Factors and Interactions

Adam J Sullivan, PhD

1/29/2018

Factors

What are Factors?

- Factors are categorical data.
- Factors contain
  - Levels
  - Can be numerical or character data

Why do we use them?

- Factors allow us to group things by category.
- Factors create dummy variables or indicator variables in our regressions.

What is an indicator variable?

- Consider the scenario where we have 3 treatments: A, B, & C
- We could have two indicator variables:
  - I(Treat_A) is
    - 1 if patient is on treatment A
    - 0 if patient is not on treatment A
  - I(Treat_B) is
    - 1 if patient is on treatment B
    - 0 if patient is not on treatment B
  - Treatment C would be both:
    - I(Treat_A) = 0
    - I(Treat_B) = 0

What does this mean in regressions?

- Indicator variables change the regression:

$$Outcome = \beta_0 + \beta_1 Age + \beta_2 I(Treat_A) + \beta_3 I(Treat_B)$$

- For a person on Treatment A:

$$Outcome = (\beta_0 + \beta_2) + \beta_1 Age$$

- For a person on Treatment B:

$$Outcome = (\beta_0 + \beta_3) + \beta_1 Age$$

- For a person on Treatment C:

$$Outcome = \beta_0 + \beta_1 Age$$

What does this mean in Regression?

- We can see that a factor leads to multiple different regression lines.
- Each line then has a different intercept than the others.
- In this regression age has the same effect, just the baseline is different.

Are there different types of factors?

- We can have different types of factors
  - Nominal
  - Ordinal

Nominal Factors

- Nominal factors are factors that represent named categories.
- These are categories that do not have an intrinsic ordering.
- Examples:
    - Gender
    - Sex
    - Race/ethnicity
- We must treat these as indicator variables in models.

Ordinal Factors

- Ordinal factors are factors that represent some ordered categories.
- These factors have an intrinsic ordering.
- Examples:
  - Likert Scales (Poor, Neutral, Good)
  - BMI (Underweight, Normal, Overweight, Obese)
  - Age Groups (under 18, 18-25, 25-35, 35+)
- In regression models can be indicator variables or a trend.

Indicator Variables vs Trends

- We saw with indicator variables that we have multiple variables to represent the factor.
- Each category leads to a different regression.
- Consider this:

$$Outcome = \beta_0 + \beta_1 age + \beta_2 I(BMI = underweight) + \beta_3 I(BMI = Overweight+)$$

- We then have 3 different regressions:
  - 1 for normal BMI
  - 1 for underweight BMI
  - 1 for overweight+ BMI

Our 3 regressions

- Normal BMI

$$Outcome = \beta_0 + \beta_1 age$$

- Underweight BMI

$$Outcome = (\beta_0 + \beta_2) + \beta_1 age$$

- Overweight+ BMI

$$Outcome = (\beta_0 + \beta_3) + \beta_1 age$$

Indicator Variables vs Trends

- With a trend we allow the factor to have one slope.
- Instead of 1 category leading to a new regression, each category leads to a further increase.
- Our model

$$Outcome = \beta_0 + \beta_1 age + \beta_2 BMI$$

Our Regressions

- Normal BMI

$$Outcome = \beta_0 + \beta_1 age$$

- Underweight BMI

$$Outcome = (\beta_0 + \beta_2) + \beta_1 age$$

- Overweight+ BMI

$$Outcome = (\beta_0 + 2\beta_2) + \beta_1 age$$

What is the difference?

- You can see that it appears that we still have 3 regressions.
- indicator variable regression, each group can have a unique change from the baseline.
    - $\beta_{group=2} \neq \beta_{group=3}$
- trend regression, each group has the same difference between them

An example: PBC Data

- This data is from the Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984.
- A total of 424 PBC patients, referred to Mayo Clinic during that ten-year interval, met eligibility criteria for the randomized placebo controlled trial of the drug D-penicillamine.
- The first 312 cases in the data set participated in the randomized trial and contain largely complete data.
- The additional 112 cases did not participate in the clinical trial, but consented to have basic measurements recorded and to be followed for survival.

