Correlations,
Part & Partial Correlations, &
Multiple Linear Regressions

Correlations

Variance, Covariance, & Correlations

- Variance & Covariance
  - Importance in statistical analyses
- Covariance & Correlation
  - Relationship between them
  - Why use one or the other?
  - Both are descriptive statistics
    - Even though tests can be run on them

Assumptions in Correlations

- Assumptions made in computing correlations
  - Ordinal, interval, or ratio
  - Linear relationship*
- Assumptions make in **testing** correlations
  - Monotonic (or normal for Pearson's *r*)
  - Homoscedasic
  - No big outliers

Correlations & Error

- Correlations & Error
  - Correlations separate dispersion into variance & covariance
  - But make no assumptions about error
    - Viz., where error resides
    - Instead, both variables are assumed to be equally affected by error

Correlations & Error (cont.)

- Correlations & Error

$$r = \frac{\mathrm{Cov}(X, Y)}{\mathrm{SD}(X)\,\mathrm{SD}(Y)}$$

  - The (unshared) variances of both variables comprise the denominator
    - (This will be different for linear regression models)

Partial & Semi-Partial Correlations

- Correlations describe linear relationships between two variables
  - **Without** consideration of the influence of other variables
  - Partial and semipartial correlations account for associations with other variables

Partial & Semi-Partial Correlations (cont.)

- Partial Correlations
  - Partial correlations remove the effect of another variable from *both* of the correlated pair
- Semipartial Correlations
  - Also called "part correlations"
  - Removes the association of an other variables from *one* of the correlated pair
- N.b. that either can remove the effect of several other variables from one of the pair
  - (Or create even more complex arrangements, like canonical correlations)

Partial Correlations

- Conceptually, we:
  1. Compute correlation between **X & Y**
  2. Subtract from that the ratio of:
     - How much total variance **is**
     - and **is not** explained
     - by the correlations between between **X & Z** and between **Y & Z**

Partial Correlations (cont.)

- To wit:

$$r_{xy,z} = \frac{r_{xy} - (r_{xz} \times r_{yz})}{\sqrt{(1 - r_{xz}^2) \times (1 - r_{yz}^2)}}$$

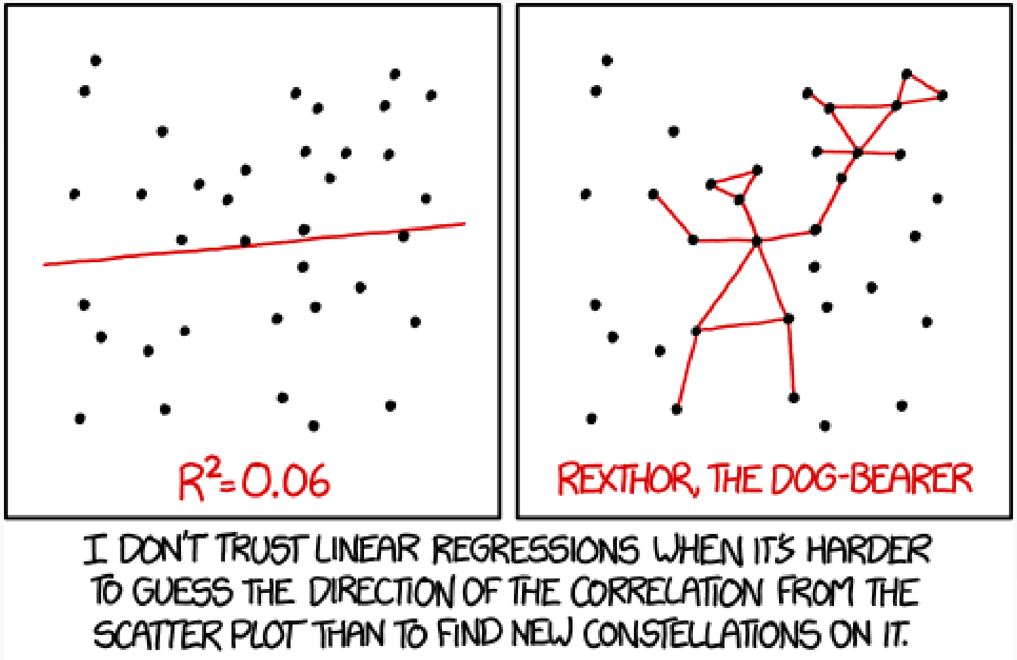
  - So, compute the correlation between X & Y
  - Remove correlations with Z
  - Divide by variance u*nexplained* by the correlations between X & Z and between Y & Z

Semipartial Correlations

- Computationally very similar to a partial correlation
  - Differers only in the denominator:

$$r_{y(x\cdot z)} = \frac{r_{xy} - (r_{xz} \times r_{yz})}{\sqrt{1 - r_{xz}^2}}$$

  - Where we only divide it by the variance unexplained in X & Z



Linear Regressions

Linear Models

- Very commonly used in inferential statistics
- Simplest form is $Y = bX$, where:
  - Y = Output / response / **criterion** / DV
  - X = Input / **predictor** / IV
  - b = Slope of X
    - If data are standardized to a normal distribution, then convention has us use β instead of b

Linear Models (cont.)

- Very commonly used in inferential statistics
- Simplest form is $Y = bX$
- However, we typically add at least two other terms: $Y = b_0 + b_1X + e$
  - Y = Response / criterion / DV
  - X = Predictor / IV
  - b_0 = y-Axis intercept
  - b_1 = Slope of X
  - e = Error

The typical null hypothesis (H_0) of "no effect" is expressed here as: $b_1 = 0$

Linear Models vs. Correlations

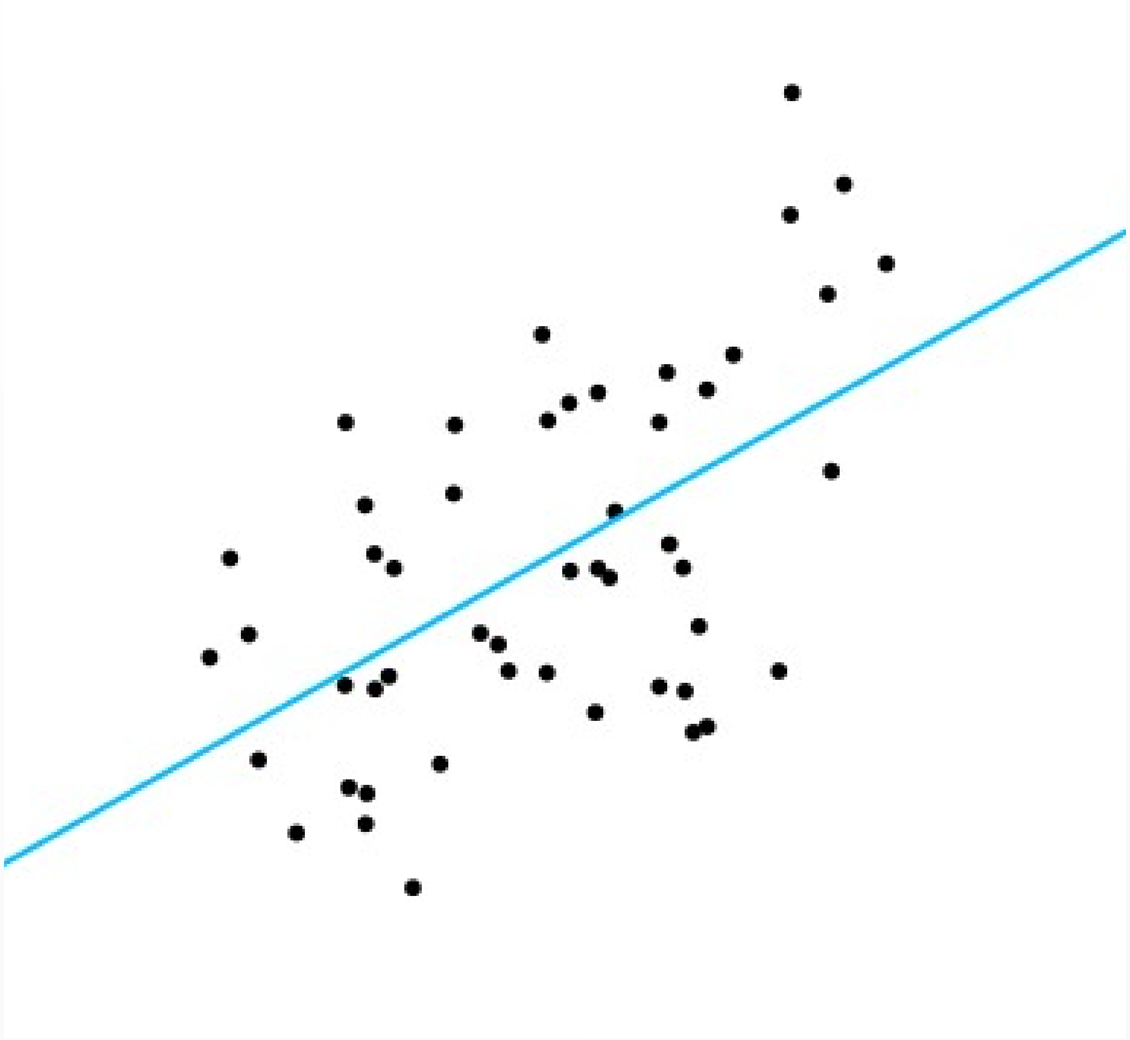
- Recall for a correlation:

