A Review of Linear Regression

Emily Hector

University of Michigan

June 18, 2019

Modeling Data

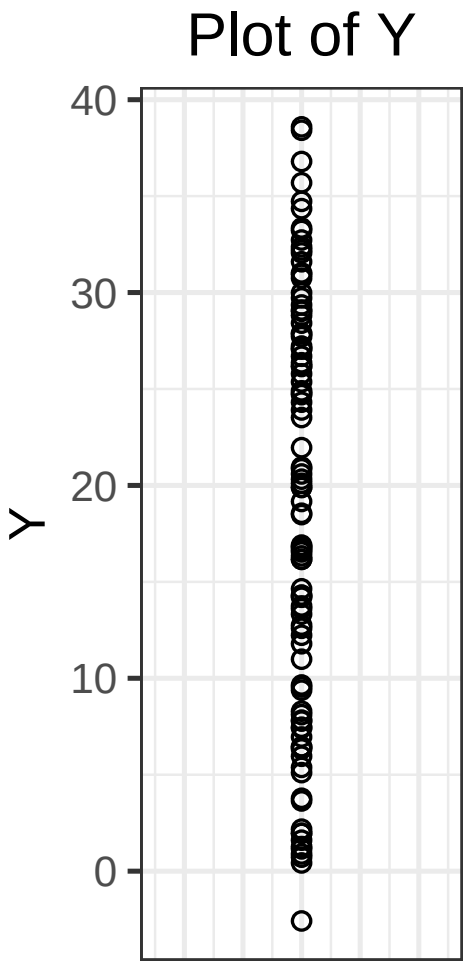
- Types of outcomes
  - Continuous, binary, counts, ...
- Dependence structure of outcomes
  - Independent observations
  - Correlated observations, repeated measures
- Number of covariates, potential confounders
  - Controlling for confounders that could lead to spurious results
- Sample size

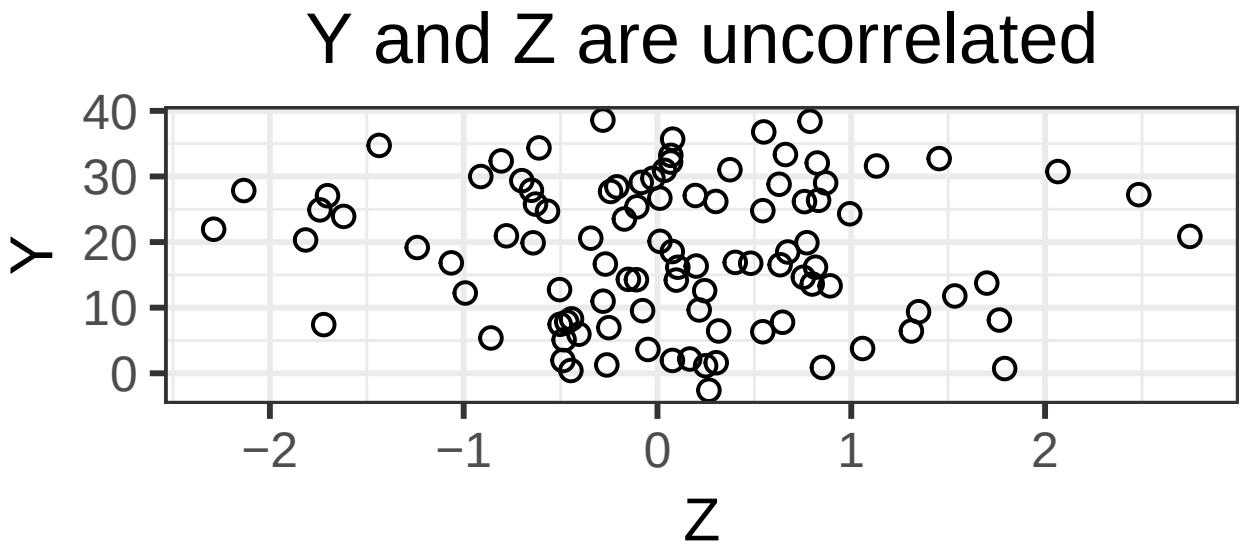
These factors will determine the appropriate statistical model to use

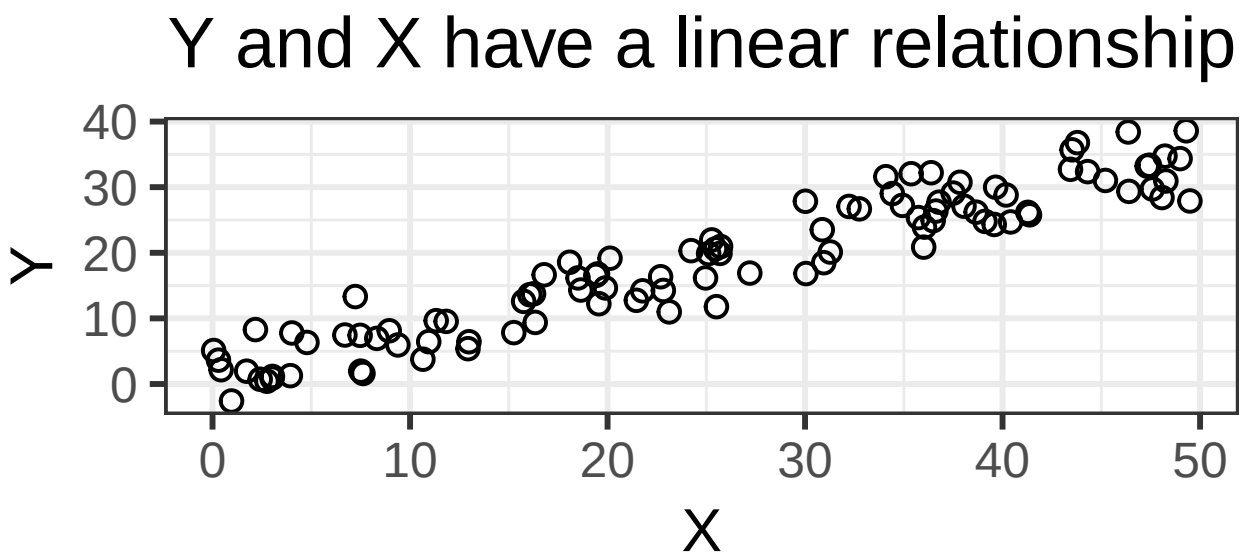
Regression

- Technique used to model and analyze data:
  - e.g. data with variables for age, sex, weight, height, socio-economic status, and diet.
- Exploits the relationships between variables to gain information about one of them through knowing values of the others.
  - e.g. the relationship between weight and diet.
- Regression can be used for estimation, hypothesis testing, prediction, and modeling causal relationships.

Can variation in outcome Y be explained by correlation with a different variable?

