# Experimental Designs

author: Maria Paniw date: February 8, 2016; UCA

## A quick overview of the type of analysis needed depending on your design

| Dependent variable (response Y) | Independent variable (predictor X) | |
|---|---|---|
| | **Continuous** | **Categorical** |
| **Continuous** | Regression | ANOVA |
| **Categorical** | Generalized linear models | Contingency tables |

We will go over four types of analyses:

- Regression (normal errors)
- ANOVA (normal errors)
- ANCOVA (normal errors)
- Regression/ANOVA/ANCOVA (non-normal errors)

## Assumptions of simple regression, ANOVA, and ANCOVA

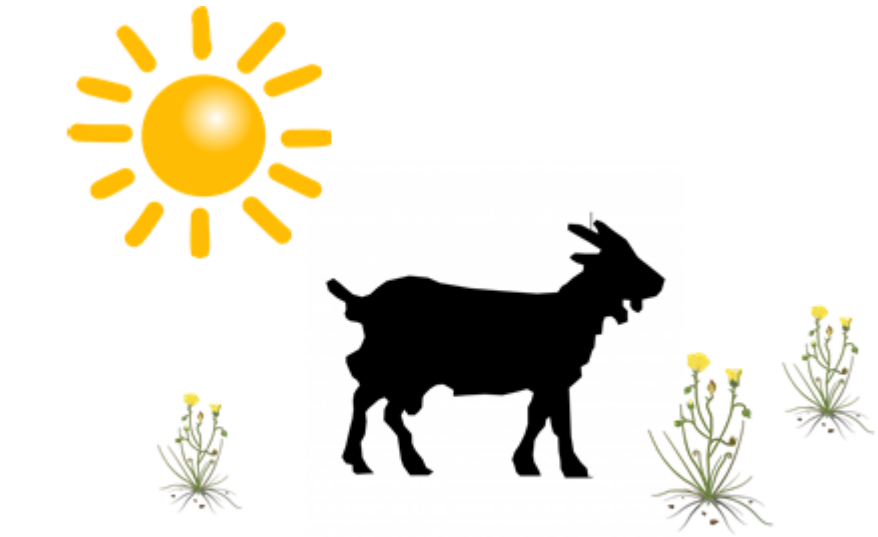
Whenever we do simple univariate statistics, we make four key assumptions:

- **Independence** of sampling units: For example, I can measure pollinator activity in 100 different plants. But if I choose my plants too close to each other, the same, *hyperactive* pollinator may go to several neighboring plants. Then, the measurements I make on the negihboring plants are not independent from one another:

Same pollinators visit my sampling units. Independence may be compromised.

- **Random** arrangement of sampling units and treatments: For example, I want to measure how grazer activity affects plant growth. I create a treatment *grazing* with two levels, *grazers allowed* and *grazers excluded*. Fo far so good. But what I fail to notice is that the sites in which grazers are allowed are in a climatically distinct zone from the ones where grazers are excluded. Then climate is a **confounding** variable and may pose a **huge problem**.

The effect of grazing and and climate are confounded:



- **Linearity** - the relationship between response and predictor is linear.
- **Normality** - the residuals of the model describing the relashionship between predictor and response and normally distributed
- **Homoscedasticity** - the variance of the residuals is constant

The assumption of homoscedasticity is typically the key aspect of simple regression and ANOVA designs.

## Steps to take in simple statistical analyses

Source: Luis Cayuela,EcoLab, Universidad de Granada, lcayuela@ugr.es.

