The first three steps in a linear regression analysis with examples in IBM SPSS.

Steve Simon

P.Mean Consulting

www.pmean.com

2. Why do I offer this webinar for free?

I offer free statistics webinars partly for fun and partly to build up goodwill for my consulting business,

- www.pmean.com/consult.html.

Also see my Facebook and LinkedIn pages

- www.facebook.com/pmean
- www.linkedin.com/in/pmean

I provide a free newsletter about Statistics, The Monthly Mean.

- www.pmean.com/news
- www.facebook.com/group.php?gid=302778306676

3. Abstract

- Abstract: This class will give you a general introduction in how to use SPSS software to compute linear regression models. Linear regression models provide a good way to examine how various factors influence a continuous outcome measure. There are three steps in a typical linear regression analysis: fit a crude model, fit an adjusted model, and check your assumptions These steps may not be appropriate for every linear regression analysis, but they do serve as a general guideline.

4. Objectives

In this class you will learn how to:

- interpret the slope and intercept in a linear regression model;
- compute a simple linear regression model; and
- make statistical adjustments for covariates.

5. Sources

Much of the material for this webinar comes from:

- Stats #03: Using SPSS to Develop a Linear Regression Model
  - www.childrens-mercy.org/stats/training/hand03.asp
- Stats #25: What Do All These Numbers Mean? Regression Coefficients.
  - www.childrens-mercy.org/stats/training/hand25.asp

6. Pop quiz #1

In a linear regression model, the slope represents

1. The estimated average change in your outcome variable when the predictor variable increases by one unit
2. The estimated average for your outcome variable in the control group
3. The estimated average for your outcome variable in the treatment group
4. The estimated average for your outcome variable when the predictor variable is zero.
5. The estimated average value for your predictor variable
6. Don’t know/not sure

7. Pop quiz #2

The linear regression model can accommodate all the following settings, except:

1. A categorical outcome variable
2. A categorical predictor variable
3. A continuous outcome variable
4. A continuous predictor variable
5. Multiple predictor variables.
6. Don’t know/not sure

8. Definitions

Categorical data is data that consist of only small number of values, each corresponding to a specific category value or label. Ask yourself whether you can state out loud all the possible values of your data without taking a breath. If you can, you have a pretty good indication that your data are categorical.

Continuous data is data that consist of a large number of values, with no particular category label attached to any particular data value. Ask yourself if your data can conceptually take on any value inside some interval. If it can, you have a good indication that your data are continuous.

9. Definitions

In a research study, the dependent variable is the variable that you believe might be influenced or modified by some treatment or exposure. It may also represent the variable you are trying to predict. Sometimes the dependent variable is called the outcome variable. This definition depends on the context of the study. In a study of prenatal care, the birthweight is an outcome or dependent variable, but in neonatology, it is more likely to be an independent variable.

10. Definitions

In a research study, an indepdent variable is a variable that you believe might influence your outcome measure. This might be a variable that you control, like a treatment, or a variable not under your control, like an exposure. It also might represent a demographic factor like age or gender.

11. Definitions

Example: A recently published research study examined the relationship of dietary fat consumption and the development of ischemic stroke in a cohort of 832 men who were free of cardiovascular disease at baseline (1966-1969) and who were followed for a twenty year period. In this study, the dependent variable was

* presence/absence of ischemic stroke

and the independent variables were:

* percentage of total fat in the diet,
* percentage of saturated fat, and
* the percentage of monounsaturated fat.

12. Interpreting regression coefficients

The linear regression model is useful when the outcome variable is continuous. An alternative, logistic regression, should be used if the outcome variable is categorical. The linear regression model can accommodate either categorical or continuous predictor variables. It can also handle multiple predictor variables.

13. Interpreting regression coefficients

When I ask most people to remember their high school algebra class, I get a mixture of reactions. Most recoil in horror. About one in every four people say they liked that class. Personally, I thought that algebra, and all the other math classes I took were great because they didn't require writing a term paper.

14. Interpreting regression coefficients

One formula in algebra that most people can recall is the formula for a straight line. Actually, there are several different formulas, but the one that most people cite is

- $Y = m X + b$

where m represents the slope, and b represents the y-intercept (we'll call it just the intercept here). They can also sometimes remember the formula for the slope:

- $m = \Delta y / \Delta x$

In English, we would say that this is the change in y divided by the change in x.

15. Interpreting regression coefficients

In linear regression, we use a straight linear to estimate a trend in data. We can't always draw a straight line that passes through every data point, but we can find a line that "comes close" to most of the data. This line is an estimate.

The linear regression model is useful when the outcome variable is continuous. An alternative, logistic regression, should be used if the outcome variable is categorical.

