

Survey Methods & Design in  
Psychology

Lecture 8  
Multiple Linear Regression – I  
(2007)

Lecturer: James Neill

---

---

---

---

---

---

---

---

Overview

- Review of correlation
- Purposes of regression
- Terminology
- Venn diagrams
- Linear regression
- Summary

---

---

---

---

---

---

---

---

Readings

- Howell (2004) – Fundamentals - Regression (Ch10)
- Howell (2007) – Methods - Correlation & Regression (Ch 9)
- Francis (2004) – Relationships Between Metric Variables - Section 3.1
- Tabachnick & Fidell (2001) – Standard & hierarchical regression in SPSS (example write-ups) – from E-reserve
- Optional: Online readings (see website)

---

---

---

---

---

---

---

---

## Purposes of Correlational Statistics

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

---

---

---

---

---

---

---

---

## Purposes of Regression

- **Explanatory**  
- e.g., study habits -> academic performance
- **Predictive**  
- e.g., demographics -> life expectancy

---

---

---

---

---

---

---

---

## Review of correlation ( $r$ )

- Shared variance



---

---

---

---

---

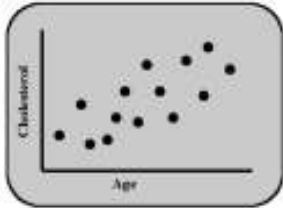
---

---

---

## Review of correlation ( $r$ )

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



---

---

---

---

---

---

---

---

## Review of correlation ( $r$ )

- Covariance is the sum of cross-products
- Correlation is the 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

---

---

---

---

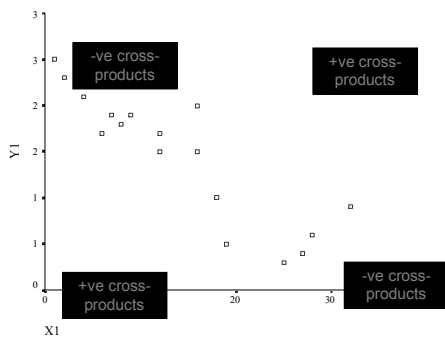
---

---

---

---

## Review of correlation ( $r$ )



---

---

---

---

---

---

---

---

## What is regression analysis?

- An extension of correlation
- A way of measuring the relationship between two or more variables.
- Used to calculate the extent to which one variable changes (DV) when other variable(s) change (IV(s)).
- Used to help understand possible causal effects of one variable on another.

---

---

---

---

---

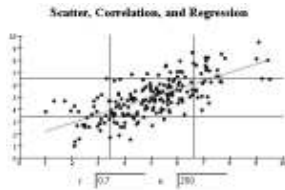
---

---

---

## What is linear regression (LR)?

- Involves:
  - one predictor (IV) and
  - one outcome (DV)
- Explains a relationship using a straight line fit to the data.



---

---

---

---

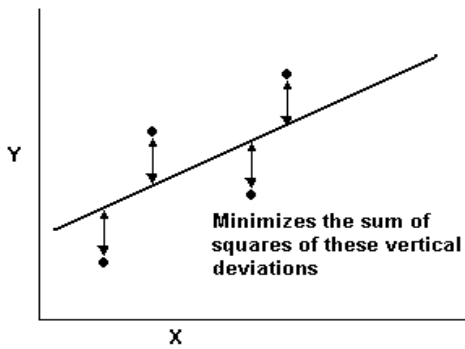
---

---

---

---

## Least Squares Criterion



---

---

---

---

---

---

---

---

## Type of Data

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

---

---

---

---

---

---

---

---

## Dummy variables

- Regression can also use nominal or ordinal IVs, but they must first be made into DUMMY VARIABLES
- Dummy variables have dichotomous coding, i.e. either having ( $x=0$ ) or not having ( $x=1$ ) a characteristic

---

---

---

---

---

---

---

---

## Dummy variables

- Dummy variables are dichotomous variables created from a higher level variable
- E.g., Religion (1 = Christian; 2 = Muslim; 3 = Atheist) -> can be recoded into three dummy variables
  - Christian (0 = no; 1 = yes)
  - Muslim (0 = no; 1 = yes)
  - Atheist (0 = no; 1 = yes)

