Intro to Linear Regression

Overview

- Review of Underlying Concepts in Inferential Statistics
- Understanding Linear Models
- Terms in Linear Models
- An Example
- Further Considerations

Review of Underlying Concepts

Review of Underlying Concepts

- 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

Review of Underlying Concepts (cont.)

- Assumptions made in computing correlations
  - Measurement level is correctly conceived (ordinal, interval, ratio, etc.)
  - Relationship is linear
- Assumptions made in testing their significance
  - Monotonicity
    - For Pearson's r, also that variables are normally distributed & homoscedastic
      - And that the variables are bivariate normal
  - No big outliers

Review of Underlying Concepts (cont.)

- Partial & Semipartial Correlations
  - Semipartial correlations remove the effect of another variable from **one** of the correlated pair
  - Could remove the effect of several other variables from one of the pair
    - (Or create even more complex arrangements, like canonical correlations)

Review of Underlying Concepts (cont.)

- Correlations & Error
  - Correlations separate dispersion into variance & covariance
    - But make no assumptions about where error comes from
  - However, when testing significance of Pearson's r, error is assumed to be normally distributed

Review of Underlying Concepts (end)

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

Understanding Linear Models

Basic Concepts

- Simplest form of a linear relationship is $Y = bX$
  - Where:
    - Y = **Outcome** / response / criterion / DV
    - X = **Predictor** / input / IV
    - b = Slope of X
      - The typical null hypothesis (H_0) of "no effect" is expressed here as: $b_1 = 0$
      - (If data are standardized, the convention is to write instead of b)

Basic Concepts (cont.)

- However, we typically add: $Y = b_0 + b_1X_1 + e$
  - Y = Outcome
  - b_0 = Value of X at y-axis intercept
  - b_1 = Slope of X
  - X_1 = Predictor X_1
  - e = Error

Linear Models vs. Correlations

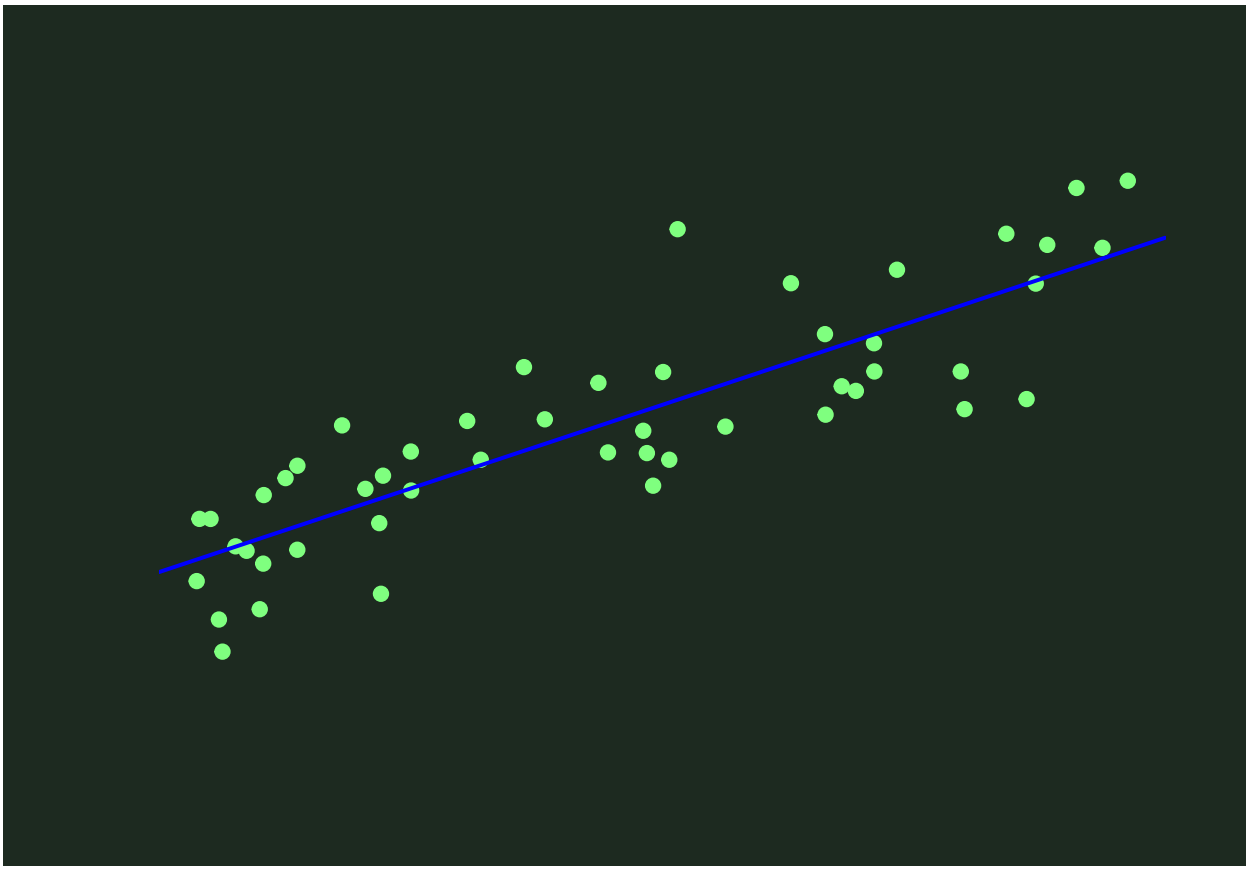
- For a correlation:

$$r = \frac{\text{Cov}(X_1, Y)}{\text{SD}(X_1)\text{SD}(Y)}$$

- 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
    - I.e., without worrying about where the axes are—or where unshared variance is coming from

Linear Models vs. Correlations (cont.)

