Statistics for Medical Research (Part III)

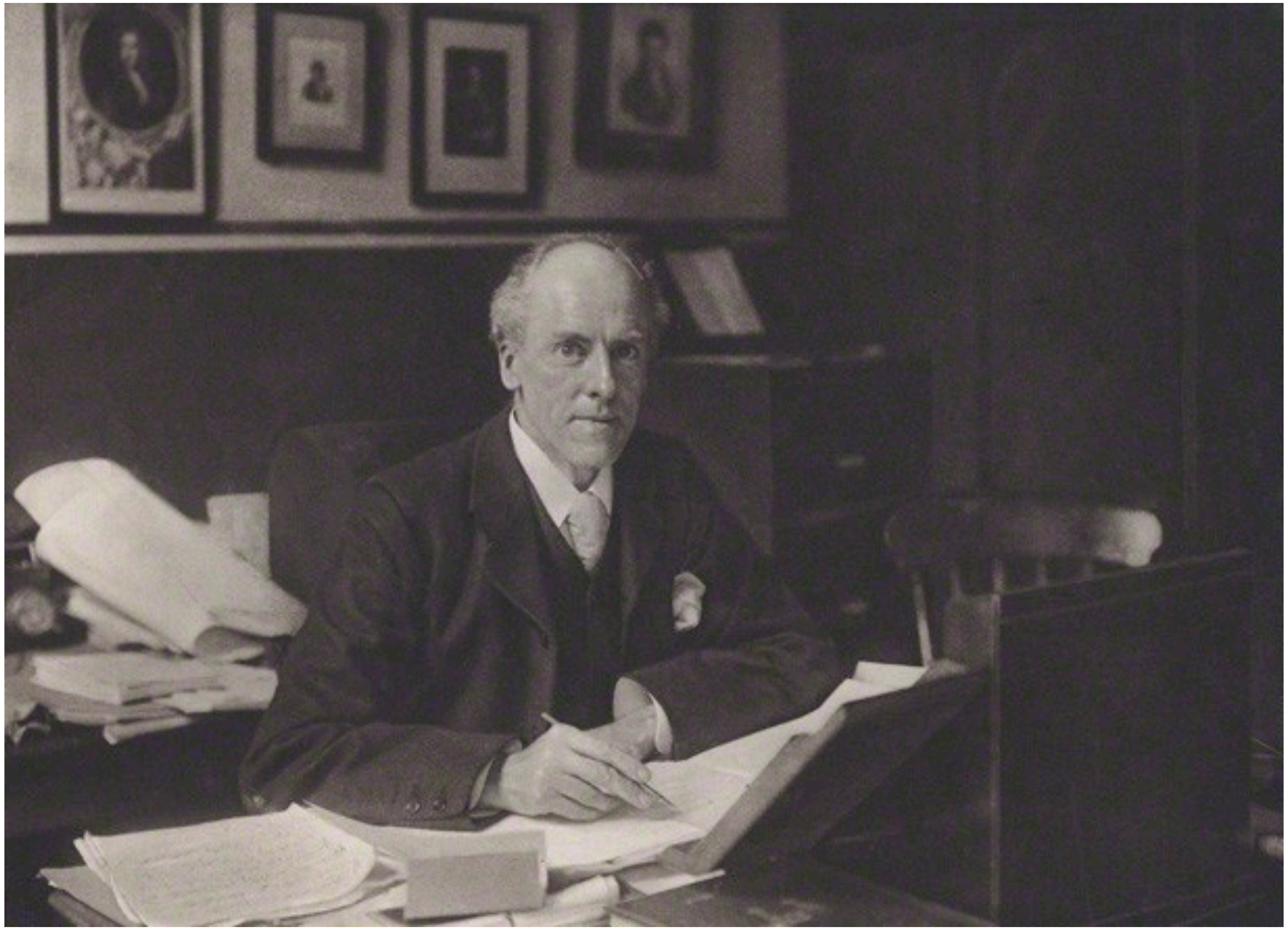
Correlational Analysis
Regression Analysis

Dr. Wong Kai Choi

Department of Psychiatry
Pamela Youde Nethersole Eastern Hospital

26th April 2016

Correlational Analysis



Karl Pearson at work in 1910

Pearson's correlation

- Also called "Pearson product-moment correlation coefficient"
- It investigate the strength of a linear relationship between two continuous variables
- It is used when neither variable can be assumed to predict the other
- It gives an estimate, the correlation coefficient and a p value
- A confidence interval can be calculated

Pearson's correlation

- Assumption
  - The relationship is linear
  - Normal distribution
    - For significant test – at least one variable to be normally disturbed
    - For confidence intervals – both variables should be normally distributed
  - A random sample within the range of interest

$$r = \frac{S_{XY}}{\sqrt{S_{XX}}\sqrt{S_{YY}}}$$

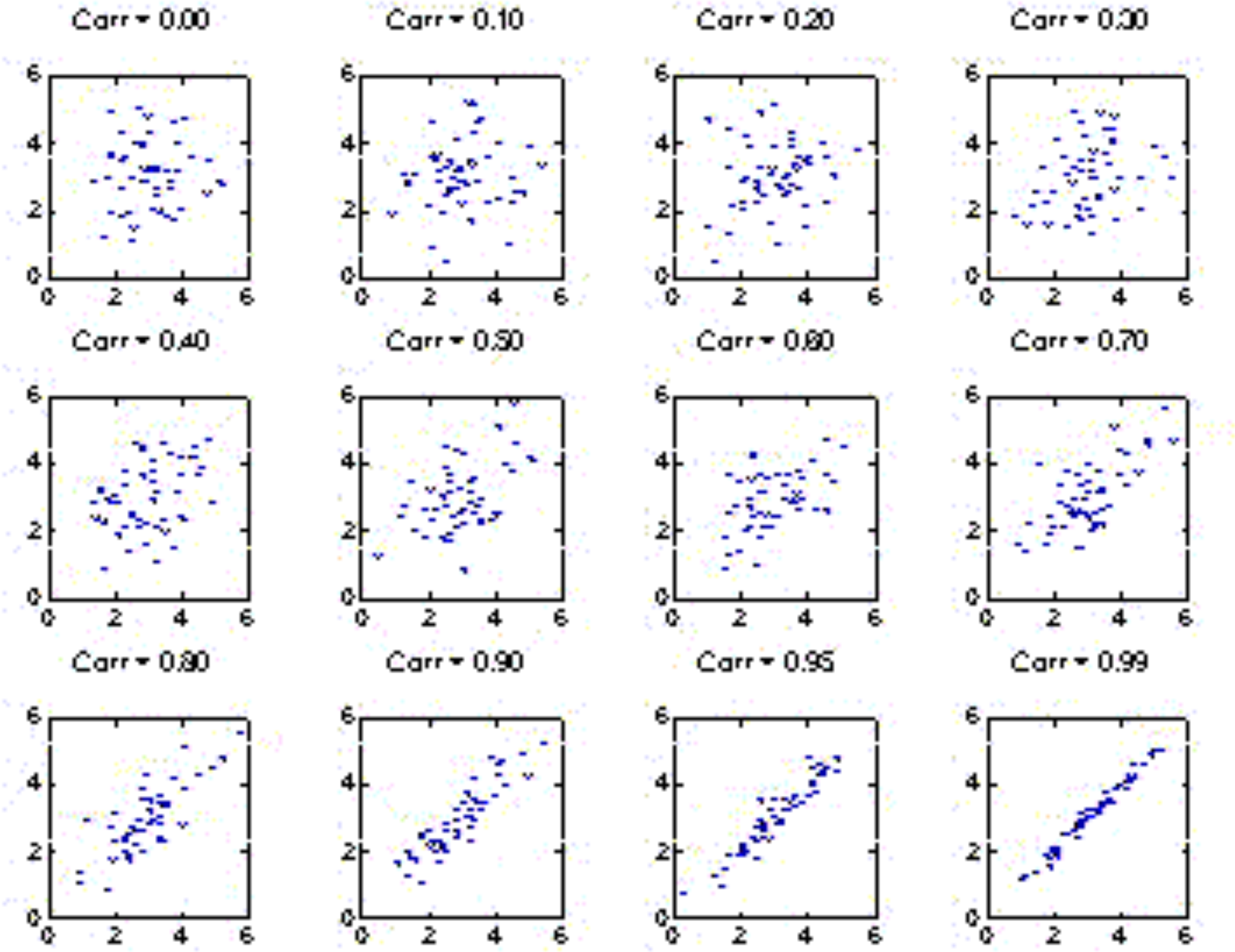
Where
$$S_{XX} = \sum x^2 - \frac{(\sum x)^2}{n}$$

