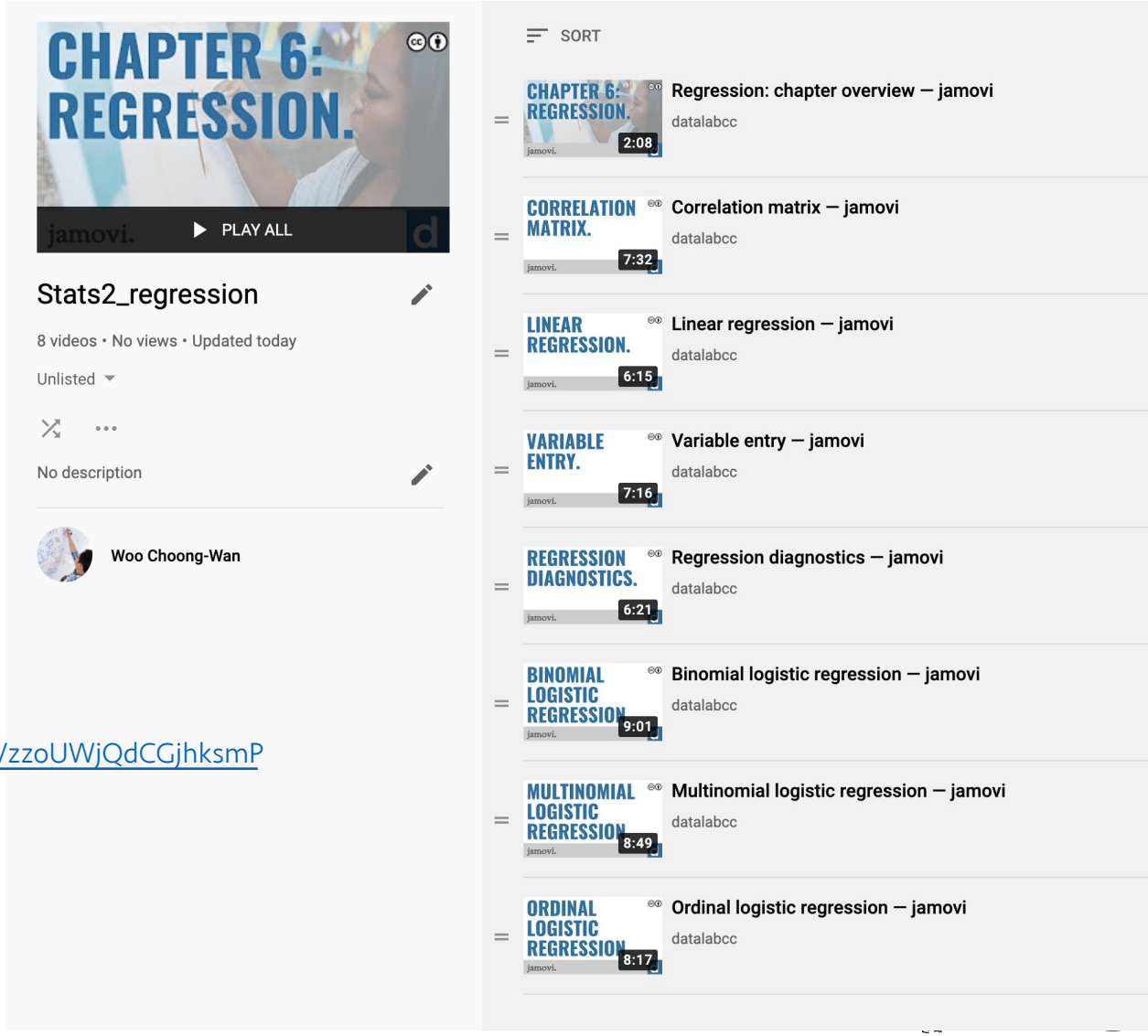
# Lecture 18
# Correlation and linear regression

CHOONG-WAN WOO | COCOAN lab | http://cocoanlab.github.io

# Regression with JAMOVI




https://www.youtube.com/playlist?list=PLXCuLG6zw7mKEkaaVzzoUWjQdCGjhksmP

# Linear relationships

- If the form of the plot looks like a line, this indicates there may be a linear relationship between the two variables.
- The relationship is strong if all the data points approximately make up a straight line.
- It is weak if the points are widely scattered about the line.

