- **THIS IS WHAT IT LOOKS LIKE WHEN YOU LAUNCH stat200 2-10-09**

stat200 2 - 10 - 09

A number of statistical routines are programmed into this Mathematica notebook file. To run them you must boot the notebook from a university lab as follows:

a. navigate to www.stt.msu.edu/~lepage

b. click on the (folder) STT200

c. click on the (program) stat200 2 - 10 - 09 (stat200 2 - 10 - 09 will launch)

d. click on the 1 + 1 line just below

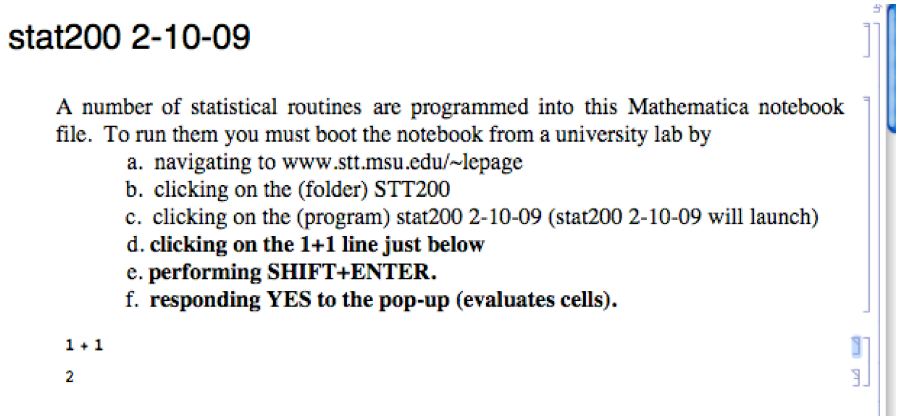
e. perform SHIFT + ENTER.

f. respond YES to the pop - up (evaluates cells).

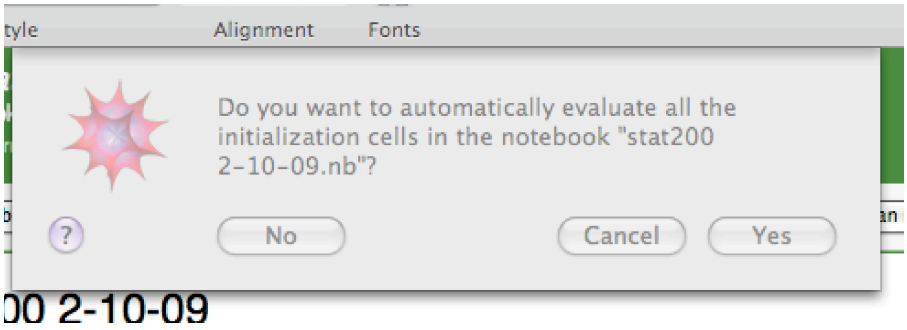
```
1 + 1
```

2

- **THIS IS WHAT IT LOOKS LIKE WHEN YOU CLICK ON THE LINE 1 + 1 AND THEN HOLD SHIFT KEY WHILE ENTERING THE RETURN KEY. NOTE THE DARKENED BRACKET TO THE RIGHT OF 1 + 1 WHICH INDICATES THAT MATHEMATICA IS WORKING ON THE CALCULATION.**



- **HERE IS THE POP - UP ASKING WHETHER YOU WISH TO EVALUATE. CLICK YES.**



- **HERE IS WHAT IT LOOKS LIKE WHEN EVALUATION IS COMPLETED. THE 1 + 1 LINE AND ITS ANSWER LINE 2 NOW HAVE NUMBERS AT THE LEFT.**

stat200 2-10-09

A number of statistical routines are programmed into this Mathematica notebook file. To run them you must boot the notebook from a university lab as follows:

a. navigate to www.stt.msu.edu/~lepage
b. click on the (folder) STT200
c. click on the (program) stat200 2-10-09 (stat200 2-10-09 will launch)
d. **click on the 1+1 line just below**
e. **perform SHIFT+ENTER.**
f. **respond YES to the pop-up (evaluates cells).**

