Correlation and Linear Regression

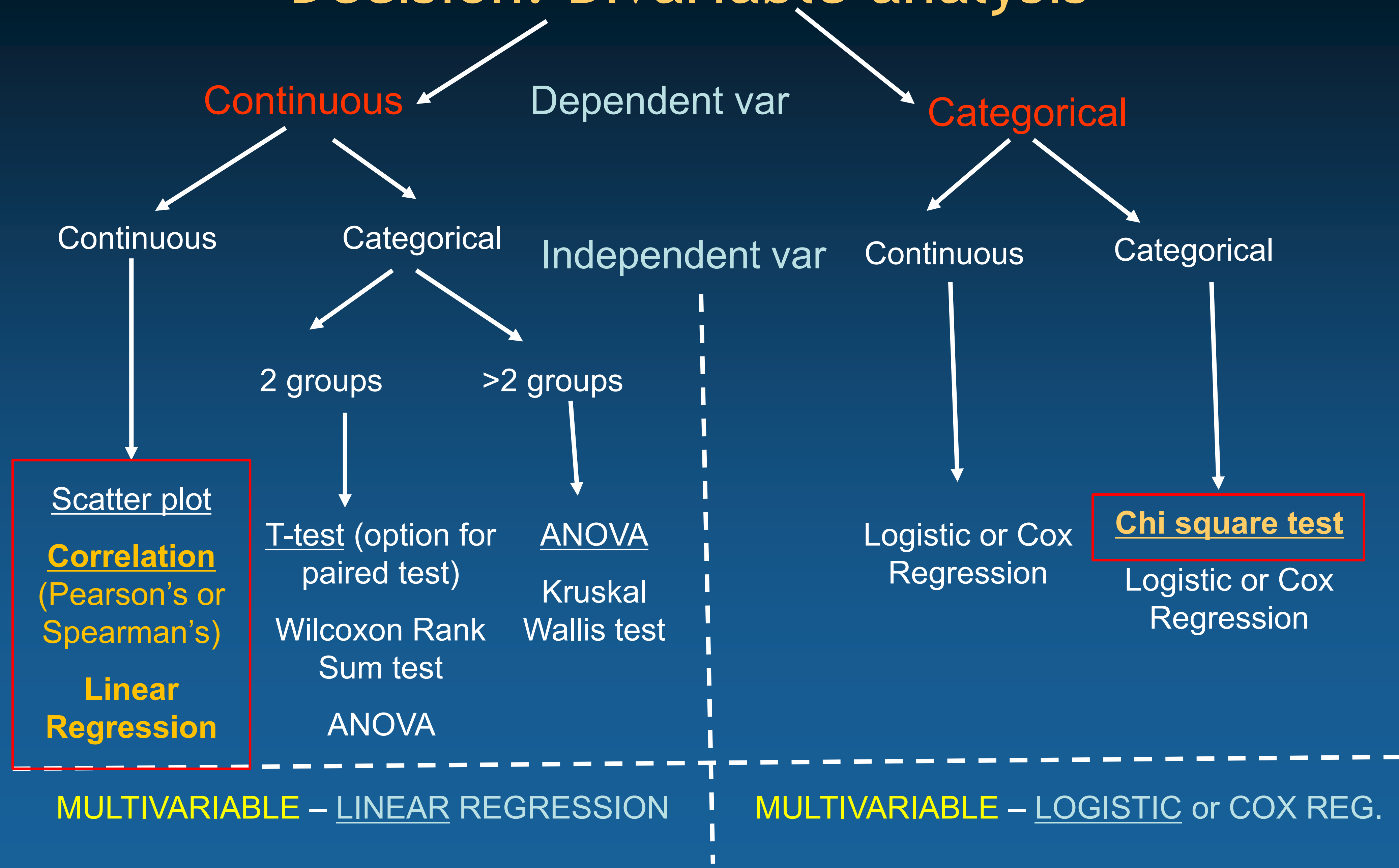
Ilir Agalliu MD, Sc.D
Associate Professor
Dept. of Epidemiology & Population Health
Albert Einstein College of Medicine

Feb 14th, 2024

Outline

- Chi-Square Test (from last session)
- Correlations
- Simple Linear Regression
  - The Model
  - The Method of Least Squares
  - Evaluation of the Model
- Multivariate Linear Regression
- Examples

Decision: Bivariable analysis



Statistical Tests - Categorical Variables
Chi-square (x^2) test

- Compares the proportion of individuals with a certain characteristic or exposure among two or more groups
- Generally used for 2 x 2 or n x n (contingency) tables
- Each cell is mutually exclusive
- Can be used for two or more independent groups

- H_0 : p1 = p2
- H_A : p1 ≠ p2 (two-sided)
- p – denotes proportion

Chi-Square Test

Assume we wish to compare proportions of two birth weight groups by maternal hypertension during pregnancy

Birth weight group * History of hypertension Crosstabulation

			History of hypertension		Total
			No	Yes	
Birth weight group	>2500 gm	Count	125	5	130
		% within Birth weight group	96.2%	3.8%	100.0%
		% within History of hypertension	70.6%	41.7%	68.8%
	< 2500 gm	Count	52	7	59
		% within Birth weight group	88.1%	11.9%	100.0%
		% within History of hypertension	29.4%	58.3%	31.2%
Total		Count	177	12	189
		% within Birth weight group	93.7%	6.3%	100.0%
		% within History of hypertension	100.0%	100.0%	100.0%

$$X^2_{(df)} = \Sigma \text{ (Obs - Exp)}^2 \text{ / Exp}$$

Need to calculate expected values

Calculation of Expected Values

Hypertension

Birth-weight	No	Yes	Total
>2500	$\frac{(a+b)*(a+c)}{T}$	$\frac{(a+b)*(b+d)}{T}$	a+b
<2500	$\frac{(c+d)*(a+c)}{T}$	$\frac{(b+d)*(c+d)}{T}$	c+d
Total	a+c	b+d	T

Expected a = ((a+b)*(a+c))/T

Chi-Square Test

Birth weight group * History of hypertension Crosstabulation

			History of hypertension		Total
			No	Yes	
Birth weight group	>2500 gm	Count	125	5	130
		Expected Count	121.7	8.3	130.0
	< 2500 gm	Count	52	7	59
		Expected Count	55.3	3.7	59.0
Total		Count	177	12	189
		Expected Count	177.0	12.0	189.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	4.388[a]	1	.036		
Continuity Correction[b]	3.143	1	.076		
Likelihood Ratio	4.022	1	.045		
Fisher's Exact Test				.052	.042
Linear-by-Linear Association	4.365	1	.037		
N of Valid Cases	189				

a. 1 cells (25.0%) have expected count less than 5. The minimum expected count is 3.75.

