MATHS 7107 Data Taming Tutorial Solutions

Chang Dong

2023-03-17

Multiple Regression

Load the data

First, let's make sure the tidyverse package is loaded and we'll look at the population data again this week.

Outline

This week we're putting together a lot of what we have done so far.

Our exercise this week will be selecting an appropriate model to predict annual population growth in the population data. To limit things a little, we will not consider interaction terms and the only predictors we will consider are:

1. med_age_all - the median age of all residents.
2. med_age_male - the median age of male residents.
3. med_age_female - the median age of female residents.
4. ed_index_2015 - the United Nations index of educational development, as at 2015.
5. continent - the continent the country is in.
6. inequality - the GINI coefficient measuring income inequality in the country.
7. per_urban - the percentage of the population living in urban centres.

Using these predictors (. . . or not), we want to find the best model to predict annual population growth. We can do this in a step-by-step process by choosing which predictors are significant. You can do this in one of two ways:

1. Start with an empty model, and try all possible predictors alone. Add the predictor with the smallest p-value. Using this model, try all possible remaining predictors, adding the predictor with the smallest p-value. Continue until there are no remaining significant predictors to possibly add.
2. Start with a 'full' model, with all possible predictors. Remove the predictor with the highest p-value.

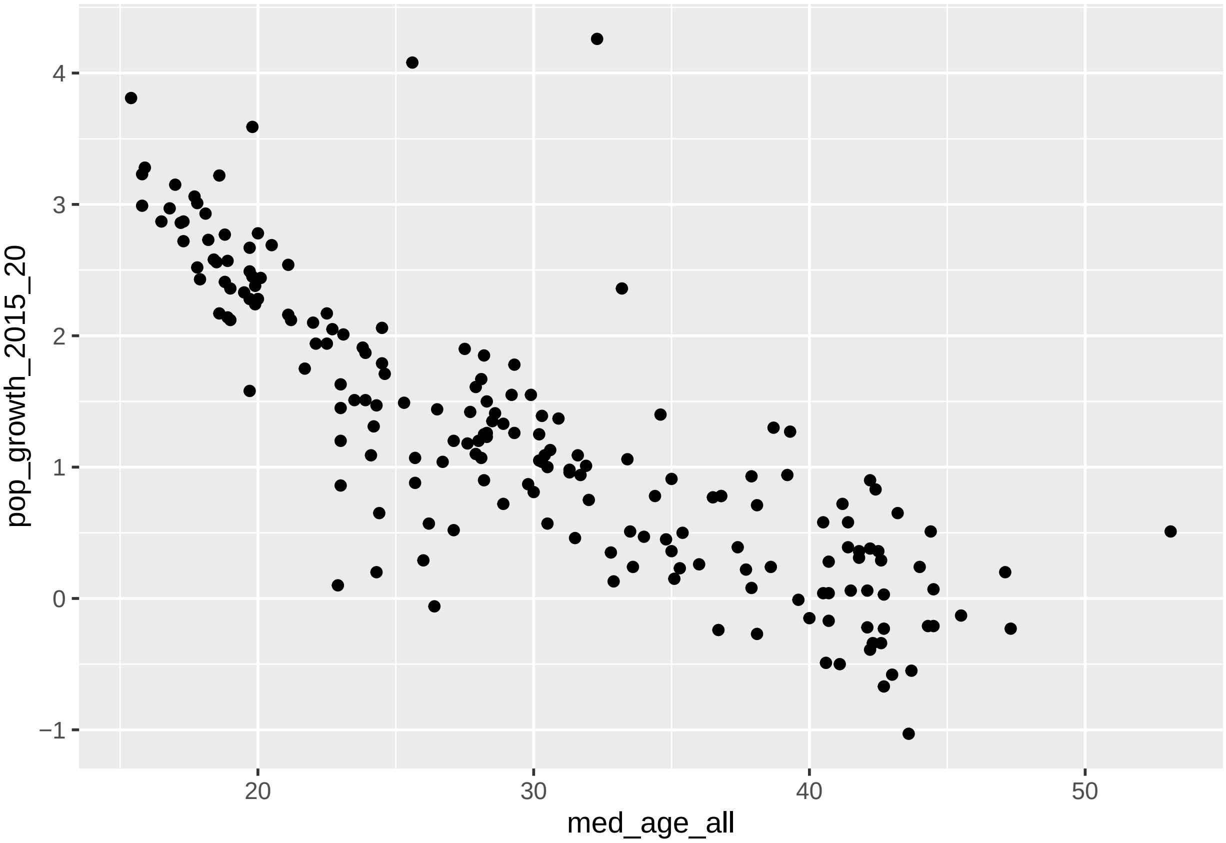
Fit the model without this predictor. Remove the predictor with the highest p-value. Continue until all predictors left in the model are significant. Today we will use method number 2.

