**Terminology**

**SLR**

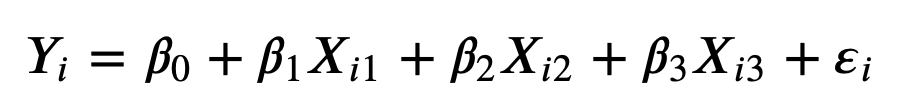
only one quantitative explanatory variable

**MLR**

more than one quantitative explanatory variable

**LM**

with one categorical explanatory variable with 2 or more levels



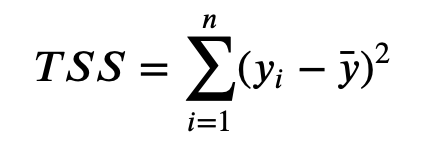
**Goodness of fit**

is our model better than “nothing”?

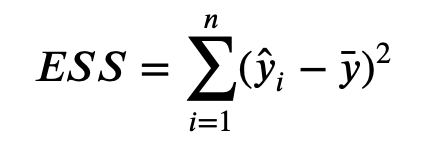
Best predictor of Y is E[Y|x] that can be estimated with the sample mean.

**TSS: Total Sum of Squares** (aka sum of squared errors)

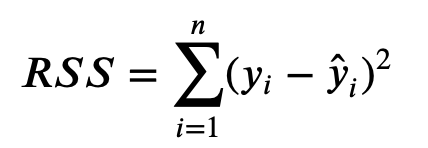
measures the average square distance between each observation and the *total* average

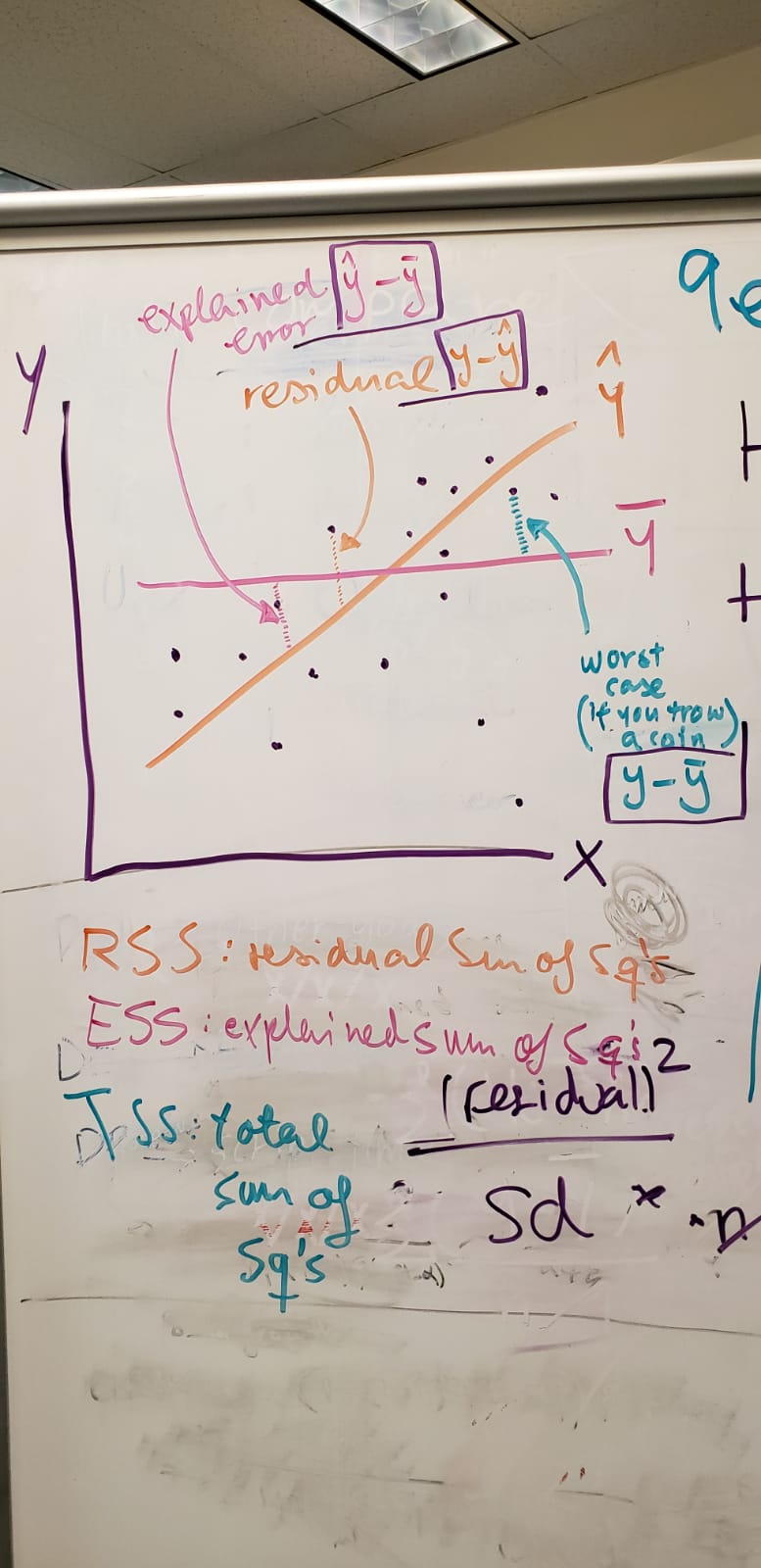
****

**ESS: Explained Sum of Squares**: (aka regression sum of squares) measures the average square distance between each predicted value and the *total* average

****

**RSS: Residual Sum of Squares**: (aka error sum of squares) measures the average square distance of the residuals (distance between each observation and its predicted value)

****



**R2: coefficient of determination**

-Measures the gain in predicting the response using the LM instead of the sample mean, relative to the total variation in the response.

