# Multiple Linear Regression

## Lecture 8

Survey Research & Design in Psychology

James Neill, 2012

## Overview

1. Readings
2. Correlation (Review)
3. Linear regression
4. LOM & dummy coding
5. Multiple linear regression
   – *R*, coefficients
   – Equation
   – Types
   – Assumptions

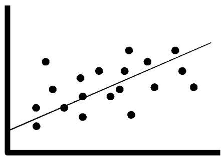


## Readings

As per next lecture

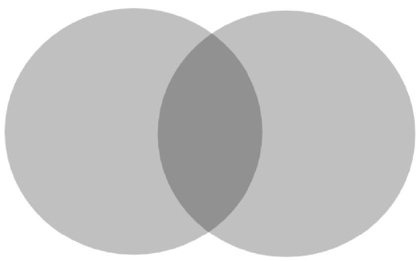
1. Howell (2009). Correlation & regression [Ch 9]
2. Howell (2009). Multiple regression [Ch 15; not 15.14 Logistic Regression]
3. Tabachnick & Fidell (2001). Standard & hierarchical regression in SPSS (includes example write-ups) [Alternative chapter from eReserve]

## Correlation (Review)



*Linear relation between two variables*

## Correlation is shared variance

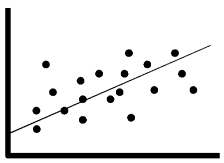


Venn diagrams are helpful for depicting relations between variables.



## Linear correlation

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



- Correlation doesn't provide a prediction equation.

## Correlation – Key points

- Covariance = sum of cross-products
- Correlation = standardised sum of cross-products, 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

7

## Purposes of correlational statistics

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

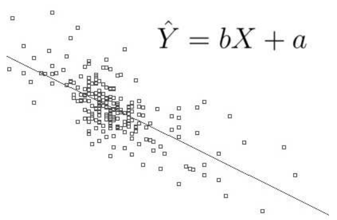
<u>Explanatory - Regression</u>
e.g., hours of study → academic grades

<u>Predictive - Regression</u>
e.g., demographics → life expectancy

8

## Linear regression

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



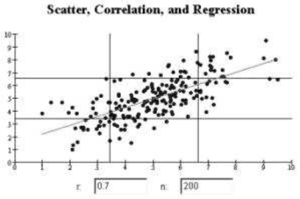
*Explains and predicts a dependent variable (DV) based on linear relations with an independent variable (IV)*

## What is linear regression (LR)?

LR involves:

- one predictor (IV) and
- one outcome (DV)

LR explains a bivariate relationship using a straight line fitted to the data.

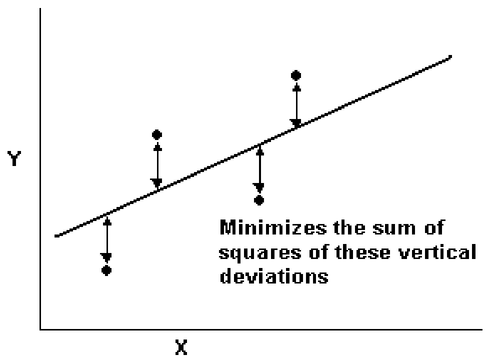



## What is linear regression (LR)?

- An extension of correlation
- Analysis of linear relations(s) between an IV and a DV.
- Calculates the extent to which the DV changes when the IV changes.
- Used to help understand possible causal effects of one variable on another.



## Least Squares Criterion



## Levels of measurement and dummy coding

## Regression: Levels of measurement

- **DV = Continuous**
  (Interval or Ratio)

- **IV = Continuous or Dichotomous**
  (may need to create dummy variables)



## Dummy variables

- To "dummy code" is to convert a more complex variable into dichotomous variables (i.e., 0 or 1)
- Dummy variables are dichotomous variables created from a variable with a higher level of measurement.