---

---

---

---

---

---

---

---

Example:  
Cigarettes & coronary heart disease



IV = Cigarette consumption



DV = Coronary Heart Disease (CHD)

---

---

---

---

---

---

---

---

Example:  
Cigarettes & coronary heart disease

- IV = Average no. of cigarettes per adult per day
- DV = Coronary Heart Disease mortality (rate of deaths per 10,000 per year due to CHD)
- Unit of analysis = Country
- How fast does CHD mortality rise with a one unit increase in smoking?

---

---

---

---

---

---

---

---

*Data*

**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 = Coronary Heart Disease Mortality per 10,000 population

---

---

---

---

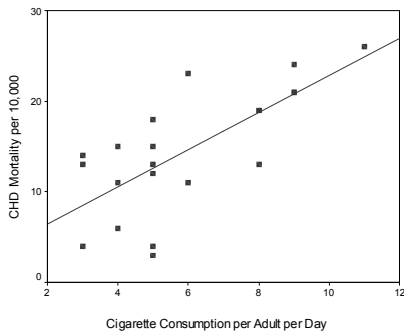
---

---

---

---

*Scatterplot with Line of Best Fit*



---

---

---

---

---

---

---

---

Linear regression formula

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

- $\hat{Y}$  (DV)  
= predicted value of  $Y$   
(annual rate of CHD mortality)
- $X$  (IV)  
= mean # of cigarettes per adult per day per country

---

---

---

---

---

---

---

---

Linear regression formula

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

- “Coefficients” are  $a$  and  $b$
- $b$  = slope  
– Change in predicted  $Y$  for one unit change in  $X$
  - $a$  = intercept  
– value of  $\hat{Y}$  when  $X = 0$

---

---

---

---

---

---

---

---

### Regression calculations

- Slope

$$b = \frac{\text{COV}_{XY}}{s_X^2}$$

- Intercept

$$a = \bar{Y} - b\bar{X}$$

---

---

---

---

---

---

---

---

### Calculations

- $\text{Cov}_{XY} = 11.13$
- $s_X^2 = 2.33^2 = 5.43$
- $b = 11.13/5.43 = 2.04$
- $a = 14.52 - 2.04*5.95 = 2.37$

---

---

---

---

---

---

---

---

### Regression coefficients - SPSS

	Coefficients <sup>a</sup>		t	Sig.
	Unstandardized Coefficients	Standardized Coefficients		
	B	Std. Error	Beta	
(Constant)	2.37	2.941		.80
Cigarette Consumption per Adult per Day	2.04	.461	.713	4.4

a. Dependent Variable: CHD Mortality per 10,000

---

---

---

---

---

---

---

---

### Making a prediction

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

---

---

---

---

---

---

---

---

### Accuracy of prediction

- 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

---

---

---

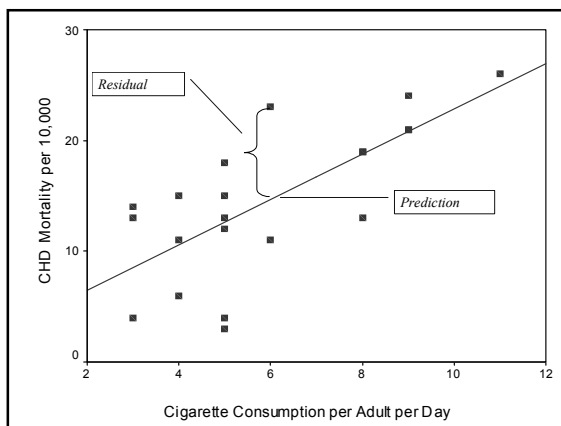
---

---

---

---

---



---

---

---

---

---

---

---

---

## Errors of prediction

- Residual variance
  - The variability of predicted values

$$s_{Y-\hat{Y}}^2 = \frac{\Sigma(Y - \hat{Y})^2}{N - 2}$$

- Standard error of estimate
  - The standard deviation of predicted values

---

---

---

---

---

---

---

---

## Standard error of estimate

$$s_{Y-\hat{Y}} = \sqrt{\frac{\Sigma(Y - \hat{Y})^2}{N - 2}}$$

- A common measure of the accuracy of our predictions
  - We want it to be as small as possible.
  - It has an inverse relationship to  $r^2$  (i.e., when  $r^2$  is large, the standard error of the estimate will be small, and vice-versa)

