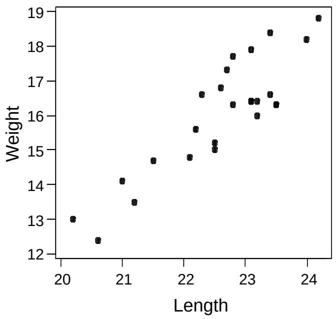
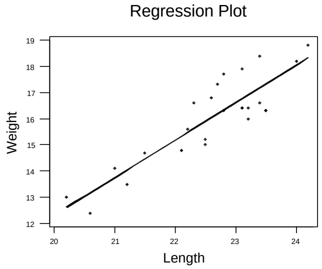
Scatter Diagrams

- Scatter diagrams are used to demonstrate **correlation** between two quantitative variables.
- Often, this correlation is **linear**.
- This means that a straight line model can be developed.



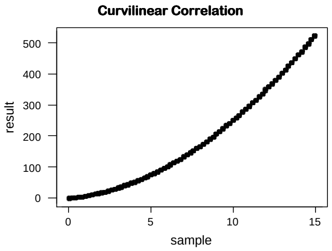
Correlation Classifications

- Correlation can be classified into three bas categories
- Linear
- Nonlinear
- No correlation



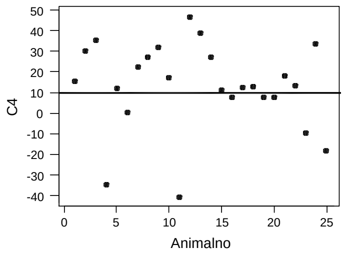
Correlation Classifications

- Two variables may be correlated but not through a linear model.
- This type of model is called non-linear
- The model might be one of a curve.



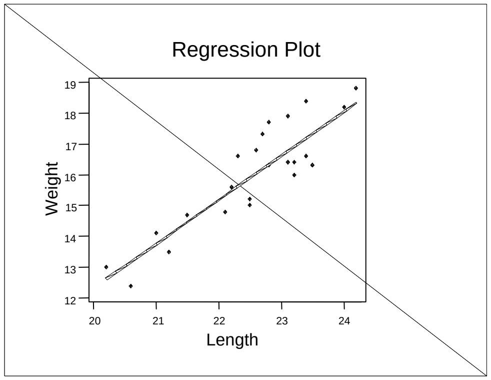
Correlation Classifications

- Two quantitative variables may not be correlated at all



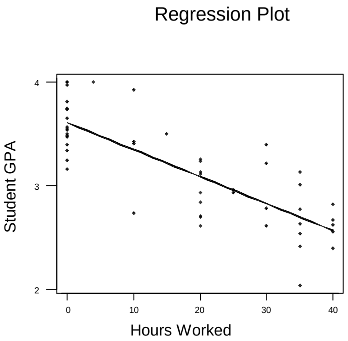
Linear Correlation

- Variables that are correlated through a linear relationship can display either positive or negative correlation
- Positively correlated variables vary directly.



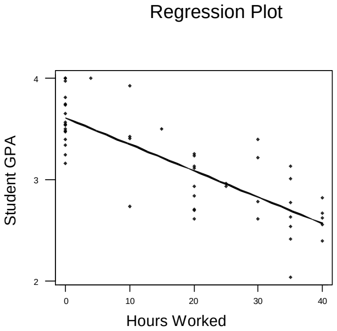
Linear Correlation

- Negatively correlated variables vary as opposites
- As the value of one variable increases the other decreases



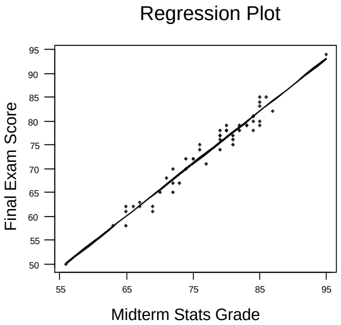
Strength of Correlation

- Correlation may be strong, moderate, or weak.
- You can estimate the strength be observing the variation of the points around the line
- Large variation is **weak** correlation



Strength of Correlation

- When the data is distributed quite close to the line the correlation is said to be **strong**
- The correlation type is independent of the strength.



The Correlation Coefficient

- The strength of a **linear relationship** is measured by the correlation coefficient
- The sample correlation coefficient is given the symbol “**r**”
- The population correlation coefficient has the symbol “ρ”.

