Topic materials:

Dr Raquel Iniesta
Department of Biostatistics and Health Informatics

Narration and contribution:

Zahra Abdula

Improvements:

Nick Beckley-Hoelscher
Kim Goldsmith
Sabine Landau

Institute of Psychiatry, Psychology and Neuroscience

Module Title: Introduction to Statistics

Session Title: Prediction and Model Fit

Topic title: Multiple regression with several explanatory variables: Adjusting for confounders

Learning Outcomes

After working through this session you should be able to:

- Use the multiple linear regression model as a tool for prediction.
- Use multiple linear regression models to obtain predicted values of dependent variables given a regression equation and values of the independent variables.
- Assess the fit of your model / quality of your prediction model.
- Understand the difference between the standard coefficient of determination R^2 and its adjusted version R^2_{adj}.

Multiple Linear Regression Model: Prediction

We can formulate the model in terms of prediction

The researcher's ultimate goal is to be able to predict the value for a **dependent variable** given **a set of other variables**.

Independent variables can also be known as
Explanatory variables and also as
Predictor variables

A multiple linear regression model can help us find the **factors useful** for the clinician to **predict…**

E.g. weight that a person can reach if he/she does not follow recommendations on habits like diet, water, exercise.

Example: Using the Model to Predict

$y = 72 - 4x_1 - 2x_2 + \varepsilon$

		p-value
Slope for x_1 (β_1)	-4	0.01
Slope for x_2 (β_2)	-2	0.03

Where:

y=weight;
x_1=frequency of exercise per week;
x_2=frequency of vegetables per day;

Use the model to predict the weight for a person who exercises 3 times a week and normally has vegetables 2 times a day, i.e.

x_1=3
x_2=2

y = 72 – (4 x 3) – (2 x 2)
y = 72 – 12 – 4
$\hat{y} = 56\text{kg}$

The model predicts a weight of 56kg for a person who does physical activity 3 times a week and normally has vegetables 2 times a day.

R^2 – The Coefficient of Determination

- The **coefficient of determination**, denoted $\boldsymbol{R^2}$ and pronounced R-Squared, is a statistical measure of how well the regression line/hyperplane approximates the real data points.

- It is also known as a measure of **goodness of fit:** The goodness of fit of any statistical model describes how well it fits a set of observations.

- $\boldsymbol{R^2 = 1 - \frac{SS_{res}}{SS_{tot}}}$ where SS = sum of squares, **res** = residuals (or errors) and **tot** = total

- R^2 ranges from 0 to 1.

  - R^2 of 0 indicates poor fit; the regression line would be perfectly horizontal.

  - R^2 of 1 indicates perfect fit; the regression line/hyperplane fit exactly to all data points.

R^2 – continued

- R^2 measures the fit of the model both in simple and multiple linear regression.

- In **simple linear regression $R^2 = r^2$**, where r is the Pearson correlation.

- In a context of regression where we are assessing **associations between variables**, R^2 is often interpreted as the proportion of the variance in the dependent variable that is "explained" by the independent variables in the model.
  - In our earlier example, this would be the proportion of variance in the weight that is explained by frequency of exercise and hours of free time.
  - R^2 of 0 indicates that none of the variance in y is explained.
  - R^2 of 1 indicates that 100% of the variance in y is explained.

- In a context of **prediction analysis**, R^2 is often interpreted as how well the model will be able to predict values of Y based on observed values for the independent variables x_i; with higher values of R^2 indicating better prediction.

What R^2 Does Not Indicate

R^2 **does not indicate** whether:

- the independent variables are a **cause** of the changes in the dependent variable;
  - (we can only say the variables are associated, not that one causes the other)
- the correct type of regression was used;
- the most appropriate set of independent variables have been chosen;
- there are enough data points to make a solid conclusion.