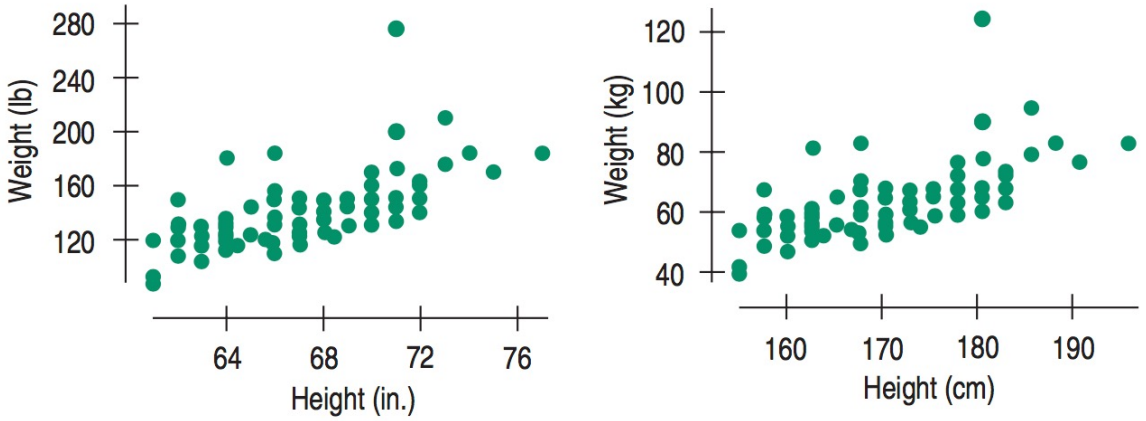
Slide credit: Martin Lindquist

# Correlation

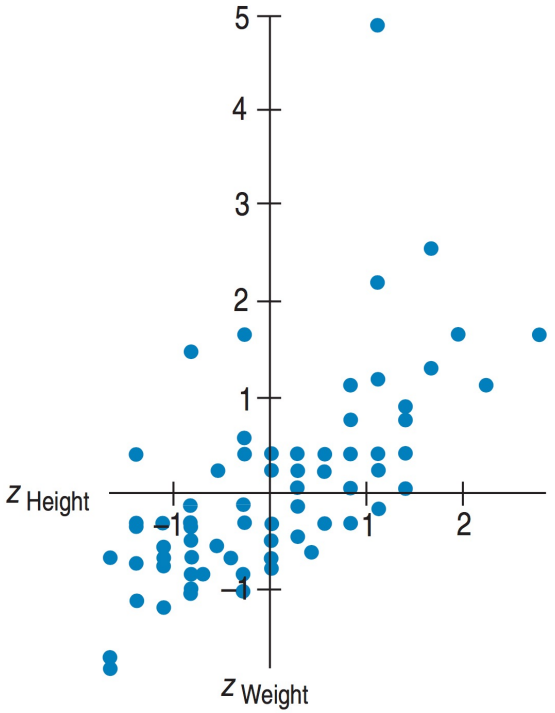
- We want a numerical summary that can be used to measure the strength of a linear relationship.
- The correlation is a measure of strength and direction of a linear relationship between two quantitative variables.
- Correlations are usually denoted by *r*.

