Lesson 7: SLR: Checking model assumptions

Nicky Wakim

2025-01-2~~7~~9

Learning Objectives

1. Describe the model assumptions made in linear regression using ordinary least squares
2. Determine if the relationship between our sampled X and Y is linear
3. Use QQ plots to determine if our fitted model holds the normality assumption
4. Use residual plots to determine if our fitted model holds the equality of variance assumption

Let’s remind ourselves of one model we have been working with

- We have been looking at the association between life expectancy and female literacy rate
- We used OLS to find the coefficient estimates of our best-fit line

Population model:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

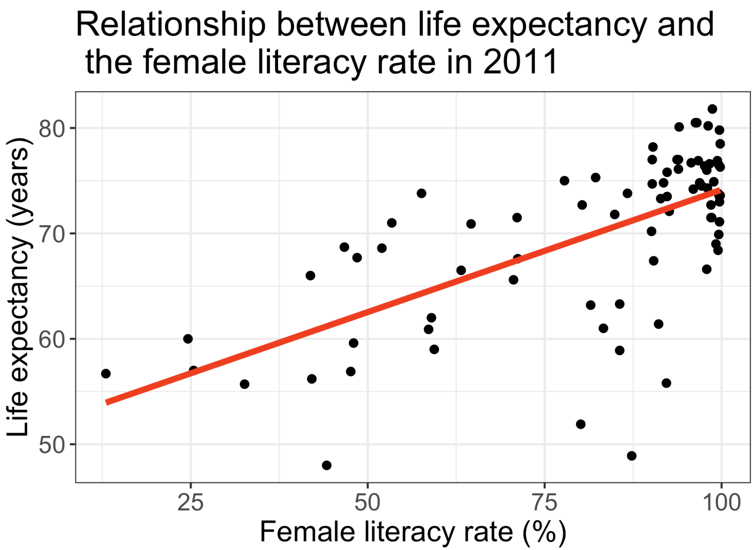
$$LE = \beta_0 + \beta_1 FLR + \varepsilon$$

Estimated model:

term	estimate	std.error	statistic	p.value
(Intercept)	50.93	2.66	19.14	0.00
FemaleLiteracyRate	0.23	0.03	7.38	0.00

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot X$$

$$\longrightarrow \widehat{\text{LE}} = 50.9 + 0.232 \cdot \text{FLR}$$

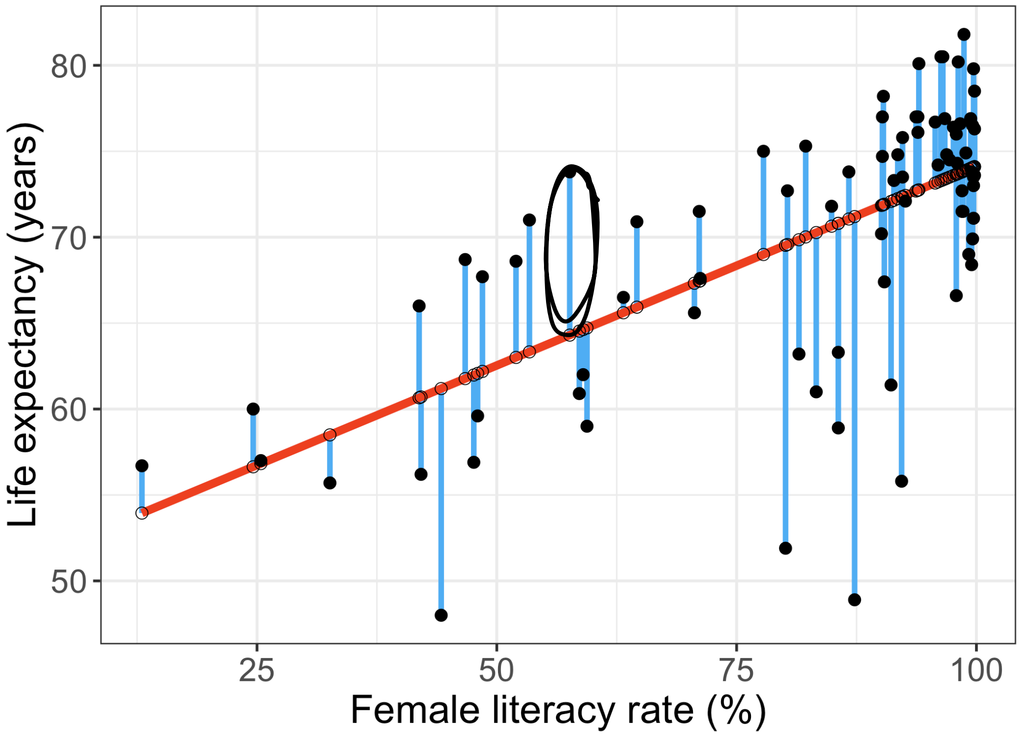


Our residuals will help us a lot in our diagnostics and assumptions!

- The **residuals** $\hat{\epsilon}_i$ are the vertical distances between
  - the observed data (X_i, Y_i)
  - the fitted values (regression line) $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$

$Y_i - \hat{Y}$

$$\hat{\epsilon}_i = Y_i - \widehat{Y}_i, \text{ for } i = 1, 2, \ldots, n$$



Learning Objectives

1. Describe the model assumptions made in linear regression using ordinary least squares
2. Determine if the relationship between our sampled X and Y is linear
3. Use QQ plots to determine if our fitted model holds the normality assumption
4. Use residual plots to determine if our fitted model holds the equality of variance assumption

Least-squares model assumptions: LINE

These are the model assumptions made in ordinary least squares:

- **[L] Linearity** of relationship between variables
- **[I] Independence** of the Y values
- **[N] Normality** of the Y's given X (or residuals)
- **[E] Equality** of variance of the residuals (homoscedasticity) σ^2

Note: These assumptions are baked into the *population model*. We look at the *population parameters* when we discuss these assumptions, but we use the *estimated model* with our data to check *if the assumptions are held up*.

L: Linearity

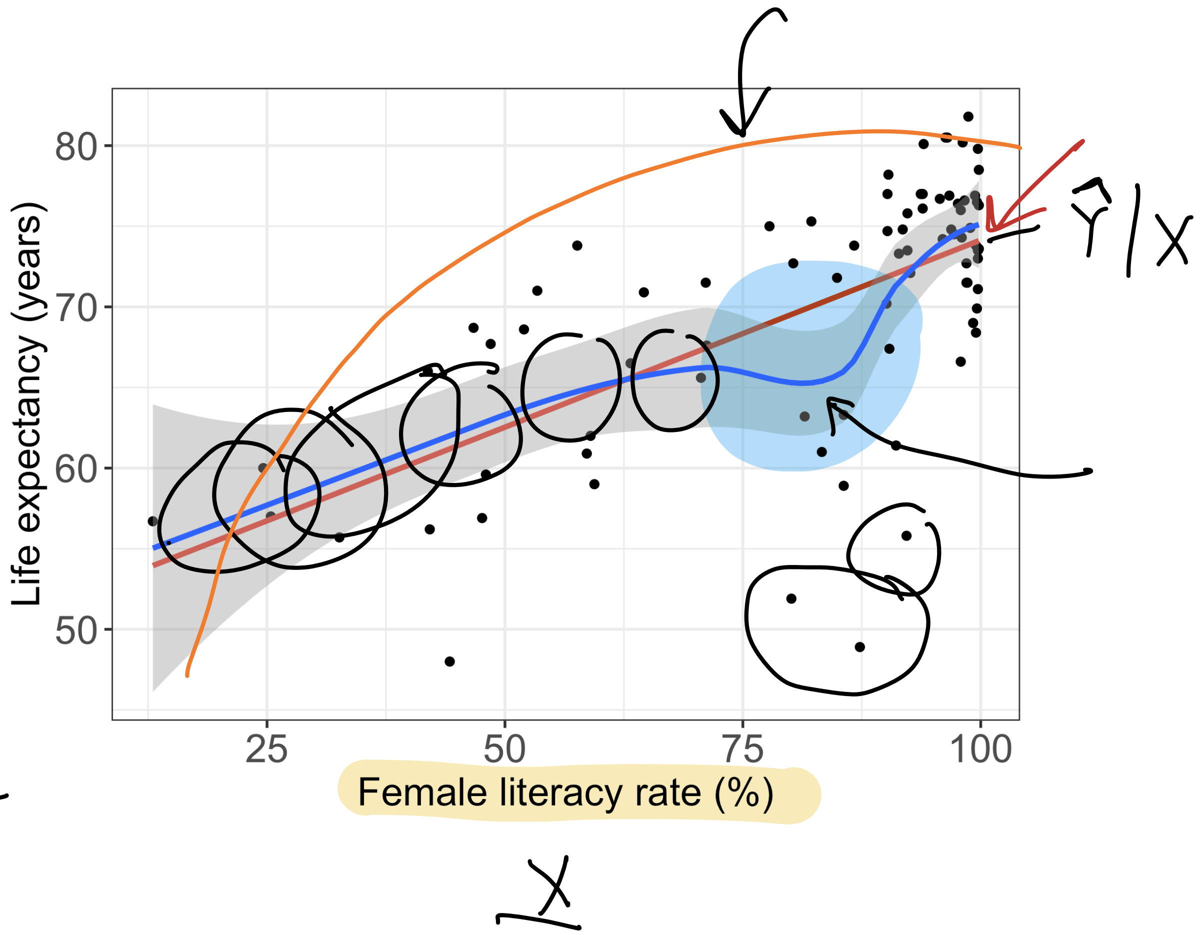
- The relationship between the variables is linear (a straight line):
    - The mean value of Y given X, $\mu_{y|x}$ or $E[Y|X]$, is a straight-line function of X