Visualise the data

Produce the appropriate plots to compare each predictor variable with the response variable. Consider the relationship in each plot.

```
pacman::p_load("tidyverse")
population <- read.csv("population.csv")
ggplot(population, aes(y = pop_growth_2015_20, x = med_age_all)) + geom_point()
```

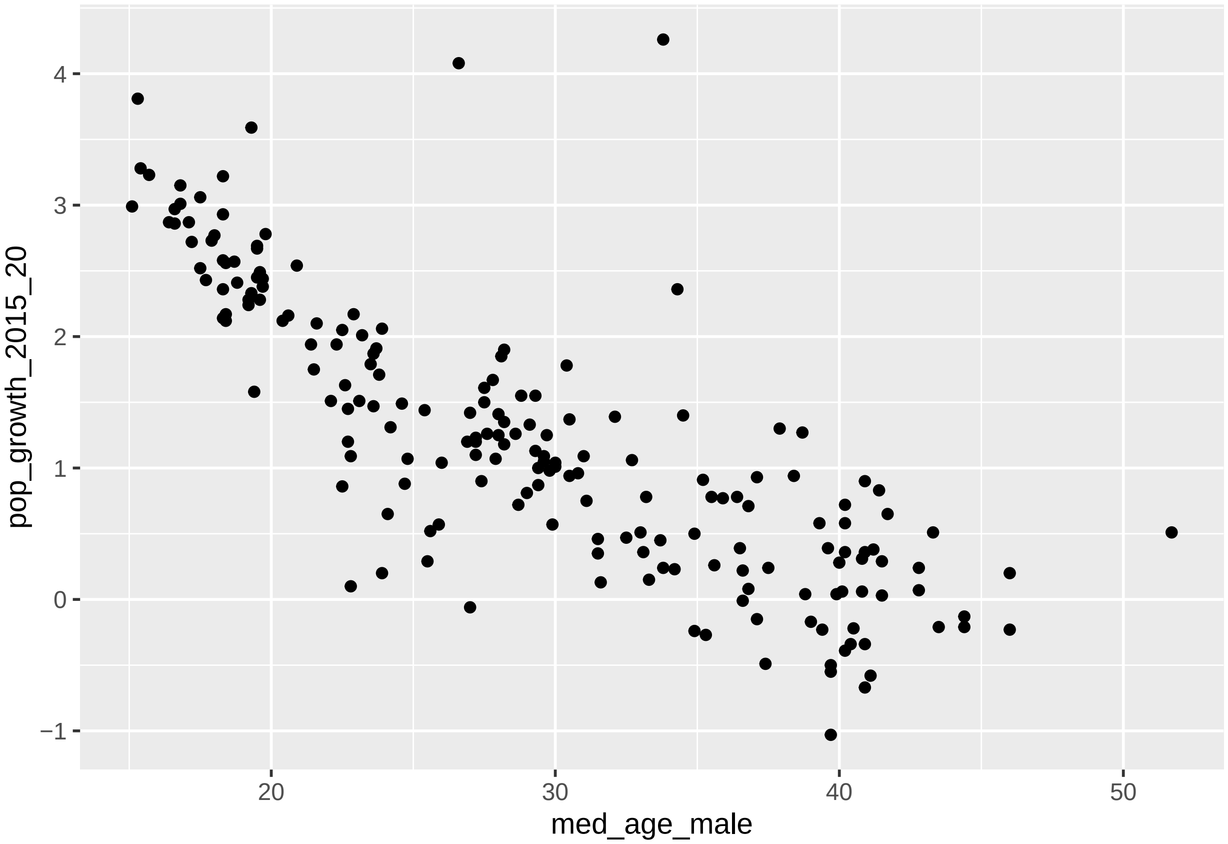
```
## Warning: Removed 78 rows containing missing values (`geom_point()`).
```



There appears to be a moderate, negative, linear relationship between the median age of all residents and population growth.

```
ggplot(population, aes(y = pop_growth_2015_20, x = med_age_male)) + geom_point()
```

