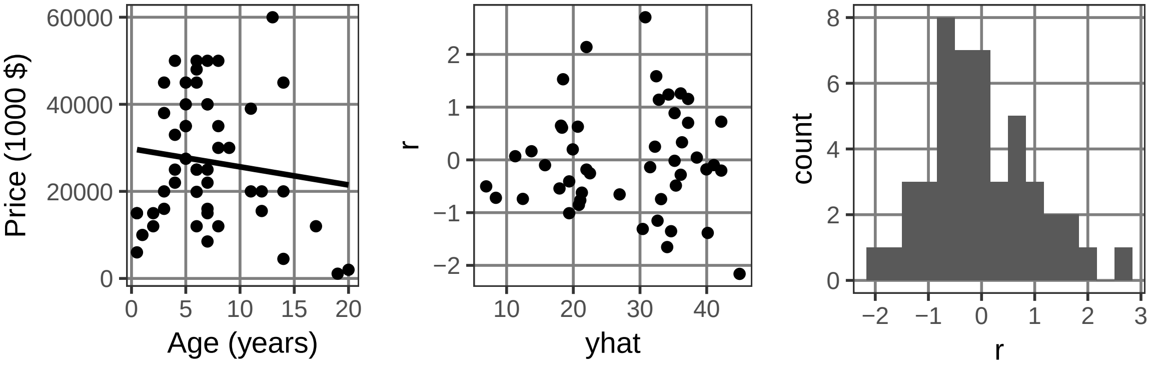
Using the words we learned in this class, answer all questions succinctly; you will lose points for rambling. If you provide `R` code to answer a problem, provide all the necessary code and indicate clearly which variable(s) contain the answer.

1. Circle one: Relative to the median, the mean is _______________ susceptible to potential outliers.

```
less                                      # A
as                                        # B
more                                      # C
not                                       # D
bannana                                   # E
```

2. Suppose you are trying to precict the sale `Price` (thousands of $) of a race horse, using the variable `Age` in years, `Sex`, and `Height` in hands. Consider the following analysis on this dataset, named `horse`.

```
##                   Estimate Std. Error    t value    Pr(>|t|)
## (Intercept)   -9.0460555 79.7625689 -0.1134123 0.910244084
## Sexm        -145.4375992 98.2359598 -1.4804925 0.146205012
## Age           -0.9159778  0.4362411 -2.0997056 0.041802287
## Sexf:Height    2.1439289  5.1088011  0.4196540 0.676877161
## Sexm:Height   11.7726328  4.0027730  2.9411193 0.005300758
```



(a) Write the equation of the fitted regression line.

(b) Interpret two slopes in the context of these data.

(c) Interpret the intercept for male horses in the context of these data.

(d) What values does the variable $Sexm$ take on within this model? Why?

(e) What is the predicted value of the average price of a 7 year old, 17 hand high, male horse?

(f) What is the predicted value of the average price of a 17 year old, 7 hand high, female horse?

(g) Without doing any calculations, comment on predicting the price of a horse at the following values of the variable `age`: $52, 15$. Do you expect one estimate to be more appropriate than the other, given these data? Why?

(h) Comment of the validity of all of the multiple linear regression assumptions.

(i) From the following output, interpret the confidence interval for the slope coefficient on height for female horses in context.

```
##                   2 %   98 %
## (Intercept) -178.10 160.01
## Sexm        -353.65  62.77
## Age           -1.84   0.01
## Sexf:Height   -8.68  12.97
## Sexm:Height    3.29  20.26
```

(j) Assuming a model has already been fit in `R`, and is named `model`, provide the `R` code to reproduce one of the above plots in this question.

3. Suppose you took a random sample of finches from two islands San Cristobal and Santa Cruz, in the Galapagos and measured each bird's `winglength` in millimeters and recorded your data in the dataset named `finch`.

   (a) What type of test is appropriate to test the difference in mean wing length between the islands?

(b) Set up, in symbols, the appropriate hypothesis test that would test the difference in mean wing length between the islands.

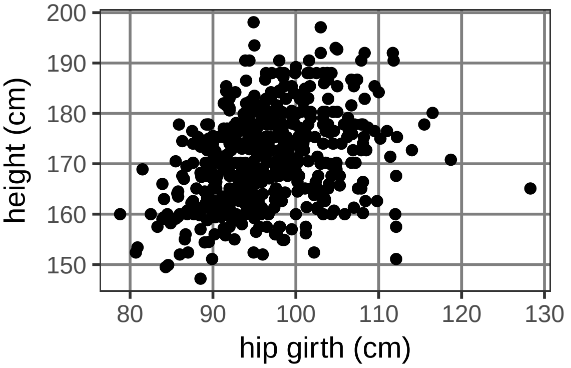
(c) If you found the 95% confidence interval of the difference, $\mu_{SL} - \mu_{SZ}$, to be $(-1.96, 3.49)$ millimeters, how would you conclude your hypothesis test above? Explain.

(d) Interpret the above confidence interval in context.

(e) Would a 98% confidence interval be wider or narrower compared to the 95% confidence interval given above?

(f) If your sample size increased, but the confidence level stayed the same at 95%, would your confidence interval be wider or narrower? Explain.

4. The scatterplot and regression output below show the relationship between height and hip girth, both measured in centimeters, of 507 physically active individuals.



```
##
## Call:
## lm(formula = hgt ~ male + hip.gi, data = bdim)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.2443  -4.1410  -0.1572   4.4581  21.3072
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 133.04869    4.21495  31.566  < 2e-16 ***
## maleTRUE     12.17089    0.58567  20.781  < 2e-16 ***
## hip.gi        0.33270    0.04386   7.585 1.61e-13 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.509 on 504 degrees of freedom
## Multiple R-squared:  0.5232,Adjusted R-squared:  0.5213
## F-statistic: 276.5 on 2 and 504 DF,  p-value: < 2.2e-16
```

(a) Write down two (hypothetical) rows of what this dataset might look like, specifying appropriate varible names and recorded data for each observation.

(b) Describe the relationship between hip girth and height. Hint: I'm looking for three key words/ideas the describe the association between two numerical variables.

(c) What is the predicted height of a male when hip girth is equal to 0?

(d) Interpret the adjusted R^2 in context of these data.

(e) Provide `R` code to reproduce the p-value for the variable `hip.gi`.

(f) Set up the corresponding hypothesis test to go along with the p-value you just created.

5. The bootstrap method aims to estimate the sampling distribution. From an estimate of the sampling distribution, we can quantify uncertainty in our estimates.

(a) Explain how the bootstrap method works, detailing each step of the algorithm.

(b) State what the boostrap method enables us to calculate to better help us quantify uncertainty in our estimates.

6. The (sample) estimate of the (population) mean is the solution to the following minimization problem (which has roots in likelihood),

$$\bar{x} = \arg\min_{\mu} \sum_{i=1}^{n} (x_i - \mu)^2.$$

(a) Circle one: We discussed in class that squared residuals provide too much weight on

```
hot dogs                         # A
observations                     # B
in general                       # C
the samlpe mean                  # D
potential outliers               # E
```

(b) Write an alternative to squared residuals that will partially remedy the problem above.

```
f <- function(mu, x) {
                                # code here
}
```

(c) Explain conceptually what maximum likelihood is used for in statistics and how we can use `R` to help. For full points, mention which function in `R` we use and draw a plot to aid your explanation.