Lesson 14: MLR Model Diagnostics

Nicky Wakim

2024-03-13

Learning Objectives

1. Apply tools from SLR (Lesson 6: SLR Diagnostics) in MLR to **evaluate LINE assumptions**, including residual plots and QQ-plots
2. Apply tools involving standardized residuals, leverage, and Cook's distance from SLR (Lesson 7: SLR Diagnostics 2) in MLR to **flag potentially influential points**
3. Use Variance Inflation Factor (VIF) and it's general form to **detect and correct multicollinearity**

Regression analysis process

Model Selection

- Building a model
- Selecting variables
- Prediction vs interpretation
- Comparing potential models

Model Fitting

- Find best fit line
- Using OLS in this class
- Parameter estimation
- Categorical covariates
- Interactions

Model Evaluation

- Evaluation of model fit
- Testing model assumptions
- Residuals
- Transformations
- Influential points
- Multicollinearity

Model Use (Inference)

- Inference for coefficients
- Hypothesis testing for coefficients
- Inference for expected Y given X
- Prediction of new Y given X

Let's remind ourselves of the final model

- Our **final model** contains
  - Female Literacy Rate `FLR`
  - CO2 Emissions in quartiles `CO2_q`
  - Income levels in groups assigned by Gapminder `income_levels1`
  - World regions `four_regions`
  - Membership of global and economic groups `members_oecd_g77`
  - Food Supply `FoodSupplykcPPD`
  - Clean Water Supply `WaterSupplePct`

▶ Display regression table for final model

term	estimate	std.error	statistic	p.value
(Intercept)	39.877	4.889	8.157	0.000
FemaleLiteracyRate	−0.073	0.047	−1.555	0.125
CO2_q(0.806,2.54]	1.099	1.914	0.574	0.568
CO2_q(2.54,4.66]	−0.292	2.419	−0.121	0.904
CO2_q(4.66,35.2]	−0.595	2.524	−0.236	0.814
income_levels1Lower middle income	5.441	2.343	2.322	0.024
income_levels1Upper middle income	6.111	2.954	2.069	0.043
income_levels1High income	7.959	3.277	2.429	0.018
four_regionsAmericas	9.003	2.050	4.391	0.000
four_regionsAsia	5.260	1.637	3.213	0.002
four_regionsEurope	6.855	2.871	2.387	0.020
WaterSourcePrct	0.166	0.066	2.496	0.015
FoodSupplykcPPD	0.004	0.002	1.825	0.073
members_oecd_g77oecd	1.119	2.674	0.418	0.677
members_oecd_g77others	1.047	2.511	0.417	0.678

It's a lot to visualize

- Part of the reason why we discussed model diagnostics in SLR was so that we could have accompanying visuals to help us understand

- With 7 variables in out final model, it is hard to visualize outliers and influential points

- I highly encourage you revisit Lesson 6 and 7 (SLR Diagnostics) to help understand these notes

Remember our friend `augment()`?

- Run `final_model` through `augment()` (`final_model` is input)
  - So we assigned `final_model` as the output of the `lm()` function
- Will give us values about each observation in the context of the fitted regression model
  - cook's distance (`.cooksd`), fitted value (`.fitted`, $\widehat{Y}_i$), leverage (`.hat`), residual (`.resid`), standardized residuals (`.std.resid`)

```
aug = augment(final_model)
head(aug) %>% relocate(.fitted, .resid, .std.resid, .hat, .cooksd, .after = LifeExp
```

```
# A tibble: 6 × 14
  LifeExpectancyYrs .fitted .resid .std.resid  .hat  .cooksd FemaleLiteracyRate
              <dbl>   <dbl>  <dbl>      <dbl> <dbl>    <dbl>              <dbl>
1              56.7    61.5 -4.78      -1.43  0.327 0.0663                 13  
2              76.7    75.3  1.38       0.387 0.227 0.00293                95.7
3              60.9    58.6  2.30       0.684 0.320 0.0147                 58.6
4              76.9    74.7  2.21       0.620 0.238 0.00799                99.4
5              76      76.9 -0.879     -0.233 0.145 0.000614               97.9
6              73.8    74.6 -0.796     -0.214 0.168 0.000618               99.5
# ℹ 7 more variables: CO2_q <fct>, income_levels1 <fct>, four_regions <fct>,
```

RDocumentation on the `augment()` function.

