Topics to be Covered

- Simple linear regression
  - When predictor is a continuous variable
  - When predictor is a categorical variable
- Model diagnosis

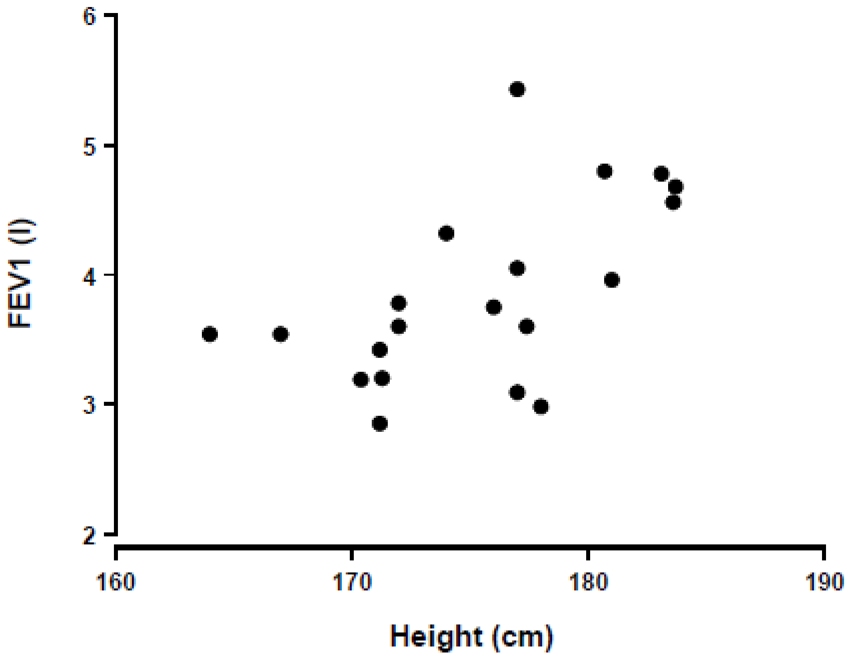
Motivating Example

- The following table gives data collected by a group of medical students in a physiology class. The objective is to assess association between height and FEV1 (amount of air exhaled during the first second of the forced breath)

Height	FEV1	Height	FEV1	Height	FEV1
164.0	3.54	172.0	3.78	178.0	2.98
167.0	3.54	174.0	4.32	180.7	4.80
170.4	3.19	176.0	3.75	181.0	3.96
171.2	2.85	177.0	3.09	183.1	4.78
171.2	3.42	177.0	4.05	183.6	4.56
171.3	3.20	177.0	5.43	183.7	4.68
172.0	3.60	177.4	3.60		

A Scatterplot

<u>About the data set</u>: 20 students with variables Height and FEV1



A **scatterplot** displays the **form, direction,** and **strength** of the relationship between two **quantitative** variables.

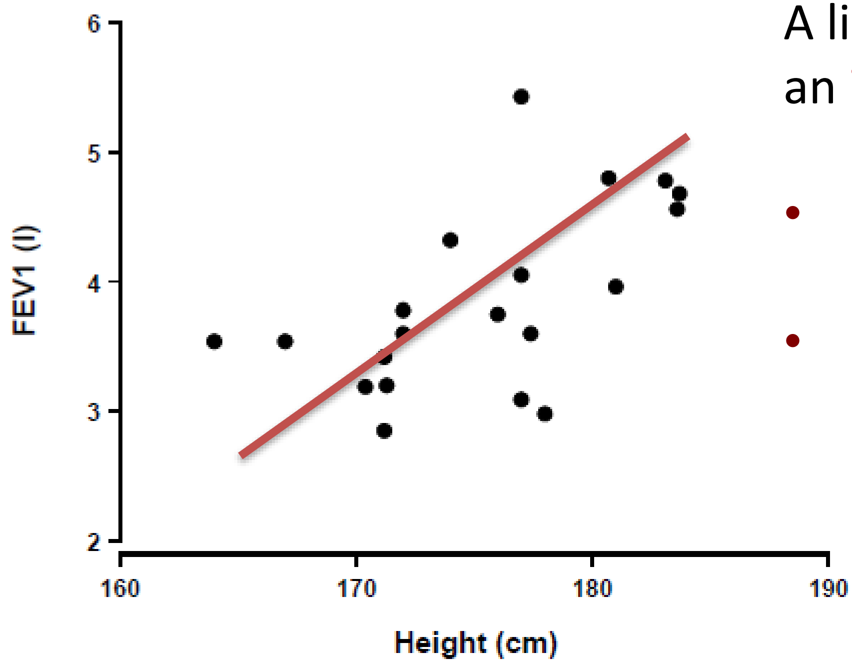
Form: linear relationship (described by a straight line)

