(7) Linear Regression Analysis

Dr. Wan Nor Arifin

Biostatistics and Research Methodology Unit
Universiti Sains Malaysia
wnarifin@usm.my / wnarifin.github.io

Last update: Jul 16, 2023

Outlines

- Introduction
- Simple Linear Regression
- Multiple Linear Regression

Learning outcomes

- Understand the concept behind simple and multiple linear regressions
- Understand and able to interpret the results of simple and multiple linear regressions

Introduction

Introduction

- Linear regression is a statistical method to model linear relationship between:
  - outcome: a numerical variable
  - predictors / independent variables: numerical, categorical variables
- Common in medical and health sciences
- Associated factors of cholesterol level, fasting glucose, BMI, stress level etc

Introduction

- Model the linear relationship

numerical outcome = numerical predictors + categorical predictors

Simple Linear Regression

Simple Linear Regression

- Linear regression is a statistical method to model linear relationship between:
  - outcome: a numerical variable
  - ONE predictor / independent variable: a numerical / categorical variable

Simple Linear Regression

- Model the linear relationship

$$numerical\ outcome = intercept + coefficient \times predictor$$

Simple Linear Regression

Research objective:

To determine the associated factor of cholesterol level

Research question:

Is this factor associated with cholesterol level?

Example

- Sample size: 200
- Outcome: cholesterol level in mmol/L
- Independent variable: DBP in mm/Hg

