

Lesson 15: MLR Model Diagnostics

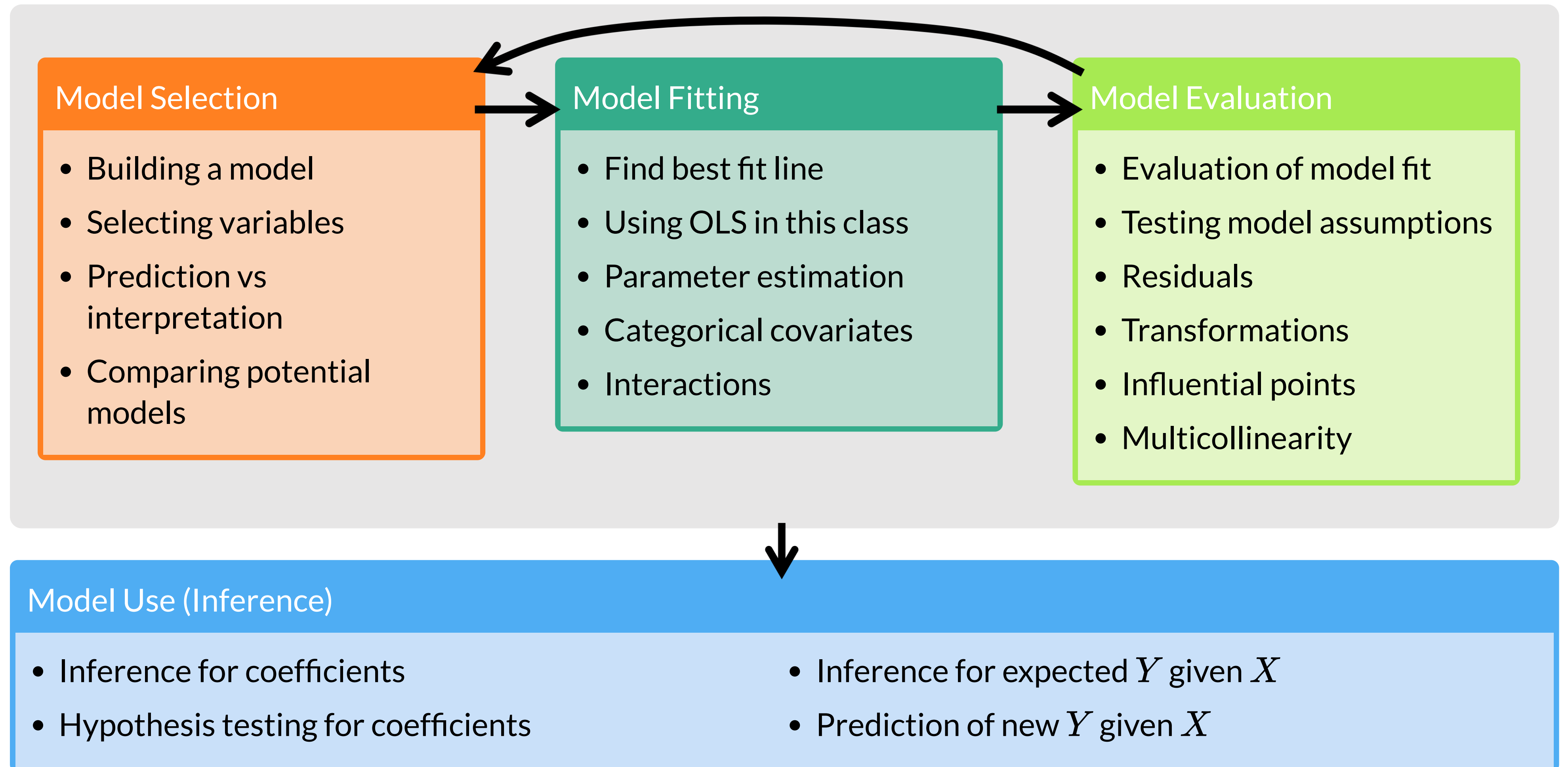
Nicky Wakim

2026-03-04

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
 - Cell phones per 100 people
 - Basic sanitation (quartiles)
 - Freedom status
 - Income level
 - Vaccination rate
 - Happiness score
 - No interactions

► Display regression table for final model

Characteristic	Beta	95% CI	p-value
Cell phones per 100 people	0.01	-0.02, 0.04	0.7
Basic sanitation (%)			
[21.6,61.1]	—	—	
(61.1,92.8]	4.5	2.0, 7.0	<0.001
(92.8,98]	7.0	4.2, 9.8	<0.001
(98,100]	9.0	5.8, 12	<0.001
Freedom status			
NF	—	—	
PF	1.2	-0.72, 3.2	0.2
F	-0.35	-2.9, 2.2	0.8
Income level			
Low income	—	—	
Lower middle income	-0.25	-3.4, 2.9	0.9
Upper middle income	-0.11	-4.2, 4.0	>0.9
High income	3.5	-1.8, 8.8	0.2
Vaccination rate (%)	0.03	-0.08, 0.14	0.6
Happiness score	0.14	0.03, 0.25	0.014

Abbreviation: 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 6 variables in our final model, it is hard to scatterplots to visualize linearity, 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 (`.cooksd`), \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, .cooksd, .after = life_exp)
```

