Introduction to Linear Regression

Dr Cameron Hurst
cphurst@gmail.com

CEU and DAMASAC, Khon Kaen University

20th August, 2558

What we will cover....

1. Background
    - Data types
    - Correlation analysis
    - Linear regression and Biostatistical modelling
2. Simple Linear Regression
    - Introduction
    - SLR example
    - SLR model assumptions
3. Multi-variable Linear Regression
    - Motivating example
    - Additional issues: Contribution of Xs
    - Additional issues: Parsimony
    - Additional issues: Multicollinearity
    - Confounding

Conventions

Same as always:

Note:.....

Things to note will occur in a green box

Pitfalls:.....

Common mistakes and things to watch out for will occur in a red box

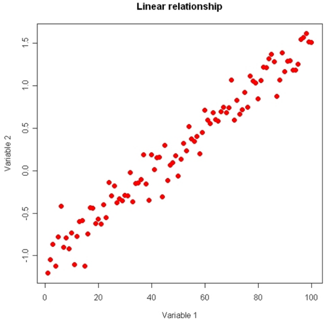
R SYNTAX:....

Most (important) R syntax will be in purple boxes and be in `courier` font. This will help you find it easily when you have to refer back to these notes.

Data for linear regression

- Need our data to be quantitative / numerical / continuous
- Basic test: If data can meaningfully be portrayed on a scatter plot and the form of the relationship is (more or less) linear

☺



Life, the universe and regression

Regression underpins most statistical methods in the discipline of biostatistics

For example:

1. General (Normal) Linear Models: Linear regression and ANOVA
2. Generalized linear models: Logistic regression, Poisson Regression etc.
3. Survival analysis method: Proportional hazards (Cox) regression
4. Methods for longitudinal/spatial data: Linear Mixed Models, Generalized Estimating equations, Generalized Linear Mixed Models...

Pearson's correlation analysis

- Denoted by r (sample statistic), and ρ (population parameter).
- Won't go into calculations for r (understand what it means).
- Takes values between -1 and $+1$ inclusive.
- Measures the strength of **linear** association between two continuous variables

I will only spend about 5 minutes on this very simple method

Propoerties of Pearson's correlation coeffcient, r

- Values of r close to -1 or $+1$ indicate a strong (negative or positive) **linear** relationship
- Values of r close to zero indicate little **linear** relationship
- Even if r close to zero, there still may be a strong relationship in the form of a curve (a non-linear relationship)

Significance test: Pearson's correlation coef., ρ

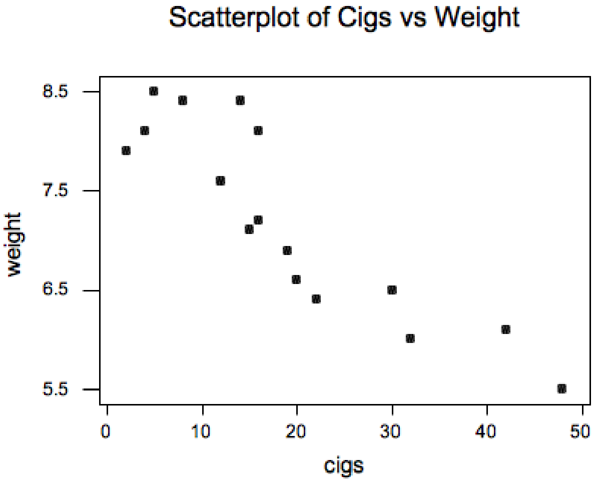
$H_0 : \rho = 0$ (There is no linear relationship between x_1 and X_2)
$H_A : \rho \neq 0$ (There is a linear relationship between x_1 and X_2

- ρ (Greek $\Rightarrow$ Population parameter)
- Conclusion: Significant linear correlation (i.e. $\rho \neq 0$) if p-value < 0.05

Motivating example

Recent studies suggest that smoking during pregnancy affects the birth weights of newborn infants. A sample of 16 women smokers recorded the average number of cigarettes they smoked per day and the birth weight of their child.

Birthweight vs Cigarettes consumed



USING YOUR EYEBALLS: What do you think??

Analysis

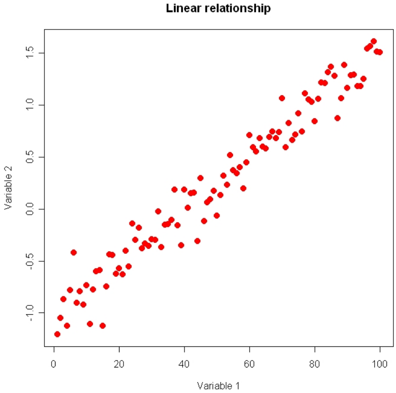
```
Correlation of cigs and weight = -0.884,
P-Value <0.001
```

R= -0.884 suggests WHAT type of relationship????

