Announcements

- HW 5 due Thursday, 11/1
- Lab 7 due Sunday, 11/4
- Project Proposal due 11/13
  - Example posters in office
Packages

```r
library(knitr)
library(broom)
library(dplyr)
library(tibble)
library(ggplot2)
library(cowplot)
```
Examples of Time Series Data
Gas Prices

120 Month Average Retail Price Chart

Regular Gas Price (US $/G)

Date (Month/Day)

USA Average
Durham

gasbuddy.com

©2018 GasBuddy.com
Global Land-Ocean Temperature Index

Source: climate.nasa.gov

NASA.gov
Stocks

Apple's Stock Price
Popular Music

Google Music Timeline
Time Series

- Regression assumes independent errors across observations.

- When data is ordered over time, errors in one period may influence error in another period.
  - Called time series data.

- Assume observations measured at equally spaced time points.

- We will do a brief introduction to time series analysis.
  - Take STA 444: Statistical Modeling of Spatial and Time Series Data for more in-depth study of the subject.
Example: Detecting Melanoma

- Incidence of melanoma (skin cancer) is related to solar radiation

- **Question:** Is there evidence that melanoma incidence is related to sunspot activity in the same year or to sunspot activity in the previous year?

- **Data:** Age-adjusted melanoma incidence among males from the Connecticut Tumor Registry 1936 - 1972 and annual sunspot activity
  - ex1514 data in Sleuth3 package
Example: Detecting Melanoma

- **Year:** 1936-1972

- **Melanoma:** Age-adjusted melanoma incidence per 100,000 males

- **Sunspot:** Measure of annual sunspot activity
Example: Detecting Melanoma

```
library(Sleuth3)
cancer_sun <- ex1514
glimpse(cancer_sun)
```

## Observations: 37
## Variables: 3
## $ Year <int> 1936, 1937, 1938, 1939, 1940, 1941, 1942, 1943, 1944,...
## $ Melanoma <dbl> 1.0, 0.9, 0.8, 1.4, 1.2, 1.0, 1.5, 1.9, 1.5, 1.5, 1.5,...
## $ Sunspot <int> 40, 115, 100, 80, 60, 40, 23, 10, 10, 25, 75, 145, 13,...
Data Exploration

Melanoma vs. Sunspot

Sunspot vs. Year

Melanoma vs. Year
Data Exploration

Melanoma vs. Sunspot

Sunspot vs. Year

Melanoma vs. Year
Melanoma vs. Sunspot

```r
model <- lm(Melanoma ~ Sunspot, data=cancer_sun)
kable(tidy(model), format="markdown", digits=3)
```

<table>
<thead>
<tr>
<th>term</th>
<th>estimate</th>
<th>std.error</th>
<th>statistic</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>2.522</td>
<td>0.371</td>
<td>6.796</td>
<td>0.000</td>
</tr>
<tr>
<td>Sunspot</td>
<td>0.003</td>
<td>0.004</td>
<td>0.715</td>
<td>0.479</td>
</tr>
</tbody>
</table>

```r
glance(model)$r.squared
```

```r
## [1] 0.01439734
```
Residuals vs. Sunspot

No concerning pattern
Residuals vs. Year

serial correlation!
Measuring of Serial Correlation

- We can compute the correlation in residuals at time $t$ and time $t - k$

- $k$: how far back you wish to go
  - This is called the lag

- Example: Lag $k$ autocorrelation coefficient:

$$r_k = \frac{c_k}{c_0} = \frac{\sum_{t=1+k}^{n} (\text{res}_t \times \text{res}_{t-k})/(n - 1)}{\sum_{t=1}^{n} \text{res}_t^2/(n - 1)}$$
Autocorrelation

- We will focus on Lag 1 autocorrelation, i.e. \( k = 1 \)

\[
r_1 = \frac{c_1}{c_0} = \frac{\sum_{t=2}^{n} (\text{res}_t \times \text{res}_{t-1})/(n - 1)}{\sum_{t=1}^{n} \text{res}_t^2/(n - 1)}
\]

