ENV 710

interactions

Duke | NICHOLAS SCHOOL OF THE ENVIRONMENT

roadmap

- recap
- download from Week 10: `ht.dat.csv`

where we are

multivariate linear models

interactions
centering/scaling explanatory variables
random effects and mixed models

general*ized* linear models

BUSINESS

The Financial Perks of Being Tall

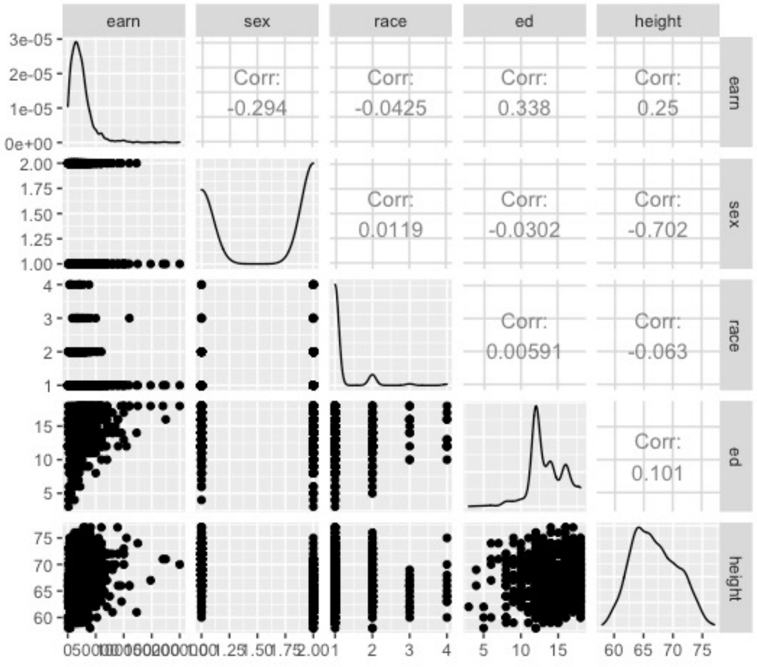
An extra inch correlates with an estimated $800 in increased annual earnings.

Taller people are RICHER: Scientists find biggest leap in income takes place between 5'4" and 5'6" - but over 6' has no effect

- On average, every extra inch in height earns a man $800 (£510) a year
- A 4-5 inch difference is associated with salary increase of 9-15 per cent
- Nutrition people received as a child - which affects both height and intelligence - is an important factor in determining salary, study finds

1 – height & education

Do earnings increase with height, education, or the interaction between height and education for a worker?



the data

```
ht.dat <- read.csv(”ht.dat.csv", header = T)
```

https://www.dailymail.co.uk/sciencetech/article-3088322/Taller-people-RICHER-Scientists-biggest-leap-income-takes-place-5-4-5-6-6-no-effect.html

1 – height & education

Do earnings increase with height, education, or the interaction between height and education for a worker?

the models

- model with just main effects
- model with interaction between height and ed

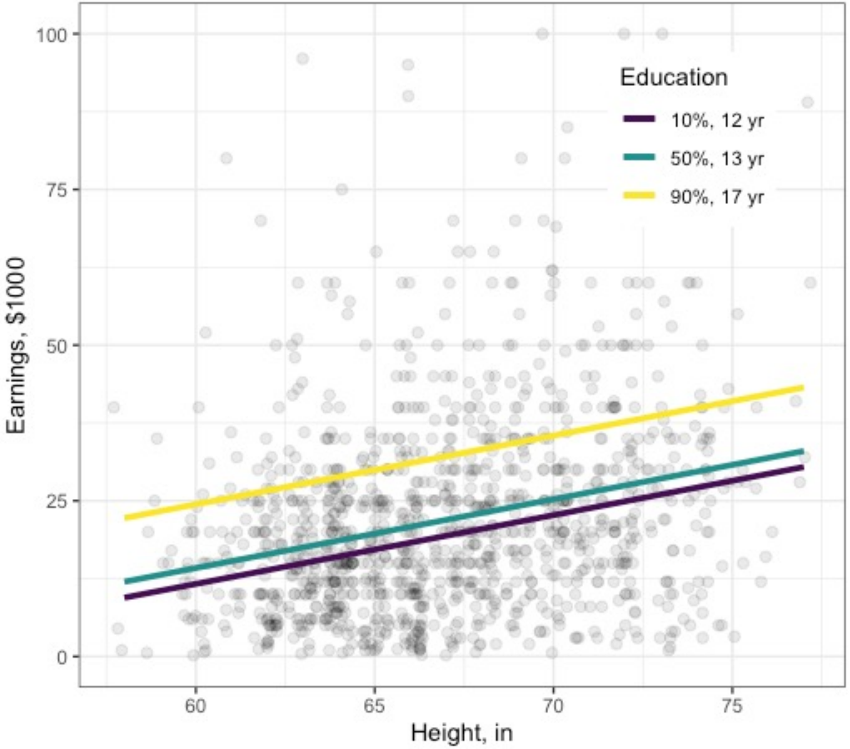
- which model is better? why?
- compare models with AIC and partial F-test (this is overkill)
- interpret the interaction

interaction effect occurs when the effect of one variable depends on the value of another variable.