What if our variables have a non-linear relationship?

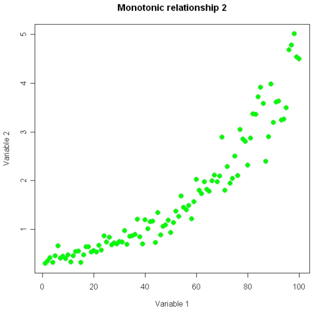
- Pearson correlation can only detect linear relationships between variables.
- Techniques are available for some non-linear relationships
- One such method is **Spearman's correlation coefficient** which can detect relationships which are (at least) monotonic

Monotonic reltaionships: Linearity

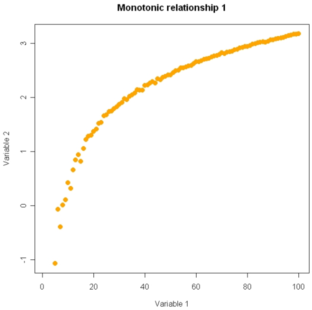


We can think of a linear relationship as walking up (or down) a hill with a constant slope. A linear relationship is **ONE** example of a montotonic relationship

Monotonic reltaionships



Monotonic relationship 2



Monotonic relationship 1

Still **always** walking uphill (or always downhill), but slope can change

Association Vs Causation

- Only if substantive theory (i.e. the science) suggests a causal relationship between variables do we have grounds to use regression analysis
  - i.e. One or more independent variables [IVs] explain a single outcome/dependant variable [DV]
- Otherwise, correlation analysis is all we can use. i.e. We are restricted to talking about associative relationships.
- Cross-sectional studies???

Regression analysis to modelling

- To understand (linear) regression and to understand how MOST other statistical modelling techniques are variants of regression we need to consider the regression **model**
- Models are the mathematical representation (and simplification) of the system under study

Linear Regression Model

Simple Linear Regression: One explanatory variable (X) related to outcome(Y)

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Multi-variable Linear Regression: Y is a linear function of Xs

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \ldots + \beta_{k-1} X_{i,k-1} + \epsilon_i$$

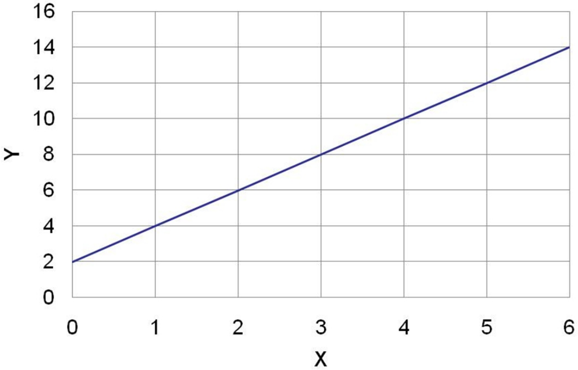
Simple Linear Regression

Simple Linear Regression: **One** explanatory variable related to a response (dependant) variable in a **linear** way.

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Linear: No matter where on X axis, Y-X relationship the same.

Linearity



Rate of change in Y, is constant over entire X domain

Steps in regression analysis

1. Estimate regression equation ('model') i.e. obtain estimates of βs (Software)
2. Assess model adequacy and test hypothesis regarding whether X explains Y
   (a) Model significance (=significance of the single X term)
   (b) Explanatory power (R^2)
   (c) Model Validity (assumptions)
3. Prediction: Sometimes model 'good' enough to predict response variable from values of explanatory variable (rarely case in 'observational' setting).

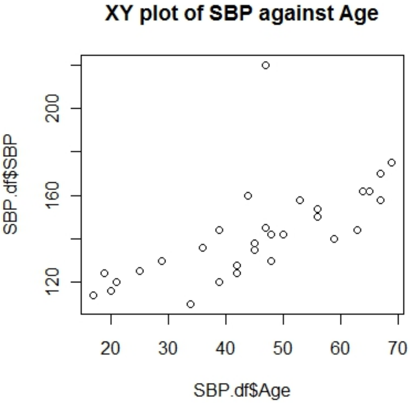
Simple 'bare bones' example

Considering Systolic blood pressure (SBP) in adults (our sample ages range from 17 - 69 years)...

Can we explain (variation in) SBP based on (variation in) Age?