- \( c_1 \) and \( c_0 \) are estimates of **autocovariance**: the covariance between the response variable and itself at two time points

- \( r_1 \) is an estimate of the **autocorrelation**: the correlation between residuals at time \( t \) and time \( t - 1 \)

  - \(-1 \leq r_1 \leq 1\)
# Melanoma vs. Sunspots - Autocorrelation

<table>
<thead>
<tr>
<th>Year</th>
<th>resid_lag1</th>
<th>resid_current</th>
</tr>
</thead>
<tbody>
<tr>
<td>1937</td>
<td>-1.644</td>
<td>-1.974</td>
</tr>
<tr>
<td>1938</td>
<td>-1.974</td>
<td>-2.028</td>
</tr>
<tr>
<td>1939</td>
<td>-2.028</td>
<td>-1.367</td>
</tr>
<tr>
<td>1940</td>
<td>-1.367</td>
<td>-1.506</td>
</tr>
<tr>
<td>1941</td>
<td>-1.506</td>
<td>-1.644</td>
</tr>
<tr>
<td>1942</td>
<td>-1.644</td>
<td>-1.093</td>
</tr>
</tbody>
</table>

```r
#calculate autocorrelation
sum(resid_current * resid_lag1) / sum(cancer_sun$Residuals^2)
```

## [1] 0.8597797

**What is one way to account for the year-to-year impact?**
# Add Year to the model
model_v2 <- lm(Melanoma ~ Sunspot + Year, data=cancer_sun)
kable(tidy(model_v2), format="html", digits=3)

<table>
<thead>
<tr>
<th>term</th>
<th>estimate</th>
<th>std.error</th>
<th>statistic</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-225.108</td>
<td>13.257</td>
<td>-16.981</td>
<td>0.00</td>
</tr>
<tr>
<td>Sunspot</td>
<td>0.001</td>
<td>0.001</td>
<td>1.074</td>
<td>0.29</td>
</tr>
<tr>
<td>Year</td>
<td>0.117</td>
<td>0.007</td>
<td>17.172</td>
<td>0.00</td>
</tr>
</tbody>
</table>

glance(model_v2)$r.squared

## [1] 0.8981024
Melanova vs. Sunspots & Year

Residuals vs. Sunspot

Residuals vs. Year

Sunspot

Year
Autocorrelation

Residuals vs. Year

# [1] "autocorrelation: 0.377"
Autoregressive Model

- There are many models that account for serial correlation in the error terms.

- Common model is the autoregressive (AR) model.

- If we have no explanatory variables, the AR model with one lag (AR(1) model) is

\[ Y_t = \beta_0 + \beta_1 Y_{t-1} + \epsilon_t \quad \epsilon_t \sim N(0, \sigma^2) \]
Example: AR(1) Model

<table>
<thead>
<tr>
<th>Year</th>
<th>Melanoma</th>
<th>melanoma_prev</th>
</tr>
</thead>
<tbody>
<tr>
<td>1937</td>
<td>0.9</td>
<td>1.0</td>
</tr>
<tr>
<td>1938</td>
<td>0.8</td>
<td>0.9</td>
</tr>
<tr>
<td>1939</td>
<td>1.4</td>
<td>0.8</td>
</tr>
<tr>
<td>1940</td>
<td>1.2</td>
<td>1.4</td>
</tr>
<tr>
<td>1941</td>
<td>1.0</td>
<td>1.2</td>
</tr>
<tr>
<td>1942</td>
<td>1.5</td>
<td>1.0</td>
</tr>
</tbody>
</table>
Example: AR(1) Model

```r
model_ar1 <- lm(Melanoma ~ melanoma_prev, data=melanoma_data)
kable(tidy(model_ar1), format="html", digits=3)
```

