STA 302/1001

Note Title 10/11/2011

Reminder Test is Tues Oct. 25

Regression

Origin of term: Galton, 19th c.

Predicting son's heights from father's heights

$$\widehat{\text{son's height}} = 26.46 + 0.61152 \text{ father's height}$$

Not $y = x$

At 68.1", model predicts son's height = father's height

Below that: sons of short (tall) fathers tend to be short (tall), but not as short as their father (tall)

Galton: "Regression to the Mean"

In any test – re-test (x) (y), expect to get a slope < 1

"the regression effect"

Why? $b_1 = r \frac{s_y}{s_x}$ if $s_y \approx s_x$

since $|r| \leq 1$

$\Rightarrow |b_1| \leq 1$

Example STA 302, in a previous year there were 2 tests (T1, T2)

For students who wrote both tests, use T1 to predict T2

Is T2 = T1 a reasonable guess? NO

	Mean	Sd
T1	62.7	18.4
T2	62.1	18.1

$$\hat{T2} = 20.27 + .667\ T1$$

If got 20% on T1, $\hat{T2} = 34$

40% on T1, $\hat{T2} = 47$

75% on T1, $\hat{T2} = 70$

90% on T1, $\hat{T2} =$ ___

On average, students who score low (high) on 1st test will score higher (lower) on second test.

Why? Students with a high T1, are there because:

- that's their skill level
- random variation caused them to score above their skill level

The regression effect happens because there are more students with skill levels near the mean than away from it.

The regression fallacy occurs when the regression effect is mistaken for a real effect

Examples - Study that enrols people with extreme values of some measurement and sees how they respond to a treatment

Expect extremes to be closer to mean when re-measured.

(Solution: use a control group)

- "While overall no effect, people whose [redacted] baseline measurement

was high, tended to go down and
... "

Anscombe example: Plot the data
- evidence of a linear relationship from the numbers is not enough

Chapter 3: Diagnostics and Transformations for Simple Linear Regression

From preface: "It makes sense to base inferences or conclusions only on valid models."

Influential Points / Outliers / Leverage Points

Observations whose inclusion or exclusion result in substantial changes to the fitted model (set of coefficients, predicted values) are said to be INFLUENTIAL (eg. NYC in crime example)

Points can be outlying in any (or all or some) of the value of the explanatory variable,

the dependent variable, or its residual.

Outliers w.r.t. the residuals represent model failure — line doesn't fit that point adequately