18. Real estate redux.

a) The price of a home that is one standard deviation above the mean size would be predicted to be 0.845 standard deviations (in other words *r* standard deviations) above the mean price.

b) The price of a home that is two standard deviations below the mean size would be predicted to be 1.69 (or 2×0.845) standard deviations below the mean price.

$\#S_y = r \#S_x$

19. Another ride.

a) The duration of a coaster whose initial drop is one standard deviation below the mean drop would be predicted to be about 0.39 standard deviations (in other words, *r* standard deviations) below the mean duration.

b) The duration of a coaster whose initial drop is three standard deviation above the mean drop would be predicted to be about 1.17 (or 3×0.39) standard deviations above the mean duration.

20. More real estate.

a) According to the linear model, the price of a home is expected to increase \$61 (0.061 thousand dollars) for each additional square-foot in size.

b)

$\widehat{Price} = 47.82 + 0.061(Sqrft)$

$\widehat{Price} = 47.82 + 0.061(3000)$

$\widehat{Price} = 230.82$

According to the linear model, a 3000 square-foot home is expected to have a price of \$230,820.

c)

$\widehat{Price} = 47.82 + 0.061(Sqrft)$

$\widehat{Price} = 47.82 + 0.061(1200)$

$\widehat{Price} = 121.02$

According to the linear model, a 1200 square-foot home is expected to have a price of \$121,020. The asking price is \$121,020 - \$6000 = \$115,020. \$6000 is the (negative) residual.

21. Last ride.

a) According to the linear model, the duration of a coaster ride is expected to increase by about 0.180 seconds for each additional foot of initial drop.

b)

$\widehat{Duration} = 64.232 + 0.180(Drop)$

$\widehat{Duration} = 64.232 + 0.180(200)$

$\widehat{Duration} = 100.232$

According to the linear model, a coaster with a 200 foot initial drop is expected to last 100.232 seconds.

c)

$\widehat{Duration} = 64.232 + 0.180(Drop)$

$\widehat{Duration} = 64.232 + 0.180(175)$

$\widehat{Duration} = 95.732$

According to the linear model, a coaster with a 150 foot initial drop is expected to last 95.732 seconds. The advertised duration is shorter, at 90 seconds.
90 seconds – 95.732 seconds = – 5.732 seconds, a negative residual.

What Can Go Wrong?

- Make sure the relationship is straight enough.
- Don't fit a straight line to a nonlinear relationship.
- Don't extrapolate beyond the data—the linear model may no longer hold outside of the range of the data.
- Beware especially of extrapolating into the future!
- Beware of lurking variables—and don't assume that association is causation.
- Don't infer that *x* causes *y* just because there is a good linear model for their relationship—association is *not* causation.
- Don't even *imply causation*.

What have we learned?

- When the relationship between two quantitative variables is fairly straight, a linear model can help summarize that relationship.
  - The regression line doesn't pass through all the points, but it is the best compromise in the sense that it has the smallest sum of squared residuals.

What have we learned? (cont.)

- The correlation, *r*, tells us several things about the regression:
  - The slope of the line is based on the correlation, adjusted for the units of *x* and *y*.
  - For each SD in *x* that we are away from the *x* mean, we expect to be *r* SDs in *y* away from the *y* mean.
  - Since *r* is always between –1 and +1, each predicted *y* is fewer SDs away from its mean than the corresponding *x* was (regression to the mean).

$\#S_y = r * S_x$

What have we learned?

- The residuals also reveal how well the model works.
  - If a plot of the residuals against predicted values shows a pattern, we should re-examine the data to see why.

What have we learned? (cont.)

- The linear model makes no sense unless the **Linear Relationship Assumption** is satisfied.
- Also, we need to check the **Straight Enough Condition** and **Outlier Condition** with a scatterplot.

What have we learned? (cont.)

- Even a good regression doesn't mean we should believe that the model says more than it really does.
  - Extrapolations from from $\bar{x}$ can lead to silly and useless predications.
  - Even an *r* near +1 doesn't indicate the *x* cause *y*. Watch our for lurking variables that may affect both *x* and *y*.

Statistics Chapter 7: Review A – KEY

1. For each term or concept, give a definition/explanation in your own words and an example (Only use your book or other resources to check when you're done or if you really get stuck!)