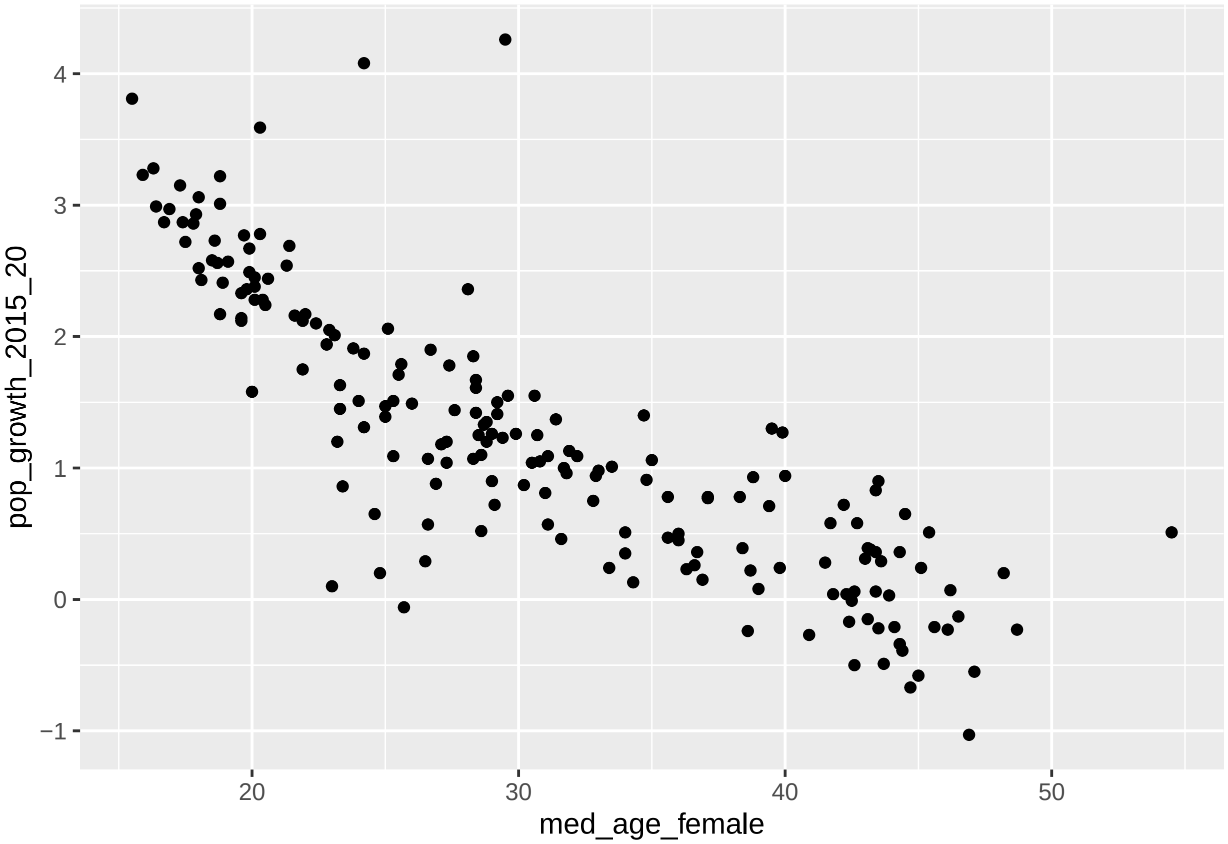
```
## Warning: Removed 78 rows containing missing values (`geom_point()`).
```



There appears to be a moderate, negative, linear relationship between the median age of male residents and population growth.

```
ggplot(population, aes(y = pop_growth_2015_20, x = med_age_female)) + geom_point()
```