Learning Objectives

1. Apply tools from SLR (Lesson 6: SLR Diagnostics) in MLR to **evaluate LINE assumptions**, including residual plots and QQ-plots

2. Apply tools involving standardized residuals, leverage, and Cook's distance from SLR (Lesson 7: SLR Diagnostics 2) in MLR to **flag potentially influential points**

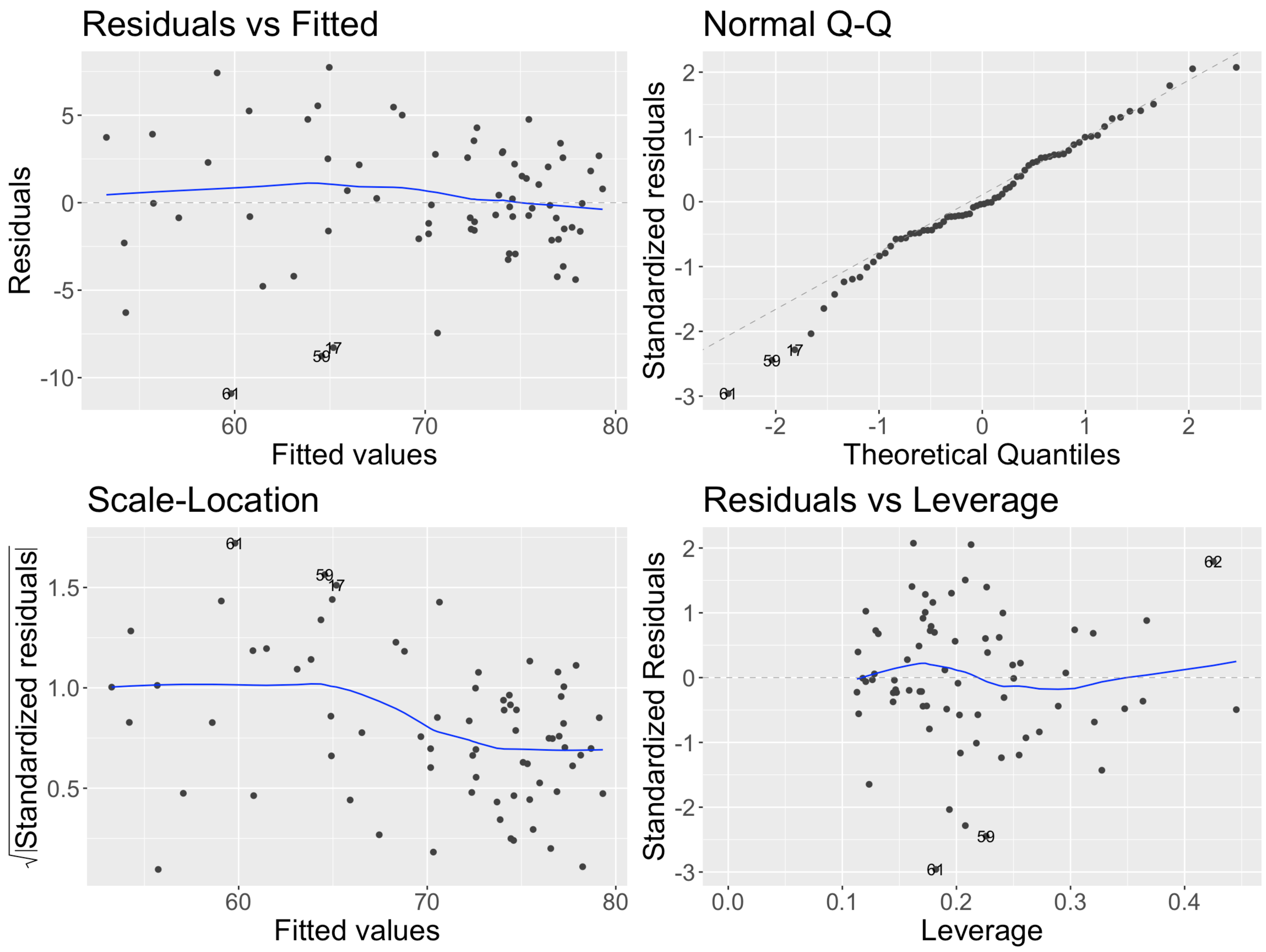
3. Use Variance Inflation Factor (VIF) and it's general form to **detect and correct multicollinearity**

Summary of the assumptions and their diagnostic tool

Assumption	What needs to hold?	Diagnostic tool
Linearity	• Relationship between **each** X and Y is linear	• Scatterplot of Y vs. X
Independence	• Observations are independent from each other	• Study design
Normality	• Residuals (and thus $Y\|X_1, X_2, \ldots, X_p$) are normally distributed	• QQ plot of residuals • Distribution of residuals
Equality of variance	• Variance of residuals (and thus $Y\|X_1, X_2, \ldots, X_p$) is same across fitted values (homoscedasticity)	• Residual plot