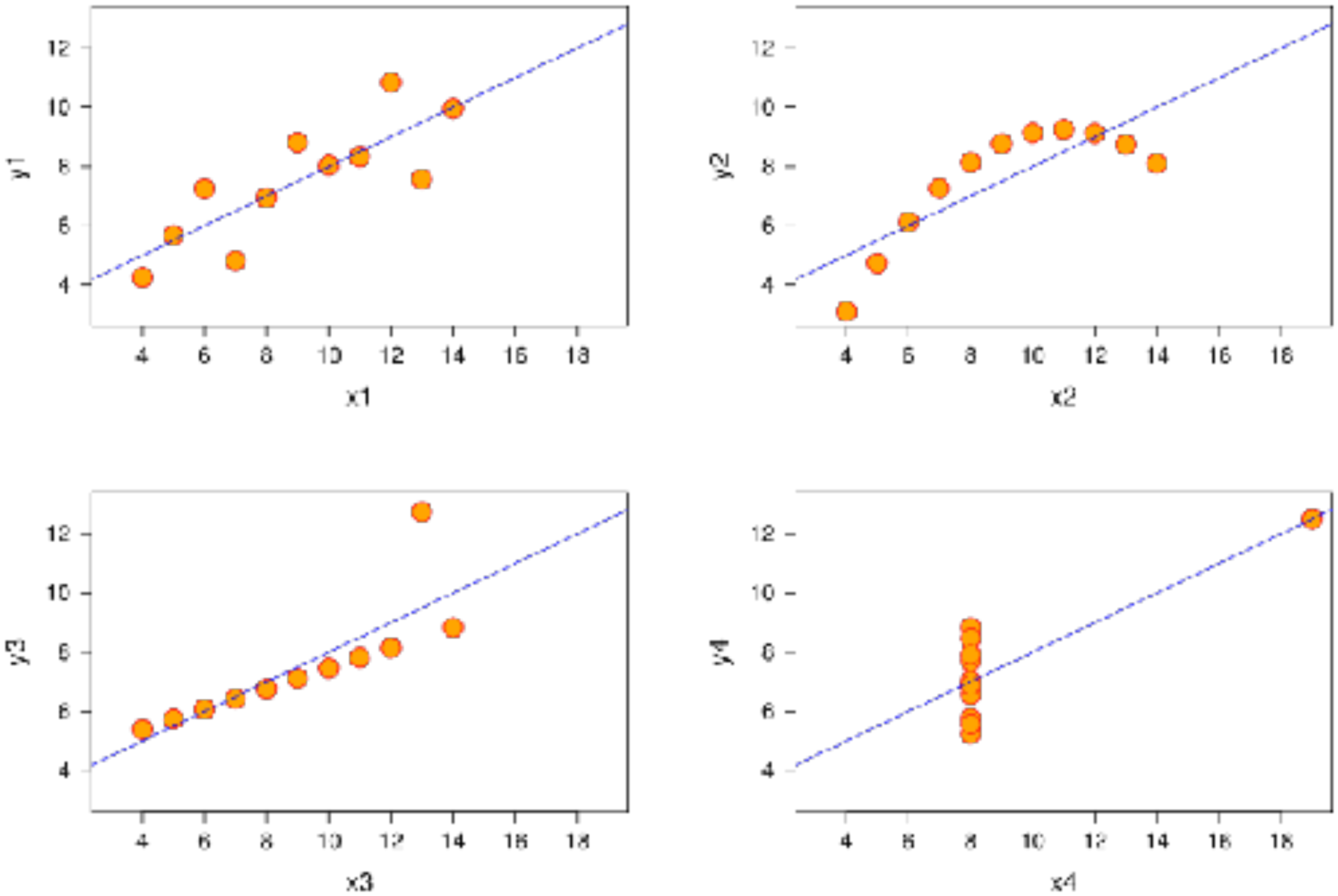
$$S_{YY} = \sum y^2 - \frac{(\sum y)^2}{n}$$

$$S_{XY} = \sum xy - \frac{(\sum x)(\sum y)}{n}$$

Interpretation of *r*

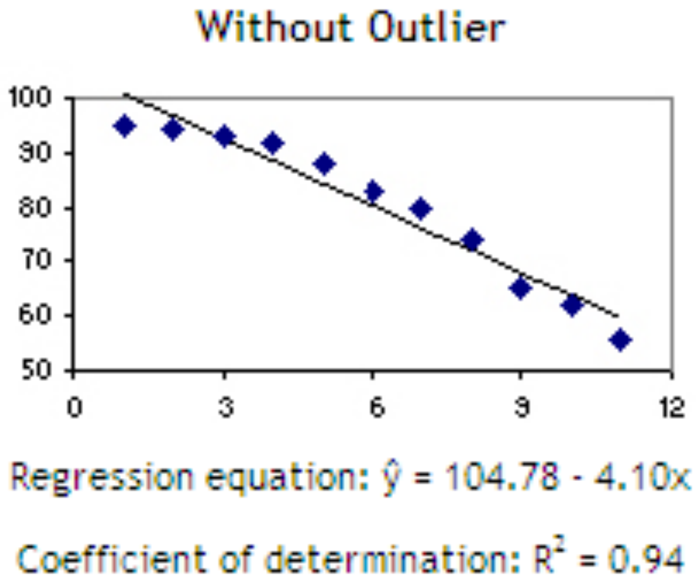
- *r* tells us how close is the linear relationship between two variables
- It lies between +1 and -1
- Negative (positive) values indicate negative (positive) linear relationship
- $r = 0$ indicate that is no linear relationship
- The closer the value +1 or -1, the stronger relationship between two variables

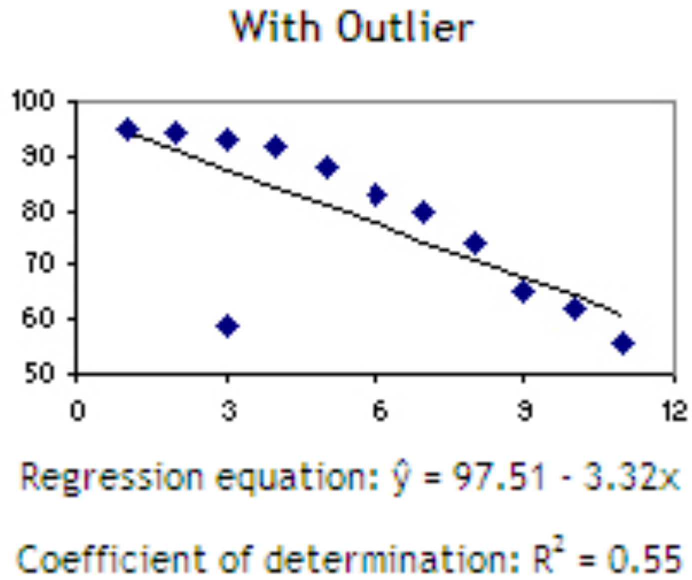

Corr = 0.00
Corr = 0.10
Corr = 0.20
Corr = 0.30
Corr = 0.40
Corr = 0.50
Corr = 0.60
Corr = 0.70
Corr = 0.80
Corr = 0.90
Corr = 0.95
Corr = 0.99


