Multiple Linear Regression I

Lecture 7

Survey Research & Design in Psychology

James Neill, 2016

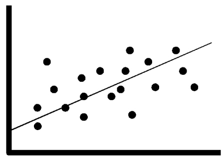
Overview

1. Correlation (Review)
2. Simple linear regression
3. Multiple linear regression
   – General steps
   – Assumptions
   – *R*, coefficients
   – Equation
   – Types
4. Summary
5. MLR I Quiz - Practice questions

Readings

1. Howitt & Cramer (2011/2014):
   – Regression: Prediction with precision [Ch 8/9] [Textbook/eReserve]
   – Multiple regression & multiple correlation [Ch 31/32] [Textbook/eReserve]
2. Tabachnick & Fidell (2013). Multiple regression (includes example write-ups) [eReserve]
3. StatSoft (2016). *How to find relationship between variables, multiple regression*. StatSoft Electronic Statistics Handbook. [Online]

Correlation (Review)



Linear relation between two variables

Purposes of correlational statistics

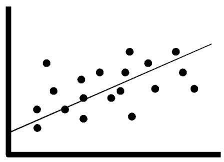
Purpose	Correlation	Factor analysis	Regression
Exploratory	√	√	
Descriptive	√	√	
Explanatory	√		√
Predictive			√

Explanatory - Regression
e.g., hours of study → academic grades

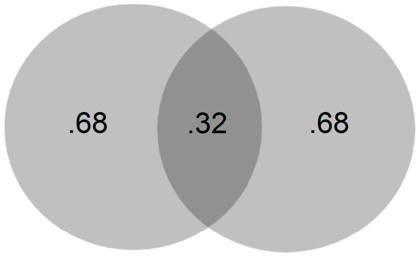
Predictive - Regression
e.g., demographics → life expectancy

Linear correlation

- Linear relations between continuous variables
- Line of best fit on a scatterplot



Correlation is shared variance



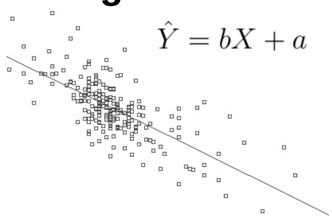
Venn diagrams are helpful for depicting relations between variables.

Correlation – Key points

- Covariance = sum of cross-products (unstandardised)
- Correlation = sum of cross-products (standardised), ranging from -1 to 1
  (sign indicates direction, value indicates size)
- Coefficient of determination (r^2) indicates % of shared variance
- Correlation does not necessarily equal causality

Simple linear regression

$$\hat{Y} = bX + a$$



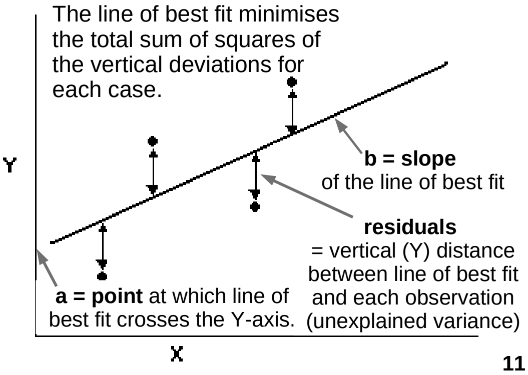
Explains and predicts a Dependent Variable (DV) based on a linear relation with an Independent Variable (IV)

What is simple linear regression?

- An extension of correlation
- Best-fitting straight line for a scatterplot between two variables. Involves:
  - a **predictor (X)** variable – also called an independent variable (IV)
  - an **outcome (Y)** variable - also called a dependent variable (DV) or criterion variable
- Uses an IV to explain/predict a DV
- Can help to understand possible causal effects of one variable on another.

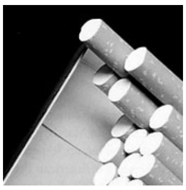
resi

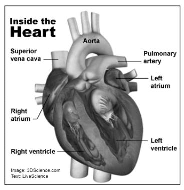
Least squares criterion



Linear Regression - Example: Cigarettes & coronary heart disease

Example from Landwehr & Watkins (1987), cited in Howell (2004, pp. 216-218) and accompanying lecture notes.





IV = Cigarette consumption

DV = Coronary Heart Disease

Linear regression - Example: Cigarettes & coronary heart disease (Howell, 2004)

Research question:
How fast does CHD mortality rise with a one unit increase in smoking?

