SLR: Model Evaluation and Diagnostics

Nicky Wakim

2023-01-29

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 the model that 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

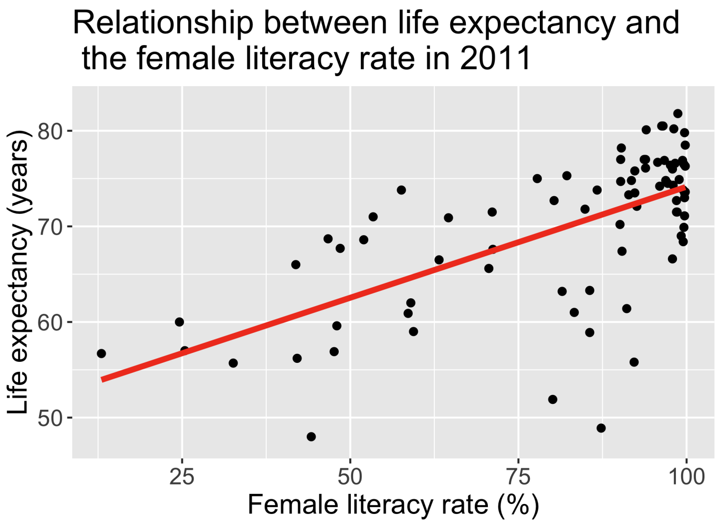
pop

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

term	estimate	std.error	statistic	p.value
(Intercept)	50.93	2.66	19.14	0.00
female_literacy_rate_2011	0.23	0.03	7.38	0.00

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

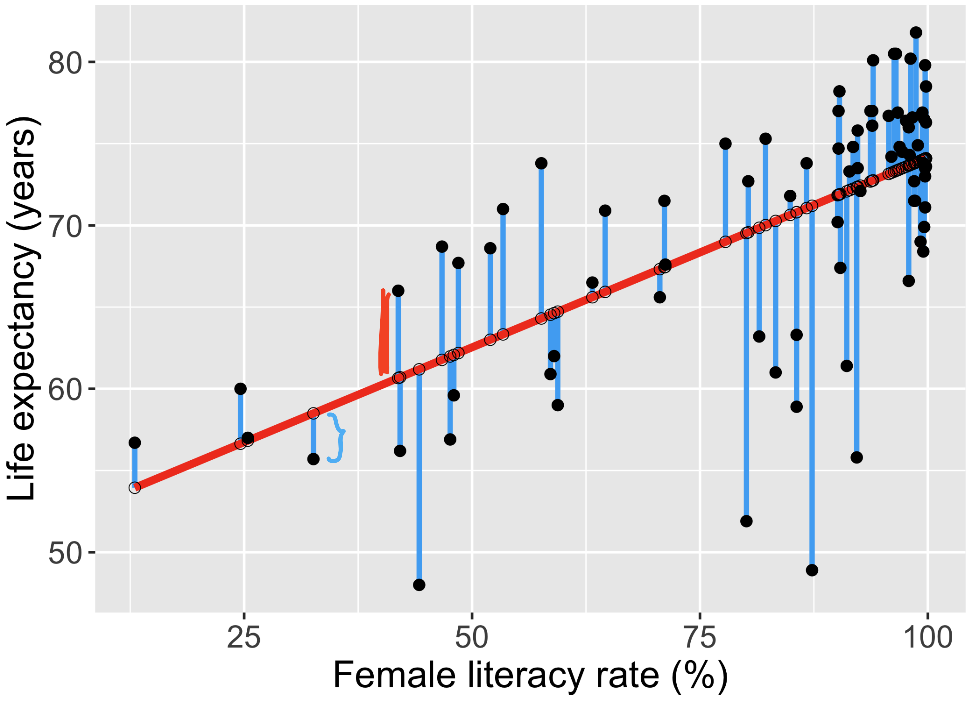
$$\widehat{\text{life expectancy}} = 50.9 + 0.232 \cdot \text{female literacy rate}$$



Our residuals will help us a lot in our diagnostics!

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

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



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

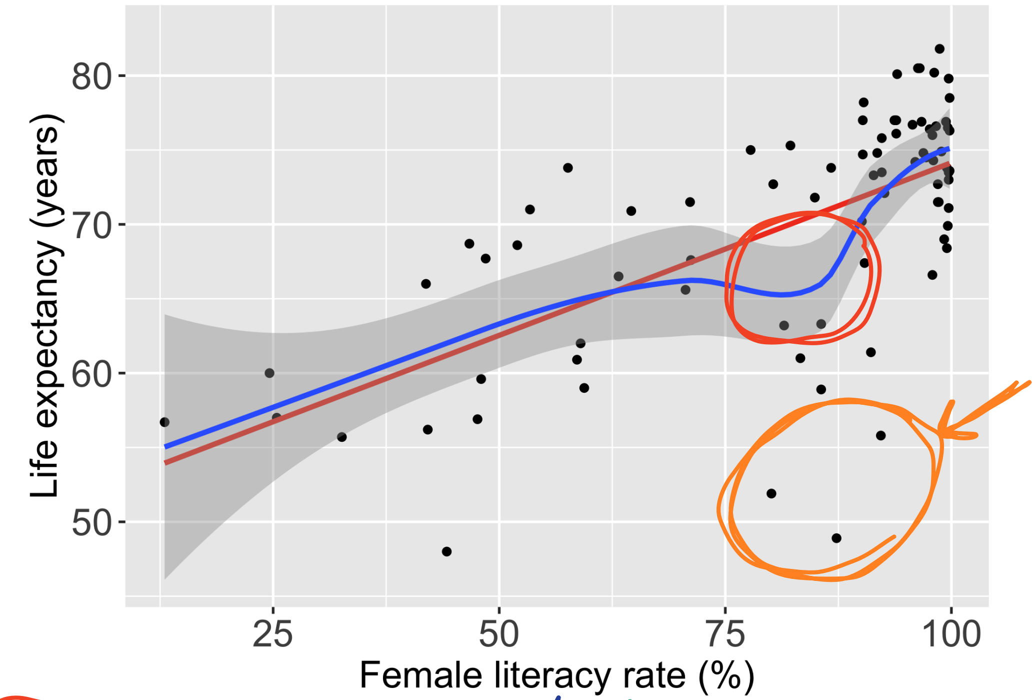
These are the model assumptions made in ordinary least squares:

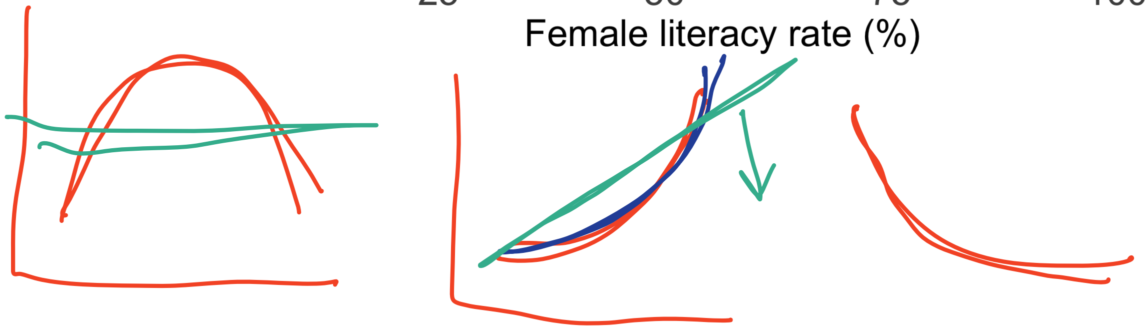
- **[L] Linearity** of relationship between variables
- **[I] Independence** of the Y values
- **[N] Normality** of the Y's given X (residuals)
- **[E] Equality** of variance of the residuals (homoscedasticity)

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

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





I: Independence of observations

$\underline{\varepsilon} \sim \underline{N}(0, \sigma^2)$

iid → independent & identically dist.

dist. normally

population

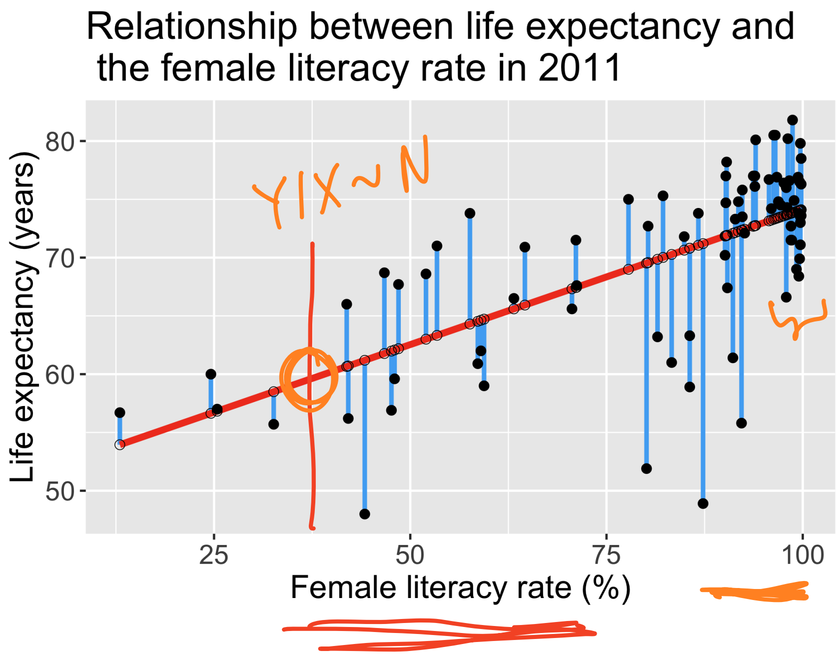
- The Y-values are statistically independent of one another
- Examples of when they are *not* independent, include
  - repeated measures (such as baseline, 3 months, 6 months)
  - data from clusters, such as different hospitals or families
- This condition is checked by reviewing the study *design* and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

N: Normality

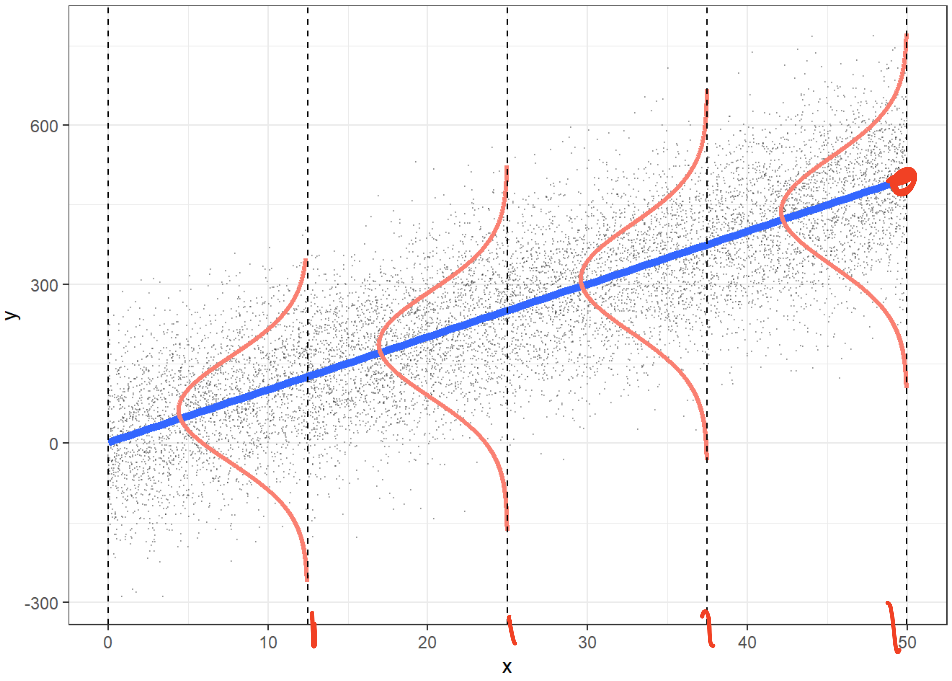
- For any fixed value of X, Y has normal distribution.
  - Note: This is not about Y alone, but $Y|X$
- Equivalently, the measurement (random) errors ϵ_i 's normally distributed
  - This is more often what we check