$$r = \frac{\text{Cov}(X, Y)}{\text{SD}(X)\,\text{SD}(Y)}$$

  - The (unshared) variances of both variables comprise the denominator
  - This is equivalent to simply drawing a line of "best fit" through the data
    - Without worrying about orientation

Linear Models vs. Correlations (cont.)

$$r = \frac{\mathrm{Cov}(X, Y)}{\mathrm{SD}(X)\,\mathrm{SD}(Y)}$$



Linear Models vs. Correlations (cont.)

- For a linear regression, we instead minimize variance in only one variable
  - Typically the criterion (outcome)
  - This assumes that all unexplained variance (error) resides in the criterion
- So, in $Y = b_0 + b_1X + e$:

$$b_1 = \frac{\mathrm{Cov}(X, Y)}{(\mathrm{SD}(Y))^2}$$

Linear Models vs. Correlations (cont.)

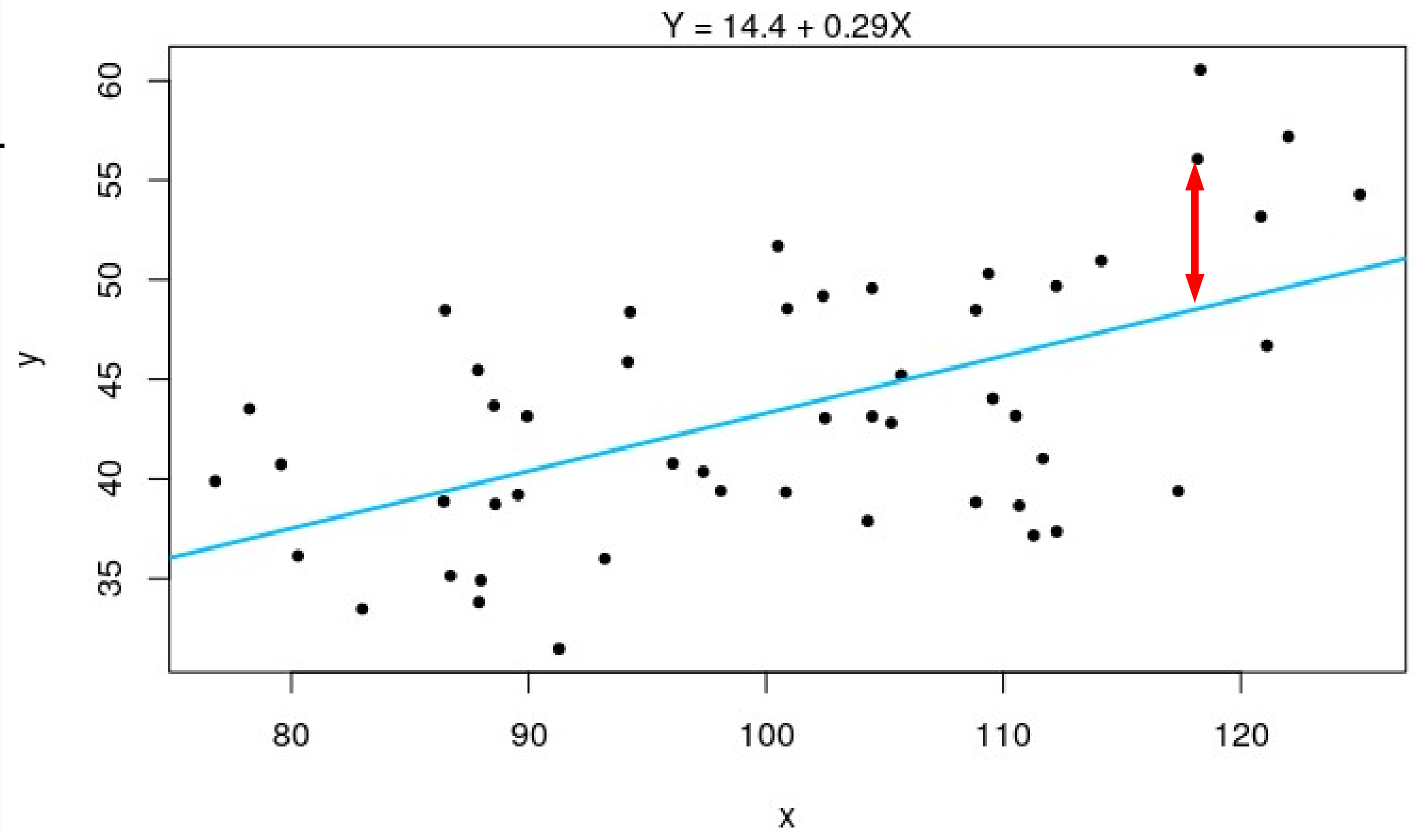
- This also means that b_1 is expressed in units of X per Y:

$$b_1 = \frac{\text{Cov}(X, \cancel{Y})}{\text{SD}(Y)\,\text{SD}(\cancel{Y})}$$

- If we standardize both variables, then the units are the same
  - (In fact, they are removed)
- And b_1 becomes equivalent to the correlation
  - (And is conventionally expressed as β_1 instead of b_1)

Linear Models vs. Correlations (cont.)