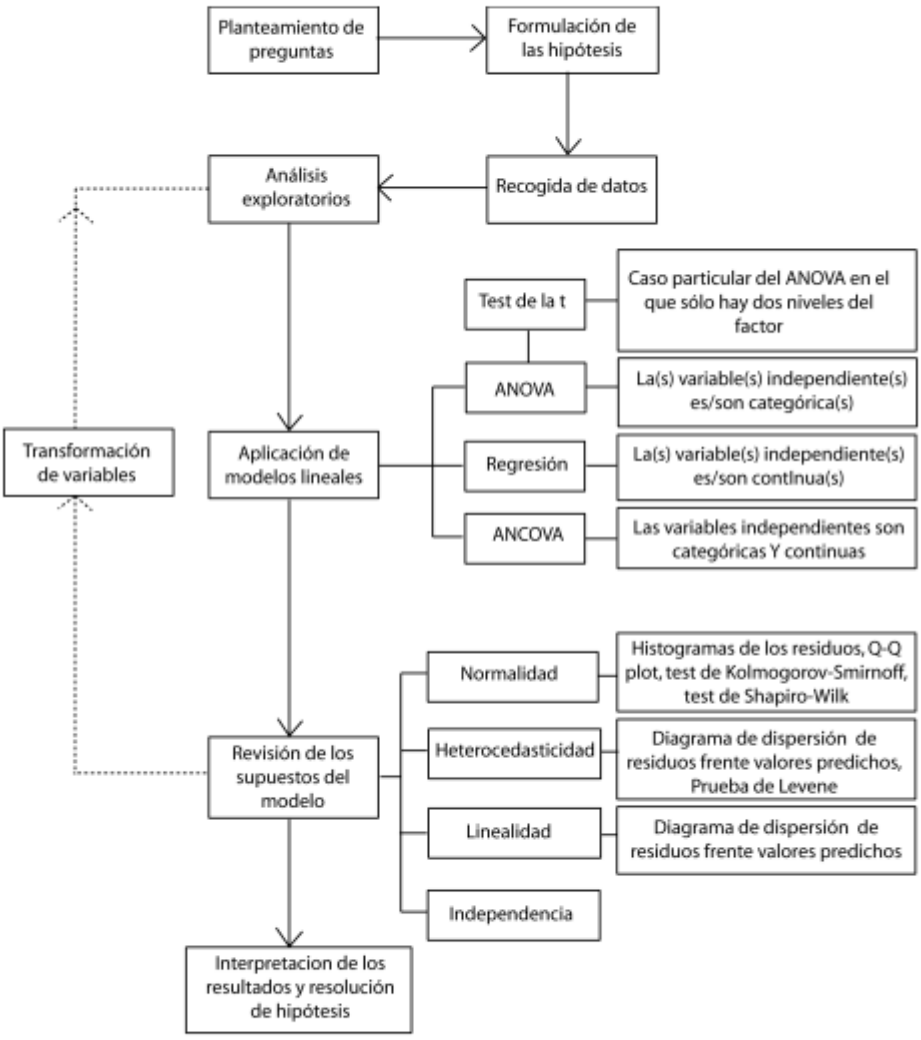



Figura 1: Esquema conceptual de los pasos que deben seguirse a la hora de ajustar un modelo lineal univariante.

## Regression

Simple regression designs are typically based on observational studies. Here, you typically analyze relationships between continuous variables, i.e., variables that can go from $-\infty$ to $\infty$

Basically, regression analysis describes the relationship between the predictor ($x$-axis) and the response ($y$-axis)

R uses the function `lm()` for linear regression analysis

## Defining a straight line

Regression, as any kind of statistical analysis, begins with a hypothesis:

> Hypothesis: Variable X drives the response Y.

Once we have our hypothesis, we must express it mathematically.

$$Y = f(X)$$

What kind of $f()$ ?

In linear regression: Y is a linear function of X, i.e.,

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

## Defining parameters

Here, $\beta_0$ is the **intercept** and $\beta_1$ is the **slope**

These are the two **parameters** of our linear model.

If the parameters are known, and our data can be entirely described by the linear relationship between X and Y, we have a deterministic function:

## Deterministic relationship between X and Y

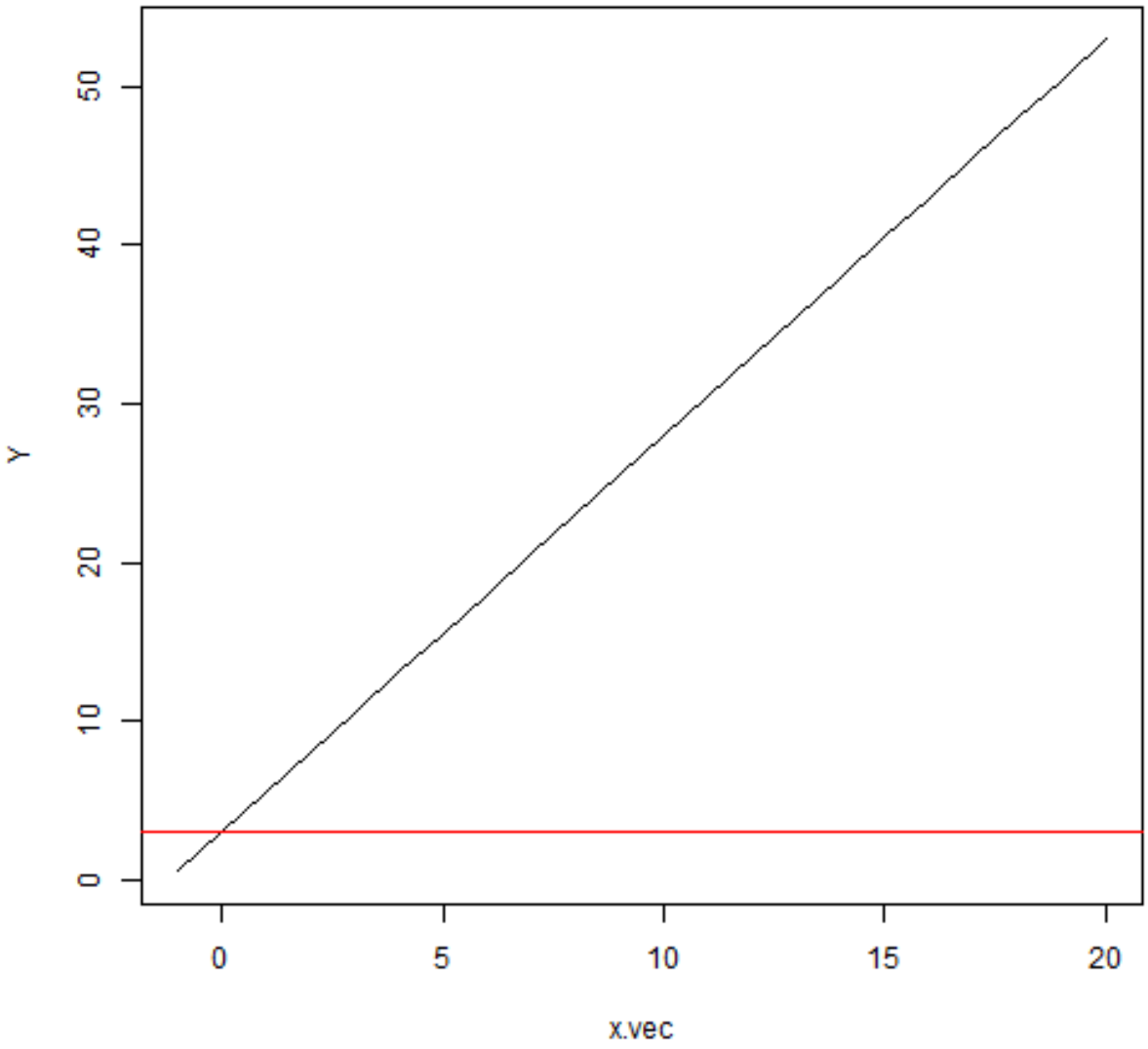
```
determ.lf = function(x) 3 + 2.5*x

x.vec = c(c(-1:20))
```

