09 - Simple Linear Regression

HCI/PSYCH 522
Iowa State University

February 17, 2022

Overview

- Simple linear regression
  - Dependent variable
  - Independent variable
  - Continuous independent variable
- Assumptions
  - Linearity
  - Normality
  - Constant Variance
  - Independence

Simple linear regression

Dependent variable

Definition

The distribution of the dependent variable depends on the values of the independent variables.

Dependent variable examples:

- Gold per minute
- Time to register
- Satisfaction

Independent variable

Definition

The independent variable affects the distribution of the dependent variable.

Independent variable examples:

- Mouse sensitivity
- Availability of a chatbot
- App being used

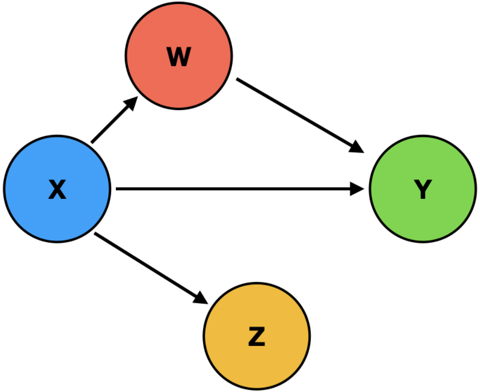
Synonyms

Terminology (all of these are [basically] equivalent):

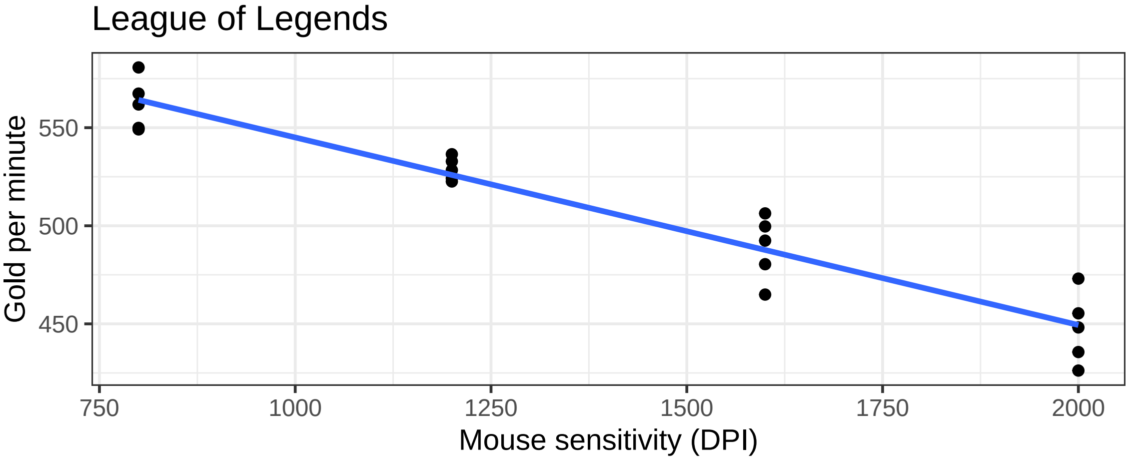
dependent	independent
response	independent
outcome	covariate
endogenous	exogenous

Independent-dependent variable

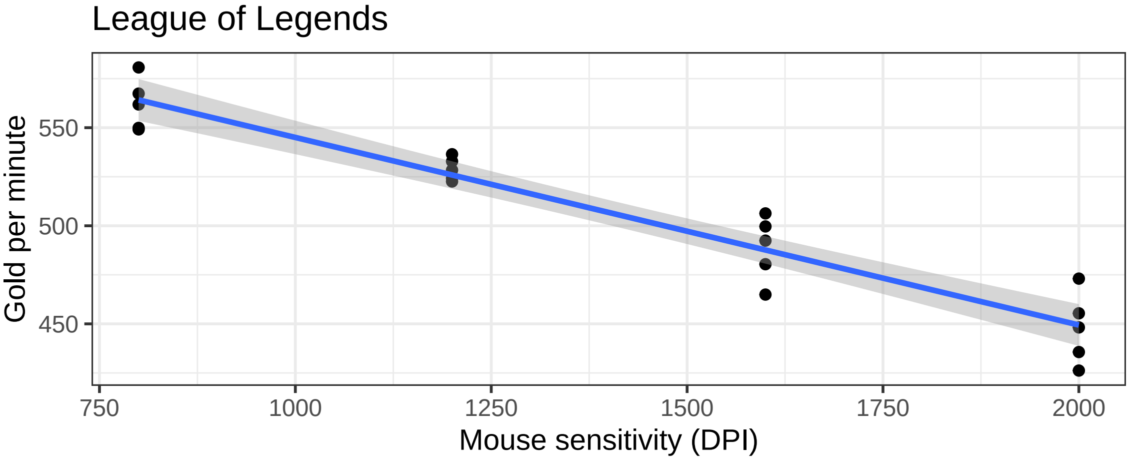
https://towardsdatascience.com/causal-inference-962ae97cefda



Continuous independent variable



Continuous independent variable



Simpe linear regression

The simple linear regression model is

$$Y_i \stackrel{ind}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2)$$

where Y_i and X_i are the dependent and independent variable, respectively, for individual i.