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**

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



population:

$\underbrace{Y|X}_{\text{Y given X}} \sim \underbrace{N}_{\text{normal}}(\overset{\text{mean}}{\beta_0 + \beta_1 X}, \underset{\text{var}}{\sigma^2})$

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
- and common variance σ^2

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)

[I] Independence of the Y values

Check that the observations are independent

[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

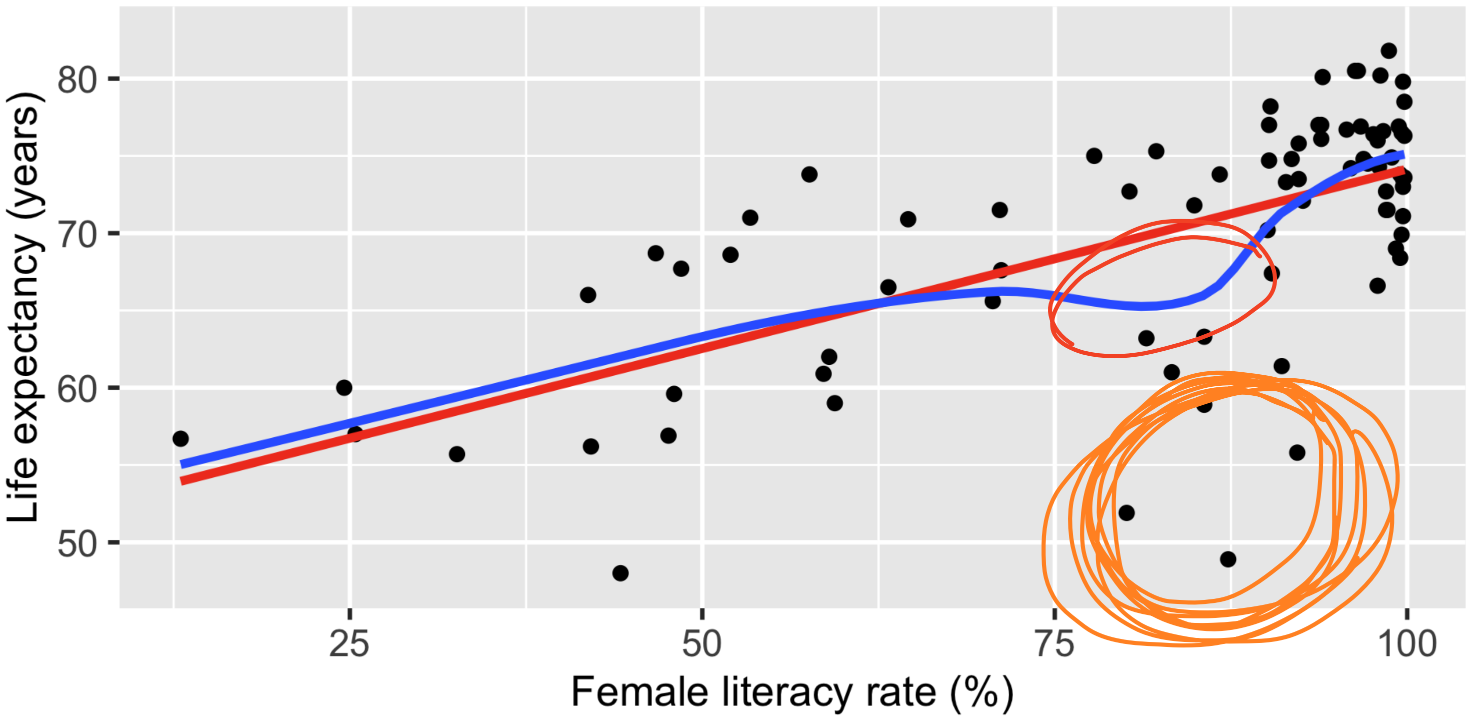
3. Use QQ plots to determine if our fitted model holds the normality assumption

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

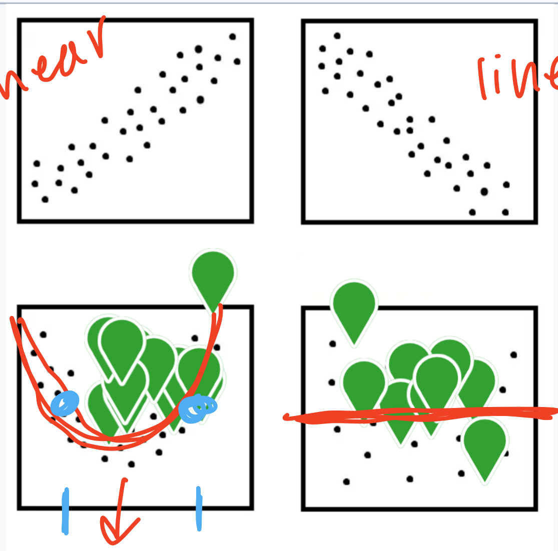
Activity is now locked.
Responses are not accepted at this time.