b. Computed only for a 2x2 table

Chi-Square Test
Can be used also for n x n tables

Birth weight group * Mothers race Crosstabulation

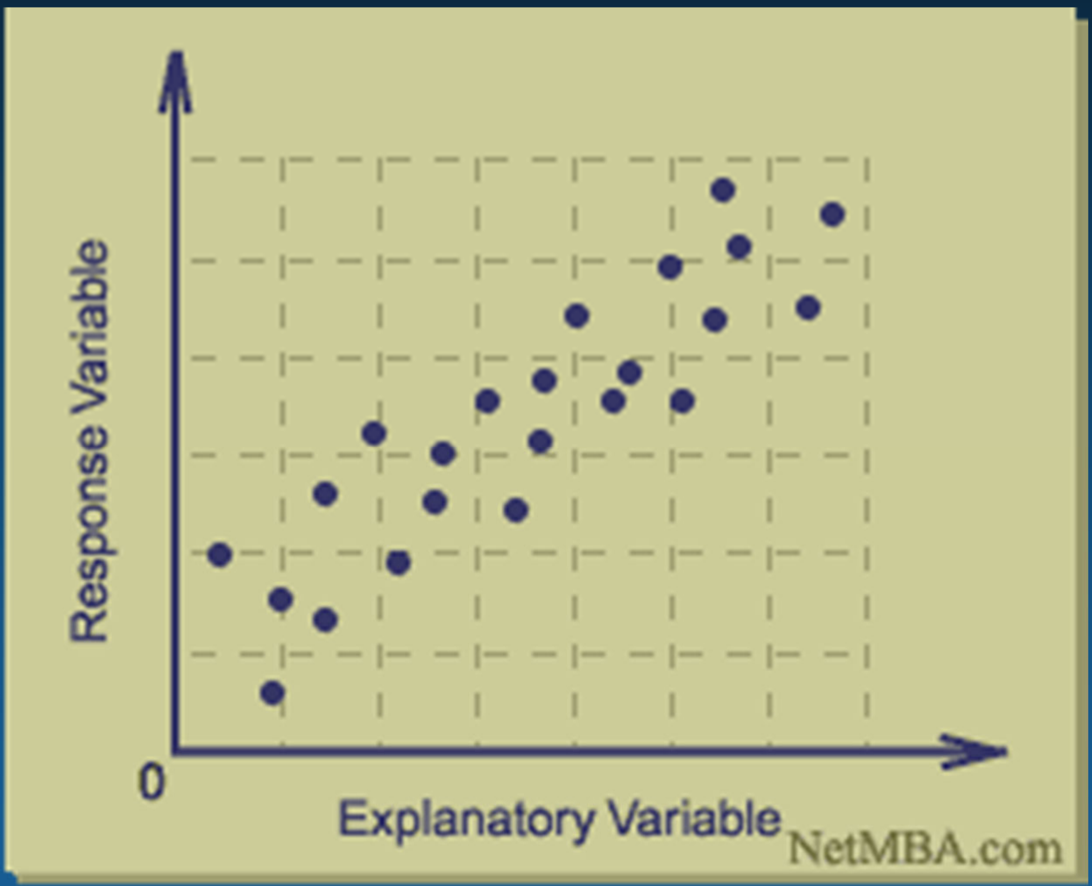
			Mothers race			Total
			White	Black	Other	
Birth weight group	>2500 gm	Count	73	15	42	130
		Expected Count	66.0	17.9	46.1	130.0
	< 2500 gm	Count	23	11	25	59
		Expected Count	30.0	8.1	20.9	59.0
Total		Count	96	26	67	189
		Expected Count	96.0	26.0	67.0	189.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	5.005[a]	2	.082
Likelihood Ratio	5.010	2	.082
Linear-by-Linear Association	3.570	1	.059
N of Valid Cases	189		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 8.12.

Correlation



- Shows the relationship between two continuous variables
- Identifies
  - Direction
  - Shape
  - Outliers
- Does NOT provide a quantitative estimate of their association

Correlation Analysis

Correlation is a measure of the statistical relationship between two variables

- PEARSONS: Normally distributed variables
- SPEARMAN's: Not-normally distributed variables
- Quantitative estimate (range from -1 to 1)