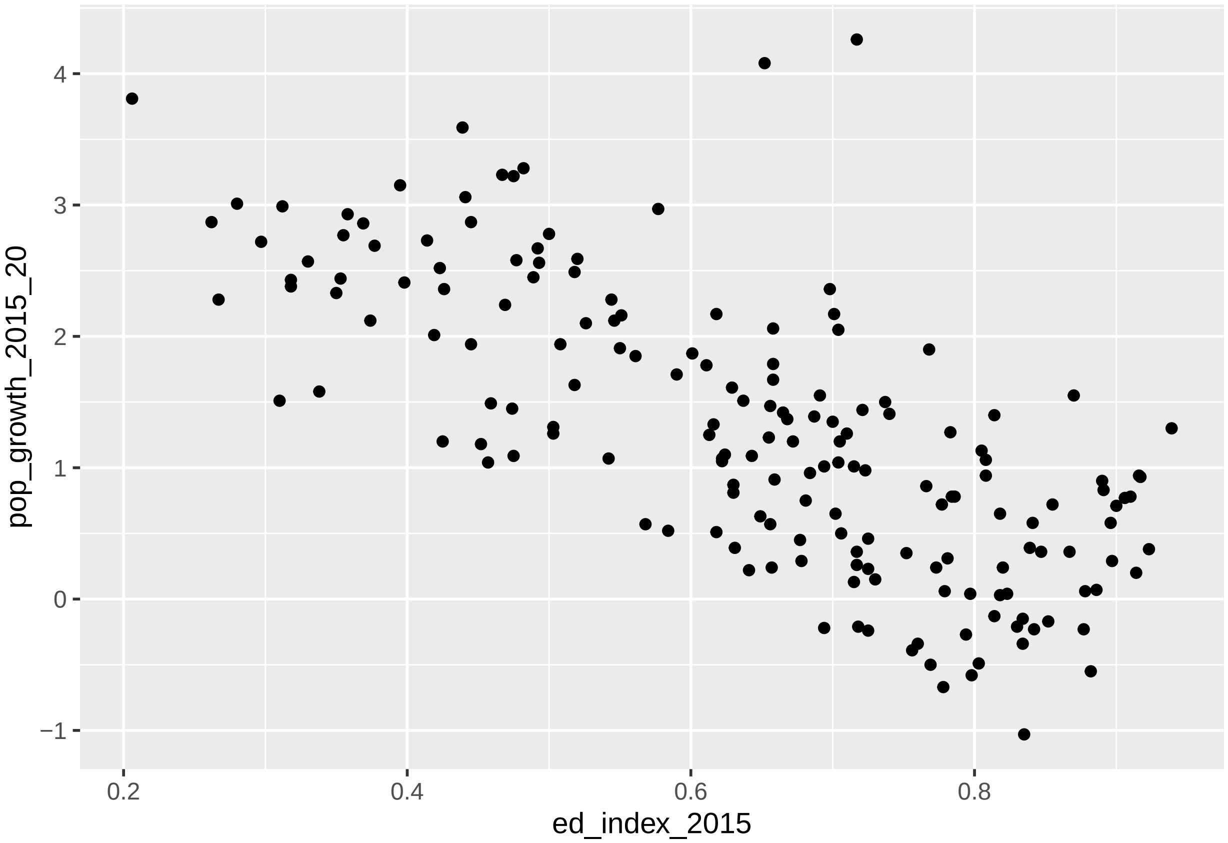
```
## Warning: Removed 78 rows containing missing values (`geom_point()`).
```



There appears to be a moderate, negative, linear relationship between the median age of female residents and population growth.

```
ggplot(population, aes(y = pop_growth_2015_20, x = ed_index_2015)) + geom_point()
```

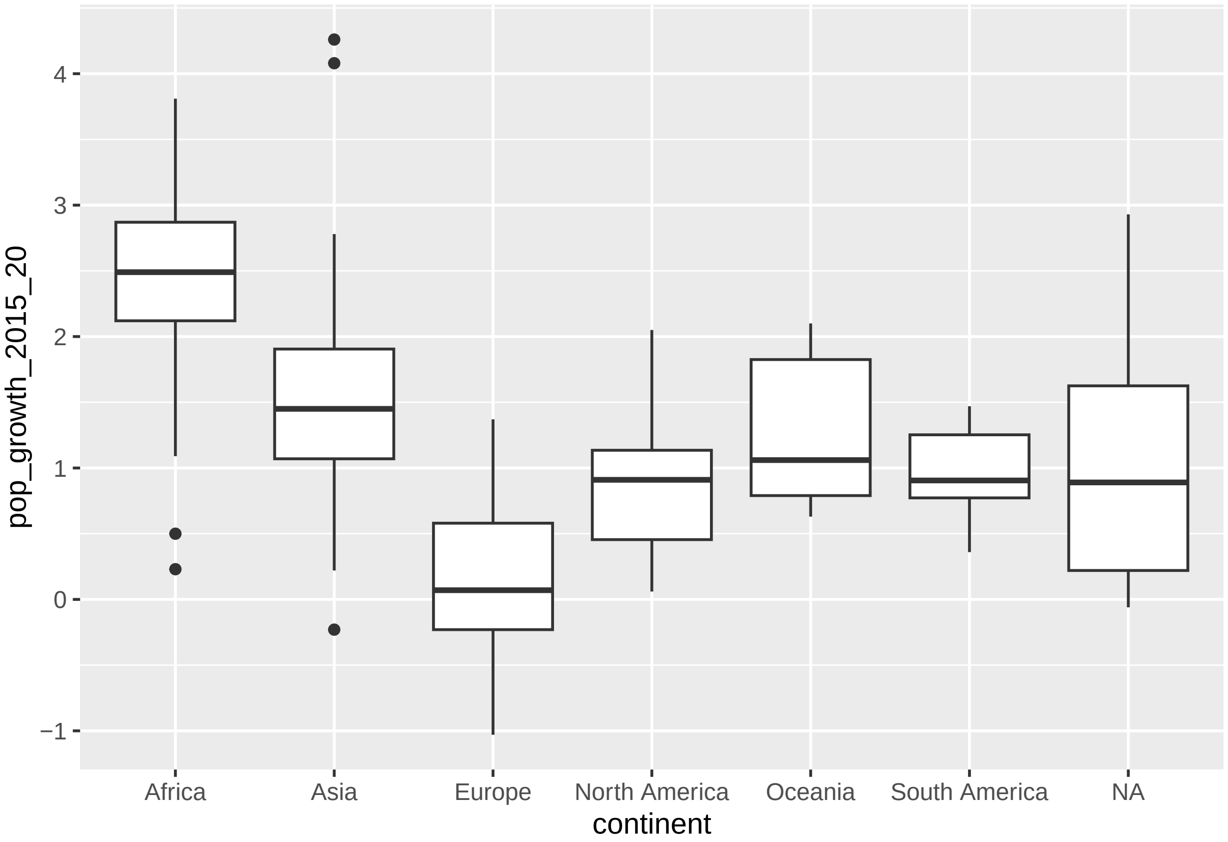
```
## Warning: Removed 92 rows containing missing values (`geom_point()`).
```