---

---

---

---

---

---

---

---

## $r^2$ as % of variability which is predictable

- Define Sum of Squares

$$r^2 = \frac{SS_Y - SS_{regression}}{SS_Y}$$

- The remaining error divided by the original error

---

---

---

---

---

---

---

---

### 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
  - $b = 0$
  - $a = 0$
  - population correlation ( $\rho$ ) = 0

---

---

---

---

---

---

---

---

### Testing Slope and Intercept

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

a. Dependent Variable: CHD Mortality per 10,000

---

---

---

---

---

---

---

---

**Example:  
Ignoring Problems & Distress**

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

---

---

---

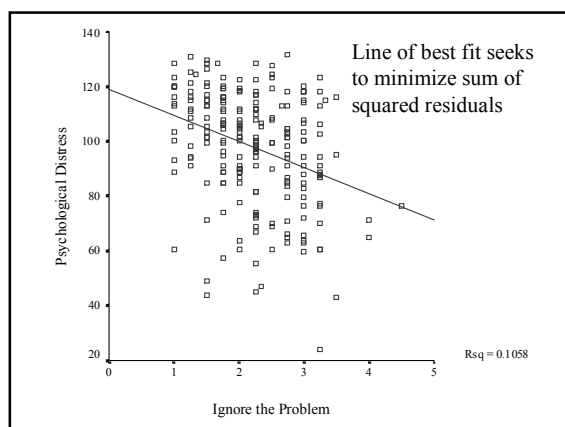
---

---

---

---

---




---

---

---

---

---

---

---

---

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.325 <sup>a</sup>	.106	.102	19.4851

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

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

---

---

---

---

---

---

---

---

**ANOVA<sup>a</sup>**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9789.888	1	9789.888	25.785	.000 <sup>b</sup>
	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 and Psychological Distress is 0%.

---

---

---

---

---

---

---

---

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Std. Error	Beta	t	Sig.
		B	Std. Error				
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 also a significant a or constant (Y-intercept).
- Ignoring Problems is a significant predictor of Psychological Distress

---

---

---

---

---

---

---

---

**Linear Regression Equation**

$$Y = a + bx + e$$

X = predictor value  
Y = predicted value  
a = Y-axis intercept  
b = slope of line of best fit (regression coefficient)  
e = error

$$PD = 119 - 9.5 * Ignore$$


---

---

---

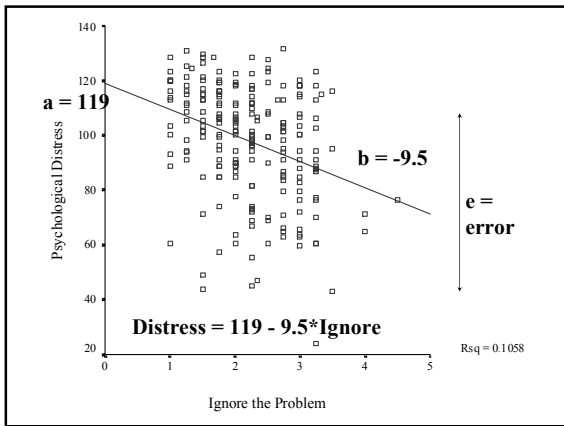
---

---

---

---

---




---

---

---

---

---

---

---

---

### Summary

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

---

---

---

---

---

---

---

---

### 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.

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

---

---

---

---

---

---

---

---

# Survey Methods & Design in Psychology

## Lecture 9 Multiple Linear Regression – II (2007)

Lecturer: James Neill

---

---

---

---

---

---

---

---

### Overview

- LR vs MLR
- MLR with Examples
- Assumptions
- General Strategy
- Summary

---

---

---

---

---

---

---

---

### Readings

- Howell – Fundamental Statistics – MLR (Ch11)
- Howell – Statistical Methods – MLR (Ch15) – but not 15.14 Logistic Regression
- Francis – Intro SPSS Windows – MLR (Section 5.1)
- Tabachnick & Fidell (2001) – Standard & hierarchical regression in SPSS (example write-ups) – E-reserve