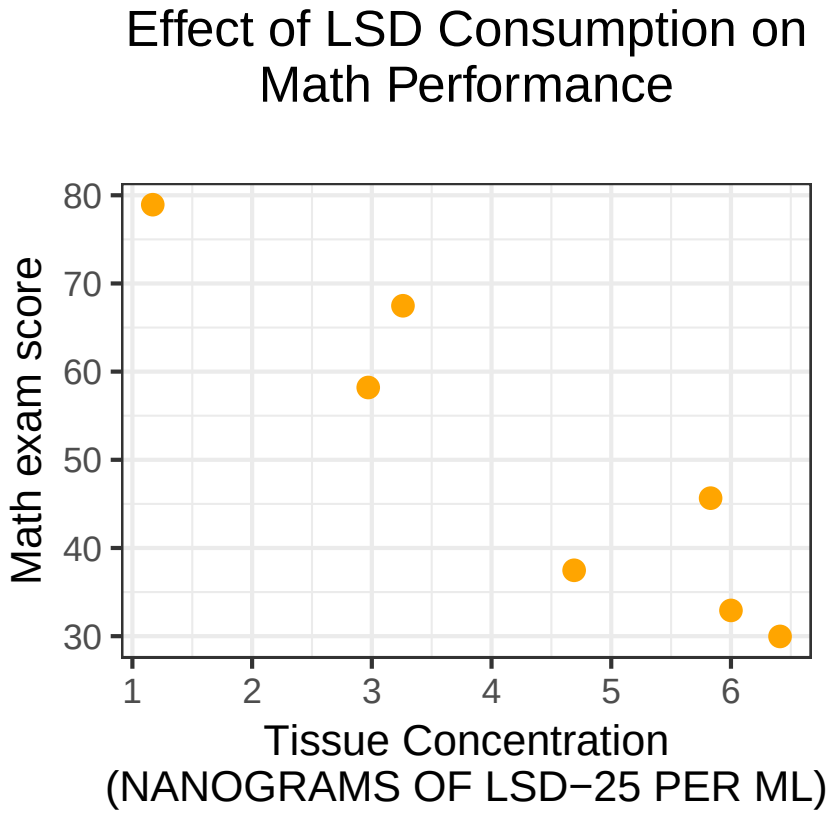




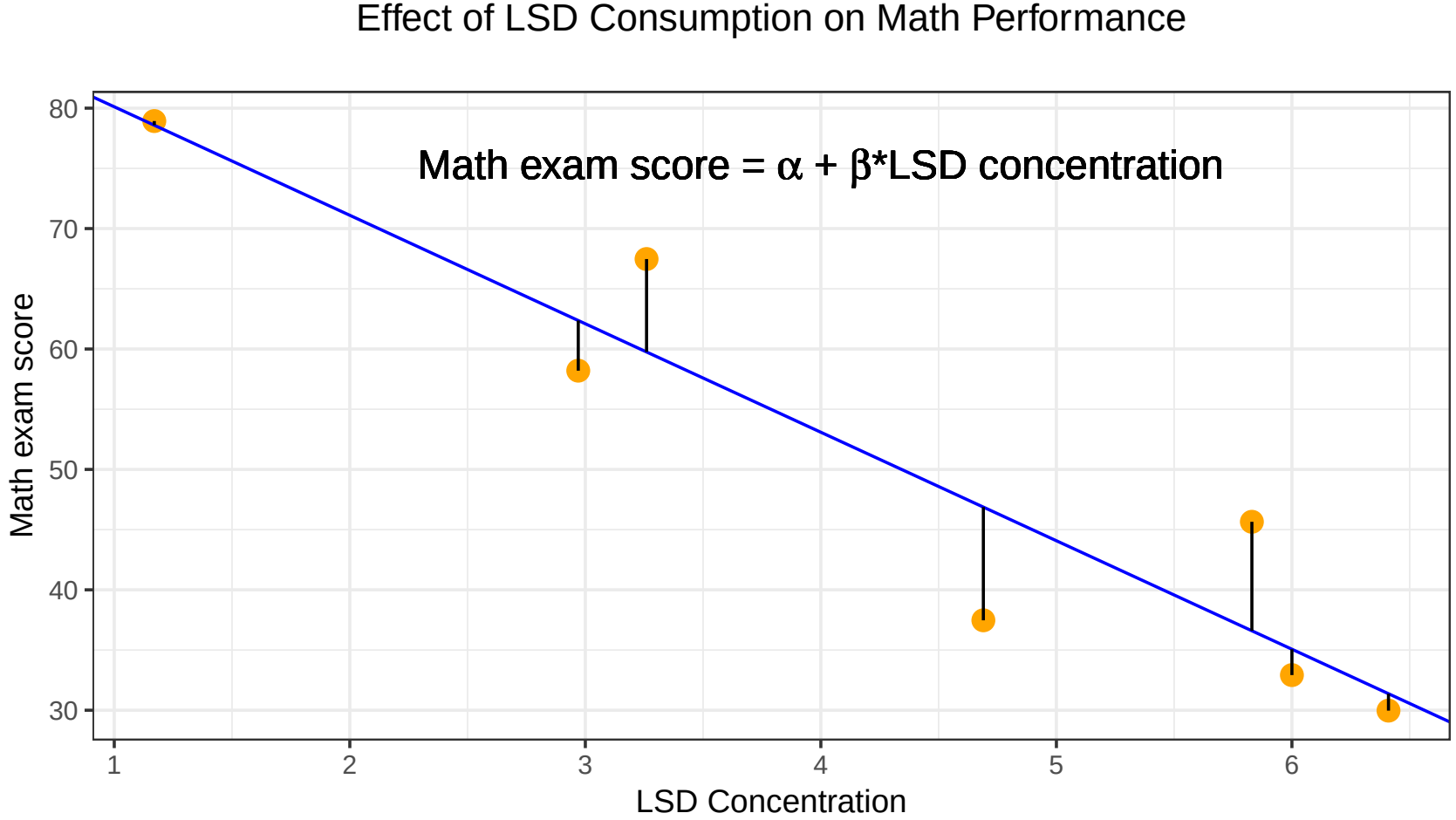


Do Psychedelics Affect Math Performance?

- Wagner, Agahajanian, and Bing (1968). Correlation of Performance Test Scores with Tissue Concentration of Lysergic Acid Diethylamide in Human Subjects. Clinical Pharmacology and Therapeutics, Vol.9 pp635-638
- Group of volunteers was given LSD, their mean scores on math exam and tissue concentrations of LSD were obtained at n=7 time points.



http://www.stat.ufl.edu/ winner/data/



- α is the intercept, β is the slope.
- ϵ_i =distance between the regression line and the observed value for the ith data point.

Least Squares Regression

- Intuitively, we want a line that is "close" to as many data points as possible. $\rightarrow$ Minimize the total distance between the regression line and the data points.
- Let $\widehat{y}_i = \alpha + \beta x_i$ describe the linear relationship between predicted value $\widehat{y}_i$ and x_i.
- Let $e_i = \widehat{y}_i - y_i$ be the difference between the predicted and observed values for the ith data point.
- Minimize the objective function

$$\phi(\alpha, \beta) = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} \left(\widehat{y}_i - y_i\right)^2 = \sum_{i=1}^{N} \left(\alpha + \beta x_i - y_i\right)^2 .$$

- The quantity $\sum_{i=1}^{N} e_i^2$ is often called Residual or Error Sum of Squares (SS_{res}).

Least Squares Regression

Solve for α and β that minimize

$$\phi(\alpha, \beta) = \sum_{i=1}^{N} (\alpha + \beta x_i - y_i)^2 .$$

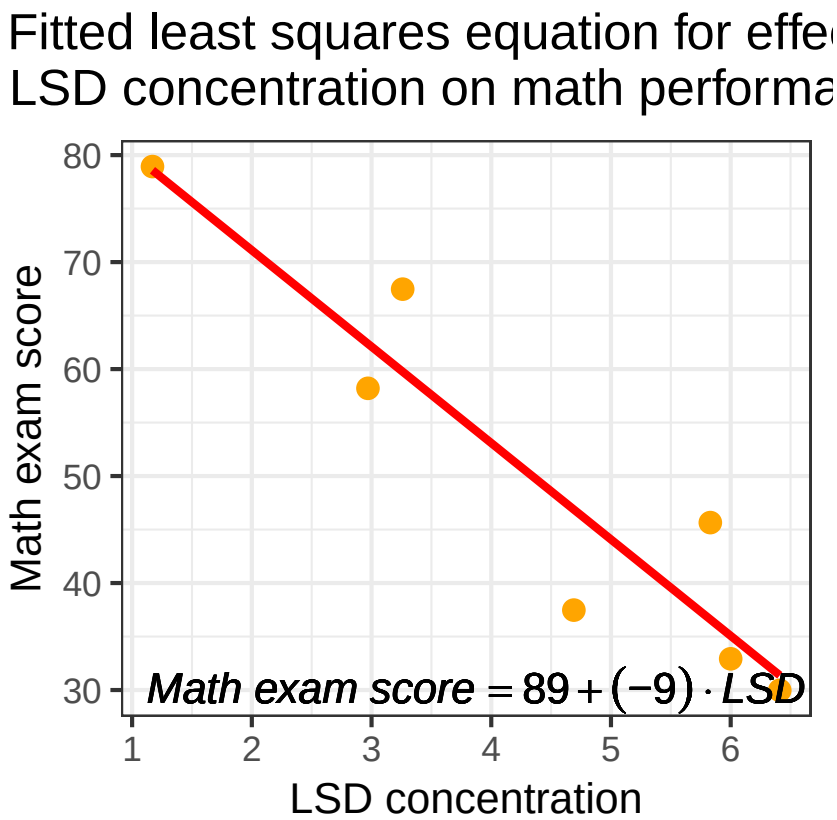
Calculus: take partial derivatives, set equal to zero and solve:

$$\begin{aligned} 0 &= \frac{\partial}{\partial \alpha}\phi \\ &= 2\sum_{i=1}^{N}(\alpha + \beta x_i - y_i) \\ \Rightarrow \alpha &= \bar{y} - \beta \bar{x}. \end{aligned}$$

$$\begin{aligned} 0 &= \frac{\partial}{\partial \beta}\phi = 2\sum_{i=1}^{N}(\alpha + \beta x_i - y_i)x_i \\ &= (\bar{y} - \beta \bar{x})\sum_{i=1}^{N} x_i + \beta \sum_{i=1}^{N} x_i^2 - \sum_{i=1}^{N} x_i y_i \\ \Rightarrow \beta &= \frac{\sum_{i=1}^{N} x_i y_i - \frac{1}{N}\left(\sum_{i=1}^{N} y_i\right)\left(\sum_{i=1}^{N} x_i\right)}{\sum_{i=1}^{N} x_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} x_i\right)^2}. \end{aligned}$$

What can you do with a least squares regression line?

- A 1 unit increase in LSD concentration results in an average decrease of ≈ 9 points on the math score.
- 70% is the predicted math score for having tissue concentration of 2.12 nanograms of LSD-25/mL.
- We have quantified the relation between the explanatory variable and the outcome.



