1.4 Simple Linear Regression Model

We introduce the Simple Linear Regression model via the *University Admissions* example in R

University Admissions Example

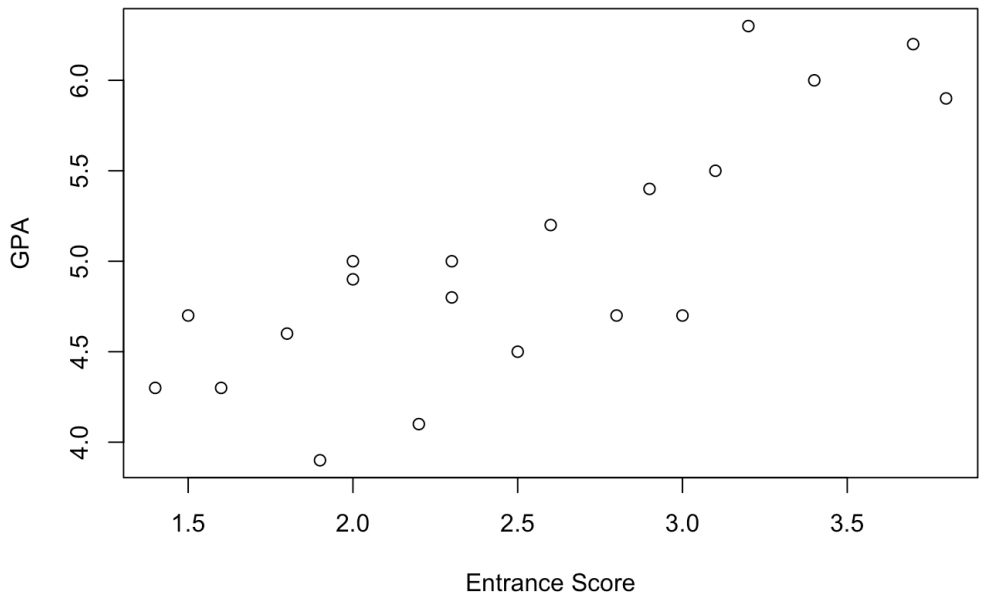
The director of admissions of a small college administered a newly designed entrance test to 20 students selected *at random* from the freshman class in a study to determine whether a student's grade point average (GPA) at the end of the freshman year (y) can be **predicted** from the entrance test score (x). The results of the study are summarized in the `admissions.txt` file (you can download the file here).

```
admissions = read.table("data/ch1/admissions.txt", header=FALSE)

# The data has no header, so we enter our own labels:
colnames(admissions)  <- c("entrance_score", "gpa")
```

We start with a scatter plot of the data to decide whether a linear relationship is a reasonable model:

```
plot(admissions$entrance_score, admissions$gpa, xlab="Entrance Score", ylab="GPA"
```



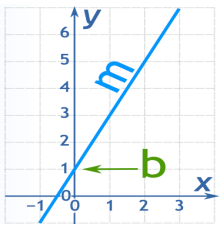
Based on the scatterplot above:

- What conclusions can we draw?
- Which variable depends on the other?
- How can we initially describe the data? Is there a trend?

> Without fitting *any* statistical model, we can say that there is a mild positive *linear* trend in the data, and there seems to be a dependence between GPA and Entrance Score, but we cannot quantify this relationship.

1.4.1 Statistical vs. Mathematical Relationship

Mathematically, a *straight line* is defined as follows:



$$y = \underbrace{m}_{slope} x + \underbrace{b}_{intercept}$$

The intercept b indicates the value of y when the variable x is 0 and the slope

m is the rate of change of y per unit change in x. This is a **deterministic** relationship, which means that there is *no uncertainty*.

A **Regression Line** is a **statistical** relationship between two variables y (the dependent) and x (the independent), which means that there is **uncertainty** arising from the *randomness* of the underlying phenomenon. The y (dependent variable) tends to vary with the (independent variable) x in a systematic (linear) fashion, but this relationship is not exact. There is a "scattering" of points around the line which we denote by adding an *error term* in the equation:

$$y = \underbrace{\beta_0}_{intercept} + \underbrace{\beta_1}_{slope} x + \text{error}$$

1.4.2 Simple Linear Regression Model & Assumptions

In **Simple Linear Regression**, we have *one* response $\mathbf{y}$ and *one* predictor $\mathbf{x}$, and therefore the data come in pairs:

x_1	y_1
x_2	y_2
...	...
x_n	y_n

The *Simple Linear Regression (SLR)* model is defined as follows:

SLR Model

$$\underbrace{y_i}_{\text{dependent var at level } i} = \beta_0 + \beta_1 \underbrace{x_i}_{\text{known constant}} + \underbrace{\varepsilon_i}_{\text{random error}},$$

where $\varepsilon_i \sim IID(0, \sigma^2)$, $i = 1, \ldots, n$ (n: sample size)[5]. The **intercept** β_0, the **slope** β_1, and the **error variance** σ^2 are the model *parameters*.

