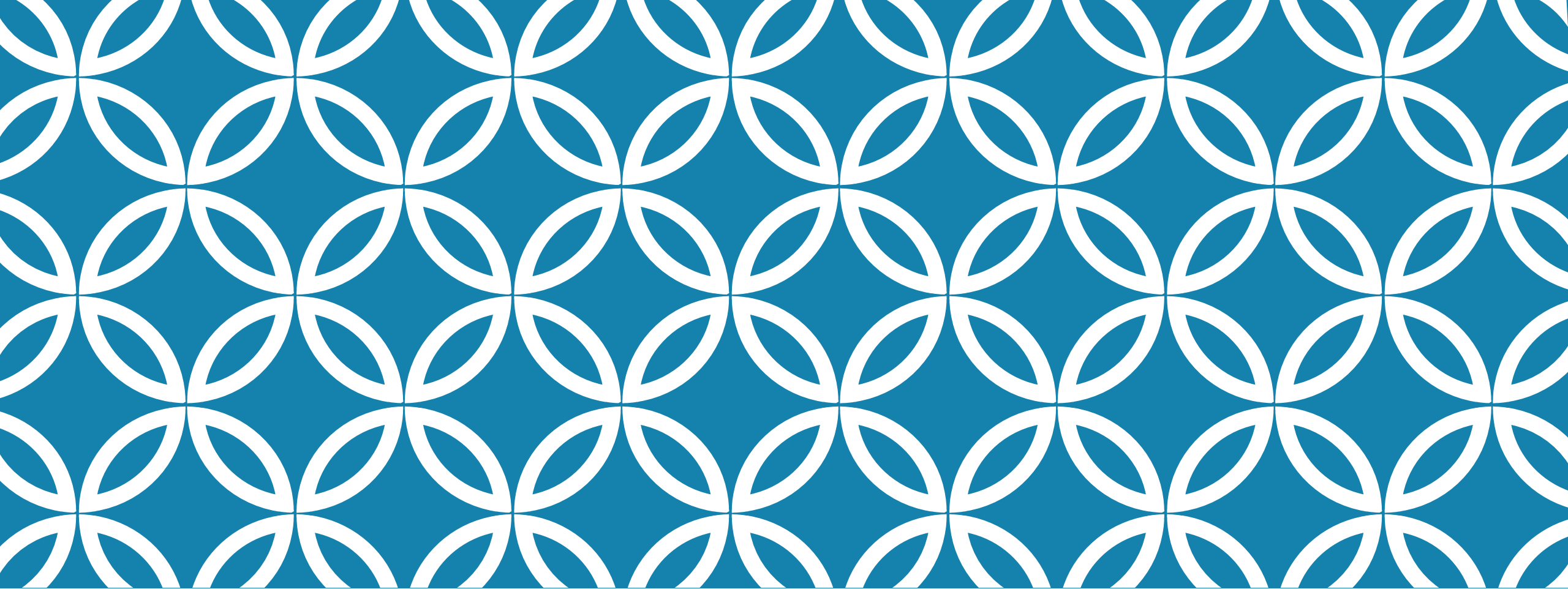


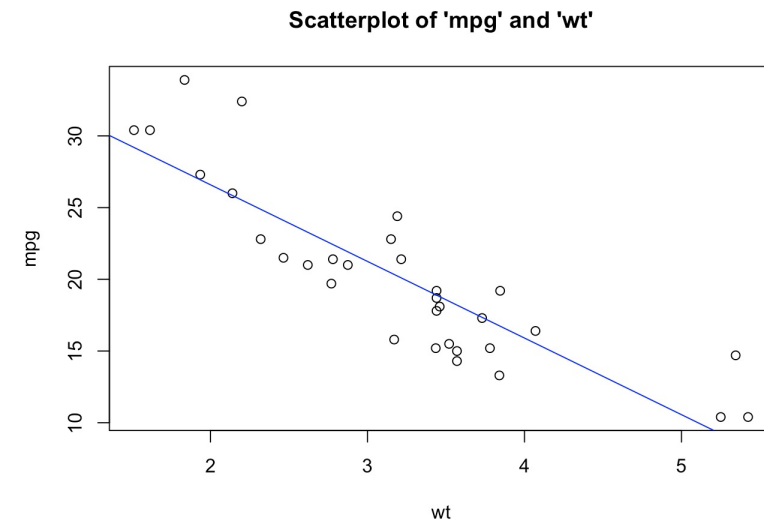


LINEAR/OLS REGRESSION

RESEARCH PROBLEM

Correlation: relationship between X and Y

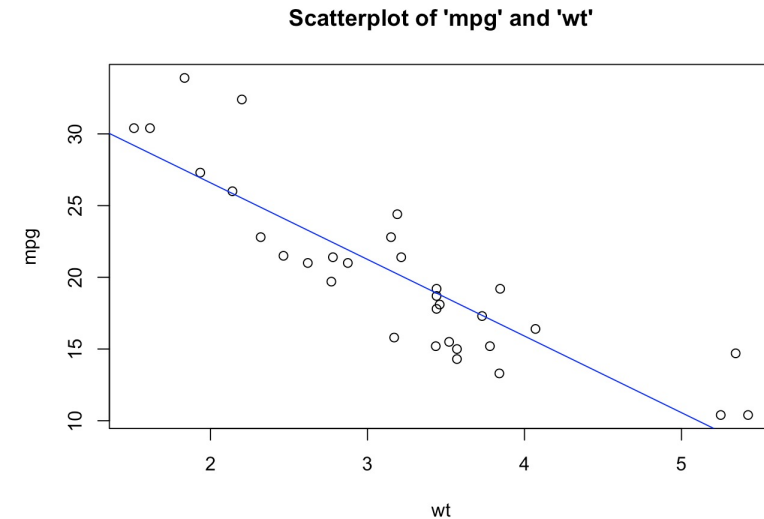
We use regression line (line of best-fit) to demonstrate how closely related X and Y were on scatterplot



RESEARCH PROBLEM

Equation of best-fit line (regression line) gives us a “best guess” for the value of Y at any given X value

- Gives “predicted” values of Y, at any given X
- Generated from all (X,Y) pairs



LOGIC OF REGRESSION LINE AND ANALYSIS

We can use regression line equation to predict (or explain variation in) values of Y ?

LOGIC OF REGRESSION LINE AND ANALYSIS

We can use values of X to predict values of Y

- Clearest illustration would be in correlation, where we use regression line to demonstrate relationship between two interval ratio variables
- But the logic of using values of a predictor (X) variable to predict values of the outcome variable (Y) can be extended to different kinds levels of measurement:
 - Nominal/Ordinal predictor (X), Interval-Ratio outcome (Y) – still regression
 - Nominal/Ordinal predictors ($X_1, X_2, X_3, \dots X_m$), Interval-Ratio outcome (Y) – still regression, but multiple
 - Nominal/Ordinal predictor (X), Nominal/Ordinal outcome (Y) – still regression, but of the Logistic/logit family

REGRESSION LINE EQUATION

Back to two interval-ratio variables – we can use algebra to predict a straight line relationship (values of Y for any given X)

Recall the primary school straight-line equation:

$$Y = mX + b$$

Y : value of the outcome variable

m : slope associated with X variable – for every 1-unit change in X , there is an m unit change in Y

X : value of X variable

b : value of the intercept – the Y value when $X == 0$

REGRESSION LINE EQUATION

This equation is based on information from X and Y values, and is often written as:

$$Y = a + bX$$

Y: value of the outcome variable

b: slope associated with X variable – for every 1-unit change in X, there is an b unit change in Y