$r = \frac{\text{Cov}(X_1, Y)}{\text{SD}(X_1)\text{SD}(Y)}$; here, $r = .86$:

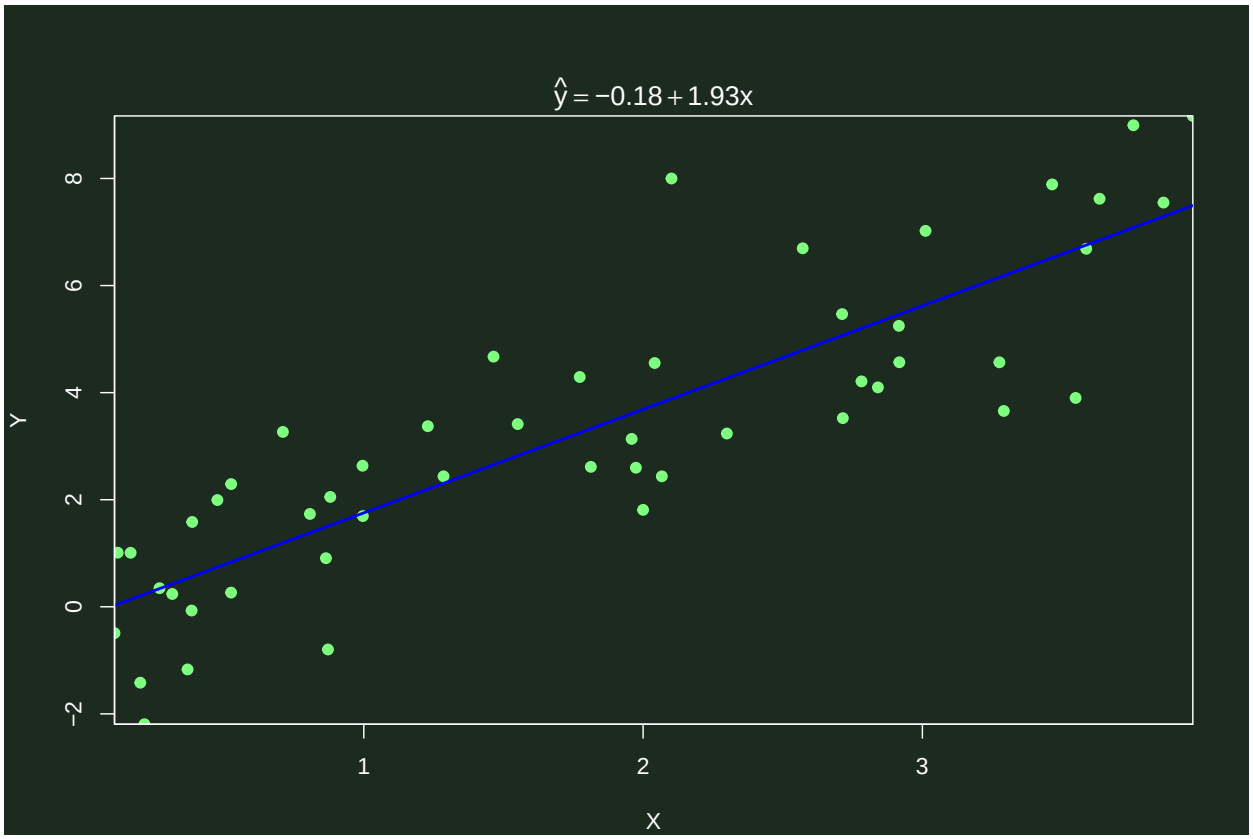


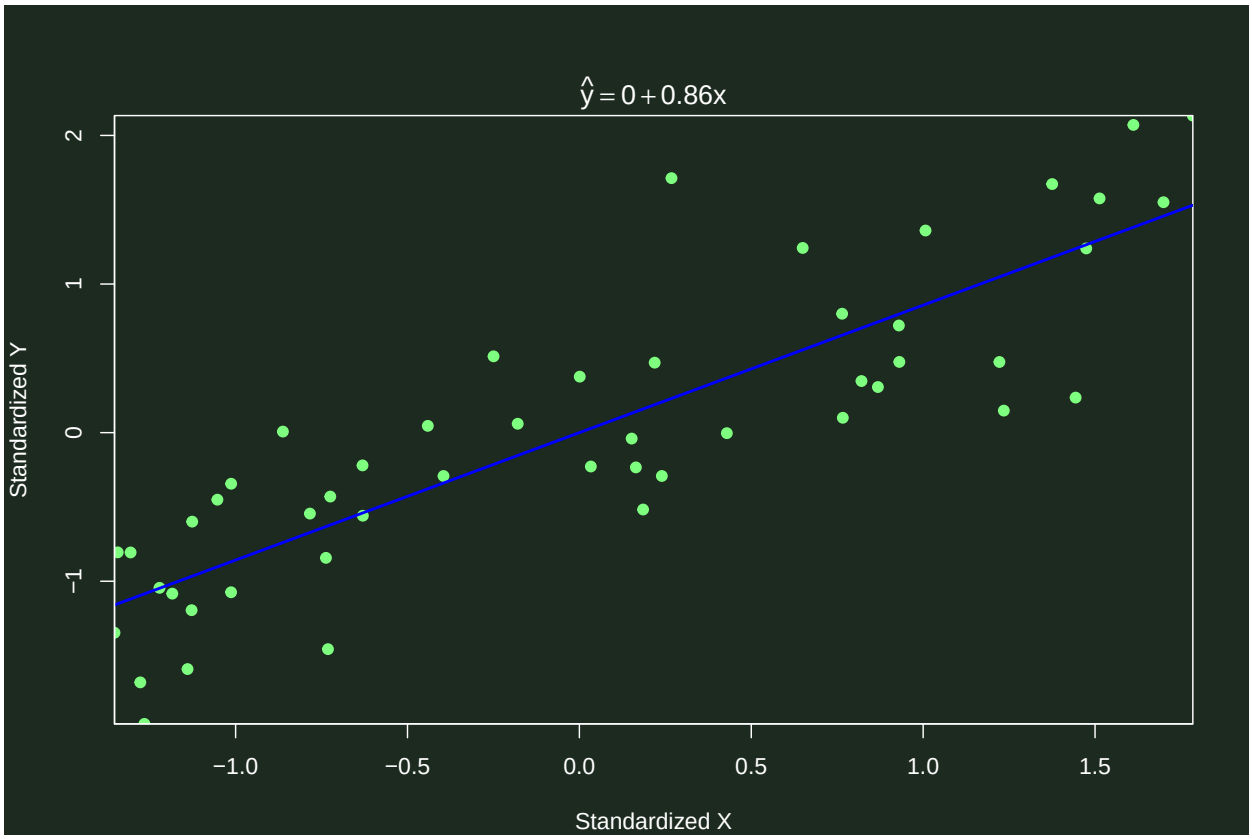
Linear Models vs. Correlations (cont.)

- For a linear regression, we instead minimize variance for only one variable
  - Typically the outcome
  - This assumes that all variance (error) resides in the outcome
- So, in $Y = b_0 + b_1 X + e$:

$$b_1 = \frac{\text{Cov}(X_1, Y)}{(\text{SD}(Y))^2}$$