Regression with more than one covariate

- An outcome rarely ever depends on just a single explanatory variable.
- Might be interested in the effect of multiple variables or want to control for the effect of a confounding variable.

Example:

- Height determined by sex, age, genetics.
- The effect of individual genetic variants on gene expression is of primary interest but need to account for confounders age, sex, batch, etc.

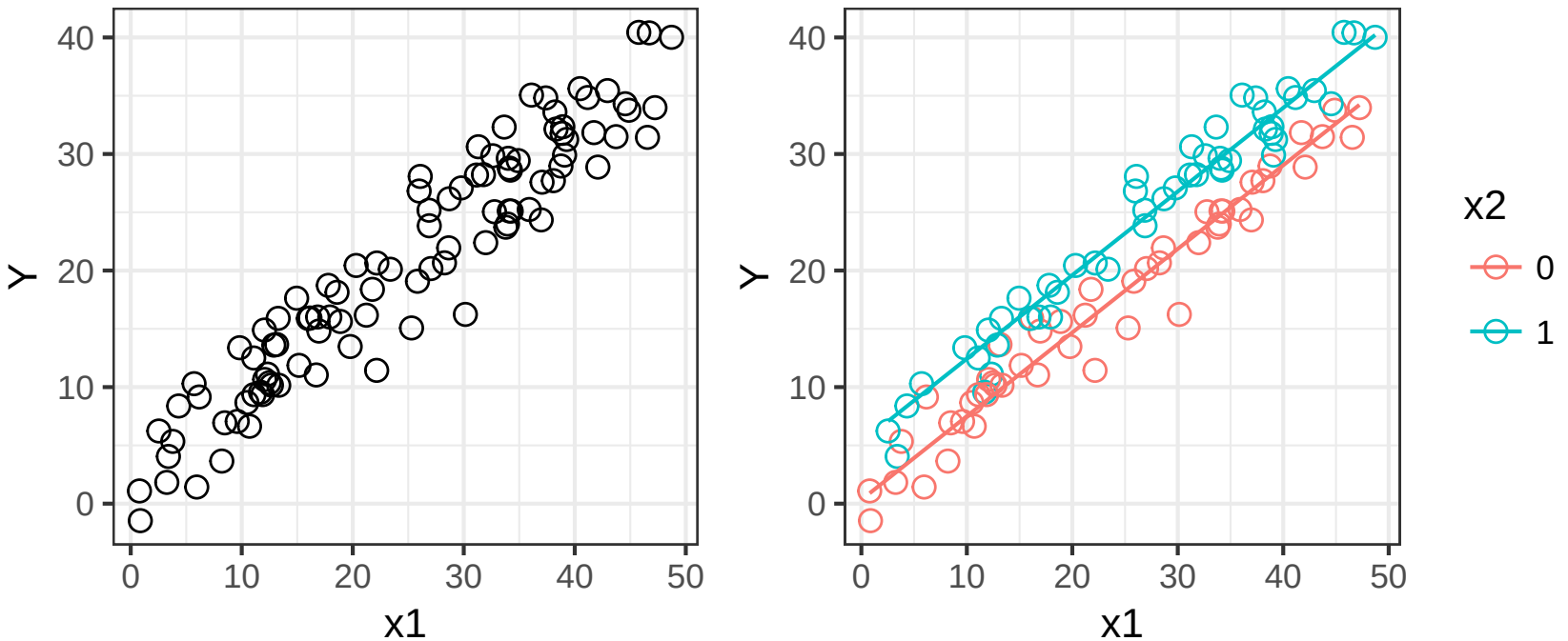
- Multiple linear regression models the mean outcome value on more than one explanatory variable.
- Provides the framework for interaction of explanatory variables.

Multiple linear regression

Linear regression with one continuous covariate x_1 and one dichtomous/binary covariate $x_2 \in \{0, 1\}$:

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2}.$$

- The regression equation is two parallel lines, each with slope β_1.
- Data points with $x_2 = 0$ have intercept of α.
- Data points with $x_2 = 1$ have intercept of $\alpha + \beta_2$.

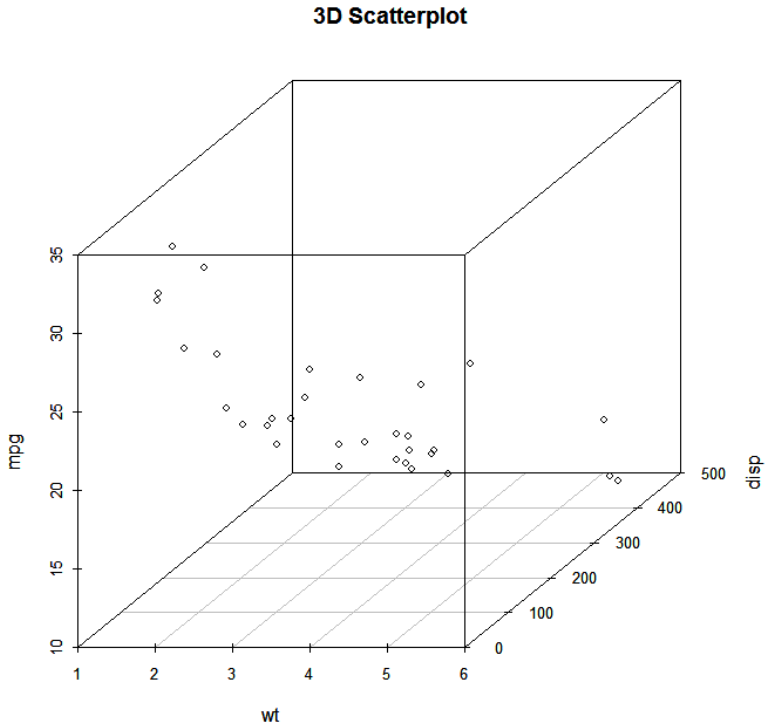


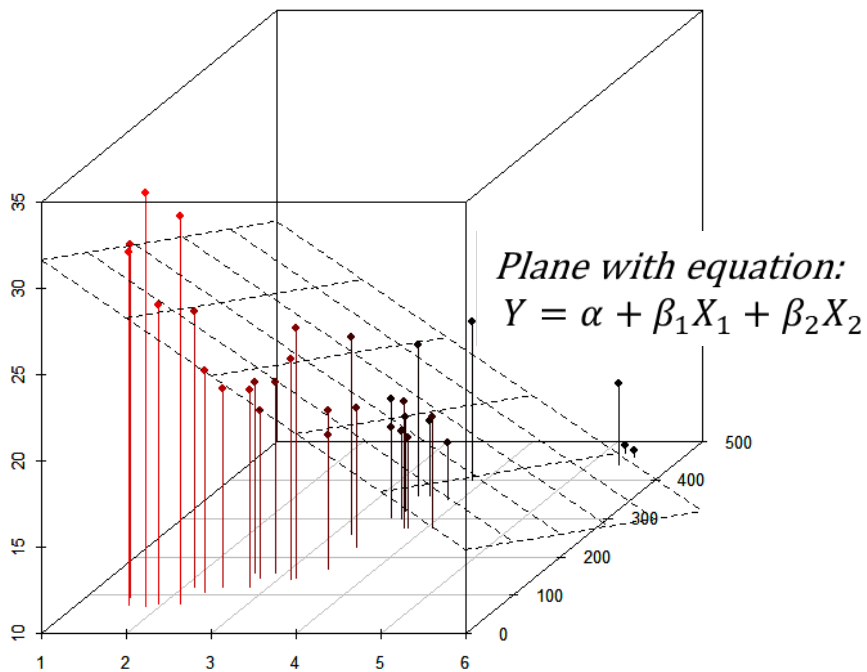
Multiple linear regression

Linear regression with two continuous covariates x_1 and x_2:

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2}.$$

- Now the regression line isa plane through points in 3D space.
- Least squares distances are from the points to the surface of the plane.





http://www.statmethods.net/graphs/scatterplot.html

Matrix notation

Writing the least squares equations in matrix notation allows for derivation of parameters when there is more than one explanatory variable.

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \ \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{N2} & \dots & x_{Np} \end{bmatrix}, \ \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \ \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix}.$$

Linear equation: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.

- Matrix notation is much easier to code for computational estimation of regression parameters for arbitrary models.
- The same matrix algebra code works for 1 covariate or 100 covariates.