1 – height & education

Do earnings increase with height, education, or the interaction between height and education for a worker?

```
c0 <- lm(earn ~ height + ed, data = ht.dat)
 summary(c0)

Call:
lm(formula = earn ~ height + ed, data = ht.dat)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -85294.7     9239.3  -9.232  < 2e-16 ***
height        1104.6      135.2   8.167 8.02e-16 ***
ed            2556.2      216.0  11.836  < 2e-16 ***
---

Residual standard error: 17840 on 1186 degrees of freedom
Multiple R-squared:  0.1616,	Adjusted R-squared:  0.1602
F-statistic: 114.3 on 2 and 1186 DF,  p-value: < 2.2e-16
```



1 – height & education

```
# model
c0 <- lm(earn ~ height + ed, ht.dat)

# equations for lines @ 10,50,90% education
cfs <- coef(c0)
eq1=function(x){(cfs[1] + cfs[2]*x + cfs[3]*quantile(ht.dat$ed, 0.5))/1000}
eq2 =function(x){(cfs[1] + cfs[2]*x + cfs[3]*quantile(ht.dat$ed, 0.1))/1000}
eq3 =function(x){(cfs[1] + cfs[2]*x + cfs[3]*quantile(ht.dat$ed, 0.9))/1000}

# plot of earnings/1000 x height
ggplot(data = ht.dat, aes(x = height, y = earn/1000)) +
  geom_jitter(shape = 19, size = 2, fill = my_jco[3], alpha = 0.1) +
   stat_function(fun = eq1, geom="line", linewidth = 1.3, aes(colour = "50%, 13 yr")) +
   stat_function(fun = eq2, geom="line", linewidth = 1.3, aes(colour = "10%, 12 yr")) +
   stat_function(fun = eq3, geom="line", linewidth = 1.3, aes(colour = "90%, 17 yr")) +
       theme_bw() + theme(legend.position = c(.8, .8)) + ylim(c(0, 100)) +
        labs(y = "Earnings, $1000", x = "Height, in") +
         scale_color_viridis("Education", discrete = TRUE)
```

I – height & education

Do earnings increase with the height and education or interaction between height and education for a worker?

```
c2 <- lm(earn ~ height*ed, data = ht.dat)
 summary(c2)


Call:
lm(formula = earn ~ height * ed, data = ht.dat)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 38433.60   48668.60   0.790  0.42986
height       -744.22     726.70  -1.024  0.30599
ed          -6699.50    3581.35  -1.871  0.06164 .
height:ed     138.16      53.36   2.589  0.00974 **
---

Residual standard error: 17800 on 1185 degrees of freedom
Multiple R-squared:  0.1664,        Adjusted R-squared:  0.1642
F-statistic: 78.83 on 3 and 1185 DF,  p-value: < 2.2e-16
```