Results

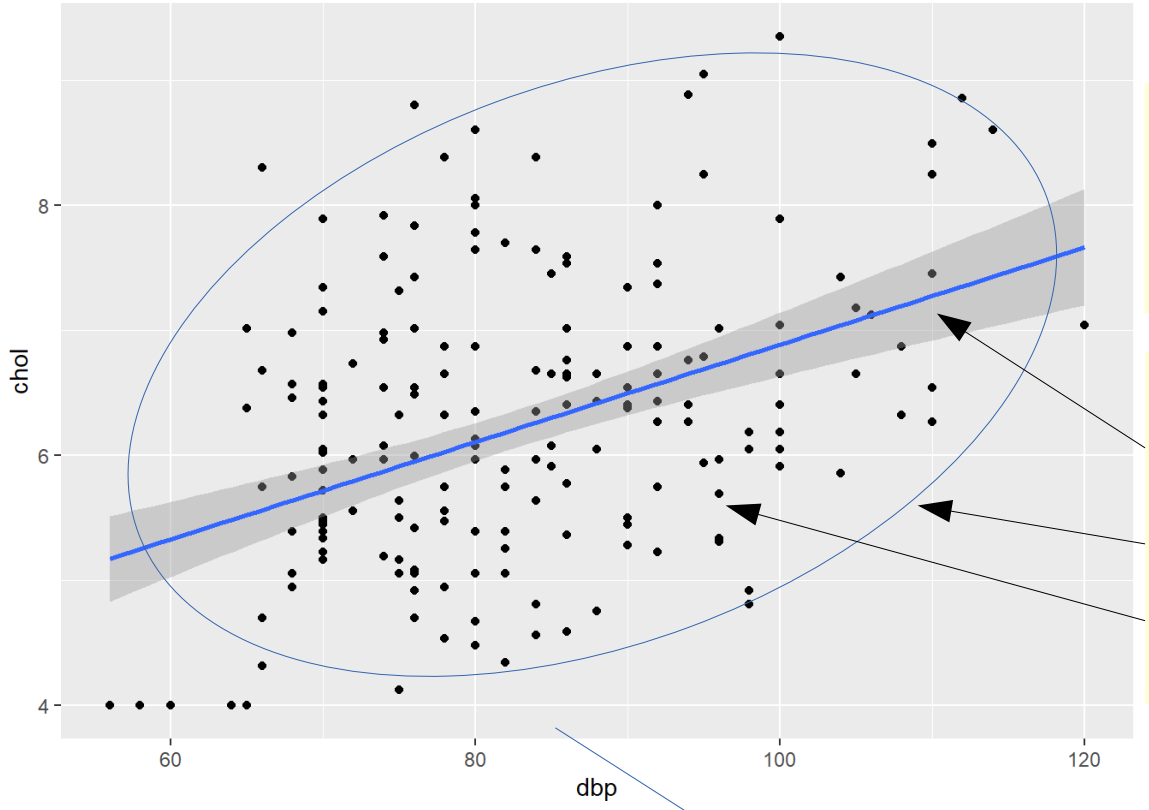
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.995134   0.492092   6.087 5.88e-09 ***
dbp         0.038919   0.005907   6.589 3.92e-10 ***
```

coefficient

P-value

Model Fit

Scatterplot: Cholesterol vs DBP



R^2 = 0.18

0%: the predictor does not explain the outcome at all

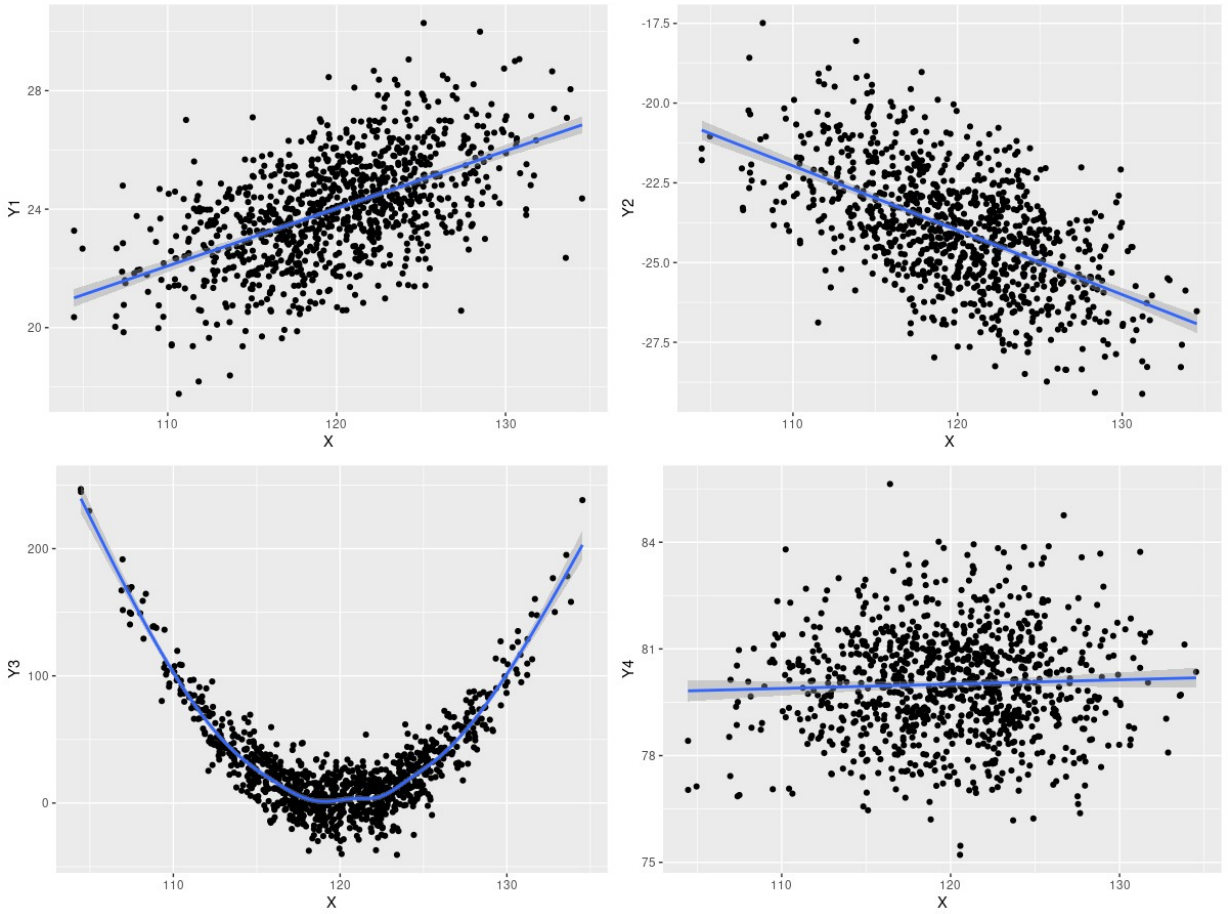
100%: the predictor explains the outcome perfectly

Assumptions:

- linear – straight line trend
- normality – oval shape
- equal variance – equal scatter of dots above and below the line

Positive linear relationship – slope upwards
Negative linear relationship – slope downwards

Scatter Plot Patterns



Guess the patterns

Results

Table X: Factor associated with cholesterol level ($n = 200$)

Factor	b (95% CI)[a]	P-value
DBP (mmHg)	0.04 (0.03, 0.02)	<0.001

DBP = diastolic blood pressure, [a] Simple linear regression ($R^2 = 0.18$)

$$Cholesterol\ Level = 3.00 + 0.04 \times DBP$$

1 mmHg increase in DBP = 0.04 mmol/L increase in Cholesterol level

10 mmHg increase in DBP = 0.4 mmol/L increase in Cholesterol level (10 x 0.04)

Multiple Linear Regression

Multiple Linear Regression

- Linear regression is a statistical method to model linear relationship between:
  - outcome: a numerical variable
  - MORE than one predictors / independent variables: numerical and categorical variables

Multiple Linear Regression

- Model the linear relationship