There appears to be a moderate, negative, linear relationship between the United Nations index of educational development (as at 2015) and population growth.

```
ggplot(population, aes(y = pop_growth_2015_20, x = continent)) + geom_boxplot()
```

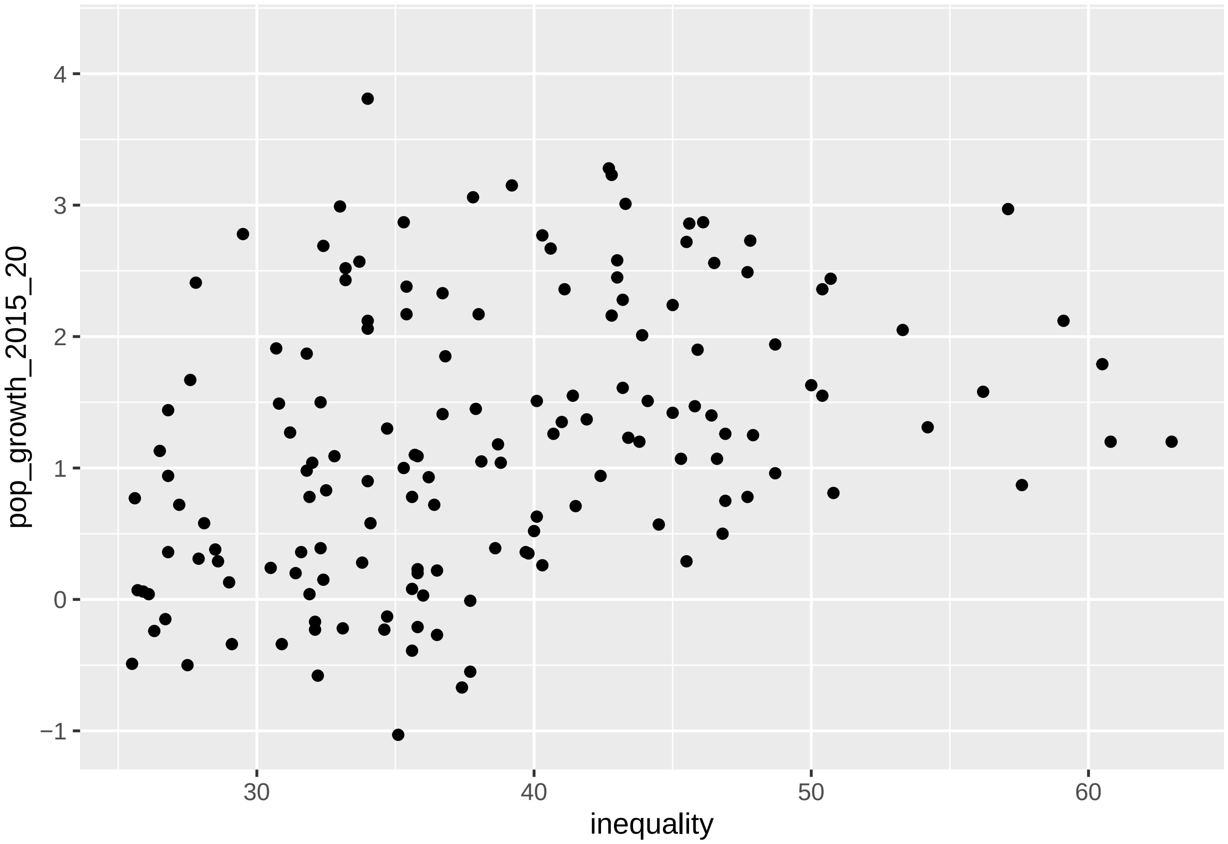
```
## Warning: Removed 70 rows containing non-finite values (`stat_boxplot()`).
```



