Week 7: Linear regression III

Monica Alexander

28/02/2021

This markdown file contains all the code used in the lecture.

Interactions

Read in data:

```
library(tidyverse)
library(here)
country_ind <- read_csv(here("data/country_indicators.csv"))
gss <- read_csv(here("data/gss.csv"))
```

TFR and life expectancy

Filter data and create an indicator variable based on country region:

```
country_ind_2017 <- country_ind %>%
  filter(year==2017) %>%
  mutate(dev_region = ifelse(region=="Developed regions", "yes", "no"))
```

Regression:

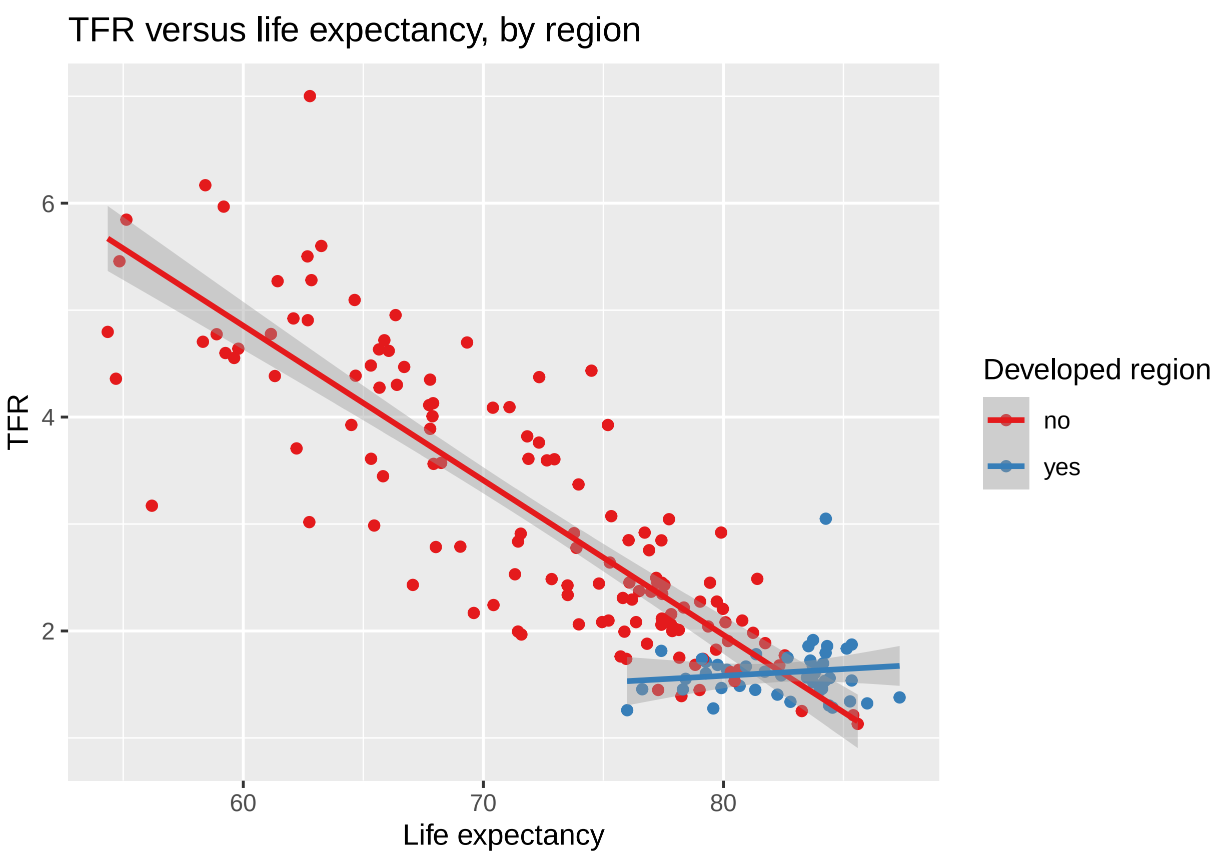
```
mod <- lm(tfr ~ life_expectancy + dev_region + life_expectancy*dev_region, data = country_ind_2017)
summary(mod)
```

```
##
## Call:
## lm(formula = tfr ~ life_expectancy + dev_region + life_expectancy *
##     dev_region, data = country_ind_2017)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.23326 -0.29618 -0.02426  0.28744  2.54832
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    13.52646    0.52158  25.933  < 2e-16 ***
## life_expectancy                -0.14454    0.00722 -20.019  < 2e-16 ***
## dev_regionyes                 -12.95159    2.91594  -4.442 1.59e-05 ***
```

```
## life_expectancy:dev_regionyes   0.15711    0.03557   4.417 1.76e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6164 on 172 degrees of freedom
## Multiple R-squared:  0.7784, Adjusted R-squared:  0.7745
## F-statistic: 201.4 on 3 and 172 DF,  p-value: < 2.2e-16
```

