Lecture 12: more Chapter 5, Section 3 Relationships between Two Quantitative Variables; Regression

- Equation of Regression Line; Residuals
- Effect of Explanatory/Response Roles
- Unusual Observations
- Sample vs. Population
- Time Series; Additional Variables

Looking Back: *Review*

- **4 Stages of Statistics**
  - Data Production (discussed in Lectures 1-4)
  - Displaying and Summarizing
    - Single variables: 1 cat,1 quan (discussed Lectures 5-8)
    - Relationships between 2 variables:
      - Categorical and quantitative (discussed in Lecture 9)
      - Two categorical (discussed in Lecture 10)
      - Two quantitative
  - Probability
  - Statistical Inference

Review

- Relationship between 2 quantitative variables
  - Display with scatterplot
  - Summarize:
    - Form: linear or curved
    - Direction: positive or negative
    - Strength: strong, moderate, weak

    If form is linear, correlation r tells direction and strength.

    Also, equation of least squares regression line lets us predict a response $\hat{y}$ for any explanatory value x.

Least Squares Regression Line

Summarize linear relationship between explanatory (*x*) and response (*y*) values with line $\hat{y} = \boxed{b_0} + \boxed{b_1}x$ that minimizes sum of squared prediction errors (called *residuals*).

- **Slope:** predicted change in response *y* for every unit increase in explanatory value *x*
- **Intercept:** where best-fitting line crosses *y*-axis (predicted response for *x*=0?)

Example: *Least Squares Regression Line*

- **Background**: Car-buyer used software to regress price on age for 14 used Grand Am's.

```
The regression equation is
Price = 14690 - 1288 Age
```

- **Question:** What do the slope (-1,288) and intercept (14,690) tell us?
- **Response:**
  - **Slope:** For each additional year in age, predict price
  - **Intercept:** Best-fitting line

Example: *Extrapolation*

- **Background**: Car-buyer used software to regress price on age for 14 used Grand Am's.

```
The regression equation is
Price = 14690 - 1288 Age
```

- **Question:** Should we predict a new Grand Am to cost \$14,690-\$1,288(0)=\$14,690?
- **Response:**

Definition

- **Extrapolation:** using the regression line to predict responses for explanatory values outside the range of those used to construct the line.

Example: *More Extrapolation*

- **Background**: A regression of 17 male students' weights (lbs.) on heights (inches) yields the equation
$$\hat{y} = -438 + 8.7x$$
- **Question:** What weight does the line predict for a 20-inch-long infant?
- **Response:**

Expressions for slope and intercept

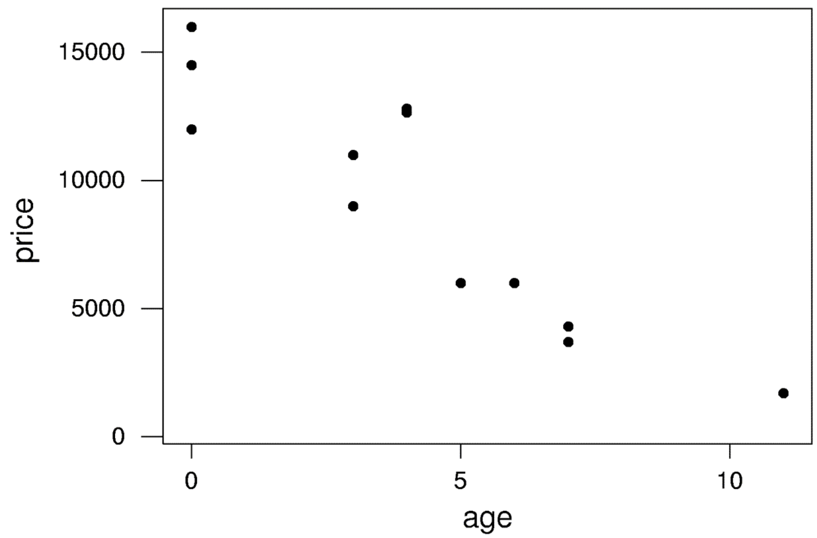
Consider slope and intercept of the least squares regression line $\hat{y} = b_0 + b_1 x$