r = 0 -> *no correlation*, r > 0 -> positive *correlation*

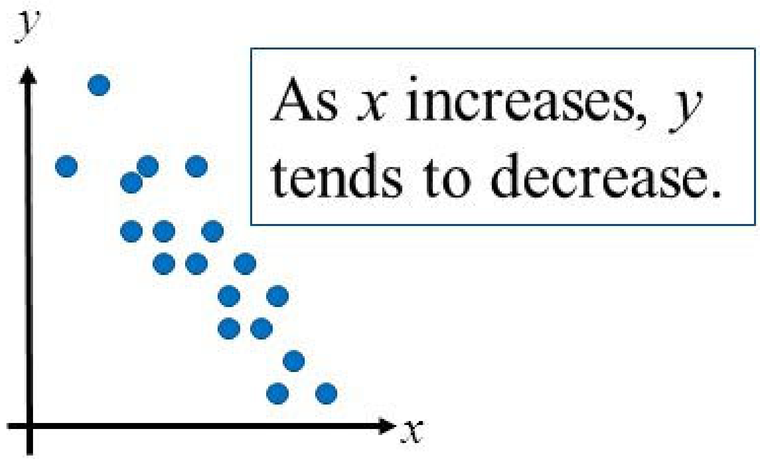
r < 0 -> *negative correlation*

r = 0.8 to 1 -> strong correlation

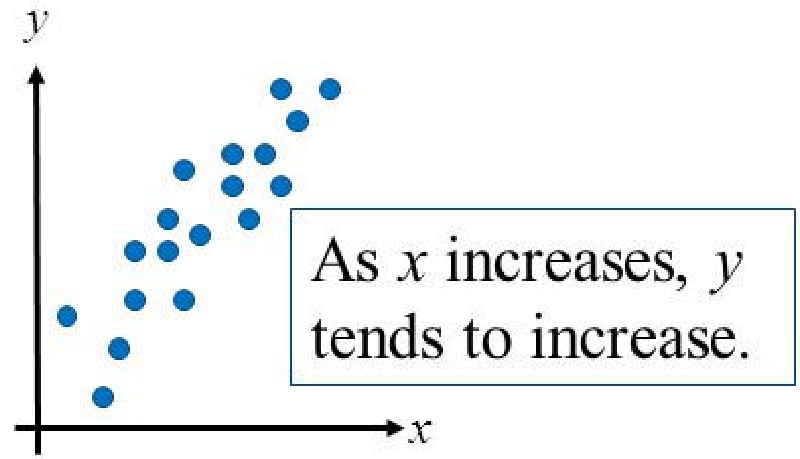
r = 0.6; p= 0.001; 95% CI: 0.4, 0.8

Does not differentiate between dependent and independent variables

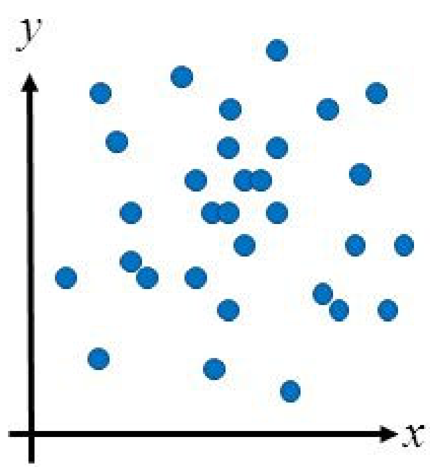
Types of Correlation



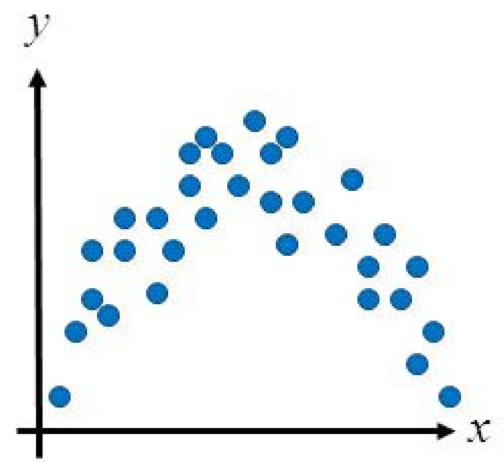
Negative Linear Correlation



Positive Linear Correlation



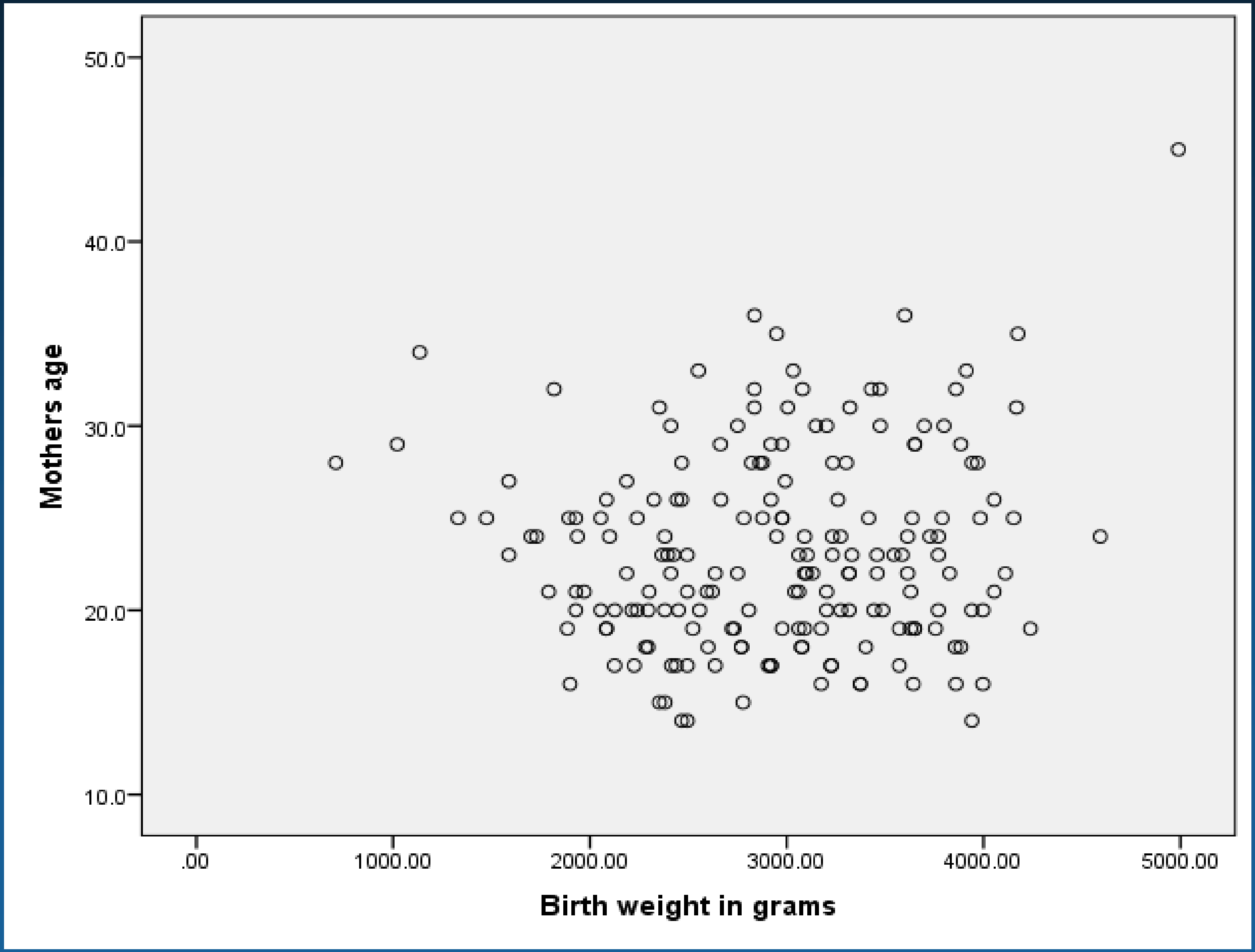
No Correlation



Nonlinear Correlation

Correlation - Example

- Is there a correlation between mother's age and baby's weight at birth?



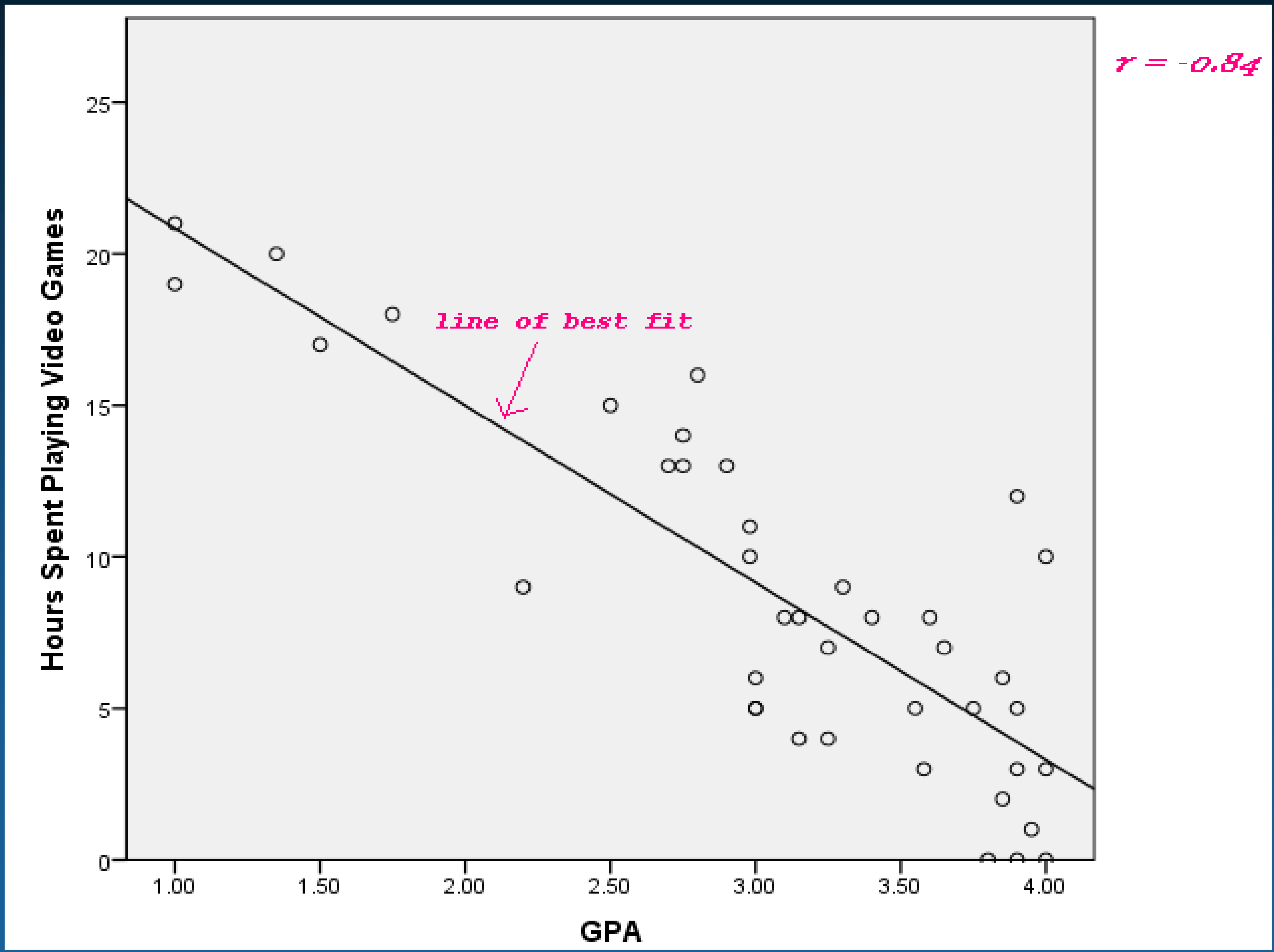
Correlation - Example

- Is there a correlation between mother's age and baby's weight at birth?

