Linear correlation and linear regression

Recall: Covariance

$$\mathrm{cov}(x,y) = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{n-1}$$

Interpreting Covariance

cov(X,Y) > 0 $\longrightarrow$ X and Y are positively correlated

cov(X,Y) < 0 $\longrightarrow$ X and Y are inversely correlated

cov(X,Y) = 0 $\longrightarrow$ X and Y are independent

Correlation coefficient

- Pearson's Correlation Coefficient is standardized covariance (unitless):

$$r = \frac{\mathrm{cov}\,ariance(x, y)}{\sqrt{\mathrm{var}\,x}\,\sqrt{\mathrm{var}\,y}}$$
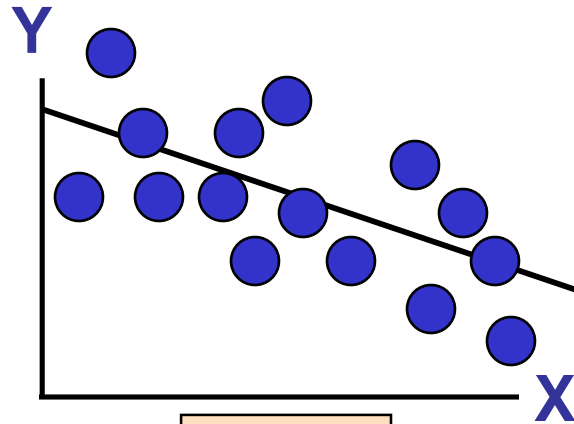
Correlation

- Measures the relative strength of the *linear* relationship between two variables

- Unit-less

- Ranges between −1 and 1

- The closer to −1, the stronger the negative linear relationship

- The closer to 1, the stronger the positive linear relationship

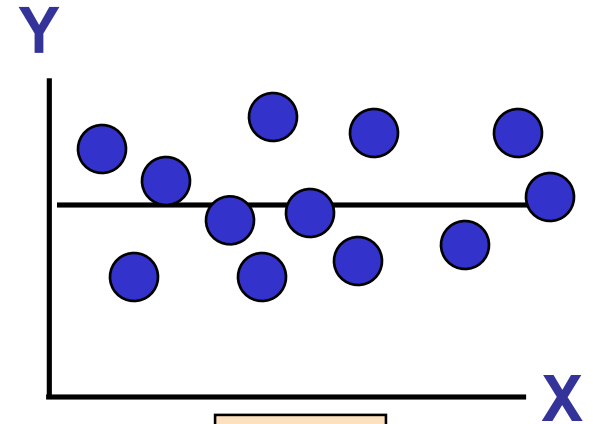- The closer to 0, the weaker any positive linear relationship

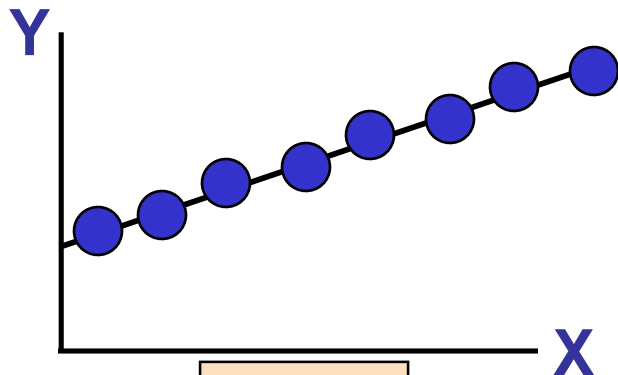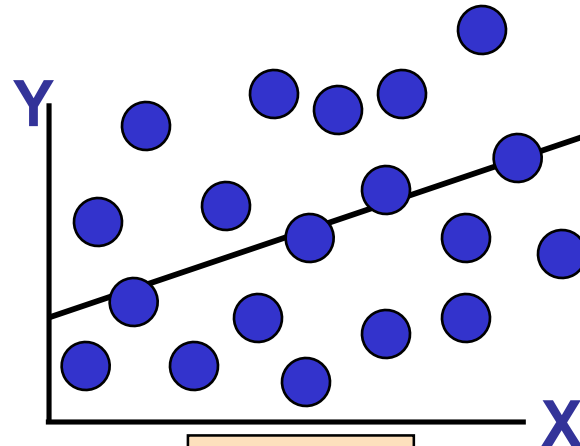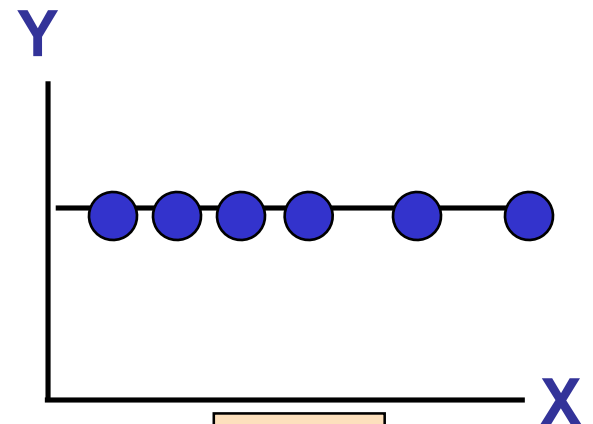# Scatter Plots of Data with Various Correlation Coefficients
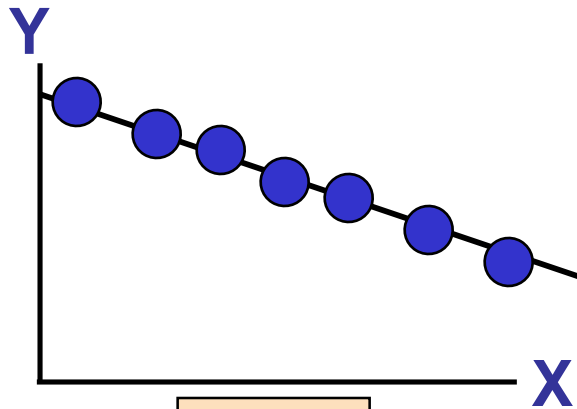
r = -1

