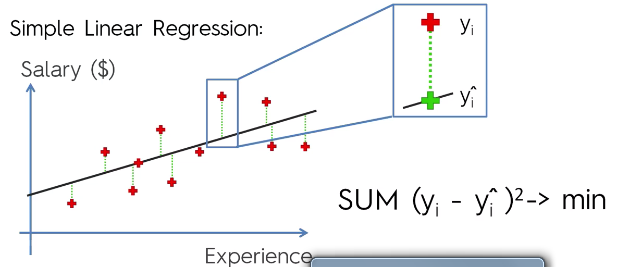
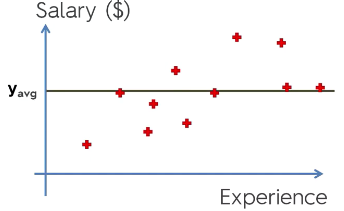
***R Squared***

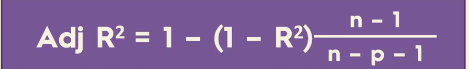
* Remember that linear regression is concerned with creating the model with the minimum **ordinary least squares** (sum of *squared differences* between predictions and actuals)



* 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

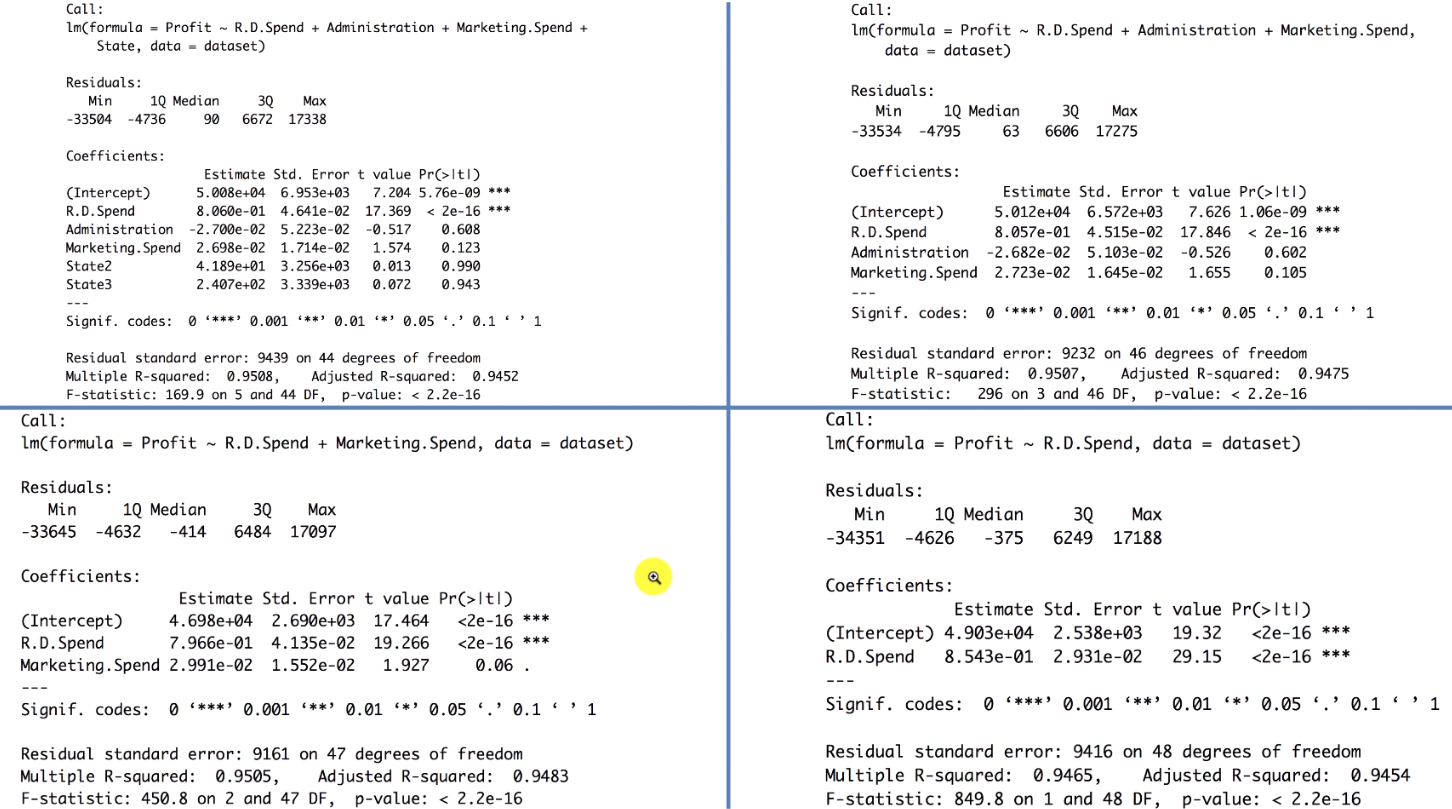


* The sum of all of *these* differences squared is the **total sum of squares (SStot)**
* **R2 = 1 – (SSres / SStot)** 🡪
* **There will always be a SStot**
* Want to fit function to minimize SSres, so a good model gives R2 values = 1 – small / SStot 🡺 1 – small = 1, i.e. values close to 1
* R2 can be *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.