Adjusted R^2 as a Measure for Model Selection

Adjusted R^2 (denoted R^2_{adj}) is a modified version of R^2 that adjusts for the **number of independent variables *p*** in the model:

$$\boldsymbol{R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1}}$$

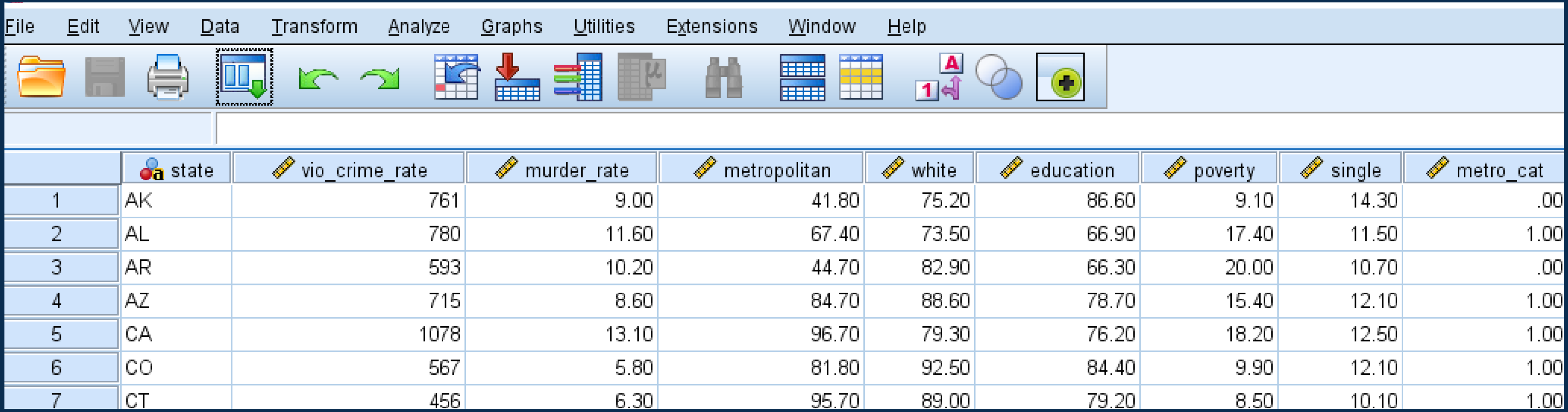
R^2_{adj} takes account of the phenomenon whereby R^2 **increases** every time an **extra independent variable** is added regardless of whether this added variable adds substantially to the explanation of dependent variable variance.

R^2_{adj} increases only when the increase in R^2 (due to the inclusion of a new independent variable) is more than one would expect to see by chance.

$R^2_{\text{adj}}$is considered to be a **better indicator for model selection:** between different models, the one with **higher** R^2_{adj} is the one that better fits the data, and should be selected.

SPSS Slide

Download the data that we are going to use during the lecture. The dataset is the **lecture_7_data.sav**.



	state	vio_crime_rate	murder_rate	metropolitan	white	education	poverty	single	metro_cat
1	AK	761	9.00	41.80	75.20	86.60	9.10	14.30	.00
2	AL	780	11.60	67.40	73.50	66.90	17.40	11.50	1.00
3	AR	593	10.20	44.70	82.90	66.30	20.00	10.70	.00
4	AZ	715	8.60	84.70	88.60	78.70	15.40	12.10	1.00
5	CA	1078	13.10	96.70	79.30	76.20	18.20	12.50	1.00
6	CO	567	5.80	81.80	92.50	84.40	9.90	12.10	1.00
7	CT	456	6.30	95.70	89.00	79.20	8.50	10.10	1.00

The dataset contains data from 51 US states, measuring the crime rates and background measures for each state with respect to their

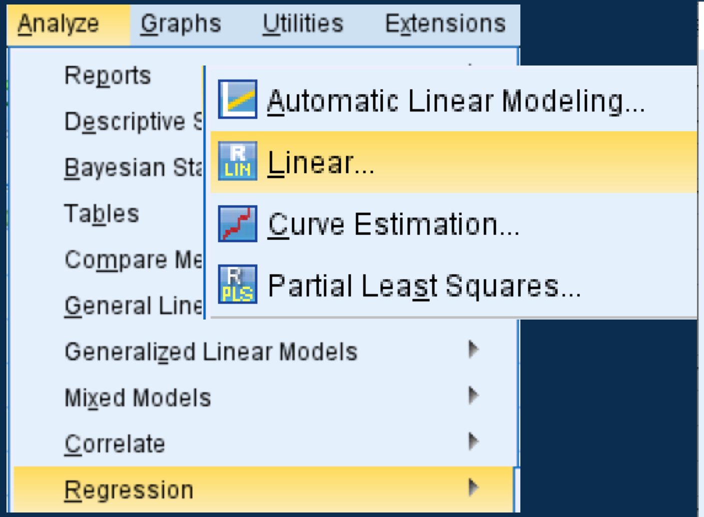
- **violent crime**: per 100,000 population
- **murder**: per 100,000 population
- **poverty**: percent below the poverty line
- **single**: percentage of lone parents