$$\begin{aligned} \textit{numerical outcome} = \textit{intercept} \\ + \textit{coefficients} \times \textit{numerical predictors} \\ + \textit{coefficients} \times \textit{categorical predictors} \end{aligned}$$

Multiple Linear Regression

Research objective:

To determine the associated factors of cholesterol level

Research question:

Are these factors associated with cholesterol level?

Example

- Sample size: 200
- Outcome: cholesterol level in mmol/L
- Independent variables:
  - DBP in mm/Hg
  - Race: Malay, Chinese, Indian

Results

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.870859   1.245373   3.911 0.000127 ***
dbp          0.029500   0.006203   4.756 3.83e-06 ***
bmi         -0.038530   0.028099  -1.371 0.171871
racechinese  0.356642   0.181757   1.962 0.051164 .
raceindian   0.724716   0.190625   3.802 0.000192 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

coefficients

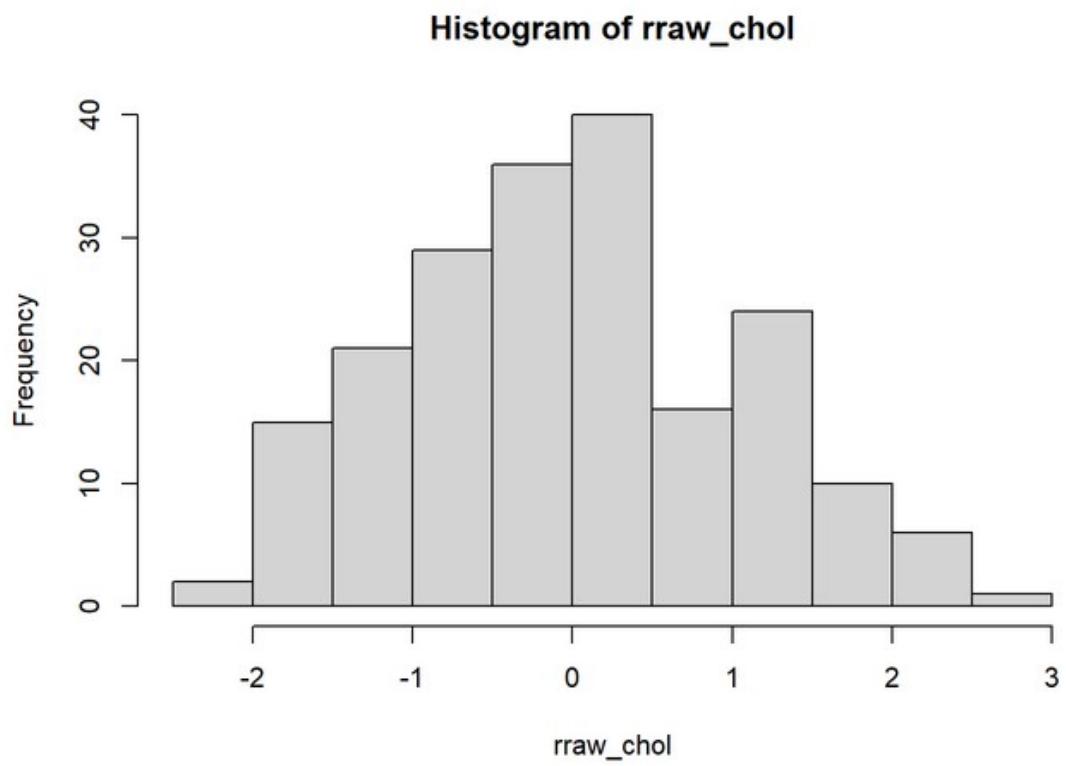
P-values

Race: Malay, Chinese, Indian → Dummy variables / coding

Race = Malay as baseline comparison
RaceChinese: Yes =1 / No = 0
RaceIndian: Yes = 1 / No = 0

Model Fit

Histogram: Raw residuals



*residuals = predicted line values – true observations

<u>Adjusted</u> R^2 = 0.22

0%: the <u>predictors</u> do not explain the outcome at all

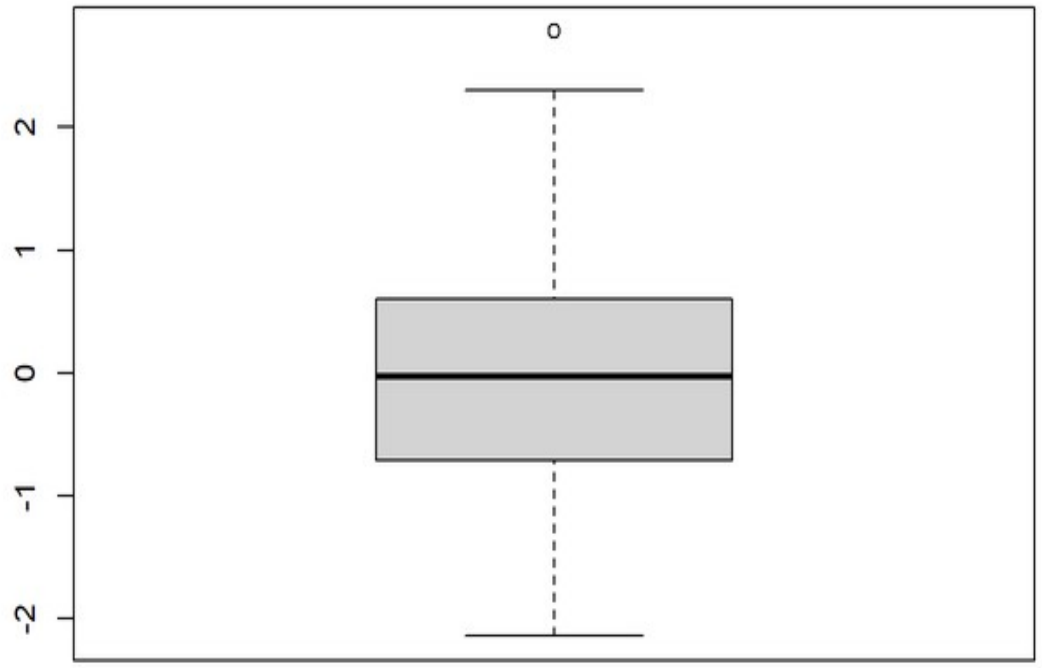
100%: the <u>predictors</u> explain the outcome perfectly

Assumptions:

- Normality of residuals
  - **Histogram**
  - Boxplot
- Linearity
  - Normality