In[4]:= **1 + 1**

Out[4]= 2

■ HERE ARE SOME OF THE COMMANDS SHOWN IN THE MATHEMATICA FILE. THE DEAL WITH MULTIPLE LINEAR REGRESSION (MLR).

regrtable[x,y] returns a table illustrating calculations of $\overline{x}$, $\overline{y}$, $\overline{x^2}$, $\overline{y^2}$, $\overline{xy}$.

regrstats[x, y] returns $\overline{x}$, $\overline{y}$, s_x, s_y, r, and the slope of the regression line

$$r \frac{s_y}{s_x} = r \frac{\hat{\sigma}_y}{\hat{\sigma}_x} .$$

regrplot[x,y] returns the plot of (x, y) pairs overlaid with the regression line.

betahat0[list x, list y] returns the least squares intercept and slope for a straight line fit of the model $y = b_0 + b_1 x + \epsilon$.

betahat[matrix x, list y] returns the least squares coefficients $\hat{\beta}$ for a fit of the matrix model $y = x\beta + \epsilon$.

resid0[list x, list y] returns the estimated errors $\hat{\epsilon} = y - x\hat{\beta}$ (see **betahat0** above).

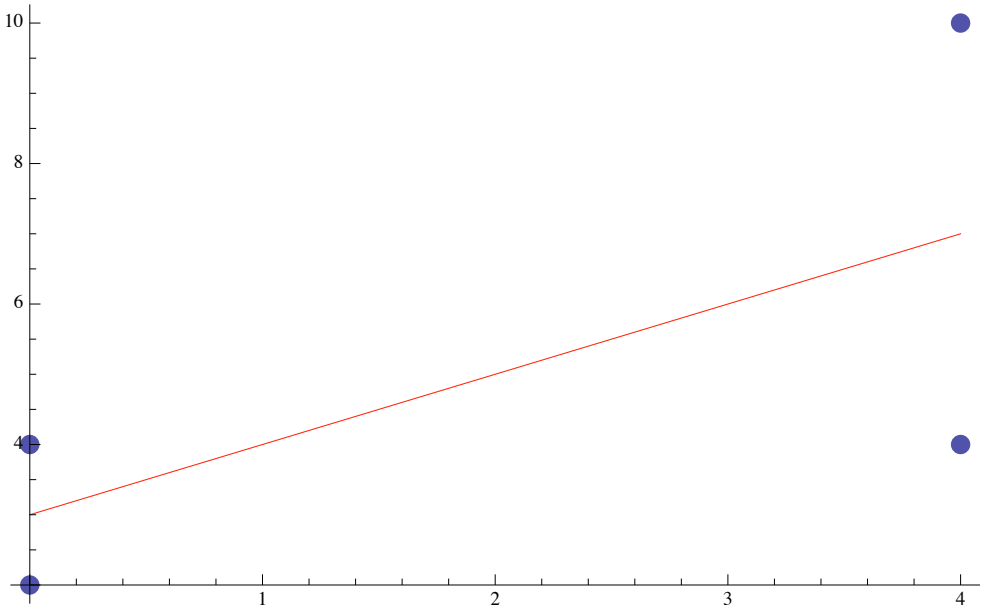
resid[matrix x, list y] returns the estimated errors $\hat{\epsilon} = y - x\hat{\beta}$ (see **betahat** above).

R[matrix x, list y] returns the **multiple correlation** R between the fitted values $x\hat{\beta}$ and data y. in the matrix model setup.

