Investigating relationships

Two categorical variables

Are boys more likely to prefer maths and science than girls?

Variables:

- Favourite subject (Nominal)
- Gender (Binary/ Nominal)

Summarise using %'s/ stacked or multiple bar charts

Test: Chi-squared

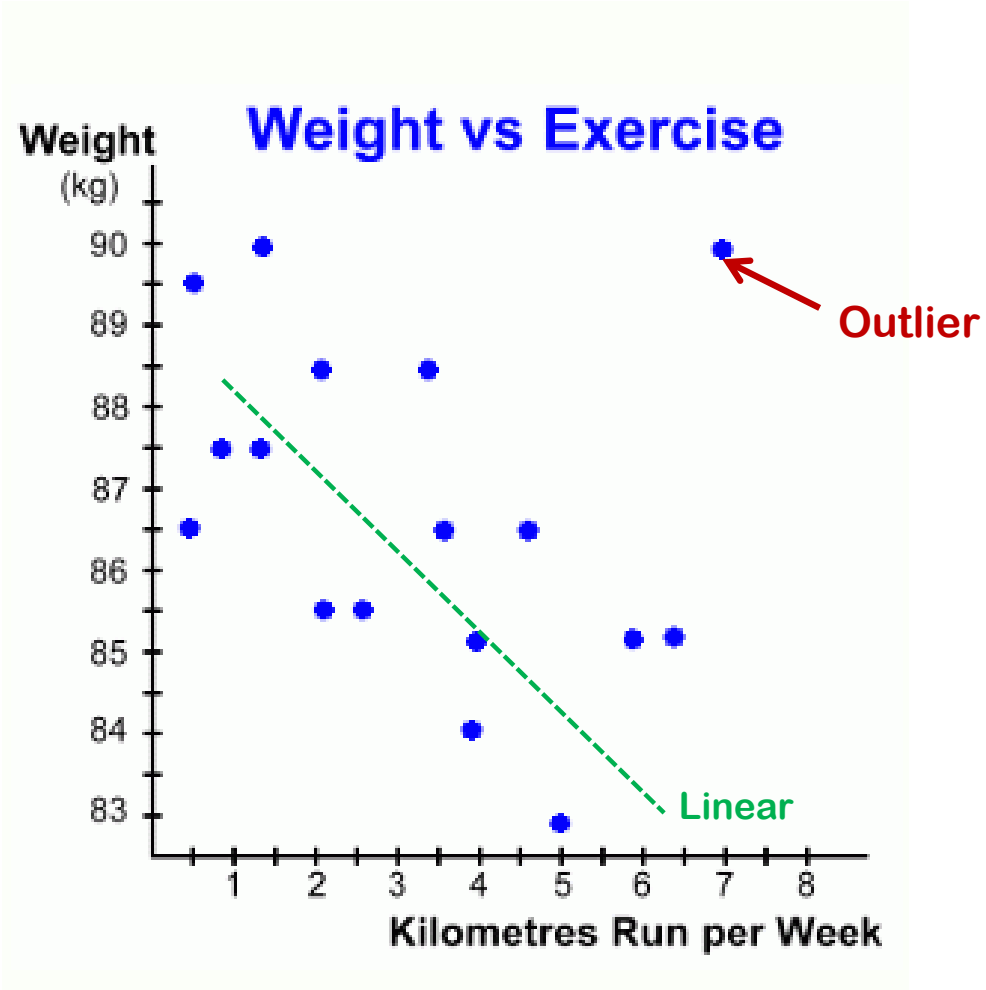
Tests for a relationship between two categorical variables

Scatterplot

Relationship between two scale variables:

- Explores the way the two co-vary: (correlate)
  - Positive / negative
  - Linear / non-linear
  - Strong / weak
- Presence of outliers
- Statistic used:

r = correlation coefficient

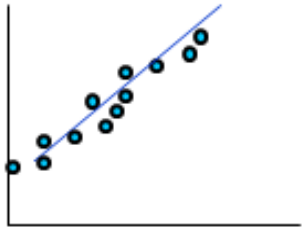


Correlation Coefficient r

- **Measures strength of a relationship between two continuous variables**

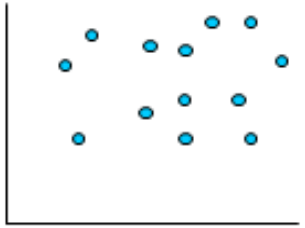
$-1 \leq r \leq 1$

Strong positive linear relationship



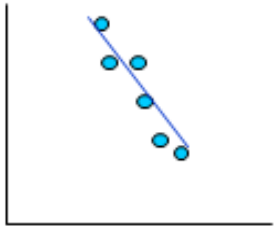
r = 0.9

No linear relationship



r = 0.01

Strong negative linear relationship



r = -0.9

Correlation Interpretation

An interpretation of the size of the coefficient has been described by Cohen (1992) as:

Correlation coefficient value			Relationship
-0.3 to +0.3			Weak
-0.5 to -0.3	or	0.3 to 0.5	Moderate
-0.9 to -0.5	or	0.5 to 0.9	Strong
-1.0 to -0.9	or	0.9 to 1.0	Very strong

Cohen, L. (1992). Power Primer. Psychological Bulletin, 112(1) 155-159

Does chocolate make you clever or crazy?

- A paper in the New England Journal of Medicine claimed a relationship between chocolate and Nobel Prize winners