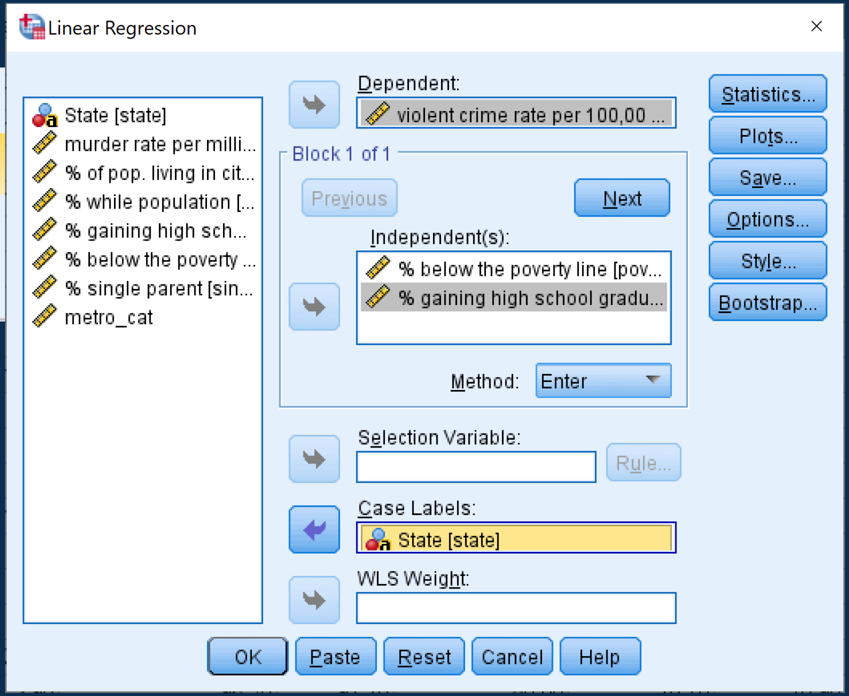
SPSS Slide: 'how to'

Researchers believe, in the population from which our data came, the % below the poverty line and % gaining a high school graduation have and effect on the Violent Crime rate

<u>Step 1</u>) Computing **R^2** for a multiple linear regression model with dependent variable 'crime' and independent variables 'poverty' and 'education' from practical_7_data.sav data

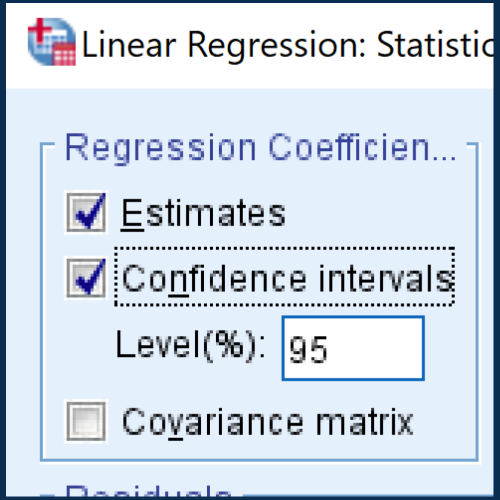
Use **Analyse -> Regression -> Linear**





Put '**crime**' in 'dependent', and '**poverty**' and '**education**' in 'independent'.

Click **Statistics**, select '**Confidence intervals**'.



SPSS Interpretation Slide

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound
1	(Constant)	345.852	1026.638		.337	.738	-1719.478	2411.181
	% below the poverty line	23.927	14.763	.347	1.621	.112	-5.774	53.627
	% gaining high school graduation	-1.502	11.239	-.029	-.134	.894	-24.112	21.109

a. Dependent Variable: violent crime rate per 100,000 population

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.369[a]	.136	.100	280.763

a. Predictors: (Constant), % gaining high school graduation, % below the poverty line

The linear multiple regression model has an R^2_{adj} of 0.100. Poverty and education explained 10.0% of the variance in violent crime.

