Case Study 5

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.3     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.4     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.0
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```
library(GGally)
```

```
Registered S3 method overwritten by 'GGally':
  method from
  +.gg   ggplot2
```

```
library(plotly)
```

```
Attaching package: 'plotly'

The following object is masked from 'package:ggplot2':

    last_plot
```

```
The following object is masked from 'package:stats':

    filter

The following object is masked from 'package:graphics':

    layout
```

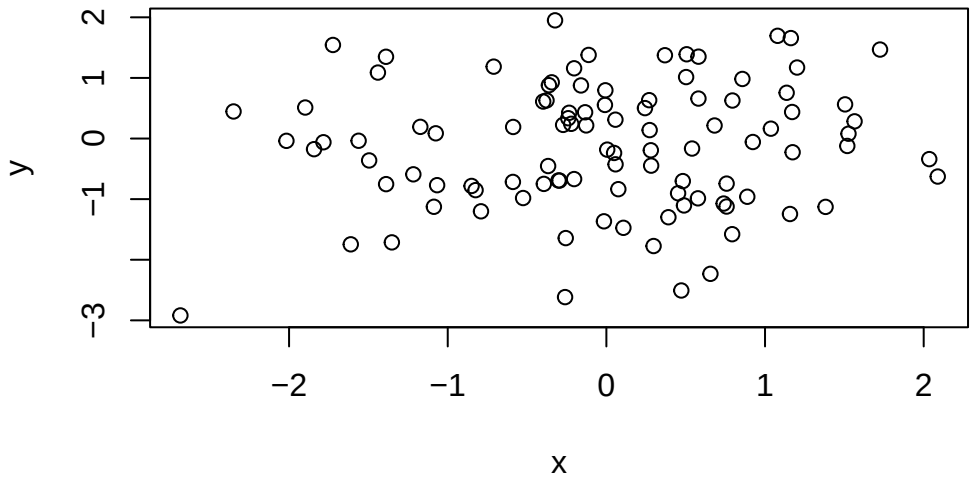
```r
library(DiagrammeR)
```

Goal

What is the effect of the study conditions on the relationship between the pre-test and posttest?

Normal Residuals

```r
normdat <- data.frame(x=rnorm(100),y=rnorm(100))
plot(y~x,data=normdat)
```



A little bit of TeX (LaTeX)

TEX and LATEX

Commands & Groups

- `$` and `$$`
- `\` – starts a command
- `{}` — gives you a gruop

Subscripts and superscripts

- Subscript `_` b_0, x_{ij}
- Superscript `^` R^2, X^{-1}

Greek letters and other commands

Greek letters are \ followed by the name θ, Θ

`\sqrt` $\sqrt{2\pi}$

Note log (in roman type)

Sums and Products

$$\sum_{i=1}^{N} x_i$$

Fractions

$\frac{1}{2}$

Bold and roman

`\text` to get roman `\textbf` or `\boldsymbol` to get bold.

Model Selection

Maximum Likelihood

Likelihood is the probability of the data given the model and parameters.

$$P(Y|\mathbf{X}, M, \theta) = \prod P(Y_i|x_i, M, \theta)$$

The *maximum likelihood* estimate of the parameters, $\hat{\theta}$ is the values of the parameters that maximizes the likelihood.

Often look at the *log likelihood*

$$L(Y|\mathbf{X}, M, \theta) = \sum \log P(Y_i|x_i, M, \theta)$$

For normal errors

$\log P(Y|X, \beta) \propto (Y - \hat{Y})^2$

For normal errors, MLE = Least Squares

Base and Saturated Models

Base Model: Needs to have all variables related to our research question.

Null Model: Just intercept

`post_scaled ~ pre_scaled + Cond_code` (compare to without `Cond_code`)

Other variables are to soak up variance.

Maximum or Saturated Model: Model will all variables we might consider.

`names(data)`

Forward Selection

Start with Minimum Model

Add variable with highest correlation with residuals.

Look at change in R^2

Stop when ~~no~~ minimal improvement.

In R, use `add1()` or `update()`

Reverse Selection

Start with saturated model.

Drop terms with non-significant slopes.

Stop just before fit becomes noticeably worse.

In R use `drop1()` or `update()`

Nested Models and F-test

Model 1 is nested in Model 2 $M_1 \subset M_2$ if every term in Model 1 is also in Model 2.

Difference in log likelihoods has approximately chi-squared. For normal model we can do an ANOVA F-test.

Stepwise Regression

Goes forwards and backwards, adding new variables and removing old ones. Usually defines an "F to enter" and "F to leave".

Evaluating Model Fit

```
normdat <- data.frame(y=rnorm(100),x=rnorm(100),
                      x1=rnorm(100),x2=rnorm(100))
mod1 <- lm(y~x,normdat)
mod2 <- lm(y~x+x1,normdat)
mod3 <- lm(y~x+x1+x2,normdat)
```

Adjusting R-squared

```
summary(mod1)$r.squared
```

```
[1] 0.004896345
```

```
summary(mod2)$r.squared
```

```
[1] 0.00677835
```

```
summary(mod3)$r.squared
```

```
[1] 0.02109727
```

```
summary(mod1)$adj.r.squared
```

```
[1] -0.005257774
```

```
summary(mod2)$adj.r.squared
```

```
[1] -0.01370045
```

```
summary(mod3)$adj.r.squared
```

```
[1] -0.009493439
```

Cross Validation

Split data into *training* and *test* data.

Do model search on training data

Do hypothesis testing of test data.

K-fold cross validation – break data into K groups. K times fit to $K-1$ groups and test on the remaining ones (average over the K times).

Leave one out (LOO) – N-fold cross validation.

Three stage – Split training data into training and test groups.

Deviance

Deviance is $-2 \log$ likelihood $= D$

Want to pick model with smallest deviance.

AIC

$$AIC = 2p + D$$

p is number of parameters (predictors).

Related to LOO Also called Mallow's C_p.

BIC

$$BIC = p\ln(N) + D$$

Related to minimum description length.

Also, DIC, WAIC, …

Box's Maxim

Box (1987). "Essentially, all models are wrong, but some are useful. However, the approximate nature of the model must always be borne in mind…"

Box (1976) "Since all models are wrong …" "… the scientist cannot obtain the 'correct' one by excessive elaboration." "… the scientist must be alert to what is importantly wrong."

"The map is not the terrain".

Occam's Window and Model Averaging

Adrian Raftery's idea:

Search for the best model, but keep the k best models.

In Bayesian framework, can create a posterior distribution over models.

(Weighted) Average of predictions is better than prediction from any single model.

ACED model for non-control students

```
library(tidyverse)
library(DescTools)
library(GGally)
library(plotly)
```

ACED Data

```
ACEDextract <- read_csv("ACED_extract1.csv",na="-999")
```

```
Rows: 290 Columns: 29
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (7): SubjID, Session, Cond_code, Sequencing, Feedback, Gender, Level_Code
dbl (22): Correct, Incorrect, Reamaining, ElapsedTime, Race, pre_scaled, pos...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
ACEDextract$Session <- factor(ACEDextract$Session)
ACEDextract$Cond_code <- factor(ACEDextract$Cond_code)
ACEDextract$Sequencing <- factor(ACEDextract$Sequencing)
ACEDextract$Feedback <- factor(ACEDextract$Feedback)
ACEDextract$Gender <- factor(ACEDextract$Gender)
ACEDextract$Race <- factor(ACEDextract$Race,1:8)
ACEDextract$Level_Code <- factor(ACEDextract$Level_Code)
```

```
summary(ACEDextract)
```

```
   SubjID              Session           Cond_code     Sequencing
 Length:290         c02    : 18   adaptive_acc :81   Adaptive:158
 Class :character   c03    : 18   adaptive_full:77   Linear  : 72
 Mode  :character   c04    : 18   control      :60   NA's    : 60
                    c13    : 18   linear_full  :72
                    c01    : 17
                    (Other):141
                    NA's   : 60
    Feedback      Correct        Incorrect       Reamaining      ElapsedTime
 Accuracy: 81   Min.   : 2.00   Min.   : 7.00   Min.   : 0.000   Min.   :  52
 Full    :149   1st Qu.:14.00   1st Qu.:29.00   1st Qu.: 0.000   1st Qu.:1542
 NA's    : 60   Median :21.00   Median :39.00   Median : 0.000   Median :1939
                Mean   :22.47   Mean   :37.13   Mean   : 3.274   Mean   :1894
                3rd Qu.:30.00   3rd Qu.:46.00   3rd Qu.: 0.000   3rd Qu.:2323
                Max.   :51.00   Max.   :59.00   Max.   :43.000   Max.   :2693
```