---

---

---

---

---

---

---

---

**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 remaining 50% of 'unexplained' variance?
- What about effects of other possible predictors like age, exercise levels and cholesterol?
- The extension of LR to multiple predictors is MLR

---

---

---

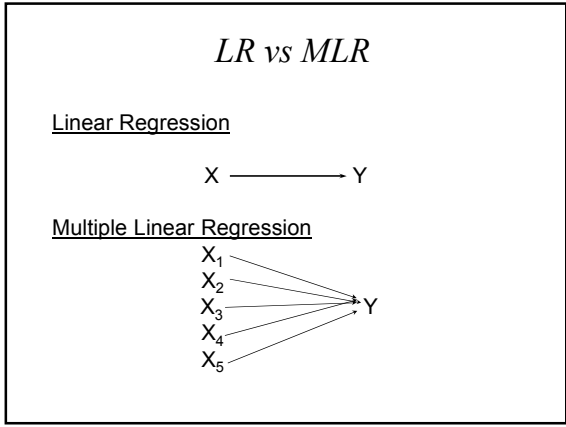
---

---

---

---

---




---

---

---

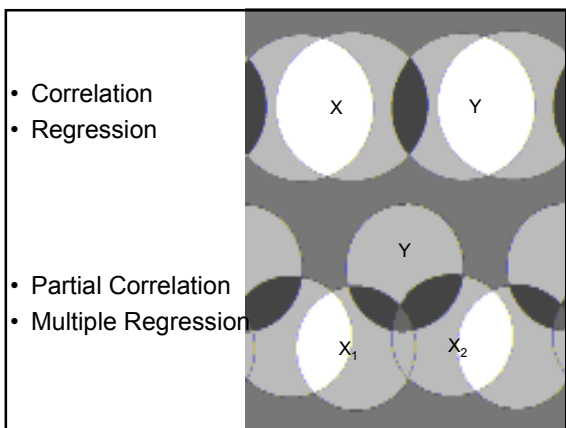
---

---

---

---

---




---

---

---

---

---

---

---

---

### What is 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 = FEV
- Weights each predictor

---

---

---

---

---

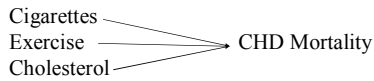
---

---

---

### Research question

- Do number of cigarettes ( $IV_1$ ), exercise ( $IV_2$ ) and cholesterol ( $IV_3$ ) predict CHD mortality (DV)?



---

---

---

---

---

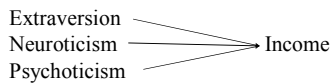
---

---

---

### Research question

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



---

---

---

---

---

---

---

---

### Research question

- “Does the number of years of psychological study (IV<sub>1</sub>) and the number of years of counseling experience (IV<sub>2</sub>) predict clinical psychologists’ effectiveness in treating mental illness (DV)?”

Study ———→ Effectiveness

Experience ———→ Effectiveness

---

---

---

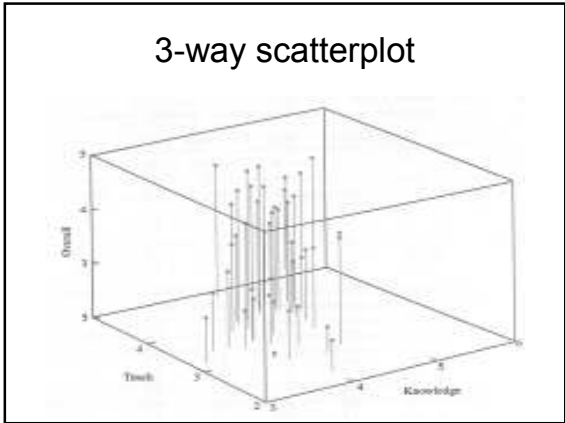
---

---

---

---

---




---

---

---

---

---

---

---

---

### Regression equation

$$Y = b_1X_1 + b_2X_2 + \dots + b_iX_i + a + e$$

- Y = observed dependent scores
- b<sub>i</sub> = unstandardised regression coefficients (the Bs in SPSS)
- X<sub>1</sub> to X<sub>i</sub> = independent variable scores
- a = Y axis intercept
- e = error (residual)