12
10
8
6
4
y1
4
6
8
10
12
14
16
18
x1
y2
x2
y3
x3
y4
x4

Outlier

- If outlier is removed, r is closer to +1 or -1

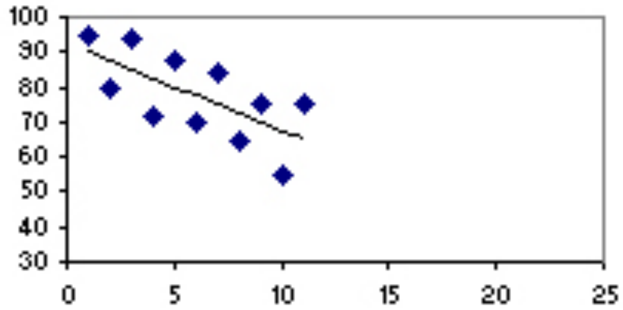




Influential point

- If influential point is removed, *r* is closer to 0

Without Outlier

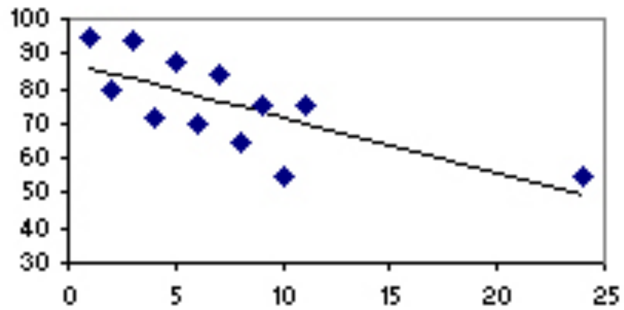


Regression equation: $\hat{y}$ = 92.54 - 2.5x

Slope: b_0 = -2.5

Coefficient of determination: R^2 = 0.46

With Outlier

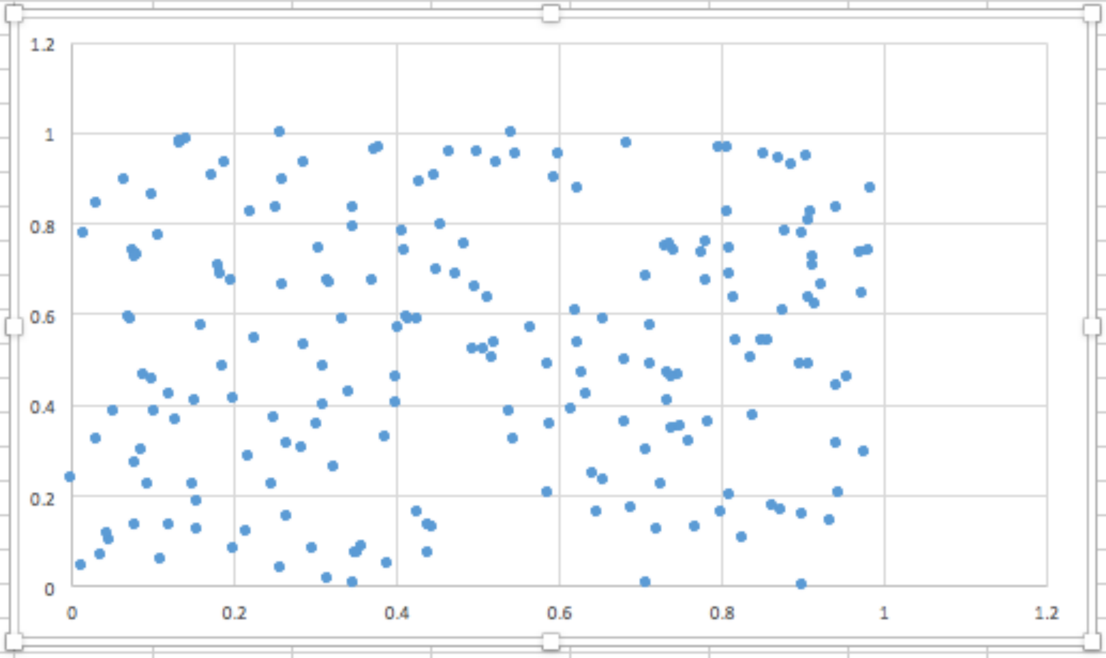


Regression equation: $\hat{y}$ = 87.59 - 1.6x

Slope: b_0 = -1.6

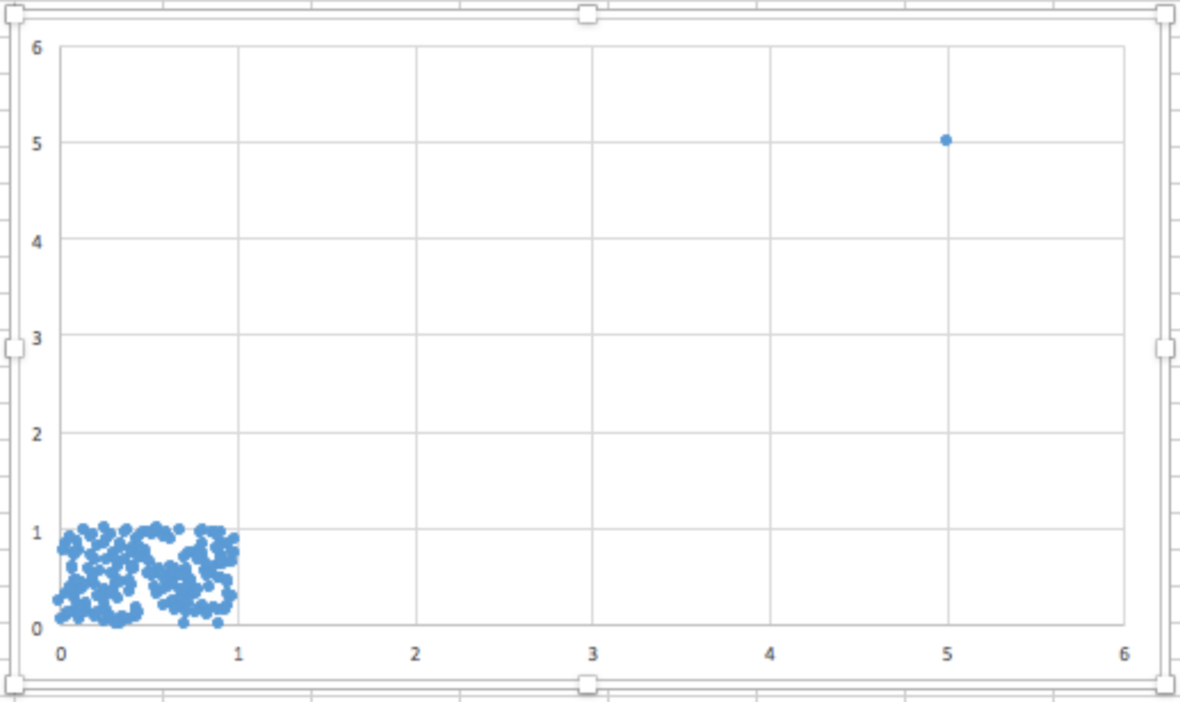
Coefficient of determination: R^2 = 0.52

0.78369356	0.36104484
0.58695968	0.48633706
0.28589949	0.53184907
0.01497872	0.77511621
0.19987698	0.08037643
0.37244902	0.96122321
0.21565872	0.12070993
0.25978464	0.89339156
0.31418253	0.01416516
0.31903115	0.66590474
0.89944036	0.77819152
0.09951408	0.45669738
0.43851104	0.13401461
0.52043599	0.53641071
0.06467871	0.89599262
0.28486375	0.30544432
0.45106513	0.69811702
0.12045844	0.13545435
0.64699505	0.16076445
0.45435319	0.79467602
0.7203749	0.12570875
0.93571737	0.14113002



0.10509867

0.01497872	0.77511621
0.19987698	0.08037643
0.37244902	0.96122321
0.21565872	0.12070993
0.25978464	0.89339156
0.31418253	0.01416516
0.31903115	0.66590474
0.89944036	0.77819152
0.09951408	0.45669738
0.43851104	0.13401461
0.52043599	0.53641071
0.06467871	0.89599262
0.28486375	0.30544432
0.45106513	0.69811702
0.12045844	0.13545435
0.64699505	0.16076445
0.45435319	0.79467602
0.7203749	0.12570875
0.93571737	0.14113002
5	5



0.5949429

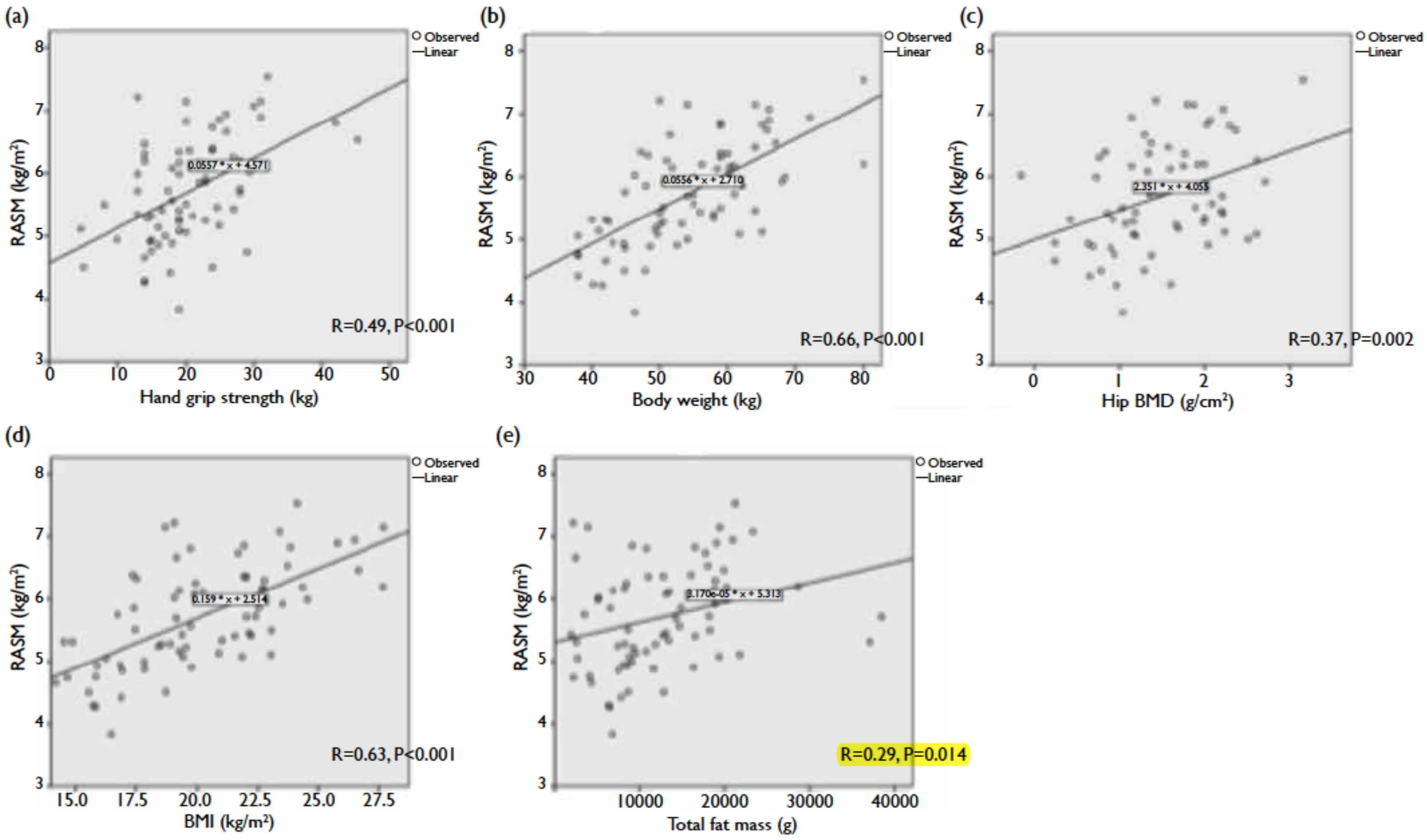


FIG 3. Correlation of relative appendicular skeletal muscle mass index (RASM) and (a) hand grip strength, (b) body weight, (c) hip body mass density (BMD), (d) body mass index (BMI), and (e) total fat mass in male patients