■ HERE IS A FAMILIAR STRAIGHT LINE REGRESSION ANALYSIS.

```
toyx = {0, 0, 4, 4}
{0, 0, 4, 4}
toyy = {2, 4, 4, 10}
{2, 4, 4, 10}
```

```
regrplot[toyx, toyy]
```




```
betahat0[toyx, toyy]
```

```
{3., 1.}
```

```
resid0[toyx, toyy]
```

```
{-1., 1., -3., 3.}
```

```
regrstats[toyx, toyy]
```

```
{2., 5., 2.3094, 3.4641, 0.666667, 1.}
```

- **Here is the same toy example, but set up as a multiple linear regression.**

```
toymatrixx = {{1, 0}, {1, 0}, {1, 4}, {1, 4}}
```

```
{{1, 0}, {1, 0}, {1, 4}, {1, 4}}
```

```
betahat[toymatrix, toyy]
```

```
{3., 1.}
```

```
resid[toymatrixx, toyy]
```

```
{-1., 1., -3., 3.}
```

```
R[toymatrixx, toyy]
```

```
0.666667
```

Multiple correlation R generalizes correlation r to the case of more than one independent variable. It is defined to be the correlation of the fitted values $\hat{y}$ with the dependent variable y. R always ranges in [0, 1]. In the straight line case we have the relation R[$\hat{y}$, y] = | r[x,y] |. Here is an example using y_i = 2009 tax of i-th subject as dependent variable in a multiple linear regression on independent variables $x_{1\,i}$= 2008 tax and $x_{2\,i}$= 2007 tax for subject i. If we use the model

$$y = b_0 + b_1 x_1 + b_2 x_2 + \text{error}$$

then there are three variables (constant variable 1 is included in the model). This is also a toy example. We have only four subjects (properties) and there are three model coefficients being estimated.

```
taxx = {{1, 4354, 4645}, {1, 7226, 8349},
   {1, 1278, 1344}, {1, 6628, 6981}}
```

```
{{1, 4354, 4645}, {1, 7226, 8349},
 {1, 1278, 1344}, {1, 6628, 6981}}
```

```
MatrixForm[taxx]
```

$$\begin{pmatrix} 1 & 4354 & 4645 \\ 1 & 7226 & 8349 \\ 1 & 1278 & 1344 \\ 1 & 6628 & 6981 \end{pmatrix}$$

```
taxy = {4878, 8654, 1693, 7446}
```

```
{4878, 8654, 1693, 7446}
```

```
betahat[taxx, taxy]
```

```
{271.079, 0.198682, 0.830957}
```

```
resid[taxx, taxy]
```

```
{-117.934, 9.58735, 51.1993, 57.1478}
```

```
R[taxx, taxy]
```

```
0.999651
```

- **The following matrix contains quantities needed to estimate margins of error for each of b_0, b_1, b_2.**

```
MatrixForm[betahatCOV[taxx, taxy]]
```

$$\begin{pmatrix} 29\,253.9 & -20.6757 & 14.342 \\ -20.6757 & 0.100653 & -0.0881197 \\ 14.342 & -0.0881197 & 0.0778522 \end{pmatrix}$$

Suppose the sample size n = 4 was instead n = 400. If the errors in the y data (departures from the model) were independent samples from a normal distribution with mean 0 then we would be entitled to employ the following 95% CI for each of the coefficients b_0, b_1, b_2.

```
In[1]:= 271.079 + {-1, 1} 1.96 Sqrt[29 253.9]
Out[1]= {-64.1549, 606.313}

In[2]:= 0.198682 + {-1, 1} 1.96 √0.100653
Out[2]= {-0.423145, 0.820509}

In[3]:= 0.830957 + {-1, 1} 1.96 √0.0778522
Out[3]= {0.284078, 1.37784}
```

The CI above are uselessly WIDE. Had they really been formed out of n = 400 samples (instead of only n = 4) they would have been far narrower.

It is important to realize that the above CI do not outwardly display the role of n. It is instead the case that the role of n is concealed inside the diagonal numbers 29253.9, 0.100653, 0.0778522 of the above 3 by 3 array.

I suggest that you get into a lab, such as B100 Wells, at the earliest convenience.

See if you can reproduce some of these calculations.

We'll work some examples, live, in lecture this Friday.