Linear Models vs. Correlations (cont.)





More About the Equation

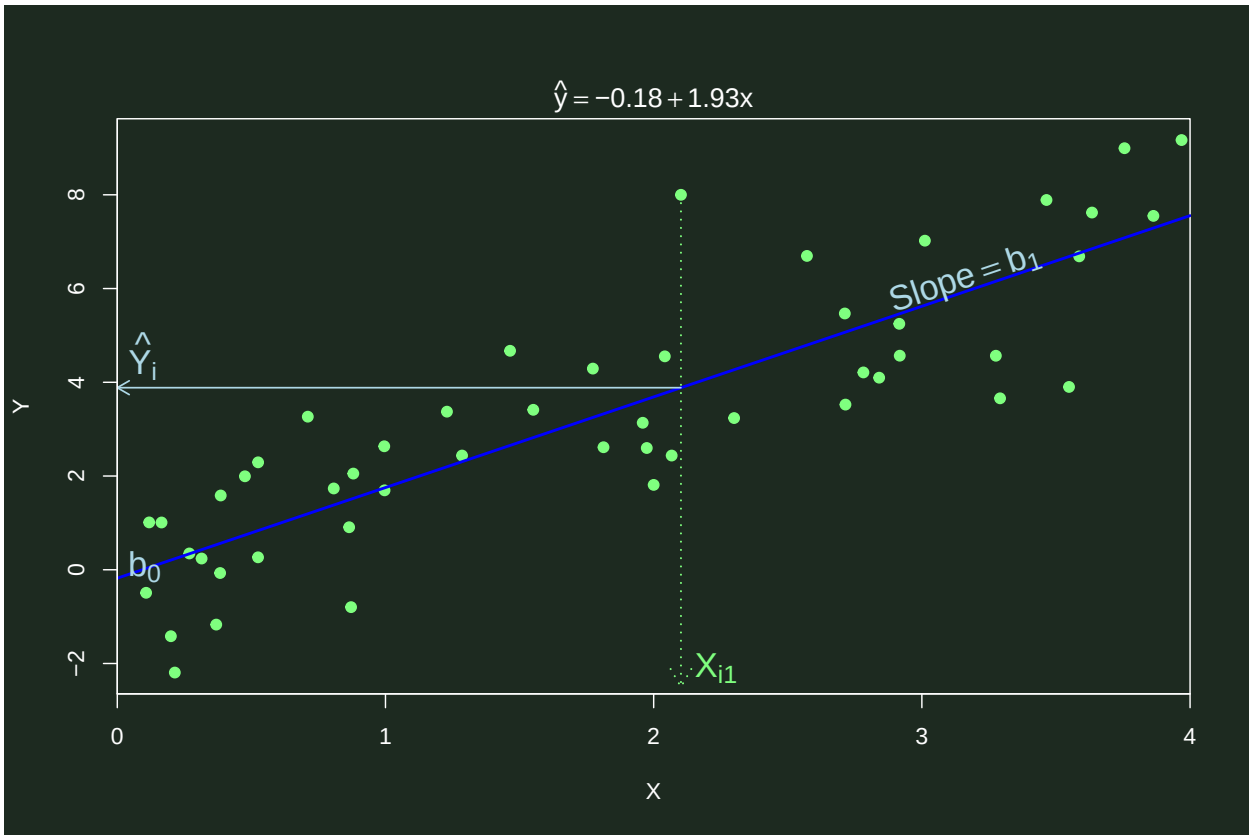
- $Y = b_0 + b_1 X_1 + e$
- Note again that error is separated out
  - And placed on the side with the predictor
- Implications:
  - The value of X_1 per se is without error
    - Because error is separated out (as e)
  - The intercept, slope, & error can be estimated separately
    - And their covariances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

