Final Exam
Thursday, July 27th

Bring to the exam:

Two red scantron forms (form F-1712)
#2 pencil
Ruler/straight edge
Scientific calculator

REVIEW GUIDE POSTED!!!!!
Assignment #4 due July 25th by 4pm
TODAY'S AGENDA

• Continue/Finish Module 4: Climate Variability
• Review last class
• The NAO
• The Arctic Oscillation (AO)
• The Antarctic Oscillation (AAO)
• Begin Module 5: Climate Chance
  • Least-squares
  • Linear regression
  • Hypothesis testing
Module #4
Climate Variability
How much of the change is natural? How much anthropogenic?
Climate

• A useful working definition: climate is the mean (or average) of the weather over time.
• Climate is what you expect, while weather is what you get.

Climate of Santa Cruz
Climate

• A useful time interval to use for defining *climate variability* and *climate change* is a full calendar year.
• In this way we average in time over a complete cycle of the four seasons (usually referred to as the *seasonal cycle*).
Covariance: Definition

• When studying climate we are often interested in time series, of say temperature, from two or different locations.
• The extent to which the temperatures at the two locations co-vary is of particular interest.
• This can be measured by the \textit{covariance}.
• Suppose \( S_i \) and \( M_i \) represent temperature measurements at two different locations, where \( i=1,2,...,N \) and \( N \) is the length of each time series.
• The \textit{covariance} between \( S \) and \( M \) is defined as:

\[
\text{cov}(S, M) = \frac{1}{(N-1)} \sum_{i=1}^{N} (S_i - \overline{S})(M_i - \overline{M})
\]
Covariance

- $\text{cov}(S, M)$ is a measure of how variations in $S$ and $M$, about their mean values, are related.
- Covariance, $\text{cov}(S, M)$, is very closely related to correlation, $r$.
- In fact:

\[
r = \frac{\text{cov}(S, M)}{\sigma_S \cdot \sigma_M}
\]

- However, unlike $r$, the covariance $\text{cov}(S, M)$ can have any value.
An Example: Sydney and Melbourne Temperatures
Variance and Covariance

Annual Mean Max. Temp. Anomalies

- It should be obvious that:
  \[ \text{cov}(S, S) = \sigma_S^2 = 0.52^\circ C^2 \]
  the variance of Sydney temperatures.

- And:
  \[ \text{cov}(M, M) = \sigma_M^2 = 0.37^\circ C^2 \]
  the variance of Melbourne temperatures.

- Furthermore:
  \[ \text{cov}(S, M) = \text{cov}(M, S) = 0.26^\circ C^2 \]
The Covariance Matrix

• Variance and covariance information for Sydney and Melbourne can be conveniently displayed in the form of a 2×2 matrix, \( C \), so that:

\[
C = \begin{pmatrix}
\text{cov}(S, S) & \text{cov}(M, S) \\
\text{cov}(S, M) & \text{cov}(M, M)
\end{pmatrix}
\]

or

\[
C = \begin{pmatrix}
\sigma_S^2 & \text{cov}(M, S) \\
\text{cov}(S, M) & \sigma_M^2
\end{pmatrix} = \begin{pmatrix}
0.52 & 0.26 \\
0.26 & 0.37
\end{pmatrix}
\]

• The diagonal elements of \( C \) are the variances of \( S \) and \( M \) and the off-diagonal elements are the covariances.
A Quick Refresher of Matrices and Vectors

Some definitions:
• If $a_{i,j}=a_{j,i}$ (i.e. the matrix is unchanged if the rows and columns are interchanged), then $A$ is referred to as a symmetric matrix.
• Covariance matrices are always symmetric matrices.
Sydney and Melbourne Temperatures: Variance and Covariance

Annual Mean Max. Temp. Anomalies

Warmer than the mean

Cooler than the mean

Time (years)

Sydney
Melbourne

\[
\text{Temp. Anomaly (°C)}
\]

\[
\begin{align*}
\text{cov}(S, S) &= \sigma_S^2 = 0.52 \degree C^2 \\
\text{cov}(M, M) &= \sigma_M^2 = 0.37 \degree C^2 \\
\text{cov}(S, M) &= \text{cov}(M, S) = 0.26 \degree C^2
\end{align*}
\]

\[
C = \begin{pmatrix}
\sigma_S^2 & \text{cov}(M, S) \\
\text{cov}(S, M) & \sigma_M^2
\end{pmatrix}
\]

\[
C = \begin{pmatrix}
0.52 & 0.26 \\
0.26 & 0.37
\end{pmatrix}
\]

Notice that \( C \) is square and symmetric

Sydney-Melbourne covariance matrix
Eigenvectors and Eigenvalues

• There are some *very special* vectors, $e$, that when multiplied by a square matrix $C$ remain unchanged, and just multiplied by a scalar, $\lambda$.

\[
Ce = \begin{pmatrix}
    c_{1,1} & c_{1,2} \\
    c_{2,1} & c_{2,2}
\end{pmatrix}
\begin{pmatrix}
    e_1 \\
    e_2
\end{pmatrix} = \lambda \begin{pmatrix}
    e_1 \\
    e_2
\end{pmatrix}
\]

• Or $Ce = \lambda e$

• For an $m \times m$ matrix, there are only $m$ such vectors $e$ and $m$ scalars $\lambda$ that have this property.

• $e$ is called an *eigenvector*, and $\lambda$ is an *eigenvalue*. 
A Table Summary of the Principal Components of Sydney and Melbourne Temperatures

<table>
<thead>
<tr>
<th></th>
<th>( e_1 )</th>
<th>( e_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sydney</td>
<td>0.8</td>
<td>-0.6</td>
</tr>
<tr>
<td>Melbourne</td>
<td>0.6</td>
<td>0.8</td>
</tr>
<tr>
<td>Eigenvalue</td>
<td>( \lambda_1 = 0.72 \degree C^2 )</td>
<td>( \lambda_2 = 0.17 \degree C^2 )</td>
</tr>
<tr>
<td>% Variance</td>
<td>81%</td>
<td>19%</td>
</tr>
</tbody>
</table>

Often it is more convenient to present principal component information in the form of a table.
Principal Component Analysis of Sydney and Melbourne Temperatures

**Annual Mean Max. Temp. Anomalies**

- **Principal component analysis** is a name given to the study of the eigenvectors of the covariance matrix, $C$.
- Recall that the eigenvectors, $e$, are special column vectors that are unchanged when multiplied by a matrix, except that they are multiplied by a scalar, the eigenvalue $\lambda$.
  \[
  Ce = \lambda e
  \]
- Since $C$ is a $2\times2$ matrix, it has two eigenvectors and two eigenvalues.
- So we have:
  \[
  Ce_1 = \lambda_1 e_1; \quad Ce_2 = \lambda_2 e_2
  \]
- It is usual to arrange the eigenvectors in order of increasing eigenvalue, so that:
  \[
  \lambda_1 > \lambda_2
  \]
Principal Component Analysis of Sydney and Melbourne Temperatures

Annual Mean Max. Temp. Anomalies

- The covariance matrix for Sydney and Melbourne annual mean daily max. temperature anomalies is:

\[ C = \begin{pmatrix} 0.52 & 0.26 \\ 0.26 & 0.37 \end{pmatrix} \]