```
# A tibble: 6 × 13
  life_exp .fitted .resid .std.resid .hat .cooksd cell_phones_100 BS_q
  <dbl>   <dbl> <dbl>   <dbl> <dbl>   <dbl>   <dbl> <fct>
1    62.6    62.8 -0.166  -0.0506 0.342  0.000111    56.3 (61.1,92.0]
2    76.1    75.1  0.954   0.251  0.110  0.000647    98.4 (98,100]
3    73.4    80.3 -6.88   -1.87   0.166  0.0581    196. (98,100]
4    75.4    75.1  0.267   0.0700 0.109  0.0000499   131. (98,100]
5    73.7    75.5 -1.88   -0.486 0.0829 0.00178    131. (98,100]
6    71.4    71.0  0.374   0.0981 0.108  0.0000971   108. (92.8,98]
# i 5 more variables: freedom_status <fct>, income_level_4 <fct>,
# vax_rate <dbl>, happiness_score <dbl>, .sigma <dbl>
```

[RDocumentation on the `augment\(\)` function.](#)

Learning Objectives

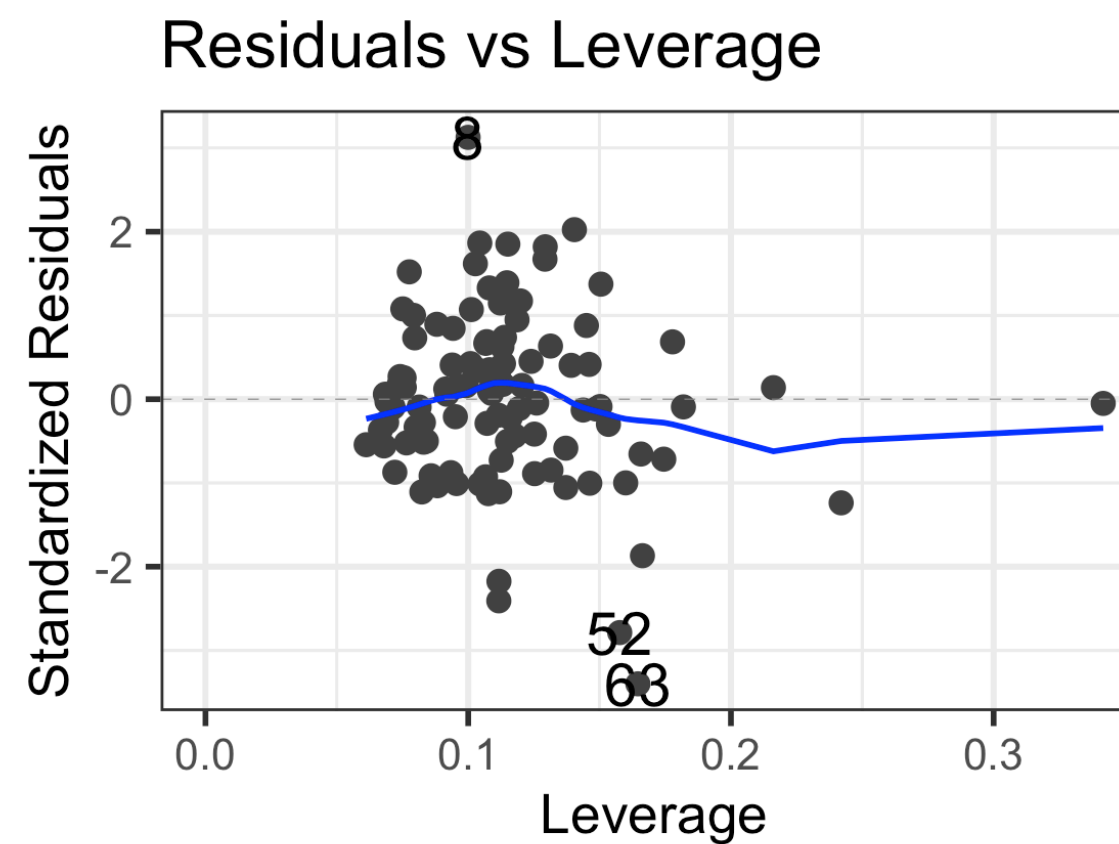
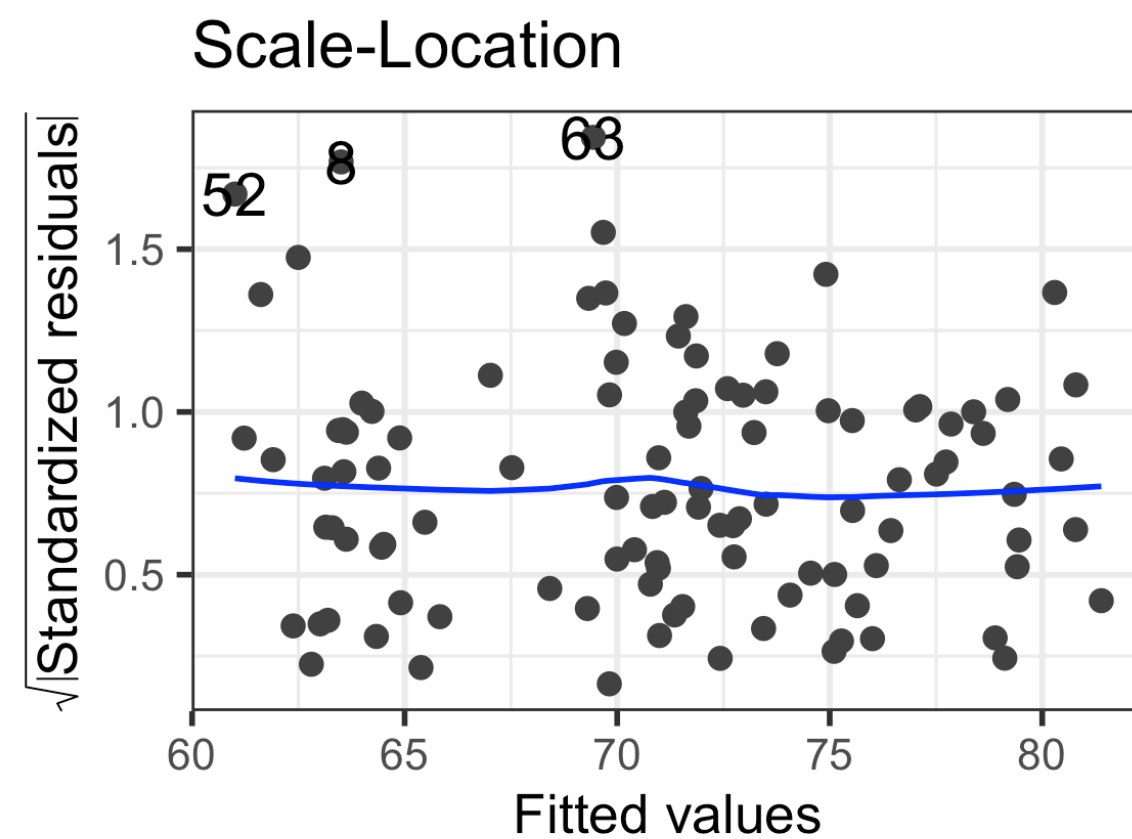
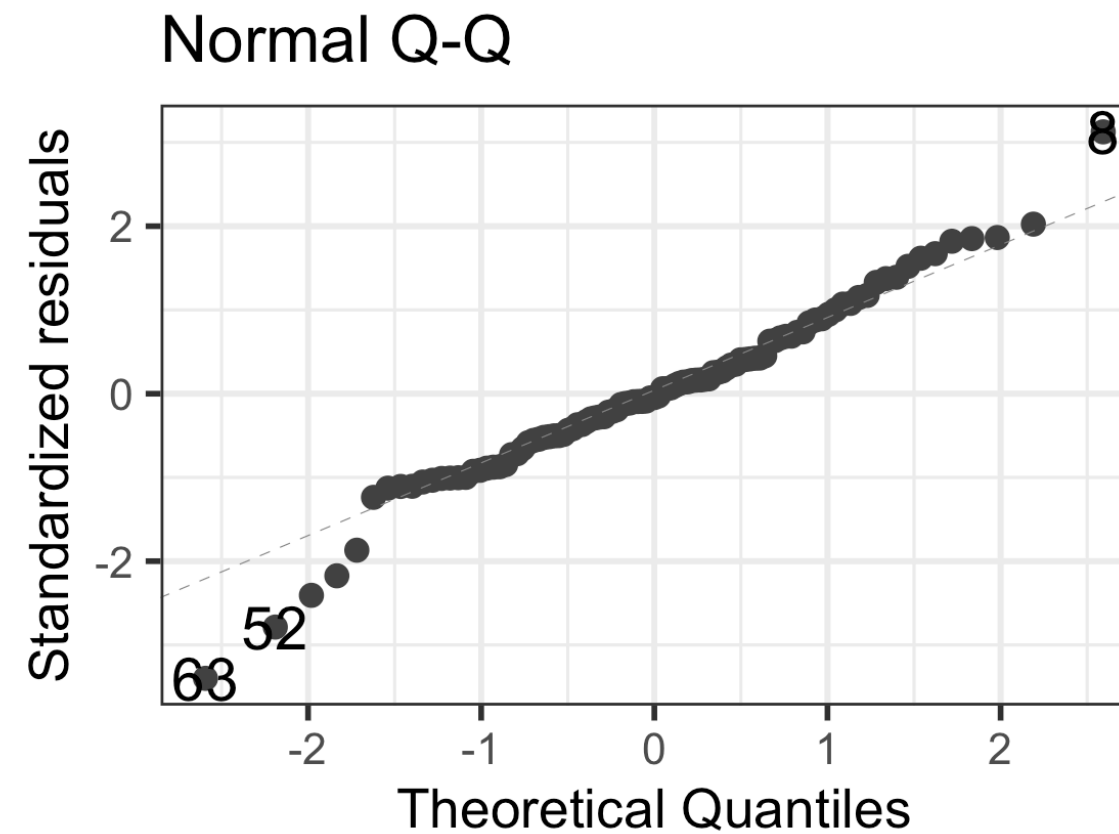
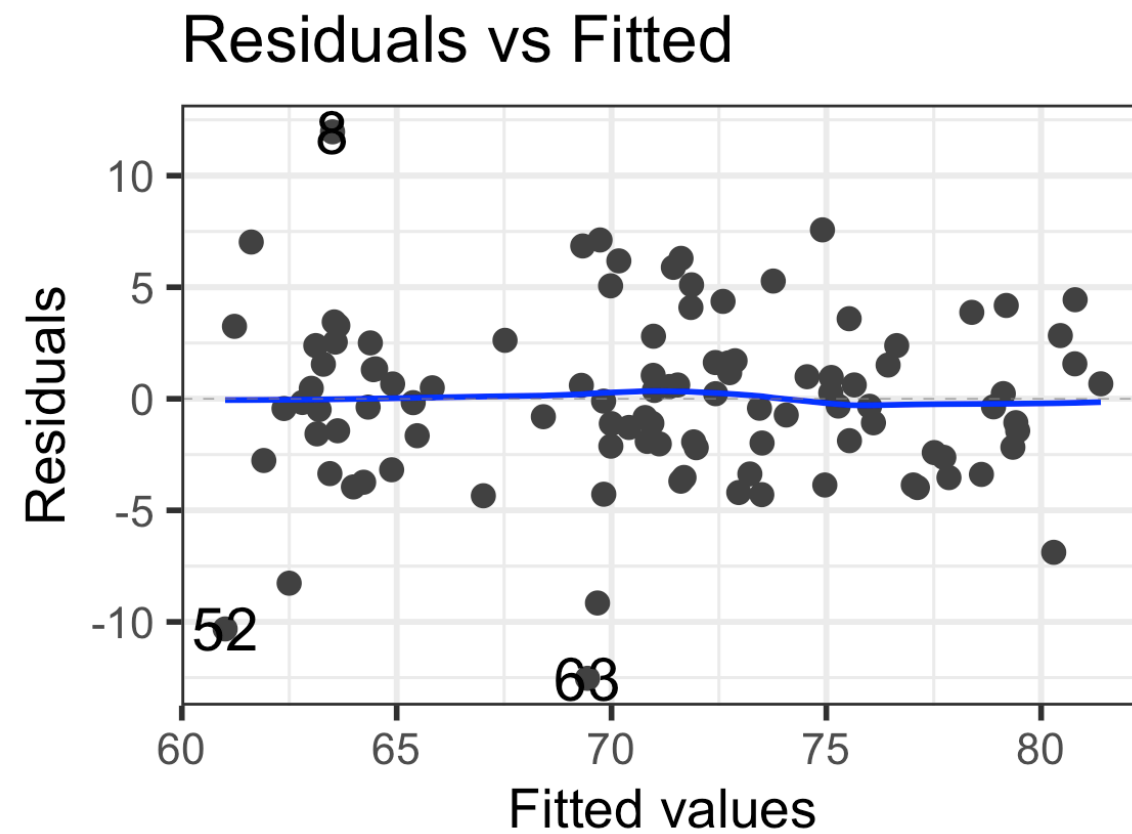
1. Apply tools from SLR (Lesson 6: SLR Diagnostics) in MLR to **evaluate LINE assumptions**, including residual plots and QQ-plots
2. Apply tools involving standardized residuals, leverage, and Cook's distance from SLR (Lesson 7: SLR Diagnostics 2) in MLR to **flag potentially influential points**
3. 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	<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 residualsDistribution of residuals
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))
```



