Public Health 501
Class 10

Linear Regression

Pearson's Correlation

Spearman's Rank Correlation

Goals

- Understand when to use correlation
- Be able to interpret correlation in terms of form, direction, and strength
- Know how to compute and interpret the Pearson correlation coefficient and perform hypothesis testing
- Be able to describe the limitations in interpretation: causality, outliers, and restriction of range
- Distinguish between relationships and predictions
- Understand the fundamentals of linear regression, how to use it, and perform hypothesis testing
- Know when and how to use the Spearman correlation coefficient

Comparing 2 Variables to Each Other

- Up until now, we have compared:
  - Continuous vs. categorical
    - Independent sample t-test (2 groups)
    - ANOVA (>2 groups)

      Normally distributed
    - Sign test, Wilcoxon signed rank, Wilcoxon rank sum
    - Kruskal-Wallis

      Not Normally distributed
  - Categorical vs. categorical
    - Chi-square tests
- But… how do we compare two continuous variables to each other?

Setting the Scene

- *Your veterinary nutritionist friend has noticed that many of her 4-legged patients have put on a few kilograms*
- *Being an astute practitioner, she also noticed that the owners of these pets tend to overindulge them*
- *She develops an index of over-indulgence, the PET-ME* score, and attempts to relate it to dog BMI*

* ***P**eople food, **E**xercise hours, **T**oys and treats, **M**eals/day, **E**ntertainment (non exercise)*

Setting the Scene

- *But BMI and PET-ME are both interval/ratio*
  - *All the tests we've learned so far involve an interval/ratio variable and a nominal variable or 2 nominal/ordinal, i.e. categorical variables*
- *We could categorize PET-ME into low and high and do a t-test, or low, medium, high and do an ANOVA, but we would lose information*

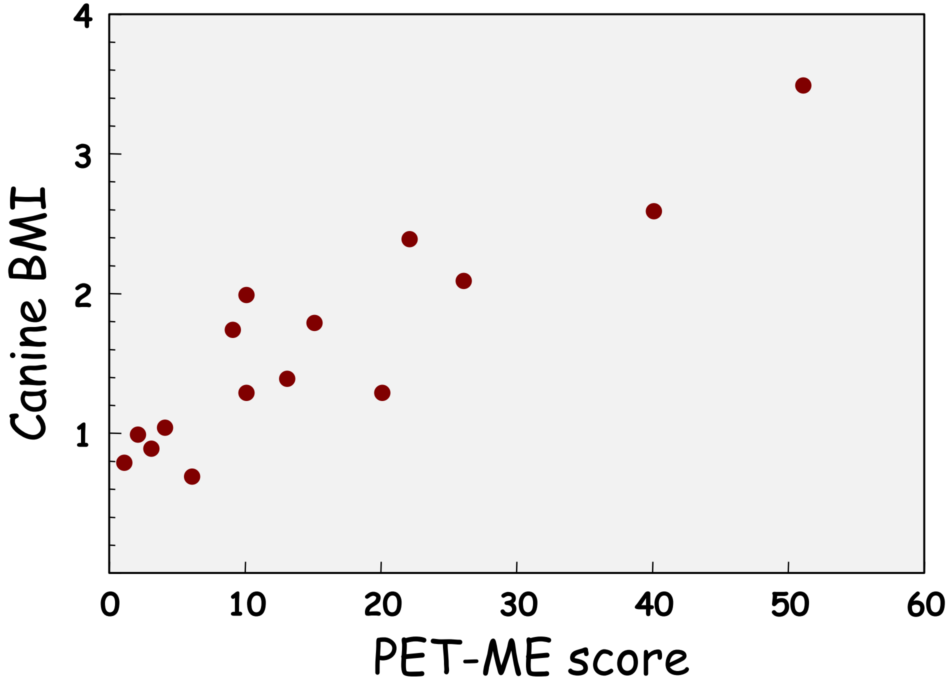
- *Is there a better way?*

Are PET-ME and BMI correlated?

- "Correlation" – used in everyday language
- Can be measured mathematically
- Correlation coefficients measure
  - Whether or not there is a relationship
    - Do BMI and PET-ME go up and down together?
  - Strength of relationship
  - Direction of relationship
- Bivariate scatter plots show correlation visually

The Study

- We enroll 15 willing clients and their purebred dogs from our practice
- We ask owners to fill out our PET-ME questionnaire
- We then weigh the dogs and calculate our BMI (observed weight/ideal weight)
- We display our data in a scatterplot

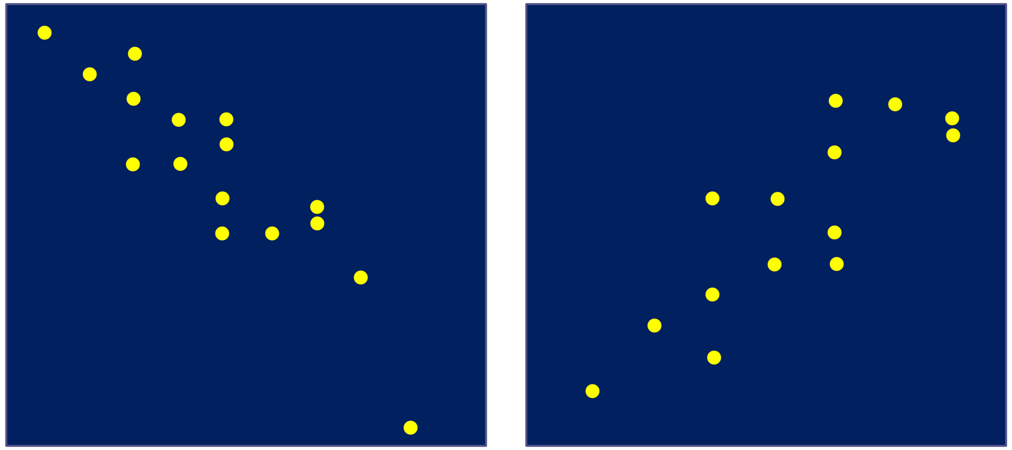


Interpreting the Scatterplot

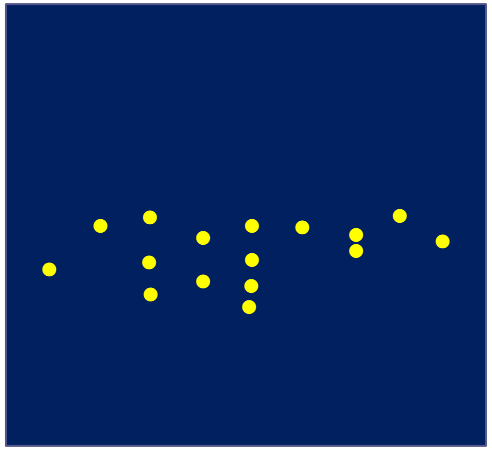
- After plotting two variables using a scatterplot, we can examine the plot for an overall pattern
- We seek to describe the relationship in terms of form, direction, and strength of the association
  - Form: linear, curves, clusters, no pattern
  - Direction: positive, negative, no direction
  - Strength: relates to how closely the points fit the "form"

Form

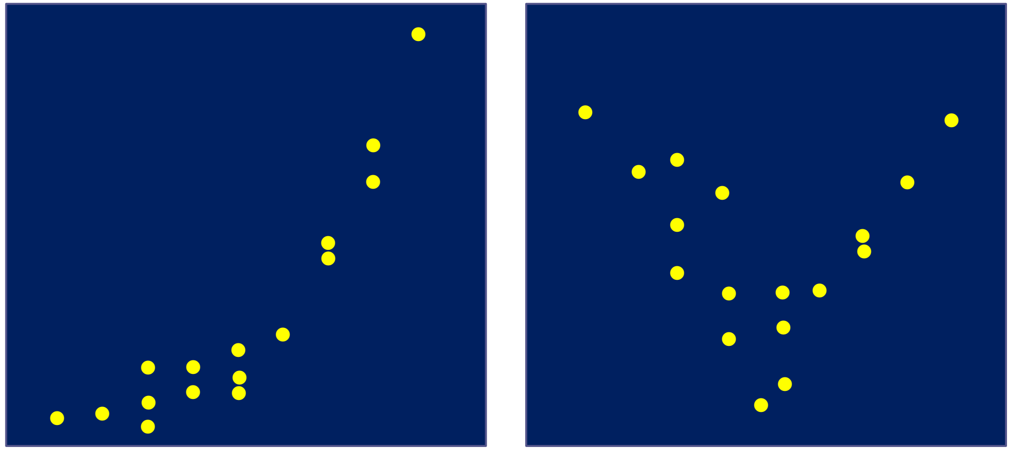
Linear

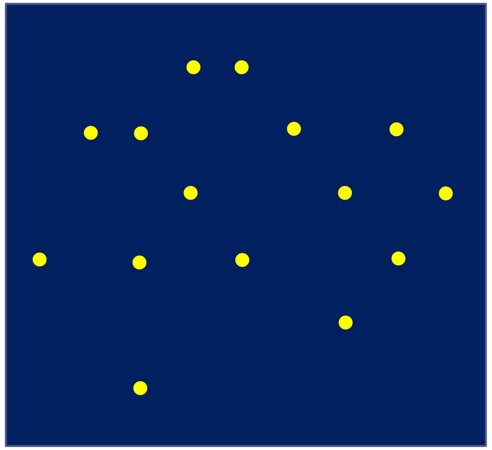


No relationship



Nonlinear

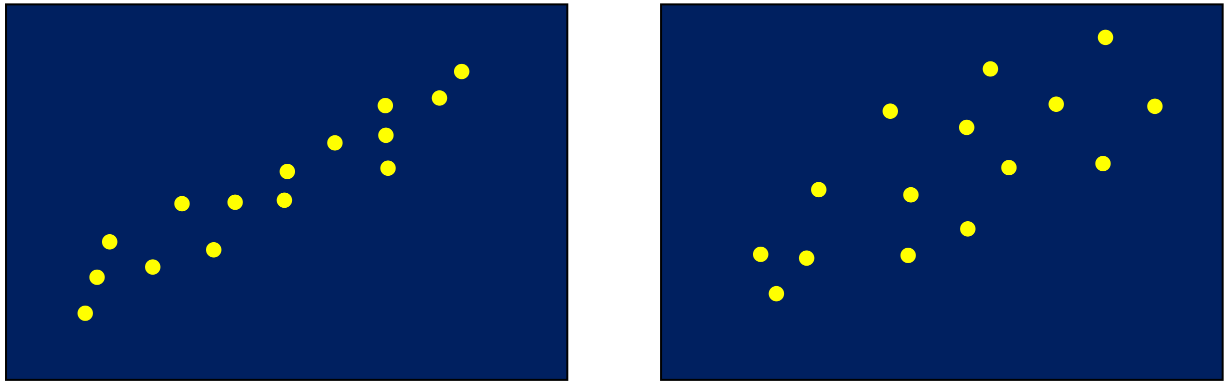




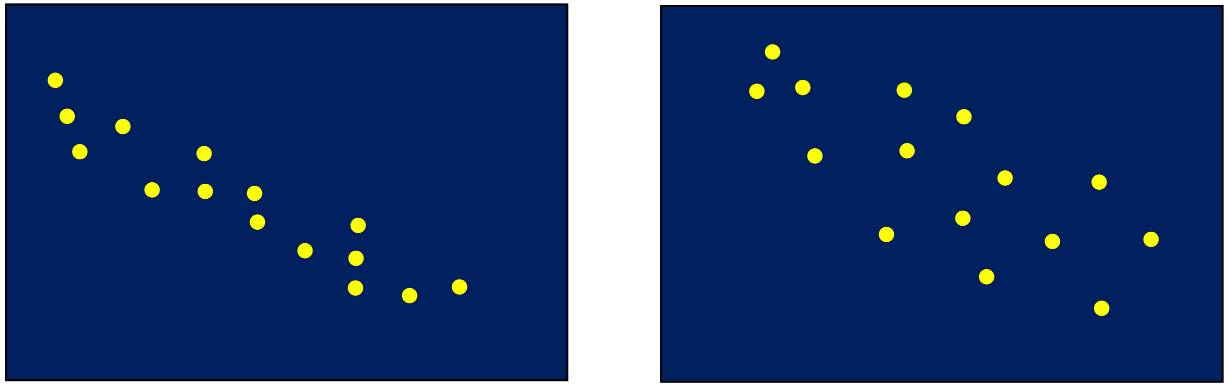
The Direction of a Correlation

- The correlation coefficient, *r*, is used to measure the strength and direction of the linear relationship, or correlation between two interval/ratio variables
- The value of *r* ranges from -1.0 to +1.0
- Values closer to ± 1.0 indicate stronger correlations
- The sign of the correlation coefficient (- or +) indicates only the direction or slope of the correlation

Direction



- Positive (or direct) Relationship: As X gets larger, so does Y



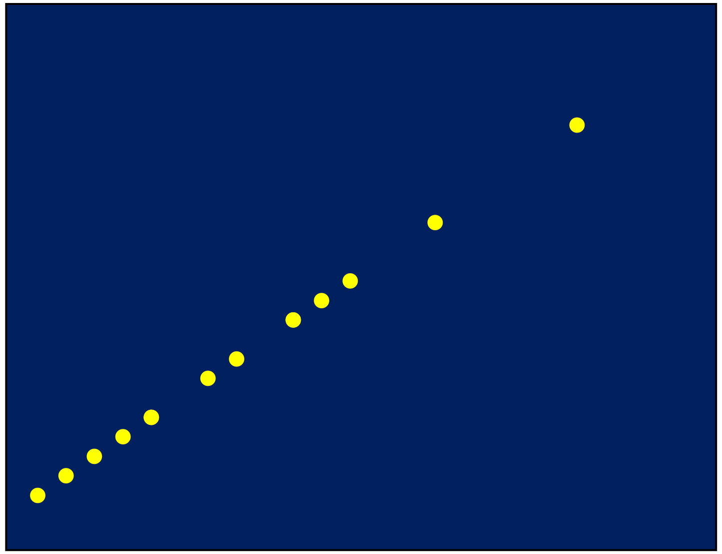
- Negative (or inverse) Relationship: As X gets larger, Y get smaller

Positive, negative, or no association?