r = -.6

r = 0

r = +1
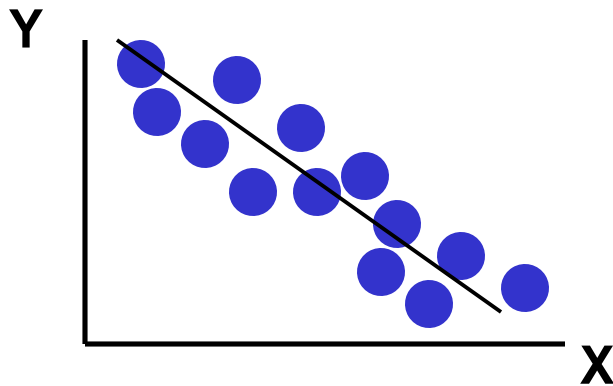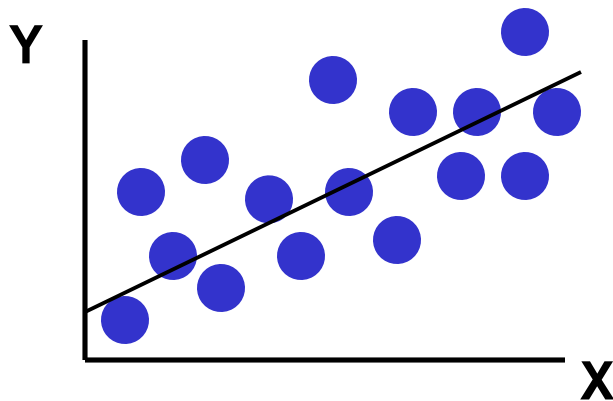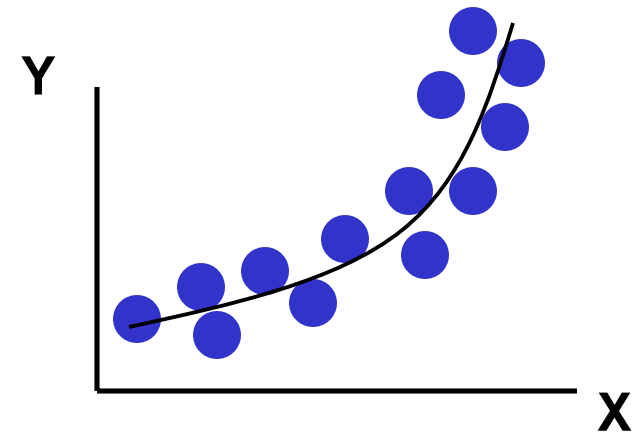
r = +.3

r = 0

Linear Correlation

Linear relationships	Curvilinear relationships
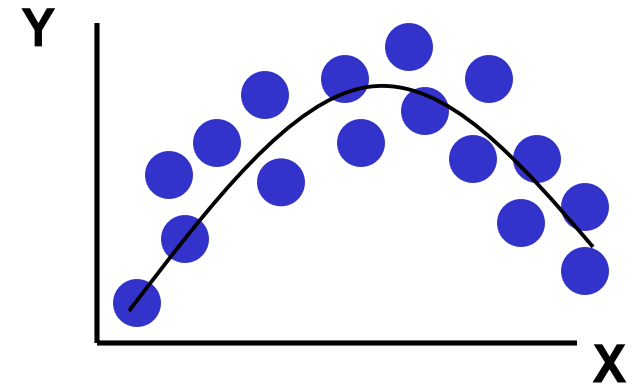
Linear Correlation