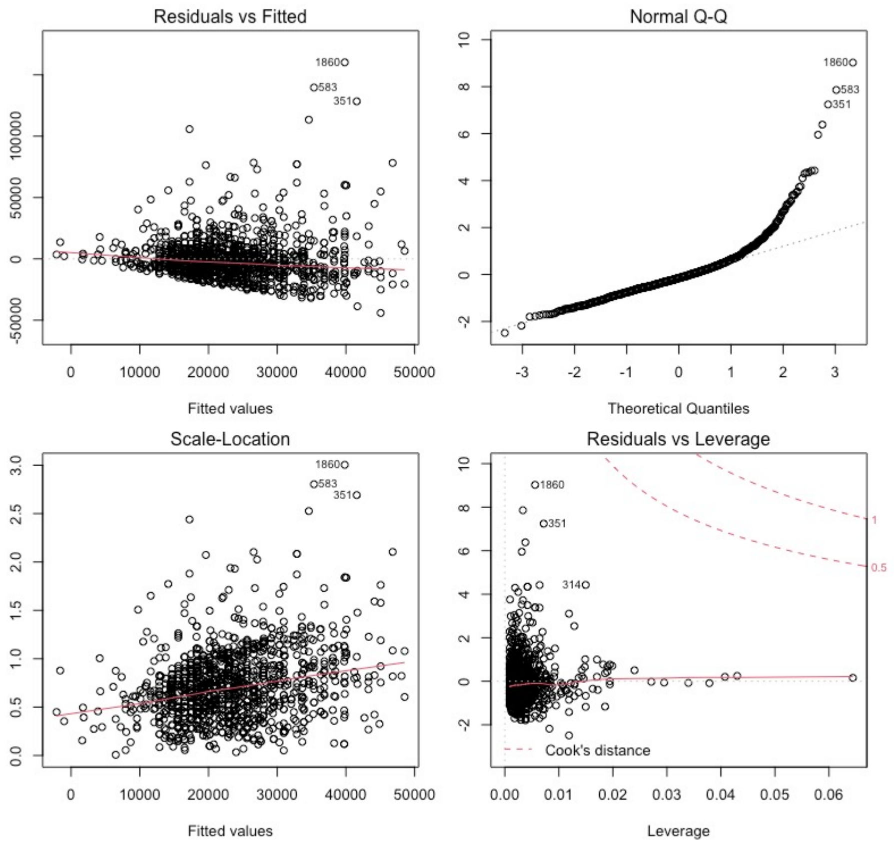
```
AIC(c0, c2)
   df      AIC
c0  4 26658.16
c2  5 26653.45

anova(c0, c2)

Analysis of Variance Table

Model 1: earn ~ height + ed
Model 2: earn ~ height * ed
  Res.Df        RSS Df  Sum of Sq      F   Pr(>F)
1   1186 3.7753e+11
2   1185 3.7541e+11  1 2123677463 6.7035 0.00974 **
---
```

I – height & education

Do earnings increase with the height and education or interaction between height and education for a worker?

```
c2 <- lm(earn ~ height*ed, data = ht.dat)
 summary(c2)
```

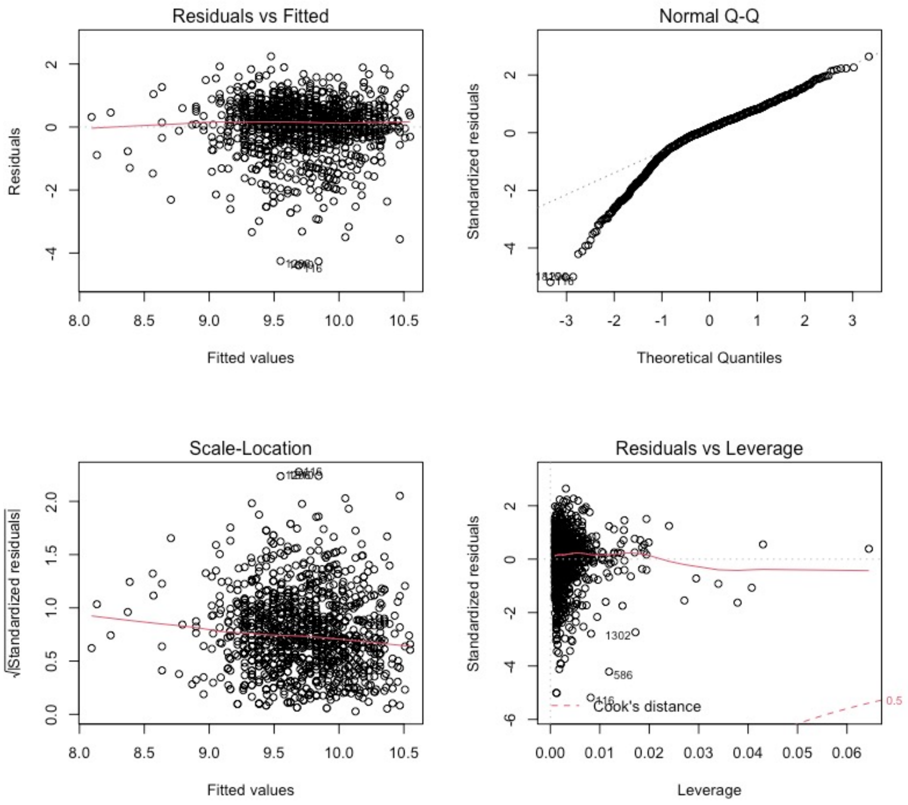
```
Call:
lm(formula = earn ~ height * ed, data = ht.dat)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 38433.60   48668.60   0.790  0.42986
height       -744.22     726.70  -1.024  0.30599
ed          -6699.50    3581.35  -1.871  0.06164 .
height:ed     138.16      53.36   2.589  0.00974 **
---

Residual standard error: 17800 on 1185 degrees of freedom
Multiple R-squared:  0.1664,  Adjusted R-squared:  0.1642
F-statistic: 78.83 on 3 and 1185 DF,  p-value: < 2.2e-16
```