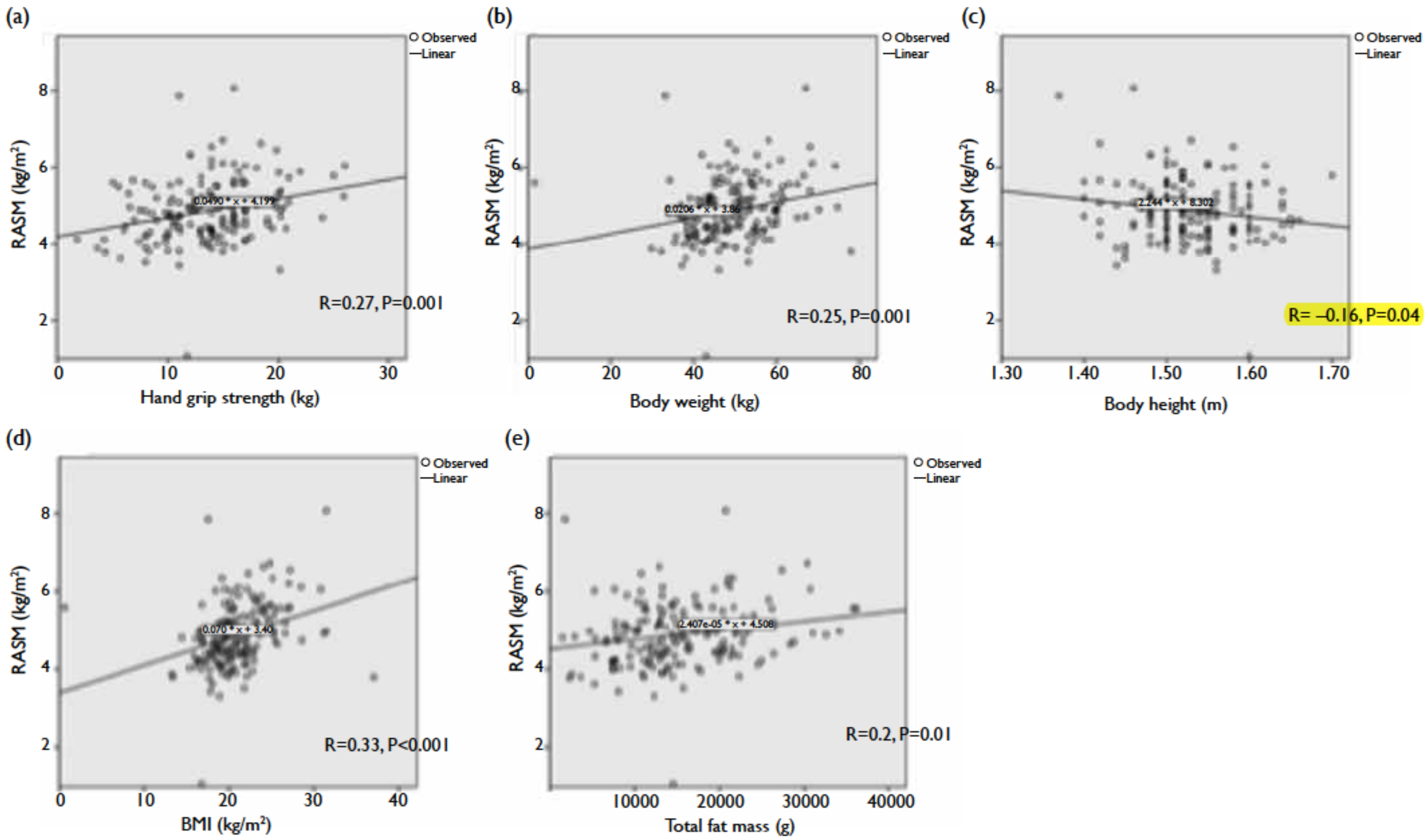


FIG 4. Correlation of relative appendicular skeletal muscle mass index (RASM) and (a) hand grip strength, (b) body weight, (c) body height, (d) body mass index (BMI), and (e) total fat mass in female patients

Test and estimate of *r*

- A significant test can be done with null hypothesis that $r = 0$
- A confidence interval of r can be calculated
- Statistical significance of r directly related to sample size
  - If sample size is large, it may be statistically significant even the relationship is weak

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

r (or n) increases, t increases, then p decreases

Values of r for the .05 and .01 Levels of Significance

$df(N-2)$	.05	.01	$df(N-2)$	.05	.01
1	.997	1.000	31	.344	.442
2	.950	.990	32	.339	.436
3	.878	.959	33	.334	.430
4	.812	.917	34	.329	.424
5	.755	.875	35	.325	.418
6	.707	.834	36	.320	.413
7	.666	.798	37	.316	.408
8	.632	.765	38	.312	.403
9	.602	.735	39	.308	.398
10	.576	.708	40	.304	.393
11	.553	.684	41	.301	.389
12	.533	.661	42	.297	.384
13	.514	.641	43	.294	.380
14	.497	.623	44	.291	.376
15	.482	.606	45	.288	.372
16	.468	.590	46	.285	.368
17	.456	.575	47	.282	.365
18	.444	.562	48	.279	.361
19	.433	.549	49	.276	.358
20	.423	.537	50	.273	.354
21	.413	.526	60	.250	.325
22	.404	.515	70	.232	.302
23	.396	.505	80	.217	.283
24	.388	.496	90	.205	.267
25	.381	.487	100	.195	.254
26	.374	.479	200	.138	.181
27	.367	.471	300	.113	.148
28	.361	.463	400	.098	.128
29	.355	.456	500	.088	.115
30	.349	.449	1000	.062	.081

The average first consultation time spent was weakly correlated with the correct asthma therapy prescribed to the patients (r=0.11, P=0.04) but there was no correlation with the correct classification of the level of asthma control. The time spent in subsequent consultations did not correlate with the correct treatment (r=0.06, P=0.25) or classification of the level of asthma control by the respondents (r=–0.002, P=0.97). The frequency of prescribing asthma medications by the doctors correlated weakly with the performance of the doctors with the correct classification of asthma severity (r=0.15, P=0.003), but not the correct prescription of asthma medications (r=0.06, P=0.22).

Variable	Pearson's correlation (r)	Significance
Age	0.12	$p < 0.05$
Depression	0.25	$p < 0.01$
Hopelessness	0.21	$p < 0.01$
Risk rescue score	0.13	$p < 0.05$

There was a clinically significant correlation between suicidal intent and age, hopelessness, depression, and lethality of the attempt (Table 2).

Spearman's correlation

- It is used when none of the 2 variables follows a normal distribution – it is assumption for Pearson's correlation.
- Null hypothesis
  - There is no tendency for one variable either to increase or to decrease as the other increases
- Assumption
  - The variables can be ranked
  - Monotonic relationship between 2 variables

Spearman's correlation

- This is calculated using same formula as for Pearson's correlation but uses the ranks of the data rather than the data values themselves
- It gives a values between -1 and +1
- p value can be obtained from statistical program