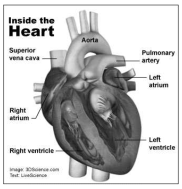
## Dummy variables – Example

- Religion
  (1 = Christian; 2 = Muslim; 3 = Atheist)
  can't be an IV in regression
  (a linear correlation 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.
- <u>More information</u> (Wikiversity)



## 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**

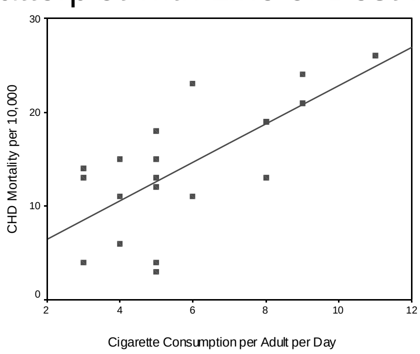
| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Cig. | 11 | 9 | 9 | 9 | 8 | 8 | 8 | 6 | 6 | 5 | 5 |
| CHD | 26 | 21 | 24 | 21 | 19 | 13 | 19 | 11 | 23 | 15 | 13 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 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





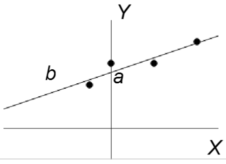

# Linear regression equation
(without error)

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

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 – Example: Equation

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

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

Co-efficients:

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



## 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



## Multiple linear regression – Example - Test for overall significance

- Sig. test of $R^2$ given by ANOVA table

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?

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

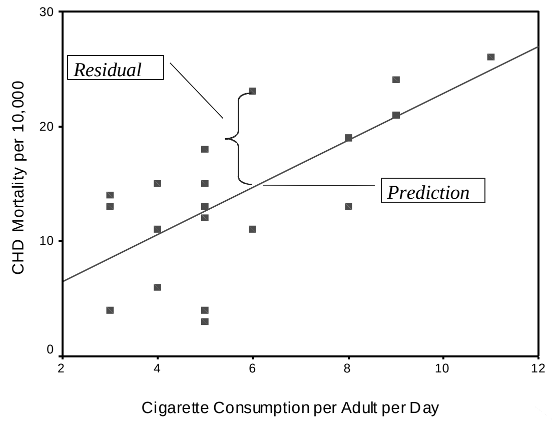
$$\hat{Y} = 2.04 * 6 + 2.37 = 14.61$$

- We predict 14.61 / 10,000 people in that country will die of coronary heart disease.



## **Linear regression - Example: Accuracy of prediction - Residual**

- Finnish smokers smoke 6 cigarettes/adult/day
- We predict 14.61 deaths /10,000
- They actually have 23 deaths / 10,000
- Our error (“residual”) = 23 - 14.61 = 8.39




# 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.



# Hypothesis testing

Null hypotheses ($H_0$):

- $a = 0$
- $b$ = 0
- population correlation (ρ) = 0

## Linear regression – Example: Testing slope and intercept

Coefficients[a]

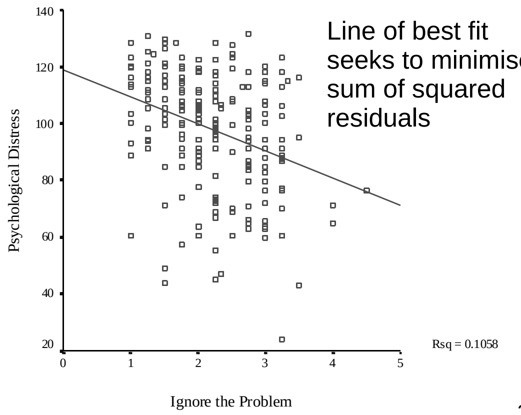
| | | Unstandardized Coefficients | | Standardized Coefficients | | |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | t | Sig. |
| *a* | (Constant) | 2.37 | 2.941 | | .80 | .43 |
| *b* | Cigarette Consumption per Adult per Day | 2.04 | .461 | .713 | 4.4 | .00 |

a. Dependent Variable: CHD Mortality per 10,000

## Linear regression - Example

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






Line of best fit seeks to minimise sum of squared residuals

## 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

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



## 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

It is unlikely that the population relationship between Ignoring Problems (IP) and Psychological Distress (PD) is 0%.