autoplot() to examine equality of variance and Normality

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

Looks like 2 obs are flagged:

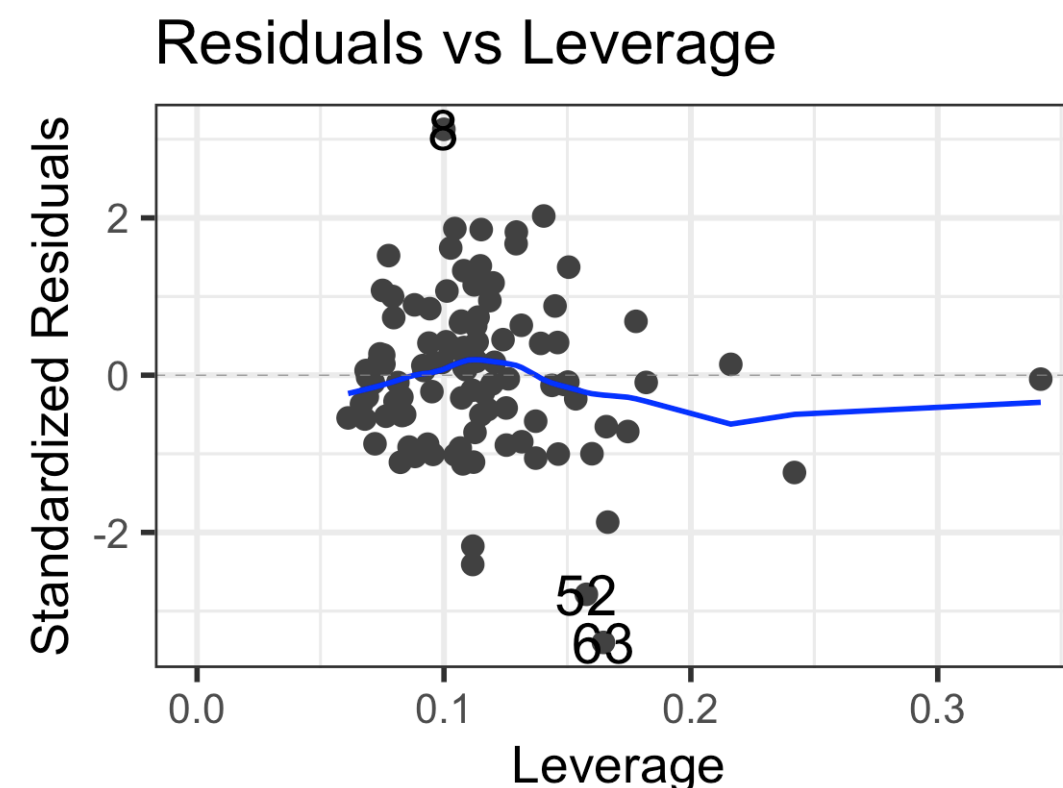