1 – height & education

Do earnings increase with the height and education or interaction between height and education for a worker?

```
c2.log <- lm(log(earn) ~ height*ed, data = ht.dat)
```



we should log-transform `earn`, but don't to keep interpretation easy for class

1 – height & education

Do earnings increase with height, education, or the interaction between height and education for a worker?

the models

- model with just main effects
- model with interaction between height and ed

- what is the predicted salary for someone who has 10 years of education and is 70 inches tall?
- what is the predicted salary for someone who has the average number of years of education and is 72 inches tall?
- create a plot showing the height x ed interaction

1 – height & education

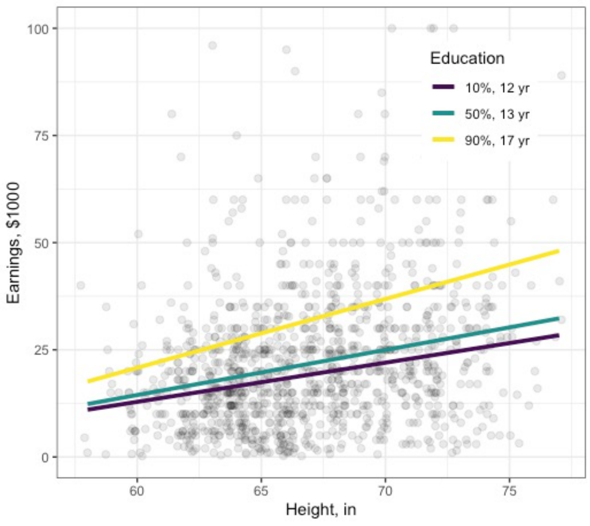
Do earnings increase with the height and education or interaction between height and education for a worker?

```
summary(c2)

Call:
lm(formula = earn ~ height * ed, data = ht.dat)


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 38433.60   48668.60   0.790  0.42986
height        -744.22     726.70  -1.024  0.30599
ed          -6699.50    3581.35  -1.871  0.06164 .
height:ed     138.16      53.36   2.589  0.00974 **
---
```

```
c.ht <- coef(c2)
c.ht

(Intercept)      height          ed   height:ed
  38433.605    -744.217   -6699.502     138.159

c.ht[1] + c.ht[2]*70 + c.ht[3]*10 + c.ht[4]*70*10

(Intercept)
   16054.67

c.ht[1] + c.ht[2]*72 + c.ht[3]*mean(ht.dat$ed) +
c.ht[4]*72*mean(ht.dat$ed)

(Intercept)
   28712.24
```

I – height & education

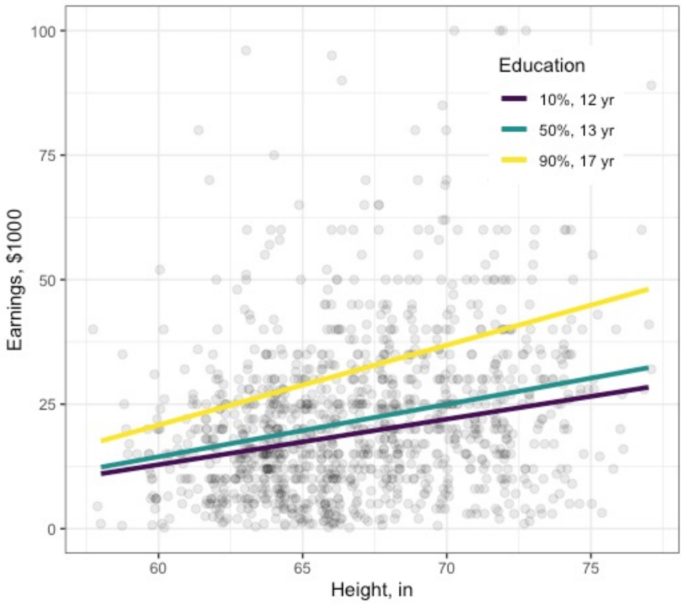
```
# extract coefficients

cfs <- coef(c2)

# set levels of education

evals <- quantile(ht.dat$ed.c, c(0.5, 0.1, 0.9))

# create equations for lines

eq1=function(x){(cfs[1] + cfs[2]*x + cfs[3]*evals[1] + cfs[4]*x*evals[1])/1000}
eq2 =function(x){(cfs[1] + cfs[2]*x + cfs[3]*evals[2] + cfs[4]*x*evals[2])/1000}
eq3 =function(x){(cfs[1] + cfs[2]*x + cfs[3]*evals[3] + cfs[4]*x*evals[3])/1000}

# plot earnings vs. height at different levels of education

p2 <- ggplot(data = ht.dat, aes(x = height, y = earn/1000)) +
       geom_jitter(shape = 19, size = 2, fill = my_jco[3], alpha = 0.1) +
          stat_function(fun = eq1, geom="line", linewidth = 1.3, aes(colour = "50%, 13 yr")) +
          stat_function(fun = eq2, geom="line", linewidth = 1.3, aes(colour = "10%, 12 yr")) +
          stat_function(fun = eq3, geom="line", linewidth = 1.3, aes(colour = "90%, 17 yr")) +
             theme_bw() + theme(legend.position = c(.8, .8)) + ylim(c(0, 100)) +
              labs(y = "Earnings, $1000", x = "Height, in") +
  scale_color_viridis("Education", discrete = TRUE)
```



