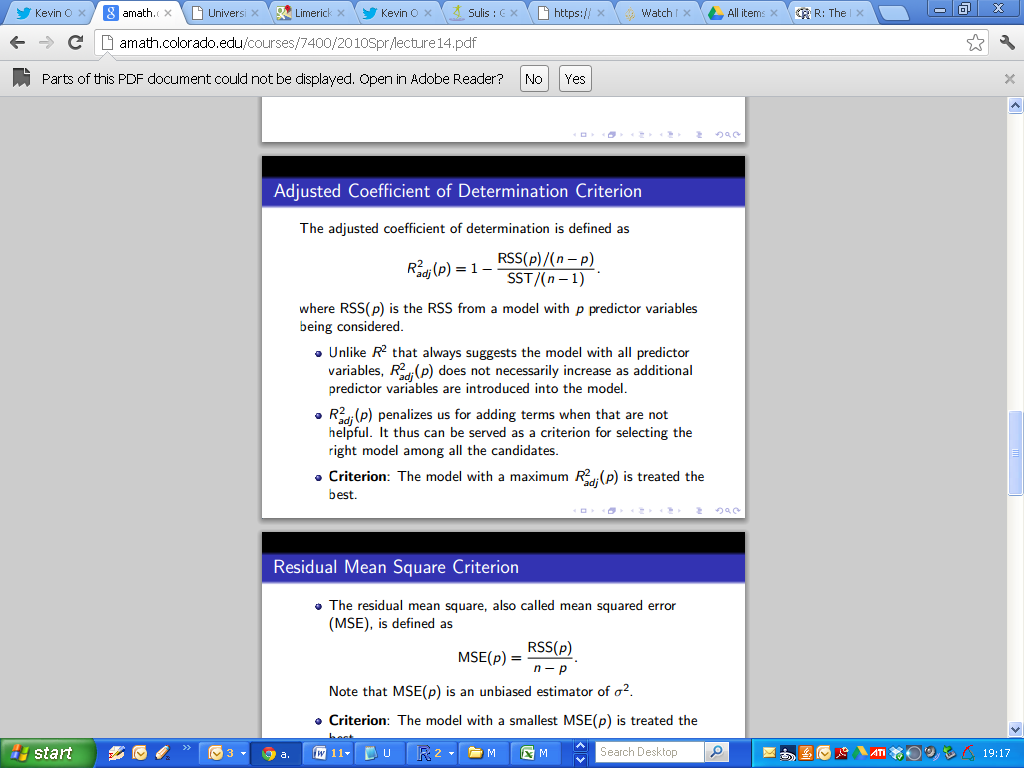
**Variable Selection Procedures**

Variable selection is intended to select the “best" subset of predictor variables. Reasons for performing variable selections are:

* We want to explain the data in the simplest way. Redundant independent variables should be removed.
* [Rule of Thumb: Among several plausible regression models, the smallest model always fits the data best. The so-called “Law of Parsimony”]
* Unnecessary predictor variables will add noise to the (precise) estimation of other quantities that interested us.
* Collinearity is caused by having too many predictor variables trying to do the same job. Removing excess predictors will aid interpretation.

**Adjusted Coefficient of Determination Criterion**

The adjusted coefficient of determination is defined as



where *RSS(p)* is the RSS from a model with p predictor variables being considered.

Unlike R2 that always suggests the model with all predictor variables, R2adj (p) does not necessarily increase as additional predictor variables are introduced into the model.

R2adj (p) penalizes us for adding terms when that are not helpful. It thus can be served as a criterion for selecting the right model among all the candidates.