Alternatively

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \qquad \epsilon_i \stackrel{ind}{\sim} N(0, \sigma^2).$$

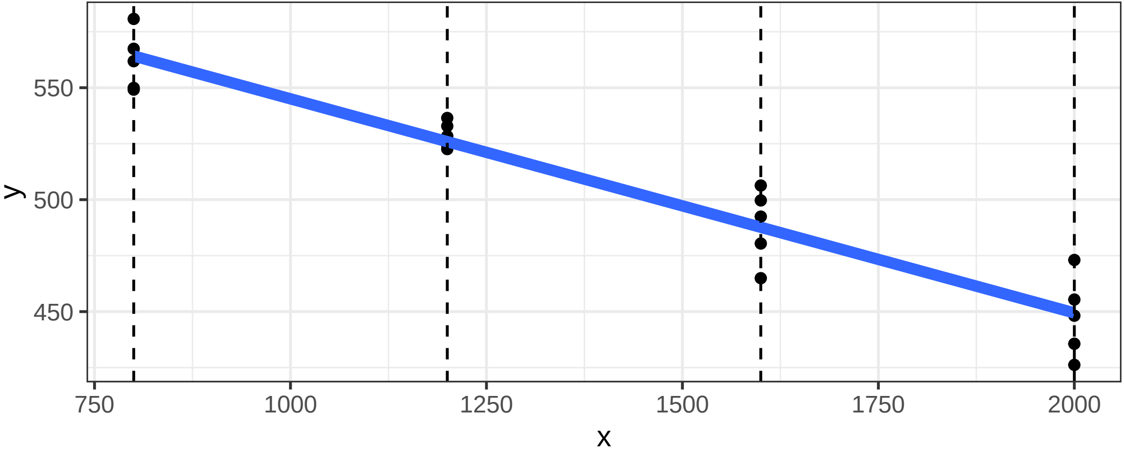
Importantly

$$E[Y_i|X_i] = \beta_0 + \beta_1 X_i$$

and

$$Var[Y_i|X_i] = \sigma^2.$$

Visualize variability



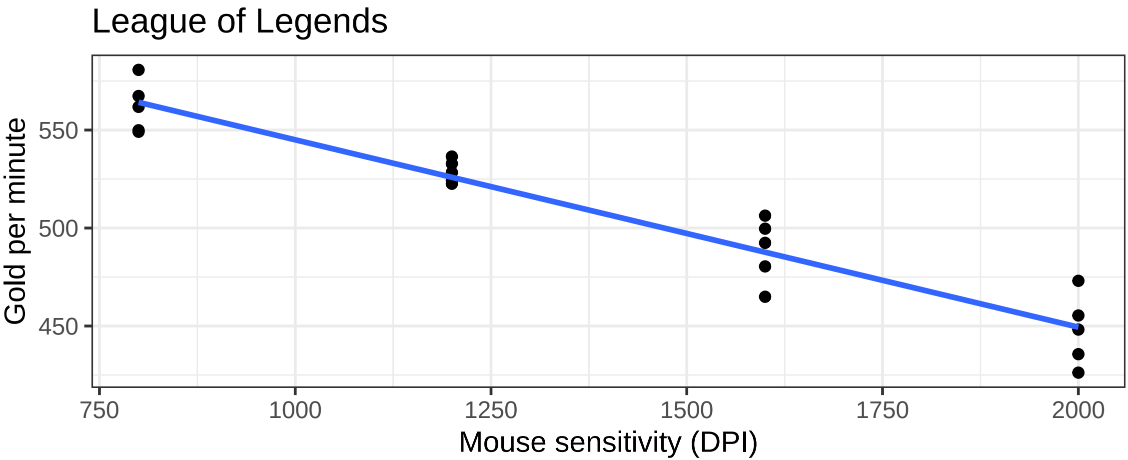
Estimate model parameters

```
m <- lm(gpm ~ sensitivity, data = mouse)
m

##
## Call:
## lm(formula = gpm ~ sensitivity, data = mouse)
##
## Coefficients:
## (Intercept)  sensitivity
##   640.63505     -0.09561
```

$$\hat{\beta}_0 = 641, \qquad \hat{\beta}_1 = -0.096$$

Fit a line


League of Legends
Gold per minute
550
500
450
750
1000
1250
1500
1750
2000
Mouse sensitivity (DPI)