Computing least squares regression parameters

Residual sum of squares:

$$\begin{aligned}\boldsymbol{S}(\boldsymbol{\beta}) &= \sum_{i=1}^{N} \epsilon_i^2 = \boldsymbol{\epsilon}^T \boldsymbol{\epsilon} \\ &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= \boldsymbol{y}^T \boldsymbol{y} - 2\boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{y} + \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{X}\boldsymbol{\beta}.\end{aligned}$$

To minimize the SS_{res}, take the derivative and set equal to 0:

$$\begin{aligned}&\frac{d}{d\boldsymbol{\beta}}\boldsymbol{S} = -2\boldsymbol{X}^T \boldsymbol{y} + 2\boldsymbol{X}^T \boldsymbol{X}\boldsymbol{\beta} = \mathbf{0}. \\ &\Rightarrow \widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T \boldsymbol{X}\right) -1\boldsymbol{X}^T \boldsymbol{y} \text{ and } \widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}.\end{aligned}$$

Statistical inference

- Notice that we did not need a model to estimate the least squares regression line. It was simply geometry.
- Least squares regression line allowed us to quantify the relationship.
- But what if we want to make inference about the relationship?
  - That is, is the effect of LSD concentration on math test scores statistically significant?

Statistical inference is drawing conclusions about a population or parameter based on observed data.

Statistical inference requires a model

We want to know if the variable x has an effect on the outcome Y.

- In statistical terms: is Y associated with x?

Linear model: $Y_i = \alpha + \beta x_i$.

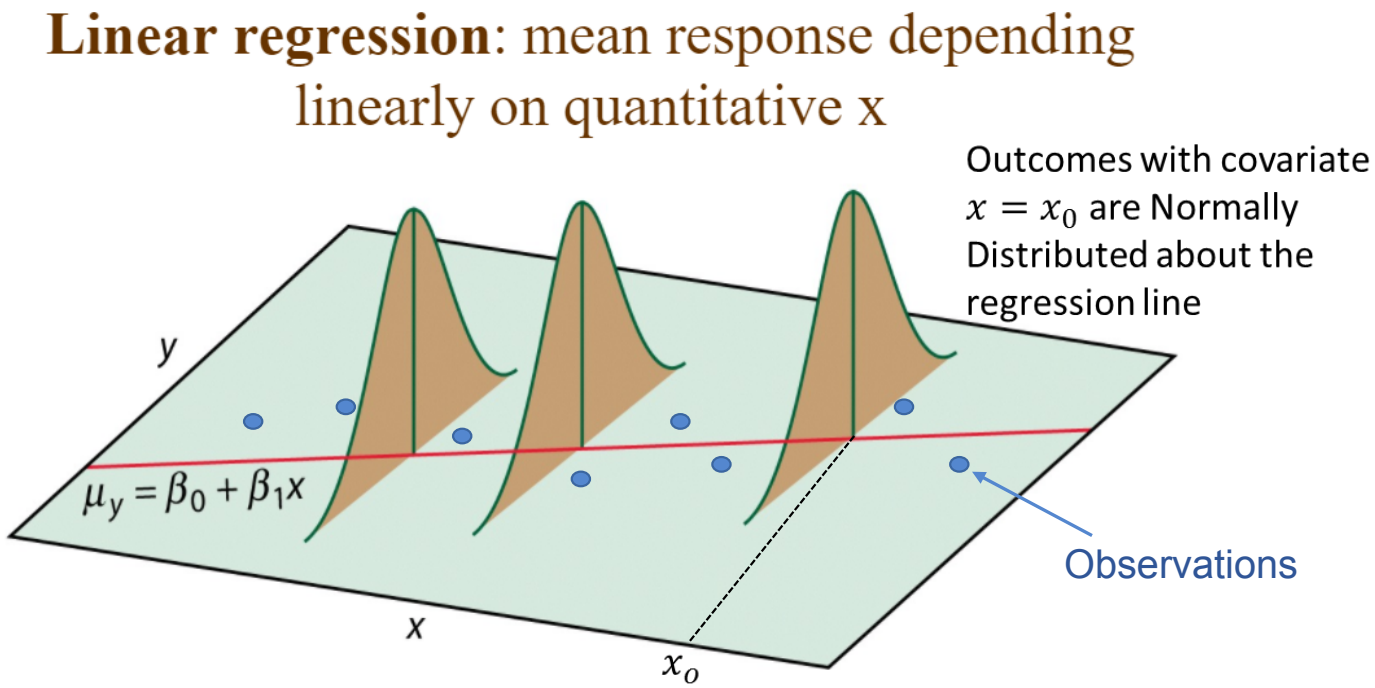
- No association would correspond to the slope parameter $\beta = 0$ in the linear regression equation.
- The β estimated from data is almost certainly no exactly zero.
- Is β non-zero because of a true association between x and Y, or simply because of unrelated variation in the outcome Y?
- We need a model to answer that question...

Linear regression statistical model

Define $Y_i = \alpha + \beta x_i + \epsilon_i$ to be the linear regression model that relates outcomes Y to covariates x.

- ϵ_i is called the **residual**, and is the quantity by which the observed value differs from the proposed model.
- Assume that $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, that is, the residuals are normally distributed with mean zero and constant variance.
- Based on properties of the normal distribution, for a fixed covariate value x_i, the corresponding outcome is distributed as $Y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2)$.
- Goal is to formally test if the slope parameter β is zero, that is we want to test the null hypothesis $H_0 : \beta = 0$.

Linear regression statistical model



- x_i is the independent variable, i.e. it is not random.
- Y_i is the dependent variable, it is random: $Y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2)$:
  - The expected value of Y is a linear function of x, but for fixed x, the variable Y differs from its expected value by a random amount.

Linear regression – likelihood approach

Given the Normal distribution assumption on the residuals, we can write a likelihood function for the observed data Y_i as follows:

- If each $Y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2)$, then the likelihood function of all the observed data is

$$\mathcal{L}(\beta, \alpha | \boldsymbol{Y}) = \prod_{i=1}^{N} \mathcal{L}(\beta, \alpha | Y_i),$$

- and the log-likelihood function is

$$\ell(\beta, \alpha | \boldsymbol{Y}) = \sum_{i=1}^{N} \log \mathcal{L}(\beta, \alpha | Y_i).$$

Linear regression – likelihood approach

- For each Y_i, the log-likelihood function is based on the pdf of a Normal random variable:

$$\mathcal{L}(\beta, \alpha | Y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(Y_i - \alpha - \beta x_i)^2}{2\sigma^2} \right\}.$$

- Then the likelihood for the full data is

$$\begin{aligned} \ell(\beta, \alpha | \boldsymbol{Y}) &= \sum_{i=1}^{N} -\frac{1}{2}\log(2\pi\sigma^2) - \frac{Y_i - \alpha - \beta x_i)^2}{2\sigma^2} \\ &= -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(Y_i - \alpha - \beta x_i)^2. \end{aligned}$$

Linear regression – likelihood approach

Given the log-likelihood function, we can compute the Maximum Likelihood Estimators (MLEs) for the parameters α, β and σ^2. Note that maximizing the log-likelihood

$$\ell(\beta, \alpha | \boldsymbol{Y}) = -\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (Y_i - \alpha - \beta x_i)^2$$

is equivalent to minimizing

$$\sum_{i=1}^{N} (Y_i - \alpha - \beta x_i)^2.$$

This is exactly the least squares regression problem!

- The least squres regression solutions are the same as the MLEs.
- Maximum likelihood estimation has several nice large sample properties that acommodate hypothesis testing.

Linear regression – analysis of variance

Analysis of Variance (ANOVA) partitions total variation in the observed outcomes into variation explained by a model and variation explained by random error.
For each observation, $Y_i - \bar{Y} = (\widehat{Y}_i - \bar{Y}) + (Y_i - \widehat{Y}_i)$.
Squaring, summing, and some algebra gives:

$$\begin{aligned} \sum_{i=1}^{N}(Y_i - \bar{Y})^2 &= \sum_{i=1}^{N}(\widehat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{N}(Y_i - \widehat{Y}_i)^2 \\ SS_T &= SS_R + SS_{res} \end{aligned}$$