X: value of X variable

a: value of the intercept – the Y value when X == 0

REGRESSION LINE EQUATION

But because the line almost never perfectly aligns the points on a scatterplot, there will be some distance between points and line – these are residuals, or error

$$Y = a + bX + e$$

Y: value of the outcome variable

b: slope associated with *X* variable – for every 1-unit change in *X*, there is an *b* unit change in *Y*

X: value of *X* variable

a: value of the intercept – the *Y* value when $X = 0$

e: error term – distance between *Y* values and the line

REGRESSION LINE EQUATION

Or, more appropriately:

$$\hat{Y} = a + bX$$

$\hat{Y}$: predicted value of the outcome variable

b : slope associated with X variable – for every 1-unit change in X , there is an b unit change in $\hat{Y}$

X : value of X variable

a : value of the intercept – the $\hat{Y}$ value when $X == 0$

REGRESSION LINE EQUATION — FINAL.PPTX

Or, more appropriately:

$$\hat{Y} = b_0 + b_1X_1$$

$\hat{Y}$: predicted value of the outcome variable

b_1 : slope associated with X variable – for every 1-unit change in X , there is an b unit change in $\hat{Y}$

X_1 : value of X variable

b_0 : value of the intercept – the $\hat{Y}$ value when $X == 0$

SIMPLE LINEAR REGRESSION

Simple because it is bivariate – equation is the straight line between only two variables

$$\hat{Y} = b_0 + b_1 X_1$$

SIMPLE LINEAR REGRESSION

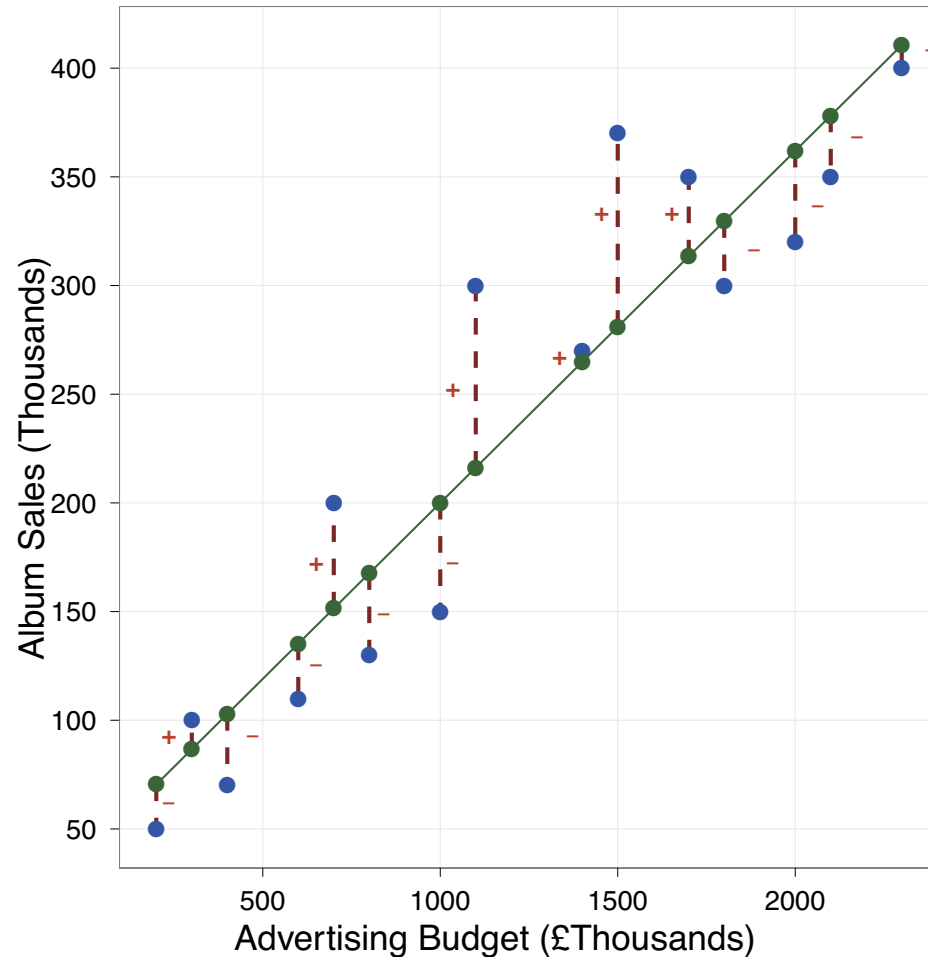


FIGURE 8.4

A scatterplot of some data with a line representing the general trend. The vertical lines (dotted) represent the differences (or residuals) between the line and the actual data

- We want to create a line that is the “best fitting” for our data...
 - Uses all the X and Y coordinates to create an equation (e.g. $Y = a + bX (+e)$) for a line that comes as close as possible to the actual coordinates (X, Y)
 - Since the regression line (in green) does not include every actual data point (in blue), the line doesn't equate to Y, it actually equates to a predicted Y value...
 - $\hat{Y} = a + bX$
 - $\hat{Y} = b_0 + b_1X_1$

SIMPLE LINEAR (ORDINARY LEAST SQUARES) REGRESSION

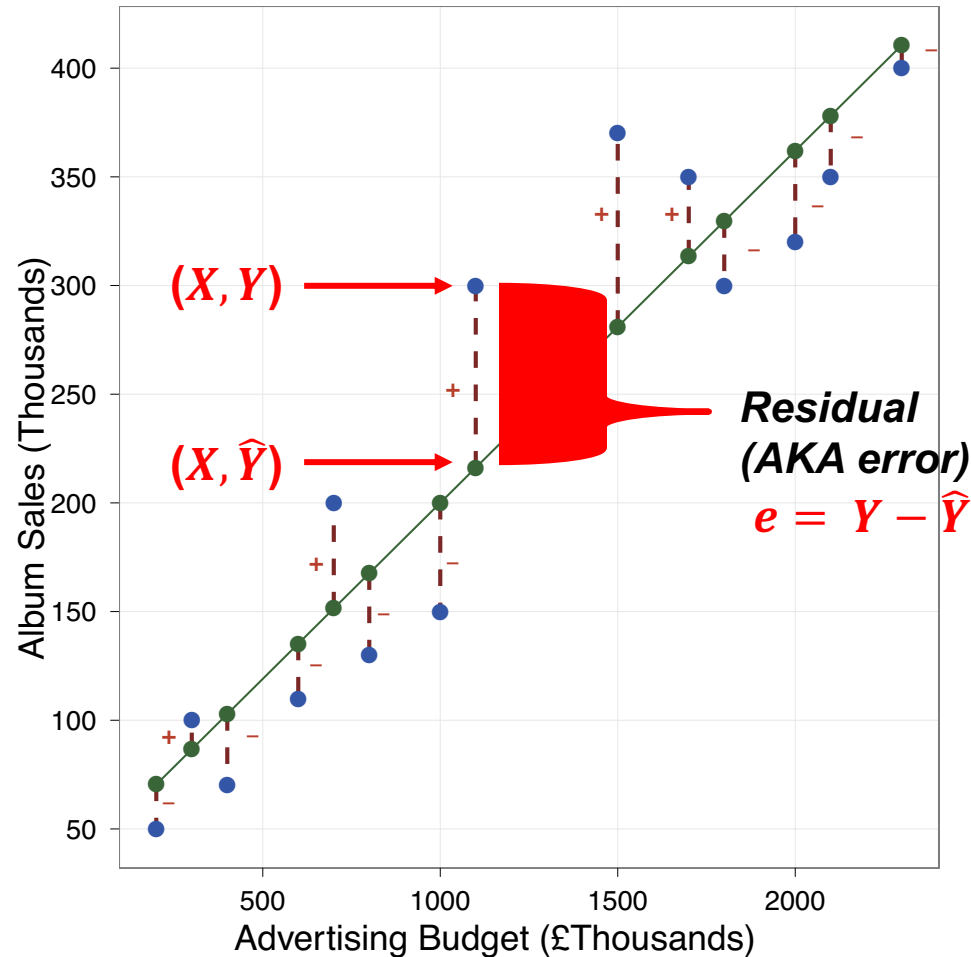


FIGURE 8.4

A scatterplot of some data with a line representing the general trend. The vertical lines (dotted) represent the differences (or residuals) between the line and the actual data

