Regression part 4: interactions

Frank Edwards

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, 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", "~
## $ citizen      <chr> "yes", "yes", "yes", "yes", "yes", "yes", "yes", "yes", "~
## $ 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", "~
## $ 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~
```

The use of regression

Sometimes we use regression to estimate causal relationships (e.g. The Mark of a Criminal Record).

Sometimes we use regression for pure prediction (e.g. election forecasts)

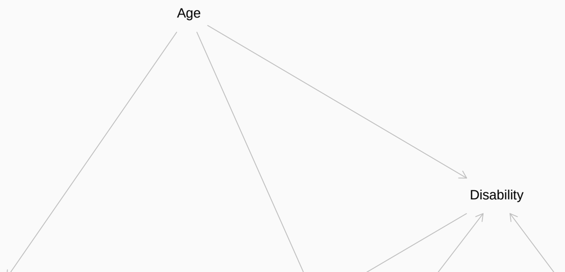
Sometimes we use regression to help us better understand and describe a process that depends on many variables.

Building a model to approximate the data generating process

1. Develop an explicit theoretical model
2. Evaluate data availability and quality
3. Experiment with model specification
4. Evaluate goodness-of-fit metrics
5. Evaluate the *predictive distribution* relative to the *empirical distribution*

So what processes *cause* income to vary across people?

```
library(dagitty)
d1 <- dagitty("dag {
            Education->Income
            Race->Income
            Gender->Income
            Age->Income
            Race->Education
            Age->Education
            Disability->Income
            Age->Disability
            Race->Disability
            Gender->Disability
            Income [outcome]
            Education [exposure]
            }")
plot(graphLayout(d1))
```



Let's check our data

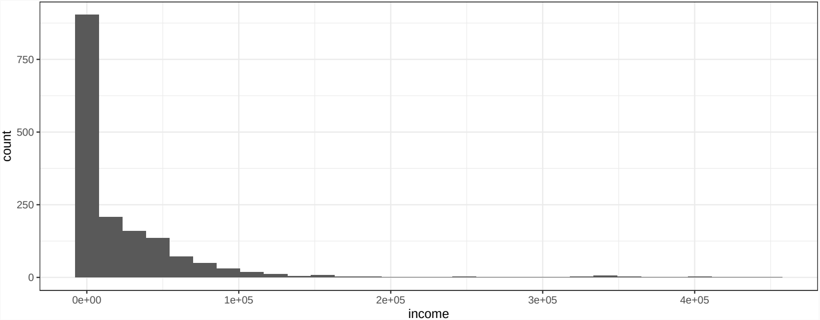
```
summary(dat$income)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0       0    3000   23600   33700  450000     377
```

```
table(dat$income > 0)
```

```
##
## FALSE  TRUE
##   729   894
```

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



Let's check our data

```
summary(dat$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   19.75   40.00   40.22   59.00   94.00
```

```
table(dat$edu)
```

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

```
table(dat$disability)
```

```
##
##   no  yes
## 1676  324
```

```
table(dat$gender)
```

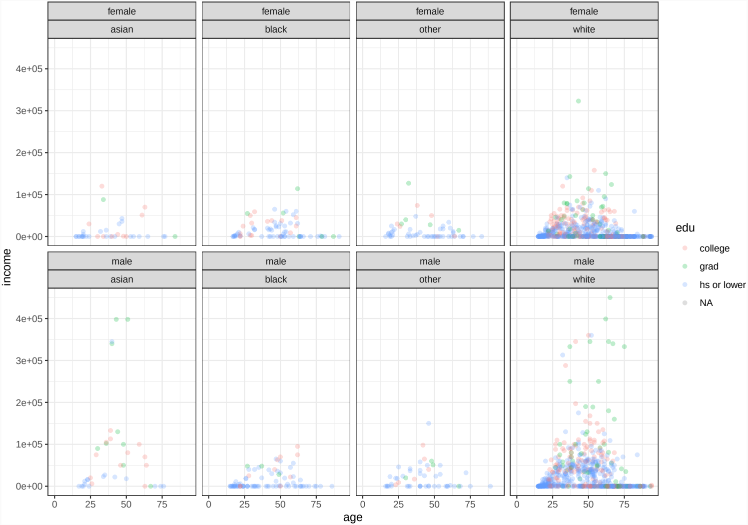
```
##
## female   male
##    969   1031
```

```
table(dat$race)
```

```
##
## asian black other white
##    87   206   152  1555
```

Let’s check our data