Direction: positive

Strength: moderate

Find a Line of Best Fit

<u>About the data set</u>: 20 students, two **continuous** variables (Heights and FEV1)



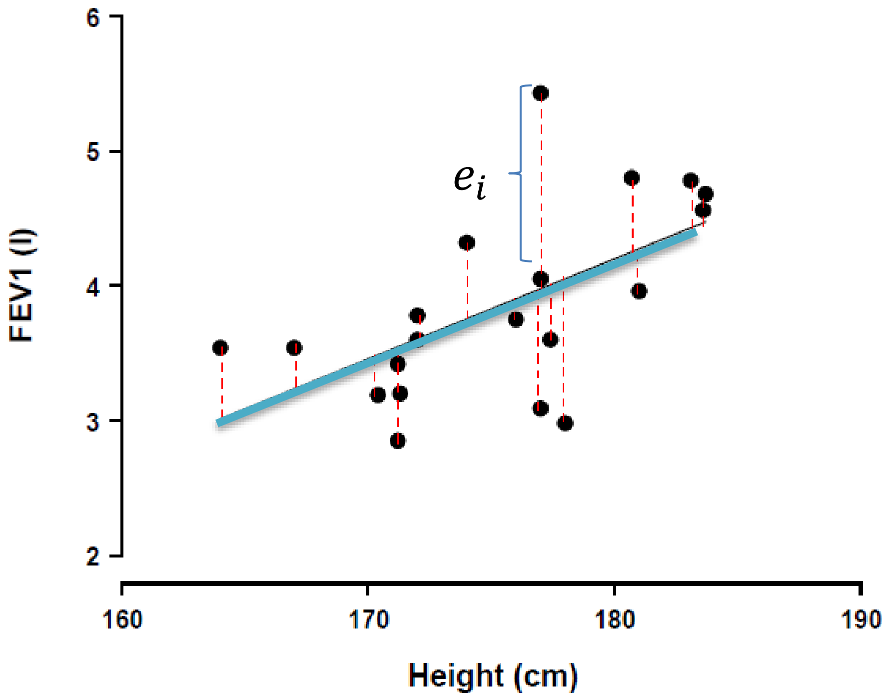
A line is described by an **intercept** and a **slope**

- **Intercept is** the value of *FEV1* when Height = 0.
- **Slope is** the change in FEV1 for each one-unit increase in *Height.*

Simple Linear Regression

- A technique to study association between two **continuous** variables: X and Y
- X= Independent variable (other names: predictors, covariates, explanatory variable)
  - Variable whose impact is to be assessed
- Y= Dependent variable (other names: outcome, response)
  - Variable on which you want to assess effect of X
- Goal is to assess the impact of X on Y

Find a Line of Best Fit



- Red dashed lines represent errors of your fit
- The **best line** to describe your data is the line which minimizes sum of squares of red lengths (least squares fit)

Sum of squares of red lengths is $\sum_{i=1}^{n} {e_i}^2$.