Plot:

```r
ggplot(aes(life_expectancy, tfr, color = dev_region), data = country_ind_2017) +
  geom_point() + geom_smooth(method = "lm") +
  ggtitle("TFR versus life expectancy, by region") +
  ylab("TFR") + xlab("Life expectancy") +
  scale_color_brewer(name = "Developed region", palette = "Set1")
```



Age at first marriage and age

Create a new age variable that is mean centered:

```r
gss <- gss %>% mutate(age_c = age - mean(age))
```

Regression without interaction:

```r
mod2 <- lm(age_at_first_marriage~ age_c + has_bachelor_or_higher, data = gss)
summary(mod2)
```

```
##
## Call:
## lm(formula = age_at_first_marriage ~ age_c + has_bachelor_or_higher,
##     data = gss)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.046  -3.379  -1.254   2.026  27.226
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               24.489886   0.116125 210.892   <2e-16 ***
## age_c                     -0.061697   0.006172  -9.996   <2e-16 ***
## has_bachelor_or_higherYes  1.982372   0.184405  10.750   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.295 on 5265 degrees of freedom
##   (15334 observations deleted due to missingness)
## Multiple R-squared:  0.04454,    Adjusted R-squared:  0.04417
## F-statistic: 122.7 on 2 and 5265 DF,  p-value: < 2.2e-16
```

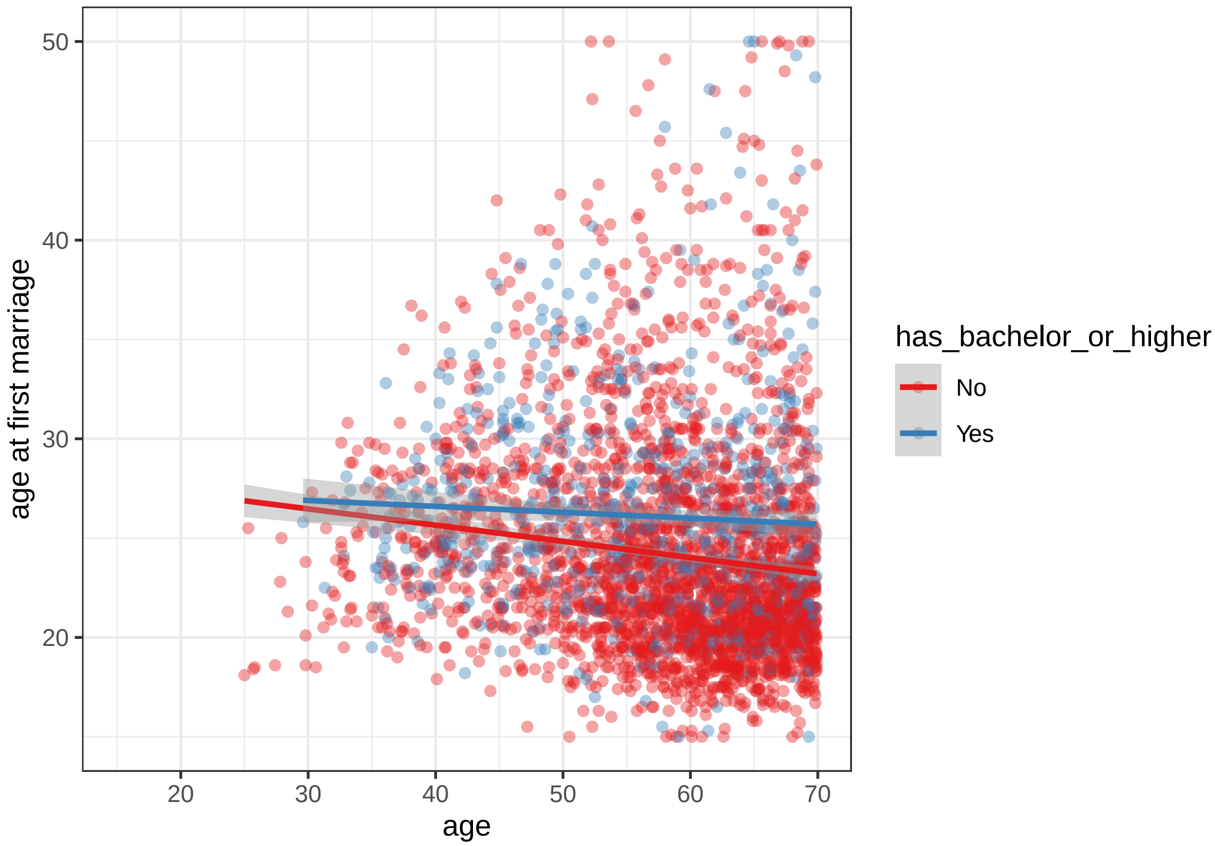
Regression with interaction:

```r
mod3 <- lm(age_at_first_marriage~ age_c*has_bachelor_or_higher, data = gss)
summary(mod3)
```

```
##
## Call:
## lm(formula = age_at_first_marriage ~ age_c * has_bachelor_or_higher,
##     data = gss)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.970  -3.372  -1.211   2.018  27.328
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      24.58532    0.12365 198.837  < 2e-16 ***
## age_c                            -0.06882    0.00694  -9.916  < 2e-16 ***
## has_bachelor_or_higherYes         1.62500    0.24374   6.667 2.88e-11 ***
## age_c:has_bachelor_or_higherYes   0.03397    0.01516   2.241   0.0251 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.293 on 5264 degrees of freedom
##   (15334 observations deleted due to missingness)
## Multiple R-squared:  0.04545,    Adjusted R-squared:  0.0449
## F-statistic: 83.54 on 3 and 5264 DF,  p-value: < 2.2e-16
```