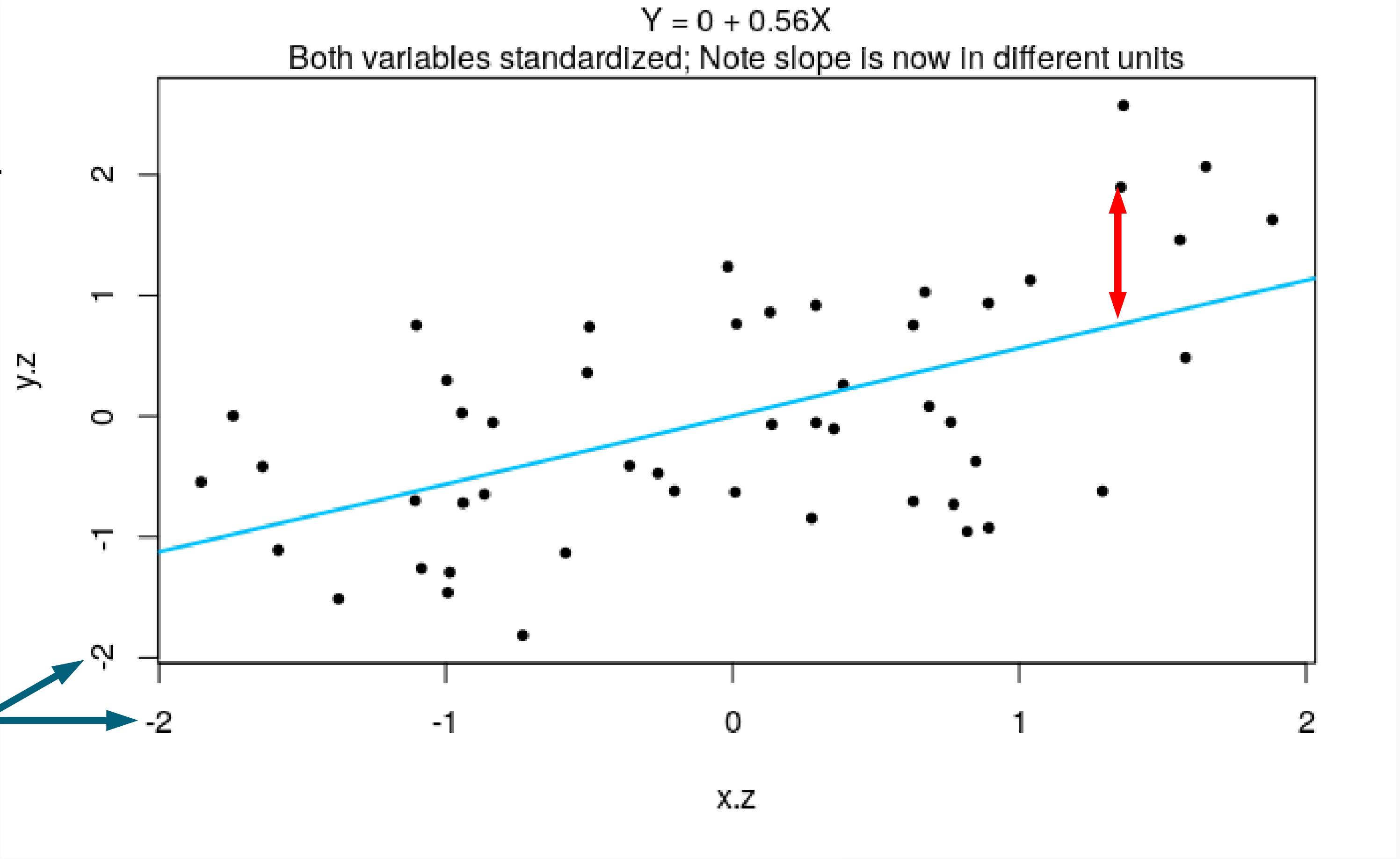
$$b_1 = \frac{\mathrm{Cov}(X, Y)}{(\mathrm{SD}(Y))^2}$$



Linear Models vs. Correlations (cont.)

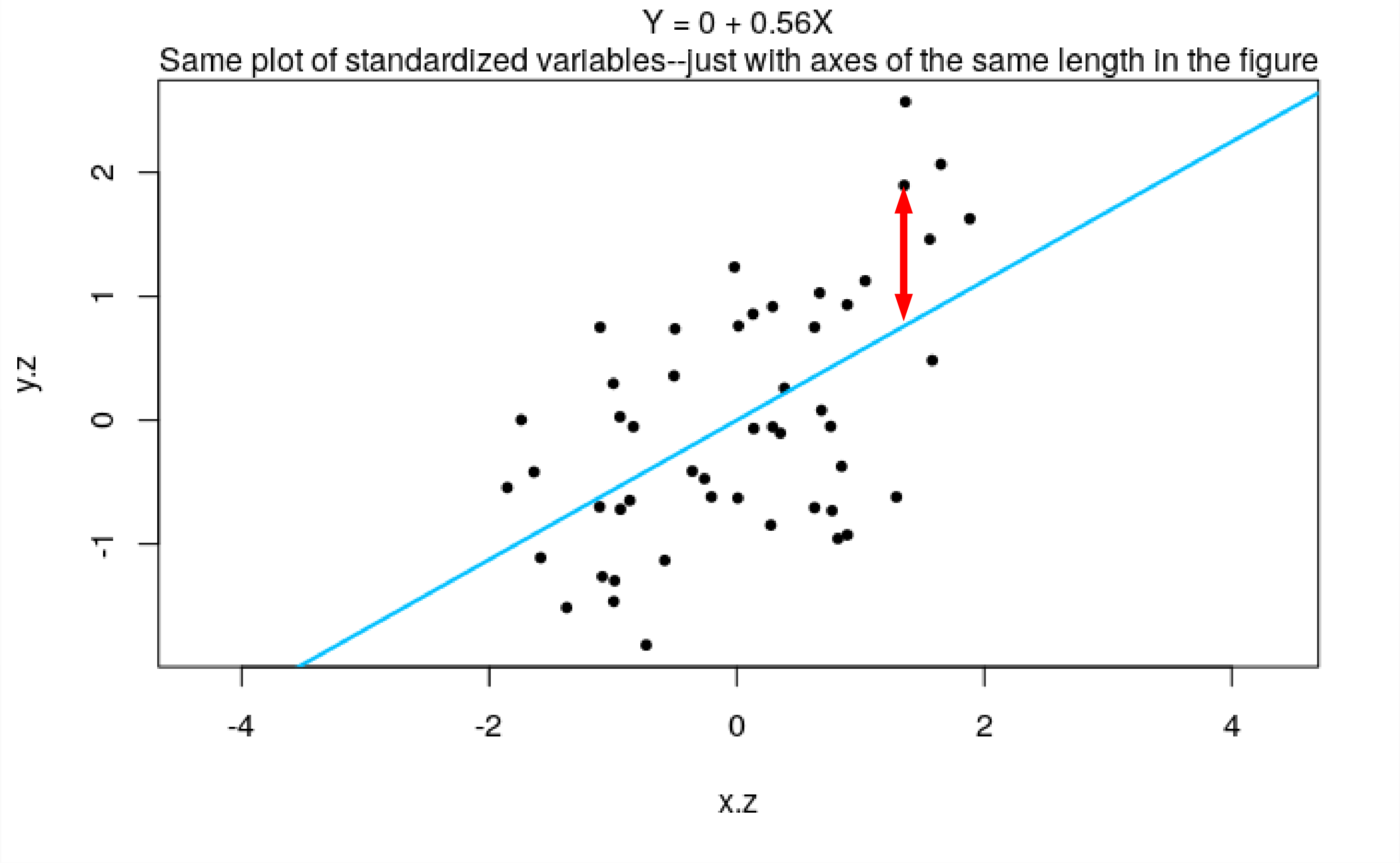
$$b_1 = \frac{\text{Cov}(X, Y)}{(\text{SD}(Y))^2}$$

Note different units in axes



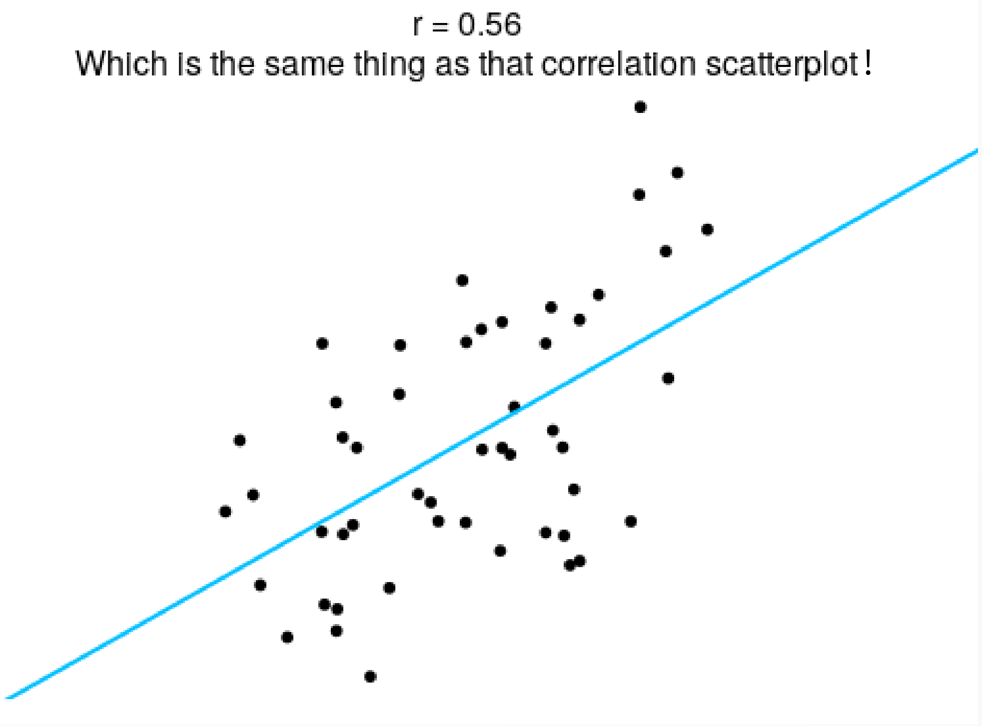
Linear Models vs. Correlations (cont.)