PBC Data

Variable	Description
age	in years
albumin	serum albumin (g/dl)
alk.phos	alkaline phosphotase (U/liter)
ascites	presence of ascites
ast	aspartate aminotransferase, once called SGOT (U/ml)
bili	serum bilirunbin (mg/dl)

PBC Data

Variable	Description
chol	serum cholesterol (mg/dl)
copper	urine copper (ug/day)
edema	0 no edema, 0.5 untreated or successfully treated
	1 edema despite diuretic therapy
hepato	presence of hepatomegaly or enlarged liver
id	case number

PBC Data

Variable	Description
platelet	platelet count
protime	standardised blood clotting time
sex	m/f
spiders	blood vessel malformations in the skin
stage	histologic stage of disease (needs biopsy)
status	status at endpoint, 0/1/2 for censored, transplant, dead

PBC Data

Variable	Description
time	number of days between registration and the earlier of death, transplantion, or study analysis in July, 1986
trt	1/2/NA for D-penicillmain, placebo, not randomised
trig	triglycerides (mg/dl)

Data

```
library(survival)
pbc
```

```
##       id time status trt      age sex ascites hepato spiders edema bili
## 1      1  400      2   1 58.76523   f       1      1       1   1.0 14.5
## 2      2 4500      0   1 56.44627   f       0      1       1   0.0  1.1
## 3      3 1012      2   1 70.07255   m       0      0       0   0.5  1.4
## 4      4 1925      2   1 54.74059   f       0      1       1   0.5  1.8
## 5      5 1504      1   2 38.10541   f       0      1       1   0.0  3.4
## 6      6 2503      2   2 66.25873   f       0      1       0   0.0  0.8
## 7      7 1832      0   2 55.53457   f       0      1       0   0.0  1.0
## 8      8 2466      2   2 53.05681   f       0      0       0   0.0  0.3
## 9      9 2400      2   1 42.50787   f       0      0       1   0.0  3.2
## 10    10   51      2   2 70.55989   f       1      0       1   1.0 12.6
## 11    11 3762      2   2 53.71389   f       0      1       1   0.0  1.4
## 12    12  304      2   2 59.13758   f       0      0       1   0.0  3.6
## 13    13 3577      0   2 45.68925   f       0      0       0   0.0  0.7
## 14    14 1217      2   2 56.22177   m       1      1       0   1.0  0.8
## 15    15 3584      2   1 64.64613   f       0      0       0   0.0  0.8
## 16    16 3672      0   2 40.44353   f       0      0       0   0.0  0.7
## 17    17  769      2   2 52.18344   f       0      1       0   0.0  2.7
## 18    18  131      2   1 53.93018   f       0      1       1   1.0 11.4
## 19    19 4232      0   1 49.56057   f       0      1       0   0.5  0.7
```

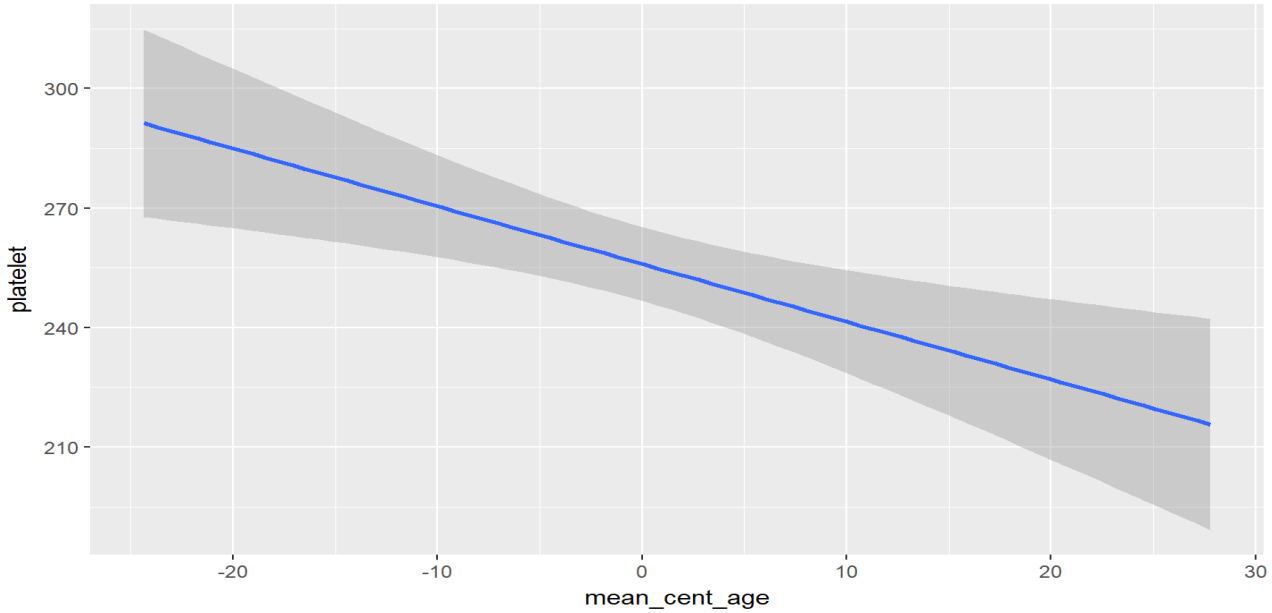
Consider Trends vs Indicators

```
library(tidyverse)
pbc <- pbc %>%
      filter(!is.na(stage)) %>%
      mutate(stage_dummy = as.factor(stage)) %>%
      mutate(mean_cent_age= age-mean(age))
```