Strong relationships
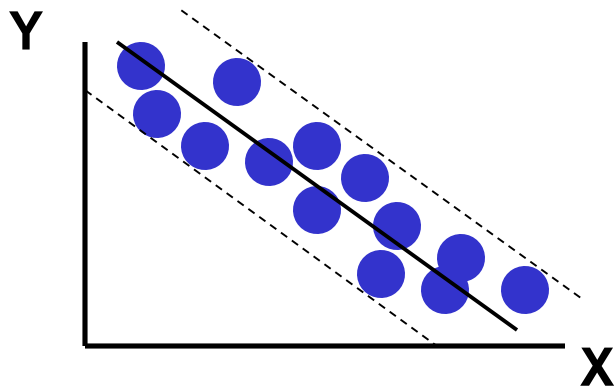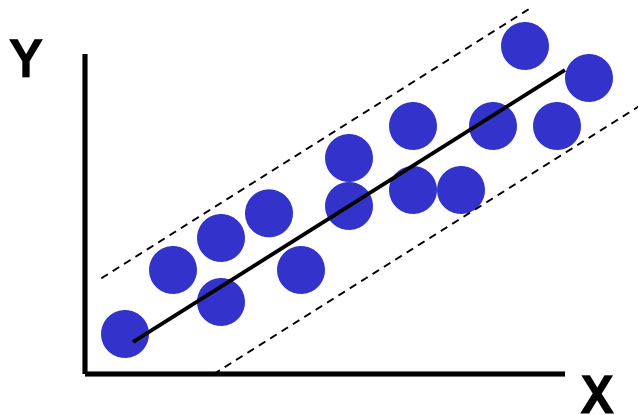
Weak relationships
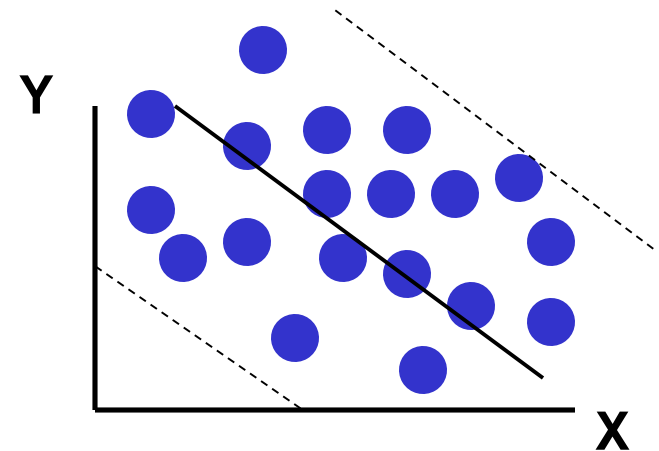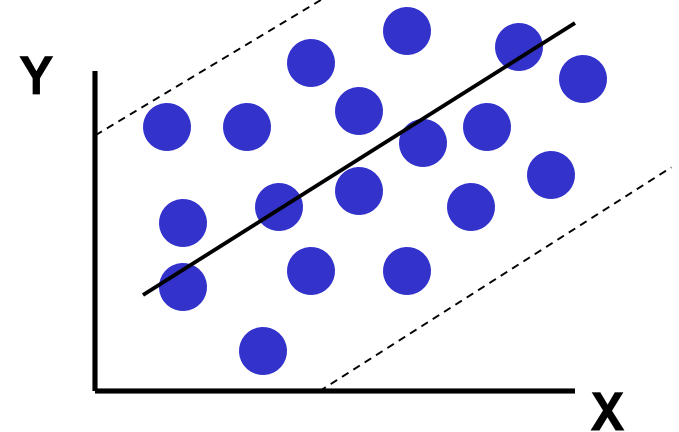
Linear Correlation

No relationship
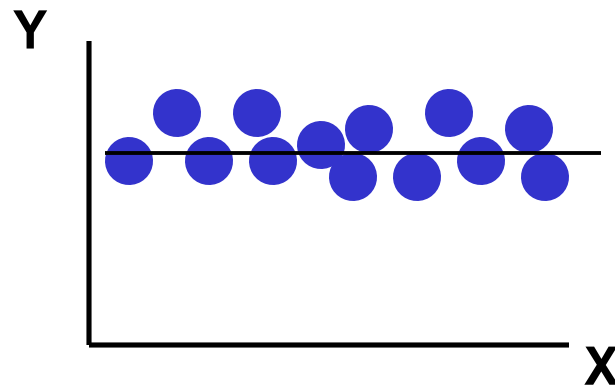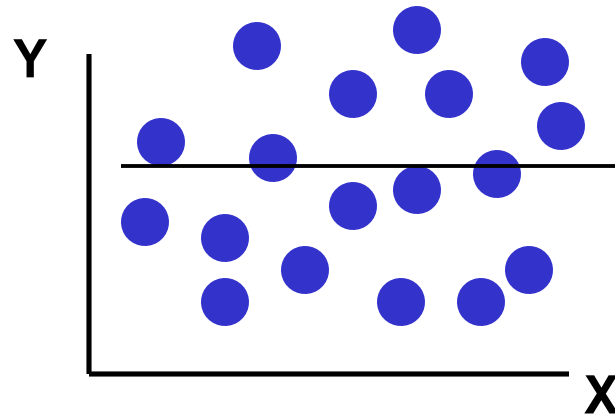
Calculating by hand...

$$\hat{r} = \frac{\text{cov}\,ariance(x, y)}{\sqrt{\text{var } x}\sqrt{\text{var } y}} = \frac{\dfrac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1}}{\sqrt{\dfrac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}\sqrt{\dfrac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n-1}}}$$

Simpler calculation formula…

$$\hat{r} = \frac{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}}{\sqrt{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}\sqrt{\dfrac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}} =$$

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{SS_{xy}}{\sqrt{SS_x SS_y}}$$

$$\hat{r} = \frac{SS_{xy}}{\sqrt{SS_x SS_y}}$$

Numerator of covariance

Numerators of variance

Linear regression

In correlation, the two variables are treated as equals. In regression, one variable is considered independent (=predictor) variable (X) and the other the dependent (=outcome) variable Y.