The model assumptions that we make so far are summarized below:

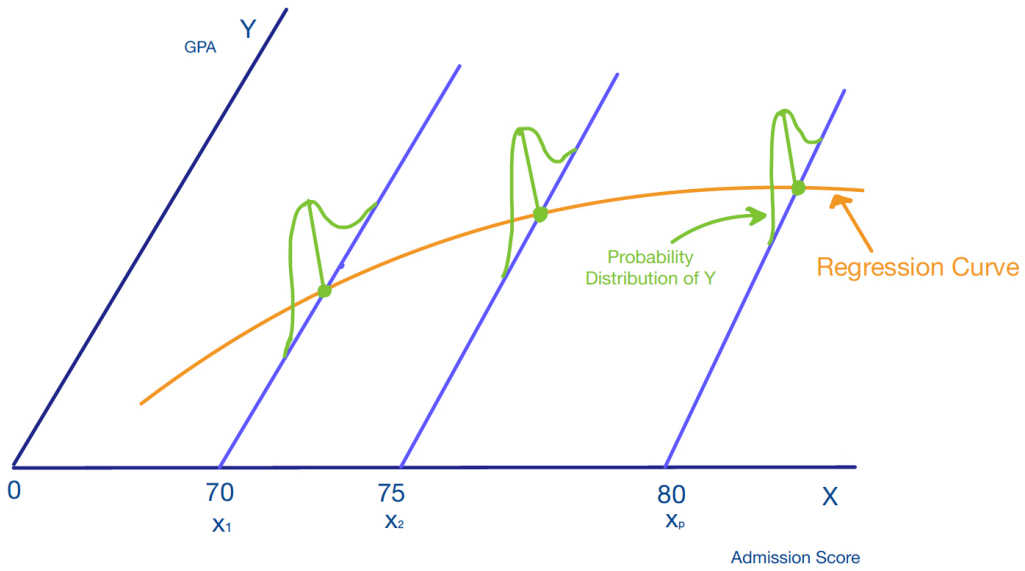
SLR Model Assumptions

The *random errors* $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are assumed to

- have **mean zero**: $\mathbb{E}(\varepsilon_i) = 0$
- be **uncorrelated**: $Cov(\varepsilon_i, \varepsilon_j) = 0, \ i \neq j,$
- be **homoscedastic**: $Var(\varepsilon_i) = \sigma^2$ does not depend on i.

Note that (so far) we made **no assumptions** on the *distribution* of the error terms. We will make such assumptions only when necessary.

A schematic **representation** of a regression model is shown below:



In a regression model describing the relationship between $\mathbf{y}$ and $\mathbf{x}$, we *assume* that

1. There is a probability distribution of $\mathbf{y}$ for every level of $\mathbf{x}$: *the probability of* $\mathbf{y}$ *happening at that level of* X.
2. Each probability distribution of $\mathbf{y}$ has a *mean* or *"center"*.

3. The means of all distributions vary in some systematic fashion (a *line* in linear regression, for example).

So, in other words, we **assume** that $\mathbf{y}$ is a *random variable that has a distribution for every level of the independent variable.* In other words, the regression curve, which describes the relation between the means of probability distributions of $\mathbf{y}$ and the level of $\mathbf{x}$, is the counterpart to the general tendency of $\mathbf{y}$ to vary systematically with $\mathbf{x}$ in a statistical relation.

Interpretation of the parameters β_1, β_0

- β_1, the slope, is the *difference in the means* of the probability distribution functions of $\mathbf{y}$ for two levels of $\mathbf{x}$ that differ by 1.
- β_0 is the *intercept* and is the mean of the probability distribution function of $\mathbf{y}$ when the level of $\mathbf{x}$ is 0. In a SLR model, the intercept may *not* have a valid/reasonable interpretation, if 0 is not a meaningful value of $\mathbf{x}$.