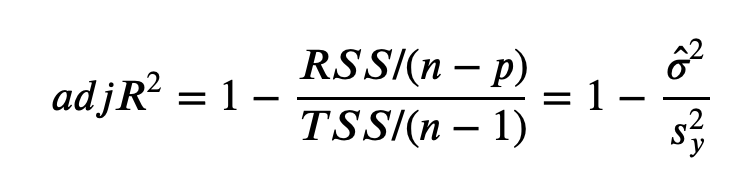
-This is sometimes refer as the proportion of variance of the response that is predictable with the estimated model

-Is computed based on in-sample observations (known as training set). It does not provide a sense of how good is our model in predicting out-of-sample cases (aka test set)!!

- R2 ranges between 0 and 1 **if** the **LM model has an intercept** and **is estimated by LS**!! Otherwise, this definition can result in negative values!

**Adjusted R2**

It takes into account the size of the model so that *R2* does not necessarily increase with additional explanatory variables.



**F test**

Formally, we are interested in testing the following hypothesis:

𝐻0:𝛽1=𝛽2=…=𝛽𝑝=0H0:β1=β2=…=βp=0

all coefficients, except for the intercept, are zero. This is sometimes known as the intercept-only model.

Now have many parameters that we want to simultaneously test if they equal zero. We can't do this with a **t-test**!!

A solution is given by an **F-test**!! 🡪 BUT we need to assume **normality of the error term** or for a **large enough sample size**, we can test a hypothesis on multiple parameters using an F-statistic that follows an F-distribution.

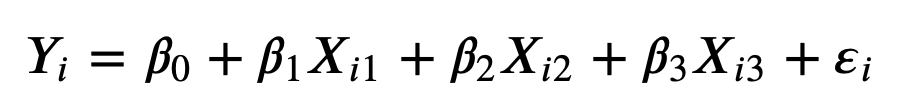
The F-statistic and its pvalue computed with the F-distribution are reported by glance() under statistic and p.value

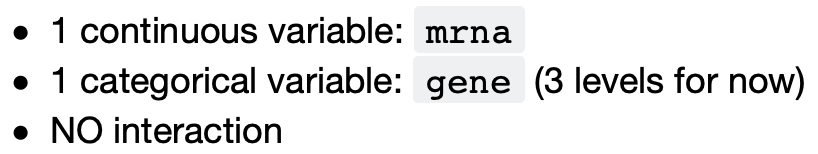
the estimate of the intercept in the null model is the sample mean of the response, the best estimate of E[Y]

