Stat 245 – Test 1

Adham Rishmawi

October 01, 2022

Casual diagram

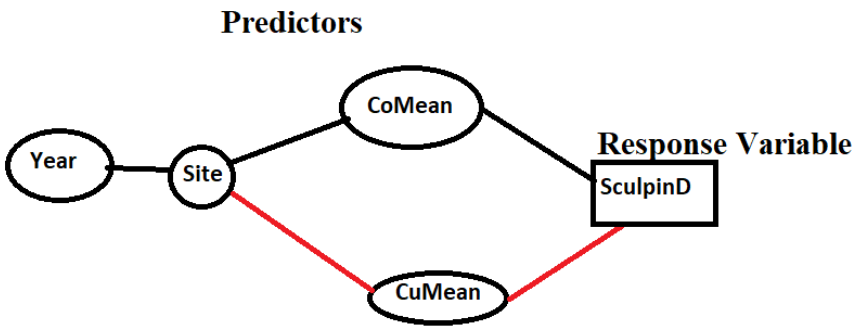


Figure 1: Casual Diagram

All arrows are assumed to go to the right –> (ex: site to CuMean and to sculpinD)

This casual diagram demonstrates that my 4 predictors are Year, Site, CoMean, CuMean and my response Variable is SculpinD. the only predicitors I excluded were Biomass and Troutd from the diagram because they do not have relevance to my response variable and would alter the results of sculpinD. Justification:

-Year: because it is a predictor that will show difference over the course of a time and prove assiocat
-Site: because it is a predictor that will show differences between different sites
and elaborate to see if location alter association between the two variables
-CoMean & CuMean: Because it will show us the total amount of copper or cobalt that will
assist in ashowing association between both desired variables.
-SculpinD: this our response variable because it will be altered based on all predictors and will be go

```
creek_data <- read.csv("https://sldr.netlify.app/data/PantherCreek.csv")
```

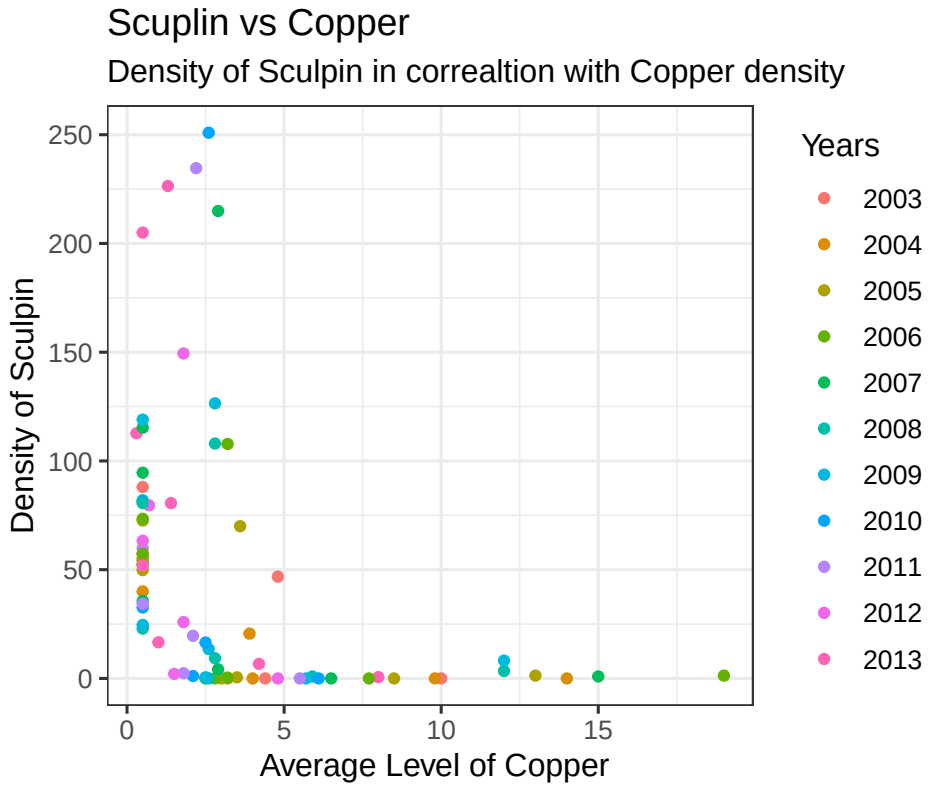
```
crelr <- lm(SculpinD ~ CuMean + CoMean + Site + Year, data = creek_data )

summary(crelr)
```

```
##
## Call:
## lm(formula = SculpinD ~ CuMean + CoMean + Site + Year, data = creek_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -98.25 -17.44  -3.35  12.15  97.92
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.236e+04  3.135e+03  -3.942 0.000195 ***
## CuMean        1.389e+00  3.616e+00   0.384 0.702027
## CoMean        8.639e-02  1.493e-01   0.579 0.564865
## SiteBD-km01   1.523e+01  4.553e+01   0.335 0.738973
## SiteDE-km0.1  1.291e+02  6.452e+01   2.001 0.049486 *
## SiteNA-km01   6.490e+01  6.323e+01   1.026 0.308463
## SitePA-km17   3.760e+01  5.537e+01   0.679 0.499458
## SitePA-km22   2.179e+01  5.632e+01   0.387 0.700096
## SitePA-km37   1.624e+02  5.376e+01   3.020 0.003575 **
## SitePA-km39   1.021e+02  6.359e+01   1.605 0.113161
## Year          6.142e+00  1.543e+00   3.981 0.000171 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.37 on 67 degrees of freedom
## Multiple R-squared:  0.7566, Adjusted R-squared:  0.7203
## F-statistic: 20.83 on 10 and 67 DF,  p-value: < 2.2e-16
```