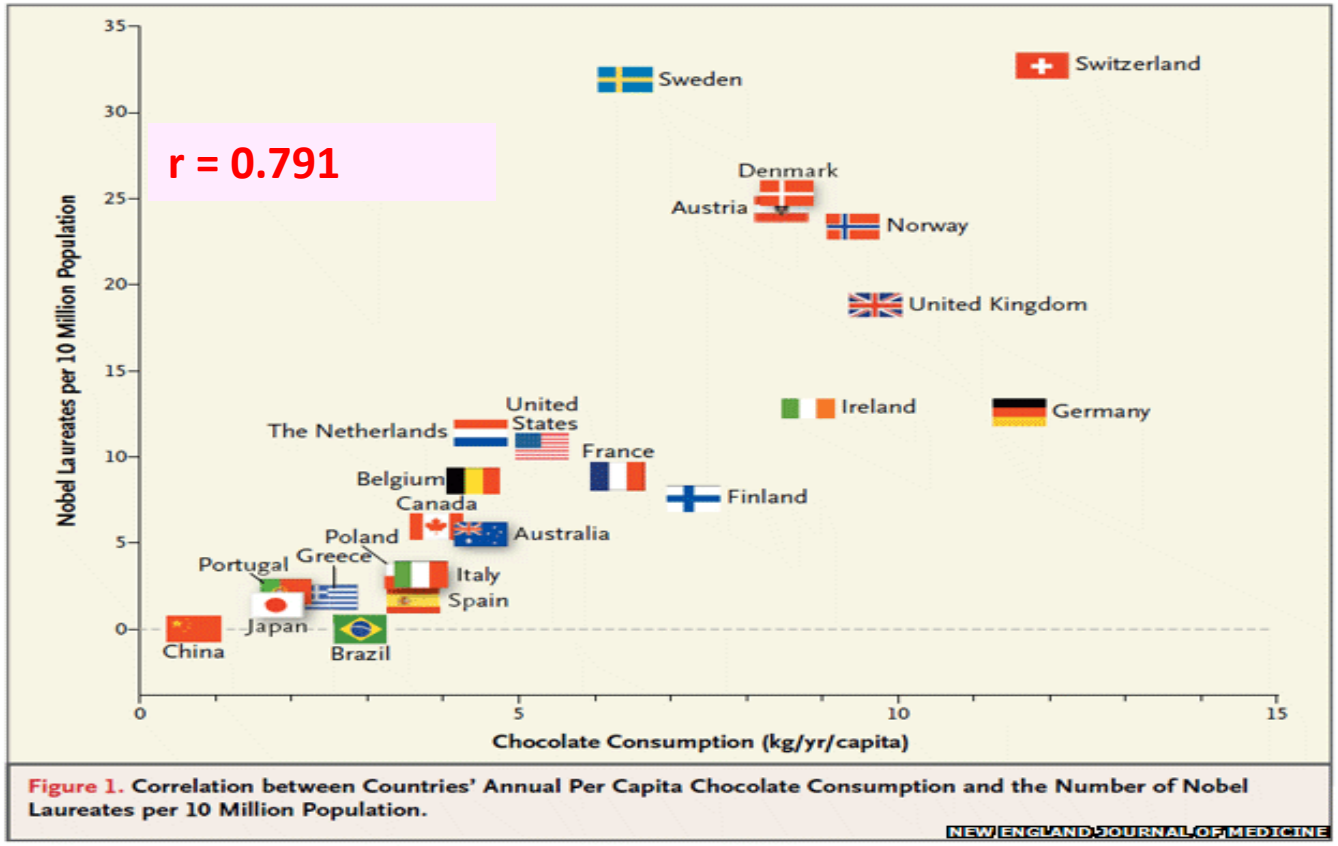


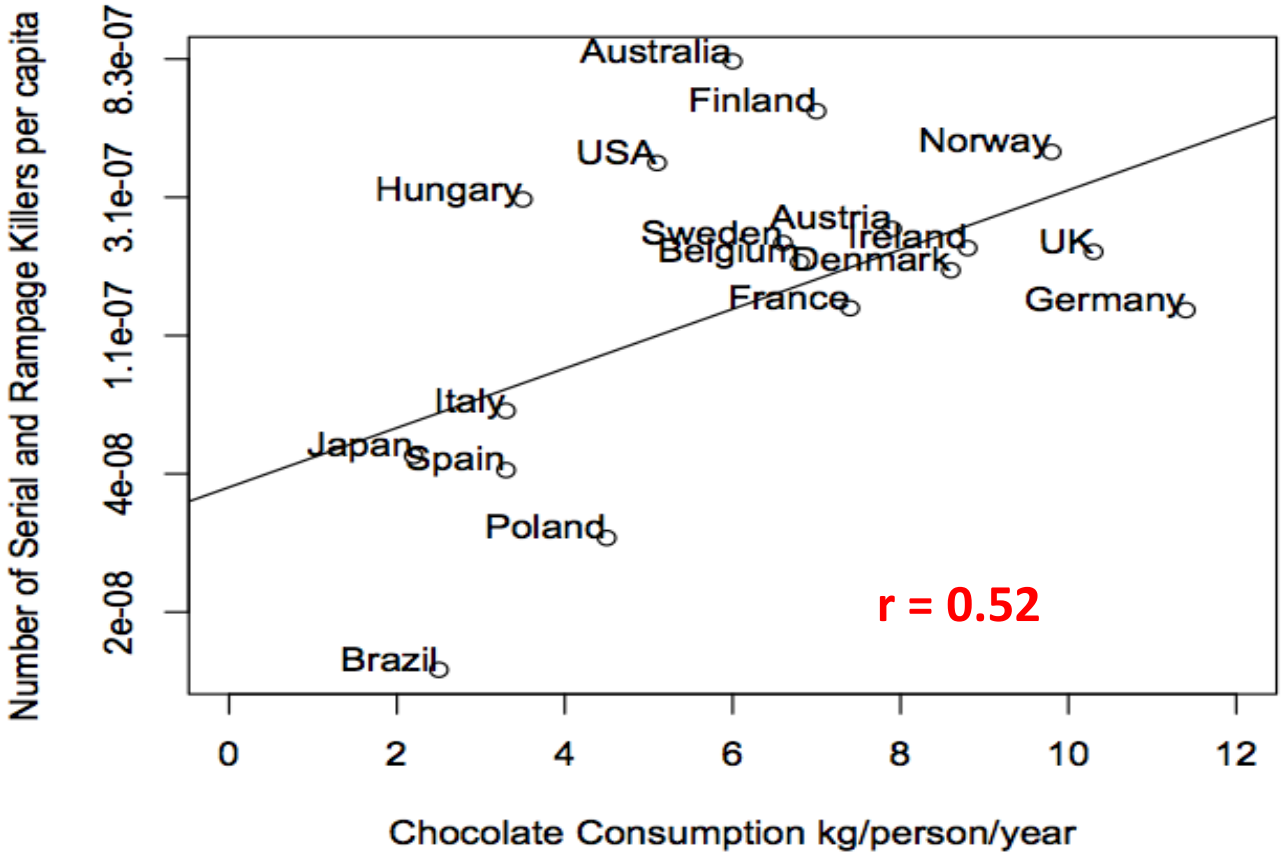
Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

NEW ENGLAND JOURNAL OF MEDICINE

http://www.nejm.org/doi/full/10.1056/NEJMon1211064

Chocolate and serial killers

- What else is related to chocolate consumption?



http://www.replicatedtypo.com/chocolate-consumption-traffic-accidents-and-serial-killers/5718.html

Hypothesis tests for r

Tests the null hypothesis that the population correlation r = 0 NOT that there is a strong relationship!

It is highly influenced by the number of observations e.g. sample size of 150 will classify a correlation of 0.16 as significant!

Better to use Cohen's interpretation

Exercise

- Interpret the following correlation coefficients using Cohen's and explain what it means