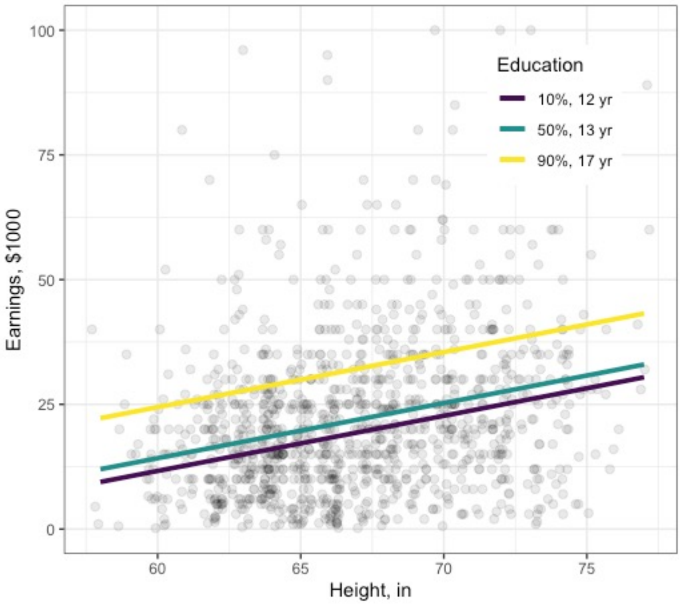
effect height depends on education (and vice versa).

1 – height & education

interaction model: increase in earnings with height is fastest with higher levels of education



main effects model: slope doesn't change with levels of education



1 – height & education

Do earnings increase with height, education, or the interaction between height and education for a worker?

the models

- center the height and education variables
- run the main effects and interaction models

- what do the coefficients mean now?
- which model is better

https://www.dailymail.co.uk/sciencetech/article-3088322/Taller-people-RICHER-Scientists-biggest-leap-income-takes-place-5-4-5-6-6-no-effect.html