## Linear regression - Example

**Coefficients[a]**

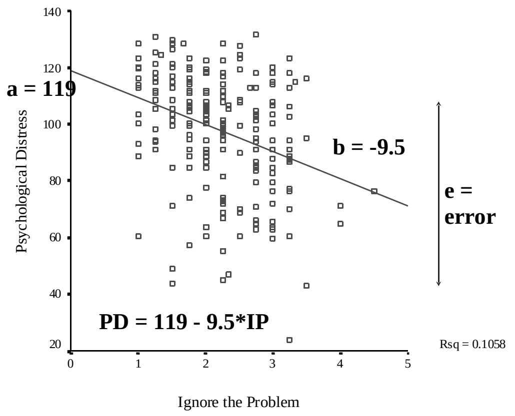
| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | 118.897 | 4.351 | | 27.327 | .000 |
| | IGNO2 ACS Time 2 - 11. Ignore | -9.505 | 1.872 | -.325 | -5.078 | .000 |

a. Dependent Variable: GWB2NEG

There is a sig. *a* or constant (Y-intercept).

IP is a significant predictor of PD

PD = 119 -9.5*Ignore



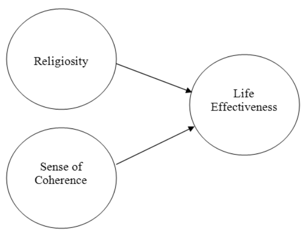

37

## Linear regression summary

- Linear regression is for *explaining* or *predicting* the linear relationship between two variables
- Y = *b*x + a + e
- Y hat = *b*x + a
  (*b* is the slope; *a* is the *Y*-intercept)

38

## Multiple Linear Regression




*Linear relations between two or more IVs and a single DV*

## 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)?
  - e.g., exercise and cholesterol?
- Single predictor: LR
  Multiple predictors: MLR

40

## Linear regression summary

Linear Regression

X Y

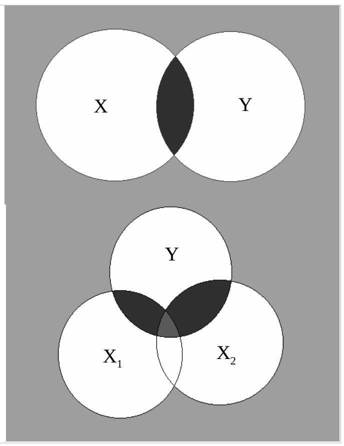
Multiple Linear Regression

$X_1$
$X_2$
$X_3$ Y
$X_4$
$X_5$

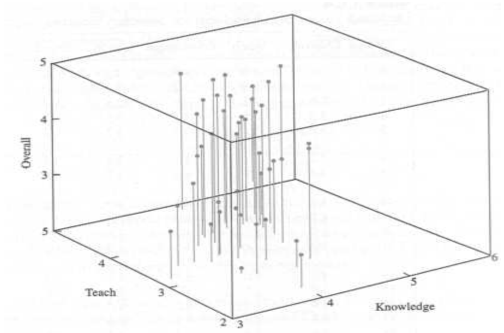
41

Correlation
Regression

Partial Correlation
MLR



## 3-way scatterplot




43

## What is multiple linear regression (MLR)?

- Use of several IVs to predict a DV
- Provides a measure of overall fit (*R*)
- Makes adjustments for inter-relationships among predictors
  - –e.g. IVs = height, gender DV = weight
- Weights each predictor (IV)

**44**

## MLR – Example Research question 1

Do these IVs:

- # of cigarettes / day ($IV_1$)
- exercise ($IV_2$) and
- cholesterol ($IV_3$)

predict

- CHD mortality (DV)?

Cigarettes
Exercise
Cholesterol

CHD Mortality

**45**

## MLR – Example
## Research question 2

To what extent do personality factors (IVs) predict income (DV) over a lifetime?

Extraversion
Neuroticism
Psychoticism

Income



## MLR - Example
## Research question 3

"Does the number of years of psychological study (IV1) and the number of years of counseling experience (IV2) predict clinical psychologists' effectiveness in treating mental illness (DV)?"