```
              NA's   :60      NA's   :60      NA's   :60       NA's   :100
   Gender         Race        Level_Code    pre_scaled     post_scaled
Female:144   7      :113   Academic:165   Min.   :27.00   Min.   :27.00
Male  :146   6      : 65   ELL     : 22   1st Qu.:44.00   1st Qu.:47.00
             3      : 43   Honors  : 38   Median :50.00   Median :54.00
             2      : 27   Part 1  : 30   Mean   :49.96   Mean   :54.58
             8      : 21   Part 2  :  8   3rd Qu.:57.00   3rd Qu.:61.00
             (Other): 18   Regular : 27   Max.   :78.00   Max.   :84.00
             NA's   :  3                  NA's   :2       NA's   :2
  Form_Order       EAP.sgp           EAP.cr           EAP.dt
Min.   :1.000   Min.   :0.0000   Min.   :0.4390   Min.   :0.4220
1st Qu.:1.000   1st Qu.:0.0020   1st Qu.:0.7252   1st Qu.:0.4230
Median :1.000   Median :0.1595   Median :1.3990   Median :0.4735
Mean   :1.486   Mean   :0.5464   Mean   :1.2752   Mean   :0.5802
3rd Qu.:2.000   3rd Qu.:0.9938   3rd Qu.:1.9378   3rd Qu.:0.7365
Max.   :2.000   Max.   :1.9980   Max.   :2.0000   Max.   :0.9520
                NA's   :60       NA's   :60       NA's   :60
    EAP.eg          EAP.exp          EAP.ext          EAP.mod
Min.   :0.0100   Min.   :0.00100   Min.   :0.0280   Min.   :0.0040
1st Qu.:0.0100   1st Qu.:0.00800   1st Qu.:0.8363   1st Qu.:0.0290
Median :0.0360   Median :0.02000   Median :1.4315   Median :0.1140
Mean   :0.4301   Mean   :0.09789   Mean   :1.3350   Mean   :0.3807
3rd Qu.:0.4560   3rd Qu.:0.06925   3rd Qu.:1.9658   3rd Qu.:0.6552
Max.   :1.9350   Max.   :1.59300   Max.   :2.0000   Max.   :1.6970
NA's   :60       NA's   :60        NA's   :60       NA's   :60
    EAP.rr          EAP.tab          EAP.vr          EAP.pic
Min.   :0.1040   Min.   :0.0180   Min.   :0.0290   Min.   :0.0110
1st Qu.:0.2440   1st Qu.:0.1273   1st Qu.:0.1492   1st Qu.:0.0350
Median :0.5195   Median :0.5095   Median :0.2750   Median :0.0830
Mean   :0.7271   Mean   :0.6885   Mean   :0.4615   Mean   :0.3114
3rd Qu.:0.9795   3rd Qu.:1.1473   3rd Qu.:0.6368   3rd Qu.:0.3415
Max.   :1.9740   Max.   :1.9740   Max.   :1.7980   Max.   :1.8950
NA's   :60       NA's   :60       NA's   :60       NA's   :60
   P.sgp..H         P.sgp..M         P.sgp..L
Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
1st Qu.:0.00000   1st Qu.:0.0020   1st Qu.:0.0810
Median :0.00000   Median :0.0575   Median :0.8415
Mean   :0.16140   Mean   :0.2235   Mean   :0.6150
3rd Qu.:0.07475   3rd Qu.:0.3975   3rd Qu.:0.9980
Max.   :0.99800   Max.   :0.9030   Max.   :1.0000
NA's   :60        NA's   :60       NA's   :60
```

Grab the non-control students

```
ACEDexp <- filter(ACEDextract,Cond_code!="Control") %>%
  na.omit() %>%
  mutate(Cond_code=factor(case_match(Cond_code,
                              "adaptive_acc"~"adaptive_acc",
                              "adaptive_full"~"adaptive_full",
                              "linear_full"~"linear_full")))
summary(ACEDexp$Cond_code)
```

```
adaptive_acc adaptive_full   linear_full
          59            64            64
```

```
summary(ACEDexp$Race)
```

```
1  2  3  4  5  6  7  8
2 21 23  3  6 42 81  9
```

Want to collapse 1, 4, 5, & 8 into other

```
ACEDexp <- mutate(ACEDexp,
                  Race=factor(case_match(as.numeric(Race),
                                         7~"Reference",
                                         6~"Focal1",
                                         3~"Focal2",
                                         2~"Focal3",
                                         c(1,4,5,8)~"Other")))
ACEDextract <- mutate(ACEDextract,
                  Race=factor(case_match(as.numeric(Race),
                                         7~"Reference",
                                         6~"Focal1",
                                         3~"Focal2",
                                         2~"Focal3",
                                         c(1,4,5,8)~"Other")))

summary(ACEDextract$Race)
```

```
  Focal1    Focal2    Focal3     Other Reference      NA's
      65        43        27        39       113         3
```

Minimum and Maximum Models

```
minMod <- post_scaled ~ pre_scaled + Sequencing + Feedback
names(ACEDexp)
```

```
 [1] "SubjID"      "Session"     "Cond_code"   "Sequencing"  "Feedback"
 [6] "Correct"     "Incorrect"   "Reamaining"  "ElapsedTime" "Gender"
[11] "Race"        "Level_Code"  "pre_scaled"  "post_scaled" "Form_Order"
[16] "EAP.sgp"     "EAP.cr"      "EAP.dt"      "EAP.eg"      "EAP.exp"
[21] "EAP.ext"     "EAP.mod"     "EAP.rr"      "EAP.tab"     "EAP.vr"
[26] "EAP.pic"     "P.sgp..H"    "P.sgp..M"    "P.sgp..L"
```

```
maxmodel <- post_scaled ~ pre_scaled + Sequencing + Feedback + Gender +
  Race + Level_Code + EAP.sgp + EAP.cr + EAP.dt + EAP.eg + EAP.ext
```

Method 1 – add

```
ACED1 <- lm(minMod,data=ACEDexp)
summary(ACED1)
```

```
Call:
lm(formula = minMod, data = ACEDexp)

Residuals:
     Min       1Q   Median       3Q      Max
-21.6934  -5.8873   0.3328   5.6252  20.4414

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      18.95130    3.55045   5.338 2.76e-07 ***
pre_scaled        0.69432    0.06639  10.458  < 2e-16 ***
SequencingLinear -0.69746    1.50707  -0.463    0.644
FeedbackFull      2.05725    1.53863   1.337    0.183
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 8.525 on 183 degrees of freedom
Multiple R-squared:  0.3779,    Adjusted R-squared:  0.3677
F-statistic: 37.06 on 3 and 183 DF,  p-value: < 2.2e-16
```

```
AIC(ACED1)
```

```
[1] 1338.125
```

```
BIC(ACED1)
```

```
[1] 1354.281
```

```
cor(residuals(ACED1),as.matrix(select(ACEDexp,where(is.numeric))))
```

```
       Correct  Incorrect Reamaining ElapsedTime   pre_scaled post_scaled
[1,] 0.4341955 -0.4356958 0.01867053   0.2809673 9.242009e-17   0.7887324
      Form_Order   EAP.sgp    EAP.cr    EAP.dt    EAP.eg   EAP.exp  EAP.ext
[1,] -0.04147316 0.4355552 0.3591018 0.4396468 0.3357967 0.2101227 0.411872
       EAP.mod    EAP.rr   EAP.tab    EAP.vr   EAP.pic  P.sgp..H  P.sgp..M
[1,] 0.4359186 0.2048478 0.3730077 0.3531783 0.3116015 0.3669614 0.2441425
      P.sgp..L
[1,] -0.434986
```

EAP.ext (extend sequence) has the highest correlation, so try adding this one next.

```
ACED2 <- update(ACED1,.~.+EAP.ext)
summary(ACED2)
```

```
Call:
lm(formula = post_scaled ~ pre_scaled + Sequencing + Feedback +
    EAP.ext, data = ACEDexp)

Residuals:
     Min       1Q   Median       3Q      Max
-19.5590  -5.4566  -0.2338   5.5469  19.6851
```

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      22.96324    3.15744   7.273 1.01e-11 ***
pre_scaled        0.42597    0.06831   6.236 3.06e-09 ***
SequencingLinear -1.12356    1.32209  -0.850    0.397
FeedbackFull      1.35442    1.35178   1.002    0.318
EAP.ext           7.83980    1.04548   7.499 2.74e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.472 on 182 degrees of freedom
Multiple R-squared:  0.5247,    Adjusted R-squared:  0.5143
F-statistic: 50.24 on 4 and 182 DF,  p-value: < 2.2e-16
```

```
AIC(ACED2)
```

```
[1] 1289.778
```

```
BIC(ACED2)
```

```
[1] 1309.164
```

```
anova(ACED1,ACED2)
```

```
Analysis of Variance Table

Model 1: post_scaled ~ pre_scaled + Sequencing + Feedback
Model 2: post_scaled ~ pre_scaled + Sequencing + Feedback + EAP.ext
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1    183 13300
2    182 10160  1    3139.2 56.232 2.736e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$SS_{all} = SS_{mod1} + SS_{mod2-mod1} + SS_e$$

Use `C(var,base=n)` to set group `n` as reference.

```r
ACED3 <- update(ACED2, .~.+C(Race,base=5))
summary(ACED3)
```

```
Call:
lm(formula = post_scaled ~ pre_scaled + Sequencing + Feedback +
    EAP.ext + C(Race, base = 5), data = ACEDexp)

Residuals:
     Min       1Q   Median       3Q      Max
-19.2866  -5.3476  -0.6474   5.7799  19.8037

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        23.03821    3.49995   6.582 5.01e-10 ***
pre_scaled          0.42115    0.07193   5.855 2.25e-08 ***
SequencingLinear   -1.12382    1.34196  -0.837    0.403
FeedbackFull        1.44319    1.36064   1.061    0.290
EAP.ext             7.76787    1.05937   7.333 7.65e-12 ***
C(Race, base = 5)1  0.49578    1.48368   0.334    0.739
C(Race, base = 5)2 -1.91016    1.79600  -1.064    0.289
C(Race, base = 5)3  2.72246    1.83947   1.480    0.141
C(Race, base = 5)4  0.19833    1.90583   0.104    0.917
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.465 on 178 degrees of freedom
Multiple R-squared:  0.536, Adjusted R-squared:  0.5152
F-statistic: 25.71 on 8 and 178 DF,  p-value: < 2.2e-16
```

