Homework discussion:

Mammalsleep:

Answer 1:
log(brain)= B_0 + B_1 log(body)

Brain= e^{B0} * body ^{B1}
$e^{2.1}$ * body $^{.75}$

Reasoning: The plot once graphed brain versus body showed a lot of bunching of data points, which is an automatic trigger to think logging both the body and brain variables.

Answer 2:
Water possum and Man are the least/most brain size adjusted for body size. They are the farthest from the calculated regression line.

Answer 3:
log(brain)= B_0 + B_1 log(body) +/- 2 σ
σ = residual standard deviation=sd(resid(regression))

Upper bound: 6.97
Lower bound: 4.605

Interval: 68 to 1068 grams

Utilities:

When choosing between two possible regression lines, the residuals and the R squared could be examined for the choice. Just because one R squared is higher than the other, it does not mean that it's the best choice. Price that is paid for the extra R squared: The model is more complex and less efficiency.

Summary: Fit is good. Simplicity is good. Common sense helps a lot.

TenMilesRace

Huge difference between the slope of both the men and women when separated compared to the slope of the aggregated data.

Why?

Gender is a confounding variable – it affects both the predictor and the response.

In this dataset, one could check by finding the mean of the net time for each sex, in which the men are on average are faster. The mean of the age for each sex also shows that the men are also a lot older. After seeing the boxplots, it's pretty good evidence than the confounding variable affect both the predictors and the responses.

Multiple variable predictors in a regression line

- Aggregation paradox
  - You get a different answer when you aggregate data instead of looking at them separately
  - There was an aggregation paradox in the TenMilesRace
  - Tends to have a confounding variable involved
- Need to figure out when to aggregate and when not to
- Video Game Reaction Time
  - $y_i = B_0 + B_1$ (faraway) + B_2 (littered) + B_3 (littered * faraway)
  - Both faraway and littered are dummy variables (0 or 1) for the first two Betas
  - Multiplying B_3 is the interactive variable (1 if it's both littered and faraway)
  - Types
    - Baseline (not faraway, not littered)
    - Main effects (including the dummy variables)
    - Interactive (both faraway and littered)
- Slice and Dice versus Main Effects
  - Slice and Dice
    - Different models for the differences
    - Not inaccurate
    - Yet much more inefficient
    - End up with a lot of groups and subgroups and ends up comparing every single little data instead of as a whole
    - Video Game Reaction: 48 parameters to estimate
    - There's an actual possibility of having more parameters to estimate than the dataset itself
  - Why bother with Main Effects?
    - You can just add another dummy variable for the added variable (subject)
    - A lot less complicated/more efficient
    - The choice of baseline is arbitrary as R will automatically adjust the coefficients for the equation to the baseline
    - Video Game Reaction: 15 parameters to estimate
- House prices
  - $y_i = B_0 + B_1 (x_i) + B_2$ (Nbhd2) + B_3 (Nhbd3)
  - Dummy variables are shifting the intercept up and down compared to baseline
  - In interaction terms, shifting the slope up or down