```
ggplot(dat, aes(y = income, x = age, color = edu, fill = edu)) + geom_point(alpha = 0.25) +
    facet_wrap(gender ~ race, ncol = 4)
```



Fitting a preliminary model

Our theory tells us that income is a function of age, disability, education, race, and gender. It doesn't tell us what form those function take though!

Let's start simple and additive

```
m0 <- lm(income ~ edu + age + race + disability + gender, data = dat)
```

This model can be written as

$$y_i = \beta_0 + \beta_1 edu_i + \beta_2 age_i + \beta_3 race_i + \beta_4 disability_i + \beta_5 gender_i + \varepsilon_i$$

Evaluating our model fit

```
m0 <- lm(income ~ edu + age + race + disability + gender, data = dat)
```

People with grad degrees, conditional on being Asian, female, and not disabled, have an expected income of 45000 + 31000.

The proportion of variation in the outcome explained by these predictors is $R^2 = 0.18$

Proportion of variance explained

The coefficient of determination, R^2, provides one measure of *goodness-of-fit*. We compute R^2 by taking the ratio of the sum of squared residuals (absolute error in our regression model) and the total sum of squares for the outcome (the sum of squared deviations from the mean).

R^2 tells us how much of the variation in our outcome is explained by the regression line $y = \beta X$ compared to the line $y = \bar{y}$

GoF basics

```r
mod1 <- lm(income ~ age, data = dat)
summary(mod1)$r.squared
```

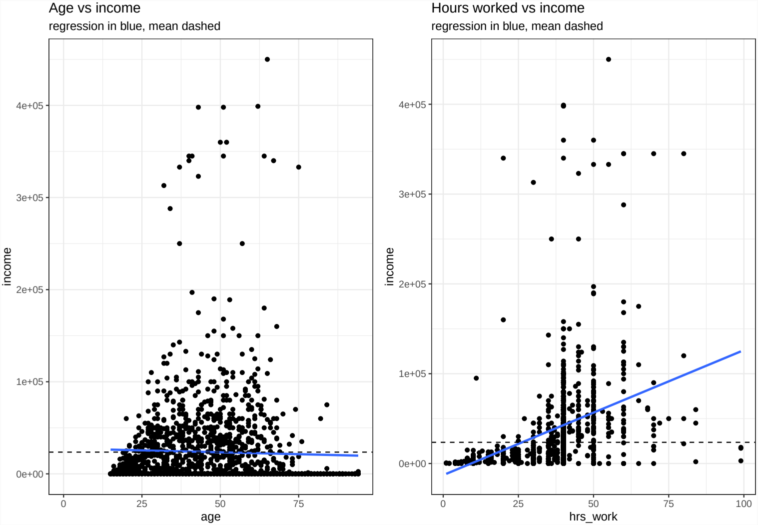
```
## [1] 0.001198718
```

```r
mod2 <- lm(income ~ hrs_work, data = dat)
summary(mod2)$r.squared
```

```
## [1] 0.1168822
```

Which model is a better fit?

GoF visualized



GoF as reduction in error

```r
## How much residual error is there in model 1?
sum(mod1$residuals^2)
```

```
## [1] 3.513102e+12
```

```r
## and how much in model 2?
sum(mod2$residuals^2)
```

```
## [1] 2.55348e+12
```

So let’s estimate and compare some models

```r
# our additive model
m0 <- lm(income ~ edu + age + race + disability + gender, data = dat)
# maybe education-> income varies by gender?
m1 <- lm(income ~ edu * gender + age + race + disability, data = dat)

summary(m0)$r.squared
```

```
## [1] 0.1840489
```

```r
summary(m1)$r.squared
```

```
## [1] 0.199946
```

So let’s estimate and compare some models

```
# maybe education-> income varies by gender and race?
m2 <- lm(income ~ edu * (gender + race) + age + disability, data = dat)

summary(m1)$r.squared
```

```
## [1] 0.199946
```

```
summary(m2)$r.squared
```

```
## [1] 0.2153131
```

So let's estimate and compare some models

```r
# maybe education-> income varies by race/gender pairs?
m3 <- lm(income ~ edu * (gender * race) + age + disability, data = dat)

summary(m3)$r.squared
```

```
## [1] 0.2261352
```

```r
summary(m2)$r.squared
```

```
## [1] 0.2153131
```

Let's go nuts

```
# maybe education-> income varies by race/gender pairs?
m4 <- lm(income ~ edu * (gender * race * age * disability), data = dat)