- **Slope:** $b_1 = r\frac{s_y}{s_x}$ so if x increases by a standard deviation, predict y to increase by r standard deviations
- **Intercept:** $b_0 = \bar{y} - b_1\bar{x}$ so when $x = \bar{x}$ predict $\hat{y} = b_0 + b_1\bar{x} = (\bar{y} - b_1\bar{x}) + b_1\bar{x} = \bar{y}$

➔the line passes through the point of averages $(\bar{x}, \bar{y})$

Example: *Individual Summaries on Scatterplot*

- **Background**: Car-buyer plotted price vs. age for 14 used Grand Ams [(4, 13,000), (8, 4,000), etc.]



- **Question:** Guess the means and sds of age and price?
- **Response:** Age has approx. mean ___ yrs, sd ___ yrs; price has approx. mean $ ___ sd $ ___

Definitions

- **Residual:** error in using regression line $\widehat{y} = b_0 + b_1 x$ to predict *y* given *x*. It equals the vertical distance *observed minus predicted* which can be written $y_i - \widehat{y}_i$
- ***s*:** denotes typical residual size, calculated as

$$s = \sqrt{\frac{(y_1 - \widehat{y}_1)^2 + \cdots + (y_n - \widehat{y}_n)^2}{n-2}}$$

Note: ***s*** *just "averages" out the residuals* $y_i - \widehat{y}_i$

Example: *Considering Residuals*

- **Background**: Car-buyer regressed price on age for 14 used Grand Ams [(4, 13,000), (8, 4,000), etc.].

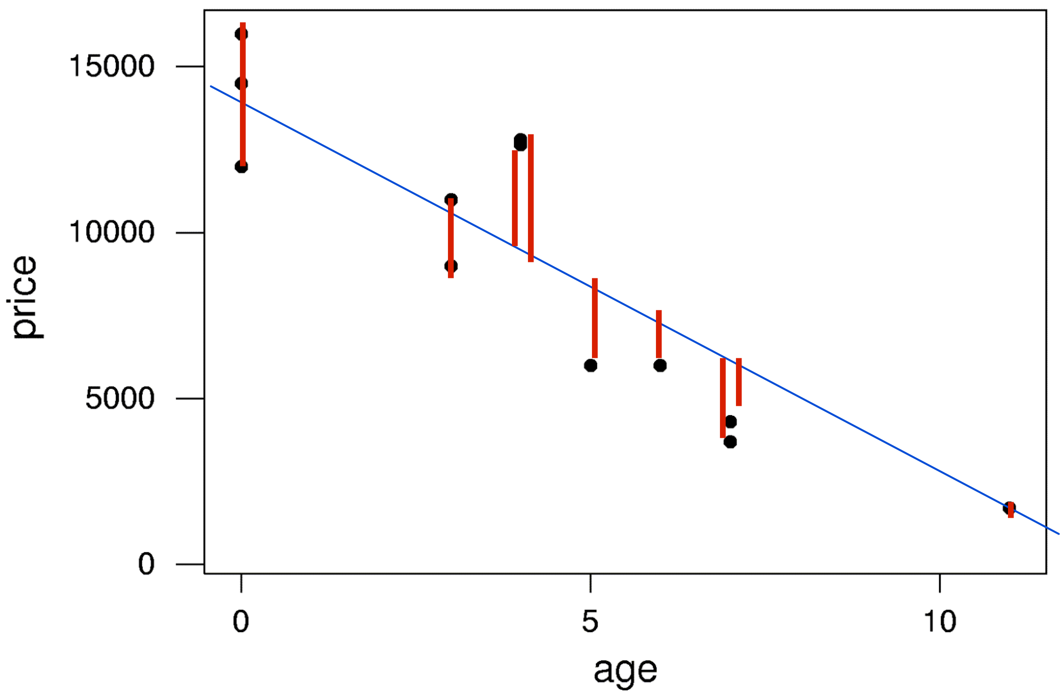
```
The regression equation is
price = 14686 - 1290 age
S = 2175          R-Sq = 78.5%      R-Sq(adj) = 76.7%
```

