HW4 has been posted.
Due in a week
Regression analysis
Two variables
(Montgomery and Runger: ch 11
Brani Vidakovic: ch 14)
Reminder
Covariance Defined

Covariance is a number quantifying average dependence between two random variables.

The covariance between the random variables $X$ and $Y$, denoted as $\text{cov}(X, Y)$ or $\sigma_{XY}$ is

$$
\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - \mu_X \mu_Y \quad (5-14)
$$

The units of $\sigma_{XY}$ are units of $X$ times units of $Y$.

Unlike the range of variance, $-\infty < \sigma_{XY} < \infty$. 
Correlation is “normalized covariance”

• Also called: Pearson correlation coefficient

\[ \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \]

is the covariance normalized to be \(-1 \leq \rho_{XY} \leq 1\)

Karl Pearson (1852–1936) English mathematician and biostatistician
Covariance and Scatter Patterns

Figure 5-13  Joint probability distributions and the sign of $\text{cov}(X, Y)$. Note that covariance is a measure of linear relationship. Variables with non-zero covariance are correlated.
Regression analysis

• Many problems in engineering and science involve a sample in which two or more variables were measured. They may not be independent from each other and one (or several) of them can be used to predict another.

• Everyday example: in most samples height and weight of people are related to each other.

• Biological example: in a cell sorting experiment the copy number of a protein may be measured alongside its volume.

• **Regression analysis** uses a sample to build a model to predict protein copy number given a cell volume.
Sir Francis Galton, (1822 -1911) was an English statistician, anthropologist, proto-geneticist, psychometrician, eugenicist, (“Nature vs Nurture”, inheritance of intelligence), tropical explorer, geographer, inventor (Galton Whistle to test hearing), meteorologist (weather map, anticyclone).

Invented both correlation and regression analysis when studied heights of fathers and sons

Found that fathers with height above average tend to have sons with height also above average but closer to the average. Hence “regression” to the mean
Two variable samples

- Oxygen can be distilled from the air
- Hydrocarbons need to be filtered out or the whole thing would go kaboom!!!
- When more hydrocarbons were removed, the remaining oxygen stays cleaner
- Except we don’t know how dirty was the air to begin with

<table>
<thead>
<tr>
<th>Observation Number</th>
<th>Hydrocarbon Level $x(%)$</th>
<th>Purity $y(%)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.99</td>
<td>90.01</td>
</tr>
<tr>
<td>2</td>
<td>1.02</td>
<td>89.05</td>
</tr>
<tr>
<td>3</td>
<td>1.15</td>
<td>91.43</td>
</tr>
<tr>
<td>4</td>
<td>1.29</td>
<td>93.74</td>
</tr>
<tr>
<td>5</td>
<td>1.46</td>
<td>96.73</td>
</tr>
<tr>
<td>6</td>
<td>1.36</td>
<td>94.45</td>
</tr>
<tr>
<td>7</td>
<td>0.87</td>
<td>87.59</td>
</tr>
<tr>
<td>8</td>
<td>1.23</td>
<td>91.77</td>
</tr>
<tr>
<td>9</td>
<td>1.55</td>
<td>99.42</td>
</tr>
<tr>
<td>10</td>
<td>1.40</td>
<td>93.65</td>
</tr>
<tr>
<td>11</td>
<td>1.19</td>
<td>93.54</td>
</tr>
<tr>
<td>12</td>
<td>1.15</td>
<td>92.52</td>
</tr>
<tr>
<td>13</td>
<td>0.98</td>
<td>90.56</td>
</tr>
<tr>
<td>14</td>
<td>1.01</td>
<td>89.54</td>
</tr>
<tr>
<td>15</td>
<td>1.11</td>
<td>89.85</td>
</tr>
<tr>
<td>16</td>
<td>1.20</td>
<td>90.39</td>
</tr>
<tr>
<td>17</td>
<td>1.26</td>
<td>93.25</td>
</tr>
<tr>
<td>18</td>
<td>1.32</td>
<td>93.41</td>
</tr>
<tr>
<td>19</td>
<td>1.43</td>
<td>94.98</td>
</tr>
<tr>
<td>20</td>
<td>0.95</td>
<td>87.33</td>
</tr>
</tbody>
</table>
\[ Y = \beta_0 + \beta_1 X + \epsilon \]

Figure 11-1 Scatter diagram of oxygen purity versus hydrocarbon level from Table 11-1.

\[ Y = 75 + 15 \cdot X + \epsilon \]
Linear regression

The **simple linear regression model** is given by

\[ Y = \beta_0 + \beta_1 X + \varepsilon \]

\( \varepsilon \) is the **random error term**

slope \( \beta_1 \) and intercept \( \beta_0 \) of the line are called **regression coefficients**

**Note:** \( Y \), \( X \) and \( \varepsilon \) are random variables

The minimal assumption: \( \mathbb{E}(\varepsilon \mid x) = 0 \) \( \Rightarrow \)

\( \mathbb{E}(Y \mid x) = \beta_0 + \beta_1 x + \mathbb{E}(\varepsilon \mid x) = \beta_0 + \beta_1 x \)
\( Y = \beta_0 + \beta_1 X + \epsilon \) \hspace{1cm} E(\epsilon | x) = 0 \quad \forall x \)

How does one find \( \beta_0 \) & \( \beta_1 \)?

\[
\text{Cov}(Y, X) = \text{Cov}((\beta_0 + \beta_1 X + \epsilon), X) = \text{Cov}(\beta_0, X) + \beta_1 \text{Cov}(X, X) + \text{Cov}(\epsilon, X)
\]

\[
\text{Cov}(\beta_0, X) = 0 \quad \text{since } \beta_0 \text{ is constant}
\]

\[
\text{Cov}(X, X) = E(X^2) - E(X)^2 = \text{Var}(X)
\]

\[
\text{Cov}(\epsilon, X) = E(\epsilon \cdot X) - E(\epsilon) \cdot E(X) = E(\epsilon \cdot X) = \sum_x \epsilon \cdot E(\epsilon | x) = 0
\]

Thus \( \beta_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \) \hspace{1cm} \beta_0 = E(Y) - \beta_1 E(X) \)
The method of least squares is used to estimate the parameters, $\beta_0$ and $\beta_1$ by minimizing the sum of the squares of the vertical deviations in Figure 11-3.

