Lab 2

Y. Samuel Wang

Intro

This lab will explore multiple linear regression and including polynomial terms.

Housing Data

In class, we fit a few models using the housing data that we've been considering in lecture. In lab, we'll take a deeper dive into the data set. First, let's load the data

```
fileName <- url("https://raw.githubusercontent.com/ysamwang/btry6020_sp22/main/lectureData/estate.csv")
housing_data <- read.csv(fileName)

head(housing_data)
```

```
##   id  price area bed bath  ac garage pool year quality style   lot highway
## 1  1 360000 3032   4    4 yes      2   no 1972  medium     1 22221      no
## 2  2 340000 2058   4    2 yes      2   no 1976  medium     1 22912      no
## 3  3 250000 1780   4    3 yes      2   no 1980  medium     1 21345      no
## 4  4 205500 1638   4    2 yes      2   no 1963  medium     1 17342      no
## 5  5 275500 2196   4    3 yes      2   no 1968  medium     7 21786      no
## 6  6 248000 1966   4    3 yes      5  yes 1972  medium     1 18902      no
```

```
View(housing_data)
```

Recall that there are 522 observations with the following variables:

- price: in 2002 dollars
- area: Square footage
- bed: number of bedrooms
- bath: number of bathrooms
- ac: central AC (yes/no)
- garage: number of garage spaces
- pool: yes/no
- year: year of construction
- quality: high/medium/low
- home style: coded 1 through 7
- lot size: sq ft
- highway: near a highway (yes/no)

There is no age data in the table, but we can compute it on our own from the year variable

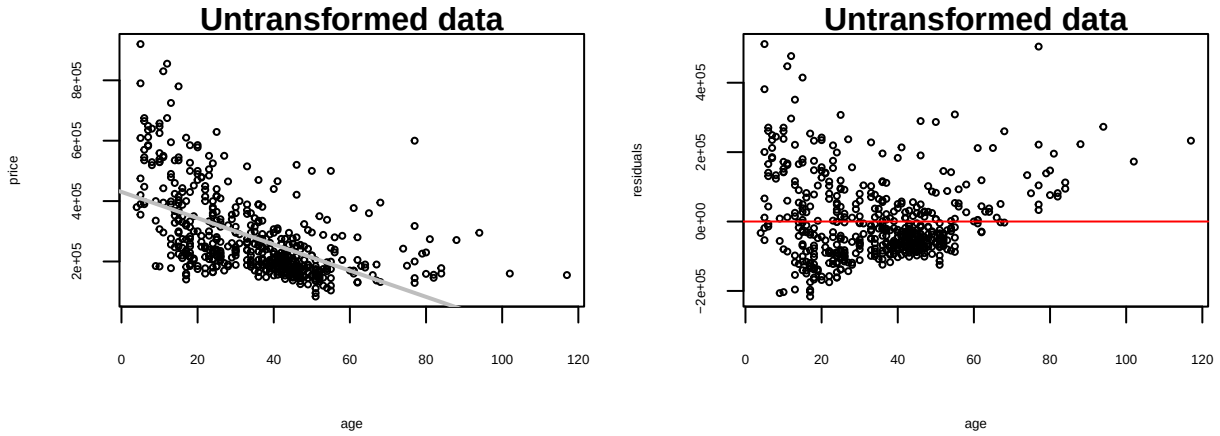
```
housing_data$age <- 2002 - housing_data$year
```

Polynomial regression

We can first fit a linear model to both the data using the age of the house.

```
reg_linear <- lm(price ~ age, data = housing_data)
```

```
par(mfrow = c(2,2), mar = c(4, 4, 1, 1))
plot(housing_data$age, housing_data$price, cex.lab = .5, cex.axis = .5,
     cex = .5, main = "Untransformed data", xlab = "age", ylab = "price")
abline(a = reg_linear$coef[1], b = reg_linear$coef[2], col = "gray", lwd = 2)
plot(housing_data$age, reg_linear$res, cex.lab = .5, cex.axis = .5,
     cex = .5, main = "Untransformed data", xlab = "age", ylab = "residuals")
abline(h = 0, col = "red")
```



Questions

- Does it look like the linear model is a good fit for the data? Why or why not?