- The eigenvalues and eigenvectors of \( C \) are:

\[ \lambda_1 = 0.72^\circ C^2 \quad e_1 = \begin{pmatrix} 0.8 \\ 0.6 \end{pmatrix} \]

\[ \lambda_2 = 0.17^\circ C^2 \quad e_2 = \begin{pmatrix} -0.6 \\ 0.8 \end{pmatrix} \]

\( e_1 \) is also called the **first principal component**

\( e_2 \) is also called the **second principal component**
A New Way of Thinking About Coordinates

• When we think about the coordinates of a point, we are used to thinking about them in terms of $x$ and $y$, relative to Cartesian axes.
• So for the starred point to the left, we represent its coordinates as $(x_1, y_1)$.
• If we drop the comma, then the coordinates of the star looks like a row vector, $(x_1 \ y_1)$.
• Now let’s write the coordinates as a column vector instead:

\[ \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} \]

• So now whenever we encounter a column vector, we can interpret it’s elements as coordinates.
Eigenvectors as Coordinates

- The Sydney-Melbourne covariance matrix, \( C \), is a 2×2 matrix because we considered the temperature at only two locations, namely Sydney and Melbourne.
- Similarly, the eigenvectors \( \mathbf{e}_1 \) and \( \mathbf{e}_2 \) are 2×1 column vectors because there are only two locations, Sydney and Melbourne.
- We just saw that we can treat the elements of a column vector as coordinates.
- In this case the coordinate axes will be the temperature anomalies \( S \) and \( M \) at the two locations, Sydney and Melbourne.
Sydney and Melbourne Temperatures as Coordinate axes: The “Scatter Plot”

- If we treat each pair of Sydney and Melbourne temperature anomalies, \( S \) and \( M \), as coordinates, and plot \( S \) vs \( M \) for each pair of temperature values, we obtain what is called a *scatter plot*.

- Notice how the resulting points form a *cloud* which slopes upwards from the lower left to the upper right.

- The orientation of the cloud is a reflection of the correlation between \( S \) and \( M \) which recall is \( r=+0.6 \).
Sydney and Melbourne Temperatures as Coordinate axes: The “Scatter Plot”

- Let’s divide the scatter plot into quadrants, centered on the origin (0,0), and label them A, B, C and D.
- Each quadrant describes different conditions at Sydney and Melbourne.
- This is sometimes called a *contingency diagram*.
- Quadrants B and C correspond to cases when Sydney and Melbourne temps co-vary *in phase*.
- Quadrants A and D correspond to cases when Sydney and Melbourne temps co-vary *out of phase*.
- Most of the points lie in quadrants B and C, so in most cases temperatures at the two locations co-vary *in phase* which is consistent with the correlation coefficient $r=+0.6$. 
The Principal Components of Sydney and Melbourne Temperatures

- Treating the principal components (i.e. eigenvectors) as coordinates of Sydney and Melbourne temperature anomalies, let’s overlay $e_1$ and $e_2$ on the scatter plot and contingency diagram.
- $e_1$ and $e_2$ are shown as the blue arrows.
- Let’s now extend the blue arrows in both directions to define a new set of axes which are the purple lines.
- Notice how the purple axis in the direction of $e_1$ passes through quadrants B and C where Sydney and Melbourne temps are in phase.
- Conversely the purple axis in the direction of $e_2$ passes through quadrants A and D where Sydney and Melbourne temps are out of phase.
The Principal Components of Sydney and Melbourne Temperatures

• The principal component $e_1$ tells us about situations when Sydney and Melbourne temperatures are in phase.

• The principal component $e_2$ tells us about situations when Sydney and Melbourne temperatures are out of phase.
The Principal Components of Sydney and Melbourne Temperatures

\[ e_1 = \begin{pmatrix} 0.8 \\ 0.6 \end{pmatrix} \]

- As we move along the purple axis defined by \( e_1 \) then for every 0.8°C change in Sydney temperature, Melbourne temperature changes by 0.6°C.

- Conversely, as we move along the purple axis defined by \(-e_1\) then for every -0.8°C change in Sydney temperature, Melbourne temperature changes by -0.6°C.
The Principal Components of Sydney and Melbourne Temperatures

\[ \mathbf{e}_2 = \begin{pmatrix} -0.6 \\ 0.8 \end{pmatrix} \]

- As we move along the purple axis defined by \( \mathbf{e}_2 \) then for every -0.6°C change in Sydney temperature, Melbourne temperature changes by 0.8°C.

- Conversely, as we move along the purple axis defined by -\( \mathbf{e}_2 \) then for every 0.6°C change in Sydney temperature, Melbourne temperature changes by -0.8°C.
The Principal Components of Sydney and Melbourne Temperatures

- The eigenvalues associated with each principal component contain useful information too.
- The total variance in the temperatures at Sydney and Melbourne is:
  \[ \sigma_S^2 + \sigma_M^2 = 0.52^\circ C^2 + 0.37^\circ C^2 = 0.89^\circ C^2 \]
- The total variance in the temperatures at Sydney and Melbourne is also equal to the sum of the eigenvalues.
  \[ \lambda_1 + \lambda_2 = 0.72^\circ C^2 + 0.17^\circ C^2 = 0.89^\circ C^2 \]
  \[ \lambda_1 + \lambda_2 = \sigma_S^2 + \sigma_M^2 \]
The Principal Components of Sydney and Melbourne Temperatures

- The eigenvalue \( \lambda_1 \) associated with principal component \( e_1 \) can be thought of as the variance associated with all the situations for which Sydney and Melbourne temperature co-vary *in phase* (*i.e.* all the points in quadrants B and C).

- The eigenvalue \( \lambda_2 \) associated with principal component \( e_2 \) can be thought of as the variance associated with all the situations for which Sydney and Melbourne temperature co-vary *out of phase* (*i.e.* all the points in quadrants A and D).
The Principal Components of Sydney and Melbourne Temperatures

- The fraction of the covariance in Sydney and Melbourne temperature accounted for along the principal component axis $e_1$ is:
  \[
  \frac{\lambda_1}{(\sigma_S^2 + \sigma_M^2)} = \frac{\lambda_1}{(\lambda_1 + \lambda_2)} = \frac{0.72}{0.89} = 0.81 \text{ (i.e. 81%)}
  \]

- The fraction of the covariance in Sydney and Melbourne temperature accounted for along the principal component axis $e_2$ is:
  \[
  \frac{\lambda_2}{(\sigma_S^2 + \sigma_M^2)} = \frac{\lambda_2}{(\lambda_1 + \lambda_2)} = \frac{0.17}{0.89} = 0.19 \text{ (i.e. 19%)}
  \]
The Principal Component Time Series

• Suppose we arrange the temperature anomalies at Sydney, $S_i$, and Melbourne, $M_i$, as the elements of a column vector, $y_i$:

\[
y_i = \begin{pmatrix} S_i \\ M_i \end{pmatrix}
\]

• As usual, $i=1$: 1859; $i=2$: 1860; $i=3$: 1861; etc.
• The vector $y_i$ can be written as a unique combination of the principal components $e_1$ and $e_2$ so that:

\[
y_i = a_i e_1 + b_i e_2
\]

• $a_i$ is called the first principal component time series, and describes situations when Sydney and Melbourne co-vary in phase.
• $b_i$ is called the second principal component time series, and describes situations when Sydney and Melbourne co-vary out of phase.
Covariance