Population Regression Coefficients

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- β_0 is the Intercept (i.e., value of y when x=0)
- We often are interested in the **slope**, β_1, which indicates the association between X and Y.
  - $\beta_1 = 0$ no association
  - $\beta_1 > 0$ positive association
  - $\beta_1 < 0$ negative association

Parameters in Regression Model

- The intercept β_0, the slope β_1, are the **unknown** parameters of the regression model.
- We rely on the random sample data to provide estimate these parameters

Least-squares regression line *True population regression line*

$$\hat{\mu}_y = b_0 + b_1 x \Longrightarrow \mu_y = \beta_0 + \beta_1 x$$

(estimated from sample)

Parameter Estimates (Theory)

- By minimizing sum of squared errors from the sample, we obtain:
  - Slope
    $$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad SE_{b1} = \frac{s}{\sqrt{\sum (x_i - \bar{x}_i)^2}}$$
  - Intercept
    $$b_0 = \bar{y} - b_1\bar{x} \qquad SE_{b0} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x}_i)^2}}$$
  - Error variance
    $$s^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$$

Analysis using R

outcome

predictor

$$FEV1 = -9.19 + 0.0744 \times Height$$

```
> fit=lm(FEV1~height, data=data)
> summary(fit)

Call:
lm(formula = FEV1 ~ height, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.07090 -0.32367  0.03446  0.31797  1.45349

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.19039    4.30644  -2.134  0.04684 *
height       0.07439    0.02454   3.031  0.00719 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5892 on 18 degrees of freedom
Multiple R-squared:  0.3379,    Adjusted R-squared:  0.3011
F-statistic: 9.187 on 1 and 18 DF,  p-value: 0.007185
```

FEV1 at zero height. Needs to be in equation but is usually of little direct interest

It quantifies relationship: FEV1 increases by 0.074 for each cm increase in height

33.79% variation in FEV1 is explained by Height.
Adj R-sq will be used in the multiple linear regression

Analysis using R

95% Confidence interval

```
> confint(fit)
                  2.5 %     97.5 %
(Intercept) -18.23788410 -0.1428935
height        0.02282546  0.1259531
```

```
> fit=lm(FEV1~height, data=data)
> summary(fit)

Call:
lm(formula = FEV1 ~ height, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.07090 -0.32367  0.03446  0.31797  1.45349

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.19039    4.30644  -2.134  0.04684 *
height       0.07439    0.02454   3.031  0.00719 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5892 on 18 degrees of freedom
Multiple R-squared:  0.3379,    Adjusted R-squared:  0.3011
F-statistic: 9.187 on 1 and 18 DF,  p-value: 0.007185
```

<u>Conclusion</u>. Higher height is associated with larger value of FEV1. FEV1 increases 0.074 (95% CI: 0.023,0.126; p=0.0072) for every one cm increase in Height.

What is R^2

- R^2 is the proportion of the variance of Y that is explained by the model. $0 \leq R^2 \leq 1$
- An R^2 value of zero indicates that a linear function of X does not predict Y at all.
- An R^2 value of one indicates perfect prediction.
- R^2 values near one are considered better.

Predictions

- Given a student of height **180** cm, what is his predicted FEV1?
- Based on the best fitted regression line, namely:

$$FEV1 = -9.19 + 0.0744 \times Height$$

- We get -9.19 + 0.07439 × **180** = 4.20

Interpolation vs. Extrapolation

- **Interpolation** means estimating Y for values of X that are between values of X that occur in the data.
  - Interpolating between X values is a legitimate use of regression.
- **Extrapolating** means estimating Y for X values greater than the largest X value, or less than the smallest X value in the data.
  - Extrapolation is dangerous. Inference should be restricted to the range of observed X values.

A Single Categorical Predictor-
One-way ANOVA

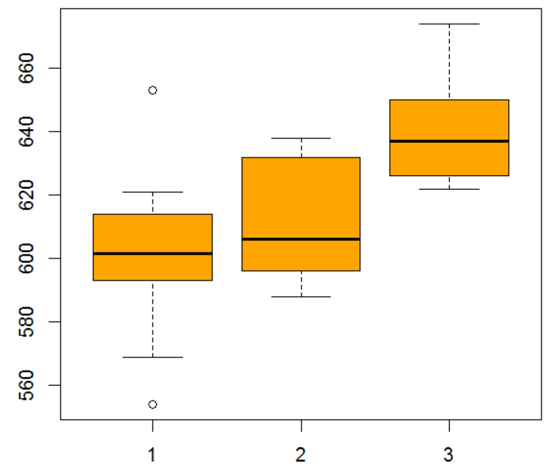
Motivating Example

y	Group
611	1
621	1
614	1
593	1
...	
632	2
631	2
588	2
607	2
596	2
...	
650	3
622	3
626	3
626	3
631	3

control

low jump

high jump



<u>The goal</u> is to compare bone density (y) across the 3 jump groups (Group)