Correlations

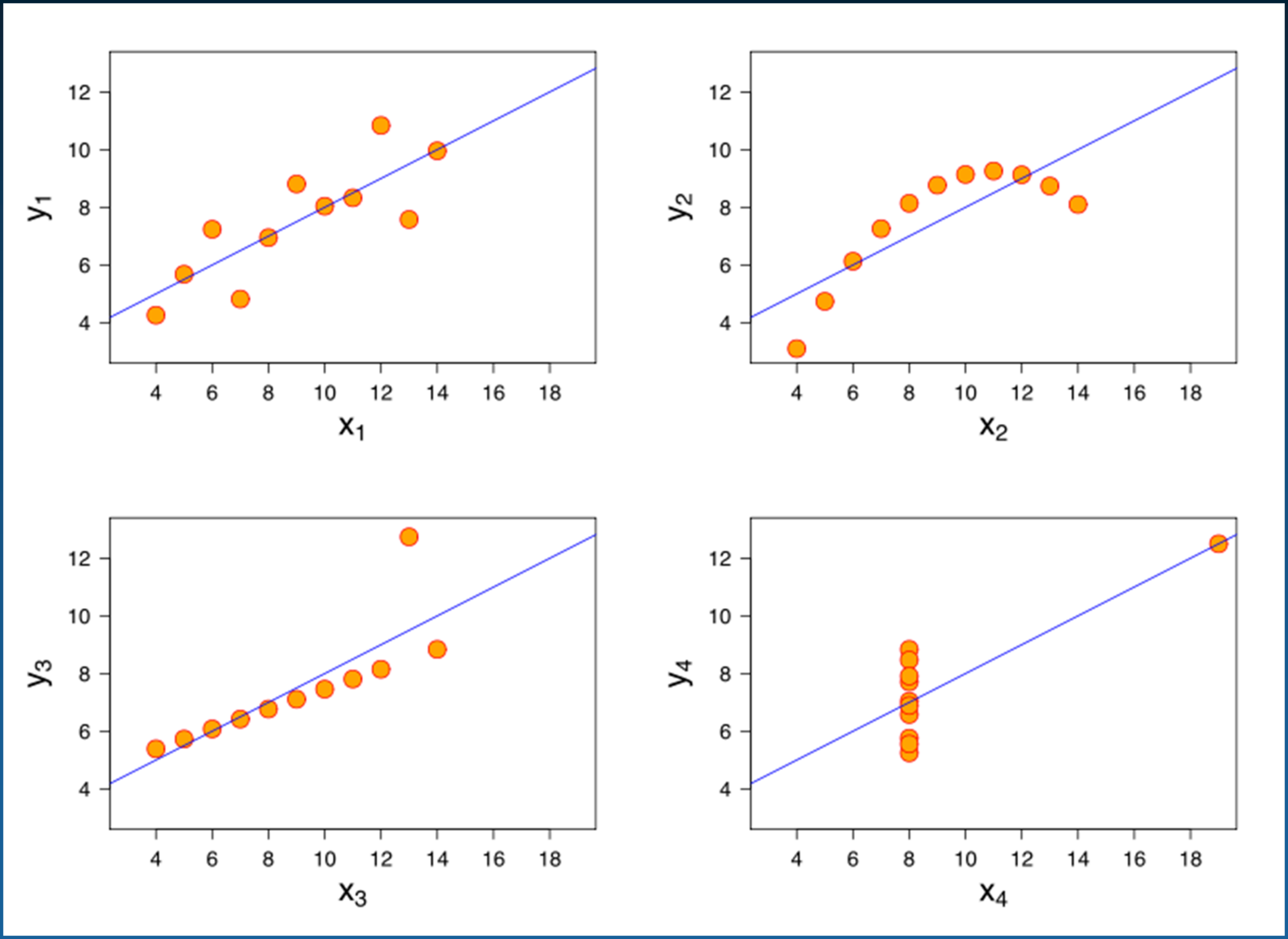
		Birth weight in grams	Mothers age
Birth weight in grams	Pearson Correlation	1	.090
	Sig. (2-tailed)		.219
	N	189	189
Mothers age	Pearson Correlation	.090	1
	Sig. (2-tailed)	.219	
	N	189	189

Hours Spent Playing Video Games & GPA
Inverse Correlation (r = -0.84; p=0.02)



Correlation is not equal to association

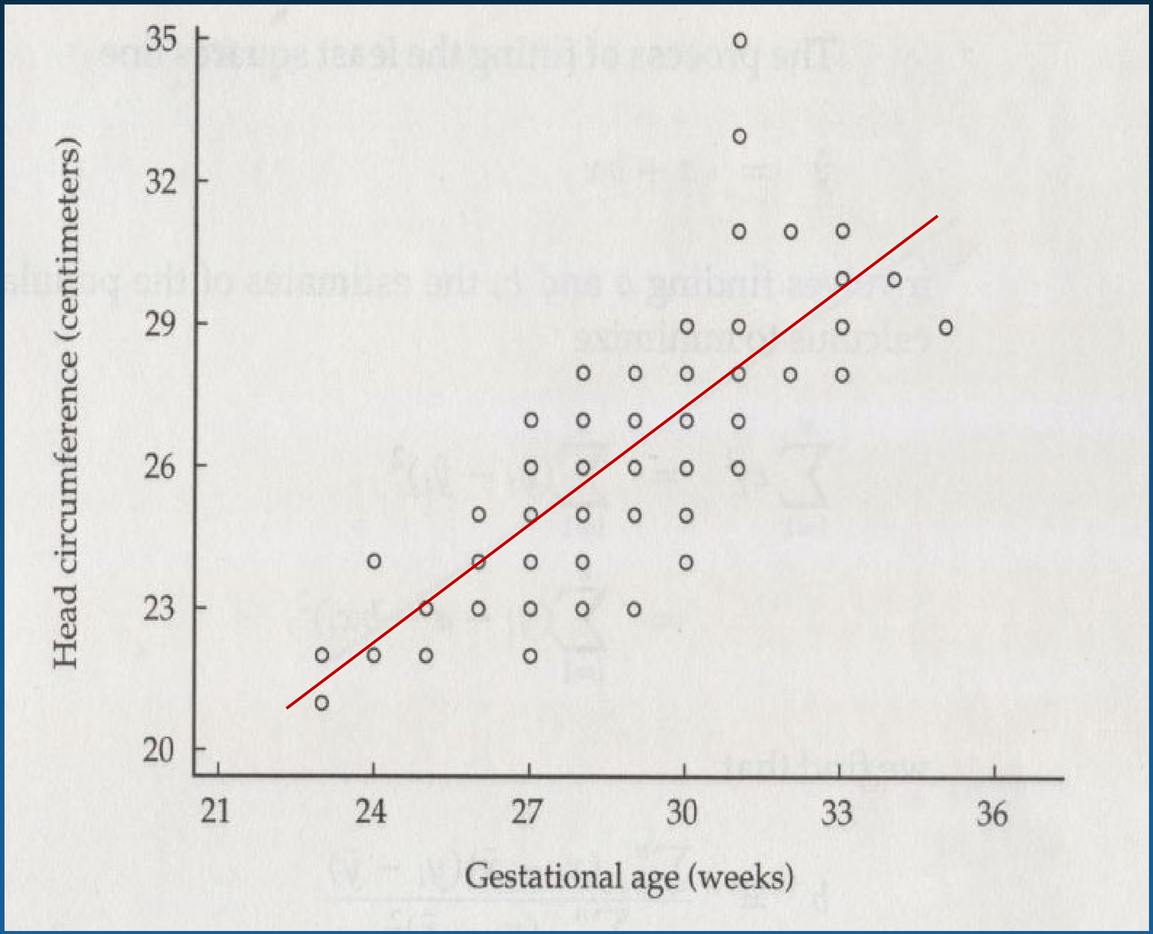
Four sets of data with the same correlation coefficient r= 0.816 (Anscombe's quartet)