Regression plot with trend:

```
library(ggplot2)
ggplot(pbc, aes(mean_cent_age, platelet, color=stage)) + geom_smooth(method="lm", se=FALSE)
```

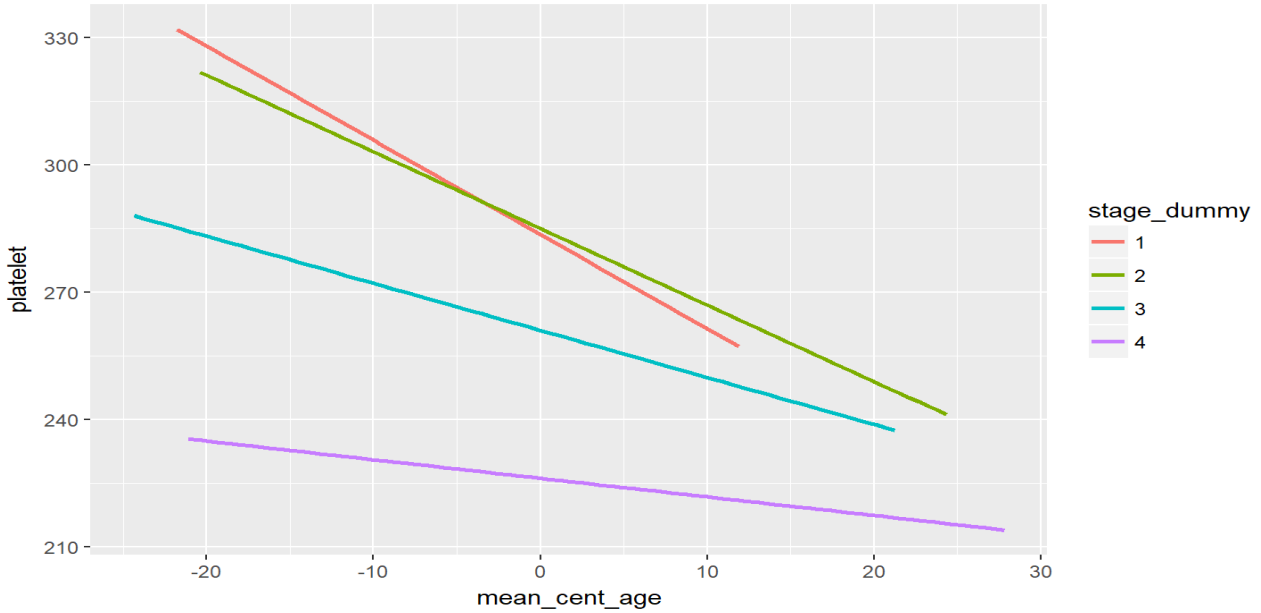
Regression plot with trend:



Regression plot with Indicators:

```
library(ggplot2)
ggplot(pbc, aes(mean_cent_age, platelet, color=stage_dummy)) + geom_smooth(method="lm")
```

Regression plot with Indicators:



Regressions: Trend

```
library(broom)
mod1 <- lm(data=pbc, platelet~mean_cent_age + stage)
tidy(mod1)
```

```
##           term   estimate  std.error statistic      p.value
## 1   (Intercept) 333.264076 16.8964124 19.723955 6.723685e-61
## 2 mean_cent_age  -1.067494  0.4485106 -2.380087 1.777869e-02
## 3         stage -25.411123  5.3457340 -4.753533 2.797793e-06
```

Interpretations

- age: For 2 people with the same disease stage, a person 1 year older has an average platelet count of 1 less than the younger person.
- stage: For 2 people of the same age, a person 1 disease stage higher has an average platelet count 25 less than the person with the lower disease stage.

Regressions: Trend

```
library(broom)
mod2 <- lm(data=pbc, platelet~age + stage_dummy)
tidy(mod2)
```

```
##          term   estimate  std.error   statistic      p.value
## 1  (Intercept) 340.217912 29.6218056 11.48538738 1.502640e-26
## 2         age  -1.031162  0.4527552 -2.27752749 2.328715e-02
## 3 stage_dummy2  -2.158798 22.9328542 -0.09413558 9.250491e-01
## 4 stage_dummy3 -26.827247 21.9497703 -1.22221082 2.223549e-01
## 5 stage_dummy4 -60.081105 22.2371484 -2.70183496 7.192199e-03
```

Interpretations

- age: For 2 people with the same disease stage, a person 1 year older has an average platelet count of 1 less than the younger person.
- stage_dummy 2: For 2 people of the same age, a person in diease stage 2 higher has an average platelet count 2 less than the person with disease stage 1.
- stage_dummy 3: For 2 people of the same age, a person in diease stage 3 higher has an average platelet count 27 less than the person with disease stage 1.
- stage_dummy 4: For 2 people of the same age, a person in diease stage 4 higher has an average platelet count 60 less than the person with disease stage 1.