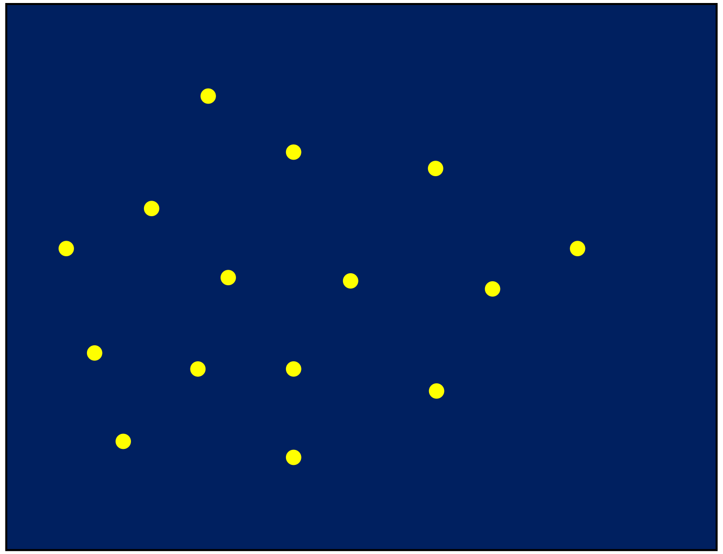
- Age and blood pressure
- Weight of car and miles per gallon
- 2 different depression scales
- Age of child and vocabulary size
- Adult height and vocabulary size
- Time since injury and disability score (higher score means more disability)

Strength

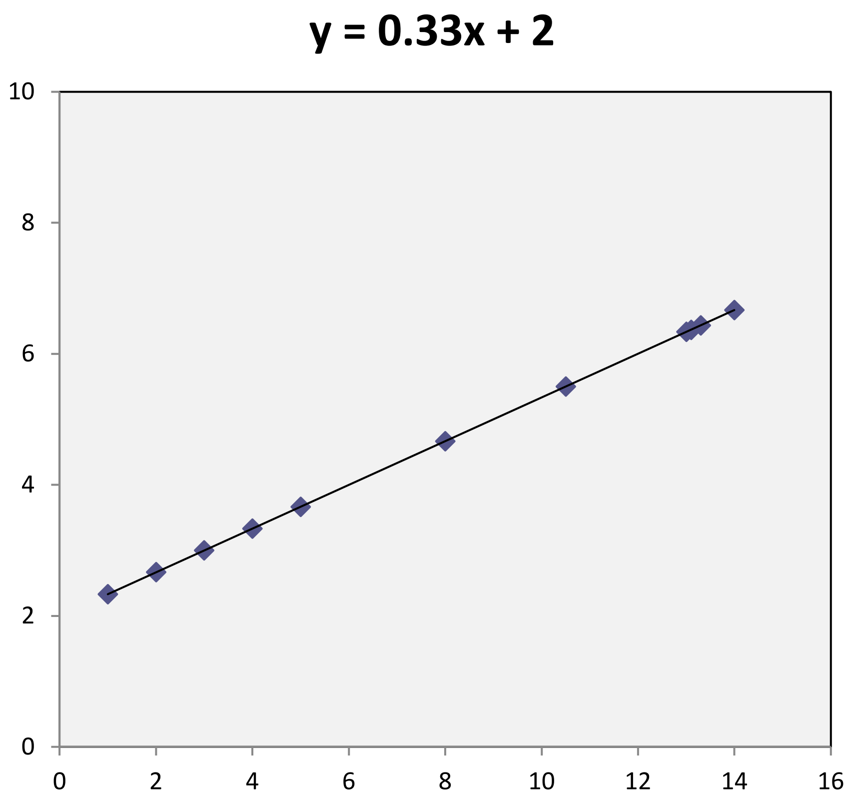
Perfect Relationship

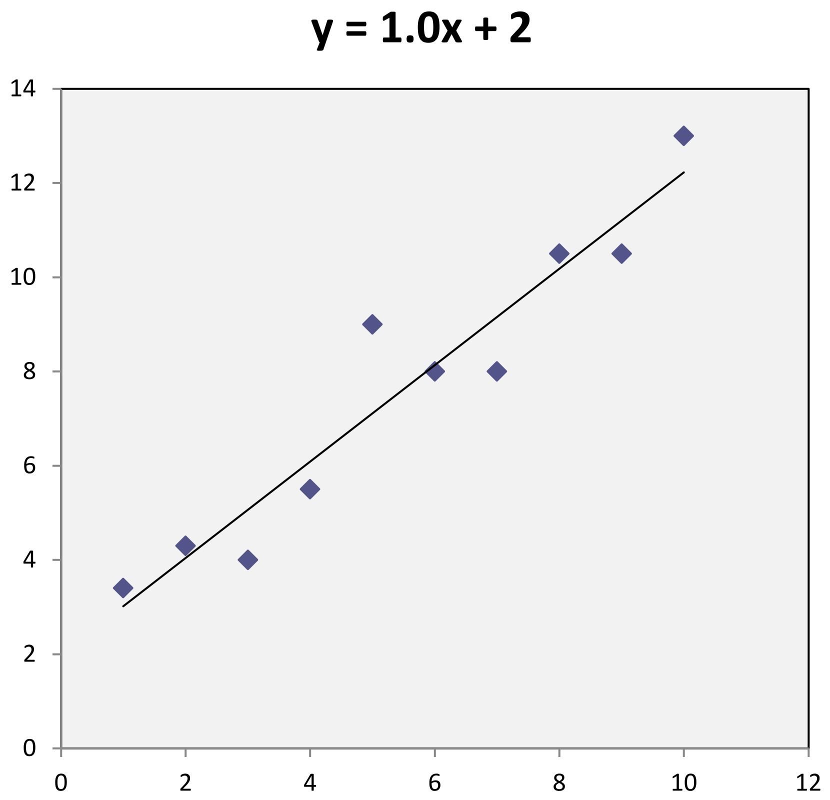


No Relationship



Which is the stronger correlation?





This one!

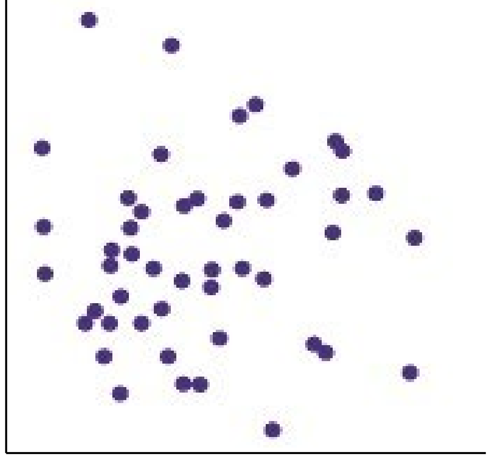
Reasons to Infer a Relationship

- The line relating the 2 is not horizontal (i.e., the slope is not zero)
- The closer the points fall to the fitted line, the stronger the relationship
  - How far the line differs from the horizontal does not tell us about the strength of the relationship

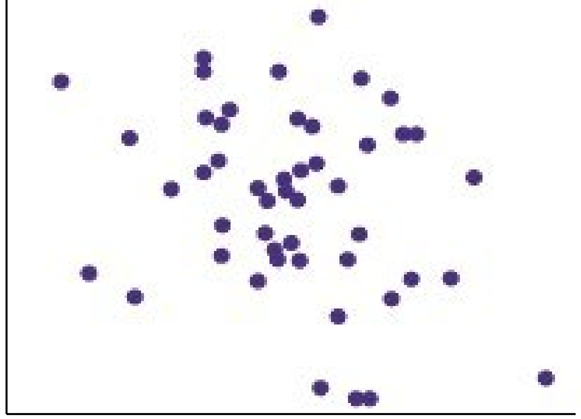
Pearson's Correlation Coefficient r

- Quantifies how close points fit a line
  - Extent to which a straight line describes the association
- Pearson's correlation coefficient
  - Population parameter: ρ (rho)
  - Sample estimate: *r*
  - Range: between -1.0 and +1.0
    - -1.0 = perfect negative correlation (line would pass through every point and association is negative)
    - +1.0 = perfect positive correlation (line would pass through every point and association is positive)
    - 0 = no correlation

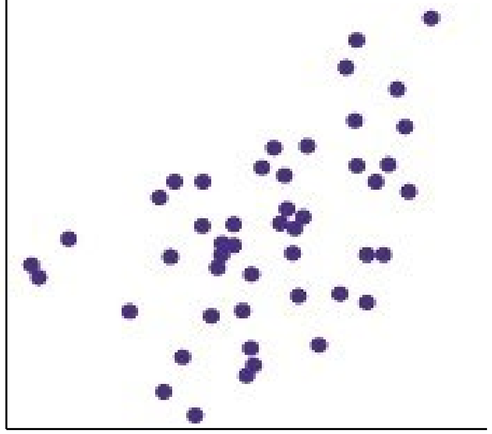
Examples of Correlation *r*



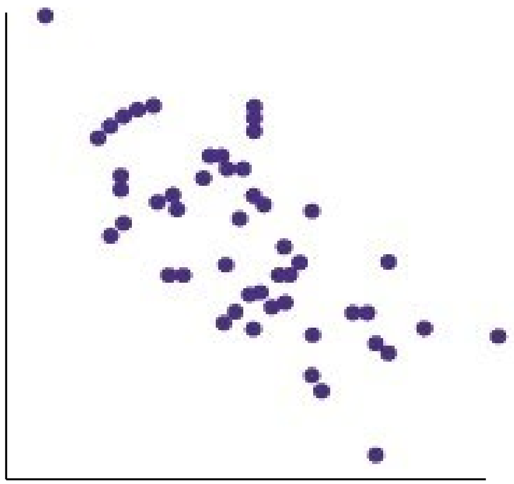
Correlation $r = 0$



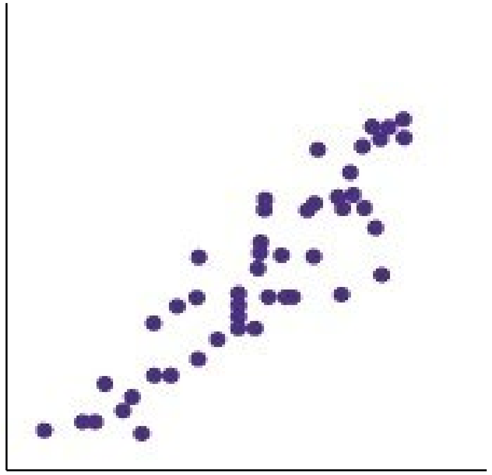
Correlation $r = -0.3$



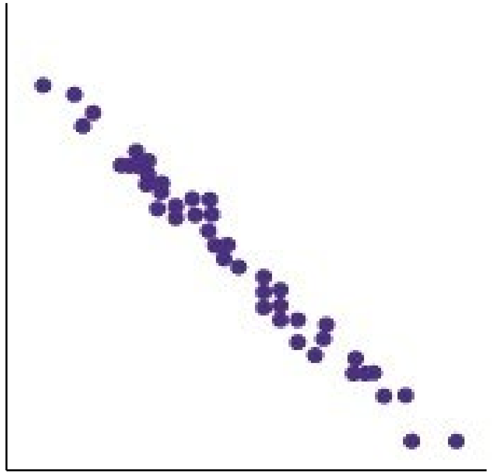
Correlation $r = 0.5$



Correlation $r = -0.7$



Correlation $r = 0.9$



Correlation $r = -0.99$

The Messy Formula

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Signal

Noise

- The numerator expresses the extent that X and Y vary together (signal)
  - Called the covariance of X and Y or cov (X,Y)
- The denominator is based on the deviations of each X from the mean of X and of each Y from the mean of Y (noise)
- When the extent to which X and Y vary together is large compared to the deviations of X and Y from their respective means, (absolute value of) r is larger

Back to our example

$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i-\bar{X})(Y_i-\bar{Y})$	$(X_i-\bar{X})^2$	$(Y_i-\bar{Y})^2$
-14.47	-0.84	12.15	209.28	0.71
-13.47	-0.64	8.62	181.35	0.41
-12.47	-0.74	9.23	155.42	0.55
-11.47	-0.59	6.77	131.48	0.35
-9.47	-0.94	8.90	89.62	0.88
-6.47	0.11	-0.71	41.82	0.01
-5.47	0.36	-1.97	29.88	0.13
-5.47	-0.34	1.86	29.88	0.12
-2.47	-0.24	0.59	6.08	0.06
-0.47	0.16	-0.07	0.22	0.03
4.53	-0.34	-1.54	20.55	0.12
6.53	0.76	4.97	42.68	0.58
10.53	0.46	4.85	110.95	0.21
24.53	0.96	23.55	601.88	0.92
35.53	1.86	66.09	1262.62	3.46
		143.27	2913.73	8.52

Mean PET-ME = 15.47

Mean BMI = 1.64

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

$$= \frac{143.3}{\sqrt{2913.7 * 8.52}}$$

$$= .909$$

If we square r to get r^2…

- Will always be positive
  - Ranges 0 to 1
  - $r^2 = 0$, no linear association
  - $r^2 = 1$, perfect linear association (negative or positive)
- r^2 called coefficient of determination
- Expresses the proportion of variance in the dependent variable (Y) explained by the independent variable (X)
  - If we know X, how much does that tell us about Y

Interpretation of *r and* r^2

- What are considered reasonable *r and* r^2?
- $r = 0.7$, $r^2 = 0.49$, viewed favorably
  - x explains about half the variance in y
- $r = 0.3$, $r^2 = 0.09$, not viewed favorably
  - only explains ~10% of the variation

Is *r* statistically significant?

Note: We estimate the population correlation ρ with the sample correlation *r*

- State the hypotheses
  - H_0: $\rho=0$ or PET-ME scores are not related to canine BMI
  - H_A: $\rho\neq0$ or PET-ME scores are related to canine BMI
- Set the criteria for a decision
  - We will compute a two-tailed test at a .05 level of significance
  - df for a correlation are n – 2
  - df for this example are 15 – 2 = 13
- Critical values can be found in the posted table: ±0.514

Critical Values of the Pearson Correlation Coefficient

Level of Significance for One-Tailed Test				
df	0.05	0.025	0.01	0.005
Level of Significance for Two-Tailed Test				
	0.10	0.05	0.02	0.01
1	.988	.997	.9995	.9999
4	.729	.811	.882	.917
8	.549	.632	.716	.765
10	.497	.576	.658	.708
13	.441	.514	.592	.641
15	.412	.482	.558	.606
:				
100	.164	.195	.230	.254

2-sided test, α=0.05, df = 13, what is the critical value?

Hypothesis Testing (continued)

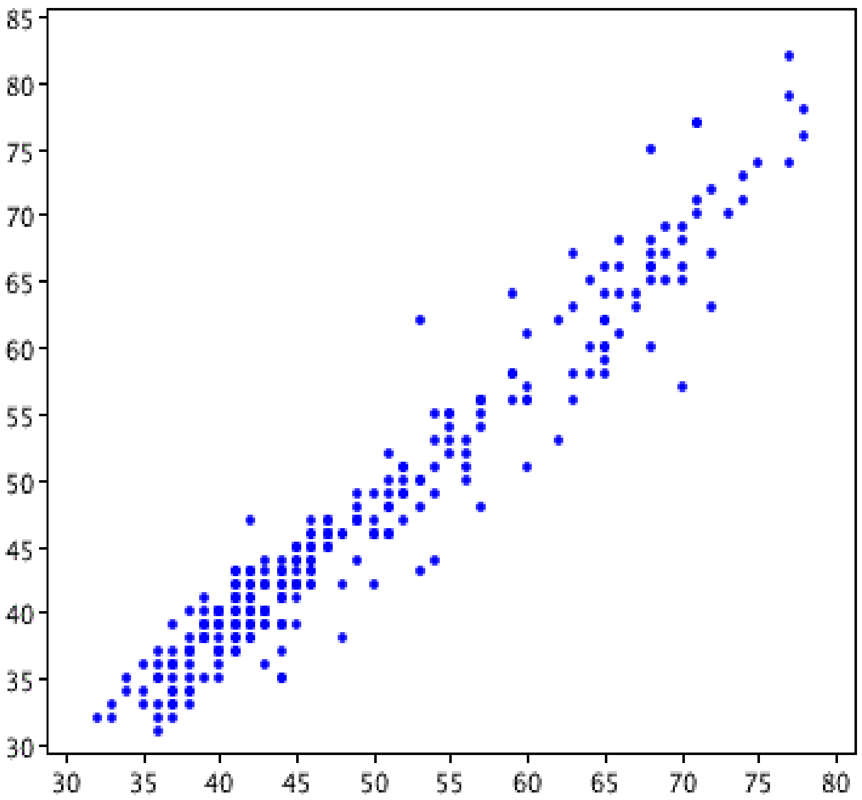
- Compute the test statistic
  - The correlation coefficient *r* is the test statistic for the hypothesis test (previously computed)
  - $r = .909$
- Make a decision
  - Because $r = .909$ exceeds the critical values ($r = \pm 0.514$), we reject the null hypothesis
  - Based on this data, we conclude that the observed PET-ME scores and canine BMI reflect a significant correlation between the two variables in the population

Assumptions for Pearson's *r*

- Both X and Y are normally distributed
- The values of Y are independent from one another
  - Value of 1 observation not related to value of another observation (e.g. no dogs from the same litter)
- The observations are drawn from a random sample
- There is equal variation in the *Y*'s across the entire range of *X*'s
  - This is called homogeneity or homoscedasticity
- There is a linear relationship