More About the Equation (cont.)

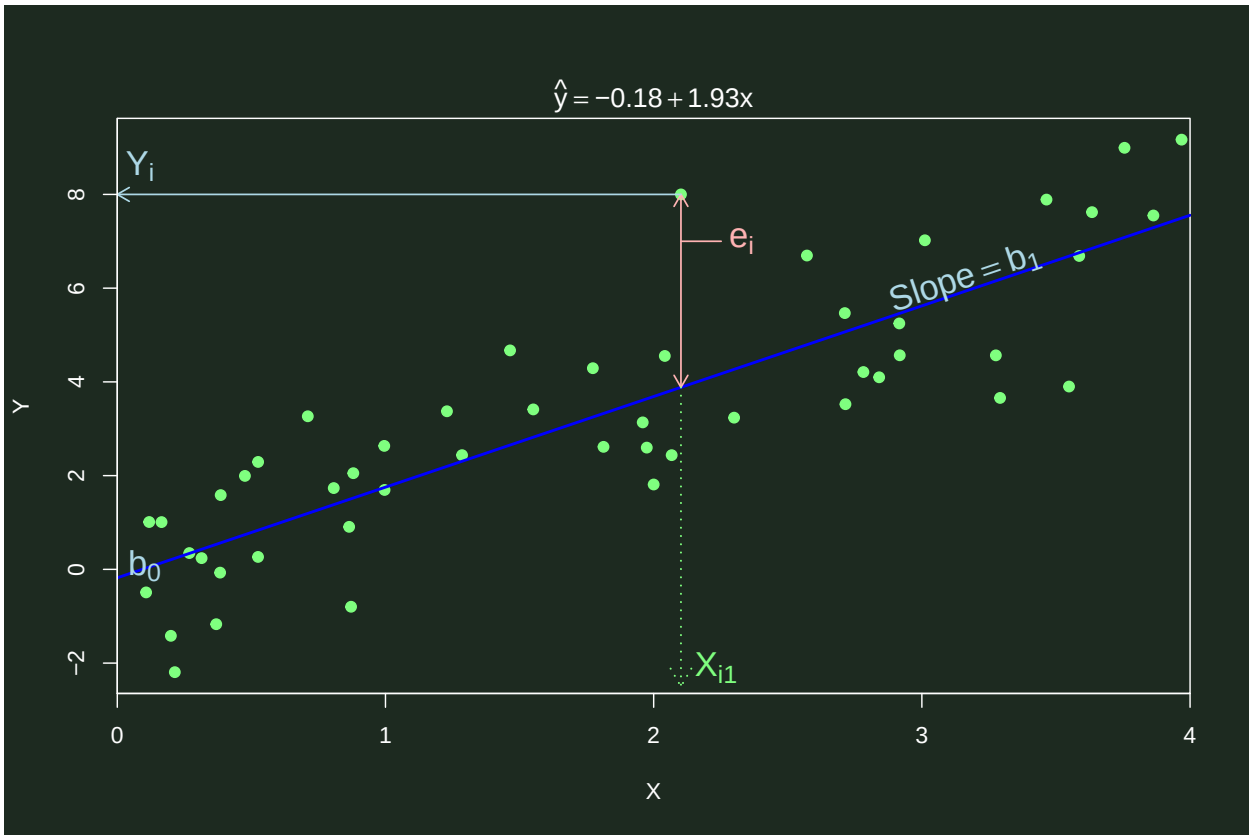
- Adding more specificity to the equation:
  $\hat{Y}_i = b_0 + b_1 X_{i1} + e_i$
  - $\hat{Y}_i$ = Predicted value of Y for participant i
  - b_1 = Slope for variable X_1
  - X_{i1} = Value on X_1 for participant i
  - e_i = Error of measurement of participant i's outcome

More About the Equation (cont.)