Case Study
Head circumference and gestational age (weeks)

A. Is there a correlation between head circumference and gestational age?

B. What is the relationship between head circumference and gestational age?



Simple Linear Regression

- Technique that is used to explore the relationship between two variables (usually continuous)
- Y: response or dependent variable
- X: predictor or explanatory variable - independent variable
- The population regression line is:

$$\mu_{y|x} = \alpha + \beta x.$$

  - α – intercept
  - β – slope (change in $\mu_{y|x}$ per <u>one unit change</u> in x)
- The regression line that we fit depends on the sample
  - $Y = \alpha + \beta x + \varepsilon$ (error term)

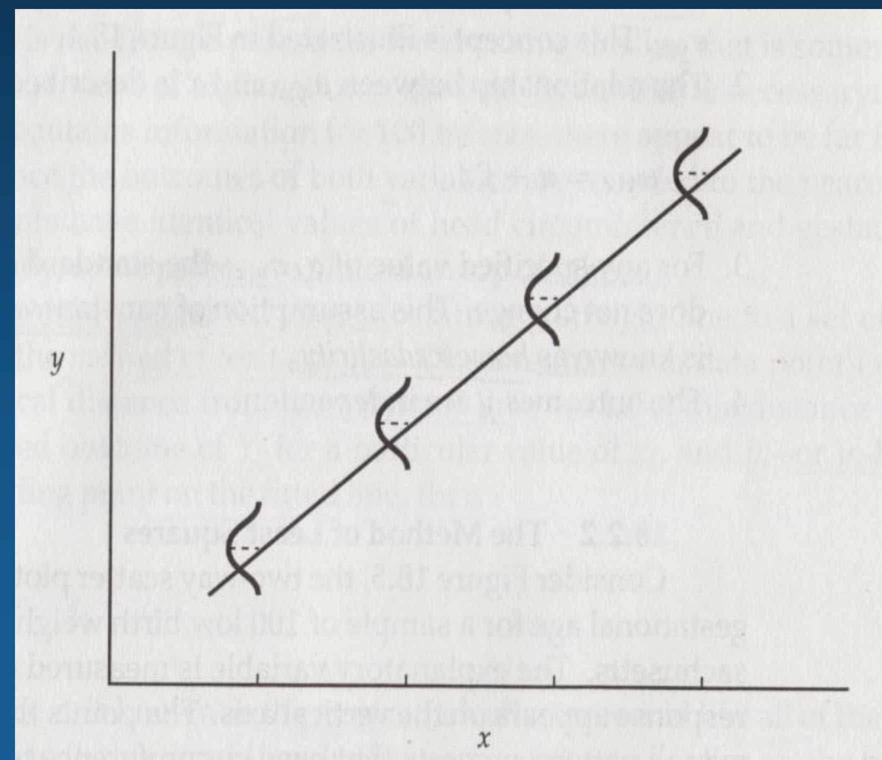
$$\hat{y} = a + bx.$$

- where $\hat{y}$ are predicted values

Assumptions of Linear Regression

- For a specific value of x, the distribution of y is normal with mean $\mu_{y|x}$ and SD $\sigma_{y|x}$
- Linear relationship
- Assumption of homoscedasticity
  - For any x, $\sigma_{y|x}$ is constant
- The outcomes y are independent
  - No correlation between Y_i

$$\mu_{y|x} = \alpha + \beta x.$$



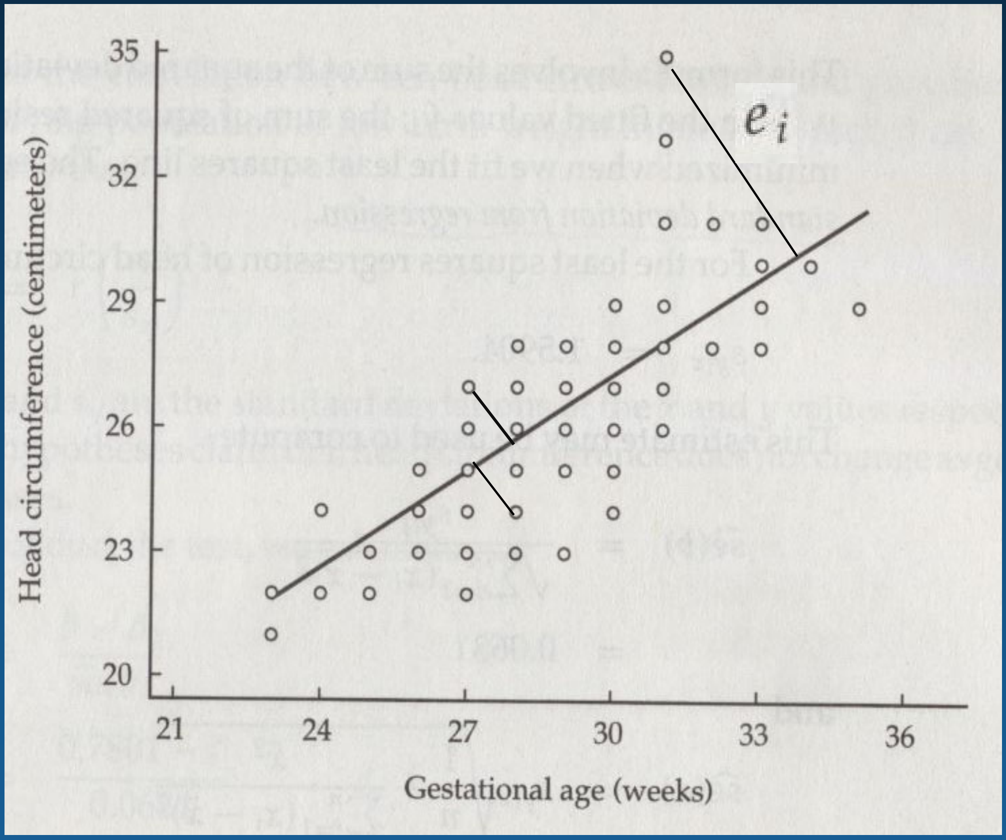
The Method of Least Squares

- When fitting a regression line not all data points will be in the line
- Residual or error is defined:

$$e_i = y_i - \hat{y}_i.$$

- Mathematical technique to fit the data in order to minimize the error or residual sum of squares

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$



Hypothesis Testing
Linear Regression

- Null hypothesis: No linear relationship between x and y

  Slope = 0 $H_0 : \beta = 0$

- Alternative hypothesis: ß ≠ 0 (2-sided)
- To conduct the test we calculate t-statistics
- $t = \dfrac{b - \beta_0}{\widehat{\text{se}}(b)}$ where $\widehat{\text{se}}(b) = \dfrac{s_{y|x}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$
- $s_{y|x}$ *standard deviation from regression*
- *t follows a t-distribution with df=n-2*