Study
Experience

Effectiveness



## MLR - Example
## Your example

Generate your own MLR research question based on some of the following variables:

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

## Regression equation

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

- $Y$ = 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)



## Multiple correlation coefficient (*R*)

- "Big R" (capitalise, i.e., *R*)
- 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 interpretation of $R^2$

- .00 = no linear relationship
- $R^2$ = .10 = small ($R$ ~ .3)
- $R^2$ = .25 = moderate ($R$ ~ .5)
- $R^2$ = .50 = strong ($R$ ~ .7)
- $R^2$ = 1.00 = perfect linear relationship

$R^2$ ~ .30 is good for social sciences



## Adjusted $R^2$

- 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 $R^2$



## Regression coefficients

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

- Y-intercept (*a*)
- Slopes (*b*):
  - Unstandardised
  - Standardised
- Slopes are the weighted loading of IV, 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 one another unless all IVs are measured on the same scale



## Standardised regression coefficients

- Beta (*b* or β) = standardised regression coefficient
- Used for comparing the relative strength of predictors
- β = *r* in LR but this is only true in MLR when the IVs are uncorrelated.



## Relative importance of IVs

- Which IVs are the most important?
- Compare the standardised regression coefficients (β's)

# Multiple linear regression - Example

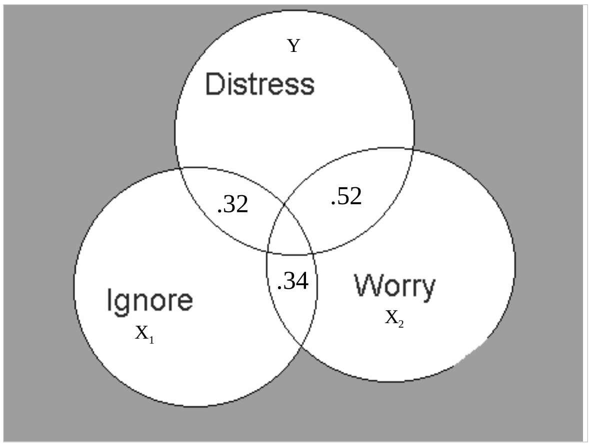
“Does ‘ignoring problems’ ($IV_1$) and ‘worrying’ ($IV_2$) predict ‘psychological distress’ (DV)”



**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 |





# 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



# 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



# 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

## Multiple linear regression - Example – Prediction equations

Linear Regression

Psych. Distress = 119 - 9.50*Ignore

$R^2$ = .11

**Multiple Linear Regression**

Psych. Distress = 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)

Internalising behaviour problems

e.g., withdrawing, anxiety, inhibited, and depressed behaviours

## Multiple linear regression – Example - Study

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

67

## 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 symptoms) – measured using the Child Behavior Checklist (CBCL)

68

Correlations

Pearson Correlation

| | 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 amongst the IVs

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

70

## Multiple linear regression –
### Example - Test for overall significance

- Shows if there is a linear relationship between all of the X variables taken together and Y
- Hypothesis:

$H_0$: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$
(No linear relationships)

$H_1$: At least one $\beta_i \neq 0$
(At least one independent variable effects Y)

71

## Test for significance: Individual variables

Shows if there is a linear relationship between each variable $X_i$ and $Y$.

Hypotheses:

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

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

72

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.

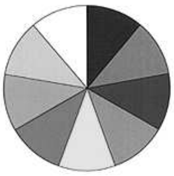
$$\hat{Y} = .038 * Wit + .273 * Stress - .074 * SocSupp + 0.477$$
$$= .038(20) + .273(5) - .074(35) + 0.477$$
$$= .012$$

76

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)



77

## 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)

78

Table 1
Bivariate correlations between variables

| | | Average life satisfaction | HDI | Log ESP/km² in |
|---|---|---|---|---|
| Average life satisfaction | Pearson cor. | 1 | | |
| | Significance | | | |
| HDI | Pearson cor. | .463 | 1 | |
| | Significance | .000 | | |
| Log ESP/km² 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.

79

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/km² 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)

80

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/km² index | 2.453 | .739 | .257 | 3.319 | .002 |

Sample size of the regression model was 50.