- **Question:** What does $s = 2{,}175$ tell us?
- **Response:** Regression line predictions not perfect:
  - x=4→predict $\hat{y}$=
    actual y=13,000→prediction error =
  - x=8→predict $\hat{y}$=
    actual y=4,000→prediction error =
  - Typical size of 14 prediction errors is (dollars)

Practice: 5.56a p.197

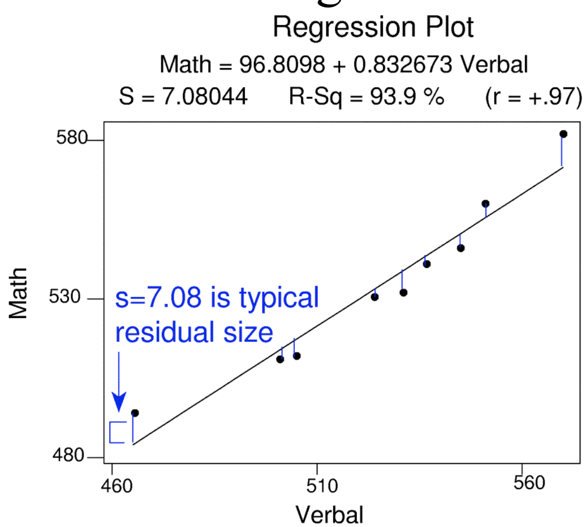
Example: *Considering Residuals*

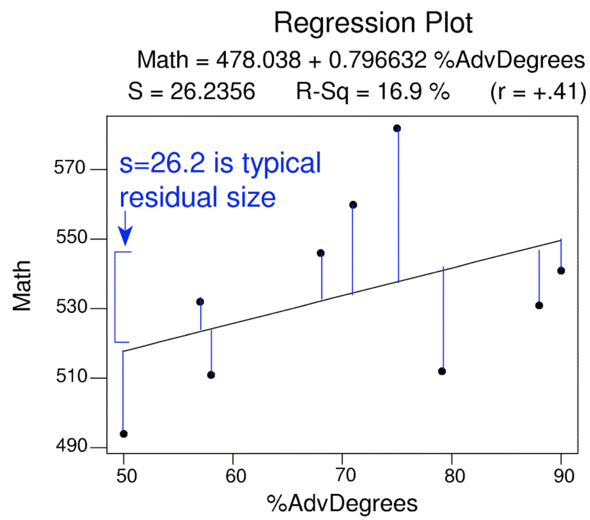
- Typical size of 14 prediction errors is s = 2,175 (dollars): Some points' vertical distance from line more, some less; 2,175 is typical distance.



Example: *Residuals and their Typical Size s*

- **Background:** For a sample of schools, regressed
  - average Math SAT on average Verbal SAT
  - average Math SAT on % of teachers w. advanced degrees





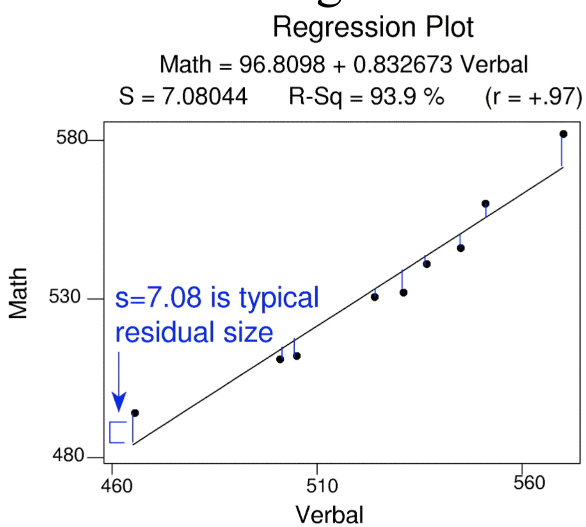
***A Closer Look:** If output reports R-sq, take its square root (+ or - depending on slope) to find r.*