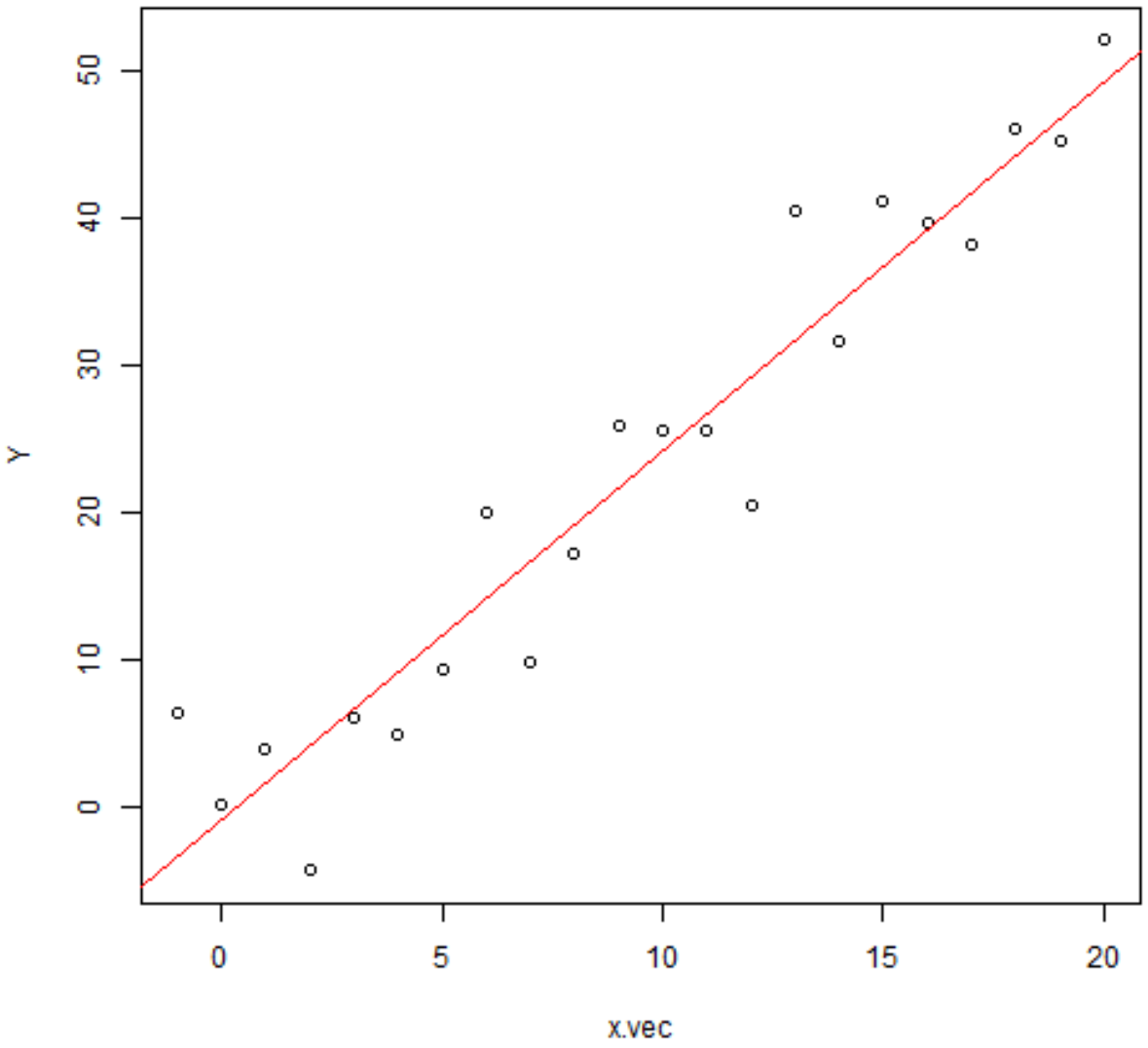
## Plot the relationship

```
plot(x=x.vec,y=determ.lf(x=x.vec),type="l", ylab="Y")

abline(h=3,col="red")
```


Y
0
10
20
30
40
50
0
5
10
15
20
x.vec

## Statistical relationship between X and Y




## Statistical relationship between X and Y

If we cannot predict our data excatly by the linear regression because we have variation in our data that is not accounted for, we turn to a statistical model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Each observation $i$ in our data (*rows*), is associated with a value for X and Y, both measured at the same replicate!

## The betas and the epsilon

$\beta_0$ and $\beta_1$ are constants

On the other hand, $\epsilon \sim \mathcal{N}(0, \sigma^2)$

$\sigma^2$ is the variance, which is 0 if the points lie perfectly on the line.

## So what's the deal with normality?

When you do linear regression, you are always asked:

> Is Y normally distributed?