What is "Linear"?

- Remember this:

- *Y=mX+B?*
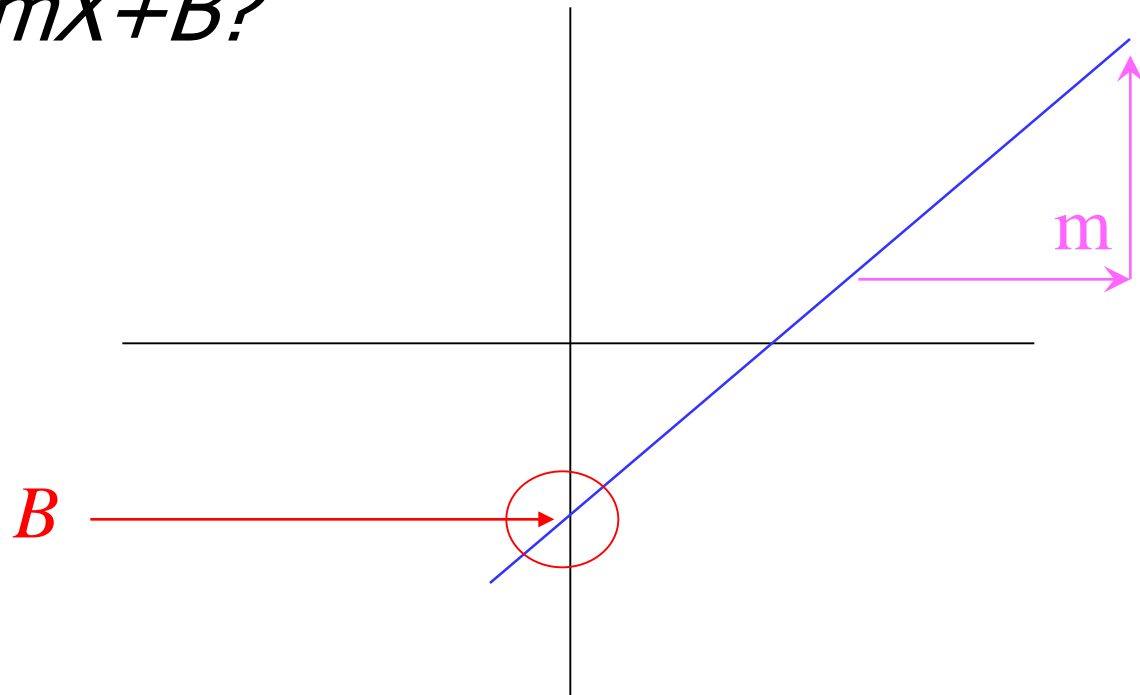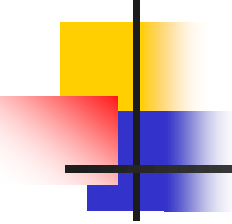
What's Slope?

A slope of 2 means that every 1-unit change in X yields a 2-unit change in Y.

Predicted value for an individual…

$$\hat{y}_i = \quad \alpha + \beta * x_i \quad + \quad \boxed{\text{random error}_i}$$

Fixed – exactly on the line

Follows a normal distribution

Assumptions (or the fine print)

- Linear regression assumes that…
    - 1. The relationship between X and Y is linear
    - 2. Y is distributed normally at each value of X
    - 3. The variance of Y at every value of X is the same (homogeneity of variances)
    - 4. The observations are independent

The standard error of Y given X is the average variability around the regression line at any given value of X. It is assumed to be equal at all values of X.
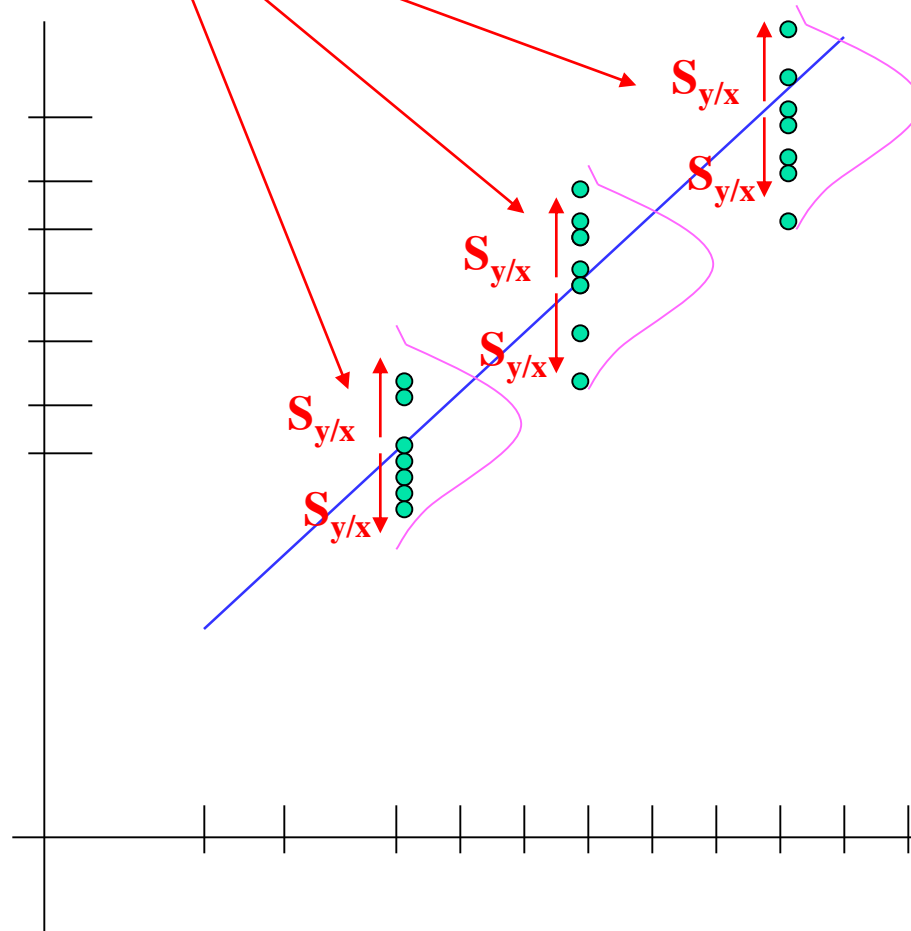
Regression Picture