**Also, F-test can be used to compare nested models!!!!**

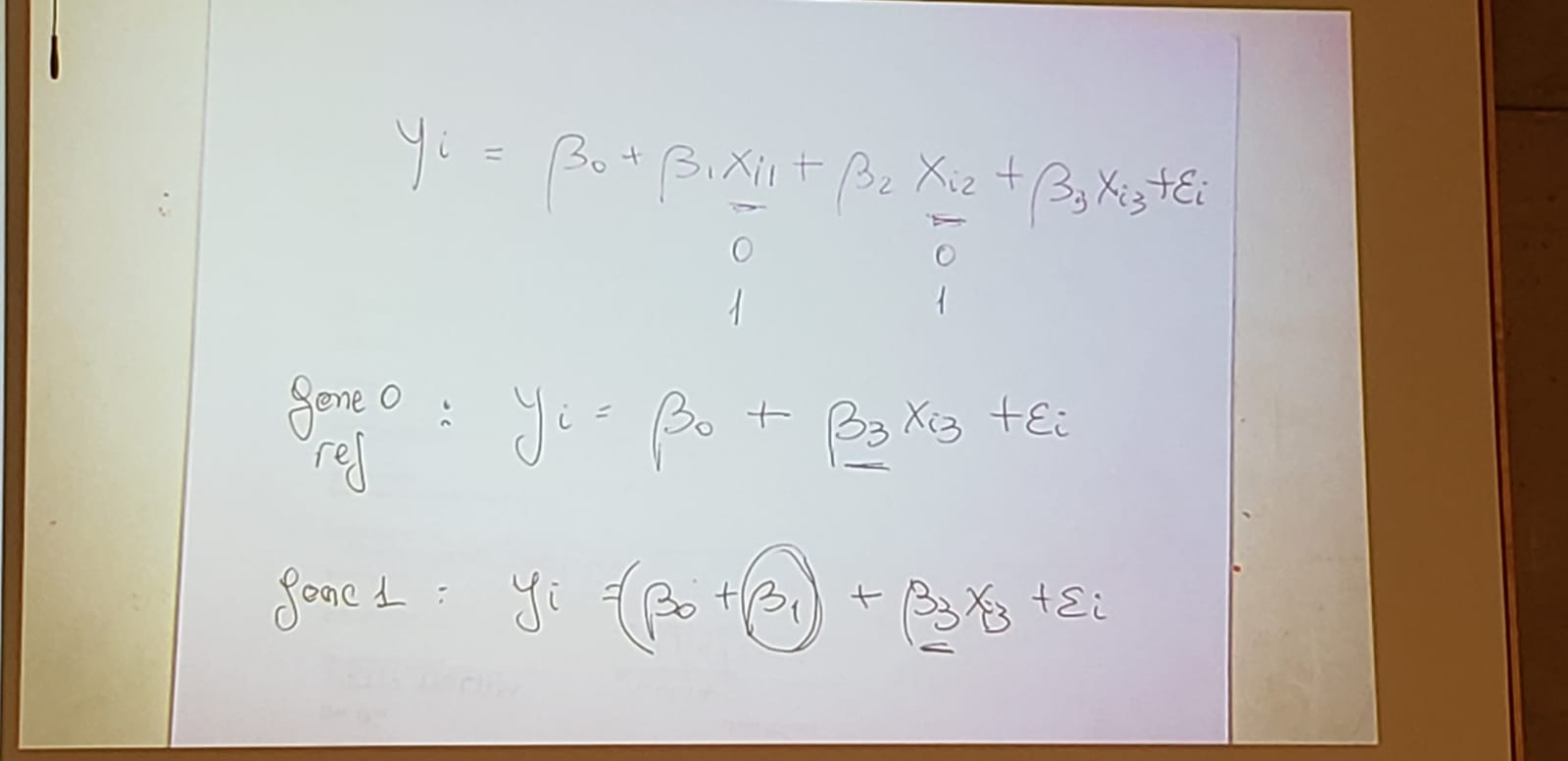
**additive model**

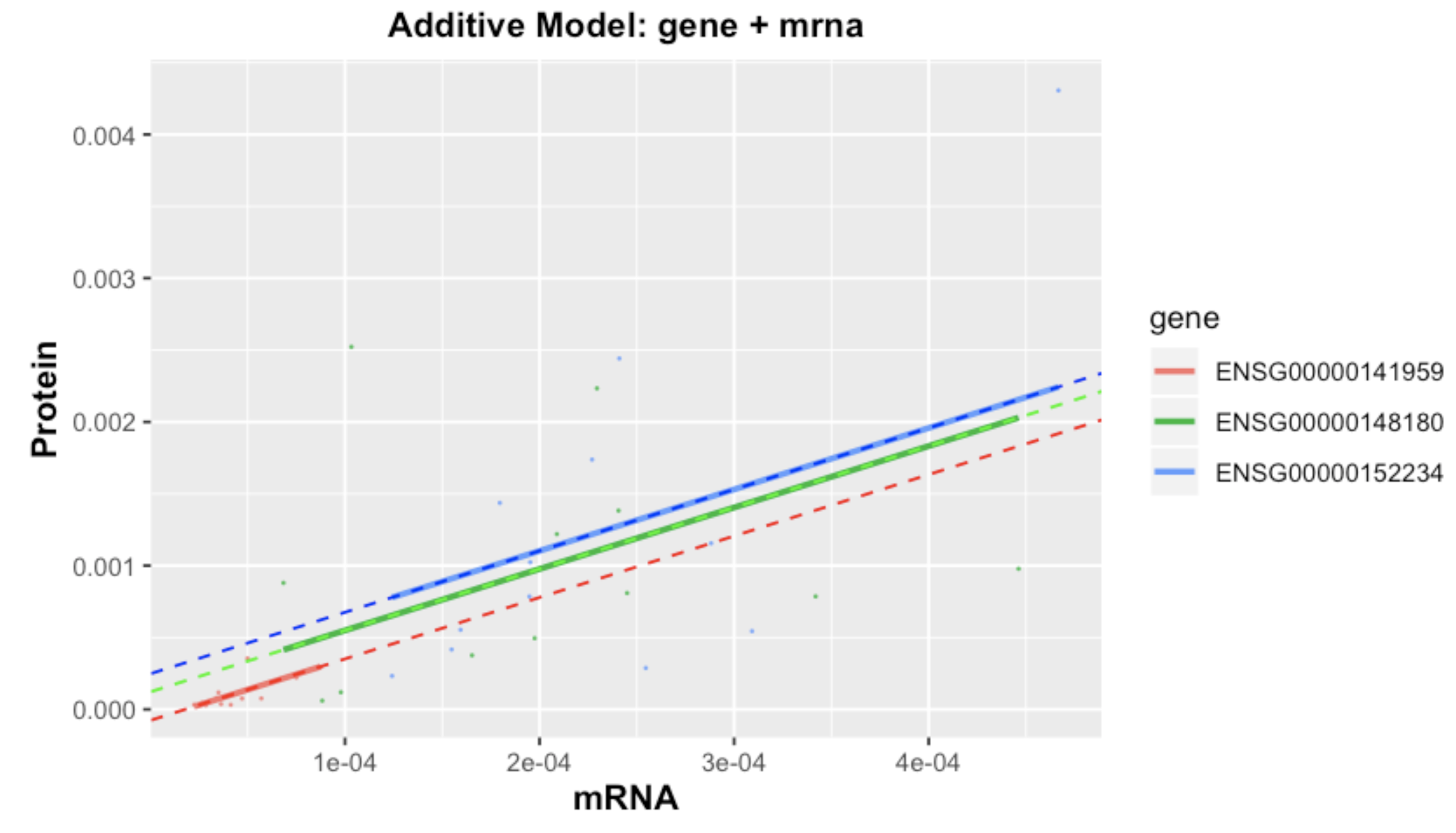
lm(wage~age+master): With a plus (+) only one slope