- **IV** = Av. # of cigs per adult per day
- **DV** = CHD mortality rate (deaths per 10,000 per year due to CHD)
- **Unit of analysis** = Country

Linear regression - Data: Cigarettes & coronary heart disease
(Howell, 2004)

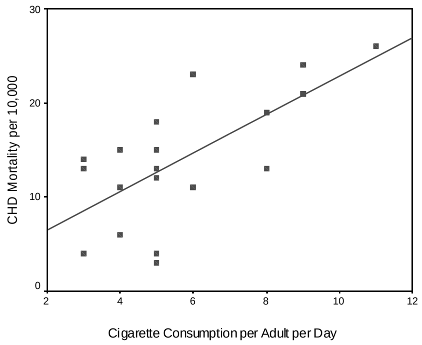
Cigarette Consumption and Coronary Heart Disease Mortality for 21 Countries

Cig.	11	9	9	9	8	8	8	6	6	5	5
CHD	26	21	24	21	19	13	19	11	23	15	13

Cig.	5	5	5	5	4	4	4	3	3	3
CHD	4	18	12	3	11	15	6	13	4	14

Cig. = Cigarettes per adult per day
CHD = Cornary Heart Disease Mortality per 10,000 population

Linear regression - Example: Scatterplot with Line of Best Fit



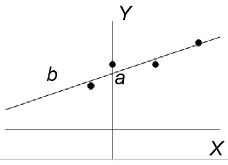
Linear regression equation
(without error)

$$\hat{Y} = bX + a$$

predicted values of *Y*

slope = rate of increase/decrease of Y hat for each unit increase in *X*

Y-intercept = level of *Y* when X is *0*.



Linear regression equation (with error)

$$Y = bX + a + e$$

X = IV values

Y = DV values

a = *Y*-axis intercept

b = slope of line of best fit

(regression coefficient)

e = error

Linear regression – Example: Equation

Variables: $\hat{Y} = bX + a$

- (DV) = predicted rate of CHD mortality
- *X* (IV) = mean # of cigarettes per adult per day per country

Regression co-efficients:

- *b* = rate of ↑/↓ of CHD mortality for each extra cigarette smoked per day
- *a* = baseline level of CHD (i.e., CHD when no cigarettes are smoked)

Linear regression – Example: Explained variance

- $r = .71$
- $R^2 = .71^2 = .51$
- Approximately 50% in variability of incidence of CHD mortality is associated with variability in smoking rates.

19

Linear regression – Example: Test for overall significance

- $R = .71$, $R^2 = .51$, $p < .05$

ANOVA[b]

	Sum of Squares	df	Mean Square	F	Sig.
Regression	454.482	1	454.48	19.59	.00[a]
Residual	440.757	19	23.198		
Total	895.238	20			

a. Predictors: (Constant), Cigarette Consumption per Adult per Day

b. Dependent Variable: CHD Mortality per 10,000

Linear regression – Example: Regression coefficients - SPSS

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
a	(Constant)	2.37	2.941		.80	.43
b	Cigarette Consumption per Adult per Day	2.04	.461	.713	4.4	.00

a. Dependent Variable: CHD Mortality per 10,000

Linear regression - Example: Making a prediction

- What if we want to predict CHD mortality when cigarette consumption is 6?

$$\hat{Y} = bX + a = 2.04X + 2.37$$
$$\hat{Y} = 2.04 * 6 + 2.37 = 14.61$$

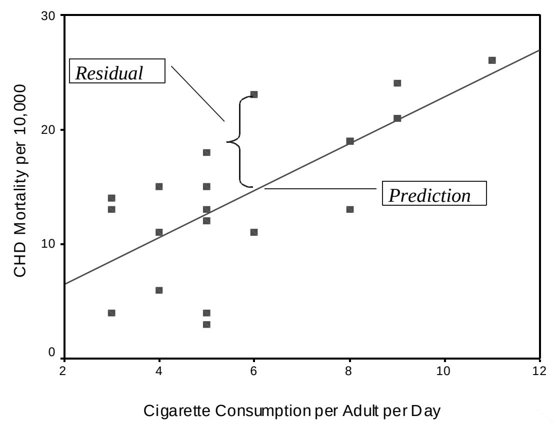
- We predict that 14.61 / 10,000 people in a country with an average cigarette consumption of 6 per person will die of coronary heart disease per annum.

22

Linear regression - Example: Accuracy of prediction - Residual

- Finnish smokers smoke 6 cigarettes/adult/day
- We predict 14.61 deaths /10,000
- But Finland actually has 23 deaths / 10,000
- Therefore, the error ("residual") for this case is 23 - 14.61 = 8.39

23



Hypothesis testing

Null hypotheses (H_0):

- *a* (Y-intercept) = 0
- *b* (slope of line of best fit) = 0

25

Linear regression – Example: Testing slope and intercept

Coefficients[a]