Knowledge Check – Prediction

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound
1	(Constant)	3.813	.334		11.411	.000	3.158	4.469
	Reading Test Score	-.080	.013	-.184	-6.000	.000	-.106	-.054
	Sex	1.410	.174	.248	8.093	.000	1.068	1.752

a. Dependent Variable: Malaise Score at Age 22

This analysis was done using the Lecture_6_data (NCDS Data) dataset.

It shows the result of fitting a multiple linear regression model with malaise score at age 22 as the dependent variable, with reading and sex as independent variables (sex coded as 0 = male, 1 = female).

Q1: Write out the regression equation, both in terms of *Y* and *X* as well as using the variable names.

Q2: What is the predicted malaise score at age 22 for a female with a reading test score of 11?

Q3: What is the predicted malaise score at age 22 for a male with a reading test score of 28?

Knowledge Check Solutions - Prediction

<u>Q1:</u>

Regression equation:

$y = 3.813 - 0.08(x_1) + 1.410(x_2)$

Malaise score at age 22 = 3.813 – (0.08 x reading score) + (1.410 x sex)

<u>Q2:</u>

Malaise at age 22 = 3.813 – (0.08 x 11) + (1.410 x 1)

Malaise score at age 22 = 4.343

<u>Q3:</u>

Malaise at age 22 = 3.813 – (0.08 x 28) + (1.410 x 0)

Malaise score at age 22 = 1.573

Knowledge Check - R^2

The Psychosis department at the IoPPN is investigating whether quality of life in people diagnosed with schizophrenia depends on a series of demographic and clinical variables. They have asked us to help them choose among different models.

Q4: Which one should they keep as the best model?

Dependent variable:
Quality of Life (QoL) measured with QOLS scale (ranging from 16 to 112)
Independent variables:
Severity of illness, age, gender (1=female), marital status (1=married)

Model	*y*	β_0	Severity β_1 (p-value)	Age β_2 (p-value)	Gender β_3 (p-value)	Marital Status β_4 (p-value)	R^2_{adj}
I	QOLS	50	-3.4 (0.01)	-2.1 (0.10)	Not included	5.1 (0.001)	0.73
II	QOLS	47	Not included	-1.8 (0.07)	1.03 (0.13)	6.2 (0.002)	0.51
III	QOLS	56	-3.1 (0.02)	Not included	Not included	5.3 (0.001)	0.85

Knowledge Check Solutions – R^2

<u>Q4:</u> Which one should they keep as the best model and why?

The best model is the model III with Severity of illness and status as the independent variables. This is because we see from the adjusted R^2 that it explains 85% of the variability in quality of life. If we compare to model I we can see that adding age decreased the adjusted R^2 – this makes sense in combination with the fact that age doesn't seem to be a significant predictor of quality of life. Model II has a lower adjusted R^2 because it is missing the important severity predictor.

Model	y	β_0	Severity β_1 (p val)	Age β_2 (p val)	Gender β_3 (p val)	Marital Status β_4 (p val)	R^2_{adj}
I	QOLS	50	-3.4 (0.01)	-2.1 (0.10)	Not included	5.1 (0.001)	0.73
II	QOLS	47	Not included	-1.8 (0.07)	1.03 (0.13)	6.2 (0.002)	0.51
III	**QOLS**	**56**	**-3.1 (0.02)**	**Not included**	**Not included**	**5.3 (0.001)**	**0.85**

References

Agresti, A., & Finlay, B. (2009).
Statistical Methods for the Social Sciences (4th ed.). New Jersey, NJ: Prentice Hall Inc.

Douglas, C., Montgomery, D. C., Peck, E. A., & Vining, G. G. (2006).
Introduction to Linear Regression Analysis. New York, NY: Wiley.

Thank you

Please contact your module leader or the course lecturer of your programme, or visit the module's forum for any questions you may have.

If you have comments on the materials (spotted typos or missing points) please contact Dr Iniesta:

Raquel Iniesta, PhD
Department of Biostatistics and Health Informatics
IoPPN, King's College London, SE5 8AF, London, UK
raquel.iniesta@kcl.ac.uk

For any other comments or remarks on the module structure, please contact one of the three module leaders of the Biostatistics and Health Informatics department:

Zahra Abdula: **zahra.abdulla@kcl.ac.uk**

Raquel Iniesta: **raquel.iniesta@kcl.ac.uk**

Silia Vitoratou: **silia.vitoratou@kcl.ac.uk**