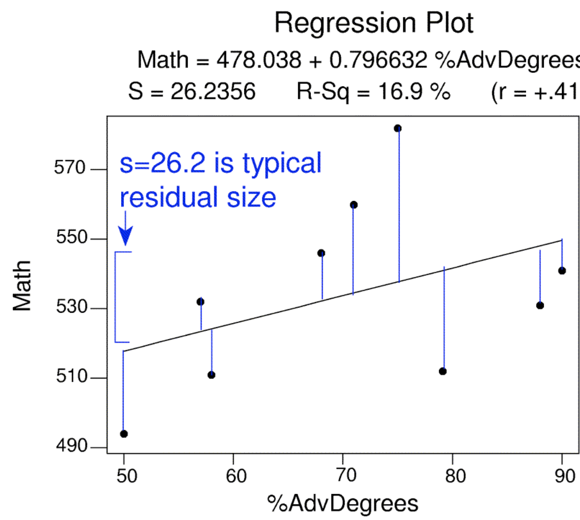
- **Question:** How are $s = 7.08$ (left) and $s = 26.2$ (right) consistent with the values of the correlation r?
- **Response:** On left $r = \sqrt{Rsq} = \sqrt{0.939} = 0.97$; relation is and typical error size is (only 7.08).

Practice: 5.70i-k p.203

Example: *Residuals and their Typical Size s*

- **Background:** For a sample of schools, regressed
  - average Math SAT on average Verbal SAT *Smaller s →better predictions*
  - average Math SAT on % of teachers w. advanced degrees



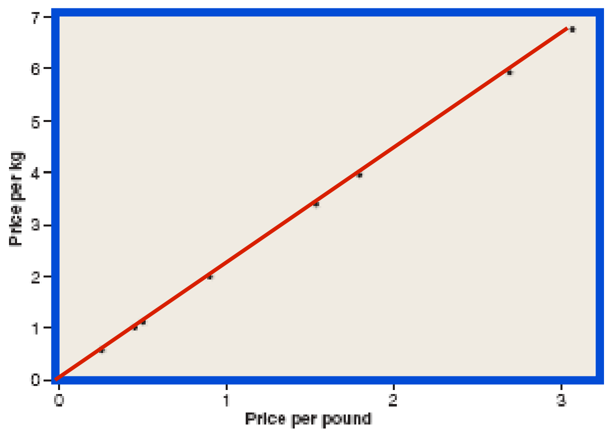


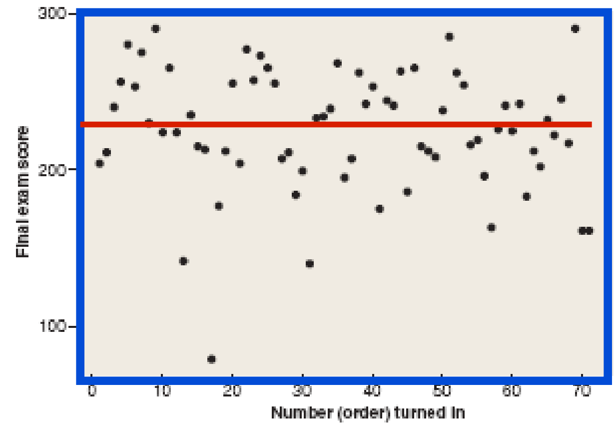
***Looking Back:** r based on averages is overstated; strength of relationship for individual students would be less.*

- **Question:** How are $s = 7.08$ (left) and $s = 26.2$ (right) consistent with the values of the correlation r?
- **Response:** On right $r =$
  relation is and typical error size is (26.2).

Example: *Typical Residual Size s close to s_y or 0*

- **Background**: Scatterplots show relationships…
  - Price per kilogram vs. price per lb. for groceries
  - Students' final exam score vs. (number) order handed in





Regression line approx. same as line at average y-value.