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

linear association

linear

linear



non-linear

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
  - 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.

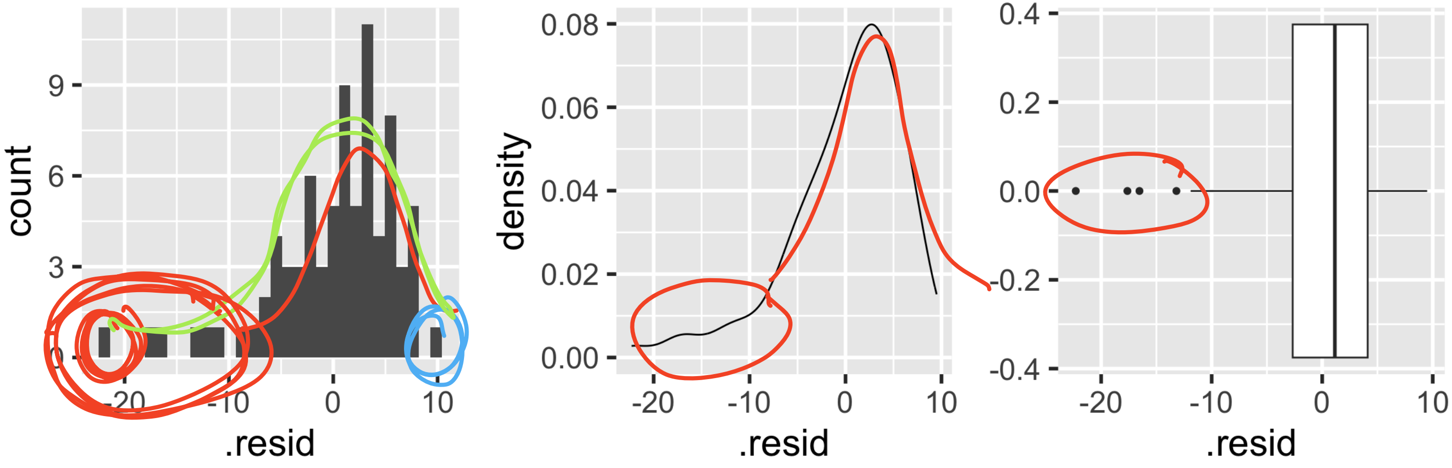
```
model1 <- lm(life_expectancy_years_2011 ~ female_literacy_rate_2011,
                 data = gapm)
aug1 <- augment(model1)

glimpse(aug1)
```

```
Rows: 80
Columns: 9
$ .rownames                 <chr> "1", "2", "5", "6", "7", "8", "14", "22", "…
$ life_expectancy_years_2011 <dbl> 56.7, 76.7, 60.9, 76.9, 76.0, 73.8, 71.0, 7…
$ female_literacy_rate_2011  <dbl> 13.0, 95.7, 58.6, 99.4, 97.9, 99.5, 53.4, 9…
$ .fitted                   <dbl> 53.94643, 73.14897, 64.53453, 74.00809, 73.…
$ .resid                    <dbl> 2.7535654, 3.5510294, -3.6345319, 2.8919074…
$ .hat                      <dbl> 0.13628996, 0.01768176, 0.02645854, 0.02077…
$ .sigma                    <dbl> 6.172684, 6.168414, 6.167643, 6.172935, 6.1…
$ .cooksd                   <dbl> 1.835891e-02, 3.062372e-03, 4.887448e-03, 2…
```