- Participant i's score on X_1 here is 2.10
- The ***predicted*** value of Y_i for participant i is:
  - $\hat{Y}_i = -0.18 + (1.93 \times 2.10) = 3.97$



- The ***actual*** value of Y_i for participant i also includes the error:
  - $Y_i = -0.18 + (1.93 \times 2.10) + 4.13 = 8.00$



More About the Equation (end)

- Adding another variable to the equation:
  $\hat{Y}_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + e_i$
  - X_{i2} = Participant i's value on the other variable (X_2) added to the model
  - b_2 = Slope for X_2
- Since there are multiple predictors (Xs) 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
  - Is a more flexible framework
  - Can model complex relationships & data structures
    - E.g., non-linear relationships & nested data
  - And can test whole models
    - And changes to the whole model when variables are added or removed

Questions Best Addressed by ANOVAs vs. Linear Models

- ANOVAs (and ANCOVAs, MANOVAs, etc.) can ask:
  - Which variable is significant?
  - Is there an interaction between variables?
- Linear regressions can *also* ask:
  - What is the best combination of variables?
  - Does a given variable—or set of variables—significantly contribute to what we already know?

Terms in Linear Models

Adding More Terms to Models

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

$$\hat{Y}_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + b_3 X_{i3} + b_4 X_{i4} + e_i$$

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

$$\hat{Y}_i = b_0 + b_1 X_{i1} ... + b_k X_{ik} + e_i$$

  - Were k is the number of variables

Adding More Terms to Models (cont.)

$\hat{Y}_i = b_0 + b_1 X_{i1} ... + b_k X_{ik} + e_i$

- We can test interactions by adding additional terms
  - E.g., $...b_1 X_{i1} + b_2 X_{i2} + \mathbf{b_3}(\mathbf{X_{i1} X_{i2}})...$
- Or test non-linear effects, also by adding terms
  - E.g., $...b_1 X_{i1} + \mathbf{b_2 X_{i1}^2}...$

Adding More Terms to Models (end)

- Just as we separated out the effects of the predictors,
  - We can separate out sources of error
    - 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 (aka hierarchical) models

Signal-to-Noise in Linear Models

- Signal-to-noise in the equation

$\hat{Y}_i = b_0 + b_1 X_{i1} ... + b_k X_{ik} + e_i$

- The variance in $\hat{Y}_i$ is divided into:
  - Changes due to the predictors
  - Changes due to "other things" (and relegated to error / noise term(s)
  - I.e., into signals and noise(s)
  - (N.b., the intercept, b_0, is a constant and not included in this partitioning of variance)

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

$\hat{Y}_i = b_0 + b_1 X_{i1} ... + b_k X_{ik} + e_i$

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

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

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

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

- I.e.,the squared sum of the differences of each instance (Y_i) from the mean ($\overline{Y}$) equals:
  - The squared sum differences of each predicted value ($\hat{Y}_i$) from the mean
  - Plus the squared sums of differences of the actual values (Y_is) from the respective predicted values

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

- Another way of saying this:

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

- Is to say this:
  Total SS = SS from Regression + SS from Error
  - Or, further condensed as:
    - $SS_{Total} = SS_{Reg.} + SS_{Error}$

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

- Using $SS_{Total} = SS_{Reg.} + SS_{Error}$,
  - We can compute the ratio of predicted to actual:
    Ratio of Predicted-to-Actual Variance $= \frac{SS_{Reg.}}{SS_{Total}}$
    - Or, equivalently as $1 - \frac{SS_{Reg.}}{SS_{Error}}$

. . .

- We typically represent this ratio **as R^2**:

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

. . .

Yep, that's what R^2 means in ANOVAs

The Equation's Terms

$\hat{Y}_i = b_0 + b_1 X_{i1} ... + b_k X_{ik} + e_i$

- Y is assumed to follow a certain distribution
  - This determines how error is modeled
  - E.g., is error usually assumed to be normally distributed
  - But both distributions can be assumed to be something else
    - E.g., logarithmic

The Equation's Terms (cont.)

$\hat{Y}_i = b_0 + b_1 X_{i1} ... + b_k X_{ik} + e_i$:

- Xs 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

The Equation's Terms (cont.)

$\hat{Y}_i = b_0 + b_1 X_{i1} ... + b_k X_{ik} + e_i$:

- For an **ANOVA**:
  - Y is assumed to be normally distributed
  - The Xs are nominal
  - And the terms are not transformed
    - Called an "identity" because they are multiplied by 1
  - This "identity transformation" looks like this:
    $\hat{Y}_i = 1 \times (b_0 + b_1 X_{i1} ... b_k X_{ik} + e_i)$

The Equation's Terms (end)

- The terms can be 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
- E.g., logistic regression uses a logarithmic (e) link:
  $\hat{Y}_i = \frac{e^{b_0+b_1X_{i1}+\cdots+b_kX_{ik}}}{1+e^{b_0+b_1X_{i1}+\cdots+b_kX_{ik}}}$
  which is more often written as:
  $\ln\frac{\hat{Y}_i}{1-\hat{Y}_i} = b_0 + b_1X_{i1} + \cdots + b_kX_{ik}$