$$b_1 = \frac{\mathrm{Cov}(X, Y)}{(\mathrm{SD}(Y))^2}$$



Linear Models vs. Correlations (final)

$$r = \frac{\mathrm{Cov}(X, Y)}{\mathrm{SD}(X)\,\mathrm{SD}(Y)}$$



Linear Models (cont.)

- $Y = b_0 + b_1 X_1 + e$
- Note:
  - Error is separated out
    - And placed on the side with the predictor
- Implications:
  - The value of *X* per se is without error
    - Because error is separated out (as *e*)
  - The intercept, slope, & error can be estimated separately
    - And their covariances with *Y* are thus separated

Linear Models (cont.)

- Adding more specificity to the equation:

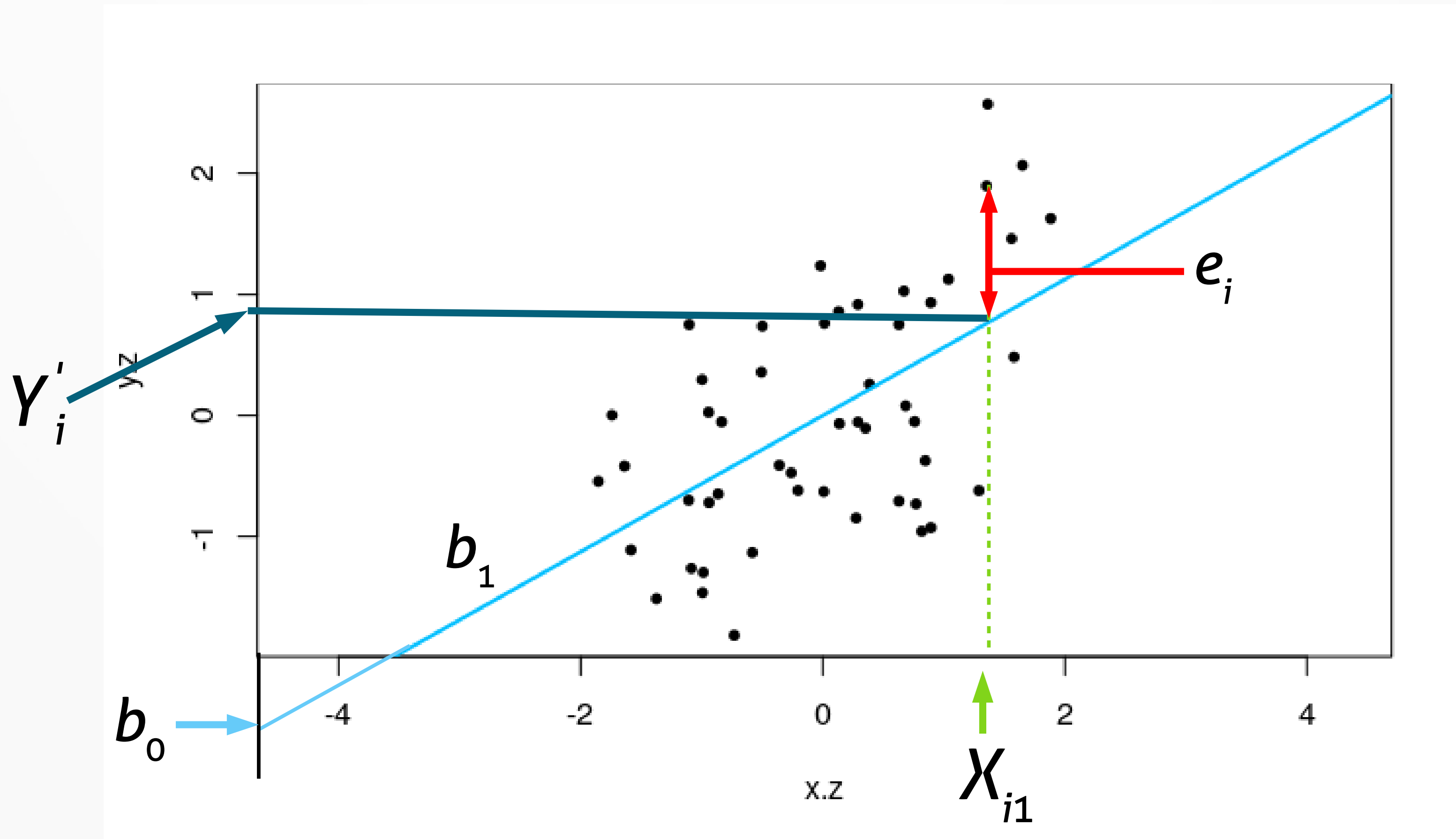
  $Y_i' = b_0 + b_1 X_{i1} + e_i$

  - Y_i' = *Predicted* value of Y for instance i
  - b_1 = Slope for variable X_1
  - e_i = Error for instance i

Linear Models (cont.)

- Adding more specificity to the equation:

  $Y'_i = b_0 + b_1 X_{i1} + e_i$



Linear Models (cont.)

- Adding another variable to the equation:

  $Y_i' = b_0 + b_1 X_{i1} + b_2 X_{i2} + e_i$

  - X_{i2} = *i*'s value on another variable added to the model
  - b_2 = Slope for variable X_2
- Since there are multiple predictors (*X*s) in this model,
  - This is called a **multiple** linear model
  - Or **multiple linear regression**

Linear Models vs. ANOVAs

- ANOVA (and ANCOVA, MANOVA, etc.)
  - Is a type of linear regression
  - Results focus on significance of variables
    - When all are present in the model together
- Linear Regression
  - A general, flexible framework
  - Results (usually) focus on significance of whole model
    - And changes in the whole model when variables are added or removed

Questions Best Addressed by Linear Models vs. ANOVAs

- ANOVA (and ANCOVA, MANOVA, etc.)
  - Which variable is significant?
  - Is there an interaction between variables?
- Linear Regression
  - What is the best combination of variables?
  - Does a given variable significantly contribute more to what we already know?
  - Can also test interactions
    - But also for *groups* of, e.g., theoretically-relevant variables

Linear Models (cont.)

- We can continue to add more variables to the model, e.g., X_3 and X_4:

$$Y_i' = b_0 + b_1 X_{i1} + b_2 X_{i2} + b_3 X_{i3} + b_4 X_{i4} + e_i$$

- When there are a lot of variables in the model, say k of them, we usually abbreviate the equation:

$$Y_i' = b_0 + b_1 X_{i1} + b_2 X_{i2} + \ldots + b_k X_{ik} + e_i$$

Linear Models (cont.)