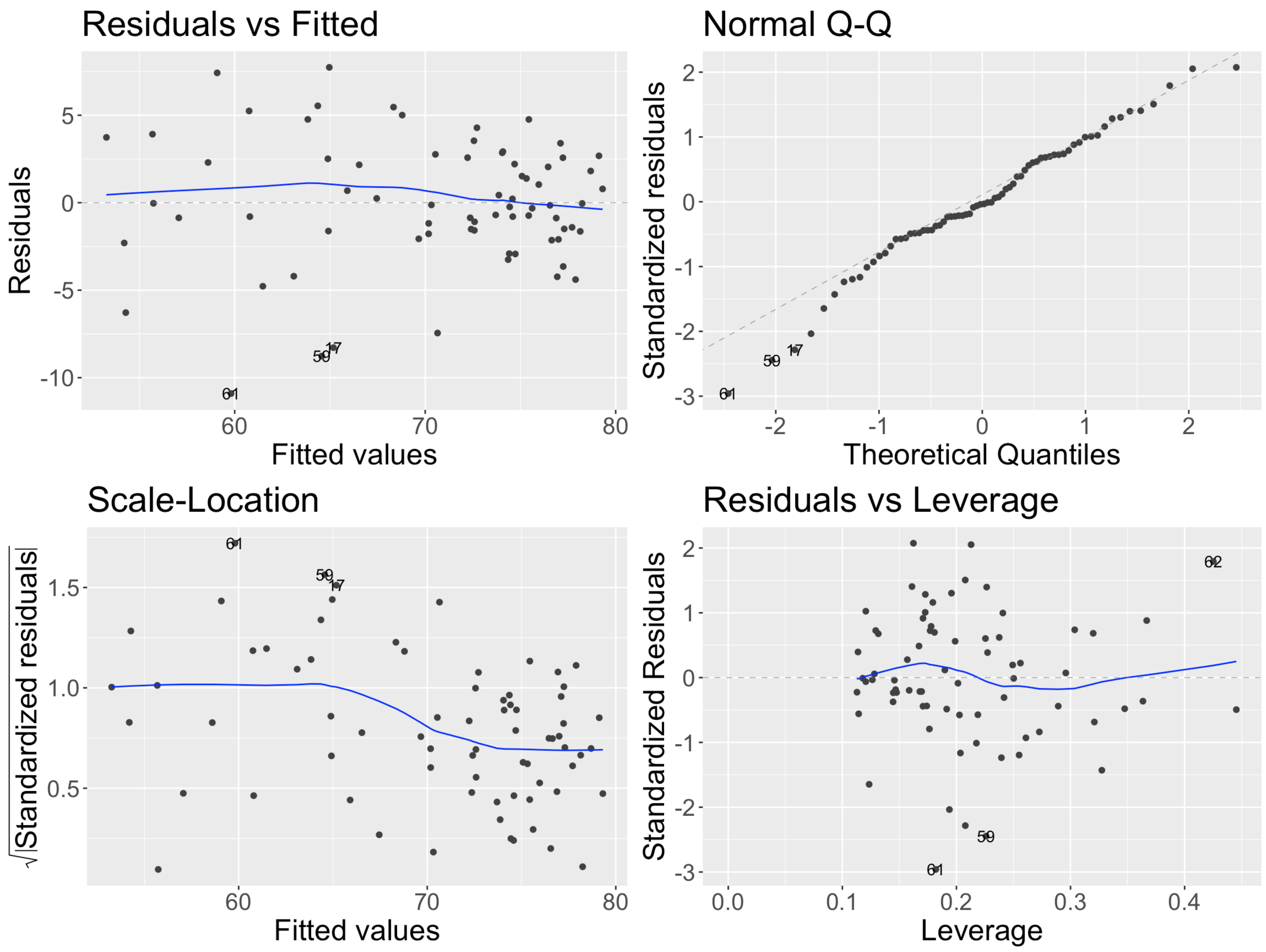
`autoplot()` to examine equality of variance and Normality

```
library(ggfortify)
autoplot(final_model) + theme(text=element_text(
```



`autoplot()` to examine equality of variance and Normality

```
library(ggfortify)
autoplot(final_model) + theme(text=element_text(
```



Looks like 3 obs are flagged:

- 17: Cote d'Ivoire
- 59: South Africa
- 61: Kingdom of Eswatini (formerly Swaziland in 2011)

Without them, QQ-plot and residual plot look good

- Points on QQ-plot are close to identity line
- Residuals have pretty consistent spread across fitted values

But don't take them out!!!