1 – height & education

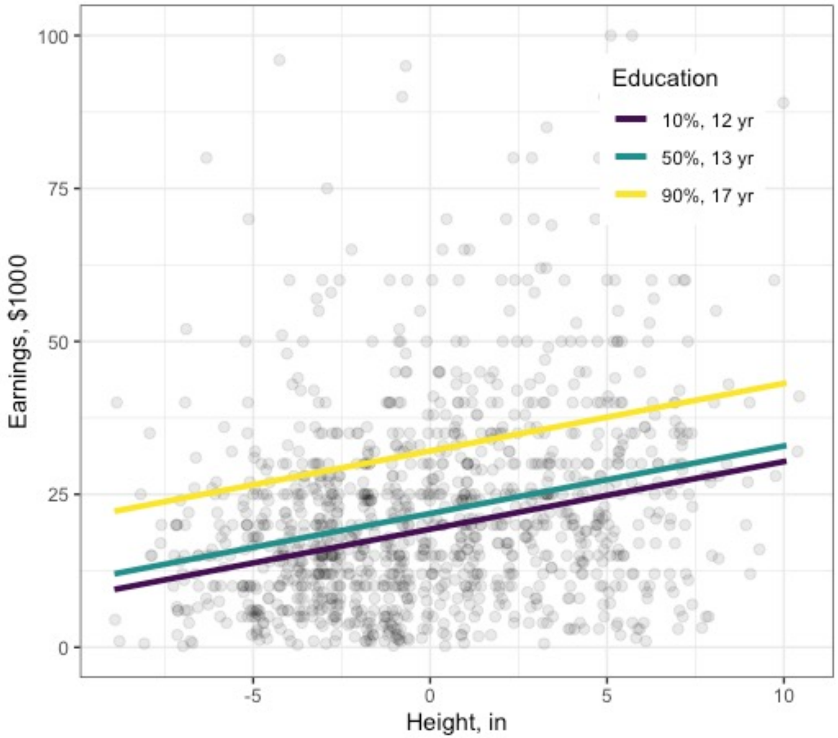
```
# standardizing with z
ht.dat$ed.z <- (ht.dat$ed - mean(ht.dat$ed))/(sd(ht.dat$ed))
ht.dat$ht.z <- (ht.dat$height - mean(ht.dat$height, na.rm = T))/ sd(ht.dat$height)

# centering on mean
ht.dat$ed.c <- ht.dat$ed - mean(ht.dat$ed)
ht.dat$ht.c <- ht.dat$height - mean(ht.dat$height)

# main effects model with centered IV's
c5 <- lm(earn ~ ht.c + ed.c, data = ht.dat)
   cfs5 <- coef(c5)

# plot with centered IV's
cfs <- coef(c5)
 eq1=function(x){(cfs[1] + cfs[2]*x + cfs[3]*quantile(ht.dat$ed.c, 0.5))/1000}
 eq2 =function(x){(cfs[1] + cfs[2]*x + cfs[3]*quantile(ht.dat$ed.c, 0.1))/1000}
 eq3 =function(x){(cfs[1] + cfs[2]*x + cfs[3]*quantile(ht.dat$ed.c, 0.9))/1000}

 p3 <- ggplot(data = ht.dat, aes(x = ht.c, y = earn/1000)) +
       geom_jitter(shape = 19, size = 2, fill = my_jco[3], alpha = 0.1) +
          stat_function(fun = eq1, geom="line", linewidth = 1.3, aes(colour = "50%, 13 yr")) +
          stat_function(fun = eq2, geom="line", linewidth = 1.3, aes(colour = "10%, 12 yr")) +
          stat_function(fun = eq3, geom="line", linewidth = 1.3, aes(colour = "90%, 17 yr")) +
             theme_bw() + theme(legend.position = c(.8, .8)) +
              labs(y = "Earnings, $1000", x = "Height, in") + ylim(c(0, 100)) +
         scale_color_viridis("Education", discrete = TRUE)
```

```
centered.x <- scale(x, scale = FALSE))

scaled.x <- scale(x)
```

1 – height & education

Do earnings increase with height, education, or the interaction between height and education for a worker?

```
c5 <- lm(earn ~ ht.c + ed.c, data = ht.dat)
 summary(c5)

lm(formula = earn ~ ht.c + ed.c, data = ht.dat)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  23149.8      517.4  44.741  < 2e-16 ***
ht.c          1104.6      135.2   8.167 8.02e-16 ***
ed.c          2556.2      216.0  11.836  < 2e-16 ***
---

Residual standard error: 17840 on 1186 degrees of freedom
Multiple R-squared:  0.1616,    Adjusted R-squared:  0.1602
F-statistic: 114.3 on 2 and 1186 DF,  p-value: < 2.2e-16
```

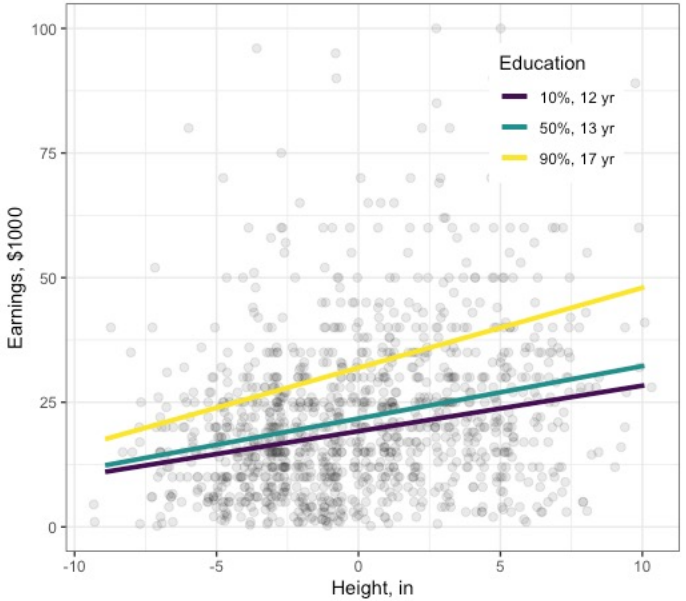


intercept is now the mean earnings at the mean height (5'6" or 67") and mean education (13.5 yrs)