16. Interpreting regression coefficients

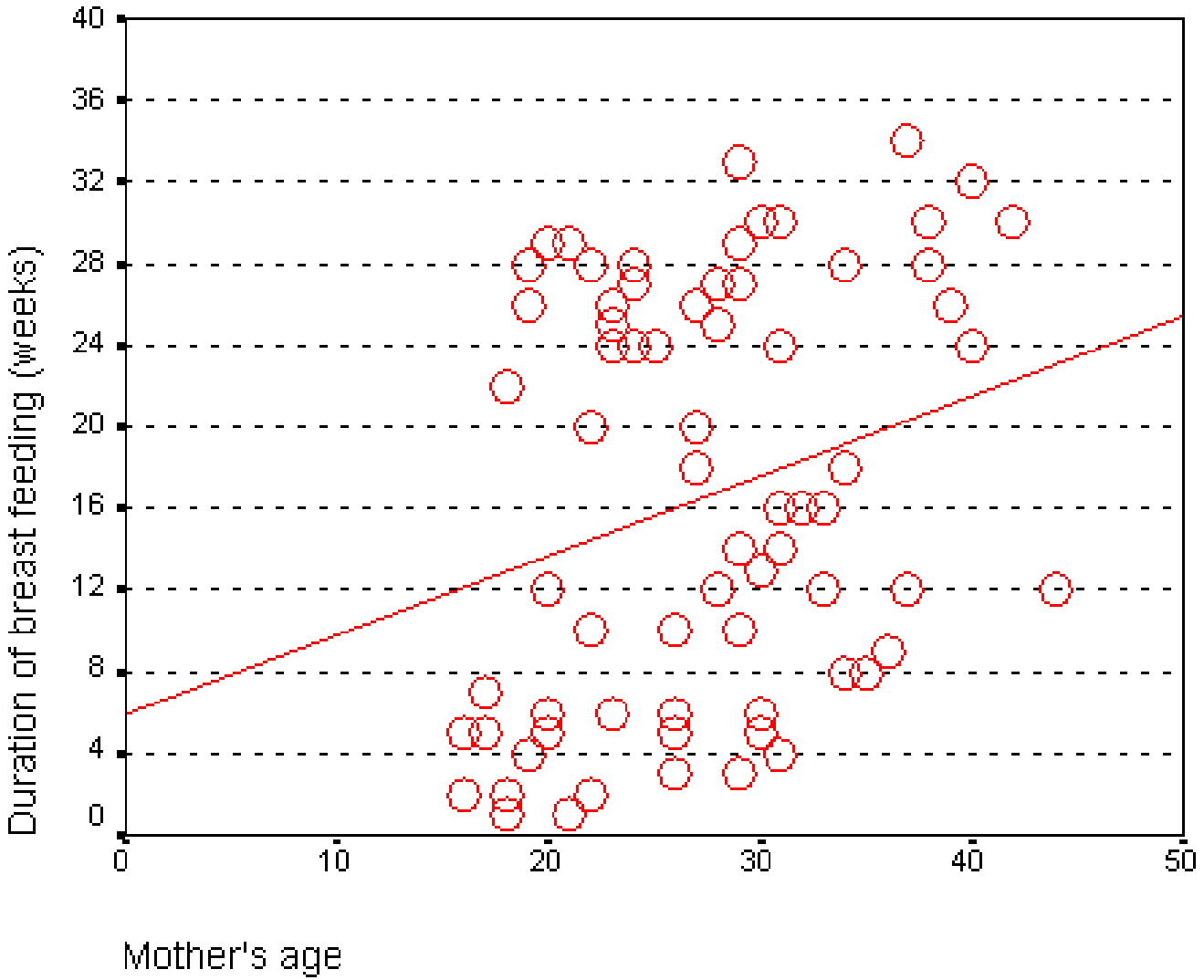
- You should interpret the slope and the intercept of this line as follows:
  - The slope represents the estimated average change in Y when X increases by one unit.
  - The intercept represents the estimated average value of Y when X equals zero.

17. Interpreting regression coefficients

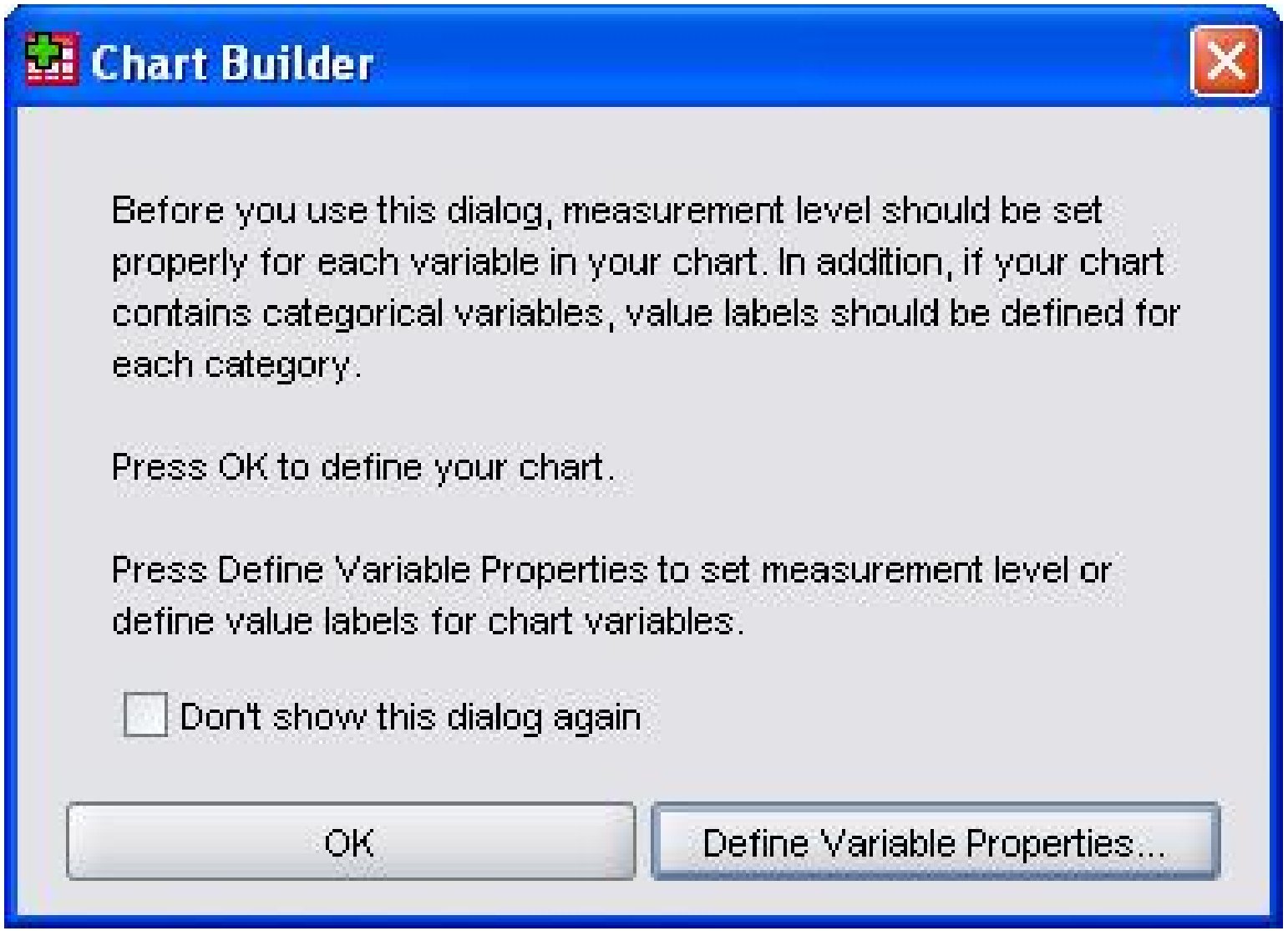
Be cautious with your interpretation of the intercept. Sometimes the value X=0 is impossible, implausible, or represents a dangerous extrapolation outside the range of the data.