- Instead, discuss what may be missing in our regression model that is not capturing the characteristics of these countries

Poll Everywhere Question 1

Learning Objectives

1. Apply tools from SLR (Lesson 6: SLR Diagnostics) in MLR to **evaluate LINE assumptions**, including residual plots and QQ-plots

2. Apply tools involving standardized residuals, leverage, and Cook's distance from SLR (Lesson 7: SLR Diagnostics 2) in MLR to **flag potentially influential points**

3. Use Variance Inflation Factor (VIF) and it's general form to **detect and correct multicollinearity**

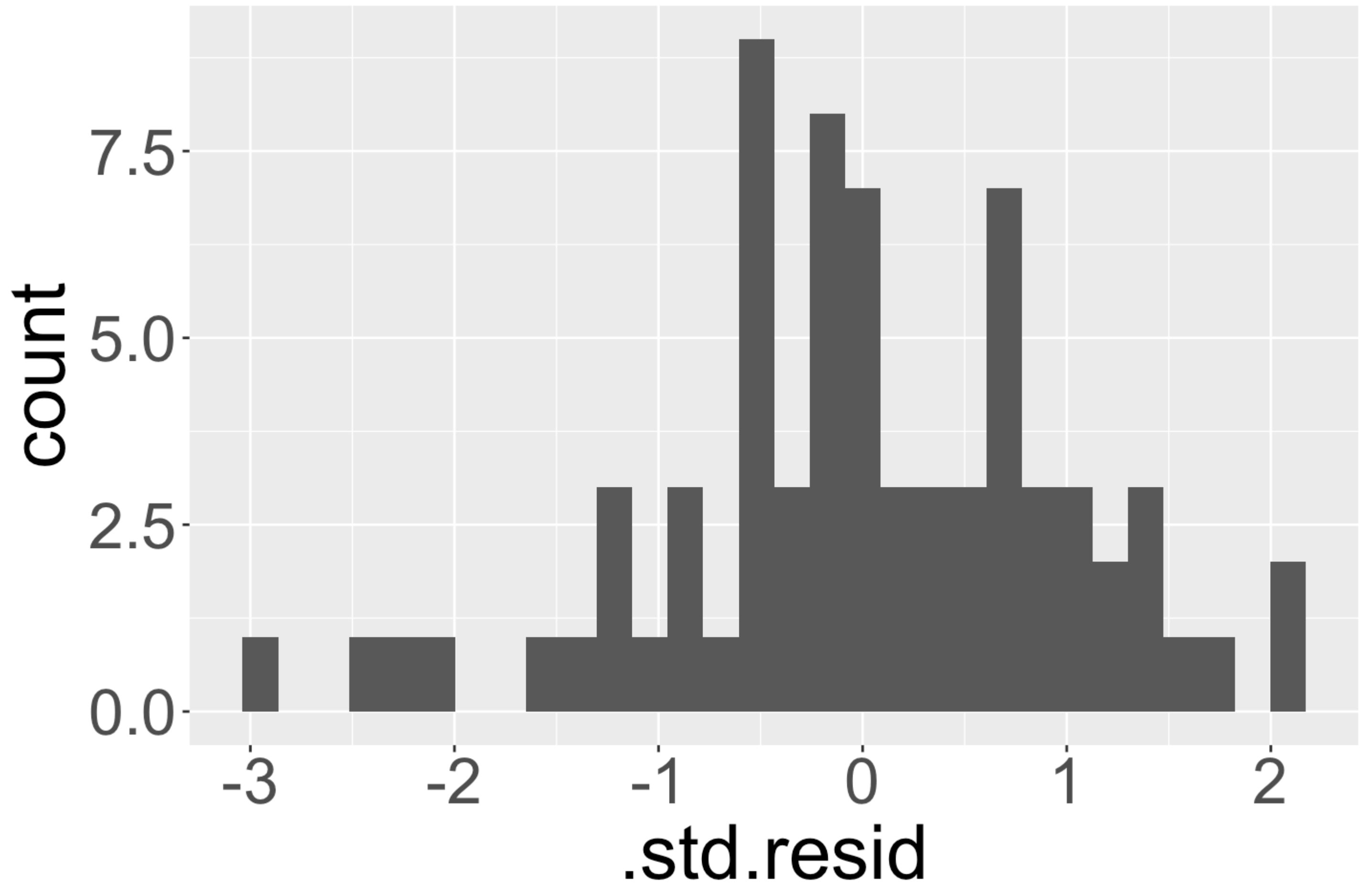
Identifying outliers

Internally standardized residual

$$r_i = \frac{\hat{\epsilon}_i}{\sqrt{\widehat{\sigma}^2(1 - h_{ii})}}$$

- We flag an observation if the standardized residual is "large"
  - Different sources will define "large" differently
  - PennState site uses $|r_i| > 3$
  - `autoplot()` shows the 3 observations with the highest standardized residuals
  - Other sources use $|r_i| > 2$, which is a little more conservative

```
ggplot(data = aug) +
  geom_histogram(aes(x = .std.resid))
```



Countries that are outliers ($|r_i| > 2$)

- We can identify the countries that are outliers

```
aug %>% relocate(.std.resid, .after = country) %>%
  filter(abs(.std.resid) > 2) %>% arrange(desc(abs(.std.resid)))
```

```
# A tibble: 6 × 15
  country   .std.resid LifeExpectancyYrs FemaleLiteracyRate CO2_q income_levels1
  <chr>          <dbl>             <dbl>              <dbl> <fct> <fct>
1 Swaziland      -2.96              48.9               87.3 (0.8… Lower middle …
2 South Af…      -2.45              55.8               92.2 (4.6… Upper middle …
3 Cote d'I…      -2.28              56.9               47.6 [0.0… Lower middle …
4 Cape Ver…       2.07              72.7               80.3 (0.8… Lower middle …
5 Sudan           2.05              66.5               63.2 [0.0… Lower middle …
