

Lesson 15: MLR Model Diagnostics

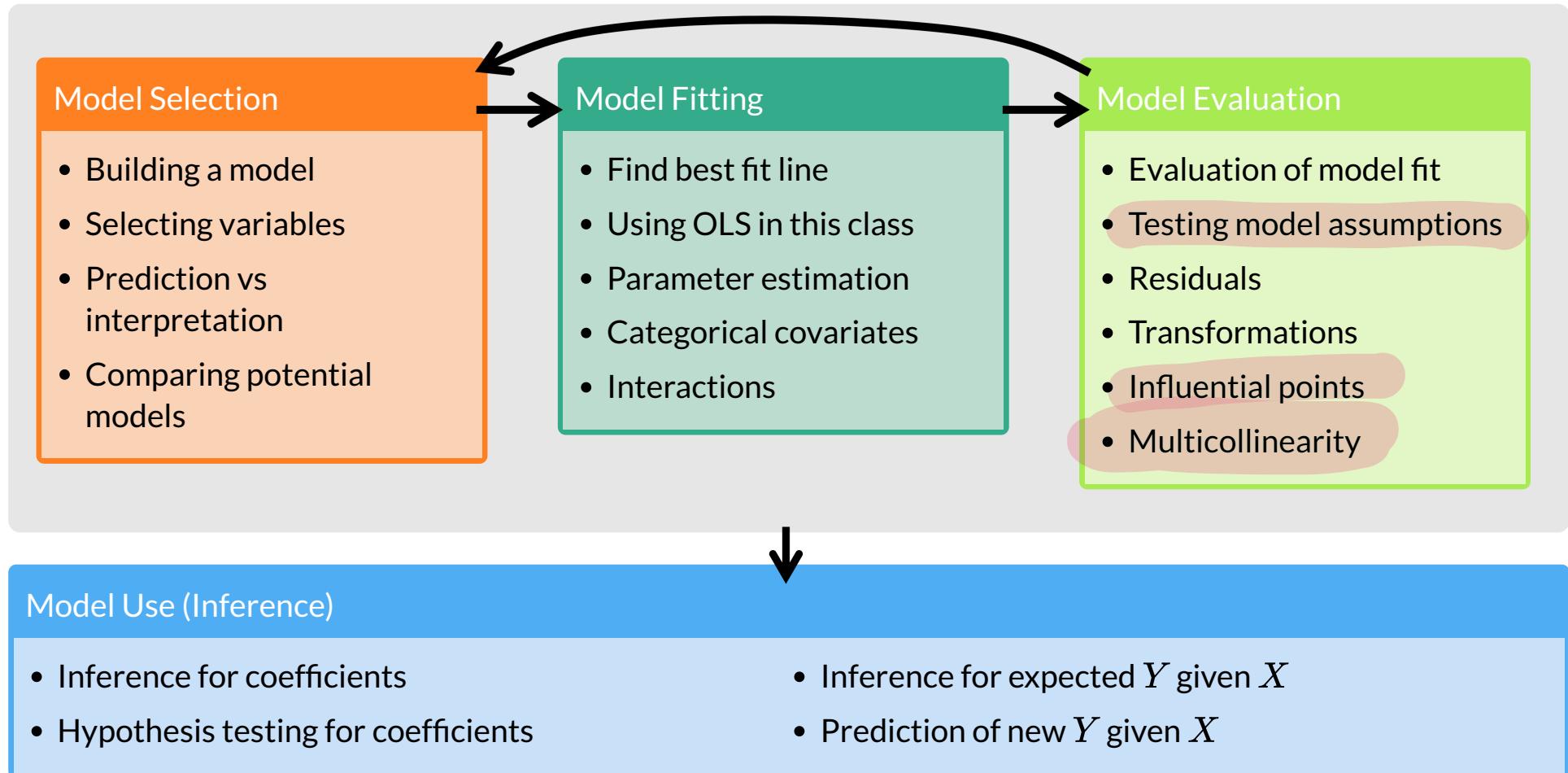
Nicky Wakim

2025-03-05

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 its general form to detect and correct multicollinearity

Regression analysis process



Let's remind ourselves of the final model

- Our **final model** contains
 - Female Literacy Rate **FLR**
 - CO2 Emissions in quartiles **C02_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**
 - No interactions