Slide credit: Martin Lindquist

# Correlation




- The units should have no effects on our measure of strength of a linear relationship.
- *z*-scores can be used: $(z_x, z_y) = \left(\frac{x - \bar{x}}{s_x}, \frac{y - \bar{y}}{s_y}\right)$



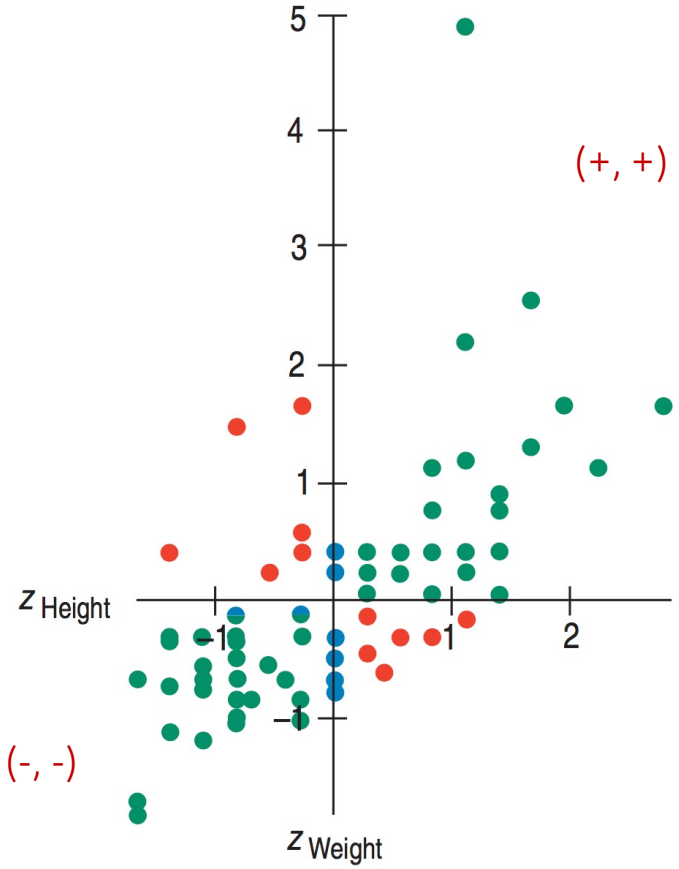
# Correlation

- Strength of linear relationship

  should be proportional to $\sum z_x z_y$

- Correlation coefficient: $r = \dfrac{\sum z_x z_y}{n - 1}$

  , where $(z_x, z_y) = \left(\dfrac{x - \bar{x}}{s_x}, \dfrac{y - \bar{y}}{s_y}\right)$



# Correlation

- Strength of linear relationship

  should be proportional to $\sum z_x z_y$

- Correlation coefficient: $r = \dfrac{\sum z_x z_y}{n-1}$

  , where $(z_x, z_y) = \left(\dfrac{x-\bar{x}}{s_x}, \dfrac{y-\bar{y}}{s_y}\right)$

- Different expressions, but mathematically same:

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}} = \frac{\sum (x-\bar{x})(y-\bar{y})}{(n-1)s_x s_y}$$

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

# Properties of Correlation

- **Sign of correlation:** the direction of the association (e.g., positive, negative)
- **Range:** *r* is always between -1 and 1.
  - When *r* = 1 all of the points lie on a straight line with a positive slope.
  - *r* < 0 indicates a negative association.
  - When *r* = -1 all points lie on a straight line with negative slope.
  - If *r* is close to 0, this indicates a very weak linear relationship.
- **Symmetry:** The correlation of x with y is the same as the correlation of y with x.
- **No units**
  - The value of *r* does not change even if units of measure are changed.
  - The correlation has no unit of measurement.
- **Only linear:** Correlation measures only the strength of a *linear* relationship.
- **Sensitive to outliers:** The correlation is sensitive to outliers.

# Other correlation measures (non-parametric):

- Kendall's Tau ($\tau$)
  - can be used for Likert-type scale data (*ordinal* variable)
  - Likert-type scale: e.g., 0 = not at all, 1 = a little, 2 = moderately, 3 = very much
  - commonly used in questionnaires or survey