---

---

---

---

---

---

---

---

### Multiple correlation coefficient ( $R$ )

- Directly analogous to simple  $r$
- Always capitalised (i.e.,  $R$ )
- Always positive
- Usually report  $R^2$  instead of  $R$
- $R^2$  = % of variance in DV explained by combined effects of the IVs
- Adjusted  $R^2$  used for estimating explained variance in a population.

---

---

---

---

---

---

---

---

### Coefficient of determination ( $R^2$ )

- Also known as the squared multiple correlation coefficient
- Usually report  $R^2$  instead of  $R$
- $R^2$  = % of variance in DV explained by combined effects of the IVs
- Analogous to  $r^2$

---

---

---

---

---

---

---

---

### Interpretation of $R^2$

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

–  $R^2 = .10$  = small

–  $R^2 = .20$  = medium

–  $R^2 = .30$  = large

---

---

---

---

---

---

---

---

### Adjusted $R^2$

- Adjusted  $R^2$  used for estimating explained variance in a population.
- As number of predictors approaches  $N$ ,  $R^2$  is inflated
- Hence report  $R^2$  and adjusted  $R^2$  particularly for small  $N$  and where results are to be generalised
- If  $N$  is small, take more note of adjusted  $R^2$

---

---

---

---

---

---

---

---

### Regression coefficients

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

- 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
- Can't be compared with one another unless all IVs are measured on the same scale

---

---

---

---

---

---

---

---

### Standardised regression coefficients

- Beta ( $b$  or  $\beta$ ) = standardised regression coefficient
- Used for comparing the relative strength of predictors
- $\beta = r$  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 ( $\beta$ 's)

---

---

---

---

---

---

---

---

### 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

---

---

---

---

---

---

---

---




---

---

---

---

---

---

---

---

Model Summary <sup>a</sup>				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.543 <sup>b</sup>	.295	.288	17.34399

a. Predictors: (Constant), Ignore the Problem, Worry  
b. Dependent Variable: Psychological Distress

---

---

---

---

---

---

---

---

ANOVA <sup>a</sup>						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	27261.12	2	13630.56	45.345	.000 <sup>b</sup>
	Residual	55275.06	217	254.714		
	Total	82536.17	219			

a. Predictors: (Constant), Ignore the Problem, Worry  
b. Dependent Variable: Psychological Distress.

---

---

---

---

---

---

---

---

Coefficients <sup>a</sup>						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
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

---

---

---

---

---

---

---

---

Coefficients <sup>a</sup>			
Model		Correlations	
		Zero-order	Partial
1	Worry	-.521	-.460
	Ignore the Problem	-.325	-.178

a. Dependent Variable: Psychological Distress

---

---

---

---

---

---

---

---




---

---

---

---

---

---

---

---

**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	Lower 95%	Upper 95%
Intercept	562.151009	516.1930837	608.108935
X Variable 1	-5.4365806	-6.169132673	-4.7040285
X Variable 2	-20.012321	-25.11620102	-14.90844

$\beta_1 = -5.44$

The 95% CI:  
 $-6.17 \leq \beta_1 \leq -4.70$

The est. average consumption of oil is reduced by between 4.7 gallons to 6.17 gallons per each increase of 1° F.

---

---

---

---

---

---

---

---

### Confidence interval for the slope

Coefficients <sup>a</sup>				
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 is reduced by between 8.5 and 14.5 units per increase of Worry units.