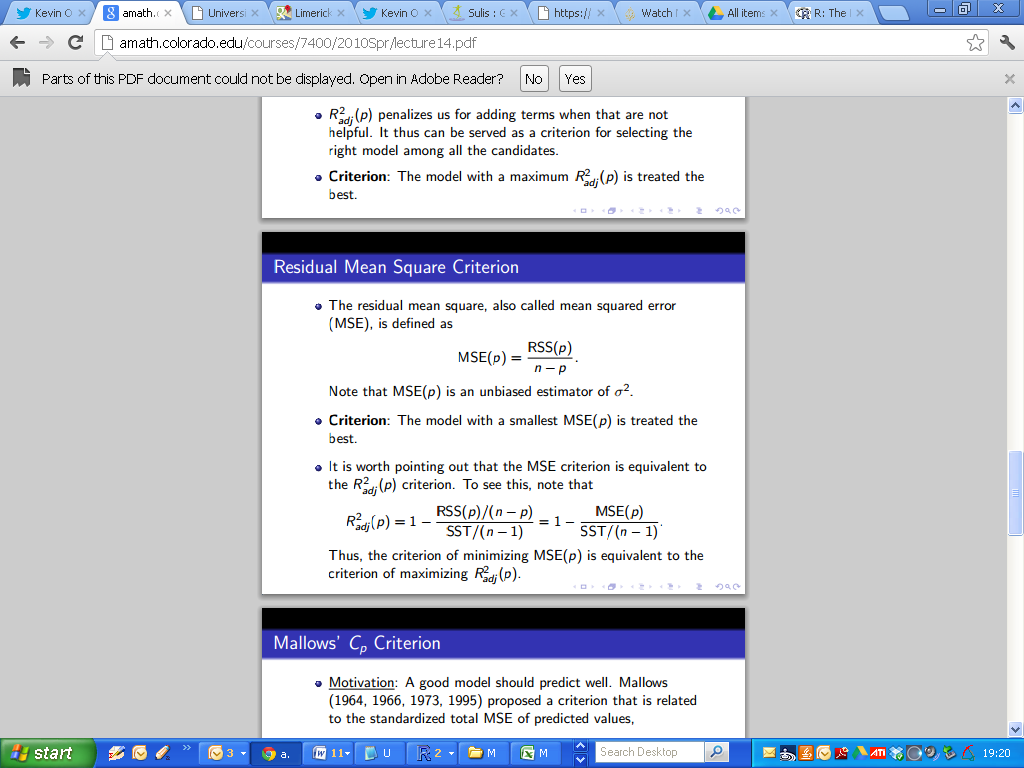
Criterion: The model with a maximum R2adj (p) is treated the best.

R2adj (p) is listed as **Adjusted R-square** in the output of the summary command.

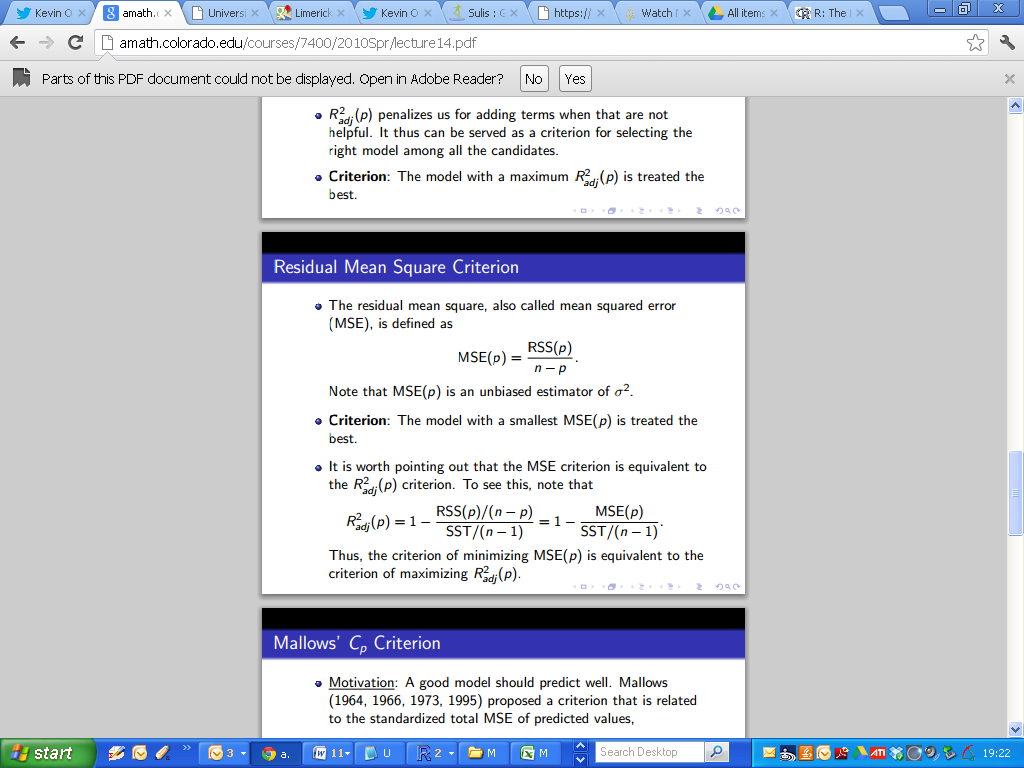
|  |
| --- |
| > summary(FitA)  ……  ……  Residual standard error: 13.82 on 28 degrees of freedom  Multiple R-squared: 0.302, Adjusted R-squared: 0.2771  F-statistic: 12.11 on 1 and 28 DF, p-value: 0.001658 |