Equation linear regression model = SculpinD = -1.236e+04 + 1.389e+00CuMean + 8.639e-02CoMean + 1.523e+01SiteBD + 1.291e+02SiteDE + 6.490e+01SiteNA + 3.760e+01SitePA-km17 + 2.179e+01SitePA-km22 + 1.624e+02SitePA-km37 + 1.021e+02SitePA-km39 + 6.142e+00 Year + ϵ where $\epsilon \sim N(0, 32.37)$.

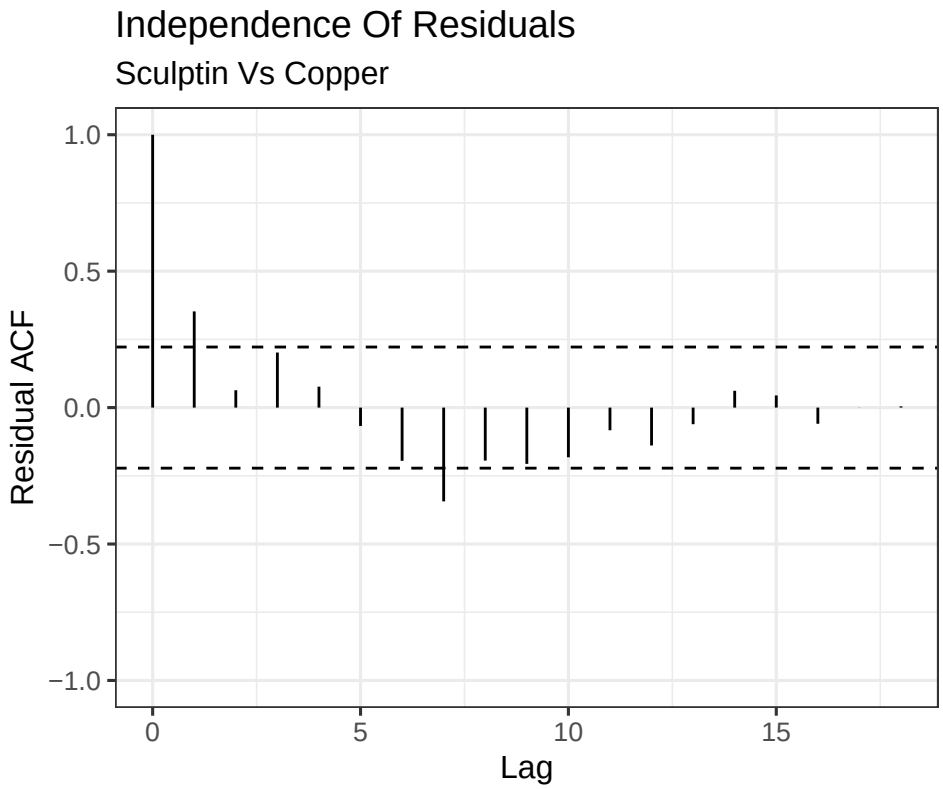
Due to the high R-squared value, we can determine that the predictors are altering the response value. hence, giving good confidence that the two variables SculpinD and CoMean are associated.

```
creek_data |>
    mutate(Year_chr = as.character(Year))|>
      gf_point(SculpinD ~ CuMean,
            data = creek_data,
            color = ~Year_chr)|>
  gf_labs(subtitle = "Density of Sculpin in correaltion with Copper density",
          title = "Scuplin vs Copper",
          x = " Average Level of Copper",
          y = "Density of Sculpin",
          color = "Years" )
```



