# Regression with multiple predictors

## Overview

* Define the multiple linear regression model as
* Interpret the estimate for the intercept as the expected value of y when all predictors are equal to 0, on average. Sometime it might not have practical meaning.
* Interpret the estimate for a slope (say ​) as "All else held constant, for each unit increase in ​, we would expect y to be higher/lower on average by  ​."
* Define collinearity as a high correlation between two independent variables such that the two variables contribute redundant information to the model -- which is something we want to avoid in multiple linear regression.
* Define model selection as identifying the best model for predicting a given response variable.
* Note that we usually prefer simpler (parsimonious) models over more complicated ones.
* Define the full model as the model with all explanatory variables included as predictors.

## Adjusted

 will increase with each explanatory variable added to the model, regardless of whether or not the added variable is a meaningful predictor of the response variable. Therefore we use adjusted , which applies a penalty for the number of predictors included in the model, to better assess the strength of a multiple linear regression model:

Note that  will only increase if the added variable has a meaningful contribution to the amount of explained variability in y, i.e. if the gains from adding the variable exceeds the penalty.

* Since k is never negative, adjusted is always less than .
* Adjusted applies a penalty for the number of predictors included in the model.
* We choose models with higher adjusted over others.

## Collinearity

* Two predictor variables are said to be collinear when they are correlated with each other.
* Since predictors are also called independent variables, so they should be independent of each other.
* Inclusion of collinear predictors (also called **multicollinearity**) complicates model estimation.

## Parsimony

* Avoid adding predictors associated with each other because often times the addition of such variable brings nothing new to the table
* Prefer the simplest best model, i.e. the **parsimonious model**
  + **Occam’s razor**: Among competing hypotheses, the one with the fewest assumptions should be selected
* Addition of collinear variables can result in biased estimates of the regression parameters
* While it’s impossible to avoid collinearity from arising in observational data, experiments are usually designed to control for correlated predictors

# Inference for MLR

## Overview

* The significance of the model **as a whole** is assessed using an F-test

Df = n-k-1 degrees of freedom.

Usually reported at the bottom of the regression output.

* the p-values associated with each predictor are **conditional** on other variables being included in the model, so they can be used to assess if a given predictor is significant, given that all others are in the model.

These p-values are calculated based on a t-distribution with n-k-1 degrees of freedom

The corresponding CI for each slope parameter of each predictor is:

* Stepwise model selection (backward or forward) can be done based on p-values (drop variables that are not significant) or based on adjusted  (choose the model with higher adjusted ).
* The general idea behind **backward-selection** is to start with the full model and eliminate one variable at a time until the ideal model is reached.
  + p-value method:

1. Start with the full model.
2. Drop the variable with the highest p-value and refit the model.
3. Repeat until all remaining variables are significant.
   * adjusted  method:
4. Start with the full model.
5. Refit all possible models omitting one variable at a time, and choose the model with the highest adjusted .
6. Repeat until maximum possible adjusted  is reached.

* The general idea behind **forward-selection** is to start with only one variable and adding one variable at a time until the ideal model is reached.
  + p-value method:

1. Try all possible simple linear regression models predicting y using one explanatory variable at a time. Choose the model where the explanatory variable of choice has the lowest p-value (the lower p-value is, the more significant this variable is).
2. Try all possible models adding one more explanatory variable at a time, and choose the model where the added explanatory variable has the lowest p-value.
3. Repeat until all added variables are significant.
   * adjusted  method:
4. Try all possible simple linear regression models predicting y using one explanatory variable at a time. Choose the model with the highest adjusted  .
5. Try all possible models adding one more explanatory variable at a time, and choose the model with the highest adjusted  .
6. Repeat until maximum possible adjusted  is reached.

* Adjusted  method is more computationally intensive, but it is more reliable, since it doesn't depend on an arbitrary significance level
* Conditions for MLR:

1. **linear relationship** between each (numerical) explanatory variable and the response

Check using

* scatterplots of  vs. each
* residuals plots of residuals vs. each : random scatter around 0. This is preferred instead of scatterplot of y vs x, because it allows for considering the other variables that are also in the model, and not just the bivariate relationship between a given x and y.

1. **nearly normal residuals with mean 0**

Check using

* a nearly normal probability plot (qqnorm + qqline)
* histogram of residuals (hist)

1. **constant variability of residuals**

Check using

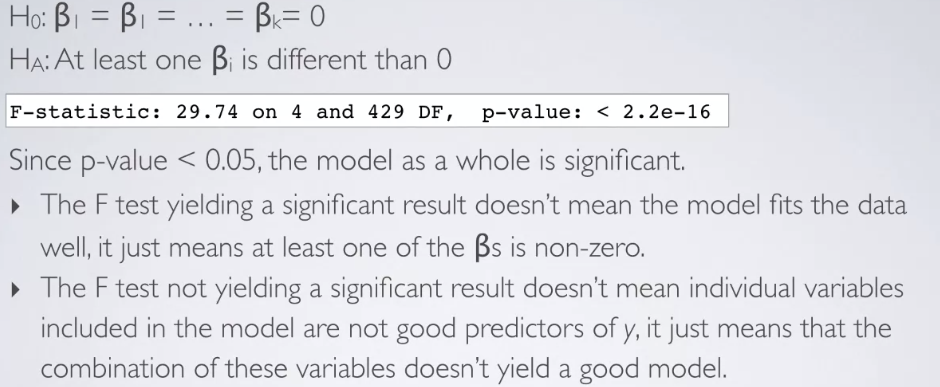
* residuals plots of residuals vs. . This is preferred because it allows for considering the entire model (with all explanatory variables) at once. We would like to see residuals randomly scattered in a band with a constant width around 0 (no fan shape).
* residuals vs. each . This helps us to identify unusual observations easily.