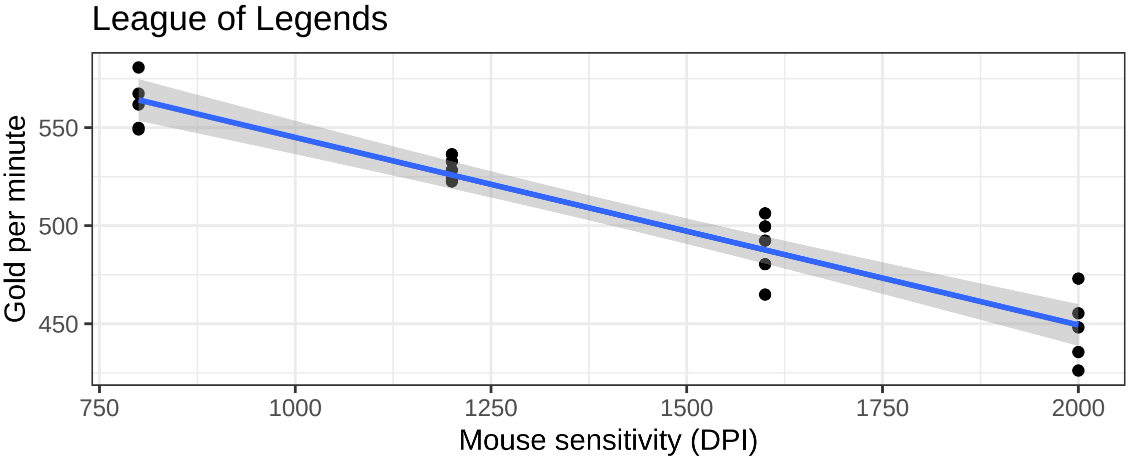
Credible intervals

```
confint(m)

##                    2.5 %       97.5 %
## (Intercept) 619.7064093 661.56368201
## sensitivity  -0.1098489  -0.08136859
```

A 95% CI for β_0 is (620, 662).
A 95% CI for β_1 is (-0.11, -0.081).

Uncertainty in the line



Interpretation

$$E[Y_i|X_i] = \beta_0 + \beta_1 X_i$$

When $X_i = 0$ (when mouse sensitivity is 0), $E[Y_i]$ (expected gold per minute) is 641 with a 95% CI of (620, 662).

For every 1 increase in X_i (mouse sensitivity increases by 1), the expected increase in Y_i (gold per minute) is -0.096 with a 95% CI of (-0.11, -0.081).

For every 400 increase in X_i (mouse sensitivity increases by 1), the expected increase in Y_i (gold per minute) is -40 with a 95% CI of (-44, -32).

Regression summary

```
summary(m)

##
## Call:
## lm(formula = gpm ~ sensitivity, data = mouse)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.2125  -8.8834   0.6222   7.8498  23.6453
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 640.635046   9.961643   64.31  < 2e-16 ***
## sensitivity  -0.095609   0.006778  -14.11 3.59e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.56 on 18 degrees of freedom
## Multiple R-squared:  0.917,Adjusted R-squared:  0.9124
## F-statistic:   199 on 1 and 18 DF,  p-value: 3.589e-11
```

Simple linear regression model assumptions

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \qquad \epsilon_i \stackrel{ind}{\sim} N(0, \sigma^2).$$

Assumptions:

- Linearity
- Normality
- Constant variance
- Independence

Many plots will be based off residuals:

$$r_i = \hat{\epsilon}_i = Y_i - \hat{Y}_i = Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i.$$

Linearity assumption

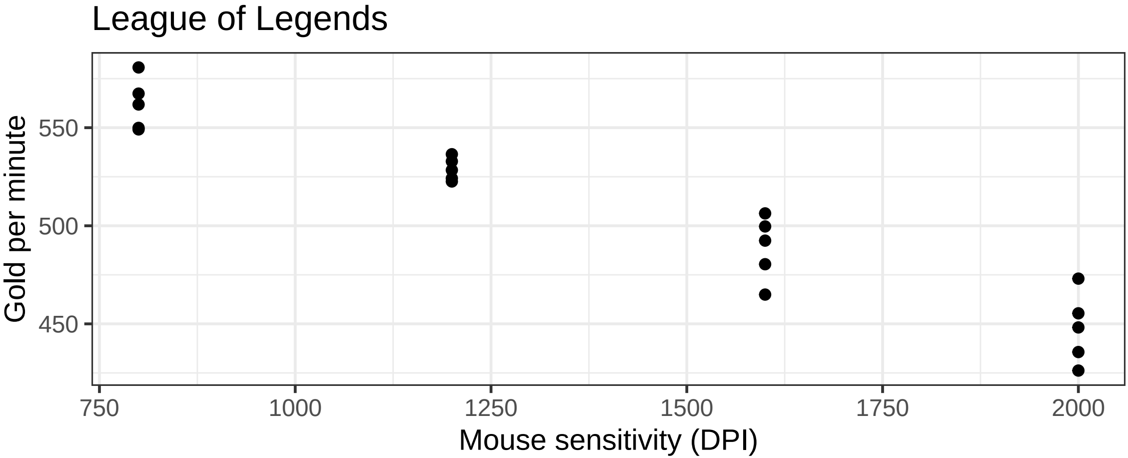
Linear relationships between expected value of the dependent variable and the independent variable:

$$E[Y_i|X_i] = \beta_0 + \beta_1 X_i$$

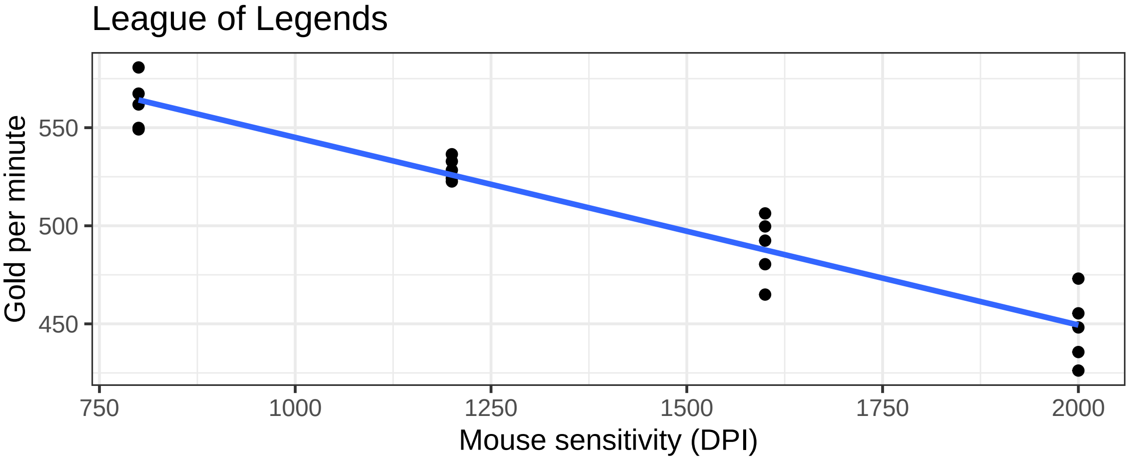
Look at

- Independent variable vs dependent variable
- Residuals vs predicted value

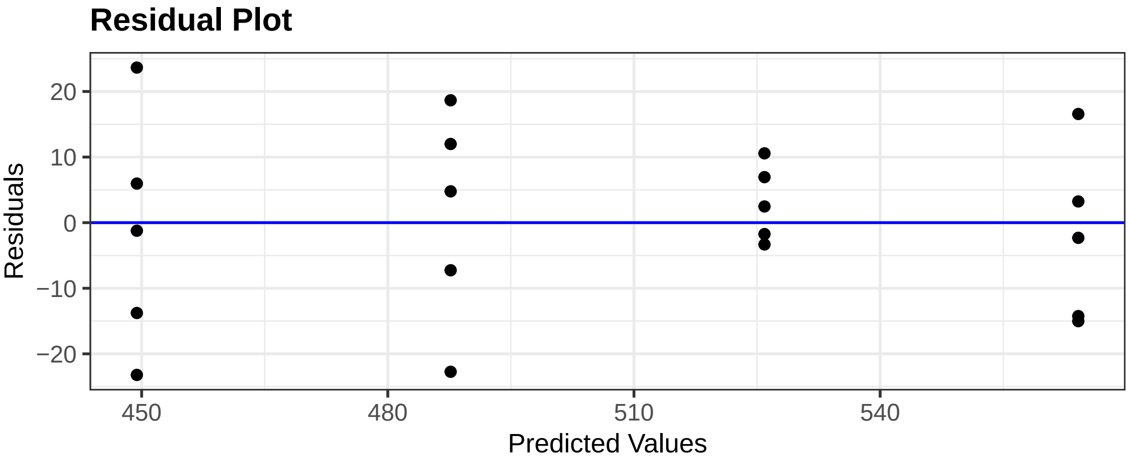
Linear assumption is valid



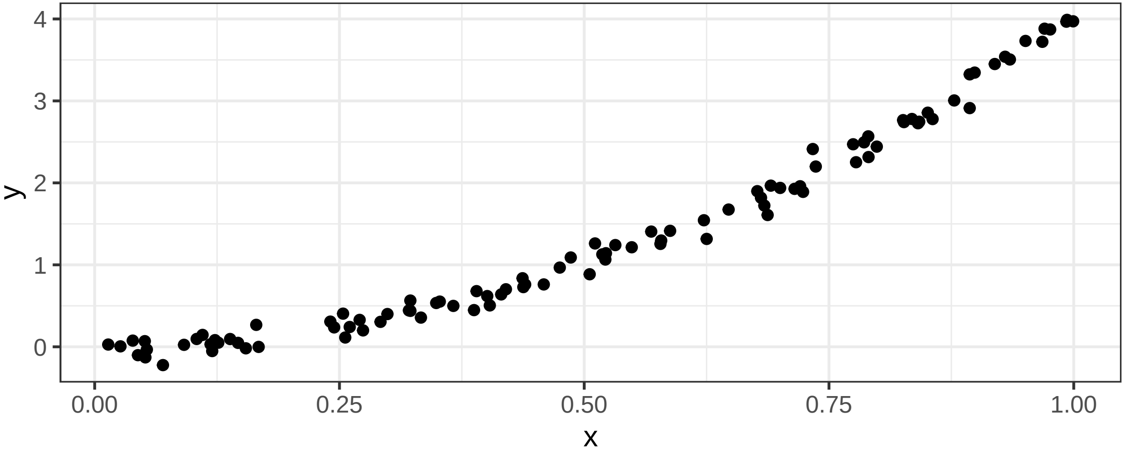