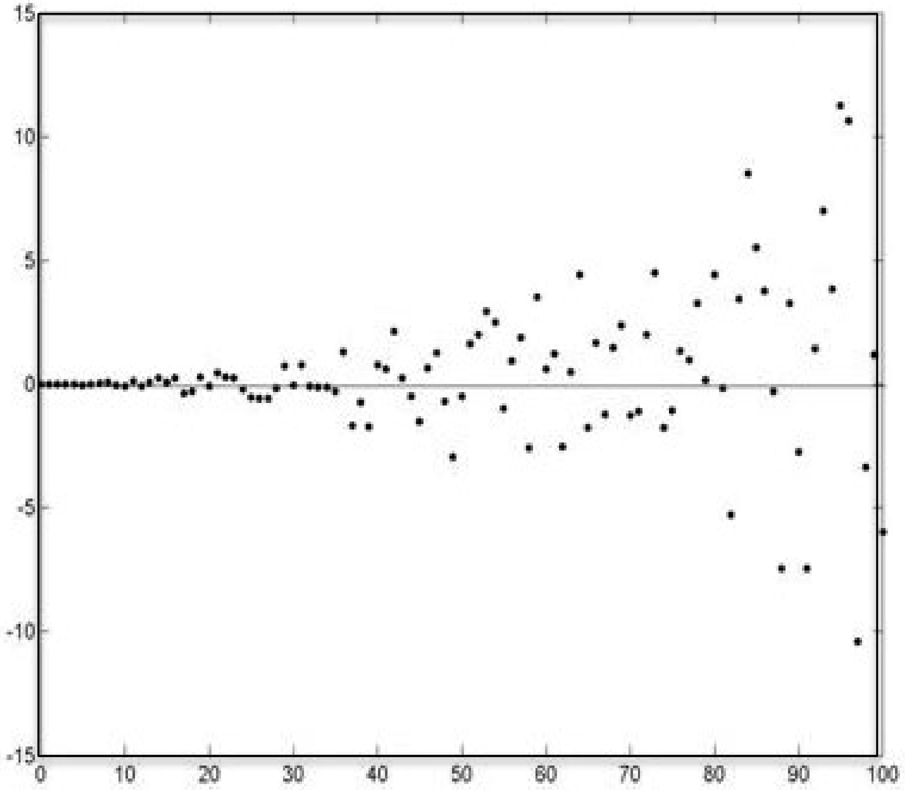
Homoscedasticity vs Heteroscedasticity

Homoscedastic



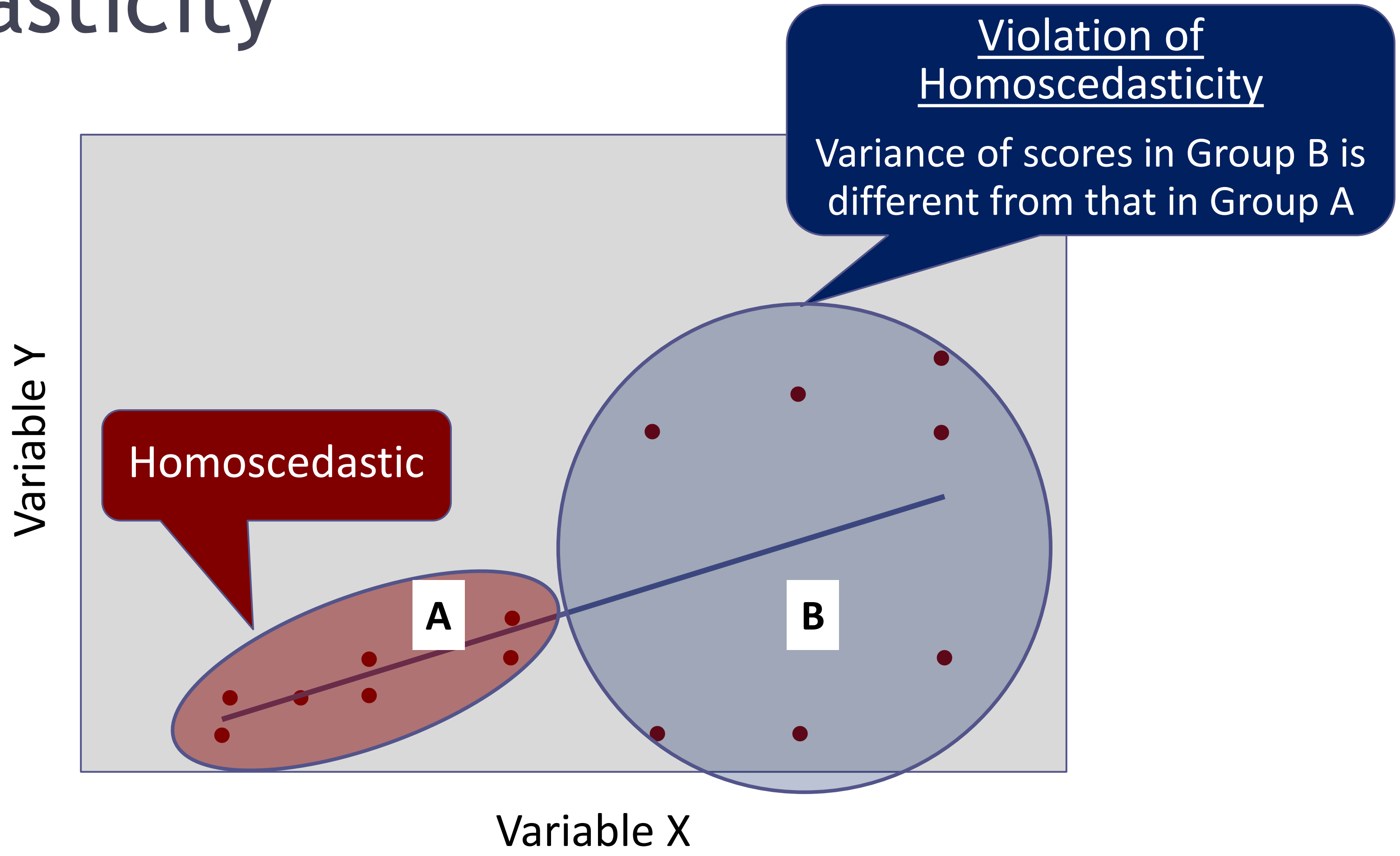
Linear Regression appropriate

Heteroscedastic



Linear Regression not appropriate

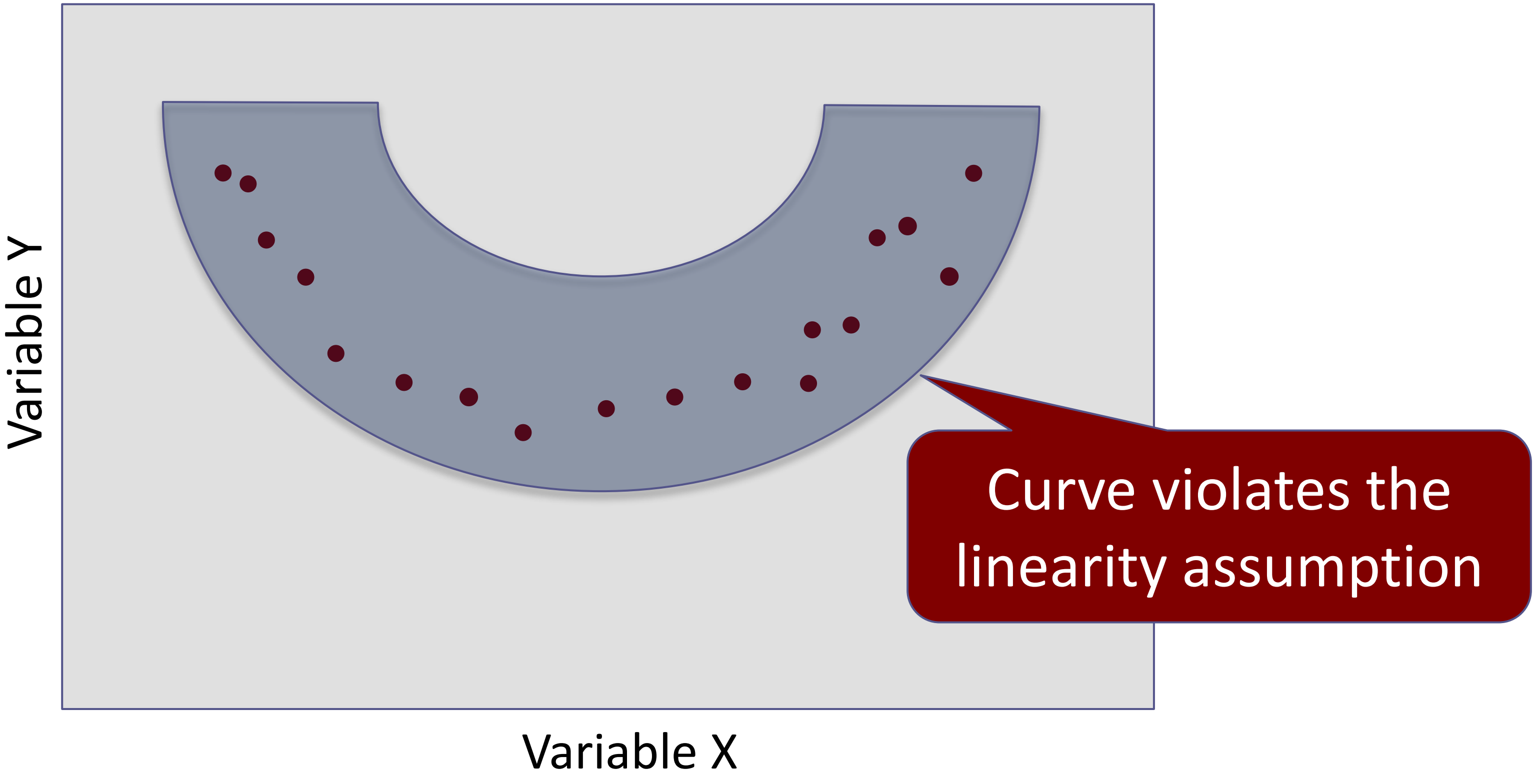
Homoscedasticity



- Homoscedasticity is assumed
- The variability in Y is consistent along the regression line

Linearity

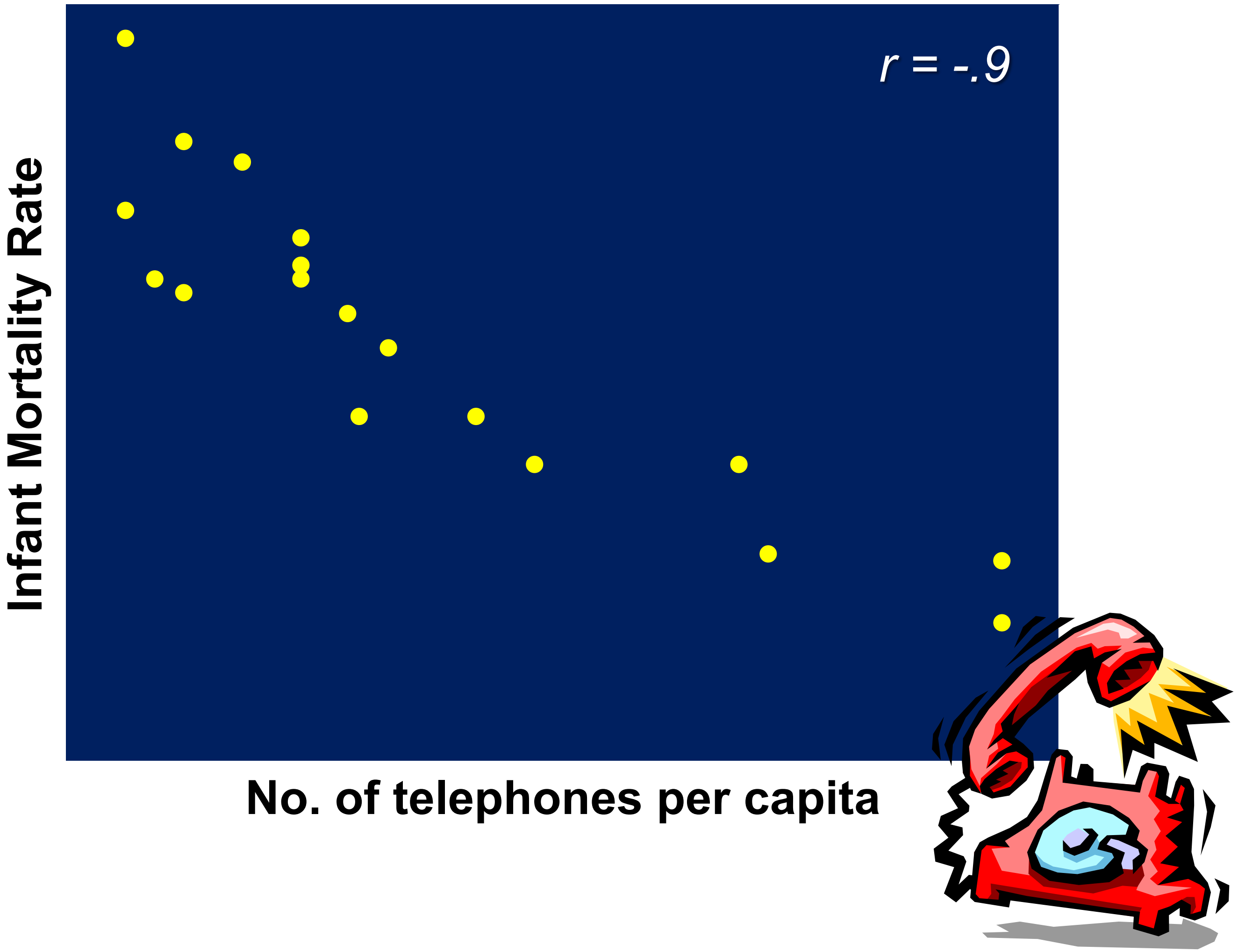
- Linearity is assumed
- The best way to describe the relationship between the two variables is by using a straight line



Limitations of Interpretation

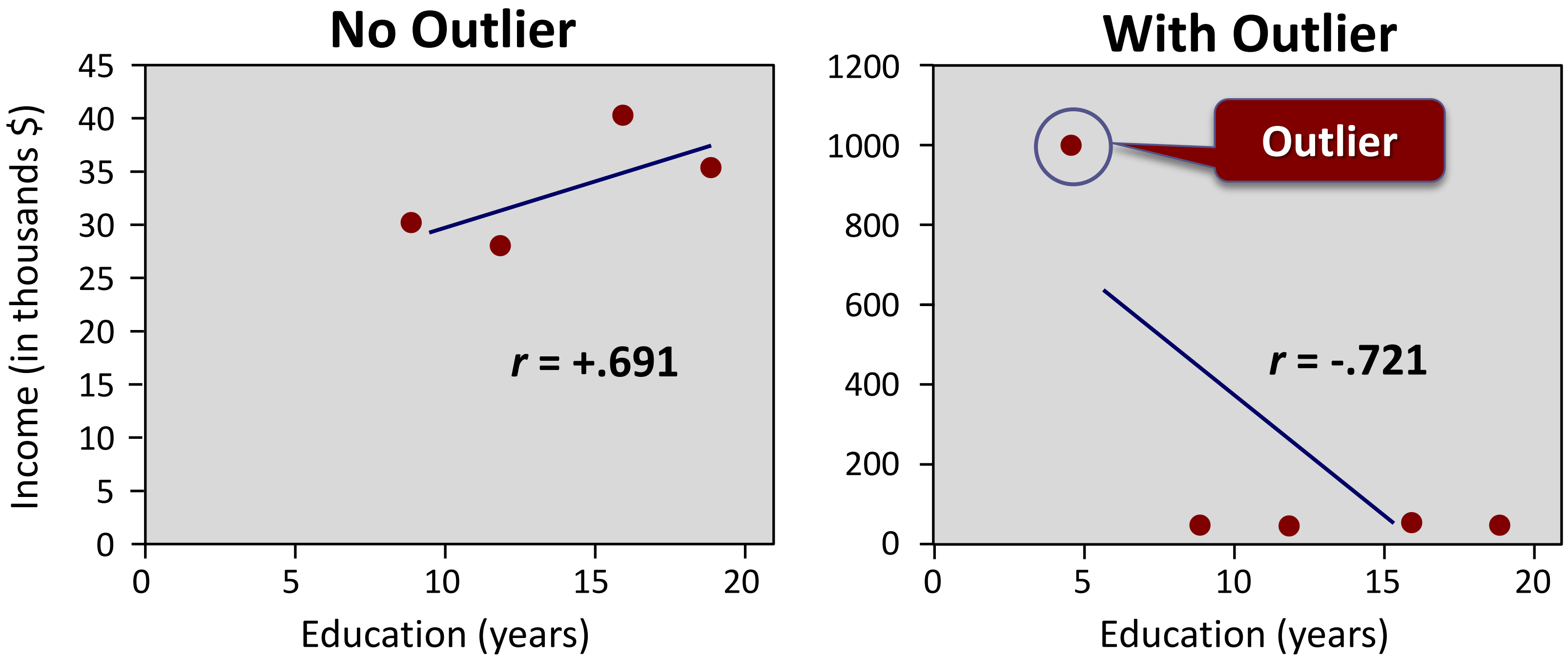
- Correlation does not equal causation
- Just because you can predict *Y* from *X* and just because the correlation is significant at $p<.0001$, ***does not*** mean *X* causes *Y*
- Equally plausible that *Y* causes *X* or yet both result from some other thing such as *Z*