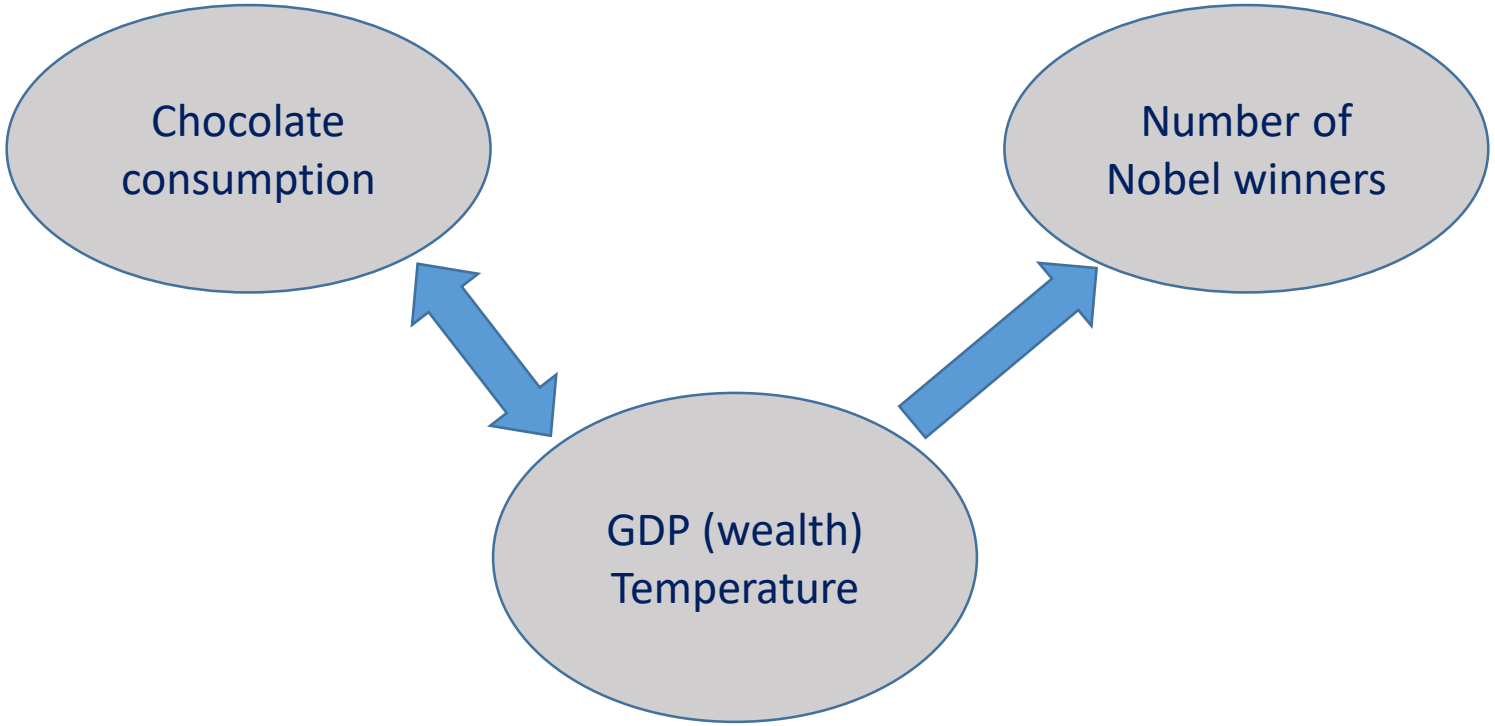
Relationship	Correlation
Average IQ and chocolate consumption	0.27
Road fatalities and Nobel winners	0.55
Gross Domestic Product and Nobel winners	0.7
Mean temperature and Nobel winners	-0.6

Exercise - solution

Relationship	Correlation	Interpretation
Average IQ and chocolate consumption	0.27	Weak positive relationship. More chocolate per capita = higher average IQ
Road fatalities and Nobel winners	0.55	Strong positive. More accidents = more prizes!
Gross Domestic Product and Nobel winners	0.7	Strong positive. Wealthy countries = more prizes
Mean temperature and Nobel winners	-0.6	Strong negative. Colder countries = more prizes.

Confounding

Is there something else affecting both chocolate consumption and Nobel prize winners?



Dataset for today

- Factors affecting birth weight of babies

id	headcirumference	length	Birthweight	Gestation	smoker	motherage
1313	12	17	5.8	33	0	24
431	12	19	4.2	33	1	20
808	13	19	6.4	34	0	26
300	12	18	4.5	35	1	41
516	13	18	5.8	35	1	20
321	13	19	6.8	37	0	28
1363	12	19	5.2	37	1	20
575	12	19	6.1	37	1	19
822	13	19	7.5	38	0	20
1081	14	21	8.0	38	0	18
1636	14	20	8.6	38	0	29

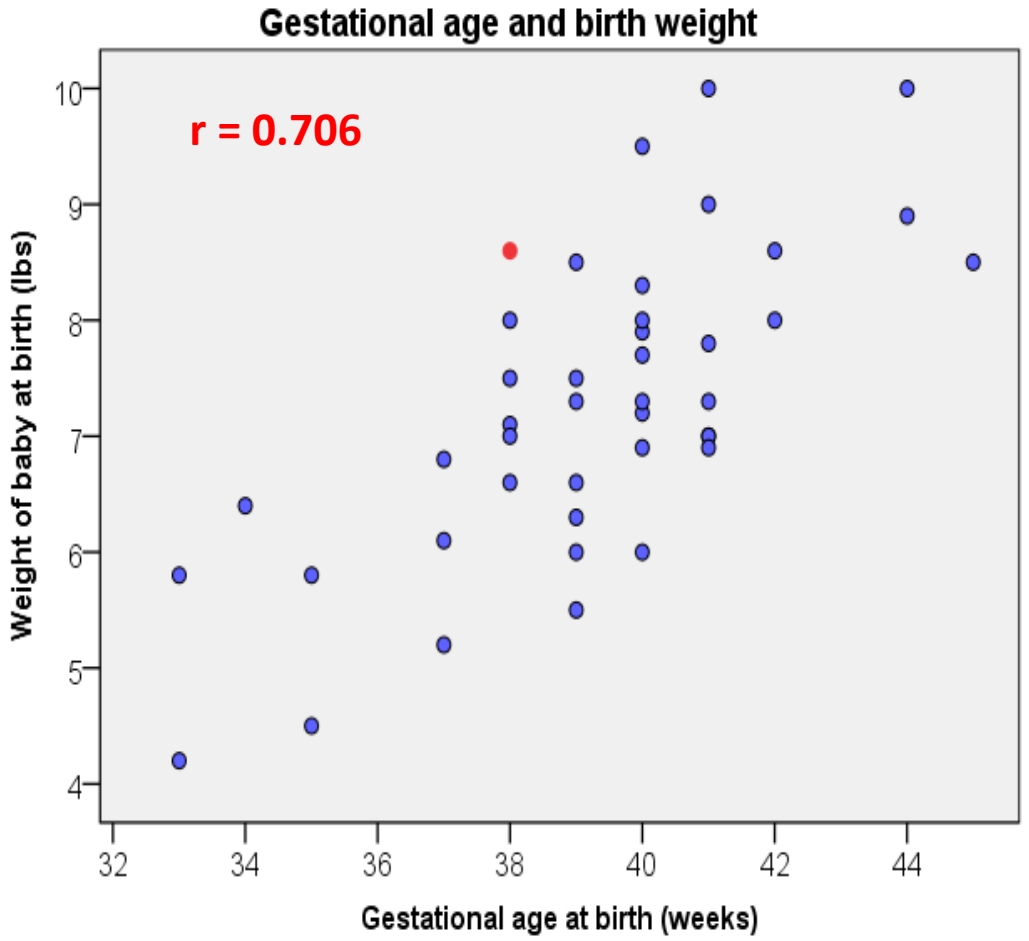
Mother smokes = 1

Standard gestation = 40 weeks

Exercise: Gestational age and birth weight

a) Describe the relationship between the gestational age of a baby and their weight at birth.