- Adding more complexity to the equation (cont.):

  $$Y_i' = b_0 + b_1 X_{i1} + b_2 X_{i2} + \ldots + b_k X_{ik} + e_i$$

  - Note the effects of predictors are separated
    - Like semipartial correlations
  - Of course, we could test interactions by adding additional terms
    - E.g., … + $b_1 X_{i1} + b_2 X_{i2} + \boldsymbol{b_3 (X_{i1} X_{i2})}$ +…
  - Or test non-linear effects, also by adding terms
    - E.g., … + $b_1 X_{i1} + \boldsymbol{b_2 X_{i1}^2}$…

Linear Models (cont.)

- Adding more complexity to the equation (cont.):

  $$Y_i' = b_0 + b_1 X_{i1} + b_2 X_{i2} + \ldots + b_k X_{ik} + e_i$$

  - Just as we separated out the effects of the predictors,
    - We can separate out sources of error (not shown)
      - E.g., per predictor / term in the model
    - We can also combine error terms
      - E.g., when we "nest" one variable into another
        - We will cover this when we discuss multilevel models

Linear Models: Signal-to-Noise

- Signal-to-noise in the equation

  $$Y_i' = b_0 + b_1 X_{i1} + b_2 X_{i2} + \ldots + b_k X_{ik} + e_i$$

  - Y_i' is the estimated value of Y_i
  - The variance in Y_i' per se can be divided into:
    - Changes due to the predictors
    - Changes due to "other things" (and relegated to error / noise term(s))
    - (N.b., the intercept, b_0, is a constant and not included in this partitioning)

Linear Models: Signal-to-Noise (cont.)

- Signal-to-noise in the equation (cont.)

$Y_i' = b_0 + b_1 X_{i1} + b_2 X_{i2} + \ldots + b_k X_{ik} + e_i$

  - The sum of squares representation of this partition into predictors & error looks like:

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i} (\hat{Y}_i - \bar{Y})^2 + \sum_{i} (Y_i - \hat{Y}_i)^2$$

- Where $\hat{Y}_i$ is the least-squares estimate of Y_i
  - I.e., that predicted by the slope of the model

Linear Models: Signal-to-Noise (cont.)

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_i (\hat{Y}_i - \bar{Y})^2 + \sum_i (Y_i - \hat{Y}_i)^2$$

- I.e.:
- The squared sum of the deviations of each instance from the mean equals:
  - The squared sum of each predicted value from the mean
    - (That predicted from all of the predictors)
  - Plus the squared sums of all other variation in *Y* from the predicted value

Linear Models: Signal-to-Noise (cont.)

- We could rewrite

$$\sum_{i=1}^{n}(Y_i-\bar{Y})^2=\sum_i(\hat{Y}_i-\bar{Y})^2+\sum_i(Y_i-\hat{Y}_i)^2$$

- As:

Total SS = SS from Regression + SS from Error

  - Or, further condensed as:

$SS_{Total} = SS_{Reg.} + SS_{Error}$

Linear Models: Signal-to-Noise (cont.)

- Using $SS_{Total} = SS_{Reg.} + SS_{Error}$
  - We can compute the ratio of predicted to actual:

$$\text{Ratio of Predicted-to-Actual Variance} = \frac{SS_{Reg.}}{SS_{Total}}$$

  - Or, equivalently:

$$\text{Ratio of Predicted-to-Actual Variance} = 1 - \frac{SS_{Reg.}}{SS_{Error}}$$

Linear Models: Signal-to-Noise (final)

- We typically represent this ratio of predicted-to-actual
  - (or total variance minus proportion of error)…
- ***as R^2***

$$R^2 = \frac{SS_{\text{Reg.}}}{SS_{\text{Total}}} = 1 - \frac{SS_{\text{Error}}}{SS_{\text{Total}}}$$

  - Yep, that's what R^2 means ☻

Linear Models (redux)

- More about the equation:

  $Y_i' = b_0 + b_1 X_{i1} + b_2 X_{i2} + \ldots + b_k X_{ik} + e_i$:
  - *Y* is assumed to follow a certain distribution
    - This determines how error is modeled
      - E.g., is error assumed to be normally distributed
  - The *X*s can be nominal, ordinal, interval, or ratio
    - This affects how those variables are modeled
    - As well as the error related to them
  - We could transform the terms on the right
    - E.g., raise them to a power or take their log

Linear Models (cont.)

- More about the equation:

$$Y_i^{'} = b_0 + b_1 X_{i1} + b_2 X_{i2} + \ldots + b_k X_{ik} + e_i :$$

  - E.g., for an ANOVA:
    - *Y* is assumed to be normally distributed
    - The *X*s are nominal
    - And the terms are not transformed
      - Called an "identity" because they are multiplied by 1

Linear Models (final)

- Less noticeable in
  $Y_i' = b_0 + b_1 X_{i1} + b_2 X_{i2} + \ldots + b_k X_{ik} + e_i$:
  - E.g., for an ANOVA:
    - Y is assumed to be normally distributed
    - The Xs are nominal
    - And the terms are not transformed
      - These terms **are** transformed in other models
      - This transformation is called a **Link Function**
      - Since it "links" the terms on the right to the predicted value of Y on the left

Types of Link Functions

Model	Distribution of *Y*	Link	Types of *X*s
ANOVA	Normal	Identity	Nominal
ANCOVA	Normal	Identity	Nominal & Interval / Ratio
Linear Regression	Normal	Identity	Interval / Ratio
Logistic Regression	Binomial	Logistic	Nominal & Interval / Ratio

Types of Link Functions (rev.)

Model	Random Component	Link	Systematic Component
ANOVA	Normal	Identity	Categorical
ANCOVA	Normal	Identity	Categorical & Continuous
Linear Regression	Normal	Identity	Continuous
Logistic Regression	Binomial	Logit	Categorical &/or Continuous

Generalized Linear Models

- That family of models is referred to as **generalized linear models**
  - ANOVAs, *t*-tests, and all other linear regressions are types of generalized linear models
  - Generalized linear models use maximum likelihood estimation (MLE) to compute terms
    - The ordinary least squares of ANOVAs, etc. is itself a specific type of MLE
      - (If assumptions are met)
    - So, yeah, it's O.K. to still use OLS & ANOVAs

Generalized Linear Models (cont.)

- N.b., confusingly, in addition to **generalized** linear models,
  - There are **general** linear models
  - "**General** linear model" simply refers to models with:
    - Normal Random Components &
    - Identity Link Functions
  - Like ANOVAs & "mulitple linear regressions"

Generalized Linear Models (final)

- Assumptions of generalized linear models:
  - Relationship between response and predictors must be expressible as a linear function
    - Can even model heteroscedasticity
  - Cases must be iid (independent & identically distributed)
  - Predictors should not be too inter-correlated (lack of multicollinearity)
  - The random & link functions should approximate the real functions

