

Multiple regression part 2

Frank Edwards

The linear regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

The linear regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

β_0 : The value of y when x is equal to zero

The linear regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

β_0 : The value of y when x is equal to zero

β_1 : The average increase in y when x increases by one unit

The linear regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

β_0 : The value of y when x is equal to zero

β_1 : The average increase in y when x increases by one unit

ε : The distance between the line $y = \beta_0 + \beta_1 X$ and the actual observed values of y . Allows us to estimate the line, even when x and y do not fall exactly on a line.

The linear regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

β_0 : The value of y when x is equal to zero

β_1 : The average increase in y when x increases by one unit

ε : The distance between the line $y = \beta_0 + \beta_1 X$ and the actual observed values of y . Allows us to estimate the line, even when x and y do not fall exactly on a line.

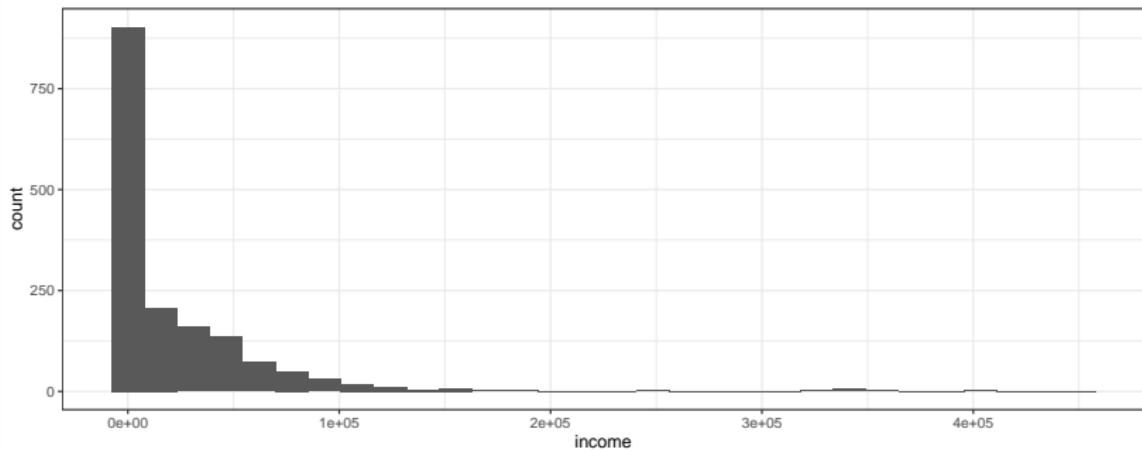
Data for today

```
dat<-read_csv("https://www.openintro.org/data/csv/acs12.csv")
glimpse(dat)

## Rows: 2,000
## Columns: 13
## $ income      <dbl> 60000, 0, NA, 0, 0, 1700, NA, NA, NA, 45000, NA, 8600, 0, ~
## $ employment   <chr> "not in labor force", "not in labor force", NA, "not in l~
## $ hrs_work     <dbl> 40, NA, NA, NA, 40, NA, NA, NA, 84, NA, 23, NA, NA, N~
## $ race         <chr> "white", "white", "white", "white", "white", "other", "wh~
## $ age          <dbl> 68, 88, 12, 17, 77, 35, 11, 7, 6, 27, 8, 69, 69, 17, 10, ~
## $ gender        <chr> "female", "male", "female", "male", "female", "female", "f~
## $ citizen       <chr> "yes", "yes", "yes", "yes", "yes", "yes", "yes", "yes", "y~
## $ time_to_work <dbl> NA, NA, NA, NA, NA, 15, NA, NA, NA, 40, NA, 5, NA, NA, NA~
## $ lang          <chr> "english", "english", "english", "other", "other", "other~
## $ married       <chr> "no", "no", "no", "no", "no", "yes", "no", "no", "no", "y~
## $ edu           <chr> "college", "hs or lower", "hs or lower", "hs or lower", "h~
## $ disability    <chr> "no", "yes", "no", "no", "yes", "yes", "no", "yes", "no", ~
## $ birth_qrtr    <chr> "jul thru sep", "jan thru mar", "oct thru dec", "oct thru~
```

Let's look at income for this ACS 2012 sample

```
ggplot(dat,  
       aes(x = income)) +  
  geom_histogram()
```