Africa has the highest median population growth whilst Europe has the lowest. Oceania has the largest variability (IQR) in population growth whilst South America has the smallest. There are outliers in both Africa and Asia.

```
ggplot(population, aes(y = pop_growth_2015_20, x = inequality)) + geom_point()
```

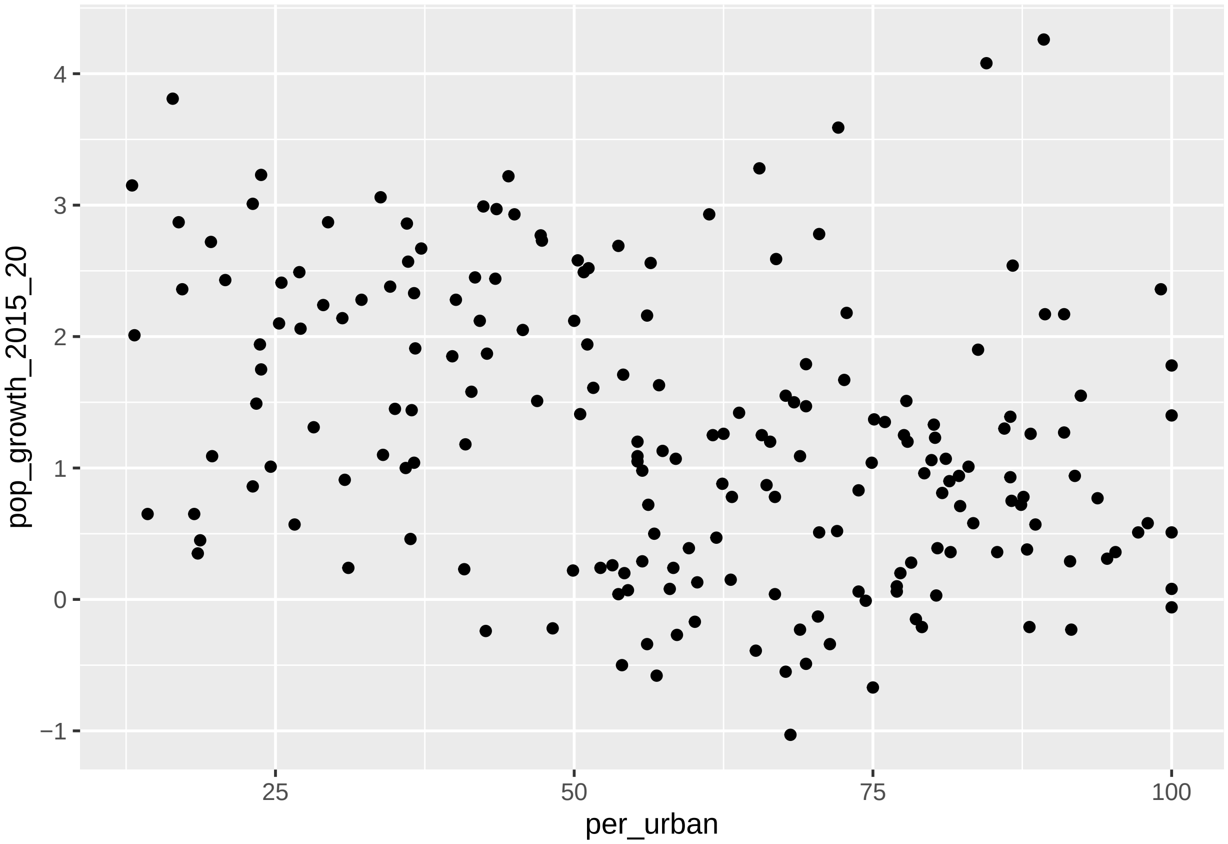
```
## Warning: Removed 115 rows containing missing values (`geom_point()`).
```



There appears to be no relationship between the GINI coefficient measuring incomme inequality in the country and population growth.

```
ggplot(population, aes(y = pop_growth_2015_20, x = per_urban)) + geom_point()
```

```
## Warning: Removed 72 rows containing missing values (`geom_point()`).
```



There appears to be no relationship between the percentage of the population living in urban centres and population growth.