► Display regression table for final model

Characteristic	Beta	95% CI ¹	p-value
Female literacy rate (%)	-0.07	-0.17, 0.02	0.13
CO2 emissions quartiles			
[0.0439,0.806]	—	—	
(0.806,2.54]	1.1	-2.7, 4.9	0.6
(2.54,4.66]	-0.29	-5.1, 4.6	>0.9
(4.66,35.2]	-0.60	-5.6, 4.5	0.8
Income levels			
Low income	—	—	
Lower middle income	5.4	0.75, 10	0.024
Upper middle income	6.1	0.20, 12	0.043
High income	8.0	1.4, 15	0.018
World region			
Africa	—	—	
Americas	9.0	4.9, 13	<0.001
Asia	5.3	2.0, 8.5	0.002
Europe	6.9	1.1, 13	0.020
Access to improved water (%)	0.17	0.03, 0.30	0.015
Food supply (kcal PPD)	0.00	0.00, 0.01	0.073
Intergovernmental group			
g77	—	—	
oecd	1.1	-4.2, 6.5	0.7
others	1.0	-4.0, 6.1	0.7

¹ CI = Confidence Interval

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 our final model, it is hard to visualize outliers and influential points
- I highly encourage you revisit the SLR lessons ([SLR: Checking model assumptions](#) and [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 (`.cooksdi`), \hat{Y}_i (`.fitted`), leverage (`.hat`), residuals (`.resid`), std residuals (`.std.resid`)

```
1 aug = augment(final_model)
2 head(aug) %>% relocate(.fitted, .resid, .std.resid, .hat, .cooksdi, .after = LifeExpectancyYrs)
```

A tibble: 6 × 14

obs	LifeExpectancyYrs	.fitted	.resid	.std.resid	.hat	.cooksdi	FemaleLiteracyRate
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

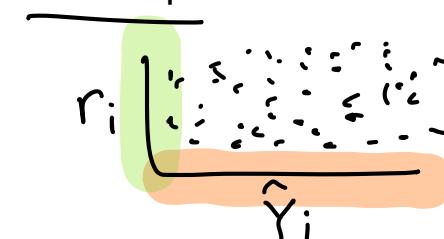
i 7 more variables: C02_q <fct>, income_levels1 <fct>, four_regions <fct>, WaterSourcePrct <dbl>, FoodSupplykcPPD <dbl>, members_oecd_g77 <chr>, .sigma <dbl>

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 its general form to **detect and correct multicollinearity**