Linear assumption is valid



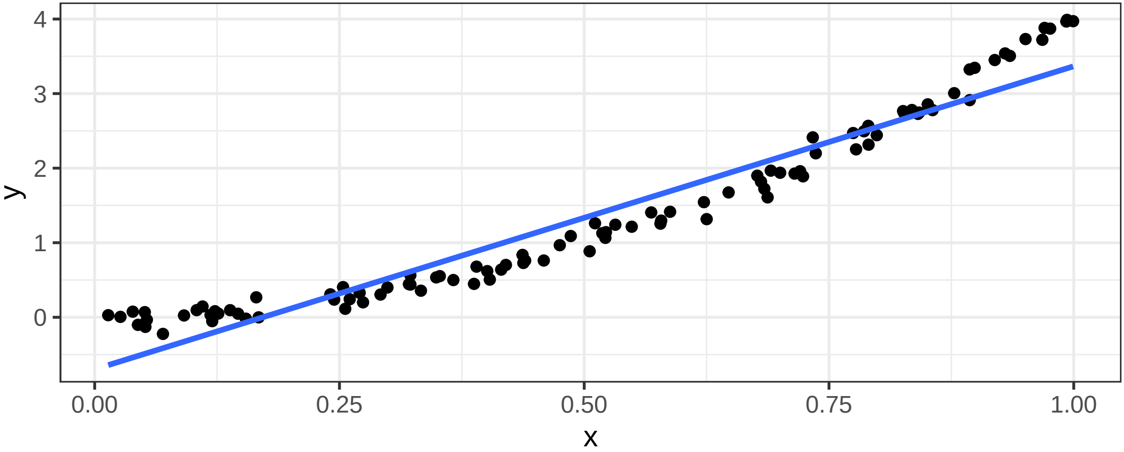
Linear assumption is valid



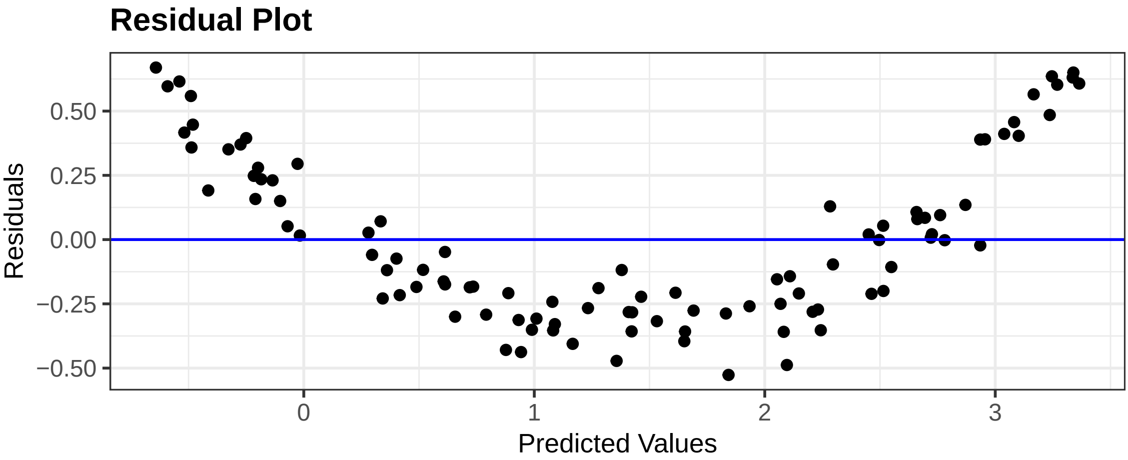
Linear assumption is NOT valid



Linear assumption is NOT valid



Linear assumption is NOT valid

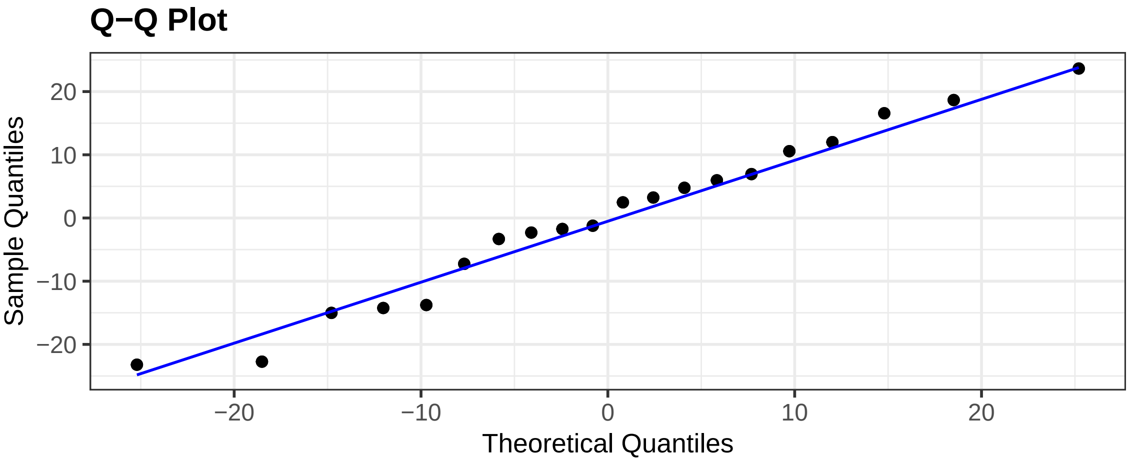