$$\hat{Y} = b_0 + b_1 X_1$$

$$(\hat{Y} = a + bX)$$
$$(\hat{Y}_i = b_0 + b_1 X_i)$$

- We want regression line that minimizes the sum of all the residuals but b/c this equals zero ($\sum(Y - \hat{Y}) = 0$), we square the residuals $\sum(Y - \hat{Y})^2$
- Linear (e.g. **ordinary**) line that minimizes (e.g. **least**) the squared residuals (**squares**)

SIMPLE LINEAR OLS REGRESSION

b_0

- Y intercept (value of Y when $X = 0$)
- Point at which the regression line crosses the Y-axis

b_1

- Regression coefficient for the predictor
 - *Amount of change in Y for each one-unit change in X*
- Slope of the regression line
- Direction/Strength of Relationship

$$\hat{Y} = b_0 + b_1 X_1$$

SIMPLE LINEAR OLS REGRESSION

$$b_1 = \frac{cov_{XY}}{var_X} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2} = \frac{\frac{\sum (X - \bar{X})(Y - \bar{Y})}{N-1}}{\frac{\sum (X - \bar{X})^2}{N-1}}$$

$$b_0 = \bar{Y} - b_1 \bar{X}_1$$

SIMPLE LINEAR OLS REGRESSION

b_0 intercept and b_1 slope are related to the covariance between X and Y (e.g. correlation)...

SIMPLE LINEAR REGRESSION (RESEARCH QUESTION)

Is variation in X related to variation in Y ?

- What is the effect of variation in X on variation in Y ?

RESEARCH PROBLEM (CONTINUED)

With correlation, and with simple linear regression, we are concerned about relationship between X and Y

- Correlation: relationship between X and Y
- Regression: how X predicts Y

However, if we want to get closer causation (e.g. cause and effect, “ X causes Y ”) in the social sciences, we must meet 3 criteria:

1. X must precede Y
2. Establish an association between X and Y
3. Ensure that the relationship is absent of spuriousness (no 3rd, 4th, or k^{th} variable can account for the relationship between X and Y)

RESEARCH PROBLEM (CONTINUED)

We must include several other X variables (additional independent variables known as control variables) to ensure that they can't remove or diminish the relationship between X and Y

By including multiple control variables, we are able to

1. measure multiple, competing variables that represent different theories or hypotheses... and
2. control for spuriousness (e.g. some unexplained third variable that could explain the relationship between X and Y)

THE MULTIPLE OLS REGRESSION MODEL

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \dots b_kX_k$$

This is one (full) multiple/OLS regression model, with k predictor/control variables

$\hat{Y}$: predicted value of the outcome variable

b_i : slope associated with i^{th} X variable – for every 1-unit change in X_i , there is an b_i unit change in $\hat{Y}$

X_i : value of i^{th} X variable... all the way up to the k^{th} variables

b_0 : value of the intercept – the $\hat{Y}$ value when $X_i == 0$

LOGIC OF MULTIPLE OLS REGRESSION

For each X variable:

- Examines how much each regression coefficient/slope (b) differs from 0.
 - Compares b to expected b in population (e.g. $b = 0$)
 - Calculates a t-statistic for each b : *Variability/difference between b in sample and expected b in population, divided by variability within the estimate of b .*
 - If b extremely different from 0, we reject H_0 that $b = 0$, and this X variable significantly predicts Y.

Overall, for (the combination of all X variables in) the regression model:

- Examines how much the regression model differs from one with no predictors (or where the slopes for each X is 0)
 - Compares "null" model (with no X variables/predictors, only the intercept) to your full model (with your X variables/predictors)
 - Calculates an F-ratio for the entire regression model: *Variability in Y explained by Xs in regression model, divided by variability in Y not explained by Xs (by intercept alone).*
 - If extremely different from 0, we reject the H_0 that $F = 0$ (fit of intercept-only null model is equal to fit of the full model). Therefore, for this model, the combination of X variables significantly predicts Y better than the null model.

MULTIPLE OLS REGRESSION (RESEARCH QUESTION)

Is variation in X related to variation in Y , in the presence of control variables?

- What is the effect of variation in X on variation in Y , while controlling for other variables?

MULTIPLE OLS REGRESSION (VARIABLE TYPES)

IVs: any level of measurement

DV: interval-ratio (e.g. continuous)

MULTIPLE OLS REGRESSION (ASSUMPTIONS)

Dependent variable is interval-ratio

Independence of observations

- Groups are not related or dependent upon each other. Case can't be in more than one group. No ties between observations
 - Examine data collection strategy ensure random sampling

Adequate sample size

- Sample large enough to have cases at various values of dependent variable for various values of each independent variable
 - $N \geq 50 + 8k$; k is # of independent variables (modified from Green 1991, as cited in Tabachnick and Fidel 2006)

Normality, Linearity, and Homoscedasticity (Homogeneity of Variance)

- Ensure that the variables (and their error terms) are normally-distributed, that they are linearly related, and that variances at levels of the dependent variable are the same for various levels of the independent variables
 - Examine residuals plot (plot of differences between actual Y values and predicted Y values), see that the residuals (points) fall equally on both sides of the predicted Y line.

Absence of Multicollinearity and Singularity

- **Multicollinearity**: two or more IVs correlate too highly with one another
 - Examine correlation matrix to see if any of the independent variables correlate at $r \approx .90$ or higher
- **Singularity**: two or more IVs represent the exact same thing
 - Examine independent variables to make sure not doubly-including variables, or are using subscales along with their aggregates (e.g. using 3 IVs that are subscales, then including another IV that is the full scale).

Absence of Outliers

- To ensure that regression line is not pulled/affected by extreme scores
 - Examine scatterplots of each independent variable with the dependent variable and identify outliers. (Can identify case and see if it significantly different from the mean/predicted Y line – using t-test or Howell's ANOVA distance - but need extreme justification to exclude cases... otherwise it looks like cherry-picking). Just explain outliers in data/method and writeup.