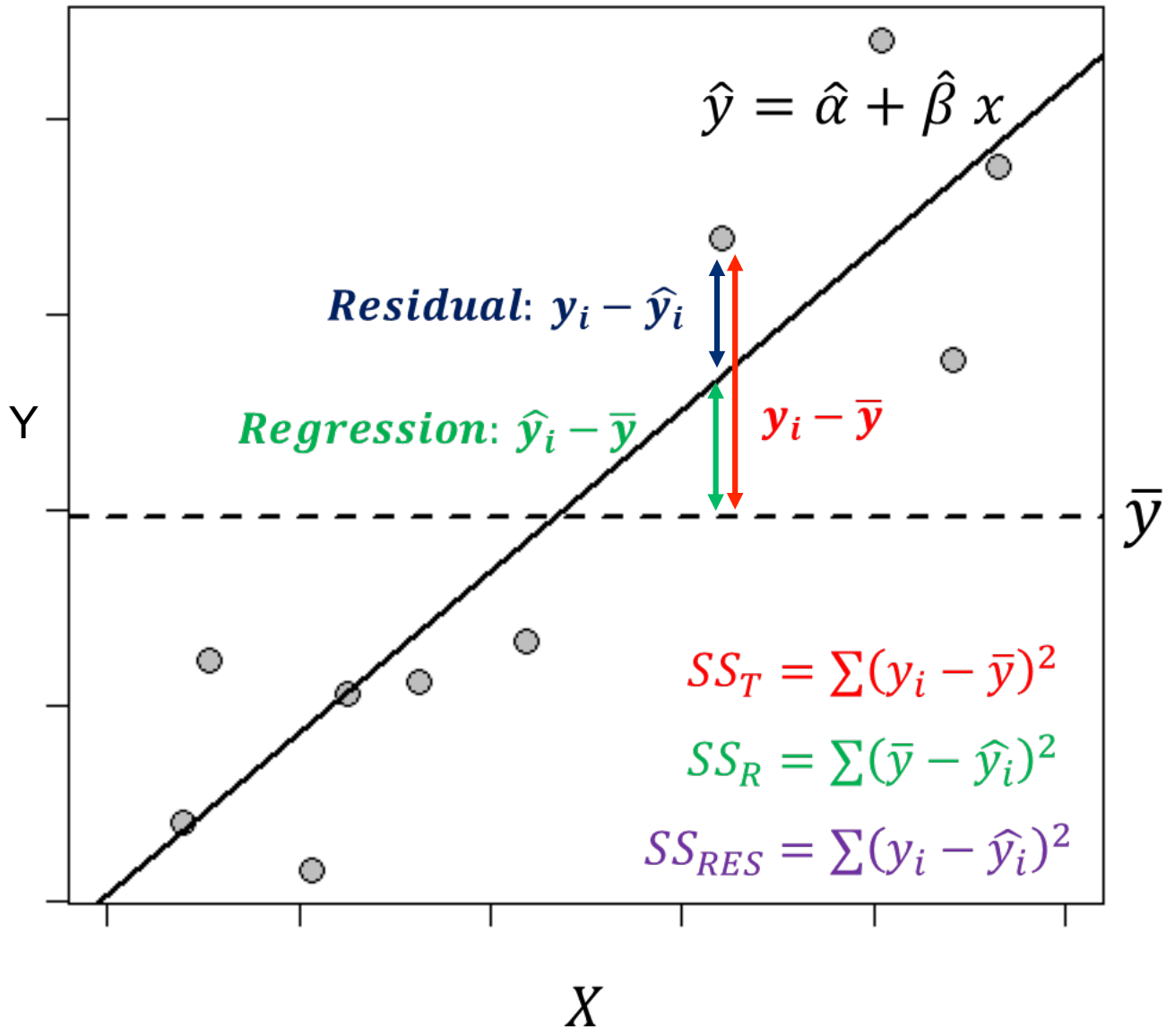
where SS_T is the Total Sum os Squares, SS_R is the Regression Sum of Squares, and SS_{res} is the Residual Sum of Squares.

ANOVA

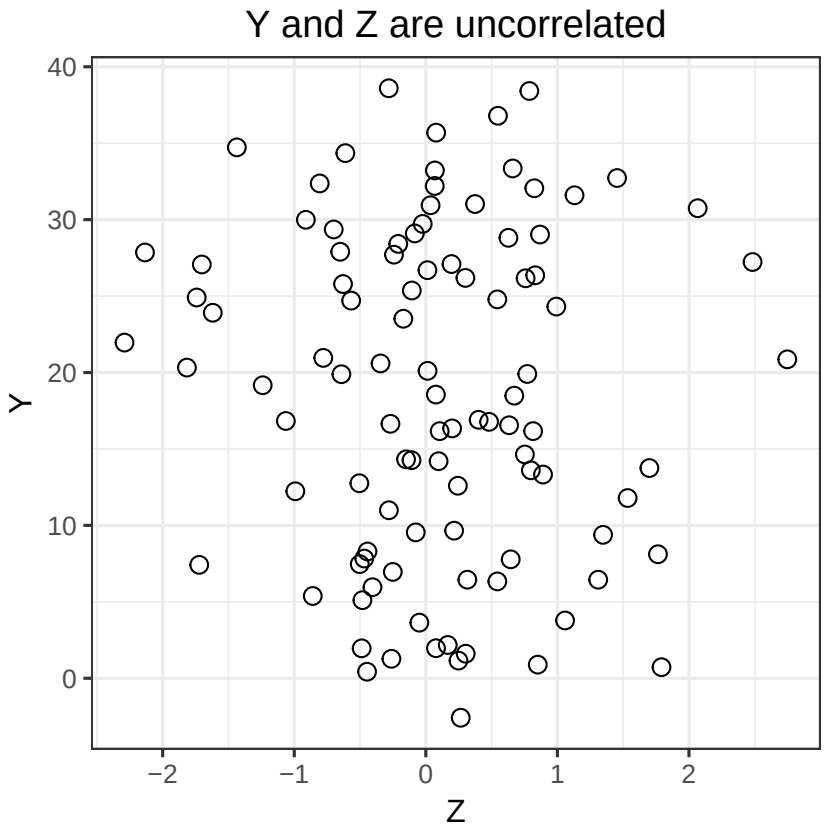
ANOVA table for testing the null hypothesis $H_0 : \beta = 0$

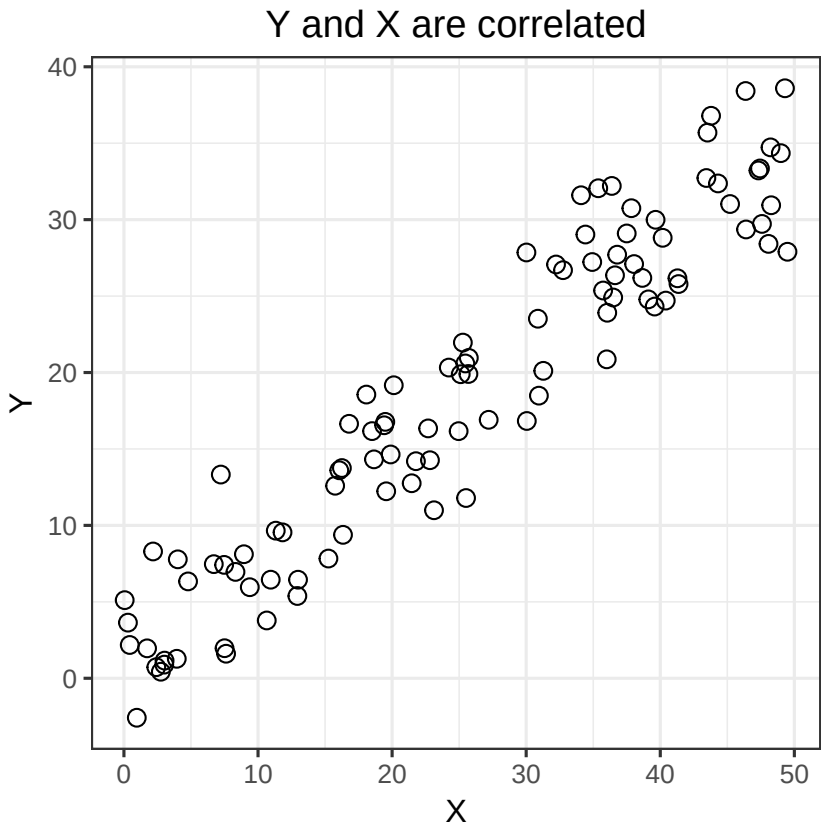
Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	$F_0 \sim F_{1,N-2}$
Regression	S_R	1	$SS_R/1$	$\frac{SS_R/1}{SS_{res}/(N-2)}$
Residual	SS_{res}	$N-2$	$SS_{res}/(N-2)$	
Total	SS_T	$N-1$		