$$\hat{y}_i = \beta x_i + \alpha$$
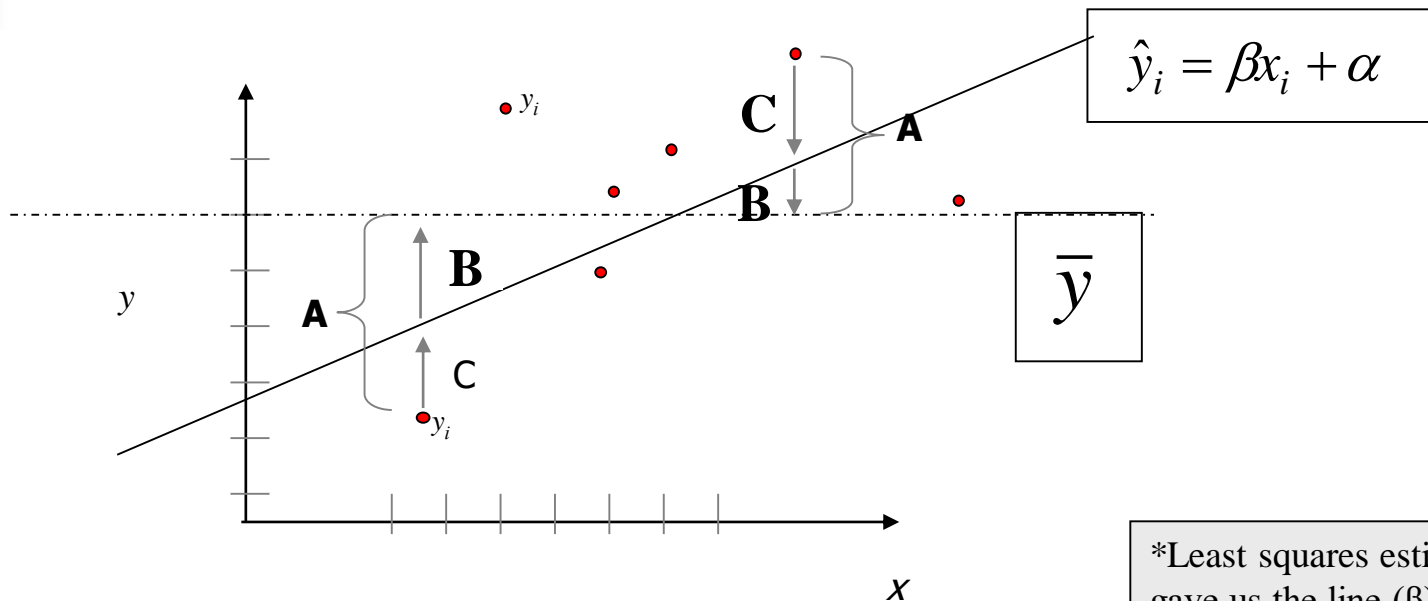
$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

*Least squares estimation gave us the line (β) that minimized C^2

$R^2 = SS_{reg}/SS_{total}$

A^2
SS_{total}
Total squared distance of observations from naïve mean of y
Total variation

B^2
SS_{reg}
Distance from regression line to naïve mean of y
Variability due to x (regression)

C^2
$SS_{residual}$
Variance around the regression line
Additional variability not explained by x—what least squares method aims to minimize

Estimating the intercept and slope: least squares estimation

** Least Squares Estimation

A little calculus....

What are we trying to estimate? **β, the slope,** from

What's the constraint? We are trying to minimize the squared distance (hence the "least squares") between the observations themselves and the predicted values , or   (<u>also called the "residuals", or left-over unexplained variability</u>)

$\text{Difference}_i = y_i - (\beta x + \alpha)$ $\text{Difference}_i^2 = (y_i - (\beta x + \alpha))^2$

Find the β that gives the minimum sum of the squared differences. How do you maximize a function? Take the derivative; set it equal to zero; and solve. Typical max/min problem from calculus....

$$\frac{d}{d\beta} \sum_{i=1}^{n} (y_i - (\beta x_i + \alpha))^2 = 2(\sum_{i=1}^{n} (y_i - \beta x_i - \alpha)(-x_i))$$

$$2(\sum_{i=1}^{n} (-y_i x_i + \beta x_i^2 + \alpha x_i)) = 0...$$

From here takes a little math trickery to solve for β...

