# MATH204 Notes

Winter ’23

Louis Meunier

notes.louismeunier.net

## Contents

# 1 Simple Linear Regression

## 1.1 Definitions

**Definition 1 (*Simple Linear Regression*)**

*Modeling of the relationship between two variables $X$ and $Y$, which assumes that $Y$ is a linear function of $X$. We can denote this:*

$$E(Y|X = x) = f(x)$$

$$Y = f(x) + \varepsilon, X = x$$

*where $\varepsilon$ is the error of the model.*

More specifically, we say that for each $i \in \{1, \ldots, n\}$, $y_i$ is the $i^{\text{th}}$ **response** (or, in other words, the response of the $i^{\text{th}}$ observation), and $x_i$ is the $i^{\text{th}}$ **independent variable**, or **covariate**. We assume that:

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

for each $i \in \{1, \ldots, n\}$. Here, $\beta_0$ and $\beta_1$ are **population parameters**, representing the $y$-intercept and slope of the model, respectively. In practice, we must estimate these variables, and they become the **estimators** $\hat{\beta}_0$ and $\hat{\beta}_1$; we thus say:

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

Similarly, $\hat{y}_i$ is the "estimation" of $y_i$, given the particular $\hat{\beta}_0$ and $\hat{\beta}_1$. The goal of regression is to find the appropriate estimators such that $\hat{y}_i \approx y_i, \forall\, i \in \{1, \ldots, n\}$. We achieve this using the **least squares criterion**, and we can rephrase our goal as finding the appropriate estimators that reduce the sum of the square of the difference between each $\hat{y}_i$ and its corresponding $y_i$. In other words, we aim to minimize:

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

This can be computed in a number of ways, but using calculus is fairly straightforward, by defining a function of $\hat{\beta}_0$ and $\hat{\beta}_1$, and minimizing it; briefly:

$$\begin{aligned}
S(\hat{\beta}_0, \hat{\beta}_1) &= \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i - y_i)^2 \\
S_{\hat{\beta}_0}(\hat{\beta}_0, \hat{\beta}_1) &= \sum [2(\hat{\beta}_0 + \hat{\beta}_1 x_i - y_i)] = 0 \\
n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i &= \sum y_i \\
S_{\hat{\beta}_1}(\hat{\beta}_0, \hat{\beta}_1) &= \sum [2(\hat{\beta}_0 + \hat{\beta}_1 x_i - y_i) x_i] = 0 \\
\hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2 &= \sum x_i y_i
\end{aligned}$$

From here, we have a system of two equations of two unknowns, which can be solved for in any number of

ways. In the end, we find that

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

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

In R, these values (among other stats) can be computed using **lm**(y~x).

## 1.2 Assumptions of the Model

- $x_i, \ldots, x_n$ are **explanatory** variables treated as independent **constants**, with negligible error in measurement.
- the error terms $\varepsilon_1, \ldots, \varepsilon_n$ are **mutually independent** random variables, with a mean $\bar{\varepsilon} = 0$ and variance of $\sigma^2$
- the terms $Y_1, \ldots, Y_n$ are **random** and **mutually independent** (as they are the sum of a random term and a constant term).

We can denote these assumptions as follows:

$$\begin{aligned} \mathrm{E}(Y_i | X_i = x_i) &= \beta_0 + \beta_1 x_i + \mathrm{E}(\varepsilon_i) \\ &= \beta_0 + \beta_1 x_i \\ \mathrm{Var}(Y_i | X_i = x_i) &= \mathrm{Var}(\beta_0 + \beta_1 x_i) + \mathrm{Var}(\varepsilon_i) \\ &= 0 + \sigma^2 \end{aligned}$$

This assumption is formally known as **homoscedasticity**.

## 1.3 Interpreting the Model

Assuming $\bar{\varepsilon} = 0$, then $\beta_1$ is the change in the mean of $Y_i$ for a one-unit change in $x_i$ (i.e., the slope). $\beta_1$ is the mean of $Y_i$ when $x_i = 0$ (i.e., the $y$-intercept).

## 1.4 Nature of Estimators: Making Inferences

When computing estimators, we are working with a sample of a larger population of data, and as a result, changing our sample can change our estimators; in this way, we can consider estimators as an "approximate representation of reality". In particular, we can use $\hat{\beta}_1$ to **infer** the value of $\beta_1$. However, this is only really helpful given a good value of $\hat{\beta}_1$.

**Definition 2 (*Unbiasedness*)**

> $\hat{\beta}_1$ *is unbiased for* $\beta_1$ *if it gives a good estimate over large number of samples, on* ***average****.*

We compute how accurate $\hat{\beta}_1$ using its standard deviation:

$$\sigma_{\hat{\beta}_1} = \sqrt{\text{Var}(\hat{\beta}_1)} = \frac{\sigma}{\sqrt{\text{S}_{xx}}}$$

However, $\sigma$ is a **population parameter**, which we typically do not know the true value of, so we must estimate it. First, take the formula for the estimation of the variance, $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y_i})^2 = \frac{\text{SSE}}{n-2} = \frac{\text{S}_{yy} - \hat{\beta}_1\text{S}_{xy}}{n-2}$$

Note that *SSE* represents the *sum of squares due to error* and $n-2$ represents the degrees of freedom. With this estimation of the variance, we can now rewrite our formula for the standard error of $\hat{\beta}_1$:

$$\hat{\sigma}_{\hat{\beta}_1} = \frac{\hat{\sigma}}{\sqrt{\text{S}_{xx}}} = \frac{\sqrt{\frac{\text{SSE}}{n-2}}}{\sqrt{\text{S}_{xx}}}$$

As before, this value can be computed with the **lm** function in R, and is the value under *"Std. Error"* in the output. The larger the $\sigma^2$ (or its estimate $\hat{\sigma}^2$):

- the more spread out the data is around the regression line
- the more uncertainty there is about the quality of the model
- the more uncertainty about the parameter estimates

## 1.5 Further Inferences about $\beta_1$

In order to make further inferences about the model, we must make the extra assumption that the error terms are a Normal random sample:

$$\varepsilon_1, \ldots, \varepsilon_n \sim \mathcal{N}(0, \sigma^2)$$

By extension, when the error terms are normally distributed, then so is $\hat{\beta}_1$:

$$\hat{\beta}_1 \sim \mathcal{N}(\beta_1, \frac{\sigma^2}{\text{S}_{xx}})$$

Note that this extra assumption is only needed for making inferences about $\beta_1$, while the previous assumptions are the minimum needed to estimate $\hat{\beta}_1$. If $\sigma^2$ were known, we could use this assumption to construct confident intervals and test hypotheses about $\beta_1$, namely:

$$T = \frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{\text{S}_{xx}}} \sim \mathcal{N}(0, 1)$$

In practice, however, as previously mentioned, $\sigma^2$ is not known, and we thus cannot use the Normal distribution. Instead, we can use the Student $t$ distribution (for reasons beyond the scope here):

$$T = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{\text{S}_{xx}}} \sim t_{n-2}$$

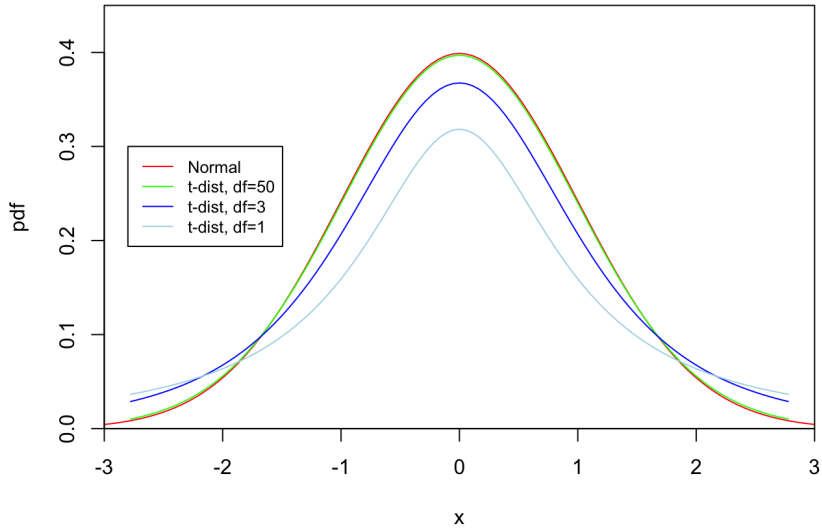


Figure 1: Normal Distribution compared to Student $t$ Distribution

When hypothesis testing, we typically set set $\mathcal{H}_0 : \beta_1 = 0$ and $\mathcal{H}_a : \beta_1 \neq 0$ to test whether a linear trend does truly exist. With this particular $\mathcal{H}_0$, we can then calculate the test statistic as follows:

$$T = \frac{\hat{\beta}_1 - \mathbf{0}}{\hat{\sigma}/\sqrt{S_{xx}}}$$

We can use the resulting value to construct a rejection region and proceed accordingly; if we reject $\mathcal{H}_0$, we are saying that the data does display a linear relationship. We can similarly use the distribution of $T$ to construct a $100 \times (1-\alpha)\%$ **confidence interval** for $\beta_1$, where $\alpha$ is our desired level of significance:

$$\text{C.I.} = \hat{\beta}_1 \pm t_{n-2,\alpha/2} \frac{\hat{\sigma}}{\sqrt{S_{xx}}}$$

Recall that $t$ is the $x$-value of the $t$-distribution chart where the area under the curve from $x$ to $\infty$ is $\alpha/2$. Note also that the $t$-value can be computed in R using the **qt**(1 – alpha/2, **df**) function, or you can compute the confidence interval directly by calling confint() on the **lm** object. Practically, we can interpret this interval by saying that our *true* $\beta_1$ is likely to be in this interval $100 \times (1-\alpha)\%$ of the time.

## 1.6 Analysis of Variance

**Definition 3 (*Analysis of Variance*)**

*Short-handed as "anova", this is a statistical method used to analyze the differences between groups, and specifically, compare the variance caused by error to the variance caused by estimation.*

## 1.7 Correlation

**Definition 4 (*Correlation*)**

*A measure of association between two random variables.*

Statistically, correlation is a measure of linear association that is symmetric, i.e., the correlation between $X$ and $Y$ is the same as that of $Y$ and $X$;

$$\text{corr}(X, Y) = \text{corr}(Y, X)$$

In linear regression swapping $X$ and $Y$ will not yield the same slope coefficients: the scales can differ and $\beta_1$ is the change of one variable in relation to the change of another.

The **correlation** between two random variables $X$ and $Y$ can be calculated:

$$r = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}$$

where $r$ is called the **correlation coefficient**, and $S_{XX}$, $S_{XY}$, and $S_{YY}$ are the same as previously defined (i.e., the sum of the product of the differences between values and the mean). We can say that $r$ has the following properties:

- it lies between $-1$ and $+1$;
- it is scaleless; a particular value of correlation is the same regardless of the scale of the variables. This is in contrast to $\hat{\beta}_1$, which is scale-dependent;
- it is an *estimator* for the **population correlation**, $\rho$. By extension, we can create a hypothesis test for lack of association between two random variables with $\mathcal{H}_0 : \rho = 0$. This test, under the assumption that the error terms are Normal is the *equivalent* of $\mathcal{H}_0 : \beta_1 = 0$:

  $$\mathcal{H}_0 : \rho = 0 \iff \mathcal{H}_0 : \beta_1 = 0$$

  The only difference in using one of these tests over the other is in the assumptions needed for the data; the first assumes that both $X$ and $Y$ are random, while the second assumes that $X$ is constant and $Y$ is random.

A $100 \times (1 - \alpha)\%$ confidence interval for $\rho$ using *Fisher's variance stabilizing $z$-transformation* can also be made. This involves first transforming $r$ to $z$:

$$z = \frac{1}{2} \ln \left( \frac{1 + r}{1 - r} \right) (= \text{arctanh}(r))$$

We can then build a C.I. for $z$:

$$(c_L, c_U) = z \pm \frac{z_{\alpha/2}}{\sqrt{n - 3}}$$

Inverting the $z$-transformation, we can then compute a $100 \times (1 - \alpha)\%$ C.I. for $\rho$:

$$\left[ \frac{e^{2c_L} - 1}{e^{2c_L} + 1}, \frac{e^{2c_U} - 1}{e^{2c_U} + 1} \right]$$

The reason for using this transformation is that, in data with $r$ near $-1$ or $+1$, distribution is highly skewed; this transformation yields a value with an approximately-Normal distribution.