1.4.3 Paremeter Estimation: *Least-Squares Method*

To *fit the line to the data* we need to estimate the parameters of the regression line (slope and intercept). So, to estimate β_0, and β_1 we use the *least-squares* method:

Consider the responses y_i and the *expected responses* $\mathbb{E}(y_i)$. We would like to minimize the difference between what we **observe**, y_i, and what we **expect**, $\mathbb{E}(y_i)$, i.e.

$$\min\Big(y_i - \mathbb{E}(y_i)\Big) \Leftrightarrow \min_{\beta_0,\beta_1} \Big(y_i - (\beta_0 + \beta_1 x_i)\Big)$$

By minimizing this **expected loss** we obtain the "best" line that will be closest to the actual data points. However, if we try to minimize

$$y_i - (\beta_0 + \beta_1 x_i)$$

this quantity may be *positive* or *negative*. So, we choose to minimize the **Residual Sum of Squares (RSS)** (or the expected square loss) instead:

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

to obtain $(\hat{\beta}_0, \hat{\beta}_1)$.

Mathematical Derivation of the LS Estimators

1. Start by taking (partial) derivatives, with respect to β_0 and β_1

$$\begin{cases} \frac{\partial \text{RSS}}{\partial \beta_0} = 0 \\ \frac{\partial \text{RSS}}{\partial \beta_1} = 0 \end{cases} \Leftrightarrow \begin{cases} -2\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \\ -2\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)\, x_i = 0 \end{cases}$$

   Using the fact that $\sum_{i=1}^{n} 1 = n$

$$\begin{cases} \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \beta_0 - \beta_1 \sum_{i=1}^{n} x_i = 0 \\ \sum_{i=1}^{n} y_i x_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0 \end{cases}$$

2. Re-arrange the equations

$$\begin{cases} n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

3. Solve the $2\times$ 2 system with respect to β_0, β_1:

Start by solving the first equation with respect to β_0

$$\beta_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \beta_1 \frac{1}{n}\sum_{i=1}^{n} x_i$$

and then plugg-it into the second one:

$$\left(\frac{1}{n}\sum_{i=1}^{n} y_i - \beta_1 \frac{1}{n}\sum_{i=1}^{n} x_i\right)\sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

$$\Leftrightarrow \frac{1}{n}\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i - \beta_1 \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

$$\Leftrightarrow n\,\frac{1}{n}\sum_{i=1}^{n} y_i \frac{1}{n}\sum_{i=1}^{n} x_i - \beta_1 n\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

$$\Leftrightarrow n\bar{y}\bar{x} - n\beta_1\bar{x}^2 + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

This gives

$$\begin{cases} \beta_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \beta_1 \frac{1}{n}\sum_{i=1}^{n} x_i \\ \beta_1\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) = \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i y_i \end{cases}$$

which leads to the *LS estimators* when solving with respect to β_0 and β_1:

Least Squares Estimators

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

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

There are many ways that we can write the LS estimators, by re-arranging the terms in the equation above and using identities such as:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

$$\sum_{i=1}^{n}(x_i - \bar{x})\bar{x} = 0$$

A popular representation of the slope estimator is via the following quantities:

$$S_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y})$$
$$S_{xx} = \sum_i (x_i - \bar{x})^2$$
$$S_{yy} = \sum_i (y_i - \bar{y})^2$$

Alternative Representation of $\hat{\beta}_1$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

Recall the definition of the sample correlation

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

Using this, we can also express the *slope* of the SLR as

$$\hat{\beta}_1 = r_{xy} \sqrt{\frac{S_{yy}}{S_{xx}}}$$

Let us discuss now how we can obtain the LS estimators and the fitted line in R by re-visiting the University Admissions example:

University Admissions Example (Revisited)

In our previous example, the fitted regression line is obtained as follows:

```
admissions.lm = lm(gpa~entrance_score, admissions)
summary(admissions.lm)
```