$$\mu_{y|x} = \underline{\beta_0 + \beta_1 \cdot X}$$

$\underline{Y}$

$$Y^3 = \beta_0 + \beta_1 X + \varepsilon$$

$$Y = \beta_0 + \beta_1 e^X + \varepsilon$$

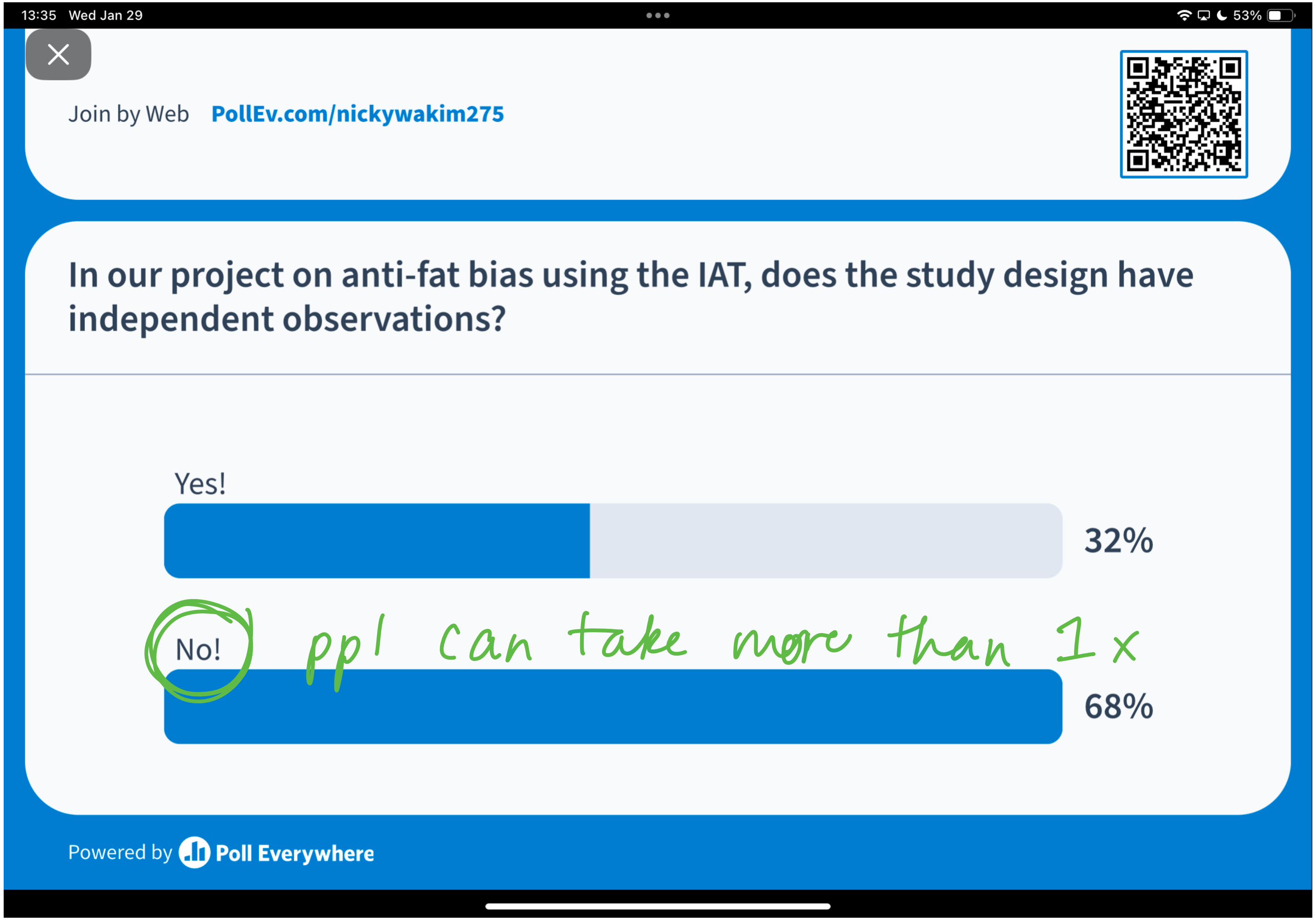


I: Independence of observations

- The Y-values are statistically independent of one another
- Examples of when they are *not* independent, include
    - repeated measures (such as baseline, 3 months, 6 months)
    - data from clusters, such as different hospitals or families
- This condition is checked by reviewing the study *design* and not by inspecting the data

- How to analyze data using regression models when the Y-values are not independent is covered in BSTA 519 (Longitudinal data)

Poll Everywhere Question 1



Join by Web PollEv.com/nickywakim275

In our project on anti-fat bias using the IAT, does the study design have independent observations?

Yes! 32%

[✓ No!] 68%

ppl can take more than 1 x

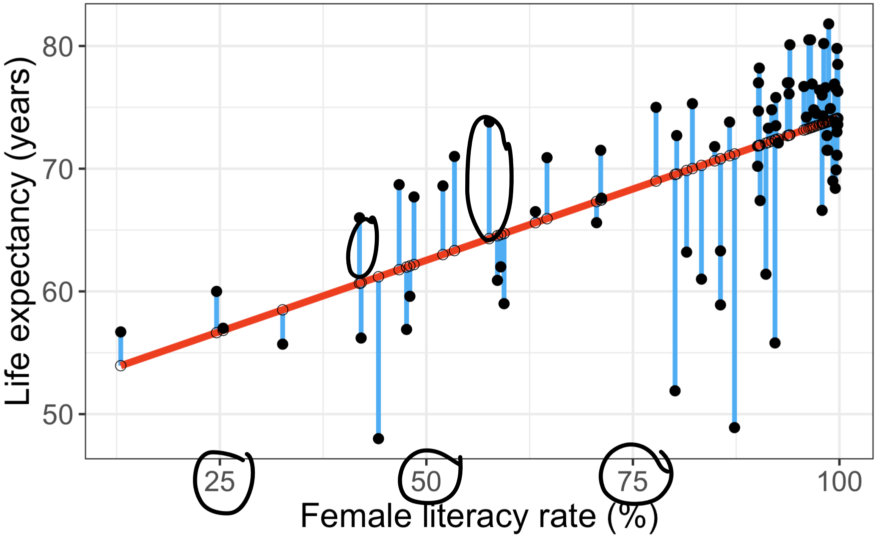
N: Normality

can't just plot y & say it's normal

- For any fixed value of X, Y has normal distribution.
  - Note: This is not about Y alone, but $Y|X$
- Equivalently, the measurement (random) errors ϵ_i 's normally distributed
  - This is more often what we check