- **Questions:** Which has $s = 0$? Which has s close to s_y?
- **Responses:** Plot on left has $s =$ no prediction errors.

  Plot on right: s close to (Regressing on x doesn't help; regression line is approximately horizontal.)

Example: *Typical Residual Size s close to s_y*

- **Background**: 2008-9 Football Season Scores

Regression Analysis: Steelers versus Opponents

The regression equation is

Steelers = 23.5 - 0.053 Opponents

S = 9.931

Descriptive Statistics: Steelers

Variable	N	Mean	Median	TrMean	StDev	SE Mean
Steelers	19	22.74	23.00	22.82	9.66	2.22

Variable	Minimum	Maximum	Q1	Q3
Steelers	6.00	38.00	14.00	31.00

Question: Since $s = 9.931$ and $s_y = 9.66$ are very close, do you expect $|r|$ close to 0 or 1?

Response: r must be close to

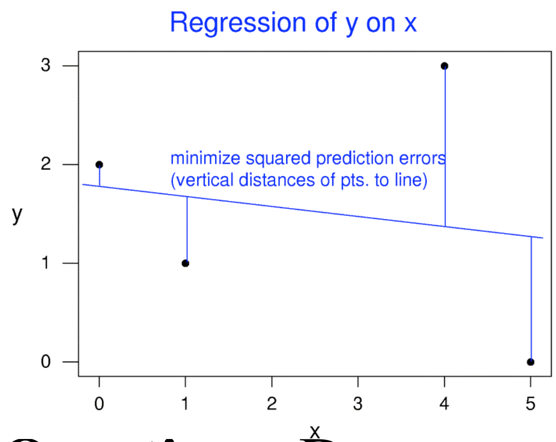
Practice: 5.59 p.198

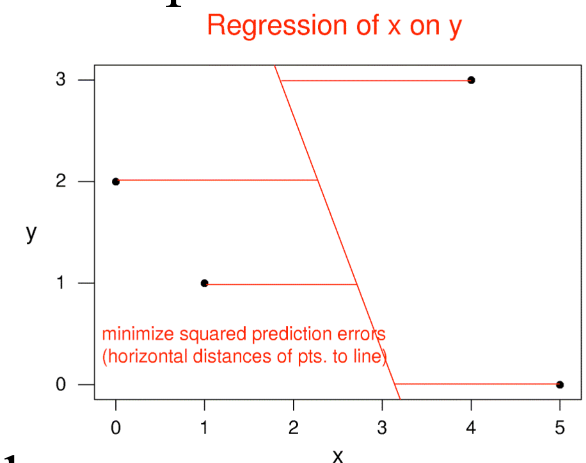
Explanatory/Response Roles in Regression

Our choice of roles, explanatory or response, does *not* affect the value of the correlation r, but it *does* affect the regression line.

Example: *Regression Line when Roles are Switched*

- **Background**: Compare regression of *y* on *x* (left) and regression of *x* on *y* (right) for same 4 points:





- **Question:** Do we get the same line regressing *y* on *x* as we do regressing *x* on *y*?
- **Response:** The lines are very different.
  - Regressing *y* on *x*: slope
  - Regressing *x* on *y*: slope

Context needed; consider variables and their roles before regressing.