As an alternative, we can also use polynomial regression. Let's include the covariate of age squared.

```
## R requires you to use I(age^2) instead of just including age^2
reg_quad1 <- lm(price ~ age + I(age^2), data = housing_data)
summary(reg_quad1)
```

```
##
## Call:
## lm(formula = price ~ age + I(age^2), data = housing_data)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -280265  -55366  -21785   49671  432273
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 552349.97   15860.35   34.83   <2e-16 ***
## age         -11922.03     795.81  -14.98   <2e-16 ***
## I(age^2)        93.34       9.26   10.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 105100 on 519 degrees of freedom
## Multiple R-squared:  0.4218, Adjusted R-squared:  0.4196
## F-statistic: 189.3 on 2 and 519 DF,  p-value: < 2.2e-16
```

The variables, age and age squared will be quite correlated, which as we will see on Wednesday can be a bad thing. So we typically will want to use a transformation of the polynomial covariates which are not as highly correlated. We will use the `poly` function which takes the covariate and the degree of the polynomial (in this case 2) and return a set of covariates which act like age and age squared, but are not correlated. It's also easier to type out instead of including a bunch of terms by hand. The coefficients aren't directly interpretable since the covariates aren't exactly age and age squared anymore, but we can see that they give the same fitted values as before.

```
reg_quad2 <- lm(price ~ poly(age,2), data = housing_data)
summary(reg_quad2)
```

```
##
## Call:
## lm(formula = price ~ poly(age, 2), data = housing_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -280265  -55366  -21785   49671  432273
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     277894       4599   60.42   <2e-16 ***
## poly(age, 2)1 -1748855     105079  -16.64   <2e-16 ***
## poly(age, 2)2  1059186     105079   10.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 105100 on 519 degrees of freedom
## Multiple R-squared:  0.4218, Adjusted R-squared:  0.4196
## F-statistic: 189.3 on 2 and 519 DF,  p-value: < 2.2e-16
```

```
sum(abs(reg_quad1$fitted.values - reg_quad2$fitted.values))
```

```
## [1] 1.461012e-08
```

We can compare the RSS of the linear model and the model which includes the quadratic term:

```
sum((housing_data$price - reg_linear$fitted.values)^2)
```

```
## [1] 6.852419e+12
```

```
sum((housing_data$price - reg_quad2$fitted.values)^2)
```

```
## [1] 5.730544e+12
```

Alternatively, we can calculate the R^2 of each model:

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

```
## [1] 0.3085985
```

```
summary(reg_quad1)$r.squared
```

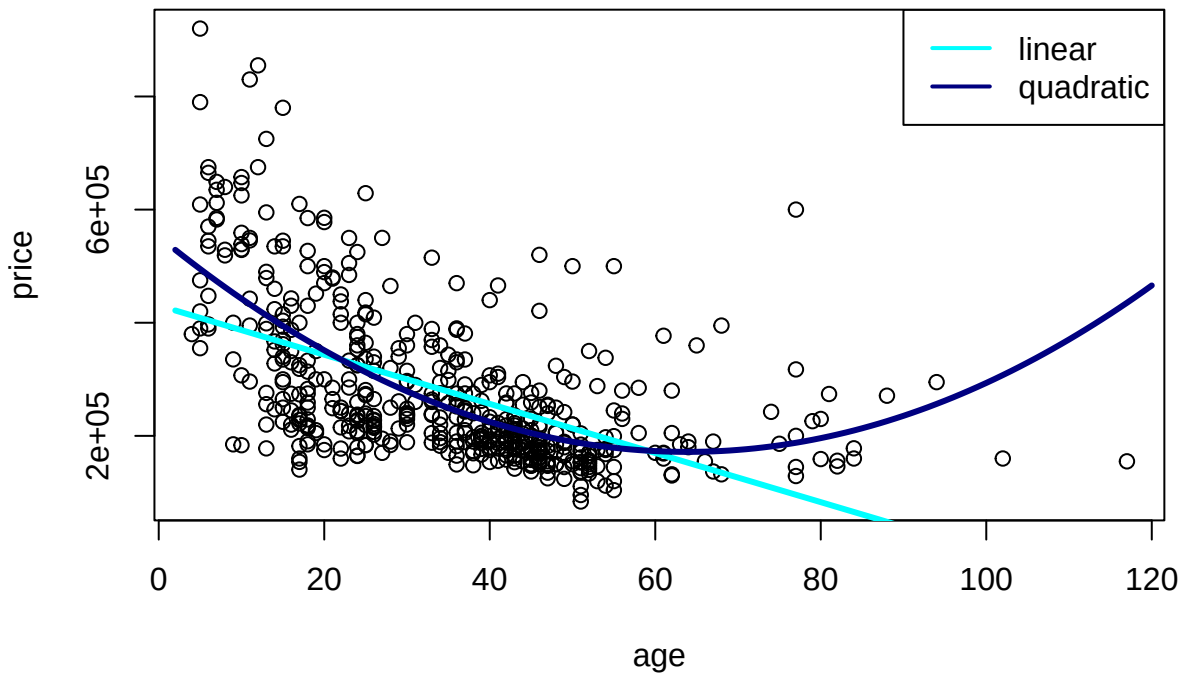
```
## [1] 0.4217944
```

```
summary(reg_quad2)$r.squared
```

```
## [1] 0.4217944
```

