Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon \]

we say that we have used \((p + 1)\) degrees of freedom due to the estimation of the \(p + 1\) parameters. However, in the future it will be more difficult to define an analogous quantity. If \(H\) is the hat matrix, then note

\[
\sum_{i=1}^{n} H_{ii} = \text{tr}(H)
\]

\[
= \text{tr}(X(X'X)^{-1}X')
\]

\[
= \text{tr}(X'X(X'X)^{-1})
\]

\[
= \text{tr}(I_p) = p.
\]

Recalling that \(H_{ii}\) is the leverage of the \(i\)th data point, then \(p = \sum_{i=1}^{n} H_{ii}\) is the total influence of all observations, so the greater the number of parameters the greater the aggregate influence of the observations. We will see later that semiparametric regression methods also fall into a class of predictors like \(\hat{y} = My\), where we can define \(\text{tr}(M)\) as the effective degrees of freedom.