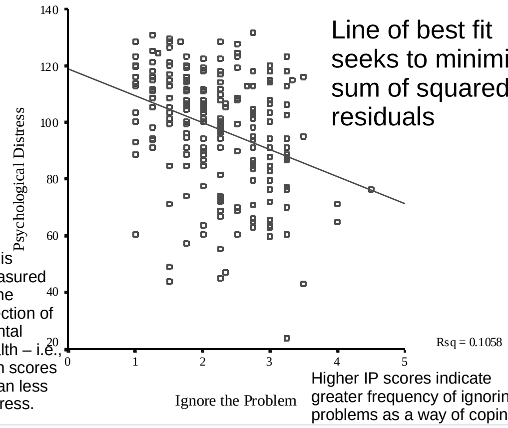
		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
a	(Constant)	2.37	2.941		.80	.43
b	Cigarette Consumption per Adult per Day	2.04	.461	.713	4.4	.00

a. Dependent Variable: CHD Mortality per 10,000

Linear regression - Example

Does a tendency to 'ignore problems' (IV) predict 'psychological distress' (DV)?

27



Line of best fit seeks to minimise sum of squared residuals

PD is measured in the direction of mental health – i.e., high scores mean less distress.

Higher IP scores indicate greater frequency of ignoring problems as a way of coping.

28

Linear regression - Example

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.325[a]	.106	.102	19.4851

a. Predictors: (Constant), IGNO2 ACS Time 2 - 11. Ignore

R = .32, R^2 = .11, Adjusted R^2 = .10

Ignoring Problems accounts for ~10% of the variation in Psychological Distress

The predictor (Ignore the Problem) explains approximately 10% of the variance in the dependent variable (Psychological Distress).

29

Linear regression - Example

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9789.888	1	9789.888	25.785	.000[a]
	Residual	82767.884	218	379.669		
	Total	92557.772	219			

a. Predictors: (Constant), IGNO2 ACS Time 2 - 11. Ignore

b. Dependent Variable: GWB2NEG

The population relationship between Ignoring Problems and Psychological Distress is unlikely to be 0% because p = .000

(i.e., reject the null hypothesis that there is no relationship)

30

Linear regression - Example

Coefficients[a]

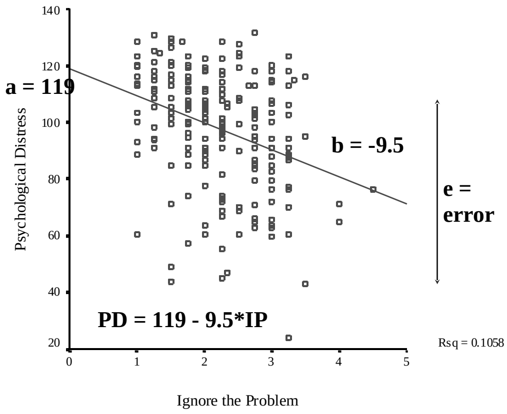
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	118.897	4.351		27.327	.000
	IGNO2 ACS Time 2 - 11. Ignore	-9.505	1.872	-.325	-5.078	.000

a. Dependent Variable: GWB2NEG

There is a sig. *a* or constant (Y-intercept) - this is the baseline level of Psychological Distress.

In addition, Ignore Problems (IP) is a significant predictor of Psychological Distress (PD).

PD = 119 - 9.5*IP

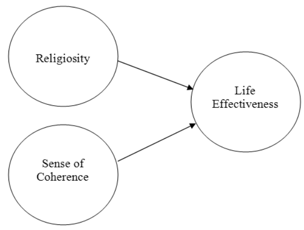


Linear regression summary

- Linear regression is for *explaining* or *predicting* the linear relationship between two variables
- Y = *b*x + a + e
- = *b*x + a

(*b* is the slope; *a* is the *Y*-intercept)

Multiple Linear Regression



Linear relations between two or more IVs and a single DV

What is multiple linear regression (MLR)? Visual model

Linear Regression

Single predictor X Y

Multiple Linear Regression

Multiple predictors X_1 X_2 X_3 X_4 X_5 Y

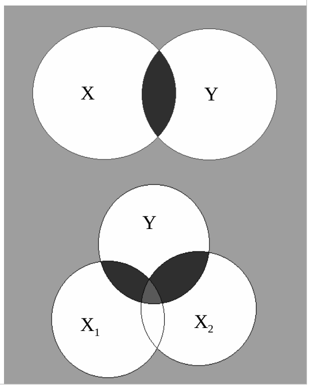
What is MLR?

- Use of several IVs to predict a DV
- Weights each predictor (IV) according to the strength of its linear relationship with the DV
- Makes adjustments for inter-relationships among predictors
- Provides a measure of overall fit (*R*)

What is MLR?