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

---

---

---

---

---

---

---

---

### Example – Effect of violence, stress, social support on internalizing behavior



---

---

---

---

---

---

---

---

### Study

- Participants were children 8 to 12 years
  - Lived in high-violence areas, USA
- Hypothesis: violence and stress lead to internalising behavior, whereas social support would reduce internalising behaviour.

---

---

---

---

---

---

---

---

## Variables

- Predictors
  - Degree of witnessing violence
  - Measure of life stress
  - Measure of social support
- Outcome
  - Internalising behaviour (e.g., depression, anxiety symptoms)

---

---

---

---

---

---

---

---

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).

---

---

---

---

---

---

---

---

## $R^2$

### Model Summary

	R	Adjusted R Square	Std. Error of the Estimate
a.	.37 <sup>a</sup>	.135	2.2198

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

---

---

---

---

---

---

---

---

### Test for overall significance

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

• Hypothesis:

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

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

---

---

---

---

---

---

---

---

### Test for overall significance

• Significance test of  $R^2$  given by ANOVA table

ANOVA<sup>b</sup>

	Sum of Squares	df	Mean Square	F	Sig.
Regression	454.482	1	454.48	19.59	.00 <sup>a</sup>
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

---

---

---

---

---

---

---

---

### 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)

---

---

---

---

---

---

---

---

Coefficients					
	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

$$\hat{Y} = b_1X_1 + b_2X_2 + b_3X_3 + b_0$$

$$= 0.038Wit + 0.273Stress - 0.074SocSupp + 0.477$$

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

---

---

---

---

---

---

---

---

### Interpretation

$$\hat{Y} = b_1X_1 + b_2X_2 + b_3X_3 + b_0$$

$$= 0.038Wit + 0.273Stress - 0.074SocSupp + 0.477$$

- Slopes for Witness and Stress are positive, but slope for Social Support negative.
- If you had subjects with identical 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.

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

---

---

---

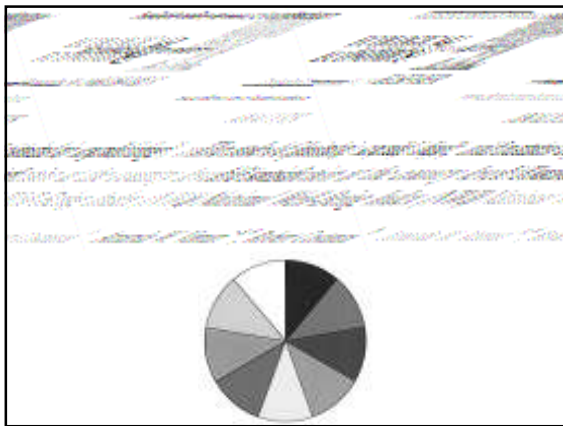
---

---

---

---

---



---

---

---

---

---

---

---

---

## Variables

- IVs:
  - Human & Built Capital (Human Development Index)
  - Natural Capital (Ecosystem services per km2)
  - Social Capital (Press Freedom)
- DV = Life satisfaction
- Units of analysis: Countries (N = 57; mostly developed countries, e.g., in Europe and America)