```r
cat("AIC mod2=",AIC(ACED2),"mod3=",AIC(ACED3),"\n")
```

```
AIC mod2= 1289.778 mod3= 1293.279
```

```r
cat("BIC mod2=",BIC(ACED2),"mod3=",BIC(ACED3),"\n")
```

```
BIC mod2= 1309.164 mod3= 1325.59
```

```r
anova(ACED1,ACED2,ACED3)
```

```
Analysis of Variance Table

Model 1: post_scaled ~ pre_scaled + Sequencing + Feedback
Model 2: post_scaled ~ pre_scaled + Sequencing + Feedback + EAP.ext
Model 3: post_scaled ~ pre_scaled + Sequencing + Feedback + EAP.ext +
    C(Race, base = 5)
  Res.Df   RSS Df Sum of Sq       F   Pr(>F)
1    183 13300
2    182 10160  1   3139.25 56.3350 2.82e-12 ***
3    178  9919  4    241.55  1.0837   0.3661
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Start with saturated model and remove

```r
ACEDm1 <- lm(maxmodel,ACEDexp)
summary(ACEDm1)
```

```
Call:
lm(formula = maxmodel, data = ACEDexp)

Residuals:
     Min       1Q   Median       3Q      Max
-16.8197  -5.3102   0.5779   4.2794  15.9426

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       35.25885   17.77837   1.983 0.048970 *
pre_scaled         0.27637    0.07413   3.728 0.000263 ***
SequencingLinear  -0.88926    1.29559  -0.686 0.493422
FeedbackFull       1.09157    1.29477   0.843 0.400394
GenderMale         0.14343    1.08888   0.132 0.895361
RaceFocal2        -1.83724    2.01036  -0.914 0.362086
RaceFocal3         2.24463    1.99435   1.125 0.261986
RaceOther          0.22718    1.96143   0.116 0.907932
RaceReference     -0.41644    1.47701  -0.282 0.778331
```

```
Level_CodeELL     -4.61036    2.77264  -1.663 0.098216 .
Level_CodeHonors   3.55763    1.58947   2.238 0.026519 *
Level_CodePart 1  -3.85135    1.95773  -1.967 0.050800 .
Level_CodePart 2  -4.15953    3.16800  -1.313 0.190980
Level_CodeRegular -3.22166    2.36030  -1.365 0.174099
EAP.sgp            6.77707   11.65218   0.582 0.561607
EAP.cr            -2.24871    1.64407  -1.368 0.173211
EAP.dt             2.34809   42.94598   0.055 0.956462
EAP.eg            -1.65969    1.63890  -1.013 0.312669
EAP.ext            3.76698    1.55166   2.428 0.016251 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.995 on 168 degrees of freedom
Multiple R-squared:  0.6155,    Adjusted R-squared:  0.5743
F-statistic: 14.94 on 18 and 168 DF,  p-value: < 2.2e-16
```

```r
ACEDm2 <- update(ACEDm1,.~.-EAP.dt)
summary(ACEDm2)
```

```
Call:
lm(formula = post_scaled ~ pre_scaled + Sequencing + Feedback +
    Gender + Race + Level_Code + EAP.sgp + EAP.cr + EAP.eg +
    EAP.ext, data = ACEDexp)

Residuals:
     Min       1Q   Median       3Q      Max
-16.8631  -5.2994   0.5934   4.2785  15.9306

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       36.20573    4.00687   9.036 3.64e-16 ***
pre_scaled         0.27650    0.07388   3.743 0.000249 ***
SequencingLinear  -0.88500    1.28942  -0.686 0.493431
FeedbackFull       1.08903    1.29012   0.844 0.399789
GenderMale         0.13798    1.08110   0.128 0.898597
RaceFocal2        -1.84069    2.00343  -0.919 0.359526
RaceFocal3         2.25825    1.97290   1.145 0.253980
RaceOther          0.22771    1.95561   0.116 0.907441
RaceReference     -0.41046    1.46861  -0.279 0.780210
```

```
Level_CodeELL     -4.59918    2.75692  -1.668 0.097122 .
Level_CodeHonors   3.56273    1.58204   2.252 0.025611 *
Level_CodePart 1  -3.84375    1.94702  -1.974 0.049992 *
Level_CodePart 2  -4.14772    3.15129  -1.316 0.189891
Level_CodeRegular -3.21546    2.35061  -1.368 0.173152
EAP.sgp            7.40416    2.04983   3.612 0.000400 ***
EAP.cr            -2.21994    1.55300  -1.429 0.154719
EAP.eg            -1.67975    1.59258  -1.055 0.293052
EAP.ext            3.79207    1.47787   2.566 0.011159 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.975 on 169 degrees of freedom
Multiple R-squared:  0.6155,    Adjusted R-squared:  0.5768
F-statistic: 15.91 on 17 and 169 DF,  p-value: < 2.2e-16
```

Stepwise Regression

possible to do both forwards and backwards

```
ACEDstep <- step(ACED2,list(lower=minMod,upper=maxmodel),
                 trace=3)
```

```
Start:  AIC=757.09
post_scaled ~ pre_scaled + Sequencing + Feedback + EAP.ext

             Df Sum of Sq     RSS    AIC
+ EAP.sgp     1    999.50  9161.0 739.73
+ EAP.dt      1    995.14  9165.4 739.82
+ EAP.eg      1    475.58  9685.0 750.13
+ Level_Code  5    832.90  9327.6 751.10
+ EAP.cr      1    144.44 10016.1 756.42
<none>                    10160.5 757.09
+ Gender      1     15.24 10145.3 758.81
+ Race        4    241.55  9919.0 760.60
- EAP.ext     1   3139.25 13299.8 805.44

Step:  AIC=739.73
post_scaled ~ pre_scaled + Sequencing + Feedback + EAP.ext +
    EAP.sgp
```

```
             Df Sum of Sq     RSS    AIC
+ Level_Code  5    632.54  8528.5 736.35
<none>                     9161.0 739.73
+ EAP.eg      1     23.75  9137.3 741.25
+ EAP.cr      1     21.16  9139.9 741.30
+ Gender      1     10.40  9150.6 741.52
+ EAP.dt      1      1.30  9159.7 741.70
+ Race        4    253.48  8907.6 742.48
- EAP.ext     1    647.80  9808.8 750.51
- EAP.sgp     1    999.50 10160.5 757.09

Step:  AIC=736.35
post_scaled ~ pre_scaled + Sequencing + Feedback + EAP.ext +
    EAP.sgp + Level_Code

             Df Sum of Sq    RSS    AIC
<none>                    8528.5 736.35
+ EAP.cr      1     77.67 8450.8 736.64
+ EAP.eg      1     43.62 8484.9 737.39
+ EAP.dt      1      0.04 8528.5 738.35
+ Gender      1      0.01 8528.5 738.35
- Level_Code  5    632.54 9161.0 739.73
+ Race        4    164.11 8364.4 740.72
- EAP.ext     1    314.09 8842.6 741.11
- EAP.sgp     1    799.14 9327.6 751.10
```

```
ACEDstep
```

```
Call:
lm(formula = post_scaled ~ pre_scaled + Sequencing + Feedback +
    EAP.ext + EAP.sgp + Level_Code, data = ACEDexp)

Coefficients:
      (Intercept)         pre_scaled   SequencingLinear       FeedbackFull
          34.1289             0.2902            -1.2780             0.9910
          EAP.ext            EAP.sgp     Level_CodeELL   Level_CodeHonors
           3.3676             4.7765            -3.4958             3.2858
 Level_CodePart 1   Level_CodePart 2  Level_CodeRegular
          -4.2284            -3.4463            -3.5683
```

```r
summary(ACEDstep)
```

```
Call:
lm(formula = post_scaled ~ pre_scaled + Sequencing + Feedback +
    EAP.ext + EAP.sgp + Level_Code, data = ACEDexp)

Residuals:
     Min       1Q   Median       3Q      Max
-18.1533  -5.0814   0.2509   4.3575  15.2667

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       34.12893    3.78576   9.015 3.27e-16 ***
pre_scaled         0.29019    0.06962   4.168 4.81e-05 ***
SequencingLinear  -1.27801    1.23932  -1.031   0.3039
FeedbackFull       0.99103    1.26789   0.782   0.4355
EAP.ext            3.36758    1.32273   2.546   0.0118 *
EAP.sgp            4.77646    1.17618   4.061 7.35e-05 ***
Level_CodeELL     -3.49579    2.64868  -1.320   0.1886
Level_CodeHonors   3.28579    1.55626   2.111   0.0362 *
Level_CodePart 1  -4.22840    1.81026  -2.336   0.0206 *
Level_CodePart 2  -3.44634    3.06876  -1.123   0.2630
Level_CodeRegular -3.56829    2.26626  -1.575   0.1172
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.961 on 176 degrees of freedom
Multiple R-squared:  0.6011,    Adjusted R-squared:  0.5784
F-statistic: 26.52 on 10 and 176 DF,  p-value: < 2.2e-16
```

Earnings Data

```r
earnings <- read_csv("https://raw.githubusercontent.com/avehtari/ROS-Examples/master/Earni
```

```
Rows: 1816 Columns: 15
-- Column specification --------------------------------------------------------
Delimiter: ","
```

```
chr  (1): ethnicity
dbl (14): height, weight, male, earn, earnk, education, mother_education, fa...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
summary(earnings)
```

```
     height          weight           male             earn        
 Min.   :57.00   Min.   : 80.0   Min.   :0.0000   Min.   :     0  
 1st Qu.:64.00   1st Qu.:130.0   1st Qu.:0.0000   1st Qu.:  6000  
 Median :66.00   Median :150.0   Median :0.0000   Median : 16000  
 Mean   :66.57   Mean   :156.3   Mean   :0.3717   Mean   : 21147  
 3rd Qu.:69.25   3rd Qu.:180.0   3rd Qu.:1.0000   3rd Qu.: 27000  
 Max.   :82.00   Max.   :342.0   Max.   :1.0000   Max.   :400000  
                 NA's   :27                                       
     earnk          ethnicity           education     mother_education
 Min.   :  0.00   Length:1816        Min.   : 2.00   Min.   : 3.00   
 1st Qu.:  6.00   Class :character   1st Qu.:12.00   1st Qu.:12.00   
 Median : 16.00   Mode  :character   Median :12.00   Median :13.00   
 Mean   : 21.15                      Mean   :13.24   Mean   :13.61   
 3rd Qu.: 27.00                      3rd Qu.:15.00   3rd Qu.:16.00   
 Max.   :400.00                      Max.   :18.00   Max.   :99.00   
                                     NA's   :2       NA's   :244     
 father_education      walk          exercise        smokenow      
 Min.   : 3.00    Min.   :1.000   Min.   :1.000   Min.   :1.000  
 1st Qu.:12.00    1st Qu.:3.000   1st Qu.:1.000   1st Qu.:1.000  
 Median :13.00    Median :6.000   Median :2.000   Median :2.000  
 Mean   :13.65    Mean   :5.303   Mean   :3.049   Mean   :1.745  
 3rd Qu.:16.00    3rd Qu.:8.000   3rd Qu.:5.000   3rd Qu.:2.000  
 Max.   :99.00    Max.   :8.000   Max.   :7.000   Max.   :2.000  
 NA's   :295                                      NA's   :1      
     tense            angry             age       
 Min.   :0.000   Min.   :0.000   Min.   :18.00  
 1st Qu.:0.000   1st Qu.:0.000   1st Qu.:29.00  
 Median :0.000   Median :0.000   Median :39.00  
 Mean   :1.421   Mean   :1.421   Mean   :42.93  
 3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:56.00  
 Max.   :7.000   Max.   :7.000   Max.   :91.00  
 NA's   :1       NA's   :1                      
```