The graph on the following page shows an example of a regression line.

18. Interpreting regression coefficients



19. A quick note about graphs in IBM SPSS



20. A quick note about graphs in IBM SPSS

- Scale (continuous variables) 
- Nominal and numeric (categorical variables) 
- Nominal and string (categorical variables) 

21. Interpreting regression coefficients

The intercept is about 6 and the slope is about 0.4.

The slope represents the estimated average change in Y when X increases by one unit. In this case, it means that the estimated average duration of breastfeeding increases by about 0.4 weeks for each additional year of the mother's age.

The intercept represents the estimated average value of Y when X equals zero.

22. Interpreting regression coefficients

The intercept (6) represents the estimated average value of Y when X equals zero. This is meaningless in this setting. If you tried to provide an interpretation, it would be something like this: The estimated average duration of breastfeeding is 6 weeks in a mother who is zero years old.

23. General linear model

If you used SPSS to calculate the linear regression model, you would get the following output.

Parameter Estimates

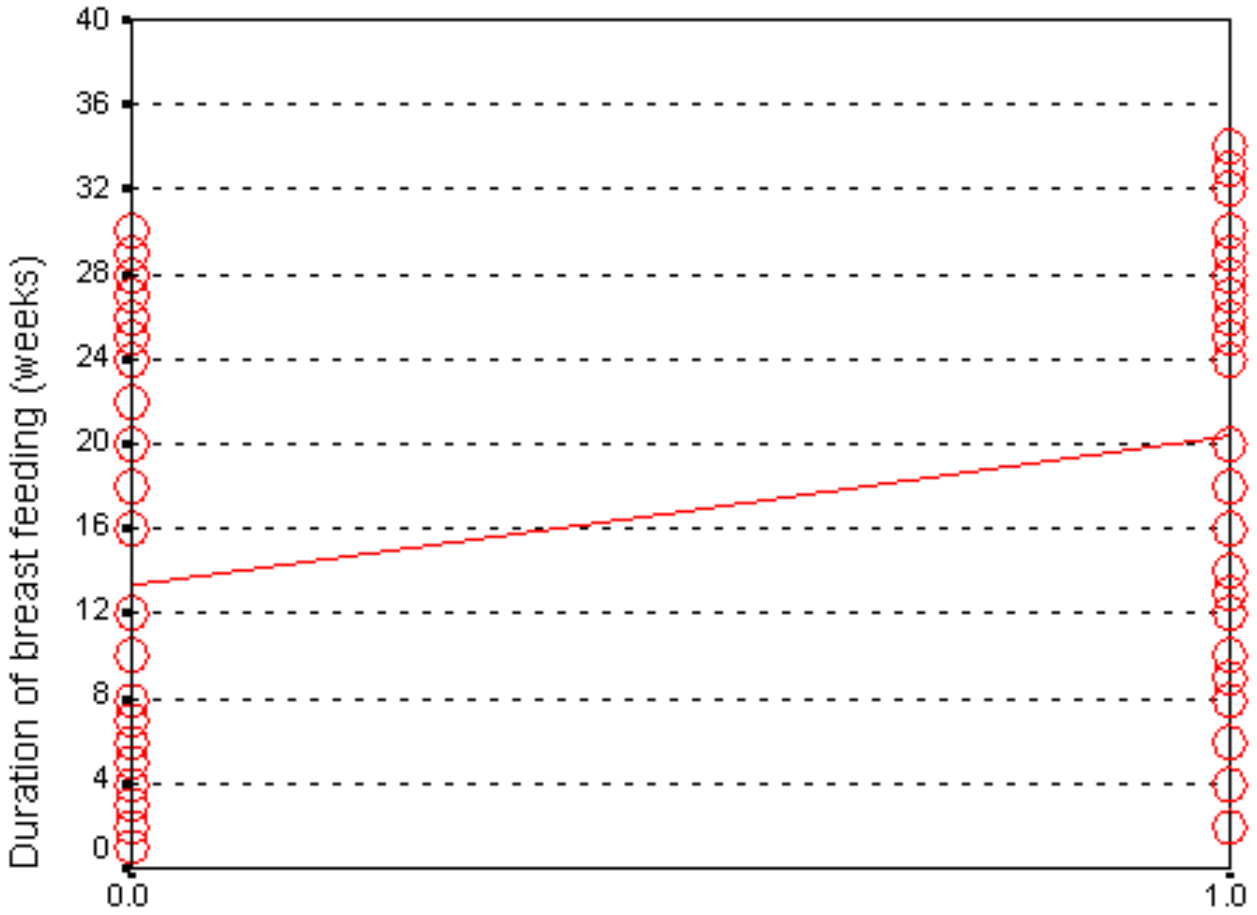
Dependent Variable: Duration of breast feeding (weeks)