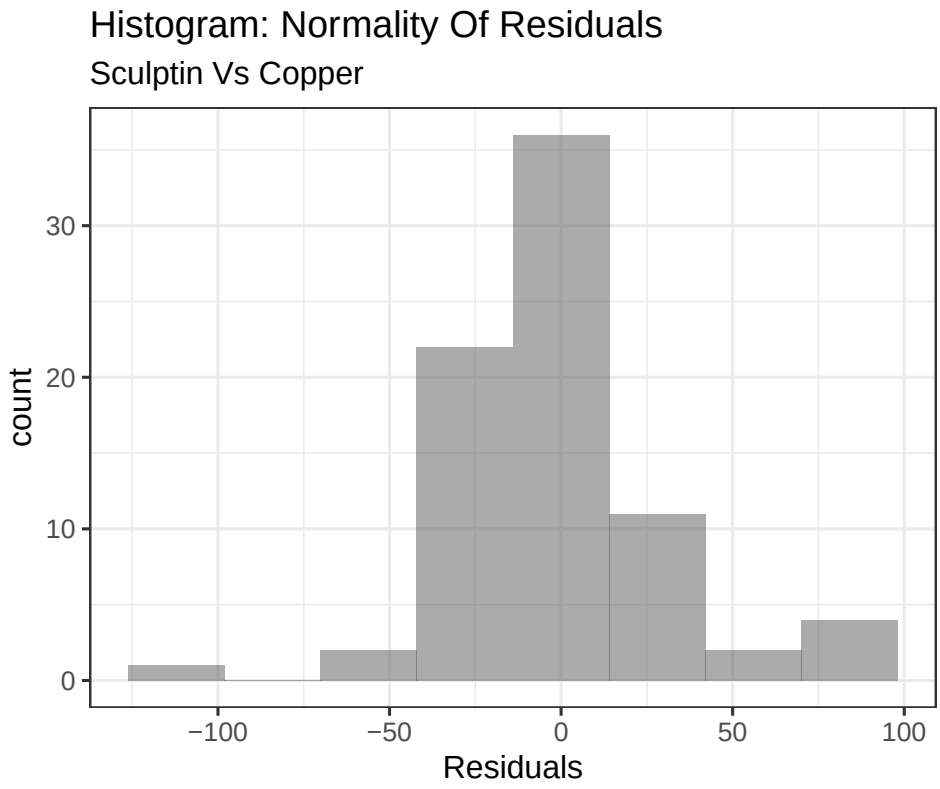
Initial assiocation between SculpinD and CoMean As we can observe from the graph, it shows that as the level of cobalt increases, the level of sculptin drastically decreases. This can be an intital sign of assiocation between both variables.

```
s245::gf_acf(~crelr)|>
gf_lims(y=c(-1,1))|>
   gf_labs(subtitle = "Sculptin Vs Copper", title = "Independence Of Residuals", )
```