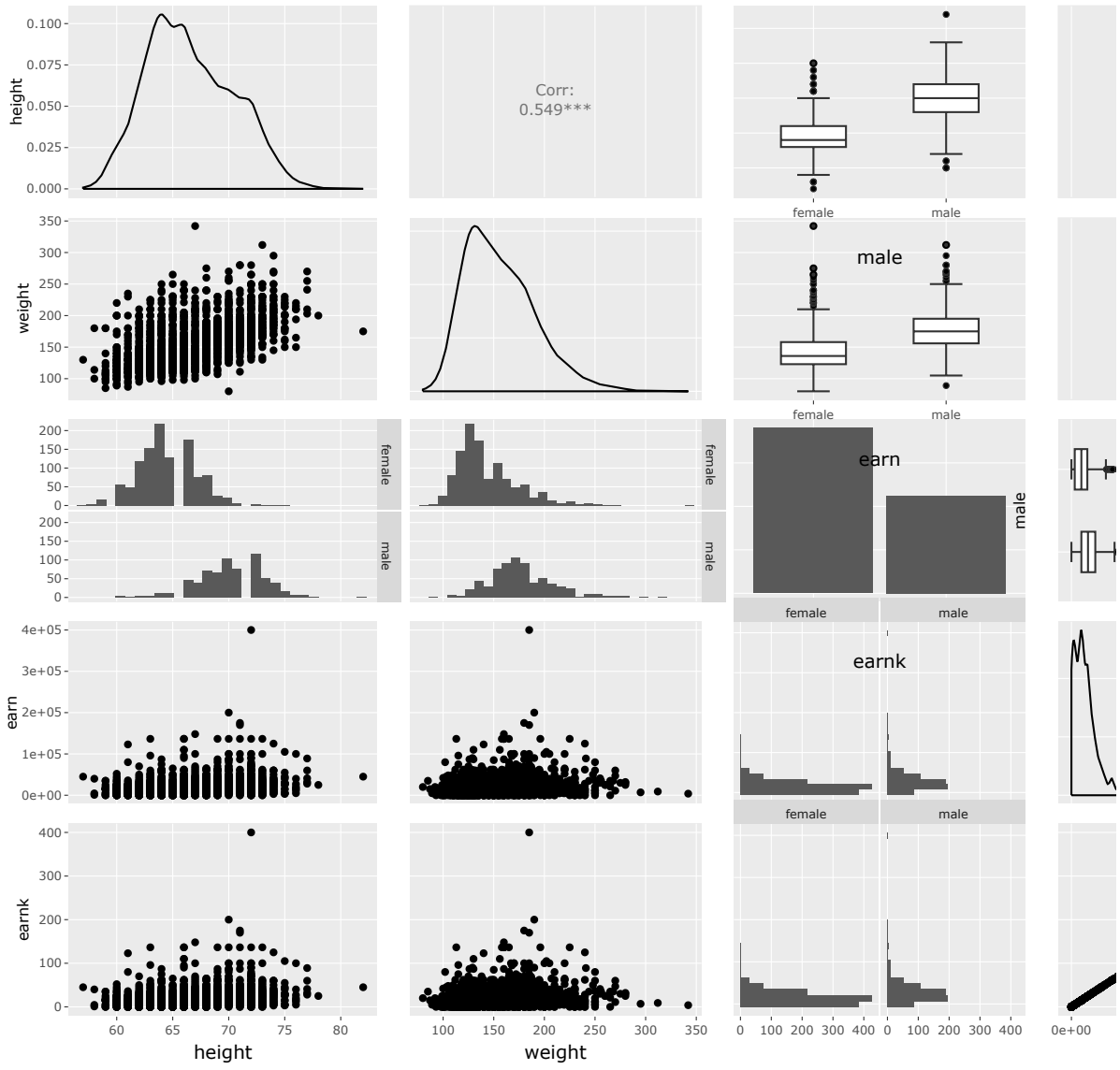
```r
earnings$male <- factor(earnings$male,labels=c("female","male"))
earnings$ethnicity <- factor(earnings$ethnicity)
earnings$smokenow <- factor(earnings$smokenow)
summary(earnings)
```

```
    height          weight          male           earn
Min.   :57.00   Min.   : 80.0   female:1141   Min.   :     0
1st Qu.:64.00   1st Qu.:130.0   male  : 675   1st Qu.:  6000
Median :66.00   Median :150.0                 Median : 16000
Mean   :66.57   Mean   :156.3                 Mean   : 21147
3rd Qu.:69.25   3rd Qu.:180.0                 3rd Qu.: 27000
Max.   :82.00   Max.   :342.0                 Max.   :400000
                NA's   :27
    earnk           ethnicity      education     mother_education
Min.   :  0.00   Black   : 180   Min.   : 2.00   Min.   : 3.00
1st Qu.:  6.00   Hispanic: 104   1st Qu.:12.00   1st Qu.:12.00
Median : 16.00   Other   :  38   Median :12.00   Median :13.00
Mean   : 21.15   White   :1494   Mean   :13.24   Mean   :13.61
3rd Qu.: 27.00                   3rd Qu.:15.00   3rd Qu.:16.00
Max.   :400.00                   Max.   :18.00   Max.   :99.00
                                 NA's   :2       NA's   :244
father_education      walk          exercise     smokenow        tense
Min.   : 3.00    Min.   :1.000   Min.   :1.000   1   : 462   Min.   :0.000
1st Qu.:12.00    1st Qu.:3.000   1st Qu.:1.000   2   :1353   1st Qu.:0.000
Median :13.00    Median :6.000   Median :2.000   NA's:   1   Median :0.000
Mean   :13.65    Mean   :5.303   Mean   :3.049               Mean   :1.421
3rd Qu.:16.00    3rd Qu.:8.000   3rd Qu.:5.000               3rd Qu.:2.000
Max.   :99.00    Max.   :8.000   Max.   :7.000               Max.   :7.000
NA's   :295                                                 NA's   :1
    angry            age
Min.   :0.000   Min.   :18.00
1st Qu.:0.000   1st Qu.:29.00
Median :0.000   Median :39.00
Mean   :1.421   Mean   :42.93
3rd Qu.:2.000   3rd Qu.:56.00
Max.   :7.000   Max.   :91.00
NA's   :1
```

```r
highlight_key(earnings) %>%
  GGally::ggpairs(columns=1:5) %>%
```

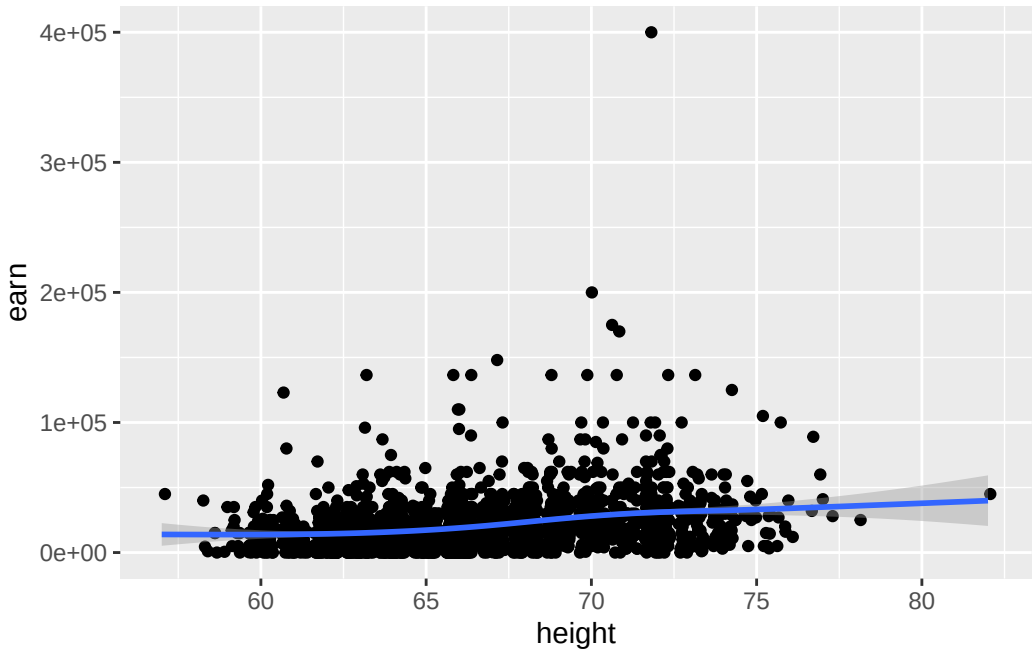
```
    ggplotly() %>%
    highlight("plotly_selected")
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
Removed 27 rows containing missing values

Warning: Removed 27 rows containing non-finite values (`stat_density()`).

`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

Warning: Removed 27 rows containing non-finite values (`stat_bin()`).

Warning: Can only have one: highlight

Warning: Removed 27 rows containing non-finite values (`stat_boxplot()`).

`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

Warning in subplot(columnList, nrows = p$nrow, margin = 0.01, shareX = TRUE, :
Must have a consistent number of axes per 'subplot' to share them.

Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
Removed 27 rows containing missing values

Warning: Can only have one: highlight

Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
Removed 27 rows containing missing values

Warning: Can only have one: highlight

PhantomJS not found. You can install it with webshot::install_phantomjs(). If it is installe
Setting the `off` event (i.e., 'plotly_deselect') to match the `on` event (i.e., 'plotly_sele
```



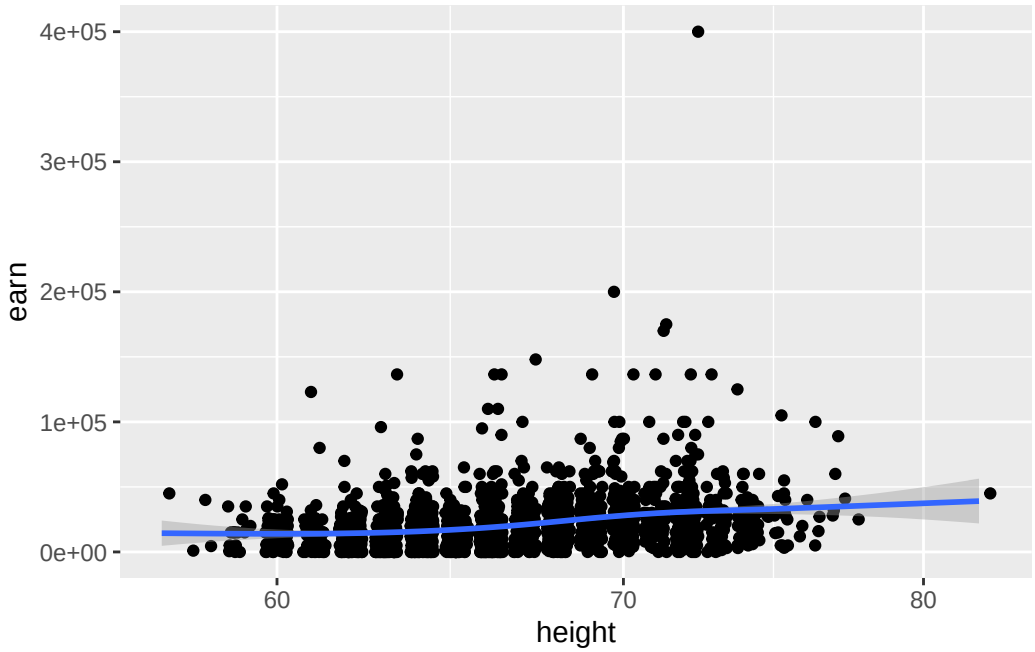

```
ggplot(earnings,aes(y=earn,x=height)) + geom_point(position="jitter") + geom_smooth()
```

```
`geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```