  - Linear pattern
  - Equal variance

Model Fit

Boxplot: Raw residuals



*residuals = predicted line values – true observations

Adjusted R^2 = 0.22

0%: the predictors do not explain the outcome at all

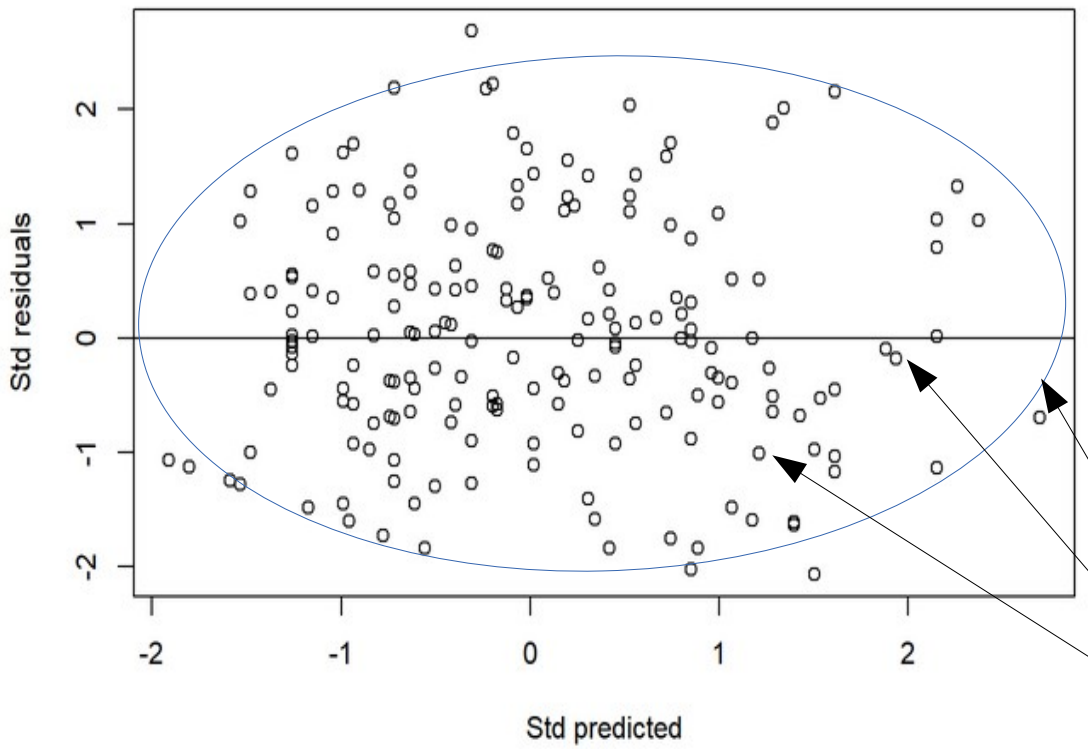
100%: the predictors explain the outcome perfectly

Assumptions:

- Normality of residuals
  - Histogram
  - **Boxplot**
- Linearity
  - Normality
  - Linear pattern
  - Equal variance

Model Fit

Scatterplot: standardized residuals vs standardized predicted values



Adjusted R^2 = 0.22

0%: the predictors do not explain the outcome at all

100%: the predictors explain the outcome perfectly

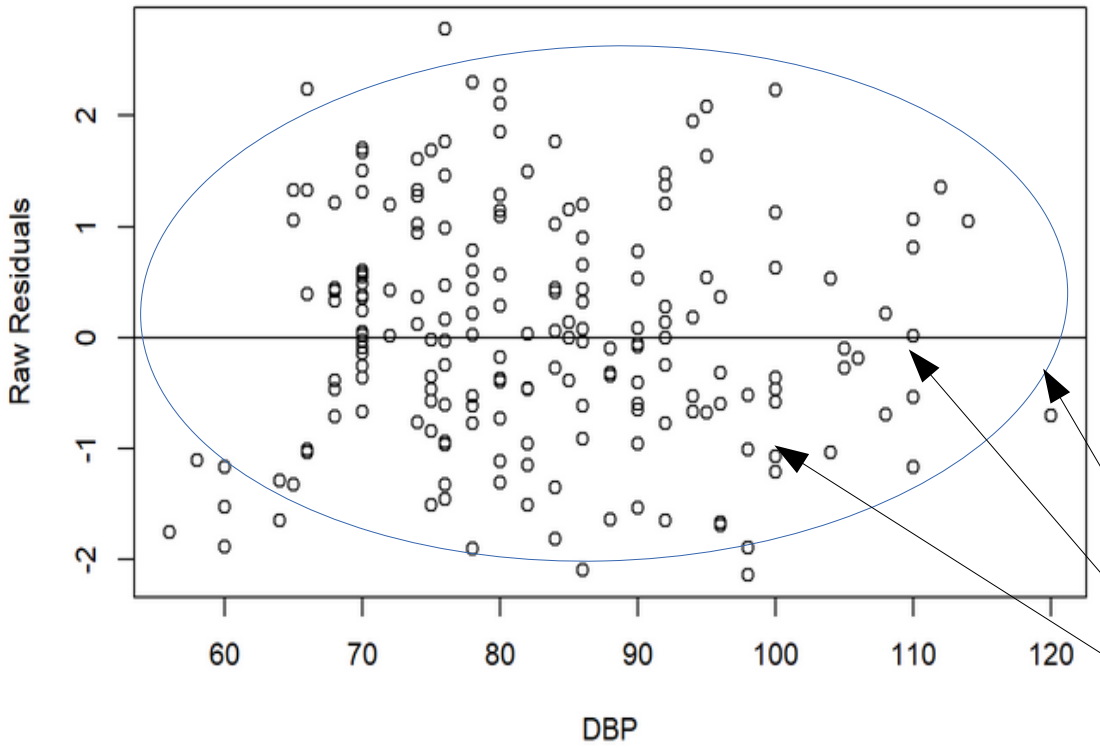
Assumptions:

- Normality of residuals
  - Histogram
  - Boxplot
- **Linearity**
  - Normality
  - Linear pattern
  - Equal variance

*residuals = predicted line values – true observations

Model Fit

Scatterplot: raw residuals vs DBP (numerical predictor)



<u>Adjusted</u> R^2 = 0.22

0%: the <u>predictors</u> do not explain the outcome at all

100%: the <u>predictors</u> explain the outcome perfectly

Assumptions:

- Normality of residuals
  - Histogram
  - Boxplot
- **Linearity**
  - *Normality*
  - *Linear pattern*
  - *Equal variance*

*residuals = predicted line values – true observations

Results

Table X: Factors associated with cholesterol level ($n = 200$)

Factors		Adjusted b (95% CI)[a]	P-value
DBP (mmHg)		0.04 (0.03, 0.02)	<0.001
Race			
	Malay	-	-