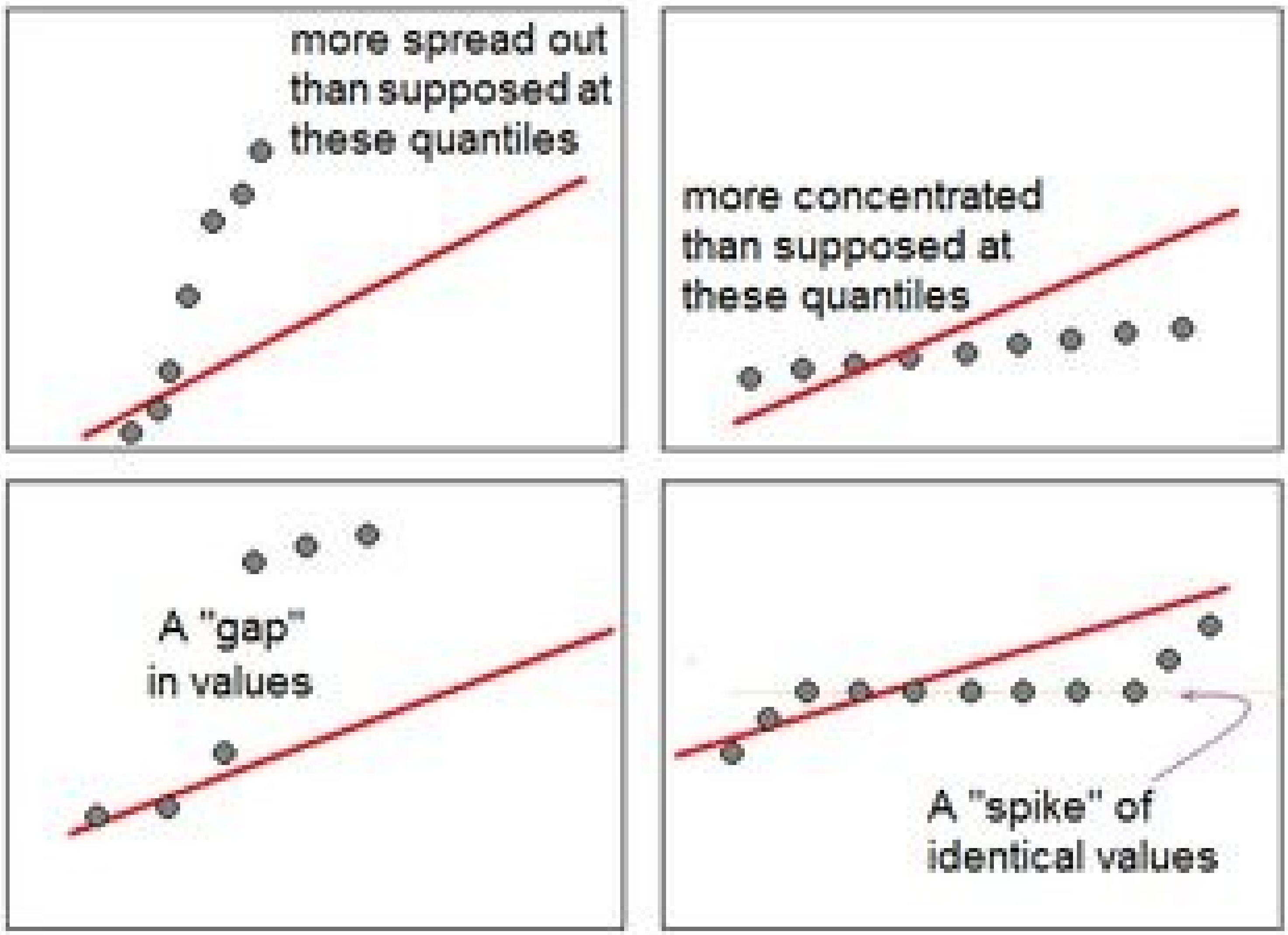
Evaluating Distributions

- The actual distribution of error & scores does not need to strictly follow the assumed distribution
  - (E.g., the actual data don't need to be completely normal)
  - But large deviations should be addressed

Evaluating Distributions: Q-Q Plots

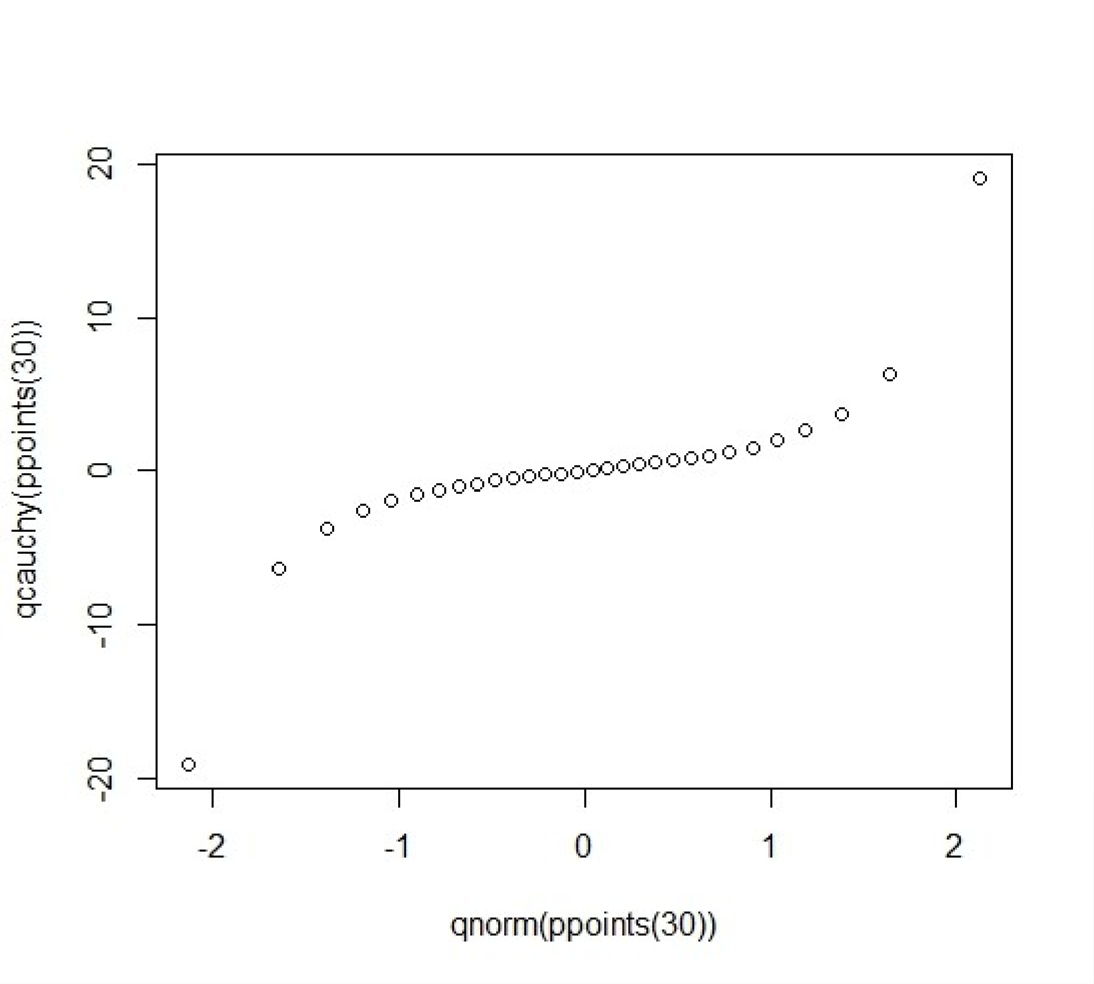
- We can use Q-Q plots to evaluate deviations from normality
  - Q-Q plots have the values of the actual data on the *y*-axis
  - And the values that each data point *would* have if they followed the given distribution on the *x*-axis
  - If all data fall on a straight line on the plot, then the data are exactly the values expected to be given that distribution

Evaluating Distributions:
Q-Q Plots (cont.)