Head circumference and gestational age

Equation takes the form

$$y = 3.91 + 0.78\,x$$

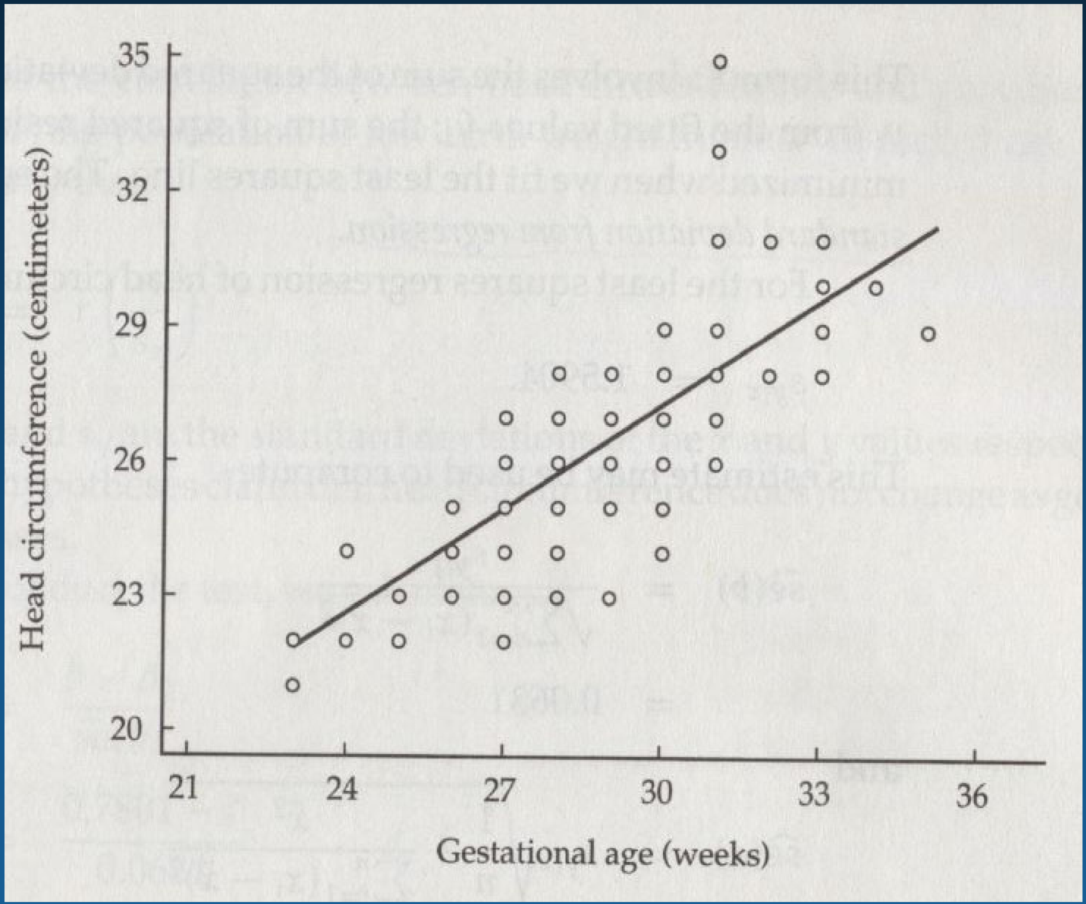
How do you interpret 3.91 and 0.78?

What is the predicted value of Y for gestational age 30 weeks?

Hypothesis testing:

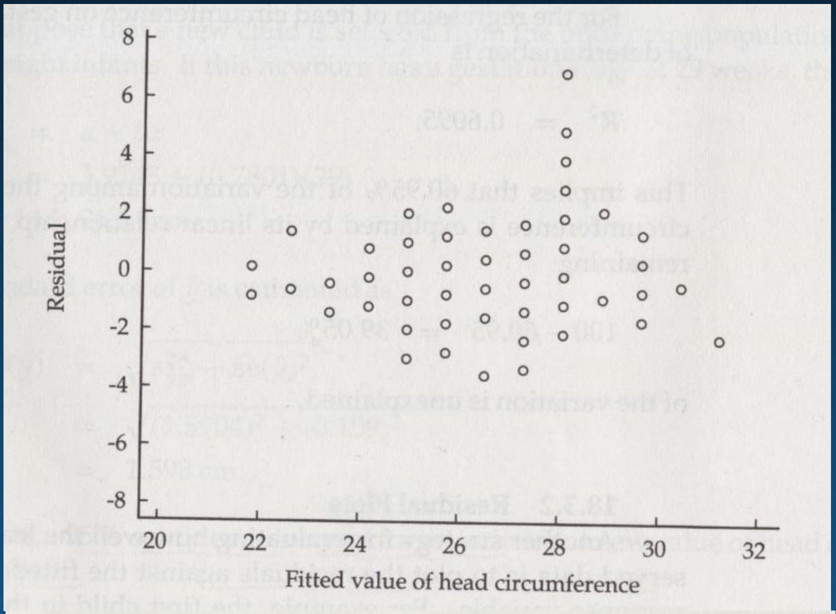
$t = 12.36$; $df=98$ and $p<0.001$

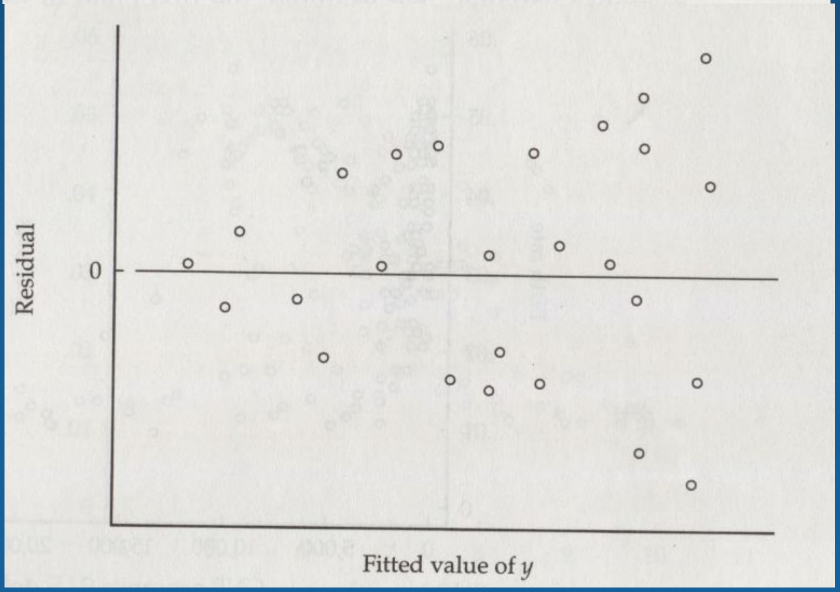
How do you interpret this?