N: Check normality with "usual" distribution plots

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

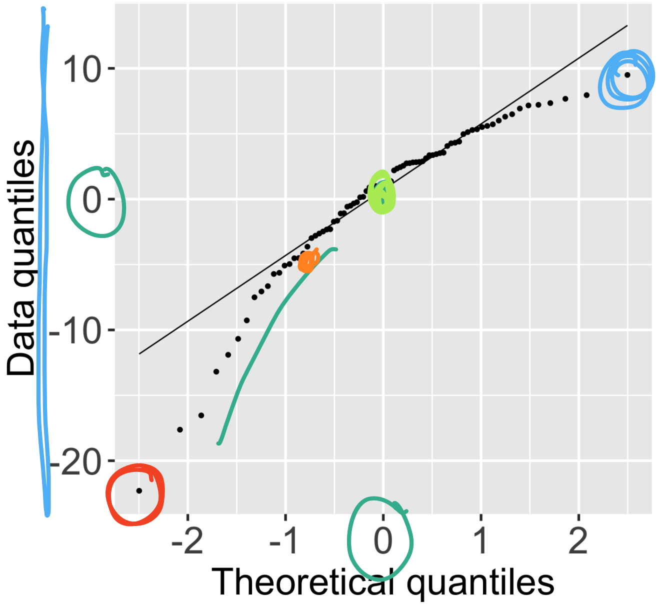
```
hist1 <- ggplot(aug1, aes(x = .resid)) + geom_histogram()

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

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

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



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