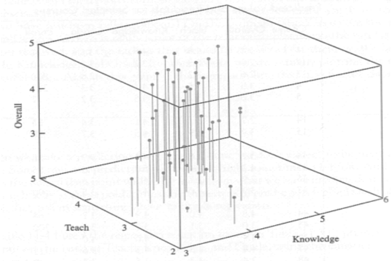
Correlation / Regression

Correlation
Partial correlation
MLR



What is MLR?

A 3-way scatterplot can depict the correlational relationship between 3 variables.



However, it is difficult to graph/visualise 4+-way relationships via scatterplot.

General steps

1. Develop a visual model and express a research question and/or hypotheses
2. Check assumptions
3. Choose type of MLR
4. Interpret output
5. Develop a regression equation (if needed)

LR → MLR example:
Cigarettes & coronary heart disease

- ~50% of the variance in CHD mortality could be explained by cigarette smoking (using LR)
- Strong effect - but what about the other 50% ('unexplained' variance)?
- What about other predictors?
  – e.g., exercise and cholesterol?

MLR – Example
Research question 1

How well do these three IVs:
- # of cigarettes / day (IV_1)
- exercise (IV_2) and
- cholesterol (IV_3)

predict
- CHD mortality (DV)?

Cigarettes
Exercise
Cholesterol

CHD Mortality

MLR – Example
Research question 2

To what extent do personality factors (IVs) predict annual income (DV)?

Extraversion
Neuroticism
Psychoticism

Income

MLR - Example
Research question 3

"Does the # of years of formal study of psychology (IV1) and the no. of years of experience as a psychologist (IV2) predict clinical psychologists' effectiveness in treating mental illness (DV)?"

Study
Experience
Effectiveness

43

MLR - Example
Your example

Generate your own MLR research question (e.g., based on some of the following variables):

- Gender & Age
- Stress & Coping
- Uni student satisfaction
  - Teaching/Education
  - Social
  - Campus
- Time management
  - Planning
  - Procrastination
  - Effective actions
- Health
  - Psychological
  - Physical

44

Assumptions

- Levels of measurement
- Sample size
- Normality (univariate, bivariate, and multivariate)
- Linearity: Linear relations between IVs & DVs
- Homoscedasticity
- Multicollinearity
  - IVs are not overly correlated with one another (e.g., not over .7)
- Residuals are normally distributed

45

Levels of measurement

- **DV = Continuous**
  (Interval or Ratio)
- **IV = Continuous or Dichotomous**
  (if neither, may need to recode into a dichotomous variable or create dummy variables)

46

Dummy coding

- "Dummy coding" converts a more complex variable into a series of dichotomous variables (i.e., 0 or 1)
- So, dummy variables are dichotomous variables created from a variable with a higher level of measurement.

47

Dummy coding - Example

- Religion
  (1 = Christian; 2 = Muslim; 3 = Atheist) can't be an IV in regression
  (a linear correlation with a categorical variable doesn't make sense).
- However, it can be dummy coded into dichotomous variables:
  - `Christian (0 = no; 1 = yes)`
  - `Muslim    (0 = no; 1 = yes)`
  - ~~`Atheist   (0 = no; 1 = yes)`~~ `(redundant)`
- These variables can then be used as IVs.
- More information (Dummy variable (statistics), Wikiversity)

48

Sample size: Some rules of thumb

- Enough data is needed to provide reliable estimates of the correlations.
- *N* >= 50 cases and *N* >= 10 to 20 as many cases as there are IVs, otherwise the estimates of the regression line are probably unstable and are unlikely to replicate if the study is repeated.
- Green (1991) and Tabachnick & Fidell (2013) suggest:
  - 50 + 8(*k*) for testing an overall regression model and
  - 104 + *k* when testing individual predictors (where k is the number of IVs)
  - Based on detecting a medium effect size (β >= .20), with critical α <= .05, with power of 80%.

Dealing with outliers

Extreme cases should be deleted or modified if they are overly influential.

- Univariate outliers - detect via initial data screening
- Bivariate outliers - detect via scatterplots
- Multivariate outliers - unusual combination of predictors – detect via Mahalanbis' distance

Multivariate outliers

- A case may be within normal range for each variable individually, but be a multivariate outlier based on an unusual combination of responses which unduly influences multivariate test results.
- e.g., a person who:
  - Is 18 years old
  - Has 3 children
  - Has a post-graduate degree

Multivariate outliers

- Identify & check unusual cases
- Use Mahalanobis' distance or Cook's *D* as a MV outlier screening procedure

Multivariate outliers

- Mahalanobis' distance (MD)
  - Distributed as χ^2 with *df* equal to the number of predictors (with critical $\alpha = .001$)
  - Cases with a MD greater than the critical value are multivariate outliers.
- Cook's *D*
  - Cases with CD values > 1 are multivariate outliers.
- Use either MD or CD
- Examine cases with extreme MD or CD scores - if in doubt, remove & re-run.