Country Statistics

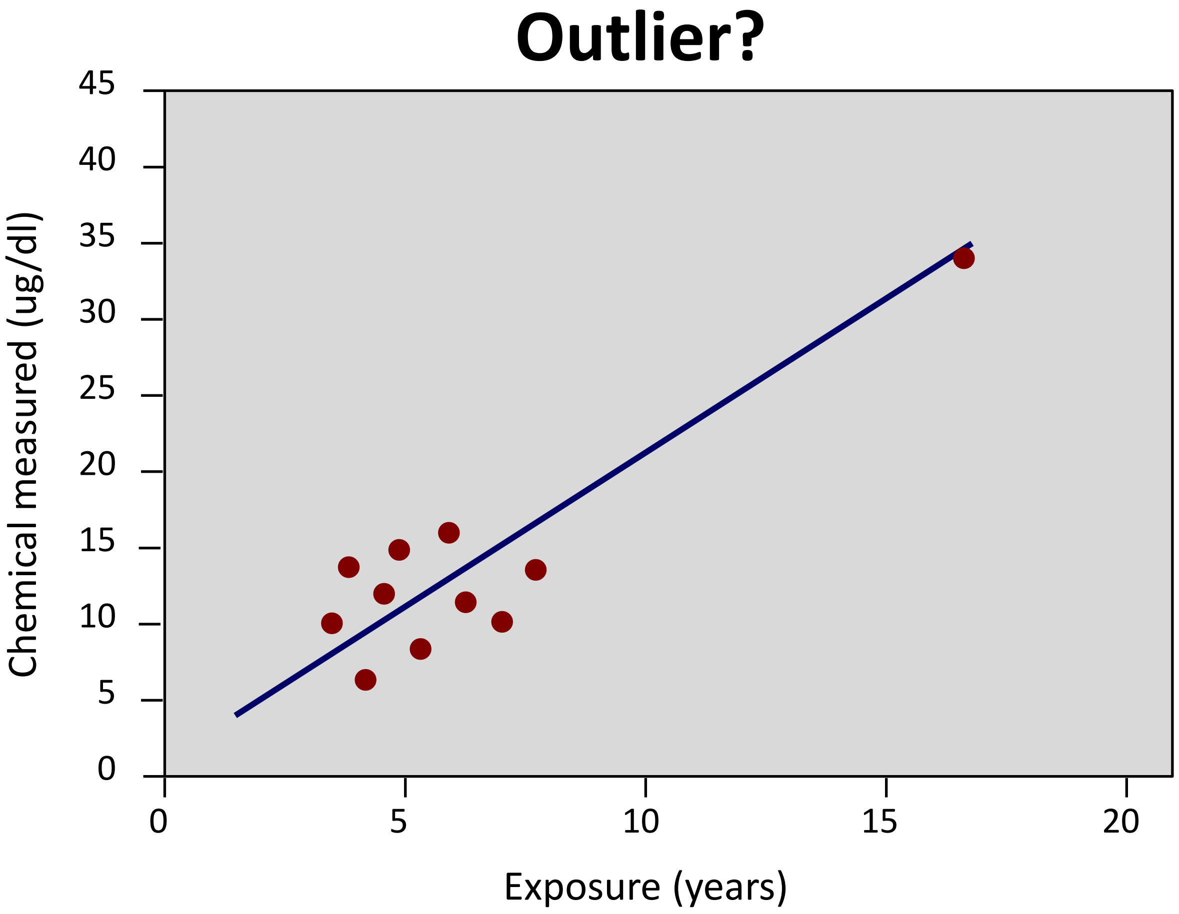


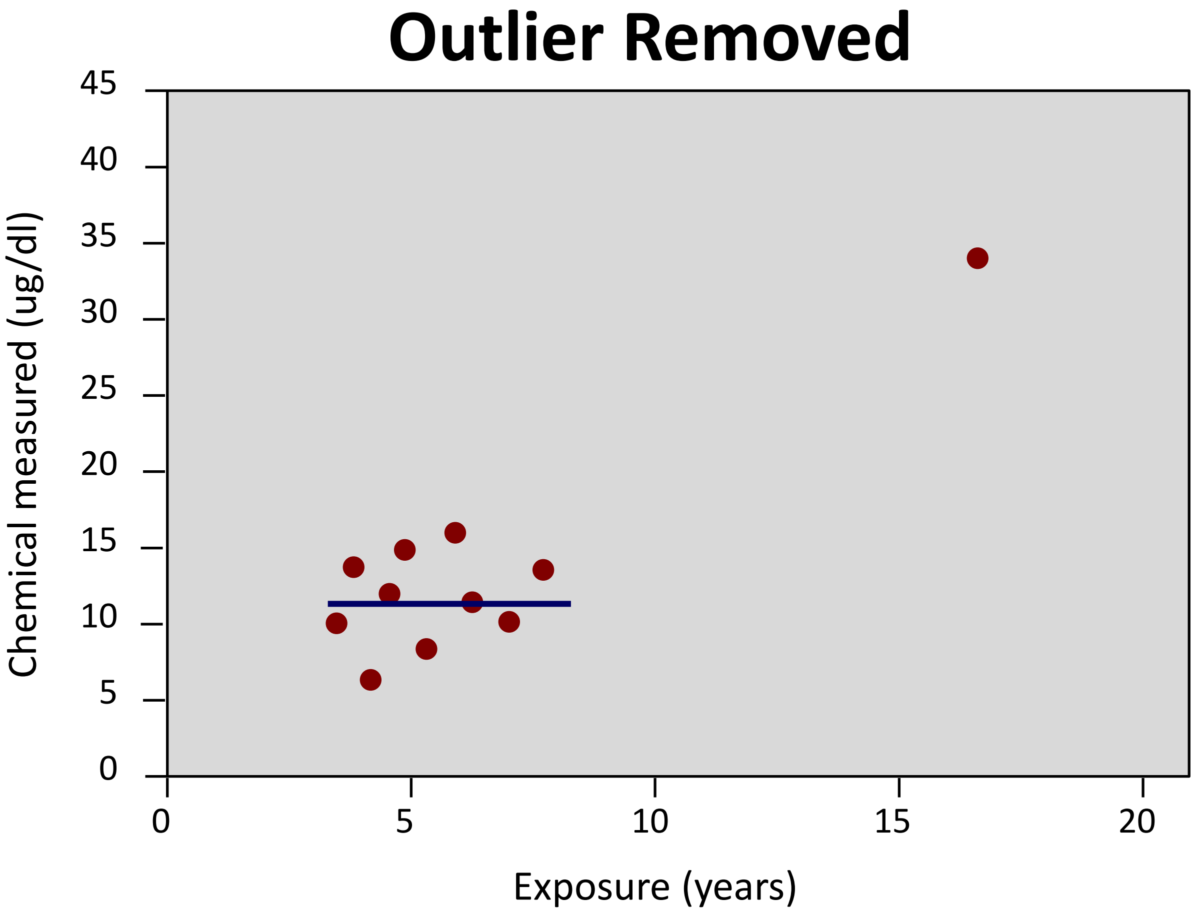
Outliers

- An outlier is a score that falls substantially above or below most other scores in a data set
- Outliers can obscure relationship between two variables by altering direction & strength of observed correlation



Outliers: A second example



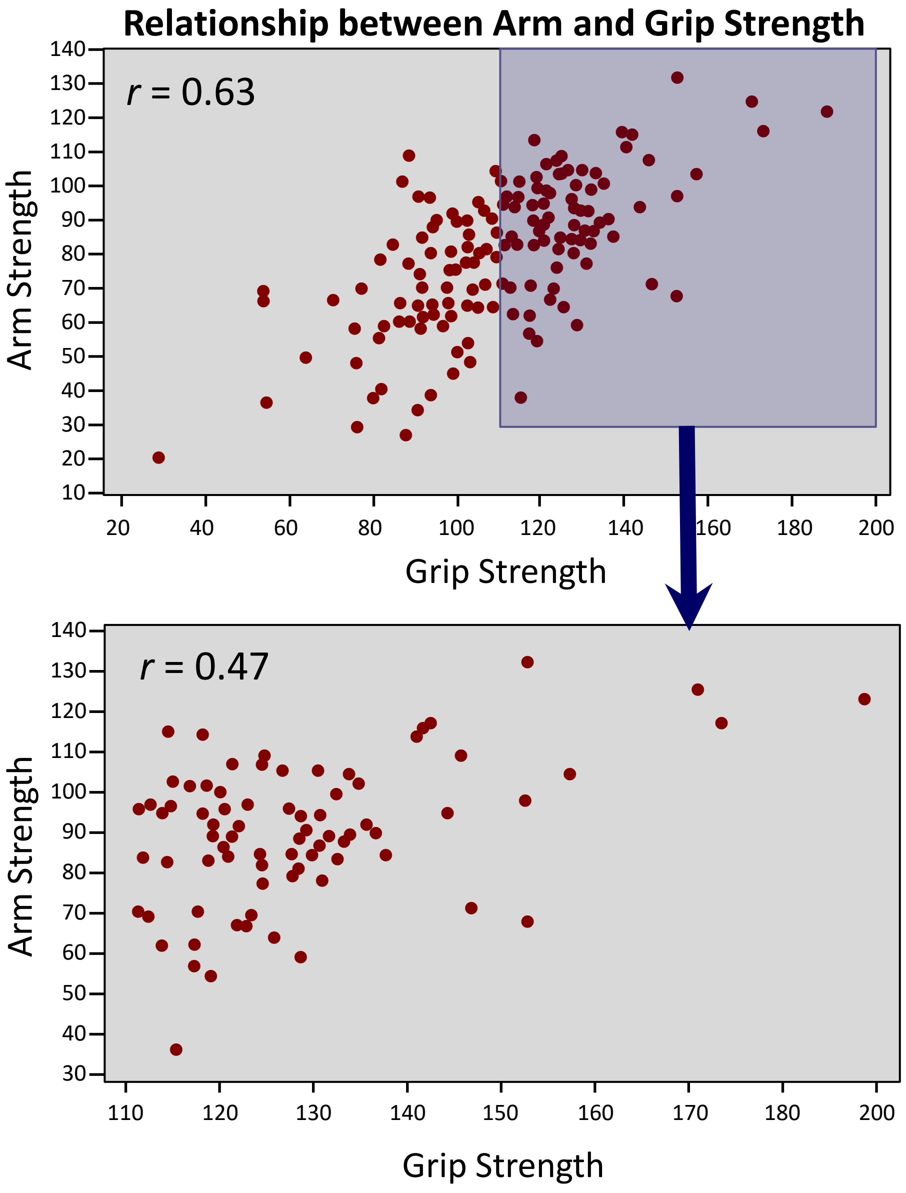


Restriction of Range

- Problem arises when range of data for one or both correlated variables in a sample is limited/restricted, compared to the range of data in the population from which the sample was selected
- When interpreting a correlation, it is important to avoid making conclusions about relationships that fall beyond the range of data measured

Restriction of Range

- Figure to right shows scatterplot of grip and arm strength for people working in physically demanding jobs
- N=147, *r*=0.63



- Lower figure shows scatterplot for 73 workers with the highest grip strength
- Note scales of x-axes are different
- Whenever a sample has a restricted range of scores, the correlation will be reduced

From Relationships to Predictions

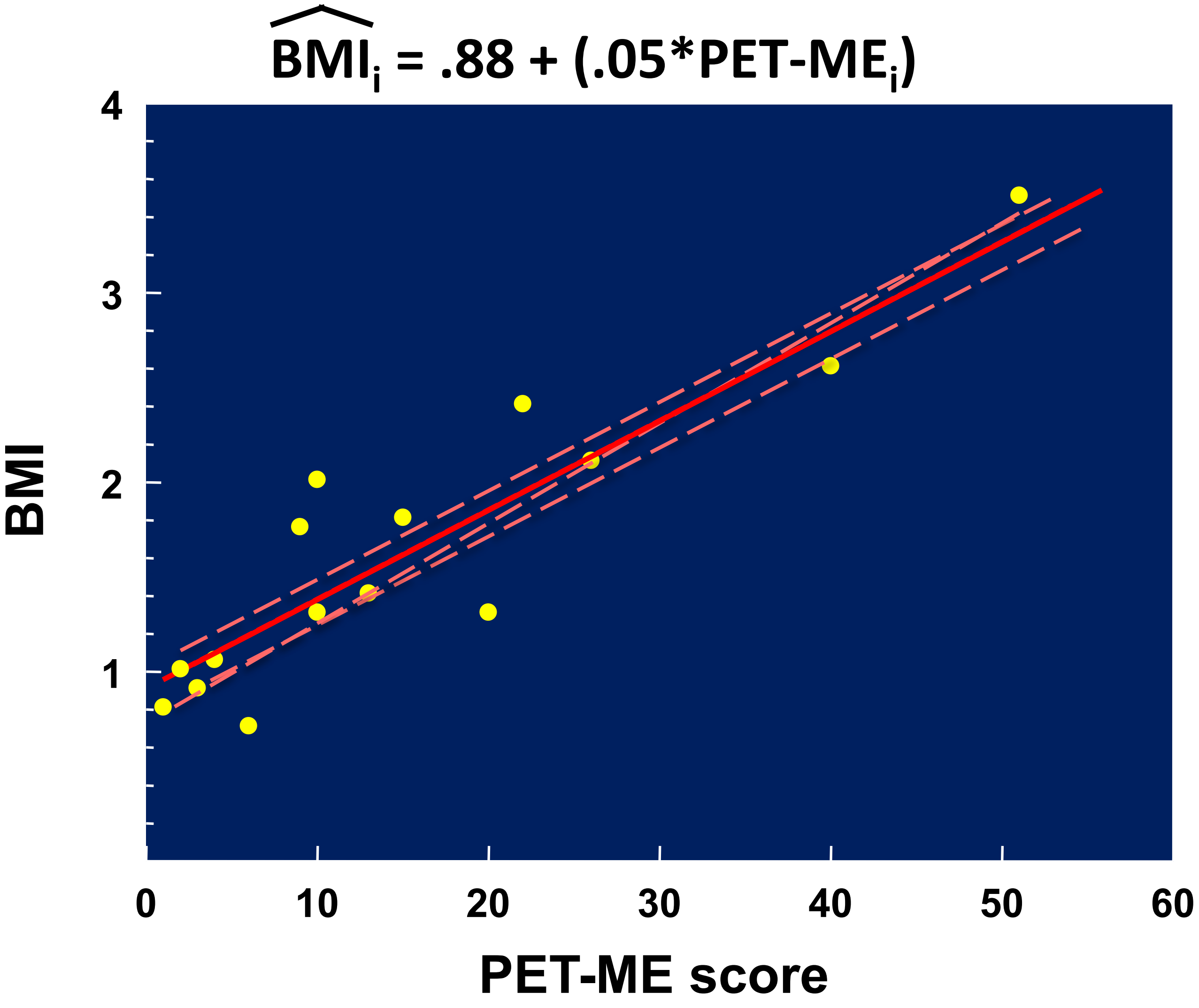
- Can also use information provided by *r* to predict values of one variable, given known values of a second variable
- Linear regression is a statistical procedure used to estimate the equation of a regression line to a set of data points
- It is used to determine the extent to which the regression equation can be used to predict values of one variable, given known values of a second variable in a population

Fundamentals of Linear Regression

You can use linear regression to answer the following questions about the pattern of data points and the significance of a linear equation:

- Is a pattern evident in a set of data points?
- Which equation of a straight line best describes this pattern?
- Are the predictions made from this equation significant?

Back to PET-ME & Canine BMI



Remember Algebra??

