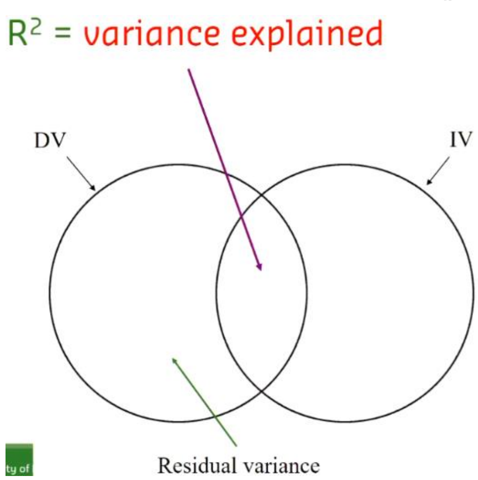
Multiple Regression:
Testing associations for a group of predictors

- Is there a linear relationship between a DV and a set of IVs? Normally we get a better prediction than when we use just 1 IV.
- Sub-questions:
  - Which of the IVs are significant predictors of the DV?
  - What is the order of importance of significant IVs?
- Practical use: Good for analyzing real world problems where experimental studies are not possible (e.g., factors from early development predicting literacy skills in school-age language impaired children)

- Explain what the multiple correlation (R) and squared multiple correlation (R^2) tell us.

Strength of Predictions:
- R – quantifies the strength of linear association between X and Y (R varies between 0 & 1)
- R is always positive, and the direction of the relationship is determined from B sign
- Accuracy of prediction of Y from X depends on the strength of the linear association.
- R^2 Quantifies the proportion of variance in one variable (Y) explained by the other (X)
- 1 – R^2 = residual variance (proportion variance not explained by the linear association)



Cohen's conventions for MR effect size:

- $f^2 = R^2 / (1 - R^2)$
- Small effect: $f^2 = .02$, $R^2 = .0196$, $R = .14$
- Medium effect: $f^2 = .15$, $R^2 = .13$, $R = .36$
- Large effect : $f^2 = .35$, $R^2 = .26$, $R = .50$

Multiple regression:

- Multiple R = the strength of linear association
  - Varies between 0 - 1
- Multiple R^2 = Proportion of variance in DV explained by set of predictors
  - Shared/common + unshared/unique
- F ratio tests H0: all population slopes (Bs) are 0
  - No (linear) relationship between DV & set of IV's

Multiple R value quantifies the strength
If at least one predictor has linear relation with outcome measure, we will get a result of variance explained.

R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination. R-squared = Explained variation / Total variation
R-squared is always between 0 and 100%:
- 0% indicates that the model explains none of the variability of the response data around its mean.
- 100% indicates that the model explains all the variability of the response data around its mean.

- Identify at least 4 important preliminary procedures or checks when undertaking a multiple regression analysis (e.g., what requirements/assumptions need to be satisfied?).

MULTIPLE REGRESSION: Certain conditions do apply – Assumptions:
- DV & IVs have normal distribution
- Linear relationship between DV and IVs
- Homoscedasticity
- Normality of residuals

Practical issues
- >10 times as many participants as variables recommended
- Minimum of 5 n Interval/ratio & dichotomous IV's
- Check for univariate & multivariate (with respect to 2+variables) outliers
- Check for multicolinearity (high correlations between IV's, i.e., > .9)
- Check for singularity (perfect correlation between IV's)
- Check assumptions (look at the standardised residual plots)

- Explain what it means when the F value associated with multiple R is significant.

The **F value** is the ratio of the mean **regression** sum of squares divided by the mean error sum of squares. Its **value** will range from zero to an arbitrarily large number. The **value** of Prob(**F**) is the probability that the null hypothesis for the full model is true (i.e., that all of the **regression** coefficients are zero). The F-test of overall significance indicates whether your linear regression model provides a better fit to the data than a model that contains no independent variables.

In general, an F-test in regression compares the fits of different linear models. Unlike t-tests that can assess only one regression coefficient at a time, the F-test can assess multiple coefficients simultaneously. The F-test of the overall significance is a specific form of the F-test. It compares a model with no predictors to the model that you specify. If the P value for the F-test of overall significance test is less than your significance level, you can reject the null-hypothesis and conclude that your model provides a better fit than the intercept-only model.

- **Null hypothesis**: The fit of the intercept-only model and your model are equal.
- **Alternative hypothesis**: The fit of the intercept-only model is significantly reduced compared to your model.

Model Summary [b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.927[a]	.860	.841	8.6678

a. Predictors: (Constant), Chronological Age (months), Nonverbal Short Scale IQ, Expressive:Formulated Sentences Stnd Score, Receptive: Concepts&Directions Stnd Score, Word Attack Stndrd Score

b. Dependent Variable: Short Scale Reading Stndrd Score

What is the strength of linear relationship between the DV and set of IVs? ← R = 0.927

How much variance in the DV is explained by the set of IVs? ← R^2 = 0.86

Std. Error of Estimate = how much the actual score deviates from predicted score on average

Statistical conclusion: A significant amount of variance in reading scores is explained by the IVs, and we can reject H0 (population slopes of all IVs are zero), F(5, 37) = 45.40, p < .001.