We can also plot the fitted prices for each model. For this, we will use the `predict` function. The `predict` function takes an `lm` object and a data frame of covariate observations. It then computes the predicted value of the covariate observations based on the coefficients estimated in the `lm` object.

```
plot(housing_data$age, housing_data$price, xlab = "age", ylab = "price")
lines(2:120, predict(reg_linear, data.frame(age = 2:120)),
      col = "cyan", lwd = 3)
lines(2:120, predict(reg_quad1, data.frame(age = 2:120)),
      col = "navy", lwd = 3)
legend("topright", col = c( "cyan", "navy"),
       legend = c("linear", "quadratic"), lwd = 2)
```



Question:

- With your neighbors, discuss which model you would use if you were fitting the data?
- What if you were trying to explain this model to a collaborator?
- What if you were just trying to predict what you should sell your house for?
- What if the house you are selling is 150 years old?

Can we improve the quadratic model? Let's see if we can just fit higher polynomials to the data. Using a 3rd degree polynomial is called a cubic and using a 4th degree polynomial is called a quartic.

```
reg_cubic <- lm(price ~ poly(age,3), data = housing_data)
reg_quartic <- lm(price ~ poly(age,4), data = housing_data)

summary(reg_quad2)
```

```
##
## Call:
## lm(formula = price ~ poly(age, 2), data = housing_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -280265  -55366  -21785   49671  432273
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     277894       4599   60.42   <2e-16 ***
## poly(age, 2)1 -1748855     105079  -16.64   <2e-16 ***
## poly(age, 2)2  1059186     105079   10.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 105100 on 519 degrees of freedom
## Multiple R-squared:  0.4218, Adjusted R-squared:  0.4196
## F-statistic: 189.3 on 2 and 519 DF,  p-value: < 2.2e-16
```

```
summary(reg_cubic)
```

```
##
## Call:
## lm(formula = price ~ poly(age, 3), data = housing_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -293191  -55948  -22012   47322  420616
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     277894       4502  61.728  < 2e-16 ***
## poly(age, 3)1 -1748855     102857 -17.003  < 2e-16 ***
## poly(age, 3)2  1059186     102857  10.298  < 2e-16 ***
## poly(age, 3)3  -500361     102857  -4.865 1.52e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 102900 on 518 degrees of freedom
## Multiple R-squared:  0.4471, Adjusted R-squared:  0.4439
## F-statistic: 139.6 on 3 and 518 DF,  p-value: < 2.2e-16
```

