Mathematics 231

Lecture 10

Liam O'Brien

1

Announcements

- Reading
  - Today — M&M 2.5 — 142-151
  - Next class — M&M 2.3 — 119-121
    M&M 2.4 — 125-132
    Supplemental — Regression to the Mean

2

Regressions Gone Bad

- Outliers and influential observations (not the same thing)
- Extrapolation
- Aggregation: Ecological Correlations

3

Outliers and Influential Points

- **Outlier:** In regression, a point that lies far from the fitted line, often producing a large residual.
- **Influential point:** A point whose removal would markedly change the position of the regression line.

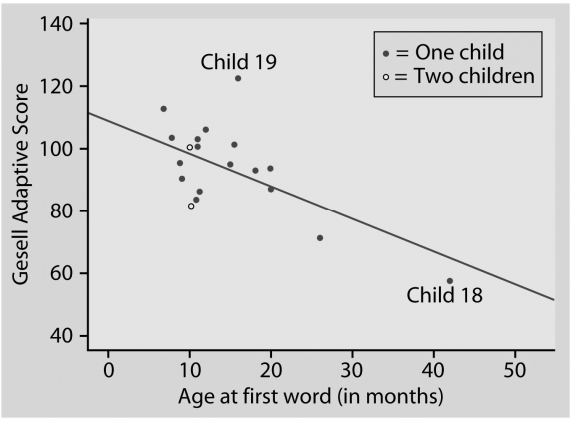
4

Example: Cognitive Ability in Children

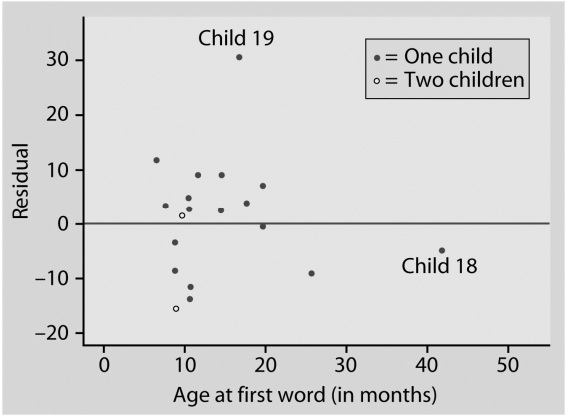
TABLE 2.9 Age (in months) at first word and Gesell score

Child	Age	Score	Child	Age	Score
1	15	95	12	9	96
2	26	71	13	10	83
3	10	83	14	11	84
4	9	91	15	11	102
5	15	102	16	10	100
6	20	87	17	12	105
7	18	93	18	42	57
8	11	100	19	17	121
9	8	104	20	11	86
10	20	94	21	10	100
11	7	113			

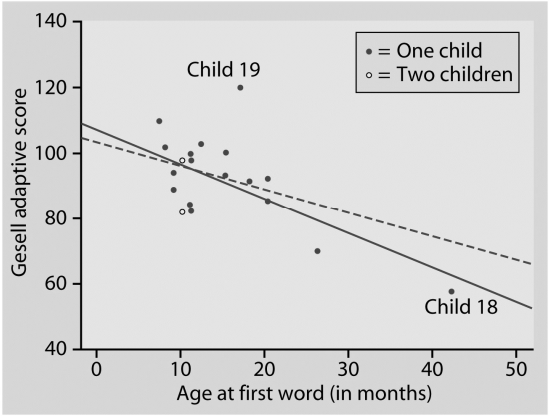
5



6



7



8

Example: Cognitive Ability in Children

- Child 18 is an outlier in the X direction – it is an influential point.
- Child 19 is an outlier in the Y direction.
- The regression line is more likely to be heavily influenced by an outlier in the X direction.
- Regression including Child 18: $R^2 = 0.41$
- Regression excluding Child 18: $R^2 = 0.11$

9

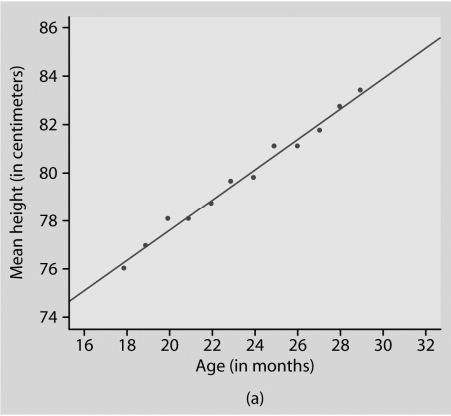
Extrapolation

- **Extrapolation:** Predicting Y for values of X outside the range of the observed data.
- Example: Heights in cm of children 18-32 months.

  Linear regression of height on age:

  height = 64.9 + 0.635 (age)
- Can we use this regression model to predict the height for a 25-year old woman?

10



11

Example: Heights of Children

- Example: Heights in cm of children 18-32 months.

  Linear regression of height on age:

  height = 64.9 + 0.635 (age)
- Can we use this regression model to predict the height for a 25-year old woman?

  height = 64.9 + 0.635 (25*12) = 255cm

  8'4"

12

Aggregation

- **Aggregation:** Associations based on averaged data.
- **Problem:**
  - Averaging smoothes out the variation resulting in very high correlations (and R^2)
  - Trend in averaged data cannot be extrapolated to individual data.