Plot:

```
ggplot(gss %>% drop_na(has_bachelor_or_higher) %>% filter(age<70), aes(x = age, y = age_at_first_marriag
  geom_point(alpha = 0.4)+
  geom_smooth(method = "lm")+
  theme_bw()+
  scale_color_brewer(palette = "Set1")+
  labs(y = "age at first marriage")
```



Problems

For this section, I actually simulated some data. Here's the code to do this. You don't need to understand it, but if you're interested, let me know and I can go through it next week.

```
ages <- 20:70

eps <- rnorm(length(ages)*100, sd = 1)

df <- tibble(age = rep(ages, each = 100),
       truth = log(3000) + 0.01*(age-45)-0.002*(age-45)^2,
       eps = eps,
       income = truth + eps,
       yrs = sample(5:15, length(ages)*100, replace = TRUE),
       sample = income-mean(income) + yrs-10+rnorm(length(ages)*100)>0)
```

```r
# final dataset just showing those surveyed
d <- df %>%
  mutate(income = ifelse(sample, exp(income), NA)) %>%
  select(age, yrs, income)

d
```

```
## # A tibble: 5,100 x 3
##      age   yrs income
##    <int> <int>  <dbl>
##  1    20     6    NA
##  2    20    13   330.
##  3    20     7    NA
##  4    20    12   536.
##  5    20     5    NA
##  6    20    15   748.
##  7    20    14   544.
##  8    20     5    NA
##  9    20     9    NA
## 10    20    12   358.
## # ... with 5,090 more rows
```