$$\varepsilon_i \sim N(0, \sigma^2)$$

Relationship between life expectancy and the female literacy rate in 2011

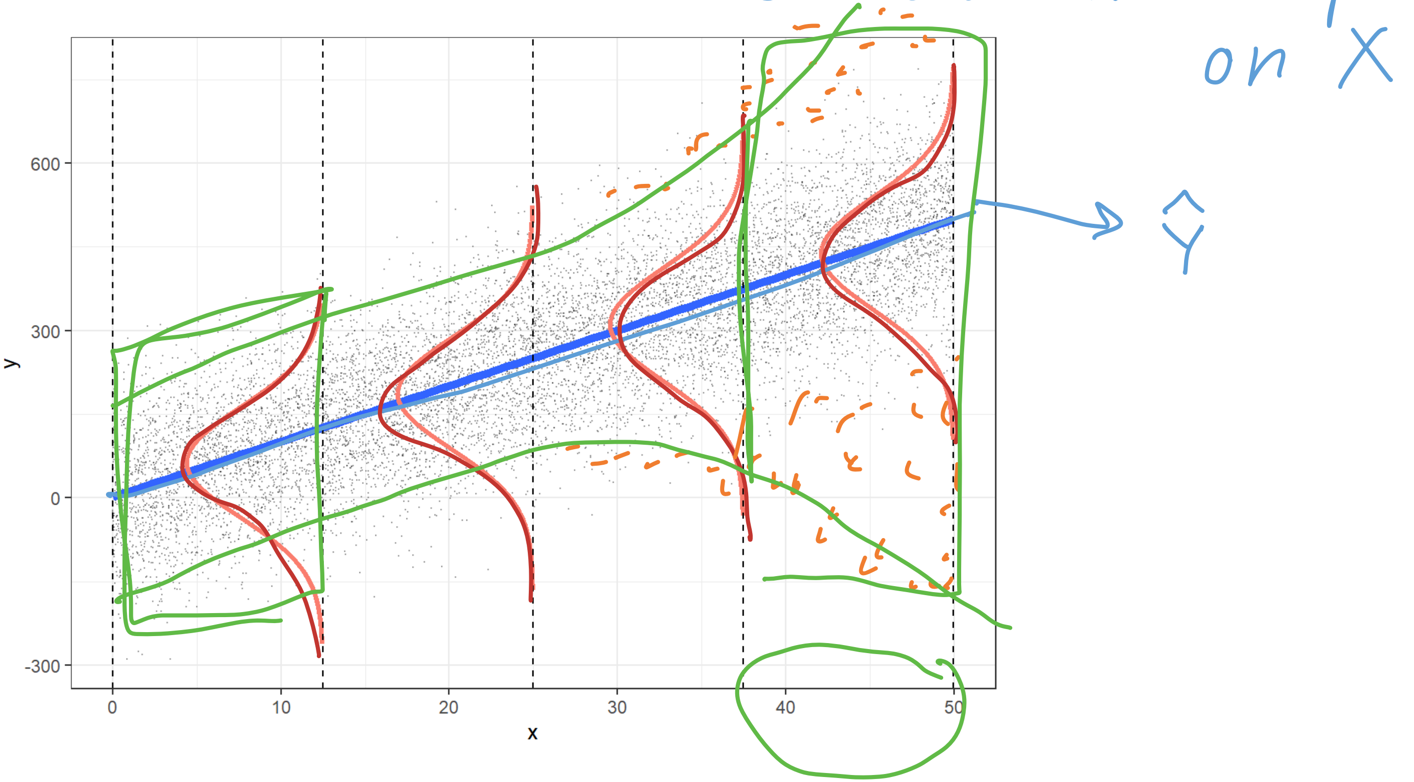


E: Equality of variance of the residuals

- The variance of Y given X ($\sigma^2_{Y|X}$), is the same for any X
    - We use just σ^2 to denote the common variance
- This is also called **homoscedasticity**

$\varepsilon_i \sim N(0, \sigma^2)$

σ^2 does not depend on X



Summary of LINE model assumptions

- Y values are independent (check study design!)

The distribution of Y given X is

- normal
- with mean $\mu_{y|x} = \beta_0 + \beta_1 \cdot X$
- and common variance σ^2 ✓

This means that the residuals are

- normal
- with mean = 0 $\quad Y_i - \hat{Y}$
- and common variance σ^2

In mathematical form:

$$Y_i|X \overset{\text{i.i.d.}}{\sim} N(\underbrace{\beta_0 + \beta_1 X}_{\mu_{y|x}}, \sigma^2)$$

$$\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$$

iid: independent & identically distributed

How do we determine if our model follows the LINE assumptions?

[L] **Linearity** of relationship between variables

Check if there is a linear relationship between the mean response (Y) and the explanatory variable (X)

X

[I] **Independence** of the Y values

Check that the observations are independent

via study design

[N] **Normality** of the Y's given X (residuals)

Check that the responses (at each level X) are normally distributed

- Usually measured through the residuals

[E] **Equality** of variance of the residuals (homoscedasticity)

Check that the variance (or standard deviation) of the responses is equal for all levels of X

- Usually measured through the residuals

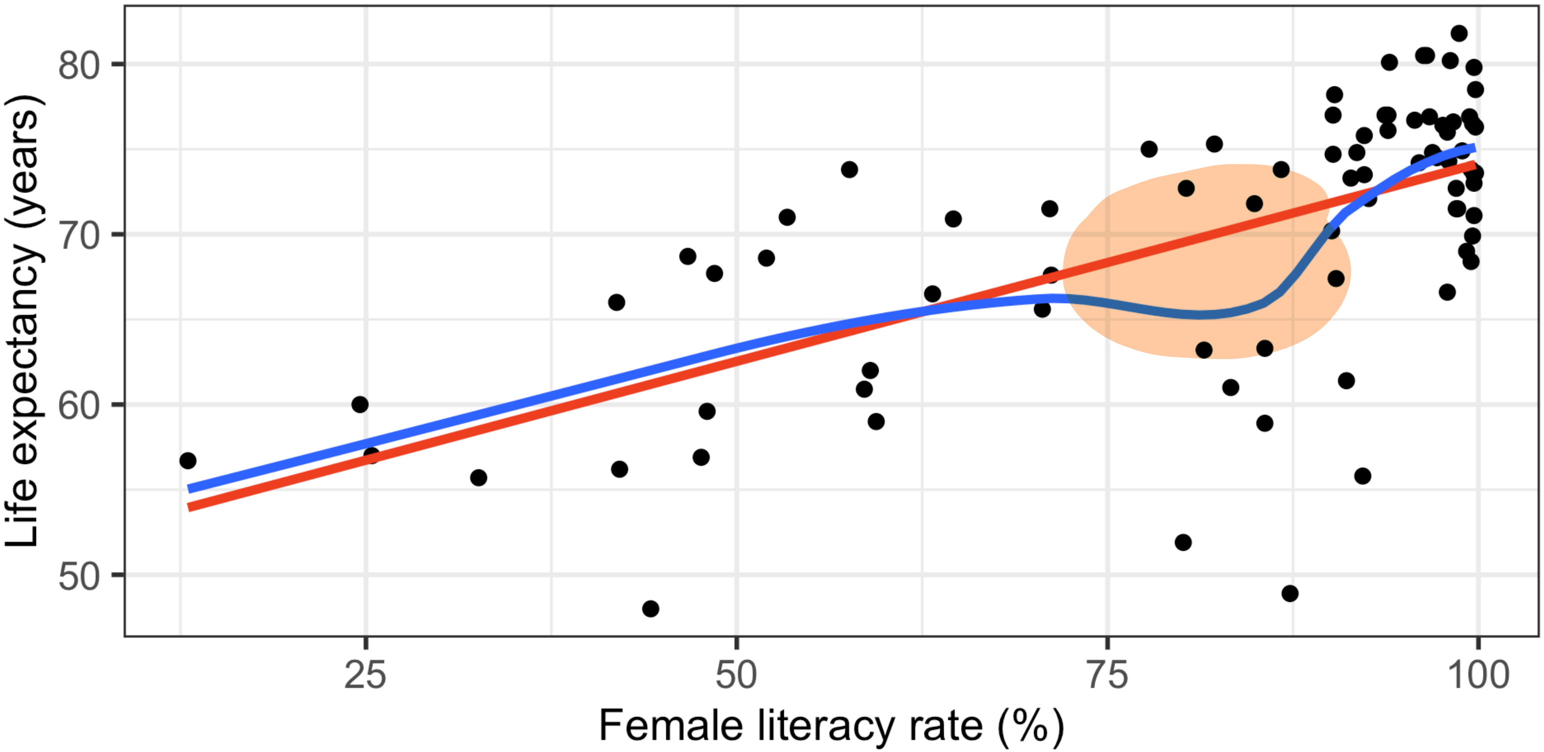
Learning Objectives

1. Describe the model assumptions made in linear regression using ordinary least squares
2. Determine if the relationship between our sampled X and Y is linear
3. Use QQ plots to determine if our fitted model holds the normality assumption
4. Use residual plots to determine if our fitted model holds the equality of variance assumption