Generalized Linear Models

- That very general 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 typically 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)

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 you already know.
    - I.e., those with:
      - Normally-distributed, `iid` variables
      - Identity link functions
    - Like ANOVAs & multiple linear regressions

Generalized Linear Models (end)

- Assumptions of generalized linear models:
  - Relationship between response and predictors must be expressible as a linear function
    - But many can model heteroscedasticity well
  - Cases must be independent of each other
    - But predictors should not be too inter-correlated (lack of multicollinearity)
  - The random & link functions should approximate the real functions

An Example

Predicting BMI from Sex & Neighborhood Safety

- Predicting body mass index levels among adolescents from:
  1. Whether an adolescent is biologically female
  2. Whether they feel their neighborhood is safe
- via SPSS (v. 29)

Data Used

- From the National Longitudinal Study of Adolescent to Adult Health (Add Health)
  - Using the prepared add_health.sav dataset
  - Since data are longitudinal, only the first instance (wave) of data collection was used
    - Selected via:
      1. `Data` > `Select Cases...`
      2. Under `If condition is satisfied`, added `Wave = 1` to select only the first wave

Descriptives

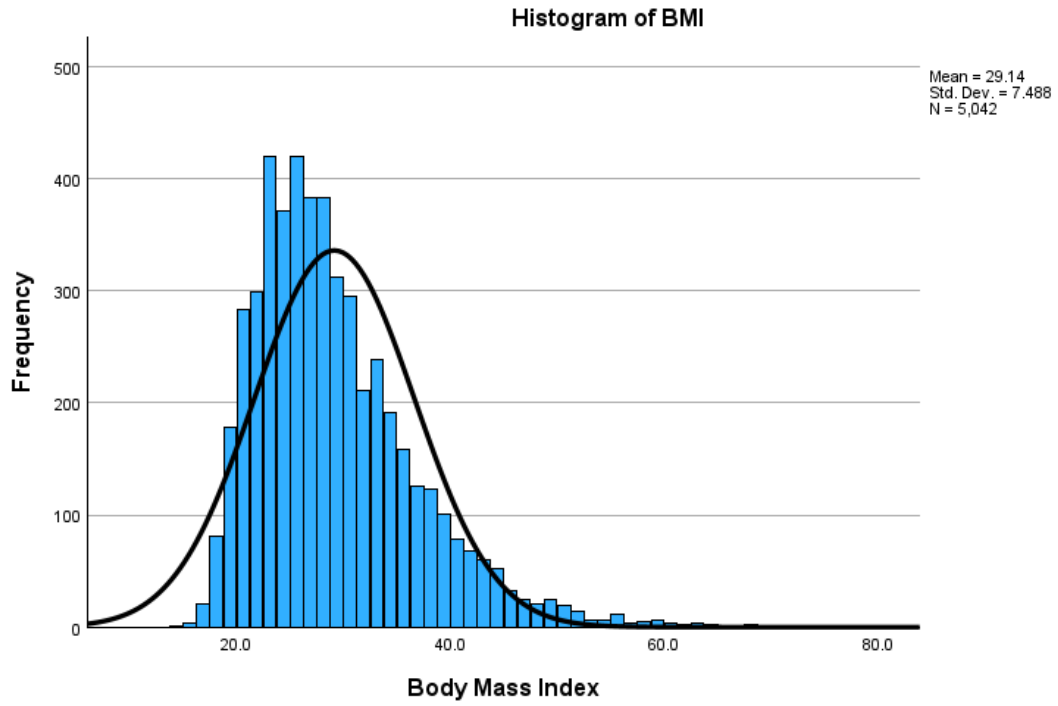
Descriptive Statistics

	N	Minimum	Maximum	Mean	
	Statistic	Statistic	Statistic	Statistic	Std. Error
Body Mass Index	5042	14.4	70.3	29.144	.1055
Biological Sex	6503	0	1	.52	.006
Feel Safe in Neighborhood	6468	0	1	.90	.004
Valid N (listwise)	5025				

- The mean BMI (29.144) was nearly obese
- Since `Bio_Sex` was coded 0 = `Male` & 1 = `Female`, 52% of the participants were biologically female
- And 90% reported feeling safe in their neighborhood
- About 77% (5025/6503) of cases had data on all three variables