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	5.920	4.580	1.292	.200	-3.195	15.035
MOM_AGE	.389	.162	2.399	.019	6.626E-02	.712

24. Interpreting regression coefficients

- When the predictor variable is categorical, your interpretation changes slightly.
  - The slope represents the estimated average change in Y when you switch from one group to the other.
  - The intercept represents the estimated average value of Y for the group coded as zero.

25. Interpreting regression coefficients



26. Interpreting regression coefficients

The intercept is about 13 and the slope is about 7.

The slope represents the estimated average change in Y when you switch from one group to the other. In this case, it means that the estimated average duration of breastfeeding increases by about 7 weeks when you switch from the control group to the treatment group.

The intercept represents the estimated average value of Y when X equals zero.

27. Interpreting regression coefficients

The intercept represents the estimated average value of Y for the group coded as zero. In this case it means that the estimated average duration of breastfeeding is 13 weeks in the control group.

28. Interpreting regression coefficients

If you change the coding (0=treatment group and 1=control group) then the interpretation changes.

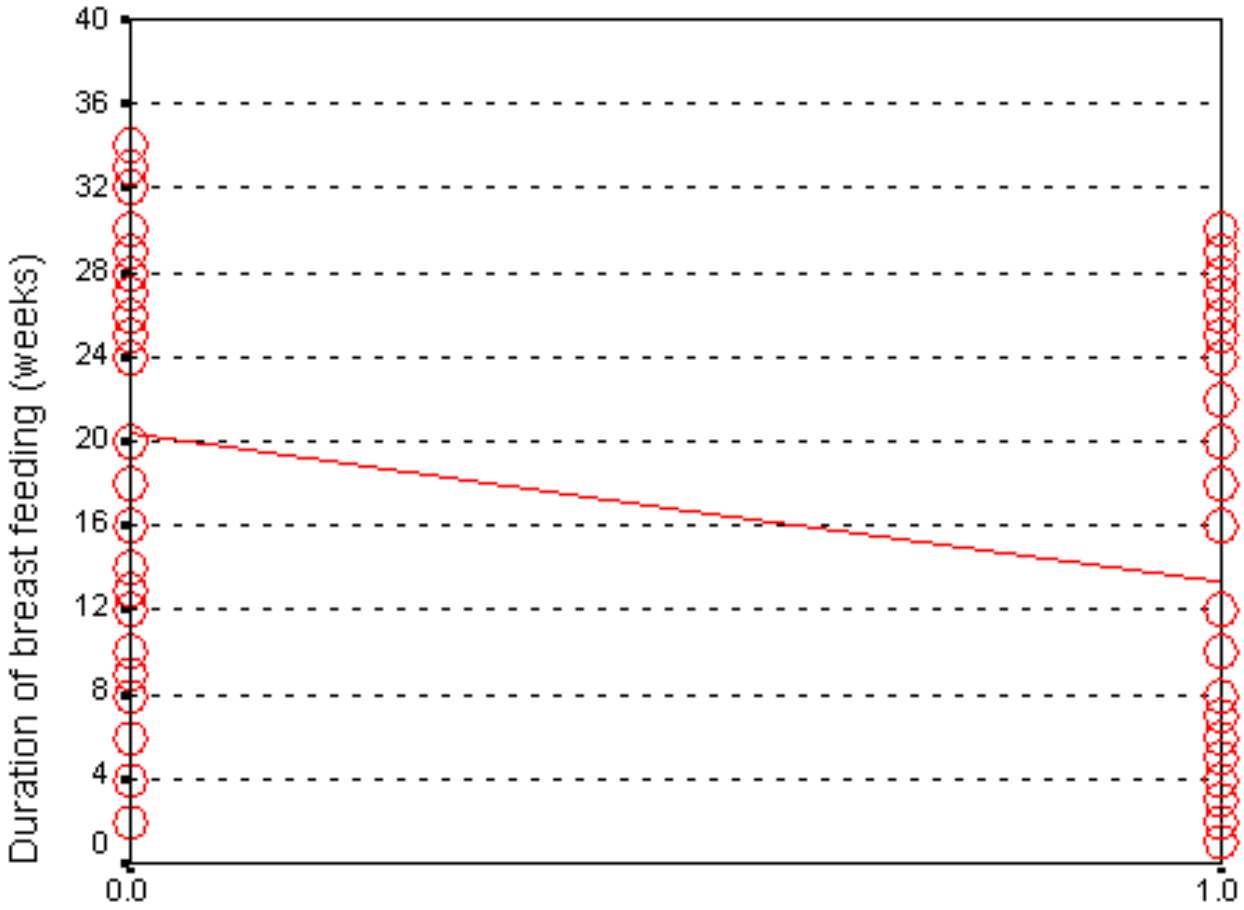
Parameter Estimates

Dependent Variable: Duration of breast feeding (weeks)