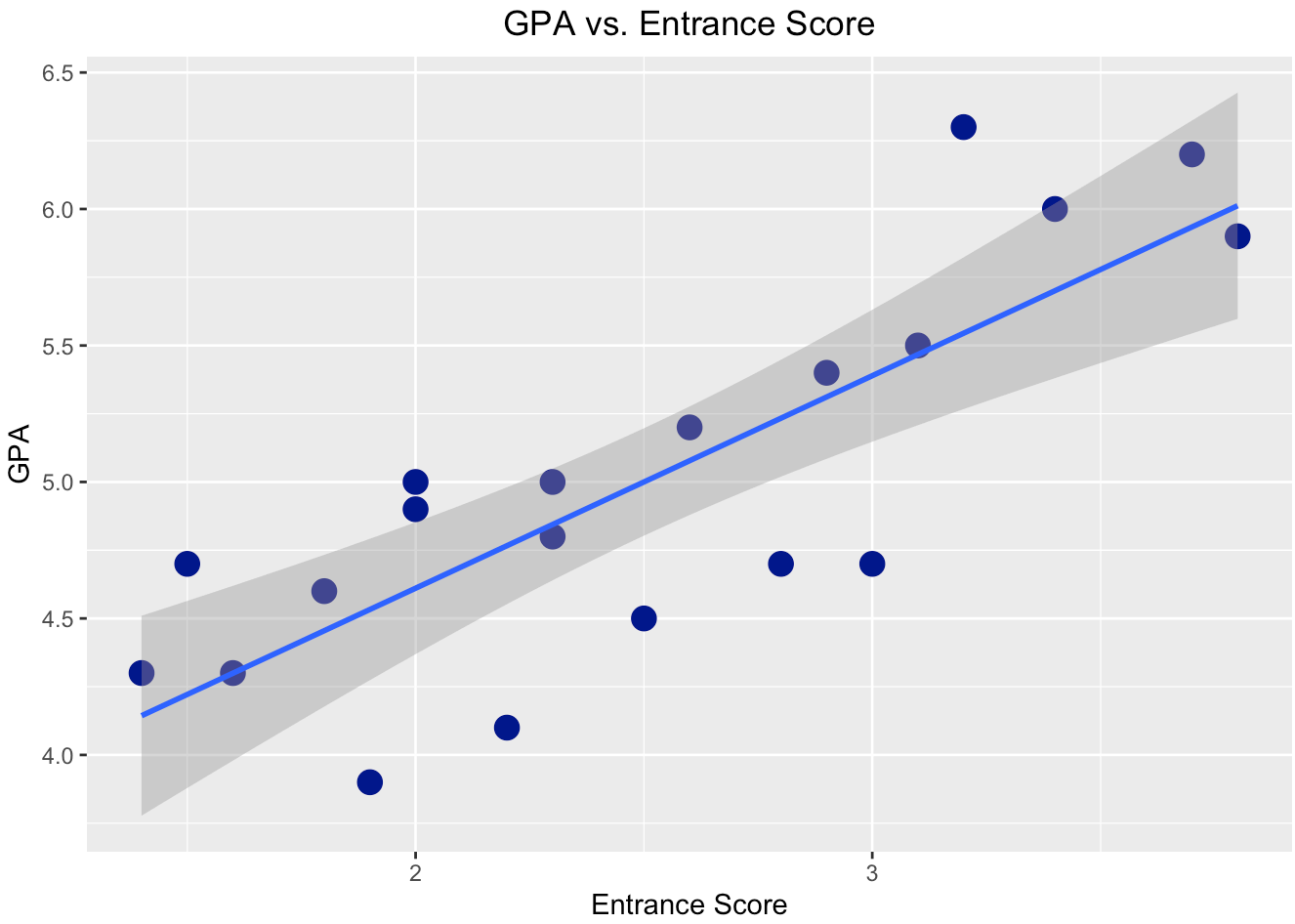
```
##
## Call:
## lm(formula = gpa ~ entrance_score, data = admissions)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6892 -0.2090  0.1054  0.2717  0.7551
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.0539     0.3467   8.809 6.05e-08 ***
## entrance_score   0.7785     0.1335   5.831 1.60e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4188 on 18 degrees of freedom
## Multiple R-squared:  0.6538, Adjusted R-squared:  0.6346
## F-statistic:    34 on 1 and 18 DF,  p-value: 1.597e-05
```

The fitted regression line is written as

$$(\text{GPA}) = 3.0539 + 0.7785 \cdot (\text{Entrance Score})$$

and can be plotted using the `ggplot2` library as follows:

```
library(ggplot2)
scatterplot = ggplot(admissions, aes(entrance_score, gpa)) +
  geom_point(size=4, color='darkblue') +
  labs(title="GPA vs. Entrance Score",y="GPA",x="Entrance Score") +
  theme(legend.position = "none") +
  theme(plot.title = element_text(hjust = 0.5))+
   geom_smooth(method=lm)
plot(scatterplot)
```



1.4.4 Sample Correlation & Linear Regression

In the previous discussion, we saw that there is a relation between the sample correlation and the estimated slope of the regression line. So, let's try understand the meaning of this relationship.