Term/Concept	*Definition/Explanation*	*Example*
Model	Our best guess at explaining the linear relationship between two variables.	$\widehat{Cost} = 65 + 0.25 MinutesUsed$
Predicted Value	An estimate based on known data.	Use the cost model above to predict the cost of a cell phone bill when 300 minutes are used.
Conditions to use a linear model	Straight Enough Condition and Outlier Condition	If there appears to be an outlier or outliers OR if the data does not appear to be straight, we cannot use the linear model.
Residuals	The distance between actual and predicted values.	If you use the cost model above to predict the value of 100 minutes used on a cell phone plan, you would get \$90. If the actual value was \$80, then the residual would be 80 – 90 = -\$10.
Regression towards the mean	Predictors give estimates that are closer to their means than the predictor is to its own mean.	One-hit-wonders: a band has an amazing first album but future albums are mediocre at best.
Slope of a linear model (what it is and how to interpret it in context)	The slope tells us about the association of the linear relationship – if it is positive or negative. To interpret it, we use the units provided in the context of the scenario.	In the cell phone example above, the slope would be \$0.25/minute. This means the model predicts that every minute of use is associated with an increase of \$0.25.
Intercept of a linear model (what it is and how to interpret it in context)	The intercept is the starting point for our model. Sometimes it has meaning and sometimes it does not – depending on the context.	In the above cell phone example, the intercept has meaning. That is, it is the starting value of the cost for the cell phone plan predicted by the model.

Term/Concept	*Definition/Explanation*	*Example*
Causation vs. Association	Causation means an explanatory variable caused a change in a response variable. This can only be done in experiments. Association just implies that there is a link between two variables.	Causation: cigarettes cause cancer. Association: studying is associated with better grades.
Lurking variable	A lurking variable is a variable that is behinds the scenes in a relationship between two other variables. The lurking variable produces a change in the response variable and this change can be wrongly attributed to the explanatory variable.	Sunburns and ice cream sales. There is a strong association between the variables but the lurking variable is the season. In summer there are more ice cream sales and more sunburns.
Extrapolation (and its dangers)	Going outside the data to predict values.	Using gas prices from the 1950s to predict gas prices in 2010.

2. Sally created the following model using data taken in an experiment in her laboratory.

$$\widehat{Number\ of\ Bacteria} = 34,219 + 532.1 hours$$

a. Identify the explanatory and response variables she is measuring.

The explanatory variable is the number of hours and the response variable is the number of bacteria.

b. What should Sally have done with the data before finding this model?

She should have checked the Outlier Condition and Straight Enough Condition.

c. Which are true?

- This model predicts how many bacteria there are after a certain number of hours. TRUE
- This model predicts how many hours a certain number of bacteria have been alive. FALSE
- This model produces (*x,y*) points of (*number of bacteria, number of hours*) FALSE
- This model produces (*x,y*) points of (*number of hours, number of bacteria)* TRUE

d. Describe the slope and intercept of this model in context

Slope: The slope is the predicted increase in bacteria per hour, 532.1bacteria/hour.
Intercept: The intercept is the predicted amount of bacteria at the beginning of the experiment, 34,219.

e. Calculate the residual for this data point (2, 37642).

$\widehat{Number\ of\ Bacteria} = 34219 + 532.1(2) = 35283.2$. $Residual = 37642 - 35283.2 = 2358.2$ bacteria.

f. Describe what a positive residual means in this context

A positive residual means that the model predicts fewer bacteria than there actually were at some time in the experiment.

3. Data shows a strong positive association between ice cream sales and the number of cases of sunburn in a small town.

   a. Describe what this means in language a 10 year old would understand.

      One possible explanation: During the summer, when the sun is out and it's warm, people enjoy eating ice cream more than they do at other times of the year. People also get more sunburned in the summer than at other times of the year. So, we notice that as ice cream sales go up, it's usually summer, and so people are more likely to get sunburned.

   b. Does this mean that eating ice cream causes sunburn?

      No. Just an association. This is not an experiment. There is no cause-and-effect relationship.

   c. What might be a lurking variable in this situation? Explain.

      People tend to stay outdoors more in the summer than at other times of the year. The sun is out. People tend to get sunburned more often.

4. Here is some fictional data for a set of cars showing the number of oil changes and the annual cost of repairs for each car. (from http://illuminations.nctm.org/LessonDetail.aspx?ID=L298

Oil changes	3	5	2	3	1	4	6	4	3	2	0	10	7
Repair costs ($)	300	300	500	400	700	400	100	250	450	650	600	0	150

   a. Justify the use of a linear model for this data.

      Using a scatterplot, the data appear to fulfill the Outlier Condition and the Straight Enough Condition.

   b. Describe the association between the number of oil changes and repair costs for these cars.

      As the number of oil changes a car receives increases, the annual cost of repairs tends to decrease.

   c. Create a model to predict the annual repair costs for a car based on the number of oil changes in that year.

$$\widehat{Costs} = 650.27 - 73.07Changes$$

   d. Describe the slope and intercept of your model in context.

      slope: For every oil change, the model predicts a decrease in annual repair costs of $73.07.

      intercept: A car that receives no oil changes is predicted to have $650.27 in annual repair costs.

   e. Predict the annual repair costs for a car that has 8 oil changes that year.

$$\widehat{Costs} = 650.27 - 73.07(8) = \$65.71$$

   f. Predict the number of oil changes for a car that had $85 in repairs that year.

$$\widehat{Changes} = 8.0674 - 0.0114Costs$$

$$\widehat{Changes} = 8.0674 - 0.0114(85) = 7.1 \text{ oil changes}$$