Normality

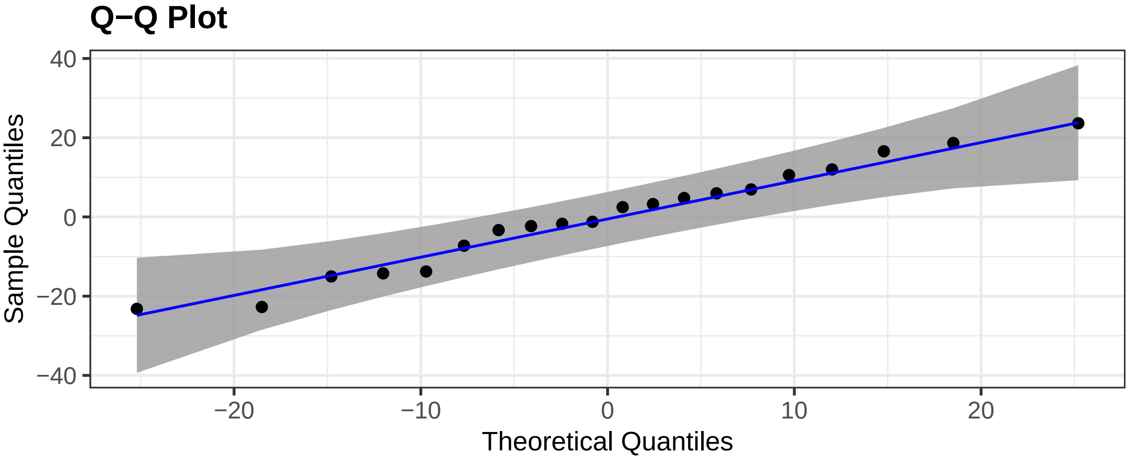
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \qquad \epsilon_i \stackrel{ind}{\sim} N(0, \sigma^2).$$

Best diagnostic is a QQ-plot

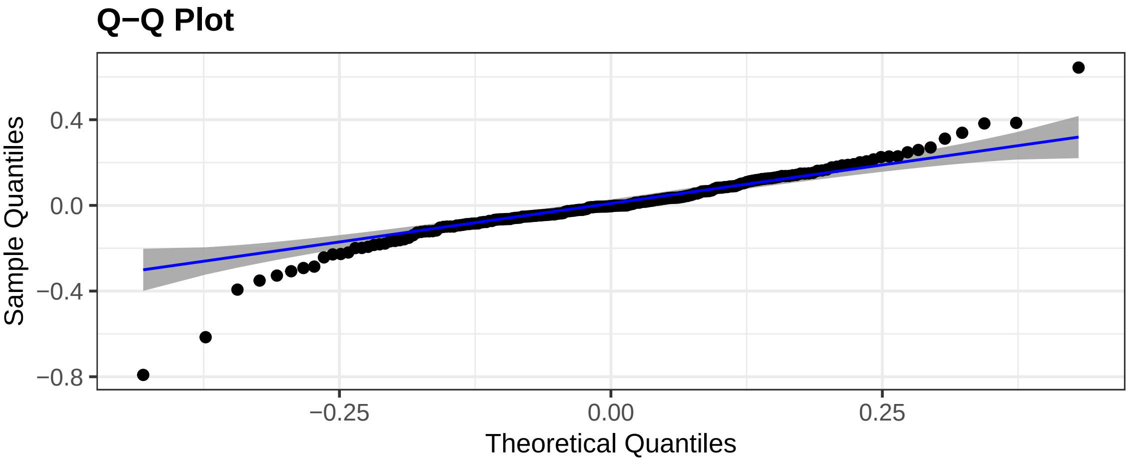
QQ-plot (normality is valid)