OK, what could cause variation in income?

```
glimpse(dat)

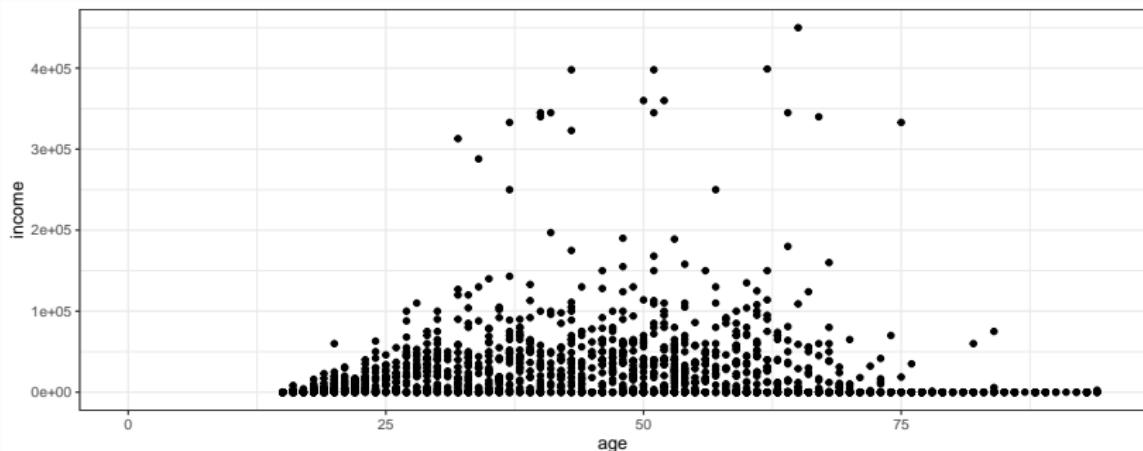
## Rows: 2,000
## Columns: 13

## $ income <dbl> 60000, 0, NA, 0, 0, 1700, NA, NA, NA, 45000, NA, 8600, 0, ~
## $ employment <chr> "not in labor force", "not in labor force", NA, "not in l~
## $ hrs_work <dbl> 40, NA, NA, NA, NA, 40, NA, NA, NA, 84, NA, 23, NA, NA, N~
## $ race <chr> "white", "white", "white", "white", "white", "other", "wh~
## $ age <dbl> 68, 88, 12, 17, 77, 35, 11, 7, 6, 27, 8, 69, 69, 17, 10, ~
## $ gender <chr> "female", "male", "female", "male", "female", "female", "f~
## $ citizen <chr> "yes", "yes", "yes", "yes", "yes", "yes", "yes", "yes", "y~
## $ time_to_work <dbl> NA, NA, NA, NA, NA, 15, NA, NA, NA, 40, NA, 5, NA, NA, NA~
## $ lang <chr> "english", "english", "english", "other", "other", "other", ~
## $ married <chr> "no", "no", "no", "no", "no", "yes", "no", "no", "no", "n~
## $ edu <chr> "college", "hs or lower", "hs or lower", "hs or lower", "hs~
## $ disability <chr> "no", "yes", "no", "no", "yes", "yes", "no", "yes", "no", ~
## $ birth_qtrr <chr> "jul thru sep", "jan thru mar", "oct thru dec", "oct thru~
```

Visual checks

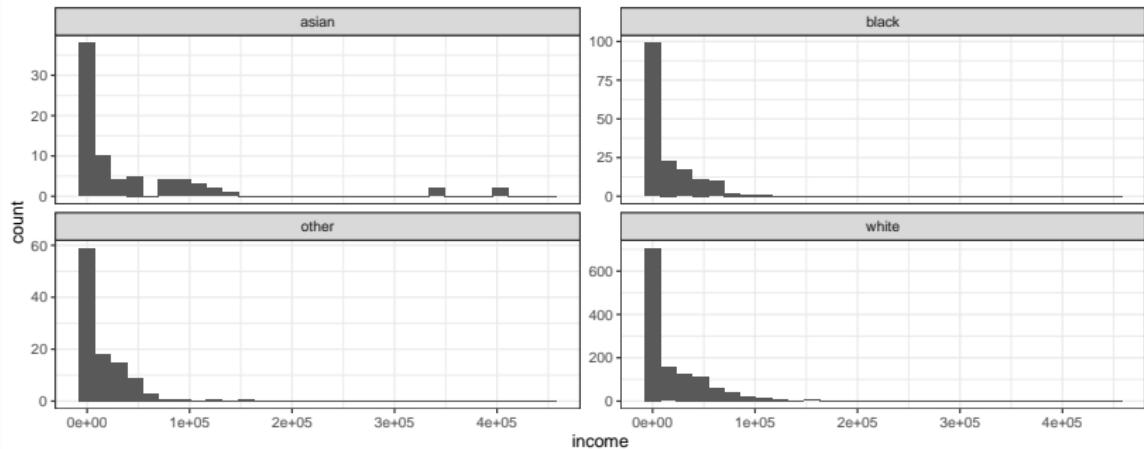
Causation requires association (though it's not always unconditional!).
Looks promising here!

```
ggplot(dat,  
       aes(y = income, x = age)) +  
  geom_point()
```



