Simple Linear Regression

Regression models are used to study the relationship of a *response* variable and one or more *predictors*.

The response is also called the *dependent variable*, and the predictors are called *independent variables*.

In the *simple linear regression* model, there is only one predictor.
Empirical Models

Example: Oxygen and Hydrocarbon levels

Production of oxygen

Response: \( Y \), purity of the produced oxygen (%);

Predictor: \( x \), level of hydrocarbons in part of the system (%).

In R

```r
oxygen <- read.csv("Data/Table-11-01.csv")
# scatter plot:
plot(Purity ~ HC, oxygen)
```
The relationship between $x$ and $y$ is roughly linear, so we assume that

$$ Y = \beta_0 + \beta_1 x + \epsilon $$

for some coefficients $\beta_0$ (the intercept) and $\beta_1$ (the slope), where $\epsilon$ is a random noise term (or random error).

The noise term $\epsilon$ is needed in the model, because without it the data points would have to fall exactly along a straight line, which they don’t.

This is an empirical model, rather than a mechanistic model, because we have no physical or chemical mechanism to justify it.
What can we say about $\beta_0$ and $\beta_1$?

By eye, the slope $\beta_1$ appears to be around \( \frac{96 - 90}{1.5 - 1.0} = 12 \), and $\beta_0$ appears to be around $90 - 1.0 \times \beta_1 = 78$.

In R:
```
abline(a = 78, b = 12)
```

This line over-predicts most of the data points with lower HC percentages, so perhaps it can be improved.
For any candidate values $b_0$ and $b_1$, the predicted value for $x = x_i$ is $b_0 + b_1 x_i$, and the observed value is $y_i$, $i = 1, 2, \ldots, n$.

The residual is

$$e_i = \text{observed} - \text{predicted} = y_i - (b_0 + b_1 x_i), \quad i = 1, 2, \ldots, n.$$  

The candidate values $b_0$ and $b_1$ are good if the residuals are generally small, and in particular if

$$L(b_0, b_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2$$

is small.
Least Squares Estimates

The *best* candidates (in this sense) are the values of $b_0$ and $b_1$ that give the lowest value of $L(b_0, b_1)$.

They are the *least squares estimates* $\hat{\beta}_0$ and $\hat{\beta}_1$ and can be shown to be

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$
In R, the function `lm()` calculates least squares estimates:

```r
summary(lm(Purity ~ HC, oxygen))
```

The line labeled (Intercept) shows that $\hat{\beta}_0 = 74.283$, and the line labeled HC shows that $\hat{\beta}_1 = 14.947$.

This line does indeed fit better than our initial candidate:

```r
abline(a = 74.283, b = 14.947, col = "red")
# check the sums of squared residuals:
L <- function(b0, b1) with(oxygen,
    sum((Purity - (b0 + b1 * HC))^2))
L(b0 = 78, b1 = 12)  # 27.8985
L(b0 = 74.283, b1 = 14.947)  # 21.24983
```
Estimating $\sigma$

The regression equation also involves a third parameter $\sigma$, the standard deviation of the noise term $\epsilon$.

The least squares residuals are $e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$ so the residual sum of squares is

$$SS_E = \sum_{i=1}^{n} e_i^2.$$

Because two parameters were estimated in finding the residuals, the residual degrees of freedom are $n - 2$, and the estimate of $\sigma^2$ is

$$\hat{\sigma}^2 = MS_E = \frac{SS_E}{n - 2}.$$
Other Estimates

The sum of squared residuals is not the only criterion that could be used to measure the overall size of the residuals.

One alternative is

\[ L_1(b_0, b_1) = \sum_{i=1}^{n} |e_i| = \sum_{i=1}^{n} |y_i - (b_0 + b_1x_i)|. \]

Estimates that minimize \( L_1(b_0, b_1) \) have no closed form representation, but may be found by linear programming methods.

They are used occasionally, but least squares estimates are generally preferred.
Sampling Variability

The 20 observations in the oxygen data set are only one sample, and if other sets of 20 observations were made, the values would be at least a little different.

The least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ would therefore also vary from sample to sample.

We assume that there are *true parameter values* $\beta_0$ and $\beta_1$ and that if we carried out many experiments at a given level $x$, the responses $Y$ would average out to

$$\beta_0 + \beta_1 x.$$
The *standard error* measures how far an estimate might typically deviate from the true parameter value.

*Estimated* standard errors may be calculated for $\hat{\beta}_0$ and $\hat{\beta}_1$, and are shown in the R output.

They are used to set up confidence intervals:

$$\hat{\beta}_1 \pm t_{\alpha/2,\nu} \times \text{estimated standard error}$$

is a $100(1 - \alpha)$ confidence interval for $\beta_1$, where $\nu = n - 2$ is the degrees of freedom for Residuals.
We can also test hypotheses: for example, there is no relationship between $x$ and $y$ if $\beta_1 = 0$.

To test $H_0 : \beta_1 = 0$, we use

$$t_{\text{obs}} = \frac{\hat{\beta}_1}{\text{estimated standard error}}$$

and find the $P$-value as usual, as the probability of finding as large a value of $|t_{\text{obs}}|$ if $H_0$ were true.
The null hypothesis is not always $\beta_1 = 0$. For instance, previous operations might have suggested that $\beta_1 = 12$.

To test $H_0 : \beta_1 = 12$, we use

$$t_{\text{obs}} = \frac{\hat{\beta}_1 - 12}{\text{estimated standard error}}$$

and in this case $|t_{\text{obs}}| = 2.238$, and the probability that $|T| \geq 2.238$ is 0.038.

So we would reject this hypothesis at the $\alpha = 0.05$ level, but not at the 0.01 level.