Multiple t tests vs. One-way ANOVA

- We could look at separate ***t* tests** to compare each <u>pair</u> of means to see if they are different:

- What is the advantage of using o**ne-way ANOVA?**

Regression Equation

- ANOVA model is a regression model
- A categorical variable with *I* levels needs *I*-1 dummy variables to represent it

If *I*=3, we need to create two dummy variables: z_1 and z_2

$$\hat{y} = b_0 + b_1 z_i + b_2 z_2$$

z1	z2
0	0
0	0
0	0
0	0
1	0
1	0
1	0
1	0
0	1
0	1
0	1
0	1

b_0 *control (reference group)*

$b_0 + b_1$ *low jump grp*

$b_0 + b_2$ *High jump grp*

$\boldsymbol{b_1}$ is difference between low jump group and control

$\boldsymbol{b_2}$ is difference between high jump group and control

Analysis using R

"group" is a factor

```
> model = lm(y ~ group,  data = data)
> # obtain parameter estimates
> summary(model)

Call:
lm(formula = y ~ group, data = data)

Residuals:
   Min     1Q Median     3Q    Max
 -47.1  -13.3   -3.3   12.5   51.9

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  601.100      6.826  88.064  < 2e-16 ***
group2        11.400      9.653   1.181 0.247912
group3        37.600      9.653   3.895 0.000584 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.58 on 27 degrees of freedom
Multiple R-squared:  0.3714,    Adjusted R-squared:  0.3249
F-statistic: 7.978 on 2 and 27 DF,  p-value: 0.001895
```

Pairwise group comparisons

```
> library(emmeans)
> paircom=emmeans(model, specs = pairwise ~ group, adjust = "none")
> # get confidence intervals
> confint(paircom)
$emmeans
 group emmean   SE df lower.CL upper.CL
 1        601 6.83 27      587      615
 2        612 6.83 27      598      627
 3        639 6.83 27      625      653

Confidence level used: 0.95

$contrasts
 contrast        estimate   SE df lower.CL upper.CL
 group1 - group2    -11.4 9.65 27    -31.2     8.41
 group1 - group3    -37.6 9.65 27    -57.4   -17.79
 group2 - group3    -26.2 9.65 27    -46.0    -6.39
```

$y=601.1+11.4 \times z1 + 37.6 \times z2$

11.4 is the difference between low jump group and control
37.6 is the difference between high jump group and control

<u>Conclusion</u>. High jump group has higher bone density than controls (difference in mean bone density is 37.6, 95% CI: 17.8, 57.4; p=0.00058).

Advantage of ANOVA over pair-wise t-tests

- Pair-wise comparisons using t-tests is cumbersome and is not easy to summarize when the number of groups/populations compared is large.
- ANOVA is more powerful than the t-tests since it uses all data to estimate the error (residual variability).

Checking Assumptions

Before you can trust the results of inference, you must check the conditions for inference one by one.

- ✓ The relationship is linear (for continuous predictor)
- ✓ Normality of the residuals
- ✓ Equality of variance (variance of the responses is the same for all values of *x*).

Remember that the Y's need NOT be normal, but residuals should be normal. (e.g., both Y and X may be skewed, but residuals may be normal).