XY scatter plot

Eyeball data: Before anything look at relationship



Using R for SLR: Data input

R syntax: Read in data and generate scatter plot

```
setwd("D:/myR")

SBP.df<-read.csv("Bloodpressure.csv")

plot(x=SBP.df$Age, y=SBP.df$SBP, main="Plot:  SBP vs Age")
```

1. Set working directory
2. Read in data and dump to data frame
3. Plot SBP against age
    - Include title on plot

Using R for SLR: Regression analysis

R syntax: Run a simple regression analysis

```
my.SLR<-lm(SBP~Age, data=SBP.df)
summary(my.SLR)
anova(my.SLR)
```

1. Run regression
2. Show βs and R-squared
3. Test significance of OVERALL model

Key point:

Note: In LINEAR regression analysis, $Y \sim X$ represents the model $Y = \beta_0 + \beta_1 X_1 + \epsilon$

You will see this time and time again in R

Output 1:

```
> summary(my.SLR)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  98.7045    10.0142   9.856 1.32e-10 ***
Age           0.9697     0.2102   4.613 7.99e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1

Residual standard error: 17.32 on 28 degrees of freedom
Multiple R-squared: 0.4318,Adjusted R-squared: 0.4115
F-statistic: 21.28 on 1 and 28 DF,  p-value: 7.991e-05
```

Ouput2:

```
> anova(my.SLR)
Analysis of Variance Table

Response: SBP
          Df Sum Sq Mean Sq F value    Pr(>F)
Age        1 6385.0  6385.0  21.277 7.991e-05 ***
Residuals 28 8402.4   300.1
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1
```

Step 1: Estimate linear of best fit

Remember the simple linear regression model

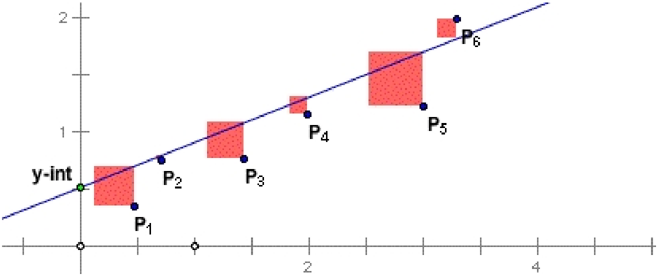
$$Y_i = b_0 + b_1 X_i$$

In this case (from Output 1),

$$SBP_i = 98.7 + 0.97 Age_i$$

Note here that b (the sample estimates) rather than β (the population parameters) are used

Least squares: How linear regression estimates the line of best fit (β_0 and β_1)



The line of best fit is the (straight) line whose distance from all points is minimized (smallest sum of red squares possible)

Interpreting coefficients: β_0, β_1

β_0 is y-intercept (b_0 is the sample estimate)
Value of Y when X = 0
The SBP will be **???** if you are zero years old (newborn baby).

β_1 is the slope (b_1 is the sample estimate)
The change in Y for each unit change in X
As you age 1 year we would expect (i.e. on average) your SBP to change by **???**.

Step2a: Significance Tests for Model = Test of $\beta_1 = 0$

TWO HYPOTHESES:
1. The p-values are for tests that the POPULATION intercept is significantly different from zero.
For β_0

$H_0 : \beta_0 = 0$
$H_A : \beta_0 \neq 0$ (MOSTLY..who cares?)

In words:
H_0 : The SBP of new born babies is zero
H_A : The SBP of new born babies differs from zero

Relevant???

Step2a: Significance Tests for Model = Test of $\beta_1 = 0$

2. The p-values are for tests that the POPULATION slope is significantly different from zero.
For β_1

$H_0 : \beta_1 = 0$
$H_A : \beta_1 \neq 0$

In words:
H_0 : Age does not explain variation in SBP
H_A : Age DOES explain variation in SBP

Step2b: Assessing the model: The Coefficient of Determination, R^2

Represented by R^2 (measures goodness of model fit)
Do you think it's related to Pearson's corr coefficient: r?
(Literally represents the square of Pearson's corr. coefficient)

R^2 **measures the percentage of variability in Y that is explained by X**

Interpret R^2 for the Blood pressure data(Output 1):
$R^2 = 43.2\%$ or $R^2 = 0.432$ (as proportion)
Hint: Write it down (by hand) $\ggg$

Other considerations

- The p-value in the analysis of variance table is equivalent to a test for the slope $= 0$ when using a single predictor variable.
- That is, in **Simple** Linear Regression, the significance of the overall model is always the **same** as the significance of the (single) explanatory variable

Step2c: Simple linear regression assumptions

Three main assumptions. First two are easy, the third requires a little more thought.

