# Descriptive regression and model fit

Frank Edwards

# The use of regression

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

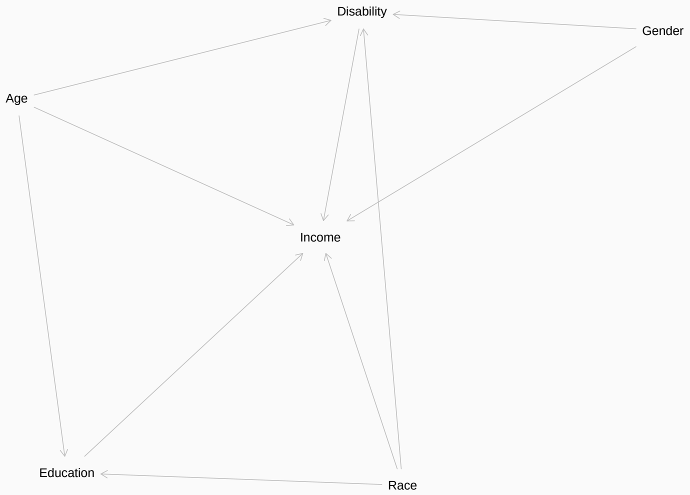
Sometimes we use regression for pure prediction (e.g. election forecasts)

**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?



# Let’s check our data

```
dat<-read_csv("https://www.openintro.org/data/csv/acs12.csv")
### subset to in labor force
dat <- dat |>
  filter(employment != "not in labor force")
glimpse(dat)
```

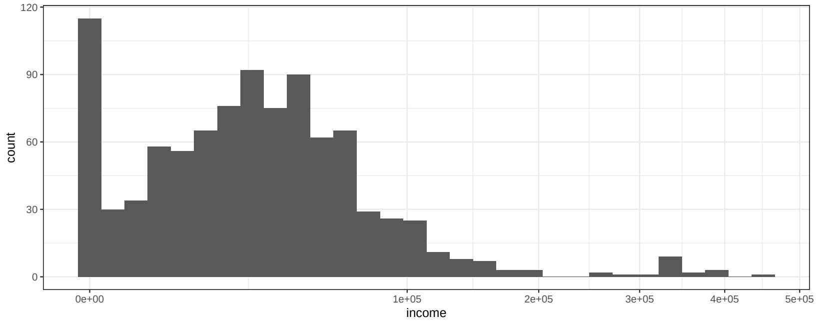
```
## Rows: 949
## Columns: 13
## $ income       <dbl> 1700, 45000, 8600, 33500, 4000, 19000, 3400, 0, 140000, 0~
## $ employment   <chr> "employed", "employed", "employed", "employed", "employed~
## $ hrs_work     <dbl> 40, 84, 23, 55, 8, 35, 25, NA, 40, 8, 23, 72, 40, 50, 35,~
## $ race         <chr> "other", "white", "white", "white", "white", "white", "wh~
## $ age          <dbl> 35, 27, 69, 52, 67, 36, 40, 27, 35, 31, 32, 35, 51, 50, 2~
## $ gender       <chr> "female", "male", "female", "male", "female", "female", "~
## $ citizen      <chr> "yes", "yes", "yes", "yes", "yes", "yes", "yes", "yes", "~
## $ time_to_work <dbl> 15, 40, 5, 20, 10, 15, NA, NA, 30, 20, 45, 25, 10, 40, 10~
## $ lang         <chr> "other", "english", "english", "english", "english", "eng~
## $ married      <chr> "yes", "yes", "no", "yes", "yes", "yes", "no", "no", "no"~
## $ edu          <chr> "hs or lower", "hs or lower", "hs or lower", "hs or lower~
## $ disability   <chr> "yes", "no", "no", "no", "no", "no", "yes", "no", "no", "~
## $ birth_qrtr   <chr> "jul thru sep", "oct thru dec", "jul thru sep", "apr thru~
```

# The distribution of income among those in the labor forceit ad

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0    6600   25200   39808   50000  450000
```

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



# Let’s check our data

```r
dat |> group_by(race, gender) |>
  summarize(n = n()) |>
  knitr::kable()
```

| race | gender | n |
|---|---|---:|
| asian | female | 14 |
| asian | male | 28 |
| black | female | 48 |
| black | male | 48 |
| other | female | 34 |
| other | male | 35 |
| white | female | 324 |
| white | male | 418 |

# 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 with R^2