EDA

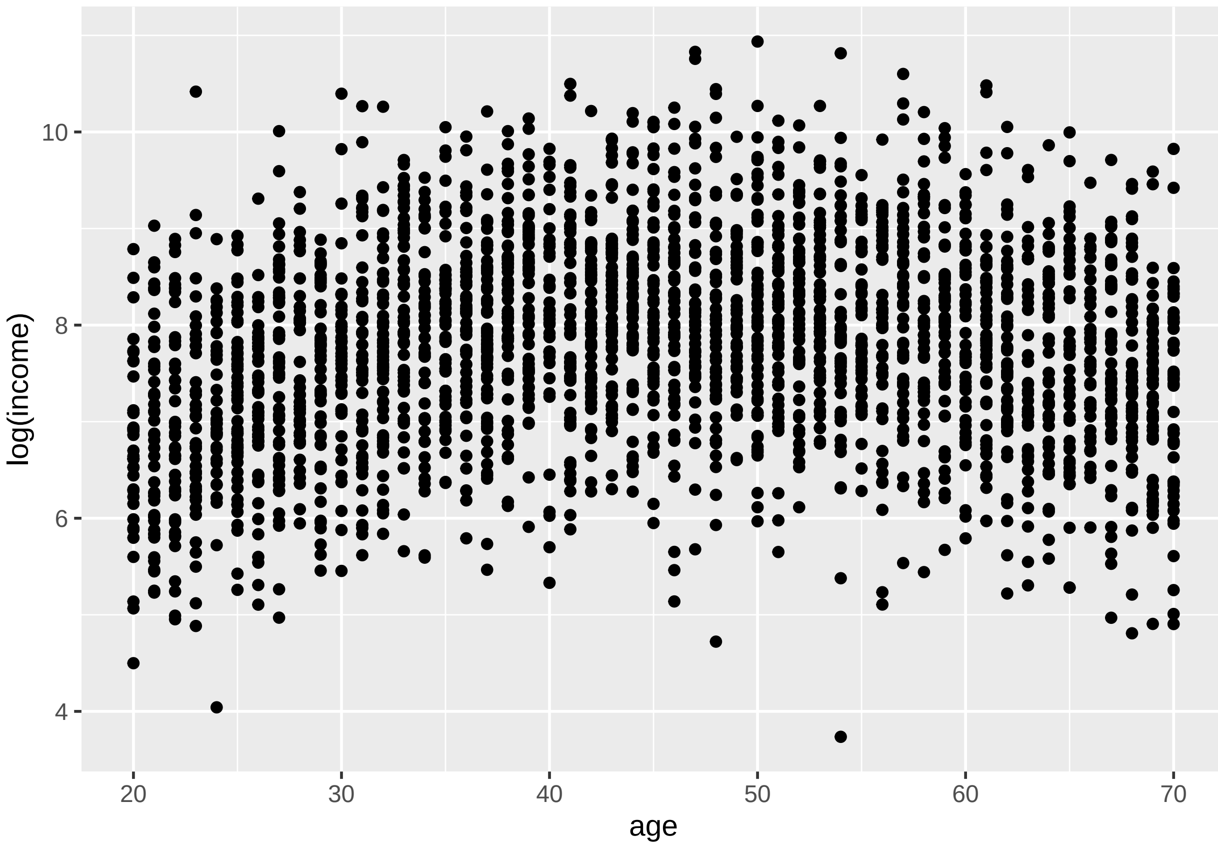
Summary stats:

```r
d %>%
  group_by(yrs) %>%
  summarize(mean_log_income = mean(log(income), na.rm = TRUE),
            n = n(),
            n_income_missing = sum(is.na(income)))
```

```
## # A tibble: 11 x 4
##      yrs mean_log_income     n n_income_missing
##    <int>           <dbl> <int>            <int>
##  1     5           NaN     469              469
##  2     6           NaN     424              424
##  3     7            10.0   492              487
##  4     8             8.84  474              433
##  5     9             8.44  472              355
##  6    10             8.31  445              214
##  7    11             7.90  457              111
##  8    12             7.70  503               43
##  9    13             7.68  452                5
## 10    14             7.60  451                1
## 11    15             7.57  461                0
```

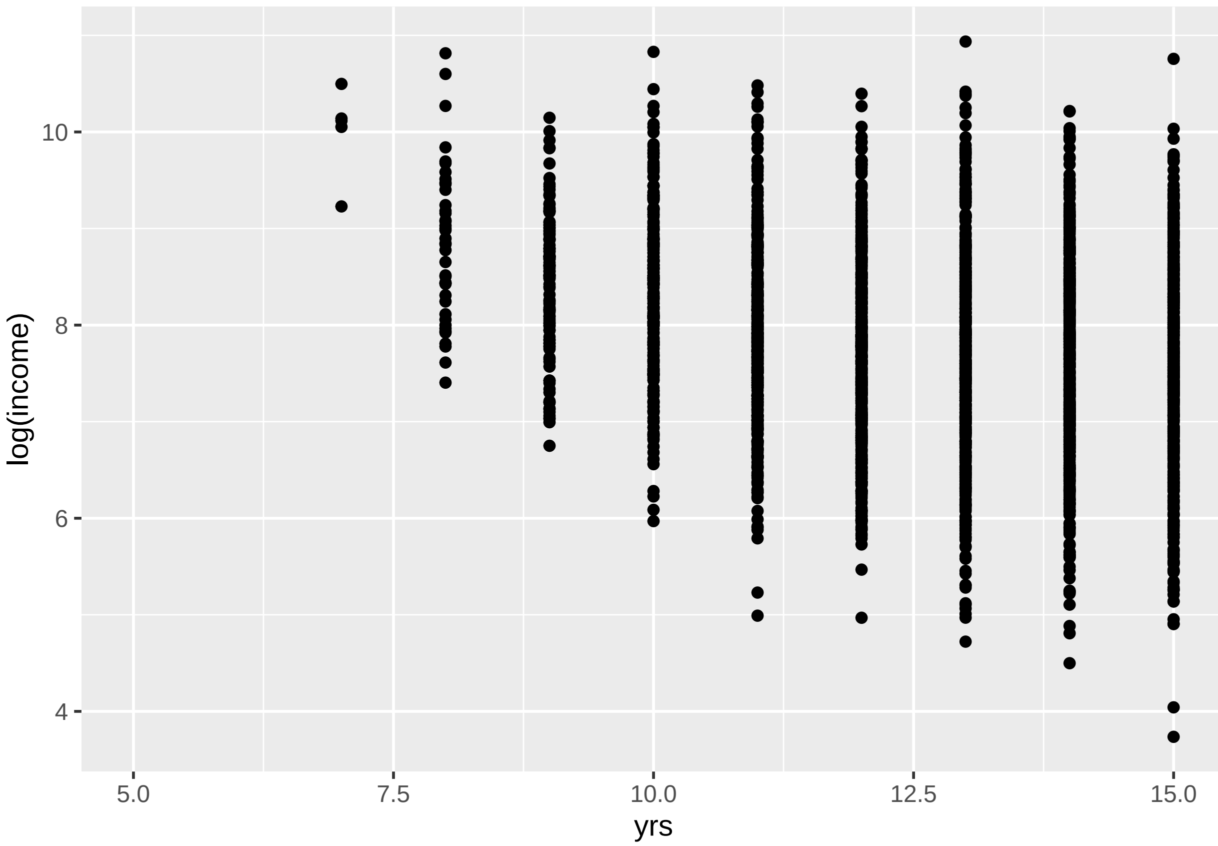
Age versus log income

```r
d %>%
  ggplot(aes(age, log(income))) + geom_point()
```