1. Y (dep. var.) and X (expl. var.) are linearly related.
2. Ys are serially independent
3. The remaining part of Y (the residual) is normally distributed around zero and with a constant variance:

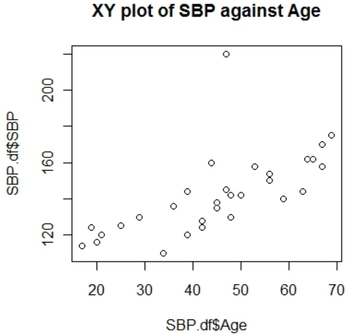
$$\epsilon \sim N(0, \sigma^2)$$

Step2c: Simple linear regression assumptions

1. Y (dep. var.) and X (expl. var.) are linearly related.
   Why would we use a linear model otherwise?
2. Ys are serially independent
3. The remaining part of Y (the residual) is normally distributed around zero and with a constant variance:

$$\epsilon \sim N(0, \sigma^2)$$

Linearity assumption: XY scatter plot



If we eyeball the data, and the data appear (approximately) linearly related.....

Step2c: Simple linear regression assumptions

1. Y (dep. var.) and X (expl. var.) are linearly related.
2. Ys are serially independent
3. The remaining part of Y (the residual) is normally distributed around zero and with a constant variance:

$$\epsilon \sim N(0, \sigma^2)$$

Serial independence assumption

- Usually we can answer this question by just thinking about the study design
- In most longitudinal studies the data are correlated
  E.g. My SBP today will be correlated with my SBP yesterday
- In many cross-sectionally designs, independence assumption safe
- One exception to this is in studies that contain a clustering design effect
  - E.g.1 Physical activity behaviour of people living in the same area: seeing other people jog may mean I am more likely to jog myself
  - E.g.2 Sets of patients treated in groups(clusters) defined by teams/clinics/hospitals

Step2c: Simple linear regression assumptions

1. Y (dep. var.) and X (expl. var.) are linearly related.
2. Ys are serially independent
3. The remaining part of Y (the residual) is normally distributed around zero and with a constant variance:

$$\epsilon \sim N(0, \sigma^2)$$

Definition of residuals

The residual is the difference between our model prediction of y, $\hat{y}_i$, and what we observe y to be, y_i

That is, $\epsilon_i = y_i - \hat{y}_i$

Investigating the $\epsilon \sim N(0, \sigma^2)$ assumption and how it might be violated can tell us a lot about what's going on.

This investigation is called: **RESIDUAL ANALYSIS**

Step2b: Residuals assumption

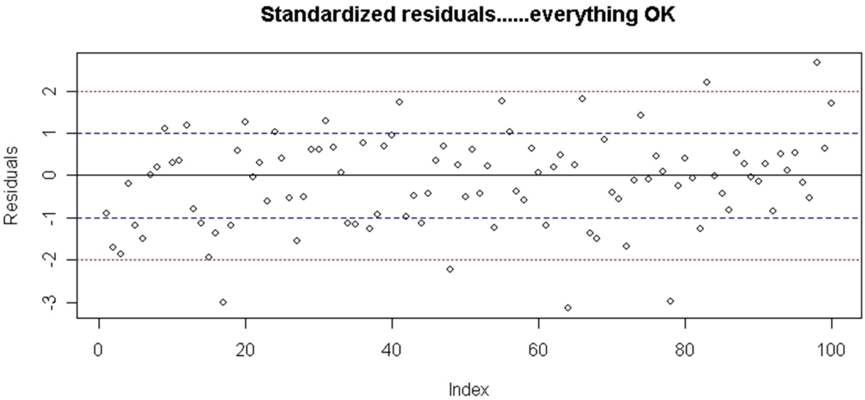
The mathematical statement:

$$\epsilon \sim N(0, \sigma^2)$$

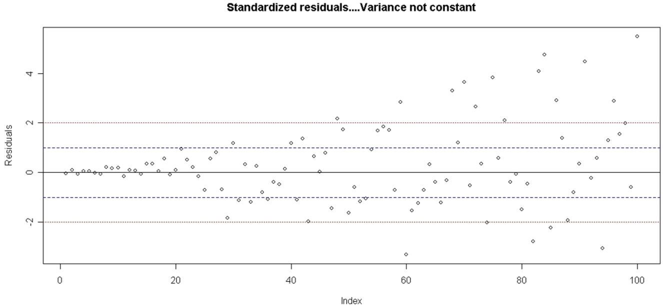
has a number of 'sub-statements':

1. Residuals (errors) are normally distributed
2. Residuals have a mean of zero
3. Residuals have a constant variance(aka:homoscedacisity)

Residuals are normally distributed + Residuals have mean of zero

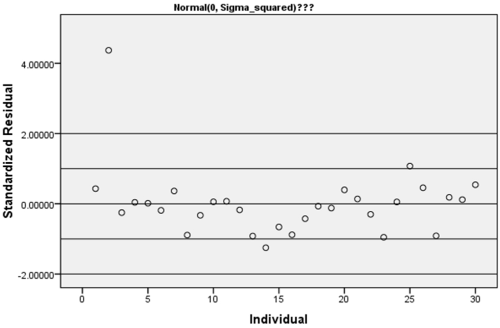


Example of a Violation of $\epsilon \sim N(0, \sigma^2)$: Residuals without constant variance (Heteroscedasticity)



What can we do? Transform? Weighted Least Squares approach?