	Chinese	0.36 (0.00, 0.72)	0.050
	Indian	0.71 (0.34, 1.1)	<0.001

DBP = diastolic blood pressure, Race = Malay (baseline), Chinese, Indian

[a] Multiple linear regression ($R^2 = 0.22$)

$$Cholesterol = 3.30 + 0.03 \times DBP + 0.36 \times Race\ (Chinese) + 0.71 \times Race\ (Indian)$$

- 1 mmHg increase in DBP = 0.03 mmol/L increase in Cholesterol level, keeping other variables constant
- If Chinese = 0.36 mmol/L higher Cholesterol level as compared to Malay, keeping other variables constant
- If Indian = 0.71 mmol/L higher Cholesterol level as compared to Malay, keeping other variables constant

Quiz

- Describe the purpose of analysis by linear regression
- Compare simple and multiple linear regression analyses

Quiz

Table 4. Factors predicting the ADDQOL-18 average weighted impact score among T2DM patients (n = 180)

Model*	SLR[a]			MLR[b]			
	b value[c]	95% CI	*P*-value	*adj. b*[d] value	95% CI	t-stat	*P*-value
Constant				**−6.82**	**−8.64, −4.99**	**−7.38**	**0.00**
Age (year)	0.05	0.02, 0.09	0.002	0.05	0.02, 0.08	2.90	0.004
Female vs Male	−0.68	−1.34, −0.03	0.041	–	–	–	–
Secondary education versus No/ primary education	−0.45	−1.90, 0.20	0.175	–	–	–	–
Tertiary education versus No/ primary education	−0.01	−0.77, 0.75	0.980				
Staying alone versus Staying with others	1.17	0.10, 2.24	0.032	–	–	–	–
HbA1C (%)	−0.22	−0.36, −0.07	0.004	–	–	–	–
Insulin users versus Non-insulin users	−0.96	−1.61, 0.60	0.004	−0.84	−1.48, 0.20	−2.57	0.011