6 Vanuatu        -2.04              63.2               81.5 [0.0… Lower middle …
# ℹ 9 more variables: four_regions <fct>, WaterSourcePrct <dbl>,
```

Leverage h_i

- Values of leverage are: $0 \leq h_i \leq 1$
- We flag an observation if the leverage is "high"
  - **Only good for SLR:** Some textbooks use $h_i > 4/n$ where n = sample size
  - **Only good for SLR:** Some people suggest $h_i > 6/n$
  - **Works for MLR:** $h_i > 3p/n$ where p = number of regression coefficients

```r
aug = aug %>% relocate(.hat, .after = FemaleLiteracyRate)
aug %>% arrange(desc(.hat))
```

```
# A tibble: 72 × 15
   country       LifeExpectancyYrs FemaleLiteracyRate  .hat CO2_q income_levels1
   <chr>                     <dbl>              <dbl> <dbl> <fct> <fct>
 1 Mexico                     75.8               92.3 0.445 (2.5… Upper middle …
 2 Tajikistan                 69.9               99.6 0.425 [0.0… Lower middle …
 3 Bosnia and H…              76.9               96.7 0.367 (4.6… Upper middle …
 4 Uzbekistan                 69                 99.2 0.363 (2.5… Lower middle …
 5 Bangladesh                 71                 53.4 0.347 [0.0… Lower middle …
 6 Afghanistan                56.7               13   0.327 [0.0… Low income
 7 Zimbabwe                   51.9               80.1 0.321 [0.0… Low income
```

Countries with high leverage ($h_i > 3p/n$)

- We can look at the countries that have high leverage: there are NONE

```r
n = nrow(gapm2); p = length(final_model$coefficients) - 1
aug %>%
  filter(.hat > 3*p/n) %>%
  arrange(desc(.hat))
```

```
# A tibble: 0 × 15
# ℹ 15 variables: country <chr>, LifeExpectancyYrs <dbl>,
#   FemaleLiteracyRate <dbl>, .hat <dbl>, CO2_q <fct>, income_levels1 <fct>,
#   four_regions <fct>, WaterSourcePrct <dbl>, FoodSupplykcPPD <dbl>,
#   members_oecd_g77 <chr>, .fitted <dbl>, .resid <dbl>, .sigma <dbl>,
#   .cooksd <dbl>, .std.resid <dbl>
```