• The concept of **covariance** can be extended to situations where we are interested in how time series at *more than* two locations co-vary.
• In fact we may consider **any number** of time series, and in studies of climate it is common to consider many thousands of time series altogether!
• However, as our next step, we will consider just three time series.
An Example: Sydney, Melbourne & Perth

Temperatures

- Sydney (S)
- Melbourne (M)
- Perth (P)
An Example: Sydney, Melbourne & Perth Temperatures (1945-2013)

Time Series of Temperature Anomalies

- We will denote annual mean daily max. temperature anomalies at Sydney as $S_i$
- We will denote annual mean daily max. temperature anomalies at Melbourne as $M_i$
- We will denote annual mean daily max. temperature anomalies at Perth as $P_i$
- At each location, the time series span the period 1945-2013, so that $i=1: 1945$; $i=2: 1946$; $i=3: 1947$; ...; $i=69: 2013$. 
The Sydney-Melbourne-Perth Covariance Matrix

• Variance and covariance information for Sydney, Melbourne and Perth can be displayed in the form of a 3×3 matrix, \( C \), so that:

\[
C = \begin{pmatrix}
\text{cov}(S, S) & \text{cov}(M, S) & \text{cov}(P, S) \\
\text{cov}(S, M) & \text{cov}(M, M) & \text{cov}(P, M) \\
\text{cov}(S, P) & \text{cov}(M, P) & \text{cov}(P, P)
\end{pmatrix}
\]

• The 3×3 covariance matrix describes the covariance between the time series at all possible location pairs.

• Recall though that:

\[
\text{cov}(S, S) = \sigma_S^2; \quad \text{cov}(M, M) = \sigma_M^2; \quad \text{cov}(P, P) = \sigma_P^2; \\
\text{cov}(S, M) = \text{cov}(M, S); \quad \text{cov}(S, P) = \text{cov}(P, S); \quad \text{etc}
\]
The Sydney-Melbourne-Perth Covariance Matrix

\[
C = \begin{pmatrix}
\text{cov}(S, S) & \text{cov}(M, S) & \text{cov}(P, S) \\
\text{cov}(S, M) & \text{cov}(M, M) & \text{cov}(P, M) \\
\text{cov}(S, P) & \text{cov}(M, P) & \text{cov}(P, P)
\end{pmatrix}
\]

\[
C = \begin{pmatrix}
\sigma_S^2 & \text{cov}(M, S) & \text{cov}(P, S) \\
\text{cov}(S, M) & \sigma_M^2 & \text{cov}(P, M) \\
\text{cov}(S, P) & \text{cov}(M, P) & \sigma_P^2
\end{pmatrix}
\]

\[
C = \begin{pmatrix}
0.31 & 0.23 & 0.19 \\
0.23 & 0.46 & 0.26 \\
0.19 & 0.26 & 0.60
\end{pmatrix}
\]

Notice that \(C\) is square and symmetric, \(C_{i,j} = C_{j,i}\)
Principal Component Analysis of the Sydney-Melbourne-Perth Covariance Matrix

The 3×3 covariance matrix for Sydney-Melbourne-Perth will have 3 eigenvectors (or principal components) and 3 eigenvalues, summarized in the table below.

<table>
<thead>
<tr>
<th></th>
<th>$e_1$</th>
<th>$e_2$</th>
<th>$e_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sydney</td>
<td>0.43</td>
<td>0.39</td>
<td>0.81</td>
</tr>
<tr>
<td>Melbourne</td>
<td>0.58</td>
<td>0.57</td>
<td>-0.58</td>
</tr>
<tr>
<td>Perth</td>
<td>0.69</td>
<td>-0.72</td>
<td>-0.02</td>
</tr>
<tr>
<td>Eigenvalue</td>
<td>$\lambda_1=0.93^\circ C^2$</td>
<td>$\lambda_2=0.29^\circ C^2$</td>
<td>$\lambda_3=0.15^\circ C^2$</td>
</tr>
<tr>
<td>% Variance</td>
<td>68%</td>
<td>22%</td>
<td>10%</td>
</tr>
</tbody>
</table>
Fraction of Variance Explained by each Principal Component

- Recall that the total variance equals the sum of the eigenvalues:

\[ \sigma_S^2 + \sigma_M^2 + \sigma_P^2 = \lambda_1 + \lambda_2 + \lambda_3 \]

- So the fraction of the total variance explained by principal component \( e_1 \) is:

\[ \frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3} \]

- The fraction of the total variance explained by principal component \( e_2 \) is:

\[ \frac{\lambda_2}{\lambda_1 + \lambda_2 + \lambda_3} \]

- The fraction of the total variance explained by principal component \( e_3 \) is:

\[ \frac{\lambda_3}{\lambda_1 + \lambda_2 + \lambda_3} \]
Visualizing the principal components in more than 2-dimensions

• In the previous example, where we considered only Sydney and Melbourne, it was useful to create a *scatter plot* and a *contingency diagram*.
• This allowed us to explore the *in phase* and *out of phase* relationship at Sydney and Melbourne described by the principal components.
• In the present example, however, it would be difficult to visualize a 3-dimensional *scatter plot* and *contingency diagram*, so we must use another approach to present the principal component information.
Visualizing the principal components in more than 2-dimensions

The elements of the principal components will be represented as circles. Red (blue) circles denote positive (negative) numbers, and the circle diameter is proportional to the relative size of each element.

\[ e_1 = \begin{pmatrix} 0.43 \\ 0.58 \\ 0.69 \end{pmatrix} \]

- Notice that all three elements are positive which means that the temperatures at each location co-vary in phase.
- A \(+0.43^\circ C\) \((-0.43^\circ C\) temperature anomaly at Sydney is accompanied by a \(+0.58^\circ C\) \((-0.58^\circ C\) anomaly at Melbourne, and a \(+0.69^\circ C\) \((-0.69^\circ C)\) anomaly at Perth.
For $e_1$

$$\frac{\Delta S}{0.43} = \frac{\Delta M}{0.58} = \frac{\Delta P}{0.69}$$
Visualizing the principal components in more than 2-dimensions

The elements of the principal components will be represented as circles. Red (blue) circles denote positive (negative) numbers, and the circle diameter is proportional to the relative size of each element.

\[
e_2 = \begin{pmatrix} 0.39 \\ 0.57 \\ -0.72 \end{pmatrix}
\]

- A +0.39°C (-0.39°C) temperature anomaly at Sydney is accompanied by a +0.57°C (-0.57°C) anomaly at Melbourne, and a -0.72°C (+0.72°C) at Perth.
- Sydney and Melbourne temperatures co-vary in phase with each other, but co-vary out of phase with temperatures at Perth.
For $e_2$

\[
\frac{\Delta S}{0.39} = \frac{\Delta M}{0.57} = \frac{\Delta P}{-0.72}
\]
Principal Component (PC) Time Series

- Arrange time series for $S$, $M$, and $P$ in a vector $\mathbf{y}$:
  \[
  \mathbf{y}_i = \begin{pmatrix} S_i \\ M_i \\ P_i \end{pmatrix}
  \]
  for $i=1$: 1945; $i=2$: 1946; $i=3$: 1947; \ldots; $i=69$: 2013.