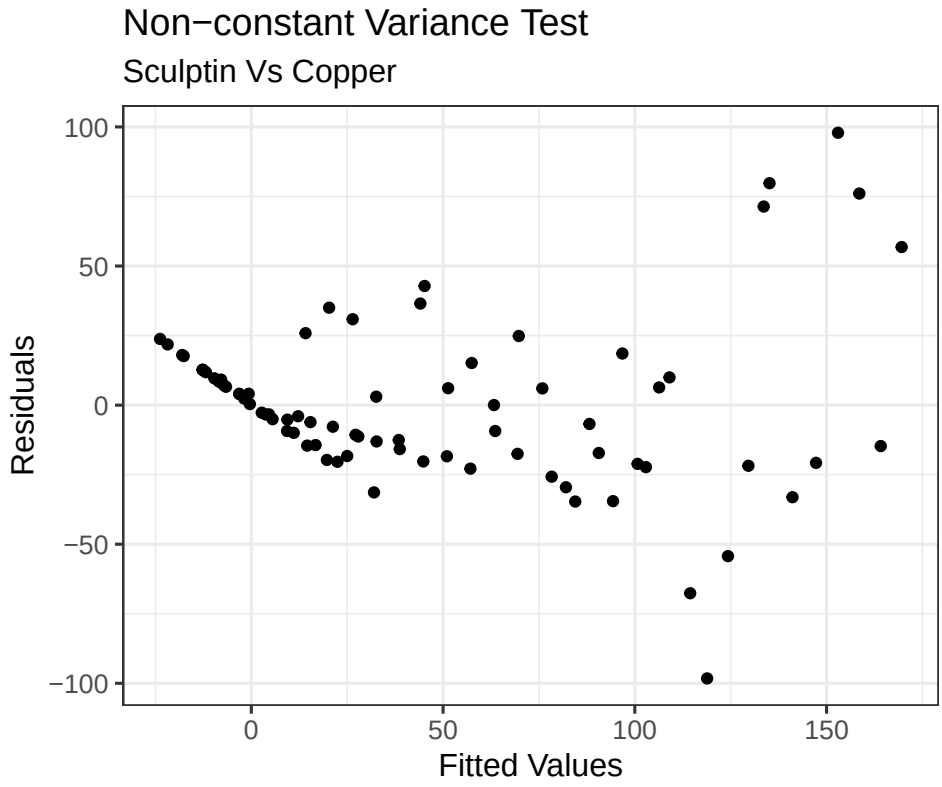
This graph failed the ACf test and shows us that these two variables most likely have no assiocation. It failed because the lines went beyond the dashed lines after the first line. (However, for the sake of the test we will do all tests!)

```
creek_data <- creek_data |>
mutate(preds = predict (crelr),
resids = resid(crelr))
gf_histogram(~resids, data = creek_data, bins=8)|>
  gf_labs(subtitle = "Sculptin Vs Copper", title = "Histogram: Normality Of Residuals", x = "Residuals"
```



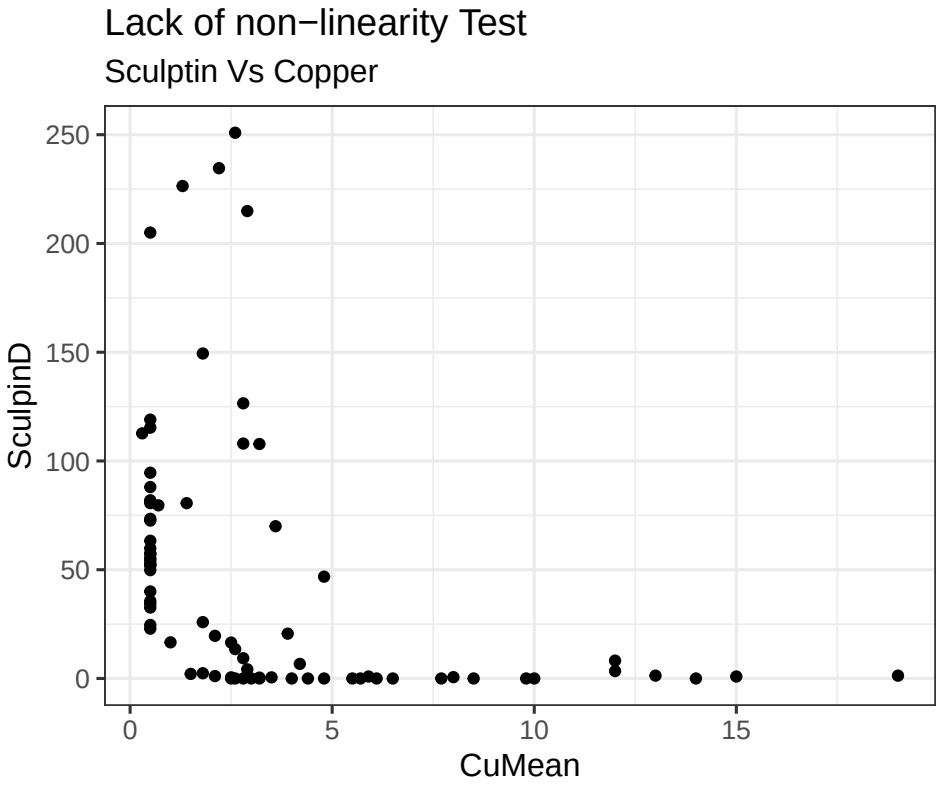
This graph doesnt pass because the histogram is mostly skewed right and fails to fall between the necessary boundaries. I do not see a even distrubution that would give me the confidence to say this passed! *grade this one*

```
gf_point(resid(crelr)~fitted(crelr))|>
gf_labs(x = 'Fitted Values', y = 'Residuals')|>
  gf_labs(subtitle = "Sculptin Vs Copper", title = "Non-constant Variance Test", )
```