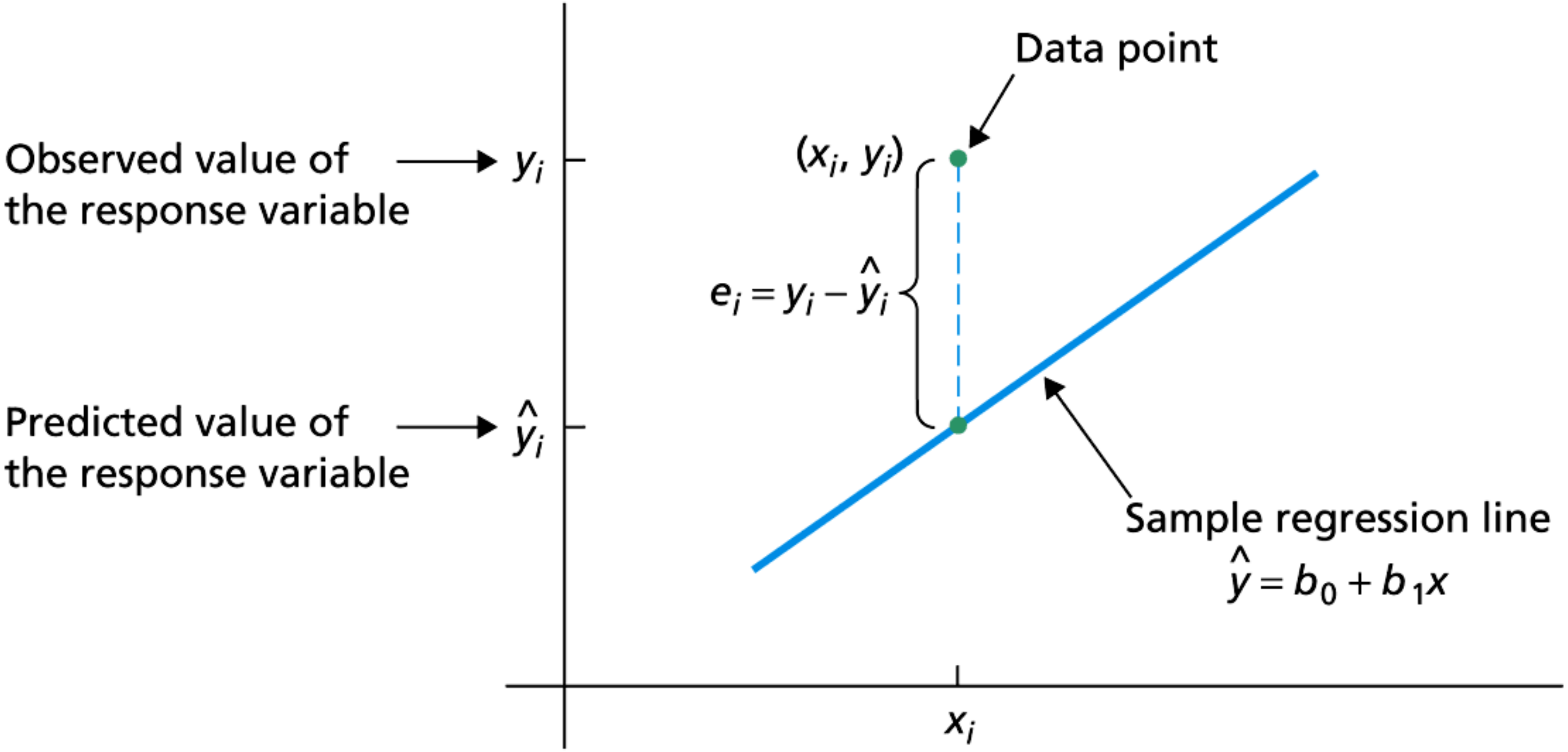
Simple linear regression

Simple Linear Regression

- Estimate the nature of linear relationship between two continuous variables
  - Dependent (response) variable
  - Independent (explanatory) variables
- The calculation based on least squared method – It minimize the sum of the squares of these residual (= observed value – fitted values)

Adrien-Marie Legendre
(1752-1833)

Carl Friedrich Gauss
(1777-1855)


Data point
Observed value of the response variable
y_i
(x_i, y_i)
$e_i = y_i - \hat{y}_i$
Predicted value of the response variable
$\hat{y}_i$
Sample regression line
$\hat{y} = b_0 + b_1 x$
x_i

Simple Linear Regression

- Slope or regression coefficient is given by
  - $y = a + bx$
  - $b = \frac{S_{XY}}{S_{XX}} = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$
- The line goes through the mean point: (x, y)
- Therefore the intercept is given by: $a = \bar{y} - b\bar{x}$

Interpretation of the equation

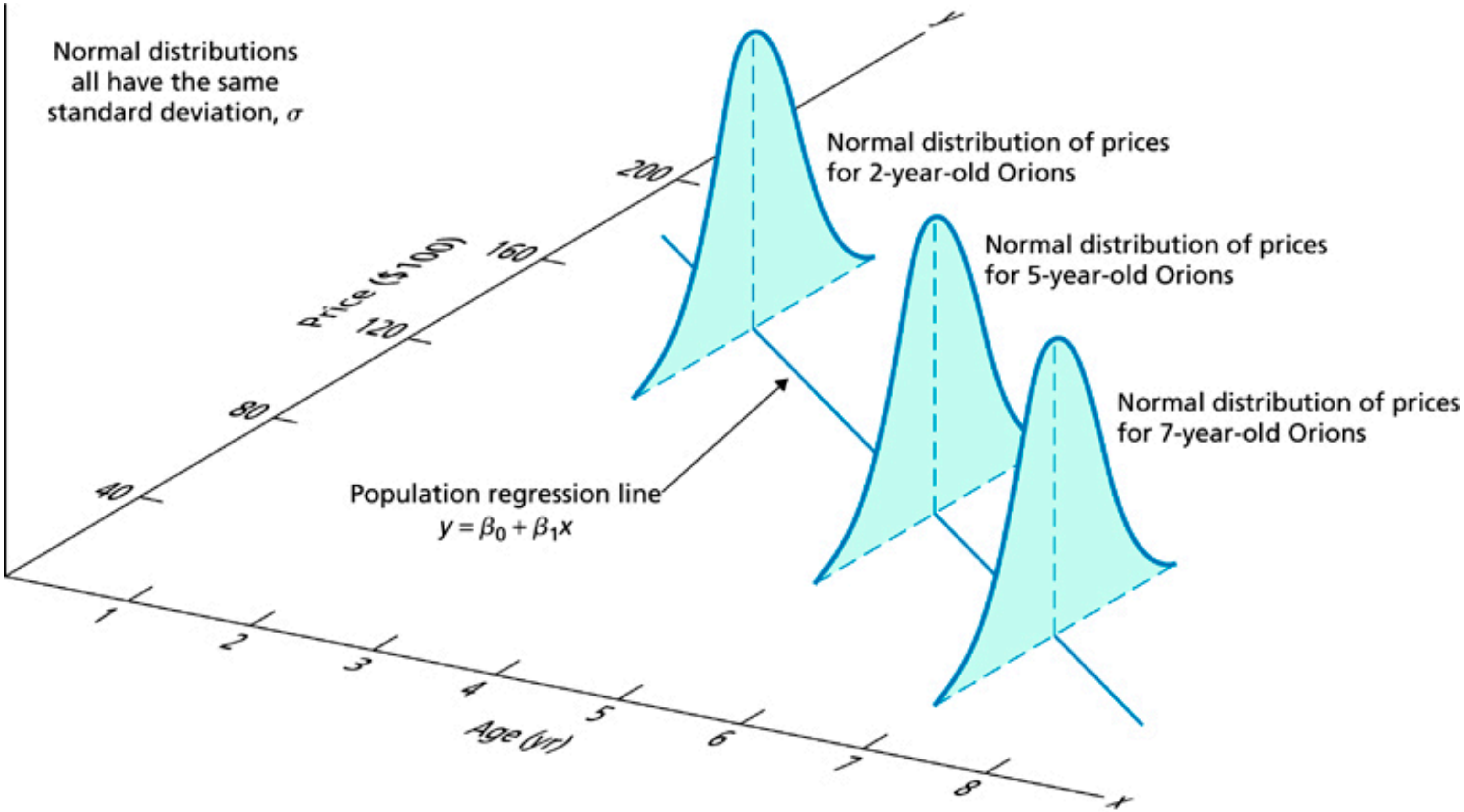
- Regression coefficient gives the change in the outcome (y) for a unit change in the predictor variable (x)
- The intercept gives the value of y when x is 0
- The line gives the mean or expected value of y for each value of x

Simple Linear Regression

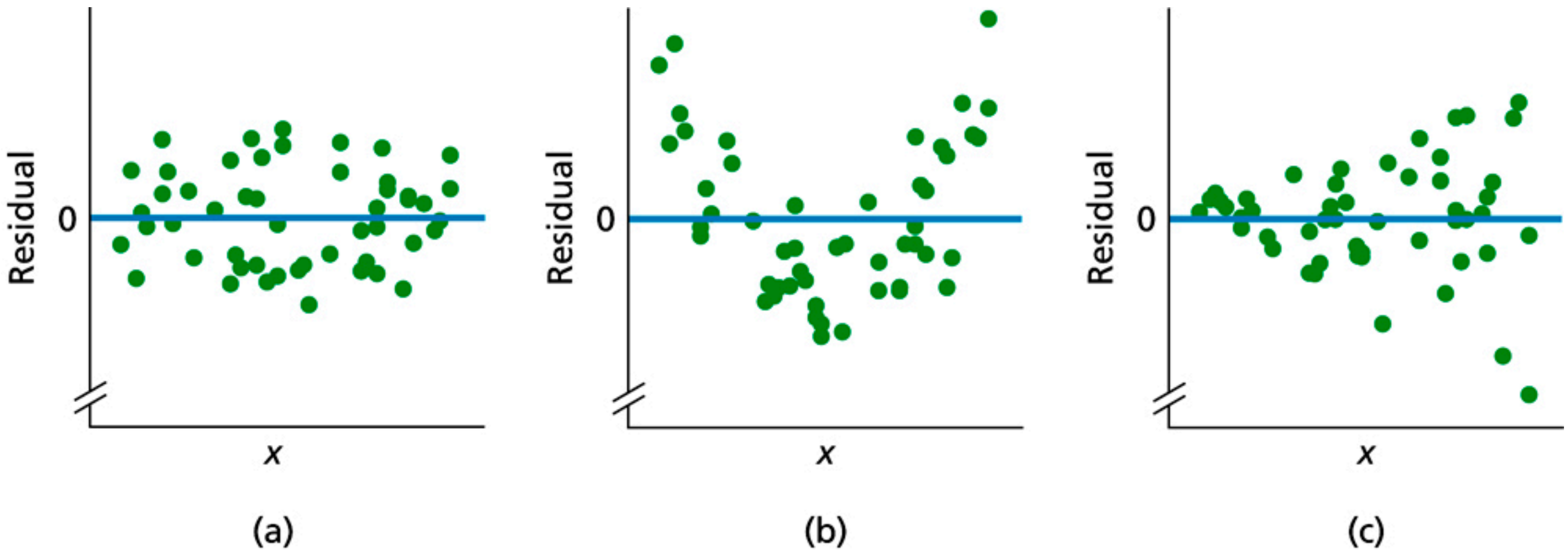
- If there is no relationship between x and y, the true regression coefficient b will be 0
- Can be tested using a form of t test
- The regression coefficient b is a useful summary to show how the two variables related
- 95% confidence interval can be calculated for b
- The equation of the line can be used for prediction