'Backwards' model selection

First, let's start by building the full linear model.

Then you will need to look at the p-values. Remember the p-value for the categorical variable continent will be given by an ANOVA (using the anova() function), while the other p- values will be given by the model's summary (using the summary() function):

```
lm_pop <- lm(pop_growth_2015_20 ~ med_age_all +
               med_age_male +
               med_age_female +
               ed_index_2015 +
               inequality +
               per_urban +
               continent ,
               data = population)
summary(lm_pop)
```

```
##
## Call:
## lm(formula = pop_growth_2015_20 ~ med_age_all + med_age_male +
##     med_age_female + ed_index_2015 + inequality + per_urban +
##     continent, data = population)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max 
## -1.10263 -0.21505 -0.00065  0.25987  0.99146 
## 
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)    
## (Intercept)             4.690524   0.262505  17.868  < 2e-16 ***
## med_age_all            -0.025372   0.191213  -0.133 0.894638    
## med_age_male            0.108867   0.109219   0.997 0.320671    
## med_age_female         -0.183018   0.090181  -2.029 0.044392 *  
## ed_index_2015           0.355051   0.399974   0.888 0.376300    
## inequality             -0.014109   0.005076  -2.780 0.006225 ** 
## per_urban               0.005880   0.002096   2.805 0.005788 ** 
## continentAsia          -0.372219   0.108024  -3.446 0.000761 ***
## continentEurope        -0.309949   0.155930  -1.988 0.048878 *  
## continentNorth America -0.490619   0.125728  -3.902 0.000150 ***
## continentOceania       -0.210812   0.211572  -0.996 0.320849    
## continentSouth America -0.548619   0.136542  -4.018 9.74e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.3632 on 134 degrees of freedom
##   (121 observations deleted due to missingness)
## Multiple R-squared:  0.889,  Adjusted R-squared:  0.8798 
## F-statistic: 97.53 on 11 and 134 DF,  p-value: < 2.2e-16
```

```
anova(lm_pop)
```

```
## Analysis of Variance Table
## 
## Response: pop_growth_2015_20
##                 Df  Sum Sq Mean Sq  F value    Pr(>F)    
## med_age_all      1 129.723 129.723 983.5398 < 2.2e-16 ***
## med_age_male     1   4.723   4.723  35.8076 1.876e-08 ***
## med_age_female   1   0.710   0.710   5.3861 0.0218080 *  
## ed_index_2015    1   0.105   0.105   0.7946 0.3742982    
## inequality       1   1.808   1.808  13.7055 0.0003114 ***
## per_urban        1   0.782   0.782   5.9258 0.0162370 *  
## continent        5   3.646   0.729   5.5294 0.0001160 ***
## Residuals      134  17.674   0.132                       
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Consider the p-values. Can any of the predictor variables be removed? If so, which one? Remove it.

Repeat this process until all variables in your model are significant.

So our p-value for continent is 0.000116, while our p-values for the quantitative variables are:

- 0.895 for med_age_all,
- 0.321 for med_age_male,
- 0.044 for med_age_female,
- 0.376 for ed_index_2015,
- 0.006 for inequality, and

9 • 0.006 for per_urban.

So med_age_all is not significant, after the other variables are considered. That makes sense, since the overall median age is going to be pretty covered by the median age of men and of women - we probably don't need all three.

Let's have a look at a model without it:

```
lm_pop <- lm(pop_growth_2015_20 ~ med_age_male +
               med_age_female +
               ed_index_2015 +
               inequality +
               per_urban +
               continent ,
               data = population)
summary(lm_pop)
```

```
##
## Call:
## lm(formula = pop_growth_2015_20 ~ med_age_male + med_age_female +
##     ed_index_2015 + inequality + per_urban + continent, data = population)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.10324 -0.21501  0.00005  0.26037  0.99152
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               4.693304   0.260714  18.002  < 2e-16 ***
## med_age_male              0.094834   0.027181   3.489 0.000655 ***
## med_age_female           -0.194530   0.024525  -7.932 7.30e-13 ***
## ed_index_2015             0.360926   0.396066   0.911 0.363774
## inequality               -0.014136   0.005053  -2.797 0.005905 **
## per_urban                 0.005835   0.002062   2.830 0.005369 **
## continentAsia            -0.373034   0.107456  -3.472 0.000696 ***
## continentEurope          -0.310380   0.155328  -1.998 0.047702 *
## continentNorth America -0.489838   0.125132  -3.915 0.000143 ***
## continentOceania         -0.209317   0.210502  -0.994 0.321822
## continentSouth America -0.547893   0.135935  -4.031 9.25e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3618 on 135 degrees of freedom
##   (121 observations deleted due to missingness)
## Multiple R-squared:  0.8889, Adjusted R-squared:  0.8807
## F-statistic: 108.1 on 10 and 135 DF,  p-value: < 2.2e-16
```