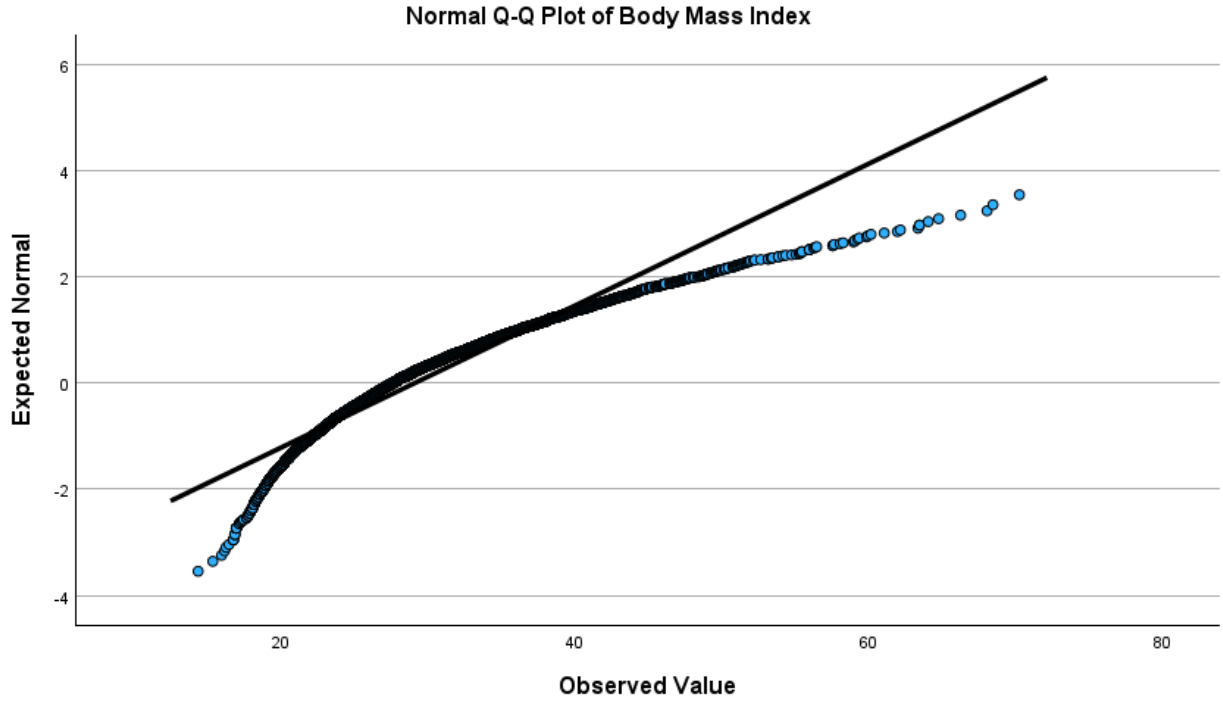
Descriptives (cont.)

- BMI was positively skewed
- So, those with exceptionally high BMIs affect the results more



Descriptives: Q-Q Plot

- `Analyze` > `Explore...` > `Plots` > `Normality plots with tests`
- That skew—and a limited lower range—caused some deviations from normality



Descriptives: 2 × 2 Table

- `Analyze` > `Descriptives Statistics` > `Crosstabs...`
- The number of adolescents who felt safe in their neighborhood is not significantly different between the sexes

		Feel Safe in Neighborhood (0) No	(1) Yes	Total
Biological Sex	0	314	2818	3132
	1	361	2975	3336
Total		675	5793	6468

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1.095[a]	1	.295		
Continuity Correction[b]	1.011	1	.315		
Likelihood Ratio	1.096	1	.295		
Fisher's Exact Test				.309	.157
Linear-by-Linear Association	1.094	1	.296		
N of Valid Cases	6468				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 326.86.

b. Computed only for a 2x2 table

Correlations

- The correlations also reflect the weak relationship between `Bio_Sex` & `Feel_Safe_in_Nghbrhd`
- Feeling safe—but not sex—significantly correlated with BMI

Correlations

		Body Mass Index	Biological Sex	Feel Safe in Neighborhood
Body Mass Index	Pearson Correlation	1	.026	-.058**
	Sig. (2-tailed)		.066	<.001
	N	5042	5041	5025
Biological Sex	Pearson Correlation	.026	1	-.013
	Sig. (2-tailed)	.066		.296
	N	5041	6503	6468
Feel Safe in Neighborhood	Pearson Correlation	-.058**	-.013	1
	Sig. (2-tailed)	<.001	.296	
	N	5025	6468	6468

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

Correlations with CIs

- The 95% confidence intervals (and correlations themselves) are slightly different when using Fisher's r-to-z transformation versus bootstrapping
- Given the deviations from normality, bootstrapping is preferable here

95% confidence intervals generated from Fisher's r-to-z transformation:

Correlations

				Statistic		
Variable	Variable2	Correlation	Count	Lower C.I.	Upper C.I.	Notes
BMI	Bio_Sex	.026	5041	-.002	.053	
	Feel_Safe_in_Nborhood	-.058	5025	-.085	-.030	

Missing value handling: PAIRWISE, EXCLUDE. C.I. Level: 95.0

95% confidence intervals generated from bootstrapping:

Correlations

				Biological Sex	Feel Safe in Neighborhood
Body Mass Index	Pearson Correlation			.025	-.058**
	Sig. (2-tailed)			.078	<.001
	N			5025	5025
	Bootstrap[c]	Bias		.000	.000
		Std. Error		.014	.016
		95% Confidence Interval	Lower	-.003	-.092
			Upper	.053	-.028

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

c. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Linear Regression

- Conducted via:
  1. `Analyze` > `Regression` > `Linear`
  2. `BMI` as `Dependent`
  3. `Bio_Sex` & `Feel_Safe_in_Nghbrhd` as predictors in `Block 1 of 1`

Linear Regression (cont.)