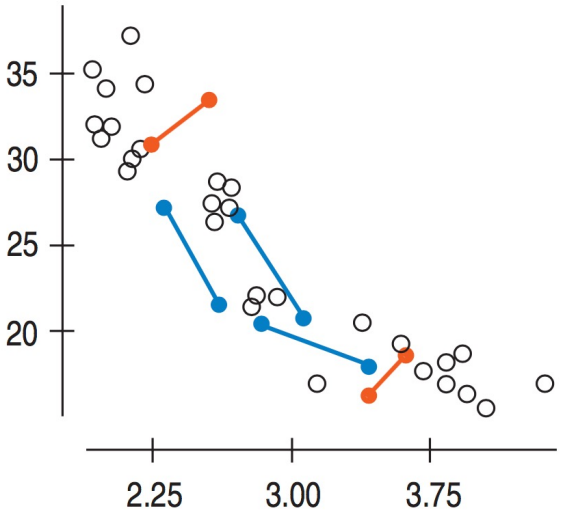
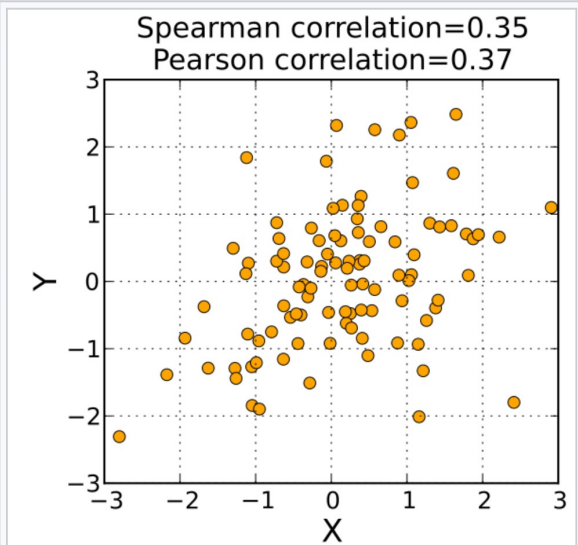



**FIGURE 7.6**
For each pair of points, Kendall's tau records whether the slope between them is positive (red), negative (blue), or zero.

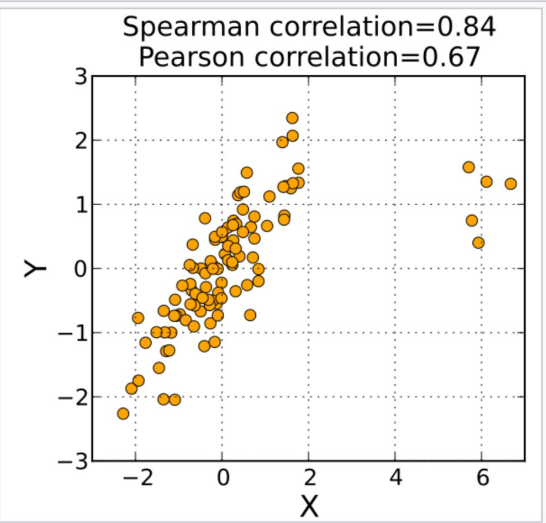
$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{n(n-1)/2}$$

# Other correlation measures (non-parametric):

- Spearman's Rho ($\rho$)
  - replaces the original data with their ranks within each variable
  - then, calculate the correlation




When the data are roughly elliptically distributed and there are no prominent outliers, the Spearman correlation and Pearson correlation give similar values.



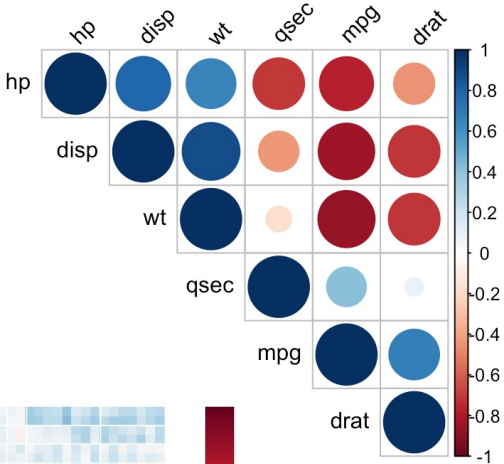

The Spearman correlation is less sensitive than the Pearson correlation to strong outliers that are in the tails of both samples. That is because Spearman's rho limits the outlier to the value of its rank.