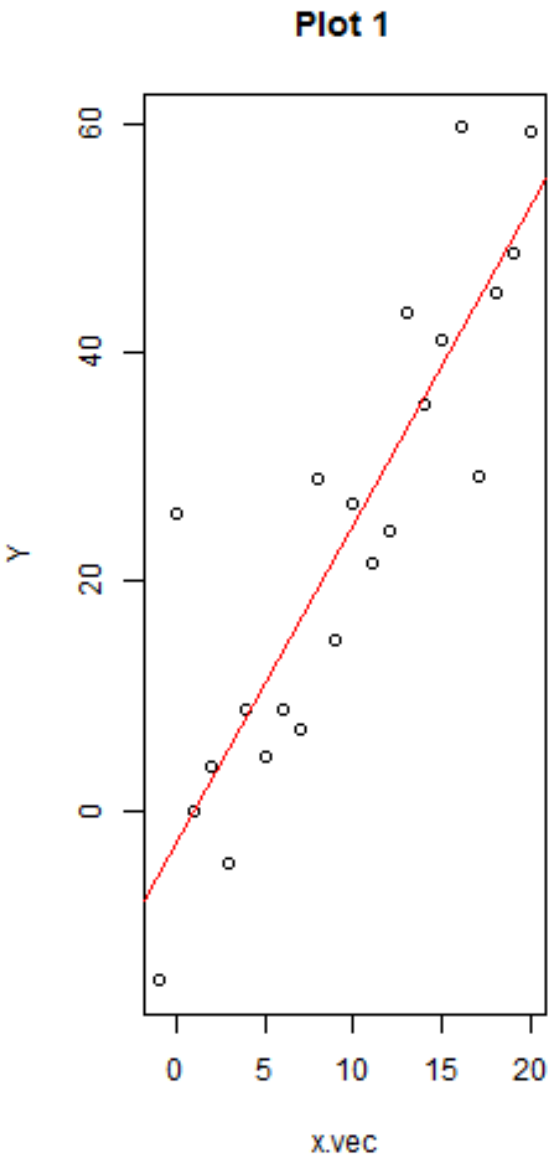
This is because, shifting the equation for linear regression a bit:

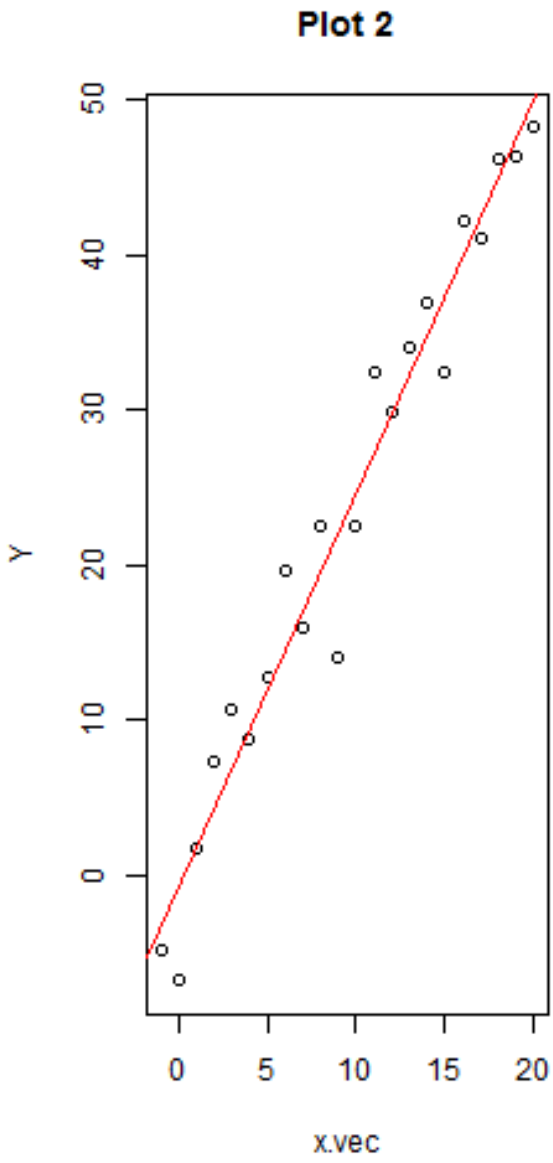
$$Y \sim \mathcal{N}(\beta_0 + \beta_1 X, \sigma^2)$$

## The points make the variance

The more spread out the points, the larger $\sigma^2$

In plot 1, $\sigma^2$ is larger than in plot 2




## If points have a noise component, how do we fit a line?

The key concept of linear regression:

The residual sum of squares!

$$RSS = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

where $Y_i$ is the observed repsonse at point $i$ and $\hat{Y}_i$ is the predicted value from the regression equation.

The **best fit** regression line minimizes RSS.