1. **independence of residuals** (and hence observations)

Check using

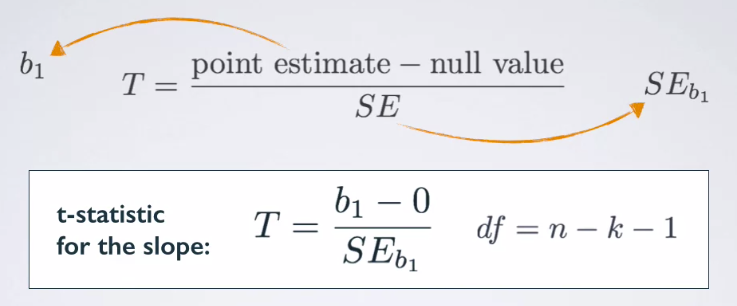
* a scatterplot of residuals vs. order of data collection (will reveal non-independence if data have time series structure)
* if not, think about how the data are sampled.
* No model is perfect, but even imperfect models can be useful

## Inference for the model as a whole



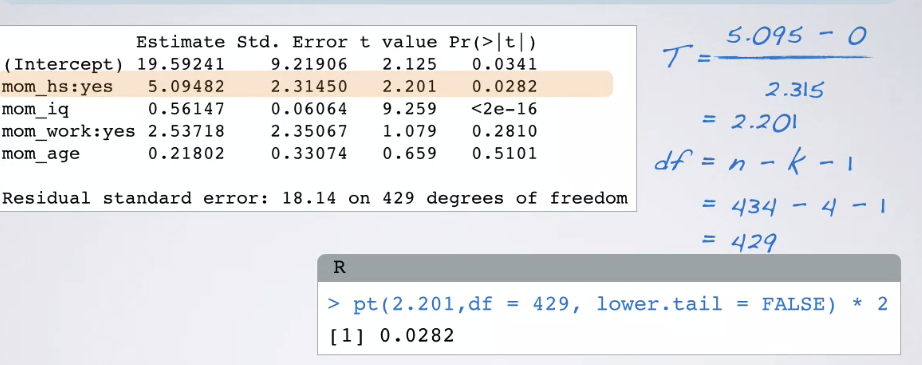
## Hypothesis testing for individual slope

Use a t-statistic in inference in inference for regression

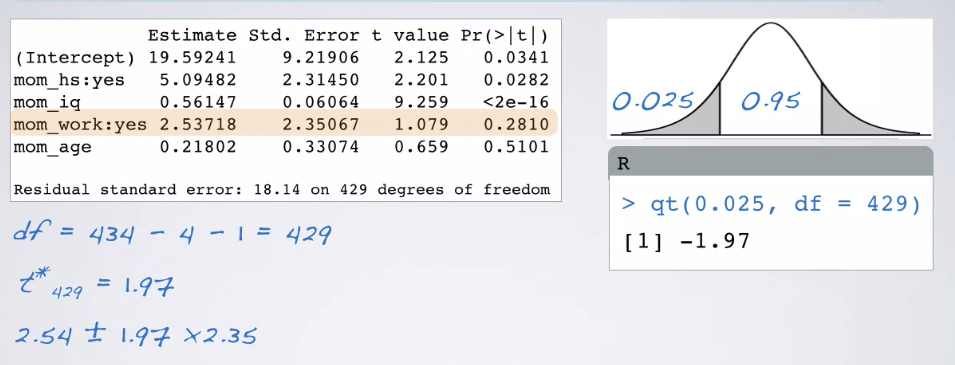


For MLR, df = n-k-1

For simple linear regression, df = n-1-1 = n-2



## Confidence intervals for slopes



**Interpretation**:

We are 95% confident that, all else being equal, the model predicts that children whose moms worked during the first three years of their lives score 2.09 points lower to 7.17 points higher than those whose moms did not work.

# Model selection

Criteria for stepwise model selection:

* P-value, adjusted
* AIC, BIC, DIC, Bayes factor, Mallow’s

If we have a categorical variable with multiple levels, we cannot drop some of the levels of that variable and keep others. We either need to decide that to keep the entire variable as a whole or drop it as a whole. If there is at least one level that is significant, we would keep the entire variable.

## Adjusted vs. p-value

* P-value: if we are trying to figure out which predictors are significant.
* Adjusted : If we want more reliable predictions.

# Model Diagnostics

4 assumptions for MLR