QQ-plot (normality is valid)



QQ-plot (normality is NOT valid)

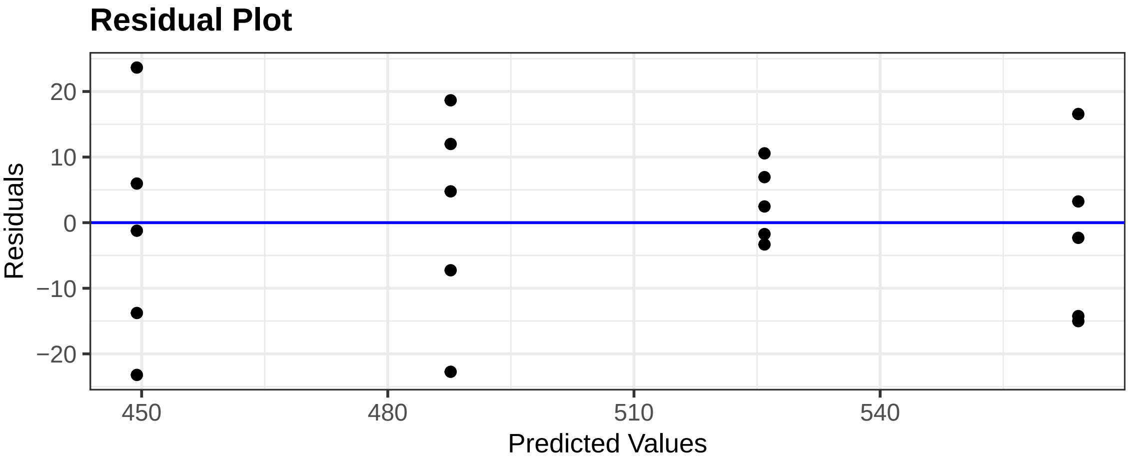


Constant variance

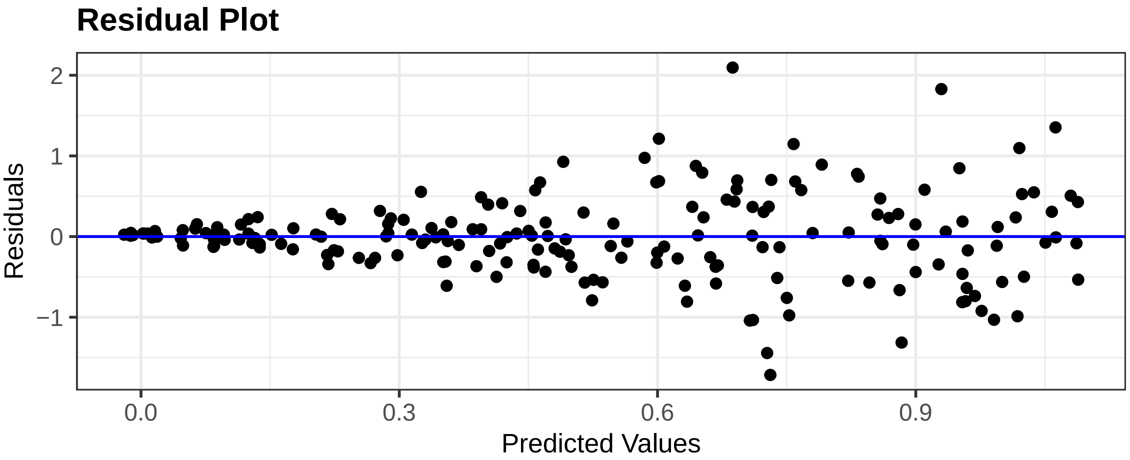
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \qquad \epsilon_i \stackrel{ind}{\sim} N(0, \sigma^2).$$