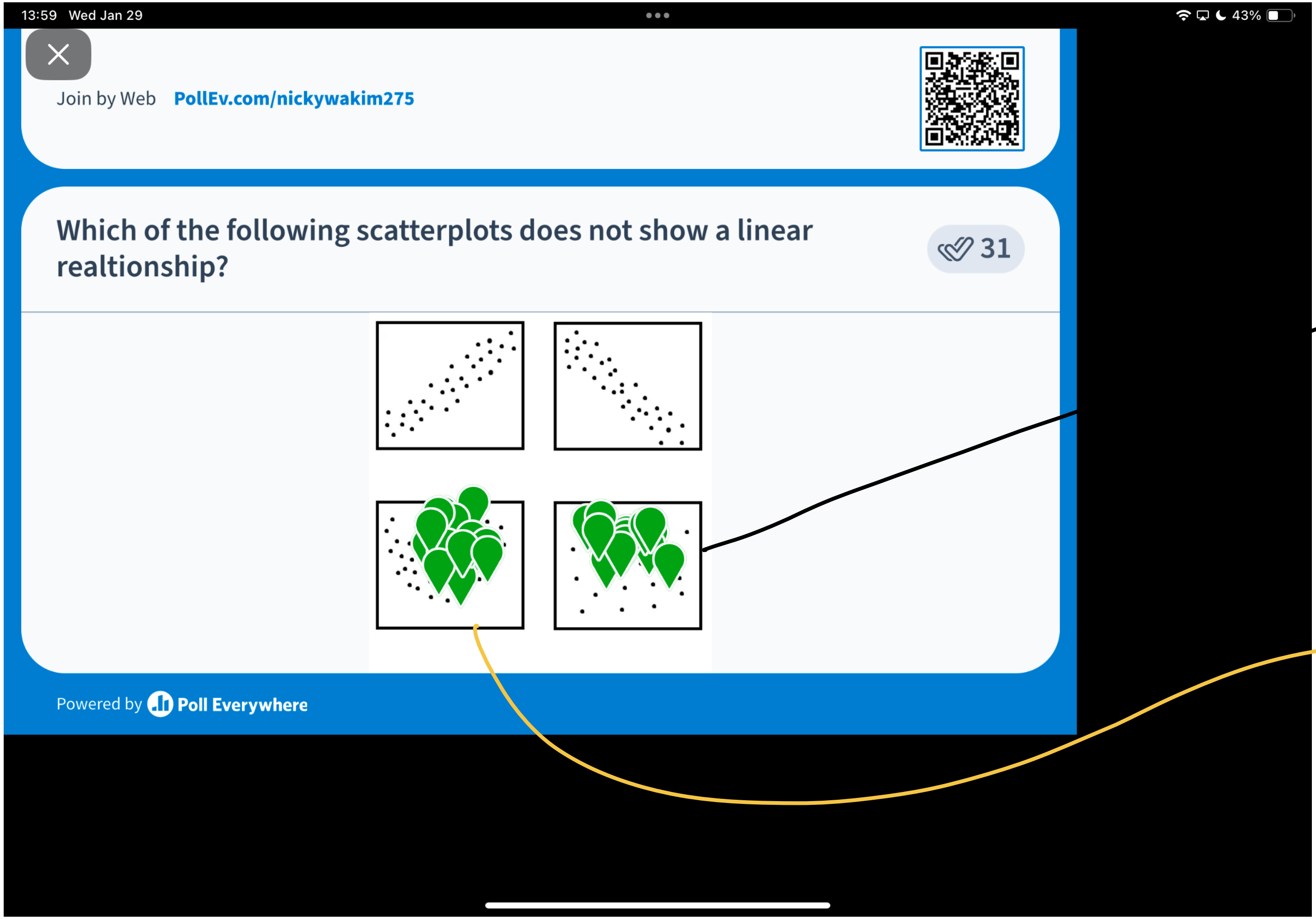
L: Linearity of relationship between variables

Is the association between the variables **linear**?

- **Diagnostic tool:** Scatterplot of X vs. Y



Poll Everywhere Question 2



13:59 Wed Jan 29

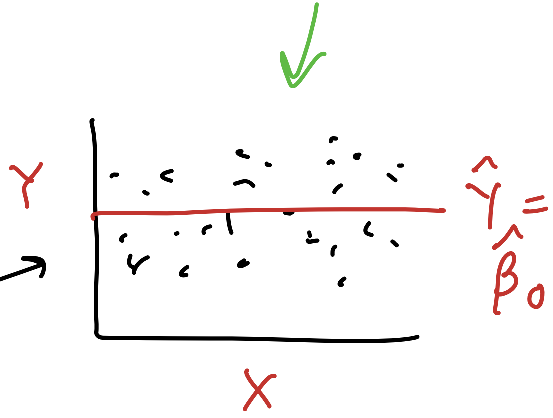
43%

Join by Web PollEv.com/nickywakim275

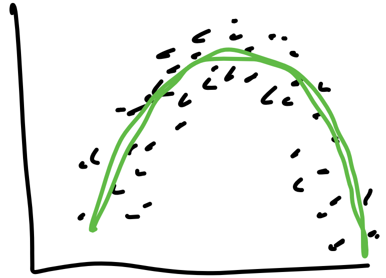
Which of the following scatterplots does not show a linear realtionship?

31

Powered by Poll Everywhere



$\hat{Y} = \hat{\beta}_0$



I: Independence of the residuals (Y values)

- **Are the data points independent of each other?**

- **Diagnostic tool:** reviewing the study *design* and not by inspecting the data

Learning Objectives

1. Describe the model assumptions made in linear regression using ordinary least squares
2. Determine if the relationship between our sampled X and Y is linear
3. Use QQ plots to determine if our fitted model holds the normality assumption
4. Use residual plots to determine if our fitted model holds the equality of variance assumption

N: Normality of the residuals

- We need to check if the errors/residuals (ϵ_i's) are normally distributed

- **Diagnostic tools:**
  - Distribution plots of residuals histogram
  - QQ plots of residuals

- Extra resource on how QQ plots are made

N: Extract model’s residuals in R

- First extract the residuals’ values from the model output using the `augment()` function from the `broom` package.
- Get a tibble with the orginal data, as well as the residuals and some other important values.

```r
model1 <- lm(LifeExpectancyYrs ~ FemaleLiteracyRate,
                 data = gapm)
aug1 <- augment(model1)

glimpse(aug1)
```

```
Rows: 80
Columns: 8
$ LifeExpectancyYrs  <dbl> 56.7, 76.7, 60.9, 76.9, 76.0, 73.8, 71.0, 76.9, 58.…
$ FemaleLiteracyRate <dbl> 13.0, 95.7, 58.6, 99.4, 97.9, 99.5, 53.4, 96.7, 85.…
$ .fitted            <dbl> 53.94643, 73.14897, 64.53453, 74.00809, 73.65980, 7…
$ .resid             <dbl> 2.7535654, 3.5510294, -3.6345319, 2.8919074, 2.3402…
$ .hat               <dbl> 0.13628996, 0.01768176, 0.02645854, 0.02077123, 0.0…
$ .sigma             <dbl> 6.172684, 6.168414, 6.167643, 6.172935, 6.176043, 6…
$ .cooksd            <dbl> 1.835891e-02, 3.062372e-03, 4.887448e-03, 2.400993e…
$ .std.resid         <dbl> 0.48238134, 0.58332052, -0.59972251, 0.47579667, 0.…
```