ANOVA – graphical interpretation

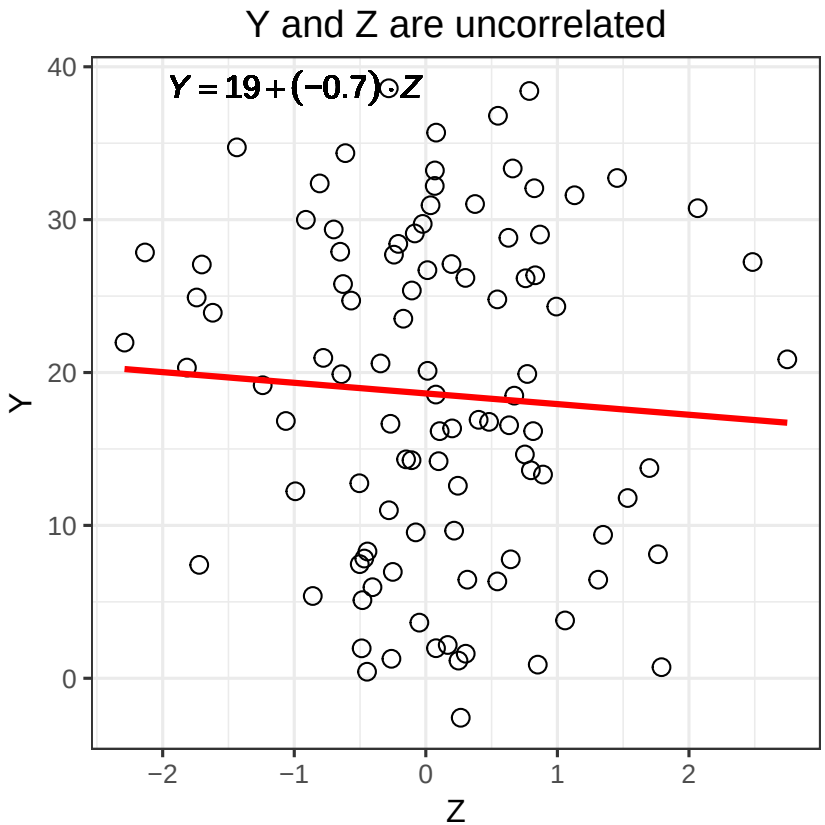


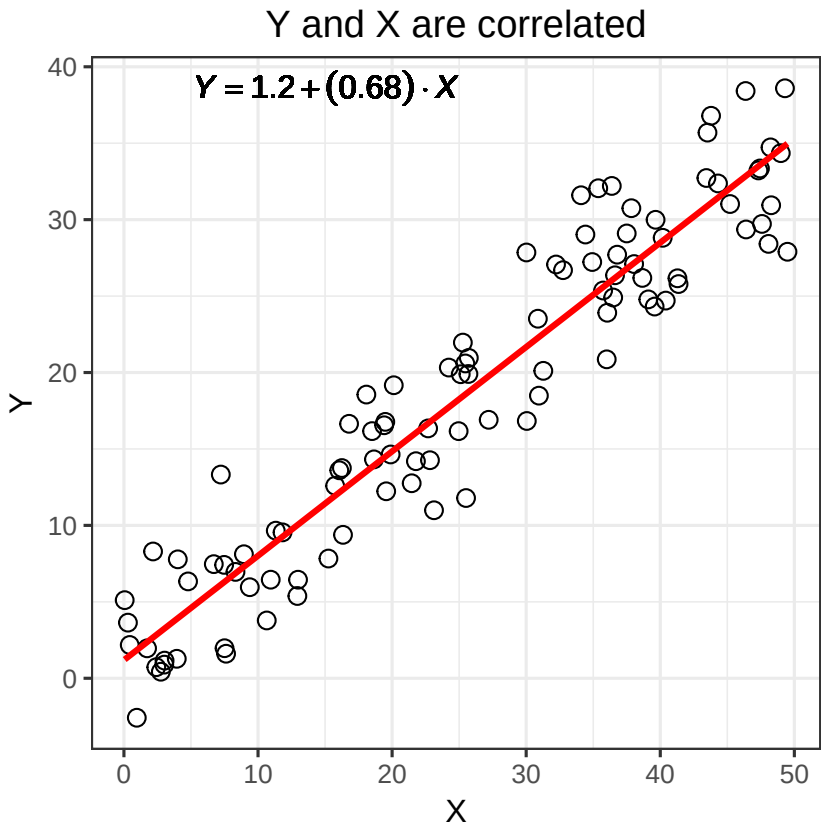
ANOVA – graphical interpretation



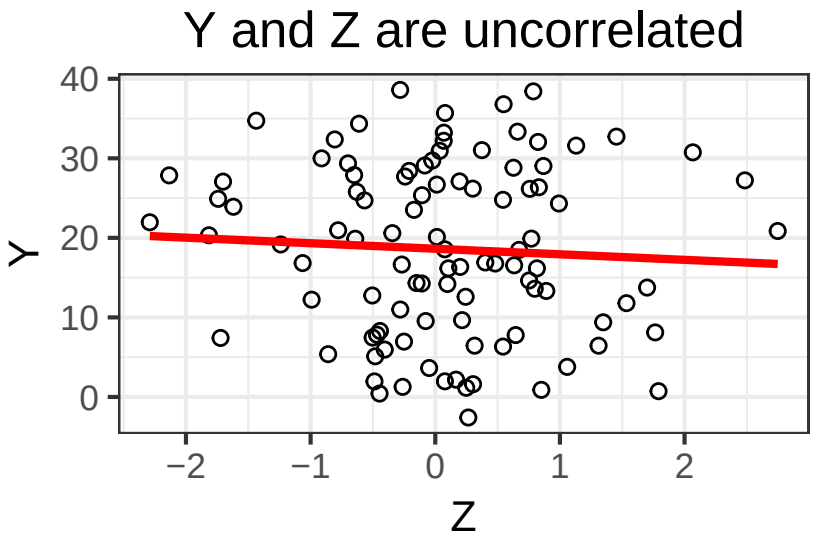


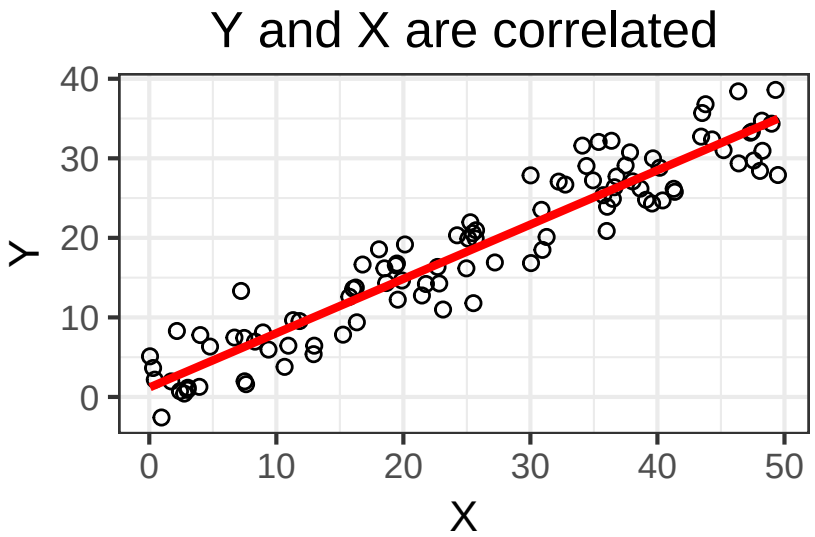
ANOVA – graphical interpretation





ANOVA – graphical interpretation





Source	Sum of Squares	F_0
SS_R	42.47	0.36
SS_{res}	1.2×10^4	
SS_T	1.2×10^4	$p = 0.55$

Source	Sum of Squares	F_0
SS_R	10^4	890.54
SS_{res}	1148.1	
SS_T	1.2×10^4	$p = 6 \times 10^{-51}$

Asumptions of linear regression

We have made multiple assumptions about the outcomes along the way.

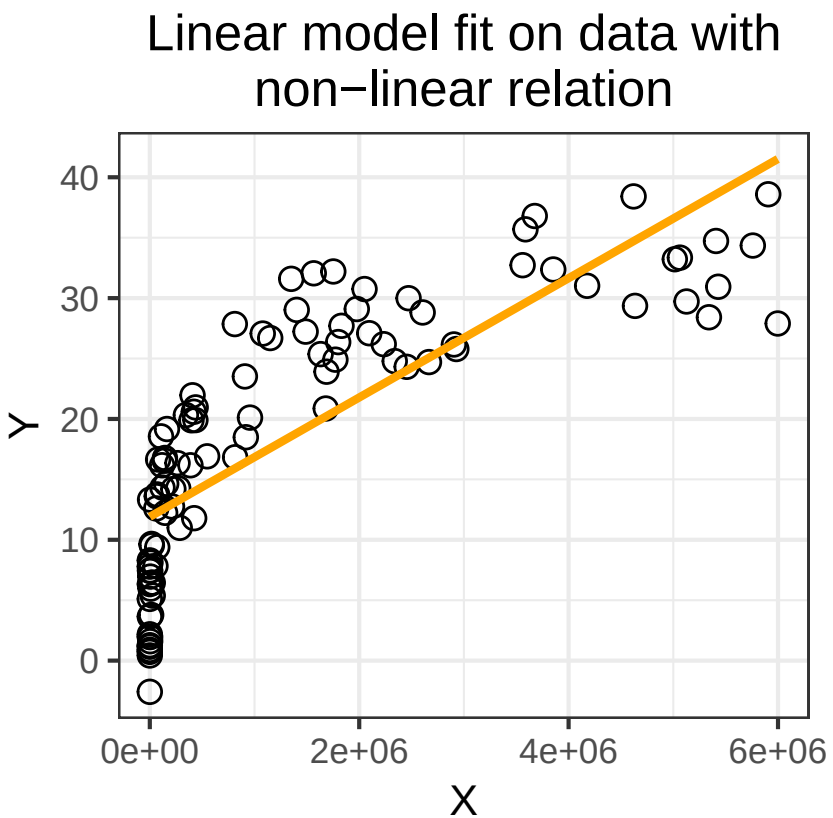
- What exactly are they?
- How do we check them?
- What if they are violated?
  - Inference can be incorrect.
  - Extensions to simple linear regression