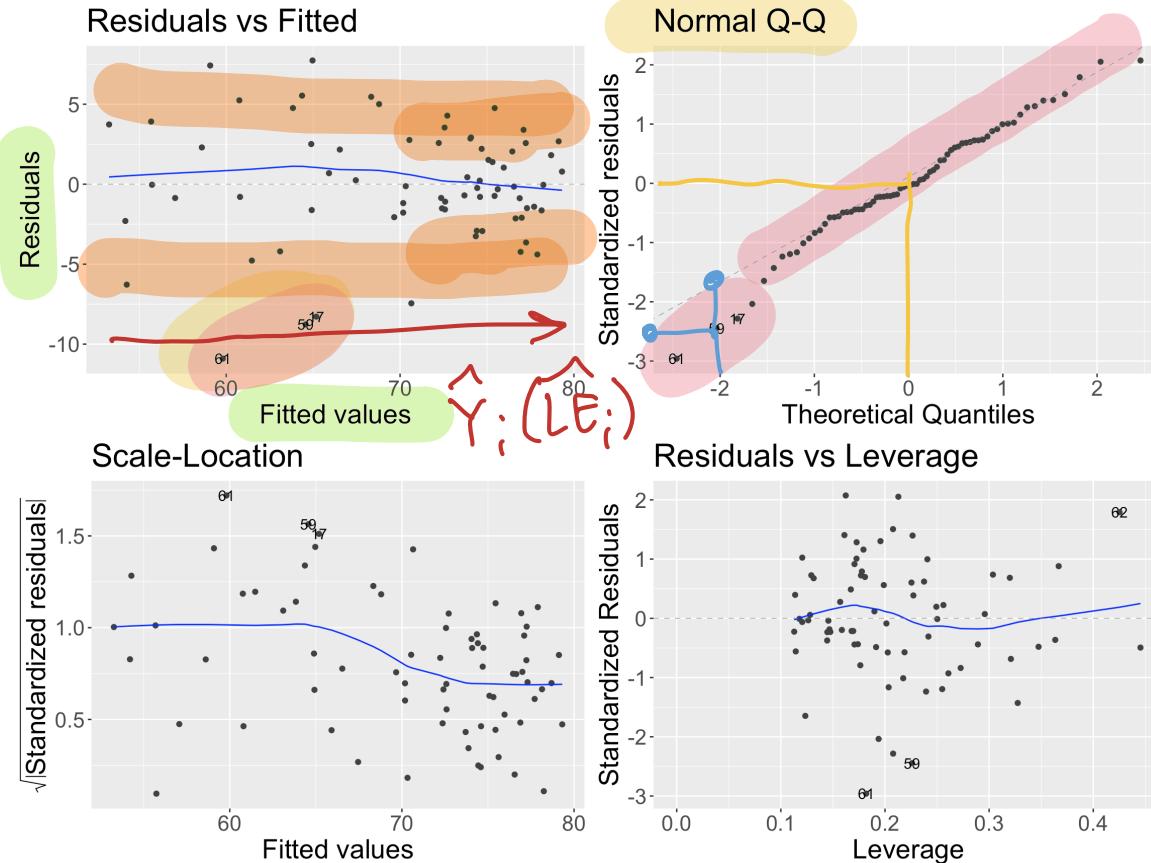
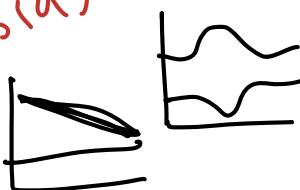
Summary of the assumptions and their diagnostic tool

Assumption	What needs to hold?	Diagnostic tool
Linearity	<ul style="list-style-type: none">Relationship between each X and Y is linear <p><i>really only for numeric X's</i></p>	<ul style="list-style-type: none">Scatterplot of Y vs. X
Independence	<ul style="list-style-type: none">Observations are independent from each other	<ul style="list-style-type: none">Study design
Normality	<ul style="list-style-type: none">Residuals (and thus $Y X_1, X_2, \dots, X_p$) are normally distributed	<ul style="list-style-type: none">QQ plot of residuals<u>Distribution of residuals</u>
Equality of variance	<ul style="list-style-type: none">Variance of residuals (and thus $Y X_1, X_2, \dots, X_p$) is same across fitted values (homoscedasticity)	<ul style="list-style-type: none">Residual plot 

autoplot() to examine equality of variance and Normality

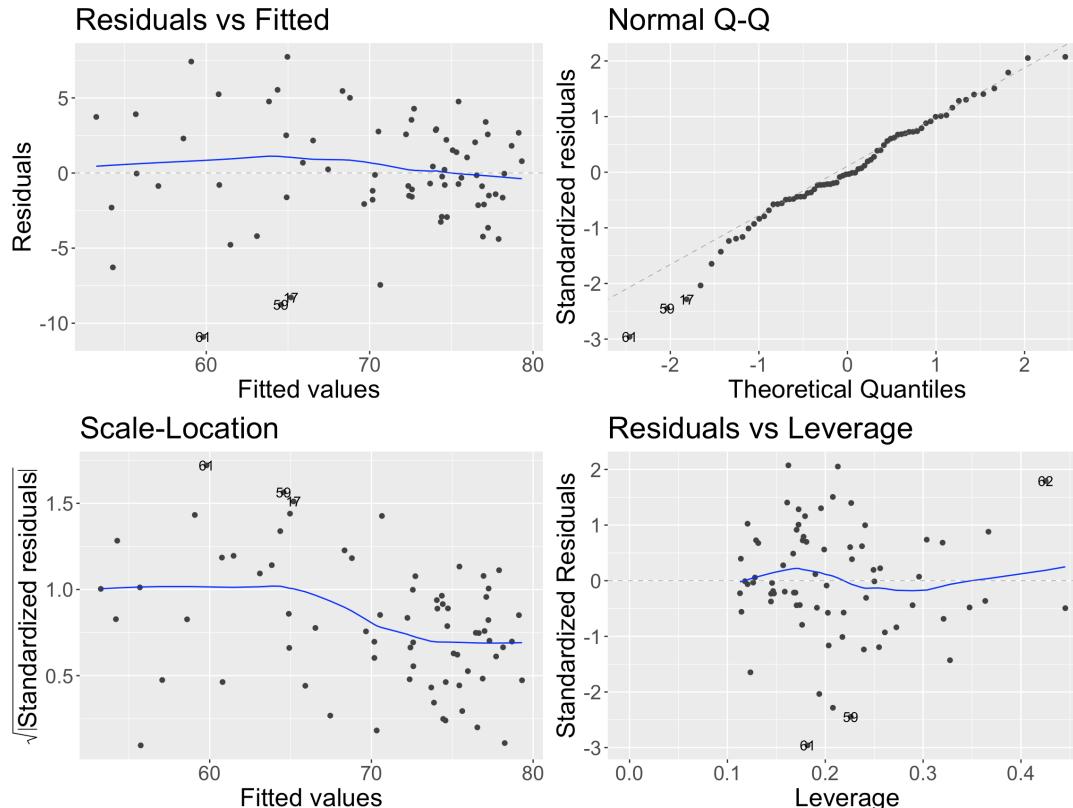
```
1 library(ggfortify)  
2 autoplot(final_model) + theme(text=element_text(size=20))
```

