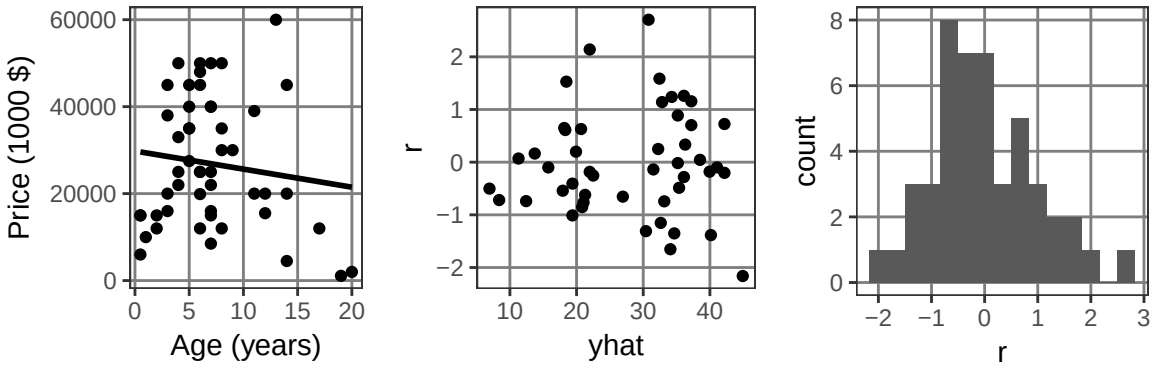
Using the words we learned in this class, answer all questions succinctly; you will lose points for rambling. If you provide `R` code to answer a problem, provide all the necessary code and indicate clearly which variable(s) contain the answer.

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

```
less                                    # A
as                                      # B
more                                    # C answer
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.

$\hat{Price}/1000 = -9.05 - 145.44 * Sexm - 0.92 * Age + 2.13 * Sexf : Height + 11.77 * Sexm : H$

(b) Interpret two slopes in the context of these data. For every 1 year increase in age, the price decreases by .92 thousands of dollars, holding all else constant. For every 1 hand increase in a male horses height, the price increases by 2.14 thousands of dollars, holding all else constant.

(c) Interpret the intercept for male horses in the context of these data. When all numerical explanatory variables are 0, the average price of a female horse is -9.05 thousands of dollars.

(d) What values does the variable *Sexm* take on within this model? Why? 1 or 0 because it is an indicator variable.

(e) What is the predicted value of the average price of a 7 year old, 17 hand high, male horse? -9.05 - 145.44 - 0.92*7 + 11.77*17

(f) What is the predicted value of the average price of a 17 year old, 7 hand high, female horse? -9.05 - 0.92*7 + 2.13*17

(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? 15 is reasonable since we have data around 15, but 52 is not since this would be extapolation.

(h) Comment of the validity of all of the multiple linear regression assumptions. Normality looks good by histogram. Linearity seems ok by scatter plot, no real pattern. Constant variation is not great since a slight cone shape is going on. We'd be content with the assumption of independence if the data were randomly collected. Age and height might be highly correlated, which would break multicolinearity, so we should check for it.

(i) From the following output, interpret the confidence interval for the slope coefficient on height for female horses in context. We are 96% confident that for every one hand increase in a femalse horse's height, the price decreases by between 8.68 and 12.97 thousands of dollars, holding all else constant.

```
##                   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.

```
ggplot(horse, aes(age, price)) + geom_point() + geom_smooth(method="lm", se=FAL
df <- data.frame(yhat=fitted(model), r=rstandard(model))
ggplot(df, aes(yhat, r)) + geom_point()
ggplot(df, aes(r)) + geom_histogram(binwidth=1/3)
```

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? Two sample t-test.

(b) Set up, in symbols, the appropriate hypothesis test that would test the difference in mean wing length between the islands. $H_0 : \mu_{sl} = \mu_{sz}$ versus $H_1 : \mu_{sl} \neq \mu_{sz}$ at $\alpha = 0.05$

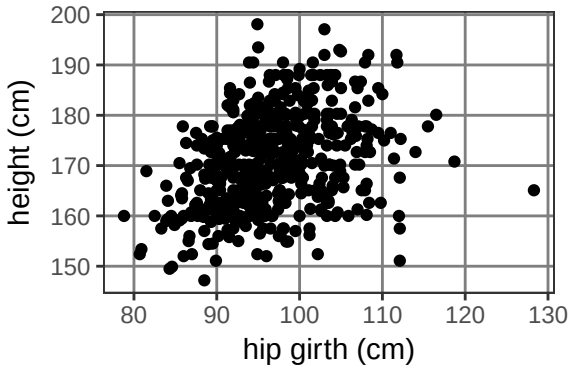
(c) If you found the 95% confidence interval of the difference, $\mu_{SL} - \mu_{SZ}$, to be $(-1.96, 3.49)$ millimeters, how would you conclude your hypothesis test above? Explain. Fail to reject H_0 since 0 is contained within the confidence interval.

(d) Interpret the above confidence interval in context. We are 95% confident that the true difference in mean wing lengths, between Galapagos finches on the islands San Cristobal and Santa Cruz, is between -1.96 and 3.49 – since 0 is contained within the interval we don't care which way we subtracted.

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

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

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. hgt hig.gi male
170 90 1
180 100 0

(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. positive, medium, linear associatoin

(c) What is the predicted height of a male when hip girth is equal to 0? 133.05 + 12.17

(d) Interpret the adjusted R^2 in context of these data. 52% of the variation in height can be explained by this linear model.

(e) Provide `R` code to reproduce the p-value for the variable `hip.gi`. 2*(1 - pt(abs(.3327/04386), df=504))

(f) Set up the corresponding hypothesis test to go along with the p-value you just created. $H_0 : \beta_{hip.gi} = 0$ versus $H_1 : \beta_{hip.gi} \neq 0$ at $\alpha = 0.05$

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. Repeatedly resample the observation, uniformly and with replacement. With each resample, calculate the statistic of interest. Estimate standard error by calculating the standard deviation of the bootstrap statistics.

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

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 answer
```

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

```
f <- function(mu, x) {
abs((x - mu)^2)# 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.

Maximum likelihood is a technique to estimate population parameters. R can help us find the maximum of arbitrary functions. Specifically, optim is R's minimizaton function. Picture.