Is there a difference?

- Yes!!
- If we look between diease stage 1 and 2 the difference is on average 2 in the model with dummy variables.
- In the trend model the difference between any 2 stages is on average 25.

Is this difference Significant?

- We can test for significane with the F-test

```
anova(mod1,mod2)
```

```
## Analysis of Variance Table
##
## Model 1: platelet ~ mean_cent_age + stage
## Model 2: platelet ~ age + stage_dummy
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1    398 3383405
## 2    396 3370016  2     13389 0.7866 0.4561
```

- Based on our test, the trend gives us just as much information.

How about R^2

```
library(broom)
glance1 <- glance(mod1)[,c(1:2)]
glance2 <- glance(mod2)[,c(1:2)]
bind_rows(glance1, glance2)
```

```
##    r.squared adj.r.squared
## 1 0.07740521    0.07276905
## 2 0.08105604    0.07177377
```

Interaction

Interaction

- The definition of interaction is the direct effect that one kind of particle has on another.
- This is similar to how we view it in statistics.
- When there is an interaction, the effect of one variable is different in one group than in another.
- For example, if we feel there is a interaction between sex and treatment, then the effect of ones treatment is directly related to what their sex is.

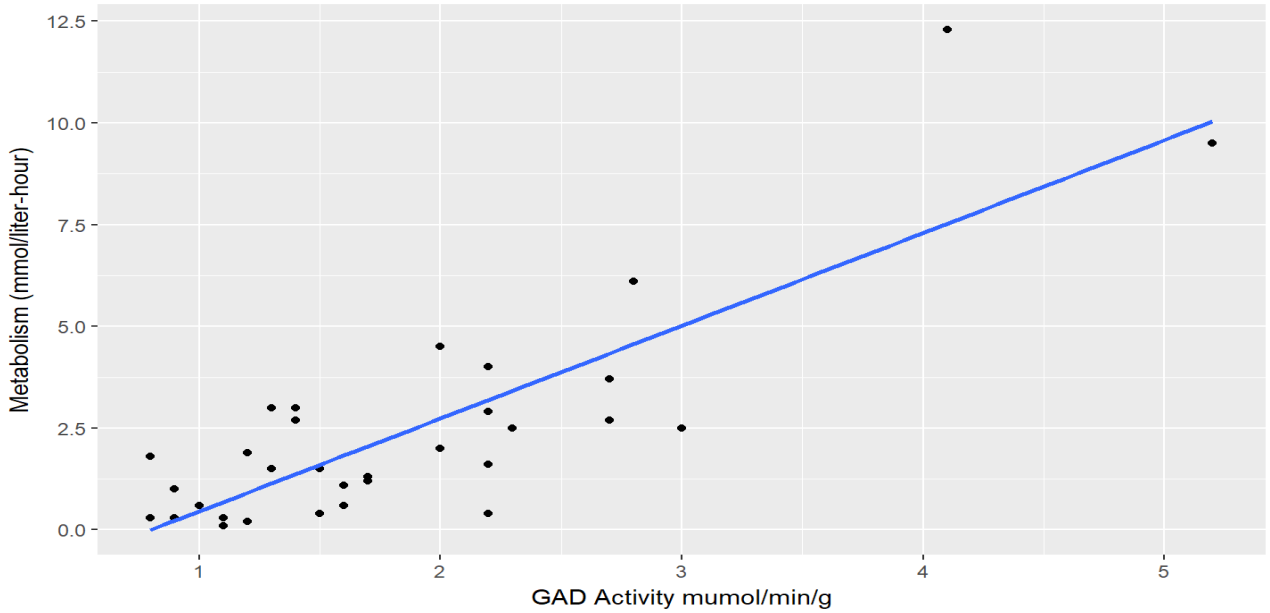
Example with Categorical Interaction

- This data is from 18 women and 14 men to investigate a certain theory on why women exhibit a lower tolerance for alcohol and develop alcohol–related liver disease more readily than men.
- This data is from The Statistical Sleuth: A Course in Methods of Data Analysis