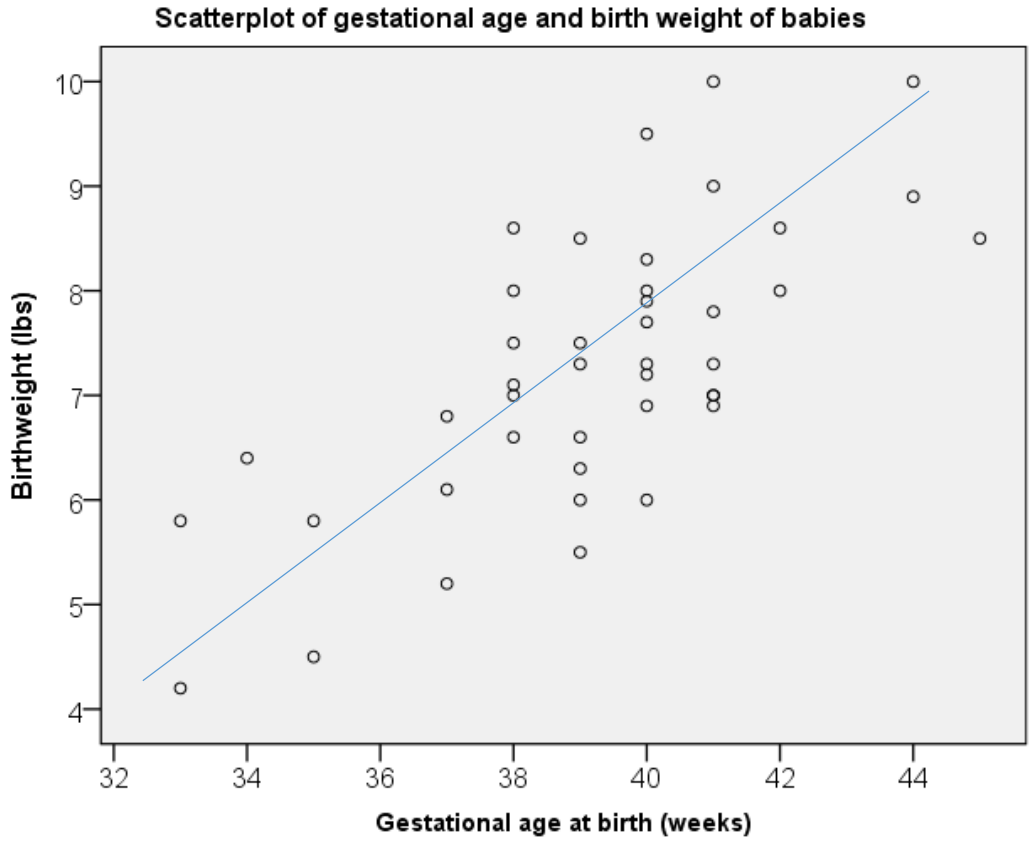
b) Draw a line of best fit through the data (with roughly half the points above and half below)



Exercise - Solution

Describe the relationship between the gestational age of a baby and their weight at birth.

There is a strong positive relationship which is linear



Regression: Association between two variables

- Regression is useful when we want to

a) *look for significant relationships* between two variables

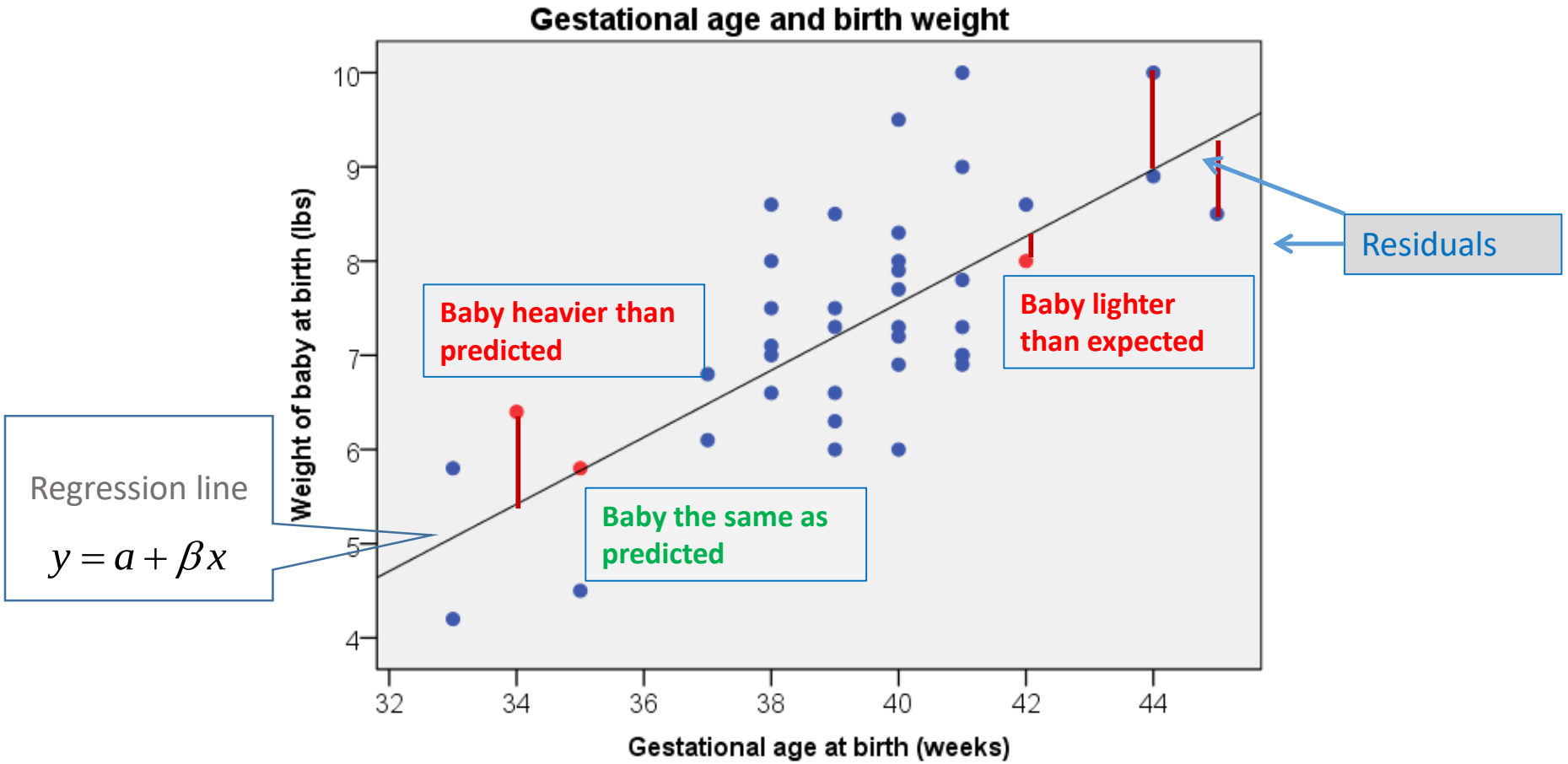
b) *predict* a value of one variable for a given value of the other

It involves estimating the line of best fit through the data which minimises the sum of the squared residuals

What are the residuals?