## How doe we estimate the two parameters in the model?

The sum of squares of a variable X ($SS_X$) measures the squared deviation of each observation $X_i$ from the mean of all the observations $\bar{X}$ :

$$SS_X = \sum_{i=1}^{n}(X_i - \bar{X})^2$$

Dividing this SS by ($n$-1), where $n$ is our sample size, gives us the sample variance:

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

## In a linear regression, we have at least two variables of course

Thus, we have to measure the sample covariance!

In this case, we have a **sum of cross products**:

$$SS_{XY} = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

Which, divided by ($n$-1), gives us the sample **covariance**:

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

## The slope and intercept

$\beta_0$ and $\beta_1$ are the estimated as:

$$\hat{\beta}_1 = \frac{s_{xy}}{s_X^2} = \frac{SS_{XY}}{SS_X}$$

and

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$$

## The error term

The last thing we need to estimate is $\epsilon_i$.

Remember that $\epsilon = \mathcal{N}(0, \sigma^2)$

So, we need an estimate of $\sigma^2$ to get $\epsilon_i$:

$$\hat{\sigma}^2 = \frac{RSS}{n-2} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n}\left[Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)\right]^2}{n-2}$$

## SS as fundamental principle of parametric analyses

In parametric analysis, we want to partition the SS (sum of squares) into different components.

Conceptually, we are looking at our response, $Y$, and $SS_Y = \sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2$ is the total variance that we are trying to partition into its components:

The $RSS$ is the random component. The remaining variation, $SS_{reg}$, is the regression relatioship $Y_i = \beta_0 + \beta_1 X_i$

So, the total variance is: $SS_Y = SS_{reg} + RSS$

## Is the relationship between X and statistically significant?

$H_0 : Y_i = \beta_0 + \epsilon_i$

$H_a : Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

We are testing the significance of the slope, as this is the only paramter that differs!

========== You can consider single-factor regression (only one predictor X) as an ANOVA table (Gotelli & Ellison. 2008. A Primer of Ecological Statistics):

**TABLE 9.1 Complete ANOVA table for single factor linear regression**

| Source | Degrees of freedom (df) | Sum of squares (*SS*) | Mean square (*MS*) | Expected mean square | F-ratio | *P*-value |
|---|---|---|---|---|---|---|
| Regression | 1 | $SS_{reg} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$ | $\frac{SS_{reg}}{1}$ | $\sigma^2 + \beta_1^2 \sum_{i=1}^{n} X^2$ | $\frac{SS_{reg}/1}{RSS/(n-2)}$ | Tail of the F distribution with $1, n-2$ degrees of freedom |
| Residual | $n-2$ | $RSS = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ | $\frac{RSS}{(n-2)}$ | $\sigma^2$ | | |
| Total | $n-1$ | $SS_Y = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ | $\frac{SS_Y}{(n-1)}$ | $\sigma_Y^2$ | | |

The first column gives the source of variation in the data. The second column gives the degrees of freedom (df) associated with each component. For a simple linear regression, there is only 1 degree of freedom associated with the regression, and $(n-2)$ associated with the residual. The degrees of freedom total to $(n-1)$ because 1 degree of freedom is always used to estimate the grand mean of the data. The single factor regression model partitions the total variation into a component explained by the regression and a remaining ("unexplained") residual. The sums of squares are calculated using the observed $Y$ values ($Y_i$), the mean of the $Y$ values ($\bar{Y}$), and the $Y$ values predicted by the linear regression model ($\hat{Y}_i$). The expected mean squares are used to construct an F-ratio to test the null hypothesis that the variation associated with the slope term ($\beta_1$) equals 0.0. The $P$-value for this test is taken from a standard table of F-values with 1 degree of freedom for the numerator and $(n-2)$ degrees of freedom for the denominator. These basic elements (source, df, sum of squares, mean squares, expected mean square, F-ratio, and $P$-value) are common to all ANOVA tables in regression and analysis of variance (Chapter 10).

## What about multiple regression?

If we have more than one predictor, the RSS can be expressed in the following way:

$$\sum_{i=1}^{n}\left\{Y_i - \left(\beta_0 + \sum_{j=1}^{p}\beta_j x_{i,j}\right)\right\}^2$$

We have now a vector (length = $j$) of $\beta_j$ depicting the slopes associated with each predixtor $X_j$

Things are a bit more complicated now, and we are getting to the core of linear models - creating the **model matrix** and doing some matrix algebra to get our $\beta$.