In R, you can find the correlation coefficient using the **cor**() function on two vectors.

## 1.8 Determination

**Definition 5 (*Determination*)**

*A measure of the proportion of variance in $Y$ explained by the model (by $X$, that is).*

In simple linear regression, the **coefficient of determination** ($R^2$) is defined:

$$R^2 = 1 - \frac{\text{SSE}}{S_{YY}}$$

We can show that if:

- $X$ is *not* linearly associated with $Y$, then $\text{S}_{YY} = \text{SSE} \implies R^2 = 0$
- $X$ is linearly associated with $Y$, $\text{SSE} < S_{YY} \implies 0 < R^2 < 1$

Also note that (as the notation implies) the coefficient of determination is also the square of the correlation coefficient. Recall that $\text{SSE} = \text{S}_{YY} - \hat{\beta}_1 \text{S}_{XY}$, and $\hat{\beta}_1 = \frac{\text{S}_{XY}}{\text{S}_{XX}}$, therefore we can write

$$\text{SSE} = \text{S}_{YY} - \frac{\text{S}_{XY}}{\text{S}_{XX}}\text{S}_{XY} \implies \frac{\text{SSE}}{\text{S}_{YY}\text{S}_{XX}} = 1 - \frac{(\text{S}_{XY})^2}{\text{S}_{XX}}$$

And rewrite $R^2$ accordingly:

$$R^2 = 1 - \frac{\text{SSE}}{\text{S}_{YY}} = \frac{(\text{S}_{XY})^2}{\text{S}_{XX}\text{S}_{YY}} = \left(\frac{\text{S}_{XY}}{\sqrt{\text{S}_{XX}\text{S}_{YY}}}\right)^2 = r^2$$

The determination coefficient is denoted in the **summary**() function.

Note also that $R^2$ is a measure of the strength of a model, while $r$ is a measure of the strength of the association between $X$ and $Y$; knowing which to use depends on the context.

## 1.9 Estimation and Prediction

Once a regression line has been fitted from data, the model can be used to

- estimate the **mean response** $y_0$ for a particular $x_0$ value;
- predict a **new individual value** of the response for a particular $x_0$ value.

In both of these cases, the numerical values will be the same, while the the nature of uncertainty will differ.

### 1.9.1 Estimaing the Mean Response

For estimators $\hat{\beta}_0$, $\hat{\beta}_1$, and value $x_0$, we have

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

This estimate $\hat{y}_0$ is **unbiased**;

$$E(\hat{y}_0) = \beta_0 + \beta_1 x_0$$

It can be shown that its variance is

$$\text{var}(\hat{y}_0) = \sigma^2 \left\{ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}} \right\}$$

As before, the value of $\sigma^2$ is usually unknown and is estimated with $\hat{\sigma}^2$, resulting in an estimation of the variance of $\hat{y}_0$:

$$\widehat{\text{var}}(\hat{y}_0) = \hat{\sigma}^2 \left\{ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}} \right\}$$

From here, under the assumption of Normally distributed error terms, a $100 \times (1 - \alpha)\%$ confidence interval for the mean value $y_0$ is thus:

$$\hat{y}_0 \pm t_{n-2,\alpha/2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}}$$

### 1.9.2 Predicting a Value

We can (very similarly) use the model to predict some individual new value, $Y_0$, for $x_0$. The variability associated with this new value is estimated

$$\widehat{\text{var}}(Y_0) = \hat{\sigma}^2 \left\{ 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}} \right\}$$

Again, if the error terms are Normally distributed, we can find a $100 \times (1 - \alpha)\%$ confidence interval for an individual: $Y_0$

$$\hat{y}_0 \pm t_{n-2,\alpha/2} \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}}$$

The confidence interval when predicting a value for a particular $x_0$ is always wider than when just estimating the mean response for that $x_0$ (*for reasons outside the scope of this course...*).

**Caution:** it is important to note that using the least squares value for estimation/prediction for values of $x$ that fall outside of the range of values used to create the model may lead to errors, as there is no guarantee that a similar linear relationship will hold for values outside of the range of the data. To **extrapolate** from a regression model, we must make the assumption that nothing will change in the future with respect to the mean and standard deviation of the model, and that the linear relationship will persist.

## 1.10 Residual Analysis

Previously, appropriate assumptions for the error terms $(\varepsilon_1, \ldots, \varepsilon_n)$ were made; validations of these assumptions will be done using estimates, $(\hat{\varepsilon}_i, \ldots, \hat{\varepsilon}_n)$ called **residuals**. Recall that fitted values for our model are computed:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \, , i \in \{1, \ldots, n\}$$

This is the best estimate of the mean of $Y_i$; therefore, the best guess to the *unobservable* value of $\varepsilon_i$ is

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

This is the $i$th residual.

Because $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n$ are estimates of $\varepsilon_1, \ldots, \varepsilon_n$, these estimates can be used to check our assumptions that $\varepsilon_1, \ldots, \varepsilon_n$ are:

- normally distributed;
- have mean 0;
- have the same variance $\sigma^2$;
- independent;

While there are formulaic methods to check these assumptions, we will instead use graphical methods, which are far easier to interpret and use. One way to analyze residuals is by **standardizing** them; the $i$th standardized residual is equal to

$$\hat{\varepsilon}_i^{std} = \hat{\varepsilon}_i / \hat{\sigma}$$

where $\hat{\sigma}$ is the estimate of $\sigma$ from regression.

### 1.10.1 Normallity

The first thing we can check is the Normality of the residuals, which can be done using a histogram and a Q-Q plot, which can created with R, using a combination of **hist**, **qqnorm**, and **qqline**.

But how can we check *for sure?* We can use the standardized residuals to examine the "significance" of the magnitude of a residual, using the Empirical rule that 95% of observations are within 2 standard deviations of the mean, etc.. Since standardized residuals have a mean of 0 and standard deviation of 1, we can say that residuals that are larger than three (absolute value), then they are possible **regression outliers**.

What should we do in the case that these outliers do exist? Technically, the data is assumed to be a random sample from the population, and if the data point is not "special", it should *not* be excluded, as there is no way of knowing if it would be an outlier compared to the entire population. Naturally, there are some cases where it is logically clear that a particular point should be removed.

We can also redo our regression calculation after removing the outliers, and compare the results to the original regression. If the results are similar, then we can be confident that the outliers were not affecting the regression.

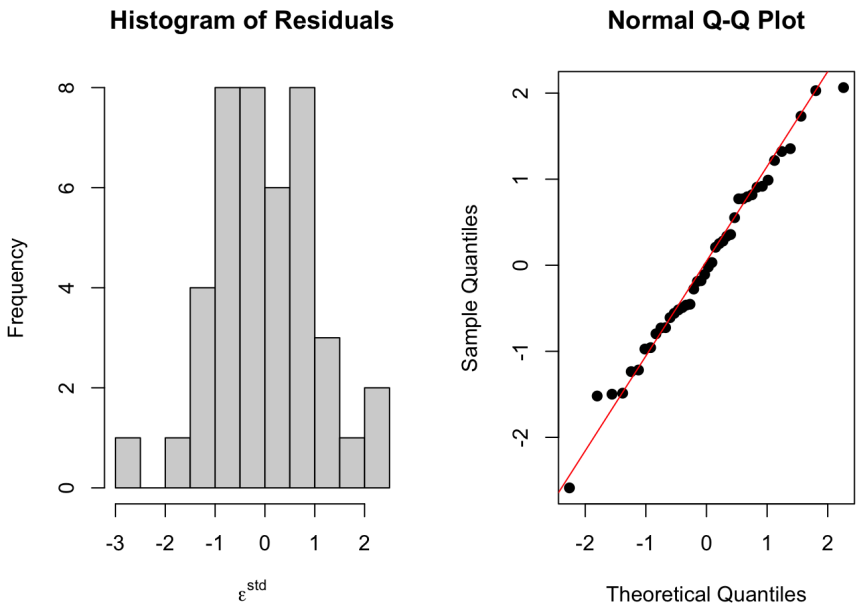


Figure 2: Example Histogram & Q-Q plot of residuals with corresponding line

### 1.10.2 Mean and Homoscedasticity

The next thing to check is the assumption that for all $i \in \{1, \ldots, n\}$,

$$\mathrm{E}(\varepsilon_i) = 0, \quad \mathrm{var}(\varepsilon_i) = \sigma^2$$

We can begin to check these assumptions by plotting the residuals against the fitted values. This way, we usually see one of two patterns:

1. If the residuals do *actually* have a mean of 0, then we should not see any residuals that vary as a function of the fitted mean
2. If the residuals all have the same variance, then we should see variability across all the fitted values

In general, if there is a pattern in the residuals, then the model is not appropriate for the data. In terms of simple linear regression, one can try to improve the model by adding more **polynomial terms** to it, such as in the model:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

We can also analyze whether the variance of the residuals stays constant in a number of ways; often, if the variance increases as a function of the residuals, despite a linear relationship, then the model is not appropriate for the data. It is also possible to see a more "football"-shaped pattern, indicating that the variance of the residuals is larger in the middle than at the ends. In either case, this violates the assumption of homoscedasticity; this, in turn, is called **heteroscedasticity**.

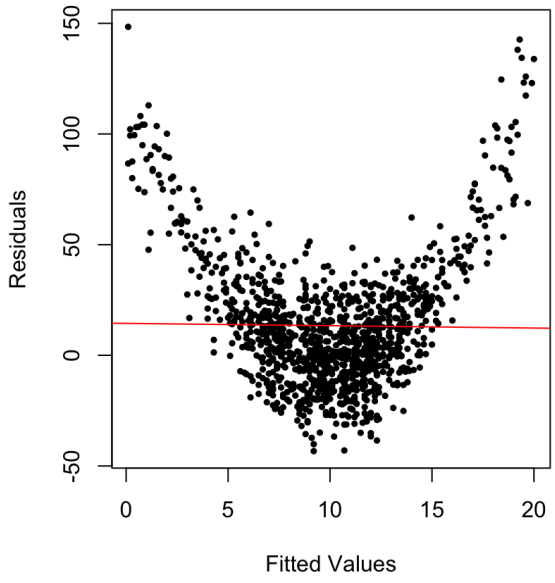


Figure 3: Example of residuals against fitted values; in this case, there is a pattern in the residuals, indicating that the model is not appropriate for the data. Intuitively, one can probably tell that a polynomial model would be more appropriate for this data.

## 1.11 Polynomial Regression

Often times, it is clear that a linear regression model will not be appropriate for some given data, i.e.,

$$\mathrm{E}(Y|X = x) \neq \beta_0 + \beta_1 x$$

In this case, it is possible to try and fit a **polynomial regression model** to the data. In this case, we might assume that $\forall\, i \in \{1, \ldots, n\}$,

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i,$$

or, equivalently,

$$\mathrm{E}(Y|X = x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

This is specifically a **quadratic model**, but more general, polynomial regression specifies that

$$\mathrm{E}(Y|X = x) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p$$

where the integer $p$ is the largest power (degree) of $x$ in the model.

It should be noted that *all intermediate powers need not be present*; in this case, some $\beta_i$ would equal 0 where $i < p$.

To find the appropriate polynomial regression models, we use the same least squares criterion as in simple

linear regression, i.e., we seek $\beta_0, \ldots, \beta_p$ such that

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_i x_i - \cdots - \beta_p x_i^p)^2$$

is minimized.

As before, the respective estimates $\hat{\beta}_0, \ldots, \hat{\beta}_p$ can be computed with R, by passing extra terms to the **lm** function: **lm**(y ~ x + **I**(x^2) + ... + **I**(x^p)). The same functions and tests can be used from the output of this function as in the case of simple linear regression.

When fitting a polynomial regression model;

- it is important to be sensitive to both overfitting local features of the data and also to interpretation;
- the more complex the model, the less plausibly we can justify trying to estimate with a relatively small amount of data (adding excessive polynomial terms to a model);

Generally, it is also preferable to *not* have a quadratic term without a linear model.