- Recall that we can express $\mathbf{y}_i$ as a unique combination of $\mathbf{e}_1$, $\mathbf{e}_2$, and $\mathbf{e}_3$.
  \[
  \mathbf{y}_i = a_i \mathbf{e}_1 + b_i \mathbf{e}_2 + c_i \mathbf{e}_3
  \]
  where $a_i$ is the 1$^{st}$ principal component (PC) time series, $b_i$ is the 2$^{nd}$ PC time series, and $c_i$ is the 3$^{rd}$ PC time series.
Assignment #4: Principal Component Analysis for Sydney, Melbourne, Perth, Darwin and Alice Springs
Assignment #4: Principal Component Analysis for Sydney, Melbourne, Perth, Darwin and Alice Springs

Time Series of Temperature Anomalies

- Annual mean daily max. temperature anomalies at Sydney, $S_i$
- Annual mean daily max. temperature anomalies at Melbourne, $M_i$
- Annual mean daily max. temperature anomalies at Perth, $P_i$
- Annual mean daily max. temperature anomalies at Alice Springs, $A_i$
- Annual mean daily max. temperature anomalies at Darwin, $D_i$

- At each location, the time series spans the period 1947-2013, so that $i=1: 1947$; $i=2: 1947$; $i=3$: 1948; . . . ; $i=67$: 2013.
The Sydney-Melbourne-Perth-Alice-Darwin 5×5 Covariance Matrix

\[ C = \begin{pmatrix}
\sigma_S^2 & \text{cov}(M,S) & \text{cov}(P,S) & \text{cov}(A,S) & \text{cov}(D,S) \\
\text{cov}(S,M) & \sigma_M^2 & \text{cov}(P,M) & \text{cov}(A,M) & \text{cov}(D,M) \\
\text{cov}(S,P) & \text{cov}(M,P) & \sigma_P^2 & \text{cov}(A,P) & \text{cov}(D,P) \\
\text{cov}(S,A) & \text{cov}(M,A) & \text{cov}(P,A) & \sigma_A^2 & \text{cov}(D,A) \\
\text{cov}(S,D) & \text{cov}(M,D) & \text{cov}(P,D) & \text{cov}(A,D) & \sigma_D^2
\end{pmatrix} \]

\[ C = \begin{pmatrix}
0.31 & 0.22 & 0.19 & 0.30 & 0.11 \\
0.22 & 0.44 & 0.25 & 0.30 & 0.09 \\
0.19 & 0.25 & 0.60 & 0.19 & 0.09 \\
0.30 & 0.30 & 0.19 & 0.84 & 0.19 \\
0.11 & 0.09 & 0.09 & 0.19 & 0.17
\end{pmatrix} \]

Notice that C is square and symmetric

\[ C_{i,j} = C_{j,i} \]
Principal Component Analysis of the Sydney-Melbourne-Perth-Alice-Darwin Covariance Matrix

<table>
<thead>
<tr>
<th>City</th>
<th>e$_1$</th>
<th>e$_2$</th>
<th>e$_3$</th>
<th>e$_4$</th>
<th>e$_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sydney</td>
<td>0.37</td>
<td>0.01</td>
<td>0.18</td>
<td>0.74</td>
<td>-0.53</td>
</tr>
<tr>
<td>Melbourne</td>
<td>0.43</td>
<td>0.19</td>
<td>0.81</td>
<td>-0.28</td>
<td>0.20</td>
</tr>
<tr>
<td>Perth</td>
<td>0.43</td>
<td>0.77</td>
<td>-0.45</td>
<td>-0.13</td>
<td>-0.01</td>
</tr>
<tr>
<td>Alice Springs</td>
<td>0.67</td>
<td>-0.60</td>
<td>-0.31</td>
<td>-0.30</td>
<td>-0.07</td>
</tr>
<tr>
<td>Darwin</td>
<td>0.20</td>
<td>-0.08</td>
<td>-0.11</td>
<td>0.51</td>
<td>0.82</td>
</tr>
<tr>
<td>Eigenvalue</td>
<td>$\lambda_1$=1.37</td>
<td>$\lambda_2$=0.51</td>
<td>$\lambda_3$=0.23</td>
<td>$\lambda_4$=0.15</td>
<td>$\lambda_5$=0.10</td>
</tr>
</tbody>
</table>
Visualizing principal components with 5-dimensions

The elements of the principal components will be represented as circles. Red (blue) circles denote positive (negative) numbers, and the circle diameter is proportional to the relative size of each element.

What is the physical interpretation of this principal component?
For $e_1$

\[
\frac{\Delta S}{0.37} = \frac{\Delta M}{0.43} = \frac{\Delta P}{0.43} = \frac{\Delta A}{0.67} = \frac{\Delta D}{0.20}
\]
Covariance Matrices with a High Dimension

- With the advent of earth orbiting satellites in the early 1980s, long time series of temperature (and other climate variables) are available at adjacent locations due to the repeat nature of satellite orbits.
Covariance Matrices with a High Dimension

• If there are $N$ such time series of temperature, then the resulting covariance matrix will be $N \times N$.
• Modern day satellites can measure surface temperature with $\sim 5$km resolution, meaning the $N$ will be very large!
• Nevertheless, we can still perform principal component analysis by computing the eigenvectors and eigenvalues of the resulting $N \times N$ covariance matrix.
• However, while there will be $N$ pairs of eigenvalues and eigenvectors, it is generally found that the first few principal components account for most of the total variance.
• Therefore, even though $N$ is large, we need only trouble ourselves with looking at a few principal components.
• How then do we visualize a principal component vector with a very large dimension (e.g. $N \sim 100s$ or 1000s!)?
Visualizing Principal Components Covariance with a High Dimension

• In the previous examples, we used red and blue circles of varying diameter to represent multi-dimensional principal components.
• However, when $N$ is very large this approach won’t work well at all since all the circles will overlap and merge together.
• *Instead we can draw contours through the elements of the principal components that correspond to the different geographical locations.*
• *Thus we can represent the principal components graphically as a color map.*
• Another common name used for the principal components in climate science is *Empirical Orthogonal Functions (EOFs).*
• Eigenvectors =

• Principal Components (PCs) =

• Empirical Orthogonal Functions (EOFs)
ENSO as a Principal Component

The El Niño Southern Oscillation can also be identified using principal component analysis.

Time series of 1st principal component of Sea Surface Temp. (SST) and surface currents (SC).

1st principal component e₁ of Sea Surface Temp. (SST) and surface currents (SC).

SST – colored shading; SC - arrows
The Pacific Decadal Oscillation (PDO) & The North Pacific Gyre Oscillation (NPGO)

$e_1$ for N. Pacific SST

$e_2$ for N. Pacific SST

Principal component time series for $e_1$

Principal component time series for $e_2$
The PDO

Notice the stronger than normal Aleutian Low when PDO is positive.

Notice the weaker than normal Aleutian Low when PDO is negative.
### Table 1: summary of Pacific and North American climate anomalies associated with extreme phases of the PDO.