MULTIPLE OLS REGRESSION (ASSUMPTIONS)

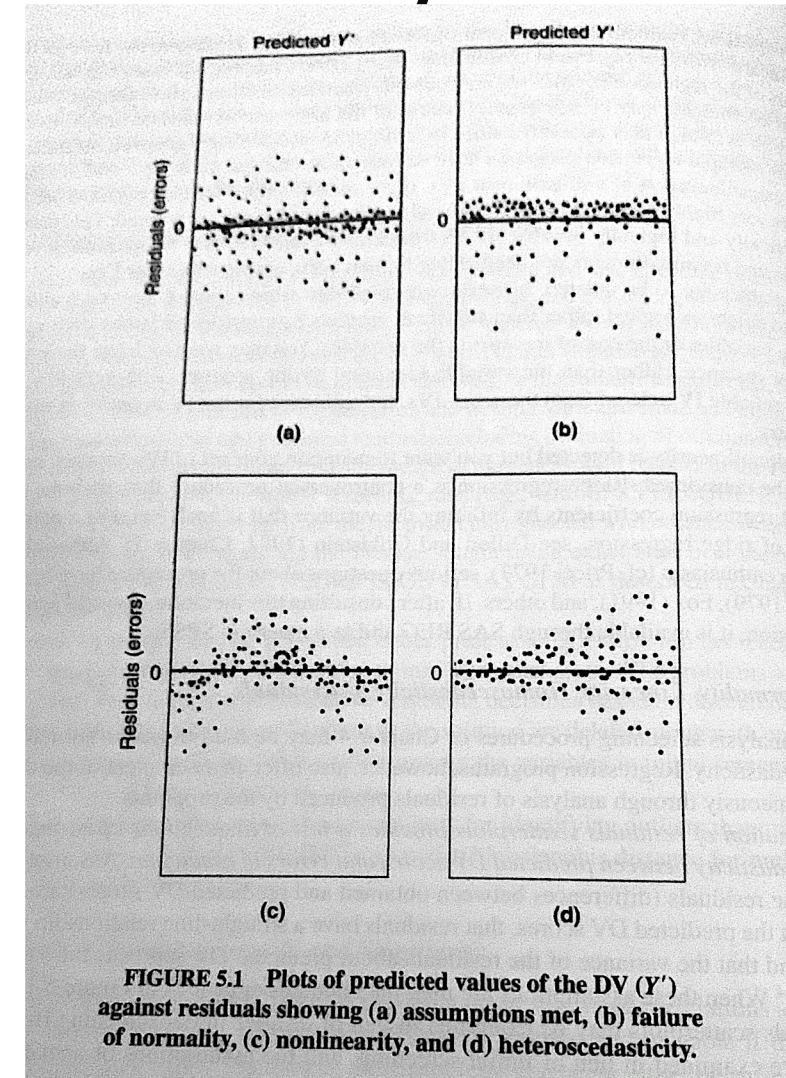
Absence of Multicollinearity

	x_1	x_2	x_3	x_6	x_7	x_8	x_9	x_{10}
x_2	0.945							
x_3	0.989	0.964						
x_6	0.659	0.772	0.653					
x_7	-0.781	-0.643	-0.746	-0.301				
x_8	0.855	0.797	0.864	0.425	-0.663			
x_9	0.801	0.718	0.788	0.316	-0.668	0.885		
x_{10}	0.946	0.883	0.943	0.521	-0.718	0.948	0.902	
x_{11}	0.835	0.727	0.801	0.417	-0.855	0.686	0.651	0.772

MULTIPLE OLS REGRESSION (ASSUMPTIONS)

Linearity, Normality, Homoscedasticity

- Residuals plot ($e = Y - \hat{Y}$)
- **Linearity:** Variables are linearly related
 - Nonlinearity is indicated when residuals exhibit non- or curvilinearity about the zero/predicted Y line.
- **Normality:** Error terms are normally distributed
 - Non-normality is indicated when most residuals fall either above or below the zero/predicted Y line. There is a hard stop indicated by the residuals
- **Homoscedasticity:** Variability in values of one variable is roughly the same at all values of the other variables
 - Heteroscedasticity indicated when residuals "fan out" like the bell of a trumpet
 - Heteroscedasticity is not fatal to analysis of ungrouped data, but should be recognized/accounted for.
 - Violations can sometimes be dealt with by transforming the DV (squared, square root, natural logarithm, etc), or using more stringent alpha.



MULTIPLE OLS REGRESSION (HYPOTHESES)

Null hypothesis (H_0)

- **For each X:** In the presence of control variables, there is no relationship between **an** X variable and the Y variable, in the population. Slope representing the relationship is zero.
- **For the model:** All of the slopes for the X variables in the model are 0.
 - $H_0: \beta_i = 0$
 - $H_0: \beta_1 = \beta_2 = \beta_3 = \dots \beta_k = 0$

Research hypothesis (H_1)

- **For each X:** In the presence of control variables, there is a relationship between **an** X variable and the Y variable, in the population. Slope representing relationship is not zero.
- **For the model:** All of the slopes for the X variables in the model are not 0.
 - $H_1: \beta_i \neq 0$
 - $H_1: \beta_1 \neq 0; \beta_2 \neq 0; \beta_3 \neq 0; \dots \beta_k \neq 0$

HYPOTHESIS TESTING IN MULTIPLE OLS REGRESSION MODELS

For each X variable:

- Examines how much each regression coefficient/slope (b) differs from 0 (e.g. no relationship between X and Y).
- Calculates a t-statistic for each:

$$t = \frac{b_i}{se_{b_i}} = \frac{b_i - 0}{\frac{\sqrt{\frac{Res\ SS}{N-2}}}{SD_X \sqrt{N-1}}} = \frac{b_i - 0}{\frac{\sqrt{\frac{\sum(Y - \hat{Y})^2}{N-2}}}{SD_X \sqrt{N-1}}} = \frac{b_i - 0}{\sqrt{\frac{\sum(Y - \hat{Y})^2}{N-k}}}{\sqrt{\sum(X - \bar{X})^2}}$$

$$df = N - k - 1$$

(for one variable, this equates to $N - 2$)

If $|t_{\text{obtained}}| \geq |t_{\text{critical}}|$, then...

- Relationship between X and Y is so extremely different from 0 (no relationship) that we can't blame it on sampling error, therefore... H_0 is probably not true, so...
- t_{obtained} is in rejection region
- Reject H_0 (that $b = 0$)
- $p \leq \alpha$
- Statistically significant relationship between X and Y (X significantly predicts Y)

HYPOTHESIS TESTING IN MULTIPLE OLS REGRESSION MODELS

Overall, for the regression model:

- Examines how much variability in Y is explained by variation in the combination of X variables in the regression model, compared to the model without predictors
- Compares "null" model (with no X variables/predictors, only the intercept) to your full model (with your X variables/predictors)
- Calculates an F-ratio for the comparison:

$$F = \frac{\frac{Reg\ SS}{k-1}}{\frac{Res\ SS}{N-k}} = \frac{\frac{\sum(\hat{Y}-Y)^2}{k-1}}{\frac{\sum(Y-\hat{Y})^2}{N-k}} = \frac{(N-k-1)R^2}{k(1-R^2)}$$

$$df_{Reg} = k - 1$$
$$df_{Res} = N - k$$

If $F_{obtained} \geq F_{critical}$, then...

- The difference between the full model and the null model (a model with no predictors, only any intercept) is so extremely different from 0 (no difference) that we can't blame it on sampling error, therefore... H_0 is probably not true, so...
- $F_{obtained}$ is in rejection region
- Reject H_0 (that $F = 0$)
- $p \leq \alpha$
- Statistically significant difference between Null Model (intercept only) and the current full model with all the predictors (the combination of X variables in the full model significantly predicts Y better than the null model)

MODEL FIT: HOW GOOD IS THE MODEL AT EXPLAINING/PREDICTING Y?

