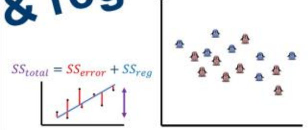
SE

# Correlation & regression

**PART 2 details of the procedure**




### Regression/correlation

Goal: identify a non-random relationship between X and Y.

- Check assumptions.
  e.g., linearity, independence.
- Estimate parameters.
  $Y = \alpha + \beta X$, $r$, $R^2$
- Test for stat. significance.
  ANOVA, t-test
- Calculate confidence intervals.
- Interpret results.


### Assumptions of correlation and linear regression.

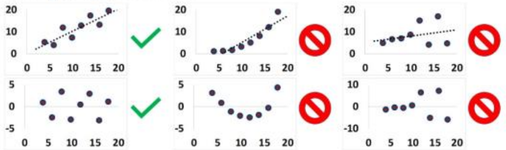
Independence of data points is hard to tell just from the data. We must know the nature of the values.

Treatment (i.e., X) variables being fixed is a weak assumption and technically distinguishes correlation and regression studies.

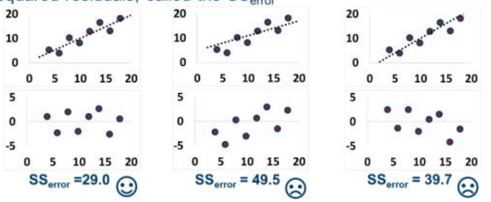
Linearity, normal distribution of residuals, and equal variance along X range done by plotting data and residuals to judge fit to assumptions.

### Assumptions of correlation and linear regression.

Linearity, normal distribution of residuals, and equal variance along X range done by plotting data first and residuals again after getting best fit line.



The "best" fit line is the one that minimizes the sum of the squared residuals, called the $SS_{error}$




### Estimating the best fit equation, $Y = a + bX$

The slope will be: $b = \dfrac{SP_{XY}}{SS_X}$

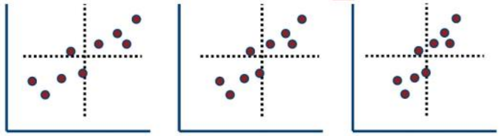
The best-fit line also passes through the center point $(\bar{X}, \bar{Y})$.

We can therefore rearrange the equation below to solve for the Y-intercept.

$\bar{Y} = a + b\bar{X}$ ➡ $a = \bar{Y} - b\bar{X}$

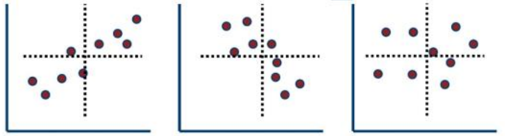
### Estimating the best fit equation, $Y = a + bX$

The slope will be: $b = \dfrac{SP_{XY}}{SS_X}$ $SP_{XY} = \sum (X_i - \bar{X})(Y_i - \bar{Y})$ $SS_X = \sum (X_i - \bar{X})^2$



### Estimating the best fit equation, $Y = a + bX$

The slope will be: $b = \dfrac{SP_{XY}}{SS_X}$ $SP_{XY} = \sum (X_i - \bar{X})(Y_i - \bar{Y})$ $SS_X = \sum (X_i - \bar{X})^2$



### Estimating the best fit equation, $Y = a + bX$

Make a table of the data with columns for X and Y values. Calculate:
Sum of squares for the X values, $SS_X$
Sum of squares for the Y values, $SS_Y$
Sum of the cross-products, $SP_{XY}$

| | X | Y | SP |
|---|---|---|---|
| | 3 | 6 | 6 |
| | 5 | 8 | 0 |
| | 7 | 13 | 8 |
| sum | 15 | 27 | 14 |
| mean | 5 | 9 | |
| SS | 8 | 26 | |

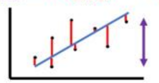
$SS_X = \sum (X_i - \bar{X})^2$

$SS_Y = \sum (Y_i - \bar{Y})^2$ $SP_{XY} = \sum (X_i - \bar{X})(Y_i - \bar{Y})$

The "best" fit line is the one that minimizes the sum of the squared residuals.