Interpreting r

- The sign of the correlation coefficient tells us the direction of the linear relationship
  - ⌦ If r is negative (<0) the correlation is negative. The line **slopes** down
  - ⌦ If r is positive (> 0) the correlation is positive. The line **slopes** up

Interpreting r

- The size (magnitude) of the correlation coefficient tells us the strength of a **linear** relationship
  - ⌦ If $| r | \geq 0.90$ implies a strong linear association
  - ⌦ For $0.65 < | r | < 0.90$ implies a moderate linear association
  - ⌦ For $| r | \leq 0.65$ this is a weak linear association

Cautions

- The correlation coefficient only gives us an indication about the strength of a **linear relationship**.
- Two variables may have a strong curvilinear relationship, but they could have a “weak” value for r

Fundamental Rule of Correlation

- Correlation **<u>DOES NOT</u>** imply causation
  - Just because two variables are highly correlated does not mean that the explanatory variable "**<u>causes</u>**" the response

- Recall the discussion about the correlation between sexual assaults and ice cream cone sales

Setting

- A chemical engineer would like to determine if a relationship exists between the extrusion temperature and the strength of a certain formulation of plastic. She oversees the production of 15 batches of plastic at various temperatures and records the strength results.

The Study Variables

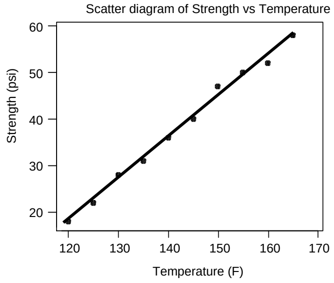
- The two variables of interest in this study are the strength of the plastic and the extrusion temperature.
- The independent variable is extrusion temp. This is the variable over which the experimenter has control. She can set this at whatever level she sees as appropriate.
- The response variable is strength. The value of "strength" is thought to be "dependent on" temperature.

The Experimental Data

Temp	120	125	130	135	140
Str	18	22	28	31	36
Temp	145	150	155	160	165
Str	40	47	50	52	58

The Scatter Plot

- The scatter diagram for the temperature versus strength data allows us to deduce the nature of the relationship between these two variables



What can we conclude simply from the scatter diagram?

Conclusions by Inspection

- Does there appear to be a relationship between the study variables?
- Classify the relationship as: Linear, curvilinear, no relationship
- Classify the correlation as positive, negative, or no correlation
- Classify the strength of the correlation as strong, moderate, weak, or none

Computing r

$$r = \frac{1}{\boxed{n-1}} \Sigma \left\{ \left(\boxed{\frac{x-\bar{x}}{s_x}} \right) \left(\boxed{\frac{y-\bar{y}}{s_y}} \right) \right\}$$

df

z-scores
for x data

z-scores
for y data

Computing r

$$r = \frac{1}{n-1} \Sigma \left[\left(Z_x \right) \left(Z_y \right) \right]$$

Computing r - Example

See example handout for the plastic strength versus extrusion temperature setting

Classifying the strength of linear correlation

•The strength of a linear correlation between the response and the explanatory variable can be assigned based on r

➢These classifications are discipline dependent

Classifying the strength of linear correlation

For this class the following criteria are adopted:

➢ If $|r| \geq 0.90$ then the correlation is strong

➢ If $|r| \leq 0.65$ then the correlation is weak

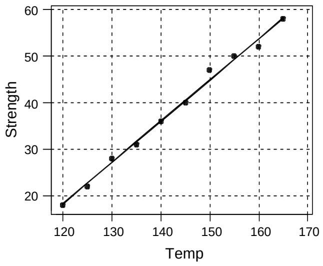
➢ If $0.65 < |r| < 0.90$ then the correlation is moderate

Scatter Diagrams and Statistical Modeling and Regression

- We’ve already seen that the best graphic for illustrating the relation between two quantitative variables is a scatter diagram. We’d like to take this concept a step farther and, actually develop a mathematical model for the relationship between two quantitative variables

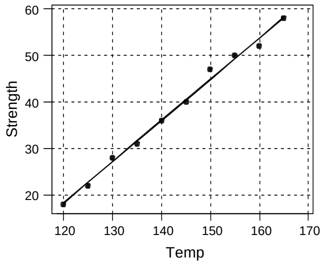
The Line of Best Fit Plot

- Since the data appears to be linearly related we can find a straight line model that fits the data better than all other possible straight line models.
- This is the Line of Best Fit (LOBF)



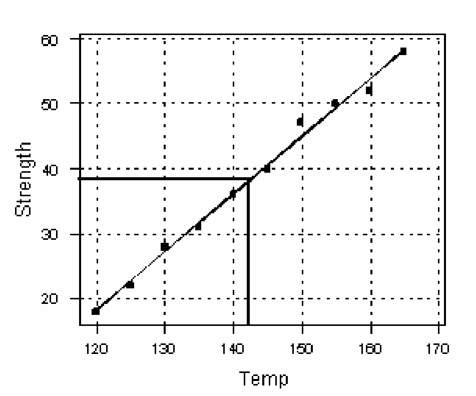
Using the Line of Best Fit to Make Predictions

- Based on this graphical model, what is the predicted strength for plastic that has been extruded at 142 degrees?