<table>
<thead>
<tr>
<th>Climate Anomalies</th>
<th>Warm Phase PDO</th>
<th>Cool Phase PDO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ocean surface temperatures in the northeastern and tropical Pacific</td>
<td>Above average</td>
<td>Below average</td>
</tr>
<tr>
<td>October-March northwestern North American air temperatures</td>
<td>Above average</td>
<td>Below average</td>
</tr>
<tr>
<td>October-March southeastern US air temperatures</td>
<td>Below average</td>
<td>Above average</td>
</tr>
<tr>
<td>October-March southern US/Northern Mexico precipitation</td>
<td>Above average</td>
<td>Below average</td>
</tr>
<tr>
<td>October-March Northwestern North America and Great Lakes precipitation</td>
<td>Below average</td>
<td>Above average</td>
</tr>
<tr>
<td>Northwestern North American spring time snow pack and water year (October-September) stream flow</td>
<td>Below average</td>
<td>Above average</td>
</tr>
<tr>
<td>Winter and spring time flood risk in the Pacific Northwest</td>
<td>Below average</td>
<td>Above average</td>
</tr>
</tbody>
</table>

---

### Influence of PDO on Climate

Similar to that of ENSO in many ways

<table>
<thead>
<tr>
<th>PDO Indicators</th>
<th>PDO positive phase</th>
<th>PDO negative phase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pacific Northwest, British Columbia, and Alaska</td>
<td>Above average</td>
<td>Below average</td>
</tr>
<tr>
<td>Mexico to South-East US</td>
<td>Below average</td>
<td>Above average</td>
</tr>
<tr>
<td>Precipitation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alaska coastal range</td>
<td>Above average</td>
<td>Below average</td>
</tr>
<tr>
<td>Mexico to South-Western US</td>
<td>Above average</td>
<td>Below average</td>
</tr>
<tr>
<td>Canada, Eastern Siberia and Australia</td>
<td>Below average</td>
<td>Above average</td>
</tr>
<tr>
<td>India summer monsoon</td>
<td>Below average</td>
<td>Above average</td>
</tr>
</tbody>
</table>
Influence of PDO on Surface Temperatures

PDO positive phase (reverse for the negative phase).

Red indicates warmer than normal; Blue indicates cooler than normal
Influence of PDO on Precipitation

PDO positive phase (reverse for the negative phase).

Red indicates drier than normal; Blue indicates wetter than normal
The PDO During the Last Century
The PDO During the Last Millenium
A Global View of the Pacific Decadal Oscillation (PDO)

Notice how the PDO has a similar temperature structure to ENSO in the tropical Pacific. The positive (negative) phase of the PDO tends to add to (subtract from) the El Niño pattern.
The PDO During the Last Century
The Pacific Decadal Oscillation (PDO) & The North Pacific Gyre Oscillation (NPGO)

Principal component time series for $e_1$

$e_1$ for N. Pacific SST

Principal component time series for $e_2$

$e_2$ for N. Pacific SST
The North Pacific Gyre Oscillation (NPGO) and Sea Surface Height

The positive phase of NPGO sea surface height

Red=higher than normal sea level.

Blue=lower than normal sea level

The 2nd principal component of sea surface height, e2
What will the *positive* phase of the NPGO do to existing sea level and gyre strength?

For the *positive* phase of NPGO sea surface height:

Red = *higher* than normal sea level.

Blue = *lower* than normal sea level

This corresponds to stronger gyre circulations.

The 2\textsuperscript{nd} principal component of sea surface height, $e_2$
What will the **negative** phase of the NPGO do to existing sea level and gyre strength?

For the **negative** phase of NPGO sea surface height:

- **Red** = lower than normal sea level.
- **Blue** = higher than normal sea level

This corresponds to weaker gyre circulations.

The 2\textsuperscript{nd} principal component of sea surface height, \( e_2 \)
The North Atlantic Oscillation Revisited

• The North Atlantic Oscillation (NAO) can also be analyzed using principal component analysis.
• In fact, all of the major atmospheric teleconnection patterns, including the PNA, emerge as leading principal components of the sea level pressure anomalies.
• In this case the covariance between time series of sea level pressure is computed to yield the covariance matrix $\mathbf{C}$.
• If the *annual mean* sea level pressure anomalies over the North Atlantic Ocean are used to compute $\mathbf{C}$, the low-frequency nature of the NAO can be explored.
The principal component, $e_1$, of annual mean sea level pressure

Variance explained = 32%

Positive phase of NAO:
Red dashed=negative
Blue=positive

The principal component time series of $e_1$
The Arctic Oscillation (AO)

• Principal component analysis of the covariance matrix of atmospheric pressure anomalies over the entire northern hemisphere reveals the presence of an annular mode of variability called the Arctic Oscillation (AO).

• The leading principal component $e_1$ of the northern hemisphere surface pressure anomalies showing the structure of the Arctic Oscillation:
  - The AO accounts for 19% of the surface pressure variance.
  - The positive phase of the AO is characterized by lower than normal pressures over the north pole, and higher than normal pressures over the North Pacific and North Atlantic.
  - The negative phase of the AO is characterized by higher than normal pressures over the north pole, and lower than normal pressures over the North Pacific and North Atlantic.
The AO and the Jet Streams

- There are typically two jet streams in the northern hemisphere.
- A “**subtropical jet stream**” where the downward branch of the Hadley and Ferrel cells meet.
- A “**subpolar jet stream**” located above the upward branch of the Ferrel cell.
Jet Stream Cirrus over Canada
Jet Stream Cirrus over the Middle East
The Arctic Oscillation (AO)

The AO can affect winter time weather and climate of North America, Europe and Asia by shifting the storm tracks.

- During the **positive** phase of the AO, the *subpolar jet stream* is **farther north than normal**, meaning that storm systems travel further north of the usual path.
- During these time **N. America, Europe, Siberia** and **East Asia** experience fewer outbreaks of cold Arctic air than normal.
- During the **negative** phase of the AO, the *subpolar jet stream* and storm tracks are **farther south than normal**, meaning that these same locations experience more outbreaks of cold polar air.
The Arctic Oscillation Principal Component Time Series

Principal Component Time Series for the AO

The AO Index
The Antarctic Oscillation (AAO)

- Principal component analysis of the covariance matrix of atmospheric pressure anomalies over the entire southern hemisphere reveals the presence of an annular mode of variability called the Antarctic Oscillation (AAO), also known as the Southern Annual Mode (SAM).

- The leading principal component $e_1$ of the southern hemisphere 750mb geopotential height (mid-troposphere) anomalies showing the structure of the Antarctic Oscillation (AAO).
- The AAO accounts for 27% of the 750mb pressure variance.
- The positive (negative) phase of the AAO is characterized by lower (higher) than normal pressures over the south pole, and higher (lower) than normal pressures over the South Pacific, Southern Indian Ocean and South Atlantic.
The Antarctic Oscillation (AAO)

- The AAO has been linked to changes in the strength of the circumpolar vortex in the stratosphere.

Leading EOF (27%) shown as regression map of 700mb height (m)

The Circumpolar Vortex

Copyright © 2004 Pearson Prentice Hall, Inc.
The Antarctic Oscillation (AAO)

- The circumpolar vortex forms in the SH winter, and influences the ozone hole.

