Data Analysis

Week 3

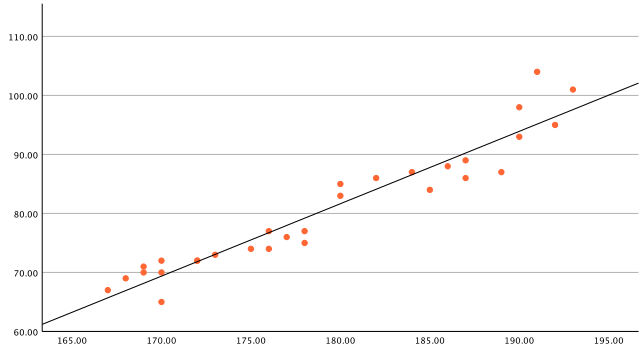
Linear regression

- A simple linear regression assesses the linear relationship between two continuous variables to predict the value of a dependent variable based on the value of independent variable.

- Provides insights into the following:
  - Whether the linear regression between the variables is statistically significant
  - Determine how much of the variation in the dependent variable is explained by the independent variable
  - Understand the direction and magnitude of any relationship
  - Predict values of the dependent variables based on different value(s) of the independent variable

Assumptions of linear regression

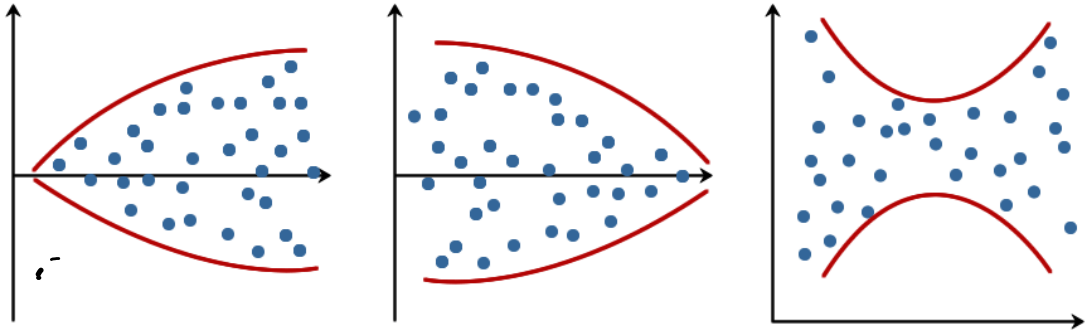
- There needs to be linear relationship between dependent and independent variable.



- There should be independence of observations. This can be checked using the Durbin-Watson statistic (the value should be around 2).