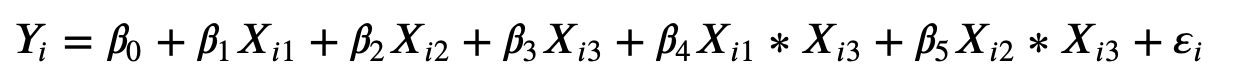
Results: parallel lines with different intercept but same slope.

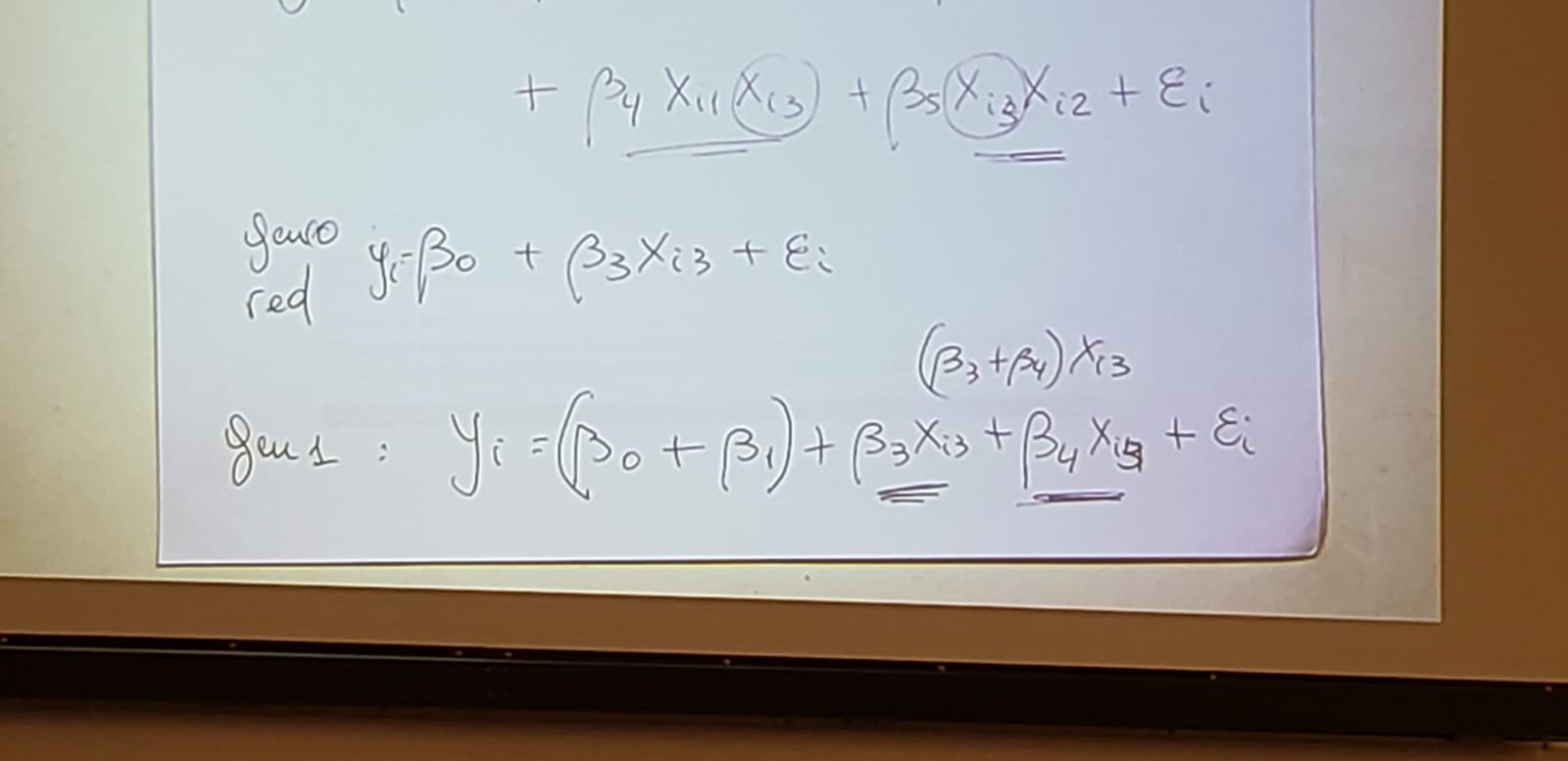


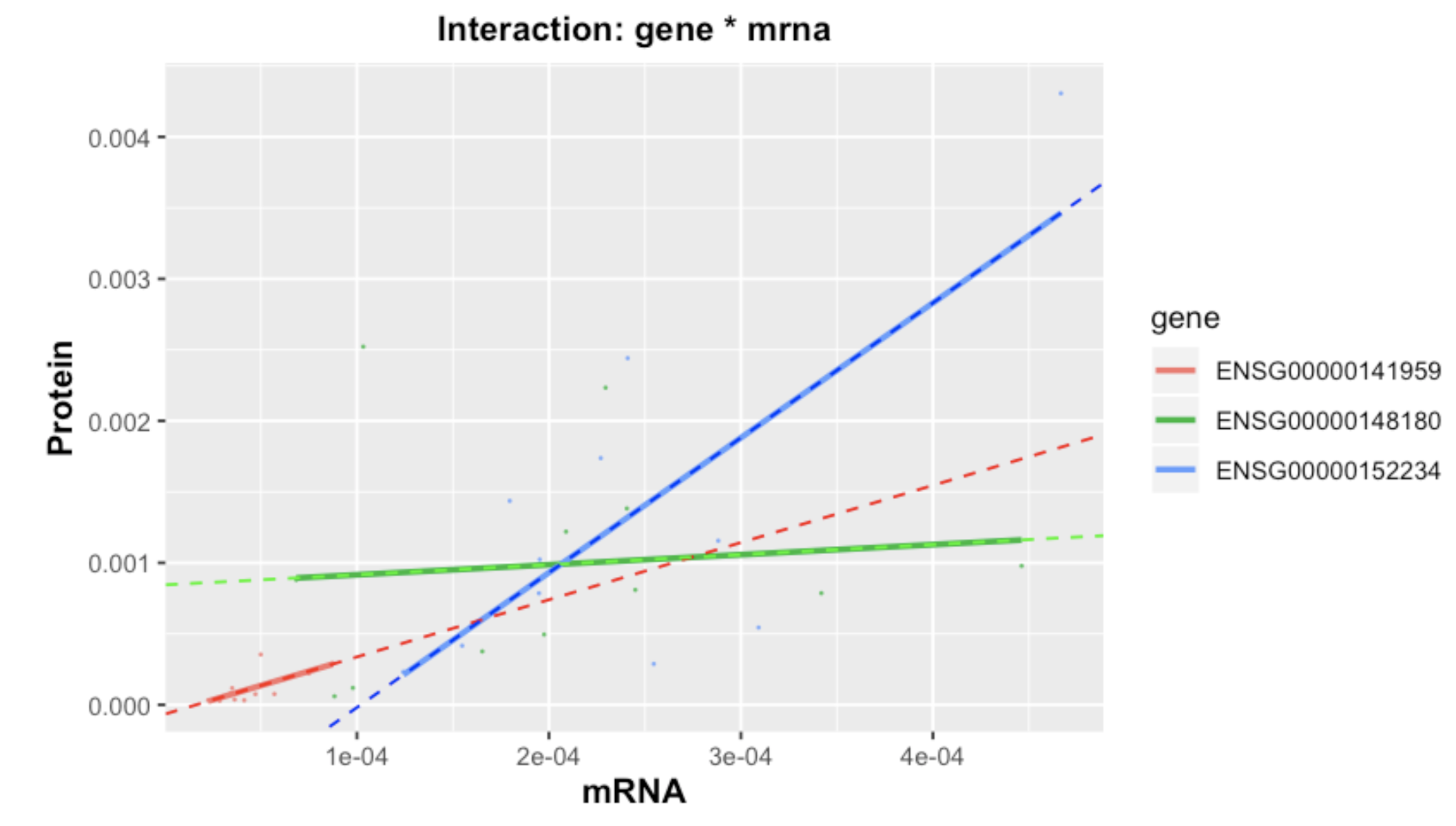