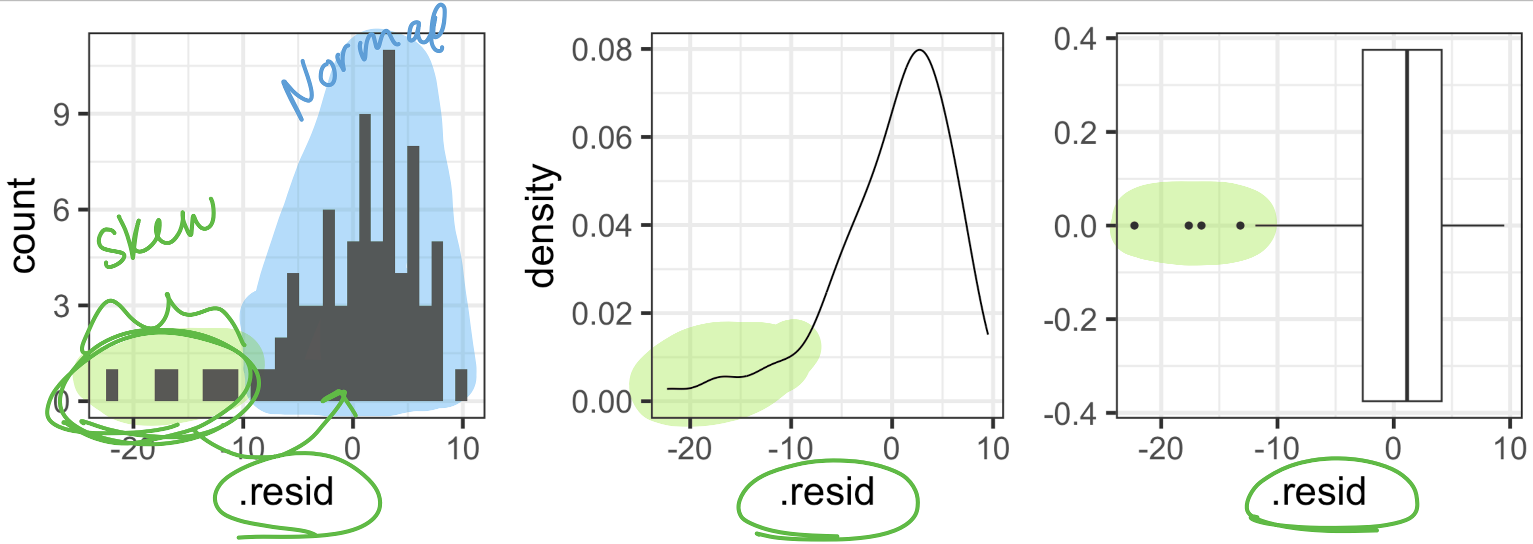
$\hat{Y}_i | X$

$\hat{\varepsilon}_i$

N: Check normality with distribution plots of residuals (1/2)

Note that below I save each figure as an object, and then combine them together in one row of output using `grid.arrange()` from the `gridExtra` package

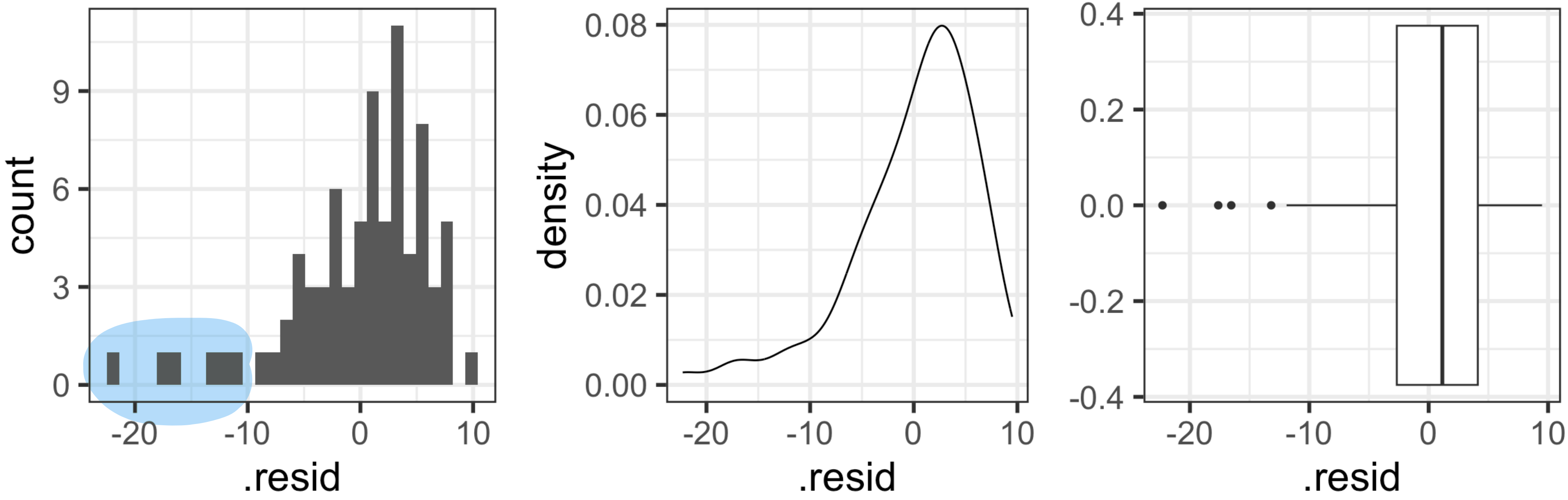
```
hist1 <- ggplot(aug1, aes(x = .resid)) + geom_histogram()

density1 <- ggplot(aug1, aes(x = .resid)) + geom_density()

box1 <- ggplot(aug1, aes(x = .resid)) + geom_boxplot()

grid.arrange(hist1, density1, box1, nrow = 1)
```



N: Check normality with distribution plots of residuals (2/2)

- So do these plots of the residuals look normal?

```
grid.arrange(hist1, density1, box1, nrow = 1)
```

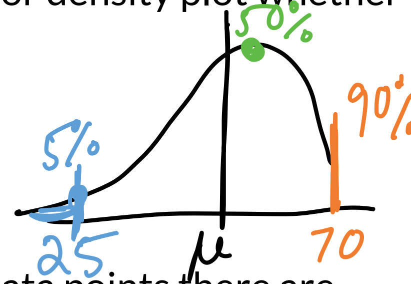


- My assessment: Looks like our residuals could be normal if we did not have those values around -20

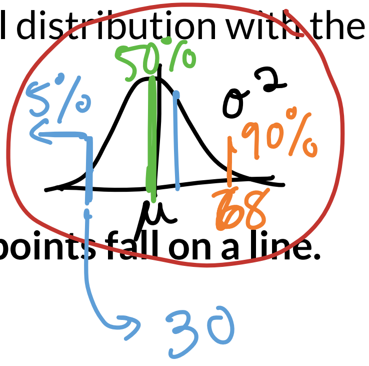
might consider more covariates in model

N: Normal QQ plots (QQ = quantile-quantile)

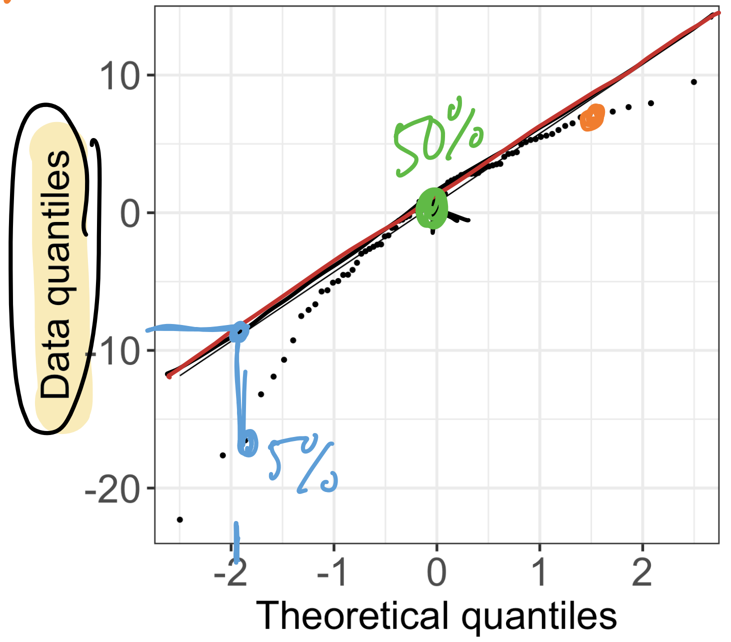
- It can be tricky to eyeball with a histogram or density plot whether the residuals are normal or not
- QQ plots are often used to help with this