Stratospheric ozone is formed in the tropics and transported to the poles by the Brewer-Dobson circulation in the stratosphere.
Latitudinal Distribution of Ozone

Notice the pronounced seasonal cycle

Ozone column depth (DU)

Copyright © 2004 Pearson Prentice Hall, Inc.
Ozone Production – The Chapman Mechanism

(1) $\text{O}_2 + \text{UV photon} \rightarrow \text{O} + \text{O}$

(2) $\text{O} + \text{O}_2 + \text{M} \rightarrow \text{O}_3 + \text{M}$

(3) $\text{O}_3 + \text{photon} \rightarrow \text{O}_2 + \text{O}$

(4) $\text{O} + \text{O}_3 \rightarrow 2\text{O}_2$

M = another molecule to carry away excess energy
The photon in reaction 3 involves visible radiation not UV.
Reaction 1 can only occur in the stratosphere while 3 can occur all the way down to the surface. This is why $\text{O}_3$ located in stratosphere.
The destruction of stratospheric ozone is accelerated by chlorine from CFCs.
The Ozone Hole

• During the SH polar winter, chlorine gas, released from chlorofluorocarbons (CFCs), builds up in the very cold and dark stratosphere.
• When the sun rises above the horizon in the spring, the chlorine gas is photodissociated by solar radiation forming Cl⁻ ions.
• The Cl⁻ ions very quickly destroy ozone via a catalytic reaction, and the ozone hole forms.
• The figure on the left shows stratospheric ozone concentration during Sept, 2000. The region of blue indicates ozone depletion, and 2000 was the largest ozone hole on record.
• In the winter and spring, the polar vortex is very strong and prevents new ozone formed in the tropics from replenishing the ozone destroyed near the pole.
The Ozone Hole is a very dynamic feature

KNMI/NIVR/NASA

www.knmi.nl 2004-09-27
The Double Ozone Hole of 2002
Primarily over the Southern Ocean
The Antarctic Oscillation (AAO)

- The **positive** phase of the AAO **strengthens** the circumpolar vortex further inhibiting the transport of ozone to the polar region.
The Antarctic Oscillation Principal Component Time Series

Principal Component Time Series for the AAO

The AAO Index
BREAK!
Module #5
Climate Change
Detailed notes on the class website: module5.pdf

Module #4
Climate Change
Part I

• Straight lines
• The method of “least-squares”
• Linear regression
The Equation of a Straight Line

Recall the equation of a straight line:

\[ y = mx + c \]
Some Example Straight Lines

\[ y = mx + c \]
Data that Approximate a Straight Line

• Suppose now that $x$ and $y$ are two climate variables, such as CO$_2$ concentration and global average temperature.
• If the two variables fall on a straight line, then it is likely (although not guaranteed) that they are related in some way.
An Illustrative Example: The Acceleration due to Gravity

- Consider the speed, \( v \), of a ball released from rest, falling under the influence of gravity in a vacuum.
- From Newton’s laws of motion, the speed \( v \) is given by:

\[
v = gt
\]

where \( g \) is the acceleration due to gravity and \( t \) is time.

- Recall the equation of a straight line: \( y = mx + c \)
- If we replace \( y \) by \( v \), and \( x \) by \( t \), and choose \( c=0 \), then the equation for \( v \) is identical to that of a straight line.
- Therefore measurements of \( v \) and \( t \) should fall on a straight line with a slope \( g \).
An Illustrative Example: The Acceleration due to Gravity

Measurements of speed and time

• Measurements of speed \( v \) and time \( t \) will be subject to instrument errors and human errors.
• Because of such errors, the measurements of \( v \) and \( t \) will not fall exactly on a straight line.
• The degree to which the measurements depart from a straight line depends on the accuracy of measurements of speed and time.
• The figure shows what measurements made 100 years ago might have looked like using the technology of the day.
• The measurement errors in speed have a standard deviation of 6 m/s.
• The measurement errors in time have a standard deviation of 1 second.

Notice how the measurements of speed and time are scattered about a straight line due to errors in the measurements.
The PDFs of Measurement Errors

The mean error is zero for both speed and time.

Standard deviation:
- $\sigma = 6 \text{ m/s}$ for errors in speed
- $\sigma = 1 \text{ s}$ for errors in time
Finding the best-fit straight line to measurements: The method of least-squares

- Suppose that we don’t know the acceleration due to gravity.
- How well can we estimate it from the measurements?
- We expect our data to fall on a straight line.
- So if we can draw a straight line that passes close to all the data, we could estimate the acceleration due to gravity from the slope of the line.
- Unfortunately, there are an infinite number of possible straight lines!
- However, there is one which is “best” based on some criterion.
- The criterion usually used is to find the straight line that yields the minimum squared distance between the data and the straight line.
Finding the best-fit straight line to measurements: The method of least-squares
Finding the best-fit straight line to measurements: The method of least-squares

\[ \text{distance} = v - gt \]
Finding the best-fit straight line to measurements: The method of least-squares

Measurements of speed and time

- Let’s label the pairs of $t$ and $v$ as $(t_1, v_1)$, $(t_2, v_2)$, $(t_3, v_3)$, etc.
- Using the **minimum squared distance** criterion, the problem of finding the “best” straight line becomes one of finding the acceleration due to gravity, $g$, that yields the smallest value of:

$$ J = \frac{1}{N} \sum_{i=1}^{N} (v_i - gt_i)^2 $$

where $N$ is the total number of measurement pairs.

- Identifying the acceleration due to gravity, $g$, in this way is referred to as finding the “**least-squares straight line fit to the data**.”
Finding the best-fit straight line to measurements: The method of least-squares

The figure shows how $J$ given by:

$$J = \frac{1}{N} \sum_{i=1}^{N} (v_i - gt_i)^2$$

varies for different choices of $g$ between 8 ms$^{-2}$ and 12 ms$^{-2}$.

- With the “least-squares” criterion in mind, the smallest value of $J$ occurs for a value of $g$ close to 10 ms$^{-2}$.
- There are two ways that we could estimate the “best” value of $g$:
  1. Very carefully determine the value of $g$ from the graph as corresponding to the smallest value of $J$.
  2. Find the value of $J$ that corresponds to the point on the curve where the slope (i.e. the gradient) exactly vanishes.
Finding the best-fit straight line to measurements: The method of least-squares

- The slope of the curve is given by $dJ/dg$:

$$
\frac{dJ}{dg} = -\frac{2}{N} \sum_{i=1}^{N} t_i (v_i - gt_i)
$$

- The slope of the curve vanishes where $dJ/dg=0$, in which case:

$$
-\frac{2}{N} \sum_{i=1}^{N} t_i (v_i - gt_i) = 0
$$

and:

$$
g = \frac{\sum_{i=1}^{N} t_i v_i}{\sum_{i=1}^{N} t_i^2}
$$

- Using the data for $v_i$ and $t_i$ this formula gives a value of $g=9.81 \text{ ms}^{-2}$ which is the generally accepted value of the acceleration due to gravity at the earth surface.
Finding the best-fit straight line to measurements: The method of least-squares

• Using the “least-squares best fit” value for g, the “best fit straight line” is given by:

\[ v = 9.81t \]

• Based on the “least-squares” criterion, this is the best choice of straight line from the infinite number of possible choices.
Finding the best-fit straight line to measurements: The method of least-squares

- If the measurement errors are larger, the “scatter” of the measurements increases.
- The figure shows the measurements that result when the standard deviation of errors in measurements of speed and time are increased to 12 ms\(^{-1}\) and 2 seconds respectively.
- The “least-squares best fit straight line” is almost the same as the previous case.
Linear Regression: Melbourne vs Sydney Temperatures

- Suppose that we have a sequence of measurements from two locations and we want to determine if they are related in some way?
- We can use the “least-squares” method to explore this question.
- This is referred to as linear regression analysis.
- As an example, consider again the annual mean daily maximum temperature anomalies (i.e. the mean seasonal cycle has been removed) at Sydney ($S$) and Melbourne ($M$).
- Time series of the anomalies $S$ and $M$ are shown in the figure.
Linear Regression: Melbourne vs Sydney Temperatures (1859-2013)