Identifying points with high Cook's distance

The Cook's distance for the i^{th} observation is

$$d_i = \frac{h_i}{2(1 - h_i)} \cdot r_i^2$$

where h_i is the leverage and r_i is the studentized residual

- No countries with high Cook's distance

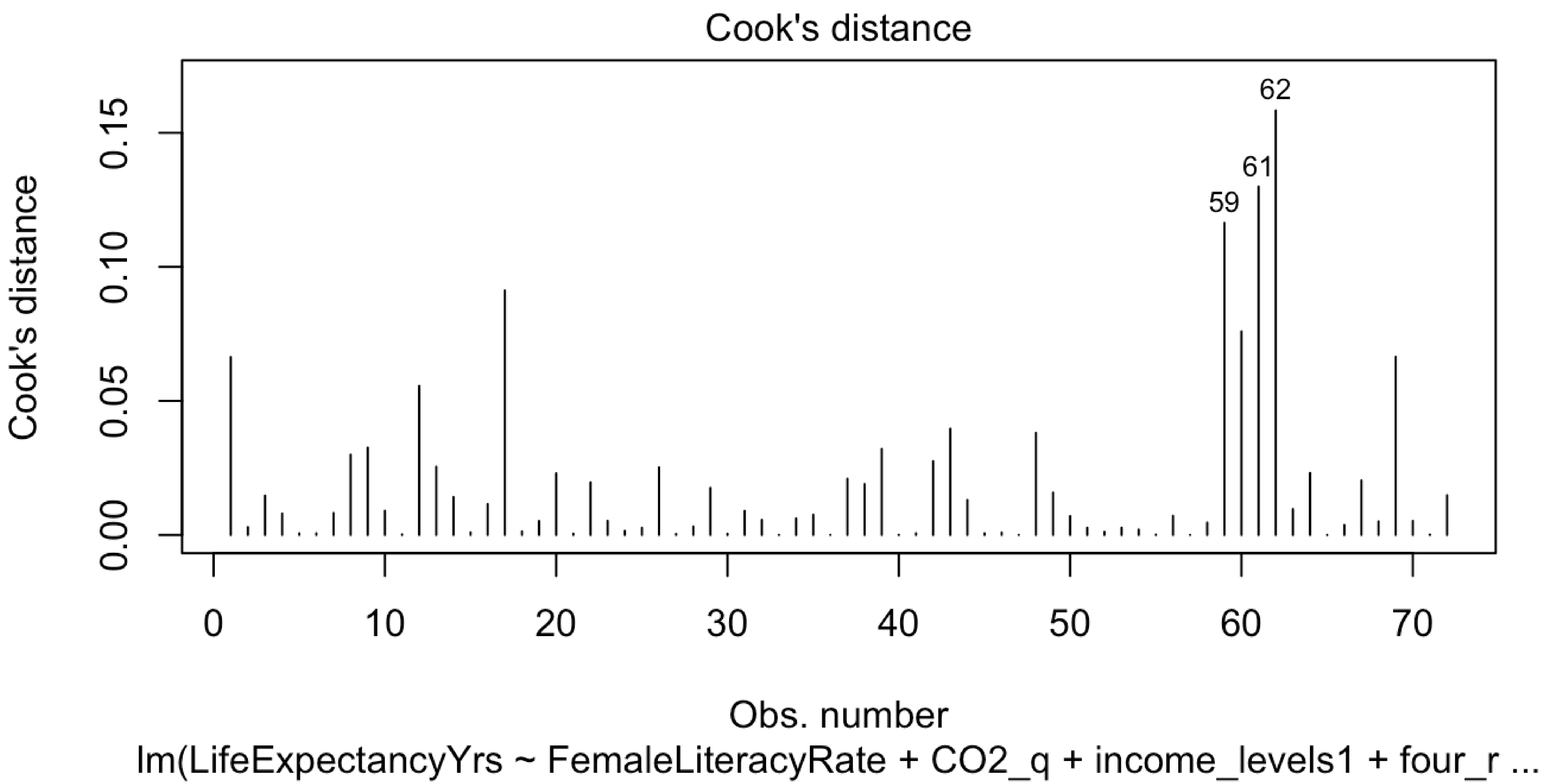
- Another rule for Cook's distance that is not strict:
  - Investigate observations that have $d_i > 1$
- Cook's distance values are already in the augment tibble: `.cooksd`

```r
aug = aug %>% relocate(.cooksd, .after = country)
aug %>% arrange(desc(.cooksd)) %>% filter(.cooksd > 1)
```

```
# A tibble: 0 × 15
# ℹ 15 variables: country <chr>, .cooksd <dbl>, LifeExpectancyYrs <dbl>,
#   FemaleLiteracyRate <dbl>, .hat <dbl>, CO2_q <fct>, income_levels1 <fct>,
#   four_regions <fct>, WaterSourcePrct <dbl>, FoodSupplykcPPD <dbl>,
#   members_oecd_g77 <chr>, .fitted <dbl>, .resid <dbl>, .sigma <dbl>,
#   .std.resid <dbl>
```

Plotting Cook's Distance

```
# plot(model) shows figures similar to autoplot()
# adds on Cook's distance though
plot(final_model, which = 4)
```



How do we deal with influential points?