I – height & education

Do earnings increase with height, education, or the interaction between height and education for a worker?

```
summary(c6)

lm(formula = earn ~ ht.c * ed.c, data = ht.dat)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 23020.45     518.59  44.390  < 2e-16 ***
ht.c         1121.57     135.08   8.303 2.74e-16 ***
ed.c         2546.81     215.48  11.819  < 2e-16 ***
ht.c:ed.c     138.16      53.36   2.589  0.00974 ** 
---

Residual standard error: 17800 on 1185 degrees of freedom
Multiple R-squared:  0.1664,    Adjusted R-squared:  0.1642 
F-statistic: 78.83 on 3 and 1185 DF,  p-value: < 2.2e-16
```



1 – height & education

Do earnings increase with height, education, or the interaction between height and education for a worker?

```
Model 1: earn ~ height + ed
Model 2: earn ~ height * ed
Model 3: earn ~ ht.c + ed.c
Model 4: earn ~ ht.c * ed.c
```

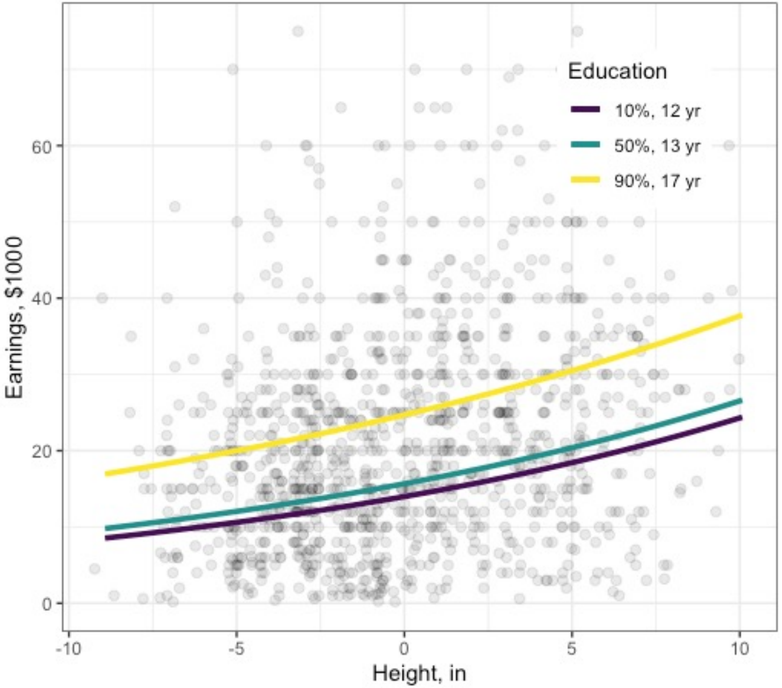
```
AIC(c0, c2, c5, c6)

   df      AIC
c0  4 26658.16
c2  5 26653.45
c5  4 26658.16
c6  5 26653.45
```

centering and scaling do not change the fit of the model to the data, choice of best model or R^2

```
c(summary(c0)$adj.r.squared, summary(c2)$adj.r.squared,
summary(c5)$adj.r.squared, summary(c6)$adj.r.squared)

[1] 0.1602305 0.1642497 0.1602305 0.1642497
```

I – height & education

```
c7 <- lm(log(earn/1000) ~ ht.c*ed.c, data = ht.dat)

cfs <- coef(c7)
evals <- quantile(ht.dat$ed.c, c(0.5, 0.1, 0.9))

eq1=function(x){exp(cfs[1] + cfs[2]*x + cfs[3]*evals[1] +
                cfs[4]*x*evals[1])}
eq2 =function(x){exp(cfs[1] + cfs[2]*x + cfs[3]*evals[2] +
                 cfs[4]*x*evals[2])}
eq3 =function(x){exp(cfs[1] + cfs[2]*x + cfs[3]*evals[3] +
                 cfs[4]*x*evals[3])}
```




```
ggplot(data = ht.dat, aes(x = ht.c, y = earn/1000)) +
  geom_jitter(shape = 19, size = 2, fill = my_jco[3], alpha = 0.1) +
   stat_function(fun = eq1, geom="line", linewidth = 1.3, aes(colour = "50%, 13 yr")) +
   stat_function(fun = eq2, geom="line", linewidth = 1.3, aes(colour = "10%, 12 yr")) +
   stat_function(fun = eq3, geom="line", linewidth = 1.3, aes(colour = "90%, 17 yr")) +
        theme_bw() + theme(legend.position = c(.8, .8)) +
         labs(y = "Earnings, $1000", x = "Height, in") + ylim(c(0, 75)) +
        scale_color_viridis("Education", discrete = TRUE)
```