R^2

- The proportion of variance in Y accounted for by the regression model.
- Convert proportion to percent, then is percent of variance in Y explained by variance in the combination of X variables in the regression model

REPORTING MULTIPLE OLS REGRESSION

Report

- The test used
 - The variables used in the analysis
 - The variables that are significantly associated with Y
 - For each significant X , its unit-change effect on Y
 - List all variables unrelated (e.g. not significantly related) to Y
 - Report omnibus model comparison and model fit statistics
-
- “The OLS regression model demonstrates the effect of $X_1, \dots, X_k$ on Y . In this model, $(X_1, \dots, X_k)$, are significantly related to Y . First, for every 1-unit change in X_i , there is an associated b_i -unit decrease/increase in Y . (...)”
 - The overall model is/is not a significant improvement in predicting variability in Y relative to the null model: $F(\underline{?}, \underline{?}) = \underline{?}, p \underline{?} .05$.
 - The model fit statistic, $R^2 = \underline{?}$, indicates that the combination of predictors $(X_1, \dots, X_k)$ explains $\underline{?}$ Percent of the variation in Y .

EXAMPLE FROM SPSS OUTPUT: PREDICTING OVERALL COURSE RATINGS

Variable	Variable Description	Values				
		1	2	3	4	5
OVERALL	<i>overall quality of lectures</i>	bad				exceptional
TEACH	<i>teaching skills of instructor</i>	bad				exceptional
EXAM	<i>quality of tests and exams</i>	bad				exceptional
KNOWLEDGE	<i>instructor's perceived knowledge of subject matter</i>	bad				exceptional
GRADE	<i>students expected grade in course</i>	F	D	C	B	A
ENROLL	<i>enrollment of the course</i>	bad				exceptional

EXAMPLE FROM SPSS OUTPUT: PREDICTING OVERALL COURSE RATINGS

Section 15.1 Multiple Linear Regression 499

Descriptive Statistics

	Mean	Std. Deviation	N
OVERALL	3.550	.614	50
TEACH	3.664	.532	50
EXAM	3.808	.493	50
KNOWLEDG	4.176	.408	50
GRADE	3.486	.351	50
ENROLL	88.00	145.06	50

Correlations

	OVERALL	TEACH	EXAM	KNOWLEDG	GRADE	ENROLL
OVERALL	Pearson Correlation Sig. (2-tailed) N	1.000 .804** 50	.596** .000 50	.682** .000 50	.301* .034 50	-.240 .094 50
TEACH	Pearson Correlation Sig. (2-tailed) N	.804** .000 50	1.000 .000 50	.720** .000 50	.526** .001 50	.469** .001 50
EXAM	Pearson Correlation Sig. (2-tailed) N	.596** .000 50	.720** .000 50	1.000 .001 50	.451** .000 50	.610** .000 50
KNOWLEDG	Pearson Correlation Sig. (2-tailed) N	.682** .000 50	.526** .000 50	.451** .001 50	1.000 .118 50	.224 .376 50
GRADE	Pearson Correlation Sig. (2-tailed) N	.301* .034 50	.469** .001 50	.610** .000 50	.224 .118 50	1.000 -.337* .017 50
ENROLL	Pearson Correlation Sig. (2-tailed) N	-.240 .094 50	-.451** .001 50	-.558** .000 50	-.128 .376 50	1.000 .017 50

** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.869 ^a	.755	.728	.320

^a Predictors: (Constant), ENROLL, KNOWLEDGE, GRADE, TEACH, EXAM

(continued)

Exhibit 15.1 Correlations

500 Chapter 15 Multiple Regression

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	13.934	5	2.787	27.184	.000 ^b
	Residual	4.511	44	.103		
	Total	18.445	49			

^a Predictors: (Constant), ENROLL, KNOWLEDG, GRADE, TEACH, EXAM
^b Dependent Variable: OVERALL

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	Constant	-1.195	.631		-1.893	.065		
	TEACH	.763	.133	.662	5.742	.000	.418	2.391
	EXAM	.132	.163	.106	.811	.422	.325	3.081
	KNOWLEDG	.489	.137	.325	3.581	.001	.675	1.482
	GRADE	-.184	.165	-.105	-1.114	.271	.620	1.614
	ENROLL	5.255E-04	.000	.124	1.347	.185	.653	1.530

^a Dependent Variable: OVERALL

Exhibit 15.1 (continued)

Thus, we can write

$$\hat{Y} = -1.195 + 0.763\text{Teach} + 0.132\text{Exam} + 0.489\text{Knowledge} - 0.184\text{Grade} + 0.0005\text{Enroll}$$

This is the standard regression equation and can be used to obtain a predicted value of (Overall) for any specific set of values of the predictors. For example, for Course 1, substituting the data from Table 15.1, we have

$$\hat{Y} = -1.195 + 0.763(3.8) + 0.132(3.8) + 0.489(4.5) - 0.184(3.5) + 0.0005(21) = 3.773^3$$

Because Course 1 actually obtained a mean Overall lecture rating of 3.4, our residual error of prediction, for that course is $e = Y - \hat{Y} = 3.4 - 3.773 = -0.373$. For Course 1, the course with the extreme observations,

$$\hat{Y} = -1.195 + 0.763(2.2) + 0.132(1.9) + 0.489(3.9) - 0.184(2.8) + 0.0005(80) = 2.546$$

In this case, the actual Overall score was 2.6, for a residual of $2.6 - 2.546 = 0.054$. So the Course with extreme predictors was good. That is not to say, however, that we should be satisfied that extreme values are not a problem in this example. It may well be that observation pulled the regression surface toward itself and, thus, fits well as a result. From our regression equation, we can see that if all other variables were held constant, the predicted value of Y would be high.

EXAMPLE FROM SPSS OUTPUT: PREDICTING OVERALL COURSE RATINGS

Multicollinearity:

		Correlations					
		OVERALL	TEACH	EXAM	KNOWLEDG	GRADE	ENROLL
OVERALL	Pearson Correlation	1.000	.804**	.596**	.682**	.301*	-.240
	Sig. (2-tailed)	.	.000	.000	.000	.034	.094
	N	50	50	50	50	50	50
TEACH	Pearson Correlation	.804**	1.000	.720**	.526**	.469**	-.451**
	Sig. (2-tailed)	.000	.	.000	.000	.001	.001
	N	50	50	50	50	50	50
EXAM	Pearson Correlation	.596**	.720**	1.000	.451**	.610**	-.558**
	Sig. (2-tailed)	.000	.000	.	.001	.000	.000
	N	50	50	50	50	50	50
KNOWLEDG	Pearson Correlation	.682**	.526**	.451**	1.000	.224	-.128
	Sig. (2-tailed)	.000	.000	.001	.	.118	.376
	N	50	50	50	50	50	50
GRADE	Pearson Correlation	.301*	.469**	.610**	.224	1.000	-.337*
	Sig. (2-tailed)	.034	.001	.000	.118	.	.017
	N	50	50	50	50	50	50
ENROLL	Pearson Correlation	-.240	-.451**	-.558**	-.128	-.337**	1.000
	Sig. (2-tailed)	.094	.001	.000	.376	.017	.
	N	50	50	50	50	50	50

** . Correlation is significant at the 0.01 level (2-tailed).
* . Correlation is significant at the 0.05 level (2-tailed).