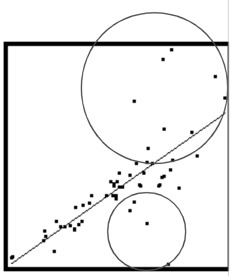
Normality & homoscedasticity

Normality

- If variables are non-normal, this will create heteroscedasticity

Homoscedasticity

- Variance around the regression line should be the same throughout the distribution
- Even spread in residual plots



Multicollinearity

- **Multicollinearity** – IVs shouldn't be overly correlated (e.g., over .7) – if so, consider removing one.
- **Singularity** - perfect correlations among IVs.
- Leads to unstable regression coefficients.

Multicollinearity

Detect via:

- **Correlation matrix** - are there large correlations among IVs?
- **Tolerance statistics** - if < .3 then exclude that variable.
- **Variance Inflation Factor (VIF)** – if < 3, then exclude that variable.
- VIF is the reciprocal of Tolerance (so use one or the other – not both)

Causality

- Like correlation, regression does not tell us about the causal relationship between variables.
- In many analyses, the IVs and DVs could be swapped around – therefore, it is important to:
  - Take a theoretical position
  - Acknowledge alternative explanations

Multiple correlation coefficient (*R*)

- "Big R" (capitalised)
- Equivalent of *r*, but takes into account that there are multiple predictors (IVs)
- Always positive, between 0 and 1
- Interpretation is similar to that for *r* (correlation coefficient)

Coefficient of determination (R^2)

- "Big R squared"
- Squared multiple correlation coefficient
- Usually report R^2 instead of *R*
- Indicates the % of variance in DV explained by combined effects of the IVs
- Analogous to r^2

Rule of thumb for interpretation of R^2

- .00 = no linear relationship
- .10 = small (R ~ .3)
- .25 = moderate (R ~ .5)
- .50 = strong (R ~ .7)
- 1.00 = perfect linear relationship

R^2 ~ .30 is good for social sciences

Adjusted R^2

- R^2 is explained variance in a sample.
- Adjusted R^2 is used for estimating explained variance in a population.
- Report R^2 and adjusted R^2
- Particularly for small N and where results are to be generalised, take more note of adjusted R^2

Multiple linear regression – Test for overall significance

- Shows if there is a linear relationship between all of the X variables taken together and Y
- Examine F and p in the ANOVA table to determine the likelihood that the explained variance in Y could have occurred by chance

Regression coefficients

- Y-intercept (*a*)
- Slopes (*b*):
  - Unstandardised
  - Standardised
- Slopes are the weighted loading of each IV on the DV, adjusted for the other IVs in the model.

Unstandardised regression coefficients

- *B* = unstandardised regression coefficient
- Used for regression equations
- Used for predicting Y scores
- But can't be compared with other *B*s unless all IVs are measured on the same scale

Standardised regression coefficients

- Beta (β) = standardised regression coefficient
- Useful for comparing the relative strength of predictors
- $\beta = r$ in LR but this is only true in MLR when the IVs are uncorrelated.

Test for significance: Individual variables

Indicates the likelihood of a linear relationship between each variable X_i and Y occurring by chance.

Hypotheses:

H_0: $\beta_i = 0$ (No linear relationship)

H_1: $\beta_i \neq 0$ (Linear relationship between X_i and Y)

Relative importance of IVs

- Which IVs are the most important?
- To answer this, compare the standardised regression coefficients (β's)

67

Regression equation

$$Y = b_1x_1 + b_2x_2 + \ldots + b_ix_i + a + e$$

- Y = observed DV scores
- b_i = unstandardised regression coefficients (the *B*s in SPSS) - slopes
- x_1 to x_i = IV scores
- a = Y axis intercept
- e = error (residual)

68

Multiple linear regression - Example

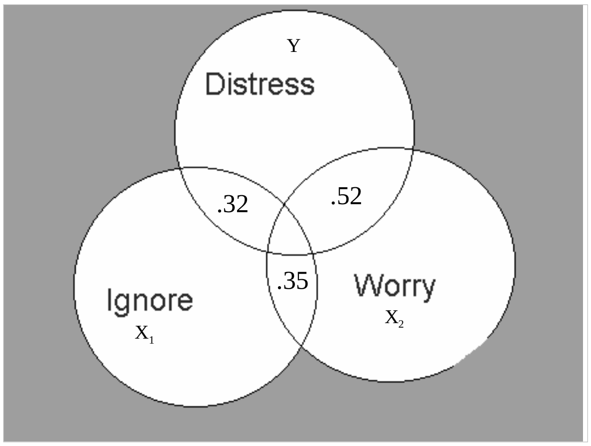
"Does 'ignoring problems' (IV_1) and 'worrying' (IV_2) predict 'psychological distress' (DV)"