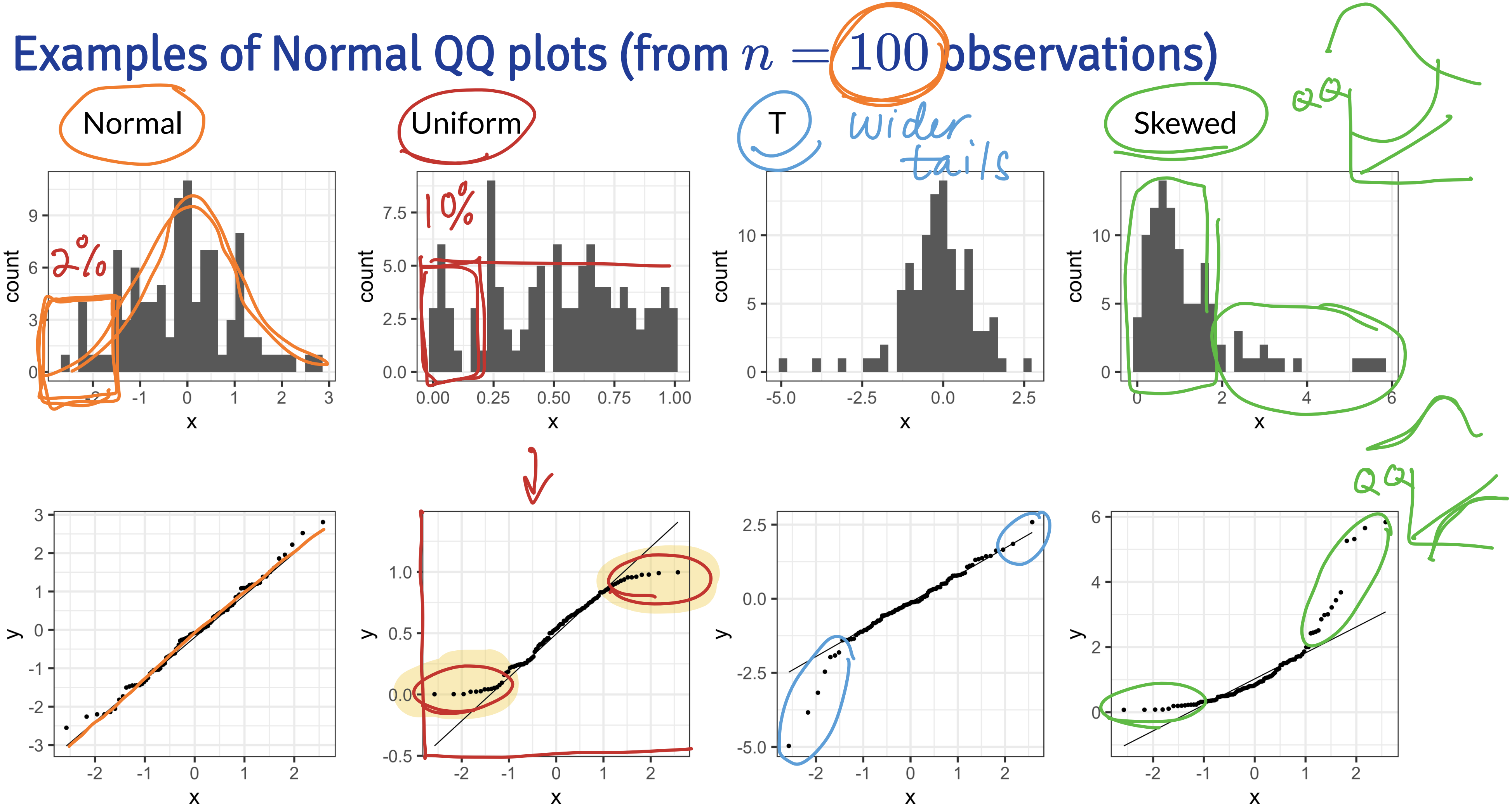
- *Vertical axis*: **data quantiles**
  - data points are sorted in order and
  - assigned quantiles based on how many data points there are
- *Horizontal axis*: **theoretical quantiles**
  - mean and standard deviation (SD) calculated from the data points
  - theoretical quantiles are calculated for each point, assuming the data are modeled by a normal distribution with the mean and SD of the data



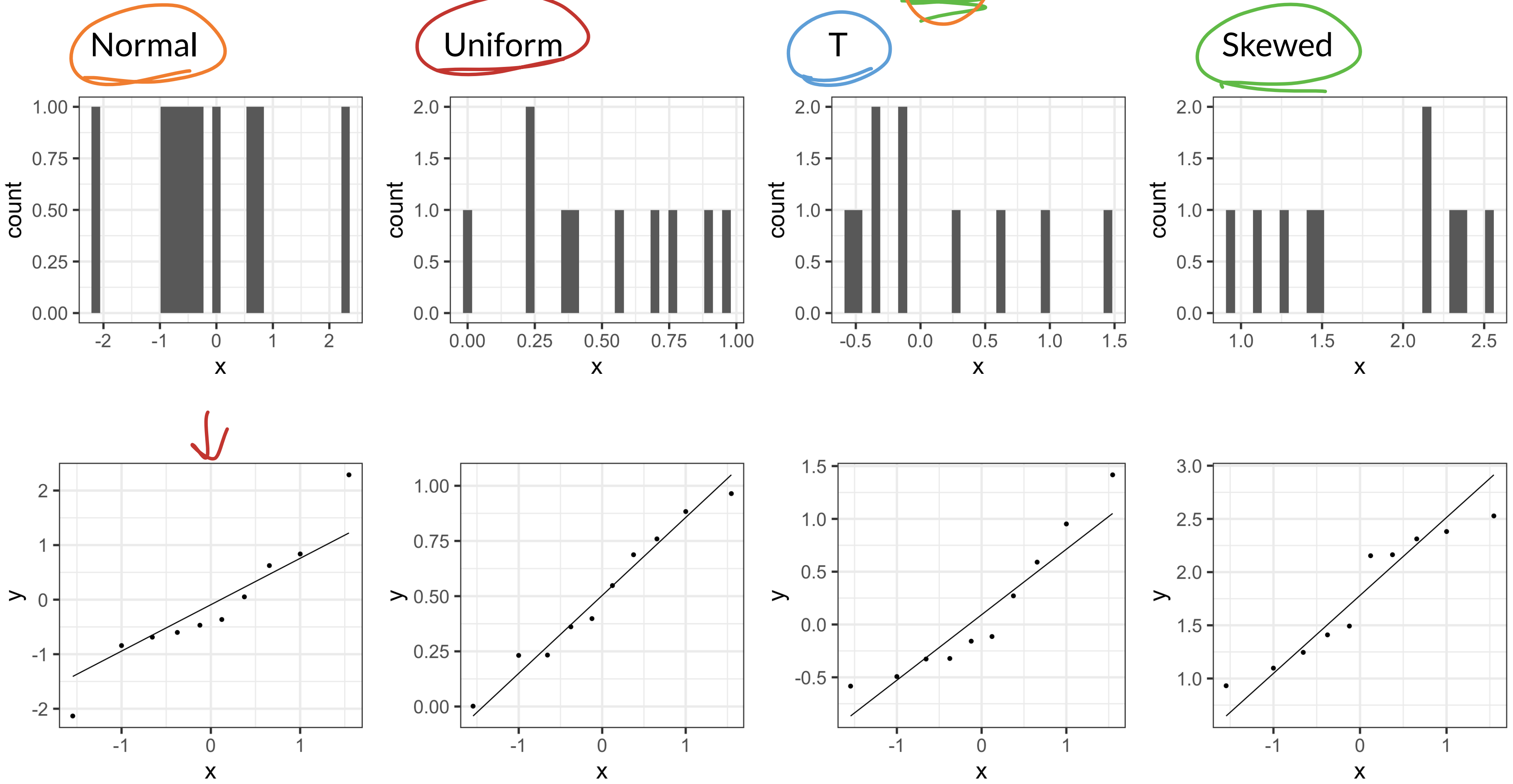
- **Data are approximately normal if points fall on a line.**