- The combination of `Bio_Sex` & `Feel_Safe_in_Nghbrhd` did not explain much of the variance in `BMI` scores
  - The R^2 was .004; adjusted for number of terms in the model, it was .003
  - This combination of variables thus only accounted for about 0.3% – 0.4% of the total variance in BMIs
    - The high standard error, however, indicates that replications may find rather different R^2s

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.061[a]	.004	.003	7.4992

a. Predictors: (Constant), Feel Safe in Neighborhood, Biological Sex

Linear Regression (cont.)

- Nonetheless, the model was significant
  - The intercept & sample size both surely helped
- This ANOVA source table presents the effect of the overall model
  - Like first testing a variable in an ANOVA before conducting *post hoc*s, this helps protect against over-interpreting

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1859.991	2	929.996	16.537	<.001[b]
	Residual	499677.905	8885	56.238		
	Total	501537.896	8887			

a. Dependent Variable: Body Mass Index

b. Predictors: (Constant), Feel Safe in Neighborhood, Biological Sex

Linear Regression (cont.)

- Both biological sex & feeling safe in one's neighborhood both significantly predicted BMI
- The standardized for sex means its effect size was close to "small"
  - It is "medium" for feeling safe (q.v. ^{2} criteria in this table)
- The positive effect for sex means those identifying as female (1s) tended to have higher BMIs than those identifying as male (0s)
- The negative value for feeling safe means those who felt safe (1s) tended to have lower BMIs than those who didn't (0s)

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound
1	(Constant)	30.205	.259		116.415	<.001	29.696	30.714
	Biological Sex	.325	.159	.022	2.036	.042	.012	.637
	Feel Safe in Neighborhood	-1.383	.258	-.057	-5.356	<.001	-1.890	-.877

a. Dependent Variable: Body Mass Index

Interpretting the Effects

- Writing these results in linear equation form:

$$\hat{\text{BMI}} = 30.205 + (0.325 \times \text{Sex}) + (-1.383 \times \text{Feeling Safe})$$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	30.205	.259		116.415	<.001	29.696	30.714
	Biological Sex	.325	.159	.022	2.036	.042	.012	.637
	Feel Safe in Neighborhood	-1.383	.258	-.057	-5.356	<.001	-1.890	-.877

a. Dependent Variable: Body Mass Index

Interpretting the Effects (cont.)

$$\hat{\text{BMI}} = 30.205 + (0.325 \times \text{Sex}) + (-1.383 \times \text{Feeling Safe})$$

- Since `Bio_Sex` was coded 0 = `Male` & 1 = `Female`
  - And `Feel_Safe_in_Nghbrhd` as 0 = `No` & 1 = `Yes`,
- We predict that the BMI
  - For a **male** (0)
  - Who does **not** feel safe (0)
  - Is 30.205:

$$\begin{aligned}\hat{\text{BMI}} &= 30.205 + (0.325 \times 0) + (-1.383 \times 0) \\ &= 30.205 + 0 + 0 \\ &= 30.205\end{aligned}$$

Interpretting the Effects (cont.)

$$\hat{\text{BMI}} = 30.205 + (0.325 \times \text{Sex}) + (-1.383 \times \text{Feeling Safe})$$

- The predicted BMI for
  - A **female** (1)
  - Who does **not** feel safe (0)
  - Is 30.530:

$$\begin{aligned}\hat{\text{BMI}} &= 30.205 + (0.325 \times 1) + (-1.383 \times 0) \\ &= 30.205 + 0.325 + 0 \\ &= 30.530\end{aligned}$$

Interpretting the Effects (cont.)

$$\hat{\text{BMI}} = 30.205 + (0.325 \times \text{Sex}) + (-1.383 \times \text{Feeling Safe})$$

- The predicted BMI for
  - A **male** (0)
  - Who **does** feel safe (1)
  - Is 29.147:

$$\begin{aligned}\hat{\text{BMI}} &= 30.205 + (0.325 \times 0) + (-1.383 \times 1) \\ &= 30.205 + 0 - 1.383 \\ &= 29.147\end{aligned}$$

- Etc.

Further Considerations

Multicollinearity

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

Multicollinearity (cont.)

- Problems caused by multicollinearity:
  - Parameter estimates of multicollinear terms can be unstable
    - And even reverse sign
  - Reduces the power of the whole model
    - Because the parameter estimates are less precise

Multicollinearity (cont.)

- Multicollinearity doesn't typically affect the whole model's R^2
  - Or the model's good-of-fit statistics
- It mostly impairs interpretation of individual predictors
- Can be tested with variance inflation factor (VIF)
  - VIF ranges from 1 to ∞
  - Where values >10 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 test/compare overall model fit
  - Use another analysis
    - E.g., canonical correlations or principal component analysis

Multicollinearity (end)

- Multicollinearity is typically not a concern if the variables with high multicollinearity are:
  - Control variables
  - Intentional products of other variables
    - E.g., interaction terms, raised to a power, 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
    - Through both false positives (Type 1) & false negatives (Type 2)
- 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 Games

- Space Invaders
- Lunar Lander
- Asteroids
  - Doesn't work well on Firefox
- Tempest
- Star Wars
- Battlezone
- Elite