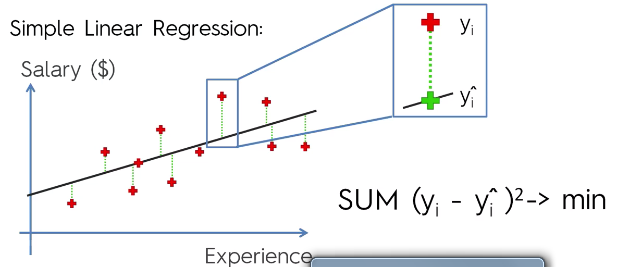
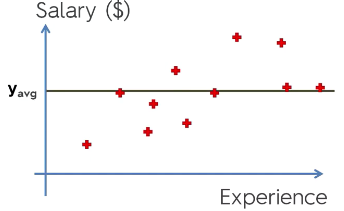
***R Squared***

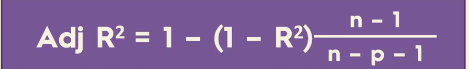
* Remember that linear regression is concerned with creating the model with the minimum **ordinary least squares** (sum of *squared differences* between predictions and actuals)



* This is also known as the **sum of squared residuals (SSres)**
* Now look at the differences between the average y value and the actual y values

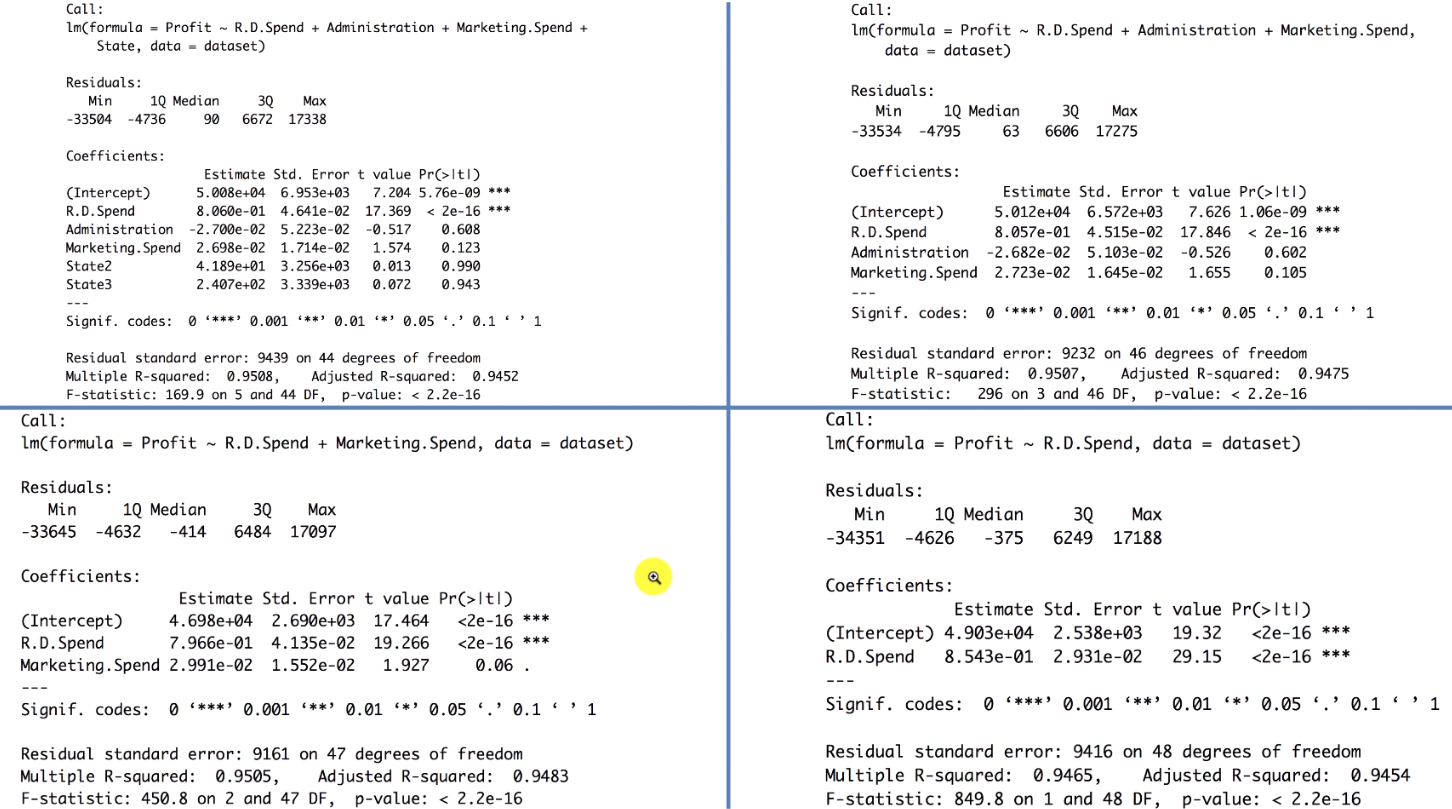
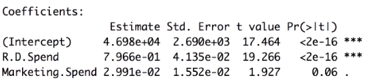