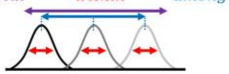
Conceptually this is like partitioning the overall $SS_{total}$ in the Y direction into $SS_{error}$ and $SS_{regression}$

$SS_{total} = SS_{error} + SS_{reg}$



A direct parallel with the ANOVA method of partitioning the overall $SS_{total}$ into $SS_{within}$ and $SS_{among}$

$SS_{total} = SS_{within} + SS_{among}$



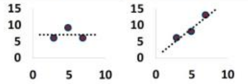
### Estimating the best fit equation, $Y = a + bX$

$b = \dfrac{SP_{XY}}{SS_X}$

| | X | Y | SP | X | Y | SP |
|---|---|---|---|---|---|---|
| | 3 | 6 | 2 | 3 | 6 | 6 |
| | 5 | 9 | 0 | 5 | 8 | 0 |
| | 7 | 6 | -2 | 7 | 13 | 8 |
| sum | 15 | 21 | 0 | 15 | 27 | 14 |
| mean | 5 | 7 | | 5 | 9 | |
| SS | 8 | 6 | | 8 | 26 | |

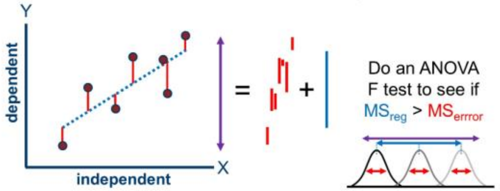
$b = 0/8 = 0$
$a = 7 - 0(5) = 7$

$b = 14/8 = 1.75$
$a = 9 - 1.75(5) = 0.25$



### Significance testing of the slope, ANOVA.

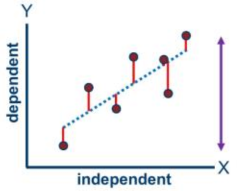
$SS_{total} = SS_{error} + SS_{reg}$




Do an ANOVA F test to see if $MS_{reg} > MS_{error}$

### Significance testing of the slope, ANOVA.

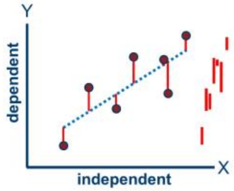
$SS_{total} = SS_{error} + SS_{reg}$




$SS_Y = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$

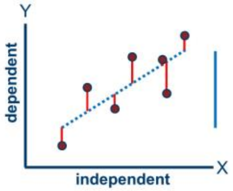
### Significance testing of the slope, ANOVA.

$SS_{total} = SS_{error} + SS_{reg}$




$SS_{error} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$

$SS_{total} = SS_{error} + SS_{reg}$

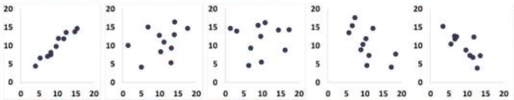



$SS_{reg} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$

### The coefficient of determination, $R^2$

Goal: measures strength of the relationship between two variables.

$R^2 = \dfrac{SS_{reg}}{SS_{total}} = r^2$

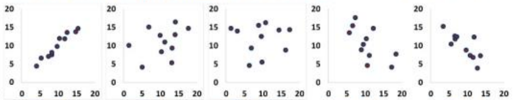



Has a direct interpretation as the proportion of variance in Y explained by variance in X. Ranges from 0 to 1.

### The correlation coefficient, r

Goal: measures strength, and direction, of the relationship between two variables.

$r = \dfrac{SP_{XY}}{\sqrt{SS_X SS_Y}}$




No direct quantitative interpretation. Ranges from -1 to 1.

### Significance testing of the slope, t-test.

$H_0$: $\beta = \beta_0$
$H_A$: $\beta \neq \beta_0$

$t_{calc} = \dfrac{b - \beta_0}{SE(slope)} = \dfrac{b - \beta_0}{SE_b}$

$df = n - 2$

Tests whether variance in X explains variance in Y differently from $\beta_0$ slope.