Residuals for SBP data:



Step 3: Prediction

Interpolation:

(a) Predict the SBP for somebody who is 50 years old

✓

Extrapolation:

(a) Predict SBP for a five year old
(b) Predict SBP for a 85 year old

⊘

Finally, do you think the model is good enough (i.e. R^2) to make predictions?????

Recap:

Three steps in simple linear regression analysis:

1. Estimate equation (find b_0 and b_1)
2. Assess adequacy of model
   - Hypothesis tests (significance)
   - Explanatory power (R-squared)
   - Assumptions (especially residuals)
3. (if appropriate) Use 'good' model to make predictions

Multi-variable Linear Regression

Now we will consider the case where we have **More than one explanatory variable**

- The **Multivariable** linear regression model is exactly the same as the Simple linear regression model just with additional explanatory variables.

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \ldots + \beta_{k-1} X_{i,k-1} + \epsilon_i$$

- Each explanatory variable has a slope associated with it.

Steps in MLR analysis

New steps are in **bold** (i.e. Specific to MLR)

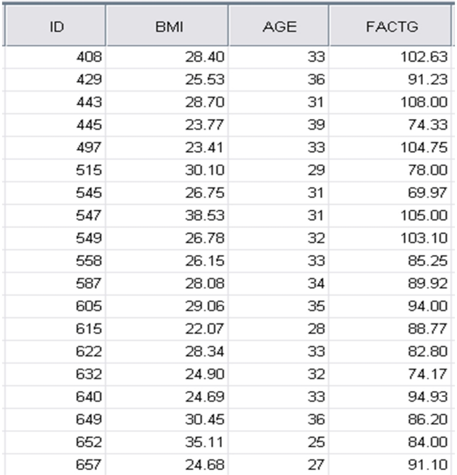
1. Estimate regression equation ('model')
2. Significance
   (a) **OVERALL** Model significance (ANOVA F test)
   (b) **Consider significance of individual covariates (Xs)**
   (c) Explanatory power (**adjusted R-sq**)
   (d) Model Validity (assumptions)
      - Residuals
      - Independence
      - **Multicollinearity**

   (e) **Parsimony**
3. If model good enough, make predictions

MLR: Motivating example

- Dataset containing 3 variables: **BMI** (Body Mass Index) , **Age** and **pf-QoL**, a Physical functioning sub-scale of the Functional Assessment of Cancer Therapy-General questionnaire. A physical quality of life measure for people undergoing treatment for cancer.
- We suspect that Age and BMI can explain variation in pf-QoL

The data.....



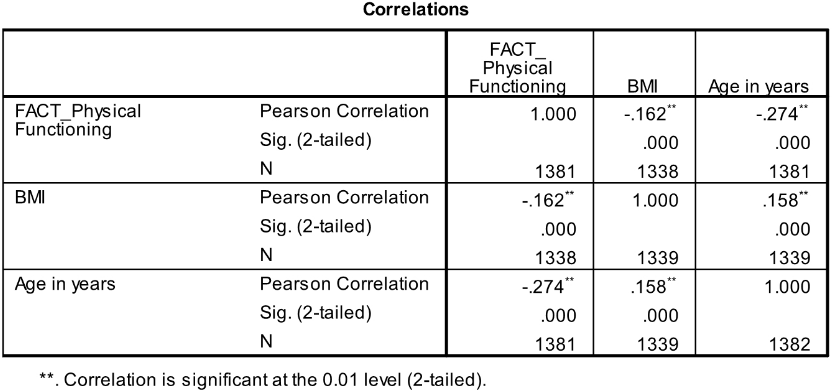
ID	BMI	AGE	FACTG
408	28.40	33	102.63
429	25.53	36	91.23
443	28.70	31	108.00
445	23.77	39	74.33
497	23.41	33	104.75
515	30.10	29	78.00
545	26.75	31	69.97
547	38.53	31	105.00
549	26.78	32	103.10
558	26.15	33	85.25
587	28.08	34	89.92
605	29.06	35	94.00
615	22.07	28	88.77
622	28.34	33	82.80
632	24.90	32	74.17
640	24.69	33	94.93
649	30.45	36	86.20
652	35.11	25	84.00
657	24.68	27	91.10

A quick word on model selection

- A whole other (very important) topic is how we decide which combination of variables should be considered in our (final) model
- Not within scope to discuss here
- Also, we are only considering two (potential) predictors, (BMI and Age) so not too complicated
- We will just FORCE our predictors into the model
- BUT you should be aware other 'Model selection' strategies available (e.g. Stepwise, Best subset, **Purposeful selection of covariates** etc.)