* The sum of all of *these* differences squared is the **total sum of squares (SStot)**
* **R2 = 1 – (SSres / SStot)** 🡪
* **There will always be a SStot**
* Want to fit function to minimize SSres, so a good model gives R2 values = 1 – small / SStot 🡺 1 – small = 1, i.e. values close to 1
* R2 can be *negative* if SSres actually first the data *worse than the average line* (difficult to do)
* **Adjusted R2** gives a penalty for the number of predictors used in a model
* R2 *cannot* decrease if we add more predictors, so we could add more predictors and increase our R2, but this would be a misleading indicator of quality of the model
* W/ more predictors, the SStot would not change, since this is based on the y values
* However, our SSres will be even smaller than before, and if a new predictor cannot decrease SSres (makes model worse), it gets a coefficient of about 0
* It will never be *exactly* 0, b/c there will always be some slight random correlation between the predictor and the outcome
* Therefore we must penalize our measure of quality for each predictor added, which gives us Adj. R2



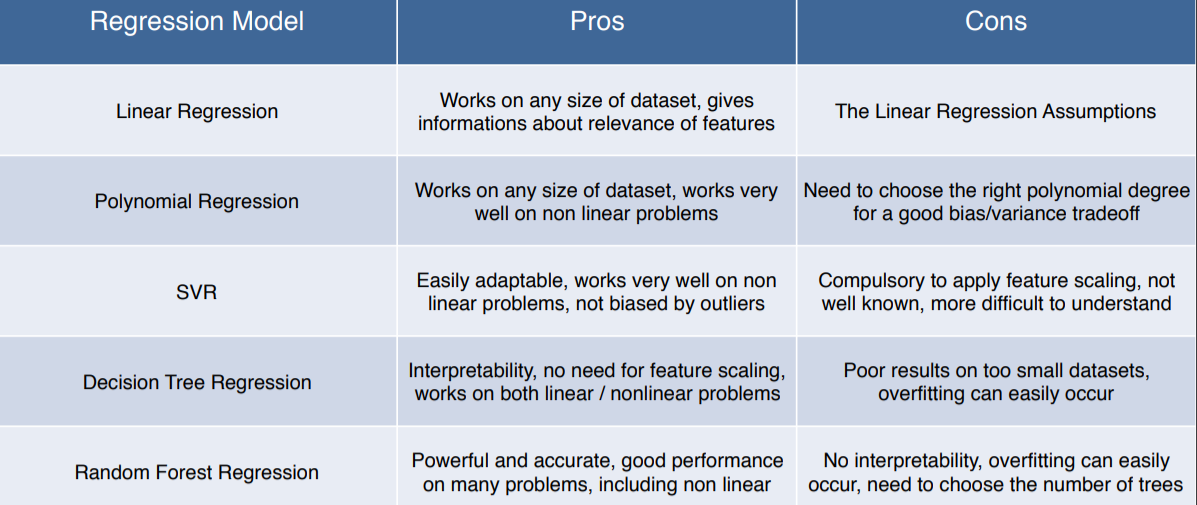
* p = predictors/regressors
* n = sample size
* W/ more p’s (predictors), that denominator *decreases*, so that final ratio *increases*, and therefore our entire 2nd half *increases*, so 1 minus that 2nd half will *decrease*, penalizing the additional p’s
* Also, when normal R2 increase, 1 – R2 decreases, and our Adj. R2 increases
* So there is a constant battle between the increasing R2 from adding p’s and the increasing penalties for adding p’s
* If a predictor is not good (i.e. does not increase R2), then we see a larger penalization in Adj. R2
* But if it helps a lot, the increase in R2 can overcome the penalization factor