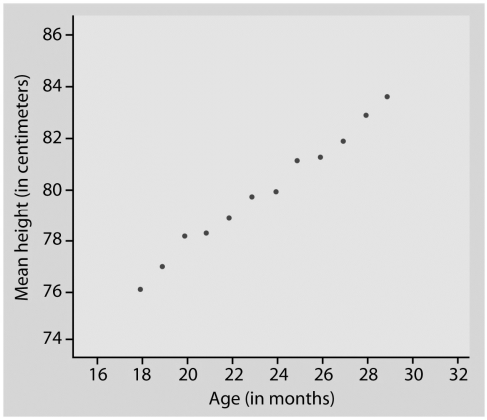
13

Example: The Kalama Kids…

TABLE 2.7 Mean height of Kalama children

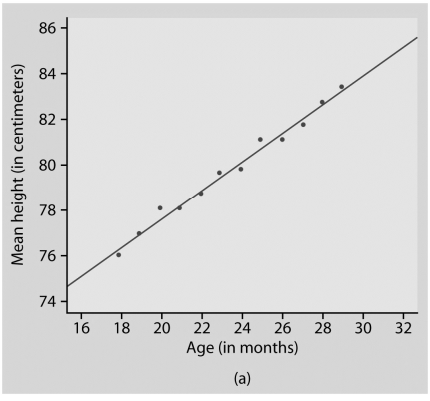
Age *x* in months	Height *y* in centimeters
18	76.1
19	77.0
20	78.1
21	78.2
22	78.8
23	79.7
24	79.9
25	81.1
26	81.2
27	81.8
28	82.8
29	83.5

14



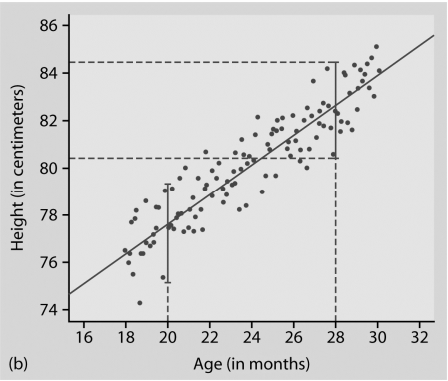
15

$R^2 = 0.989$



16

$R^2 = 0.849$



17

Aggregation

- The scatterplot of just the averages hides much of the variability in the data.
- In general, regression models that use aggregate data overstate the strength of the association (larger R^2).
- Also, combining different groups can lead to results that differ from the results that would be obtained by studying each group separately.

18

Ecological Fallacy

- **Ecological Fallacy:** Assuming relationship observed for aggregate data (groups) necessarily holds for individuals.
- Aggregate data is often easier to obtain than individual-level data and might offer clues about individual trends.
- At best though, aggregate data can only weakly support inferences about individual data.

19

Example: Fat Intake and Breast Cancer

- In countries where there is a high fat diet that dominates, the breast cancer death rates are higher.
- So women who eat more fatty foods are at a higher risk for breast cancer, right?
- Studies using individual-level data do not support an association between the two.

20

Example: Economists' Salaries

- Strong positive correlation exists between years of education and salary for economists in business firms
- Also, strong correlation exists between years and education and salary for economists in academia
- However, when all economists are considered, there is a **negative** correlation.
- What happened?

21

Relationships Between Two Categorical Variables

- If we want to relate two quantitative variables we can so via a scatter plot, correlation, or regression.
- These methods don't work for relating two categorical variables.
- Consider high blood pressure and its relationship to oral contraceptive use.

22

Marginal Distribution

- The marginal distribution of high blood pressure ignores the effects of all other variables.

High Blood Pressure	
Yes	200 (8.3%)
No	2200 (91.7%)
Total	2400 (100%)

Oral Contraceptive User	
Yes	800 (33.3%)
No	1600 (66.7%)
Total	2400 (100%)

23

Joint Distribution

- The joint distribution relates the two variables together. Often in a two-way table.

		OC User		
		Yes	No	Total
HBP	Yes	64 (2.67%)	136 (5.67%)	200
	No	736 (30.7%)	1464 (61.0%)	2200
	Total	800	1600	2400

24

Conditional Distribution

- The conditional distribution fixes the value of one of the variables, and looks at the distribution(s) of the other(s).
- Conditional on OC users:

OC Yes	HBP
Yes	64 (8.00%)
No	736 (92.0%)
Total	800

OC No	HBP
Yes	64 (8.5%)
No	736 (91.5%)
Total	1600

25

Three-Way Tables

- Three-way tables look at the relationship among three categorical variables.
- Generate two two-way tables by splitting over the categories of one of the variables.
- Simpson's Paradox: Associations in subgroups reverse direction when data are combined across subgroups.

26

HBP and OC use

Age 18-34		OC User		
		Yes	No	Total
HBP	Yes	36	16	52
	No	564	384	948
	Total	600	400	1000

Age 18-34		OC User		
		Yes	No	Total
HBP	Yes	28	120	148
	No	172	1080	1252
	Total	200	1200	1400

27