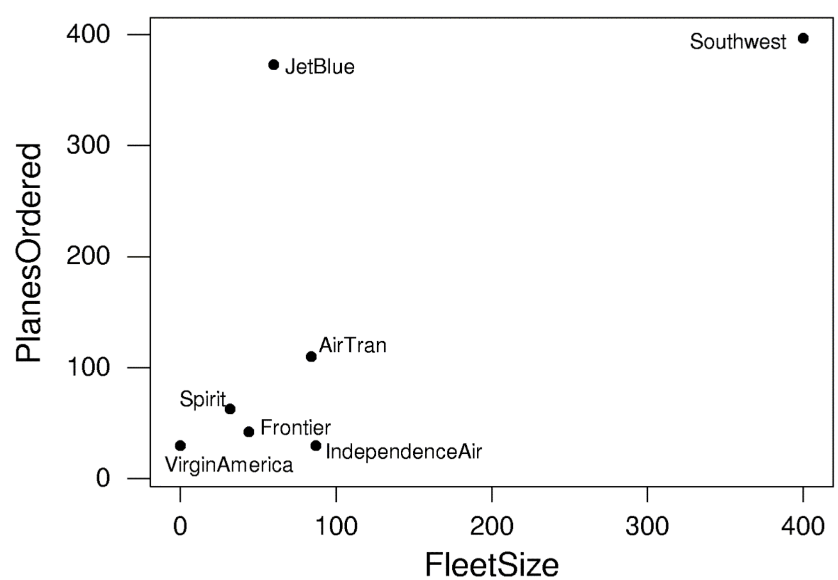
Practice: 5.60b p.198

Definitions

- **Outlier:** (in regression) point with unusually large residual
- **Influential observation:** point with high degree of influence on regression line.

Example: *Outliers and Influential Observations*

- **Background**: Exploring relationship between orders for new planes and fleet size. (r=+0.69)



- **Question:** Are Southwest and JetBlue outliers or influential?
- **Response:**
  - Southwest: (omit it→slope changes a lot)
  - JetBlue: large residual; omit it→r increases to +0.97)

Example: *Outliers and Influential Observations*

- **Background**: Exploring relationship between orders for new planes and fleet size. ($r = +0.69$)

```
Unusual Observations
Obs   FleetSiz   PlanesOr      Fit     SE Fit    Residual    St Resid
  6        400      397.0    398.1      127.1        -1.1     -0.04 X
  7         60      373.0    115.2       51.7       257.8      2.16R
R denotes an observation with a large standardized residual
X denotes an observation whose X value gives it large influence.
```

- **Question:** How does Minitab classify Southwest and JetBlue?
- **Response:**
  - **Southwest:** (marked in Minitab)
  - **JetBlue:** (marked in Minitab)

Influential observations tend to be extreme in ***horizontal*** *direction.*

Practice: 5.70g p.203

Definitions

- **Slope β_1:** how much response *y* changes in general (for entire population) for every unit increase in explanatory variable *x*
- **Intercept β_0:** where the line that best fits all explanatory/response points (for entire population) crosses the *y*-axis

Looking Back: *Greek letters often refer to population parameters.*

Line for Sample vs. Population

- **Sample:** line best fitting sampled points: predicted response is

$$\hat{y} = b_0 + b_1 x$$

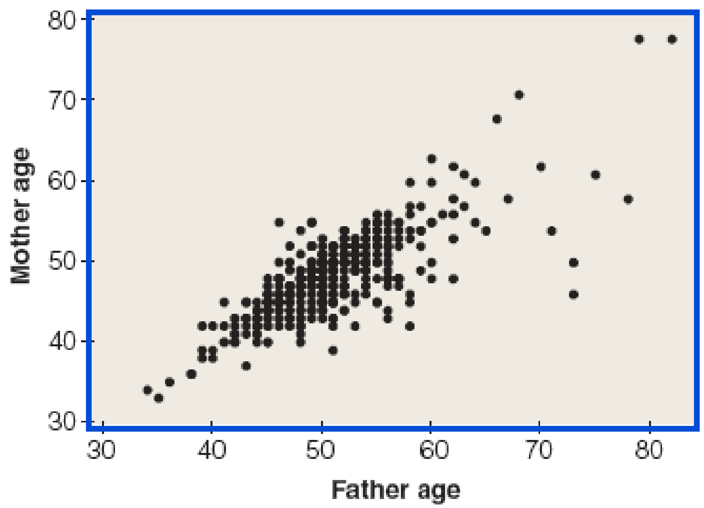
- **Population:** line best fitting all points in population from which given points were sampled: mean response is