Evaluation of the Model

Residual Plots





R^2: Coefficient of Determination

Interpreted as proportion of the variability in y that is explained by linear regression of y on x

$R^2 = 0.61$ for head circumference and gestational age

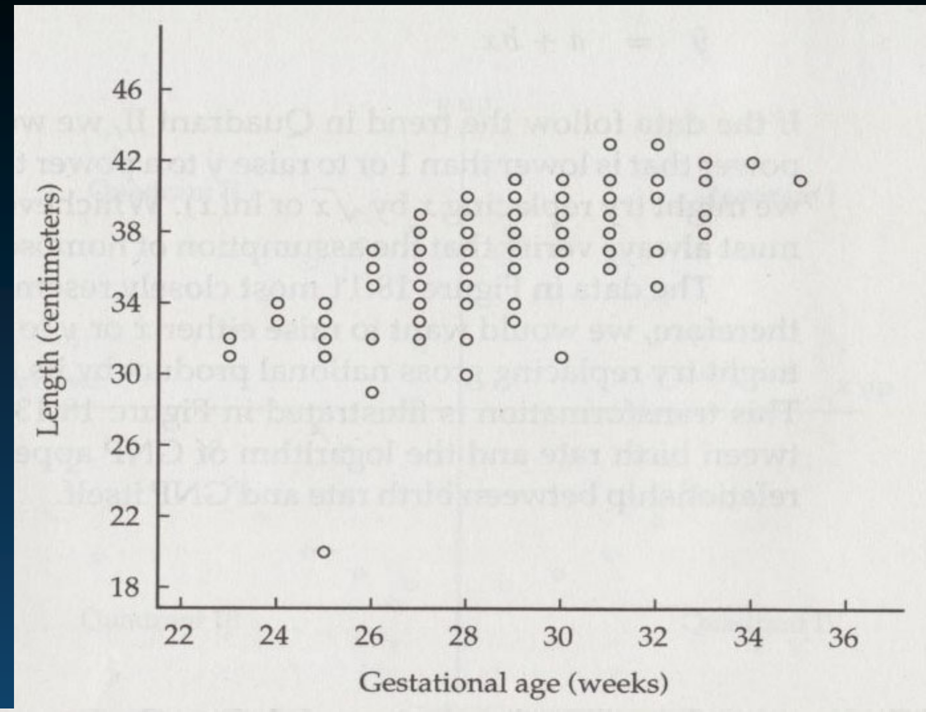
Residual Plots

Detect outliers

Failure of assumption of homoscedascity →

Might suggest other patterns of association (maybe non-linear)

Another Example
Length and gestational age (weeks)



Source	SS	df	MS
Model	575.73916	1	575.73916
Residual	687.02084	98	7.01041674
Total	1262.76	99	12.7551515

Number of obs	=	100
F(1,98)	=	82.13
Prob > F	=	0.0000
R-square	=	0.4559
Adj R-square	=	0.4504
Root MSE	=	2.6477

Variable	Coefficient	Std. Error	t	P>|t|	[95% Conf.	Interval]
length						
gestage	0.9516035	0.1050062	9.062	0.000	.7432221	1.159985
_cons	9.3281740	3.0451630	3.063	0.003	3.285149	15.3712

Multiple Linear Regression

What if we want to assess simultaneously the effect of two or more predictor variables on a continuous outcome?

Consider the following research question

- What is the association between baby's length and gestational age (week) as well as mother having hypertension (pre-eclampsia) during pregnancy?

- We can extend simple linear regression to accommodate two or more independent variables:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q + \epsilon$$

- Same assumptions apply also for multiple linear model
- Use the least square method to fit the model

Example: Multiple Linear Regression

What is the association between baby's length and gestational age (week) as well as mother having pre-eclampsia (toxemia)?

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \text{ (error term)}$$

where - x_1 continuous variable (gestage)

- x_2 indicator/dummy variable (tox: yes=1, no=0)

Table 19.2 Stata output displaying the regression of length on gestational age and toxemia

Source	SS	df	MS
Model	619.253622	2	309.626811
Residual	643.506378	97	6.63408638
Total	1262.76	99	12.7551515

Number of obs	=	100
F(2,97)	=	46.67
Prob > F	=	0.0000
R-square	=	0.4904
Adj R-square	=	0.4799
Root MSE	=	2.5757

length	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
gestage	1.069883	.1121039	9.544	0.000	.8473879	1.292378
tox	−1.777381	.6939918	−2.561	0.012	−3.154763	−.3999997
_cons	6.284326	3.191824	1.969	0.052	−.050561	12.61921

Example: Multiple Linear Regression Indicator variable

$\hat{y} = 6.28 + 1.07\,x_1 - 1.78\,x_2$

What is the equation for mothers <u>with toxemia</u>?

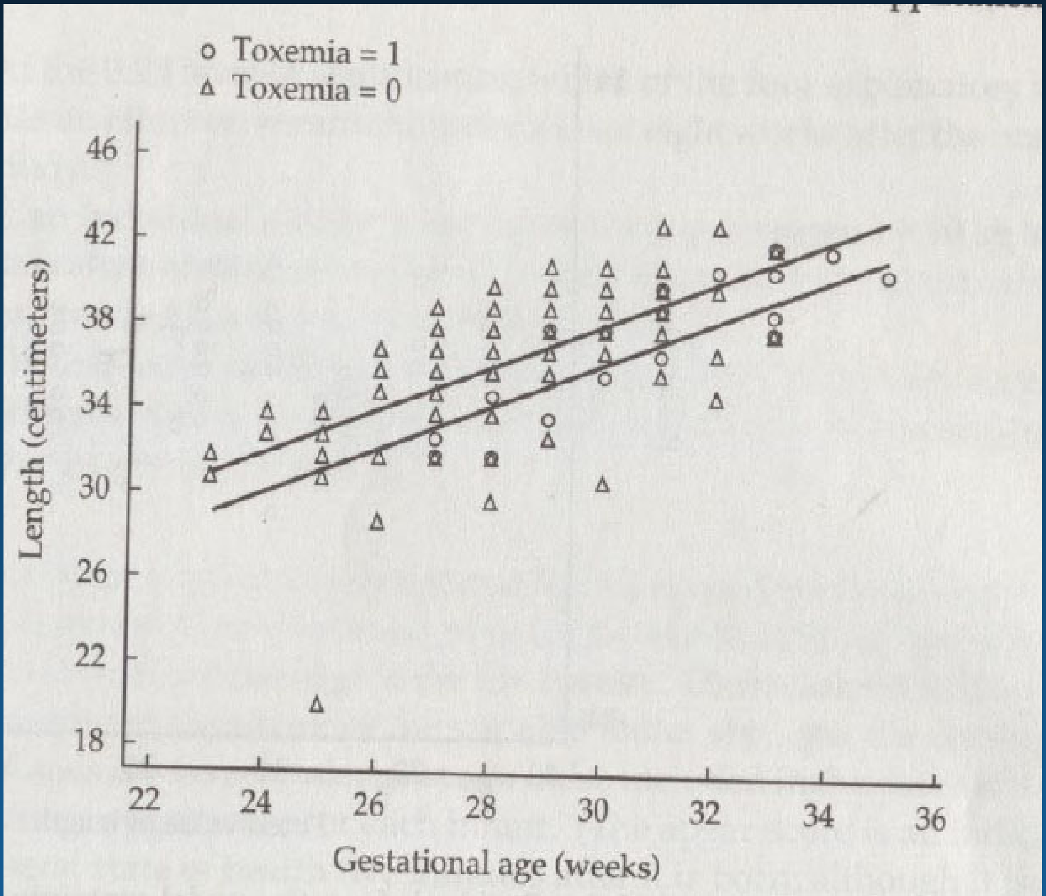
$y = 6.28 + 1.07\,x_1 - 1.78\,(1)$

$y = 4.50 + 1.07\,x_1$

What is the equation for mothers <u>without toxemia</u>?

$y = 6.28 + 1.07\,x_1 - 1.78\,(0)$

$y = 6.28 + 1.07\,x_1$



What is the predicted value of length ($\hat{y}$) for a baby born 32 weeks of age and having a mother with pre-eclampsia?

Linear Regression – Another Example

Is there a relationship between baby birth weight and maternal hypertension during pregnancy, after adjusting for age and smoking?