One complication versus No complication	−0.34	−1.04, 0.37	0.346				
≥2 complications versus No complication	−1.18	−2.18, −0.18	0.021	–	–	–	–
Had hospital admission versus No hospital admission	−1.24	−2.34, −0.14	0.027	–	–	–	–

*Model only included variables with $P < 0.25$
[a]Simple Linear Regression; [b]Multiple Linear regression using Stepwise method
[c]Crude regression coefficient; [d]Adjusted regression coefficient
MLR Final Model: R^2: 0.09; Adjusted R^2: 0.08; Model F statistic: 8.58, $P < 0.001$; The model was reasonably fit; No interaction between independent variables; No multicollinearity problem

Jusoh, Z., Tohid, H., Omar, K., Muhammad, N. A., & Ahmad, S. (2018). Clinical and sociodemographic predictors of the quality of life among patients with type 2 diabetes mellitus on the east coast of Peninsular Malaysia. The Malaysian journal of medical sciences: MJMS, 25(1), 84.

Quiz

Table 6. Predictors of caregivers' satisfaction with the health care management of children with ASD at tertiary care ($n = 227$).

Variables	Mean Satisfaction Score (SD)	Simple Linear Regression		Multiple Linear Regression	
		b [a] (95% CI)	p-value	b [a] (95% CI)	p-value
Caregiver with medical problems					
No	31.25 (0.34)	0	0.013		
Yes	26.45 (3.37)	−4.80 (−8.55, −1.05)		−6.09 (−9.32, −2.85)	<0.001
Presence of sleeping problems					
No	30.71 (0.54)	0			
Yes	31.64 (0.64)	0.92 (−0.79, 2.65)	0.291	1.65 (0.09, 3.11)	0.035
Offered support group post-diagnosis					
No	30.79 (0.40)	0			