```
ggplot(earnings,aes(y=earn,x=height)) + scale_x_log10() + geom_point(position="jitter") +
```

```
`geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



```r
learn <- lm(log(earn+1) ~ height, data=earnings, na.action=na.omit)
summary(learn)
```

```
Call:
lm(formula = log(earn + 1) ~ height, data = earnings, na.action = na.omit)

Residuals:
    Min      1Q  Median      3Q     Max
-9.9506  0.0256  0.8789  1.5725  4.0685

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.03600    1.22056  -3.307 0.000963 ***
height       0.19160    0.01831  10.467  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.988 on 1814 degrees of freedom
Multiple R-squared:  0.05695,   Adjusted R-squared:  0.05643
F-statistic: 109.6 on 1 and 1814 DF,  p-value: < 2.2e-16
```

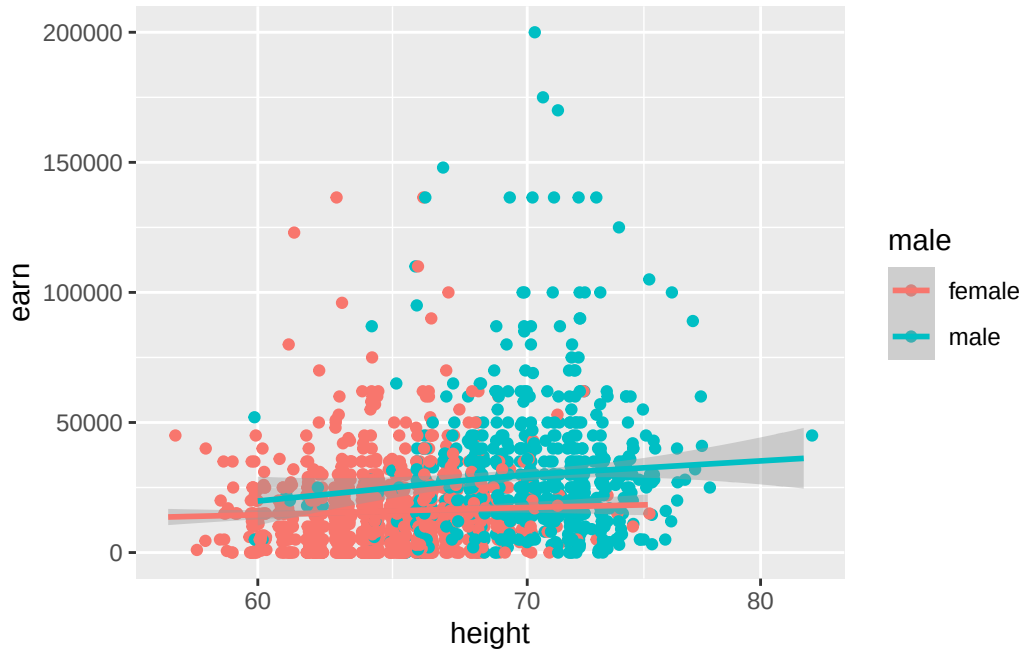
```r
lny = -4 + .19*c(66,67)
exp(lny)
```

```
[1] 5115.344 6185.728
```

Male–Female Interaction

```r
ggplot(earnings[earnings$earnk<350,],aes(y=earn,x=height,color=male)) + scale_x_log10() +
```

```
`geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```




```
learng <- lm(log(earn+1) ~ height + male, data=earnings, na.action=na.omit)
summary(learng)
```

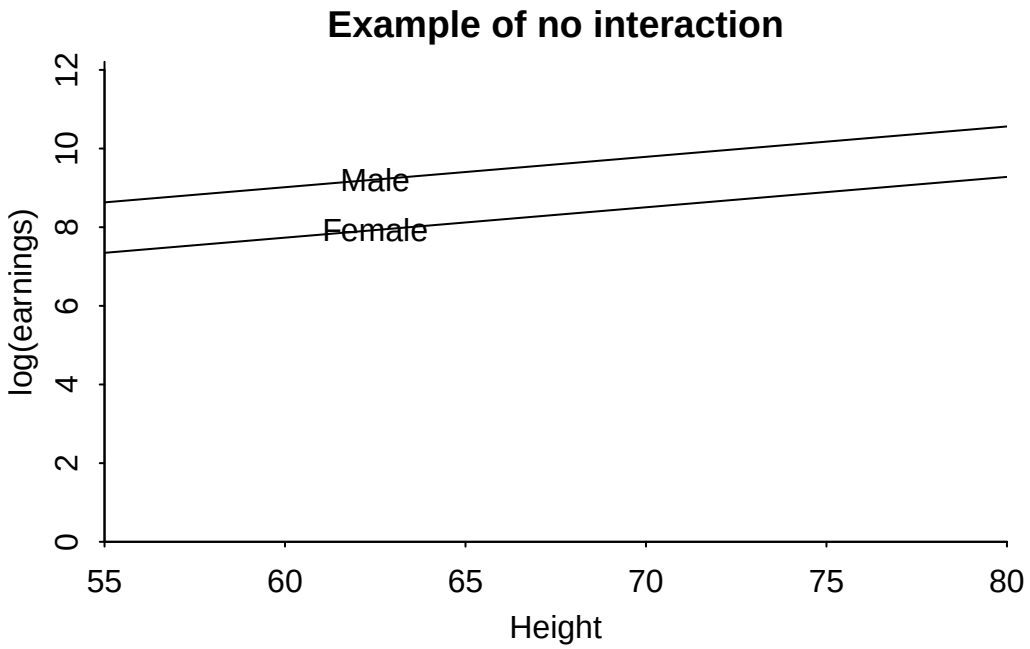
```
Call:
lm(formula = log(earn + 1) ~ height + male, data = earnings,
    na.action = na.omit)

Residuals:
     Min       1Q   Median       3Q      Max
-10.0210  -0.0155   0.7896   1.6499   3.9084

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.10018    1.65325   1.875  0.06093 .
height       0.07723    0.02560   3.017  0.00259 **
malemale     1.28267    0.20294   6.320 3.27e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.957 on 1813 degrees of freedom
Multiple R-squared:  0.07729,   Adjusted R-squared:  0.07627
```

```
F-statistic: 75.93 on 2 and 1813 DF,  p-value: < 2.2e-16
```

```
learngi <- lm(log(earn+1) ~ height * male, data=earnings, na.action=na.omit)
summary(learngi)
```

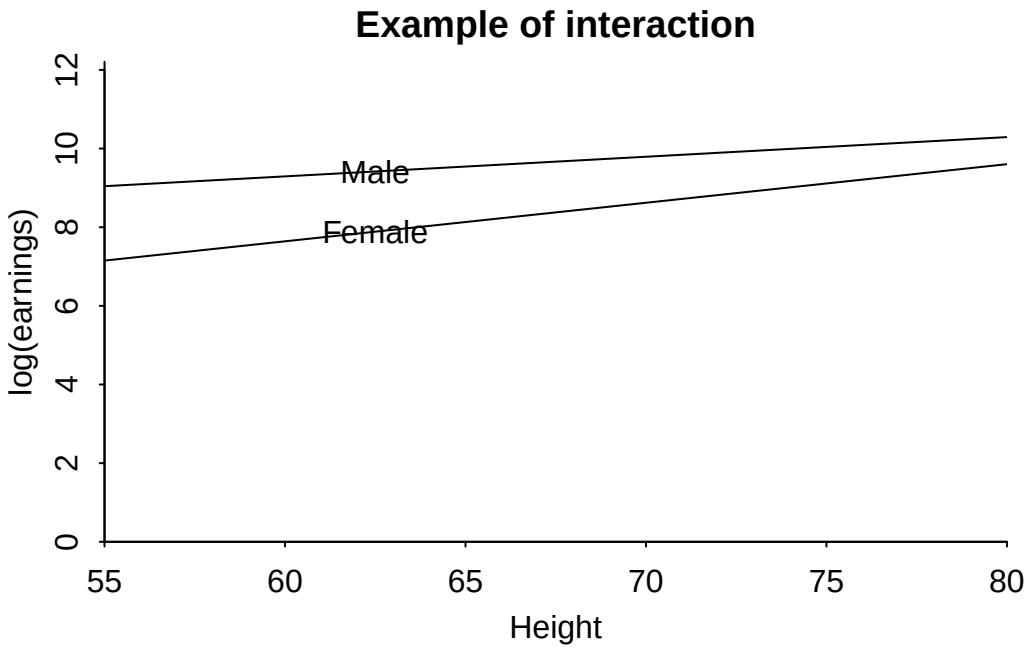
```

Call:
lm(formula = log(earn + 1) ~ height * male, data = earnings,
    na.action = na.omit)

Residuals:
    Min      1Q  Median      3Q     Max
-9.9415 -0.0061  0.7785  1.6670  3.9811

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.75623    2.19395   0.800  0.42353
height          0.09808    0.03399   2.885  0.00396 **
malemale        4.54088    3.50230   1.297  0.19495
height:malemale -0.04815    0.05167  -0.932  0.35153
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.957 on 1812 degrees of freedom
Multiple R-squared:  0.07773,   Adjusted R-squared:  0.0762
F-statistic:  50.9 on 3 and 1812 DF,  p-value: < 2.2e-16
```