69

Correlations

	Psychological Distress	Worry	Ignore the Problem
Psychological Distress	1.000	-.521	-.325
Worry	-.521	1.000	.352
Ignore the Problem	-.325	.352	1.000
Psychological Distress	.	.000	.000
Worry	.000	.	.000
Ignore the Problem	.000	.000	.
Psychological Distress	220	220	220
Worry	220	220	220
Ignore the Problem	220	220	220

70



Multiple linear regression - Example

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.543[a]	.295	.288	17.34399

a. Predictors: (Constant), Ignore the Problem, Worry

b. Dependent Variable: Psychological Distress

Together, Ignoring Problems and Worrying explain 30% of the variance in Psychological Distress in the Australian adolescent population (R^2 = .30, Adjusted R^2 = .29).

72

Multiple linear regression - Example

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	27281.12	2	13640.558	45.345	.000[a]
	Residual	65276.66	217	300.814		
	Total	92557.77	219			

a. Predictors: (Constant), Ignore the Problem, Worry

b. Dependent Variable: Psychological Distress

The explained variance in the population is unlikely to be 0 (p = .00).

Multiple linear regression - Example

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	138.932	4.680		29.687	.000
	Worry	-11.511	1.510	-.464	-7.625	.000
	Ignore the Problem	-4.735	1.780	-.162	-2.660	.008

a. Dependent Variable: Psychological Distress

Worry predicts about three times as much variance in Psychological Distress than Ignoring the Problem, although both are significant, negative predictors of mental health.

Multiple linear regression - Example – Prediction equations

Linear Regression

PD (hat) = 119 – 9.50*Ignore

R^2 = .11

Multiple Linear Regression

PD (hat) = 139 - .4.7*Ignore - 11.5*Worry

R^2 = .30

	B
(Constant)	138.932
Worry	-11.511
Ignore the Problem	-4.735

Confidence interval for the slope

Coefficients[a]

Model		Standardized Coefficients	95% Confidence Interval for B	
		Beta	Lower Bound	Upper Bound
1	(Constant)		129.708	148.156
	Worry	-.464	-14.486	-8.536
	Ignore the Problem	-.162	-8.242	-1.227

a. Dependent Variable: Psychological Distress

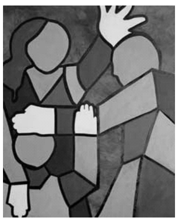
Mental Health (PD) is reduced by between 8.5 and 14.5 units per increase of Worry units.

Mental Health (PD) is reduced by between 1.2 and 8.2 units per increase in Ignore the Problem units.

Multiple linear regression - Example

Effect of violence, stress, social support on internalising behaviour problems

Kliewer, Lepore, Oskin, & Johnson, (1998)



Multiple linear regression – Example - Study

- Participants were children:
  - 8 - 12 years
  - Lived in high-violence areas, USA
- **Hypotheses**:
  - Violence and stress → ↑ internalising behaviour
  - Social support → ↓ internalising behaviour.

Multiple linear regression – Example - Variables

- **Predictors**
  - –Degree of witnessing violence
  - –Measure of life stress
  - –Measure of social support
- **Outcome**
  - –Internalising behaviour (e.g., depression, anxiety, withdrawal symptoms) – measured using the Child Behavior Checklist (CBCL)

Correlations

Pearson Correlation

Correlations amongst the IVs	Amount violenced witnessed	Current stress	Social support	Internalizing symptoms on CBCL
Amount violenced witnessed				
Current stress	.050			
Social support	.080	-.080		
Internalizing symptoms on CBCL	.200*	.270*	-.170	

*. Correlation is significant at the 0.05 level (2-tailed).

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

Correlations between the IVs and the DV

R^2

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.37[a]	.135	.108	2.2198

a. Predictors: (Constant), Social support, Current stress, Amount violenced witnessed

Coefficients[a]

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
(Constant)	.477	1.289		.37	.712
Amount violenced witnessed	.038	.018	.201	2.1	.039
Current stress	.273	.106	.247	2.6	.012
Social support	-.074	.043	-.166	-2	.087

a. Dependent Variable: Internalizing symptoms on CB

Regression equation

$$\hat{Y} = b_1 X_1 + b_2 X_2 + b_3 X_3 + b_0$$
$$= 0.038 Wit + 0.273 Stress - 0.074 SocSupp + 0.477$$

- A separate coefficient or slope for each variable
- An intercept (here its called b_0)

Interpretation

$$\hat{Y} = b_1 X_1 + b_2 X_2 + b_3 X_3 + b_0$$
$$= 0.038 Wit + 0.273 Stress - 0.074 SocSupp + 0.477$$

- Slopes for Witness and Stress are +ve; slope for Social Support is -ve.
- Ignoring Stress and Social Support, a one unit increase in Witness would produce .038 unit increase in Internalising symptoms.