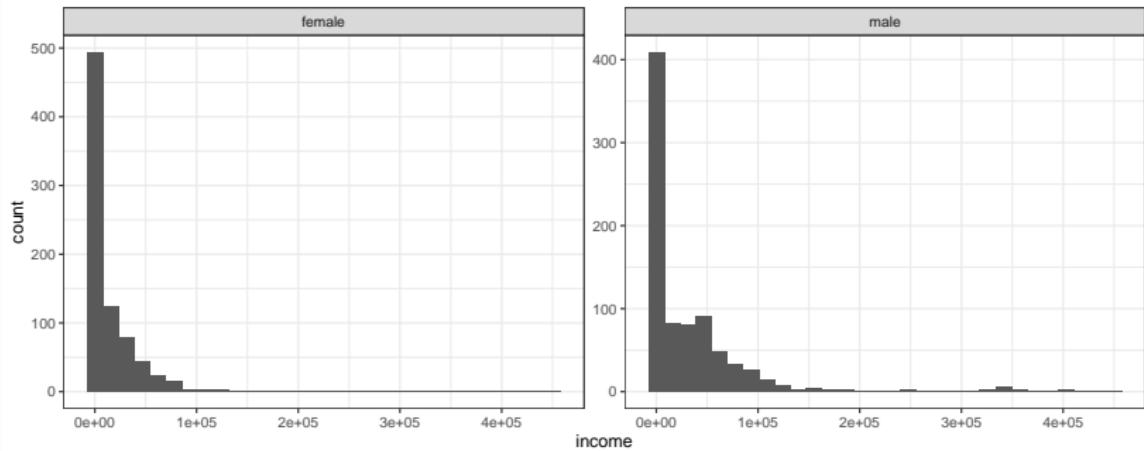
Visual checks

```
ggplot(data,  
       aes(x = income)) +  
  geom_histogram() +  
  facet_wrap(~race, scales = "free_y")
```



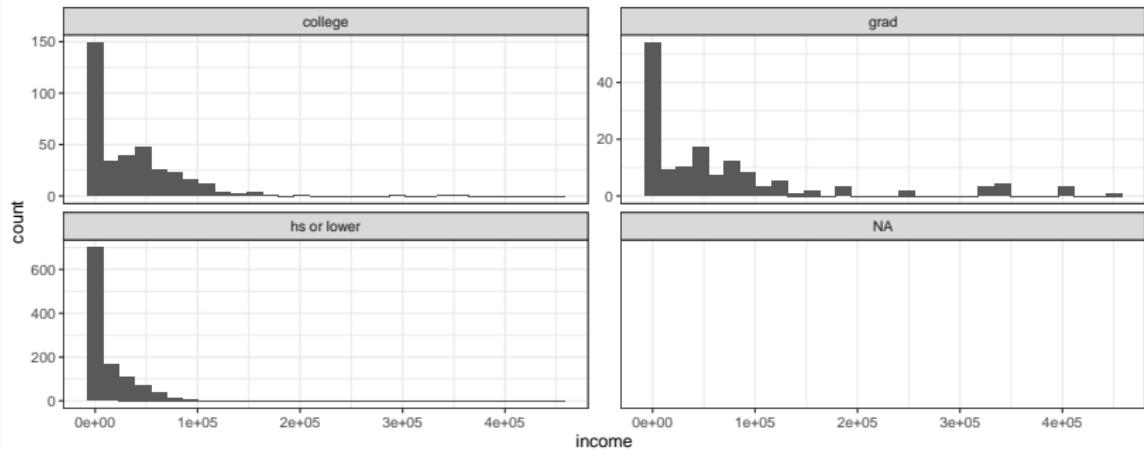
Visual checks

```
ggplot(data,  
       aes(x = income)) +  
  geom_histogram() +  
  facet_wrap(~gender, scales = "free_y")
```