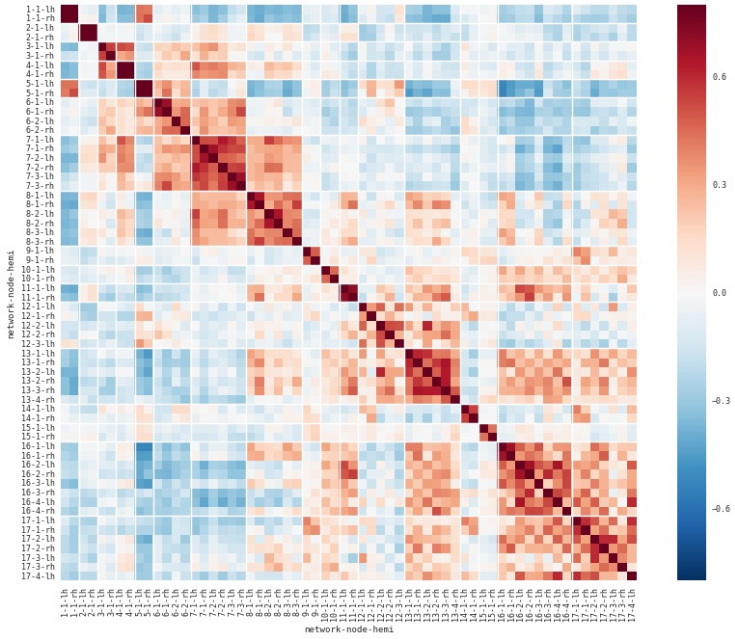
https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

# Correlation table

| | Assets | Sales | Market Value | Profits | Cash Flow | Employees |
|---|---|---|---|---|---|---|
| Assets | 1.000 | | | | | |
| Sales | 0.746 | 1.000 | | | | |
| Market Value | 0.682 | 0.879 | 1.000 | | | |
| Profits | 0.602 | 0.814 | 0.968 | 1.000 | | |
| Cash Flow | 0.641 | 0.855 | 0.970 | 0.989 | 1.000 | |
| Employees | 0.594 | 0.924 | 0.818 | 0.762 | 0.787 | 1.000 |

**TABLE 7.1**
A correlation table of data reported by *Forbes* magazine for large companies. From this table, can you be sure that the variables are linearly associated and free from outliers?



# Assessing regression model: $R^2$

- Correlation: strength and direction
- To evaluate how well a regression model does, direction won't matter that much.
- $R^2$: ranges between 0 and 1
- tells us the fraction of the data's variation accounted for by the model

- $R^2 = 1 - \frac{Sum\ of\ squared\ residuals}{Sum\ of\ squared\ deviation\ from\ the\ mean} = 1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}$
  - In the linear model, $R^2$ is same with $r^2$.
  - 1-$r^2$: the fraction of the original variation left in the residuals

- How big should $R^2$ be? *It depends! Data type, field, etc.*
- What is more important between $b$ and $R^2$? *It depends! Data type, field, research question, etc.*

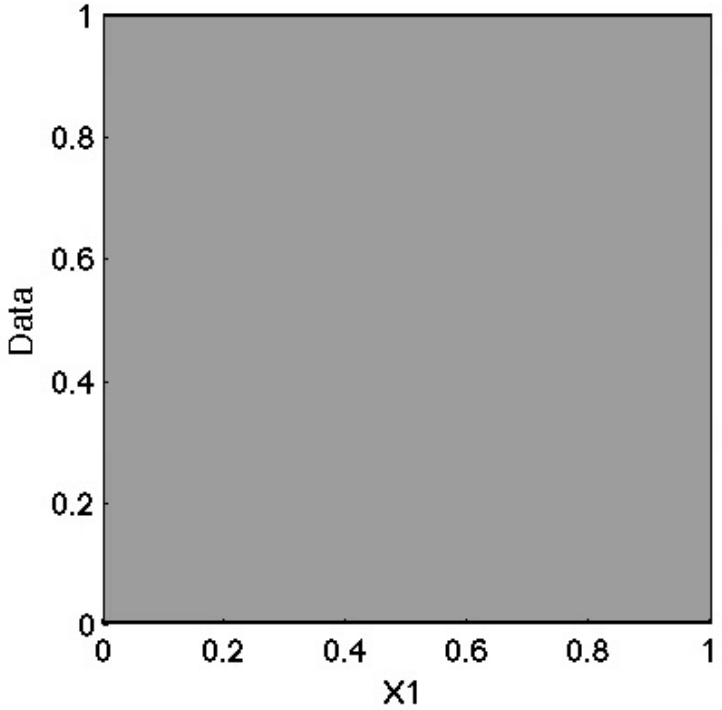
# Stepwise regression and hierarchical regression