Correlation

Let's start by perusing the correlation matrix:



Correlations

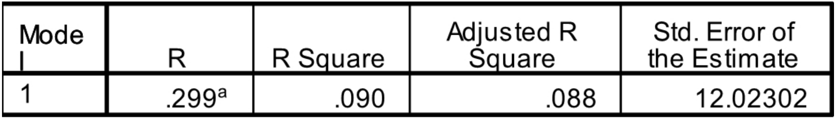
		FACT_ Physical Functioning	BMI	Age in years
FACT_Physical Functioning	Pearson Correlation	1.000	-.162**	-.274**
	Sig. (2-tailed)		.000	.000
	N	1381	1338	1381
BMI	Pearson Correlation	-.162**	1.000	.158**
	Sig. (2-tailed)	.000		.000
	N	1338	1339	1339
Age in years	Pearson Correlation	-.274**	.158**	1.000
	Sig. (2-tailed)	.000	.000	
	N	1381	1339	1382

**. Correlation is significant at the 0.01 level (2-tailed).

From correlation matrix...

- **Age** seems to be (somewhat) negatively correlated with **pf-QoL** suggesting that the older people are (undergoing cancer therapy), the less their physical quality of life
- **BMI** also seems to be negatively correlated with **pf-QoL**
- Note (for later) that **BMI** (an X variable) also seems to correlate to **Age** (another X variable)

Results of MLR: Explanatory power



Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.299[a]	.090	.088	12.02302

The (unadjusted)$R^2 = 0.09$
The **adjusted**-$R^2 = 0.088$

The overall model explains 8.8% of the variation in physical Quality of Life.

Why Adjusted-R^2??

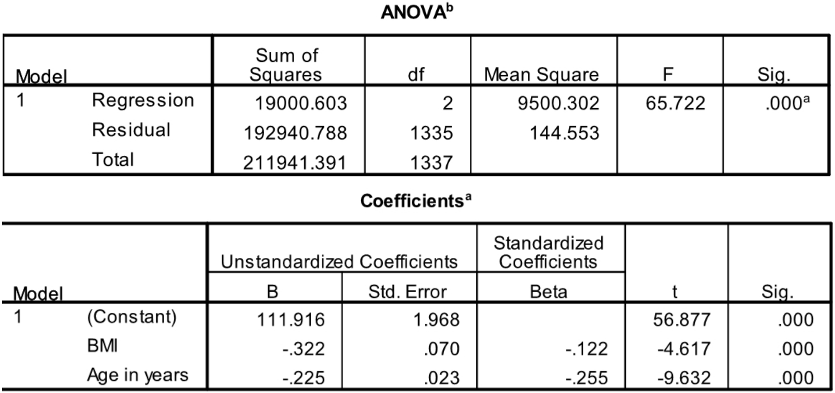
Additional issues in MLR: Adjusted-R^2

- **Explanatory power(R^2)**: Both SLR and MLR produce R^2 values. However, we have to account (penalize) for the number of variables used to explain Y. So in MLR we use Adjusted-R^2

- Adjusted-R^2 adjusts for the number of explanatory variables used to explain Y

- Non-adjusted R^2 becomes increasingly (upwardly) biased with increased number of Xs. That is, it overestimates the explanatory power of the model

MLR significance: Overall model and individual predictors



ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	19000.603	2	9500.302	65.722	.000[a]
	Residual	192940.788	1335	144.553		
	Total	211941.391	1337			

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	111.916	1.968		56.877	.000
	BMI	-.322	.070	-.122	-4.617	.000
	Age in years	-.225	.023	-.255	-9.632	.000

Interpretation

ANOVA table: Overall model is significant (F $=$ 65.7, p$<$0.05)

From 'coefficient' table
$H_0 : \beta_0 = 0$ [WHO CARES]
$H_0 : \beta_{BMI} = 0$ (t=-4.617, p$<$0.05)
Reject Ho. BMI explains variation in physical functioning in this population. $b_{BMI} = -0.322 \Rightarrow$ As BMI goes up a single unit, (on average) pf-QoL goes down 0.322 units
$H_0 : \beta_{Age} = 0$ (t= -9.632, p$<$0.05)
Reject Ho. Age explains variation in physical functioning. In this case, $b_{Age} = -0.225 \Rightarrow$ every year older the patient gets, (on average) their pf-QoL decreases by 0.225 units

Contribution of individuals predictors

- Since we have more than one explanatory variable, useful knowing which (significant) variables contribute more in explaining variation in the response variable.