```
##
## Call:
## lm(formula = income ~ edu + age + race + disability + gender,
##     data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -115914  -23209   -4333   12883  332880
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     39810.9     8960.6   4.443 9.93e-06 ***
## edugrad         45547.5     5671.5   8.031 2.88e-15 ***
## eduhs or lower -18364.0     3603.7  -5.096 4.19e-07 ***
## age               603.3      108.9   5.540 3.93e-08 ***
## raceblack      -38705.4     9023.7  -4.289 1.98e-05 ***
## raceother      -39660.9     9520.7  -4.166 3.39e-05 ***
## racewhite      -29874.8     7720.1  -3.870 0.000116 ***
## disabilityyes  -16771.6     5452.3  -3.076 0.002158 **
## gendermale      22421.5     3165.0   7.084 2.74e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48220 on 940 degrees of freedom
## Multiple R-squared:  0.247,  Adjusted R-squared:  0.2406
## F-statistic: 38.54 on 8 and 940 DF,  p-value: < 2.2e-16
```

## Proportion of variance explained

The coefficient of determination, $R^2$, provides one measure of *goodness-of-fit*.

$$R^2 = \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

$R^2$ tells us how much of the variation in $y$ 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
```

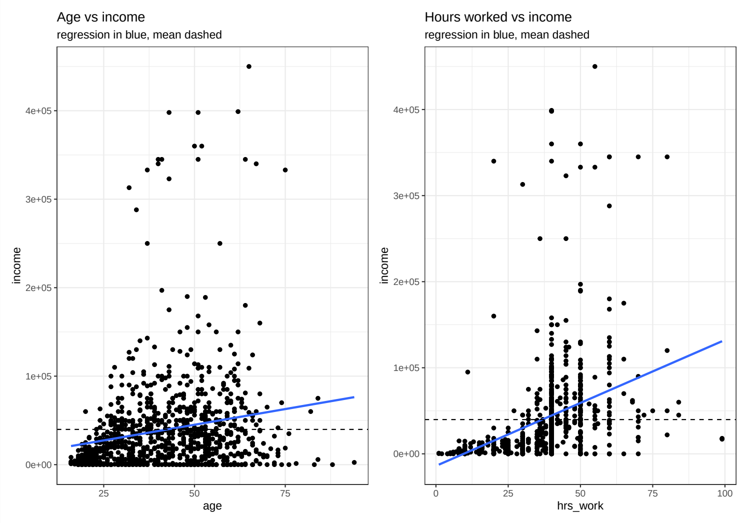
```
## [1] 0.03610956
```

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

```
## [1] 0.1174225
```

Which model is a better fit?

# Have we improved our fit compared to guessing the mean (dotted line)?



# GoF as reduction in error

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

```
## [1] 2.797424e+12
```

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

```
## [1] 2.481356e+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] 0.2469889
```

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

```
## [1] 0.263837
```

# So let's estimate and compare some models

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

summary(m1)$r.squared
```

```
## [1] 0.263837
```

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

```
## [1] 0.2769639
```

# So let’s estimate and compare some models

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

summary(m3)$r.squared
```

```
## [1] 0.287483
```

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

```
## [1] 0.2769639
```

# 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] 0.287483
```

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