The null hypothesis test we use to determine the appropriate fit of a polynomial regression model is, appropriately,

$$\mathcal{H}_0 : \beta_p = 0,$$

where $p$ is the highest degree of the regression. If the hypothesis is not rejected for, say, $p = 2$, then we cannot say that a polynomial regression is any better than a linear regression.

# 2 Multiple Regression

**Multiple regression** is an extension of simple linear regression, in which a response variable $Y$ is modeled as a function of *several* covariates $X_1, \ldots, X_K$, rather than just one. Using multiple regression is helpful when want to assess the association between various covariates and the response variable, or, similarly, we want to assess how the addition of more than one covariate affects the quality of prediction of the model.

## 2.1 Comparison with Simple Linear Regression

Several of the issues from simple linear regression are still present in multiple regression;

- **Parameter estimation:** there are just more parameters; how do we estimate and interpret them?
- **Hypothesis testing:** how do the hypothesis tests changes, if at all?
- **Diagnosis of residuals:** what assumptions are made underlying the model?

New issues also arise in multiple regression:

- **Model selection:** which covariates do we include in the model?

- **Simultaneous hypotheses:** can we test hypotheses about multiple parameters at once?
- **Multi-colinearity:** what happens if the covariates are associated with each other?
- **Qualitative covariates:** can we use qualitative covariates in the model?
- **Interactions:** it may be possible that association of one covariate with response depends on the values of other covariates?

## 2.2 Model Description

In multiple regression, it is still assumed that the covariates have a linear association with the response variable. Assuming there are $K$ covariates of interest, $X_1, \ldots, X_l$, and a single response variable $Y$, the regression model assumes that for each $i \in \{1, \ldots, n\}$,

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{iK} + \varepsilon_i.$$

i.e., the response variable can be written as a *non-random* linear function of the covariates (plus, as always, some random error). This model has $n - (K + 1)$ degrees of freedom, as it has $K + 1$ parameters to estimate (the $K$ regression coefficients and the intercept) and $n$ observations.

The model is **linear in the parameters** but not necessarily in the covariates. To specify, all of the following are examples of linear models;

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, \\ Y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}^2 + \varepsilon_i, \\ \ln(Y_i) &= \beta_0 + \beta_1 e^{x_{i1}} + \beta_2 x_{i2}^2 + \beta_3 \sqrt{x_{i3}} + \varepsilon_i. \end{aligned}$$

**Note that in many of these cases, the covariates are somehow transformed to "create" a linear relation.** For instance, in the model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}^2 + \varepsilon_i,$$

the covariates are $X_1$ and $X_2^2$; while $Y_i$ is not linear with $X_2$, it *is* linear with $X_2^2$. Similar rationale applies to the other examples. In general, linear regression, whether multiple or simple, refers to linearity in the **coefficients** of the model, not necessarily in the covariates.

Non-linear models are also possible, where the coefficients appear, for instance, as powers of the covariates.

## 2.3 Assumptions of the Model

The multiple regression model states that for all $i \in \{1, \ldots, n\}$,

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{iK} + \varepsilon_i.$$

The assumptions of the model are that

1. $\mathrm{E}(\varepsilon_1) = \cdots = \mathrm{E}(\varepsilon_n) = 0$;
2. $\mathrm{var}(\varepsilon_1) = \cdots = \mathrm{var}(\varepsilon_n) = \sigma^2$;
3. $\varepsilon_1, \ldots, \varepsilon_n$ are mutually independent.

To make further inferences (confidence intervals, test), particularly when working with small sample sizes, we must additionally assume that

$$\varepsilon_1, \ldots, \varepsilon_n \sim \mathcal{N}(0, \sigma^2)$$

## 2.4 Interpretation

Interpretation of the coefficients of a multiple regression model is more complicated than in simple linear regression. $\beta_j$ represents the increase in the mean of $Y_i$ observed for a single unit increase in $x_{ij}$, **holding all other variables** $x_{i1}, \ldots, x_{iK}$ **constant**. $\beta_j$ is now the slope of the regression **plane** in the direction of $x_j$ (a plane with $K + 1$ dimensions; $K$ covariates and the intercept).

$\beta_j$ is not *just* the association of $X_j$ with $Y$, but is actually the association of $X_j$ with $Y$ while accounting for the associations of all the other covariates with $Y$.

## 2.5 Parameter Estimation

The estimate $\hat{\beta}$ of $\beta$ minimizes the sum of the squared distances between the predicted values of the value, namely,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{iK};$$

and the actual response values $y_i$ for $i \in \{1, \ldots, n\}$. i.e., $\hat{\beta}$ minimizes

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

Just as before, the **lm** (...) function can be used in R to estimate the parameters, but simply adding more covariates to the model.

## 2.6 Matrix Formulation of the Model (extra)

We can derive a matrix formulation of the multiple linear regression model as follows. Take

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

in which

- $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ is an $n \times 1$ column vector;
- $\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1K} \\ 1 & x_{21} & \ldots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{nK} \end{bmatrix}$ is an $n \times (K+1)$ matrix (this is the **design matrix**);
- $\beta = (\beta_0, \ldots, \beta_K)^T$ is a $p \times 1$ vector where $p = K + 1$;
- $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is an $n \times 1$ vector.

All together, this can be written;

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1K} \\ 1 & x_{21} & \ldots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{nK} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_K \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Note that this framework also fits to simple linear regression, by taking $p = 1$.

Looking at the design matrix part of this framework, each $x_{ij}$ corresponds to a particular $\beta_j$, with the first column of the design matrix corresponding to the intercept $\beta_0$ (all 1's).

As before, the goal is to find the estimate $\hat{\beta}$ to minimize the sum of the squared distances between the predicted values. In this new matrix view, it can be shown that

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = (\hat{\beta}_0, \ldots, \hat{\beta}_K)^T.$$

## 2.7 Inference

(Recall the following, in the case of simple linear regression). $\hat{\beta}$ is unbiased for $\beta$ (i.e. $\mathrm{E}(\hat{\beta}) = \beta$), and thus for each ordinary least squares coefficient, $\mathrm{E}(\beta)_j = \beta_j$. To get an (unbiased) estimate of the variance of $\beta$, $\sigma^2$, we must divide SSE by $n-(K+1)$, where $K+1$ is the number of coefficients estimated in the mode, i.e. $\beta_0, \ldots, \beta_K$. Thus, we have

$$\hat{\sigma}^2 = \frac{1}{n-(K+1)} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \frac{\mathrm{SSE}}{n-(K+1)}.$$

Note, too, that $n - (K+1) = n - 2$

Using these properties, we can create hypothesis tests and confidence intervals for each respective regression coefficient. Typically, our tests are of the form

$$\mathcal{H}_0 : \beta_j = 0, \quad \mathcal{H}_a : \beta_j \neq 0.$$

In this case, $\mathcal{H}_0$ represents no association between $Y$ and $X_j$, after adjusting for all other covariates in the model. **Note:** failing to reject $\mathcal{H}_0$ does not mean that $X_j$ is **NOT** associated with $Y$; it simply means there is no association *after adjustment.*

### 2.7.1 Confidence Intervals

To form a $100 \times (1-\alpha)\%$ confidence interval for $\beta_j$, we use

$$\hat{\beta}_j \pm t_{n-(K+1),\alpha/2} \times \hat{\sigma}_{\hat{\beta}_j}.$$

As before, this interval can be interpreted as containing $\beta_j$ $100 \times (1-\alpha)\%$ of the time.

Note that there is an inherent problem with multiple testing; we can only interpret the Type I error for each coefficient individually

## 2.8 Measuring the Fit of the Model

The *multiple coefficient of determination*, $R^2$, is defined

$$R^2 = 1 - \frac{\text{SSE}}{\text{S}_{YY}} \in [0,1]$$

As before, the $R^2$ value can be interpreted as the proportion of the variance in $Y$ that is explained by the model. *However*, unlike before, $R^2$ can no longer be interpreted as the squared correlation between $Y$ and a covariate, because there are several covariates. $R^2$ is, in some sense, a summary of how strongly the response variable is linearly associated to the linear combination of the covariates.

The $R^2$ value, however cannot be used to determine the best combination of covariates, since the $R^2$ value for a model with one extra variable will always be at least as large as the $R^2$ value for the "original" model. Instead, $R^2$ should be *adjusted* to take into account this fact.

The **adjusted coefficient of determination**, or, commonly, adjusted $R^2$ is defined

$$\begin{aligned} R_a^2 &= 1 - \frac{n-1}{n-(K+1)}\left(\frac{\text{SSE}}{\text{S}_{YY}}\right) = 1 - \frac{n-1}{n-K-1}(1-R^2) \\ &= \frac{n-1}{n-K-1}R^2 - \frac{K}{n-K-1}. \end{aligned}$$

Note that $R_a^2 < R^2$, and $R_a^2$ cannot be 'forced' to be 1 by adding more variables. As always, $R_a^2$ and $R^2$ are *sample statistics*.

## 2.9 Testing the Global Fit

To test if the fitted model is actual doing anything to help with the prediction, we can use an overall hypothesis test for the regression model. The null hypothesis;

$$\mathcal{H}_0 : \beta_1 = \cdots = \beta_K = 0,$$

and the alternative hypothesis;

$$\mathcal{H}_a : \text{at least one of } \beta_j \text{ is not equal to zero } (\exists\, \beta_j \neq 0).$$

We can test this hypothesis using the **F-test**;

$$F = \frac{(\mathrm{S}_{YY} - \mathrm{SSE})/K}{\mathrm{SSE}/\left[n-(K+1)\right]} = \frac{R^2/K}{(1-R^2)/\left[n-(K+1)\right]} = \frac{\mathrm{MSR}}{\mathrm{MSE}},$$

where MSR is the mean square for the regression model. Under the above null hypothesis, the $F$-statistic will have $F$ distribution with $K$ and $n-(K+1)$ degrees of freedom. $\mathcal{H}_0$ is rejected for large values of $F$; for $F > F_{\alpha,K,n-(K+1)}$.

**Note:** rejecting $\mathcal{H}_0$ does not say which of the slope coefficients is unlikely to be equal to 0, only that there is evidence that they are *all* 0. It is also possible to reject the above null hypothesis without rejecting a single null hypothesis of the form $\mathcal{H}_0 : \beta_j = 0$; in this case, the only conclusion we can draw is that the model is doing "something". However, if $\mathcal{H}_0$ isn't rejected, then none of the covariates will yield significant Student $t$-tests.

## 2.10 Prediction

Given a set of values for the covariates $x_0 = (x_{10}, \ldots, x_{K0})$, the value of $Y$ predicted for the model for this set of covariate values is

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{10} + \cdots + \hat{\beta}_K x_{K0}.$$

This formula could be used to estimate

- the mean of observations at a given set of covariate values;
- a new individual observation at a given set of covariate values.

Calculating prediction intervals for multiple regression is quite straightforward using the **predict** (...) function.

As in simple linear regression, one must be careful about making predictions outside the range of values in the dataset. In the case of multiple regression, this must be true for *each* covariate.

## 2.11 Interactions

In multiple regression, two variables, $X_1$ and $X_2$, are said to **interact** when the relationship between $X_1$ and the response variable $Y$ depends on the value of the second variable $X_2$, and vice versa. We can allow for interactions in the model by using the **product of the two covariates** as an additional linear predictor. For instance, if $X_1$ and $X_2$ are thought to interact, the model can be written as

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \boldsymbol{\beta_3 x_{i1} x_{i2}} + \varepsilon_i.$$

In this case, $\beta_3$ is the coefficient for the interaction between $X_1$ and $X_2$.

Doing this serves the intended purpose in the model because it can be rewritten the following two ways,

$$Y_i = \beta_0 + (\beta_1 + \beta_3 x_{i2})x_{i1} + \beta_2 x_{i2} + \varepsilon_i,$$

and

$$Y_i = \beta_0 + (\beta_2 + \beta_3 x_{i1})x_{i2} + \varepsilon_i.$$

Thus, the slope coefficient for $X_1$ depends on the values of $\beta_1$, $\beta_3$, and $X_2$, and similarly, the slope coefficient for $X_2$ depends on the values of $\beta_2$, $\beta_3$, and $X_1$.

Note the following:

1. If an interaction term is included in the model, then we want to test for the presence of that interaction first.
2. Once an interaction is shown to exist, we *cannot* interpret the main effects anymore. An overall association between an independent variable and the response cannot be discussed because the association always depends on the level of the other variable.
3. Only if we *fail* to reject the null hypothesis of no interaction will we want to try to interpret the main effects.