---

---

---

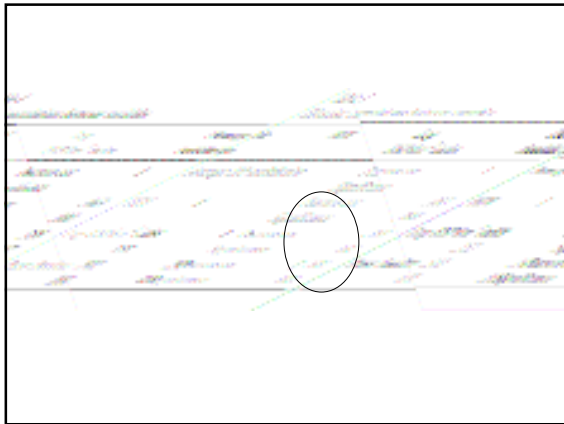
---

---

---

---

---




---

---

---

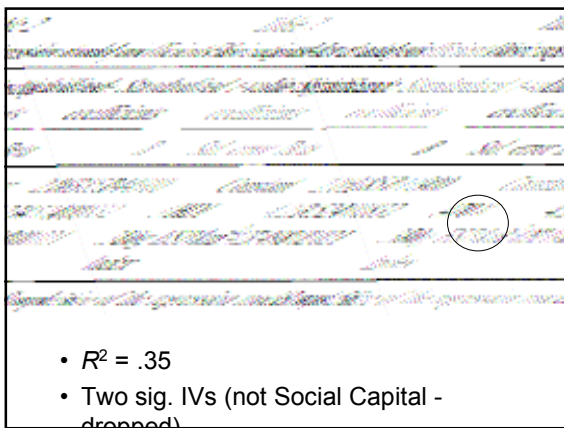
---

---

---

---

---




---

---

---

---

---

---

---

---

Table 4  
Revised regression models coefficients for national-level analysis

Standardized coefficients	t-value	Unstandardized coefficients	Std. Error	Standardized coefficients	Beta	t-value	Significance
		B				t	Sig.
Constant	-2.781	22008.799		Constant	2.781	2008.799	
TIID1	10.038	87500.884		TIID1	10.038	2600.581	.7
TI.log ESP/km <sup>2</sup> index	3.319	45802.739		TI.log ESP/km <sup>2</sup> index	3.319	5002.739	.2

Sample size of the regression model was 50.

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

---

---

---

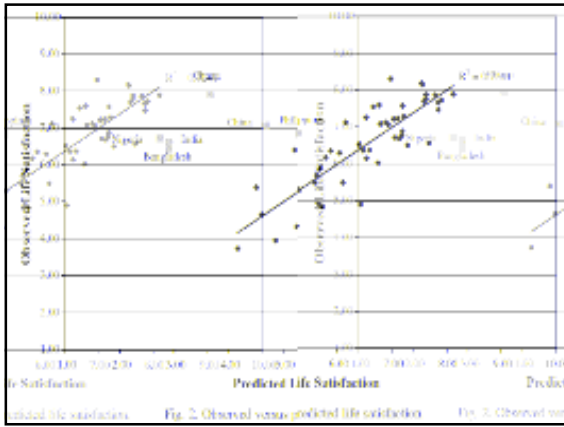
---

---

---

---

---




---

---

---

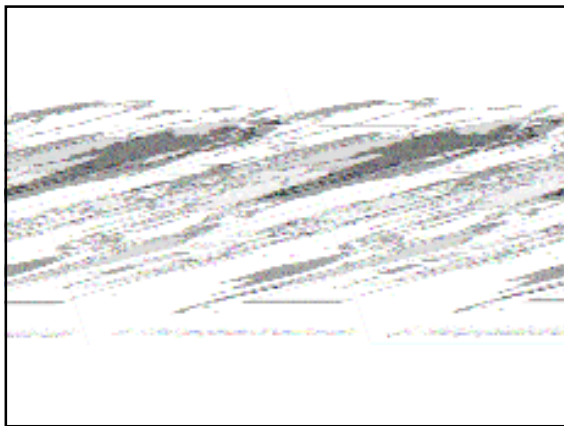
---

---

---

---

---




---

---

---


---

---

---

---

---



### Types of MLR

- Standard or direct (simultaneous)
- Hierarchical or sequential
- Stepwise (forward & backward)
- Forward addition
- Backward elimination

---

---

---

---

---

---

---

---

### Direct or Standard

- All predictor variables are entered together
- 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)

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

---

---

---

---

---

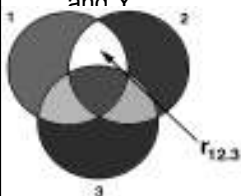
---

---

---

### Partial correlations

- Partial correlation -  $r$  between X and Y after controlling for (partialling out) the influence of a 3rd variable from both X and Y



- e.g., Does:
- # years of marriage predict (IV)
  - marital satisfaction (DV)
  - after # children is controlled for?

