The Coefficient of Determination

- The coefficient of determination R^2 is used in the context of statistical models whose main purpose is the prediction of future outcomes on the basis of other related information. It is the proportion of variability in a data set that is accounted for by the statistical model. It provides a measure of how well future outcomes are likely to be predicted by the model.
- R^2 is a statistic that will give some information about the goodness of fit of a model. In regression, the R^2 coefficient of determination is a statistical measure of how well the regression line approximates the real data points. An R^2 of 1.0 indicates that the regression line perfectly fits the data.
- In the case of simple linear regression, the coefficient of determination is equivalent to the squared value of the Pearson correlation coefficient. (Consider this to be co-incidental, rather than a definition).

The Adjusted Coefficient of Determination

- Adjusted R^2 (often written as and pronounced "R bar squared") is a modification of R^2 that adjusts for the number of predictor variables in a model. Adjusted R^2 is used to compensate for the addition of variables to the model.
- As more independent variables are added to the regression model, unadjusted R^2 will generally increase but there will never be a decrease. This will occur even when the additional variables do little to help explain the dependent variable.
- To compensate for this, adjusted R^2 is corrected for the number of independent variables in the model, increases only if the new term improves the model more than would be expected by chance.
- If too many predictor variables are being used, this will be reflected in a reduced adjusted R^2. The adjusted R^2 can be negative, and will always be less than or equal to R^2.
- The result is an adjusted R^2 than can go up or down depending on whether the addition of another variable adds or does not add to the explanatory power of the model. Adjusted R^2 will always be lower than unadjusted.
- **Important** Adjusted R square is generally considered to be a more accurate goodness-of-fit measure than R square. It has become standard practice to report the adjusted R^2, especially when there are multiple models presented with varying numbers of independent variables.

Information Criterions

We define two types of information criterion: the Akaike Information Criterion (AIC) and the Schwarz's Bayesian Information Criterion (BIC). The Akaike information criterion is a measure of the relative **goodness of fit** of a statistical model.

$$\text{AIC} = 2p - 2\ln(L)$$

- p is the number of predictor variables in the model.
- L is the value of the Likelihood function for the model in question.
- For AIC to be optimal, n must be large compared to p.

An alternative to the AIC is the Schwarz BIC, which additionally takes into account the sample size n.

$$\text{BIC} = p \ln n - 2 \ln(L)$$

When using the AIC (or BIC) for selecting the optimal model, we choose the model for which the AIC (or BIC) value is lowest.

Akaike Information Criterion

- Akaike's information criterionis a measure of the goodness of fit of an estimated statistical model. The AIC was developed by Hirotsugu Akaike under the name of "an information criterion" in 1971.
- The AIC is a ***model selection*** tool i.e. a method of comparing two or more candidate regression models. The AIC methodology attempts to find the model that best explains the data with a minimum of parameters. (i.e. in keeping with the law of parsimony)
- The AIC is calculated using the "likelihood function" and the number of parameters . The likelihood value is generally given in code output, as a complement to the AIC. (*Likelihood function is not on our course*)
- Given a data set, several competing models may be ranked according to their AIC, with the one having the lowest AIC being the best. (Although, a difference in AIC values of less than two is considered negligible).

Model Metrics for Logistic Regression Models

- In order to understand how much variation in the dependent variable can be explained by a logistic regression model (the equivalent of R^2 in multiple regression), you should consult ***Model Summary*** statistics.
- Although there is no close analogous statistic in logistic regression to the coefficient of determination R^2 the Model Summary Table provides some approximations.
- Logistic regression does not have an equivalent to the R-squared that is found in OLS regression; however, many researchers have tried to come up with one.
- The SPSS output table below contains the *Cox & Snell R Square* and *Nagelkerke R Square* values, which are both methods of calculating the explained variation. These values are sometimes referred to as ***pseudo*** R^2 ***values*** (and will have lower values than in multiple regression).
- However, they are interpreted in the same manner, but with more caution. Therefore, the explained variation in the dependent variable based on our model ranges from 24.0% to 33.0%, depending on whether you reference the Cox & Snell R^2 or Nagelkerke R^2 methods, respectively.
- Nagelkerke R^2 is a modification of Cox & Snell R^2, the latter of which cannot achieve a value of 1. For this reason, it is preferable to report the Nagelkerke R^2 value.
- The Nagelkerke modification that does range from 0 to 1 is a more reliable measure of the relationship.
- Nagelkerkes R^2 will normally be higher than the Cox and Snell measure.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	102.088[a]	.240	.330

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Figure 1: SPSS output

- Cox and Snells R-Square attempts to imitate multiple R-Square based on likelihood, but its maximum can be (and usually is) less than 1.0, making it difficult to interpret. Here it is indicating that 55.2% of the variation in the dependent variable is explained by the logistic model.

Pseudo R-squares

- Cox & Snell R Square and Nagelkerke R Square are two measures from the **pseudo R-squares** family of measures.
- There are a wide variety of pseudo-R-square statistics (these are only two of them). Because this statistic does not mean what R-squared means in OLS regression (the proportion of variance explained by the predictors), we suggest interpreting this statistic with great caution.

Cox & Snell R Square

Cox and Snell's R-Square is an attempt to imitate the interpretation of multiple R-Square based on the likelihood, but its maximum can be (and usually is) less than 1.0, making it difficult to interpret. It is part of SPSS output.

Nagelkerke's R-Square

- Nagelkerkes R^2 is part of SPSS output in the Model Summary table and is the most-reported of the R-squared estimates.
- In our case it is 0.737, indicating a moderately strong relationship of 73.7% between the predictors and the prediction.
- Nagelkerke's R-Square is a further modification of the Cox and Snell coefficient to assure that it can vary from 0 to 1. Nagelkerke's R-Square will normally be higher than the Cox and Snell measure. It is part of SPSS output and is the most-reported of the R-squared estimates.