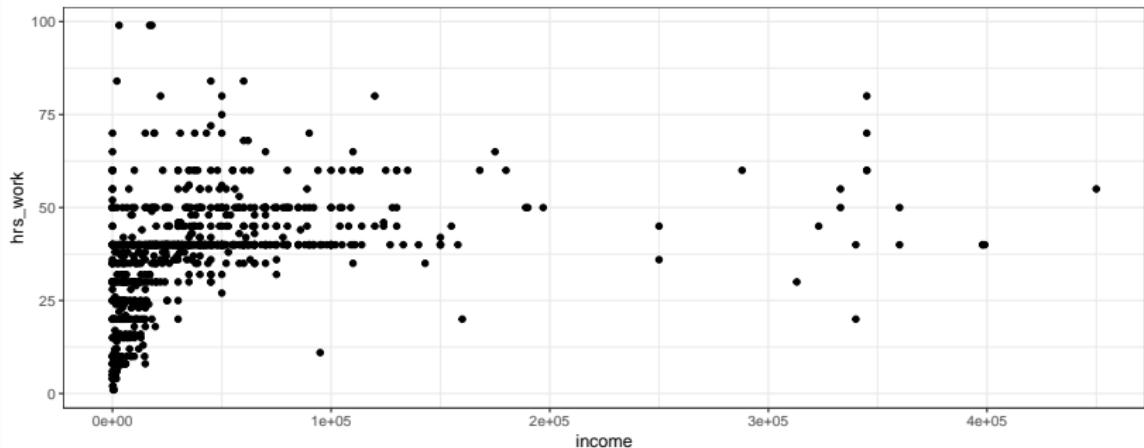
Visual checks

```
ggplot(data,  
       aes(x = income)) +  
  geom_histogram() +  
  facet_wrap(~edu, scales = "free_y")
```



Visual checks

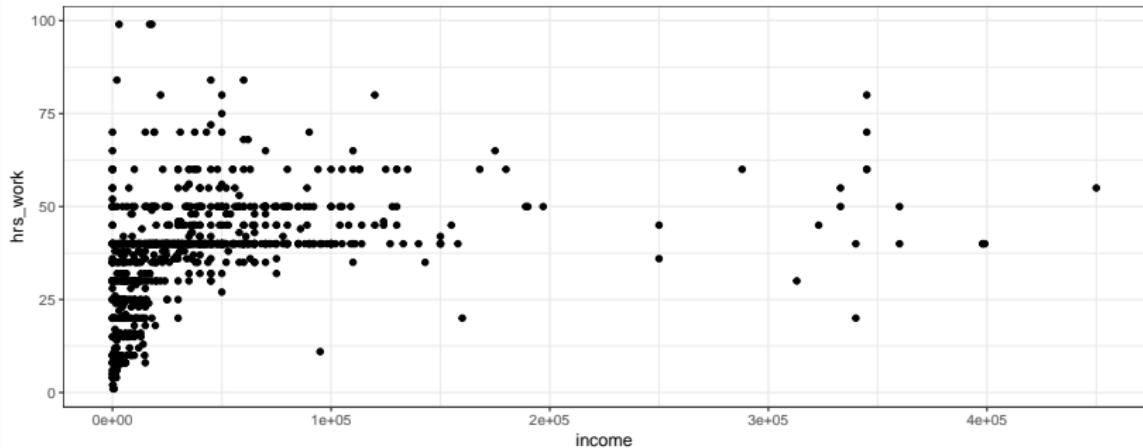
```
ggplot(dat,  
       aes(x = income, y = hrs_work)) +  
  geom_point()
```



Let's build a **causal model** to formalize what we think causes variation in income across the population.

Visual checks

```
ggplot(dat,  
       aes(x = income, y = hrs_work)) +  
  geom_point()
```



Let's build a **causal model** to formalize what we think causes variation in income across the population.

To do this, we'll use *Directed Acyclic Graphs*, or *DAGs* for short.

Let's start with a simple model

Based on our deep scientific knowledge we suspect that *hours worked t* has direct effects on *income I*

```
library(dagitty)
d1<-dagitty("dag {
    t->I
    t [exposure]
    I [outcome]
    }")
plot(graphLayout(d1))
```



Basic features of a DAG

DAGs contain *nodes* that represent variables, and *edges* that represent causal relationships between variables. In this case, we have two nodes, hours worked and income, and one edge, representing the effect of time spent working on income.

```
plot(graphLayout(d1))
```



Basic features of a DAG

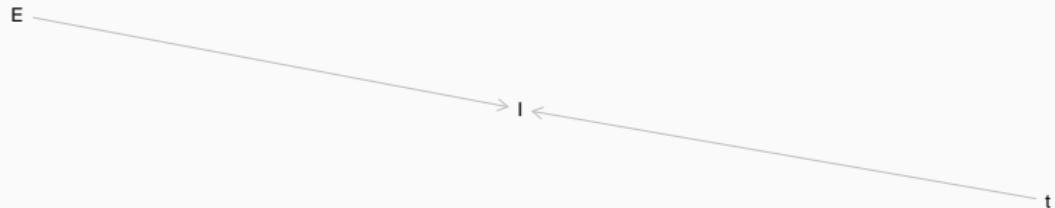
DAGs compactly represent our theoretical models. What is the theory presented here and do we believe it is adequate?



Adding complexity

Let's add level of education to our model. What theoretical relationships does this model suggest?

```
d2<-dagitty("dag {  
  E->I  
  t->I  
  E [exposure]  
  I [outcome]  
 }")  
  
plot(graphLayout(d2))
```



The importance of theory

Which model is more plausible?

