Chapter 5

Regression

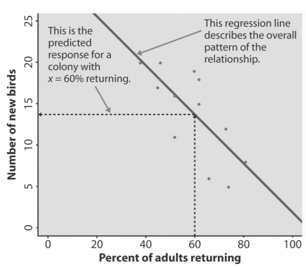
Linear Regression

- Objective: To *quantify* the linear relationship between an explanatory variable (x) and response variable (y).
- We can then ***predict*** the average response for all subjects with a given value of the explanatory variable.

Prediction via Regression Line

Number of new birds and Percent returning

Example: predicting number (y) of new adult birds that join the colony based on the percent (x) of adult birds that return to the colony from the previous year.



Least Squares

- Used to determine the "best" line
- We want the line to be as close as possible to the data points in the vertical (*y*) direction (since that is what we are trying to predict)
- **Least Squares**: use the line that minimizes the sum of the squares of the vertical distances of the data points from the line

Least Squares Regression Line

- Regression equation: $\hat{y} = a + bx$
  - ***x*** is the value of the explanatory variable
  - ***"y-hat"*** is the average value of the response variable *(predicted response for a value of x)*
  - note that ***a*** and ***b*** are just the intercept and slope of a straight line
  - note that ***r*** and ***b*** are not the same thing, but their signs will agree

Prediction via Regression Line

Number of new birds and Percent returning

- The regression equation is
  y-hat = 31.9343 - 0.3040*x*
  - ***y-hat*** is the average number of new birds for all colonies with percent ***x*** returning
- For all colonies with 60% returning, we ***predict*** the average number of new birds to be 13.69:
  31.9343 - (0.3040)(60) = 13.69 birds
- Suppose we know that an individual colony has 60% returning. What would we ***predict*** the number of new birds to be for just that colony?

Regression Line Calculation

- Regression equation: $\hat{y} = a + bx$

$$b = r\frac{s_y}{s_x}$$

$$a = \bar{y} - b\bar{x}$$

where s_x and s_y are the standard deviations of the two variables, and r is their correlation

Regression Calculation Case Study

Per Capita Gross Domestic Product and Average Life Expectancy for Countries in Western Europe

Regression Calculation Case Study

Country	Per Capita GDP (x)	Life Expectancy (y)
Austria	21.4	77.48
Belgium	23.2	77.53
Finland	20.0	77.32
France	22.7	78.63
Germany	20.8	77.17
Ireland	18.6	76.39
Italy	21.5	78.51
Netherlands	22.0	78.15
Switzerland	23.8	78.99
United Kingdom	21.2	77.37

Regression Calculation Case Study

Linear regression equation:

$\bar{x} = 21.52 \quad \bar{y} = 77.754 \quad r = 0.809$

$s_x = 1.532 \quad s_y = 0.795$

$$b = r\frac{s_y}{s_x} = (0.809)\left(\frac{0.795}{1.532}\right) = 0.420$$

$$a = \bar{y} - b\bar{x} = 77.754 - (0.420)(21.52) = 68.716$$

$$\hat{y} = 68.716 + 0.420x$$

Coefficient of Determination (R^2)

- Measures usefulness of regression prediction
- **R^2** (or **r^2**, the square of the correlation): measures what fraction of the variation in the values of the response variable (y) is explained by the regression line
  - r=1: R^2=1: regression line explains all (100%) of the variation in y
  - r=.7: R^2=.49: regression line explains almost half (50%) of the variation in y

Residuals

- A ***residual*** is the difference between an observed value of the response variable and the value predicted by the regression line:

$$\textit{residual} = y - \hat{y}$$

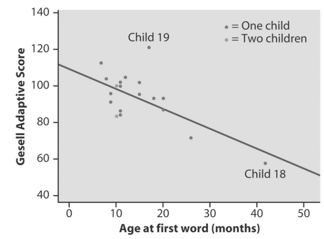
Residuals

- A ***residual plot*** is a scatterplot of the regression residuals against the explanatory variable
  - used to assess the fit of a regression line
  - look for a “random” scatter around zero

Case Study

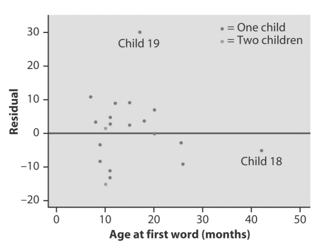
Gesell Adaptive Score and Age at First Word

Draper, N. R. and John, J. A. “Influential observations and outliers in regression,” *Technometrics*, Vol. 23 (1981), pp. 21-26.