- As in previous modules, we will denote Sydney temperature anomalies as $S_i$ and Melbourne temperature anomalies as $M_i$
  - $i=1$: 1859; $i=2$: 1860; $i=3$: 1861; etc.
- Recall the scatter plot from Module 4 where we use Melbourne temperature as the $y$-axis coordinate and Sydney temperature as the $x$-axis coordinate.
- The scatter plot for Melbourne vs Sydney temperature is shown in the figure.
- As we noted in Module 4, $S_i$ and $M_i$ form a “cloud” of points that slopes upward from the lower left to the upper right.
- $S_i$ and $M_i$ are positively correlated ($r=0.6$), and principal component analysis reveals that 81% of the covariance is explained by a warming trend.
Linear Regression: Melbourne vs Sydney Temperatures (1859-2013)

- The upward slope of the cloud from the lower left to the upper right suggests that the temperatures at the two locations are related.
- This is not surprising given that the two cities are only separated by a distance of 1000 km.
- Nevertheless, it is of interest to quantify how much of the temperature variation at one location, say Melbourne, can be explained (or “predicted”) using temperature variations at the other location, Sydney.
- This is called linear regression and is based on the equation of a straight line:

\[ y = mx + c \]
Linear Regression: Melbourne vs Sydney Temperatures (1859-2013)

- In the language of linear regression, the variable $x$ is referred to as the predictor, while the variable $y$ is referred to as the predictand.

\[ y = mx + c \]

- In the example considered here, we will consider Sydney temp. anomalies $S_i$ to be the predictor and the Melbourne temp. anomalies $M_i$ to be the predictand.

\[ M = mS + c \]
In other words, we will explore how well we can predict (or explain) the temperature anomalies at Melbourne using the temperature anomalies at Sydney.

To do this we will use the equation of a straight line:

\[ M = mS + c \]

The problem in linear regression is to identify the “best” choice of \( m \) and \( c \) using the “least-squares” criterion.
Linear Regression: Melbourne vs Sydney Temperatures (1859-2013)

- Specifically, we want to find the choice of $m$ and $c$ in the equation:

$$M = mS + c$$

that correspond to the *smallest value* of:

$$J = \frac{1}{N} \sum_{i=1}^{N} \left(M_i - mS_i - c\right)^2$$

- The solution in this case is more complicated than before since there are two unknown parameters $m$ and $c$.
- The idea is the same though, but now we find the values of $m$ and $c$ that *simultaneously* correspond the case where the gradient of $J$ vanishes, *i.e.*:

$$\frac{dJ}{dm} = 0; \quad \frac{dJ}{dc} = 0$$
Linear Regression: Melbourne vs Sydney Temperatures (1859-2013)

- R can very conveniently solve this problem for us!
- The resulting least-squares solution gives $m=0.5$ and $c=0$.
- So the resulting “least-squares best fit straight line” is given by:

$$M = 0.5S$$
Linear Regression: Melbourne vs Sydney Temperatures (1859-2013)

• The “scatter” of the measurements about the straight line \( M=0.5S \) tells us a lot about how robust the straight line relationship between Sydney and Melbourne temperatures is.

• The gradient of the straight line \( m \) and the intercept \( c \) can be treated as random numbers described by normal probability density functions (PDF).

• The standard deviation of the PDFs for \( m \) and \( c \) tell us how reliable our estimates of \( m \) and \( c \) really are.

• The standard deviation for \( m \) is 0.05.

• The standard deviation for \( c \) is 0.04.

• So we can write \( m=0.5\pm0.05 \) and \( c=0\pm0.04 \).

• The standard deviations are both quite small, indicating that our estimates of \( m \) and \( c \) are reliable.
Linear Regression: Melbourne vs Sydney Temperatures (1859-2013)

- The variance of Melbourne temperature $\sigma^2=0.36°C^2$ (see Modules 1 & 4).
- However, if we use the equation $M=0.5S$ to “predict” Melbourne temperature based on Sydney temperature, the variance of the predicted Melbourne temperature is only $0.13°C^2$.
- The fraction of the actual Melbourne temperature variance explained by the equation $M=0.5S$ is $0.13/0.36=0.36$, or 36%. This statistic is called the coefficient of determination $r^2$.
- The coefficient of determination is simply the squared correlation coefficient $r$ between the two time series.
- For $M$ and $S$: $r=0.60$ and $r^2=(0.6)^2=0.36$. 

“least-squares best fit straight line
## Linear Regression: Melbourne vs Sydney Temperatures (1859-2013)

### Summary:

- The coefficient of determination is $r^2 = 0.36$
- The correlation coefficient between Sydney and Melbourne temperature is $r = 0.6$.

\[ M = 0.5S \]

“least-squares best fit straight line”
Important Concepts

• The equation of a straight line.
• The method of “least-squares”
• The least-squares best fit straight line that passes through the measurements
• Linear regression
• The coefficient of determination.
BREAK!
Part II

- More on linear regression
- Hypothesis testing
- Climate change
Data that Approximate a Straight Line

• Suppose that $x$ and $y$ are two climate variables (e.g. CO$_2$ concentration and global average temperature).
• If the two variables fall on a straight line, then it is likely (although not guaranteed) that they are related in some way.
Linear Regression: The SOI vs the Niño3.4 index

- Recall from Module 3 that the Southern Oscillation Index (SOI) is a useful indicator of El Niño and La Niña conditions in the atmospheric circulation in the tropical Pacific.
- Similarly, the Niño3.4 Index is a useful indicator of El Niño and La Niña conditions in the ocean circulation in the tropical Pacific.
Linear Regression: The SOI vs the Niño3.4 index

- Suppose that we now treat the Niño3.4 Index and the SOI as coordinates, and plot corresponding values of each that occur at the same time as individual points.
- This yields a scatter plot, just like the one we encountered before.
- In this case, the points form a cloud that slopes upward from the lower right to the upper left.
- This suggests that the Niño3.4 Index and the SOI are related.
- Of course we expect that they are related, because they describe different aspects of the same phenomenon, the El Niño Southern Oscillation (ENSO).
- Nevertheless, let’s perform a linear regression analysis to see what we can learn.
- Can we use the Niño3.4 Index as a predictor of SOI?
Linear Regression: The SOI vs the Niño3.4 index

- We will treat the Niño3.4 Index as the **predictor** and the SOI as the **predictand**.
- Using **linear regression** we are trying to find the “**best**” straight line relationship of the form:
  \[ \text{SOI} = m \cdot \text{Nino3.4} + c \]

  where \( m \) is the slope (or gradient) of the straight line, and \( c \) is the intercept.

- Before we actually do the analysis, let’s see what we might expect to find.
- **What do you expect the sign of the slope (gradient) \( m \) to be?**
- **What do you expect the sign of the intercept \( c \) to be?**
- **What is the criterion that we must use to identify the “**best**” straight line that passes through the cloud of data points?**
Linear Regression: The SOI vs the Niño3.4 index

The “least-squares best fit straight line” is shown in the figure, where $m=-1.36$ and $c=0.26$.

In other words,

$$\text{SOI} = -1.36 \cdot \text{Nino3.4} + 0.26$$

Taking into account the expected errors in $m$ and $c$ we find that $m=-1.36\pm0.05$, and $c=0.26\pm0.04$.

The standard deviation of $m$ is small (3%) so the estimate of $m$ is reliable.