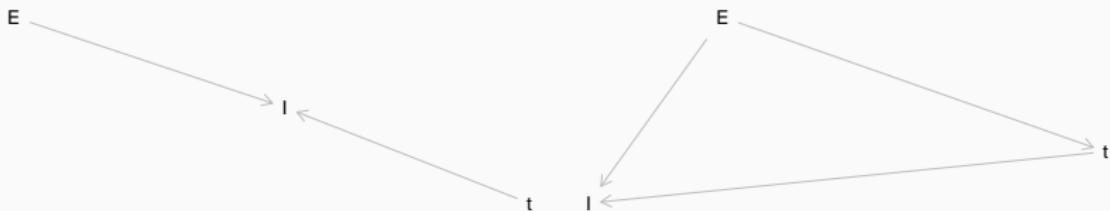
```
d3<-dagitty("dag {  
    E->I  
    t->I  
    E->t  
    E [exposure]  
    I [outcome]  
 }")  
  
plot(graphLayout(d3))
```



Confounding

We cannot obtain a valid estimate of the effect of t on I if DAG 3 is correct, unless we adjust for E .

This is a case of *confounding*. A relationship between two variables X and Y is confounded when a third variable Z also causes X and Y .



Let's try it: unconditional linear relationship

```
m0<-lm(income ~ hrs_work,
         data = dat)
summary(m0)

##
## Call:
## lm(formula = income ~ hrs_work, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -121855  -23313   -9189    7245  386372
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -12905.5     4983.1   -2.59  0.00975 **
## hrs_work     1391.5      123.6   11.25 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 51650 on 957 degrees of freedom
##   (1041 observations deleted due to missingness)
## Multiple R-squared:  0.1169, Adjusted R-squared:  0.116
## F-statistic: 126.7 on 1 and 957 DF,  p-value: < 2.2e-16
```

Let's try it: additive linear relationship

```
m1<-lm(income ~ hrs_work +
         edu,
         data = dat)
summary(m1)

##
## Call:
## lm(formula = income ~ hrs_work + edu, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -115673 -19681   -6409   10353  340318
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1314.8     5437.7   0.242   0.809
## hrs_work    1198.1     116.4  10.297 < 2e-16 ***
## edugrad     42472.0    5588.7   7.600 7.08e-14 ***
## eduhs or lower -18597.7   3578.8  -5.197 2.48e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
## Residual standard error: 48140 on 955 degrees of freedom
## (1041 observations deleted due to missingness)
## Multiple R-squared:  0.2346, Adjusted R-squared:  0.2322
## F-statistic:  97.6 on 3 and 955 DF,  p-value: < 2.2e-16
```

Multiple regression (regression with more than 1 predictor)

We can generalize the linear regression

$$Y = \beta_0 + \beta_1 X + \varepsilon$$
$$\varepsilon \sim N(0, \sigma^2)$$

Multiple regression (regression with more than 1 predictor)

We can generalize the linear regression

$$Y = \beta_0 + \beta_1 X + \varepsilon$$
$$\varepsilon \sim N(0, \sigma^2)$$

as

$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \varepsilon$$
$$\varepsilon \sim N(0, \sigma^2)$$

Where k is the number of predictor variables we include in the model. Our only constraint is that k must be smaller than the number of observations n in our data.

Our model for income

Our theoretical model tells us that if we want to learn about $t \rightarrow I$, we must adjust for the effects that E has on both t and I .

We tried this with the model:

$$E(\text{income}) = \beta_0 + \beta_1 \text{hrs} + \beta_2 \text{edu}$$

(Keep in mind that **edu** is going to be treated as the number of categories in the variable - 1 extra parameters).

Interpreting this model

```
table(dat$edu)

##
##      college      grad hs or lower
##      359          144        1439

summary(m1)

##
## Call:
## lm(formula = income ~ hrs_work + edu, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -115673 -19681   -6409   10353  340318 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1314.8     5437.7   0.242   0.809    
## hrs_work    1198.1     116.4   10.297  < 2e-16 ***
## edugrad    42472.0    5588.7   7.600 7.08e-14 ***
## eduhs or lower -18597.7   3578.8  -5.197 2.48e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
## Residual standard error: 48140 on 955 degrees of freedom
## (1041 observations deleted due to missingness)
## Multiple R-squared:  0.2346, Adjusted R-squared:  0.2322 
## F-statistic:  97.6 on 3 and 955 DF,  p-value: < 2.2e-16
```