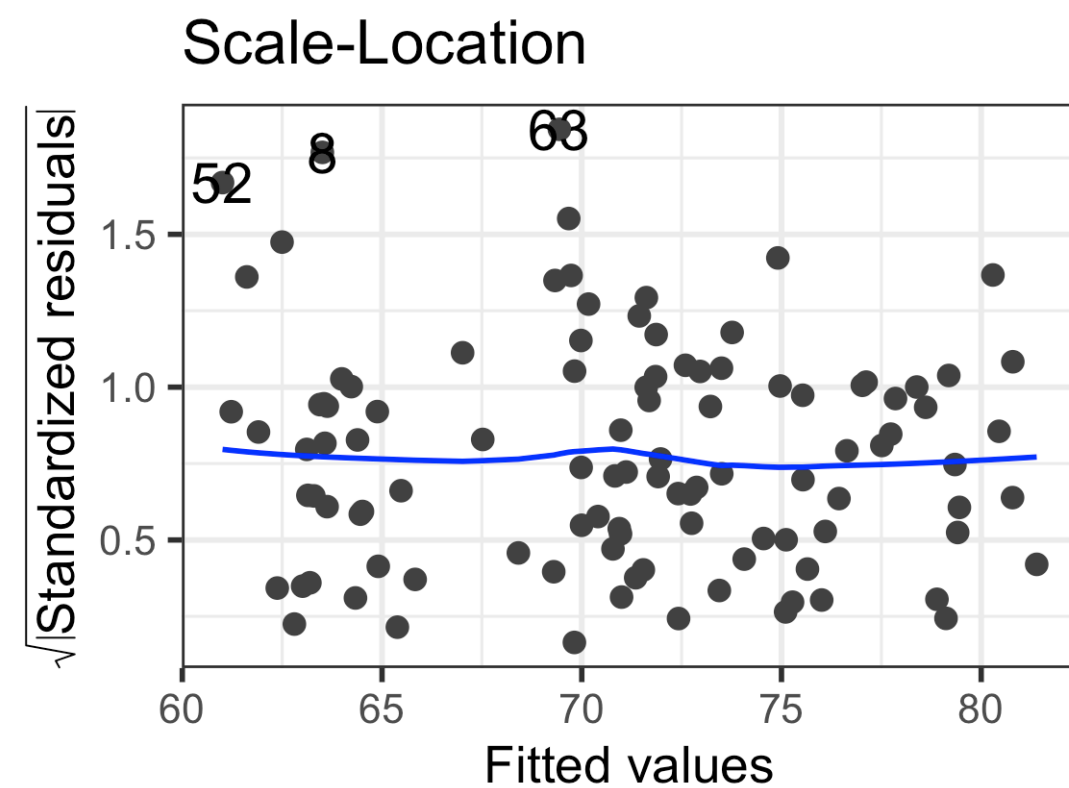
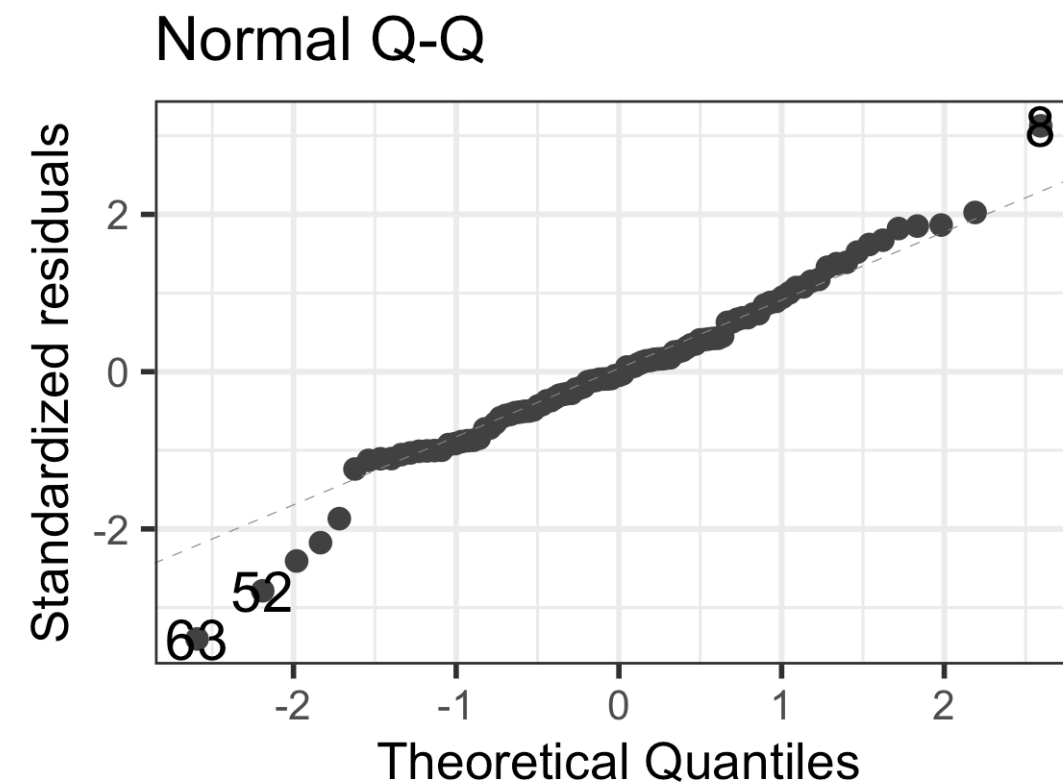
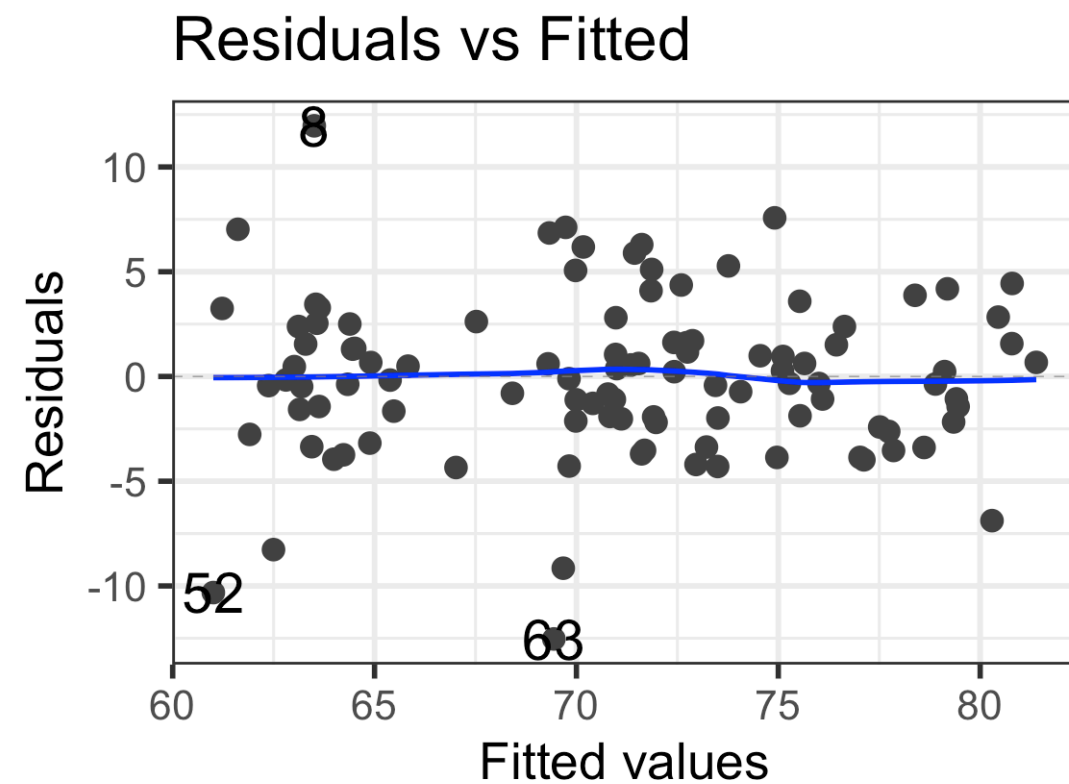
- 8: Bangladesh
- 52: Lesotho
- 66: Malawi