University Admissions Example (Revisited)

Suppose that you **only** have the following information available:

	Mean	Variance
Entrance Score	2.5	0.52
GPA	5	0.48

and that $\text{Corr}(\text{GPA}, \text{Entrance Score}) = 0.81$

*If you knew a student with entrance score **4.3**, could you **guess** their GPA?*

The *unit-free, location/scale invariant*[6] version of the GPA ($\mathbf{y}$) and the *unit-free, location/scale invariant* version of the Entrance Score ($\mathbf{x}$) have the following relationship

$$\frac{\mathbf{y} - \mu_y}{\sigma_y} \approx r_{xy} \frac{\mathbf{x} - \mu_x}{\sigma_x}$$

Write the *sample* equivalent of this expression

$$\frac{y - \bar{y}}{\sqrt{S_{yy}}} \approx r_{xy} \frac{x - \bar{x}}{\sqrt{S_{xx}}}$$

Re-arrange the terms to obtain

$$y \approx \underbrace{\left(\bar{y} - r_{xy}\sqrt{\frac{S_{yy}}{S_{xx}}}\bar{x}\right)}_{=\hat{\beta}_0} + \underbrace{\left(r_{xy}\sqrt{\frac{S_{yy}}{S_{xx}}}\right)}_{=\hat{\beta}_1} x$$

Therefore, the (best) guess for this student's GPA is

$$y \approx \left(5 - 0.81\sqrt{\frac{0.48}{0.52}}\right) + \left(0.81\sqrt{\frac{0.48}{0.52}}\right) 4.3 = 7.57$$

1.4.5 Fitted Values & Residuals

Fitted Values

Given $\hat{\beta}_0$, $\hat{\beta}_1$, the LS estimates of the regression coefficients, we call

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

the **fitted value** (or **predicted value**) at x_i, or the **prediction of** y_i.

Residuals

The i**th residual** is the difference between y_i (*observed value*) and its prediction (*fitted value*):

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

The Residuals satisfy a set of very useful properties:

1. $\sum_i r_i = 0$
2. $RSS = \sum_i r_i^2$ is a minimum
3. $\sum_i y_i = \sum_i \hat{y}_i$
4. $\sum_i x_i r_i = 0$
5. $\sum_i \hat{y}_i r_i = 0$
6. The regression line *always* goes through the point $(\bar{x}, \bar{y})$.

Proof of (1):

$$\begin{aligned}
\sum_i r_i &= \sum_i (y_i - \hat{y}_i) \\
&= \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \\
&= \sum_i y_i - \sum_i \hat{\beta}_0 - \hat{\beta}_1 \sum_i x_i \\
&= n\bar{y} - n\hat{\beta}_0 - \hat{\beta}_1 n\bar{x} \\
&= n\bar{y} - n(\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 n\bar{x} = 0
\end{aligned}$$

■

There is also another way to prove (1):

We can start with the derivative of RSS with respect to β_0, i.e.

$$\frac{\partial \text{RSS}}{\partial \beta_0} = 0 \Leftrightarrow \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$$

Now, recall what $\hat{\beta}_0$ and $\hat{\beta}_1$ are: the quantities that minimize RSS or in other words, the quantities that satisfy the equation

$$\frac{\partial RSS}{\partial \beta_0}\Big|_{\hat{\beta}_0, \hat{\beta}_1} = 0$$

or equivalently

$$\sum_{i=1}^{n}\left(y_i - \underbrace{\hat{\beta}_0 - \hat{\beta}_1 x_i}_{=\hat{y}_i}\right) = 0$$

$$\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0 \Leftrightarrow \sum_{i=1}^{n} r_i = 0.$$

■

Proof of (2):

This is straightforward since

$$RSS = \sum_i r_i^2 = \sum_i (y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i)$$

and $\hat{\beta}_0$, $\hat{\beta}_0$ are the ones that minimize $\sum_i (y_i - \beta_0 + \beta_1 x_i)^2$.

■

Estimator of σ^2

The **error variance** is estimated by

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_i r_i^2$$

and the degrees of freedom (df) of the residuals are

$$(\text{sample size}) - (\text{no. of parameters}) = n - 2$$

1.4.6 Goodness of Fit: R-square

The **Total Variation** in the response y is measured by the *Total Sum of Squares (TSS)*, can be decomposed as follows:

$$\begin{aligned}\sum_i (y_i - \bar{y})^2 &= \sum_i (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 = \sum_i (r_i + \hat{y}_i - \bar{y})^2 \\ &= \sum_i r_i^2 + \sum_i (\hat{y}_i - \bar{y})^2 \\ TSS &= RSS + FSS\end{aligned}$$

the *Residual Sum of Squares (RSS)* and the *Fitted value Sum of Squares (FSS)*.