- The straight line formula:
  - $y = a + bx$ (or $y = mx + b$)
  - where *a* is the intercept or the value of y when x equals zero
  - and b is the slope or the amount of change in y for one unit change in x

Rewriting the equation

- Change *a* to β_0
- Change *b* to β_1
- With our example the equation now reads:

Called "hat"

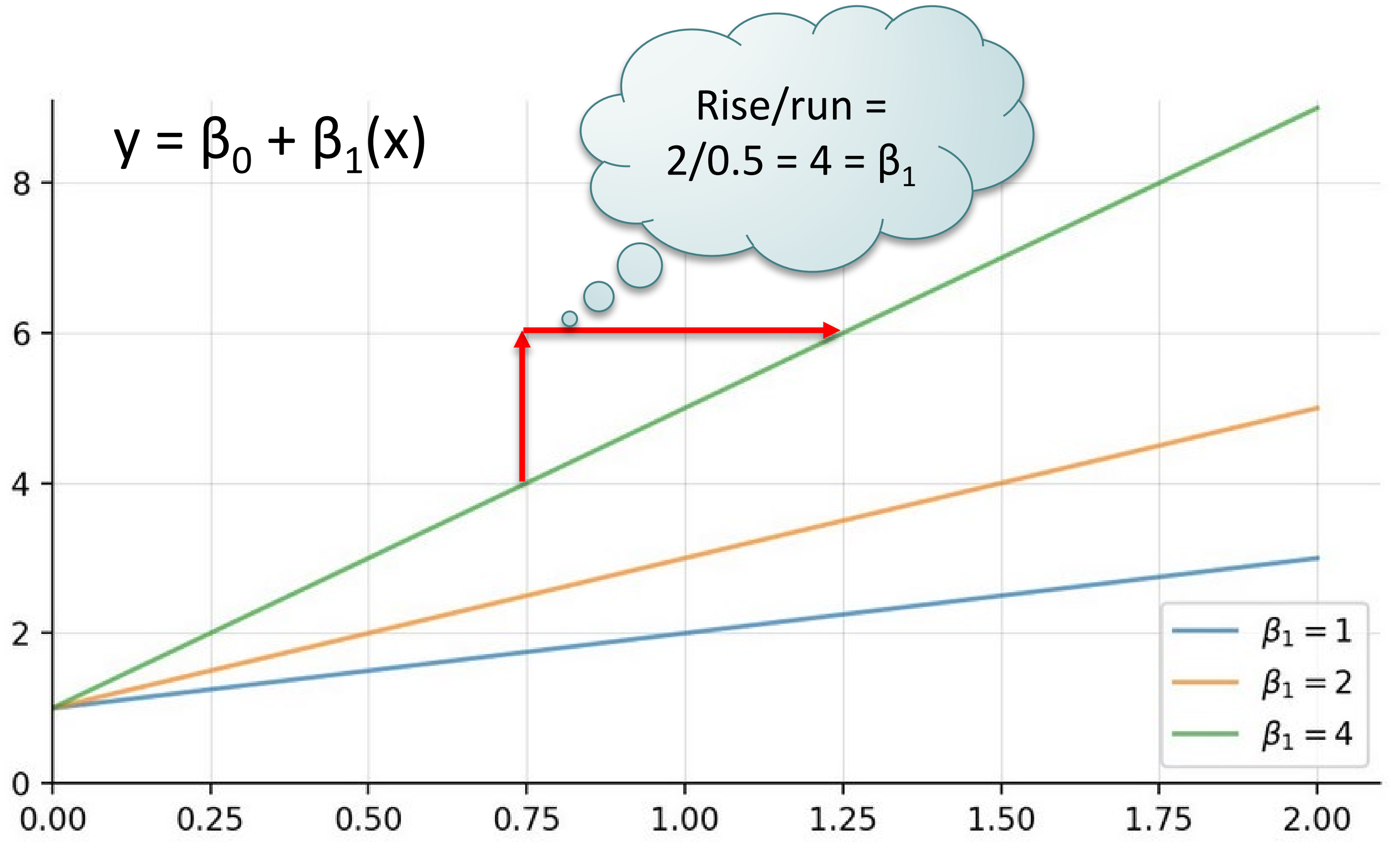
$$\widehat{BMI} = \beta_0 + \beta_1(PET\text{-}ME)$$

- A "hat" over a variable means estimate rather than the original value

Interpreting the slope

- Slope = change in y/change in x = ‘rise over run’
- If slope = 2
  - For a 1 unit change in x, how much does y change?
    - Increases by 2 units
- If slope = -2
  - For a 1 unit change in x, how much does y change?
    - Decreases by 2 units

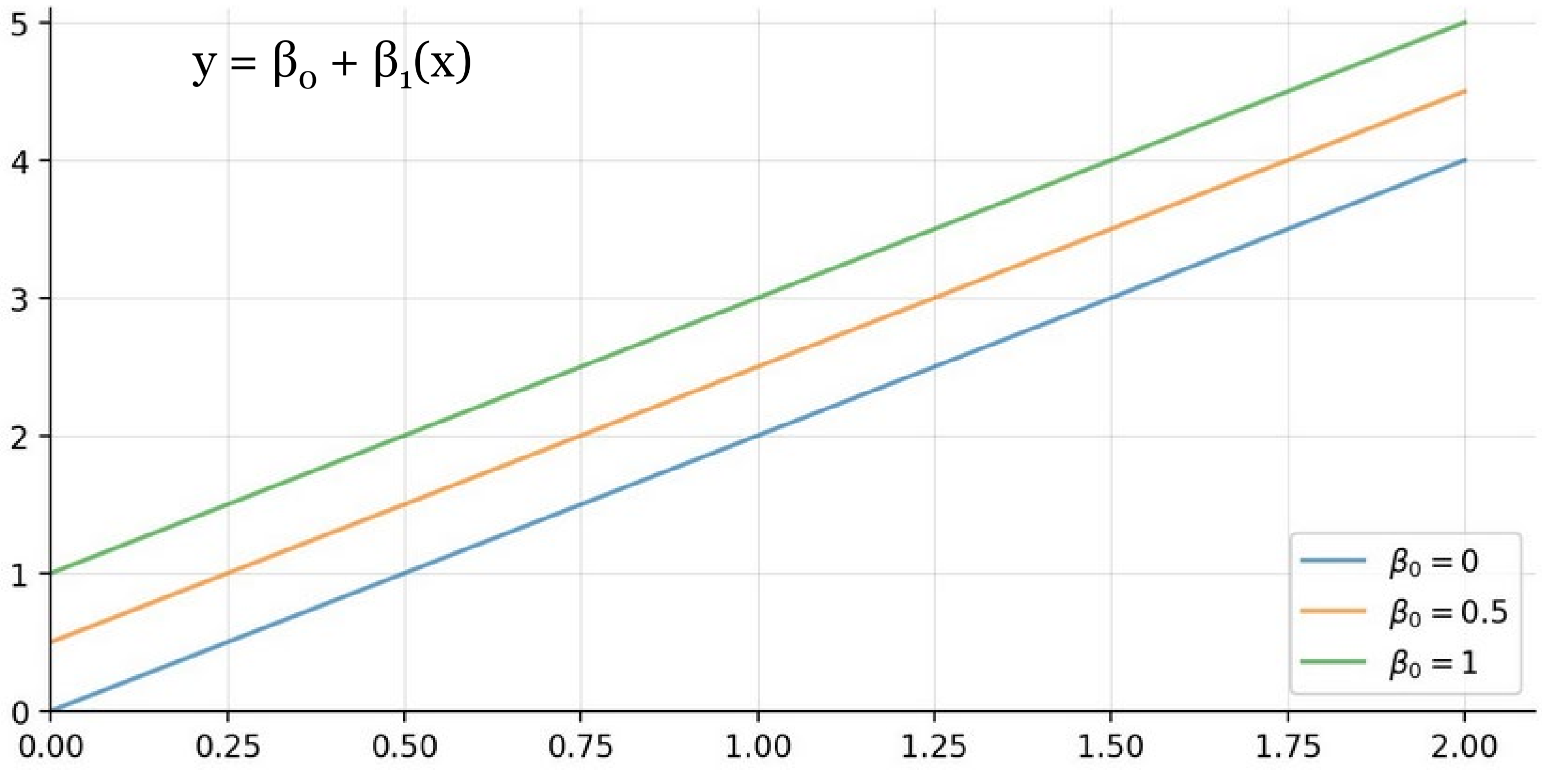
Coefficient = β_1 = Slope



Different coefficient values for the linear model: y = 1 + Beta1x

https://towardsdatascience.com/linear-regression-in-real-life-4a78d7159f16

Y Intercept = β_0



Different intercept values for the linear model: y = Beta0+ 2x

https://towardsdatascience.com/linear-regression-in-real-life-4a78d7159f16

Using the Equation

- $\widehat{BMI}_i = .88 + (.05 * \text{PET-ME}_i)$
- PET-ME scores ranged from about 1 to 60
- BMI for dogs = body weight/ideal body weight
- If PET-ME = 2, what is $\widehat{BMI}$?
  - $\hat{y} = 0.88 + 0.05(2) = 0.88 + 0.1 = 0.98$
- If PET-ME = 10, $\hat{y} = 0.88 + 0.05(10) = 1.38$
- If PET-ME = 15, $\hat{y} = 0.88 + 0.05(15) = 1.63$

Using the Equation

- To see how well the line estimates BMI of a study participant based on dog's PET-ME score
- To predict BMI based on any PET-ME score, even if no dog in the study had that score

Estimates or predictions of BMI for any PET-ME value

Dog #	PETME	BMI	$\widehat{BMI}$
1	1	0.80	0.93
2	2	1.00	0.98
3	3	0.90	1.03
4	4	1.05	1.08
5	6	0.70	1.17
6	9	1.75	1.32
7	10	2.00	1.37
8	10	1.30	1.37
9	13	1.40	1.52
10	15	1.80	1.62
11	20	1.30	1.86
12	22	2.40	1.96
13	26	2.10	2.16
14	40	2.60	2.85
15	51	3.50	3.39

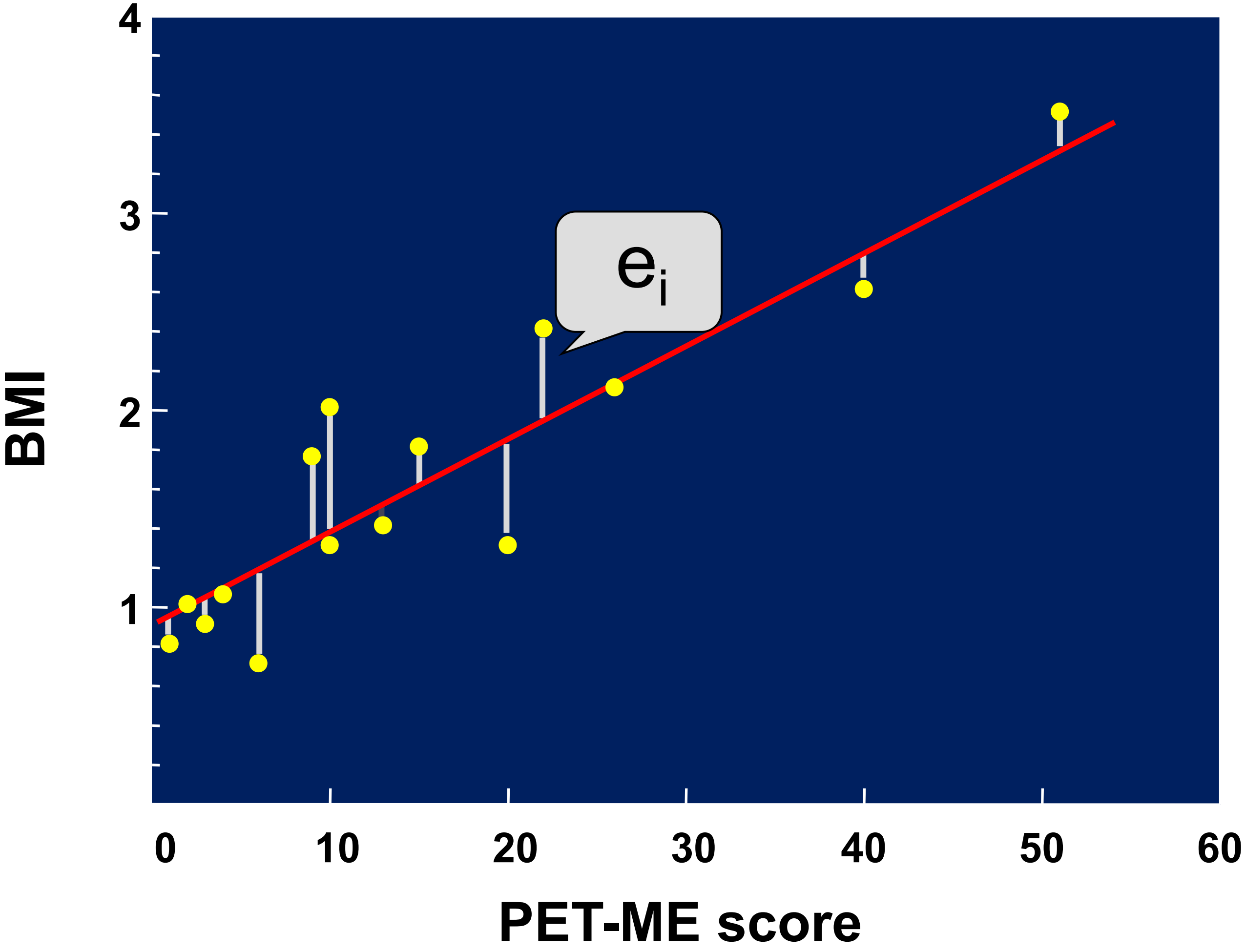
$$\widehat{BMI} = \beta_0 + \beta_1(PET\text{-}ME)$$

$$\widehat{BMI} = 0.88 + .05(PET\text{-}ME)$$

Selecting values for β_0 and β_1 to best fit the line

- Strategy used is to adjust the value in such a way as to
  - Maximize the variance resulting from the fitted line or
  - Minimize the variance resulting from deviations from the fitted line
- Optimal solution determined by least squares analysis

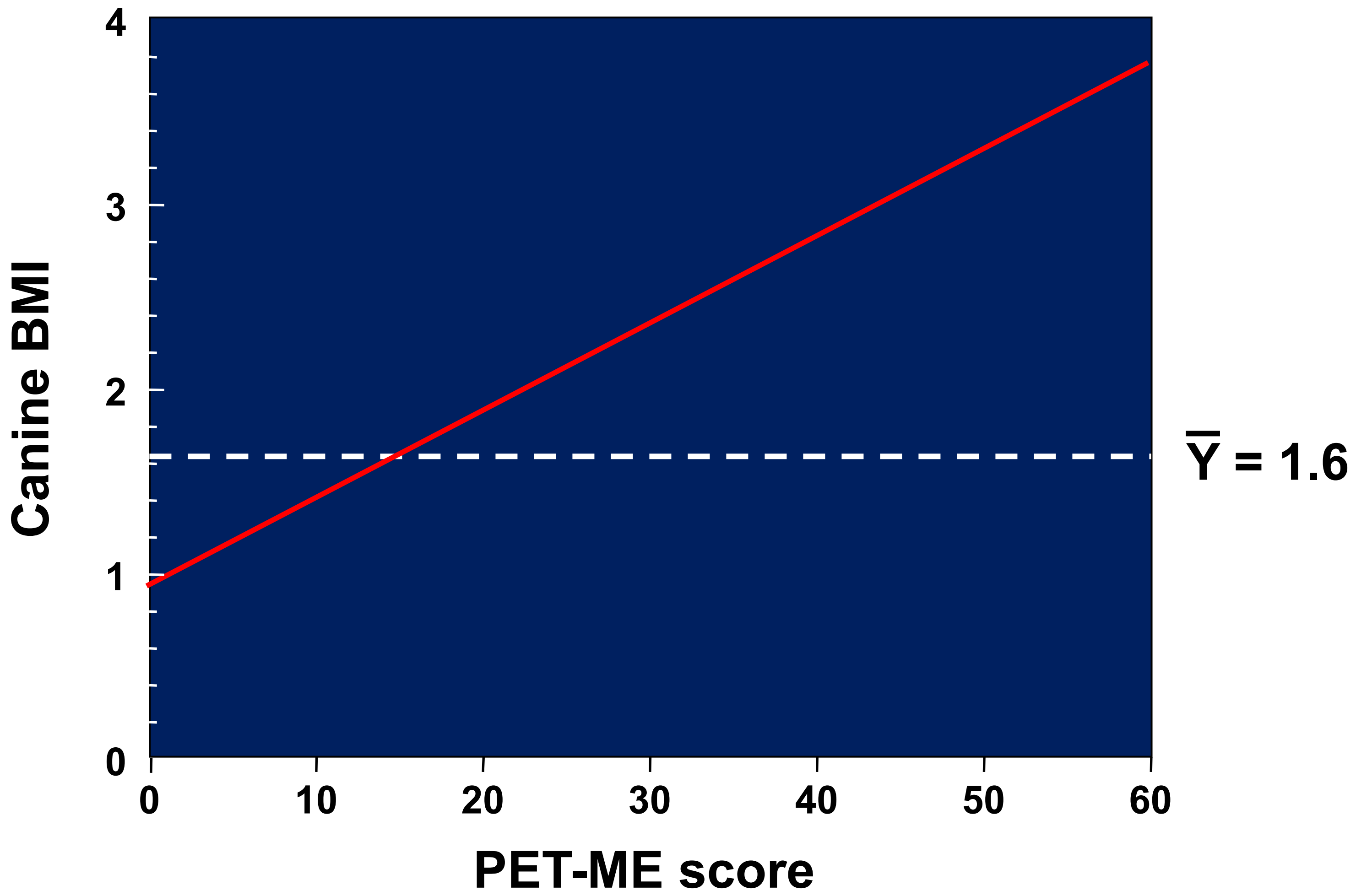
Minimizing Σe_i^2



But is the line statistically significant?