**Figure 11-3** Deviations of the data from the estimated regression model.
Traditional notation

Definition

The least squares estimates of the intercept and slope in the simple linear regression model are

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

\[
\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \left( \frac{\sum_{i=1}^{n} y_i \left( \sum_{i=1}^{n} x_i \right)}{n} \right)}{\sum_{i=1}^{n} x_i^2 - \left( \frac{\sum_{i=1}^{n} x_i}{n} \right)^2}
\]

where \( \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \) and \( \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \).
11-2: Simple Linear Regression

Definition

The least squares estimates of the intercept and slope in the simple linear regression model are

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

(11-7)

\[
\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \left( \sum_{i=1}^{n} y_i \right) \left( \sum_{i=1}^{n} x_i \right)}{n^2} - \frac{\sum_{i=1}^{n} x_i^2 - \left( \sum_{i=1}^{n} x_i \right)^2}{n^2}
\]

(11-8)

\[
\text{Cov}(X, Y) \quad \frac{\text{Var}(X)}{n}
\]

where \( \bar{y} = (1/n) \sum_{i=1}^{n} y_i \) and \( \bar{x} = (1/n) \sum_{i=1}^{n} x_i \).
11-4.2 Analysis of Variance Approach to Test Significance of Regression

The analysis of variance identity is

\[ \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \]  

(11-24)

Symbolically,

\[ SS_T = SS_R + SS_E \]  

(11-25)
11-7: Adequacy of the Regression Model

11-7.2 Coefficient of Determination ($R^2$)

**VERY COMMONLY USED**

- The quantity

\[
R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}
\]

is called the **coefficient of determination** and is often used to judge the adequacy of a regression model.

- $0 \leq R^2 \leq 1$;
- We often refer (loosely) to $R^2$ as the amount of variability in the data explained or accounted for by the regression model.
11-7.2 Coefficient of Determination (R²)

- For the oxygen purity regression model,
  \[ R^2 = \frac{SS_R}{SS_T} = \frac{152.13}{173.38} = 0.877 \]
  
- Thus, the model accounts for 87.7% of the variability in the data.
Estimating $\sigma^2_{\epsilon}$

An unbiased estimator of $\sigma^2_{\epsilon}$ is

\[ \hat{\sigma}^2_{\epsilon} = \frac{SSE}{n - 2} \]  

where $SSE$ can be easily computed using

\[ SSE = SS_T - \hat{\beta}_1 S_{xy} \]
11-3: Properties of the Least Squares Estimators

• Slope Properties

\[ E(\hat{\beta}_1) = \beta_1 \]

\[ V(\hat{\beta}_1) = \frac{\hat{\sigma}_\epsilon^2}{S_{xx}} = \frac{\hat{\sigma}_\epsilon^2}{n \hat{\sigma}_x^2} \]

• Intercept Properties

\[ E(\hat{\beta}_0) = \beta_0 \quad \text{and} \quad V(\hat{\beta}_0) = \hat{\sigma}_\epsilon^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right] = \hat{\sigma}_\epsilon^2 \left[ 1 + \frac{\mu_x^2}{\hat{\sigma}_x^2} \right] \times \frac{1}{n} \]
Figure 11-5 The null hypothesis $H_0: \beta_1 = 0$ is accepted.
Figure 11-6: The null hypothesis $H_0: \beta_1 = 0$ is rejected.
11-4.1 Use of Z-tests for large n

An important special case of the hypotheses of Equation 11-18 is

\[ H_0: \beta_1 = 0 \]
\[ H_1: \beta_1 \neq 0 \]

These hypotheses relate to the significance of regression. Failure to reject \( H_0 \) is equivalent to concluding that there is no linear relationship between \( X \) and \( Y \).
11-4: Hypothesis Tests in Simple Linear Regression

\[ H_0: \beta_1 = 0 \]
\[ H_1: \beta_1 \neq 0 \]

\[ Z = \frac{\hat{\beta}_1}{\hat{\sigma}_\beta / \sqrt{n}} \]

Choose \( \alpha \)
(e.g., \( \alpha = 0.05 \))
(For 95% confidence in rejecting \( H_0 \))

Reject \( H_0 \) if \( |Z| > Z_{\alpha/2} = 1.96 \)
11-4.1 Use of $t$-tests for smaller $n$.

The number of degrees of freedom is $n-2$.

One can always fit a straight line through two points so one needs $n \geq 3$. 
11-4: Hypothesis Tests in Simple Linear Regression

\[ H_0: \beta_1 = 0 \]
\[ H_1: \beta_1 \neq 0 \]

\[ T = \beta_1 \] - [Equation]

Choose \( \alpha \)
\( (e.g., \alpha = 0.05) \)
for a 95% confidence in rejecting \( H_0 \)

\[ t_{\alpha/2, n-2} \] is such that

\[ 1 - \frac{\alpha}{2} = \text{cdf} \left( t_{\alpha/2, n-2} \right) \]

Reject \( H_0 \) if
\[ |T| > t_{\alpha/2, n-2} \]
QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE

Credit: XKCD comics