---

---

---

---

---

---

---

---

### Forward selection

- The 'best' predictor variables are entered, one by one, if they reach a criteria (e.g.,  $p < .05$ )
- Best predictor =  $x$  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 both forward and 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.
- Means that redundant predictors will be 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

- IVs = metric (interval or ratio) or dichotomous
- DV = metric (interval or ratio)
- Linear relations exist between IVs & DVs
- IVs are not overly correlated with one another (multicollinearity- e.g., over .7)
- Normality enhances solution
- Homoscedasticity
- Ratio of cases to IVs; total *N*:
  - Min. 5:1; > 20 cases total
  - Ideal 20:1; > 100 cases total

---

---

---

---

---

---

---

### 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...

---

---

---

---

---

---

---

## Multivariate outliers

- Can use Mahalanobis distance or Cook's  $D$  as a MV outlier screening procedure
- A case may be within normal range on all variables, but represent a multivariate outlier which unduly influences multivariate test results

e.g., a person who:

- Is 19 years old
- Has 3 children
- Has an undergraduate degree

• Identify & check unusual cases

---

---

---

---

---

---

---

---

## Multivariate outliers

- Mahalanobis distance is distributed as  $\chi^2$  with d. of f. equal to no. of predictors ( $\alpha = .001$ )
- If any cases have a Mahalanobis distance greater than critical level -> multivariate outlier.
- Cook's  $D$  - identifies influential variables, values  $>1$  are considered unusual.

---

---

---

---

---

---

---

---

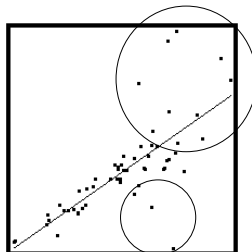
## Normality & homoscedasticity

### Normality

- If non-normality, there will be heteroscedasticity

### Homoscedasticity

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



---

---

---

---

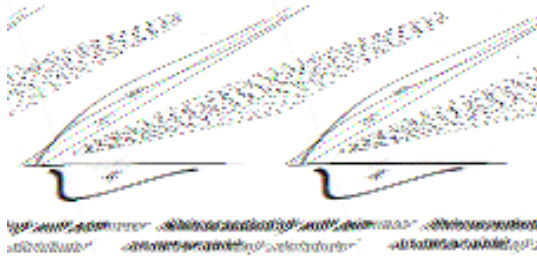
---

---

---

---

## Homoscedasticity



Heteroscedasticity is not fatal  
- but it weakens the analysis.

---

---

---

---

---

---

---

---

## Multicollinearity

- Multicollinearity - high correlations (e.g., over .7) between IVs.
- Singularity - perfect correlations among IVs.
- This 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) - looking for  $< 3$ , otherwise exclude variable.

---

---

---

---

---

---

---

---

## Causality

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

---

---

---

---

---

---

---

---

## General MLR strategy

- Check assumptions
- Conduct MLR – choose type
- Interpret the output
- Develop a regression equation

---

---

---

---

---

---

---

---

## 1. Check assumptions

- Assumptions (Xs not correlated, X-Y linear relations, normal distributions, homoscedasticity)
- Check histograms (normality)
- Check scatterplots (linearity & outliers)
- Check correlation table (linearity & collinearity)
- Check influential outlying cases (mv outliers)
- Check residual plots

---

---

---

---

---

---

---

---

## 2. Conduct MLR

Conduct a multiple linear regression:

- Standard
- Hierarchical
- Stepwise
- Forward
- Backward

---

---

---

---

---

---

---

---

## 3. Interpret the results

Interpret the technical and psychological meaning of the results, based on:

- Overall amount of  $Y$  predicted ( $R$ ,  $R^2$ , Adjusted  $R^2$ , the statistical significance of  $R$ )
- Changes in  $R$  and  $F$  change if hierarchical.
- Coefficients for IVs  
Standardised and unstandardised regression coefficients for IVs in each model ( $b$ ,  $B$ ).
- Relations between  $X$  predictors ( $r$ )
- Zero-order and partial correlations for each IV in each model (draw a Venn diagram)

---

---

---

---

---

---

---

---

## 4. Regression equation

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

---

---

---

---

---

---

---

---

## References

- 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.
- 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, 119-133.

---

---

---

---

---

---

---

---