- Test whether slope = 0 (horizontal line)
- If no association, what would our best guess be for y for any value of x?
  - Mean of y

Is best line fit the data better than a horizontal line ?



Hypotheses

- Null hypothesis (H_0)
  - There is no linear relationship between PET-ME scores and BMI
  - H_0: $\beta_1 = 0$
- Alternative Hypothesis (H_A)
  - There is a linear relationship between PET-ME scores and BMI
  - H_A: $\beta_1 \neq 0$
- For 1-sided test, H_A: $\beta_1 > 0$ or $\beta_1 < 0$

Testing the “best” fit line

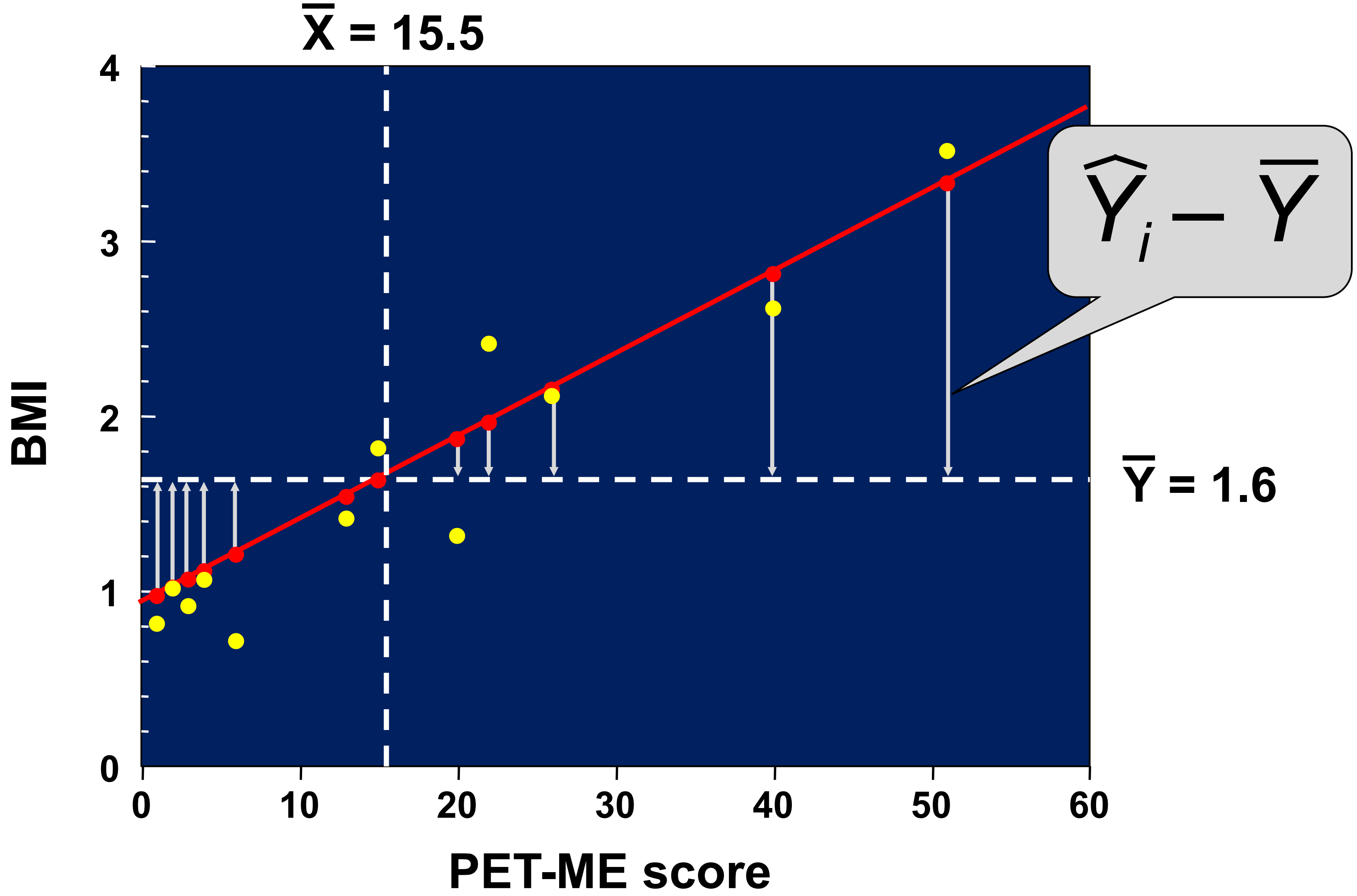
- We will be using Sums of Squares (SS) for our test
- The SS from the regression results from the difference between the fitted points and the horizontal line through the mean of *X* and *Y*

$$SS_{\text{regression}} = \Sigma(\hat{Y}_i - \bar{Y})^2$$

Signal

- Tells us how far the predicted values differ from the overall mean

Graphically depicted as



Sum of Squares of the residual

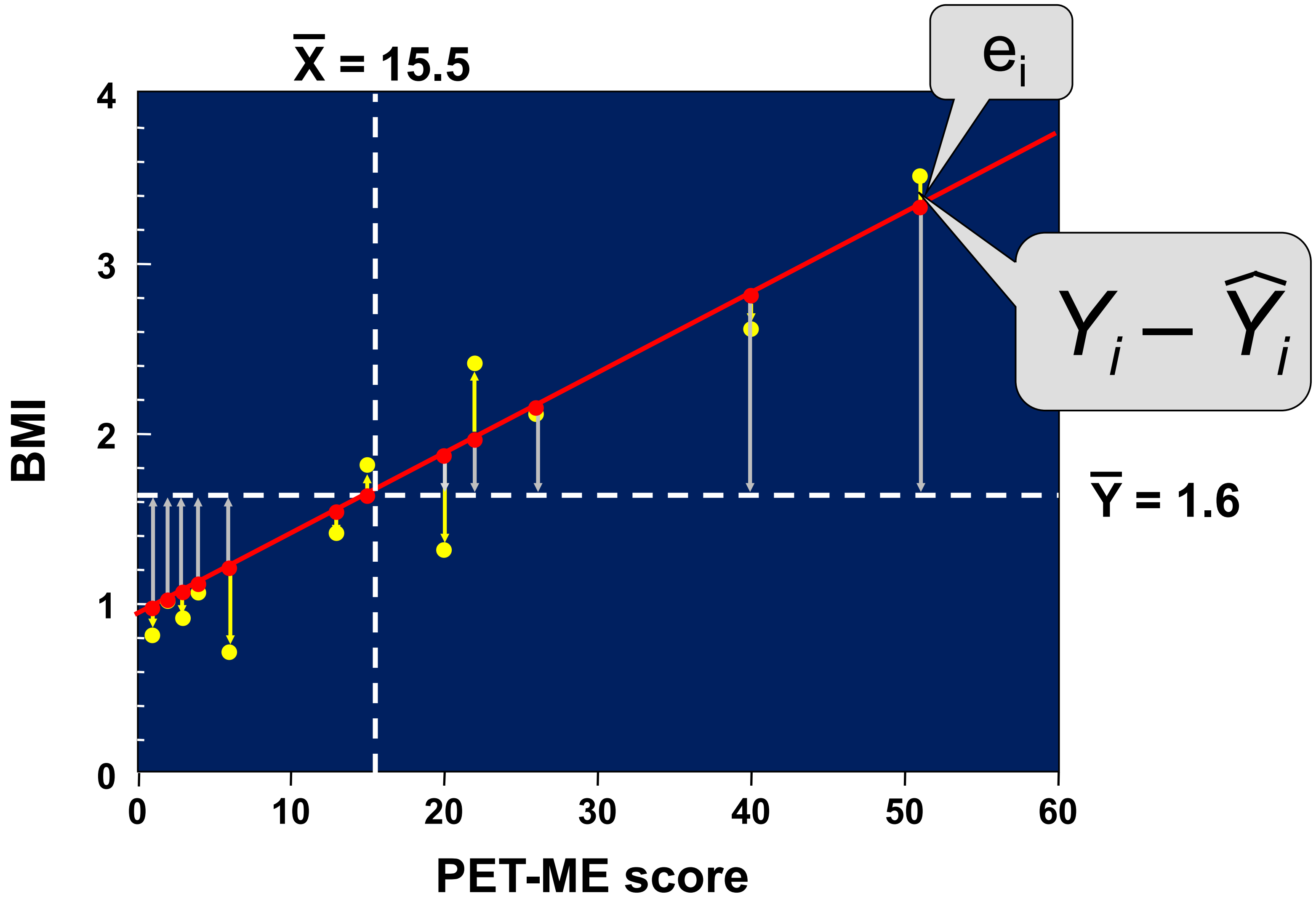
- Reflects the difference between the original data and the fitted line

$$SS_{residual} = \Sigma(Y_i - \hat{Y}_i)^2$$

Noise

- This captures the error between the estimate and the actual data
- Expresses the variance that remains after the regression is over

Graphically depicted as



Plugging in the numbers

$$SS_{reg} = (1.4 - 1.64)^2 + (1.08 - 1.64)^2 + \cdots$$
$$(2.16 - 1.64)^2 + (1.96 - 1.64)^2$$
$$= 7.04$$

$$\widehat{Y}_i - \overline{Y}$$

$$SS_{res} = (2.0 - 1.4)^2 + (1.05 - 1.08)^2 + \cdots$$
$$(2.1 - 2.16)^2 + (2.4 - 1.96)^2$$
$$= 1.48$$

$$Y_i - \widehat{Y}_i$$

Regression Table

Regression table: PET-ME (IV) and BMI (DV)

	df	*SS*	*MS*	*F*	*Pr >F*
Regression	1	7.04			
Residual	13	1.48			
Total	14	8.52			

- <u>Degrees of freedom</u>
  - Regression: df=1
  - Residual: df = n-2
- <u>Test statistic = F</u>
  - $F_{1,\ 13}$

Regression table: PET-ME (IV) and BMI (DV)

	df	*SS*	*MS*	*F*	*Pr >F*
Regression	**1**	**7.04**	**7.04**	**62.03**	**<.0001**
Residual	**13**	**1.48**	**0.11**		
Total	**14**	**8.52**			

- MS = mean square
  - Sum of squares/df
  - Regression MS = 7.04/1 = 7.04
  - Residual MS = 1.48/13 = 0.11
- Test statistic = F
  - $F_{1,13}$ = 7.04/0.11 = 62.03
  - P<.0001

Summary of the Math

- Regression sum of squares: how far the fitted line is from a slope of 0, a horizontal line
  - How much of y is explained by the fitted line
- Residual sum of squares: how far observed y’s are from fitted y’s
- If regression SS is large compared to residual SS, slope is statistically significantly different from 0

Back to r^2

- We got it by squaring *r*, correlation coefficient
  - Proportion of variance in the dependent variable (Y) explained by the independent variable (X)
  - Proportion of variance explained by the regression line
- Another way to calculate it:

$$r^2 = \frac{\overset{\text{Explained Variation}}{SS_{reg}}}{\underbrace{\boxed{SS_{reg} + SS_{res}}}_{\text{Total Variation}}}$$

- Can get *r* by taking square root

Our example . . . .

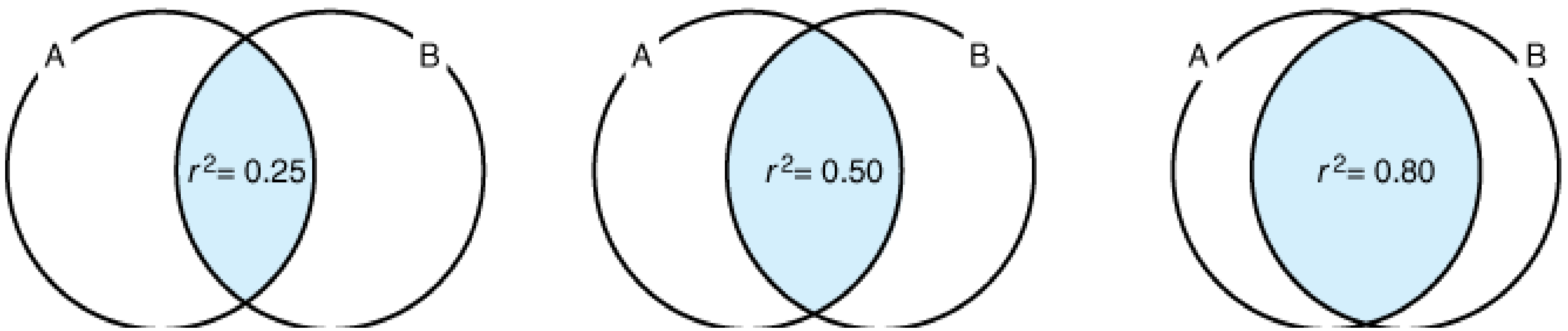
$$r^2 = \frac{SS_{reg}}{SS_{reg} + SS_{res}}$$

$$r^2 = \frac{7.04}{7.04 + 1.48}$$

$$= .827$$

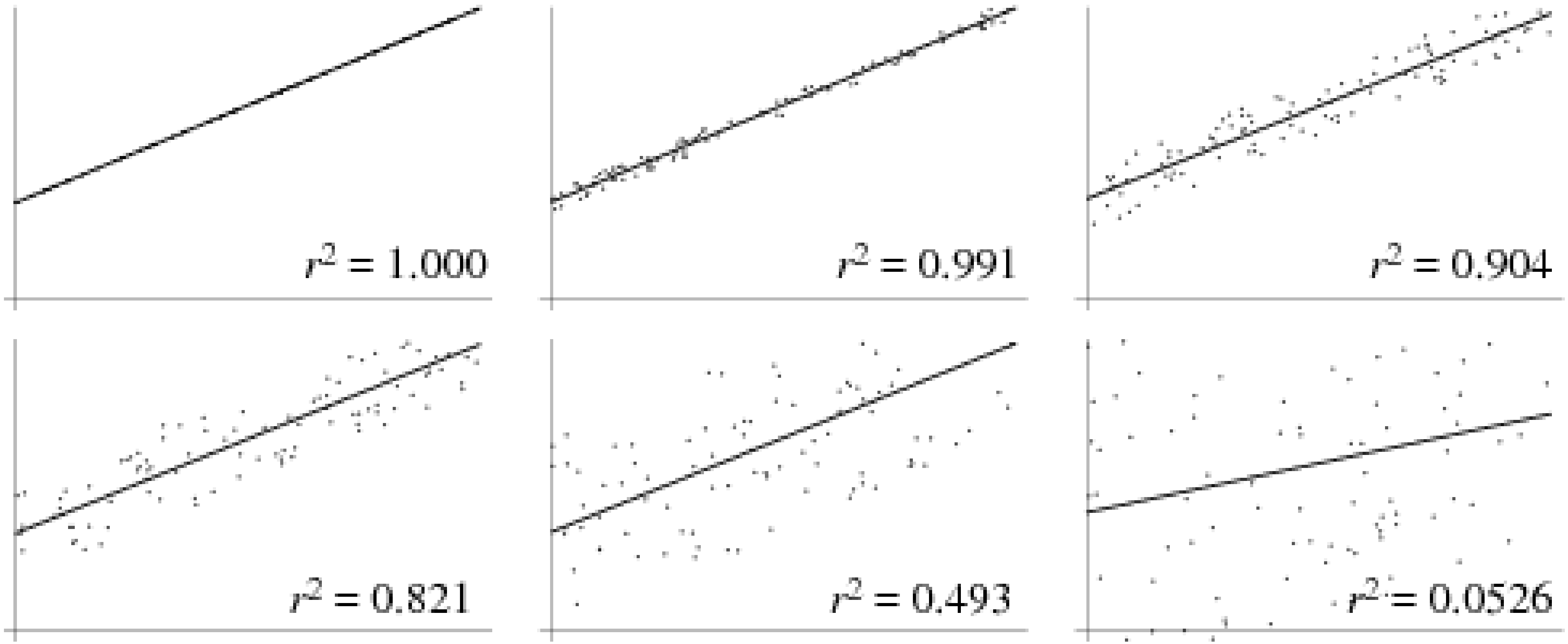
What does r^2 mean?