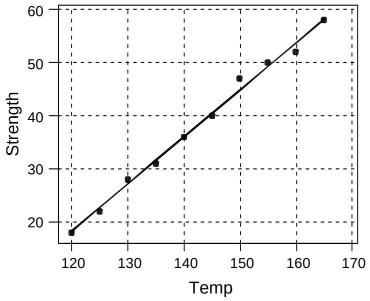
Using the Line of Best Fit to Make Predictions

- Given a value for the predictor variable, determine the corresponding value of the dependent variable graphically.
- Based on this model we would predict a strength of appx. 39 psi for plastic extruded at 142 F



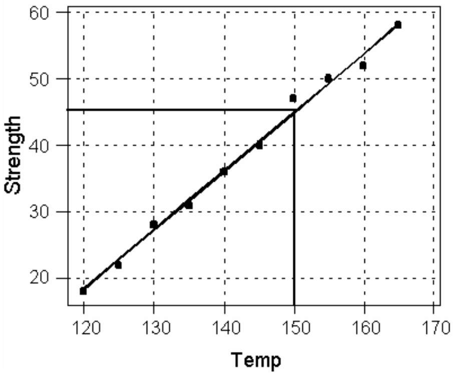
Using the Line of Best Fit to Make Predictions

- Based on this graphical model, at what temperature would I need to extrude the plastic in order to achieve a strength of 45 psi?



Using the Line of Best Fit to Make Predictions

- Locate 45 on the response axis (y-axis)
- Draw a horizontal line to the LOBF
- Drop a vertical line down to the independent axis
- The intercepted value is the temp. required to achieve a strength of 45 psi

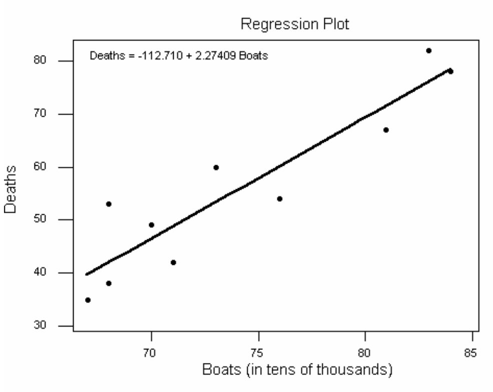


Computing the LSR model

- Given a LSR line for bivariate data, we can use that line to make predictions.
- How do we come up with the best linear model from all possible models?

Bivariate data and the sample linear regression model

- For example, look at the fitted line plot of powerboat registrations and the number of manatees killed.
- It appears that a linear model would be a good one.



$$\hat{y} = b_o + b_1 x$$

The straight line model

- Any straight line is completely defined by two parameters:
  - ⌧ The slope – steepness either positive or negative
  - ⌧ The y-intercept – this is where the graph crosses the vertical axis

The Parameter Estimators

- In our model "b_0" is the estimator for the intercept. The true value for this parameter is β_0
- "b_1" estimates the slope. The true value for this parameter is β_1

Calculating the Parameter Estimators

- The equation for the LOBF is:

$$\hat{y} = b_o + b_1 x$$

Calculating the Parameter Estimators

- To get the slope estimator we use:

$$b_1 = \frac{n\,\Sigma\left(x\,y\right) - \Sigma x \cdot \Sigma y}{n\,\Sigma\left(x^2\right) - \left(\Sigma x\right)^2}$$

$$or$$

$$b_1 = r\left(\frac{s_y}{s_x}\right)$$

Computing the Intercept Estimator

- The intercept estimator is computed from the variable means and the slope:

$$b_0 = \bar{y} - b_1\bar{x}$$

- Realize that both the slope and intercept estimated in these last two slides are really point estimates for the true slope and y-intercept

Revisit the manatee example

Look at the summary statistics and correlation coefficient data from the manatee example

```
Variable   N    Mean    SEMean   StDev
Boats      10   74.10    2.06     6.51
Deaths     10   55.80    5.08    16.05
```

<u>Minitab correlation coefficient output</u>

Correlations: Boats, Deaths

```
Pearson correlation of Boats and Deaths = 0.921
P-Value = 0.000
```

Computing the estimators

So the slope is:

$$b_1 = r\left(\frac{s_y}{s_x}\right)$$

$$b_1 = 0.921\left(\frac{16.05}{6.51}\right) = 2.27$$

Computing the estimators

And the intercept is calculated using the slope information along with the variable means:

$$\begin{aligned} b_0 &= \bar{y} - b_1\bar{x} \\ &= 55.8 - 2.27(74.1) \\ &= -112.4 \end{aligned}$$

Put it together

- In general terms any old linear regression equation is:

  response = intercept + slope(predictor)

  - Specifically for the manatee example the sample regression equation is:

    Deaths = -112.7 + 2.27(boats)

The slope estimate

- b_1 is the estimated slope of the line
- The interpretation of the slope is, “The amount of change in the response for every one unit change in the independent variable.”