Residual Plot: Case Study

Gesell Adaptive Score and Age at First Word

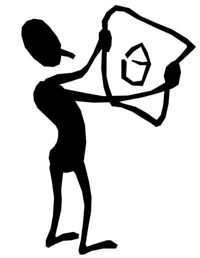


Outliers and Influential Points

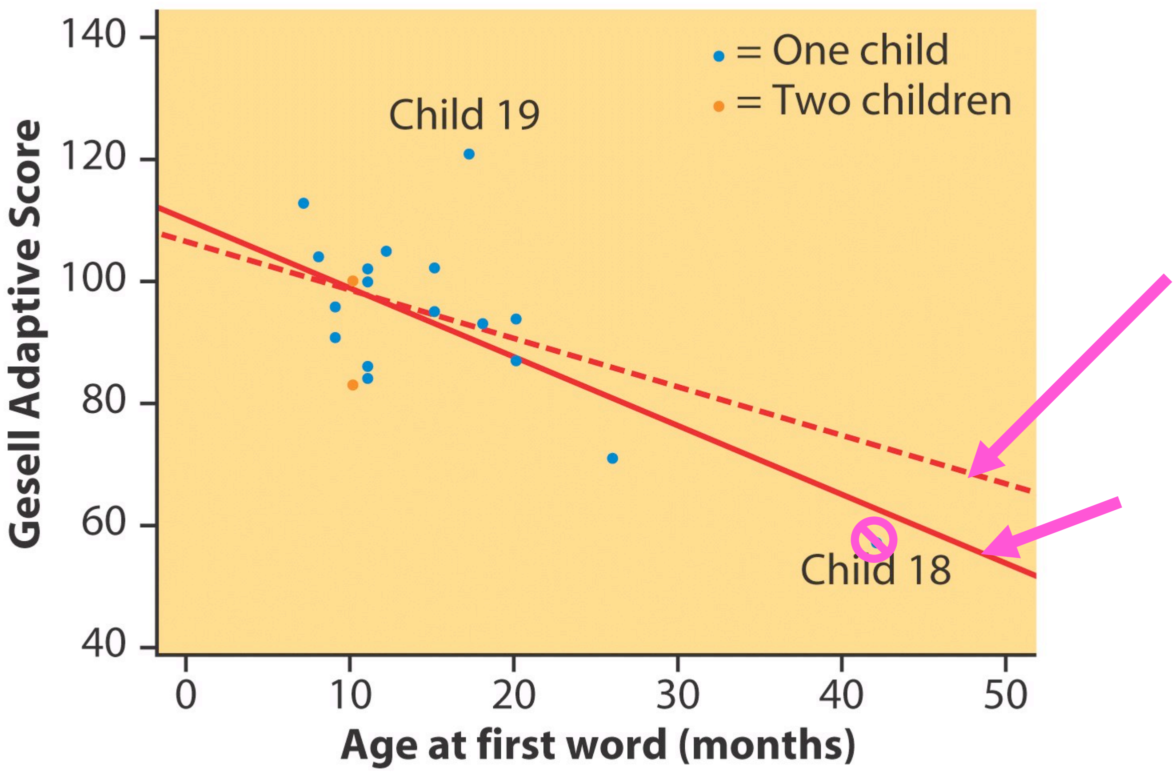
- An ***outlier*** is an observation that lies far away from the other observations

  outliers are often *influential* for the least-squares regression line, meaning that the removal of such points would markedly change the equation of the line

Outliers: Case Study



Gesell Adaptive Score and Age at First Word



r^2 = 11%

r^2 = 41%

Cautions about Correlation and Regression

- only describe <u>linear</u> relationships
- are both affected by outliers
- always plot the data before interpreting
- beware of *extrapolation*
  - predicting outside of the range of *x*
- beware of *lurking variables*
  - have important effect on the relationship among the variables in a study, but are not included in the study
- association does not imply causation

BPS - 5th Ed. Chapter 5 18

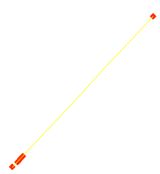
Caution: Beware of Extrapolation

- Sarah's height was plotted against her age
- Can you predict her height at age 42 months?
- Can you predict her height at age 30 years (360 months)?

BPS - 5th Ed. Chapter 5 19

Caution: Beware of Extrapolation

- Regression line: y-hat = 71.95 + .383 x
- height at age 42 months? y-hat = 88
- height at age 30 years? y-hat = 209.8
  - She is predicted to be 6' 10.5" at age 30.



BPS - 5th Ed. Chapter 5 20

Caution: Beware of Lurking Variables

Meditation and Aging

(*Noetic Sciences Review*, Summer 1993, p. 28)

- Explanatory variable: observed meditation practice (yes/no)
- Response: level of age-related enzyme
  - general concern for one's well being may also be affecting the response (and the decision to try meditation)

BPS - 5th Ed. Chapter 5 21

Caution: Correlation Does *Not* Imply Causation

Even very strong correlations may not correspond to a real causal relationship (changes in *x* actually causing changes in *y*).

(correlation may be explained by a lurking variable)

BPS - 5th Ed. Chapter 5 22

Caution: Correlation Does *Not* Imply Causation

Social Relationships and Health

House, J., Landis, K., and Umberson, D. "Social Relationships and Health," *Science*, Vol. 241 (1988), pp 540-545.

- Does lack of social relationships cause people to become ill? *(there was a strong correlation)*
- **<u>Or</u>**, are unhealthy people less likely to establish and maintain social relationships? *(reversed relationship)*
- **<u>Or</u>**, is there some <u>other factor</u> that predisposes people both to have lower social activity and become ill?

BPS - 5th Ed. Chapter 5 23

Evidence of Causation

- Other considerations:
  - The association is strong
  - The association is *consistent*
    - The connection happens in repeated trials
    - The connection happens under varying conditions
- A properly conducted ***experiment*** establishes the connection *(chapter 9)*

BPS - 5th Ed. Chapter 5 24