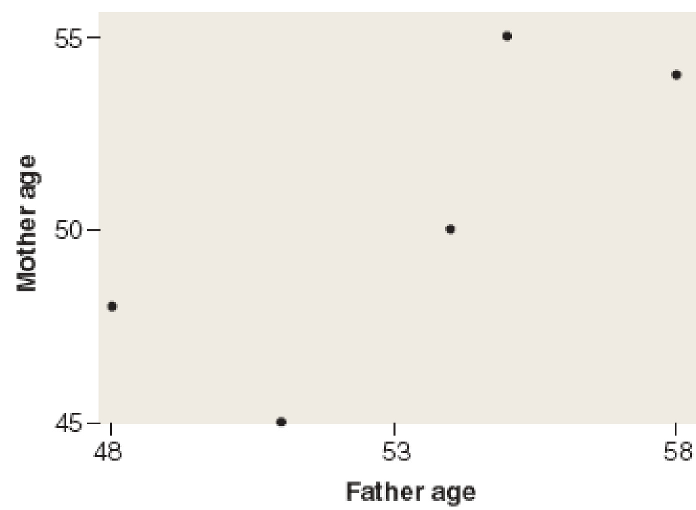
$$\mu_y = \beta_0 + \beta_1 x$$

A larger sample helps provide more evidence of a relationship between two quantitative variables in the general population.

Example: *Role of Sample Size*

- **Background**: Relationship between ages of students' mothers and fathers both have $r = +0.78$, but sample size is over 400 (on left) or just 5 (on right):





- **Question:** Which plot provides more evidence of strong positive relationship in population?
- **Response:** Plot on ____________________

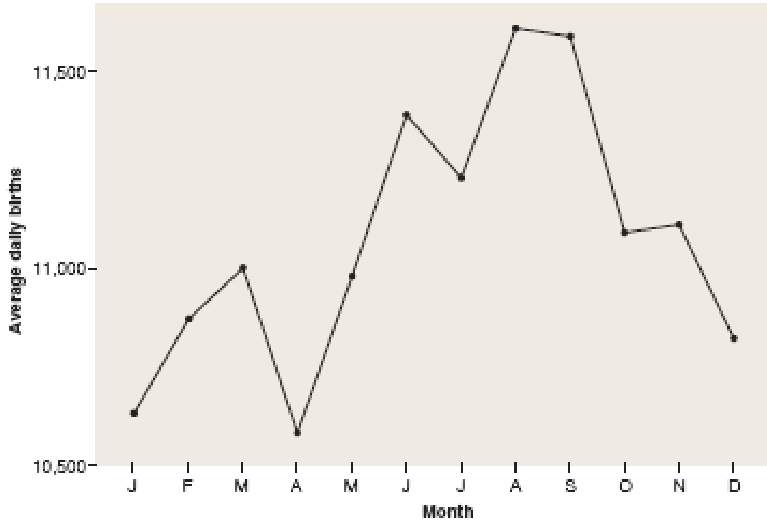
Can believe configuration on occurred by chance.

Time Series

If explanatory variable is time, plot one response for each time value and "connect the dots" to look for general trend over time, also peaks and troughs.

Example: *Time Series*

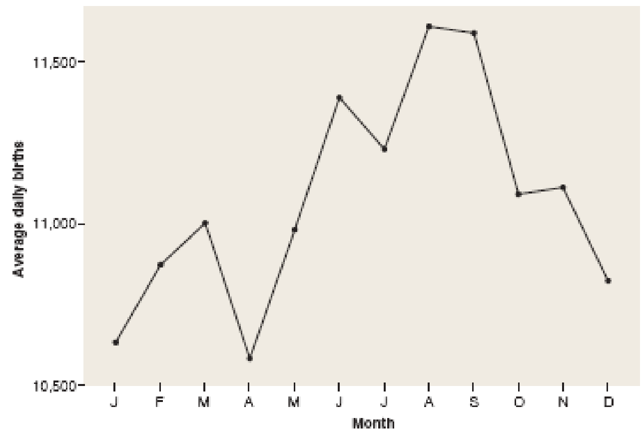
- **Background**: Time series plot shows average daily births each month in year 2000 in the U.S.:



- **Question:** Where do you see a peak or a trough?
Response: Trough in ________ peak in ____________

Example: *Time Series*

- **Background**: Time series plot of average daily births in U.S.