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	20.368	1.569	12.983	.000	17.246	23.491
[FEED_TYP=Control]	-7.050	2.142	-3.292	.001	-11.312	-2.788
[FEED_TYP=Treatmen]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

29. Interpreting regression coefficients



30. Interpreting regression coefficients

Suppose you have two independent variables. The linear regression model (sometimes called a multiple linear regression model in this case) has three estimated values:

- Intercept
- Slope for the first independent variable
- Slope for the second independent variable

31. Interpreting regression coefficients

- You should interpret the two slopes and the intercept as follows:
  - The slope for X1 represents the estimated average change in Y when X1 increases by one unit and X2 is held constant. There is a similar interpretation for the slope for X2.
  - The intercept represents the estimated average value of Y when X1 and X2 both equal zero.

32. Interpreting regression coefficients

Parameter Estimates

Dependent Variable: Age when bf stopped

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	12.961	5.146	2.519	.014	2.719	23.203
MOM_AGE	.249	.165	1.510	.135	-7.919E-02	.577
[FEED_TYP=1]	-5.972	2.241	-2.664	.009	-10.434	-1.511
[FEED_TYP=2]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

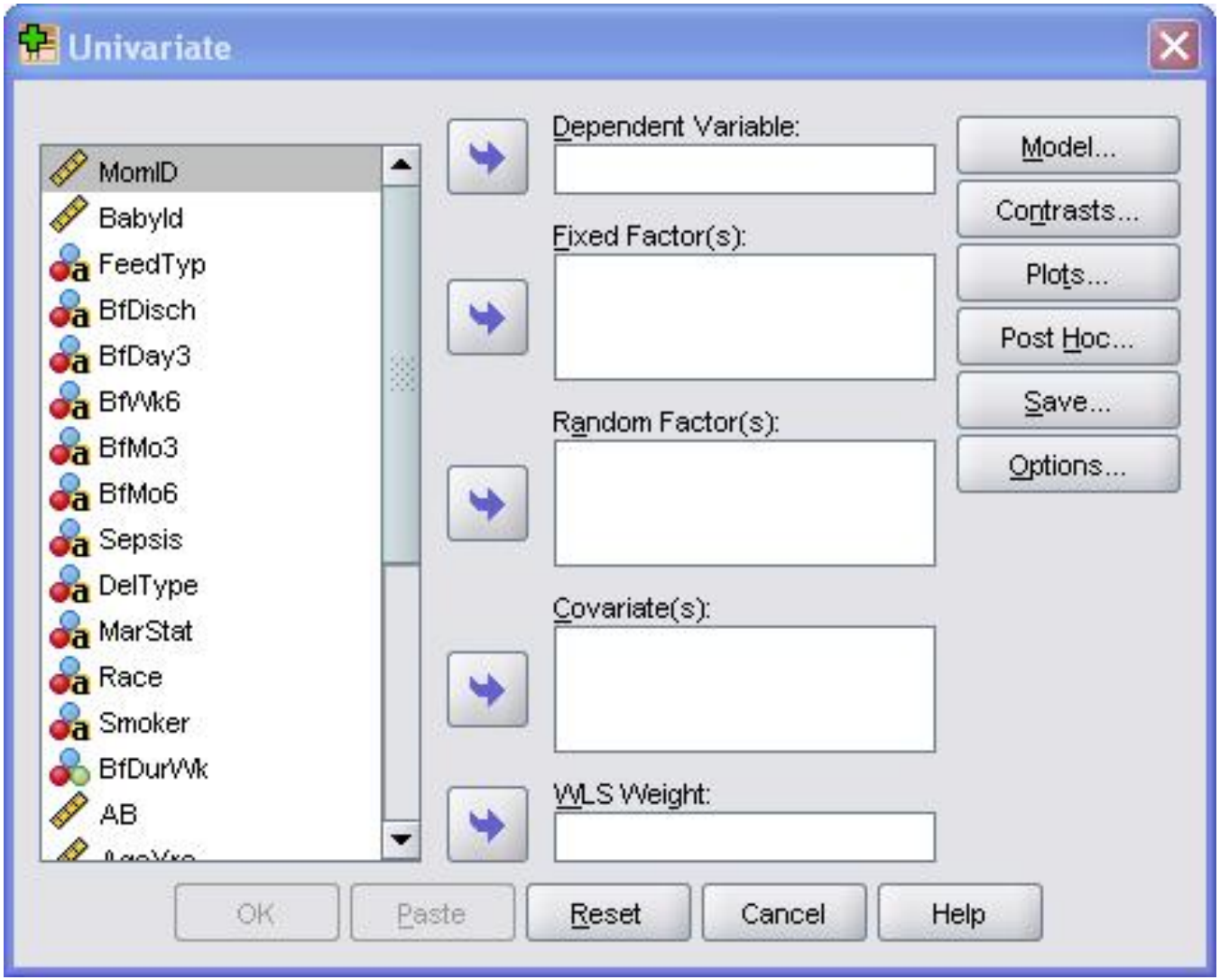
33. General linear model

If you use IBM SPSS, I recommend that you run the general linear model to fit a linear regression line. The general linear model is very flexible and can incorporate many statistical models into one procedure:

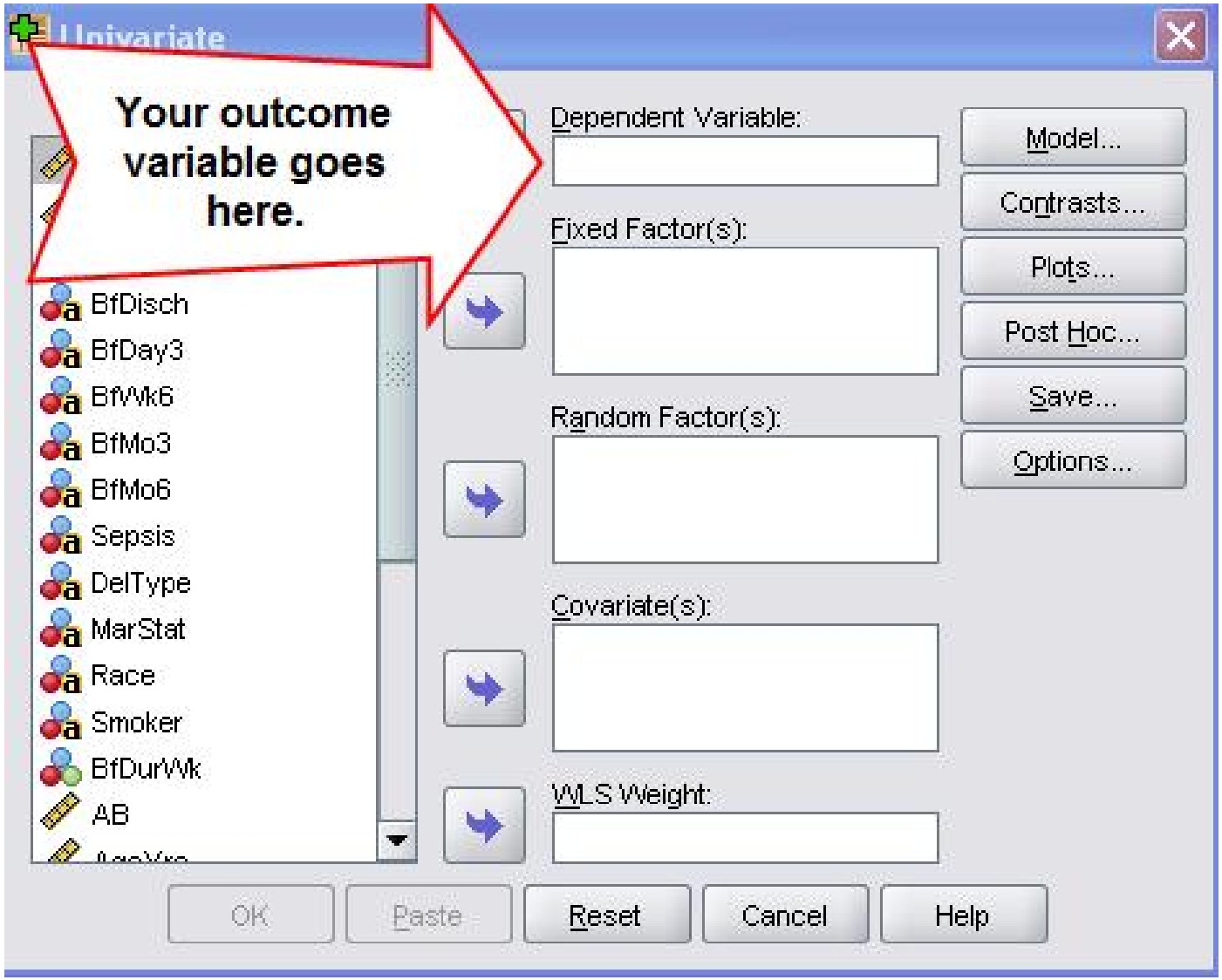
- T-test
- Analysis of Variance (ANOVA)
- Linear regression
- Analysis of covariance (ANCOVA)

Note that the GENERAL linear model is not the same as the GENERALIZED linear model.