```
oldpar <- par(mar=c(3,3,2,1), mgp=c(1.7,.5,0), tck=-.01)
plot(c(55,80),c(0,log(200000)), type="n", xaxs="i", yaxs="i",
  xlab="Height", ylab="log(earnings)", bty="l", main="Example of no interaction")
lines(c(55,80),coef(learng)['(Intercept)']+coef(learng)["malemale"] +
        c(55,80)*coef(learng)["height"])
lines(c(55,80),coef(learng)['(Intercept)']+
        c(55,80)*coef(learng)["height"])
text(62.5, coef(learng)['(Intercept)']+coef(learng)["malemale"] +
        62.5*coef(learng)["height"], "Male")
text(62.5, coef(learng)['(Intercept)']+
        62.5*coef(learng)["height"], "Female")
```




```
par(oldpar)
```

```
oldpar <- par(mar=c(3,3,2,1), mgp=c(1.7,.5,0), tck=-.01)
plot(c(55,80),c(0,log(200000)), type="n", xaxs="i", yaxs="i",
  xlab="Height", ylab="log(earnings)", bty="l", main="Example of interaction")
lines(c(55,80),coef(learngi)['(Intercept)']+coef(learngi)["malemale"] +
        c(55,80)*(coef(learngi)["height"]+coef(learngi)["height:malemale"]))
lines(c(55,80),coef(learngi)['(Intercept)']+
        c(55,80)*coef(learngi)["height"])
text(62.5, coef(learngi)['(Intercept)']+coef(learngi)["malemale"] +
        62.5*(coef(learngi)["height"]+coef(learngi)["height:malemale"]), "Male")
text(62.5, coef(learngi)['(Intercept)']+
        62.5*coef(learngi)["height"], "Female")
```

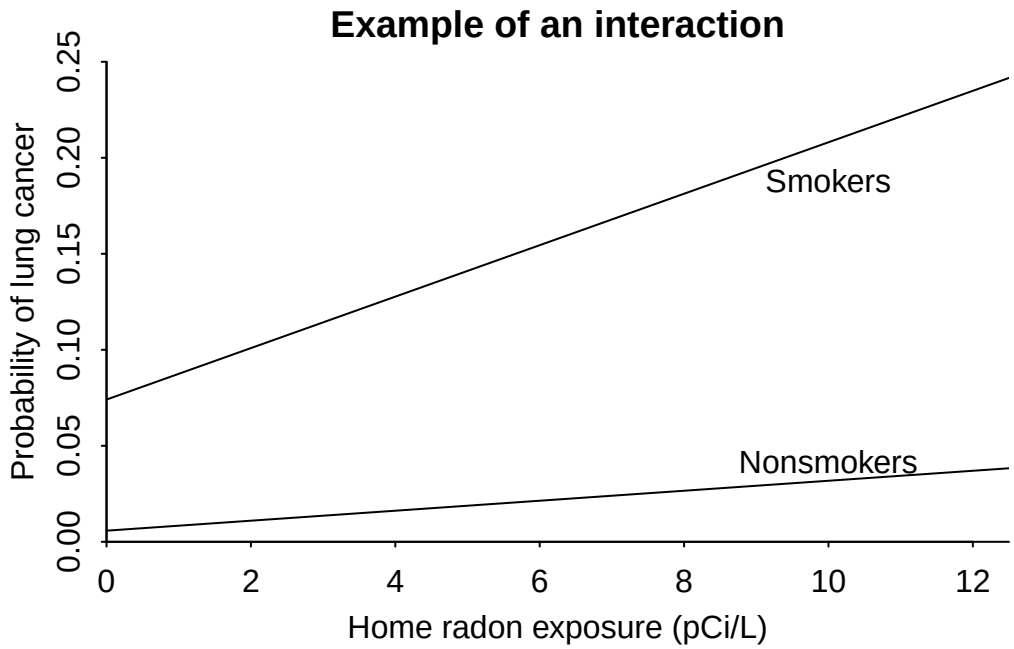



```
par(oldpar)
```

Asbestos and cancer Example

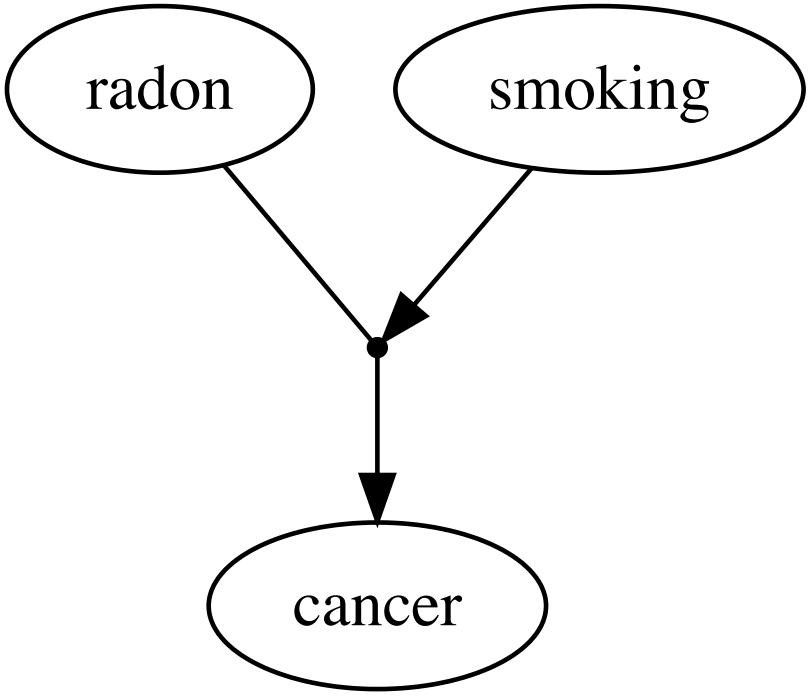
Example from Gelman, Hill & Vehtari, Chapter 1.

```
oldpar <- par(mar=c(3,3,2,1), mgp=c(1.7,.5,0), tck=-.01)
plot(c(0,12.5),c(0,.25), type="n", xaxs="i", yaxs="i",
  xlab="Home radon exposure (pCi/L)", ylab="Probability of lung cancer", bty="l", main="Ex
lines(c(0,20),.07409+c(0,20)*.0134)
lines(c(0,20),.00579+c(0,20)*.0026)
text(10, .07409+10*.0134 - .02, "Smokers")
text(10, .00579+10*.0026 + .01, "Nonsmokers")
```

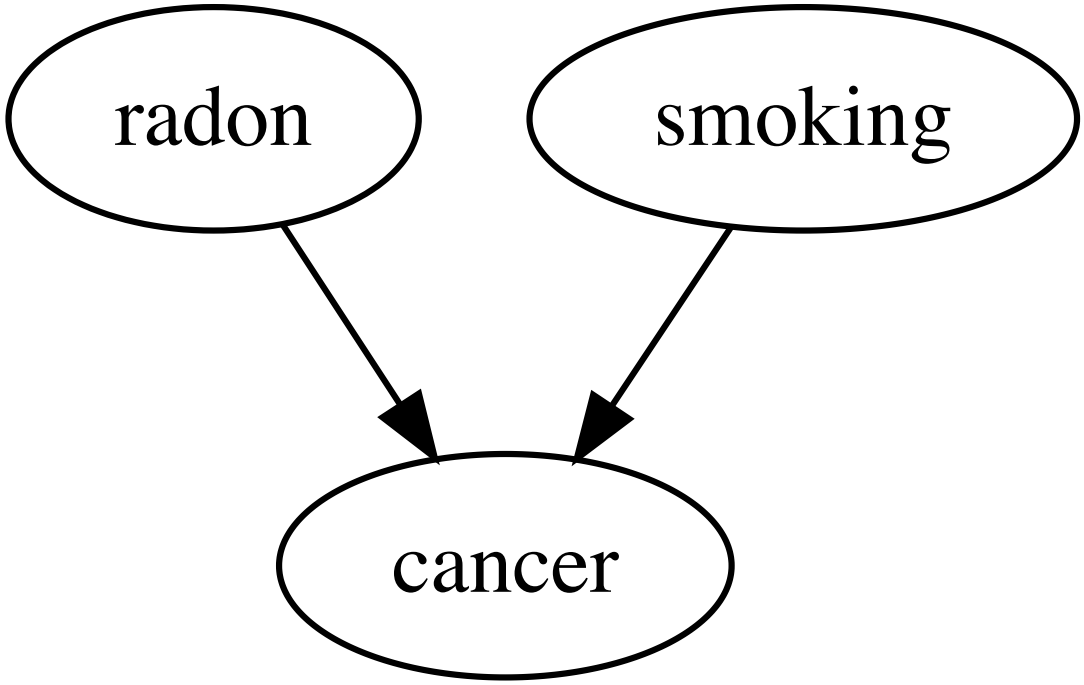



```
par(oldpar)
```

```
DiagrammeR::grViz('
digraph rs {
  i [label="" shape="point"]
  radon -> i [arrowhead="none"]
  i -> cancer
  smoking -> i
}
')
```




```
DiagrammeR::grViz('
digraph rs {

  radon -> cancer
  smoking -> cancer
}
')
```



This is compatible with `cancer ~ asbestos + smoking` and `cancer ~ asbestos * smoking`

This is a *moderator*

Moderators and Mediators

Path Diagram

Nodes (vertices) represent variables.

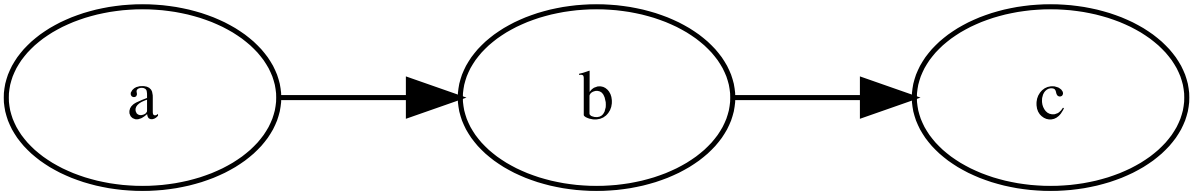
Arrows go from predictor to predicted; often used to represent hypothesized causes.

Mediation Model

A mediator goes in between