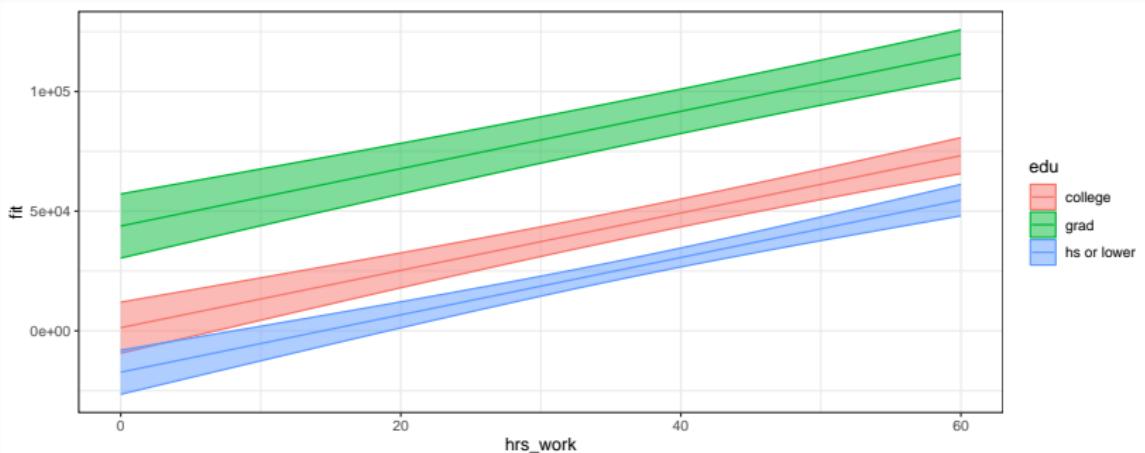
To understand what the model says, let's visualize y-hat

```
# set up prediction data with values of interest
hrs_work<-0:60
edu<-c("college", "hs or lower", "grad")
pred_dat<-expand_grid(hrs_work, edu)
# generate expected values and CI, join pred_dat
e_y<-predict(m1,
              newdata = pred_dat,
              interval = "confidence") |>
  bind_cols(pred_dat)
# inspect
head(e_y)

## # A tibble: 6 x 5
##       fit     lwr      upr hrs_work edu
##     <dbl>   <dbl>    <dbl>    <int> <chr>
## 1  1315. -9356.  11986.      0 college
## 2 -17283. -26515. -8051.      0 hs or lower
## 3  43787.  30390.  57184.      0 grad
## 4   2513. -7968.  12994.      1 college
## 5 -16085. -25110. -7060.      1 hs or lower
## 6  44985.  31752.  58218.      1 grad
```

Visualizing model expectations

```
ggplot(e_y,  
       aes(y = fit,  
           ymin = lwr,  
           ymax = upr,  
           x = hrs_work,  
           fill = edu,  
           color = edu)) +  
  geom_ribbon(alpha = 0.5) +  
  geom_line()
```



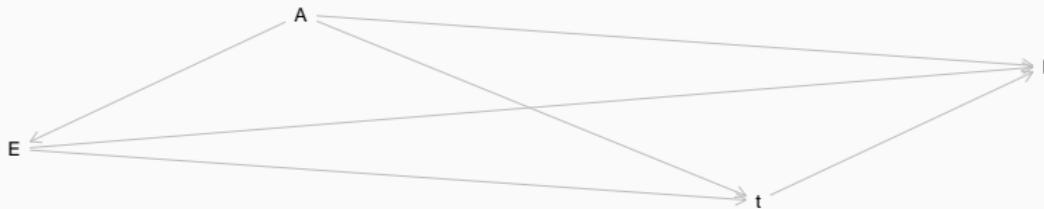
Multiple regression basics

1. Categorical predictors act as intercepts, or differences in level
2. Continuous predictors act as slopes

Regression with more than one slope and multiple intercepts

Maybe we think age also plays a role. Let's assume this causal model, where A is age. Now, we have to condition on A and E to close all *back door* paths between t and I and adjust for confounding

```
d4<-dagitty("dag {  
  E->I  
  t->I  
  E->t  
  A->I  
  A->t  
  A->E  
  t [exposure]  
  I [outcome]  
 }")  
  
plot(graphLayout(d4))
```