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

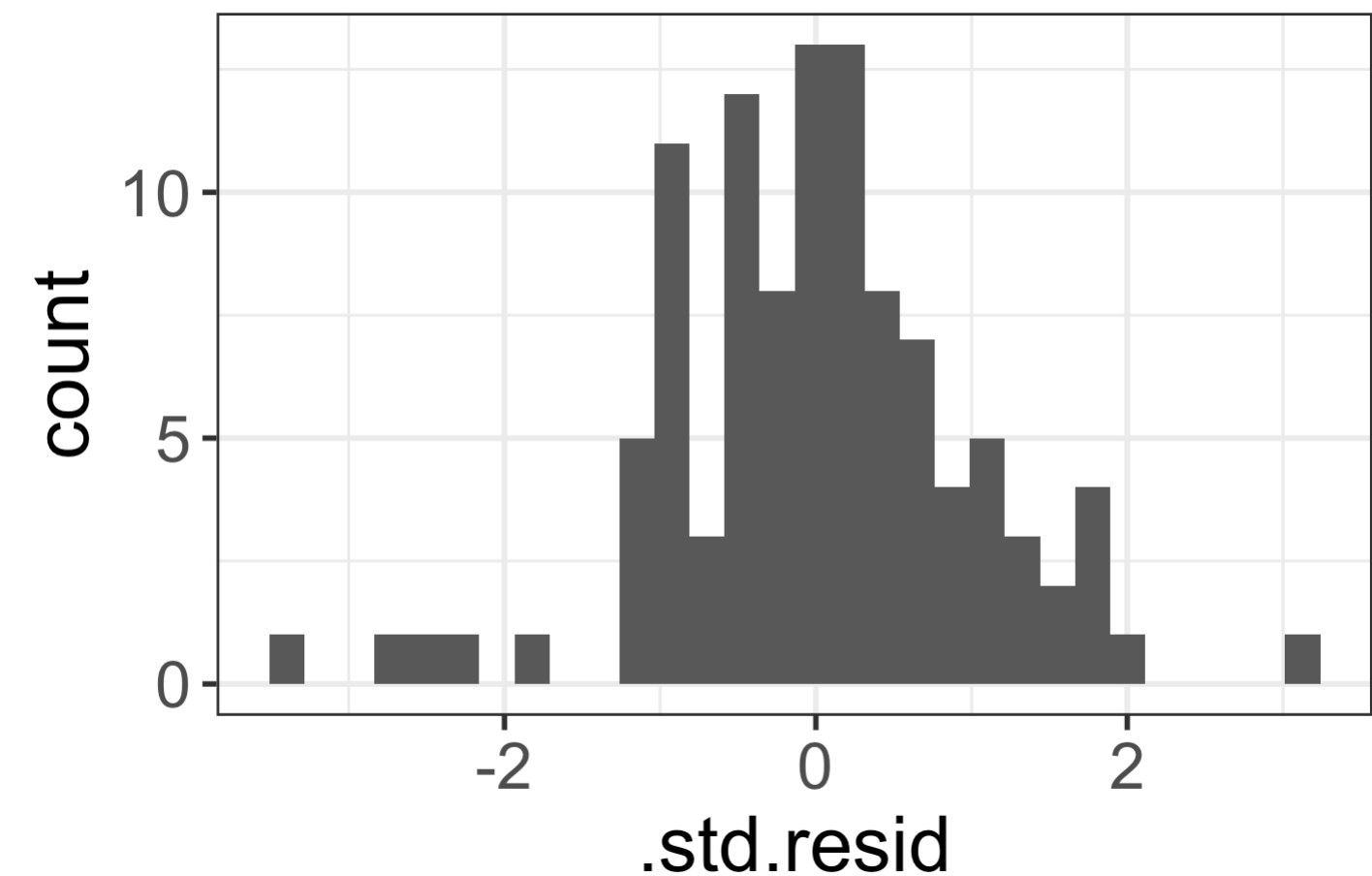
Identifying outliers

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 = territory) %>%  
2   filter(abs(.std.resid) > 3) %>% arrange(desc(abs(.std.resid)))
```

```
# A tibble: 2 × 14
```

```
territory  .std.resid life_exp cell_phones_100 BS_q      freedom_status  
<chr>      <dbl>    <dbl>         <dbl> <fct>      <fct>  
1 Mozambique -3.40    56.9          45.8 (61.1,92.8] PF  
2 Bangladesh  3.13    75.5          110. [21.6,61.1] PF  
# i 8 more variables: income_level_4 <fct>, vax_rate <dbl>,  
# happiness_score <dbl>, .fitted <dbl>, .resid <dbl>, .hat <dbl>,  
# .sigma <dbl>, .cooks_d <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 = cell_phones_100)
2 aug %>% arrange(desc(.hat)) %>% head(5)
```

```
# A tibble: 5 × 14
  territory    life_exp cell_phones_100  .hat BS_q  freedom_status income_level_4
  <chr>         <dbl>         <dbl> <dbl> <fct> <fct>          <fct>
1 Afghanistan  62.6           56.3 0.342 (61... NF          Low income
2 Botswana     62.7           178.  0.242 (61... F          Upper middle ...
3 Yemen        66.3           46.5 0.216 (61... NF          Low income
4 Montenegro  75.7           207.  0.182 (98,... PF          Upper middle ...
5 Gabon        66.9           123.  0.178 [21... NF          Upper middle ...
# i 7 more variables: vax_rate <dbl>, happiness_score <dbl>, .fitted <dbl>,
#   .resid <dbl>, .sigma <dbl>, .cooks_d <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: 1 × 14
```