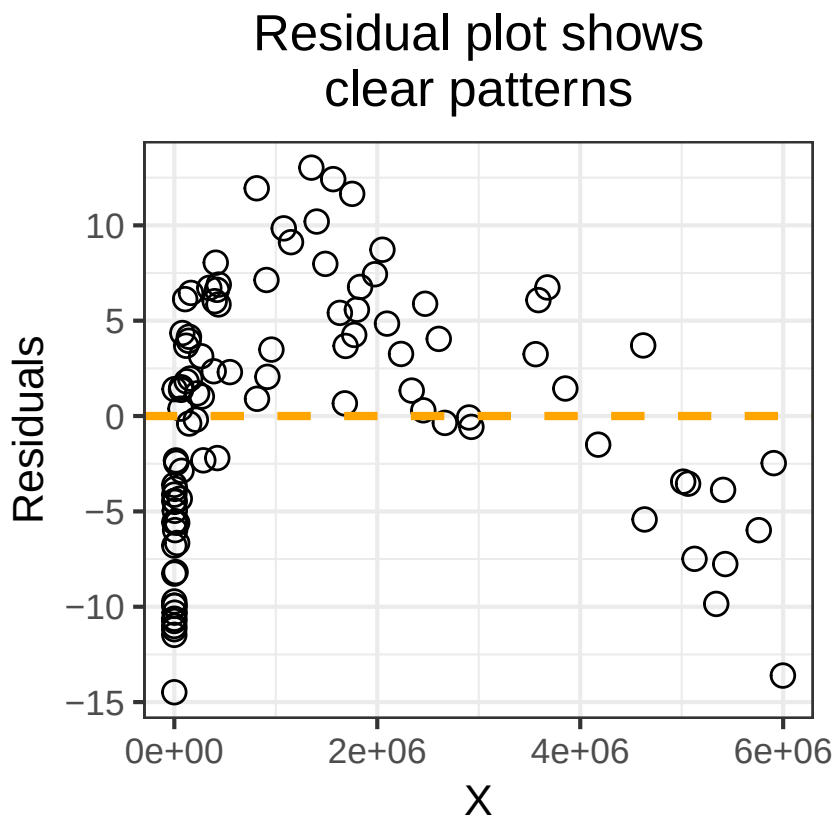
Assumptions:

1. Independence between observations.
2. Linearity between covariate and outcome.
3. Constant variance of errors (homoscedasticity).
4. Normally distributed errors.

Checking linearity assumption

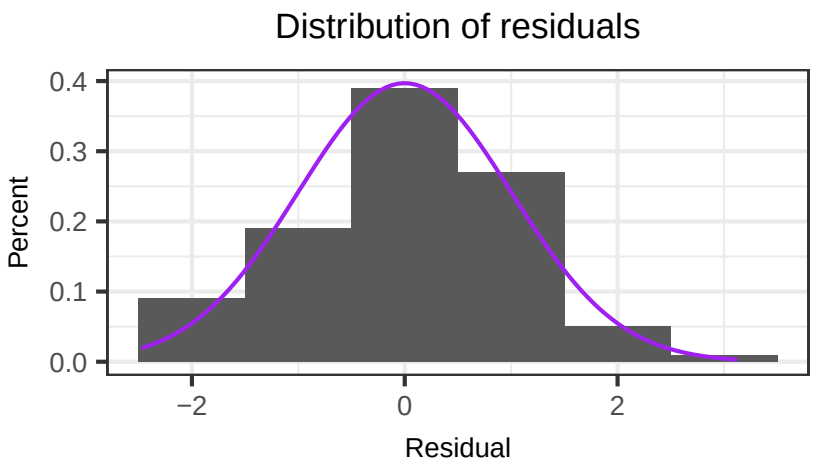
- We are assuming a linear relationship between the outcome and covariate.
- Residual plot (residuals vs X) should look like a random cloud of points around the line $Y = 0$; Any pattern may indicate non-linearity.
- Try adding non-linear terms of the covariate (e.g. x^2, $logx$, ...)