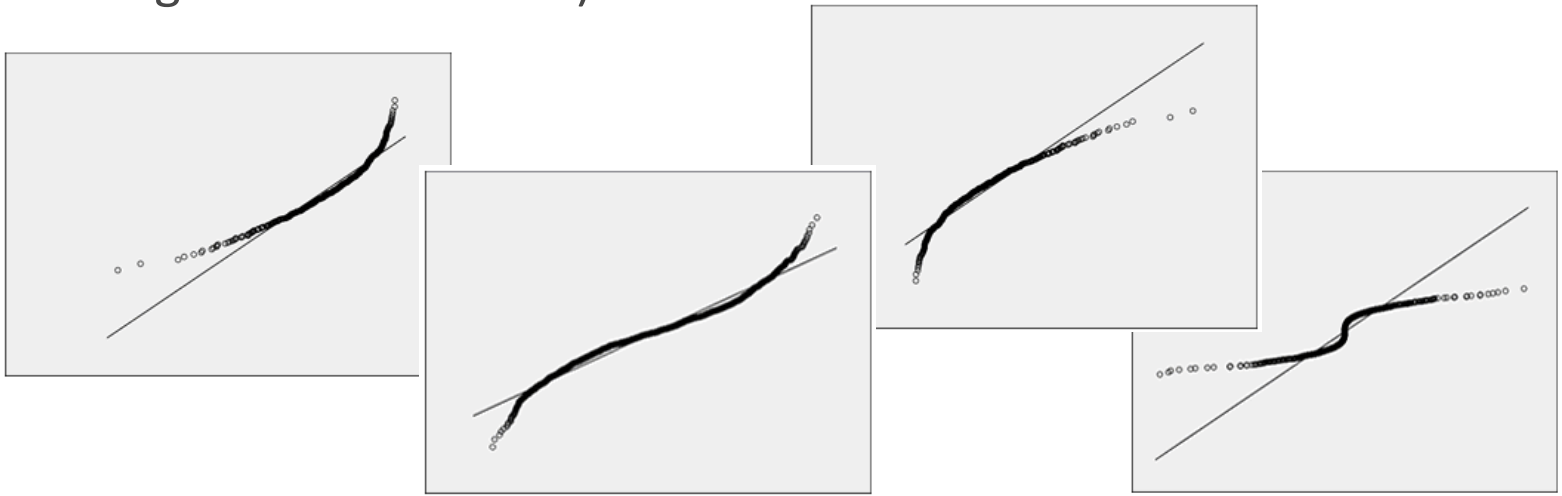
Assumptions of linear regression

- The (residual) data needs to show homoscedasticity. In other words, it needs to be randomly scattered on the plot and not follow any of the patterns below.



Assumptions of linear regression

- There should be no significant outliers.
- Residuals (errors) of the regression line should be approximately normally distributed (close to the regression line without significant deviation).



Linear regression – interpreting results

- SPSS: Analyze ⇢ Regression ⇢ Linear

When interpreting results, consider these three stages:

1. Determining whether the linear regression is a good fit for the data
2. Checking the coefficients of the regression model
3. Making predictions of the dependent variable based on the values of the independent variable

Linear regression – how well the model fits I

- Model summary table shows you information on the proportion of variance explained

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.359[a]	.129	.120	.49100	1.913

a. Predictors: (Constant), Time in minutes spent watching TV

b. Dependent Variable: Cholesterol concentration

- R^2 value represents the proportion of variance in the dependent variable that can be explained by the independent variable (in this example it's 12.9%). This is based on the **sample**.

- Adjusted R^2 provides a value that would be expected in the **population**. This is what you use.

Linear regression – how well the model fits II

- The ANOVA table informs you whether the regression model results in a statistically significantly better prediction of the dependent variable, than if you just used the mean of the dependent variable.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3.470	1	3.470	14.395	.000[b]
	Residual	23.385	97	.241		
	Total	26.856	98			

a. Dependent Variable: Cholesterol concentration

b. Predictors: (Constant), Time in minutes spent watching TV

- *P* value needs to be <.05 for the model to be statistically significant and for the relationship to be considered statistically significant linear relationship.

- Reporting: $F(1,97) = 14.40$, $p < .0005$

Linear regression – interpreting the coefficient

- Slope coefficient represents the change in the dependent variable for a unit of change in the independent variable

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-.944	1.677		-.563	.575	-4.272	2.383
	Time in minutes spent watching TV	.037	.010	.359	3.794	.000	.018	.056

a. Dependent Variable: Cholesterol concentration

- In this example, the slope is .037 which means that for every extra one minute spent watching TV, the cholesterol concentration will increase for .037.

- This only works for the range of values between the minimum and maximum value of the independent variable.

Linear regression – regression equation

- We can make predictions using the constant (intercept) and slope coefficient following the formula:

DEP. VAR = b_0 + (b_1 x IND.VAR)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-.944	1.677		-.563	.575	-4.272	2.383
	Time in minutes spent watching TV	.037	.010	.359	3.794	.000	.018	.056

a. Dependent Variable: Cholesterol concentration

- So, in this example:

CHOLESTEROL C. = -0.944 + (0.037 x TIME_TV)

Linear regression – reporting

- A linear regression established that daily time spent watching TV could statistically significantly predict cholesterol concentration, $F(1, 97) = 14.395$, $p < .001$ and time spent watching TV accounted for 12.9% ($R^2 = 0.129$) of the explained variability in cholesterol concentration. The regression equation: predicted cholesterol concentration = -0.944 + (0.037 x time spent watching tv).

Difference between groups

How would you go about answering the following?

- Is there a difference in salary between male and female doctors?
- Is there a difference in productivity amongst packers at a factory based on the use of background music?

With the current dataset in mind:

- Do smokers have lighter babies?
- Do women over 35 have lighter babies?

Difference between groups

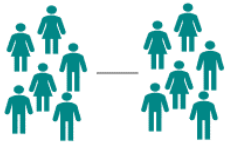
- A t-test tells us how significant the differences between groups are and it tells us if those differences (measured in means) could have happened by chance.