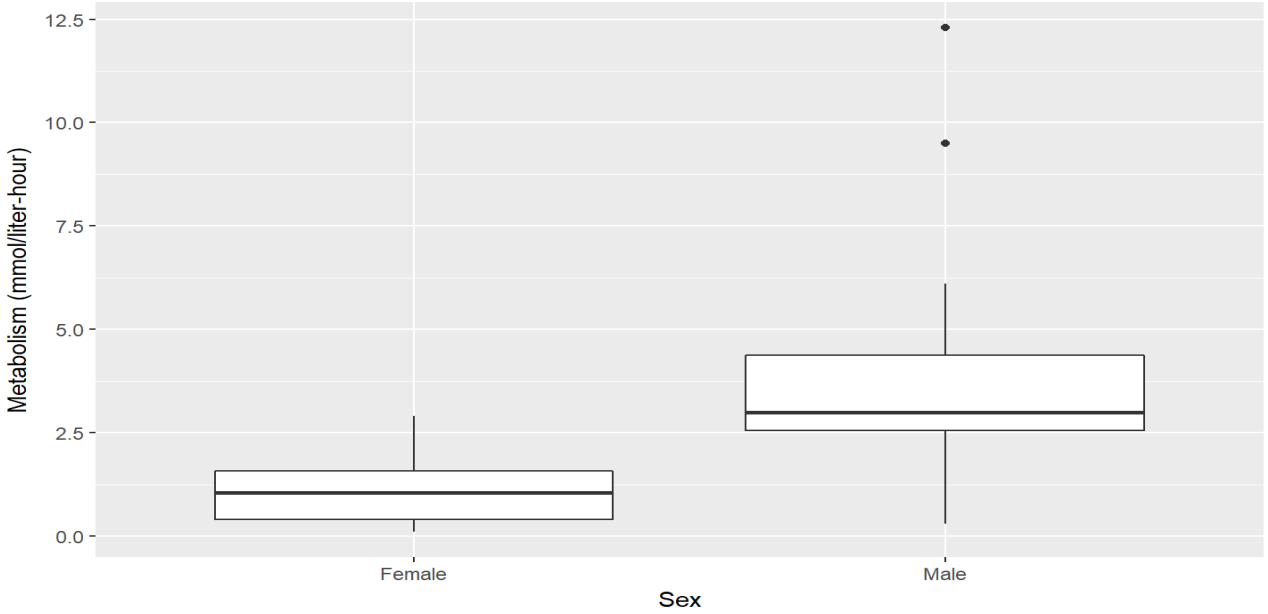
Example with Categorical Interaction

Variable Name	Description
Subject	subject number in the study
Metabol	first–pass metabolism of alcohol in the stomach (in mmol/liter-hour)
Gastric	gastric alcohol dehydrogenase activity in the stomach (in mumol/min/g of tissue)
Sex	sex of the subject
Alcohol	whether the subject is alcoholic or not

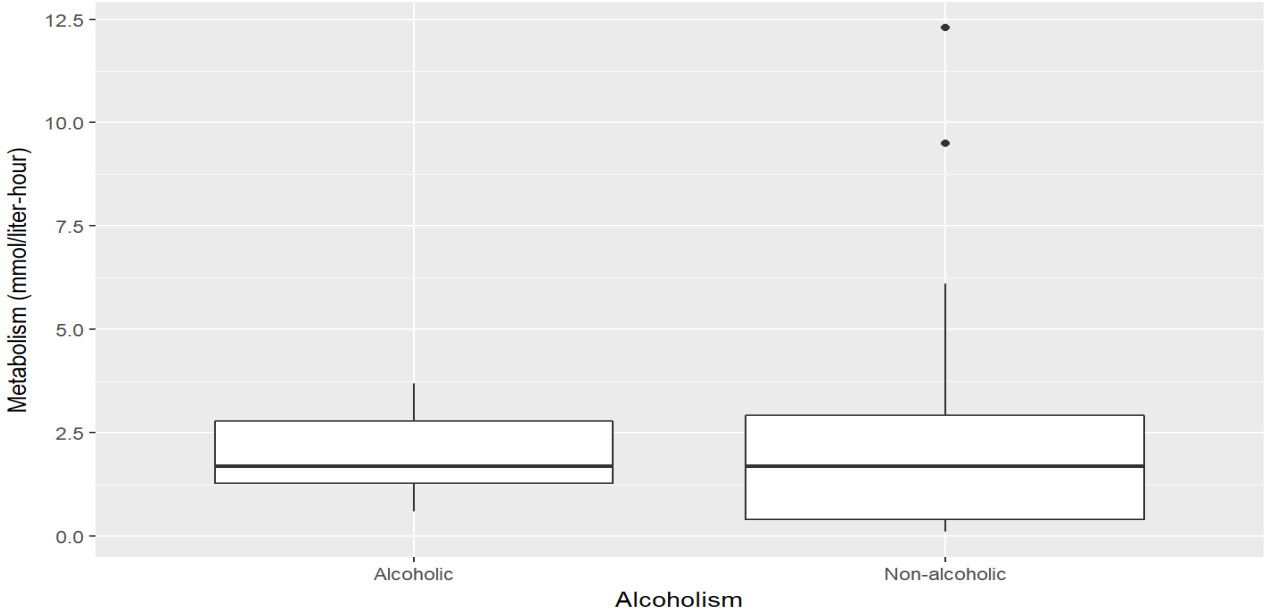
Data Exploration: Metabolism by Gastric Activity