- Coefficient of determination expresses the amount of variation in *Y* explained by knowing *X*
- In our example, r^2=.827, which means that ~83% of the variation in BMI can be explained by the PET-ME score
- Venn diagrams



Source: Dawson B, Trapp RG: Basic & Clinical Biostatistics, 4th Edition: http://www.accessmedicine.com

Examples of coefficient of determination r^2 (always positive)



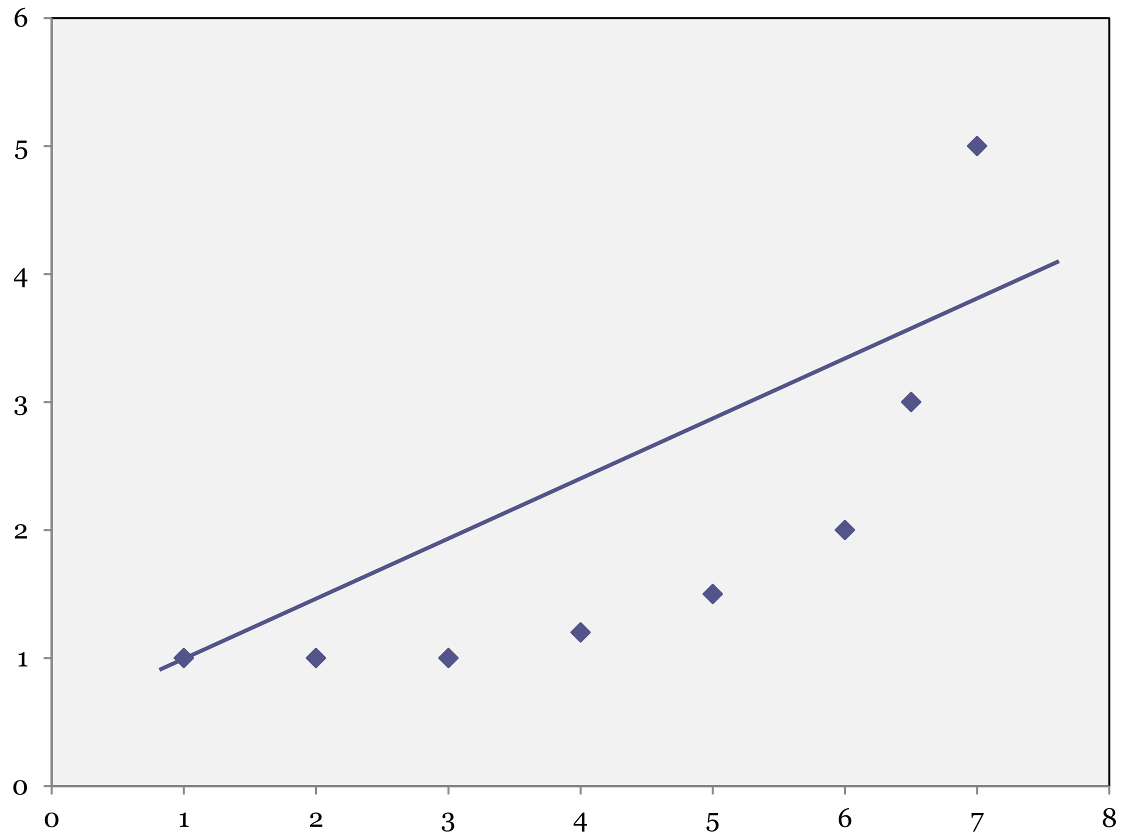
Confidence interval around *β* or *r*

- Slope from linear regression = sample estimate
- Estimate of true population parameter
- Confidence interval around slope or *r* tells us where true value is likely to be
- Example: r=0.32, 95% CI 0.11 to 0.50
  - 95% probability that the population (true) correlation coefficient is between 0.11 and 0.50

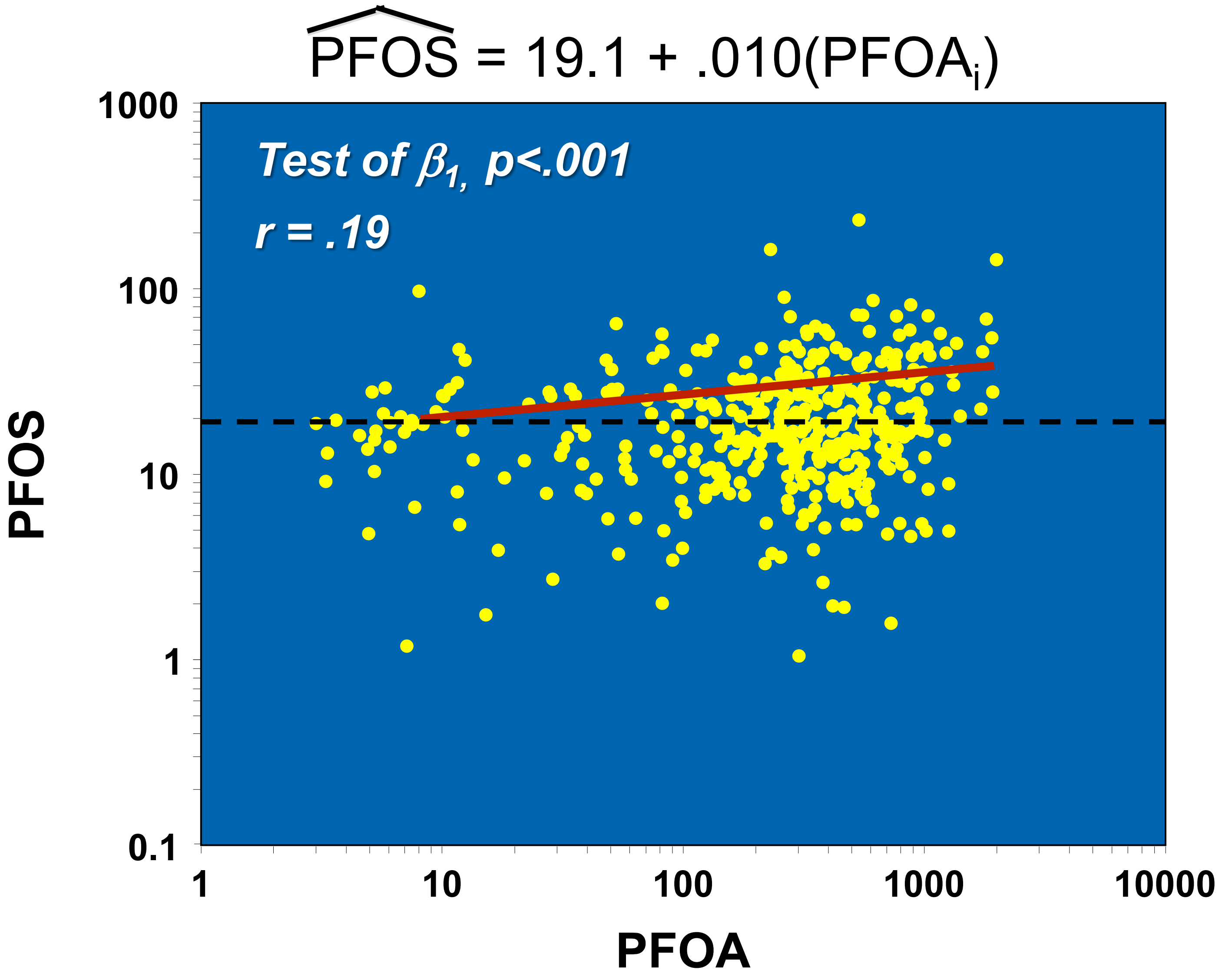
Linear regression

- Form of ‘modeling’
  - Representing reality by an equation
- Some models are good, some aren’t
- Will get equation for regression line whether or not x and y are linearly related or not
  - If not linearly related, line is misleading

Line for non-linear data: misleading



Another Conundrum



Linear Regression: Summary

- Provides 3 pieces of information
  - Whether or not there is a linear relationship (p-value)
  - Direction of relationship (sign of slope β or sign of r)
  - Strength of relationship (r or r^2)
- Testing significance of β is the same as testing significance of r
  - Same p-value for both
- Important to know which is independent and which is dependent variable

What is the Difference Between *r* and Beta?

r	Beta
• *Correlation* coefficient • Strength of relationship, how close values are to e.o. or how close values are to the line • Direction of relationship (+/-) • Range: -1 to 1	• *Regression* coefficient • Magnitude of change in Y for each change in x – Slope • Direction of relationship (+/-) • Range -∞ to +∞

When X and Y are not normally distributed

- Spearman’s rank correlation coefficient
- For skewed data or data with extreme values
- Based on ranks
  - Same formulas but use ranks instead of values
  - Also an alternative formula using differences of ranks
- Interpretation is similar
  - Testing for linear association between ranks of subjects on 2 variables

Still more Spearman's rank correlation

- r_s same interpretation as r
- Both range from -1 to 1
  - Values close to -1 and 1 indicate high correlation
  - Values close to 0 indicate lack of linear association
- r_s can be thought of as measure of the concordance of the ranks of the 2 variables

PET-ME and BMI using Spearman

PET-ME			BMI			
Score	Rank	Corrected Rank	BMI	Rank	*d*	d^2
1	1	1	0.8	2	-1.0	1.0
2	2	2	1.0	4	-2.0	4.0
3	3	3	0.9	3	0.0	0.0
4	4	4	1.1	5	-1.0	1.0
6	5	5	0.7	1	4.0	16.0
9	6	6	1.8	9	-3.0	9.0
10	8	7.5	1.3	6.5	1.0	1.0
10	7	7.5	2.0	11	-3.5	12.3
13	9	9	1.4	8	1.0	1.0
15	10	10	1.8	10	0.0	0.0
20	11	11	1.3	6.5	4.5	20.3
22	12	12	2.4	13	-1.0	1.0
26	13	13	2.1	12	1.0	1.0
40	14	14	2.6	14	0.0	0.0
51	15	15	3.5	15	0.0	0.0

d^2 = 67.5

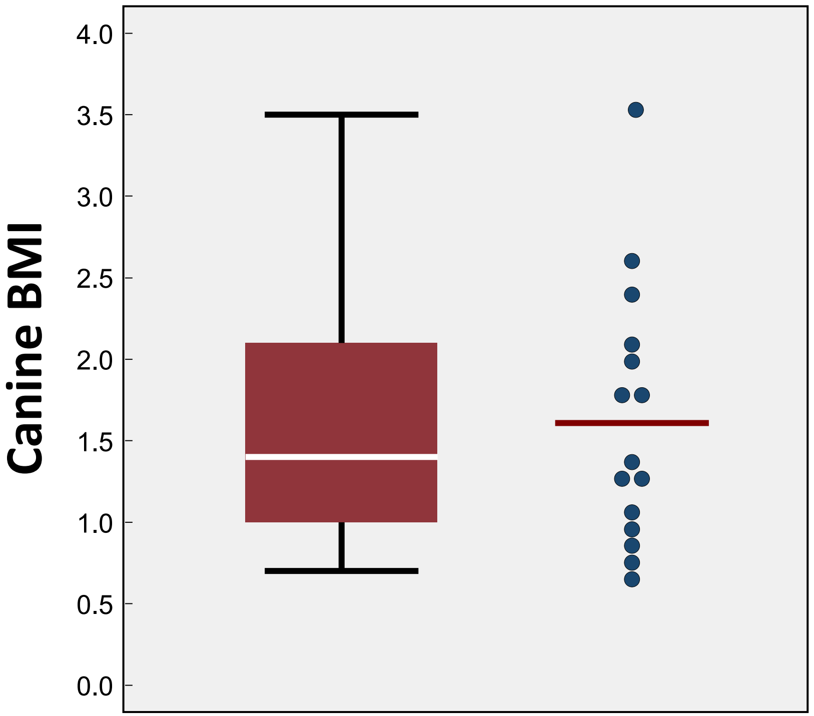
$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

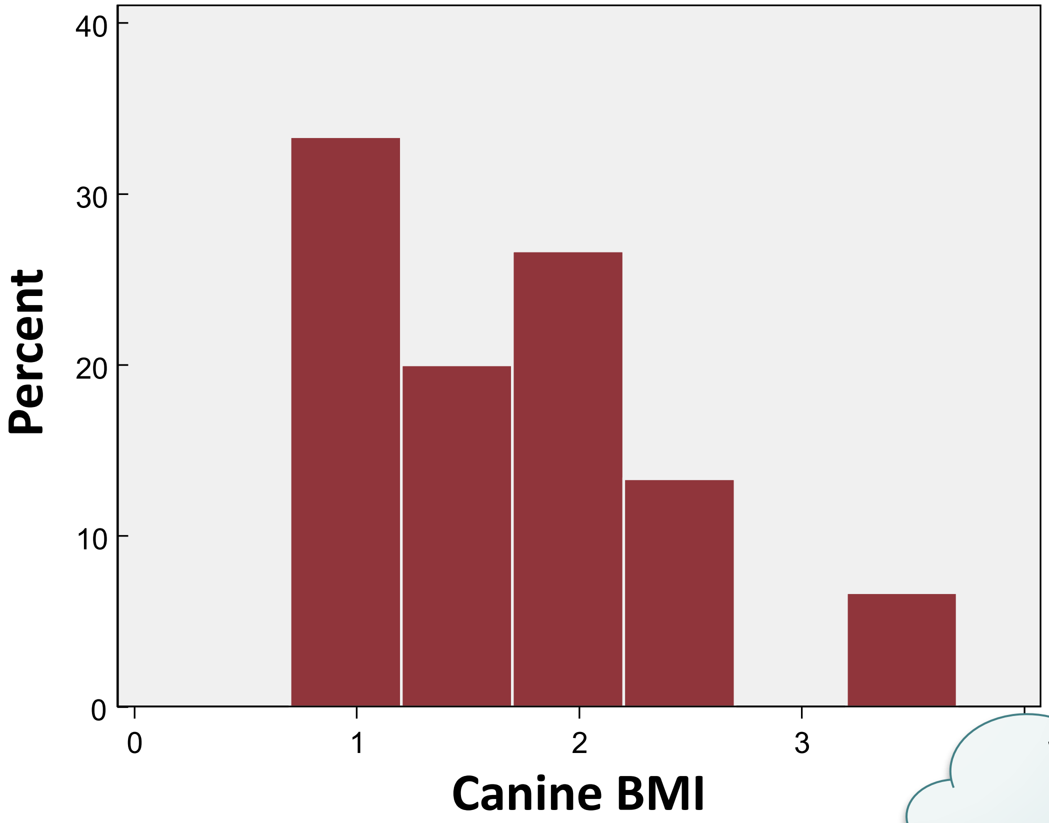
$$r_s = 1 - \frac{6(67.5)}{15(15^2-1)}$$

$$r_s = 0.879$$

- Somewhat smaller than our Pearson's *r* (.909)
- *r* could be inflated due to non-normality of data
- PET-ME not normally distributed ☹

Canine BMI – Normal?





Why we need to check first!