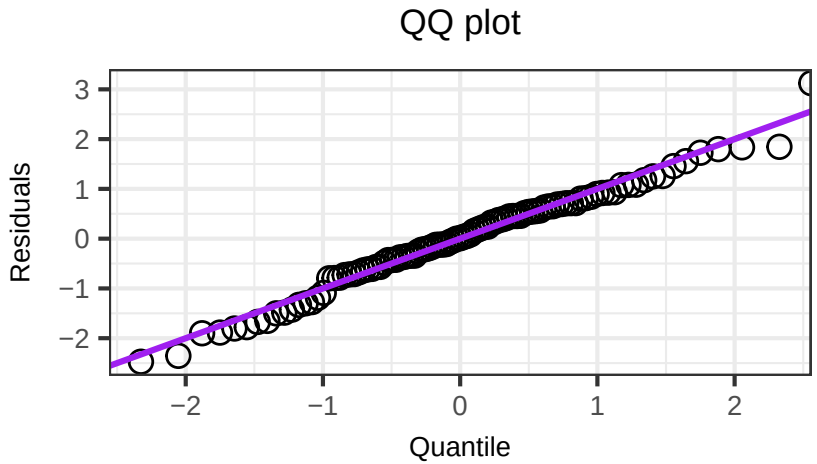


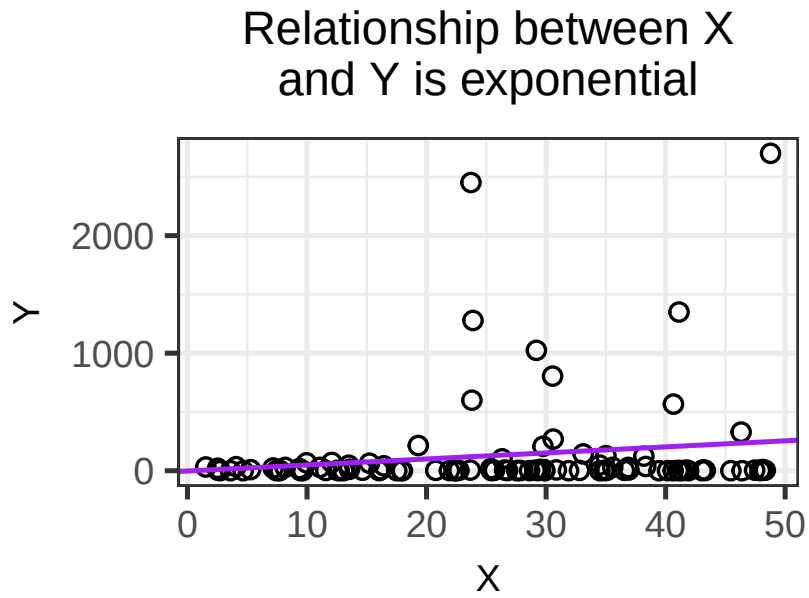


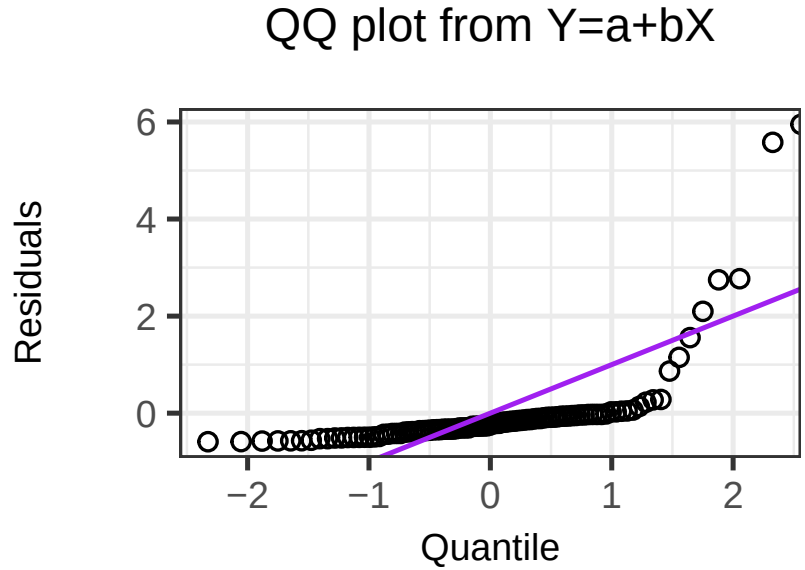
Checking normality assumption

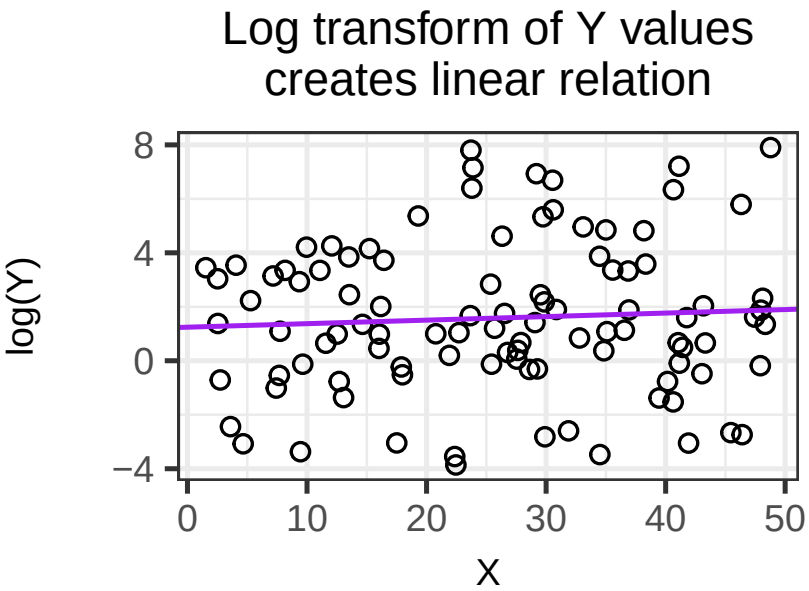
- We assume that the residuals are normally distributed.
- Can check normality assumption using a quantile-quantile plot of the residuals (QQ plot).
- A QQ plot is an ordered set of residuals plotted against the quantiles of a theoretical distribution.
- If the points follow the theoretical distribution, the points should form a line along $y = x$.
- A transformation of the outcome can often correct violations of the normality assumption (e.g. $logY$).

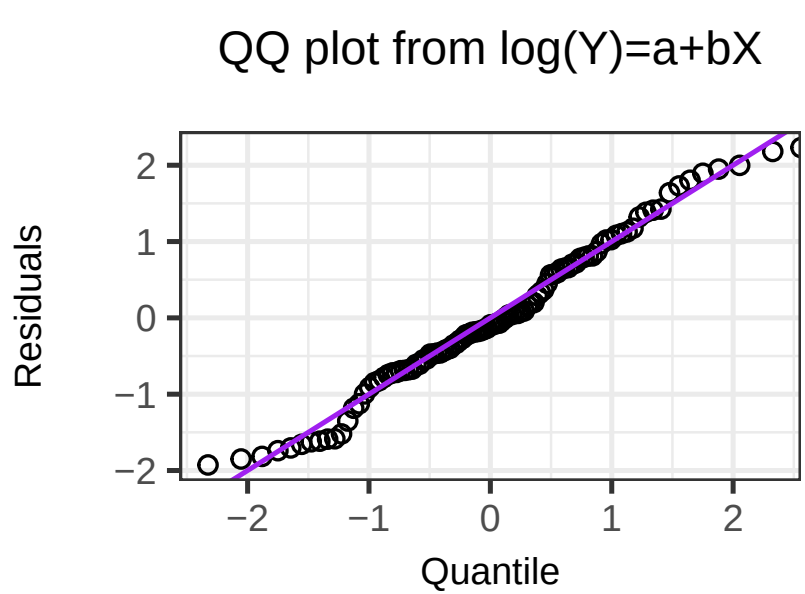





Relationship between X
and Y is exponential
Y
2000
1000
0
0
10
20
30
40
50
X

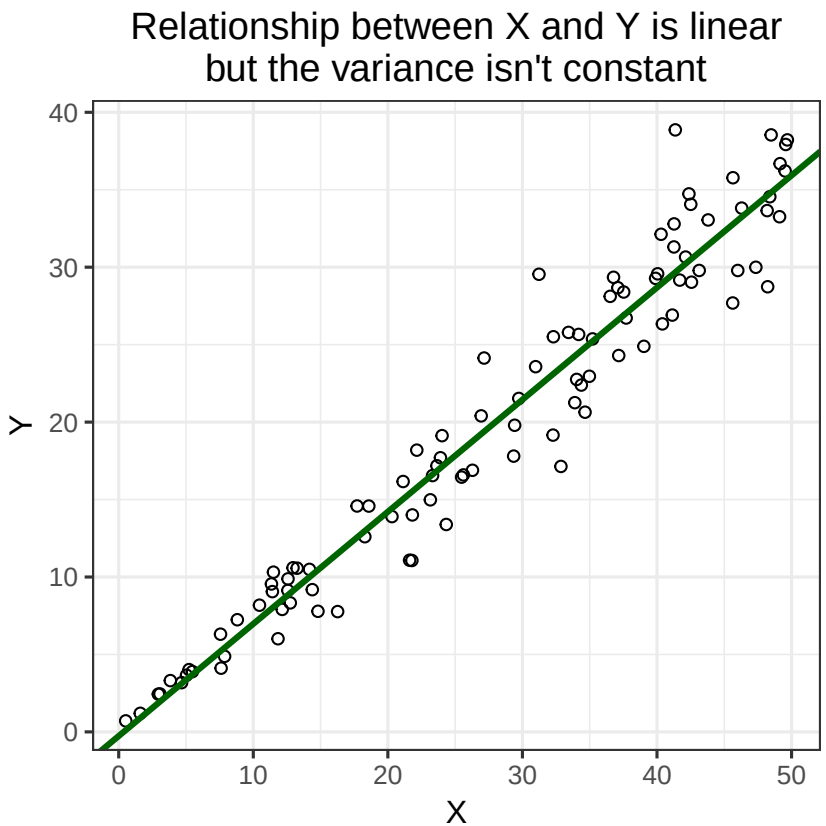

QQ plot from Y=a+bX
Residuals
6
4
2
0
−2
−1
0
1
2
Quantile

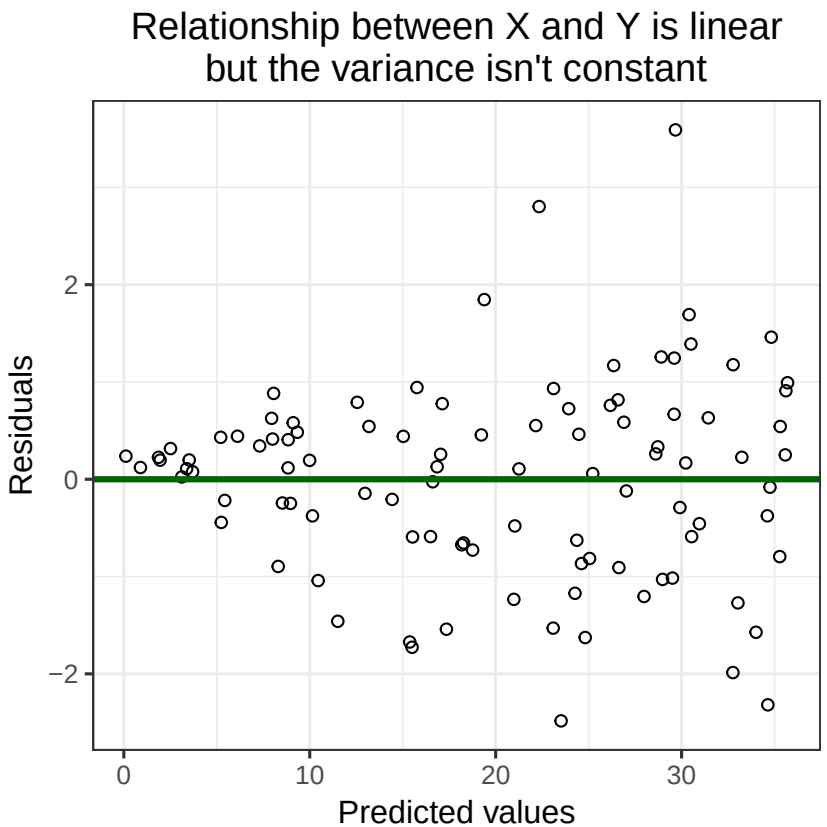

Log transform of Y values
creates linear relation
log(Y)
8
4
0
−4
0
10
20
30
40
50
X


QQ plot from log(Y)=a+bX
Residuals
2
1
0
−1
−2
−2
−1
0
1
2
Quantile

Checking constant variance assumption

- We assume that the residuals have constant variance across values of X.
- Look at plots of residuals versus predicted values, and plot of residuals vs. the covariate X.
- Use either a variable transformation or **weighted least squares regression** to fix.





Checking independence assumption

- We assume that the outcomes are independent.
- Usually this is determined by the experimental design:
  - Are measurements taken from the same people?
  - Are samples recruited from different hospitals? Are measurements from the same hospital more similar than measurements from different hospitals?
- Ignoring correlation between observations can lead to improper inference (your estimate of variance is wrong)
- **Linear mixed models** and **generalized estimating equations** allow you to take into account the correlation of the outcomes.

Summary

- Reviewed the basics of simple linear regression.

Hopefully, this lecture

- Improved your intuition on the theory behind simple linear regression.
- Made connections to ANOVA, likelihood theory, extensions to simple linear regression.