To clarify, consider the model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \varepsilon_i.$$

If it is concluded that $\beta_3 \neq 0$, then $\beta_1$ cannot be interpreted as the change in the mean of $Y$ for a one-unit change in $X_1$; we can only interpret the association of $X_1$ with the response $Y$ depending on the value of $X_2$ (and vice versa).

### 2.11.1 Testing for Interaction

For a model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \varepsilon_i,$$

the null hypothesis of no interaction is simply

$$\mathcal{H}_0 : \beta_3 = 0.$$

$\beta_3$ can be thought of as a coefficient of a "new" covariate ($X_3 = X_1 \times X_2$), and the same $t$-test can be used.

The general procedure is as follows;

1. Fit the model including the two covariates and the interaction;

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \varepsilon_i.$$

2. Conduct a global $F$-test to determine if any of the regression coefficients are different from zero;

$$\mathcal{H}_0 : \beta_1 = \beta_2 = \beta_3 = 0.$$

3. If this hypothesis is rejected, then test for *interaction* by using a Student $t$-test;

$$\mathcal{H}_0 : \beta_3 = 0.$$

If this hypothesis is rejected, then stop. Otherwise, re-fit the model *without* the interaction term to estimate $\beta_1$ and $\beta_2$ normally.

**Note that** if $\mathcal{H}_0 : \beta_3 = 0$ is rejected, we still need to keep the main effect terms in the model, although they are not interpreted the same way. Although we can't interpret the main effect terms in the present of an interaction, we *cannot* drop a main effect covariate from the model due to an insignificant $p$-value. This is because in a model with an interaction, all three of the model terms ($\beta_1 X_1$, $\beta_2 X_2$, $\beta_3 X_1 X_2$) work together to describe the interaction, and we therefore cannot take any of them out, or we would no longer be modeling their interaction.

In general, interactions should be used with *moderation*, and only when they model real relationships between the variables at hand. Otherwise, they can lead to overfitting the data.

## 2.12 Qualitative Data

### 2.12.1 Two Groups

Given "qualitative" data in a model (i.e. yes/no, true/false, etc.), we can assign a value to each category, i.e.

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

The group coded $Z = 0$ is called the **reference** group for the analysis. Using this model,

- the group coded $Z = 1$ has a mean level of $\beta_0 + \beta_1$;
- the group coded $Z = 0$ has a mean level of $\beta_0$.

By extension, $\beta_1$ is the difference in the mean of the response between the two groups. Fitting this model, we find that, more simply,

$$\hat{\beta}_0 = \bar{Y}_0, \text{ and } \hat{\beta}_1 = \bar{Y}_1 - \bar{Y}_0.$$

To test the significance of this model, we use the hypothesis

$$\mathcal{H}_0 : \beta_1 = 0 \iff \mu_1 = \mu_0; \quad \mathcal{H}_a : \beta_1 \neq 0 \iff \mu_1 \neq \mu_0,$$

where $\mu_i$ are the true population means for $Z = 1$ and $Z = 0$, respectively. Thus, you can test this either using a $t$-test on the difference of the means or a $t$-test on the coefficient $\beta_1$ (using the fitted model).

### 2.12.2 More than Two Groups

When there are more than two groups that need to be coded, we *cannot* use the same idea as above, where each of the $n$ groups gets a numerical code $1, \ldots, n$, since this would not result in the intended difference in

the values of the corresponding coefficients in the model. Instead, we have to use multiple variables for each "subgrouping" of groups, called **indicator** or **dummy** variables. Say there are three groups; we can code them

$$Z_1 = \begin{cases} 1 & \text{if in group 2} \\ 0 & \text{otherwise} \end{cases}; \quad Z_2 = \begin{cases} 1 & \text{if in group 3} \\ 0 & \text{otherwise} \end{cases},$$

with the corresponding regression model

$$Y_i = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \varepsilon_i.$$

Here, the expected values of the response for each group would be;

- Group 1 ($Z_1 = 0$, $Z_2 = 0$): $\mu_1 = \beta_0$
- Group 2 ($Z_1 = 1$, $Z_2 = 0$): $\mu_2 = \beta_0 + \beta_1$
- Group 3 ($Z_1 = 0$, $Z_2 = 1$): $\mu_3 = \beta_0 + \beta_2$

Interpretation of coefficients in this case becomes rather complicated; for three variables,

- $\beta_0$ is the mean of the first group;
- $\beta_1$ is the difference in mean between the first and second groups;
- $\beta_2$ is the difference in mean between the first and third groups;

and the difference in mean between the second and third groups is

$$\mu_3 - \mu_2 = (\beta_0 + \beta_2) - (\beta_0 + \beta_1) = \beta_2 - \beta_1.$$

As a result, various hypotheses can be used to test different aspects of the model, for instance

- $\beta_1 = 0$: Groups 1, 2 have the same mean
- $\beta_2 = 0$: Groups 1, 3 have the same mean
- $\beta_1 = \beta_2 = 0$: Groups 1, 2, 3 have the same mean
- $\beta_1 = \beta_2$: Groups 2, 3 have the same mean

The overall $F$-test corresponds to

$$\mathcal{H}_0 : \beta_1 = \beta_2 = 0 \Leftrightarrow \mathcal{H}_0 : \mu_1 = \mu_2 = \mu_3.$$

The other various tests can be done using the $t$-test on the corresponding coefficient.

### 2.12.3 Mixing Qualitative and Quantitative Variables

Take a model that has a qualitative variable ($Z$, equal to 1 if in group A, and 0 if in group B) and a quantitative variable ($X$),

$$Y_i = \beta_0 + \beta_1 z_i + \beta_2 x_i + \varepsilon_i.$$

When $Z = 0$ (group B), then this model becomes

$$Y_i = \beta_0 + \beta_2 x_i + \varepsilon_i,$$

and if $Z = 1$ (group A), the model becomes

$$Y_i = (\beta_0 + \beta_1) + \beta_2 x_i + \varepsilon_i.$$

Notice that the slope of both of these new lines is $\beta_2$, which represents the mean response of $Y$ for a one unit change in $X$ *regardless* of $Z$. The $y$-intercepts are different, however.

As before, changes in $Y$ may be different depending on the value of $Z$, so we should take into account the potential interaction between the qualitative and quantitative variables, i.e.

$$Y_i = \beta_0 + \beta_1 z_i + \beta_2 x_i + \beta_3 z_i x_i + \varepsilon_i,$$

where if there is an interaction, $\beta_3 \neq 0$. Using this model, if $Z = 0$, then $Y_i = \beta_0 + \beta_2 x_i + \varepsilon$, and if $Z = 1$, then $Y_i = (\beta_0 + \beta_1) + (\beta_2 + \beta_3)x_i + \varepsilon$. These now have *different* slopes; the change in $Y$ for a change in $X$ is $\beta_2$ when $Z = 0$, and $(\beta_2 + \beta_3)$ for $Z = 1$.

In general, the interaction term should always be included (and tested for), and then removed and re-fitted if it is not significant.

## 2.13 Comparing Nested Models

Some model $M_0$ is said to be **nested** in model $M_1$ if $M_0$ contains a subset of the covariates included in $M_1$. For instance, take

$$M_1 : Y_i = \beta_0 + \beta_1 z_i + \beta_2 x_i + \beta_3 z_i x_i + \varepsilon_i,$$

and

$$M_0 : Y_i = \beta_0 + \beta_2 x_i + \varepsilon_i.$$

Note that setting $\beta_1 = \beta_3 = 0$ in $M_1$ yields $M_0$; as such, we can use this idea to test the hypothesis

$$\mathcal{H}_0 : \beta_1 = \beta_3 = 0$$

to test if these extra covariates yields any statistically significant improvement in the model.

Generally, assume that these models (the *null* and *alternative* models, respectively) can be written as

$$M_0 : Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_g x_g + \varepsilon,$$
$$M_1 : Y = \beta_0 + \beta_1 x_1 + \ldots \beta_g x_g + \beta_{g+1} x_{g+1} + \cdots + \beta_k x_k + \varepsilon.$$

Then, the null hypothesis states that $M_1$ does not improve significantly compared to $M_0$;

$$\mathcal{H}_0 : \beta_{g+1} = \beta_{g+2} = \cdots = \beta_k = 0.$$

This is a **simultaneous** test of the parameters of the larger model; $\mathcal{H}_a$ is that at least one of these parameters is non-zero.

Note too that, since $M_1$ always contains more covariates than $M_0$, it is always true that $\text{SSE}_{M_0} \geq \text{SSE}_{M_1}$. We can say, loosely,

- if $\text{SSE}_{M_0} - \text{SSE}_{M_1}$ is "large", then $M_1$ explains more of the variance than $M_0$;
- if $\text{SSE}_{M_0} - \text{SSE}_{M_1}$ is "small", then the additional terms in $M_1$ do not contribute much to the model.

To put actual numbers to "large"/"small", we use

$$F = \frac{(\text{SSE}_{M_0} - \text{SSE}_{M_1})/(k-g)}{\text{SSE}_{M_1}/\{n-(k+1)\}}.$$

If $\mathcal{H}_0$ is true, and the usual model assumptions hold, then $F$ has a distribution

$$\mathcal{F}(k-g, n-(k+1)).$$

If the observed $F$ is large, we could reject $\mathcal{H}_0$ and conclude that at least one of the extra model terms is not equal to zero.

## 2.14 Other Issues with Multiple Regression

1. **Prediction beyond observed range of covariates:** If we have no data in the region where we are trying to predict, its unsure whether the regression model fits the data in this range.
2. **Multicollinearity:** The covariates must be sufficiently different from each other to be able to actually estimate their associations with the response accurately; if two covariates are highly correlated, for instance, it can be difficult to determine the association of each individual covariate. In this situation, the regression model is said to be subject to **multicollinearity**.
3. **Model errors correlation:** The model errors may be correlated, even though they were assumed to be independent at the start. This problem tends to happen with data measured over time.

# 3 Categorical Data

## 3.1 Multinomial Distribution

Take a random qualitative, random variable $C$ with $k$ possible values, $\{c_1, \ldots, c_k\}$ (called classes, categories, etc.). Suppose

$$\Pr(C = c_1) = p_1, \ldots, \Pr(C = c_k) = p_k.$$

Note that we must have that $p_1 + \cdots + p_k = 1$, where each $p_i$ represents the probability of observing a particular $c_i$.

Say we take $n$ independent trials, measuring $C$ each time, taking $X_i$ to represent the number of times that a particular $c_i$ is observed. The set of these random variables $X_1, \ldots, X_k$ is said to have a **multinomial distribution**.

$X_1, \ldots, X_k$ are integer-values, and $X_1 + \cdots + X_k = n$, and it can be shown that

$$\Pr(X_i = n_1, \ldots, X_k = n_k) = \frac{n!}{n_1! \cdots n_k!} p_1^{n_1} \cdots p_k^{n_k},$$

where $n_1 + \cdots + n_k = n$. These values, $n_1, \ldots, n_k$, are sometimes called **cell counts**.

## 3.2 Chi-Square Test

Take $X_1, \ldots, X_k$ to be a set of random variables with multinomial distribution. It can be shown that for each $i \in \{1, \ldots, k\}$, $\mathrm{E}(X_i) = np_i$. This is called the **expected count** in a random sample (of size $n$); however, the **observed count** is $X_i = n_i$ for the $i$th category.

If $p_i$ is the *true* probability of drawing an observation from category $i$, then

$$n_i - \mathrm{E}(X_i) = n_i - np_i$$

is the **deviance** between the observed and expected counts.

If we want to test the hypotheses

$$\mathcal{H}_0 : p_1 = p_1^*, \ldots, p_k = p_k^*; \quad \mathcal{H}_a : p_i \neq p_i^* \text{ for at least one } i \in \{1, \ldots, k\},$$

where $p_i^*$ represents the hypothesized probability of $c_i$, we use the test-statistic

$$X^2 = \sum_{i=1}^{k} \frac{(n_i - np_i^*)^2}{np_i^*},$$

where $\chi^2$ is called **Pearson's chi-square statistic**. In other words, the null hypothesis represents the case where the likelihood of all categories is equal, and thus the category has no effect on the observed response. It is often re-written as

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

with $O$ and $E$ standing for *observed* and *expected* cell values, respectively. $X^2$ can be thought of as the sum of the squared deviations under the hypothesized model.