Despite the flat edge on the left side, I would gracely say that we could say that this test passed! I can also see that this plotted graph demonstrates a pattern which is consistent with a graph that has constant variance. *grade this* (I have observed non constant variance which means that the points are distributed more across the y axis)

```
gf_point(SculpinD ~ CuMean,
         data = creek_data)|>
  gf_labs(subtitle = "Sculptin Vs Copper", title = "Lack of non-linearity Test", )
```



due to the "L" shaped slope, we can see a lack of linearity in our scatter plot. This makes our test failed?!

```
fake_data <- expand.grid(CuMean = seq( from = 0 ,
                                       by  = 0.3 ,
                                       to = 13 ),
                         Year = 2006,
                          CoMean = 0 ,
                           Site = "PA-km39" )
```

```
preds <- predict(crelr,
                 newdata = fake_data,
                 se.fit = TRUE)
```

```
fake_data <- fake_data |>
  mutate( pred = predict(crelr,
                         newdata = fake_data))

glimpse(fake_data)
```

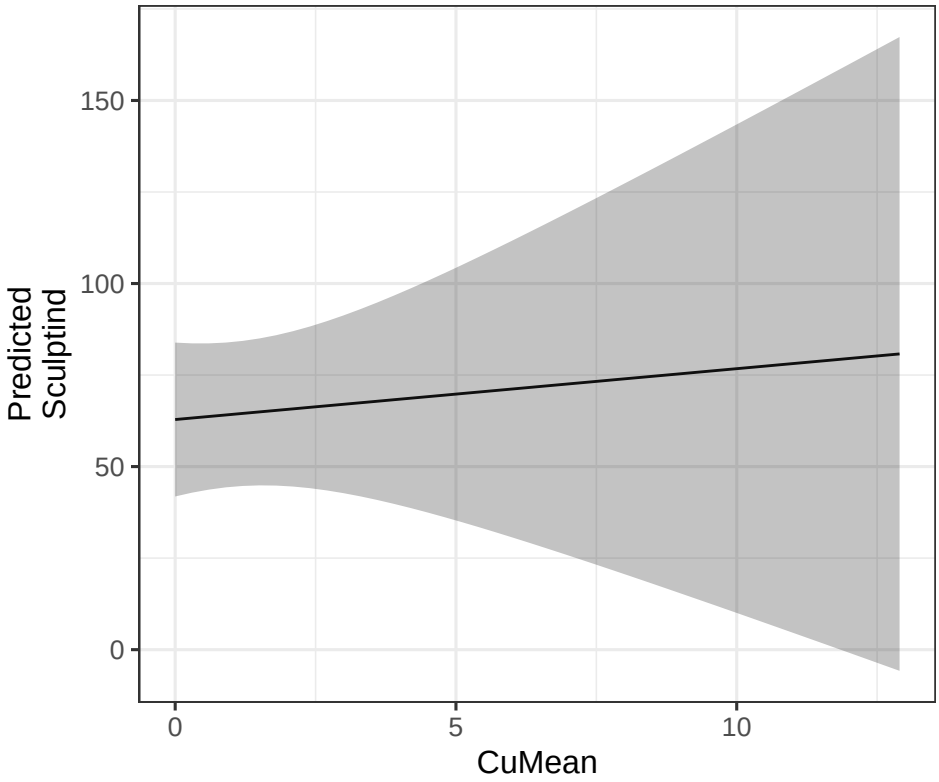
```
## Rows: 44
## Columns: 5
## $ CuMean <dbl> 0.0, 0.3, 0.6, 0.9, 1.2, 1.5, 1.8, 2.1, 2.4, 2.7, 3.0, 3.3, 3.6~
## $ Year   <dbl> 2006, 2006, 2006, 2006, 2006, 2006, 2006, 2006, 2006, 2006, 200~
## $ CoMean <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ Site   <fct> PA-km39, PA-km39, PA-km39, PA-km39, PA-km39, PA-km39, PA-km39, ~
## $ pred   <dbl> 62.85032, 63.26711, 63.68391, 64.10070, 64.51749, 64.93428, 65.~
```

```
fake_data <- fake_data |>
  mutate (pred = preds$fit,
          pred.se = preds$se.fit)
```

```
fake_data <- fake_data |>
  mutate(CI_lower = pred - 1.96*pred.se,
         CI_upper = pred + 1.96*pred.se)
glimpse(fake_data)
```

```
## Rows: 44
## Columns: 8
## $ CuMean   <dbl> 0.0, 0.3, 0.6, 0.9, 1.2, 1.5, 1.8, 2.1, 2.4, 2.7, 3.0, 3.3, 3~
## $ Year     <dbl> 2006, 2006, 2006, 2006, 2006, 2006, 2006, 2006, 2006, 2006, 2~
## $ CoMean   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ Site     <fct> PA-km39, PA-km39, PA-km39, PA-km39, PA-km39, PA-km39, PA-km39~