Simple Linear Regression

- Assumption
  - The relationship is linear
  - The distribution of the residuals is normal
  - The variance of the residual ~~(outcome)~~ of y is constant over x
- Predication
  - Within sample prediction
  - Prediction outside the sample



Normal distributions
all have the same
standard deviation, σ
Normal distribution of prices
for 2-year-old Orions
Normal distribution of prices
for 5-year-old Orions
Normal distribution of prices
for 7-year-old Orions
Population regression line
$y = \beta_0 + \beta_1 x$
Price ($100)
40
80
120
160
200
y
1
2
3
4
5
6
7
8
x
Age (yr)


Residual
0
x
(a)
Residual
0
x
(b)
Residual
0
x
(c)

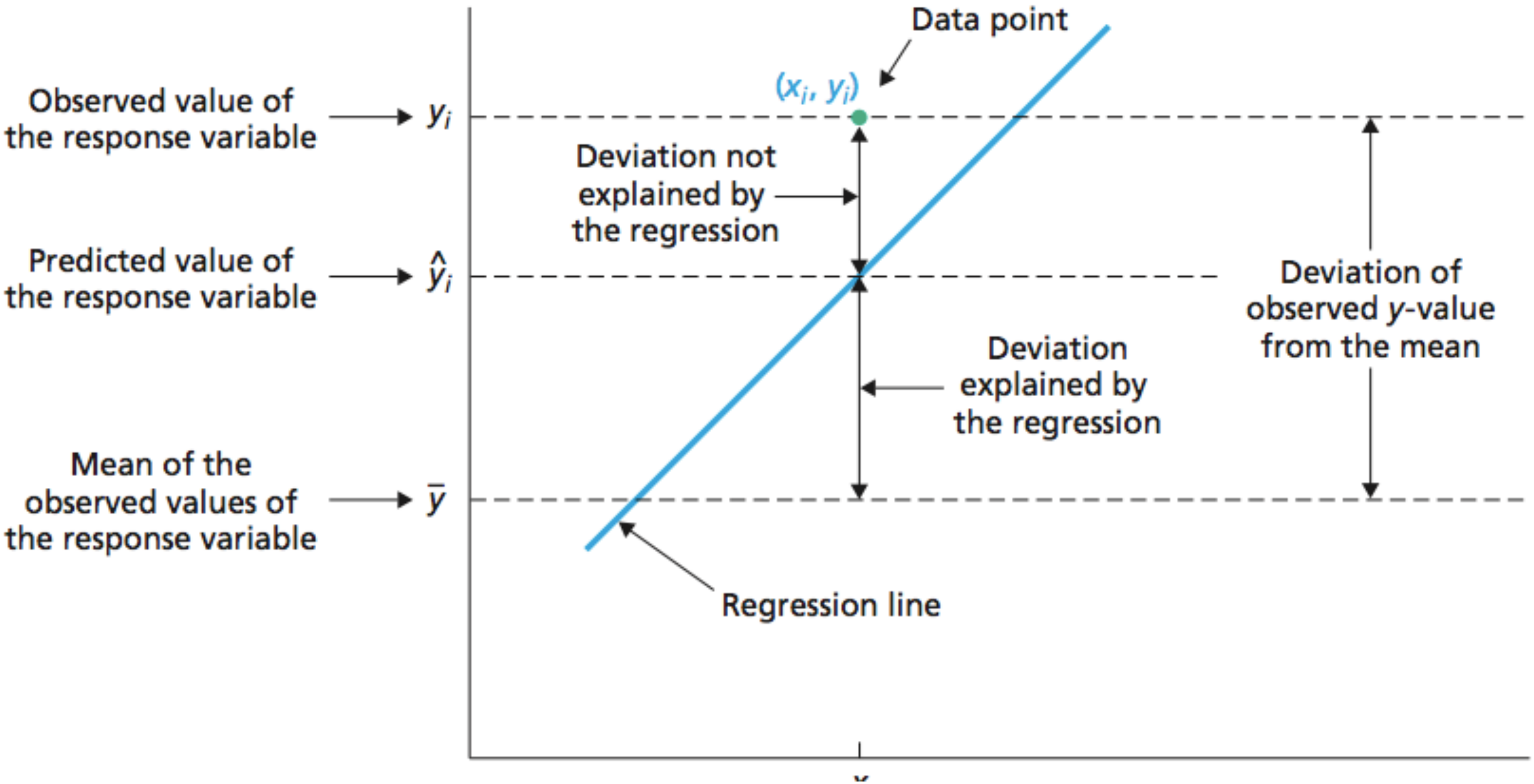
Coefficient of determination, R^2

Sums of Squares in Regression

Total sum of squares, *SST*: The total variation in the observed values of the response variable: $SST = \Sigma(y_i - \bar{y})^2$.

Regression sum of squares, *SSR*: The variation in the observed values of the response variable explained by the regression: $SSR = \Sigma(\hat{y}_i - \bar{y})^2$.

Error sum of squares, *SSE*: The variation in the observed values of the response variable not explained by the regression: $SSE = \Sigma(y_i - \hat{y}_i)^2$.


Data point
(x_i, y_i)
Observed value of the response variable
y_i
Deviation not explained by the regression
Predicted value of the response variable
$\hat{y}_i$
Deviation of observed y-value from the mean
Deviation explained by the regression
Mean of the observed values of the response variable
$\bar{y}$
Regression line

Coefficient of Determination

The **coefficient of determination, r^2,** is the proportion of variation in the observed values of the response variable explained by the regression. Thus,

$$r^2 = \frac{SSR}{SST}.$$

Interpretation: the proportion variance of dependent variable can be explained by the variance of independent variable(s).

Multivariate Analysis

Multivariate?

- When there is more than one predictor variable on predicting the outcome variable
  - E.g., head size (predictor) and recall (criterion)
  - Recall Performance (in words) = 0.32 (head size in cm) + 1.67
  - ...but how about other variables, such as IQ?
- Linear combination of variables

$$Y' = a + b_1(x_1) + b_2(x_2) + \ldots + b_k(x_k) + \epsilon$$

  - k is number of predictor variables

Multivariate Statistics: the basics

- To examine the relationship between variables, e.g., multiple regression
  - The relationship between predictor variables (can be discrete—e.g., gender, or continuous—e.g., income) and a continuous outcome variable
  - Predicting the performance in a certain task (e.g., memory) based on subjects' performance in other tasks (e.g., attention) and group status (e.g., showing hypertension)

Multivariate Statistics: the basics

- Costs
  - Requires larger sample sizes
- Benefits
  - Increases precision using multiple measures of a construct
  - Including more variables to capture genuine relationships in the multidimensional and multicausal real world

Multivariate Statistics: Data screening

- Relationships with other variables
  - **Multicollinearity** – strong inter-correlations among predictor variables ($|r| > 0.8$)
    - Observed in inter-correlation matrices
    - Redundancy among the predictor variables
  - Tolerance $1 - R^2$ ("Collinearity Statistics")
    - Larger = Better
    - If < 0.2 — multicollinear

Multivariate Statistics: Data screening

- How to tackle multicollinearity problems?
  - To combine closely related variables into a composite variable (e.g., averaging)—this requires some theoretical justifications

Multiple Regression

Multiple Regression

- To assess the relationship between <u>one</u> **<u>continuous</u>** outcome variable (Y) and several discrete/or continuous predictor variables (X)
  - The best *prediction* of a outcome variable (Y') from the combination of predictor variables

    $$Y' = a + b_1(x_1) + b_2(x_2) + \ldots + b_k(x_k) \ + \epsilon$$

  - b = unstandardized regression coefficient of each predictor variable
    - Show how much Y' would change with a one unit increase in x
  - Regression analyses—to come up with b values that make Y (data) and Y' (prediction) as close as possible

Multiple Regression

- Standard
  - All variables are simultaneously entered into the regression equation
  - The regression coefficient of **each** predictor is estimated while {holding constant/ partialling out/ controlling for} the other predictor variables

Multiple Regression

- Sequential (block entry or hierarchical)
  - Predictor variables are entered in two or more steps
  - To evaluate the relationships between a **subset** of predictor variables and an outcome variable, after controlling for the effects of other subsets of predictor variables on the outcome variable
    - Based on theories
  - Determine the significance of the change in R^2 at each step to see if each newly added predictor makes a significant **improvement** in the predictive power of the regression equation

Multiple Regression: Example 1

$$JobSatisfaction = 2.411 + (0.512)(SocialSupport) + (-0.109)(Stress) + (0.106)(Income)$$