The distribution of $X^2$ under the null hypothesis is approximately $\chi^2_\nu$, where $\nu = k - 1$, the degrees of freedom. We can thus for a rejection region given a significance level of $\alpha$,

$$\{X^2 > \chi^2_{\alpha,(k-1)}\}.$$

The $p$-value, similarly, would be

$$p = \Pr\{\chi^2_{(k-1)} > X^2\}.$$

For the test to be valid, the following must be true:

- The counts must represent a random sample from the population;
- The same size $n$ is large enough, so that for every cell the expected count is $\geq 5$.

## 3.3 Contingency Tables

Suppose the variables $X$ and $Y$ are observed for a sample of $n$ subjects, where $X$ has the values $j \in \{1, \ldots, r\}$ and $Y$ has the values $k \in \{1, \ldots, c\}$. In other words, for a subject $i \in \{1, \ldots, n\}$, $(X_i, Y_i) = (j, k)$ is observed. This data can be represented in what is called a **contingency table**.

A $d$-way contingency table contains all the counts of all possible combinations of levels of $d$ qualitative random variables; for instance, for a 2-way table for variables $X$ and $Y$ with $r$ rows and $c$ columns, the table would look like:

| | | | $Y$ | | |
|---|---|---|---|---|---|
| $X$ | 1 | 2 | $\cdots$ | c | Total |
| 1 | $n_{11}$ | $n_{12}$ | $\cdots$ | $n_{1c}$ | $n_{1\bullet}$ |
| 2 | $n_{21}$ | $n_{22}$ | $\cdots$ | $n_{2c}$ | $n_{2\bullet}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| $r$ | $n_{r1}$ | $n_{r2}$ | $\cdots$ | $n_{rc}$ | $n_{r\bullet}$ |
| Total | $n_{\bullet 1}$ | $n_{\bullet 2}$ | $\cdots$ | $n_{\bullet c}$ | $n$ |

By convention,

$$n_{j\bullet} = n_{j1} + \cdots + n_{jc}, \quad n_{\bullet k} = n_{1k} + \cdots + n_{rk},$$

where the sums $n_{1\bullet}, \ldots, n_{r\bullet}$ and $n_{\bullet 1}, \ldots, n_{\bullet c}$ are called **marginal**, or observed, counts. Note that the sum of all entries is the sample size, i.e.

$$n_{11} + \cdots + n_{rc} = n,$$

and similarly, the sum of the marginalized sums is as well,

$$n_{1\bullet} + \cdots + n_{r\bullet} = n_{\bullet 1} + \cdots + n_{\bullet c} = n.$$

This can also clearly be seen in the table under both the "Total" row and column.

Equivalently, a contingency table can be written in terms of the probabilities of each cell, where, similarly, $p_{1\bullet}, \ldots, p_{j\bullet}$ and $p_{\bullet 1}, \ldots, p_{\bullet k}$ are called **marginal** probabilities:

$$Y$$

| $X$ | 1 | 2 | $\cdots$ | c | Total |
|---|---|---|---|---|---|
| 1 | $p_{11}$ | $p_{12}$ | $\cdots$ | $p_{1c}$ | $p_{1\bullet}$ |
| 2 | $p_{21}$ | $p_{22}$ | $\cdots$ | $p_{2c}$ | $p_{2\bullet}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| $r$ | $p_{r1}$ | $p_{r2}$ | $\cdots$ | $p_{rc}$ | $p_{r\bullet}$ |
| Total | $p_{\bullet 1}$ | $p_{\bullet 2}$ | $\cdots$ | $p_{\bullet c}$ | 1 |

The sum of all the probabilities is 1;

$$p_{11} + \cdots + p_{rc} = p_{1\bullet} + \cdots + p_{r\bullet} = p_{\bullet 1} + \cdots + p_{\bullet c} = 1$$

### 3.3.1 Testing Indepdence

To test whether $X$ and $Y$ are independent, we are essentially testing **stochastic independence** between $X$ and $Y$, meaning

$$p_{jk} = \Pr(X = j \text{ and } Y = k) = \Pr(X = j)\Pr(Y = k), \forall\, j \in \{1, \ldots, r\}, \forall\, k \in \{1, \ldots, c\}.$$

Specifically,

$$\begin{array}{llll} p_{11} = p_{1\bullet}p_{\bullet 1} & p_{12} = p_{1\bullet}p_{\bullet 2} & \cdots & p_{1c} = p_{1\bullet}p_{\bullet c} \\ p_{21} = p_{2\bullet}p_{\bullet 1} & p_{22} = p_{2\bullet}p_{\bullet 2} & \cdots & p_{2c} = p_{2\bullet}p_{\bullet c} \\ \vdots & \vdots & \ddots & \vdots \\ p_{r1} = p_{r\bullet}p_{\bullet 1} & p_{r2} = p_{r\bullet}p_{\bullet 2} & \cdots & p_{rc} = p_{r\bullet}p_{\bullet c} \end{array}$$

Under the assumption of independence (our $\mathcal{H}_0$), the expected counts in a random sample of size $n$ is given as

$$E_{jk} = np_{jk} = np_{j\bullet}p_{\bullet k},$$

$\forall\, j \in \{1, \ldots, r\}$ and $\forall\, k \in \{1, \ldots, c\}$. Specifically (similar to above...),

$$\begin{array}{llll} E_{11} = np_{1\bullet}p_{\bullet 1} & E_{12} = np_{1\bullet}p_{\bullet 2} & \cdots & E_{1c} = np_{1\bullet}p_{\bullet c} \\ E_{21} = np_{2\bullet}p_{\bullet 1} & E_{22} = np_{2\bullet}p_{\bullet 2} & \cdots & E_{2c} = np_{2\bullet}p_{\bullet c} \\ \vdots & \vdots & \ddots & \vdots \\ E_{r1} = np_{r\bullet}p_{\bullet 1} & E_{r2} = np_{r\bullet}p_{\bullet 2} & \cdots & E_{rc} = np_{r\bullet}p_{\bullet c} \end{array}$$

The marginal probabilities here (for the population) are unknown, and are estimated

$$\hat{p}_{j\bullet} = n_{j\bullet}/n \quad \text{and} \quad \hat{p}_{\bullet k} = n_{\bullet k}/n,$$

which are simply the proportions in the sample. Thus, we can find the expected counts (under assumption of independence),

$$\hat{E}_{jk} = n\hat{p}_{j\bullet}\hat{p}_{\bullet k} = n(n_{j\bullet}/n)(n_{\bullet k}/n) = n_{j\bullet}n_{\bullet k}/n,$$

$\forall\, j \in \{1, \ldots, r\}$ and $\forall\, k \in \{1, \ldots, c\}$. This can then be used to find the individual estimated expected counts for each $X_i$ and $Y_k$.

The test-statistic in this case is given as

$$X^2 = \sum_{j=1}^{r} \sum_{k=1}^{c} \frac{(n_{jk} - (n_{j\bullet} n_{\bullet k}/n))^2}{n_{j\bullet} n_{\bullet k}/n}.$$

Under the $\mathcal{H}_0$ of independence, $X^2 \sim \chi^2_{(r-1)(c-1)}$, i.e., d.f. $= (r-1)(c-1)$. This allows us to construct a $\alpha$-level rejection region of

$$\text{RR} = \{X^2 > \chi^2_{[\alpha,(r-1)(c-1)]}\}.$$

Similarly to earlier, the following conditions must hold for this test to be valid:

- The counts represent a random sample from the population (multinomial experiment, $r \times c$ possible outcomes)
- The sample size $n$ is large enough such that for *every* cell the estimated expected count $\geq 5$

## 3.4 Chi-Square Test Caveats

Pearson's test relies on a chi-square approximation of the distribution of $X^2$. This approximation has some caveats:

- If the expected number of cell counts is small for any one of the cells ($< 5$), the approximation is poor. In this case, other tests (**Fisher's exact test**) can be used.
- The observations must be mutually independent and identically distributed; for paired qualitative data, we can use **McNemar's test**.

### 3.4.1 Fisher's Exact Test

Fisher's Exact Test is a test for independence of bivariate qualitative data (i.e. contingency tables). It is "exact" as it does not rely on a large-sample.

It is generalized to the Fisher-Freeman-Halton test for contingency tables larger than $2 \times 2$.

### 3.4.2 McNemar's Test

McNemar's Test can be used with frequency counts collected from matched-pairs experiments (i.e. two measures for a particular individual). An example table for this type of table would be:

| | Response 2 | | |
|---|---|---|---|
| Response 1 | Yes | No | Total |
| Yes | $n_{11}$ | $n_{12}$ | $n_{1\bullet}$ |
| No | $n_{21}$ | $n_{22}$ | $n_{2\bullet}$ |
| Total | $n_{\bullet 1}$ | $n_{\bullet 2}$ | $n$ |

The question of interest for such data is whether the proportion of **Response 1: Yes** is the same as the proportion of **Response 2: Yes**, which cannot be answered using chi-square test.

Let $p_1$ denote the true, population probability of Response 1: Yes, and $p_2$ the true, population probability of Response 2: Yes; we then want to test

$$\mathcal{H}_0 : p_1 = p_2; \quad \mathcal{H}_a : p_1 \neq p_2.$$

As before, $\hat{p}_1 = n_{1\bullet}/n$ and $\hat{p}_2 = n_{\bullet 1}/n$, and thus

$$\begin{aligned}\hat{p}_1 - \hat{p}_2 &= n_{1\bullet}/n - n_{\bullet 1}/n \\ &= (n_{11} + n_{12})/n - (n_{11} + n_{21})/n \\ &= (n_{12} - n_{21})/n.\end{aligned}$$

McNemar's test is based on the conditional binomial distribution of $(n_{12}, n_{21})$, giving the test statistic

$$Q_M = \frac{(n_{12} - n_{21})^2}{(n_{12} + n_{21})},$$

which has an approximate chi-square distribution with 1 degree of freedom.

For contingency tables larger than $2 \times 2$, the Stuart-Maxwell test and Bhakpar tests are generalizations.

# 4 Nonparametric Statistics

**Nonparametric statistics** are helpful as in they do not rely on the distribution of the sampled population.

For instance, the median, $\eta$, of a sample can be a nonparametric statistic, say

$$\mathcal{H}_0 : \eta = \eta_0.$$

## 4.1 Wilcoxon Test for Independent Samples

Take $X_1, \ldots, X_{n_1}$ be a random sample from population 1, and $Y_1, \ldots, Y_{n_2}$ be a random sample from population 2, and assume the two samples are independent. The *Wilcoxon rank sum test* can be used to test the hypothesis that the probability distributions associated with the two populations are equivalent. It goes as follows:

1. Put all the observations for both groups together, as if they were from the same population, for a group of size $n = n_1 + n_2$.
2. Order them from smallest to largest.
3. Rank them from smallest to largest, ie, the smallest gets rank 1, the largest gets rank $n$. If there are ties, take the averages of the ranks and assign it to each observation.

4. Separate the ranks by group and sum the ranks of each group;

$$W_X = \text{ sum of the ranks associated with the observations } X_1, \ldots, X_{n_1}$$

$$W_Y = \text{ sum of the ranks associated with the observations } Y_1, \ldots, Y_{n_2}$$

The test is based on the following idea;

- If there is *no difference* between the probability distribution (p.d.) of population 1 and the p.d. of population 2, the ranks of the $X$ values (and hence also the $Y$'s) are distributed randomly in the set $\{1, \ldots, n\}$.
- If the p.d. of population 1 is shifted *left* of the p.d. of population 2, the ranks of the $X$ values tend to be smaller than if randomly, and thus so will their sum.
- If the p.d. of population 1 is shifted *right* of the p.d. of population 2, the ranks of the $X$ values tend to be larger than if randomly, and thus so will their sum.

Let $D_1$ and $D_2$ denote the p.d.'s of populations 1 and 2 respectively. The null hypothesis would be $\mathcal{H}_0 : D_1 = D_2$. The relevant test statistics would be $T = W_x$ if $n_1 < n_2$, $T = W_Y$ if $n_1 > n_2$, and either if $n_1 = n_2$.