**Residual Mean Square Criterion**

The residual mean square, also called mean squared error(MSE), is defined as

****

Criterion: The model with a smallest MSE(p) is treated the best. It is worth pointing out that the MSE criterion is equivalent to the R2adj (p) criterion. To see this, note that



The Residual Mean Square is tabulated in the ANOVA table for a candidate model.

|  |
| --- |
| > anova(FitA)  Analysis of Variance Table  Response: Taste  Df Sum Sq Mean Sq F value Pr(>F)  Acetic 1 2314.1 2314.14 12.114 0.001658 \*\*  Residuals 28 5348.7 191.03  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |

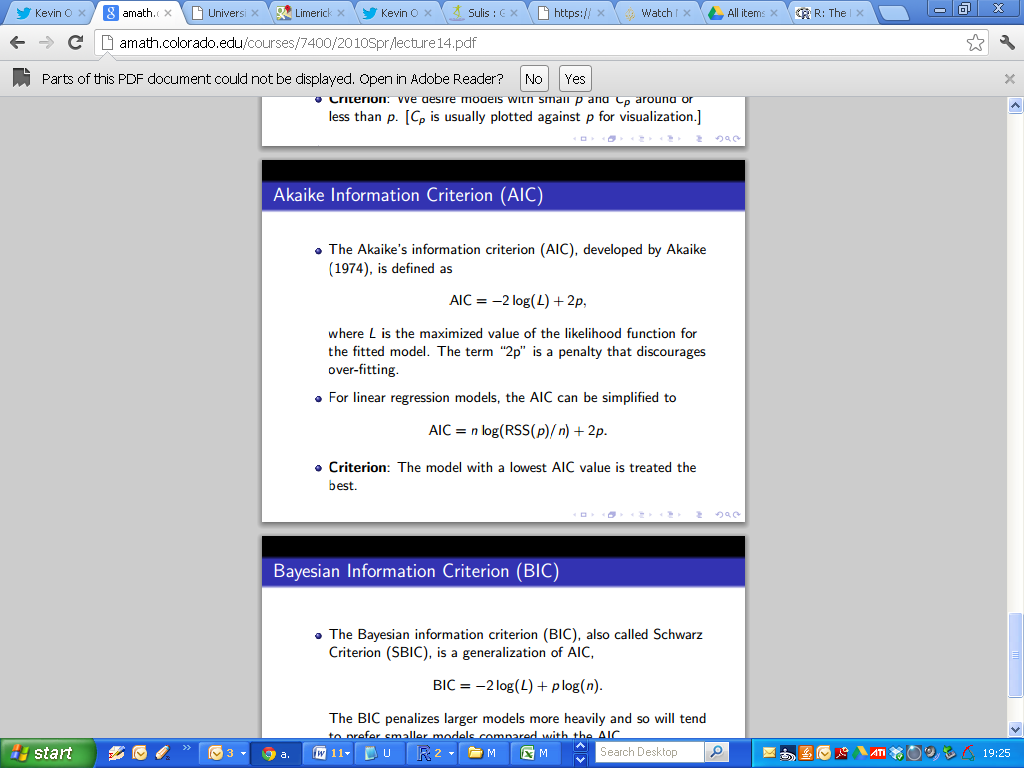
**Akaike's information criterion**

The Akaike's information criterion (AIC), developed by Akaike (1974), is defined as

***AIC = 2 log(L) + 2p;***

where ***L*** is the maximized value of the likelihood function for the fitted model. (Interesting, but not part of course). The term “***2p***" is a penalty that discourages over-fitting.

For linear regression models, the AIC can be simplified to



Criterion: The model with a lowest AIC value is treated the best.

To compute the AIC for a candidate model in R, simply specify the name of the model as an argument to the AIC() function.

|  |
| --- |
| > FitA # From the Cheese Example  Call:  lm(formula = Taste ~ Acetic, data = Cheese)  Coefficients:  (Intercept) Acetic  -61.50 15.65  > AIC(FitA)  [1] 246.6389 |