- If all other variables are held constant:
  - Predicted job satisfaction increases by 0.512 for the increase of an unit of social support
  - Predicted job satisfaction decreases by 0.109 for the increase of an unit of stress
  - Predicted job satisfaction increases by 0.106 for the increase of an unit of income

Multiple Regression: Example 1

- How important are our predictor variables?
  - Regression coefficients and tests of significance (t-test)–If the coefficient is not significantly different from zero, that predictor will serve *no use*
- **Common mistake**: relative sizes of b_i reflect the relative importance of the predictors
  - E.g., Income is more important than Social Support because *b* is larger?

$$JobSatisfaction = 2.411 + (0.512)(SocialSupport) +$$

$$(-0.109)(Stress) + (0.106)(Income)$$

Multiple Regression: Example 1

- When considering **relative** magnitudes of the coefficients, we should take into account the *SD* of the predictor variables
  - Solving the regression using *standardized* variables (*beta*)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.411	.328		7.358	.000
	social support at work	.512	.057	.502	9.019	.000
	degree of stress at work	-.109	.053	-.115	-2.053	.041
	monthly income at work	.106	.050	.102	2.112	.036

a. Dependent Variable: job satisfaction

$$JobSatisfaction_Z = (0.502)(SocialSupport_Z) + (-0.115)(Stress_Z) + (0.102)(Income_Z)$$

Multiple Regression: Example 1

$$Y' = a + b_1(x_1) + b_2(x_2) + \ldots + b_k(x_k)$$

- **Common mistake**: X_1 and Y are strongly correlated → b_1 has to be a significant regression coefficient in multiple regression models
  - A test of a predictor variable is done in the context of all other variables in the regression equation
  - After X_2, X_3, X_4, … and X_k are partialled out (or held constant), X_1 may no longer be useful for predicting the Y
  - b_1 is coefficient for regression of Y on X_1 when we *partial out* the effect of X_2, …, and X_k

Multiple Regression: Points to note

- **Sample size** considerations (rules of thumb)
  - 50 + 8 * (number of predictors), or
  - Number of predictor variables should be smaller than (the number of subjects divided by 10)
- Limitations
  - Predictability does not tell us anything about the direction of **causality** in regression analyses
    - Similar to "Correlation does not mean causation"
  - Regression assumes no measurement error
    - Must ensure **reliability** and validity of the variables

Multiple Regression: Points to note

- Should all *available* variables be entered in the regression models?
  - Adding more predictor variables can change the betas of all other predictors
  - Excludes relevant variables (e.g., intelligence) → estimation of coefficients for the remaining variables may not reflect the whole pictures
  - Includes irrelevant variables (e.g., preference to ice-cream flavor in salary and success in universities) → increase the error variance and weaken the statistical power of the model

Multiple Regression: Points to note

- Assumptions
  - Absence of multicollinearity among predictor variables
  - Normality *(residuals),* linearity, and homoscedasticity *(residuals)*
    - If the underlying function between one or more of the predictors and the criterion is *not* linear, then the betas will be biased and unreliable, so it is important to look at all bivariate plots prior to the analyses

Logistic Regression

Logistic Regression

- Can be used when the outcome **(i.e., Y)** is **dichotomous**
- Examples:
  - Does a sample have early-stage Alzheimer's disease?
  - Does a cancer patient respond to therapy?
  - Does a secondary school student smoke?
- Compare to continuous outcomes
  - Global cognitive function
  - Tumor size
  - Packs/week

Logistic Regression

- To predict the membership in discrete groups (outcome variable, e.g., binary) from several predictor variables (which can be discrete or continuous)
  - E.g., predicting the occurrence of dementia (healthy old adults vs. earliest-stage Alzheimer's diseases) based on subjects' performance in cognitive tests
  - Binary outcome variable is most common, code reference group as 0 and occurrence group as 1 (useful for odd ratio interpretation)

Logistic Regression

- Goal = to find the best **linear** combination of predictors to maximize the likelihood of obtaining the observed outcome frequencies

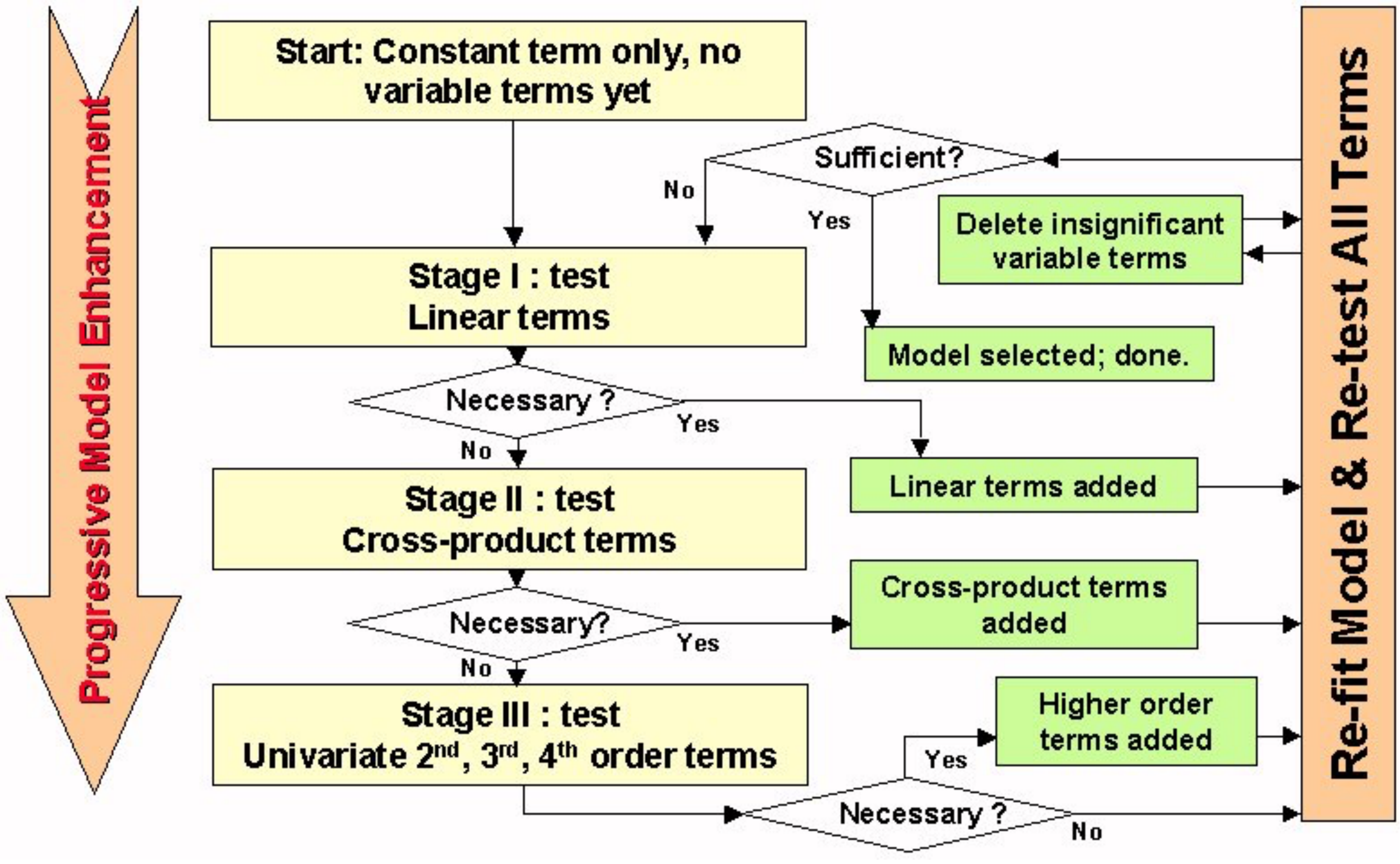
$$ln\left(\frac{p}{1-p}\right) = b_0 + b_1 x$$

  - p is the probability that the event Y occurs, p(Y=1)
  - p/(1-p) is the *odds ratio*
  - ln[p/(1-p)] is the log odds ratio, or *logit*
- Can be standard or sequential

Logistic Regression: Points to note

- Assumptions about variable distributions are not required, but may have more power if there are multivariate normality, and linearity among predictors *(it is also not required in linear multiple regression)*
- Problems may occur if too few cases relative to predictor variables *(sample size > 10 × case number)*
- Absence of **multicollinearity** across predictor variables

Is “automatic” exclusion or inclusion of independent variables by forward / backward / stepwise procedure appropriate?