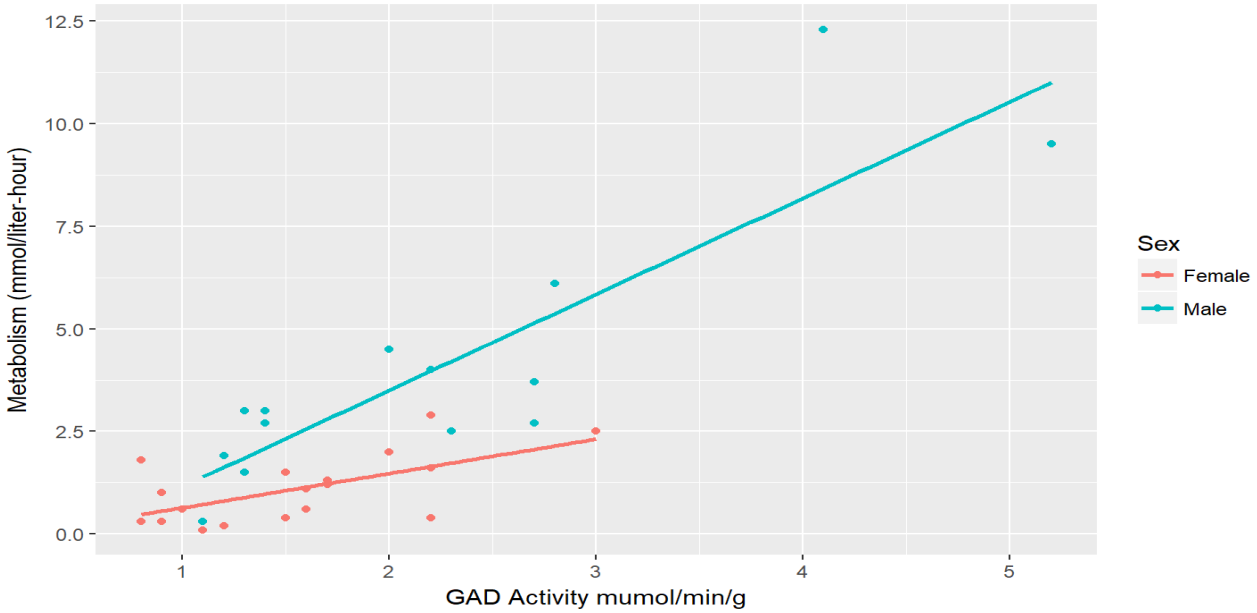
Data Exploration: Metabolism by Sex



Data Exploration: Metabolism by Alcoholism



Data Exploration: Metabolism by Gastric and Sex



Regression Model

```
##          term  estimate      p.value  conf.low  conf.high
## 1 (Intercept) -1.827084 4.282324e-03 -3.034298 -0.6198704
## 2     Gastric  2.281320 5.265880e-09  1.703790  2.8588506
```

Regression Model

```
##          term  estimate      p.value   conf.low  conf.high
## 1 (Intercept) -1.946646 7.957615e-04 -3.0097506 -0.8835407
## 2     Gastric  1.965578 4.238071e-08  1.4187842  2.5123714
## 3     SexMale  1.617444 3.649068e-03  0.5715028  2.6633860
```

Regression Model

```
##              term   estimate    p.value   conf.low conf.high
## 1     (Intercept) -0.1972691 0.80754593 -1.8405047  1.445967
## 2         Gastric  0.8369478 0.09471027 -0.1542610  1.828157
## 3         SexMale -0.9884969 0.36452374 -3.1851903  1.208197
## 4 Gastric:SexMale  1.5069236 0.01176490  0.3615822  2.652265
```

Regression Model

```
##              term   estimate      p.value   conf.low conf.high
## 1     (Intercept) -0.7504103 0.1682355862 -1.8364142 0.3355935
## 2         Gastric  1.1489074 0.0023716148  0.4433979 1.8544169
## 3 Gastric:SexMale  1.0422161 0.0001661676  0.5489507 1.5354815
```