- If $\mathcal{H}_a$: $D_1$ shifted **left** of $D_2$,

$$\text{RR} = \{T \leq T_L\} \text{ if } W_X \text{ chosen, OR } \{T \geq T_U\} \text{ if } W_Y \text{ is chosen}$$

- If $\mathcal{H}_a$: $D_1$ shifted **right** of $D_2$,

$$\text{RR} = \{T \geq T_U\} \text{ if } W_X \text{ chosen, OR } \{T \leq T_L\} \text{ if } W_Y \text{ is chosen}$$

- If $\mathcal{H}_a$: $D_1$ shifted **anywhere** of $D_2$,

$$\text{RR} = \{T \leq T_L \text{ OR } T \geq T_U\}$$

Where $T_L$ and $T_U$ are table values. The necessary conditions for this test to be valid are:

- Independent samples;
- The populations from which we are sampling have both a continuous distribution.

Rather than basing the test on $W_X$ or $W_Y$, we use the **Mann-Whitney $U$-statistic**, defined as

$$U = m_1 m_2 + \frac{m_1(m_1 + 1)}{2} - T$$

where:

- $m_1$ is the smallest of $n_1$ and $n_2$;
- $m_2 = n - m_1$;

- $T$ is the test-statistic for the Wilcoxon test.

There is also a **large-sample** version of this test, for $n_1 \geq 10$ and $n_2 \geq 10$;

$$Z = \frac{U - (n_1(n_1 + n_2 + 1))/2}{\sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}},$$

where $Z$ has an approximate $\mathcal{N}(0, 1)$ distribution.

## 4.2 Wilcoxon Test for Paired Samples

Wilcoxon's **signed rank test** can be used to test the distributions of matched-pair data. Let $X_1, \ldots, X_n$ and $Y_n, \ldots, Y_n$ be two random samples of paired observations, and let $\text{Diff}_1 = X_1 - Y_1, \ldots, \text{Diff}_n = X_n - Y_n$ be the sample of differences. Now:

1. Order the absolute value of differences from smallest to largest.
2. Rank them, after taking out all differences that are equal to 0.

Ties are handled as discussed previously; rank them as if they were consecutive, take the average of the ranks, and assign it to each observation.

To understand the relevant hypotheses, we define the following:

- $T_+$: the sum of the ranks of the differences that were *positive* before taking the absolute values.
- $T_-$: the sum of the ranks of the differences that were *negative* before taking the absolute values.

Taking $D_1$ and $D_2$ to denote the probability distributions of populations 1 and 2, respectively, we wish to test the null hypothesis $\mathcal{H}_0 : D_1 = D_2$.

- If $\mathcal{H}_a : D_1$ is shifted left of $D_2$, $T = T_+$; $RR = \{T_+ \leq T_0\}$.
- If $\mathcal{H}_a : D_1$ is shifted right of $D_2$, $T = T_-$; $RR = \{T_- \leq T_0\}$.
- If $\mathcal{H}_a : D_1$ is shifted to either the left or right of $D_2$, $T = \min(T_-, T_+)$; $RR = \{T \leq T_0\}$.

Where $T_0$ is the table value.

The necessary conditions for this test are:

- The sample differences are randomly selected from the population differences;
- The p.d. from which the sample of paired differences is taken from is continuous.

There is also a **large-sample** version of this test, for $n \geq 25$; we have

$$Z = \frac{T_+ - [n(n+1)/4]}{\sqrt{n(n+1)(2n+1)/24}},$$

where $Z$ has an approximate $\mathcal{N}(0,1)$ distribution. From here, the $RR$ and $p$-value are the same as with any test following the standard Normal distribution.

## 4.3 Kruskal-Wallis Test

The **Kruskal-Wallis Test** can be used to investigate the difference in distribution among more than two groups. It arises from what is called a **completely randomized design** (CRD), where the groups are assigned completely at random so that each subject has the same chance of being assigned; any difference is considered experimental error.

The data are assumed to be random samples from $K$ independent populations; i.e.

$$\begin{array}{l} X_{11}, \ldots, X_{1n_1} \text{ from population 1;} \\ X_{21}, \ldots, X_{2n_2} \text{ from population 2;} \\ \quad \vdots \qquad\quad \vdots \qquad\quad \vdots \qquad\quad \vdots \\ X_{K1}, \ldots, X_{Kn_K} \text{ from population K.} \end{array}$$

The hypotheses of interest are

$$\begin{array}{l} \mathcal{H}_0 : \text{The } K \text{ probability distributions are identical;} \\ \mathcal{H}_a : \text{At least two of the } K \text{ probability distributions differ in location.} \end{array}$$

The procedure of the test is very similar to that of the Wilcoxon rank sum test. The first step is to rank all of the observations as a single group, handling ties as usual. This gives us $\{(R_{11}, \ldots, R_{1n_1}), \ldots, (R_{K1}, \ldots, R_{Kn_K})\}$, the ranks from each sample.

Under $\mathcal{H}_0$, the ranks of the observations should be approximately of the same order of magnitude in each group. If $\mathcal{H}_0$ is false, the ranks in one group would be larger or smaller in magnitude than the ranks of at least one other group.

To summarize the ranks, the average rank is taken per group; for each $j \in \{1, \ldots, K\}$, $\bar{R}_j$ is the average of the pooled ranks of the $j^{\text{th}}$ group. In other words, if $R_j$ is the sum of the ranks of group $j$, then $\bar{R}_j = R_j/n_j$.

Thus, under $\mathcal{H}_0$, it is expected

$$\bar{R}_1 \approx \cdots \approx \bar{R}_K.$$

If this is true, then the rank average per group should be close to the overall average of ranks;

$$\bar{R} = (1 + \cdots + n)/n = (n+1)/2.$$

The **Kruskal-Wallis statistic** measures how much $\bar{R}_1, \ldots, \bar{R}_K$ deviate from $\bar{R}$:

$$\text{KW} = \frac{12}{n(n+1)} \sum_{j=1}^{K} n_j (\bar{R}_j - \bar{R})^2.$$

KW is similar to the total sum of squares in a one-way ANOVA, but using ranks rather than actual data. KW is also approximately distributed according to a $\chi^2$ distribution, with $K-1$ degrees of freedom (it isn't $K$ because the if the average ranks of $K-1$ of the groups are known, then the $K$th average can be deduced).

$\mathcal{H}_0$ is rejected for large values of KW, because this indicates that the average ranks of the groups are very different compared to what one would expect. The necessary conditions for the test are:

- The $K$ samples are random and independent.
- There are five or more measurements in each sample.
- The $K$ probability distributions from which the samples are drawn are continuous.

## 4.4 Friedman Test

This test addresses the analog of a matched-pairs design but for more than 2 groups. We first define the notion of "block", a group of experimental units that receive all levels of "treatments" exactly once. A **randomized block design** (RBD) has two steps:

1. Blocks are formed with each block consisting of $K$ experimental units ($K$ is the number of "treatments"). The $B$ blocks should consist of experimental units that are as similar as possible.
2. One experimental unit from each block is randomly assigned to each treatment, given $n = BK$ observations. The data would resemble the following:

| | Treatment | | |
|---|---|---|---|
| Block | 1 | $\cdots$ | $K$ |
| 1 | $Y_{11}$ | $\cdots$ | $Y_{1K}$ |
| 2 | $Y_{21}$ | $\cdots$ | $Y_{2K}$ |
| $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| $B$ | $Y_{B1}$ | $\cdots$ | $Y_{BK}$ |

A special case of RBD: $B$ subjects, each subject receives all $K$ treatments. Each subject is a block, and the experimental units are the repeat assessments on the same subject. The order of the treatments should be randomly assigned to each subject.

The **Friedman test** is based on the rank sums for each treatment.

The first step is to rank the observations within blocks.

Then, for each $j \in \{1, \ldots, K\}$, $R_j :=$ sum of the ranks *within* the $j$th treatment. Then, let

$$\bar{R}_j = R_j/B$$

be the average of the ranks within the $j$th treatment (across all blocks.) This gives us the ranks:

| | Treatment | | |
|---|---|---|---|
| Block | 1 | $\cdots$ | $K$ |
| 1 | $R_{11}$ | $\cdots$ | $R_{1K}$ |
| 2 | $R_{21}$ | $\cdots$ | $R_{2K}$ |
| $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| $B$ | $R_{B1}$ | $\cdots$ | $R_{BK}$ |
| Totals | $R_1 = \sum_{i=1}^{B} R_{i1}$ | $\cdots$ | $R_K = \sum_{i=1}^{B} R_{iK}$ |
| Means | $\bar{R}_1 = R_1/B$ | $\cdots$ | $\bar{R}_K = R_K/B$ |

Note: within each block, the sum of ranks is $K(K+1)/2$. Thus, the total sum of ranks is $BK(K+1)/2$, and the total aerage of ranks is

$$\bar{R} = \frac{B[K(K+1)/2]}{BK} = (K+1)/2.$$

Under $\mathcal{H}_0$, we expect

$$\bar{R}_1 \approx \cdots \approx \bar{R}_K \approx \bar{R} = (K+1)/2.$$

The hypotheses of interest:

$\mathcal{H}_0$ : The $K$ p.d.'s are identical;

$\mathcal{H}_a$ : At least two of the $K$ p.d.'s differ in location.

The **Friedman Statistic** measures how much $\bar{R}_1, \ldots, \bar{R}_K$ deviate from $\bar{R}$:

$$F_r = \frac{12B}{K(K+1)} \sum_{j=1}^{K} (\bar{R}_j - \bar{R})^2.$$

Under $\mathcal{H}_0$, $F_r$ is approximately distributed as $\chi^2$ with $(K-1)$ degrees of freedom. The necessary conditions are:

- Treatments are randomly assigned to the experimental units within the blocks.
- The measurements can be ranked within blocks.
- The $K$ p.d.'s from which the samples are drawn are continuous.
- Either $B$ *or* $K$ is bigger than 5.

## 4.5 Spearman Rank Correlation

Assume that we have $n$ mutually independent and identically distributed random pairs of variables

$$(X_1, Y_1), \ldots, (X_n, Y_n)$$

. The first step of this test is to rank each observation within each variable (i.e., within the $X$'s and the $Y$'s).

Let $u_i$ and $v_i$ be the ranks of observations $x_i$ and $y_i$, respectively. From here, there are two ways to compute the Spearman rank correlation. First: just take the sample *correlation coefficient* of the ranks:

$$r_s = \frac{\text{SS}_{uv}}{\sqrt{\text{SS}_{uu}\text{SS}_{vv}}},$$

where

$$\begin{aligned}
\text{SS}_{uv} &= \sum_{i=1}^{n}(u_i - \bar{u})(v_i - \bar{v}) = \sum_{i=1}^{n} u_i v_i - n\bar{u}\bar{v}, \\
\text{SS}_{uu} &= \sum_{i=1}^{n}(u_i - \bar{u})^2 = \sum_{i=1}^{n} u_i^2 - n\bar{u}^2, \\
\text{SS}_{vv} &= \sum_{i=1}^{n}(v_i - \bar{v})^2 = \sum_{i=1}^{n} v_i^2 - n\bar{v}^2.
\end{aligned}$$

If there are no ties in neither the $X$'s nor the $Y$'s, we can use

$$r_s = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n} d_i^2,$$

where $d_i = u_i - v_i$, the difference in the ranks of $X_i$ and $Y_i$.

Generally, $-1 \leq r_s \leq 1$, where:

- $r_s = -1 \implies$ perfect negative correlation;
- $r_s = 0 \implies$ no correlation;
- $r_s = 1 \implies$ perfect positive correlation.

These are all under the standard condition that random samples are drawn from continuous probability distributions.

### 4.5.1 Creating a Confidence Interval

A $(1-\alpha)100\%$ confidence interval for $\rho_s$ can be constructed using the previously discussed Fisher's variance stabilizing $z$-transformation. In short; transform $z = \frac{1}{2}\ln\left(\frac{1+r_s}{1-r_s}\right)$, created a c.i. for $z$ with $(c_L, c_U) = z \pm \frac{z_{\alpha/s}}{\sqrt{n-3}}$, then transform back to $\rho_s$ with $\left[\frac{e^{2c_L}-1}{e^{2c_L}+1}, \frac{e^{2c_U}-1}{e^{2c_U}+1}\right]$.