- No multicollinearity between IVs (TEACH, EXAM, KNOWLEDG, GRADE, ENROLL)

EXAMPLE FROM SPSS OUTPUT: PREDICTING OVERALL COURSE RATINGS

Regression equation:

Coefficients ^a								
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	Constant	-1.195	.631		-1.893	.065		
	TEACH	.763	.133	.662	5.742	.000	.418	2.391
	EXAM	.132	.163	.106	.811	.422	.325	3.081
	KNOWLEDG	.489	.137	.325	3.581	.001	.675	1.482
	GRADE	-.184	.165	-.105	-1.114	.271	.620	1.614
	ENROLL	5.255E-04	.000	.124	1.347	.185	.653	1.530

^a Dependent Variable: OVERALL

$$\hat{Y} = -1.195 + .763(TEACH) + .132(EXAM) + .489(KNOWLEDG) - .184(GRADE) + .0005(ENROLL)$$

- We can plug in various values for **TEACH** to get a estimated/predicted **OVERALL** score

EXAMPLE FROM SPSS OUTPUT: PREDICTING OVERALL COURSE RATINGS

Regression coefficient interpretation:

Coefficients ^a							
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
1	Constant	−1.195	.631	−1.893	.065		
	TEACH	.763	.133	.662	5.742	.418	2.391
	EXAM	.132	.163	.106	.811	.325	3.081
	KNOWLEDG	.489	.137	.325	3.581	.675	1.482
	GRADE	−.184	.165	−.105	−1.114	.620	1.614
	ENROLL	5.255E-04	.000	.124	1.347	.653	1.530

^a Dependent Variable: OVERALL

- Every 1 unit increase in TEACH is associated with a .763 unit increase in OVERALL

THE RELATIVE IMPORTANCE OF PREDICTORS

After assessing which t-statistics (representing the regression coefficients, b , for each predictor, X) are significant, you might want to compare which ones have greater effects on Y .

1. Which predictors ($X_1 \dots X_k$) are significantly related to Y ?
2. How much does each *significant* predictor influence change in the value of Y ?

UNSTANDARDIZED (b) VERSUS STANDARDIZED REGRESSION COEFFICIENTS (β)

We can compare the b values...but because they might all use different scales (e.g. income ranges from \$1 to 10000000, but age ranges from 18 to 78)

- Comparing the two would be like comparing apples to oranges
- To make them comparable, we have to adjust the scale of the variables.

UNSTANDARDIZED (b) VERSUS STANDARDIZED REGRESSION COEFFICIENTS (β)

To fix, we should standardize the b values... that is, turn it into something like a Z-score.

- Sets mean at 0 and SD at 1

This converts b into a standardized coefficient, β , which would be in standardized or standard deviation units.

- Interpretation for b is:
 - for every 1-unit change in X, Y changes by b units.
- Interpretation for β is:
 - for every standard deviation unit change in X, Y changes by β standard deviation unit.

UNSTANDARDIZED (b) VERSUS STANDARDIZED REGRESSION COEFFICIENTS (β)

Standardized Regression Coefficients

$$\beta = \frac{b_i SD_i}{SD_Y}$$

UNSTANDARDIZED (b) VERSUS STANDARDIZED REGRESSION COEFFICIENTS (β)

Regression equation with Standardized Coefficients

$$\hat{Y}_Z = \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \dots \beta_k Z_k$$

EXAMPLE FROM SPSS OUTPUT: PREDICTING OVERALL COURSE RATINGS

Regression equation with standardized coefficients:

Coefficients ^a								
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	Constant	−1.195	.631		−1.893	.065		
	TEACH	.763	.133	.662	5.742	.000	.418	2.391
	EXAM	.132	.163	.106	.811	.422	.325	3.081
	KNOWLEDG	.489	.137	.325	3.581	.001	.675	1.482
	GRADE	−.184	.165	−.105	−1.114	.271	.620	1.614
	ENROLL	5.255E-04	.000	.124	1.347	.185	.653	1.530

^a Dependent Variable: OVERALL

$$\hat{Y}_Z = .662(Z_{TEACH}) + .106(Z_{EXAM}) + .325(Z_{KNOWLEDG}) - .105(Z_{GRADE}) + .124(Z_{ENROLL})$$

- We can plug in various values for Z_{TEACH} to get an estimated/predicted $Z_{OVERALL}$ score

EXAMPLE FROM SPSS OUTPUT: PREDICTING OVERALL COURSE RATINGS

Regression standardized coefficients interpretation:

Coefficients ^a							
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
1	Constant	−1.195	.631	−1.893	.065		
	TEACH	.763	.133	.662	.000	.418	2.391
	EXAM	.132	.163	.106	.422	.325	3.081
	KNOWLEDG	.489	.137	.325	.001	.675	1.482
	GRADE	−.184	.165	−.105	.271	.620	1.614
	ENROLL	5.255E-04	.000	.124	.185	.653	1.530

^a Dependent Variable: OVERALL

- Every 1 SD-unit increase in **TEACH** is associated with a **.662** SD-unit **increase** in **OVERALL**

EXAMPLE FROM SPSS OUTPUT: PREDICTING OVERALL COURSE RATINGS

How much better is this regression model compared to the null model:

ANOVA ^b						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	13.934	5	2.787	27.184	.000 ^a
	Residual	4.511	44	.103		
	Total	18.445	49			

^a Predictors: (Constant), ENROLL, KNOWLEDG, GRADE, TEACH, EXAM

^b Dependent Variable: OVERALL

$$F(5,44) = 27.184, p < .05$$

- Therefore, this regression (with ENROLL, KNOWLEDG, GRADE, TEACH, and EXAM as predictors), is significantly better than the null model.

EXAMPLE FROM SPSS OUTPUT: PREDICTING OVERALL COURSE RATINGS

How well does this model predict/explain variation in OVERALL?

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.869 ^a	.755	.728	.320

^a Predictors: (Constant), ENROLL, KNOWLEDGE, GRADE, TEACH, EXAM

$$R^2 = .755$$

- Therefore, 75.5% of the variation in OVERALL is explained by this regression model (with ENROLL, KNOWLEDG, GRADE, TEACH, and EXAM as predictors).

REPORTING MULTIPLE OLS REGRESSION

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	Constant	-1.195	.631		-1.893	.065		
	TEACH	.763	.133	.662	5.742	.000	.418	2.391
	EXAM	.132	.163	.106	.811	.422	.325	3.081
	KNOWLEDG	.489	.137	.325	3.581	.001	.675	1.482
	GRADE	-.184	.165	-.105	-1.114	.271	.620	1.614
	ENROLL	5.255E-04	.000	.124	1.347	.185	.653	1.530

^a Dependent Variable: OVERALL

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	13.934	5	2.787	27.184	.000 ^a
	Residual	4.511	44	.103		
	Total	18.445	49			

^a Predictors: (Constant), ENROLL, KNOWLEDG, GRADE, TEACH, EXAM
^b Dependent Variable: OVERALL

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.869 ^a	.755	.728	.320

^a Predictors: (Constant), ENROLL, KNOWLEDGE, GRADE, TEACH, EXAM

- “The OLS regression model shows that the effects of the teaching skills of the instructor (TEACH), the quality of the exams (EXAM), the instructor’s perceived knowledge (KNOWLEDG), the student’s expected grade in the course (GRADE), and the enrollment of the course (ENROLL) on the overall quality of lectures in a course (OVERALL). In this model, TEACH and KNOWLEDG are significantly related to OVERALL. First, for every 1-unit increase in TEACH, there is an associated .763 unit increase in OVERALL. Finally, for every 1-unit increase in KNOWLEDG, there is an associated .489 unit increase in OVERALL. No other variables are significantly related to the outcome
- The overall model is a significant improvement in predicting variability in the overall quality of lectures in a course (OVERALL) relative to the null model: $F(5,44) = 27.184, p < .05$.
- The model fit statistic, $R^2 = .755$, indicates that the combination of predictors (TEACH, EXAM, KNOWLEDG, GRADE, and ENROLL) explains 75.5 percent of the variation in the overall quality of lectures in a course (OVERALL).”

