ESTIMATING A REGRESSION LINE
This course is about **REGRESSION ANALYSIS**:

- Constructing quantitative descriptions of the statistical association between $y$ (response variable) and $x$ (predictor, or explanatory variable) on the sample data.
- Introducing models, to interpret estimates and inferences on the parameters of these descriptions in relation to the underlying population.

**MULTIPLE regression**, when we consider more than one predictor variable.
Fitting a line through bi-variate sample data (as a descriptor)

Least squares fit: find the line that minimizes the sum of squared (vertical) distances from the sample points.

\[ y = \beta_0 + \beta_1 x \quad \text{generic equation of line} \]

\[
\min_{\beta_0, \beta_1} \left\{ \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 \right\} \quad \text{obj fct}
\]

Normal equations (derivatives of obj fct)

\[
\begin{align*}
\sum_{i=1}^{n} y_i &= n \beta_0 + \beta_1 \sum_{i=1}^{n} x_i \\
\sum_{i=1}^{n} x_i y_i &= \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2
\end{align*}
\]

Solution (unique)

\[
\begin{align*}
b_1 &= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \\
b_0 &= \bar{y} - b_1 \bar{x}
\end{align*}
\]
\[ \hat{y}_i = b_0 + b_1 x_i \quad \text{fitted value} \]
\[ e_i = y_i - \hat{y}_i \quad \text{residual} \]

For sample point \( i = 1 \ldots n \)

**Geometric properties** of the least square line:

\[ \sum_{i=1}^{n} e_i = 0 \]

\[ \sum_{i=1}^{n} e_i^2 = \min \]

\[ \sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} y_i \]

\[ \sum_{i=1}^{n} x_i e_i = 0 \]

\[ \sum_{i=1}^{n} \hat{y}_i e_i = 0 \]

\[ \bar{y} = b_0 + b_1 \bar{x} \]

F. Chiaromonte
Simple linear regression **MODEL:**

Assume the sample data is generated as

\[ y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \]

- \( x_i, i = 1...n \) fixed (or condition on)
- \( \varepsilon_i, i = 1...n \) random errors s.t.
  - \( E(\varepsilon_i) = 0, \forall i \) no systematic component
  - \( \text{var}(\varepsilon_i) = \sigma^2, \forall i \) constant variance
  - \( \text{cor}(\varepsilon_i, \varepsilon_j) = 0, \forall i \neq j \) uncorrelated

The values of \( y \) (given various values of \( x \)) scatter about a line, with constant variance and no correlations among the departures from the line carried by different observations…

…*quite simplistic, but very useful in many applications!*

Note: distribution of errors is unspecified for now.
If we assumed a bell-shaped distribution for the errors (which we will do later!) here is how the “population” picture would look like:

\[ E(y_i) = \beta_0 + \beta_1 x_i \]
\[ \text{var}(y_i) = \sigma^2 \]
\[ \text{cor}(y_i, y_j) = 0 \]

Interpretation of parameters:

\( \beta_1 = \text{slope} \); change in the mean of \( y \) when \( x \) changes by one unit.
\( \beta_0 = \text{intercept} \); if \( x = 0 \) is a meaningful value, mean of \( y \) when \( x = 0 \).
\( \sigma^2 = \text{error variance} \); scatter of \( y \) about the regression line.
Under the assumptions of our simple linear regression model, the slope and intercept of the least square line are (point) estimates of the population slope and intercept, with the following very important properties:

**GAUSS-MARKOV THEOREM**: Under the conditions of the simple linear regression model:

- $b_1$ and $b_0$ are unbiased estimates for $\beta_1$ and $\beta_0$

$$E(b_1) = \beta_1 \quad E(b_0) = \beta_0$$

- they are the most accurate (smallest MSE i.e. variance) among all unbiased estimates that can be computed as linear functions of the response values.

**Linearity**:

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \sum_{i=1}^{n} k_i y_i$$

$$b_0 = \bar{y} - b_1 \bar{x} = \sum_{i=1}^{n} \frac{1}{n} y_i - \sum_{i=1}^{n} k_i \bar{y} = \sum_{i=1}^{n} \tilde{k}_i y_i$$
Point estimation of error variance

\[
\hat{s}^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2
\]

SSE (error sum of squares)  
MSE for the regression line  
dof of SSE (constraints = two parameters of line)

Unbiased for \( \sigma^2 \): \( E(\hat{s}^2) = \sigma^2 \)

Point estimation of mean response

\[ \hat{E}(y \text{ at } x) = \hat{y}(x) = b_0 + b_1 x \]
**Example**: Simple linear regression for $y = \log$ length ratio between human and chicken DNA, on $x = \log$ large insertion ratio, as sampled on $n=100$ genome windows.

**Estimates from least squares**

Line parameters:
- Intercept: $0.19210$
- Slope: $0.21777$

Error variance: $0.033$ on 98 dof’s

Mean responses:
- $\hat{y}(1) = 0.41$
- $\hat{y}(2) = 0.63$  … would you trust this?
Example: Simple linear regression for $y =$ mortality rate due to malignant skin melanoma per 10 million people, on $x =$ latitude, as sampled on 49 US states.

<table>
<thead>
<tr>
<th></th>
<th>State</th>
<th>LAT</th>
<th>MORT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Alabama</td>
<td>33.0</td>
<td>219</td>
</tr>
<tr>
<td>2</td>
<td>Arizona</td>
<td>34.5</td>
<td>160</td>
</tr>
<tr>
<td>3</td>
<td>Arkansas</td>
<td>35.0</td>
<td>170</td>
</tr>
<tr>
<td>4</td>
<td>California</td>
<td>37.5</td>
<td>182</td>
</tr>
<tr>
<td>5</td>
<td>Colorado</td>
<td>39.0</td>
<td>149</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>49</td>
<td>Wyoming</td>
<td>43.0</td>
<td>134</td>
</tr>
</tbody>
</table>

Estimates from least squares

Line parameters:
Intercept: 389.19
Slope: -5.977

Error variance: 365.57 on 47 dof’s

Mean responses:  $\hat{y}(30) = 209.88$  … would you trust this?  
$\hat{y}(40) = 150.11$
Maximum likelihood estimation under normality

\[ x_i, i = 1 \ldots n \text{ fixed (or condition on)} \]

Assume (error distribution)

\[ \varepsilon_i, i = 1 \ldots n \text{ iid } N(0, \sigma^2) \]

Then

\[ y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2) \text{ and indep, } i = 1 \ldots n \]

Likelihood function

\[
L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{1}{2\sigma^2} (y_i - (\beta_0 + \beta_1 x_i))^2 \right\} \\
= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 \right\}
\]
$$\max_{\beta_0, \beta_1, \sigma^2} L(\beta_0, \beta_1, \sigma^2) \quad \text{obj fct}$$

**Solution (unique)**

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$\hat{s}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$$

$$\frac{n-2}{n} s^2$$
Some remarks

• A strong statistical association between $y$ and $x$ does not automatically imply causation (e.g. a functional relationship of linear form); $x$ can “proxy” the real causing variable, perhaps is a spurious fashion!

• In observational studies, $x$ (likewise $y$) is not controlled by the researchers; we “condition” on the observed values of $x$. In experimental studies, $x$ is controlled; we can consider the values of $x$ as fixed (although assignment of $x$ levels to units may be arranged at random in an experimental design). Experimental design facilitates causality assessment.

• Extrapolating a statistical association (e.g. a regression line) outside the range of the data on which it was “fitted” is dangerous; we don’t really know how the association would be shaped where we didn’t get to look! With an experimental design we can make sure we cover the range that is of interest.