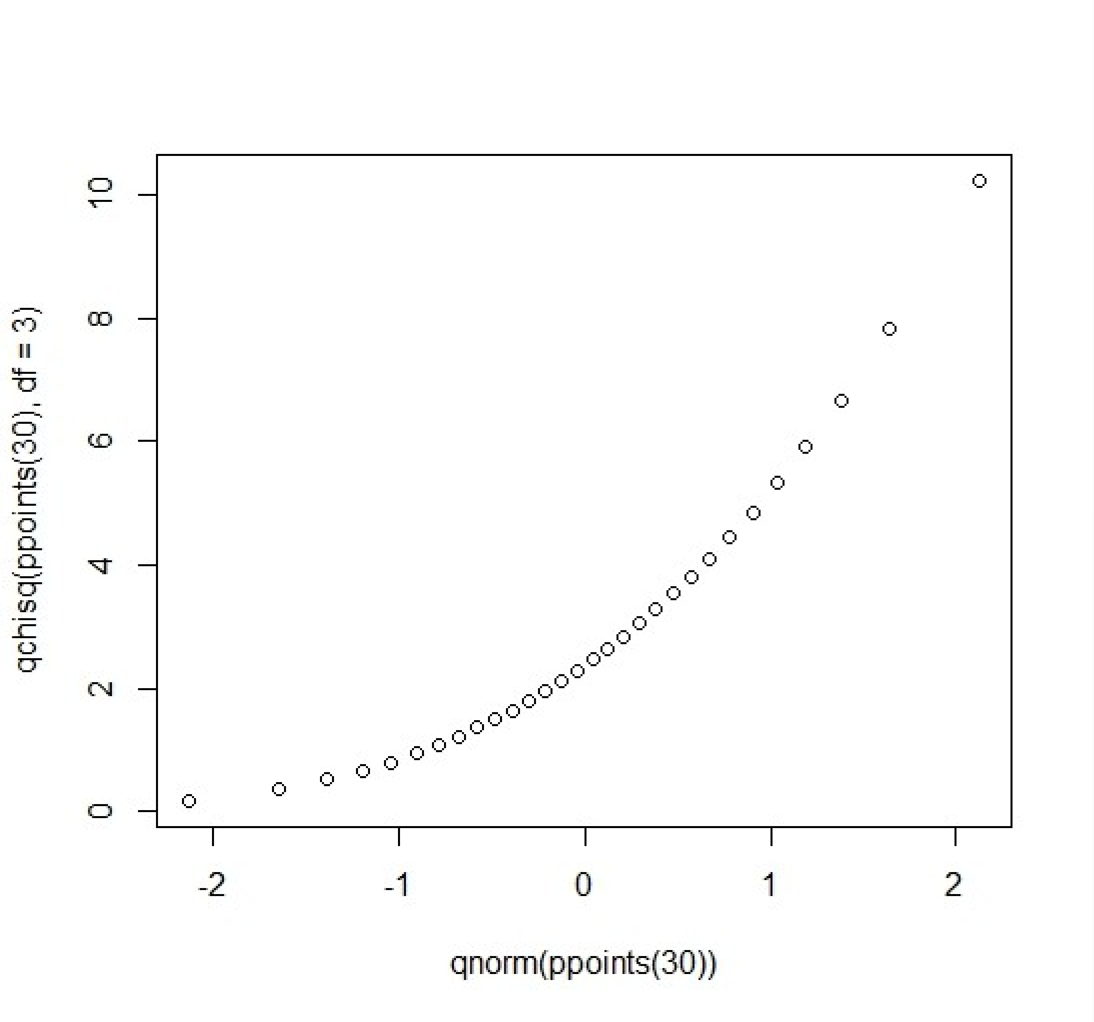
Evaluating Distributions: Q-Q Plots (cont.)

- Heavy (long) tails



Evaluating Distributions: Q-Q Plots (cont.)

- Heavy (long) tail to the right



An Example

- Predict English / language arts GPA
  - With gender
    - I.e., whether a student identifies as female
  - And special education status
    - I.e., whether a student has an individualized education program (IEP)
- Comparing ANOVA with linear regression

ELA GPA and Gender

Correlations

		ELA Grade	Female?
ELA Grade	Pearson Correlation	1	.320**
	Sig. (2-tailed)		.000
	N	248	175
Female?	Pearson Correlation	.320**	1
	Sig. (2-tailed)	.000	
	N	175	592

**. Correlation is significant at the 0.01 level (2-tailed).

ELA GPA and IEP

Correlations

		ELA Grade	Special Education Status
ELA Grade	Pearson Correlation	1	-.704**
	Sig. (2-tailed)		.000
	N	248	128
Special Education Status	Pearson Correlation	-.704**	1
	Sig. (2-tailed)	.000	
	N	128	474

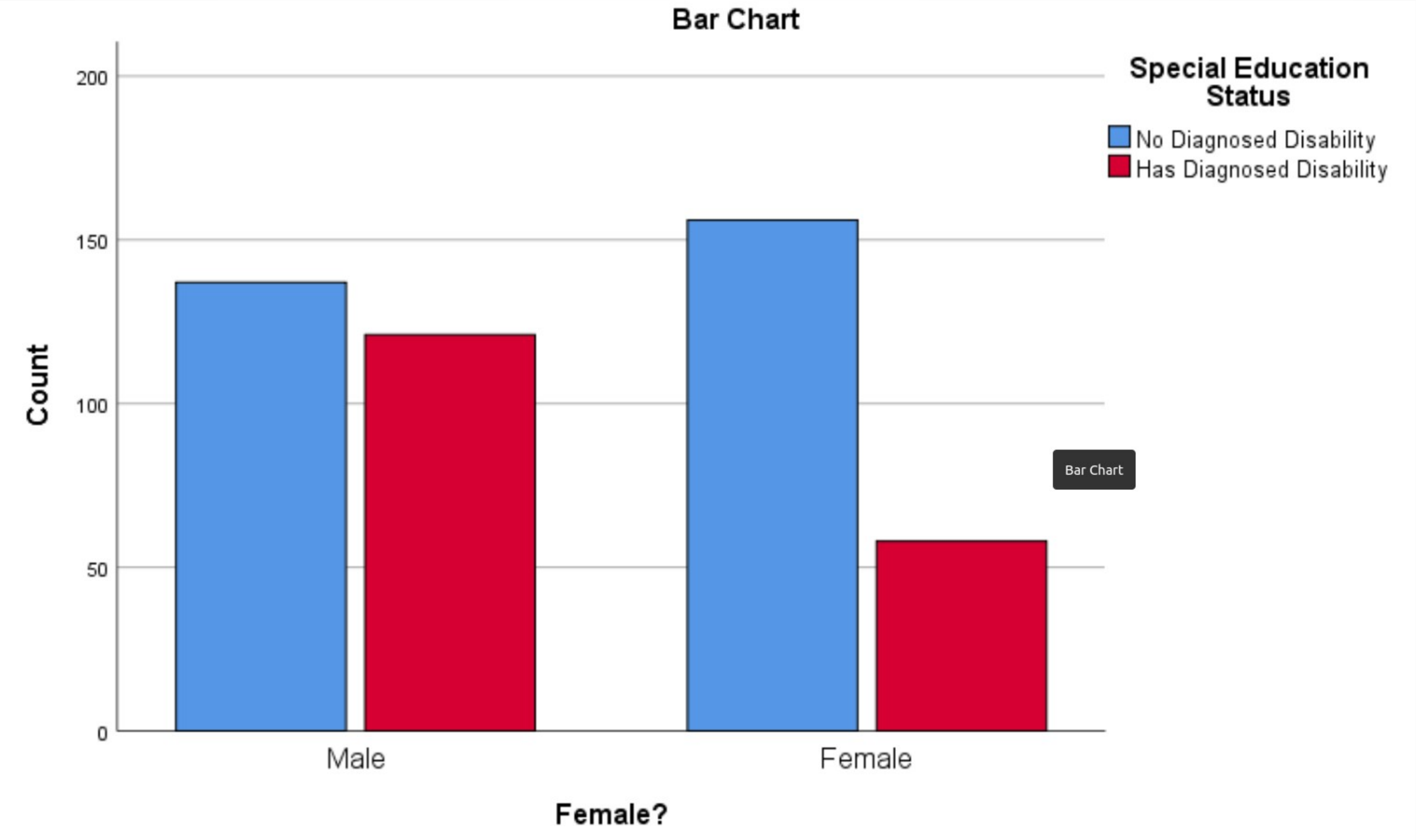
**. Correlation is significant at the 0.01 level (2-tailed).

Gender and IEP

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Female? * Special Education Status	472	70.4%	198	29.6%	670	100.0%

Gender and IEP (cont.)



Gender and IEP (cont.)

Female? * Special Education Status Crosstabulation

Count

		Special Education Status		Total
		No Diagnosed Disability	Has Diagnosed Disability	
Female?	Male	137	121	258
	Female	156	58	214
Total		293	179	472

Gender and IEP (cont.)