Remark:

In the calculations above, the cross-product term vanishes due to orthogonality:

$$\sum_i r_i(\hat{y}_i - \bar{y}) = \hat{\beta}_0 \sum_i r_i + \hat{\beta}_1 \sum_i r_i x_i - \bar{y}\sum_i r_i = 0$$

Interpreting the TSS decomposition:

- TSS measures the variation in y_i, or the uncertainty in predicting $\mathbf{y}$, when no account of the predictor variable $\mathbf{x}$ is taken. Therefore, TSS is a measure of the uncertainty in predicting $\mathbf{y}$ when $\mathbf{x}$ is *not* considered.
- RSS measures the variation in y_i when a regression model utilizing the predictor variable $\mathbf{x}$ is fitted.

A natural way of the **effect of $\mathbf{x}$ in reducing the variation in $\mathbf{y}$**, i.e. in reducing the uncertainty in predicting $\mathbf{y}$, is to express the reduction in variation as a *proportion of the total variation*:

Coefficient of Determination (R^2)

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = \frac{FSS}{TSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

R^2 is interpreted as the proportionate reduction of total variation in $\mathbf{y}$ associated with the use of the prediction variable $\mathbf{x}$.

By definition, we have that

$$0 \leq R^2 \leq 1$$

The larger R^2 is, the more the total variation of $\mathbf{y}$ is reduced by introducing the independent variable $\mathbf{x}$. The closer R^2 is to 1, the greater the degree of ***linear*** association between $\mathbf{x}$ and $\mathbf{y}$.

Recall the sample correlation coefficient r_{xy}. We can show that

$$r_{xy} = \pm\sqrt{R^2},$$

where the *sign* is the sign of the slope.

R^2 Misconceptions

1. *"A high R^2 indicates that useful predictions can be made."* This is not necessarily correct!

A Counterexample

A company manufactures refrigeration equipment as well as other replacement parts which are typically being produced in lots of varying sizes. The company needs to determine the relationship between `lot size` ($\mathbf{x}$) and `labor hours` ($\mathbf{y}$) required to producing a lot, as part of a cost improvement process. The data can be found here.

When we fit a SLR model to the data, we compute R^2 to be

$$R^2 = 0.822$$

which implies that the variation in `labor hours` is reduced by 82.2% when `lot size` is

considered. However, when we compute a *90% prediction interval for the next lot consisting of 100 units*, the interval we obtain is

$$(332,\ 507)$$

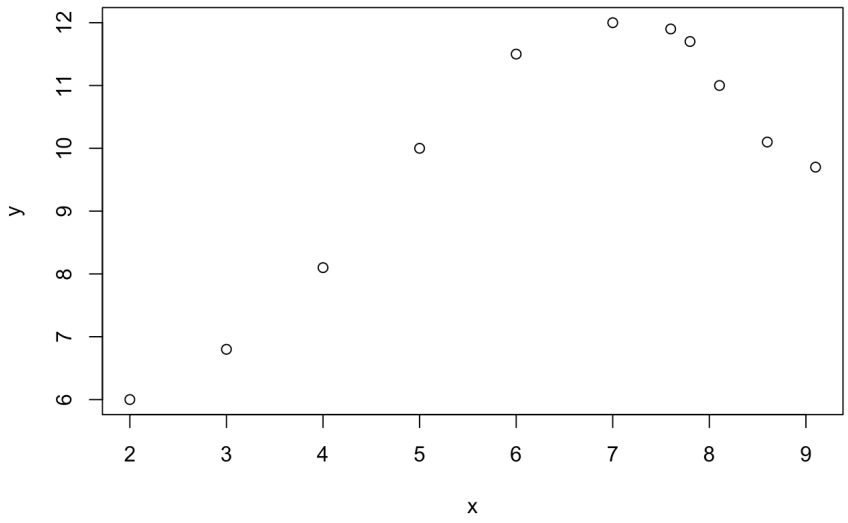
which is *not precise enough* to permit management to schedule workers effectively.

> This situation arises, because R^2 measures only a **relative** reduction of TSS and provides *no information* about the *absolute precision* for estimating a mean response or predicting a new observation.

2. *"A high R^2 indicates that the estimated regression line is a good fit."* This is not necessarily correct!