**interaction model**

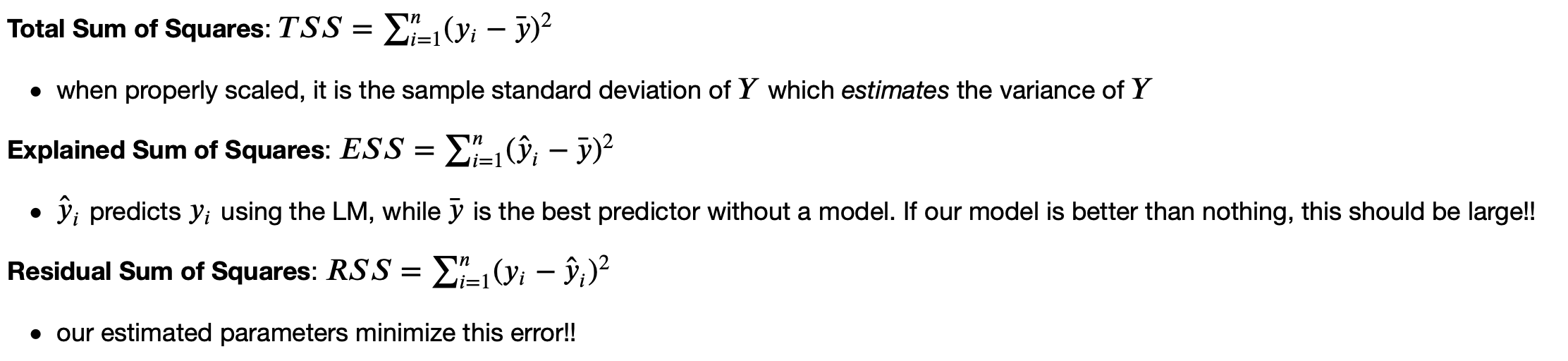
lm(wage~age\*master): With a plus (\*) that is interaction, one slope for every level