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	19.473[a]	1	.000		
Continuity Correction[b]	18.641	1	.000		
Likelihood Ratio	19.781	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	19.432	1	.000		
N of Valid Cases	472				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 81.16.

b. Computed only for a 2x2 table

ANOVA Results

Tests of Between-Subjects Effects

ELA Grade

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	47.905[a]	3	15.968	47.133	.000
Intercept	723.631	1	723.631	2135.891	.000
Gender	1.275	1	1.275	3.763	.055
Spec_Ed	34.708	1	34.708	102.445	.000
Gender * Spec_Ed	.766	1	.766	2.262	.135
Error	41.672	123	.339		
Total	1013.464	127			
Corrected Total	89.577	126			

a. R Squared = .535 (Adjusted R Squared = .523)

Linear Regression

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Female?[b]	.	Enter
2	Special Education Status[b]	.	Enter

a. Dependent Variable: ELA Grade

b. All requested variables entered.

Linear Regression (cont.)

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.320[a]	.102	.095	.71161	.102	14.379	1	126	.000
2	.727[b]	.529	.521	.51776	.426	113.013	1	125	.000

a. Predictors: (Constant), Female?

b. Predictors: (Constant), Female?, Special Education Status

c. Dependent Variable: ELA Grade

In ANOVA values are:

R^2 = .535

Adjusted R^2 = .523

Linear Regression (cont.)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7.281	1	7.281	14.379	.000[b]
	Residual	63.804	126	.506		
	Total	71.085	127			
2	Regression	37.577	2	18.788	70.087	.000[c]
	Residual	33.509	125	.268		
	Total	71.085	127			

a. Dependent Variable: ELA Grade

b. Predictors: (Constant), Female?

c. Predictors: (Constant), Female?, Special Education Status

Linear Regression (cont.)

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	2.418	.087		27.866	.000
	Female?	.479	.126	.320	3.792	.000
2	(Constant)	2.904	.078		37.257	.000
	Female?	.276	.094	.185	2.944	.004
	Special Education Status	-1.027	.097	-.667	-10.631	.000

a. Dependent Variable: ELA Grade

In Model 1:

- Intercept (b_0) is 2.418
- Value for Gender (b_1) is 0.479

Linear Regression (cont.)

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	2.418	.087		27.866	.000
	Female?	.479	.126	.320	3.792	.000
2	(Constant)	2.904	.078		37.257	.000
	Female?	.276	.094	.185	2.944	.004
	Special Education Status	-1.027	.097	-.667	-10.631	.000

a. Dependent Variable: ELA Grade

In Model 1:

- $b_0 \approx 2.4$
- $b_1 \approx 0.5$

Linear Regression (cont.)

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	2.418	.087		27.866	.000
	Female?	.479	.126	.320	3.792	.000
2	(Constant)	2.904	.078		37.257	.000
	Female?	.276	.094	.185	2.944	.004
	Special Education Status	-1.027	.097	-.667	-10.631	.000

a. Dependent Variable: ELA Grade

In Model 1:

- $Y' = 2.4 + 0.5X + e$

Dummy Variables

- In Model 1:
  - $Y' = 2.4 + 0.5X_1$
    - I.e., ignoring error
- If a student is **male**:
  - $X_1 = 0$
  - $Y' = 2.4 + 0.5(0)$
  - $Y' = 2.4 + 0$
  - $Y' = 2.4$

Dummy Variables (cont.)

- In Model 1:
  - $Y' = 2.4 + 0.5X_1$
    - I.e., ignoring error
- If a student is **female**:
  - $X_1 = 1$
  - $Y' = 2.4 + 0.5(1)$
  - $Y' = 2.4 + 0.5$
  - $Y' = 2.9$

Our analyses told us that 2.9 is significantly different than 2.4

Dummy Variables (cont.)

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	2.418	.087		27.866	.000
	Female?	.479	.126	.320	3.792	.000
2	(Constant)	2.904	.078		37.257	.000
	Female?	.276	.094	.185	2.944	.004
	Special Education Status	-1.027	.097	-.667	-10.631	.000

a. Dependent Variable: ELA Grade

In Model **2**:

- Intercept (b_0) is 2.904
- Value for Gender (b_1) is 0.276
- Value for Special Education Status is (b_2) is -1.027

Dummy Variables (cont.)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	2.418	.087		27.866	.000
	Female?	.479	.126	.320	3.792	.000
2	(Constant)	2.904	.078		37.257	.000
	Female?	.276	.094	.185	2.944	.004
	Special Education Status	-1.027	.097	-.667	-10.631	.000

a. Dependent Variable: ELA Grade

In Model **2**:

- $b_0 \approx 2.9$
- $b_1 \approx 0.3$
- $b_2 \approx -1$

Dummy Variables (cont.)

- In Model 2:
  - $Y' = 2.9 + 0.3X_1 - 1X_2$
    - I.e., ignoring error
- If a student is **male** and does **not** have an IEP:
  - $X_1 = 0$
  - $X_2 = 0$
  - $Y' = 2.9 + 0.3(0) - 1(0)$
  - $Y' = 2.9 + 0 - 0$
  - $Y' = 2.9$

Dummy Variables (cont.)

- In Model 2:
  - $Y' = 2.9 + 0.3X_1 - 1X_2$
    - I.e., ignoring error
- If a student is **male** and **does have** an IEP:
  - $X_1 = 0$
  - $X_2 = 1$
  - $Y' = 2.9 + 0.3(0) - 1(1)$
  - $Y' = 2.9 + 0 - 1$
  - $Y' = 1.9$

Dummy Variables (cont.)

- In Model 2:
  - $Y' = 2.9 + 0.3X_1 - 1X_2$
    - I.e., ignoring error
- If a student is **female** and does **not** have an IEP:
  - $X_1 = 1$
  - $X_2 = 0$
  - $Y' = 2.9 + 0.3(1) - 1(0)$
  - $Y' = 2.9 + 0.3 - 0$
  - $Y' = 3.2$

Dummy Variables (final)

- In Model 2:
  - $Y' = 2.9 + 0.3X_1 - 1X_2$
    - I.e., ignoring error
- If a student is **female** and **does have** an IEP:
  - $X_1 = 1$
  - $X_2 = 1$
  - $Y' = 2.9 + 0.3(1) - 1(1)$
  - $Y' = 2.9 + 0.3 - 1$
  - $Y' = 2.2$

Multicollinearity

- When two or more predictors share too much variance
- Two general sources:
  - **Structural**: Caused by how the model was constructed
    - E.g., adding interaction terms
  - **Data**: Caused by variables that are inherently correlated

Multicollinearity (cont.)

- Problems caused by multicollinearity:
  - Parameter estimates of multicollinear terms can be unstable
    - Significance tests of them can also fail
  - Reduces the power of the whole model
    - Because the parameter estimates are less precise

Multicollinearity (cont.)

- Multicollinearity does **not** affect predictions made by the model
  - Or the model's goodness-of-fit statistics
- Can be tested with a "variance inflation factor" (VIF)
  - VIF ranges from 1 to ∞
  - Where values >10 usually indicate problems

Multicollinearity (cont.)

- Addressing multicollinearity
  - Centering variables (subtracting the mean) can reduce structural multicollinearity (Iacobucci et al., 2016)
  - Remove one of the correlated variables
  - Only make predictions / test model fit
  - Use another analysis
    - E.g., canonical correlations or principal component analysis

Multicollinearity (final)

- Multicollinearity is typically not a concern if the variables with high multicollinearity are:
  - Control variables
  - Intentional products of other variables
    - E.g., raised to a power, an interaction, etc.
  - Dummy variables

Independence of Cases

- When one case (participant, round of tests, etc.) is correlated with another case
- Can also produce unstable parameter estimates
    - Thus affecting significance tests
        - And both Type 1 & 2 errors
- May also affect model goodness of fit
    - And not isolated to a few predictors

Independence of Cases (cont.)

- Addressing non-independence
  - Best is through research design
  - Can also model inter-dependence
    - E.g., nesting cases
      - As is done explicitly in multilevel (hierarchical) models


The End