OLS REGRESSION AND THE PROBLEM OF EXTRAPOLATION

It is true that we can use our regression equation to predict/estimate values of Y

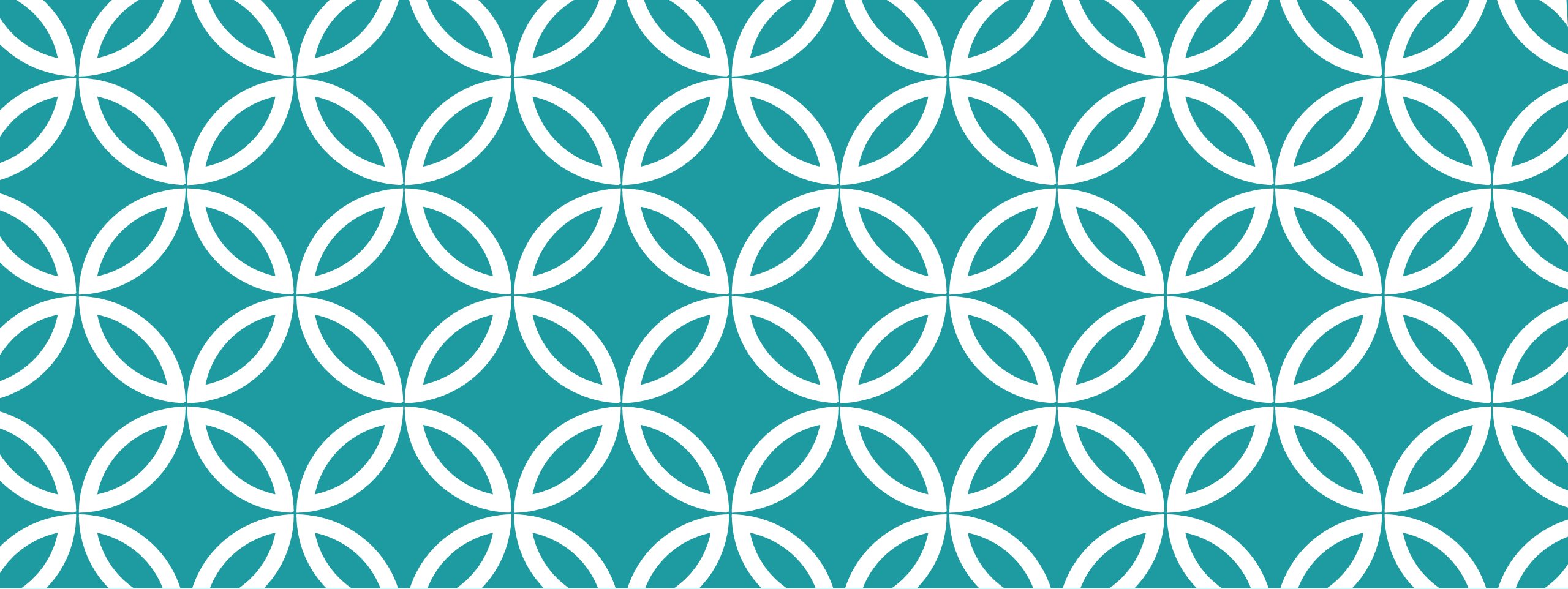
- We can plug in random values for each X to see what the predicted Y would be.

However, we can only "plug in" X values that exist within the range of X values we used for the regression.

- For example, if we used education (X), measured 0 years to 25 years, to predict income (Y), we can only plug in values for education that range between 0 and 25

We don't know if the same linear relationship between the combination of predictors and the outcome variable exists beyond the data we have available.

- Therefore, we can't predict Y values from X variables we haven't included in the model.
 - For example, plugging in 28, goes beyond the data we have... so we don't know if an X value of 28 would alter the linear relationship overall, and the magnitude of the regression slopes for other variables.



OLS REGRESSION MODELING

A NOTE ON PREDICTORS AND PARSIMONY

Parsimony is notion that a regression model includes “no more predictors than necessary” (X_1 through X_k) to explain optimal variation in Y

To select the appropriate number of predictors for your regression model

- **use theoretical/literature and practical knowledge** to figure out which predictors are most important/necessary and which are unnecessary
- consider the number of cases in data, and the suggestion: $N \geq 50 + 8k$

A NOTE ON MISSINGNESS IN OLS REGRESSION

Missingness in regression refers to whether or not an observation is missing information (a value) for any of the variables (Y or $X_1 \dots X_k$).

Regression deals with missing through listwise deletion

- If a case is missing information for either the dependent variable (Y) or any one of independent variables ($X_1 \dots X_k$), it is removed/dropped from the analysis
 - This is why you will often have fewer cases (n) in your analysis than exists in the data set.

MODELS? MODELING?

Every time we run a regression, we refer to it as a regression model.

- We are regressing Y on our predictors, X_1 through X_k

This means we will run many regression models.

This means that we can run many regression models on our data

This means we can run many different regression models regressing Y on different subsets of our predictors

MODELS? MODELING?

Approaches to Modeling

- Standard/Forced Entry
- Sequential (Nested Models/Hierarchical Regression/Blocking)
- Statistical Approaches
 - Forward
 - Backward
 - Stepwise

MODELS? MODELING?

Standard Approach/Forced Entry

- Regression model in which all predictors are included into a single model

$$\textit{Model 1: } \hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \dots b_kX_k$$

MODELS? MODELING?

Sequential (Nested Regression/Hierarchical Regression/Blocking)

- Forced Entry for a subset of predictors. Blocking predictors into sets, and modeling separate “sets”
 - Should be based on theory/literature why certain “sets” of go together or should be modeled together

$$\textit{Model 1: } \hat{Y} = b_0 + b_1X_1$$

$$\textit{Model 2: } \hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3$$

$$\textit{Model 3: } \hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \dots b_kX_k$$

- For example, **Model 1** could include the key IV whose relationship with Y you're most interested in investigating, **Model 2** could include a set of a few variables that represent an alternative theory that could also explain changes in Y (e.g. controlling for spuriousness), and **Model 3** could include all other control variables/IVs (e.g. demographic characteristics)

COMPARING SEQUENTIAL MODELS

Comparing R^2 and F for current regression model to the previous model.

- Calculate a change value for both:
 - ΔR^2 (change in R^2)
 - ΔF (change in F , sometimes also referred to as the F-increment test or F_{inc})

COMPARING SEQUENTIAL MODELS

$$\Delta R^2 = R_2^2 - R_1^2$$

- Where R_2^2 is the R^2 value for the current model, and R_1^2 is the R^2 value for the previous model.
 - It is important to note that, if your current model is your only model, then your previous model would be the null model

$$\Delta F = \frac{\frac{R_2^2 - R_1^2}{k_2 - k_1}}{\frac{1 - R_2^2}{N - k_2 - 1}}$$

- Where k_2 is the number of predictors in the current or model, and k_1 is the number of predictors value for the previous model.
 - $\Delta df_{reg} = k_2 - k_1$, $df_{res} = N - k_2 - 1$

MODELS? MODELING?

