AP STATISTICS
Notes – Scatterplots and Regressions Day 1

Archaeopteryx is an extinct beast having feathers like a bird but teeth and a long bony tail like a retile. Only six fossil specimens are known. Because these specimens differ greatly in size, some scientists think they are different species rather than individuals from the same species. If the specimens belong to the same species and differ in size because some are younger than others, there should be a positive linear relationship between the bones from all individuals. An outlier from this relationship would suggest a different species. Here are data on the lengths in centimeters of the femur (a leg bone) and the humerus (a bone in the upper arm) for the five specimens that preserve both bones.

Femur	38	56	59	64	74
humerus	41	63	70	72	84

a) Sketch the scatterplot:

explanatory variable:

response variable:

b) Find the correlation coefficient and interpret this value in context.

c) Write the least squares linear regression equation in context.

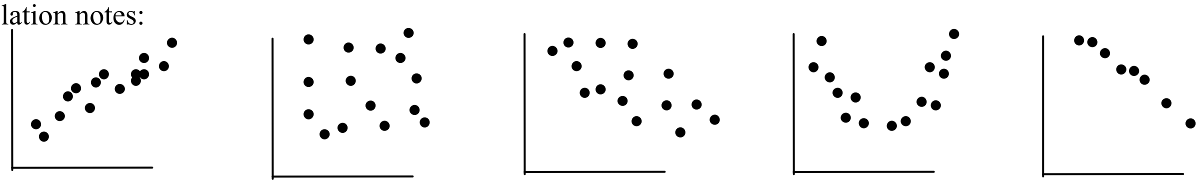
d) Interpret the slope in context.

e) Interpret the y-intercept in context:

f) Calculate and interpret the residual for the specimen with the femur length of 56 cm and humerus length of 63 cm.

g) Interpret the meaning of the r-squared value (coefficient of determination) in context.

Correlation notes:



Regression Notes

Since there are a **PLETHORA** of things that you need to ~~memorize~~ learn how to interpret in this unit, here is a quick reference (keep this safe!). *(Anywhere you see anything in "quotations" or see a blank, fill it in with the appropriate value/context/units/etc. ALWAYS interpret IN CONTEXT!)*

Slope

For each increase of 1 "unit" in "*x*", the model predicts an increase/decrease of _______ "units" in "*y*".

*(NEVER write "there **will be** an increase…" Be sure to write "**the model predicts** an increase…")*

y-intercept

The model predicts that at an "*x*" value of zero, the "*y*" value will be _______.

Residual plot

"No pattern" is a good thing! If we see a clear curve in the residual plot, that means we are using the wrong type of model. Maybe try a logarithmic or exponential model instead (we'll tackle this in chapter 10).

Correlation coefficient (*r*)

(take the square root of R-squared – if the slope is negative, make this value negative as well)

This value indicates the strength (see next sentence) and direction (positive or negative) of the linear association between "*x*" and "*y*". This value must be between -1 and +1. An *r*-value of exactly 1 (or -1) means that the points form a perfectly straight line (which never happens with real-world data).

General suggestion:

- If $|r| < 0.5$, the association is "weak"
- If $|r| > 0.8$, the association is "strong"
- If $0.5 < |r| < 0.8$, the association can be called "moderately strong" or "moderately weak"

R^2 value (coefficient of determination)

The percent of the variation in "*y*" that can be explained by the linear model for "*x*" and "*y*".

Residual ($e = y - \hat{y}$)

Observed (actual) "*y*" value minus predicted (hat) "*y*" value.
Also the vertical distance between the actual point and the regression line.

(To find the predicted ($\hat{y}$) value, plug the *x*-value of the point into the regression equation)

Standard error of residuals (s_e)

Typical difference between the observed and predicted "*y*" values for the points in this regression.

(sometimes in a regression computer printout, this is simply labeled as "s")