A Counterexample

Consider the following data set plotted below:



In this example, the R^2 is

```
## [1] 0.616264
```

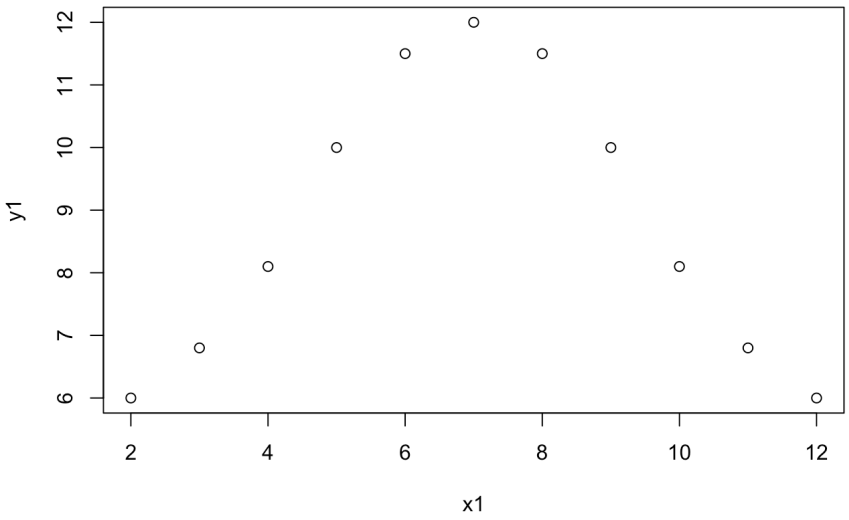
Here, the R^2 is quite high, while from the scatterplot we can see that a *straight* line is not a good fit.

> This situation arises because R^2 measures the degreeof **linear** association between X and Y, whereas the actual regression relation may be curvilinear.

3. *"An R^2 near zero indicates that $\mathbf{x}$ and $\mathbf{y}$ are not related."* This is not necessarily correct!

A Counterexample

Consider the following data plotted below:



In this example, the R^2 is

```
## [1] 3.01322e-31
```

which is extremely low. However, $\mathbf{x}$ and $\mathbf{y}$ are strongly related, but the relation here is not linear, but curvilinear.

> Same as before, this situation arises because R^2 measures the degree of **linear** association between $\mathbf{x}$ and $\mathbf{y}$, whereas the actual regression relation may not be linear.

1.4.7 Affine Transformations

Definition

This is a term that comes from Euclidean geometry: An **affine transformation** is a geometric transformation that preserves lines and parallelism.

- What does this mean?
  This says that a set of parallel lines will remain parallel after an *affine* transformation.
- Ok.. How does this relate to Regression?
  Well, if we have real numbers (which is what we have in our case) then functions $f : \mathbb{R} \to \mathbb{R}$ of the form, for example,

$$f(x) = ax + b, \quad \text{where } a,\ b \in \mathbb{R}$$

  are affine transformations of the real line.
- Is this only relevant in Geometry and the real numbers?
  No. This is a general concept: we define an affine transformation as an automorphism of an affine space (the Euclidean space is a specific affine space) onto itself while preserving both the dimension of any affine subspace and the ratios of lengths of parallel line segments. A *linear transformation* is only a specific type of affine transformations. Other types of affine transformations include rotation, similarity, reflection.

Affine Transformations on Wikipedia

Suppose we have a Simple Linear Regression model of $\mathbf{y}$ on $\mathbf{x}$, i.e.

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

What will happen to the *LS estimates* and the R^2 if we:

a. rescale y_i by $\tilde{y}_i = ay_i + b$ and then regress $\tilde{y}_i$ on x_i?

b. rescale x_i by $\tilde{x}_i = ax_i + b$ and then regress y_i on $\tilde{x}_i$?

c. regress x on y instead?

Proof of (a):

The transformed model is written as

$$\tilde{y}_i = b_0 + b_1 x_i + \varepsilon_i,$$

where $\tilde{y}_i = ay_i + b$.

To understand how the LS estimators change after an affine transformation, we will quantify the relationship between the transformed slope and intercept estimators, $\hat{b}_1$, $\hat{b}_0$, and the original ones $\hat{\beta}_1$, $\hat{\beta}_0$ respectively.

Starting with the transformed model and deriving the LS estimators, the estimated LS slope coefficient is:

$$\hat{b}_1 = r_{X\tilde{Y}} \sqrt{\frac{S_{\tilde{Y}\tilde{Y}}}{S_{XX}}}.$$

Observe that

(i) S_{XX} only depends on X, so after the affine transformation, it remains the same.