Residuals

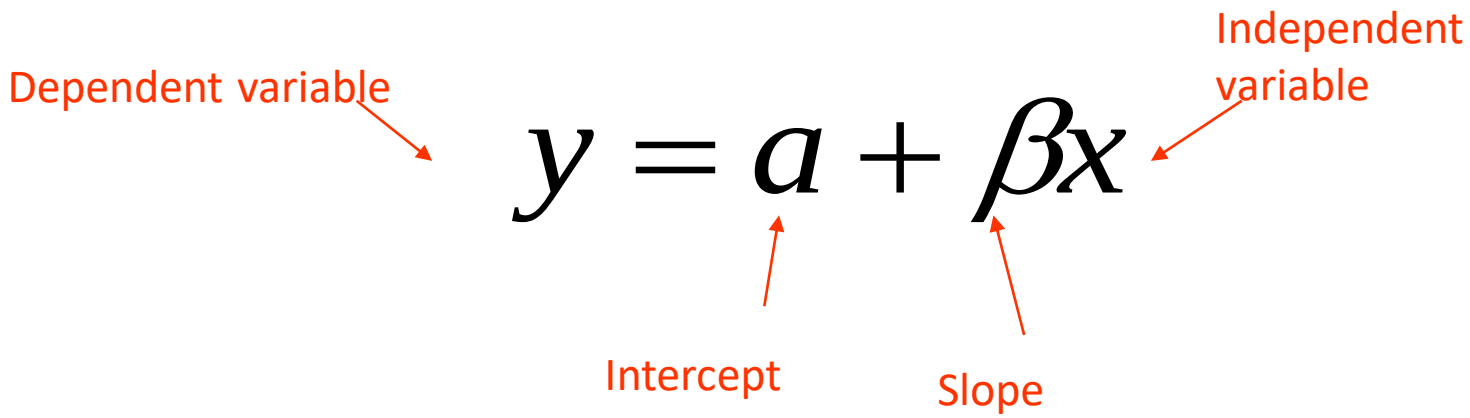
- Residuals are the differences between the observed and predicted weights



X Predictor / explanatory variable (independent variable)

Regression

Simple linear regression looks at the relationship between two Scale variables by producing an equation for a straight line of the form



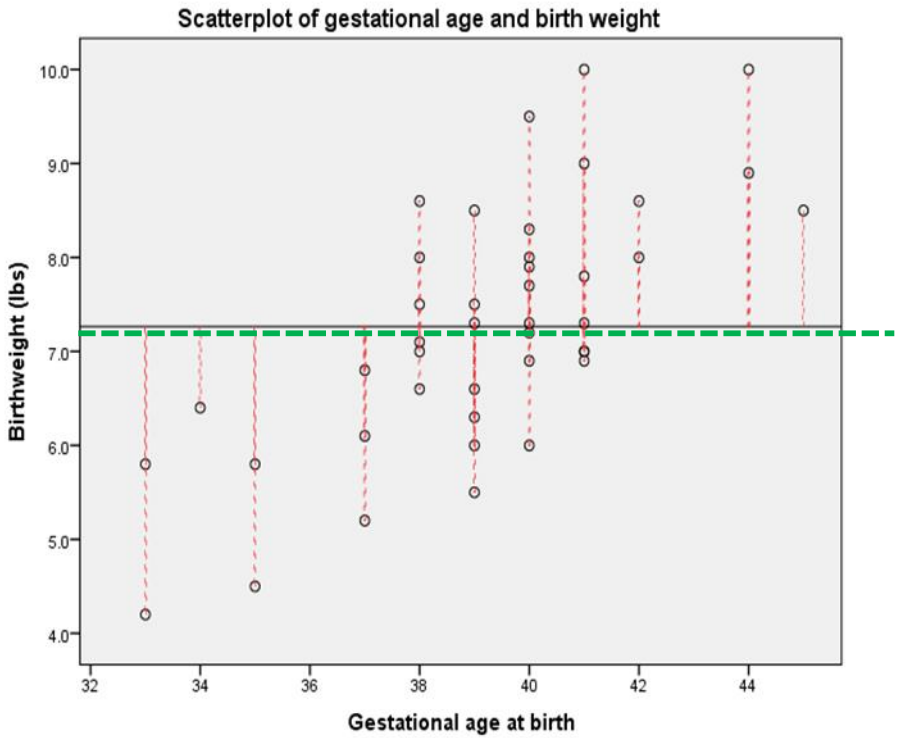
Which uses the independent variable to predict the dependent variable

Hypothesis testing

- We are often interested in how likely we are to obtain our estimated value of β if there is actually no relationship between x and y in the population

One way to do this is to do a test of significance for the slope

$$H_0 : \beta = 0$$



Output from SPSS

- Key regression table:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-6.660	2.212		-3.011	.004
	Gestational age at birth	.355	.056	.706	6.310	.000

a. Dependent Variable: Birthweight (lbs)

Y = -6.66 + 0.36x

P – value < 0.001

- As p < 0.05, gestational age is a significant predictor of birth weight. Weight increases by 0.36 lbs for each week of gestation.

How reliable are predictions? – R^2

How much of the variation in birth weight is explained by the model including Gestational age?

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.706[a]	.499	.486	.9530

a. Predictors: (Constant), Gestational age at birth

b. Dependent Variable: Birth weight (lbs)

Proportion of the variation in birth weight explained by the model $R^2 = 0.499 = 50\%$

Predictions using the model are fairly reliable.

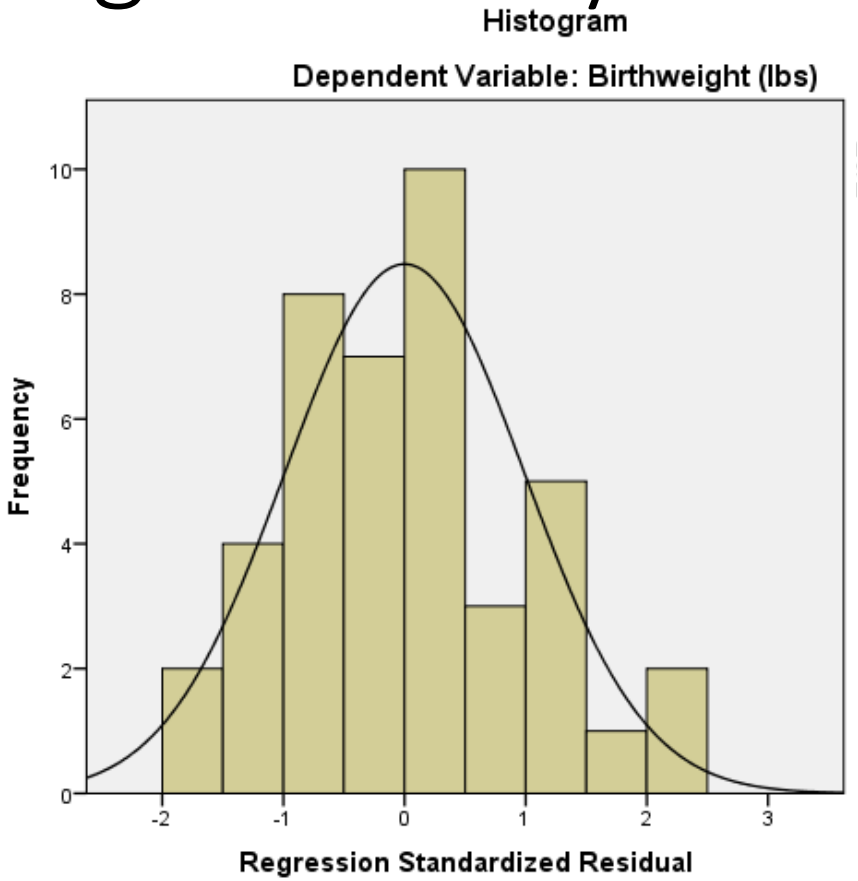
Which variables may help improve the fit of the model?

