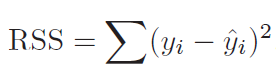
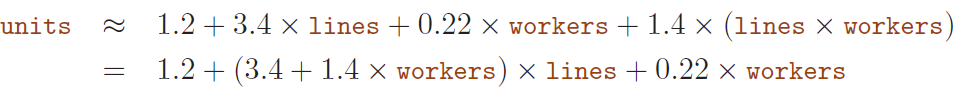
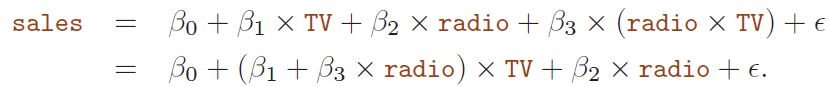
* R2 : % variation in Y explained by variation in X
* Not flexible (i.e. linear decision boundary) 🡪 low-variance, high-bias
  + Linear regression
  + KNN (with k high)

Multiple Linear Regression – 4 steps:

1. Compute F-stat
   1. H0: B1 = B2 = …. = 0
   2. Higher F-stat 🡪 reject null
      1. F-stat < 1 🡪 for sure fail to reject null
   3. Fails for: (p > n)
2. Forward/Backward selection:
   1. Step-wise add variable that results in lowest RSS
      1. 
   2. Step-wise remove variable that has largest p-value
3. Model fit
   1. R2 always increases with addition of new variables
      1. New variable allows fitting data better
   2. Use adjusted R2 to determine if variable helped
4. Predict
   1. Use confidence intervals for estimates

* Interaction term:
  + 
  + 
  + Assume 2 variables influence each other
  + True relationship b/w two variables is NOT additive
  + Include interaction term 🡪 MUST include both individual terms (regardless if p-values < 0.05)
* Problems with linear regression:
  + Errors auto-correlated
    - Underestimate true standard deviation
  + Heteroscedasticity of errors 🡪 errors follow pattern with change in x (i.e. non-constant variance)
    - Bad
    - Solution 🡪 replace Y with log(Y)
* Collinearity – highly correlated **2** independent variables
  + Hard to determine which affects the dependent variable more
  + Reduces accuracy of estimates of the regression coefficients 🡪 (+) standard error of Betas 🡪 t-stat: may fail to reject H0: Bi != 0
* Multicollinearity – highly collinearity b/w **3 independent** variables (but no collinearity b/w any two pairs)
* Testing for collinearity – **variance inflation factor (VIF)** formula
  + 1 = lowest possible value = NO collinearity
* Two solutions to collinearity:
  + (1) Drop one of collinear variables
  + (2) Combine collinear variables
* Parametric vs. Non-parametric:
  + Parametric – (linear regression)
    - Pros:
      * Easy to fit
      * Only estimate small number of coefficients
    - Cons:
      * Strong assumptions about form of f(X)
  + Non-parametric – (KNN)
    - Pros:
      * No assumption about parametric form of f(X)
    - Cons:
      * Curse of dimensionality – (+) variables 🡪 (+) distance values from each other