<table>
<thead>
<tr>
<th>term</th>
<th>estimate</th>
<th>std.error</th>
<th>statistic</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.216</td>
<td>0.186</td>
<td>1.164</td>
<td>0.253</td>
</tr>
<tr>
<td>melanoma_prev</td>
<td>0.964</td>
<td>0.063</td>
<td>15.317</td>
<td>0.000</td>
</tr>
</tbody>
</table>

```r
glance(model_ar1)$r.squared
```

```r
## [1] 0.8734264
```
Residual Plots

Residuals vs. Melanoma

Residuals vs. Year

## [1] "autocorrelation: -0.145"
Next Steps:

- We created a model that explains a lot of the variation in melanoma ($\approx 87\%$)

- We made a big improvement in terms of reducing serial correlation ($\approx 0.86$ to $\approx -0.15$)

- However, the model we created isn't ideal... why not?
AR(1) Model: One Explanatory variable

- If we want to use an explanatory variable, the AR(1) model takes the general form:

\[ Y_t = \beta_0 + \beta_1 X_t + \epsilon_t \]
\[ \epsilon_t = \alpha \epsilon_{t-1} + \delta_t, \quad \delta_t \sim N(0, \sigma^2) \]

- \( \alpha \) is the autocorrelation
  - We can estimate \( \alpha \) using \( r_1 \)
Example: Use sunspot from previous year

<table>
<thead>
<tr>
<th>Sunspot</th>
<th>sunspot_prev</th>
</tr>
</thead>
<tbody>
<tr>
<td>115</td>
<td>40</td>
</tr>
<tr>
<td>100</td>
<td>115</td>
</tr>
<tr>
<td>80</td>
<td>100</td>
</tr>
<tr>
<td>60</td>
<td>80</td>
</tr>
<tr>
<td>40</td>
<td>60</td>
</tr>
<tr>
<td>23</td>
<td>40</td>
</tr>
</tbody>
</table>
Example: Melanoma vs. sunspot_prev

```r
model_lag1 <- lm(Melanoma ~ sunspot_prev, data=lag1_data)
kable(tidy(model_lag1), format="html", digits=3)
```

<table>
<thead>
<tr>
<th>term</th>
<th>estimate</th>
<th>std.error</th>
<th>statistic</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>2.398</td>
<td>0.363</td>
<td>6.604</td>
<td>0.000</td>
</tr>
<tr>
<td>sunspot_prev</td>
<td>0.006</td>
<td>0.004</td>
<td>1.330</td>
<td>0.192</td>
</tr>
</tbody>
</table>

```r
glance(model_lag1)$r.squared
```

## [1] 0.04943778
Example: Melanoma vs. sunspot_prev

# [1] "autocorrelation: 0.841"
Example: Add Year to the Model

```r
# Control for year
model_lag1_v2 <- lm(Melanoma ~ sunspot_prev + Year, data=lag1_data)
kable(tidy(model_lag1_v2), format="html", digits=3)

<table>
<thead>
<tr>
<th>term</th>
<th>estimate</th>
<th>std.error</th>
<th>statistic</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-226.916</td>
<td>12.445</td>
<td>-18.234</td>
<td>0.000</td>
</tr>
<tr>
<td>sunspot_prev</td>
<td>0.004</td>
<td>0.001</td>
<td>3.065</td>
<td>0.004</td>
</tr>
<tr>
<td>Year</td>
<td>0.117</td>
<td>0.006</td>
<td>18.428</td>
<td>0.000</td>
</tr>
</tbody>
</table>

glance(model_lag1_v2)$r.squared

## [1] 0.9158057
```
Residual Plots

Residuals vs. Sunspot from Previous Year

Residuals vs. Year

---

# [1] "autocorrelation: 0.165"
Practice: Lynx Trappings vs. Sunspots

- ex1515 in Sleuth3 package
  - **Year:** 1821 - 1934
  - **Lynx:** Number of lynx trapped
  - **Sunspot:** Measure of sunspot activity

- **Question:** Is there evidence that the number of lynx trapped are related to sunspot activity?