The slope estimate

- In our example the estimated slope is 2.27
- This is interpreted as, “For each additional 10,000 boats registered, an additional 2.27 more manatees are killed

The intercept estimate

- Recall the sample regression model:
  “b_0” is the estimated y- intercept

$$\hat{y} = b_0 + b_1 x$$

The interpretation of the y-intercept is, “The value of the response when the control (or independent) variable has a value of 0.”

The intercept Estimate

- Sometimes this value is meaningful. For example resting metabolic rate versus ambient temperature in Centigrade (^{o}C)
- Sometimes it’s not meaningful at all.
- This is an example where the y-intercept just serves to make the model fit better. There can be no such thing as a –112.7 manatees killed

Regression Output

Use the minitab regression output for the manatee example to predict the expected number of manatees killed when the number of power boat registrations is 750,000 (x = 75)

Regression Output

- The sample regression equation is:
  ManateesKilled = -112.7 + 2.27(boats)
- So:
  ManateesKilled = -112.7 + 2.27(75) = 57.6
- This means that we expect between 57 and 58 manatees killed in a year where 750,000 power boats are registered.

Regression Output

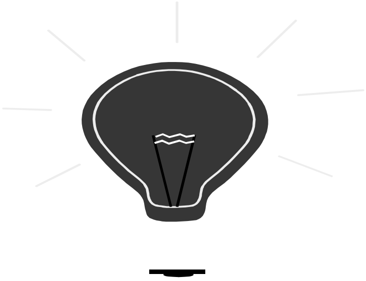
Use the minitab regression output for the manatee example to predict the expected number of manatees killed when the number of power boat registrations is 850,000 (x = 85)

Regression Output

- The sample regression equation is:
  ManateesKilled = -112.7 + 2.27(boats)
- So:
  ManateesKilled = -112.7 + 2.27(85) = 80.25
- This means that we expect between 80 and 81 manatees killed in a year where 750,000 power boats are registered.

Regression Output

STOP!! YOU HAVE VIOLATED
THE CARDINAL RULE OF REGRESSION



Cardinal Rule of Regression

- **NEVER NEVER NEVER NEVER NEVER NEVER predict a response value from a predictor value that is outside of the experimental range.**
- The only predictions we can make (statistically) are predictions for responses where powerboat registrations are between 670,000 and 840,000.
- This means that our prediction for the year when 850,000 powerboats were registered is **garbage**

Regression Estimates The coefficient of determination

- r^2 is called the **coefficient of determination**.
- r^2 is a proportion, so it is a number between 0 and 1 inclusive.
- r^2 quantifies the amount of variation in the response that is due to the variability in the predictor variable.
- r^2 values close to 0 mean that our estimated model is a poor one while values close to 1 imply that our model does a great job explaining the variation

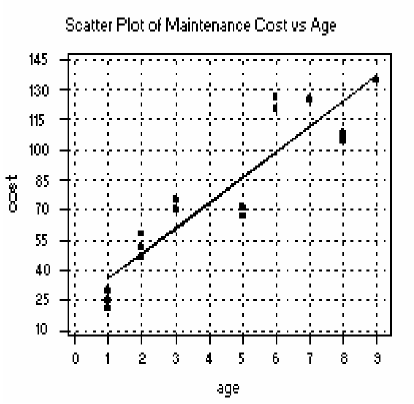
The r^2 Value

- If r^2 is, say, 0.857 we can conclude that 85.7% of the variability in the response is explained by the variability in the independent variable.
- This leaves 100 - 85.7 = 14.3% left unexplained. It’s only the unexplained variation that is incorporated into the “uncertainty”

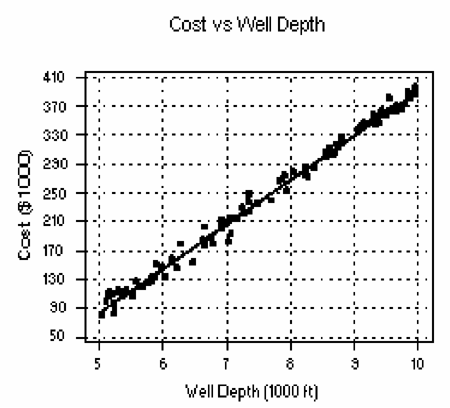
r^2 and the correlation coefficient

- r^2 is related to the correlation coefficient
- It’s just the square of r
- The interpretation as the proportion of variation in the response that is explained by the variation in the predictor variable makes it an important statistic

Scatter of Points and r^2



$r^2 = 0.848$



$r^2 = 0.992$