```
## [1] 0.321665
```

# When are we just overfitting?

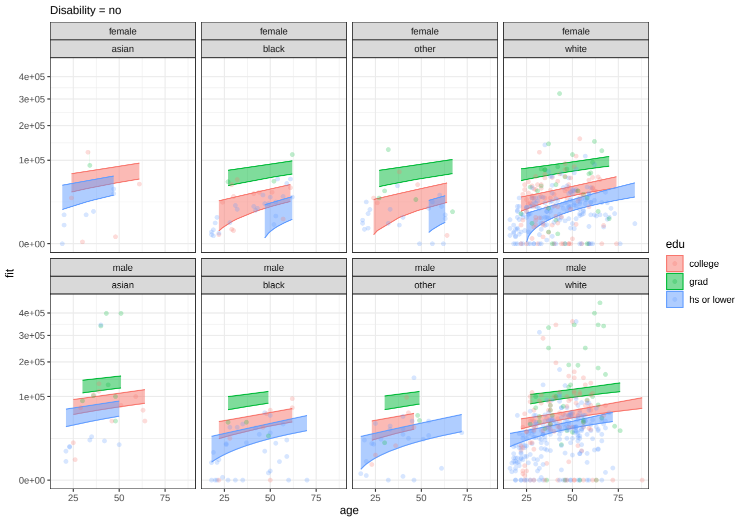
The Bayesian Information Criterion (BIC) provides a check against overfitting. It evaluates goodness of fit with a penalty for complexity (count of model parameters), based on the log-likelihood of the model. The first term $k \ln(n)$ adjusts for model complexity with $n$ as the number of observations and $k$ as the number of model parameters ($\beta$)

$$BIC = k \ln(n) - 2 \ln(\hat{L})$$

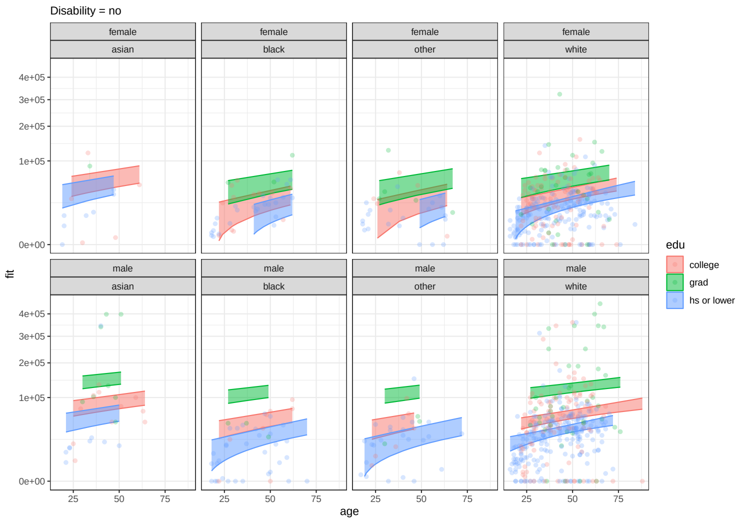
```r
BIC(m0, m1, m2, m3, m4)
```