Plot residuals vs predicted values and look for a “horn” shape pattern

Constant variance assumption is valid



Constant variance assumption is NOT valid



Independence

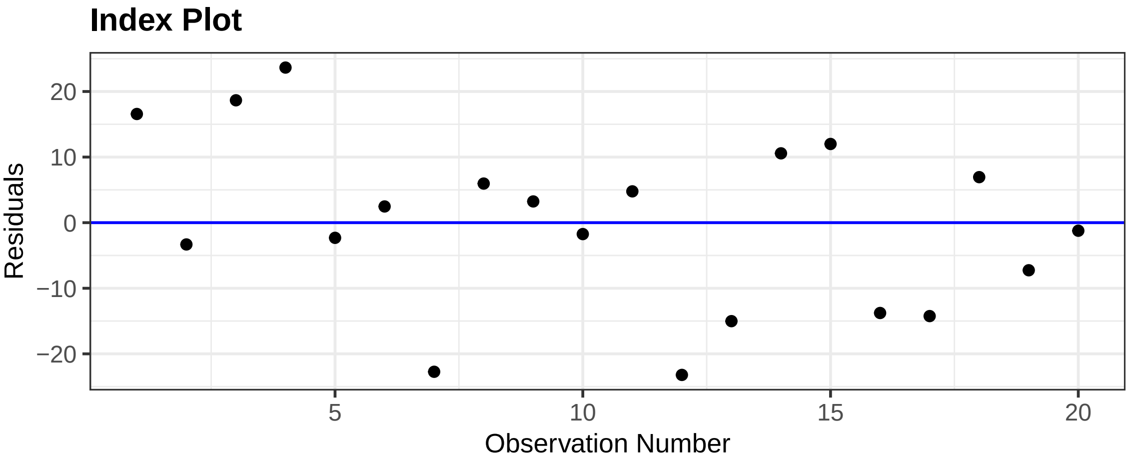
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \qquad \epsilon_i \stackrel{ind}{\sim} N(0, \sigma^2).$$

No great way to assess this assumption other than subject matter knowledge.

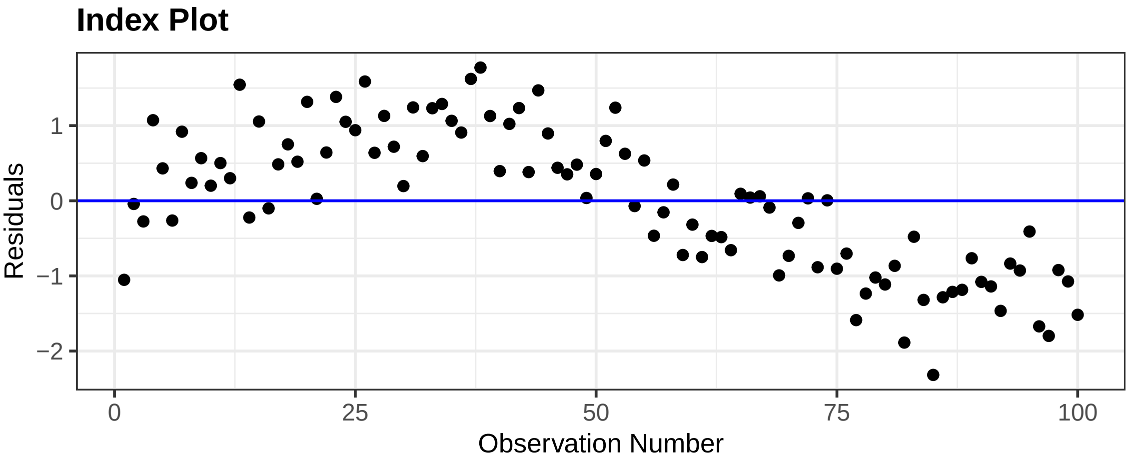
Main causes for dependence are

- temporal (residuals vs index might help)
- spatial
- clustering

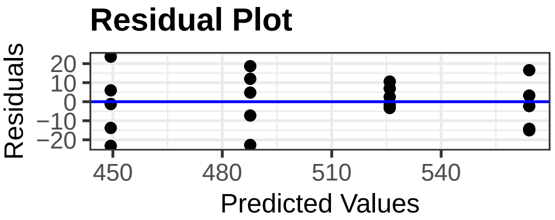
Residuals vs index (independence assumption is valid)

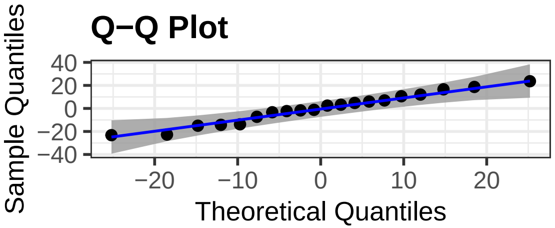


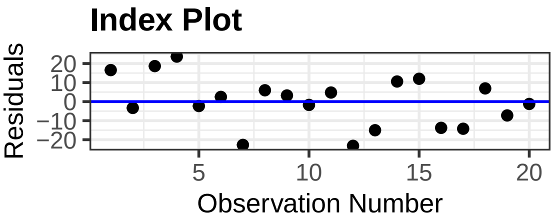
Residuals vs index (independence assumption is NOT valid)



All plots together







Summary

Simple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \qquad \epsilon_i \stackrel{ind}{\sim} N(0, \sigma^2).$$

Assumptions:

- Linearity
- Normality
- Constant variance
- Independence