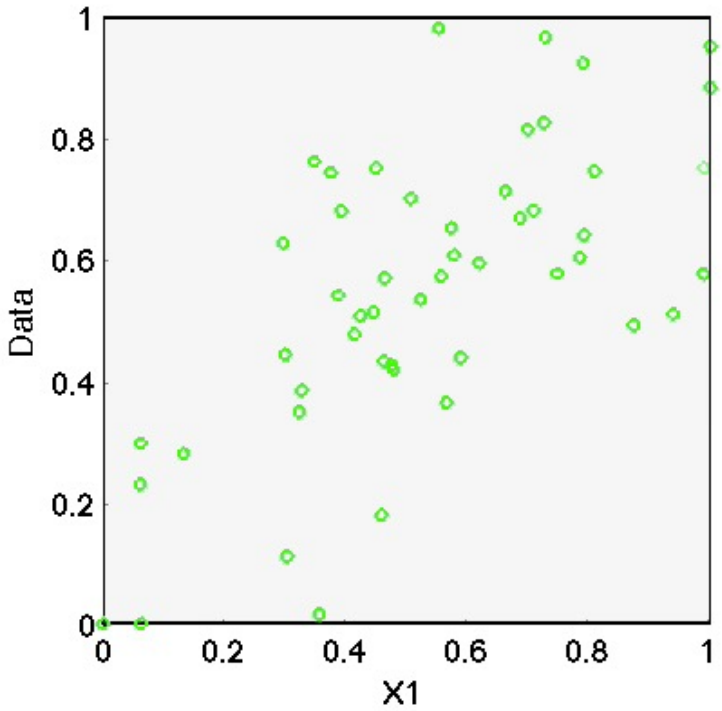
- Automatic selection?
  - **Stepwise regression**
    - At each step, a predictor is added to or removed from the model.
    - The predictor chosen to add is the one whose addition increases the $R^2$ the most.
    - The predictor chosen to remove is the one whose removal reduces the $R^2$ the least.
    - However, each step is made automatically. Thus, they can be affected by influential cases and nonlinear relationships.

  - A better strategy should be **"mimic the stepwise procedure yourself more carefully"**
    - Hierarchical regression

# Collinearity

- When we have several predictors, we have to consider how the predictors are related to each other.






Video credit: **Tor Wager**

# Variance inflation factor

- When we have several predictors, we have to consider how the predictors are related to each other.
- Predictors are **collinear**, when they are linearly related.

- Variance Inflation Factor
  - Build regression models for each predictor using the rest of predictors
  - Calculate $R^2$ for each predictor (based on the rest of predictors)