Mean 1.65, SD .78
Median 1.4, IQR 1-2.1

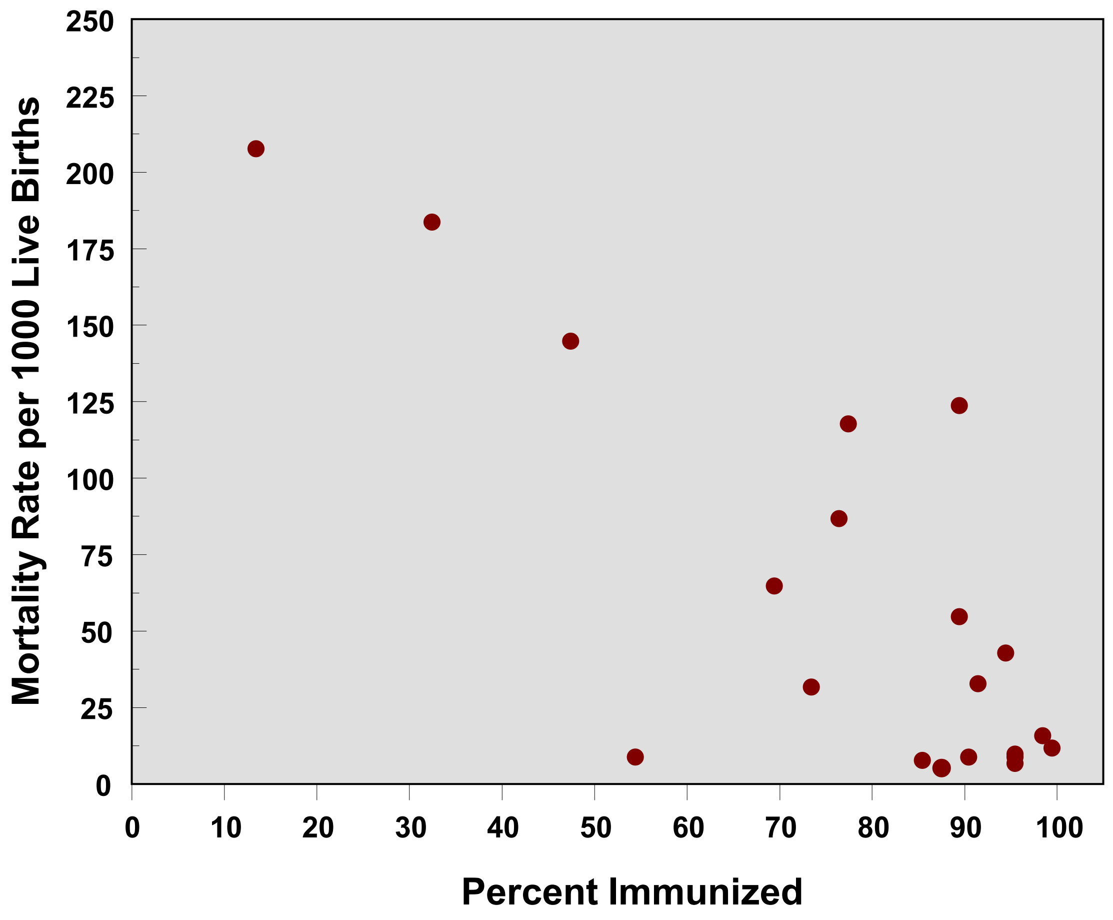
BMI is not normally distributed!

Another Example

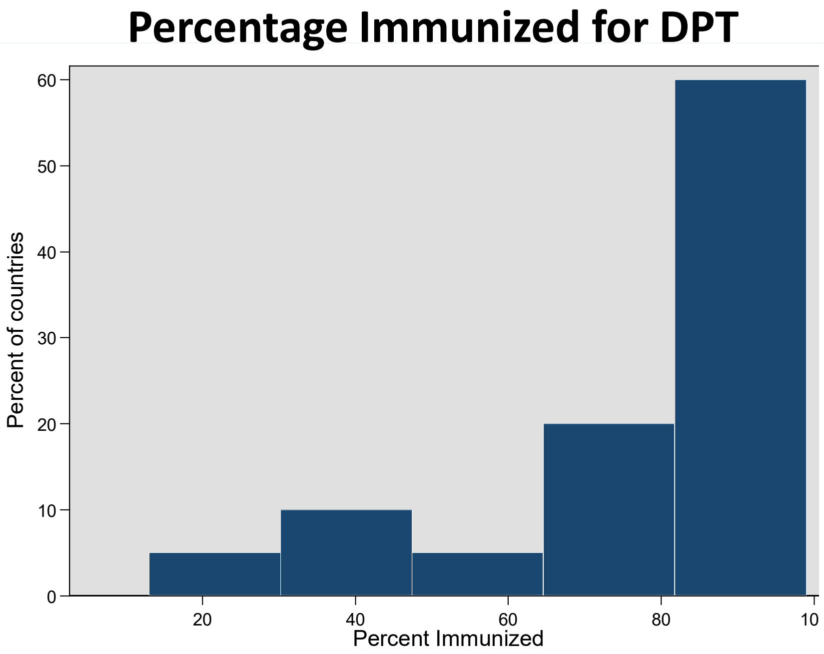
- Is the percentage of children immunized against DPT associated with mortality rate in children <5 years?
- % children immunized for DPT and mortality rates for children <5 years collected from 20 countries in 1992
- What type of analysis would we do if the mortality rates were normally distributed?
  - Linear regression, calculation of Pearson's correlation coefficient
- If not normal?
  - Spearman's rank correlation coefficient

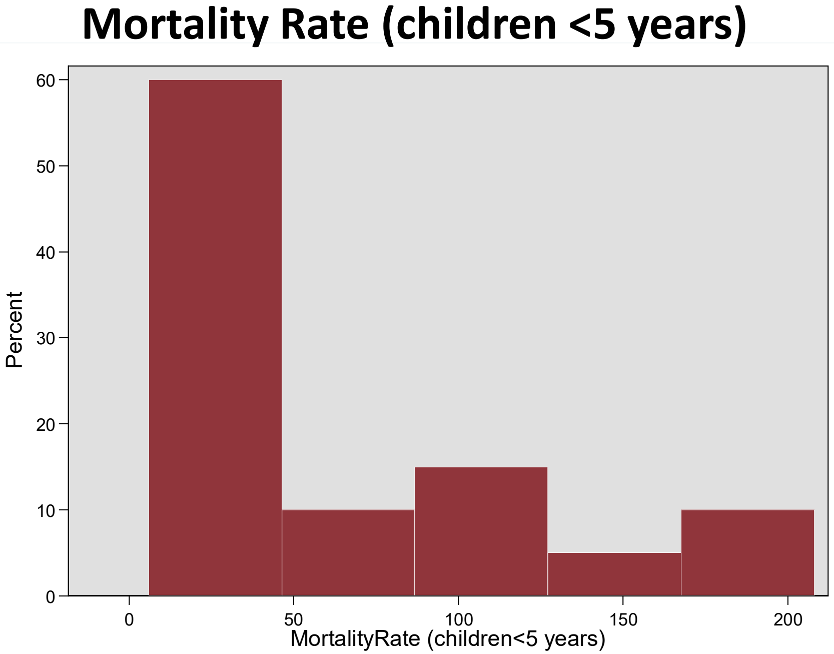
Percentage of Children Immunized against DPT & Under 5 Mortality Rates for 20 Countries (1992)

Country	Percentage immunized	Mortality rate
Ethiopia	13	208
Cambodia	32	184
Senegal	47	145
Greece	54	9
Brazil	69	65
Russia	73	32
Turkey	76	87
Bolivia	77	118
Canada	85	8
Japan	87	6
India	89	124
Egypt	89	55
UK	90	9
Mexico	91	33
China	94	43
France	95	9
Finland	95	7
Italy	95	10
Poland	98	16
Czech Republic	99	12



Checking for Normality





- Mean: 77.4
- Median: 88.0
- STD: 23.7

- Mean: 59.0
- Median: 32.5
- STD: 63.9

Pearson versus Spearman

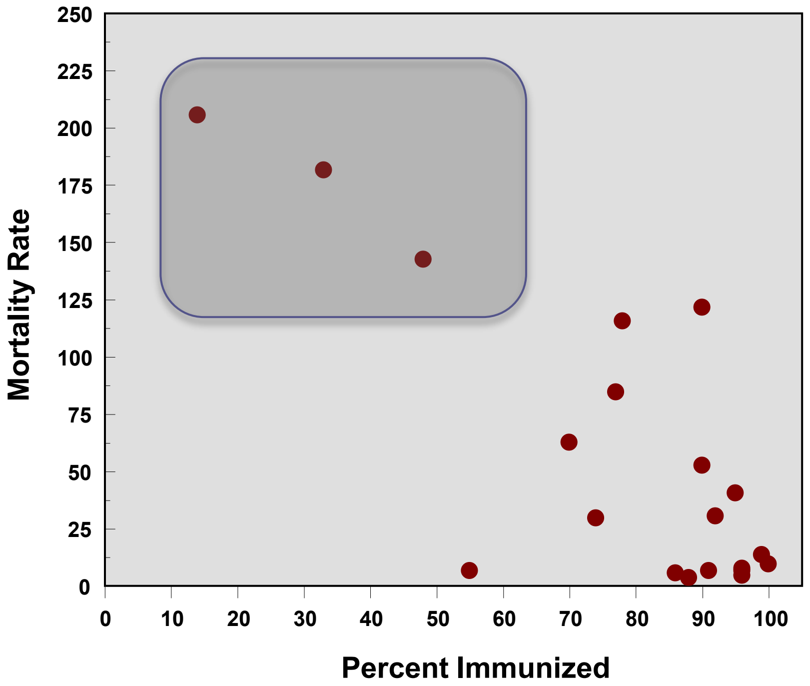
- Null & alternative hypotheses are almost identical
- Null hypothesis
  - H_0: $\rho = 0$
  - Pearson: There is no correlation between % immunized & mortality rate
  - Spearman: There is no correlation between the ranks of % immunized & mortality rate
- Alternative hypothesis
  - H_A: $\rho \neq 0$
  - Pearson: There is a correlation between % immunized & mortality rate
  - Spearman: There is a correlation between the ranks of % immunized & mortality rate

Spearman r_s

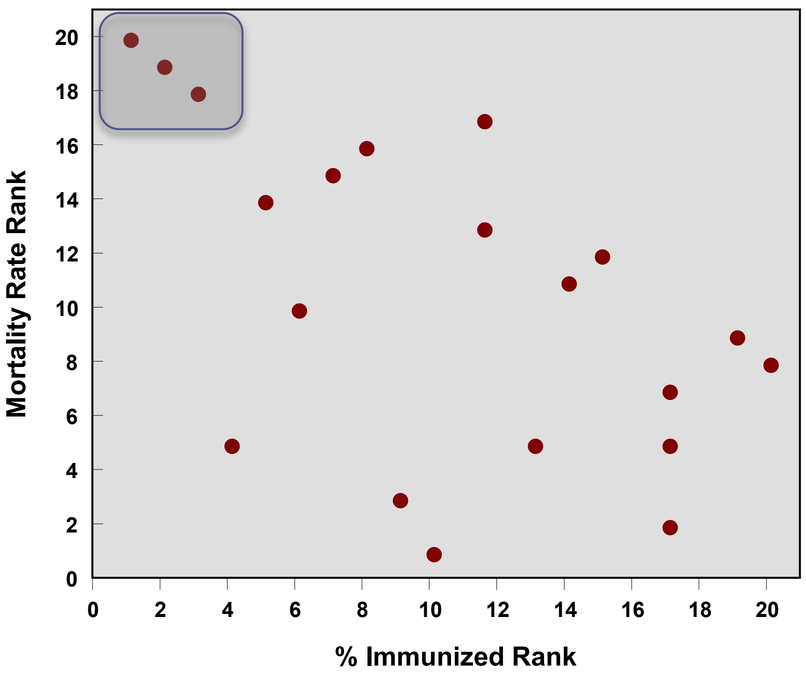
- Much less sensitive to extreme values than Pearson
  - Similar to other non-parametric tests
- Can be used when 1 or both variables are not normally distributed
- Some information lost by use of ranks instead of values

Pearson versus Spearman

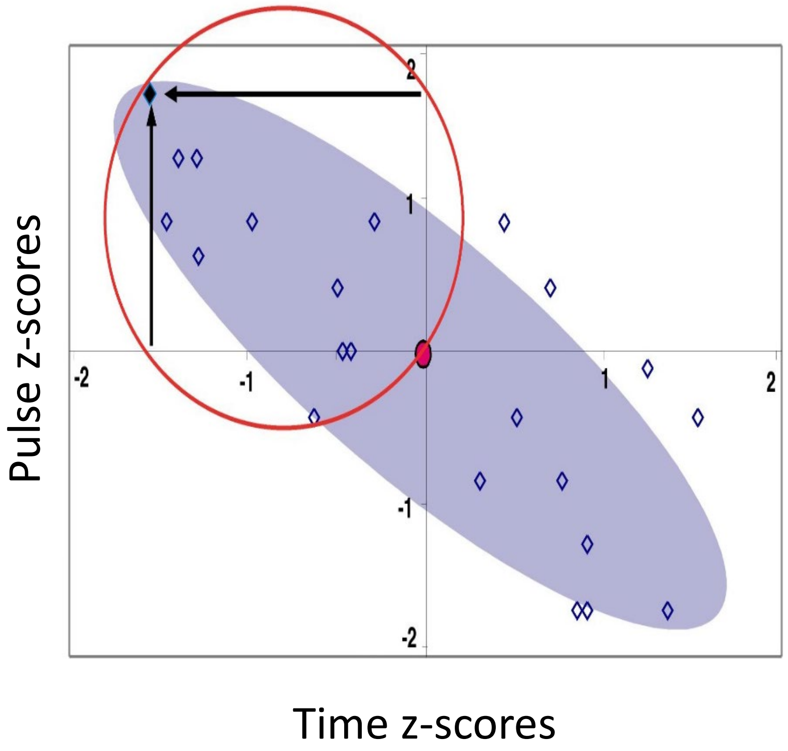
Pearson *r* = -0.79, p < 0.001

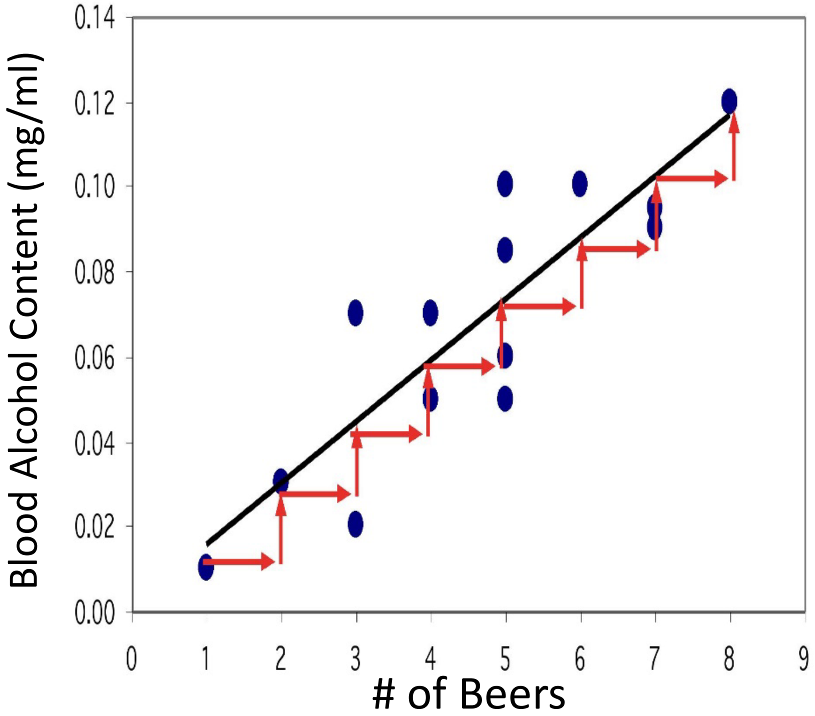


Spearman r = -0.54, p=0.01



Correlation versus Regression





The **correlation** is a measure of spread (scatter) in both the *x* and *y* directions in the linear relationship

In **regression** we examine the variation in the response variable (*y*) given change in the explanatory variable (*x*)

Summary

- Correlation assesses strength of relationship
- Related to linear regression, but different
- Regression shows overall change in Y given a change in X
- *Y=mx+b*
- Simple regression has 1 IV (x var)
  - Beta coefficient = slope
    - Explains rise over run
    - Δ in Y for every 1 unit Δ in x
  - R^2=variability of the DV explained by IV (how good our model is at predicting Y)