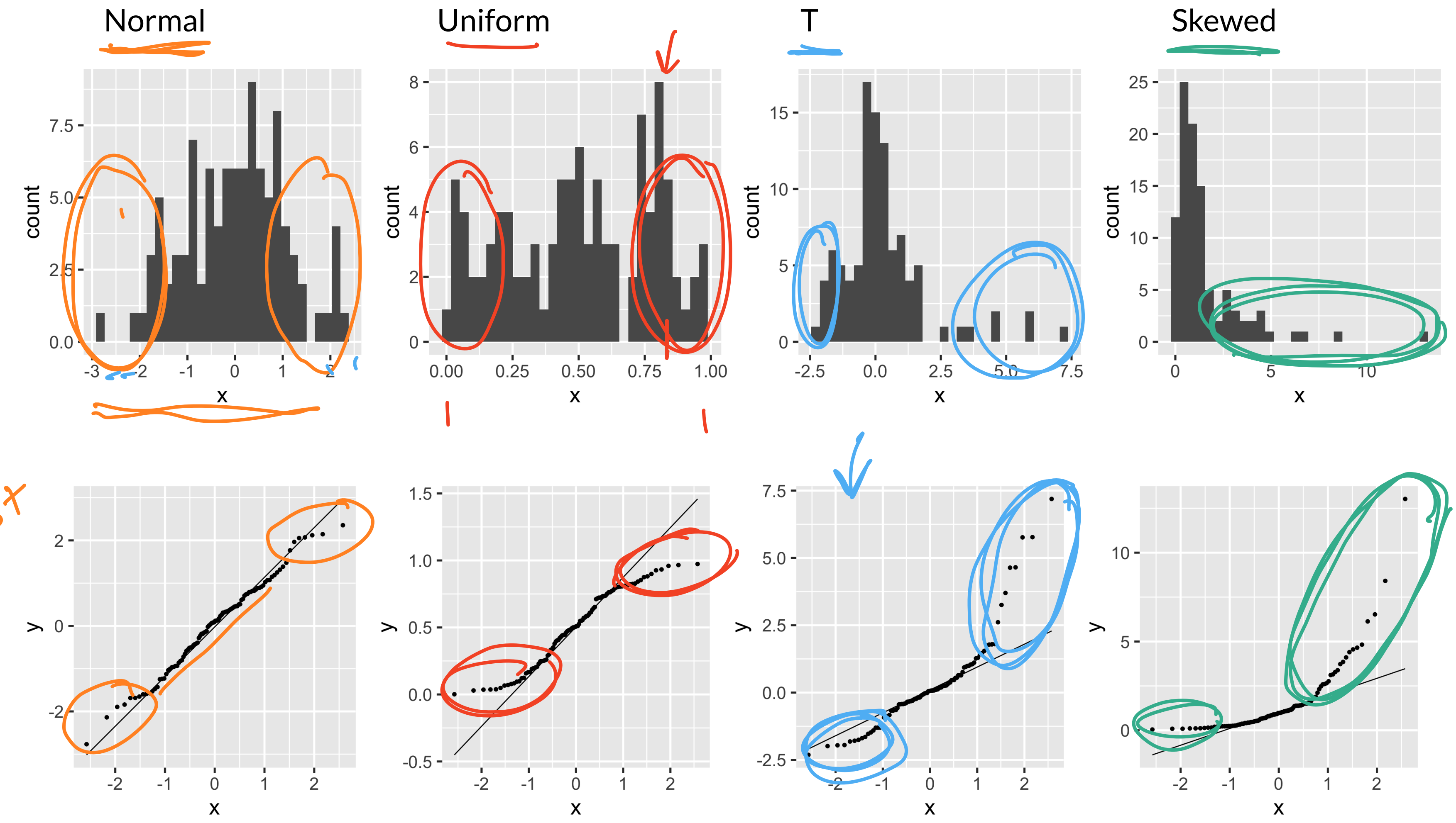
- It can be tricky to eyeball with a histogram or density plot whether the residuals are normal or not
- QQ plots are often used to help with this

- *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.**

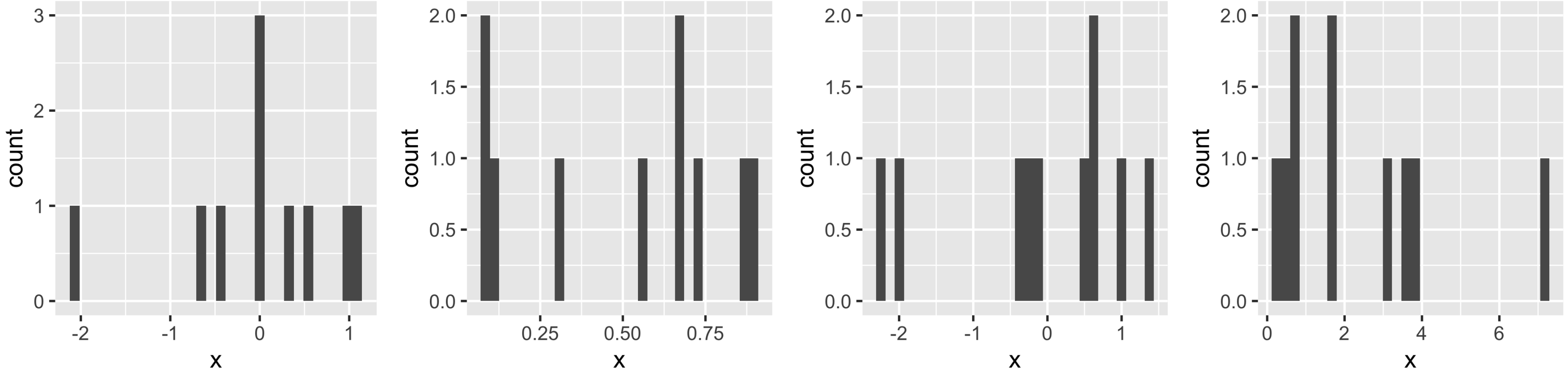
N: Examples of Normal QQ plots (from $n = 100$ observations)



QQ plot

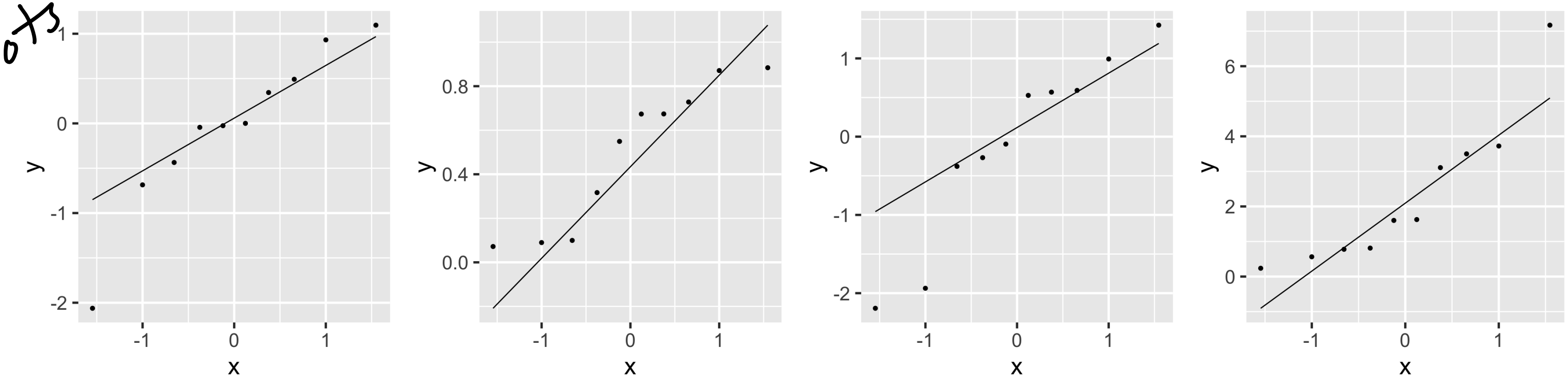