```
anova(lm_pop)
```

```
## Analysis of Variance Table
##
## Response: pop_growth_2015_20
##                 Df  Sum Sq Mean Sq  F value    Pr(>F)
## med_age_male     1 125.655 125.655 959.6820 < 2.2e-16 ***
## med_age_female   1   9.496   9.496  72.5268 2.842e-14 ***
## ed_index_2015    1   0.102   0.102   0.7775 0.3794806
## inequality       1   1.803   1.803  13.7723 0.0003007 ***
## per_urban        1   0.793   0.793   6.0586 0.0151000 *
## continent        5   3.645   0.729   5.5671 0.0001075 ***
## Residuals      135  17.676   0.131
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here, ed_index_2015 has the highest p-value, of 0.364, which is not significant. That's really interesting, one might've though education has an impact on population growth, but it looks here that it doesn't, once we account for median ages, levels of economic inequality, urbanisation and the continent.

Let's remove it.

```
lm_pop <- lm(pop_growth_2015_20 ~ med_age_male +
               med_age_female +
               inequality +
               per_urban +
               continent ,
               data = population)
summary(lm_pop)
```

```
##
## Call:
## lm(formula = pop_growth_2015_20 ~ med_age_male + med_age_female +
##     inequality + per_urban + continent, data = population)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.07316 -0.21813  0.00144  0.27410  0.98421
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              4.716501   0.259306  18.189  < 2e-16 ***
## med_age_male             0.097284   0.027030   3.599 0.000446 ***
## med_age_female          -0.192578   0.024416  -7.887 9.03e-13 ***
## inequality              -0.013727   0.005030  -2.729 0.007194 **
## per_urban                0.006631   0.001867   3.552 0.000526 ***
## continentAsia           -0.351359   0.104725  -3.355 0.001029 **
## continentEurope         -0.283793   0.152469  -1.861 0.064857 .
## continentNorth America -0.475546   0.124068  -3.833 0.000193 ***
## continentOceania       -0.162122   0.203904  -0.795 0.427948
## continentSouth America -0.534911   0.135102  -3.959 0.000121 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3616 on 136 degrees of freedom
##   (121 observations deleted due to missingness)
## Multiple R-squared:  0.8883, Adjusted R-squared:  0.8809
## F-statistic: 120.1 on 9 and 136 DF,  p-value: < 2.2e-16
```

```
anova(lm_pop)
```

```
## Analysis of Variance Table
##
## Response: pop_growth_2015_20
##                 Df  Sum Sq Mean Sq  F value    Pr(>F)
## med_age_male     1 125.655 125.655 960.8800 < 2.2e-16 ***
## med_age_female   1   9.496   9.496  72.6173 2.646e-14 ***
## inequality       1   1.687   1.687  12.8995 0.0004581 ***
## per_urban        1   1.008   1.008   7.7104 0.0062653 **
## continent        5   3.539   0.708   5.4126 0.0001425 ***
## Residuals      136  17.785   0.131
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

How could we do this in R?

First create a tibble with our new data.

```
new_country = tibble( med_age_male = 38, med_age_female = 41, inequality = 28, per_urban = 90, continen
```

Since we're interested in a particular country, rather than the average country of this character, we should make a prediction interval. This is because:

- A confidence interval is for an average country of this character
- A point prediction doesn't consider how accurate our prediction is, and it's important to know how we might be wrong.

Using R, calculate the prediction interval based on the above information. Interpret this interval in context.

```
predict(lm_pop, new_country, interval = "prediction")
```

```
##        fit        lwr     upr
## 1 0.446256 -0.2820179 1.17453
```

So we are 95% confident that annual population growth between 2015 and 2020 will be between -0.282% and 1.175% for this country.

Are there any negative numbers in the interval? If yes, are they reasonable?

Often when we see negative numbers in projections we should ask ourselves if the value is reasonable. For example, it would be unreasonable if -0.282% of the population lived in urban centres. But in this case it's perfectly reasonable for population growth to be negative, if the population is decreasing.