```
##    df      BIC
## m0 10 23219.69
## m1 12 23211.92
## m2 18 23235.98
## m3 27 23283.77
## m4 72 23545.61
```

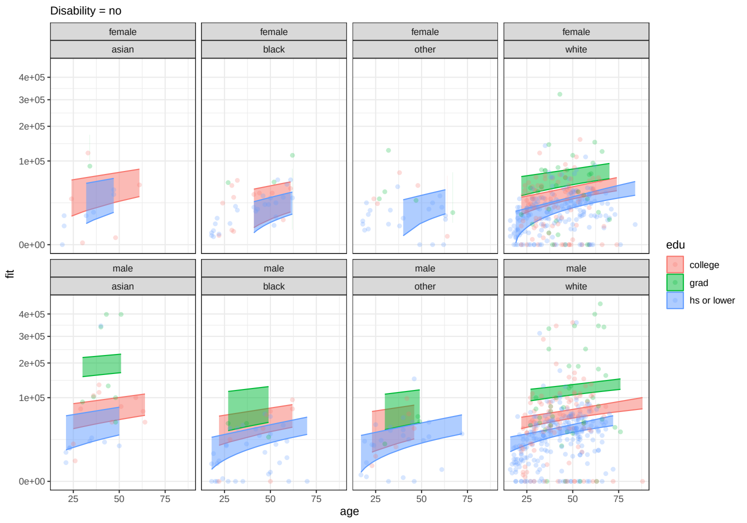
# Visualizing observed versus expected Model 0



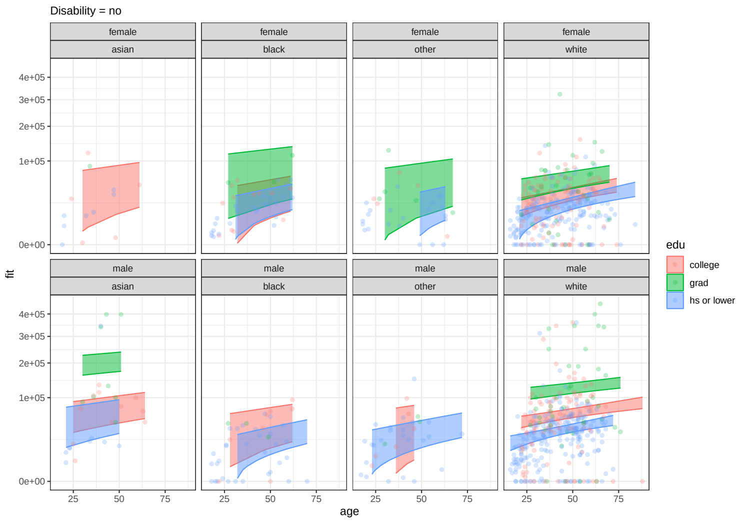
# Visualizing observed versus expected Model 1



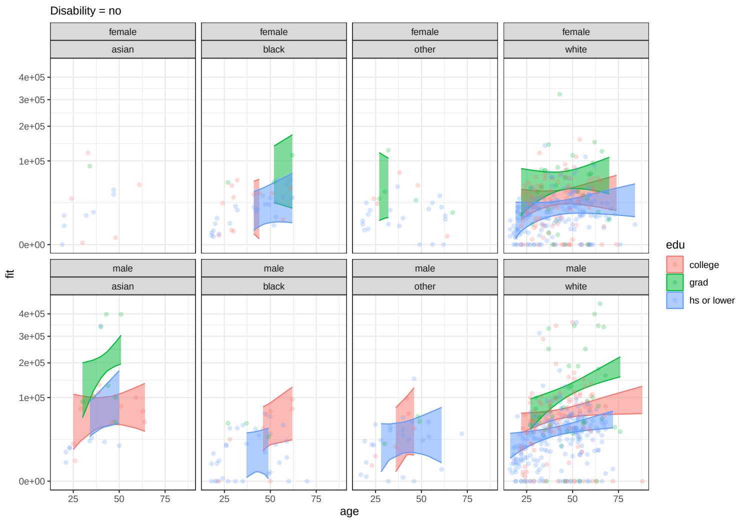
# Visualizing observed versus expected Model 2



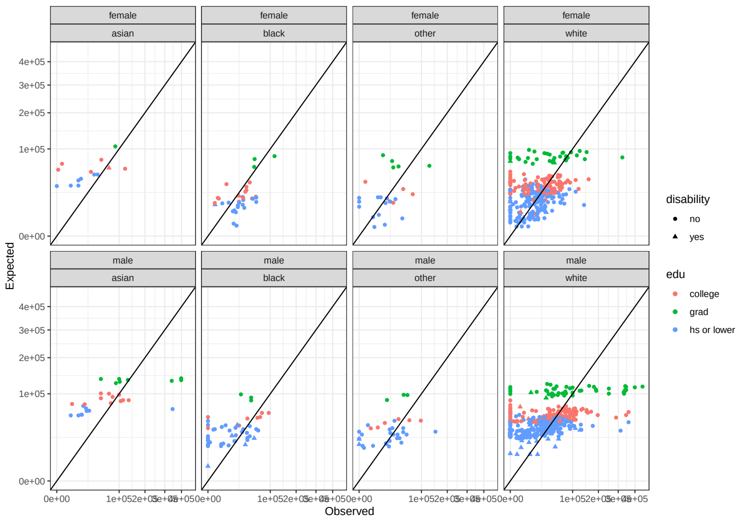
# Visualizing observed versus expected Model 3



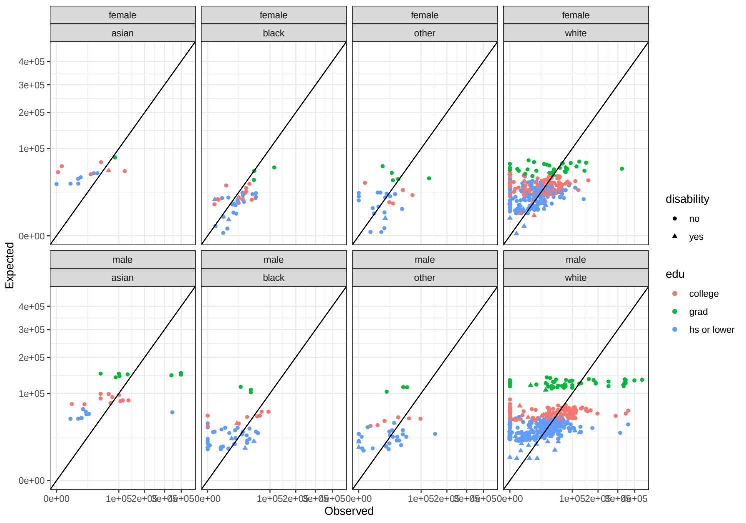
# Visualizing observed versus expected Model 4



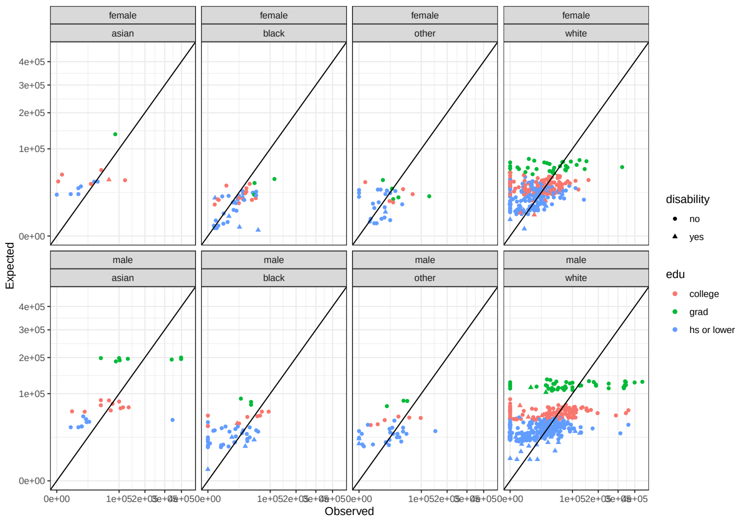
# Fitted vs observed plots can be very informative: Model 0



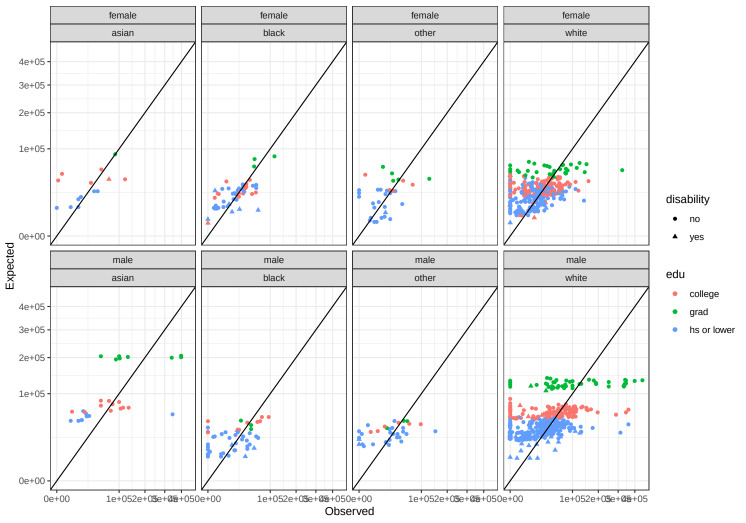
# Fitted vs observed plots can be very informative: Model 1



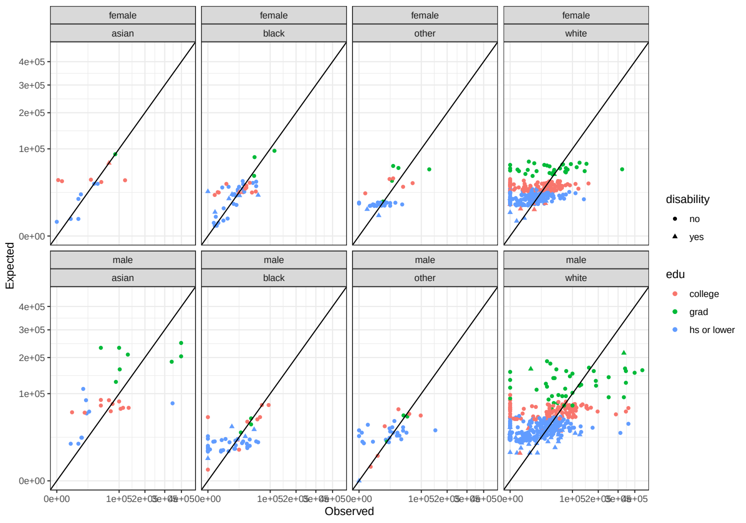
# Fitted vs observed plots can be very informative: Model 2



# Fitted vs observed plots can be very informative: Model 3



# Fitted vs observed plots can be very informative: Model 4



# Which model is best?

It depends on our target!

| | model | r2 | BIC.df | BIC.BIC |
|---|---|---|---|---|
| m0 | 0 | 0.2469889 | 10 | 23219.69 |
| m1 | 1 | 0.2638370 | 12 | 23211.92 |
| m2 | 2 | 0.2769639 | 18 | 23235.98 |
| m3 | 3 | 0.2874830 | 27 | 23283.77 |
| m4 | 4 | 0.3216650 | 72 | 23545.61 |

## General advice

When fitting a model for *descriptive* or *predictive* purposes

1. Choose predictors based on theory
2. Experiment with varying function forms (additive, interactive, nonlinear)
3. Compare goodness of fit using $R^2$, but also use BIC and other criteria robust to overfitting (leave-one-out is gold standard)
4. Evaluate expected versus observed, evaluate regression line against empirical data
5. Next time: simulate new data from your regression and evaluate it against the observed