Estimating the model

$$E(\text{income}) = \beta_0 + \beta_1 \text{hrs} + \beta_2 \text{edu} + \beta_3 \text{age}$$

```
m2<-lm(income ~ hrs_work +
  edu +
  age,
  data = dat)
tidy(m2)

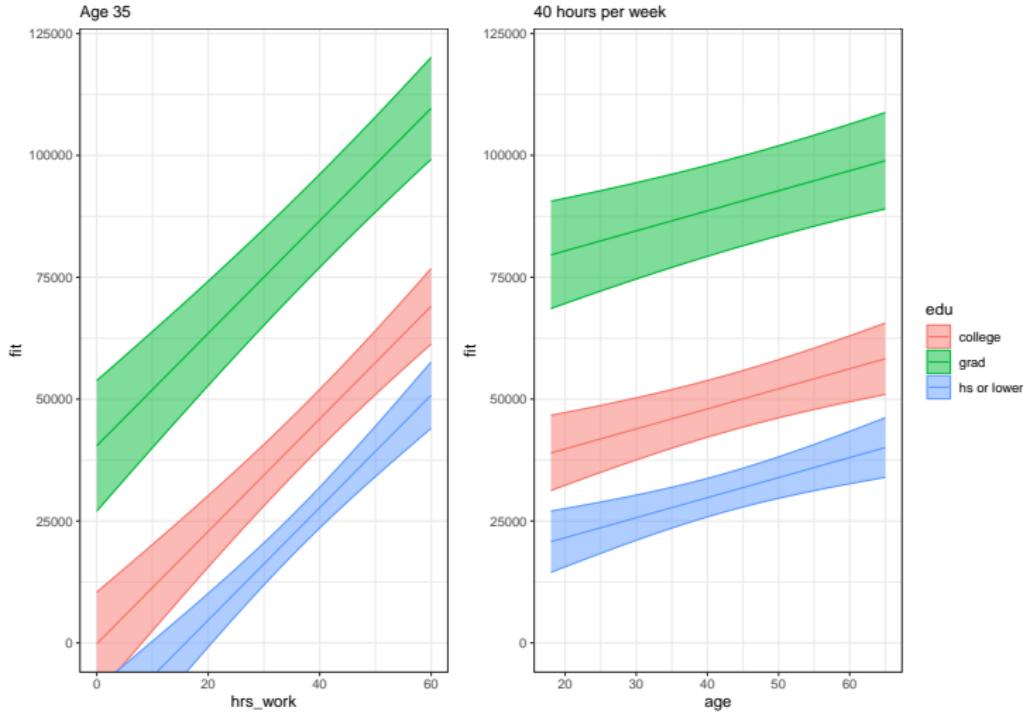
## # A tibble: 5 x 5
##   term      estimate std.error statistic p.value
##   <chr>     <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept) -14596.     6739.    -2.17 3.06e- 2
## 2 hrs_work     1154.      116.      9.95 2.99e-22
## 3 edugrad      40626.     5566.     7.30 6.13e-13
## 4 eduhs or lower -18215.    3553.    -5.13 3.58e- 7
## 5 age          411.       104.      3.94 8.67e- 5
```

Visualizing model expectations: setup

```
# set up prediction data with values of interest
hrs_work<-c(0:60)
age<-c(18:65)
edu<-c("college", "hs or lower", "grad")
pred_dat<-expand_grid(hrs_work, edu, age)
# generate expected values and CI, join pred_dat
e_y<-predict(m2,
              newdata = pred_dat,
              interval = "confidence") |>
  bind_cols(pred_dat)
# inspect
head(e_y)

## # A tibble: 6 x 6
##       fit     lwr     upr hrs_work edu      age
##   <dbl>  <dbl>  <dbl>    <int> <chr>    <int>
## 1 -7197. -18604. 4211.      0 college     18
## 2 -6785. -18118. 4547.      0 college     19
## 3 -6374. -17636. 4887.      0 college     20
## 4 -5963. -17157. 5230.      0 college     21
## 5 -5552. -16681. 5577.      0 college     22
## 6 -5141. -16209. 5927.      0 college     23
```

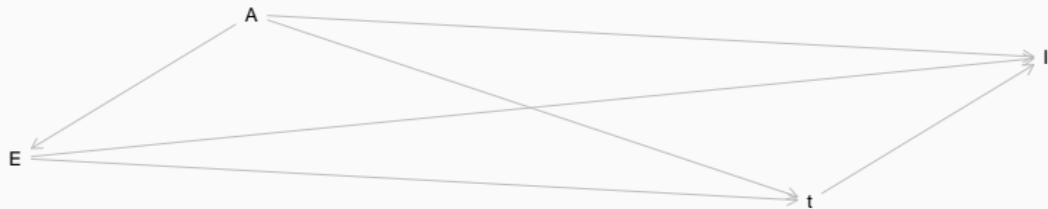
Visualizing model expectations



The difference between a DAG and a specification

This DAG can help us theorize how to adjust our models, but it does not tell us the correct regression specification. Is the relationship between A and I linear?

```
d4<-dagitty("dag {  
  E->I  
  t->I  
  E->t  
  A->I  
  A->t  
  A->E  
  t [exposure]  
  I [outcome]  
 }")  
  
plot(graphLayout(d4))
```