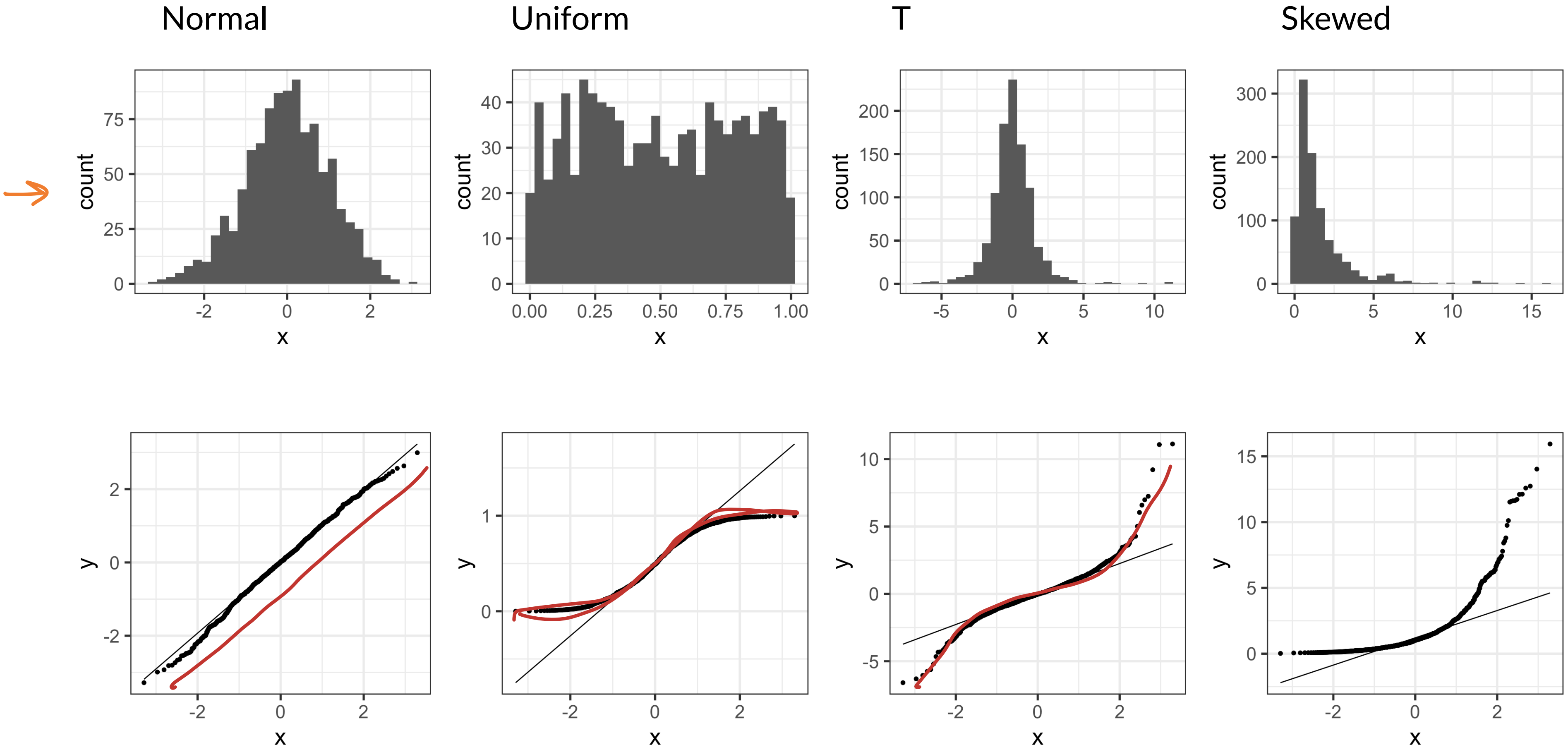
N: Examples of Normal QQ plots (from $n = 100$ observations)



N: Examples of Normal QQ plots (from $n = 10$ observations)



N: Examples of Normal QQ plots (from $n = \boxed{1000}$ observations)



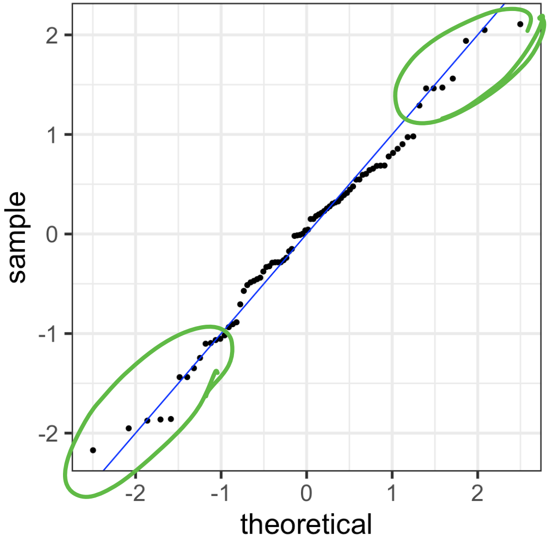
N: We can compare the QQ plots: model vs. theoretical

- QQ plot from Life Expectancy vs. Female Literacy Rate Regression

```
ggplot(aug1,
       aes(sample = .resid)) +
  stat_qq() +
  stat_qq_line()
```

- Simulated QQ plot of Normal Residuals with $n = 80$

```
ggplot() +
  stat_qq(aes(
    sample = rnorm(80))) +
  geom_abline(
    intercept = 0, slope = 1,
    color = "blue")
```



N: Shapiro-Wilk Test of Normality

- Goodness-of-fit test for the normal distribution: Is there evidence that our residuals are from a normal distribution?
- Hypothesis test:

$$
\begin{aligned}
H_0 &: \text{data are from a normally distributed population} \\
H_A &: \text{data are NOT from a normally distributed population}
\end{aligned}
$$

```r
shapiro.test(aug1$.resid)
```

```

	Shapiro-Wilk normality test

data:  aug1$.resid
W = 0.90575, p-value = 2.148e-05
```

small

test stat

Conclusion

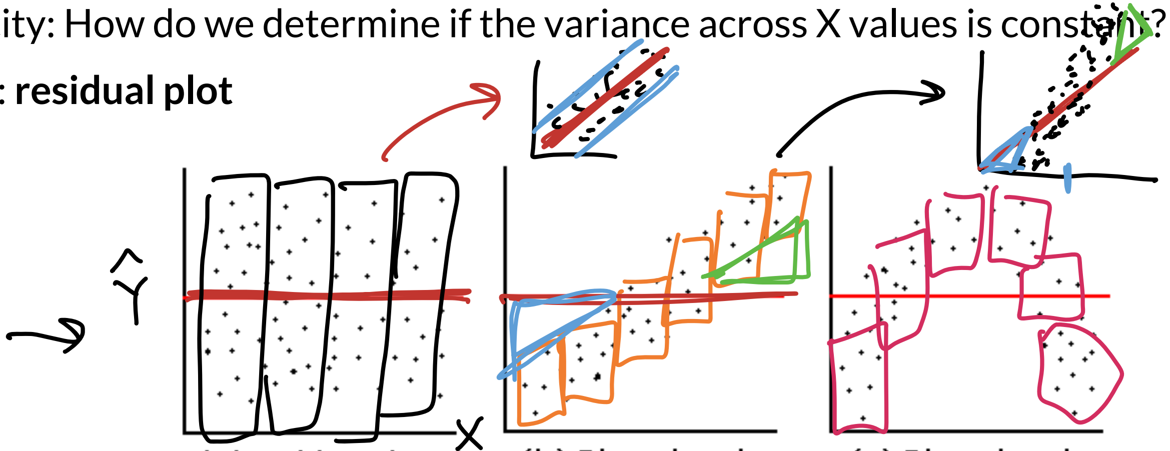
Reject the null. Data are not from a normal distribution.