Resulting formulas...

Slope (beta coefficient) =

$$\hat{\beta} = \frac{Cov(x, y)}{Var(x)}$$

Intercept=

$$\text{Calculate}: \hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}$$

Regression line always goes through the point: $(\overline{x}, \overline{y})$

Relationship with correlation

$$\hat{r} = \hat{\beta}\,\frac{SD_x}{SD_y}$$

In correlation, the two variables are treated as equals. In regression, one variable is considered independent (=predictor) variable (X) and the other the dependent (=outcome) variable Y.

Residual Analysis: check assumptions

$$e_i = Y_i - \hat{Y}_i$$

- The residual for observation i, e_i, is the difference between its observed and predicted value

- Check the assumptions of regression by examining the residuals

  - Examine for linearity assumption
  - Examine for constant variance for all levels of X (homoscedasticity)
  - Evaluate normal distribution assumption
  - Evaluate independence assumption

- Graphical Analysis of Residuals

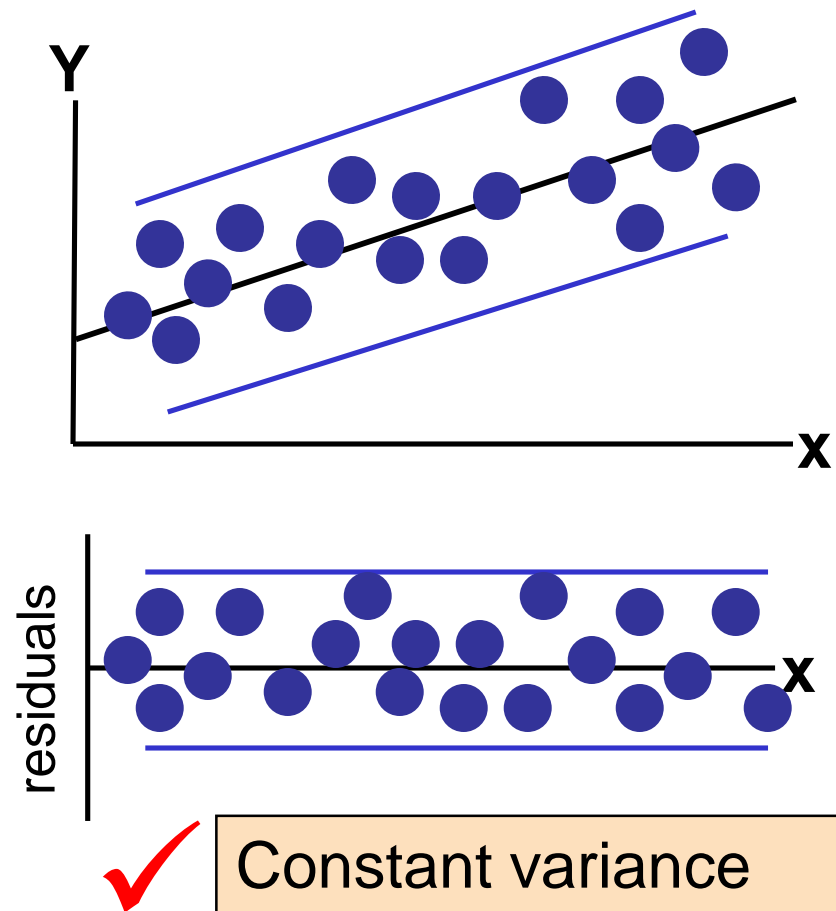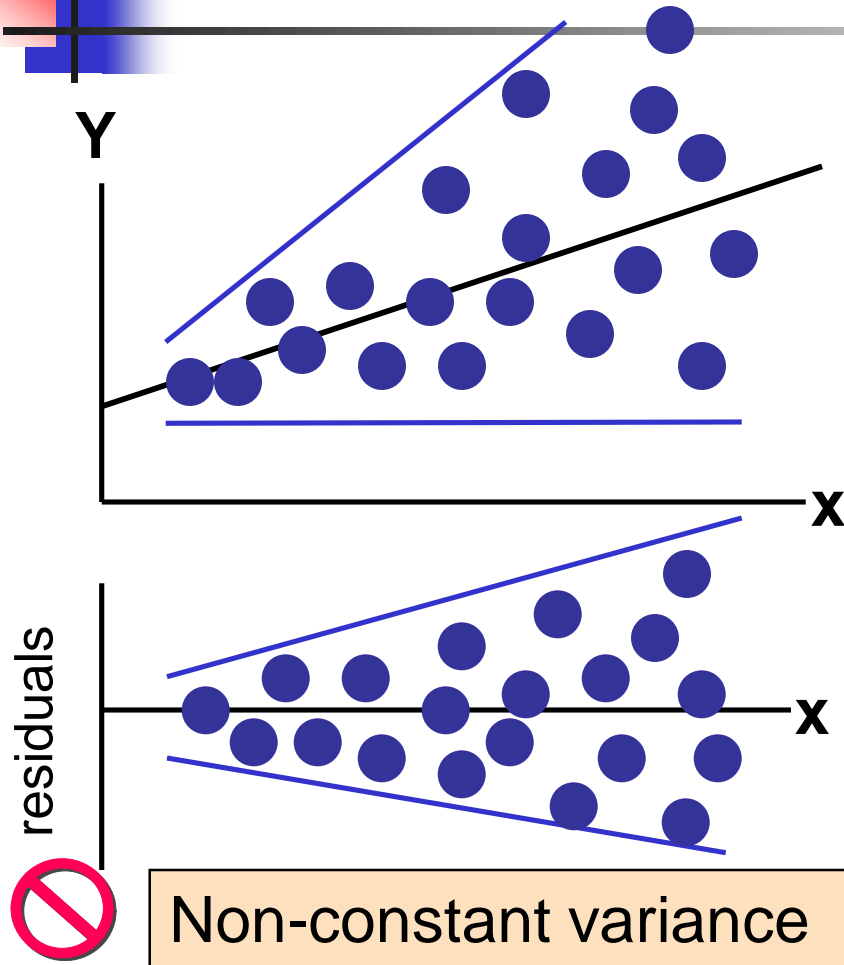 - Can plot residuals vs. X
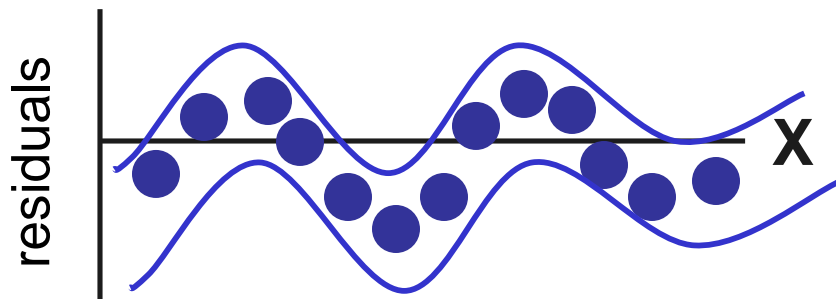
Residual Analysis for Linearity
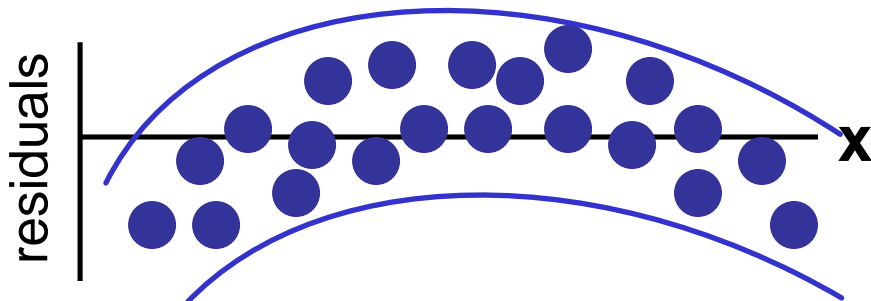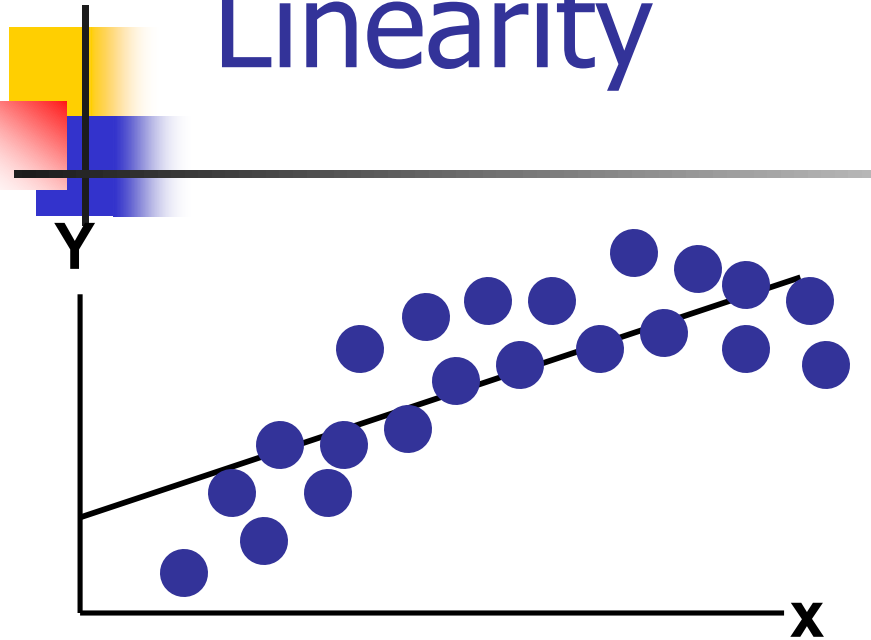
Not Linear
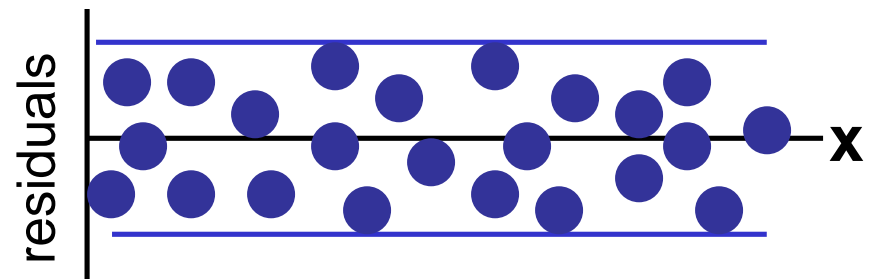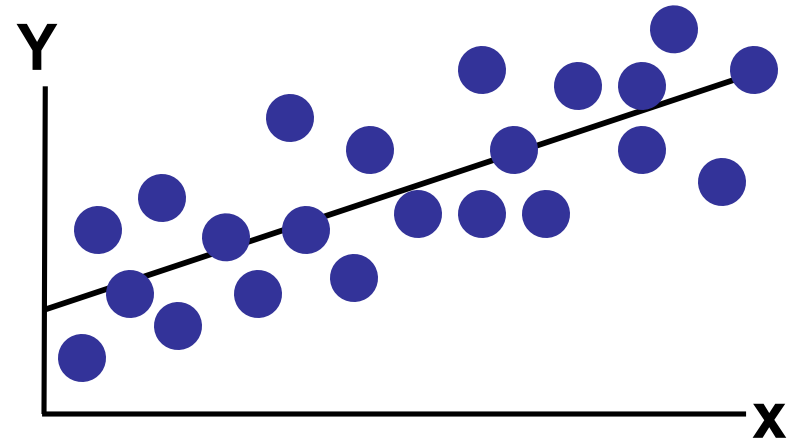
Linear