Compare models using Adjusted R^2

Assumptions for regression

Assumption	Plot to check
The relationship between the independent and dependent variables is linear.	Original scatter plot of the independent and dependent variables
Homoscedasticity: The variance of the residuals about predicted responses should be the same for all predicted responses.	Scatterplot of standardised predicted values and residuals
The residuals are independently normally distributed	Plot the residuals in a histogram

Checking normality



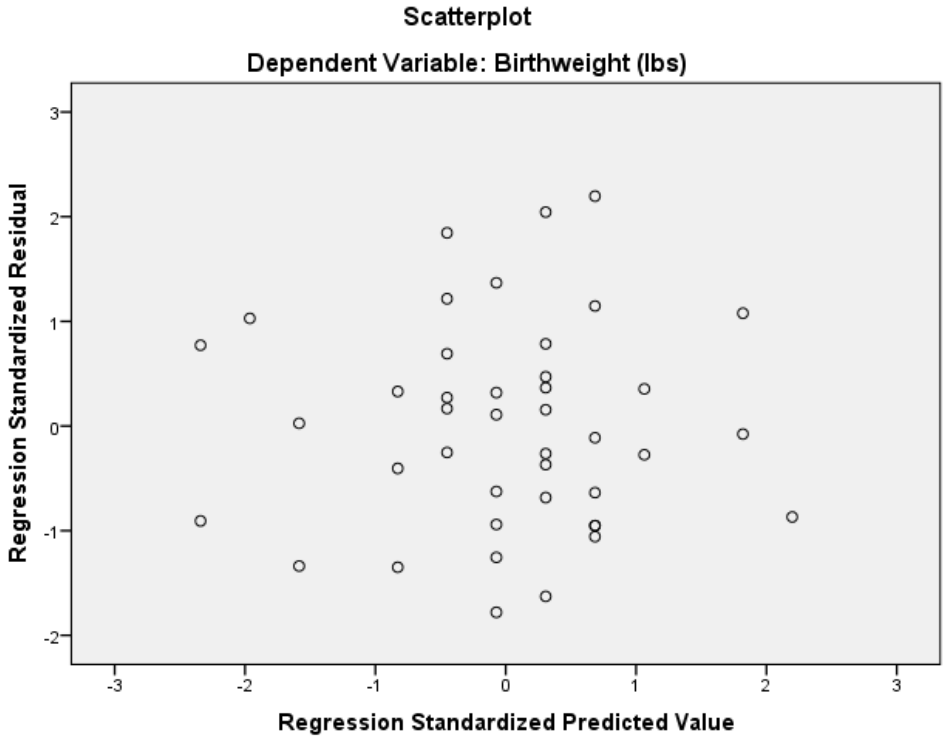
Histogram of the residuals looks approximately normally distributed

When writing up, just say ‘normality checks were carried out on the residuals and the assumption of normality was met’

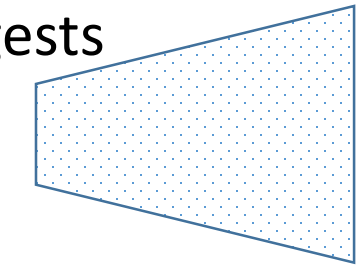
Outliers are outside ± 3

Predicted values against residuals

Are there any patterns as the predicted values increases?

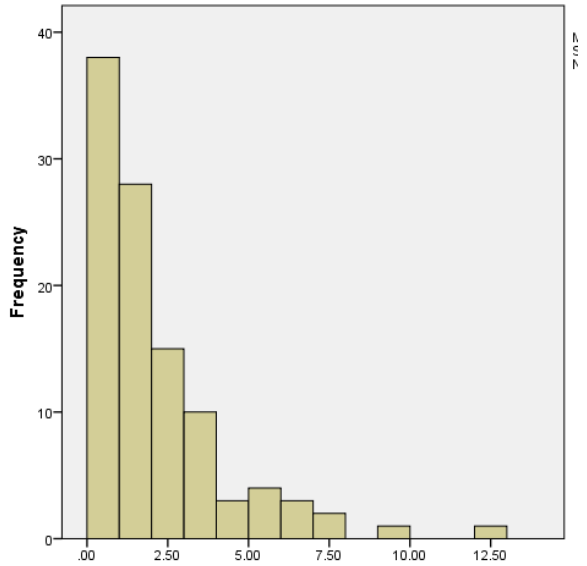


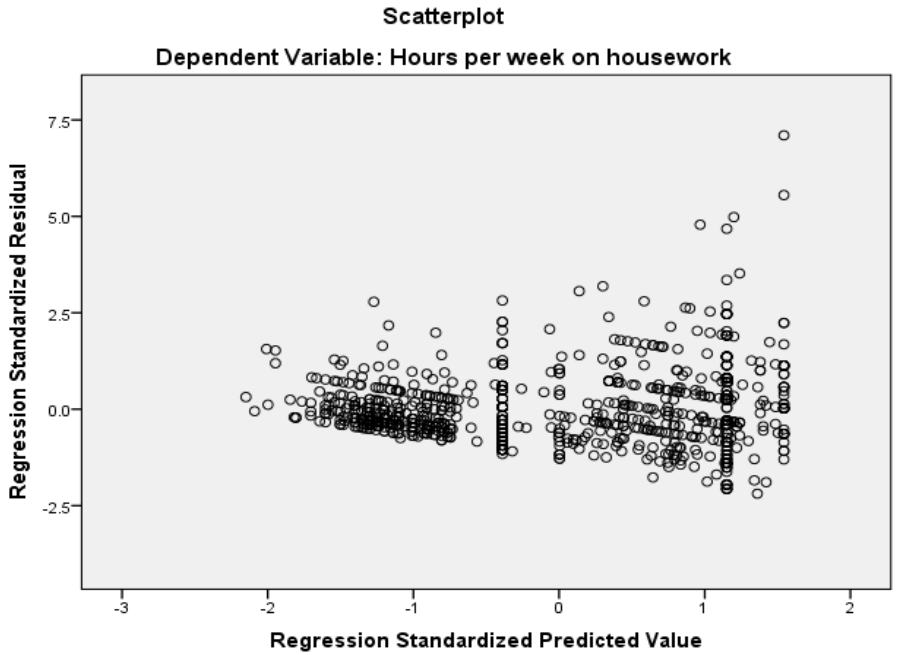
There is a problem with Homoscedasticity if the scatter is not random. A "funnelling" shape such as this suggests problems.



What if assumptions are not met?