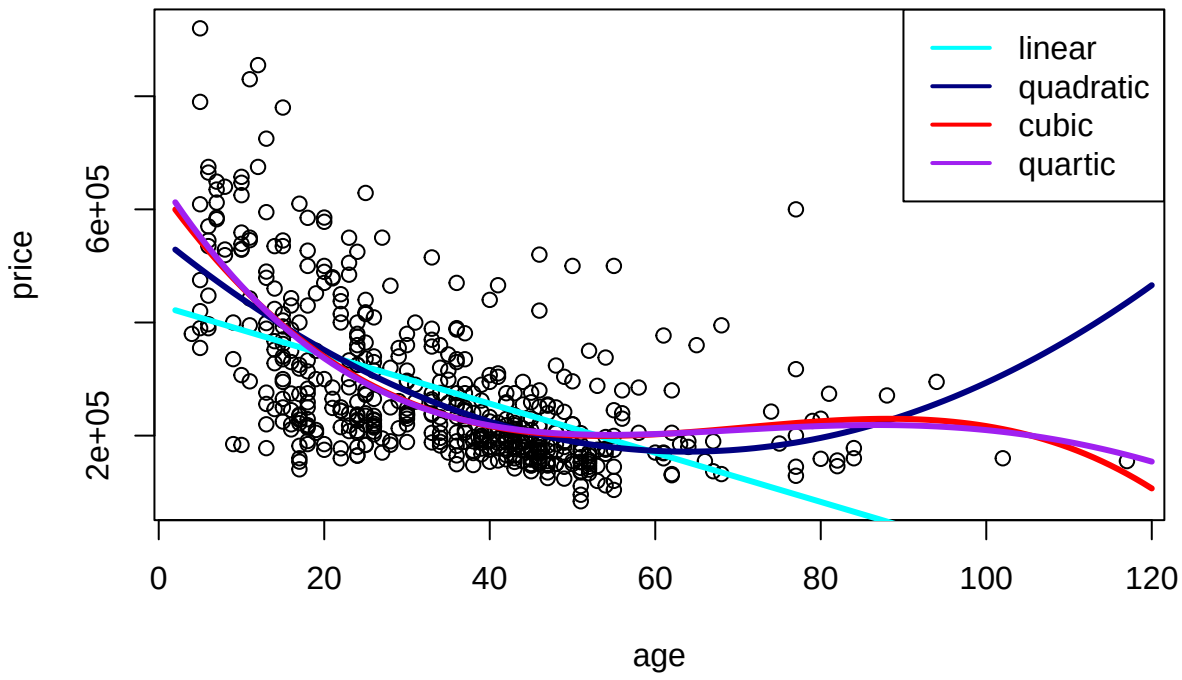
```
summary(reg_quartic)
```

```
##
## Call:
## lm(formula = price ~ poly(age, 4), data = housing_data)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -295682  -56477  -22001   47865  420627
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     277894       4504  61.696  < 2e-16 ***
## poly(age, 4)1 -1748855     102910 -16.994  < 2e-16 ***
## poly(age, 4)2  1059186     102910  10.292  < 2e-16 ***
## poly(age, 4)3  -500361     102910  -4.862 1.54e-06 ***
## poly(age, 4)4    70294     102910   0.683    0.495
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 102900 on 517 degrees of freedom
## Multiple R-squared:  0.4476, Adjusted R-squared:  0.4433
## F-statistic: 104.7 on 4 and 517 DF,  p-value: < 2.2e-16
```

```r
plot(housing_data$age, housing_data$price, xlab = "age", ylab = "price")

lines(2:120, predict(reg_linear, data.frame(age = 2:120)),
      col = "cyan", lwd = 3)
lines(2:120, predict(reg_quad1, data.frame(age = 2:120)),
      col = "navy", lwd = 3)
lines(2:120, predict(reg_cubic, data.frame(age = 2:120)),
      col = "red", lwd = 3)
lines(2:120, predict(reg_quartic, data.frame(age = 2:120)),
      col = "purple", lwd = 3)
legend("topright", col = c("cyan", "navy", "red", "purple"),
       legend = c("linear", "quadratic", "cubic", "quartic"), lwd = 2)
```



Question:

- Examine the RSS for each of the models. Each time we fit a higher order polynomial, the RSS decreases. Will this always be the case or is it just a coincidence? Why do you think so?
- How would you decide which model to use?