The standard deviation of $c$ is fairly small (15%) so the estimate of $c$ is reliable.

The coefficient of determination is $r^2=0.52$, indicating that the equation above accounts for 52% of the variance in SOI.

The correlation $r=+0.72$ or -0.72: which should we choose?

Do you remember the definition of the coefficient of determination?
A Scientific Hypothesis for the Relationship between SOI and Niño3.4

SOI = -1.36 \cdot \text{Nino3.4} + 0.26

What is the scientific explanation for this relationship?

**Positive SOI**

**Negative Niño3.4 Index**

**La Niña Conditions**

Darwin

Tahiti

Weak or reversed trade winds

Weak upwelling

Warm SST

Negative SOI

Positive Niño3.4 Index

**El Niño Conditions**

Darwin

Tahiti

Walker Circulation

Weak or reversed trade winds

Weak upwelling

Warm SST

Positive SOI

Negative Niño3.4 Index

**Positive Niño3.4 Index**

**La Niña Conditions**

Darwin

Tahiti

Walker Circulation

Strong trade winds

Vigorous upwelling

Cold SST
Hypothesis Testing

• In the previous examples of linear regression, the relationship between the predictor and the predictand is well understood.
• Sydney and Melbourne are on the same continent and fairly close together.
• SOI and the Niño3.4 Index correspond to shifts in the Walker Circulation and associated changes in ocean temperatures.
• Recall the equation of a straight line: $y=mx+c$.
• Suppose that $x$ and $y$ are two variables for which the connection is not well understood.
• If linear regression indicates that $x$ and $y$ are related, the challenge becomes one of understanding the cause and effect relationship between them.
• The equation $y=mx+c$ does not, in itself, imply cause and effect.
Hypothesis Testing

• To establish *cause and effect*, we must first pose a scientific hypothesis that explains why the two variables $x$ and $y$ are related.
• Then we must seek to prove or disprove our hypothesis.

• *Statistics can be misleading...*
Statistics can be misleading: This is a Tale of... 

... Storks 

... and babies
A Tale of Storks and Babies

- For centuries, storks have long been associated with babies and family life.
- According to Greek mythology, the Goddess Hera, the wife of Zeus, turned her rival into a stork who then tried to steal Hera’s son.
- In Egyptian mythology, the soul is represented as a stork, and return of a stork symbolizes the return of a soul.
- Hans Christian Andersen popularized storks in a fable where storks plucked sleeping babies from a pool where they lay dreaming, and delivered them to families with good children.
“New evidence for the Theory of the Stork”
Thomas Höfer, Hildegard Przyrembel, Silvia Verleger
_Pediatric and Perinatal Epidemiology_

**Summary**
Data from _Berlin_ (Germany) show a significant correlation between the increase in the stork population around the city and the increase in deliveries outside city hospitals (out-of-hospital deliveries). However, there is no correlation between deliveries in hospital buildings (clinical deliveries) and the stork population. The decline in the number of pairs of storks in the German state of _Lower Saxony_ between 1970 and 1985 correlated with the decrease of deliveries in that area. The nearly constant number of deliveries from 1985 to 1995 was associated with an unchanged stork population (no statistical significance). However, the relevance of the stork for the birth rate in that part of Germany remains unclear, because the number of out-of-hospital deliveries in this area is not well documented. A lack of statistical information on out-of-hospital deliveries in general is a severe handicap for further proof for the Theory of the Stork.
Storks in Brandenburg and the birthrates in Berlin, Germany (1990–99). Open triangles show number of clinical deliveries per year in Berlin. Open diamonds show number of out-of-hospital deliveries per year in Berlin. Number of pairs of storks are shown as full squares. Dotted lines represent linear regression trend ($y = mx + b$). For the convenience of the readers, two figures are presented. Left graph shows clinical deliveries against pairs of storks using two scalings, right graph shows numbers of out-of-hospital deliveries and pairs of storks both on the same scale. In both figures, data are from the years 1990–2000.

Could it be that storks have trouble gaining access to hospital wards?
Post-War Copenhagen  
(Dr. Gustav Fischer, Ornithologist)

- Number of babies born = $B$ (the predictand).
- Number of nesting storks = $S$ (the predictor).
- Linear regression yields: 
  $$B = 0.15S + 36$$
- The coefficient of determination $r^2 = 0.91$, meaning that 91% of the variance in the number of babies can be explained by the number of storks!
- The correlation between $S$ and $B$ is 0.95!
- The conclusion...
The Conclusion

Clearly this is absurd!
Hypotheses

• **Hypothesis #1**: The presence of storks in Copenhagen is directly responsible for the number of babies born. **Unlikely**

• **Hypothesis #2**: There is some other “hidden” variable that accounts for the high correlation between the birth rate and the number of nesting storks. **Most likely**

**The real reason**: Post-war adult migration to Copenhagen and the resulting building boom in the city led to a rise in the birth rate, and an increase in desirable nesting habitat for storks, which like to nest atop of tall buildings.
Global Climate Change: CO₂ Concentration

- Direct measurements of atmospheric CO₂ concentration have been made at **Mauna Loa Observatory** in Hawaii since 1958.
- This is the *longest* time series of direct measurements of CO₂.
- Since CO₂ is fairly well mixed in the atmosphere, we will take the concentration at Mauna Loa as representative of the global average concentration.
- During the last 55 years, CO₂ has increased from 316 ppm to ~400 ppm.
- This is an increase of 25%.
Mauna Loa Observatory, Hawaii
Global Climate Change: Global Average Temperature

- The figure shows a time series of the *global annual mean surface temperature* 1958-2013.
- During this 55 year period the *global annual mean surface temperature* has increased by approximately 0.5°C.
To what extent is the rise in global average temperature due to increasing greenhouse gas concentrations (if at all), such as CO$_2$?
Linear Regression: $\text{CO}_2$ Concentration vs Global Mean Temperature

- Suppose that we now treat $\text{CO}_2$ concentration and global mean temperature as coordinates, and generate the \textit{scatter plot} on the left.
- In this case, the points form a cloud that slopes upward from the lower left to the upper right.
- This suggests that $\text{CO}_2$ concentration and global mean temperature are related.
- If we treat $\text{CO}_2$ \textit{concentration} as the \textit{predictor} and \textit{global mean temperature}, $T$, as the \textit{predictand}, linear regression yields:

  \[ T = m \cdot \text{CO}_2 + c \]

  where $m=0.0095\pm0.0005$ and $c=10.92\pm0.18$.
- Based on the small standard deviations for $m$ and $c$ (5% and 1% respectively), we conclude that this relationship is robust.

\[ \text{Coefficient of determination: } r^2=0.87 \]
(87% of variance in $T$ can be explained by variations in $\text{CO}_2$ conc.)

\[ \text{Correlation coefficient: } r = 0.93 \]
Hypothesis:
The rise in global mean surface temperature is due primarily to increasing concentrations of carbon dioxide (and other greenhouse gases) in the atmosphere.

Is there a scientific basis for this hypothesis?

Coefficient of determination: $r^2 = 0.87$
(87% of variance in $T$ can be explained by variations in CO$_2$ conc.)

Correlation coefficient: $r = 0.93$
Important Concepts

• Linear regression and hypothesis testing
• Statistics can be misleading
Assignment #5
& optional Extra Credit (+5%)
Due July 28th by 4pm
NO EXCEPTIONS!!!!!