Assessing the Fit of Regression Models

- A well-fitting regression model results in predicted values close to the observed data values. The mean model, which uses the mean for every predicted value, generally would be used if there were no informative predictor variables. The fit of a proposed regression model should therefore be better than the fit of the mean model.
- Three statistics are used in Ordinary Least Squares (OLS) regression to evaluate model fit: R-squared, the overall F-test, and the Root Mean Square Error (RMSE). All three are based on two sums of squares: Sum of Squares Total (SST) and Sum of Squares Error (SSE). SST measures how far the data are from the mean and SSE measures how far the data are from the models predicted values. Different combinations of these two values provide different information about how the regression model compares to the mean model.

R-squared and Adjusted R-squared

The difference between SST and SSE is the improvement in prediction from the regression model, compared to the mean model. Dividing that difference by SST gives R-squared. It is the proportional improvement in prediction from the regression model, compared to the mean model. It indicates the goodness of fit of the model.

R-squared has the useful property that its scale is intuitive: it ranges from zero to one, with zero indicating that the proposed model does not improve prediction over the mean model and one indicating perfect prediction. Improvement in the regression model results in proportional increases in R-squared.

Critique of R-squared Metrics

One pitfall of R-squared is that it can only increase as predictors are added to the regression model. This increase is artificial when predictors are not actually improving the models fit. To remedy this, a related statistic, Adjusted R-squared, incorporates the models degrees of freedom. Adjusted R-squared will decrease as predictors are added if the increase in model fit does not make up for the loss of degrees of freedom. Likewise, it will increase as predictors are added if the increase in model fit is worthwhile. Adjusted R-squared should always be used with models with more than one predictor variable. It is interpreted as the proportion of total variance that is explained by the model.

There are situations in which a high R-squared is not necessary or relevant. When the interest is in the relationship between variables, not in prediction, the R-square is less important. An example is a study on how religiosity affects health outcomes. A good result is a reliable relationship between religiosity and health. No one would expect that religion explains a high percentage of the variation in health, as health is affected by many other factors. Even if the model accounts for other variables known to affect health, such as income and age, an R-squared in the range of 0.10 to 0.15 is reasonable.

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 ts 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 terms 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 (unlikely, but not impossible), 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.
- 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.

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 ts 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 modi cation of R^2 that adjusts for the number of predictor terms 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 re ected in a reduced adjusted R^2.
- The adjusted R^2 can be negative (unlikely, but not impossible), 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.
- Adjusted R square is generally considered to be a more accurate goodness-of- t 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.