Multiple Linear Regression

The rest of today's lab will have less instruction, so it is on you, as a budding statistician to provide a bit of creativity and apply what we have learned so far. In addition, we will use this data set for the module 2 assessment.

We will be looking at recent data from the UK Brexit vote. If you, aren't familiar you can read more about the whole story here http://www.vox.com/2016/6/17/11963668/brexit-uk-eu-explained.

In particular, the response variable we will be using is the percentage of individuals who voted to remain in the European Union in each local authority. We will be looking at several explanatory variables including

- Percentage of individuals born in the UK
- Percentage of individuals with no formal education beyond compulsory education
- Percentage of individuals working in manufacturing
- Percentage of individuals working in finance
- Percentage of individuals over the age of 60
- Percentage of individuals between the ages of 20 and 35

Each row in the data represents a local authority/distict in either England or Wales. The Brexit vote took place in 2016, and the explanatory variables were collected in the 2011 census. Local Authorities with missing data have been removed.

```
fileName <-
brexit.data <-
  read.csv("https://raw.githubusercontent.com/ysamwang/btry6020_sp22/main/lab2/uk_data.csv")
head(brexit.data)
```

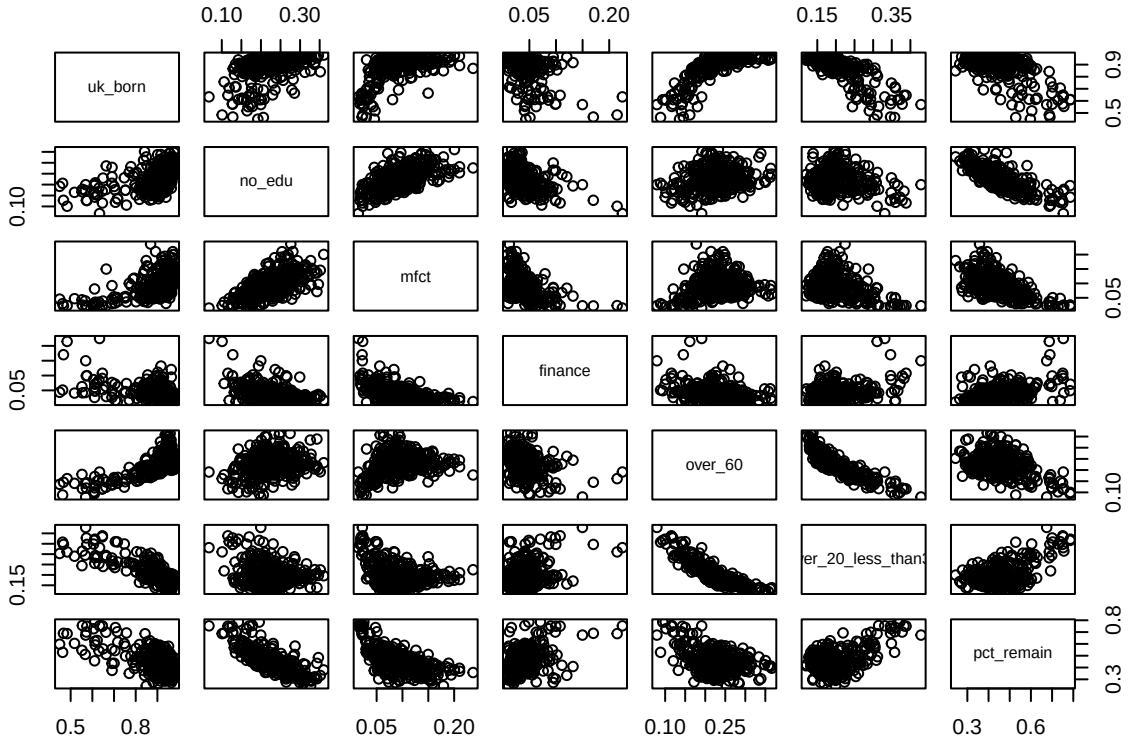
```
##              geography   uk_born    no_edu       mfct    finance   over_60
## 1           Darlington 0.9475295 0.2481226 0.09997144 0.03835639 0.2263366
## 2        County Durham 0.9675338 0.2750048 0.13156555 0.02221647 0.2360407
## 3           Hartlepool 0.9721932 0.3065959 0.11676861 0.02089125 0.2211066
## 4        Middlesbrough 0.9178539 0.2989068 0.08121437 0.02489596 0.1934081
## 5       Northumberland 0.9717588 0.2387226 0.09236833 0.02368942 0.2635178
## 6 Redcar and Cleveland 0.9776663 0.2842061 0.10318700 0.01957270 0.2520029
##   over_20_less_than35 pct_remain
## 1           0.1926604     0.4382
## 2           0.1937371     0.4245
## 3           0.1911049     0.3043
## 4           0.2263821     0.3452
## 5           0.1636817     0.4589
## 6           0.1780406     0.3381
```