across axis (\hat{Y}_i) is the spread r_i across (resid) same?



autoplot() to examine equality of variance and Normality

```
1 library(ggfortify)  
2 autoplot(final_model) + theme(text=element_text(size=12))
```



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

13:36 Mon Mar 3

X

Join by Web PollEv.com/nickywakim275

QR code

Take a look at Lesson 7 notes with influential points in SLR. What do you think will happen with the influential points comparing SLR to our final model with MLR?

21

Less influential points because more covariates might explain observations that were previously influential i...

Same influential points because more covariates might not explain observations that were previously influenti...

More influential points because more covariates explain less observations that were previously explained in SLR

Idk, can't we just find out?

Powered by  Poll Everywhere

$\sum (Y_i - \hat{Y}_i)^2$

SSE either decrease or stay same w/ more covariates in model

⇒ explain more variance or same

⇒ less inf pts or stays same

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 its general form to **detect and correct multicollinearity**

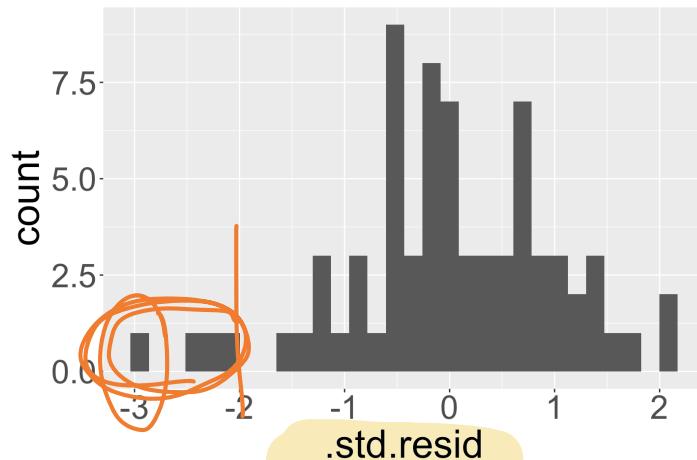
Identifying outliers

Internally standardized residual

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

- 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

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



Countries that are outliers ($|r_i| > 3$)

- We can identify the countries that are outliers

```
1 aug %>% relocate(.std.resid, .after = country) %>%  
2 filter(abs(.std.resid) > 3) %>% arrange(desc(abs(.std.resid)))  
  
# A tibble: 0 × 15  
# i 15 variables: country <chr>, .std.resid <dbl>, LifeExpectancyYrs <dbl>,  
#   FemaleLiteracyRate <dbl>, C02_q <fct>, income_levels1 <fct>,  
#   four_regions <fct>, WaterSourcePrct <dbl>, FoodSupplykcPPD <dbl>,  
#   members_oecd_g77 <chr>, .fitted <dbl>, .resid <dbl>, .hat <dbl>,  
#   .sigma <dbl>, .cooksrd <dbl>
```