## $ pred     <dbl> 62.85032, 63.26711, 63.68391, 64.10070, 64.51749, 64.93428, 6~
## $ pred.se  <dbl> 10.72109, 10.40228, 10.18954, 10.08956, 10.10571, 10.23742, 1~
## $ CI_lower <dbl> 41.83699, 42.87864, 43.71241, 44.32515, 44.71031, 44.86895, 4~
## $ CI_upper <dbl> 83.86365, 83.65559, 83.65540, 83.87624, 84.32467, 84.99962, 8~
```

```
gf_line(pred ~ CuMean,
        data = fake_data) |>
  gf_labs(y='Predicted\nSculptind') |>
  gf_ribbon(CI_lower + CI_upper ~ CuMean)
```



I was doomed from the start! Initially I established that I wanted to see if CuMean and SculptinD had any association. However, after observing multiple Assessment test's failures (like ACF and and the Histogram), I should have concluded that this test was not advisable. I had to complete all the steps because it was part of this test and in real life circumstances, I would have stopped and looked for different variables to look for association. If the assessments had passed, this type of plotted prediction would have observed that an increase in copper would led to a larger confidence band! Warning: Any Values that Are Not Shown Were Held Fixed!

```r
require(MuMIn)
crelr <- crelr |>
  update(na.action = 'na.fail')
crelr_dredge <- dredge(crelr, rank = 'BIC')
head(crelr_dredge, 4)
```

```
## Global model call: lm(formula = SculpinD ~ CuMean + CoMean + Site + Year, data = creek_data,
##     na.action = "na.fail")
## ---
## Model selection table
##    (Intrc)   CoMen  CuMen Site  Year df   logLik   BIC delta weight
## 13  -11370                   + 5.665 10 -376.184 795.9  0.00  0.793
## 14  -11640 0.06584           + 5.796 11 -376.055 800.0  4.10  0.102
## 15  -11660         0.6401    + 5.805 11 -376.163 800.3  4.32  0.092
## 16  -12360 0.08639 1.3890    + 6.142 12 -375.969 804.2  8.28  0.013
## Models ranked by BIC(x)
```

My BIC is also a sham because it relies on my Assessments all passing. If my BIC wasn't a sham, I would have looked to see which predictor had the lowest delta and that would have been line 13 in the display!

Conclusion: My work in this test can not be trusted because of initial test failure and no conclusions can be drawn.