Yes	33.94 (2.47)	3.15 (0.08, 6.23)	0.046	3.11 (0.48, 5.73)	0.021
Frequency of occupational therapy					
Once monthly or less	30.87 (0.41)	0			
Twice monthly or more	36.67 (4.04)	5.79 (0.76, 10.83)	0.025	−5.23 (1.00, 9.46)	0.016
Satisfied with frequency of appointments with speech therapist					
No	29.70 (0.61)	0			
Yes	32.62 (0.53)	2.92 (−9.25, 7.84)	<0.001	1.66 (0.20, 3.11)	0.026
Satisfied with frequency of occupational therapy appointments					
No	26.14 (1.89)	0			
Yes	31.52 (0.40)	5.38 (2.65, 8.11)	<0.001	3.82 (1.35, 6.31)	0.003
Satisfied with waiting time					
No	29.14 (0.54)	0			
Yes	32.42 (0.57)	3.28 (1.68, 4.87)	<0.001	2.55 (1.12, 3.98)	<0.001
Satisfied with doctor's knowledge and experience					
No	26.53 (0.80)	0			
Yes	31.98 (0.45)	5.46 (3.43, 7.47)	<0.001	4.39 (2.51, 6.29)	<0.001

[a] Crude regression coefficient; [b] Adjusted regression coefficient. Forward multiple linear regression method applied. Model assumptions were fulfilled. There were 2 interactions among the independent variables. No multicollinearity was detected. Coefficient of determinants, R^2 (adjusted) = 32.92%.

Nik Adib, N. A., Ibrahim, M. I., Ab Rahman, A., Bakar, R. S., Yahaya, N. A., Hussin, S., & Wan Mansor, W. N. A. (2019). Predictors of caregivers' satisfaction with the management of children with autism spectrum disorder: A study at multiple levels of health care. International journal of environmental research and public health, 16(10), 1684.

Thank You