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

81

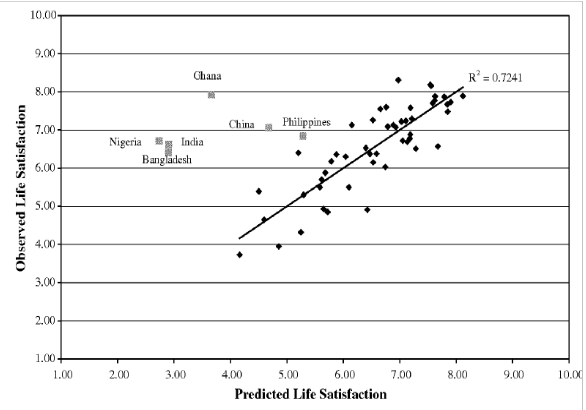



Fig. 2. Observed versus predicted life satisfaction.

## Types of MLR

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



## 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

## 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.



## 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



## 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 the smallest no. of variables.
- Redundant predictors removed.
- Computer-driven – controversial



## Which method?

- Standard: To assess impact of all IVs simultaneously
- Hierarchical: To test specific hypotheses derived from theory
- Stepwise: If goal is accurate statistical prediction – computer driven



## Assumptions

- Levels of measurement
  - IVs = metric (interval or ratio) or dichotomous
  - DV = metric (interval or ratio)
- Sample size
  - Ratio of cases to IVs; total *N*:
  - Min. 5:1; > 20 cases total
  - Ideal 20:1; > 100 cases total

## Assumptions

- Linearity
  - Linear relations exist between IVs & DVs
- Homoscedasticity
- Multicollinearity
  - IVs are not overly correlated with one another (e.g., not over .7)
- Residuals are normally distributed

91

## Dealing with outliers

- Extreme cases should be deleted or modified.
- Univariate outliers - detected via initial data screening
- Bivariate outliers – detected via scatterplots
- Multivariate outliers - unusual combination of predictors…

92

## Multivariate outliers

- Can use Mahalanobis' distance or Cook's *D* as a MV outlier screening procedure
- 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.

93

## Multivariate outliers

- e.g., a person who:
  - Is 19 years old
  - Has 3 children
  - Has a post-graduate degree
- Identify & check unusual cases

94

## Multivariate outliers

- Mahalanobis distance (MD)
  - is distributed as $\chi^2$ with df equal to no. of predictors ($\alpha$ = .001)
  - If any cases have a MD greater than critical level → multivariate outlier.
- Cook's *D*
  - If any cases have CD values >1 → multivariate outlier.
- Use one of either MD or CD

95

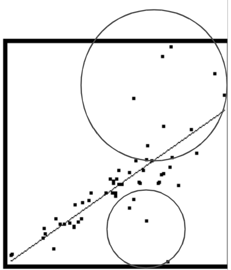
## Normality & homoscedasticity

**Normality**

- If variables are non-normal, there will be heteroscedasticity

**Homoscedasticity**

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



96

## Multicollinearity

- **Multicollinearity** - high correlations (e.g., over .7) between IVs.
- **Singularity** - perfect correlations among IVs.
- Leads to unstable regression coefficients.

97

## Multicollinearity

Detect via:

- **Correlation matrix** - are there large correlations among IVs?
- **Tolerance statistics** - if < .3 then exclude that variable.
- **Variance Inflation Factor (VIF)** - looking for < 3, otherwise exclude variable.

98

## 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

99

## General MLR strategy

1. Check assumptions
2. Choose type
3. Interpret the output
4. Develop a regression equation (if needed)

100

## 1. Check assumptions

- Levels of measurement
- Sample size
- Linearity
- Homoscedasticity
- Multicollinearity
- Multivariate outliers
- Normally distributed residuals

101

## 2. Choose type

- Standard
- Hierarchical
- Forward
- Backward
- Stepwise

102

## 3. Interpret the results

- Relations between *X* predictors (*r*)
- Amount of *Y* explained (*R*, $R^2$, Adjusted $R^2$, the statistical sig. of *R*)
  - Changes in $R^2$ and *F* change (if hierarchical)
- Coefficients for IVs - Standardised and unstandardised regression coefficients for IVs in each model (*b*, *B*).



## 4. Regression equation

- MLR is usually for explanation, sometimes prediction
- If useful, develop a regression equation for the final model.
- Interpret constant and slopes.



## Next lecture

- Review of MLR I
- 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.

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.

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