Education versus log(income)

```
d %>%
  ggplot(aes(yrs, log(income))) + geom_point()
```



Mis-specification

Run model with no squared term

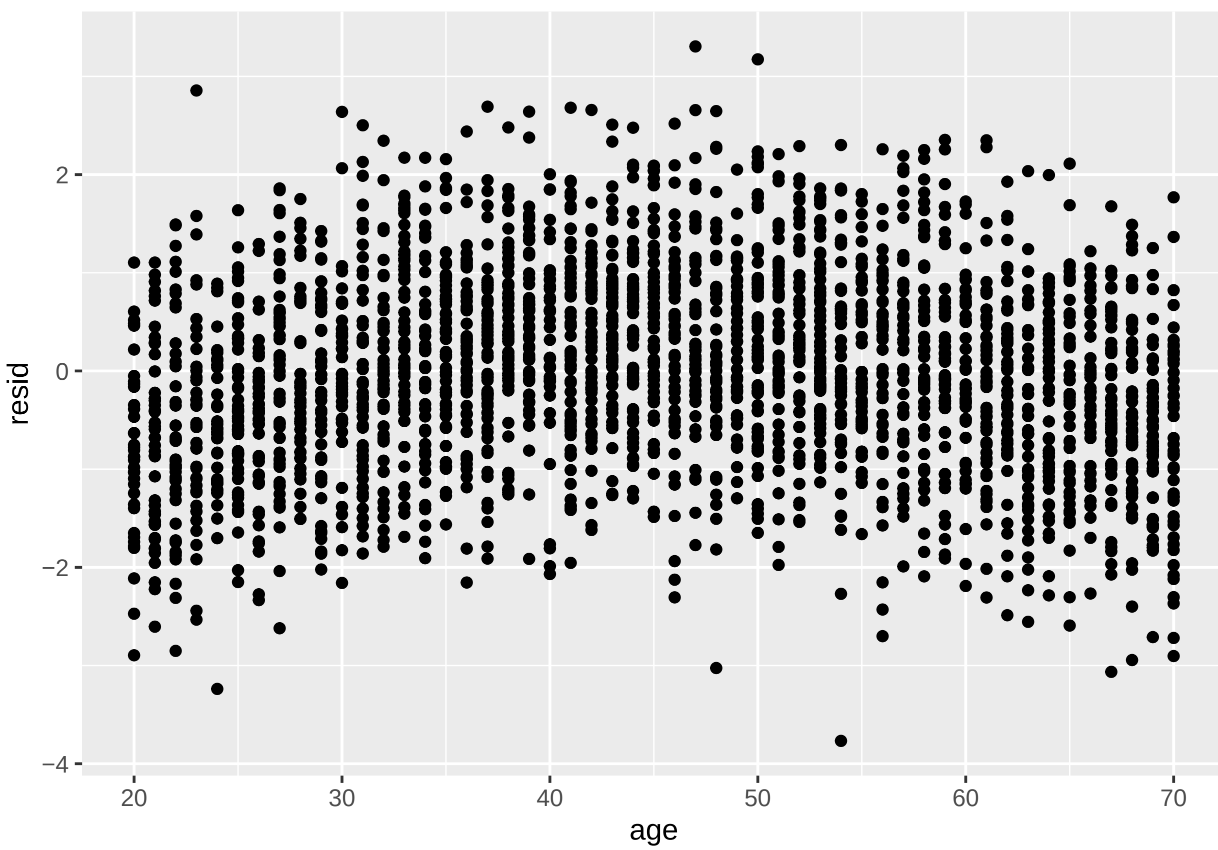
```
d <- d %>% mutate(log_income = log(income))
mod <- lm(data = d, log_income ~ age+yrs)
summary(mod)
```

```
##
## Call:
## lm(formula = log_income ~ age + yrs, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7675 -0.6833 -0.0076  0.7135  3.3051
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.263678   0.151840  61.010  < 2e-16 ***
## age          0.007469   0.001402   5.326 1.09e-07 ***
## yrs         -0.144187   0.010833 -13.310  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.014 on 2555 degrees of freedom
```

```
##   (2542 observations deleted due to missingness)
## Multiple R-squared:  0.07486,    Adjusted R-squared:  0.07413
## F-statistic: 103.4 on 2 and 2555 DF,  p-value: < 2.2e-16
```