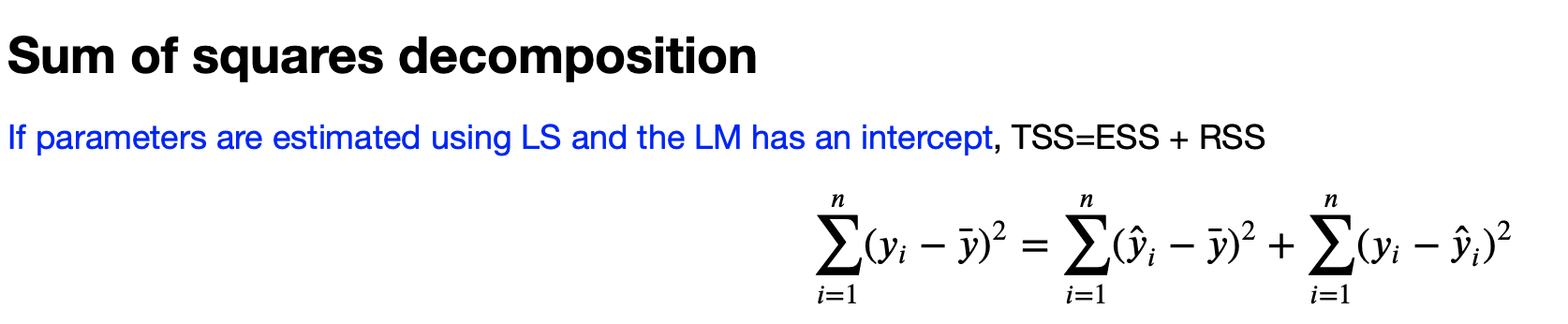


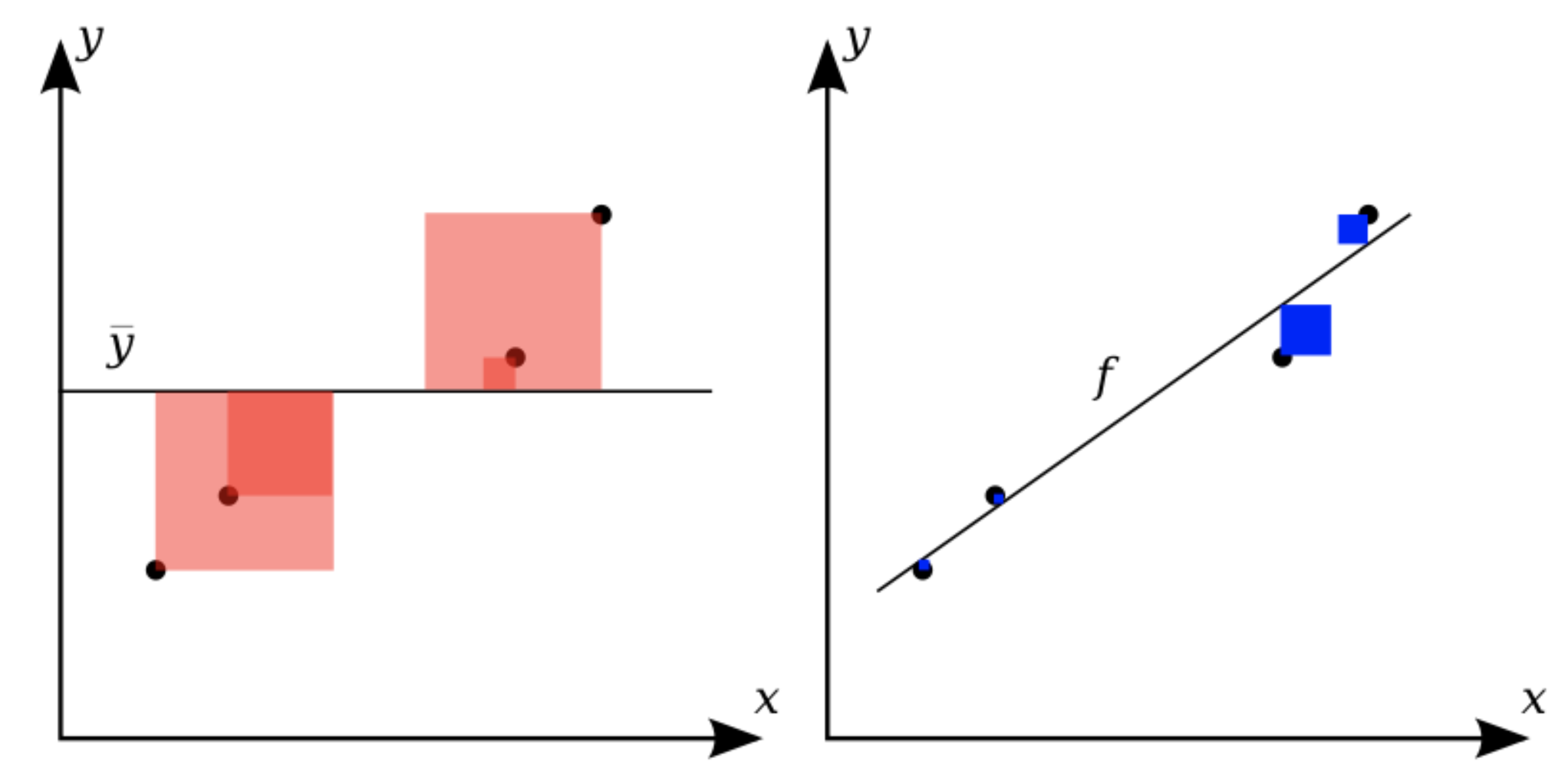


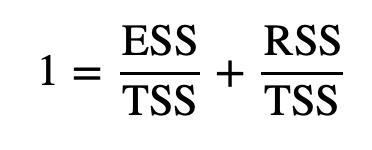
**Goodness of fit**



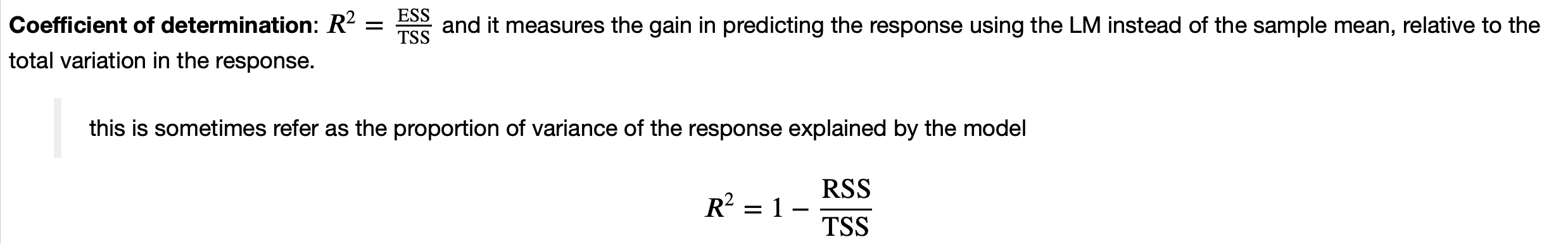
This is true with a model with intercept and Least Squares:







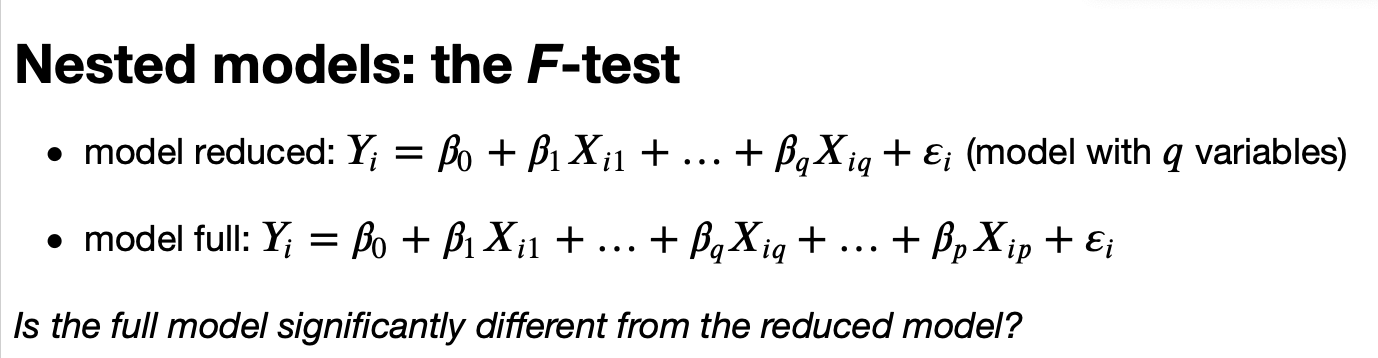
**THIS IS VERY IMPORTANT:**

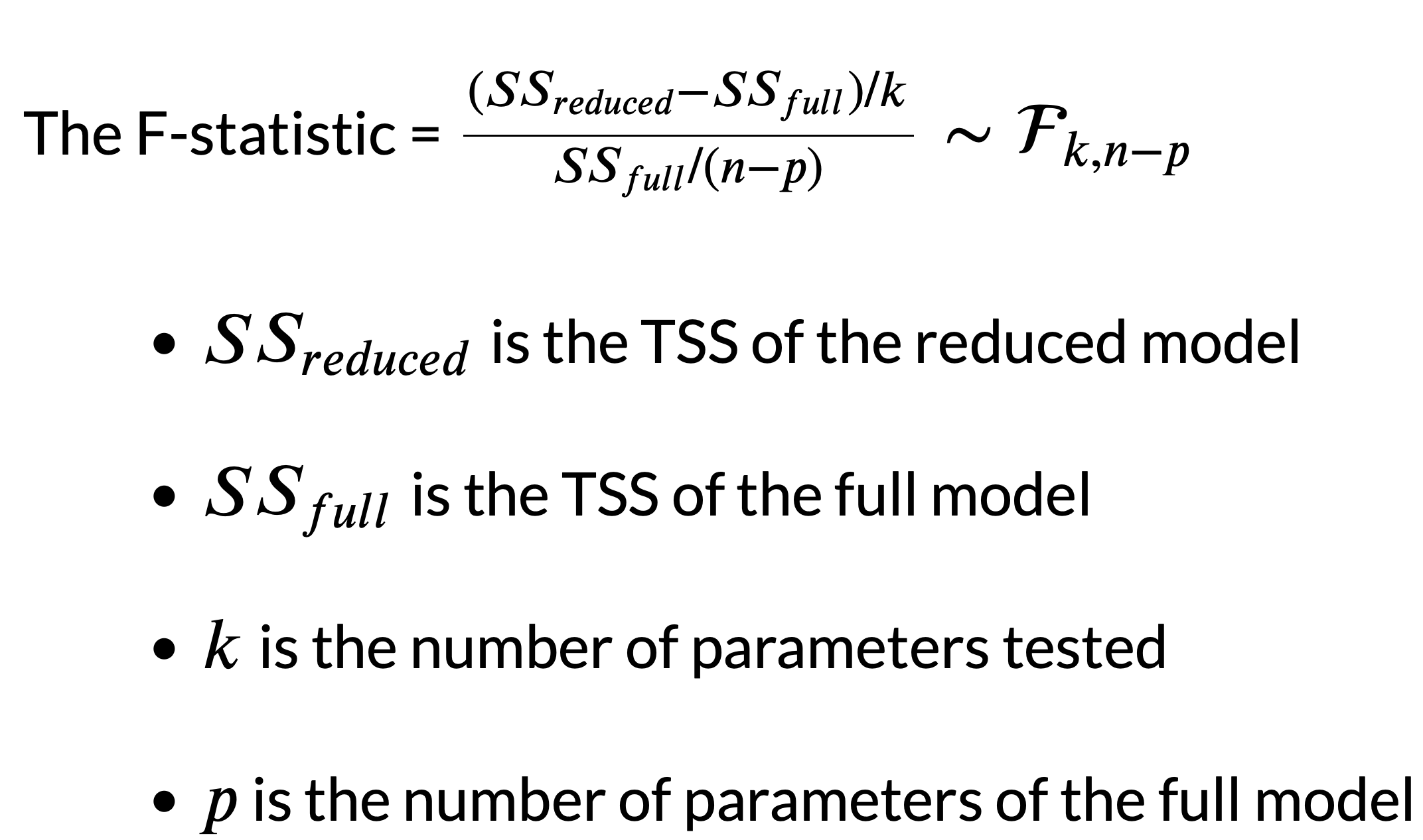


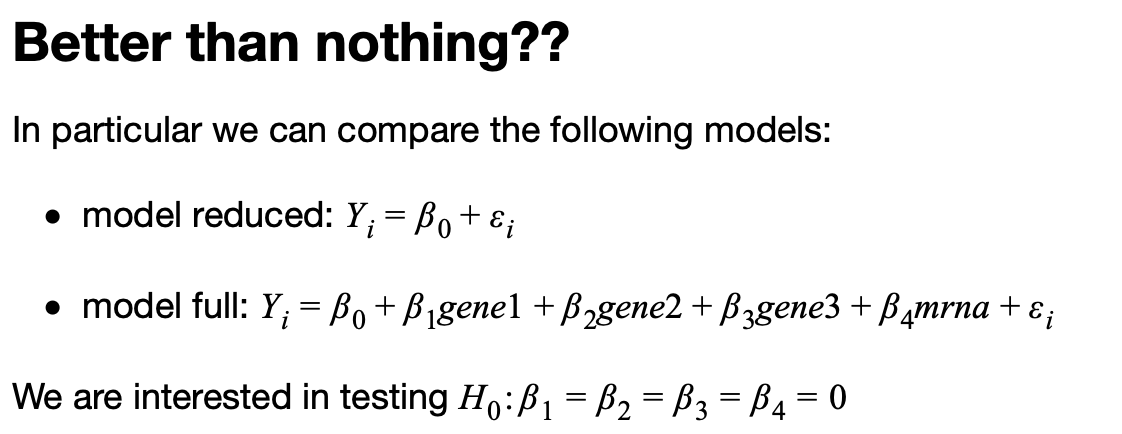
the proportion of (SAMPLE) variance of the response explained by the model.

0<R2<1

**I HAVE PROBLEMS TO UNDERSTAND THAT PART WITH R-SQUARE**

****

****

****

**The \*F\*-test above can be used to test this hypothesis**

**ANOVA FUNCTION TEST TE COMPLETE MODEL!!!**

Just a quick note on **sampling distribution**. The sampling distribution is the distribution of an **estimator**. This distribution is rarely known but in some special cases, you know the exact sampling distribution. If errors are Normal, it can be proved that the sampling distribution of the LS estimator of regression parameters is also Normal distribution (\sigma known) or Student's t (\sigma estimated from the data). If the CLT holds (it does for LS estimator under conditions stated in lect08) then the asymptotic sampling distribution is also Normal or t (asymptotic means a limiting distribution when n tends to infinity, in practice: approximate distribution for large n). Outside these cases, you don't know the sampling distribution but you can approximate it using a bootstrapping. Bootstrapping generates a long list of beta estimates so you can use the empirical distribution of these to approximate the sampling distribution.

To illustrate the CLT, we can also generate a long list of beta estimates and see the histogram to approximate the sampling distribution. The difference between this approach and bootstrapping is that in the former we sample from the population, while in bootstrapping you sample (with replacement) from the sample!!