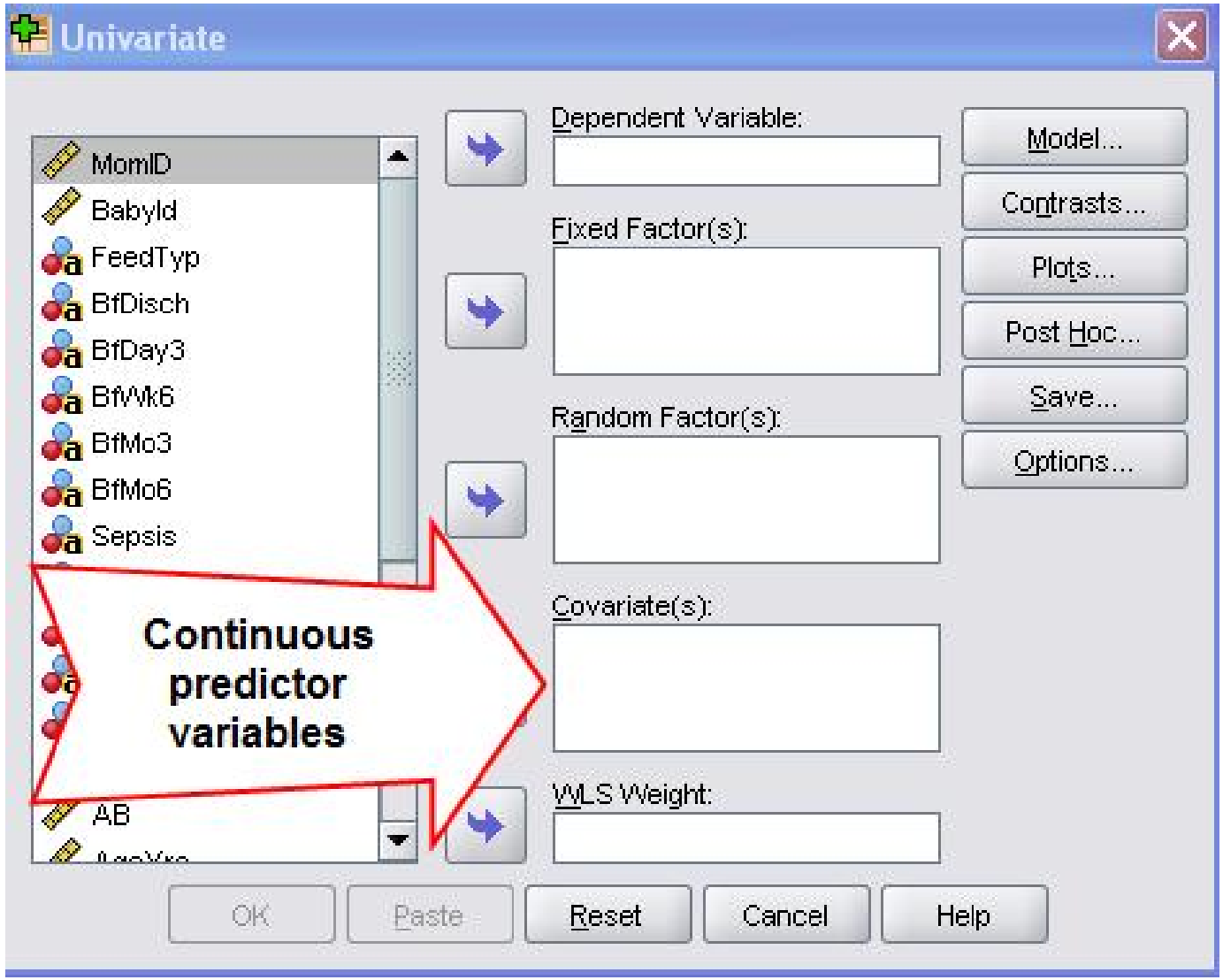
34. General linear model



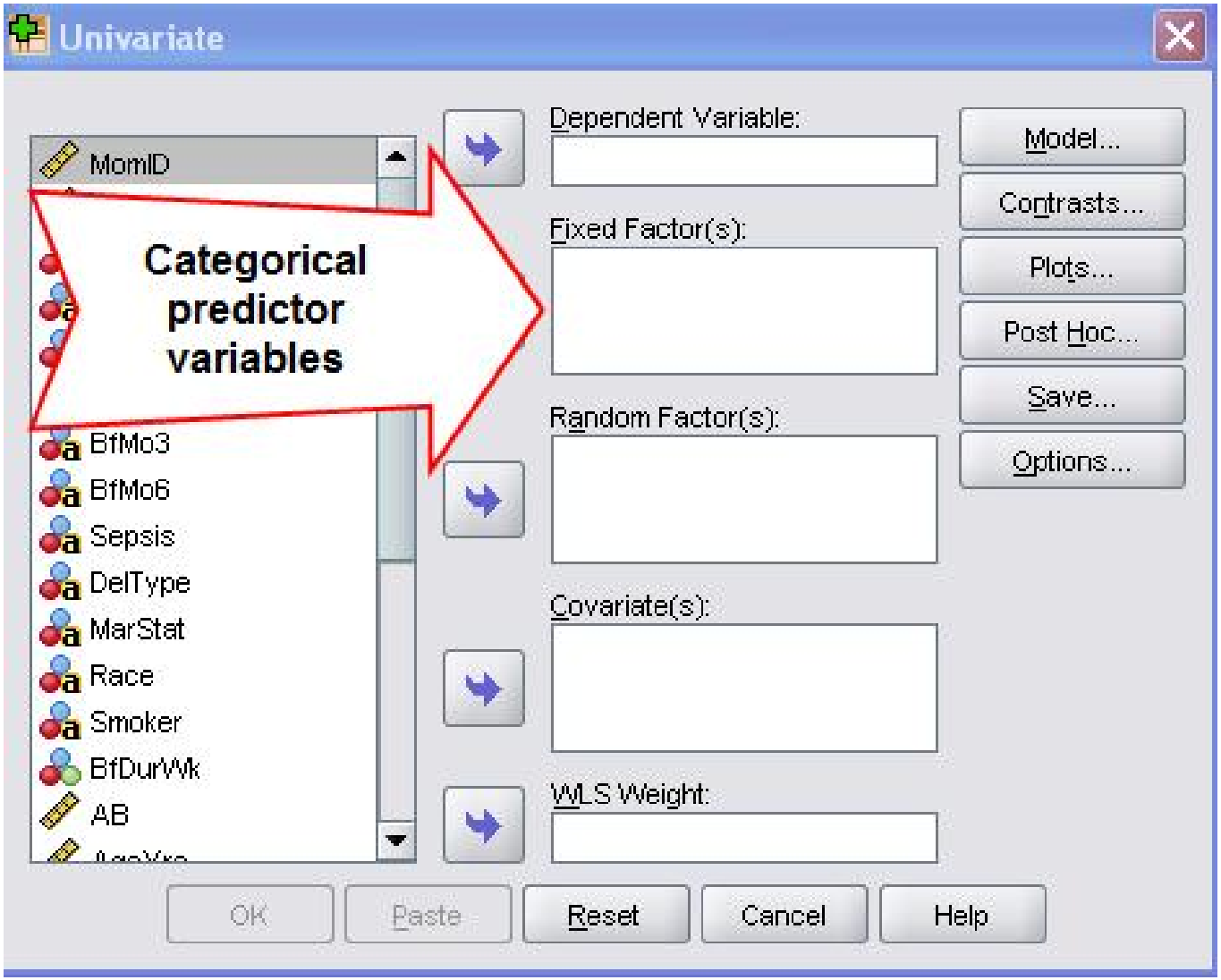
35. General linear model



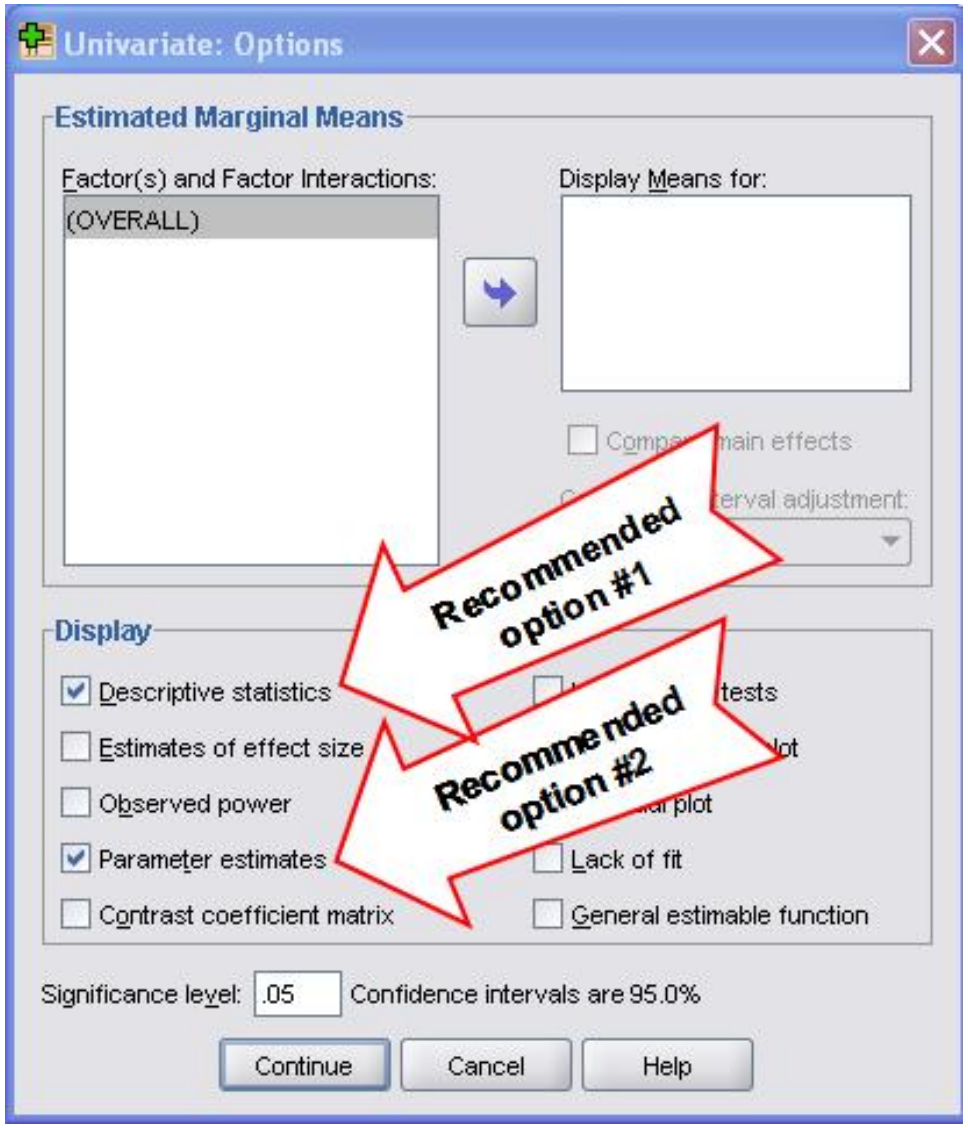
36. General linear model



37. General linear model



38. General linear model



39. Conclusion

The linear regression model is useful when the outcome variable is continuous. The linear regression model can accommodate either categorical or continuous predictor variables. It can also handle multiple predictor variables. The interpretation of the slop in a linear regression model is the estimated average change in Y when X increases by one unit.

40. Repeat of pop quiz #1

In a linear regression model, the slope represents

1. The estimated average change in your outcome variable when the predictor variable increases by one unit
2. The estimated average for your outcome variable in the control group
3. The estimated average for your outcome variable in the treatment group
4. The estimated average for your outcome variable when the predictor variable is zero.
5. The estimated average value for your predictor variable
6. Don’t know/not sure

41. Repeat of pop quiz #2

The linear regression model can accommodate all the following settings, except:

1. A categorical outcome variable
2. A categorical predictor variable
3. A continuous outcome variable
4. A continuous predictor variable
5. Multiple predictor variables.
6. Don’t know/not sure