Predictions

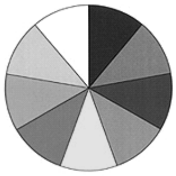
If Witness = 20, Stress = 5, and SocSupp = 35, then we would predict that internalising symptoms would be … .012.

$$\hat{Y} = .038 * Wit + .273 * Stress - .074 * SocSupp + 0.477$$
$$= .038(20) + .273(5) - .074(35) + 0.477$$
$$= .012$$

Multiple linear regression - Example

The role of human, social, built, and natural capital in explaining life satisfaction at the country level:
Towards a National Well-Being Index (NWI)

Vemuri & Costanza (2006)



Variables

- IVs:
  - –Human & Built Capital
    (Human Development Index)
  - –Natural Capital
    (Ecosystem services per km^2)
  - –Social Capital
    (Press Freedom)
- DV = Life satisfaction
- Units of analysis: Countries
  (*N* = 57; mostly developed countries, e.g., in Europe and America)

Table 1
Bivariate correlations between variables

		Average life satisfaction	HDI	Log ESP/km^2 in
Average life satisfaction	Pearson cor.	1		
	Significance			
HDI	Pearson cor.	.463	1	
	Significance	.000		
Log ESP/km^2 index	Pearson cor.	.358	.071	1
	Significance	.007	.353	
Press freedom	Pearson cor.	.502	.502	.295
	Significance	.000	.000	.000

- There are moderately strong positive and statistically significant linear relations between the IVs and the DV
- The IVs have small to moderate positive inter-correlations.

Table 2
Basic regression model coefficients for national-level analysis

	Unstandardized coefficients		Standardized coefficients	*t*-value	Significance
	B	Std. error	Beta		
Constant	1.857	.900		2.063	.044
HDI	3.524	.832	.470	4.234	.000
Log ESP/km^2 Index	3.498	1.021	.380	3.427	.001

Sample size of the regression model was 56.

- R^2 = .35
- Two sig. IVs (not Social Capital - dropped)

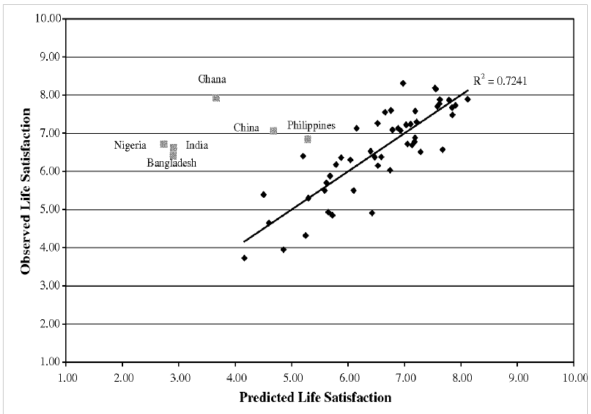


Fig. 2. Observed versus predicted life satisfaction.

Table 4
Revised regression model coefficients for national-level analysis

	Unstandardized coefficients		Standardized coefficients	*t*-value	Significance
	B	Std. error	Beta		
Constant	−2.220	.799		−2.781	.008
HDI	8.875	.884	.777	10.038	.000
Log ESP/km^2 index	2.453	.739	.257	3.319	.002

Sample size of the regression model was 50.

- R^2 = .72
  (after dropping 6 outliers)

91

Types of MLR

- Standard or direct (simultaneous)
- Hierarchical or sequential
- Stepwise (forward & backward)

92

Direct or Standard

- All predictor variables are entered together (simultaneously)
- Allows assessment of the relationship between all predictor variables and the criterion (*Y*) variable *if there is good theoretical reason for doing so.*
- Manual technique & commonly used

93

Hierarchical (Sequential)

- IVs are entered in blocks or stages.
  - Researcher defines order of entry for the variables, based on theory.
  - May enter ‘nuisance’ variables first to ‘control’ for them, then test ‘purer’ effect of next block of important variables.
- R^2 change - additional variance in *Y* explained at each stage of the regression.
  - *F* test of R^2 change.