- Standardized βs (denoted β_Z) help indicate the **relative** contribution (of the variation explained in Y) of each explanatory variable.

- In above example: it is clear that Age explains considerably more than BMI (β_Z for Age = -0.255 vs β_Z for BMI = -0.122)

β vs β_Z

Why can't we just use non-standardized β to gauge the relative importance of individual covariates (explanatory variables)????

Answer:

Additional issues in MLR: Parsimony

- In MLR we also need to consider model parsimony
- Parsimony (in MLR) is the principle of explaining the most variation with the least number of variables

REM: Occam's razor: simplest answer is often the best.

Parsimony

Consider the three models below (all of which we can assume to be 'valid' and 'significant')

Model 1: $R^2 = 0.5$
$Y_i = \beta_0 + \beta_1 X_{i,1}$
Model 2: $Adj - R^2 = 0.97$
$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3}$
Model 3: $Adj - R^2 = 0.975$
$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3} + \beta_4 X_{i,4} + \beta_5 X_{i,5}$

Which model would you select?

Additional issues in MLR: Parsimony

- A bunch of statistics that consider **both** parsimony and explanatory power are the 'Information Criteria' type statistics.
- Two well known IC stats are:
  - **AIC** (Akaike Info. Crit.)
  - **BIC** (Bayesian Info. Crit.)

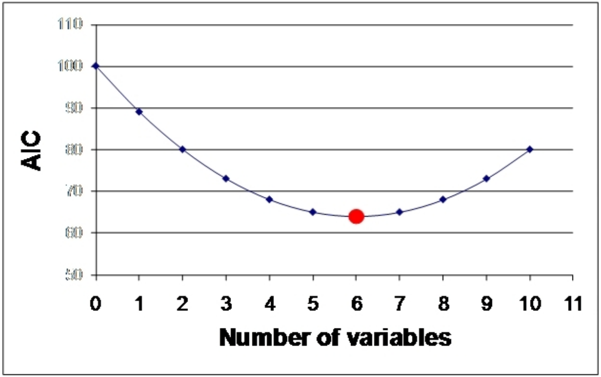
Information criterion

Basic idea:
IC = Lack of Fit(model) + penalty(num parameters)

Lack of Fit = residual = obs y - pred y
i.e. difference between model and reality (data)

Good model will have low lack of fit

AIC



Low value of IC better: Best model in this situation has 6 variables

Information criteria

- Information criteria statistics are absolute and tend not to have much meaning across studies (they are a comparative measure for a set of models predicting a particular outcome, for a particular set of data)
- However, the advantage of IC statistics is that they can be used for a wide range of models (not just linear regression) where a model can be compared to the data
- For example they are often used in Generalized linear models (e.g. Logistic regression) where there is no R^2 value

Additional issues: Multicollinearity

- One of the trickier issues that arises in MLR, especially for observational (e.g. cohort) studies
- Multicollinearity occurs when our 'so-called' independent (explanatory) variables are not independent (i.e. they are correlated)

What are the implications of multicollinearity?

First, the reason explanatory variables need to be independent is so we can attribute variation in an outcome variable **uniquely** to each explanatory variables.

For example, consider vocabulary in children:
$Vocabulary = \beta_0 + \beta_1 Age + \beta_2 ShoeSize$

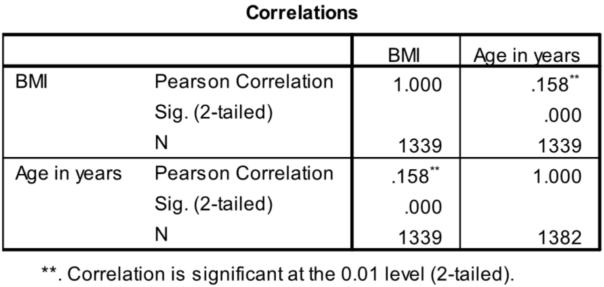
Second, multicollinearity leads to unstable βs (Specifically, inflated confidence intervals-see later)

Multicollinearity vs Confounding

- Multicollinearity **can** be the physical manifestation of confounding in statistical modelling.
- In the last example: We cannot physically separate the variation in vocabulary due to age from that explained by shoe size.
- What about in pf-QoL = f(Age, BMI) example?

How do we identify multicollinearity?

- Initially keep it simple: The correlation matrix (of X variables)



Correlations

		BMI	Age in years
BMI	Pearson Correlation	1.000	.158**
	Sig. (2-tailed)		.000
	N	1339	1339
Age in years	Pearson Correlation	.158**	1.000
	Sig. (2-tailed)	.000	
	N	1339	1382

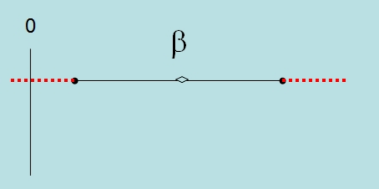
**. Correlation is significant at the 0.01 level (2-tailed).