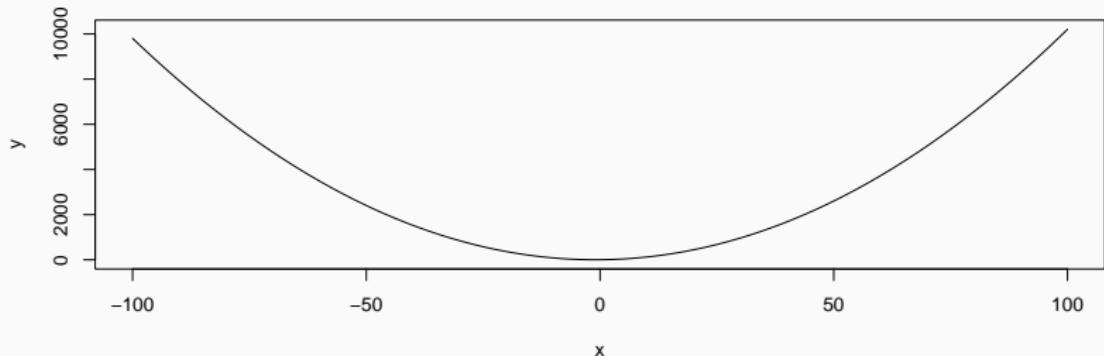


Adding complexity: quadratic terms

We know that earnings for people less than age 18 and greater than age 70 tend to be very low (or zero). We can try to use a parabola (a quadratic equation) to model this process.

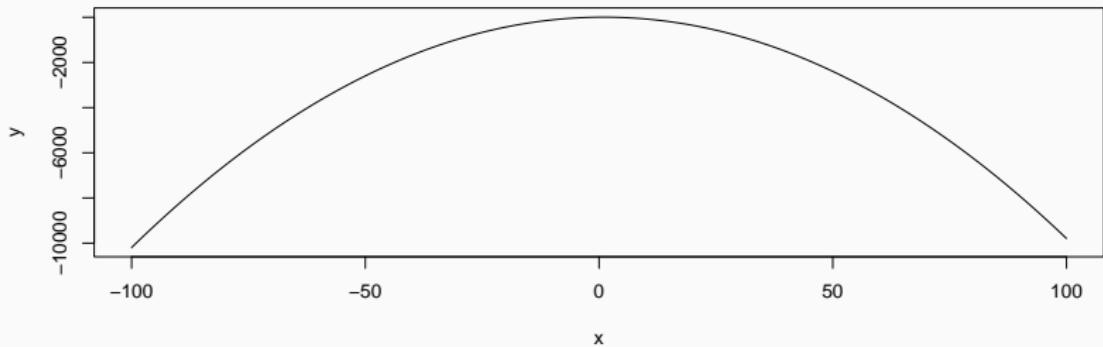
Quadratics take a form that looks like this

```
x<- -100:100  
y<- 5 + 2 * x + x^2  
plot(x, y, type = "l")
```



Adding complexity: negative sign

```
x<- -100:100  
y<- 5 + 2 * x - x^2  
plot(x, y, type = "l")
```



Fitting a quadratic term

We use the `I()` function to require R to evaluate math statements inside formula objects

```
m3<-lm(income ~ hrs_work *  
        edu +  
        (age +  
         I(age^2)),  
        data = dat)  
tidy(m3)  
  
## # A tibble: 8 x 5  
##   term            estimate std.error statistic p.value  
##   <chr>          <dbl>     <dbl>      <dbl>    <dbl>  
## 1 (Intercept) -56158.    14084.     -3.99  7.20e- 5  
## 2 hrs_work      1530.     214.       7.14  1.90e-12  
## 3 edugrad       19318.    20210.      0.956 3.39e- 1  
## 4 eduhs or lower 12569.    10480.      1.20  2.31e- 1  
## 5 age           1718.     582.       2.95  3.25e- 3  
## 6 I(age^2)      -14.4     6.49      -2.21  2.70e- 2  
## 7 hrs_work:edugrad 471.     465.       1.01  3.11e- 1  
## 8 hrs_work:eduhs or lower -792.    257.      -3.08  2.10e- 3
```

Visualizing model expectations

```
## # A tibble: 6 x 6
##       fit     lwr      upr hrs_work edu      age
##     <dbl>   <dbl>    <dbl>    <int> <chr>    <int>
## 1 -29889. -48189. -11589.     0 college    18
## 2 -28703. -46828. -10577.     0 college    19
## 3 -27545. -45523. -9567.     0 college    20
## 4 -26416. -44273. -8560.     0 college    21
## 5 -25316. -43075. -7558.     0 college    22
## 6 -24245. -41927. -6563.     0 college    23
```