**Evaluating Regression Models Performance**

* Remember our 4 models from multiple linear regression:
* Only ended up with 1 predictor, R&D Spend
* But our **step-wise regression methods** are very arbitrary
* How can we improve this method of building models and assess different situations and get other “opinions” or asses other criteria to figure out which model is best?
* We can see that although our 3rd model has an insignificant predictor (Marketing Spend), it has the highest Adjusted R2 (highest “goodness-of-fit”)
* Remember normal R2 is biased and will always increase w/ more predictors and needs to be penalized for each additional R2 value.
* For our coefficients, if a predictor’s coefficient is positive, the outcome is positively correlated w/ that predictor and will increase as the predict increases (increased profit w/ increased marketing)
* If it’s negative, the outcome is negatively correlated w/ that predictor (decreased profit w/ increased R&D spend)
* ***Magnitude*** *of coefficients is misleading*
* It depends on the *unit* *of measurement* of the variable
* 
* Instead*,* say that *R&D spend has a greater impact/effect on profit per unit of R&D spend than Marketing spend does per unit of Marketing Spend*
* One might be measures in $100k’s, the other in millions of $.
* For example, if **we keep all other variables constant**, for every unit increase of R&D spend, we increase profit by 0.7966 profit units (79 cents, b/c profit is measured in $’s)
* Coefficient’s value are what they are when keeping all other variables constant
* That’s why they change w/ different models that have different variables
* Since R&D is also measures in $’s, we can say that for every $1 increase in R&D spend, profit increases by 0.79 cents
* Marketing Spend is also in $’s, and it only leads to a 30-cent increase in profits, so we’d want to increase R&D spending (invest in R&D) over Marketing if we had the choice
* Marketing does drive a lot of *revenue,* but the *cost* incurred by Marketing may be high enough that it cuts into *profits* more than R&D spend (revenue = profits – cost)
* i.e. Marketing is eating up a lot of the revenue its creating
* Whereas R&D is creating more revenue with less costs

***Regression Summary***

* 6 regression models: What are the pros and cons of each model ?



* How do I know which model to choose for my problem?
* 1st, figure out whether the problem is linear or non-linear.
* If linear, go for Simple Linear Regression if you only have 1 feature + Multiple Linear Regression if you have several
* If the problem is non-linear, go for Polynomial Regression, SVR, Decision Tree or Random Forest.
* Then which one should you choose among these 4? 🡺 Part 10 - Model Selection.
* The method consists of using a very relevant technique that evaluates models’ performance, called **k-Fold Cross Validation**, + then picking the model that shows the best results.
* How can I improve each of these models?
* In Part 10 - Model Selection, you will find the 2nd section dedicated to **Parameter Tuning**, which allows you to improve the performance of models by **tuning** them.
* Each model is composed of two types of parameters:
* **Parameters** that are learnt (ex: coefficients in Linear Regression)
* **Hyperparameters =** parameters that are NOT learnt + are fixed values inside the model equations
* Ex: The **regularization parameter** lambda or the **penalty parameter** C
* So far, we’ve used the default value of these hyperparameters + haven't searched for their optimal value so that a model reaches even higher performance.
* Finding their optimal value is exactly what Parameter Tuning is about.