- If the residuals are heavily skewed or the residuals show different variances as predicted values increase, the data needs to be transformed
- Try taking the natural log (ln) of the dependent variable. Then repeat the analysis and check the assumptions



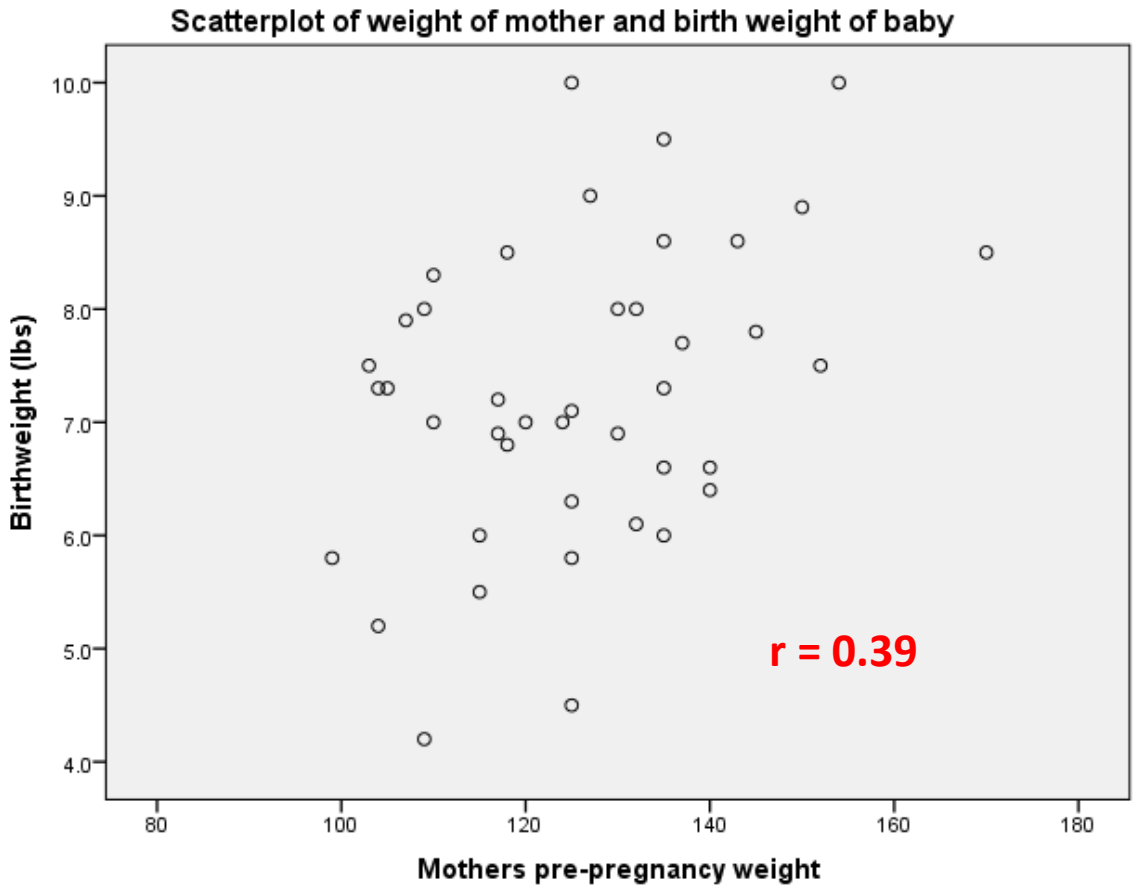


Exercise

- Investigate whether mothers pre-pregnancy weight and birth weight are associated using a scatterplot, correlation and simple regression.

Exercise - scatterplot

- Describe the relationship using the scatterplot and correlation coefficient



Regression question

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.159	1.547		2.042	.048
	Mothers pre-pregnancy weight	.033	.012	.390	2.675	.011

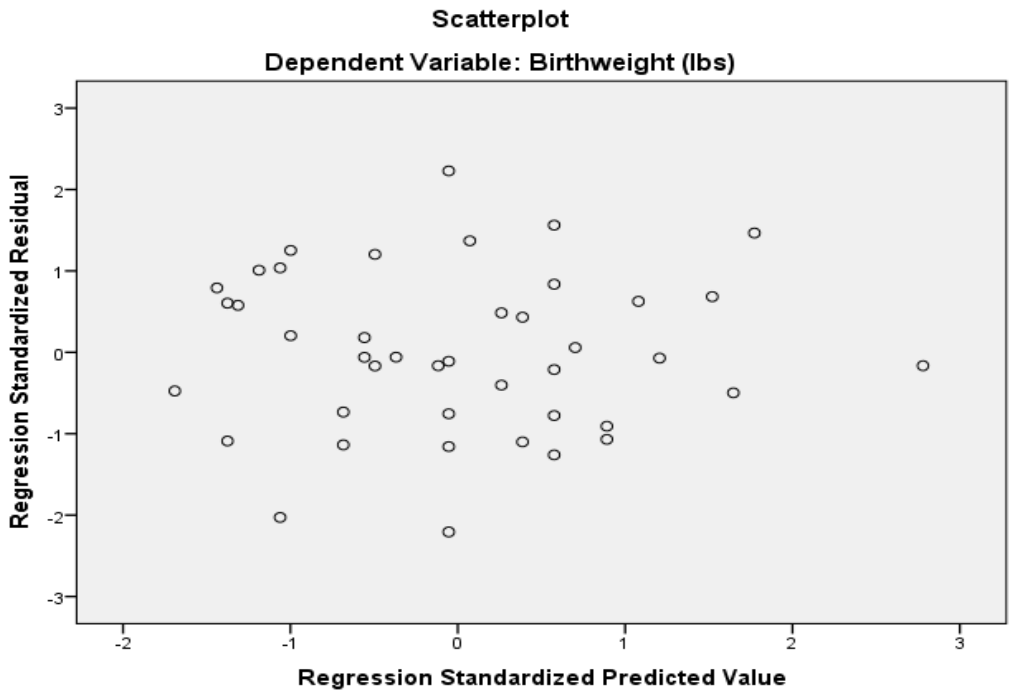
a. Dependent Variable: Birthweight (lbs)

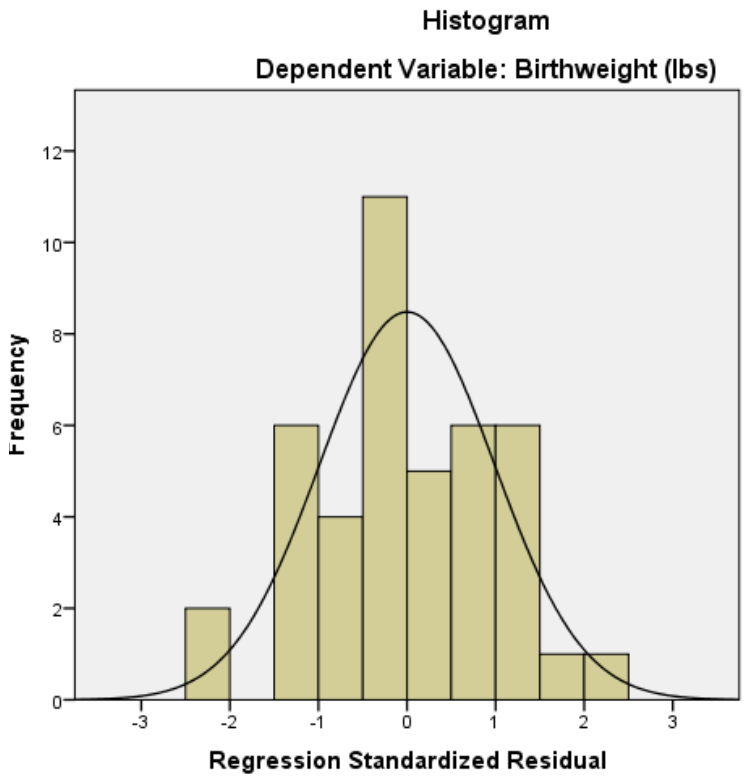
- Pre-pregnancy weight p-value:
- Regression equation:
- Interpretation:

R^2 = 0.152

Does the model result in reliable predictions?