```
DiagrammeR::grViz('
digraph abc {
  rankdir="LR"
  a->b->c
```

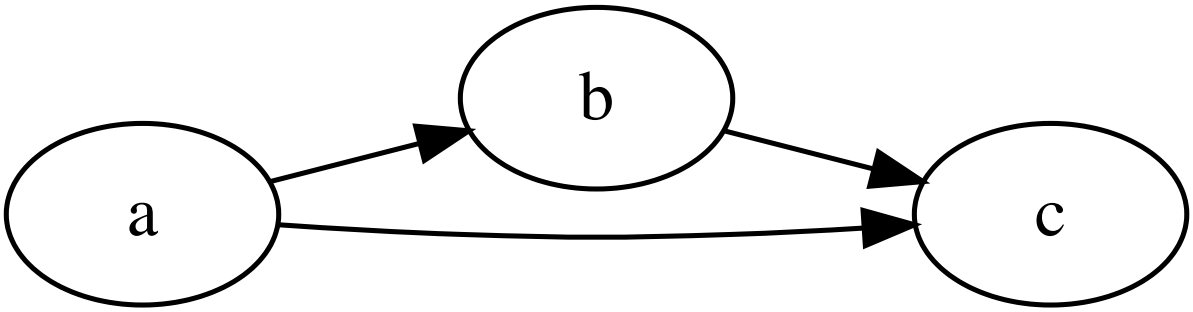
```
}
')
```



If b is removed then $a \rightarrow c$

Partial mediation

```
DiagrammeR::grViz('
digraph abc {
  rankdir="LR"
  a->b->c
  a->c
}
')
```

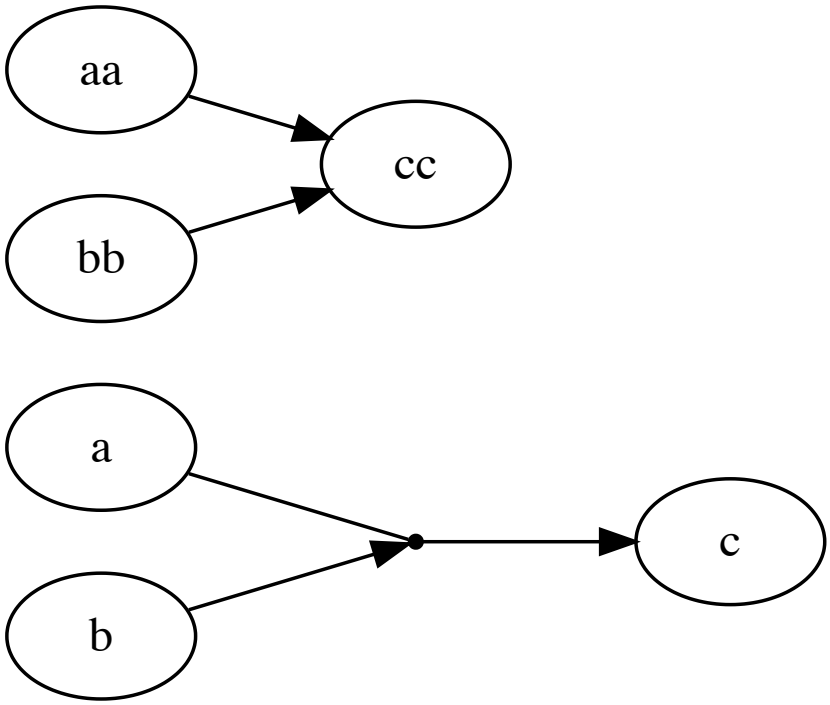


Moderators

Moderators affect the strength of the relationship between two other variables:

```
DiagrammeR::grViz('
digraph rs {
  rankdir="LR"
  i [label="" shape="point"]
  a -> i [arrowhead="none"]
  i -> c
```

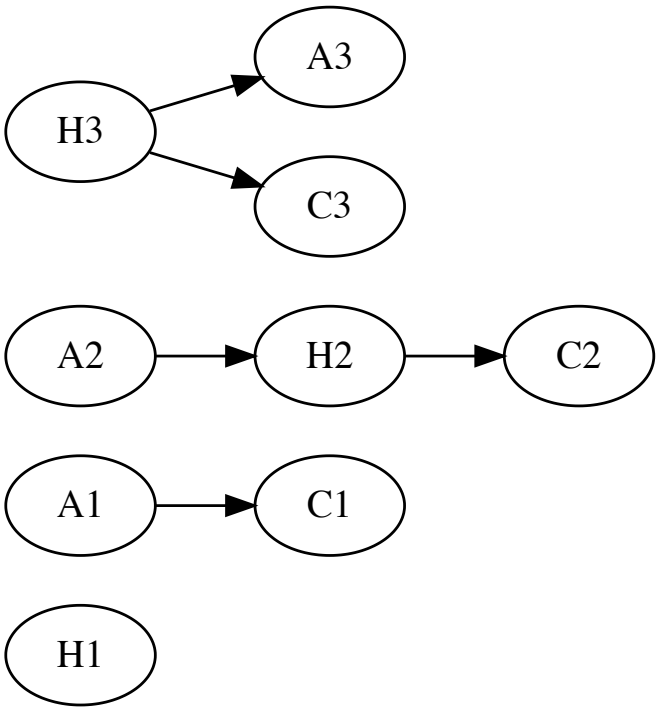
```
  b -> i

  aa -> cc
  bb -> cc
}
')
```



Hidden Variables

```
DiagrammeR::grViz('
digraph hidden {
  rankdir="LR"
  subgraph h1 {
    H1
    A1 -> C1
  }
  subgraph h2 {
    A2 -> H2
    H2 -> C2
  }
```

```
  subgraph h3 {
    H3 -> A3
    H3 -> C3
  }
}
')
```

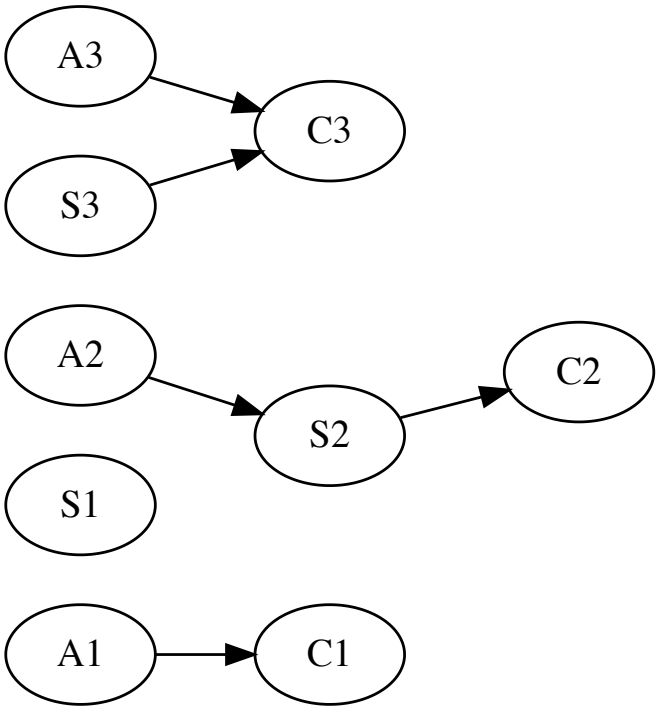


All three result in conclusions $A \rightarrow C$

Selection effect

```
DiagrammeR::grViz('
digraph se {
  rankdir="LR"
  subgraph s1 {
    A1 -> C1
    S1
  }
  subgraph s2 {
```

```
    A2 -> C2 [style="invis"]
    A2 -> S2 -> C2
  }
  subgraph h3 {
    A3 -> C3
    S3 -> C3
  }
}
')
```



Model Search

```
names(earnings)
```

```
 [1] "height"           "weight"           "male"             "earn"
 [5] "earnk"            "ethnicity"        "education"        "mother_education"
 [9] "father_education" "walk"             "exercise"         "smokenow"
[13] "tense"            "angry"            "age"
```

```r
minMod <- log(earn+1) ~ male*education
maxMod <- log(earn+1) ~ male*education + male*age + height + ethnicity + exercise + smoken
```

Standardize Variables

```r
summary(earnings)
```

```
    height          weight          male           earn
Min.   :57.00   Min.   : 80.0   female:1141   Min.   :     0
1st Qu.:64.00   1st Qu.:130.0   male  : 675   1st Qu.:  6000
Median :66.00   Median :150.0                 Median : 16000
Mean   :66.57   Mean   :156.3                 Mean   : 21147
3rd Qu.:69.25   3rd Qu.:180.0                 3rd Qu.: 27000
Max.   :82.00   Max.   :342.0                 Max.   :400000
                NA's   :27
    earnk           ethnicity      education     mother_education
Min.   :  0.00   Black   : 180   Min.   : 2.00   Min.   : 3.00
1st Qu.:  6.00   Hispanic: 104   1st Qu.:12.00   1st Qu.:12.00
Median : 16.00   Other   :  38   Median :12.00   Median :13.00
Mean   : 21.15   White   :1494   Mean   :13.24   Mean   :13.61
3rd Qu.: 27.00                   3rd Qu.:15.00   3rd Qu.:16.00
Max.   :400.00                   Max.   :18.00   Max.   :99.00
                                 NA's   :2       NA's   :244
father_education      walk          exercise     smokenow        tense
Min.   : 3.00    Min.   :1.000   Min.   :1.000   1   : 462   Min.   :0.000
1st Qu.:12.00    1st Qu.:3.000   1st Qu.:1.000   2   :1353   1st Qu.:0.000
Median :13.00    Median :6.000   Median :2.000   NA's:   1   Median :0.000
Mean   :13.65    Mean   :5.303   Mean   :3.049               Mean   :1.421
3rd Qu.:16.00    3rd Qu.:8.000   3rd Qu.:5.000               3rd Qu.:2.000
Max.   :99.00    Max.   :8.000   Max.   :7.000               Max.   :7.000
NA's   :295                                                 NA's   :1
    angry            age
Min.   :0.000   Min.   :18.00
1st Qu.:0.000   1st Qu.:29.00
Median :0.000   Median :39.00
Mean   :1.421   Mean   :42.93
3rd Qu.:2.000   3rd Qu.:56.00
Max.   :7.000   Max.   :91.00
NA's   :1
```

```
sapply(earnings,is.factor)
```

```
          height           weight             male             earn
           FALSE            FALSE             TRUE            FALSE
           earnk        ethnicity        education mother_education
           FALSE             TRUE            FALSE            FALSE
father_education             walk         exercise         smokenow
           FALSE            FALSE            FALSE             TRUE
           tense            angry              age
           FALSE            FALSE            FALSE
```