Residual Analysis for Homoscedasticity

Non-constant variance

Constant variance
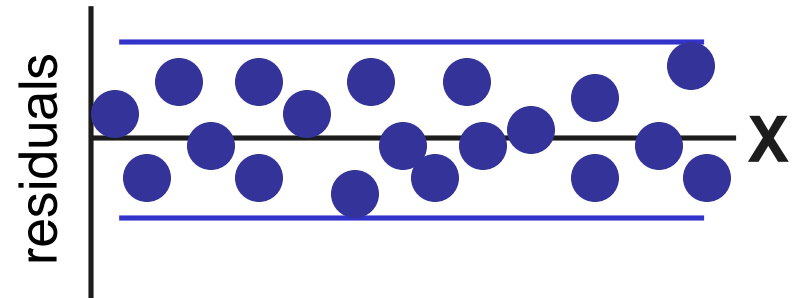
Residual Analysis for Independence

Not Independent
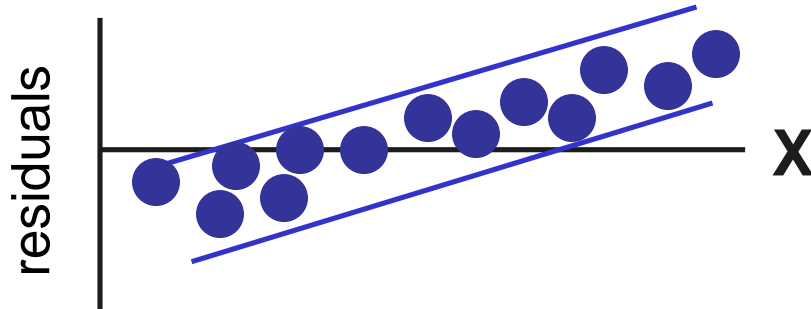
Independent

Multiple Linear Regression

- More than one predictor…

$$E(y) = \alpha + \beta_1 * X + \beta_2 * W + \beta_3 * Z \ldots$$

Each regression coefficient is the amount of change in the outcome variable that would be expected per one-unit change of the predictor, if all other variables in the model were held constant.

Other types of multivariate regression

- Multiple linear regression is for normally distributed outcomes

- Logistic regression is for binary outcomes

- Cox proportional hazards regression is used when time-to-event is the outcome

Common multivariate regression models.

Outcome (dependent variable)	Example outcome variable	Appropriate multivariate regression model	Example equation	What do the coefficients give you?
Continuous	Blood pressure	**Linear regression**	blood pressure (mmHg) = α + βsalt*salt consumption (tsp/day) + βage*age (years) + βsmoker*ever smoker (yes=1/no=0)	slopes—tells you how much the outcome variable increases for every 1-unit increase in each predictor.
Binary	High blood pressure (yes/no)	**Logistic regression**	ln (odds of high blood pressure) = α + βsalt*salt consumption (tsp/day) + βage*age (years) + βsmoker*ever smoker (yes=1/no=0)	odds ratios—tells you how much the odds of the outcome increase for every 1-unit increase in each predictor.
Time-to-event	Time-to-death	**Cox regression**	ln (rate of death) = α + βsalt*salt consumption (tsp/day) + βage*age (years) + βsmoker*ever smoker (yes=1/no=0)	hazard ratios—tells you how much the rate of the outcome increases for every 1-unit increase in each predictor.