94

Hierarchical (Sequential)

- Example
  - Drug A is a cheap, well-proven drug which reduces AIDS symptoms
  - Drug B is an expensive, experimental drug which could help to cure AIDS
  - Hierarchical linear regression:
    - Step 1: Drug A (IV1)
    - Step 2: Drug B (IV2)
    - DV = AIDS symptoms
    - Research question: To what extent does Drug B reduce AIDS symptoms *above and beyond* the effect of Drug A?
    - Examine the change in R^2 between Step 1 & Step 2

95

Forward selection

- The strongest predictor variables are entered, one by one, if they reach a criteria (e.g., p < .05)
- Best predictor = IV with the highest *r* with *Y*
- Computer-driven - controversial

96

Backward elimination

- All predictor variables are entered, then the weakest predictors are removed, one by one, if they meet a criteria (e.g., $p > .05$)
- Worst predictor = x with the lowest r with Y
- Computer-driven - controversial

Stepwise

- Combines forward & backward.
- At each step, variables may be entered or removed if they meet certain criteria.
- **Useful for developing the best prediction equation** from a large number of variables.
- Redundant predictors removed.
- Computer-driven - controversial

Which method?

- Standard: To assess impact of all IVs simultaneously
- Hierarchical: To test IVs in a specific order (based on hypotheses derived from theory)
- Stepwise: If the goal is accurate statistical prediction e.g., from a large # of variables - computer driven

Summary

Summary: General steps

1. Develop model and hypotheses
2. Check assumptions
3. Choose type
4. Interpret output
5. Develop a regression equation (if needed)

Summary: Linear regression

1. Best-fitting straight line for a scatterplot of two variables
2. Y = bX + a + e
   1. Predictor (X; IV)
   2. Outcome (Y; DV)
3. Least squares criterion
4. Residuals are the vertical distance between actual and predicted values

Summary: MLR assumptions

1. Level of measurement
2. Sample size
3. Normality
4. Linearity
5. Homoscedasticity
6. Collinearity
7. Multivariate outliers
8. Residuals should be normally distributed

Summary: Level of measurement and dummy coding

1. Levels of measurement
   1. DV = Continuous
   2. IV = Continuous or dichotomous
2. Dummy coding
   1. Convert complex variable into series of dichotomous IVs

Summary: MLR types

1. Standard
2. Hierarchical
3. Stepwise / Forward / Backward

Summary: MLR output

1. Overall fit
   1. R, R^2, Adjusted R^2
   2. F, p
2. Coefficients
   1. Relation between each IV and the DV, adjusted for the other IVs
   2. B, β, t, p, and r_p
3. Regression equation (if useful)

$$Y = b_1x_1 + b_2x_2 + ... + b_ix_i + a + e$$

Practice quiz

MLR I Quiz – Practice question 1

A linear regression analysis produces the equation Y = 0.4X + 3. This indicates that:

(a) When Y = 0.4, X = 3

(b) When Y = 0, X = 3

(c) When X = 3, Y = 0.4

(d) When X = 0, Y = 3

(e) None of the above

MLR I Quiz – Practice question 2

Multiple linear regression is a ________ type of statistical analysis.

(a) univariate

(b) bivariate

(c) multivariate

MLR I Quiz – Practice question 3

The following types of data can be used in MLR (choose all that apply):

(a) Interval or higher DV

(b) Interval or higher IVs

(c) Dichotomous Ivs

(d) All of the above

(e) None of the above

MLR I Quiz – Practice question 4

In MLR, the square of the multiple correlation coefficient, R^2, is called the:

(a) Coefficient of determination

(b) Variance

(c) Covariance

(d) Cross-product

(e) Big *R*

MLR I Quiz – Practice question 5

In MLR, a residual is the difference between the predicted Y and actual Y values.

(a) True

(b) False

Next lecture

- Review of MLR I
- Semi-partial correlations
- Residual analysis
- Interactions
- Analysis of change

References

Howell, D. C. (2004). Chapter 9: Regression. In D. C. Howell.. *Fundamental statistics for the behavioral sciences* (5th ed.) (pp. 203-235). Belmont, CA: Wadsworth.

Howitt, D. & Cramer, D. (2011). *Introduction to statistics in psychology* (5th ed.). Harlow, UK: Pearson.

Kliewer, W., Lepore, S.J., Oskin, D., & Johnson, P.D. (1998). The role of social and cognitive processes in children's adjustment to community violence. *Journal of Consulting and Clinical Psychology, 66*, 199-209.

Landwehr, J.M. & Watkins, A.E. (1987) *Exploring data*: *Teacher's edition*. Palo Alto, CA: Dale Seymour Publications.

Tabachnick, B. G., & Fidell, L. S. (2013) (6th ed. - International ed.). Multiple regression [includes example write-ups]. In *Using multivariate statistics* (pp. 117-170). Boston, MA: Allyn and Bacon.

Vemuri, A. W., & Constanza, R. (2006). The role of human, social, built, and natural capital in explaining life satisfaction at the country level: Toward a National Well-Being Index (NWI). *Ecological Economics*, *58*(1), 119-133.

Open Office Impress

- This presentation was made using Open Office Impress.
- Free and open source software.
- http://www.openoffice.org/product/impress.html