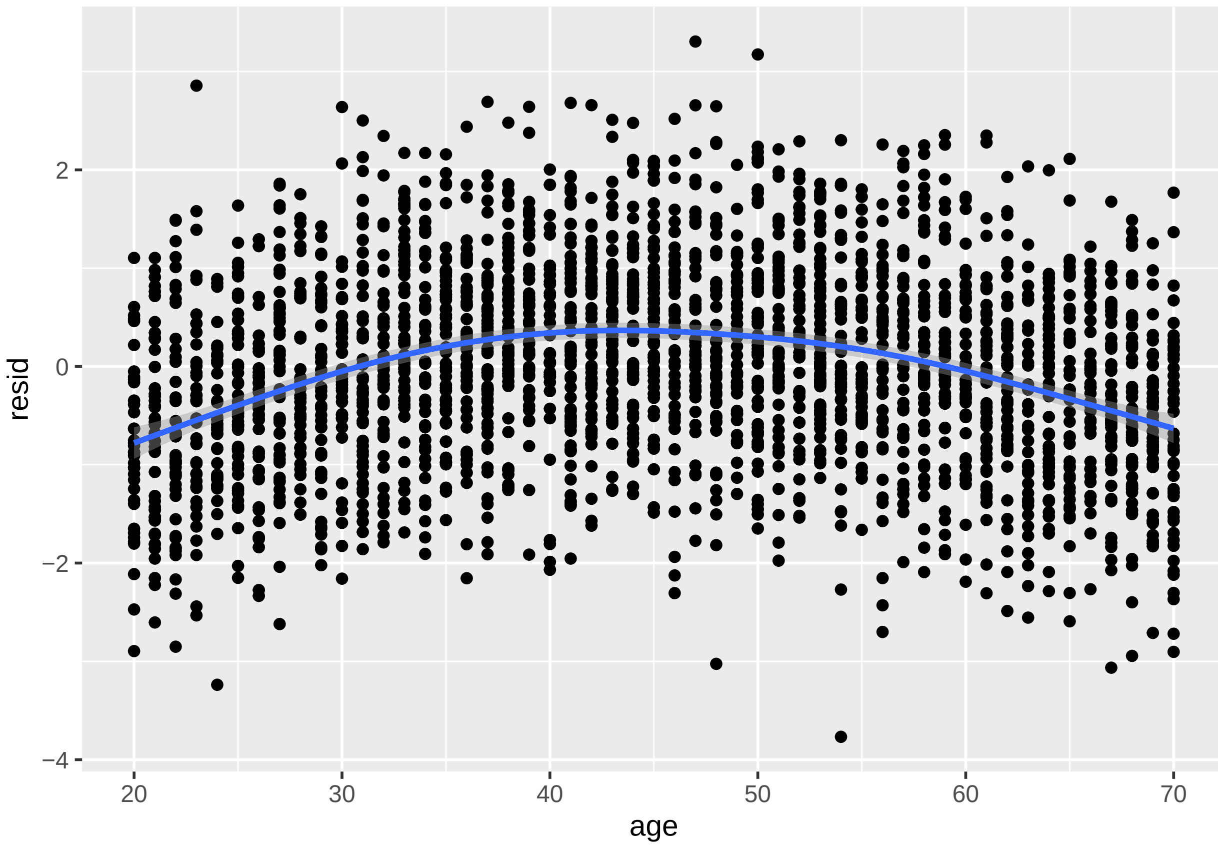
Get residuals and plot

```
df_resid <- tibble(resid = residuals(mod),
                   age = d %>%
                     drop_na() %>%
                     select(age) %>%
                     pull())
ggplot(data = df_resid, aes(age, resid)) + geom_point()
```