```
territory    life_exp cell_phones_100  .hat BS_q  freedom_status income_level_4
<chr>          <dbl>          <dbl> <dbl> <fct> <fct>          <fct>
1 Afghanistan    62.6            56.3 0.342 (61.... NF          Low income
# i 7 more variables: vax_rate <dbl>, happiness_score <dbl>, .fitted <dbl>,
#   .resid <dbl>, .sigma <dbl>, .cooksd <dbl>, .std.resid <dbl>
```

Cook's distance

- Measures the overall influence of an observation
- Attempts to measure how much **influence a single observation has** over the fitted model
 - Measures **how coefficient estimates change** when the *i*th observation is removed from the model
 - Combines leverage and outlier information

The Cook's distance for the *i*th observation

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

where h_i is the leverage and r_i is the studentized residual

- 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`

Identifying points with high Cook's distance

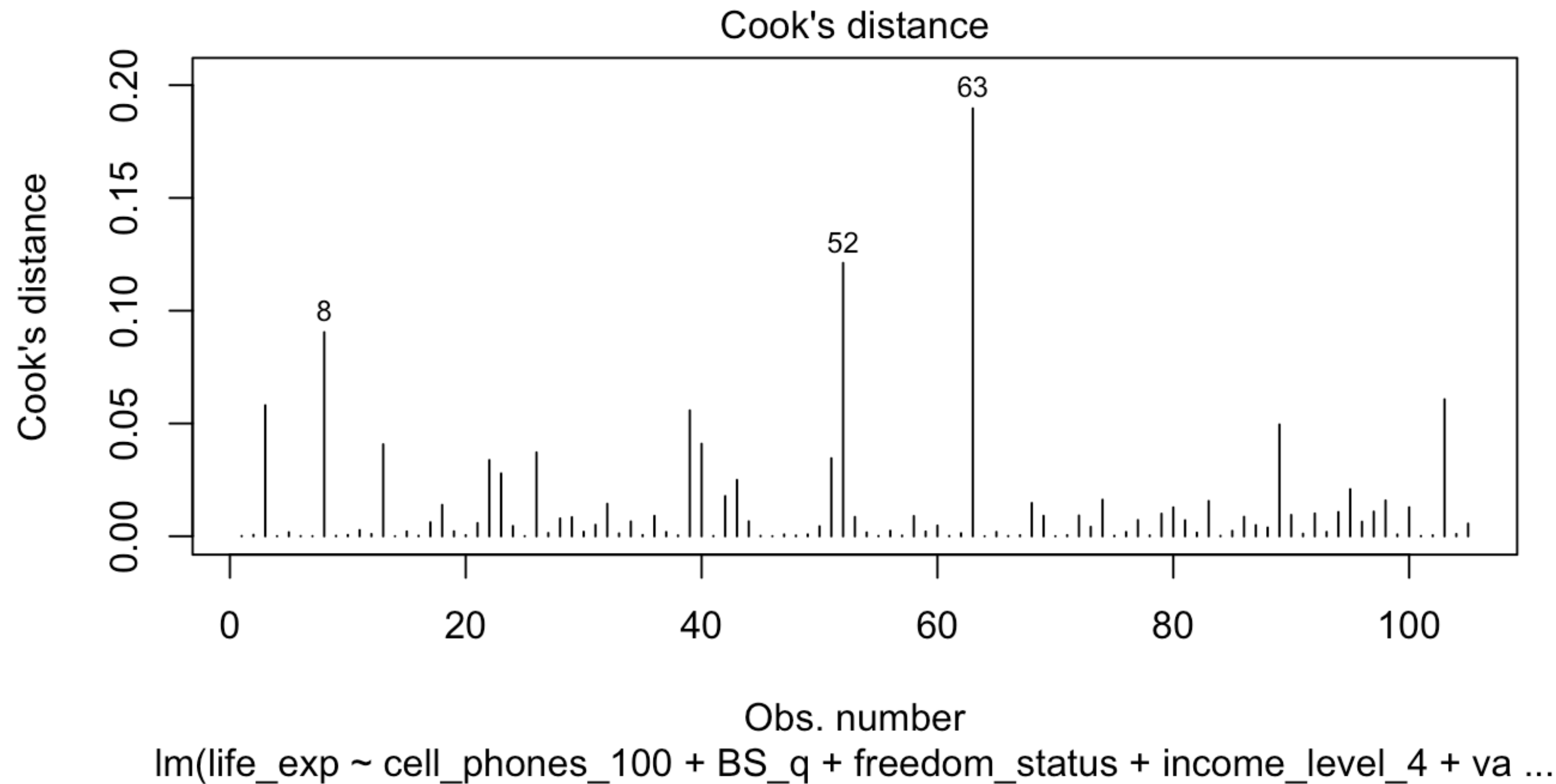
- No countries with high Cook's distance

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

```
# A tibble: 0 × 14
# i 14 variables: territory <chr>, .cooksd <dbl>, life_exp <dbl>,
# cell_phones_100 <dbl>, .hat <dbl>, BS_q <fct>, freedom_status <fct>,
# income_level_4 <fct>, vax_rate <dbl>, happiness_score <dbl>, .fitted <dbl>,
# .resid <dbl>, .sigma <dbl>, .std.resid <dbl>
```

Plotting Cook's Distance

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



- Identify 3 highest Cook's distance: 8, 52, 66 (Bangladesh, Lesotho, Malawi)

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 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)

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

Poll Everywhere Question 3

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

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**
 - 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!
 - $4 < VIF < 10$: 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 in R:

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

```
cell_phones_100          BS_q(61.1,92.8]
      1.903507          1.920420
      BS_q(92.8,98]      BS_q(98,100]
      2.428357          3.171019
freedom_statusPF        freedom_statusF
      1.517866          2.081666
income_level_4Lower middle income income_level_4Upper middle income
      3.672232          5.994002
income_level_4High income vax_rate
      6.812730          1.428216
happiness_score
      2.765792
```

- 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)
```

	GVIF	Df	GVIF^(1/(2*Df))
cell_phones_100	1.903507	1	1.379676
BS_q	2.658818	3	1.177013
freedom_status	1.660933	2	1.135241
income_level_4	5.648213	3	1.334500
vax_rate	1.428216	1	1.195080
happiness_score	2.765792	1	1.663067

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))
cell_phones_100	1.903507	1	1.379676
BS_q	2.658818	3	1.177013
freedom_status	1.660933	2	1.135241
income_level_4	5.648213	3	1.334500
vax_rate	1.428216	1	1.195080
happiness_score	2.765792	1	1.663067

- 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 35yo has centered age of -5 and a 45yo centered age of 5, but both have 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