N: Examples of Normal QQ plots (from $n = 10$ observations)

Normal Uniform T Skewed

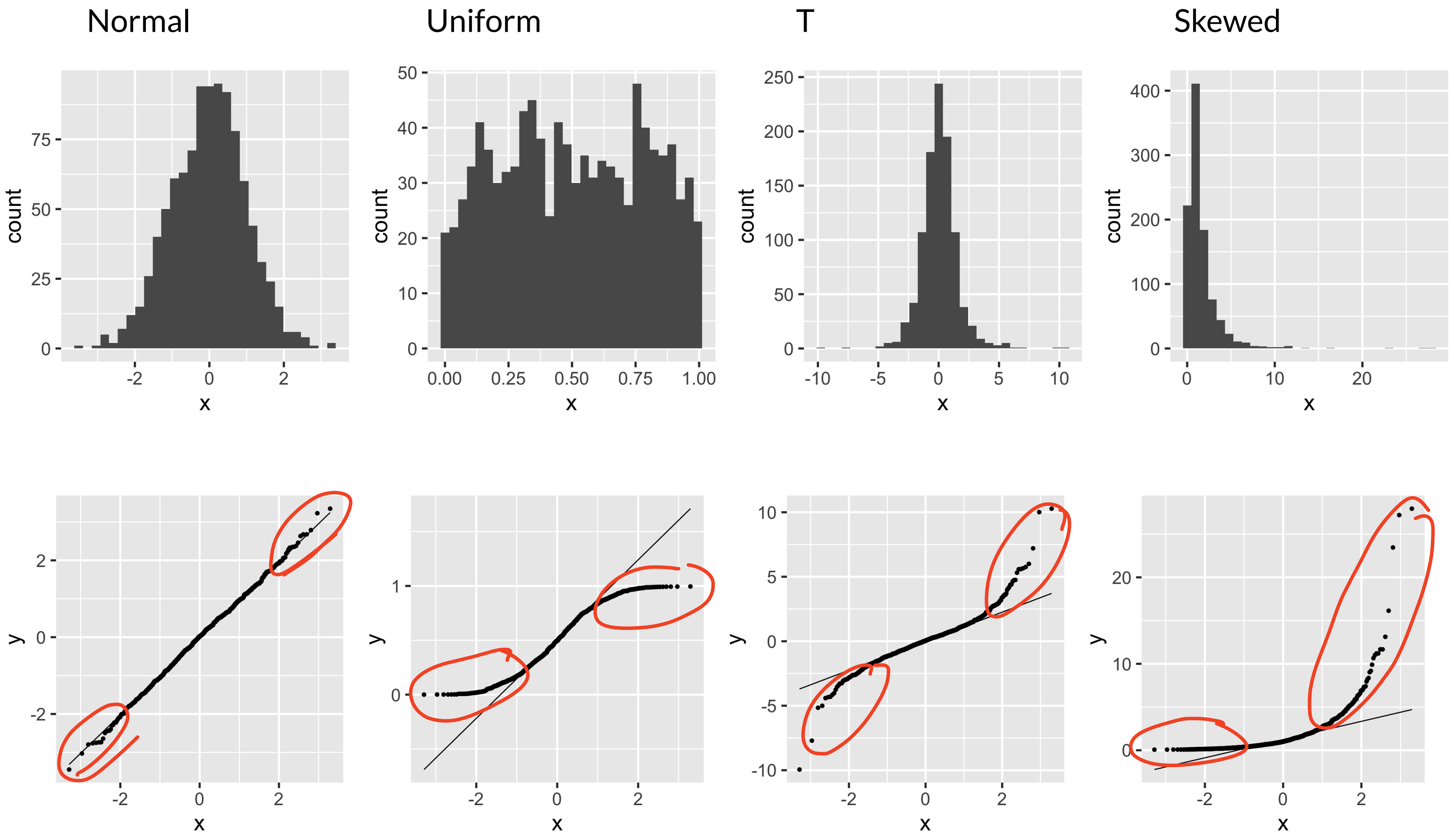


all look about the same

QQ Plots



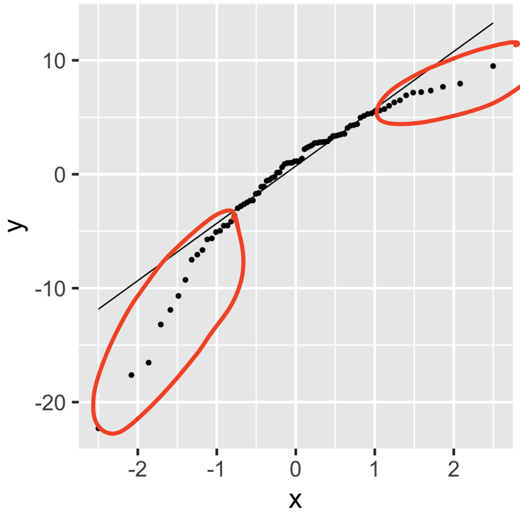
N: Examples of Normal QQ plots (from $n = 1000$ observations)



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

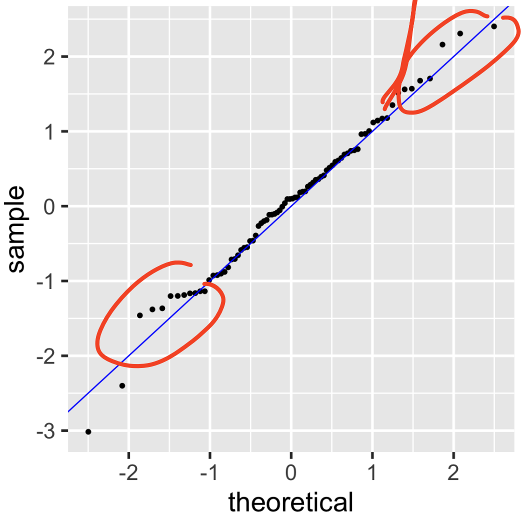
- Residuals from Life Expectancy vs. Female Literacy Rate Regression