Check the assumptions





Correlation

- Pearson's correlation = 0.39
- Describe the relationship using the scatterplot and correlation coefficient
- There is a moderate positive **linear** relationship between mothers' pre-pregnancy weight and birth weight (r = 0.39). Generally, birth weight increases as mothers weight increases

Regression

Pre-pregnancy weight p-value: $p = 0.011$

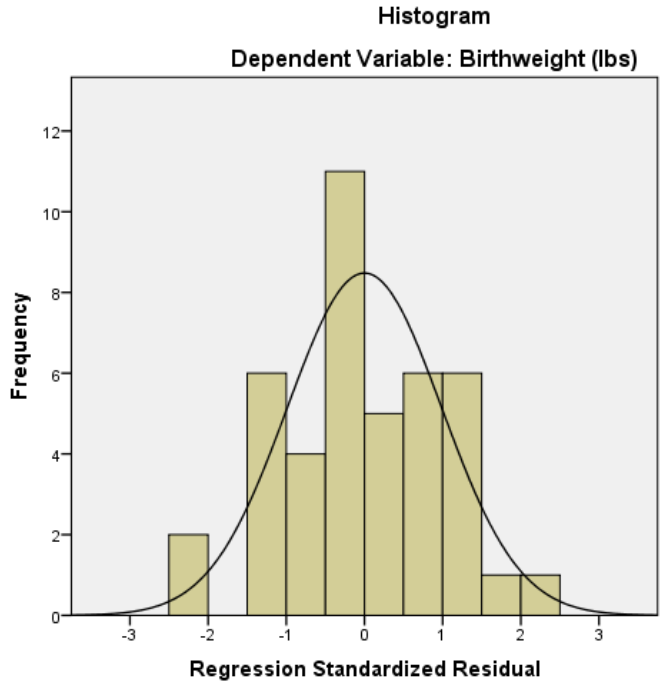
- Regression equation: $y = 3.16 + 0.03x$

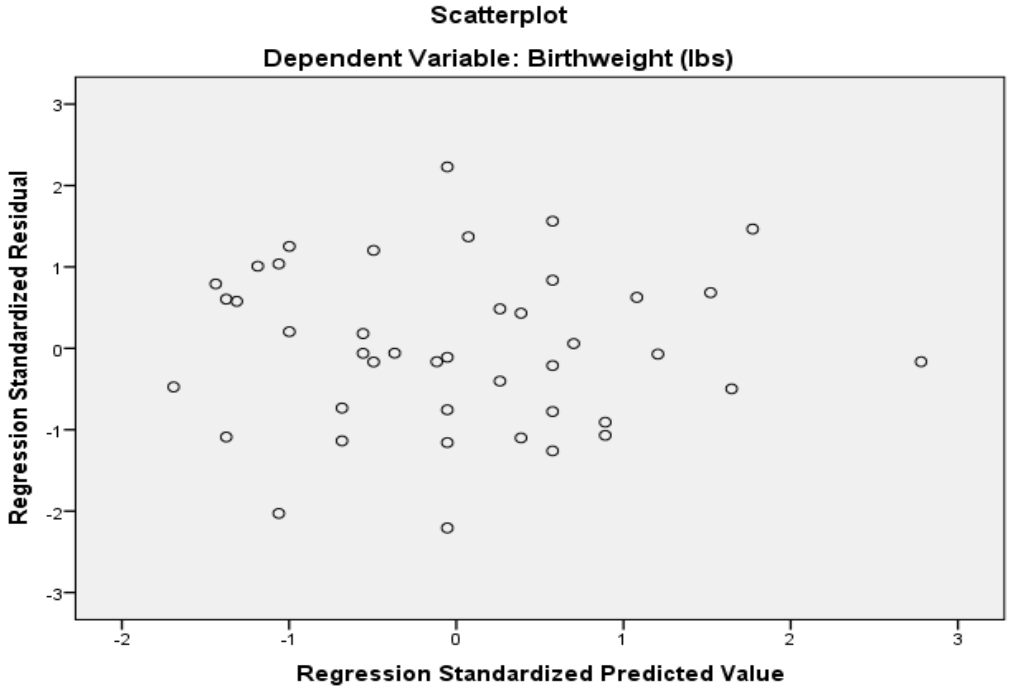
- Interpretation:
- There is a significant relationship between a mothers' pre-pregnancy weight and the weight of her baby ($p = 0.011$). Pre-pregnancy weight has a positive affect on a baby's weight with an increase of 0.03 lbs for each extra pound a mother weighs.

- Does the model result in reliable predictions?
- Not really. Only 15.2% of the variation in birth weight is accounted for using this model.

Checking assumptions

- Linear relationship
- Histogram roughly peaks in the middle
- No patterns in residuals





Multiple regression

- Multiple regression has several binary or Scale independent variables

$$y = a + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

- Categorical variables need to be recoded as binary dummy variables
- Effect of other variables is removed (controlled for) when assessing relationships

Multiple regression

What affects the number of Nobel prize winners?

Dependent: Number of Nobel prize winners

Possible independents: Chocolate consumption, GDP and mean temperature

- Chocolate consumption is significantly related to Nobel prize winners in simple linear regression
- Once the effect of a country's GDP and temperature were taken into account, there was no relationship

Multiple regression

- In addition to the standard linear regression checks, relationships BETWEEN independent variables should be assessed
- Multicollinearity is a problem where continuous independent variables are too correlated ($r > 0.8$)
- Relationships can be assessed using scatterplots and correlation for scale variables
- SPSS can also report collinearity statistics on request. The VIF should be close to 1 but under 5 is fine whereas 10 + needs checking

Exercise

- Which variables are most strongly related?

Correlations

		Birthweight (lbs)	Gestational age at birth	Maternal height	Mothers pre-pregnancy weight
Birthweight (lbs)	Pearson Correlation	1	.706**	.368*	.390*
	Sig. (2-tailed)		.000	.017	.011
	N	42	42	42	42
Gestational age at birth	Pearson Correlation	.706**	1	.231	.251
	Sig. (2-tailed)	.000		.141	.110
	N	42	42	42	42
Maternal height	Pearson Correlation	.368*	.231	1	.671**
	Sig. (2-tailed)	.017	.141		.000
	N	42	42	42	42
Mothers pre-pregnancy weight	Pearson Correlation	.390*	.251	.671**	1
	Sig. (2-tailed)	.011	.110	.000	
	N	42	42	42	42

Exercise - Solution

- Which variables are most strongly related?
- Gestation and birth weight (0.709)

- Mothers height and weight (0.671)

Mothers height and weight are strongly related. They don't exceed the problem correlation of 0.8 but try the model with and without height in case it's a problem.

- When both were included in regression, neither were significant but alone they were

Logistic regression

- Logistic regression has a binary dependent variable
- The model can be used to estimate probabilities

- Example: insurance quotes are based on the likelihood of you having an accident
- Dependent = Have an accident/ do not have accident
- Independents: Age (preferably Scale), gender, occupation, marital status, annual mileage

- Ordinal regression is for ordinal dependent variables