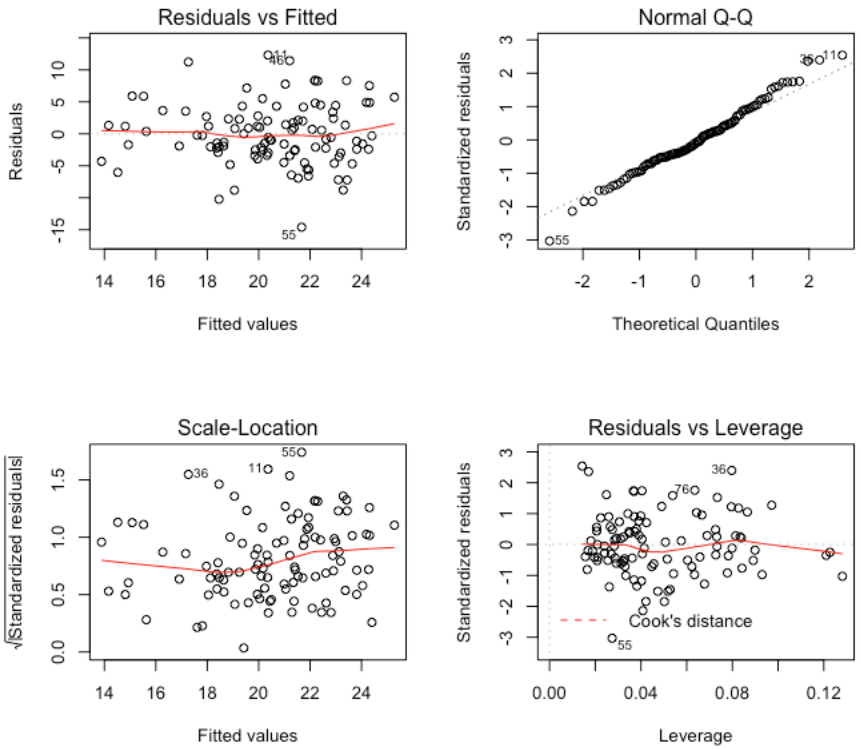
2 – tree height

Does tree height vary with precipitation, latitude and longitude in a tropical nation?

```
agb <- read.csv("agb.dat.csv", header = T)
 h1 <- lm(Ht.avg ~ Long*Lat + Temp.c, data = agb)
```

2 – tree height

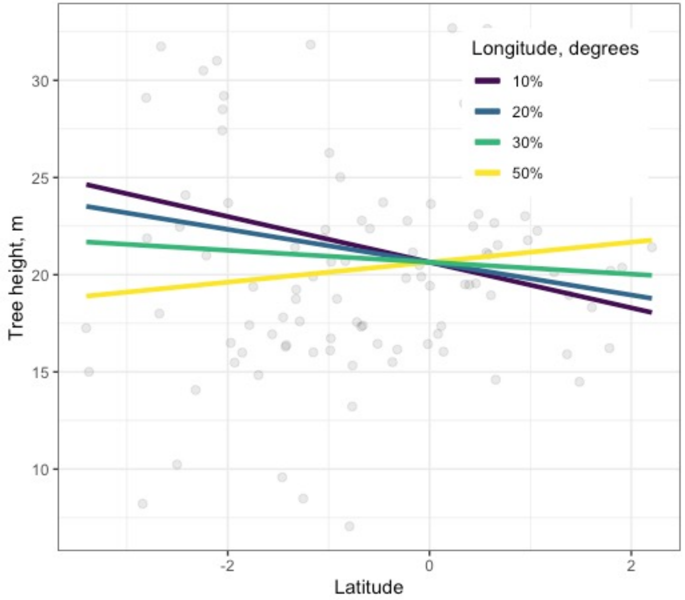
```
Call:
lm(formula = Ht.avg ~ Long * Lat + Temp.c, data = agb)

Residuals:
     Min       1Q   Median       3Q      Max
-14.6091  -2.7061  -0.4344   2.7240  12.3203

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  61.1523    17.1817   3.559 0.000573 ***
Long         -0.8763     0.4308  -2.034 0.044589 *
Lat         -11.5382     3.5056  -3.291 0.001382 **
Temp.c       -1.2355     0.5759  -2.145 0.034381 *
Long:Lat      1.0390     0.2986   3.479 0.000749 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 4.888 on 99 degrees of freedom
Multiple R-squared:  0.2207,Adjusted R-squared:  0.1892
F-statistic: 7.008 on 4 and 99 DF,  p-value: 5.207e-05
```



2 – tree height

```
agb <- read.csv("agb.dat.csv", header = T)
 h1 <- lm(Ht.avg ~ Long*Lat + Temp.c, data = agb)
```

as you go east, the effect of going north on tree height increases



Questions?