Interpretation

- We have to consider the model that we have:

$$Metabolism = \beta_0 + \beta_1 Gastric + \beta_2 Gastric * Male$$

- Females:

$$Metabolism = \beta_0 + \beta_1 Gastric$$

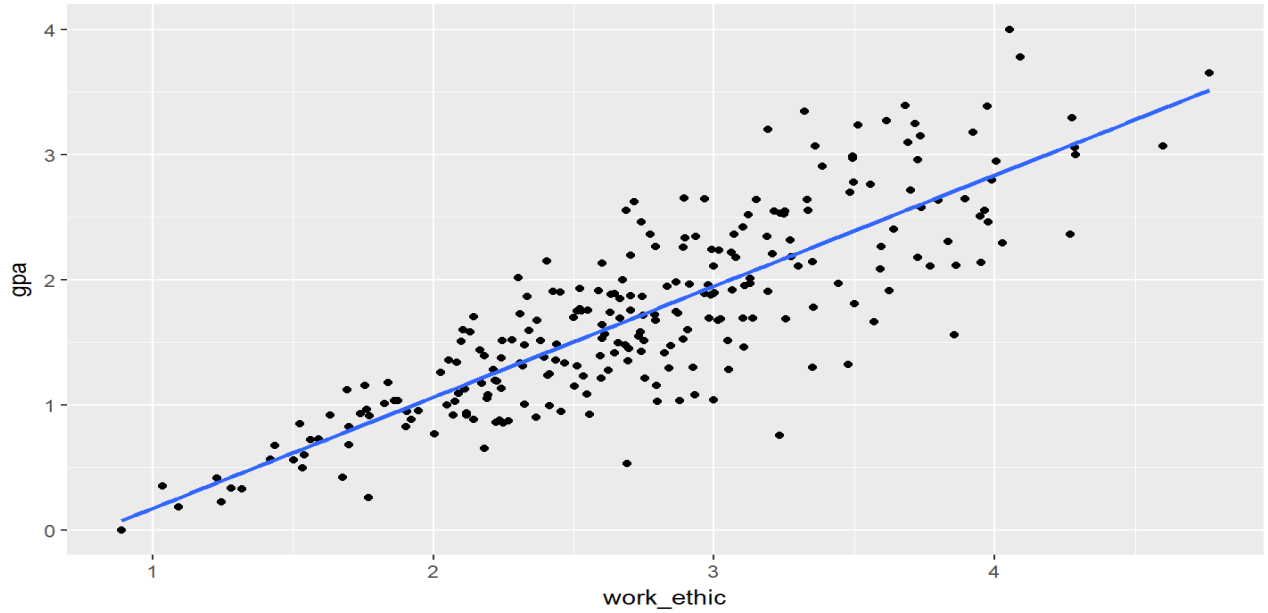
- Males:

$$Metabolism = \beta_0 + (\beta_1 + \beta_2) Gastric$$

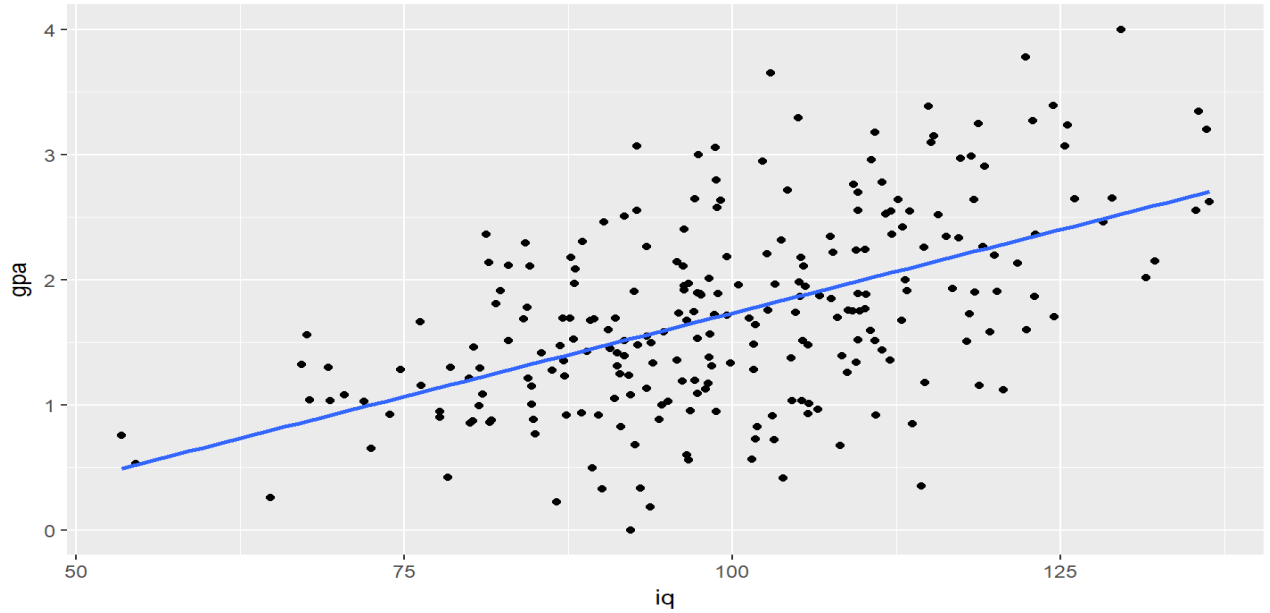
Continuous Interaction

- This is a little harder to figure out when it is happening.
- We have simulated data of GPA based on work ethic and GPA

Data Exploration



Data Exploration



Check for Interaction: Try Quantiles

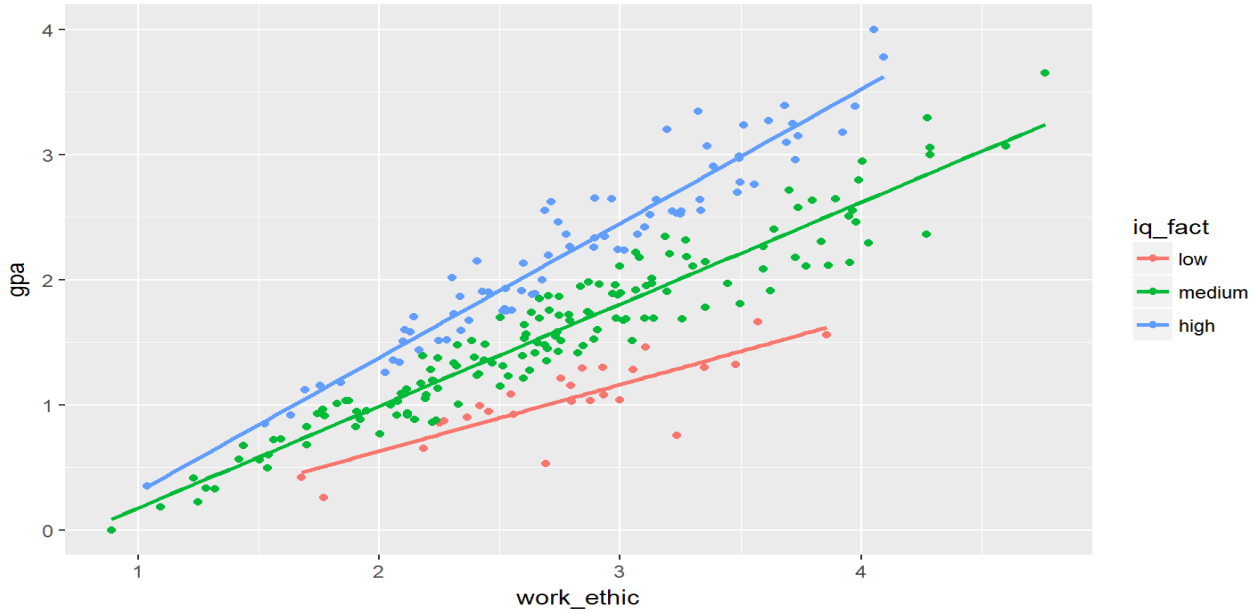
```
gpa_data %>%
  summarise(`0%`=quantile(iq, probs=0),
            `33%`=quantile(iq, probs=0.33),
            `66%`=quantile(iq, probs=0.66),
            `100%`=quantile(iq,probs=1))
```

```
##         0%      33%      66%     100%
## 1 53.44964 92.72525 106.0709 136.3098
```

Create a Factor

```
gpa_data <-gpa_data %>%
    mutate(iq_fact = cut(iq, 3, labels =c('low', 'medium', 'high')))
```

Graph Interaction



Linear Model

```
mod <- lm(data=gpa_data, gpa~work_ethic*iq)
tidy(mod, conf.int = T)[,-c(3:4)]
```

```
##           term      estimate       p.value     conf.low    conf.high
## 1   (Intercept) -0.8567222442 1.848167e-166 -0.880354261 -0.833090228
## 2    work_ethic -0.0004917628  9.046214e-01 -0.008566882  0.007583357
## 3            iq  0.0012941798  1.281443e-22  0.001058903  0.001529456
## 4 work_ethic:iq  0.0089423767 5.345782e-284  0.008861982  0.009022772
```