## 4.6 Analysis of Variance

ANOVA is a way of testing the quality of means of a response variable in $K$ populations;

$$\mathcal{H}_0 : \mu_1 = \cdots = \mu_K,$$
$$\mathcal{H}_a : \text{At least one of the } \mu_j \text{ differs from the others.}$$

This will assume the data are from a CRD.

When $K = 2$, $\mathcal{H}_0 : \mu_1 = \mu_2$, which can simply be done with a Student $t$-statistic. When $K > 2$, we could theoretically look at all pair-wise differences of means; however, this makes it very difficult to find a standard test statistic, and further, would take a long time to compute for large $K$.

If we instead assume that

- the variances of the response variable are the same for each treatment (homoscedasticity: $\sigma_1^2 = \cdots = \sigma_K^2$) and
- the response variable is Normally distributed,

the p.d. of the response variable is then the same in *each* group; except (perhaps) for the mean.

Under $\mathcal{H}_0 : \mu_1 = \cdots = \mu_K$, all observations are from the **same** distribution ($\mathcal{N}(\mu, \sigma^2)$), where

$$\mu = \mu_1 = \cdots = \mu_k \quad \text{and} \quad \sigma^2 = \sigma_1^2 = \cdots = \sigma_K^2.$$

We thus have the following setup:

$$\begin{array}{rccc} \text{Group 1:} & Y_{11}, \ldots, Y_{1n_1} & \sim & \mathcal{N}(\mu_1, \sigma^2), \\ \text{Group 2:} & Y_{21}, \ldots, Y_{2n_2} & \sim & \mathcal{N}(\mu_2, \sigma^2), \\ \vdots & \vdots & & \vdots \\ \text{Group } K\text{:} & Y_{K1}, \ldots, Y_{Kn_K} & \sim & \mathcal{N}(\mu_K, \sigma^2). \end{array}$$

We take $n = n_1 + \cdots + n_K$, the total number of observations. For each $k \in \{1, \ldots, K\}$, we can estimate $\mu_k$ given the group sample mean:

$$\bar{Y}_k = \frac{1}{n_k}(Y_{k1} + \cdots + Y_{kn_k}) = \hat{\mu}_k$$

Under $\mathcal{H}_0 : \mu_1 = \cdots = \mu_K = \mu$, an estimate of the common mean $\mu$ is given by the average of the observations, over all groups:

$$\hat{\mu} = \bar{Y} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} Y_{ki}.$$

If $\mathcal{H}_0$ is *true*, ie the treatment means are all the same, then *each* of the $\bar{Y}_k$ should be close to $\bar{Y}$. What does "close"

mean? One way to measure the distance of all of the treatment means from the overall mean is via

$$\text{SST} = \sum_{k=1}^{K} n_k(\hat{\mu}_K - \hat{\mu})^2 = \sum_{k=1}^{K} n_k(\bar{Y}_k - \bar{Y})^2,$$

the sum of the squared deviations from the overall mean, weighted by the number of observations in each group.

When SST is large, this indicates evidence against $\mathcal{H}_0$, i.e., there would be evidence that at least one of the means differs from the others, and conversely, if SST is small, it would not indicate evidence against $\mathcal{H}_0$. This naturally leads to the question of what makes a particular SST large vs small.

First, note that a measure of sampling variability can be given by the sum of squares associated with the errors, specifically,

$$\text{SSE} = \sum_{k=1}^{K}\sum_{i=1}^{n_k}(Y_{ki} - \bar{Y}_k)^2 = \sum_{k=1}^{K}(n_k - 1)S_k^2,$$

where $S_j^2$ is the sample variance in group $j$. To compare SST and SSE, we have to convert "sum of squares" to "mean of squares" by dividing each sum of squares by its degrees of freedom.

We define:

$$\text{MST} = \frac{\text{SST}}{(K-1)}; \quad \text{MSE} = \frac{\text{SSE}}{(n-K)},$$

and from here we define our test-statistic,

$$F = \frac{\text{MST}}{\text{MSE}} = \frac{\text{SST}/(K-1)}{\text{SSE}/(n-K)}.$$

If $\mathcal{H}_0$ is true, then both MST and MSE should be close to each other, i.e., F should be close to 1.

We can interpret this as the differences in treatment means should attributable to sampling error, providing little support against $\mathcal{H}_0$. If $\mathcal{H}_0$ is *not* true, then MST will be *large* on average, and so we will want to reject only for large $F$.

Overall, we have, under $\mathcal{H}_0$ and the previously stated assumptions,

$$F = \frac{\text{MST}}{\text{MSE}} \sim \mathcal{F}_{K-1,n-K}.$$

This gives us a rejection region at a significance of $\alpha$ of

$$RR = \{F > F_{\alpha,K-1,n-K}\}.$$

## 4.7 Comparing Multiple Means

ANOVA is limited by the fact that it only "detects" when at least one mean is significantly different than the rest. We may, perhaps, want to compare the means between any combination of two means. For $K$ means, there are $C = \frac{K(K-1)}{2}$ possible pair-wise comparisons that we can make. When $K = 2$, with $n_1$ and $n_2$ observations in

each group respectively, we can measure the difference in the means using

$$(\bar{Y}_1 - \bar{Y}_2) \pm t_{\alpha/2, n_1+n_2-2}\sqrt{S_p^2/n_1 + S_p^2/n_2},$$

where $S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$ is the pooled variance.

Assuming the population variances are equal between groups, we can use the MSE measure to better estimate common variance, ie

$$(\bar{Y}_i - \bar{Y}_j) \pm t_{(\alpha/2, n-K)}\sqrt{\text{MSE}/n_i + \text{MSE}/n_j}.$$

There is a problem with this method however; for each individual confidence interval we create, we can be $100 \times (1-\alpha)\%$ confident that the true difference in means will be contained in our interval, *but* we do not have the same confidence that all of the mean differences, overall, will be contained in their respective intervals. For a pseudo-math explanation:

$$\begin{aligned}
&\Pr(\text{at least one "bad" interval}) = 1 - \Pr(\text{All "good intervals"}) \\
&= 1 - \Pr(1 \text{ is good}, \ldots, \text{C is good}) \\
&= 1 - \Pr(1 \text{ is good}) \times \cdots \times \Pr(\text{C is good}) \\
&= 1 - [(1-\alpha) \times \cdots \times (1-\alpha)] = 1 - (1-\alpha)^C
\end{aligned}$$

(where a "good" interval contains the true difference, and a "bad" one does not.) Even for $K = 3$, this gives us a "bad" error rate of 14% for $\alpha = 0.05$, and becomes far worse as $K$ increases; this is known as the problem of **multiple comparisons**. We can say that we have a **comparison-wise** error rate (CER) of $\alpha$, and a **experiment-wise** error rate (EER) of $1 - (1-\alpha)^C$.

The general solution to address this issue is to adjust the CER to get a reasonable EER. One such method is the **Boneferroni method**, which is the most general, and conservative. To obtain a EER of $\alpha_E$, then we should choose our CER to be

$$\alpha = \alpha_F / C.$$

For instance, if we have $C = 5$ and $\alpha_F = 0.05$, then we should choose $\alpha = 0.01$; if we build five 99% confidence intervals, then the probability that they all contain their true target parameters will be about 95%.

Other methods include

- **Tukey's Honest Significant Difference**: only applicable when groups are of equal sizes, and is only really helpful in making pair-wise comparisons.
- **Scheffé's Method**: helpful for linear combinations of means ("contrasts"), or for pairs of means.

## 4.8 ANOVA with Randomized Block Designs

Recall that an RBD occurs when treatments are randomly assigned to units within each block. Let $B$ be the number of blocks and $K$ the number of treatments/groups; thus, there are $n = B \times K$ observations in total. Ideally, each possible ordering of treatments should appear an equal number of times in the analysis, and the orderings should be randomized to the subjects. This gives a general data structure:

| | Treatment | | | | |
|---|---|---|---|---|---|
| Block | 1 | $\cdots$ | K | Total | Mean |
| 1 | $Y_{11}$ | $\cdots$ | $Y_{1K}$ | $Y_{1\bullet} = \sum_{i=1}^{K} Y_{1i}$ | $\bar{Y}_{1\bullet}$ |
| 2 | $Y_{21}$ | $\cdots$ | $Y_{2K}$ | $Y_{2\bullet} = \sum_{i=1}^{K} Y_{2i}$ | $\bar{Y}_{2\bullet}$ |
| $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| B | $Y_{B1}$ | $\cdots$ | $Y_{BK}$ | $Y_{B\bullet} = \sum_{i=1}^{K} Y_{Bi}$ | $\bar{Y}_{B\bullet}$ |
| Total | $Y_{\bullet 1} = \sum_{j=1}^{B} Y_{j1}$ | $\cdots$ | $Y_{\bullet K} = \sum_{j=1}^{B} Y_{jK}$ | | |
| Mean | $\bar{Y}_{\bullet 1}$ | $\cdots$ | $\bar{Y}_{\bullet K}$ | | |

We can define:

$$\text{SST} = B[(\bar{Y}_{\bullet 1} - \bar{Y})^2 + \cdots + (\bar{Y}_{\bullet K} - \bar{Y})^2] = B\sum_{i=1}^{K}(\bar{Y}_{\bullet i} - \bar{Y})^2$$

$$\text{SSB} = K[(\bar{Y}_{1\bullet} - \bar{Y})^2 + \cdots + (\bar{Y}_{B\bullet} - \bar{Y})^2] = K\sum_{j=1}^{B}(\bar{Y}_{j\bullet} - \bar{Y})^2$$

$$\text{SS(Total)} = \sum_{i=1}^{K}\sum_{j=1}^{B}(Y_{ij} - \bar{Y})^2$$

$$\text{SSE} = \text{SS(Total)} - \text{SST} - \text{SSB}$$

From here, we can create an ANOVA table for RBDs:

| Source | df | SS | MS | F |
|---|---|---|---|---|
| Treat | $K-1$ | SST | $\text{MST} = \text{SST}/(K-1)$ | $\frac{\text{MST}}{\text{MSE}}$ |
| Block | $B-1$ | SSB | $\text{MSB} = \text{SSB}/(B-1)$ | $\frac{\text{MSB}}{\text{MSE}}$ |
| Error | $n-K-B+1$ | SSE | $\text{MSE} = \text{SSE}/(n-K-B+1)$ | |
| Total | $n-1$ | SS(Total) | | |

### 4.8.1 F-test for Treatment Means

We can perform an ANOVA $F$-test to compare treatment means;

$$\mathcal{H}_0 : \mu_1 = \cdots = \mu_K; \text{ vs. } \mathcal{H}_a : \text{At least two of these means differ.}$$

We have a test-statistic of

$$F = \frac{\text{SST}/(K-1)}{\text{SSE}/(n-K-B+1)} = \frac{\text{MST}}{\text{MSE}}.$$

Under $\mathcal{H}_0$, $F \sim \mathcal{F}(K-1, n-K-B+1)$. We have a rejection region of

$$RR = \{F > F_{\alpha,K-1,n-K-B+1}\}.$$

Finally, we have a $p$-value of $p = \Pr(F_{K-1,n-K-B+1} > F_{\text{obs}})$.

The necessary conditions for this test are that:

- The $B$ blocks are randomly selected and the $K$ treatments (groups) are randomly assigned to the experimental units within the blocks.
- The probability distribution of responses for each $BK$ block-treatment combinations is normal.
- The $BK$ block-treatment distributions have equal variances.

When $\mathcal{H}_0$, we can proceed with a multiple comparison of means (see previous section).

We can also conduct an $F$-test for *block* means, though this is typically of little interest; this gives

$$\mathcal{H}_0 : \tau_1 = \cdots = \tau_B \text{ vs } \mathcal{H}_a : \text{ at least one mean differs,}$$

where $\tau_j$ represents the true mean response of block $j$, where $j \in (1, \ldots, B)$. We have:

$$\begin{aligned} F &= \frac{\text{MSB}}{\text{MSE}} \\ RR &= \{F > F_{\alpha,B-1,n-K-B+1}\} \\ p &= \Pr(F_{B-1,n-K-B+1} > F_{\text{obs}}). \end{aligned}$$

The only factor that changes is the degrees of freedom; we now have the first degree of freedom $= B - 1$.