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

- We can identify the countries that are outliers

```
1 aug %>% relocate(.std.resid, .after = country) %>%  
2 filter(abs(.std.resid) > 2) %>% arrange(desc(abs(.std.resid)))  
  
# A tibble: 6 × 15  
# ... with 9 more variables: four_regions <fct>, WaterSourcePrct <dbl>,  
#   FoodSupplykcPPD <dbl>, members_oecd_g77 <chr>, .fitted <dbl>, .resid <dbl>,  
#   .hat <dbl>, .sigma <dbl>, .cooksrd <dbl>  
  
# A tibble: 6 × 15  
country .std.resid LifeExpectancyYrs FemaleLiteracyRate C02_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>,  
#   FoodSupplykcPPD <dbl>, members_oecd_g77 <chr>, .fitted <dbl>, .resid <dbl>,  
#   .hat <dbl>, .sigma <dbl>, .cooksrd <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

```
1 aug = aug %>% relocate(.hat, .after = FemaleLiteracyRate)
2 aug %>% arrange(desc(.hat)) %>% head(5)

# A tibble: 5 × 15
  country      LifeExpectancyYrs FemaleLiteracyRate .hat C02_q income_levels1
  <chr>            <dbl>             <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 He...  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 ...
# i 9 more variables: four_regions <fct>, WaterSourcePrct <dbl>,
#   FoodSupplykcPPD <dbl>, members_oecd_g77 <chr>, .fitted <dbl>, .resid <dbl>,
#   .sigma <dbl>, .cooksrd <dbl>, .std.resid <dbl>
```

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

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

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

# A tibble: 0 × 15
# i 15 variables: country <chr>, LifeExpectancyYrs <dbl>,
#   FemaleLiteracyRate <dbl>, .hat <dbl>, C02_q <fct>, income_levels1 <fct>,
#   four_regions <fct>, WaterSourcePrct <dbl>, FoodSupplykcPPD <dbl>,
#   members_oecd_g77 <chr>, .fitted <dbl>, .resid <dbl>, .sigma <dbl>,
#   .cooksdi <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

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

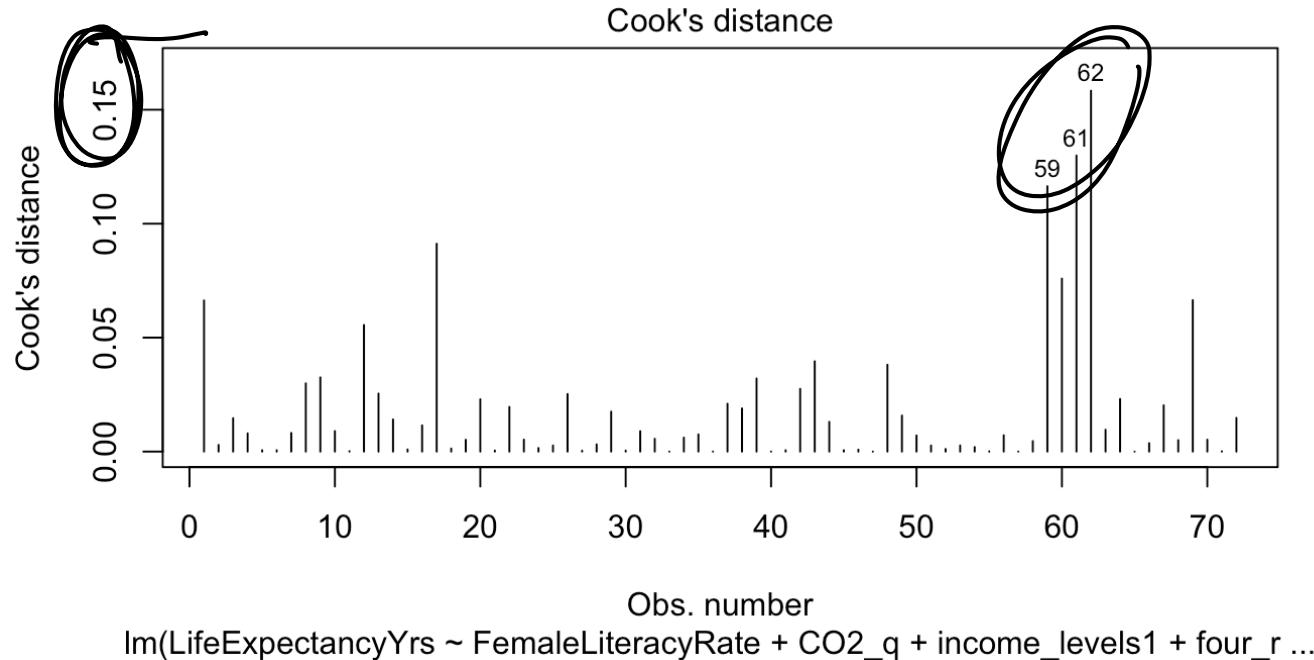
```