$SE_b = \sqrt{\dfrac{\dfrac{(SS_Y - b^2 SS_X)}{n-2}}{SS_X}}$

If $\beta_0$=0, equivalent to ANOVA

### Significance testing of the slope, t-test.

Recall: the t-test compares the difference between a hypothesized population mean and an observed sample mean, in terms of standard errors.

$H_0$: $\beta = \beta_0$
$H_A$: $\beta \neq \beta_0$

$t_{calc} = \dfrac{b - \beta_0}{SE(slope)} = \dfrac{b - \beta_0}{SE_b}$

$df = n - 2$

(It's a one-sample t-test)

### Significance testing of the slope, ANOVA.

Do an ANOVA F test to see if $MS_{reg} > MS_{error}$

$H_0$: $\beta = 0$
$H_A$: $\beta \neq 0$

$SS_{error} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$ $MS_{error} = \dfrac{SS_{error}}{n-2}$

$df_{reg} = 1$
$df_{error} = n-2$

$SS_{reg} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$ $MS_{reg} = \dfrac{SS_{reg}}{1}$

Tests whether variance in X explains variance in Y

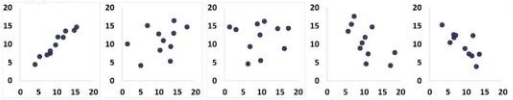
$F_{calc} = \dfrac{MS_{reg}}{MS_{error}}$

### Significance test for r (the correlation coefficient)

Can be tested using a t-test with values as shown.

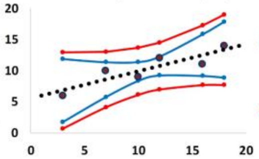
$H_0$: ρ=0
$H_A$: ρ≠0

$t_{calc} = r\sqrt{\dfrac{n-2}{1-r^2}}$ $df = n - 2$




### Confidence and prediction/inclusion intervals.

CI: where we are confident the true population relationship is.
PI: where we predict the population data lies.



CI: $\hat{Y}_i \pm t_{\frac{\alpha}{2}, n-2} \sqrt{\dfrac{(SS_Y - b^2 SS_X)}{n-2}} \sqrt{\dfrac{1}{n} + \dfrac{(X_i - \bar{X})^2}{SS_X}}$

PI: $\hat{Y}_i \pm t_{\frac{\alpha}{2}, n-2} \sqrt{\dfrac{(SS_Y - b^2 SS_X)}{n-2}} \sqrt{1 + \dfrac{1}{n} + \dfrac{(X_i - \bar{X})^2}{SS_X}}$

### Interpreting results of the analysis.

- Significant slopes in a regression analysis can imply causation, but the causation may be indirect.

- Significant slopes in a correlation analysis *may* imply causation, but other factors may be driving the pattern.

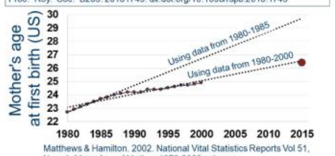
- Correlation does not imply causation, but that doesn't mean it's useless. A significant correlation is often the first step to determining causation. Also, lack of a correlation is a powerful argument against a proposed causation.

### Interpreting results of the analysis.

- A significant slope implies a non-random relationship, but is it relevant or trivial?

- The results only hold for the range of X values studied.

e.g., men have a 10% overall higher risk of cancer for each 10 cm (4") of height above 175 cm (5'9").
Nunney. 2018 Size matters: height, cell number and a person's risk of cancer. Proc. Roy. Soc. B285: 20181743. dx.doi.org/10.1098/rspb.2018.1743

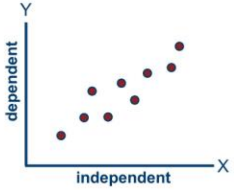



Matthews & Hamilton. 2002. National Vital Statistics Reports Vol 51, Num 1. Mean Age of Mother, 1970-2000. cdc.gov.


### Regression/correlation

Goal: identify a non-random relationship between X and Y.

- Check assumptions.
  e.g., linearity, independence.
- Estimate parameters.
  $Y = \alpha + \beta X$, $r$, $R^2$
- Test for stat. significance.
  ANOVA, t-test
- Calculate confidence intervals.
- Interpret results.


StatsExamples.com