## 4.9 Two-Way ANOVA

Suppose we wish to investigate how two factors, $A$ and $B$, affect a response variable; this is done with **factorial experiments**. If, say, $A$ and $B$ have $J$ and $K$ levels respectively, a **complete factorial experiment** is one in which every factor-level combination is used. Specifically, there must be observations at each of the $J \times K$ combinations of levels.

The $R$ observations per "treatment" or "group" are called **replications**; when $J = K$, we have a **balanced complete factorial experiment**. If there are $R$ replications for each of the groups, it follows that we have $n = J \times K \times R$ observations. This specific design can be analyzed through the use of a two-way ANOVA.

**Definition 6 (*Balanced Complete Factorial Experiment*)**

> *A* factorial experiment *is one in which there are $J$ levels of factor $A$ and $K$ levels of factor $B$; it becomes* complete *when every possible $J \times K$ combination of levels is used; it becomes* balanced *when there are $R$ replications for each of the $J \times K$ groups.*

This results in a data table resembling the following:

| | Factor B | | | |
|---|---|---|---|---|
| Factor A | 1 | $\cdots$ | $k$ | K |
| 1 | $(Y_{111}, \ldots, Y_{11R})$ | $\cdots$ | $(Y_{1k1}, \ldots, Y_{1kR})$ | $(Y_{1K1}, \ldots, Y_{1KR})$ |
| $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| $j$ | $(Y_{j11}, \ldots, Y_{j1R})$ | $\cdots$ | $(Y_{jk1}, \ldots, Y_{jkR})$ | $(Y_{jK1}, \ldots, Y_{jKR})$ |
| $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| $J$ | $(Y_{J11}, \ldots, Y_{J1R})$ | $\cdots$ | $(Y_{Jk1}, \ldots, Y_{JkR})$ | $(Y_{JK1}, \ldots, Y_{JKR})$ |

where $Y_{jkr}$ is the response value for factor $A$ at level $j$ ($j \in \{1, \ldots, J\}$) and factor B at level $k$ ($k \in \{1, \ldots, K\}$) and replication $r$ ($r \in \{1, \ldots, R\}$).

As previously discussed, interaction is often a trend we must test for in statistical analysis. In the case of two-way ANOVA, this is *always* the first test we must conduct, and in the case that no interaction exists, we can then test if the main effects are significant.

### 4.9.1 Steps for Two-Way ANOVA

**Step 1:** Test for interaction between factors A and B. If there is evidence of an interaction, we can then use aforementioned methods of multiple comparison. IF not:

**Step 2:** Test for a main effect of factor A. If one exists, we can then use aforementioned methods of multiple comparison.

**Step 3:** Repeat step 2 for factor B.

The necessary conditions for ANOVA tests are as follows:

- Random and independent samples.
- The probability distribution of responses for each $JK$ factor-level combinations is approximately normal.
- The $JK$ factor-level distributions have equal variances.

### 4.9.2 Test Statistics

The overall sum of squares for treatments is defined as

$$\text{SST} = R \sum_{j=1}^{J} \sum_{k=1}^{K} (\bar{Y}_{jk\bullet} - \bar{Y})^2,$$

where $\bar{Y}_{ij\bullet}$ is the mean of the $(j,k)^{\text{th}}$ treatment group, and $\bar{Y}$ is the sample mean.

We break SST into three sources:

$$\text{SSA} = RK\sum_{j=1}^{J}(\bar{Y}_{j\bullet\bullet} - \bar{Y})^2; \text{MSA} = \frac{\text{SSA}}{J-1}$$
$$\text{SSB} = RJ\sum_{k=1}^{K}(\bar{Y}_{\bullet k\bullet} - \bar{Y})^2; \text{MSB} = \frac{\text{SSB}}{K-1}$$
$$\text{SS(AB)} = \text{SST} - \text{SSA} - \text{SSB}; \text{MS(AB)} = \frac{\text{SS(AB)}}{(J-1)(K-1)},$$

where SSA is the SS due to A, SSB is the SS due to B, and SS(AB) is the SS due to the interaction between A and B.

For any given treatment group $(j, k)$, its sample variance is

$$S_{jk}^2 = \frac{1}{R-1}\sum_{r=1}^{R}(Y_{jkr} - \bar{Y}_{jk\bullet})^2.$$

The sum of squares of errors is defined

$$\text{SSE} = \sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{r=1}^{R}(Y_{jkr} - \bar{Y}_{jk\bullet})^2 = \sum_{j=1}^{J}\sum_{k=1}^{K}(R-1)S_{jk}^2$$
$$\text{MSE} = \frac{\text{SSE}}{n-JK}.$$

Putting this all together, the two-way ANOVA table is as follows:

| Source | df | SS | MS | $F$ |
|---|---|---|---|---|
| A | $J-1$ | SSA | $\text{MSA} = \frac{\text{SSA}}{J-1}$ | $\frac{\text{MSA}}{\text{MSE}}$ |
| B | $K-1$ | SSB | $\text{MSB} = \frac{\text{SSB}}{K-1}$ | $\frac{\text{MSB}}{\text{MSE}}$ |
| AB | $(J-1)(K-1)$ | SS(AB) | $\text{MS(AB)} = \frac{\text{SS(AB)}}{(J-1)(K-1)}$ | $\frac{\text{MS(AB)}}{\text{MSE}}$ |
| Error | $n-JK$ | SSE | $\text{MSE} = \frac{\text{SSE}}{n-JK}$ | |
| Total | $n-1$ | SST | | |

### 4.9.3 Interactions

We have the following hypotheses:

$\mathcal{H}_0$ : A and B do not interaction to affect the response

$\mathcal{H}_a$ : A and B do interact to affect the response

This has a test-statistic of

$$F = \frac{\text{SS(AB)}/\left[(J-1)(K-1)\right]}{\text{SSE}/(n-JK)} = \frac{\text{MS(AB)}}{\text{MSE}}.$$

Under $\mathcal{H}_0$, we have:

$$\begin{aligned} F &\sim \mathcal{F}_{(J-1)(K-1),\, n-JK} \\ \text{RR} &= \{F > F_{(J-1)(K-1),\, n-JK}\} \\ p &= \Pr(F_{\text{obs}} \in \text{RR}) \end{aligned}$$

### 4.9.4 Interpreting Main Effects

Take an experimental set up with $J$ levels of factor $A$ and $K$ levels of factor $B$, and $R$ replications for each of the $J \times K$ groups (ie, $n = J \times K \times R$). Say we did not reject the null hypothesis of no interaction (no strong evidence of an interaction). We then have to test the main effects of both $A$ and $B$.

**For $A$:**

$\mathcal{H}_0$ : No difference among the $J$ population mean responses due to $A$

$\mathcal{H}_a$ : At least two of the population means differ

We have a test-statistic of

$$F = \frac{\text{SSA}/(J-1)}{\text{SSE}/(n-JK)} = \frac{\text{MSA}}{\text{MSE}},$$

where $F$ has a $\mathcal{F}$-distribution with $J-1$ and $n-JK$ degrees of freedom; thus:

$$\begin{aligned} \text{RR} &= \{F > F_{\alpha,\, J-1,\, n-JK}\} \\ p &= \Pr(F_{J-1,n-JK} > F_{\text{obs}}) \end{aligned}$$

**For $B$:**

$\mathcal{H}_0$ : No difference among the $K$ population mean responses due to $B$

$\mathcal{H}_a$ : At least two of the population means differ

We have a test-statistic of

$$F = \frac{\text{SSB}/(K-1)}{\text{SSE}/(n-JK)} = \frac{\text{MSB}}{\text{MSE}},$$

where $F$ has a $\mathcal{F}$-distribution with $K-1$ and $n-JK$ degrees of freedom; thus:

$$\begin{aligned} \text{RR} &= \{F > F_{\alpha,\, K-1,\, n-JK}\} \\ p &= \Pr(F_{K-1,n-JK} > F_{\text{obs}}) \end{aligned}$$

As always, all of these tests and related conclusions are based on the assumptions of normality, constant variance, and random sampling.

# 5 Appendix

## 5.1 (Possible) Interpretations of $p$-values

| $p$-value | Evidence against $\mathcal{H}_0$ |
|---|---|
| $p < 0.001$ | Extremely Strong |
| $0.001 \leq p < 0.01$ | Very Strong |
| $0.01 \leq p < 0.05$ | Strong |
| $0.05 \leq p < 0.10$ | Modest |
| $p \geq 0.10$ | Weak |

## 5.2 Interpreting Outputs

### 5.2.1 Regression Summary

Calling **summary**() on a simple linear regression model (**lm**(y ~ x, **data**)); only showing the coefficients section of the table.

| | Estimate | Std. Error | $t$-value | $p$-value |
|---|---|---|---|---|
| (Intercept) | $\beta_0$ | $\text{SE}_{\beta_0}$ | $T_{\text{obs }0} = \beta_0/\text{SE}_{\beta_0}$ | $2 * \text{pt}(-T_{\text{obs }0}, df)$ |
| $x$ | $\beta_1$ | $\text{SE}_{\beta_1}$ | $T_{\text{obs }1} = \beta_1/\text{SE}_{\beta_1}$ | $2 * \text{pt}(-T_{\text{obs }1}, df)$ |

Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 *($\equiv p < \alpha$)*
Residual standard error: $\hat{\sigma}$ on $df$ degrees of freedom
Multiple R-squared: $R^2$, Adjusted R-squared: $R^2_{adj}$
F-statistic: $\frac{R^2/K}{(1-R^2)/(n-(K+1))}$ on $K$ and $n-(K+1)$ DF, $p$-value: $\text{pf}(F, K, n-(K+1), \text{lower.tail} = \text{FALSE})$

### 5.2.2 ANOVA

Simple case, calling **anova**() on a linear regression model.

| | DF | Sums of Squares | Mean Squares | $F$ | $p$ |
|---|---|---|---|---|---|
| $X$ | 1 | $\text{S}_{xx}$ | $\text{S}_{xx}/1$ | $\text{S}_{xx}/(\text{SSE}/(n-2))$ | $p$ |
| Residuals | $n-2$ | SSE | $\text{SSE}/(n-2)$ | | |

### 5.2.3 Two-Way ANOVA

| Source | df | SS | MS | $F$ |
|---|---|---|---|---|
| A | $J-1$ | SSA | MSA $= \frac{\text{SSA}}{J-1}$ | $\frac{\text{MSA}}{\text{MSE}}$ |
| B | $K-1$ | SSB | MSB $= \frac{\text{SSB}}{K-1}$ | $\frac{\text{MSB}}{\text{MSE}}$ |
| A:B | $(J-1)(K-1)$ | SS(AB) | MS(AB) $= \frac{\text{SS(AB)}}{(J-1)(K-1)}$ | $\frac{\text{MS(AB)}}{\text{MSE}}$ |
| Error | $n-JK$ | SSE | MSE $= \frac{\text{SSE}}{n-JK}$ | |
| Total | $n-1$ | SST | | |

## 5.3 Glossary

**Definition 1 (*Simple Linear Regression*)**

*Modeling of the relationship between two variables $X$ and $Y$, which assumes that $Y$ is a linear function of $X$. We can denote this:*

$$E(Y|X=x) = f(x)$$

$$Y = f(x) + \varepsilon, X = x$$

*where $\varepsilon$ is the error of the model.*

**Definition 2 (*Unbiasedness*)**

*$\hat{\beta}_1$ is unbiased for $\beta_1$ if it gives a good estimate over large number of samples, on **average**.*

**Definition 3 (*Analysis of Variance*)**

*Short-handed as "anova", this is a statistical method used to analyze the differences between groups, and specifically, compare the variance caused by error to the variance caused by estimation.*

**Definition 4 (*Correlation*)**

*A measure of association between two random variables.*

**Definition 5 (*Determination*)**

*A measure of the proportion of variance in $Y$ explained by the model (by $X$, that is).*

**Definition 6 (*Balanced Complete Factorial Experiment*)**

*A* factorial experiment *is one in which there are $J$ levels of factor $A$ and $K$ levels of factor $B$; it becomes* complete *when every possible $J \times K$ combination of levels is used; it becomes* balanced *when there are $R$ replications for each of the $J \times K$ groups.*

## 5.4 Summary of R Code

Available here.