Statistical Approaches

- Rely on correlation values (zero-order, semi-partial, and partial) to select the best predictors for the model for inclusion into/exclusion from the model, then builds a full model with only the best predictors
- Researcher no longer in control of model. Theoretical knowledge no longer used to build model, just computer selecting the model that yields the most significantly related variables and best model fit.

MODELS? MODELING?

Statistical Approaches

- Backward
 - Starts as forced entry, enters all variables in model
 - Then removes variables, one-by-one, until only “strongly” significant variables are left
 - Relies on pout value (if the significance of the variable’s relationship with Y exceeds the pout value, it will be thrown out of the model)
 - Selects variable with largest pout value, removes, then re-runs analysis, and removes the variable with the next highest pout value.

MODELS? MODELING?

Statistical Approaches

- Forward
 - Starts with nothing (no variables) in the model
 - Then adds variables, one-by-one, until only “strongly” significant variables are left
 - Relies on pin value (if the significance of the variable’s relationship with Y is smaller than the pin value, it will be placed into the model)
 - Selects variable with best/smallest pin value, adds, then re-runs analysis, and adds the variable with the next best/smallest pin value.

MODELS? MODELING?

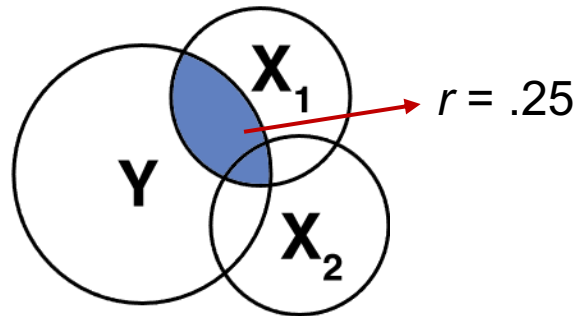
Statistical Approaches

- Stepwise
 - Starts with nothing (no variables) in the model
 - Then it goes forward: adds variables, one-by-one, until only “strongly” significant variables are left
 - Relies on pin value (if the significance of the variable’s relationship with Y is smaller than the pin value, it will be placed into the model)
 - Selects variable with best/smallest pin value, adds, then re-runs analysis, and adds the variable with the next best/smallest pin value.
 - Then, it turns around and goes backward: evaluates all variables in the model by removing, one-by-one, until only “strongly” significant variables are left
 - Relies on pout value (if the significance of the variable’s relationship with Y exceeds the pout value, it will be thrown out of the model)
 - Selects variable with largest pout value, removes, then re-runs analysis, and removes the variable with the next highest pout value.

A NOTE ON “BEST” PREDICTORS

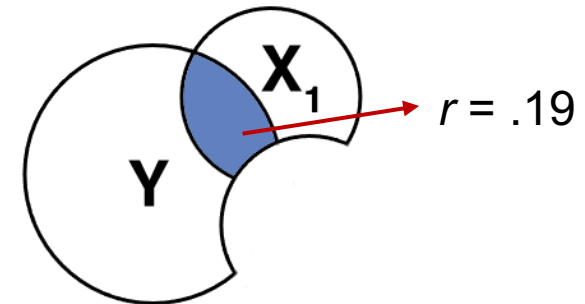
Zero-Order Correlation

- Pearson correlation coefficient
- Overlap between X_i and Y



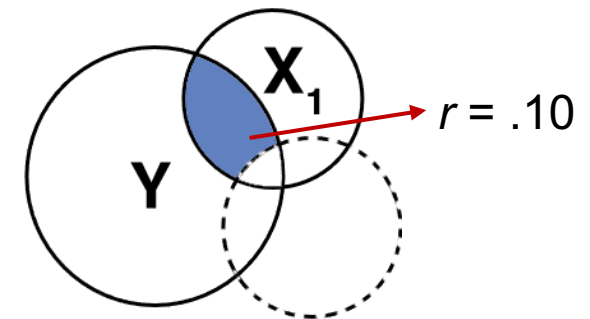
Partial Correlation

- Similar to semi-partial
- Unique overlap between X_i and Y , but removing (from Y) all that can be accounted for by other variables
- Partial correlations are bigger than semi-partial correlations because there is less Y left



Semi-Partial Correlation (Part)

- Unique overlap between X_i and Y (amount of variance that each variable accounts for), without removing (from Y) all that can be accounted for by other variables



TRANSFORMING VARIABLES FOR REGRESSION

Sometimes, the way variables are measured can create issues for our regression model.

In these cases, we have to recode our variables, or transform them into a different functional form

- Squaring
- Natural Logarithm
- Dummy Codes for Categorical Variables

CATEGORICAL VARIABLES AS DUMMY CODES

When you have categorical variables in regression, they must be assigned numerical values:

Variable	Variable Description	Values					
		1	2	3	4	5	6
Race/Ethnicity	respondent's self-identified race/ethnicity	White	Black	AAPI	Latinx	Native	Other

ID #	Race/Ethnicity
1	4
2	1
3	1
4	6
5	2
6	1
7	4
8	3
9	5

Yet, these numbers are meaningless for categorical variables, specifically nominal variables (and only sort of meaningful for ordinal variables)

- Interpretation of the race variable's slope (b) in regression would be: "a one-unit increase in race is associated with a b -unit change in Y "
- What is a "one unit increase" in race? It's not interpretable, since higher levels of race/ethnicity (?!?!) doesn't make sense.

CATEGORICAL VARIABLES AS DUMMY CODES

For categorical variables, observations either belong to a category/attribute, or they do not.

By this logic, we could create **codes** for each attribute of a variable, such that we could categorize each case as belonging/not belonging to each unique attribute.

CATEGORICAL VARIABLES AS DUMMY CODES

Dummy codes (for one variable) are new variables that each represent one of the categories/attributes of the categorical variable being recoded.

Each dummy variable is a dichotomous (binary) measure, where 0 represents not belonging to the category/attribute represented by the current dummy variable, and 1 represents belonging to that category/attribute.

The number of dummy codes required is equal to the number of attributes (k), minus 1, or...

$$D = k - 1$$

We have $k - 1$ variables because a case with all 0s represents the category/attribute that doesn't have a dummy code.

The category/attribute that doesn't have an associated dummy code is known as our reference category.

Each case will have a value for each dummy code. They will have mostly 0s and a 1 that aligns with their category.

CATEGORICAL VARIABLES AS DUMMY CODES

Therefore, with our race variable, with 6 attributes, we would need 5 dummy codes. And this...

...Becomes this

ID #	Race/Ethnicity
1	4
2	1
3	1
4	6
5	2
6	1
7	4
8	3
9	5

ID #	White	Black	AAPI	Latinx	Native
1	0	0	0	1	0
2	1	0	0	0	0
3	1	0	0	0	0
4	0	0	0	0	0
5	0	1	0	0	0
6	1	0	0	0	0
7	0	0	0	1	0
8	0	0	1	0	0
9	0	0	0	0	1

Variable	Variable Description	Values					
		1	2	3	4	5	6
Race/Ethnicity	respondent's self-identified race/ethnicity	White	Black	AAPI	Latinx	Native	Other

NEXT UP: ADDITIONAL MODELING & MEASUREMENT

Mediation, Moderation (Interactions) & Centering

Suppressor Variables