- **Questions:** How can we explain why there are…
  - Conceptions in U.S.: fewer in July, more in December?
  - Conceptions in Europe: more in summer, fewer in winter?
- **Response:**

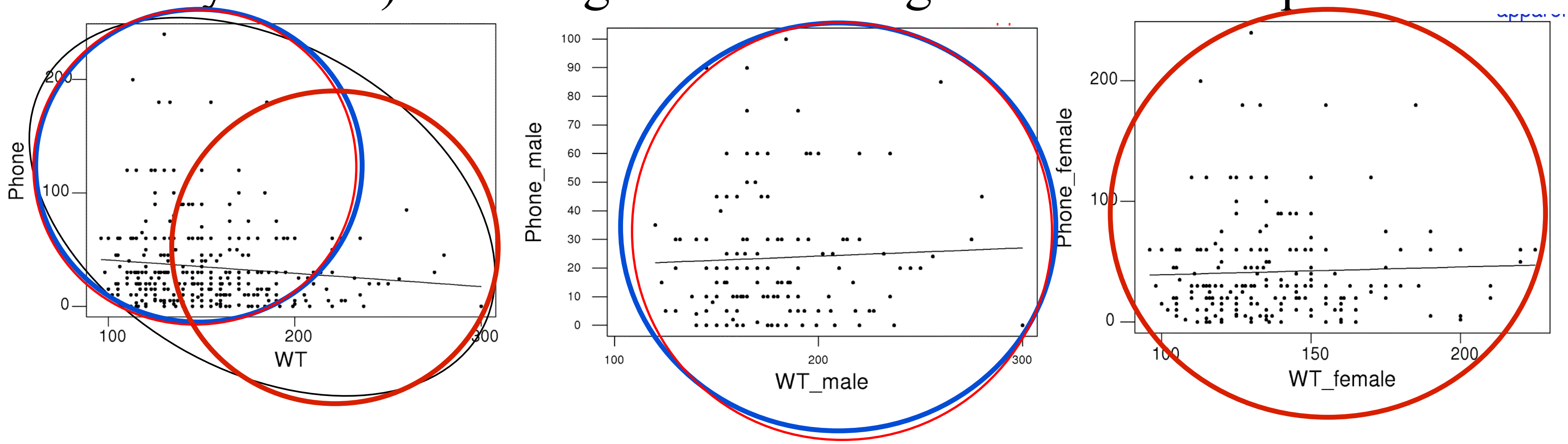
***A Closer Look:** Statistical methods can't always explain "why", but at least they help understand "what" is going on.*

Additional Variables in Regression

- **Confounding Variable:** Combining two groups that differ with respect to a variable that is related to both explanatory and response variables can affect the nature of their relationship.
- **Multiple Regression:** More advanced treatments consider impact of not just one but two or more quantitative explanatory variables on a quantitative response.

Example: *Additional Variables*

- **Background**: A regression of phone time (in minutes the day before) and weight shows a negative relationship.



- **Questions:** Do heavy people talk on the phone less? Do light people talk more?
- **Response:** is confounding variable→regress separately for → no relationship

Practice: 5.113a p.219

Example: *Multiple Regression*

- **Background**: We used a car's age to predict its price.
- **Question:** What additional quantitative variable would help predict a car's price?
- **Response:**

Lecture Summary *(Regression)*

- Equation of regression line
  - Interpreting slope and intercept
  - Extrapolation
- Residuals: typical size is s
- Line affected by explanatory/response roles
- Outliers and influential observations
- Line for sample or population; role of sample size
- Time series
- Additional variables