Fit a line

```
ggplot(data = df_resid, aes(age, resid)) + geom_point()+ geom_smooth()
```



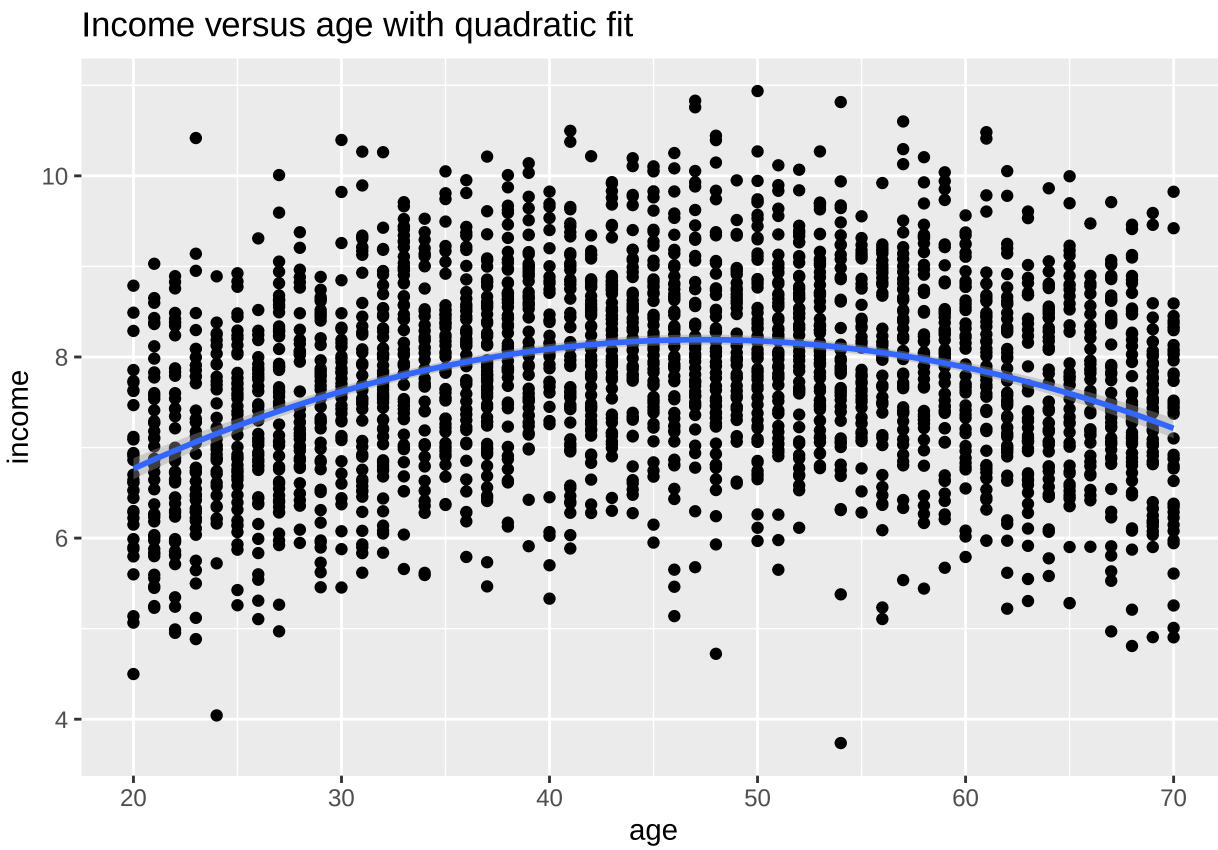
Rerun model with squared term

```
d <- d %>% mutate(age_sq = age^2)
mod2 <- lm(data = d, log_income ~ age+age_sq + yrs)
summary(mod2)
```

```
##
## Call:
## lm(formula = log_income ~ age + age_sq + yrs, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0400 -0.6513 -0.0301  0.6569  3.3797
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.729318   0.241385   23.73   <2e-16 ***
## age          0.170584   0.009074   18.80   <2e-16 ***
## age_sq      -0.001799   0.000099  -18.17   <2e-16 ***
## yrs         -0.127953   0.010236  -12.50   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9539 on 2554 degrees of freedom
##   (2542 observations deleted due to missingness)
## Multiple R-squared:  0.1807, Adjusted R-squared:  0.1798
## F-statistic: 187.8 on 3 and 2554 DF,  p-value: < 2.2e-16
```