You can check all of the conditions for regression inference by looking at graphs of the residuals, or **residual plots.**

Motivating Example

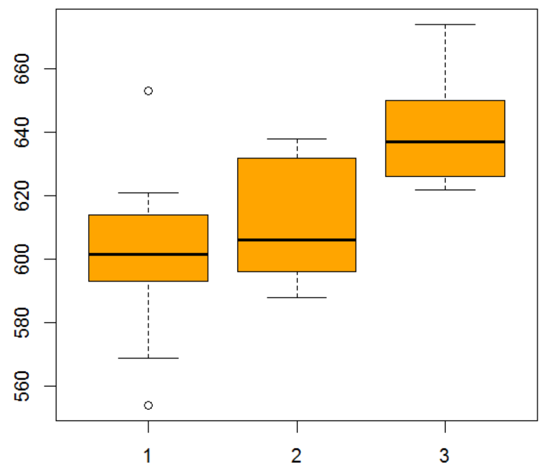
y	Group
611	1
621	1
614	1
593	1
…	
632	2
631	2
588	2
607	2
596	2
…	
650	3
622	3
626	3
626	3
631	3

control

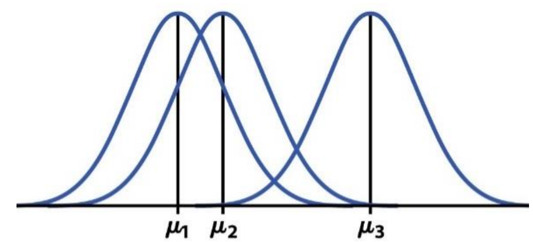
low jump

high jump

Data



Model



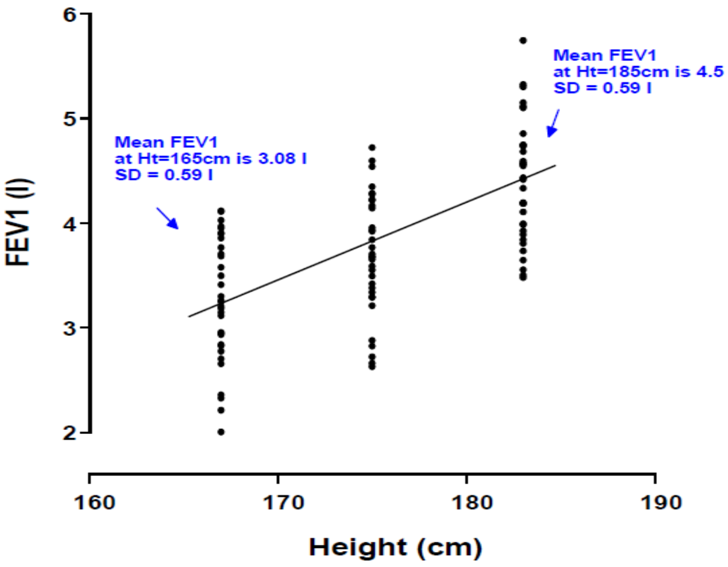
common standard deviation σ

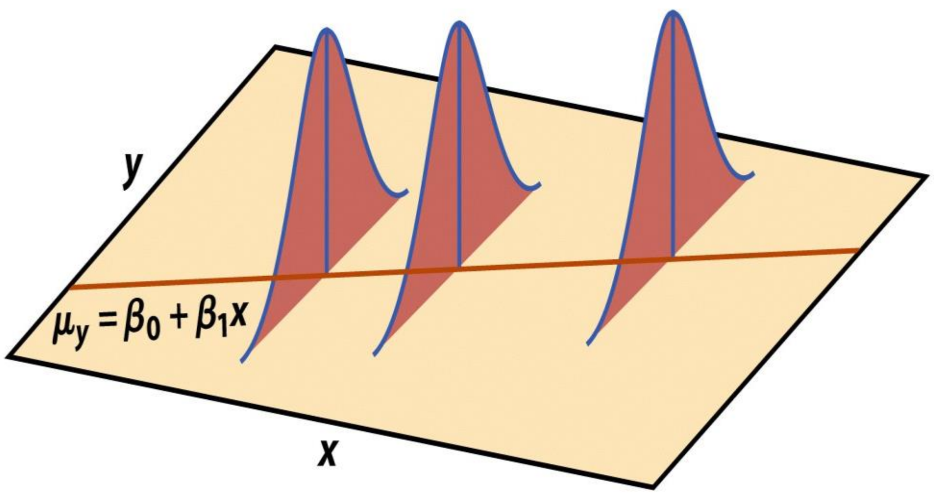
Simple Linear Regression

Theory