(ii) $S_{\tilde{Y}\tilde{Y}}$ depends on the new y, so we have

$$\begin{aligned} S_{\tilde{Y}\tilde{Y}} &= \sum_i (\tilde{y}_i - \bar{\tilde{y}})^2 \\ &= \sum_i \Big((ay_i + b) - (a\bar{y} + b) \Big)^2 \\ &= \sum_i (ay_i - a\bar{y})^2 \\ &= a^2 \sum_i (y_i - \bar{y})^2 = a^2 \; S_{YY} \end{aligned}$$

(iii) $r_{X\tilde{Y}}$ depends on the new y, so it computes as

$$\begin{aligned} r_{X\tilde{Y}} &= \frac{\sum_i (x_i - \bar{x})(\tilde{y}_i - \bar{\tilde{y}})}{\sqrt{S_{XX}}\sqrt{S_{\tilde{Y}\tilde{Y}}}} \\ &= \frac{\sum_i (x_i - \bar{x})(ay_i + b - (a\bar{y} + b))}{\sqrt{S_{XX}}\sqrt{a^2 S_{YY}}} \\ &= \frac{a\ \sum_i (x_i - \bar{x})(y_i - \bar{y})}{a\sqrt{S_{XX}}\sqrt{S_{YY}}} = r_{XY} \end{aligned}$$

Combining *(i)*, *(ii)*, *(iii)*, the new slope becomes

$$\hat{b}_1 = r_{X\tilde{Y}}\sqrt{\frac{S_{\tilde{Y}\tilde{Y}}}{S_{XX}}} = r_{XY}\sqrt{\frac{a^2 S_{YY}}{S_{XX}}} = a\ r_{XY}\sqrt{\frac{S_{YY}}{S_{XX}}} = a\ \hat{\beta}_1$$

Starting again with the transformed model the estimated LS intercept coefficient is:

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

where here $\bar{\tilde{y}}$ denotes the sample mean of the transformed $\tilde{y}$. Plugging-in the expressions for $\tilde{y}$ and $\hat{b}_1$ from before, and re-arranging the terms, we have

$$\hat{b}_0 = \underbrace{(a\bar{y} + b)}_{=\bar{\tilde{y}}} - \underbrace{a\hat{\beta}_1}_{=\hat{b}_1}\bar{x} = a\ \underbrace{(\bar{y} - \hat{\beta}_1\bar{x})}_{=\hat{\beta}_0} + b = a\hat{\beta}_0 + b$$

Let us find out what happens to the $\tilde{R}^2$ of the transformed model and the R^2 of the original one.

For the transformed model, $\tilde{R}^2$ is computed as

$$\tilde{R}^2 = r^2_{X\tilde{Y}} = r^2_{XY} = R^2$$

which means that the R-square remains unchanged.

■

1.4.8 Regression Through the Origin

The Birds Eggs Study[7]

A study is conducted to understand the relationship between the *height of a bird's egg* and its *weight.* Based on the data collected, the following regression line was obtained:

$$\text{Height } = -1.774 + 1.444 \text{ Width}$$

The questions we want to ask here are the following:

a. Is the intercept $\hat{\beta}_0 = -1.774$ meaningful here?

b. Can we fit a model without an intercept? What does it change?

The answer to (a) is **No!** The intercept has no meaning here, since there is no notion of an egg having negative height or zero weight.

The answer to (b) is **Yes**. Yes, we can choose to start with a model that has *no intercept* in a situation where we know ahead of time that the intercept is meaningless. However, this has some implications on the estimators and the analysis of the data.

Regression through the Origin Model

The **no-intercept** model is defined as

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

where ε_i satisfies the same assumptions as before.

If we calculate the LS estimator for β_1 from scratch, following the same procedure as before, we get

$$\hat{\beta}_1 = \frac{\sum_i x_i y_i}{\sum_i x_i^2}$$

which is, of course, different than the estimator of the model with intercept.

However, it is not only $\hat{\beta}_1$ that is affected.

The ordinary definition of R-square is *no longer meaningful*. If we take the R^2 formula from before, a negative R^2 is possible, since RSS may be larger than TSS. Conceptually, the ordinary R^2 measures the effect of X **after removing the effect of the intercept** by *centering* both y_i's and $\hat{y}_i$'s. For regression models with no intercept, we should *avoid centering* the ys when computing R^2.

This implies that the definition of R^2 should **change** to account for the missing intercept. So, let's start from the beginning by breaking down the *total variation* in a model with no intercept:

$$\sum_i y_i^2 = \sum_i (y_i - \hat{y}_i + \hat{y}_i)^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i \hat{y}_i{}^2$$

Then, the coefficient of linear determination is now defined as

$$\tilde{R}^2 = 1 - \frac{RSS}{\sum_i y_i^2}$$

which is different than the usual R^2. Note, that this is **not** the adjusted R^2.