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6262001.401	3	2087333.800	4.123	.007[b]
	Residual	93655051.24	185	506243.520		
	Total	99917052.65	188			

a. Dependent Variable: Birth weight in grams

b. Predictors: (Constant), Smoking during pregnancy, History of hypertension, Mothers age

Model Summary[b]

					Change Statistics				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.250[a]	.063	.047	711.50792	.063	4.123	3	185	.007

a. Predictors: (Constant), Smoking during pregnancy, History of hypertension, Mothers age

b. Dependent Variable: Birth weight in grams

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
Model		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	2824.666	239.603		11.789	.000	2351.960	3297.371
	History of hypertension	-424.465	212.287	-.142	-1.999	.047	-843.279	-5.651
	Mothers age	10.933	9.804	.079	1.115	.266	-8.409	30.276
	Smoking during pregnancy	-273.621	106.147	-.184	-2.578	.011	-483.035	-64.206

a. Dependent Variable: Birth weight in grams

Periodontal Changes in Children and Adolescents With Diabetes

Lalla et al.

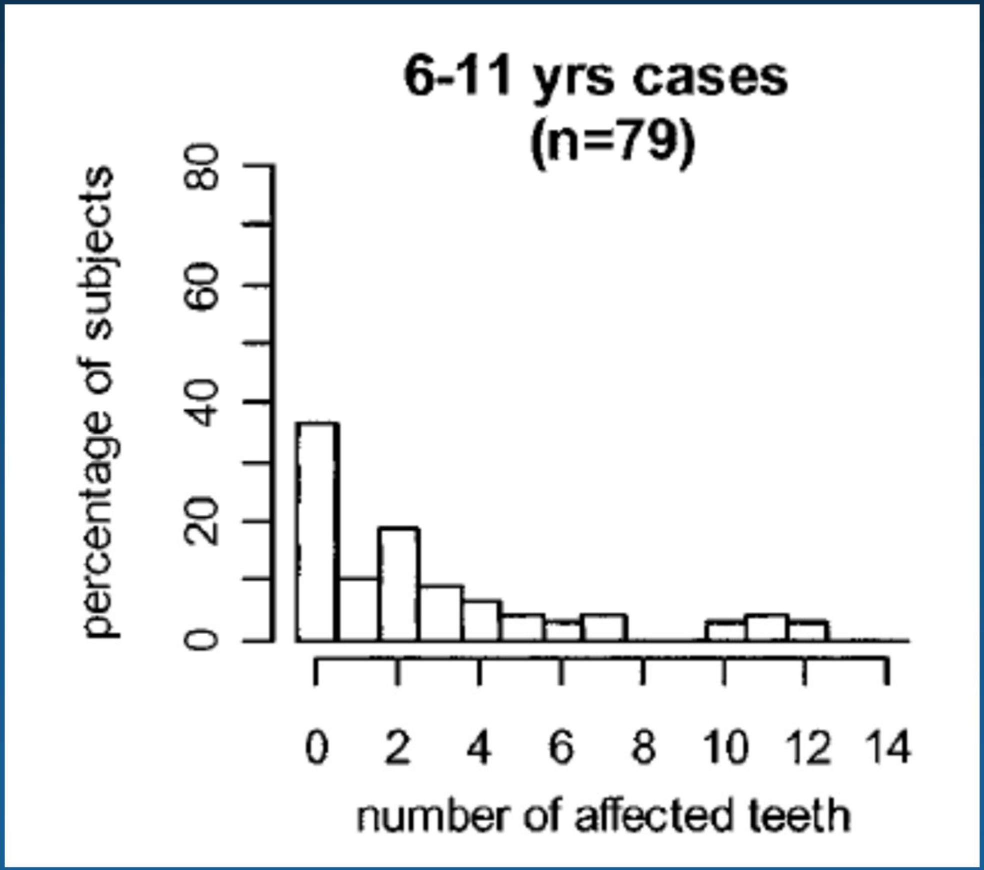
Diabetes Care, volume 29, number 2, February 2006

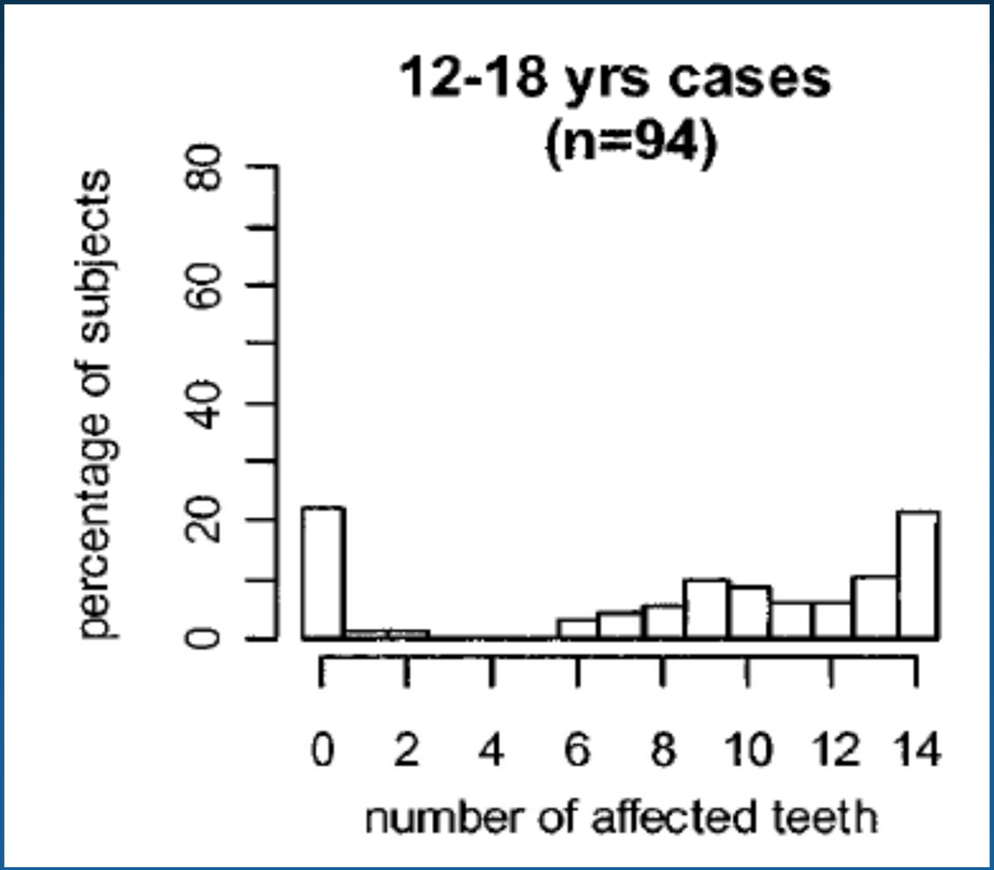
Table 4—*Estimated regression coefficients (and 95% CIs) from linear regression model* for number of affected teeth† among case subjects*

	Regression coefficient (95% CI)	*P* value
Mean A1C	0.12 (−0.25 to 0.49)	0.51
Duration of diabetes	−0.02 (−0.20 to 0.17)	0.87
Proportion of bleeding sites‡	1.72 (−0.92 to 4.36)	0.20
12–18 years age-group	5.17 (3.80–6.54)	<0.001
BMI	0.12 (0.02–0.23)	0.03

*Regression model also adjusted for sex, ethnicity, frequency of dental visits, and dental examiner. †Having at least one site with >2 mm of attachment loss. ‡Square root transformation performed to achieve a better fit.

Let's look at the distribution of outcome Nr of affected teeth Is this normally distributed?





Take Home Messages

- Correlations
  - Determines the relation between two variables
  - Doesn't differentiate the dependence

- Linear Regression
  - Used for continuous outcomes
  - Check assumptions of normality for outcome Y
  - Can accommodate both continuous and categorical independent variables
  - Goodness of Fit is indicated by R^2 and residual plots