- In the population, the association is described by

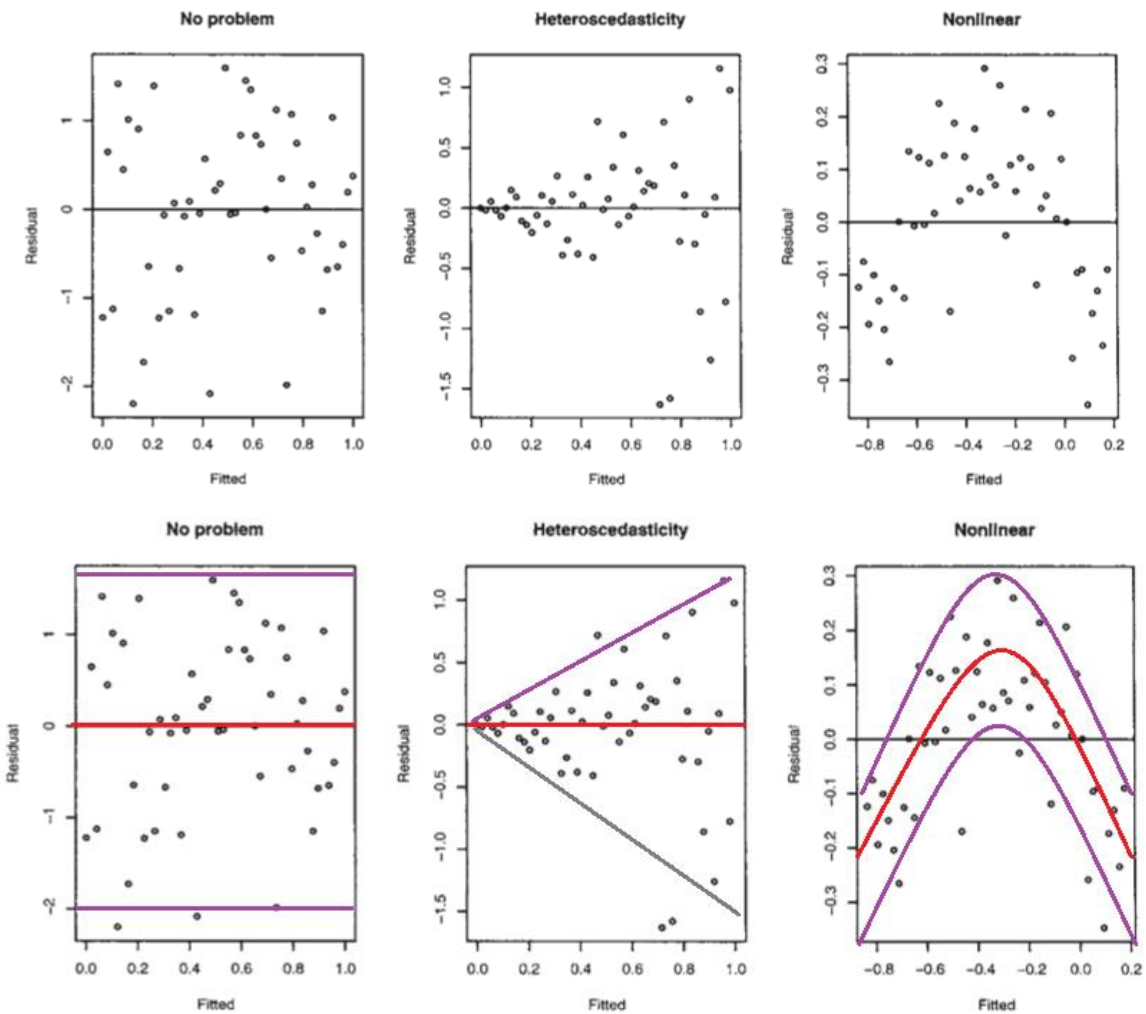
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$





where the ε_i are **independent** and **Normally** distributed $N(0,\sigma)$, and σ measures how much y vary about the regression line

Examples Residual Plots



- Residual on vertical-axis and fitted y on horizontal-axis. Ideally residuals should be a random scatter around zero.
- Residual *patterns* suggest deviations from a linear relationship or inequality of variance