# A tibble: 0 × 15
# i 15 variables: country <chr>, .cooksdi <dbl>, LifeExpectancyYrs <dbl>,
#   FemaleLiteracyRate <dbl>, .hat <dbl>, C02_q <fct>, income_levels1 <fct>,
#   four_regions <fct>, WaterSourcePrct <dbl>, FoodSupplykcPPD <dbl>,
#   members_oecd_g77 <chr>, .fitted <dbl>, .resid <dbl>, .sigma <dbl>,
#   .std.resid <dbl>
```

- 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: `.cooksdi`

Plotting Cook's Distance

- 1 `# plot(model)` shows figures similar to `autoplot()` but adds Cook's distance
- 2 `plot(final_model, which = 4)`

↳ 4th plot



- Identify 3 highest Cook's distance: 59, 61, 62 (South Africa, Kingdom of Eswatini, Tajikistan)

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? → *are they available to us?*
 - 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

Join by Web [PollEv.com/nickywakim275](https://pollev.com/nickywakim275)

"polleverywhere.com" is in full screen.
Swipe down to exit.

In our project, what might be a cause for an influential point?

Data entry error

Nonfactual information entry.

Powered by  Poll Everywhere

issue w/
LINE =
taking 2x

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 categorization or transformation (or numeric variables) 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)

steps
in
our
class

beyond
our 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 its 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**

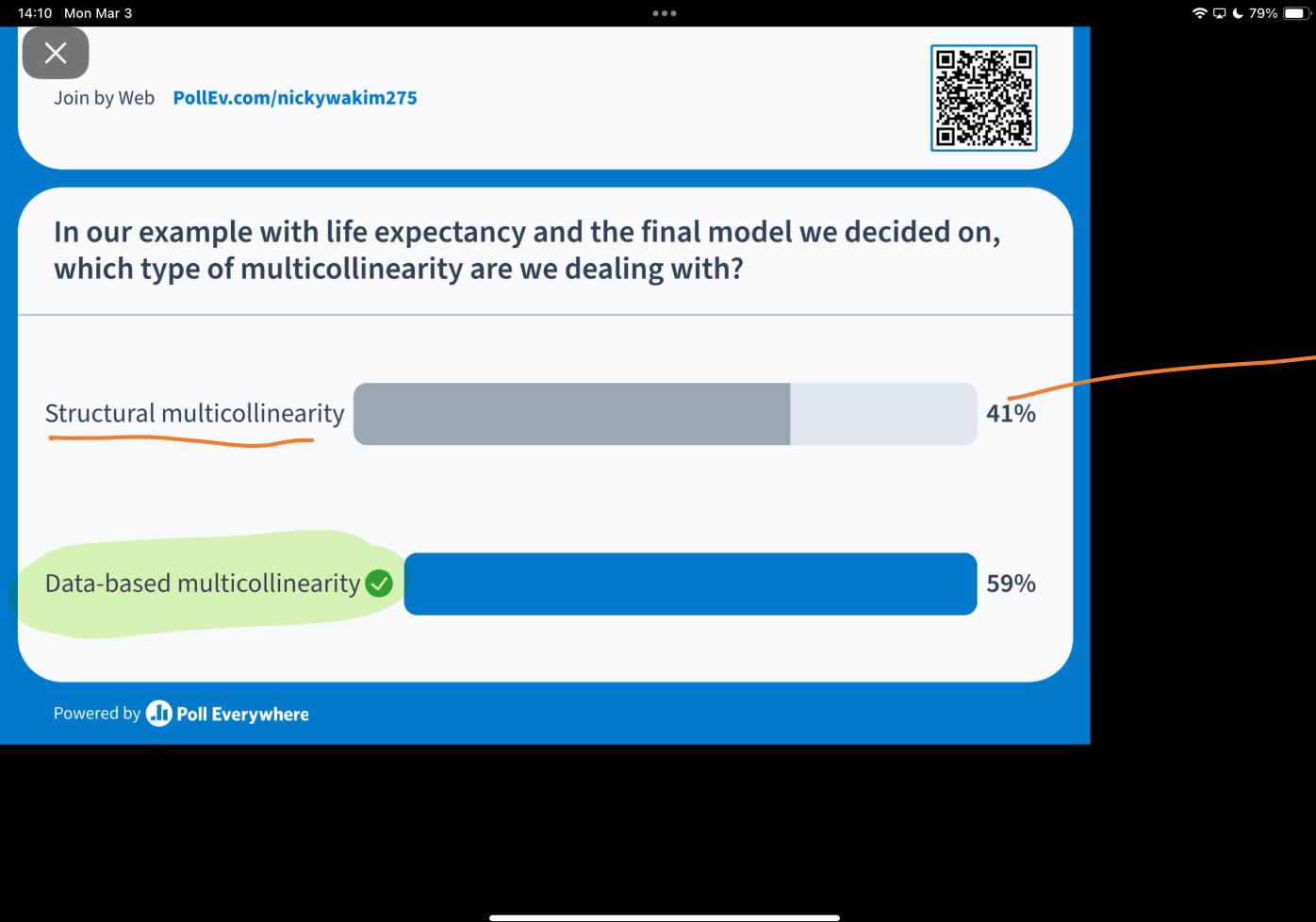
- Mathematical artifact caused by creating new covariates from other covariates
 - For example: If we have age, and decide to transform age to include age-squared
 - Then we have age and age-squared in the model: age-squared is perfectly predicted by age!

- **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
 - Or we are using two variables that are practically measuring the same thing

centering

Poll Everywhere Question 3



- no transformations on covariates
- no interactions

Why is multicollinearity a problem?

In linear regression, depending on the other predictors in the model, the following will change:

- ★ • Estimated regression coefficient of any one variable
 - Not necessarily bad, but a big change might be an issue
- ★ • Hypothesis tests for any coefficient may yield different conclusions
- Contribution of any one predictor variable in reducing sum of squared errors

] all depend
on the other
variables in
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 on **Model Selection**
- Multicollinearity is *not really an issue* when our **goal is prediction** — *accuracy NOT precision*
 - 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
 - $VIF < 4$: Good!
 - $4 < VIF < 10$: Warrant 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

precision still okay \Rightarrow CI's of $\hat{\beta}$'s are okay
Coefficient estimate CI's will be really wide

Let's apply it to our final model

- Naive way to calculate this in R:

```
1 library(rms)  
2 rms::vif(final model)
```

FemaleLiteracyRate	C02_q(0.806, 2.54]
	2.979224
C02_q(2.54, 4.66]	C02_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
WaterSourcePrct	FoodSupplykCPPD
4.824266	3.499250
members_oecd_g77oecd	members_oecd_g77others
2.720955	5.125196

- 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, $GVIF^{1/(2 \cdot df)} = \sqrt{GVIF}$

```
1 library(car)  
2 car::vif(final_model)
```

↓
 $\sqrt{10}$

→ cut off is 10

	<u>GVIF</u>	Df	<u>GVIF^(1/(2*Df))</u>
FemaleLiteracyRate	4.863139	1	2.205253
C02_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

use $GVIF^{1/2df}$

instead of

GVIF or VIF

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 $GVIF^{1/(2\cdot df)} > \sqrt{10} = 3.162$

```
1 car::vif(final_model)
```

	GVIF	Df	GVIF^(1/(2*Df))
FemaleLiteracyRate	4.863139	1	2.205253
C02_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

> 3.162?

- 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 your 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 walk 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 $\underline{VIF} > 10$ 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

