Part II	**Exploring Relationships Between Variables**
Chapter 7	**Scatterplots, Association, and Correlation**
Scatterplot ______ is plotted on the x-axis. ______ is plotted on the y-axis.	Shows the relationship between two quantitative variables on the same cases (individuals). Explanatory (*independent*/input) variable Response (*dependent*/output) variable
Once we make a scatterplot, we describe association by telling about:	**1. Form**: straight, curved, no pattern, other? **2. Direction**: + or – slope? **3. Strength**: how much scatter {how closely points follow the form} **4. Unusual Features**: outliers, clusters, subgroups?
_____ is a deliberately vague term describing the relationship between two variables. If positive then ______________	Association increases in one variable generally correspond to increases in the other.
Correlation describes the ____ and ____ of the ____ relationship between two ______ variables, without significant ______	strength direction, linear quantitative outliers.
3 conditions needed for Correlation:	1. Quantitative Variables 2. Straight Enough 3. Outlier
The correlation coefficient is found by ________________. It's value ranges from ______ , it has no ______, and is immune to changes of ________	finding the average product of the z-scores (standardized values). $$r = \frac{\sum z_x z_y}{n-1}$$ -1 to +1 units. scale or order.
Perfect correlation r = _____, occurs only when ___________.	± 1 the points lie exactly on a straight line. (you can perfectly predict one variable knowing the other)
No correlation r = ______, means that knowing one variable gives you ___________	0 no information about the other variable.
You should give the ______ and ______ of x and y along with the correlation because …	Mean Standard deviation Correlation is not a complete description of two-variable data and the its formula uses means and standard deviations in the z-scores.
Scatterplots and correlation coefficients never prove ______	 causation.
Lurking variable	A variable other than x and y that simultaneously affects both variables, accounting for the correlation between the two.
To add a categorical variable to an existing scatterplot _____	 use a different plot color or symbol for each category.
Chapter 8	**Linear Regression**

Regression to the mean	Because the correlation is always less than 1.0 in magnitude, each predicted $\hat{y}$ tends to be fewer standard deviations from its mean than its corresponding x was from its mean. ($\hat{z}_y = rz_x$)
Residual If positive If negative	Observed value – predicted value $y - \hat{y}$ Then the model makes an underestimate. Then the model makes an overestimate.
Regression line Line of best fit For standardized values For actual x and y values	The unique line that minimizes the variance of the residuals (sum of the squared residuals). $\hat{z}_y = rz_x$ $\hat{y} = b_0 + b_1x$
To calculate the regression line in real units (actual x and y values)	1. Find slope, $b_1 = \frac{rs_y}{s_x}$ 2. Find y-intercept, plug b_1 and point (x, y) [usually ($\bar{x}$, $\bar{y}$)] into $\hat{y} = b_0 + b_1x$ and solve for b_0 3. Plug in slope, b_1, and y-intercept, b_0, into $\hat{y} = b_0 + b_1x$
3 conditions needed for Linear Regression Models: /* same as correlation */	1. Quantitative Variables 2. Straight Enough – check original scatterplot & residual scatterplot 3. Outlier (clusters) –points with large residuals and/or high leverage
R^2	The square of the correlation, r, between x and y The success of the regression model in terms of the fraction of the variation of y accounted for by the model. (XX% of the variability in y is accounted for by variation in x) (differences in x explain XX% of the variability in y)
A high R^2	Does not demonstrate the appropriateness of the regression.
Looking at a ________________ is a good way to check the Straight Enough Condition. It should be _________________	a scatterplot of the residuals vs. the x-values. (appropriateness) boring: uniform scatter with no direction, shape, or outliers..
The ____ is the key to assessing how well the model fits (extracts the form).	variation in the residuals
Standard deviation of the residuals, s_e	Gives a measure of how much the points spread around the regression line.
$1 - R^2$	The fraction of the original variation left in the residuals. (The percentage of variability not explained by the regression line.)
Extrapolations	Dubious predictions of y-values based on x-values outside the range of the original data.
Chapter 9	**Regression Wisdom**
What can go wrong with regression:	1. Inferring Causation 2.Extrapolation 3.Outliers and Influential Points 4.Change in Scatterplot Pattern 5.Means (or other summaries) rather than actual data.
High leverage points With enough leverage the ____	Have x-values far from $\bar{x}$ (($\bar{x}$,$\bar{y}$) is the fulcrum) and pull more strongly on the regression line. residuals

can appear deceptively small.	
Leverage and residual produce three flavors of outliers:	1) Extreme Conformers: don’t influence model but do inflate R2 2) Large Residuals: might not influence model much but aren’t consistent with the overall form. 3) Influential Points: those that distort the model
Influential point [most menacing]	Omitting it from the data results in a very different regression model
Influential points are often difficult to detect because	They distort the model which causes their residual to be small.
The surest way to verify an outlier and its affects is to	Calculate the regression line with and without the suspect point.
A histogram of the residuals	Compliments a scatterplot of the residuals in the search for conditions, such as subsets, that may compromise the effectiveness of the regression model.
Consider comparing two or more regressions if you find	1) Points with large residuals and/or high leverage. 2) Change in Scatterplot Pattern as a result of changes over time or subsets that behave differently.
Regressions based on summaries of the data ________ Because __________	 Tend to look stronger than the regression on the original data. Summary statistics are less variable than the underlying data.
Chapter 10	**Re-expressing Data: Get It Straight!**
Re-expression	A means of altering the data to achieve the conditions/structure necessary to utilize particular summaries or models.
Several reasons to consider a re-expression:	1. Make the form of a scatterplot straighter. 2. Make the scatter in a scatterplot more consistent (not fan shaped). 3. Make the distribution of a variable (histogram) more symmetric. 4. Make the spread across different groups (boxplots) more similar.
Ladder of Powers A good starting point is _______ If all else fails ________	Orders the effects that the re-expressions have on the data 2 1 ½ 0 -½ -1 y^2 y √y log y -1/√y -1/y taking logs. try whacking the data with two logs (log x and log y).
Base 10 logs are roughly	One less than the number of digits needed to write the number.
Re-expression limitations:	1. Can’t straighten scatterplots that turn around. 2. Can’t re-express “-” data values with √ (+constant to shift > 0) 3. Minimal affect on data values far from 1-100. (-constant to shift) 4. Can’t unify multiple modes.
When discussing the accuracy or confidence of the linear regression model be sure to comment on both the __ & __	Appropriateness of the model as indicated by the residual plot Success of the model as indicated by R^2