```
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:

$$H_0 : \text{data are from a normally distributed population}$$
$$H_1 : \text{data are NOT from a normally distributed population}$$

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

```

	Shapiro-Wilk normality test

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

Conclusion

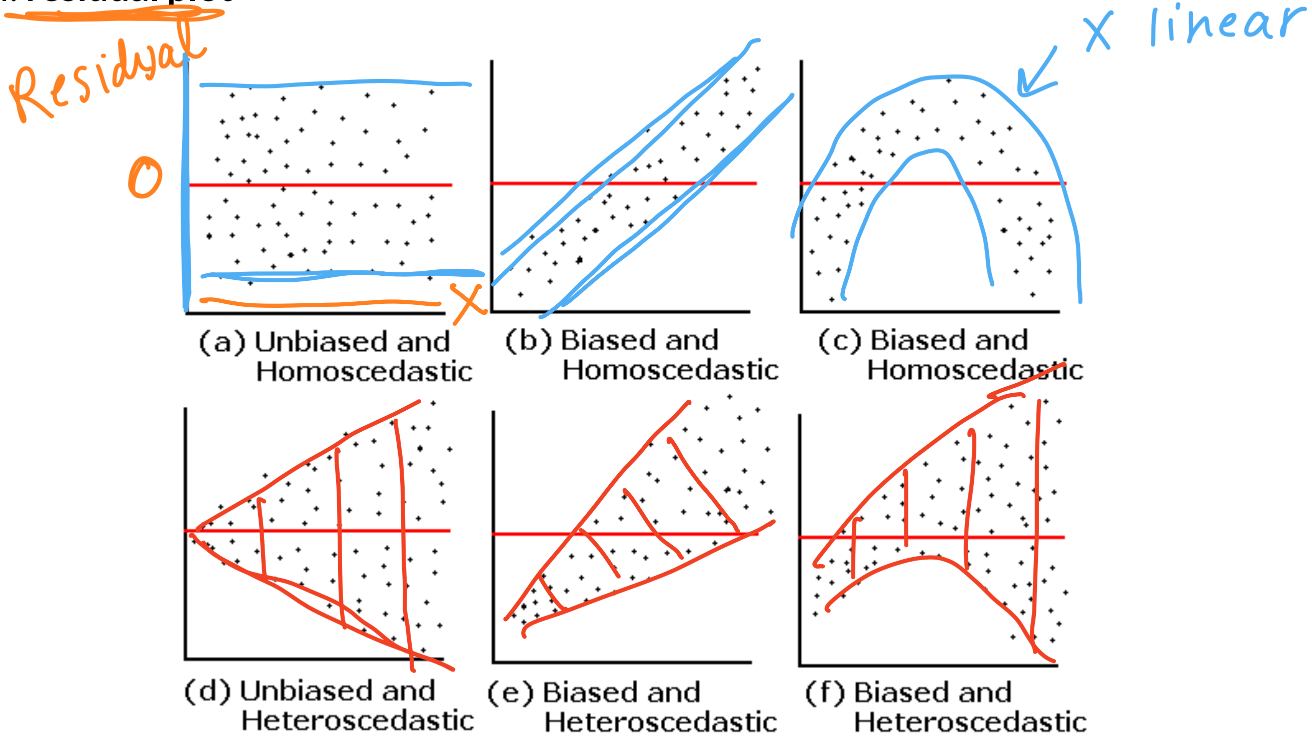
Reject the null. Data are not from a normal distribution.

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

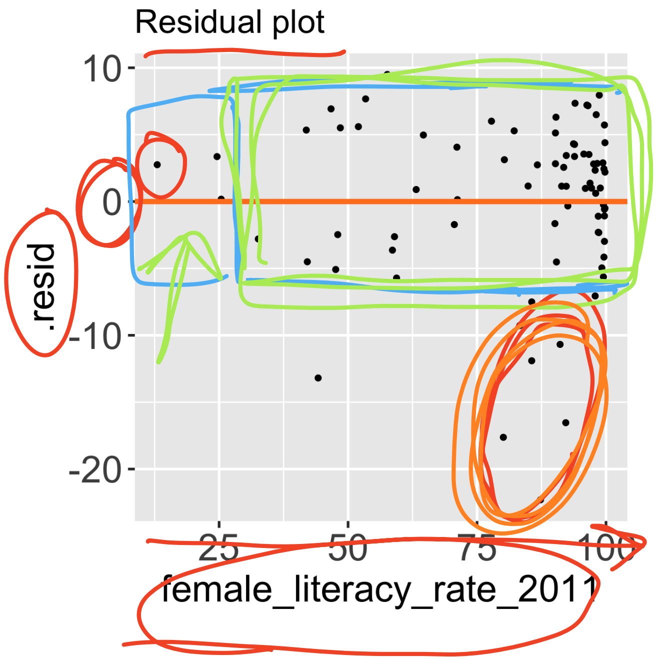
- Homoscedasticity: How do we determine if the variance across X values is constant?
- Diagnostic tool: **residual plot**



E: Creating a residual plot

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

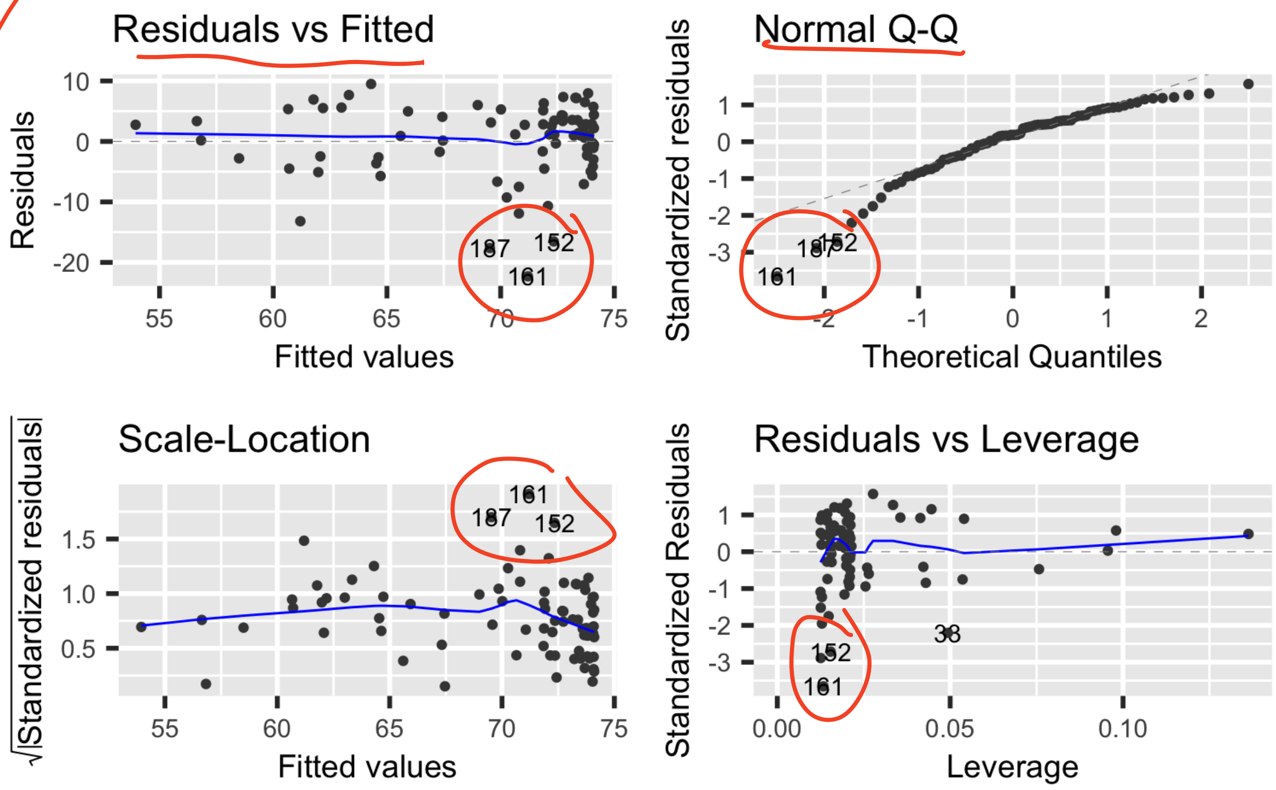
```
ggplot(aug1,
       aes(x = female_literacy_rate_2011,
           y = .resid)) +
  geom_point(size = 2) +
  geom_abline( intercept = 0, slope = 0,
    size = 2, color = "#FF8021") +
  labs(title = "Residual plot") +
  theme(axis.title = element_text(size = 30),
        axis.text = element_text(size = 30))
```



`autoplot()` can be a helpful tool

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

lm()



Summary of the assumptions and their diagnostic tool

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