summary(m3)$r.squared
```

```
## [1] 0.2261352
```

```
summary(m4)$r.squared
```

```
## [1] 0.2576525
```

When are we just overfitting?

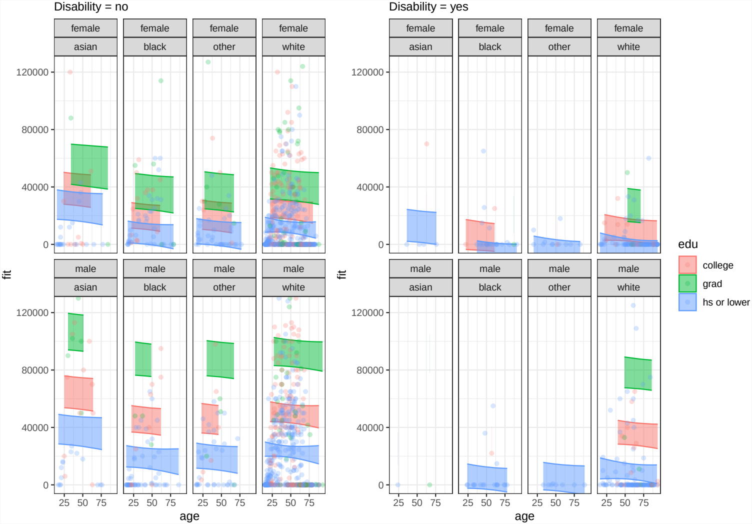
The Bayesian Information Criterion (BIC) provides a check against overfitting. It evaluates goodness of fit with a penalty for model complexity (degrees of freedom).

```
BIC(m0, m1, m2, m3, m4)
```

```
##    df      BIC
## m0 10 39238.80
## m1 12 39221.65
## m2 18 39234.53
## m3 27 39278.52
## m4 79 39595.42
```

OK - we've looked at GoF, but can our model make reasonable predictions?

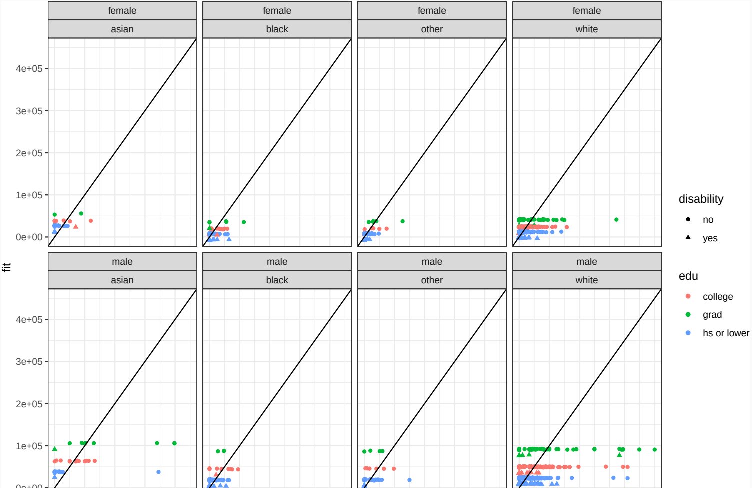
Let's check the distribution of expected values from our model(s) against the empirical distributions



Another way to look at this

Fitted vs observed plots can be very informative

```
ggplot(preds1, aes(x = income, y = fit, color = edu, shape = disability)) + geom_point() +
    geom_abline() + facet_wrap(gender ~ race, ncol = 4) + coord_cartesian(xlim = c(0,
    max(dat$income, na.rm = T)), ylim = c(0, max(dat$income, na.rm = T)))
```



Still not very satisfying, huh

Let's try again. This time, only including people with income > 0 and a multiplicative set of relationships

```
m0 <- lm(log(income) ~ edu + age + race + disability + gender, data = dat %>%
    filter(income > 0))
# maybe education-> income varies by gender?
m1 <- lm(log(income) ~ edu * gender + age + race + disability, data = dat %>%
    filter(income > 0))

BIC(m0, m1)
```

```
##    df      BIC
## m0 10 2955.528
## m1 12 2968.873
```

OK, now compare fitted vs observed

```r
preds2 <- dat %>%
    filter(income > 0) %>%
    bind_cols(predict(m1, interval = "confidence"))

ggplot(preds2, aes(x = income, y = exp(fit), color = edu, shape = disability)) +
    geom_point() + geom_abline() + facet_wrap(gender ~ race, ncol = 4) + coord_cartesian(xlim = c(0,
    max(dat$income, na.rm = T)), ylim = c(0, max(dat$income, na.rm = T)))
```