Learning Objectives

1. Describe the model assumptions made in linear regression using ordinary least squares
2. Determine if the relationship between our sampled X and Y is linear
3. Use QQ plots to determine if our fitted model holds the normality assumption
4. Use residual plots to determine if our fitted model holds the equality of variance assumption

E: Equality of variance of the residuals

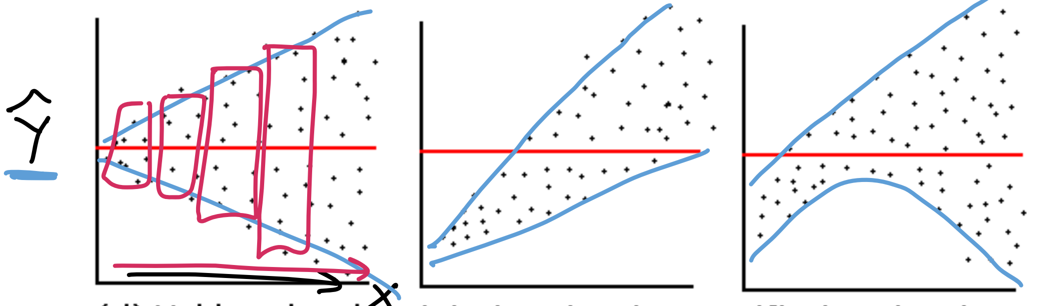
- Homoscedasticity: How do we determine if the variance across X values is constant?
- Diagnostic tool: **residual plot**



(a) Unbiased and Homoscedastic

(b) Biased and Homoscedastic

(c) Biased and Homoscedastic



(d) Unbiased and Heteroscedastic

(e) Biased and Heteroscedastic

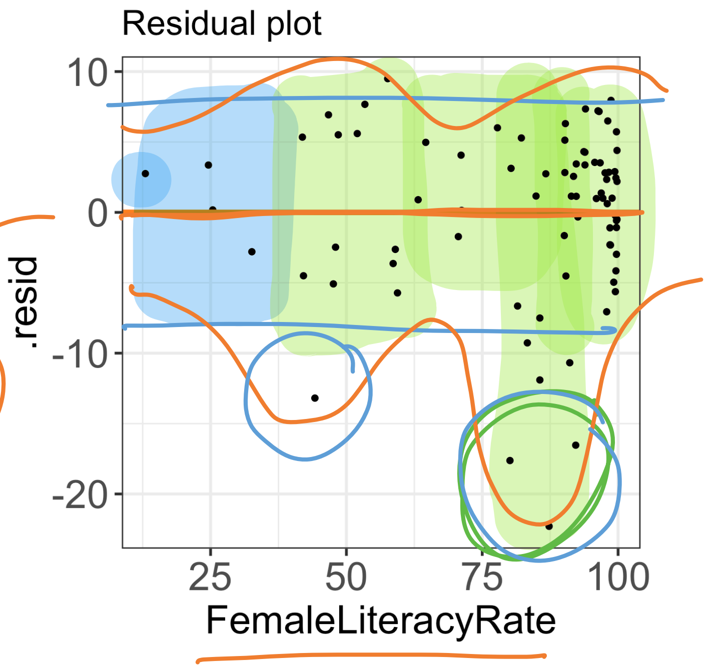
(f) Biased and Heteroscedastic

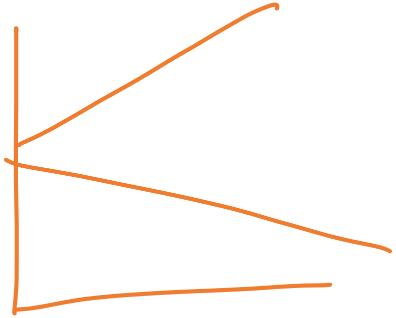
E: Creating a residual plot

- x = explanatory variable from regression model
  - (or the fitted values for a multiple regression)
- y = residuals from regression model

```r
ggplot(aug1,
          aes(x = FemaleLiteracyRate,
              y = .resid)) +
  geom_point(size = 2) +
  geom_abline( intercept = 0, slope = 0,
    size = 2, color = "#FF8021") +
  labs(title = "Residual plot") +
  theme(axis.title = element_text(size = 30),
          axis.text = element_text(size = 30))
```

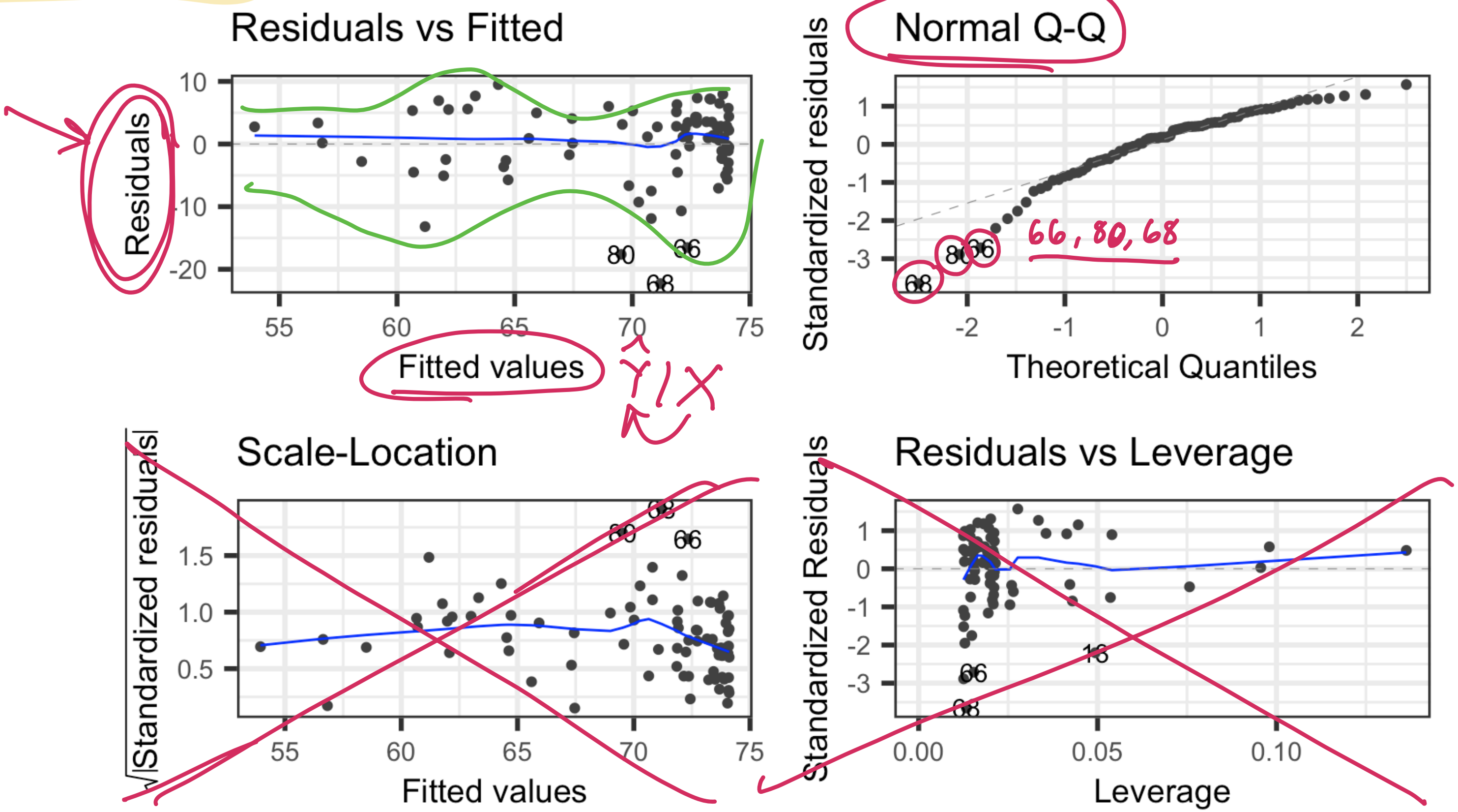
$\varepsilon_i \sim N(0, \sigma^2)$





`autoplot()` can be a helpful tool

```r
library(ggfortify)
autoplot(model1) + theme(text=element_text(size=14))
```



$\hat{Y} / X$

66, 80, 68

Summary of the assumptions and their diagnostic tool

*check if test for eq of var

Assumption	What needs to hold?	Diagnostic tool	
Linearity	• Relationship between X and Y is linear	• Scatterplot of Y vs. X	yes
Independence	• Observations are independent from each other	• Study design	yes
Normality	• Residuals (and thus $Y\|X$) are normally distributed	• QQ plot of residuals • Distribution of residuals	not really
Equality of variance	• Variance of residuals (and thus $Y\|X$) is same across X values (homoscedasticity)	• Residual plot	kinda

next steps:
- multiple linear regression - see if transformations
- check flagged countries

We didn’t really go over our options when these assumptions do not hold

- We will consider this more once we get into multiple linear regression
- For now, with SLR, when assumptions do not hold, I conclude we need to add more variables in the model

- Another note: I did not make these plots very presentable
  - Axes were left with whatever names were given to them
  - These plots are usually just for us!
  - Not really something that you include in a formal report