- If an observation is influential, we can **check data errors**:
  - Was there a data entry or collection problem?
  - If you have reason to believe that the observation does not hold within the population (or gives you cause to redefine your population)
- If an observation is influential, we can **check our model**:
  - Did you leave out any important predictors?
  - Should you consider adding some interaction terms?
  - Is there any nonlinearity that needs to be modeled?
- Basically, deleting an observation should be justified outside of the numbers!
  - If it's an honest data point, then it's giving us important information!
- **Means we will need to discuss the limitations of our model**
  - For example: Think about measurements that might help explain life expectancy that are NOT in our model
- A really well thought out explanation from StackExchange

Poll Everywhere Question 2

When we have detected problems in our model…

- We have talked about influential points
- We have talked about identifying issues with our LINE assumptions

What are our options once we have identified issues in our linear regression model?

- Are we missing a crucial measure in our dataset?
- Try a transformation if there is an issue with linearity or normality
  - Addressed in model selection
- Try a weighted least squares approach if unequal variance (oof, not enough time for us to get to)
- Try a robust estimation procedure if we have a lot of outlier issues (outside scope of class)

Learning Objectives

1. Apply tools from SLR (Lesson 6: SLR Diagnostics) in MLR to **evaluate LINE assumptions**, including residual plots and QQ-plots
2. Apply tools involving standardized residuals, leverage, and Cook's distance from SLR (Lesson 7: SLR Diagnostics 2) in MLR to **flag potentially influential points**
3. Use Variance Inflation Factor (VIF) and it's general form to **detect and correct multicollinearity**

What is multicollinearity? (adapted from parts of STAT 501 page)

So far, we've been ignoring something very important: multicollinearity

Multicollinearity

Two or more covariates in a multivariable regression model are *highly* correlated

- Types of multicollinearity
  - **Structural multicollinearity**
    - Mathematical artifact caused by creating new covariates from other covariates
    - For example: If we have age, and decide to transform age to include age-squared
      - Then we have age and age-squared in the model: age-squared is perfectly predicted by age!
  - **Data-based multicollinearity**
    - Result of a poorly designed experiment, reliance on purely observational data, or the inability to manipulate the system on which the data are collected.

Poll Everywhere Question 3

Why is multicollinearity a problem?

In linear regression…

- Estimated regression coefficient of any one variable **depends on other predictors included in the model**
  - Not necessarily bad, but a big change might be an issue
- Hypothesis tests for any coefficient may yield different conclusions **depending on other predictors included in the model**
- Marginal contribution of any one predictor variable in reducing the error sum of squares **depends on other predictors included in the model**

When there is multicollinearity in our model:

- **Precision** of the estimated regression coefficients or correlated covariates **decreases a lot**
  - Basically, **standard error increases and confidence intervals get wider**, which means we're not as confident in our estimate anymore
  - Because highly correlated covariates are not adding much more information, but are constraining our model more

Did you notice anything about all the consequences of multicollinearity?

- All consequences relate to estimating a regression coefficient **precisely**
  - Recall that precision is linked to analysis **goals of association and interpretability**
  - See Lesson 12: Model Selection

- Multicollinearity is *not really an issue* when our **goal is prediction**
  - Highly correlated covariates/predictors will not hurt our prediction of an outcome

How do we detect multicollinearity?

- **Variance inflation factors (VIF):** quantifies how much the variance of the estimated coefficient for covariate k increases
  - Increases: from SLR with only covariate k to MLR with all other covariates

- General rule of thumb
  - $4 < VIF < 10$: Warrent investigation (but most people aren't investigating this…)
  - $VIF > 10$: Requires correction
    - Influencing regression coefficient estimates

VIF

$$VIF = \frac{1}{1 - R_k^2}$$

R_k^2 is the R^2-value obtained by regressing the k^{th} covariate/predictor on the remaining predictors

Let's apply it to our final model

- Naive way to calculate this:

```
library(rms)
rms::vif(final_model)
```

```
               FemaleLiteracyRate                CO2_q(0.806,2.54]
                         4.863139                         2.979224
                CO2_q(2.54,4.66]                 CO2_q(4.66,35.2]
                         4.758904                         5.180216
income_levels1Lower middle income income_levels1Upper middle income
                         5.290718                         8.406927
        income_levels1High income             four_regionsAmericas
                         7.293148                         2.531966
                 four_regionsAsia               four_regionsEurope
                         2.096398                         7.771994
```