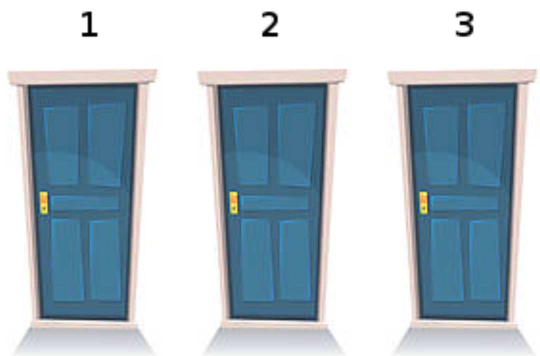

Progressive Model Enhancement
Start: Constant term only, no variable terms yet
Sufficient?
No
Yes
Delete insignificant variable terms
Stage I : test Linear terms
Model selected; done.
Necessary ?
Yes
No
Linear terms added
Stage II : test Cross-product terms
Cross-product terms added
Necessary?
Yes
No
Stage III : test Univariate 2nd, 3rd, 4th order terms
Higher order terms added
Yes
Necessary ?
No
Re-fit Model & Re-test All Terms

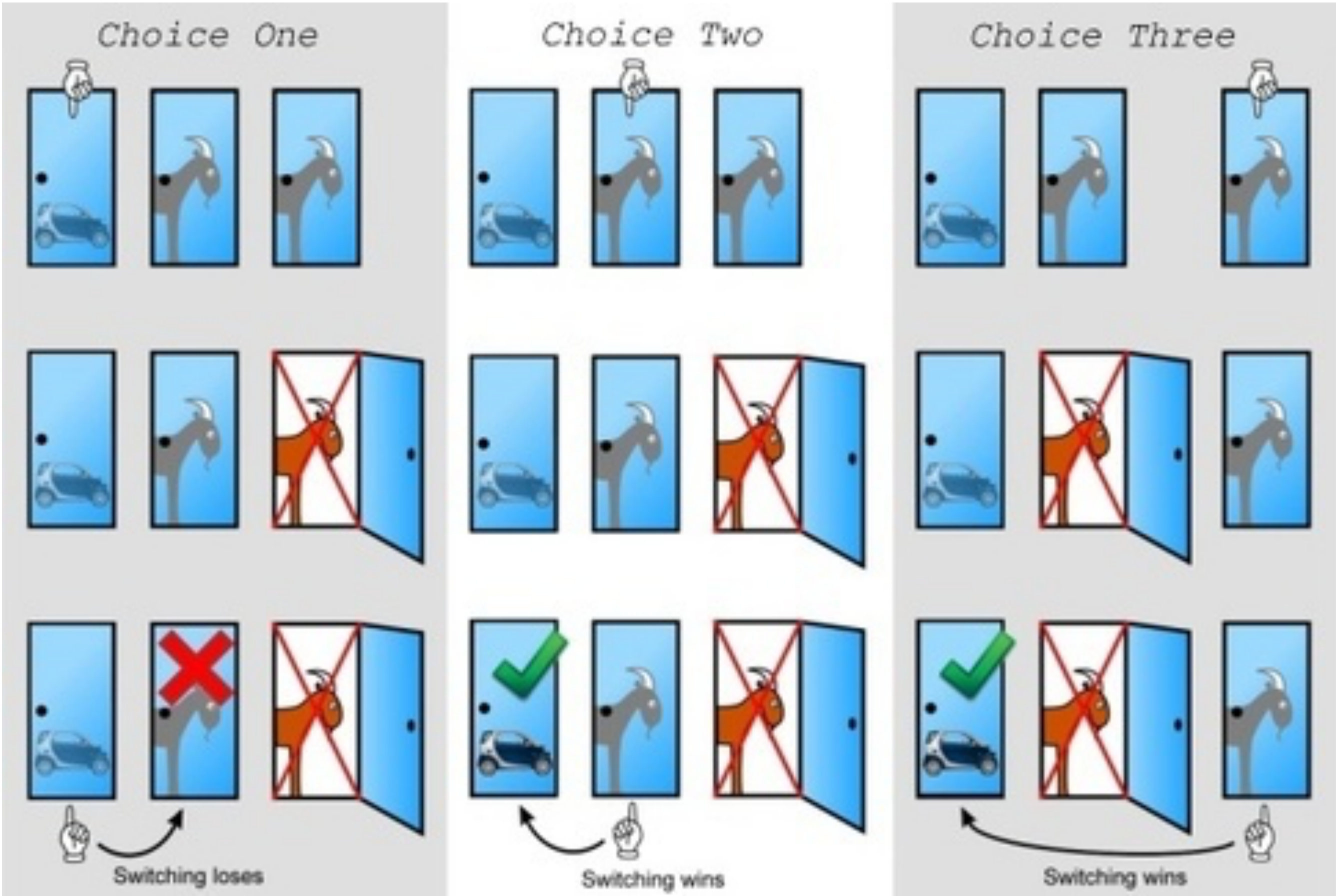
If x is a "significant" modifiable independent variable in a regression model, do the change of x result in the change of independent variable?

If independent variable x_1 and x_2 is a "significant" modifiable variable and x_1 has changed and the dependent variable was improved, will the change of x_2 further improve dependent variable "significantly" as expected in the original regression model.

Suppose you're on a game show, and you're given the choice of three doors. Behind one door is a car, behind the others, goats. You pick a door, say #1, and the host, who knows what's behind the doors, opens another door, say #3, which has a goat. He says to you, "Do you want to pick door #2?" Is it to your advantage to switch your choice of doors?

Monty Hall problem




Choice One
Choice Two
Choice Three
Switching loses
Switching wins
Switching wins

$$\Delta\chi\Delta\rho \geq \frac{\hbar}{2}$$

Uncertainty Principle

Steps for Multiple Regression Analysis

1. Checking collinearity of independent variables;
2. If there is two or more factors are highly correlated, select one to represent the whole group;
3. Use Dummy variables for categorical data;
4. If we cannot put all the independent variables in the regression model, exclude those theoretically irrelevant or those with highest p-value in bivariate analysis;
5. If all relevant variables can put into the regression model, bivariate analysis is not necessary and just select relevant independent variables;
6. Don't put the independent variables that you are not interested in the model;
7. Using Stepwise / Forward / Backward regression should be cautious and result should be interpret carefully;
8. Residual analysis after the model is established

*Table 3. Association between demographic factors with medication adherence among patients with schizophrenia.**

	Simple linear regression crude regression coefficient (95% CI)	p Value	Multiple linear regression adjusted regression coefficient (95% CI)	p Value
Age	-0.03 (-0.07 to 0.003)	0.08		
Gender	-0.80 (-1.43 to -0.18)	0.01		
Education level	-0.19 (-0.76 to 0.38)	0.51		
Marital status	-0.13 (-0.56 to 0.30)	0.54		
Occupation	-0.32 (-0.97 to 0.33)	0.32		
Income	-0.007 (-0.28 to 0.27)	0.95		
Duration of illness	-0.23 (-0.59 to 0.12)	0.19		
Frequency of admission	0.59 (-1.05 to -0.13)	0.01	-0.55 (-0.99 to -0.10)	< 0.05
Type of antipsychotic	0.08 (-0.49 to 0.66)	0.76		

Abbreviation: *CI = confidence interval*.

* *Forward, backward, and stepwise regression methods were applied. Model assumption was fulfilled. There were no interactions among dependent variables. No multicollinearity was detected. Coefficient of determination (r^2) = 0.09.*

The relationship was first examined using simple linear regression (SLR). The analysis then proceeded to multiple linear regression (MLR). All the variables that met the initial screening criteria ($p < 0.25$) were entered into MLR. After controlling for age, gender, and duration of illness, the MLR analysis showed a significant negative linear relationship between the number of admissions and total MARS score. Any admission to a psychiatric ward would reduce the total MARS score by 0.5 (adjusted $b = -0.55$; 95% confidence interval [CI], -0.99 to -0.10; $p < 0.05$; Table 3). Therefore, frequency of psychiatric admission accounted for 9% of the MARS total score variance ($r^2 = 0.09$).

Table 4. Association of symptoms, insight, and social support with medication adherence among patients with schizophrenia.*

Variable	Simple linear regression		Multiple linear regression	
	Crude regression coefficient (95% CI)	**p Value**	**Adjusted regression coefficient (95% CI)**	**p Value**
MSPSS	2.21 (0.89-3.52)	0.85		
BPRS	-0.13 (-0.24 to -0.02)	< 0.05	-0.12 (-0.24 to -0.01)	< 0.05
ITAQ	0.04 (-0.02 to 0.09)	0.19		

Abbreviations: *BPRS = Brief Psychiatric Rating Scale; CI = confidence interval; ITAQ = Insight and Treatment Attitude Questionnaire; MSPSS = Multidimensional Scale of Perceived Social Support.*

* *Forward, backward and stepwise multiple linear regression methods were applied. Model assumption was fulfilled. No multicollinearity was detected. Coefficient of determination (r^2) = 0.08.*

After controlling for insight, the MLR analysis showed a significant negative linear relationship between psychopathology and total MARS score. Increase in the total BPRS by 1 reduced the MARS total score by 0.13 (adjusted b = –0.12; 95% CI, -0.24 to -0.01; $p < 0.05$; Table 4). Therefore, the severity of schizophrenic symptoms accounted for 8% of the MARS total score variance ($r^2 = 0.08$).

We concluded that if adherence could be addressed appropriately, the number of admissions and severity of psychopathology could be improved, leading to better patient outcomes.


Q
&
A

Acknowledgement

- ***Dr. Anita Wong Sze Mui***, Associate Professor, Leader of Mathematics and Statistics Team, School of Science and Technology, The Open University of Hong Kong

- ***Mr. Ken Chan Chi Keung***, Assistant Professor, Lee Shau Kee School of Business and Administration, The Open University of Hong Kong

- ***Professor Tse Chi Shing***, Associate Professor, Department of Educational Psychology, Faculty of Education, The Chinese University of Hong Kong