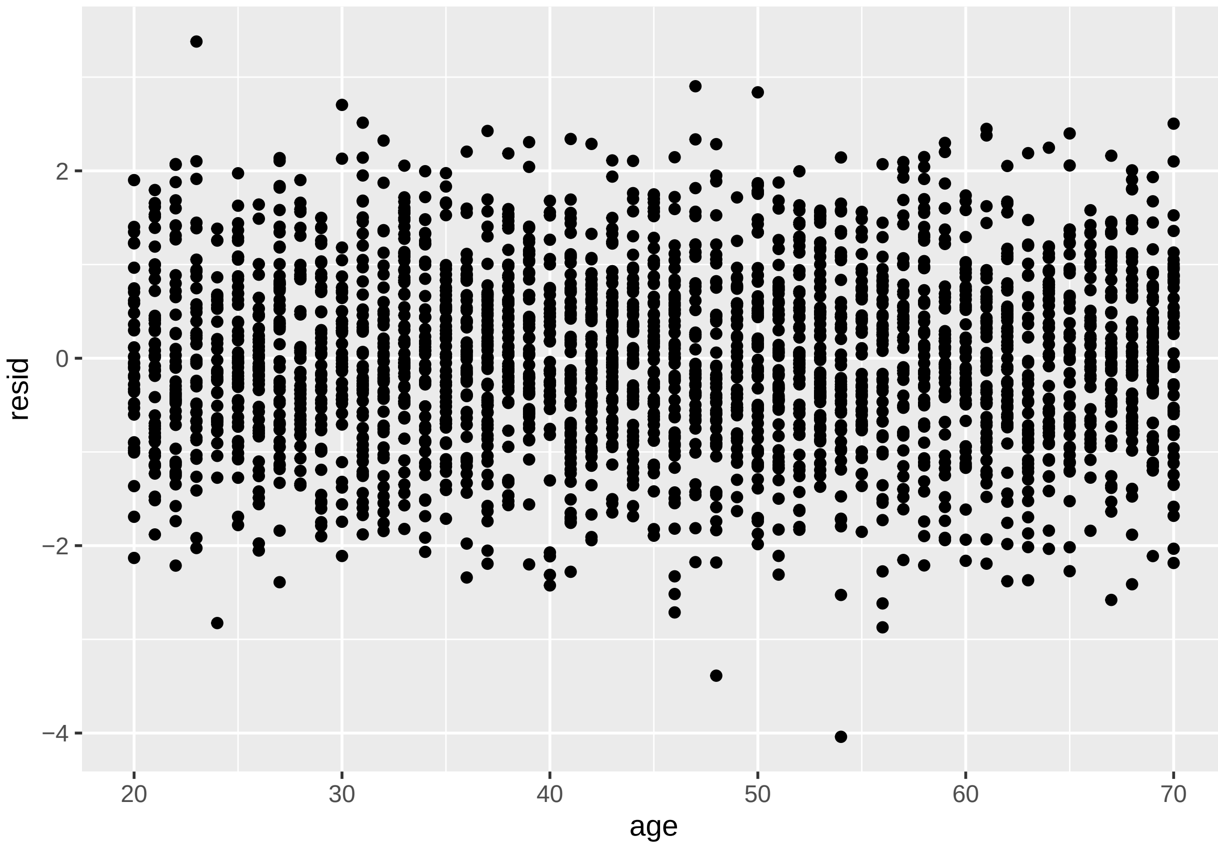
Visualize fit

```
d %>%
  ggplot(aes(age, log_income)) +
  geom_point() + geom_smooth(method = "lm", formula = y ~ poly(x,2)) +
  ylab("income") + xlab("age") + ggtitle("Income versus age with quadratic fit")
```



Redo residuals: much better

```
df_resid <- tibble(resid = residuals(mod2),
                   age = d %>%
                     drop_na() %>%
                     select(age) %>%
                     pull())
ggplot(data = df_resid, aes(age, resid)) + geom_point()
```



Non-response bias

No code here but notice that the dataset we used (`d`) was a filtered version of the full sample (`df`). If you rerun the regressions on `df` you should notice no significant association between income and years schooling.