- All $VIF < 10$
- Problem: multi-level covariates (CO2 Emissions and income level) have different VIF's even though they should be considered one variable

Let's apply it to our final model *correctly* (1/2)

- Calculate the GVIF and, more importantly, the $GVIF^{1/(2\cdot df)}$
- GVIF is the R^2-value for regressing a covariate's group indicators on the remaining covariates
  - Captures the correlation between covariates better
- $GVIF^{1/(2\cdot df)}$ helps standardize GVIF based on how many levels each categorical covariate has
  - I'll refer to this as df-corrected GVIF or standardized GVIF
  - If continuous covariate, $GVIF^{1/(2\cdot df)} = \sqrt{GVIF}$

```r
library(car)
car::vif(final_model)
```

```
                         GVIF Df GVIF^(1/(2*Df))
FemaleLiteracyRate   4.863139  1        2.205253
CO2_q                8.223951  3        1.420736
income_levels1      11.045885  3        1.492336
four_regions        13.935918  3        1.551277
WaterSourcePrct      4.824266  1        2.196421
FoodSupplykcPPD      3.499250  1        1.870628
members_oecd_g77     7.430919  2        1.651052
```

Let's apply it to our final model *correctly* (2/2)

- If continuous covariate, $GVIF^{1/(2 \cdot df)} = \sqrt{GVIF}$
- So we can square $GVIF^{1/(2 \cdot df)}$ and set VIF rules
- OR: we can correct any $GVIF^{1/(2 \cdot df)} > \sqrt{10} = 3.162$

```r
car::vif(final_model)
```

```
                       GVIF Df GVIF^(1/(2*Df))
FemaleLiteracyRate  4.863139  1        2.205253
CO2_q               8.223951  3        1.420736
income_levels1     11.045885  3        1.492336
four_regions       13.935918  3        1.551277
WaterSourcePrct     4.824266  1        2.196421
FoodSupplykcPPD     3.499250  1        1.870628
members_oecd_g77    7.430919  2        1.651052
```

- All of these covariates are okay! No multicollinearity to correct in this dataset!

But what if we do need to make corrections for multicollinearity?

- We have been dealing with **data-based multicollinearity** in our example
- If we had issues with multicollinearity, then what are our options?
  - Remove the variable(s) with large VIF
  - Use expert knowledge in the field to decide
- If one variable has a large VIF, then there is usually another one or more variables with large VIFs
  - Basically, all the covariates that are correlated will have large VIFs
- Example: our two largest GVIFs were for world region and income levels
  - Hypothetical: their $GVIF^{1/(2 \cdot df)} > 3.162$
  - Remove one of them
  - I'm no expert, but from more of a data equity lens, there's a lot of generalizations made about world regions
    - I think relying on the income level of a country might give us more information as well

What about structural multicollinearity?

- **Structural multicollinearity**
  - Mathematical artifact caused by creating new covariates from other covariates

- For example: If we have age, and decide to transform age to include age-squared
  - Then we have age and age-squared in the model: age-squared is perfectly predicted by age!
  - By having the untransformed and transformed covariate in the model, they are inherently correlated!

- **Best practice to reduce the correlation: center you covariate**
  - By centering age, we no longer have a one-to-one connection between age and age-squared
  - If centered at 40yo: a 35 yo and a 45 yo will both have centered age of 5, and age-squared of 25

- Check out the Penn State site for a work through of an example with VIFs

Summary of multicollinearity

- Correlated covariates/predictors will hurt our model's precision and interpretations of coefficients

- We need to check for multicollinearity by using VIFs or GVIFs

- If $VIF > 10$ or $GVIF^{1/(2 \cdot df)} > 3.162$, we need to do something about the covariates
  - Data based: remove one the of correlated variables
  - Structural based: centering usually fixes it

Regression analysis process

Model Selection

- Building a model
- Selecting variables
- Prediction vs interpretation
- Comparing potential models

Model Fitting

- Find best fit line
- Using OLS in this class
- Parameter estimation
- Categorical covariates
- Interactions

Model Evaluation

- Evaluation of model fit
- Testing model assumptions
- Residuals
- Transformations
- Influential points
- Multicollinearity

Model Use (Inference)

- Inference for coefficients
- Hypothesis testing for coefficients
- Inference for expected Y given X
- Prediction of new Y given X