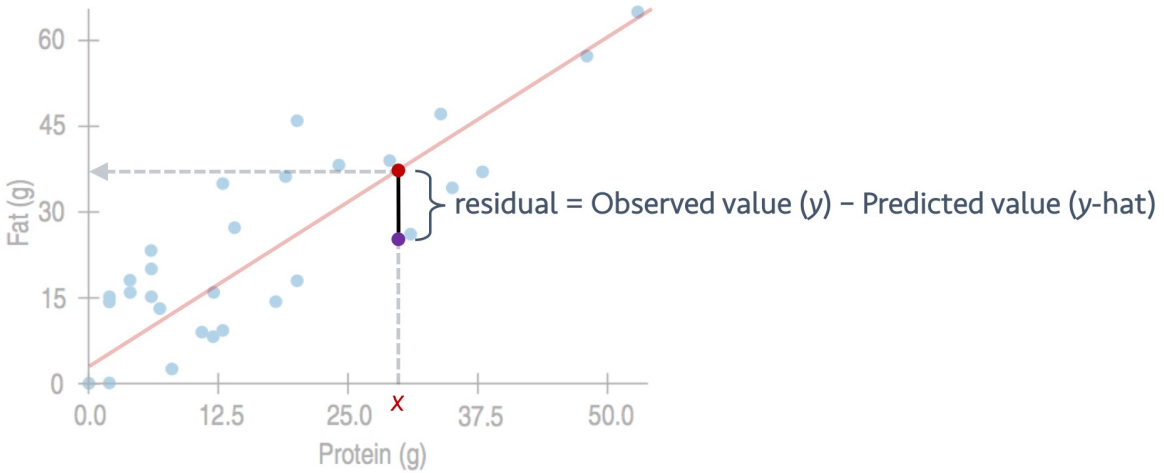
  - $\mathrm{VIF}_i = \frac{1}{1-R_i^2}$
  - High VIF means small 1-$R^2$, indicating that the variable is highly correlated with other variables (high collinearity)
  - Low VIF means large 1-$R^2$, indicating that the variable is not correlated (low collinearity).



Video credit: **Tor Wager**

# Examining the residuals

- Residuals are defined as:
- $e = y - \hat{y}$



# Examining the residuals

- Residuals are defined as:
- $e = y - \hat{y}$

- In least square regression, the sum of the residuals is always zero.
- The residuals are the variation in the data that has not been modeled.
  - ❖ DATA = MODEL + RESIDUAL

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

$$y = b_0 + b_1 x + e$$

- A residual plot is a scatter plot of the residuals against $x$ or $\hat{y}$.
- When studying the residual plot we hope to see NO pattern.

Slide partly from Martin Lindquist

# Examining the residuals: Sleep study

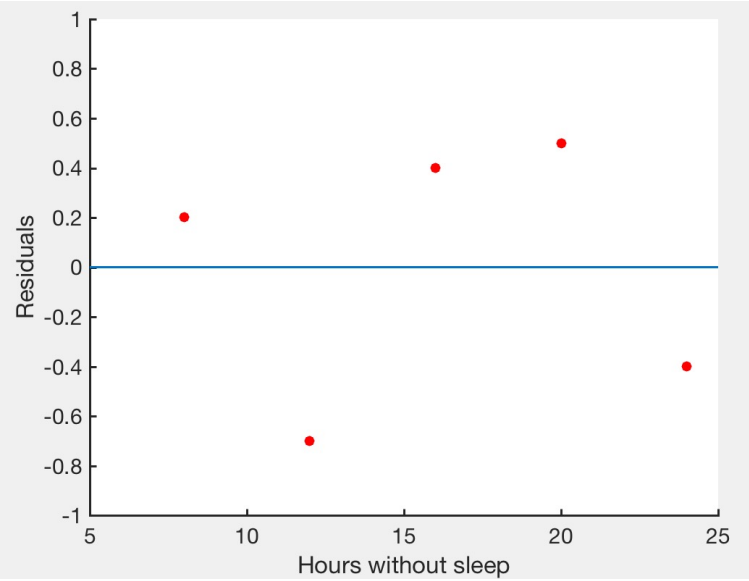
| Errors | 7 | 8 | 11 | 13 | 14 |
|---|---|---|---|---|---|
| Hours without sleep | 8 | 12 | 16 | 20 | 24 |

```
>> residuals = y-(3+0.475*x);
>> sum(residuals)

ans =

   2.6645e-15

>> scatter(x, residuals)
```




- The sum of the residuals is zero.
- should be the most boring scatterplot you've ever seen!
- shouldn't have any interesting features, direction or shape
- should stretch horizontally, with about same amount of scatter throughout
- No bends, no outliers