- In this case, explanatory variables are (weakly) correlated i.e. collinear

When is multicollinearity a problem?

A number of multicollinearity diagnostic tools. Simplest is the Variance Inflation Factor (VIF)

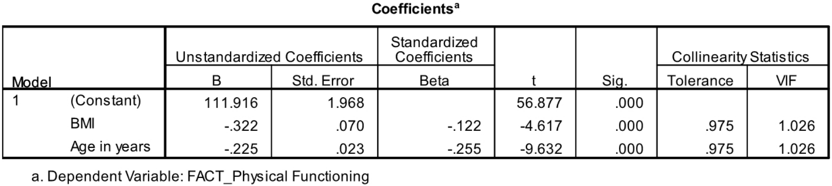
VIF indicates $\uparrow \beta$ Variances due to presence of other collinear variables in model.



Hard and fast rule: VIF $<$ 5 ☺

Motivating example: physical QoL in cancer patients

What about the correlation between Age and BMI. Does that cause a substantial problem (risk of a type II error) in our analysis?



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	111.916	1.968		56.877	.000		
	BMI	-.322	.070	-.122	-4.617	.000	.975	1.026
	Age in years	-.225	.023	-.255	-9.632	.000	.975	1.026

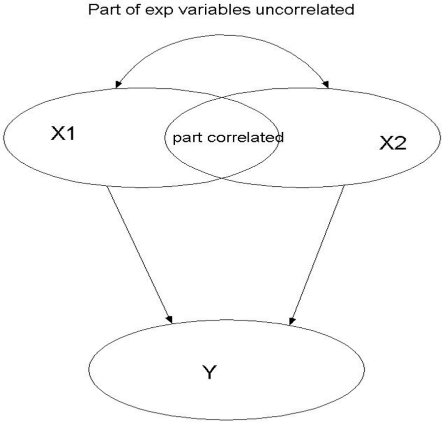
a. Dependent Variable: FACT_Physical Functioning

VIF < 5 ☺

Does multicollinearity always cause problems?

- No. Sometimes parts of the Xs correlated with each other don't relate directly to the Y variable.
- In other words, two X variables can be moderately correlated and yet the VIF (and impact of multicollinearity) low.

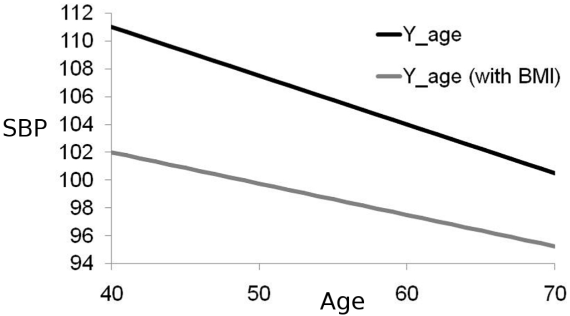
Is multicollinearity always a problems?



Confounding

- The definition of a confounder is a variable that interferes with the relationship between two others.
- A statistical definition of a confounder (in a linear regression context) is one that **changes the slope (effect) of a particular explanatory variable when the confounder is added to the model**.
- In our example, the effect of Age (on pf-QoL) may change with the addition of BMI into the model (this would make BMI a confounder)

Confounding in linear regression



As we can see, the addition of (adjustment for) the potential confounder, BMI, has slightly altered the relationship between Age and pf-QoL

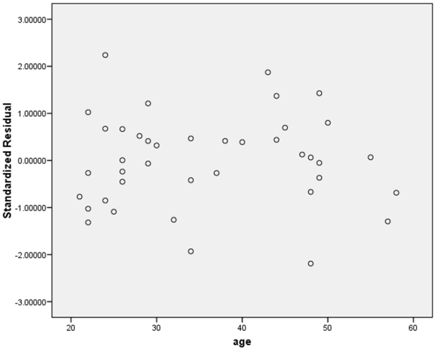
So which model is more appropriate???

What have I missed??

- In the QoL example, I have not performed a residual analysis (which should be conducted in much the same way as for SLR).
- Recall: Regression is not valid unless we can demonstrate:

$$\epsilon \sim N(0, \sigma^2)$$

Residual Analysis



OK????

Almost there: MLR with R

Works very similarly as SLR. For this example:

R syntax: Running a multivariable linear regression

```
my.model<-lm(Qol~Age+BMI, data=QoL.df)
summary(my.model)
anova(my.model)
```

THANK-YOU

Watch this space!!!!