Questions

- What direction do you think the association is between each of these variables?
- What strength do you think the association is between each of these variables?

Again, we'll use the `pairs` command to plot the many pairs of variables at once. Note that we've excluded the first column here, since that's just the name of local authority

```
pairs(brexit.data[, -1])
```



Questions

- Does this look like what you might expect?
- What sticks out?
- Do the relationships look roughly linear?

Multivariate Regression

When there are multiple variables, we still use the regular `lm` command, but we need to specify more variables in our formula. Notice now on the right hand side of the $\sim$, we have multiple variables which are separated by the + sign. We can add additional variables simply by using the + sign.

```
output <- lm(pct_remain ~ uk_born + no_edu, data = brexit.data)
summary(output)
```

```
##
## Call:
## lm(formula = pct_remain ~ uk_born + no_edu, data = brexit.data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.132640 -0.035044 -0.005769  0.030399  0.206090
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  1.01641    0.02606   39.00   <2e-16 ***
## uk_born     -0.32934    0.03220  -10.23   <2e-16 ***
## no_edu      -1.19710    0.06604  -18.13   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05556 on 341 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.6818, Adjusted R-squared:  0.6799
## F-statistic: 365.3 on 2 and 341 DF,  p-value: < 2.2e-16
```

We can see from the summary of our model that the estimated model is

$$\hat{y}_i = \hat{b}_0 + \hat{b}_{\text{uk born}} x_{i,\text{uk born}} + \hat{b}_{\text{no edu}} x_{i,\text{no edu}}$$

where $b_{uk\_born} = -.33$ and $b_{no\_edu} = -1.20$.

We can get the residuals and fitted values from the `lm` objects, and we can look at the values for specific geographic areas. For instance, "Eden" is the 23 row in the list. We can see that by using the `which` function. The function returns the index for which the statement evalues to "TRUE." This means the 23rd element of geography vector is equal to "Eden." In the residual and fitted values vector, the 23rd element corresponds to the values for "Eden"

```
which(brexit.data$geography == "Eden")
```

```
## [1] 23
```

```
output$residuals[23]
```

```
##         23
## 0.03981415
```

```
output$fitted.values[23]
```

```
##        23
## 0.4269859
```

Questions

- How would you interpret each of the estimated coefficients above?
- Does the magnitude (size) of the coefficients agree with what you would've guessed?

Now is your chance to explore the data yourself. Using the form above, fit a regression and include variables which you think might be associated with the percentage of people voting to remain in the EU. As you fit your models, check to make sure that the associations are roughly linear.

Try fitting multiple models (at least 3 or 4) and think about what makes sense to investigate and what variables might need transformations.

Questions

- Look at the R^2 value for each model. As you include more variables, what happens to the R^2 value? Does this always happen?
- When you include more variables, how do the regression coefficients change for the existing variables?

After you are done, discuss your findings with your neighbor and pat yourself on the back. Congratulations, you're on your way to being a statistician!

Questions

Questions to discuss with your neighbor.

- How did you decide which variables to include and which variables not to include?
- What is the proper interpretation of your regression coefficients?
- What are the signs of each of the coefficients?
- What are the relative sizes of the coefficients?
- Does this make sense with what we know about the world?
- What would we need to be careful about in interpreting these models?
- What other variables (that weren't available) would also be good to include?