```
facs <- sapply(earnings,is.factor)
earningz <- earnings
earningz[!facs] <- scale(earnings[!facs])
summary(earningz)
```

```
    height            weight             male           earn
 Min.   :-2.4972   Min.   :-2.2043   female:1141   Min.   :-0.9386
 1st Qu.:-0.6704   1st Qu.:-0.7599   male  : 675   1st Qu.:-0.6723
 Median :-0.1484   Median :-0.1821                 Median :-0.2284
 Mean   : 0.0000   Mean   : 0.0000                 Mean   : 0.0000
 3rd Qu.: 0.6997   3rd Qu.: 0.6845                 3rd Qu.: 0.2598
 Max.   : 4.0271   Max.   : 5.3643                 Max.   :16.8142
                   NA's   :27
    earnk             ethnicity      education       mother_education
 Min.   :-0.9386   Black   : 180   Min.   :-4.3946   Min.   :-3.2953
 1st Qu.:-0.6723   Hispanic: 104   1st Qu.:-0.4832   1st Qu.:-0.4997
 Median :-0.2284   Other   :  38   Median :-0.4832   Median :-0.1891
 Mean   : 0.0000   White   :1494   Mean   : 0.0000   Mean   : 0.0000
 3rd Qu.: 0.2598                   3rd Qu.: 0.6902   3rd Qu.: 0.7428
 Max.   :16.8142                   Max.   : 1.8636   Max.   :26.5242
                                   NA's   :2         NA's   :244
 father_education       walk            exercise       smokenow
 Min.   :-3.2768   Min.   :-1.6545   Min.   :-0.8846   1   : 462
 1st Qu.:-0.5082   1st Qu.:-0.8856   1st Qu.:-0.8846   2   :1353
 Median :-0.2006   Median : 0.2678   Median :-0.4529   NA's:   1
 Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
 3rd Qu.: 0.7222   3rd Qu.: 1.0367   3rd Qu.: 0.8423
 Max.   :26.2543   Max.   : 1.0367   Max.   : 1.7057
```

```
 NA's   :295
    tense              angry               age
 Min.   :-0.6588   Min.   :-0.6588   Min.   :-1.4529
 1st Qu.:-0.6588   1st Qu.:-0.6588   1st Qu.:-0.8120
 Median :-0.6588   Median :-0.6588   Median :-0.2293
 Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
 3rd Qu.: 0.2681   3rd Qu.: 0.2681   3rd Qu.: 0.7613
 Max.   : 2.5852   Max.   : 2.5852   Max.   : 2.8007
 NA's   :1         NA's   :1
```

Fit Baseline model

```
earn1 <- na.omit(earningz)
bearn <- lm(minMod,earn1,earn1$earnk<350,na.action=na.omit)
summary(bearn)
```

```
Call:
lm(formula = minMod, data = earn1, subset = earn1$earnk < 350,
    na.action = na.omit)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8793 -0.3811  0.2304  0.6296  3.0295

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        -0.74973    0.03318 -22.596   <2e-16 ***
malemale            0.74794    0.05424  13.788   <2e-16 ***
education           0.38293    0.03699  10.352   <2e-16 ***
malemale:education -0.07713    0.05614  -1.374     0.17
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9793 on 1436 degrees of freedom
Multiple R-squared:  0.2008,    Adjusted R-squared:  0.1991
F-statistic: 120.3 on 3 and 1436 DF,  p-value: < 2.2e-16
```

```
bearnf <- step(bearn,list(lower=minMod,upper=maxMod),trace=2)
```

```
Start:  AIC=-56.17
log(earn + 1) ~ male * education

            Df Sum of Sq    RSS      AIC
+ age        1    50.487 1326.8 -107.947
+ height     1     3.557 1373.7  -57.891
+ exercise   1     2.936 1374.3  -57.240
+ smokenow   1     2.797 1374.4  -57.095
<none>                   1377.2  -56.168
+ ethnicity  3     1.822 1375.4  -52.074

Step:  AIC=-107.95
log(earn + 1) ~ male + education + age + male:education

            Df Sum of Sq    RSS      AIC
+ height     1     7.372 1319.4 -113.971
+ male:age   1     5.640 1321.1 -112.081
+ smokenow   1     4.427 1322.3 -110.760
<none>                   1326.8 -107.947
+ exercise   1     0.100 1326.7 -106.056
+ ethnicity  3     0.670 1326.1 -102.675
- age        1    50.487 1377.2  -56.168

Step:  AIC=-113.97
log(earn + 1) ~ male + education + age + height + male:education

            Df Sum of Sq    RSS      AIC
+ male:age   1     5.596 1313.8 -118.091
+ smokenow   1     4.399 1315.0 -116.780
<none>                   1319.4 -113.971
+ exercise   1     0.056 1319.3 -112.032
+ ethnicity  3     0.241 1319.1 -108.234
- height     1     7.372 1326.8 -107.947
- age        1    54.303 1373.7  -57.891

Step:  AIC=-118.09
log(earn + 1) ~ male + education + age + height + male:education +
    male:age

            Df Sum of Sq    RSS     AIC
+ smokenow   1    4.1501 1309.6 -120.65
<none>                   1313.8 -118.09
+ exercise   1    0.1351 1313.7 -116.24
```

```
- male:age   1    5.5957 1319.4 -113.97
+ ethnicity  3    0.1503 1313.6 -112.26
- height     1    7.3283 1321.1 -112.08

Step:  AIC=-120.65
log(earn + 1) ~ male + education + age + height + smokenow +
    male:education + male:age

            Df Sum of Sq    RSS     AIC
<none>                    1309.6 -120.65
+ exercise   1    0.3052 1309.3 -118.98
- smokenow   1    4.1501 1313.8 -118.09
- male:age   1    5.3472 1315.0 -116.78
+ ethnicity  3    0.1694 1309.5 -114.83
- height     1    7.3018 1316.9 -114.64
```

```r
summary(bearnf)
```

```
Call:
lm(formula = log(earn + 1) ~ male + education + age + height +
    smokenow + male:education + male:age, data = earn1, subset = earn1$earnk <
    350, na.action = na.omit)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0458 -0.4369  0.2157  0.6437  3.1251

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        -0.58971    0.05798 -10.170  < 2e-16 ***
malemale            0.64433    0.07550   8.535  < 2e-16 ***
education           0.39254    0.03652  10.750  < 2e-16 ***
age                 0.16347    0.03412   4.791 1.84e-06 ***
height              0.10409    0.03684   2.826  0.00478 **
smokenow2          -0.12575    0.05903  -2.130  0.03332 *
malemale:education -0.08567    0.05484  -1.562  0.11850
malemale:age        0.13577    0.05615   2.418  0.01573 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.9563 on 1432 degrees of freedom
Multiple R-squared:   0.24, Adjusted R-squared:  0.2363
F-statistic: 64.61 on 7 and 1432 DF,  p-value: < 2.2e-16
```

ACED Data

Recall four conditions:

- Adaptive sequence, full feedback
- Adaptive sequence, accuracy feedback
- Linear sequence, full feedback
- Control

Interested in difference is post-test (`post_scaled`).

`post_scaled ~ Cond_code`

But, there are differences in math ability before applying treatment.

Force pretest (`pre_scaled`) into model to soak up the ability difference.

`post_scaled ~ Cond_code + pre_scaled`

Alternative is to use gain score, `post_scaled - pre_scaled`

This does not account for unreliability of measure.

Last question: is there an interaction between condition and pretest?‘

`post_scaled ~ Cond_code * pre_scaled`

```
names(ACEDextract)
```

```
 [1] "SubjID"      "Session"     "Cond_code"   "Sequencing"  "Feedback"
 [6] "Correct"     "Incorrect"   "Reamaining"  "ElapsedTime" "Gender"
[11] "Race"        "Level_Code"  "pre_scaled"  "post_scaled" "Form_Order"
[16] "EAP.sgp"     "EAP.cr"      "EAP.dt"      "EAP.eg"      "EAP.exp"
[21] "EAP.ext"     "EAP.mod"     "EAP.rr"      "EAP.tab"     "EAP.vr"
[26] "EAP.pic"     "P.sgp..H"    "P.sgp..M"    "P.sgp..L"
```

```
acedminmod <- post_scaled ~ Cond_code + pre_scaled
acedmaxmod <- post_scaled ~ Cond_code * pre_scaled + Gender + Race + Level_Code
```

Fit the Initial Model

Model Selection

Interpret the Final Model

Aptitude-Treatment Interaction (ATI)

Job Satisfaction Data

This is the data set used for the first and second homework assignments. This shows how to read it into R.

```
library(haven)
```

```
jobsat <- read_spss("../Homework/jobsat.sav")
summary(jobsat)
```

```
      ID              gender           age          environment
 Min.   :  1.00   Min.   :1.000   Min.   :19.00   Min.   : 8.0
 1st Qu.: 50.75   1st Qu.:1.000   1st Qu.:31.00   1st Qu.:17.0
 Median :100.50   Median :2.000   Median :35.00   Median :19.0
 Mean   :100.50   Mean   :1.575   Mean   :34.94   Mean   :19.5
 3rd Qu.:150.25   3rd Qu.:2.000   3rd Qu.:38.00   3rd Qu.:22.0
 Max.   :200.00   Max.   :2.000   Max.   :49.00   Max.   :28.0
  performance    preyearsalary   currentsalary       stress
 Min.   :16.00   Min.   :30.00   Min.   : 30.75   Min.   :14.00
 1st Qu.:27.00   1st Qu.:46.62   1st Qu.: 52.61   1st Qu.:22.00
 Median :30.00   Median :54.51   Median : 60.23   Median :25.00
 Mean   :29.93   Mean   :54.20   Mean   : 60.46   Mean   :24.84
 3rd Qu.:34.00   3rd Qu.:61.71   3rd Qu.: 67.79   3rd Qu.:28.00
 Max.   :43.00   Max.   :90.00   Max.   :105.04   Max.   :36.00
 jobsatisfaction     rating
 Min.   : 3.00   Min.   :0.000
 1st Qu.:18.00   1st Qu.:0.000
 Median :25.00   Median :1.000
 Mean   :24.88   Mean   :0.505
 3rd Qu.:31.00   3rd Qu.:1.000
 Max.   :44.00   Max.   :1.000
```