- SPSS: Analyze ⇢ Compare means ⇢ Independent-samples t-test or Paired-samples t-test or One-sample t-test

T-TEST

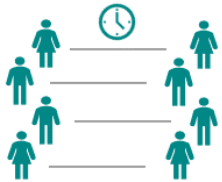
INDEPENDENT-MEANS T-TEST

Used when there are two experimental conditions and different participants were assigned to each condition



DEPENDENT-MEANS T-TEST

Used when there are two experimental conditions and the same participants took part in both conditions of the experiment



ONE SAMPLE T-TEST

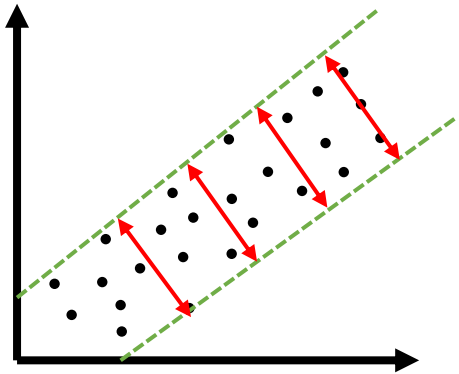
Used when testing a mean of a single group against a known mean

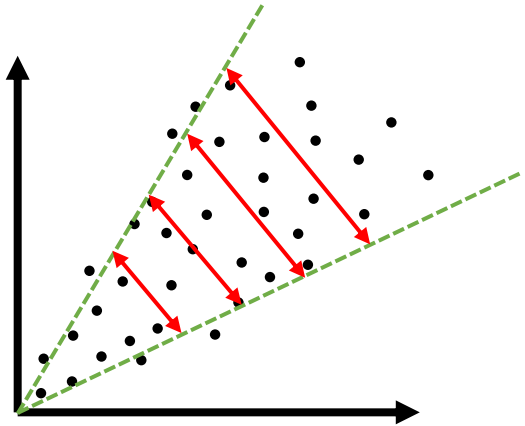
Assumptions of Independent means t-test

- The sampling distribution is (approximately) **normally distributed** (independent t-test is considered “robust” to violations of normality)
  - This can often happen for very large datasets. If the distributions are all similarly skewed, this is not as troublesome. Only strong violations of normality require a non-parametric test.
- Data are measured at least at the interval level
- Scores are **independent** (they come from different people)
- Variances in these populations are roughly equal – **homogeneity of variance** (Levene’s test > 0.05)

Homogeneity of Variance

- Homogeneity of variance is the assumption that the spread of scores is roughly equal in different groups of cases, or more generally that the spread of scores is roughly equal at different points on the predictor variable.





Assumptions of Dependent means t-test

- The sampling distribution of the differences* between scores should be (approximately) **normally distributed** (dependent t-test is considered “robust” to violations of normality)
  - This can often happen for very large datasets. If the distributions are all similarly skewed, this is not as troublesome.
- Data are measured at least at the interval level

*difference is calculated as:

DIFFERENCE = EXPERIMENTAL/POST-VALUES – CONTROL/PRE-VALUES

T – score

- Ratio between the difference between two groups and the difference within the groups
- The larger the t-score, the more difference there is between groups. The smaller the t-score, the more similarity there is between the groups.

  LARGE T-SCORE = GROUPS ARE DIFFERENT

  SMALL T-SCORE = GROUPS ARE SIMILAR

- A T– score of 3 means that the groups are three times as different from each other as they are within each other.

Effect size

- Effect size is an attempt to provide a measure of the practical significance of the result
- Tells us how much impact an intervention had

Reporting a t-test

- An independent-samples t-test was run to determine if there were differences in birth weight for babies from smoking and non-smoking mothers. The birth weight was shown to be higher for non-smoker mothers (*M* = 3.51, SD = 0.518) than for smoker mothers (*M* = 3.13, SD = 0.631), a statistically significant difference, M = 0.38, 95% CI[.01-.74], *t*(42) = 2.093, *p* = .043, d = .58.

References

- Cohen, J. (1988). *Statisticsl power analysis for the behavioral sciences* (2nd ed.). New York: Psychology Press.
- Cook, R.D. & Weisberg, S. (1982). *Residuals and influence in regression*. New York: Chapman & Hall.
- Draper, N.R. & Smith, H. (1998). *Applied regression analysis* (3rd ed.). New York: Wiley.
- Field, A. (2009). *Discovering statistics using SPSS*. London: Sage.
- Sheskin, D.J. (2011). *Handbook of parametric and nonparametrics statistical procedures* (5th ed.). Boca Raton: Chapman & Hall/CRC Press.