Linear Regression Part 1

DSA 6000: Data Science and Analytics, Fall 2019

Wayne State University

An interesting article

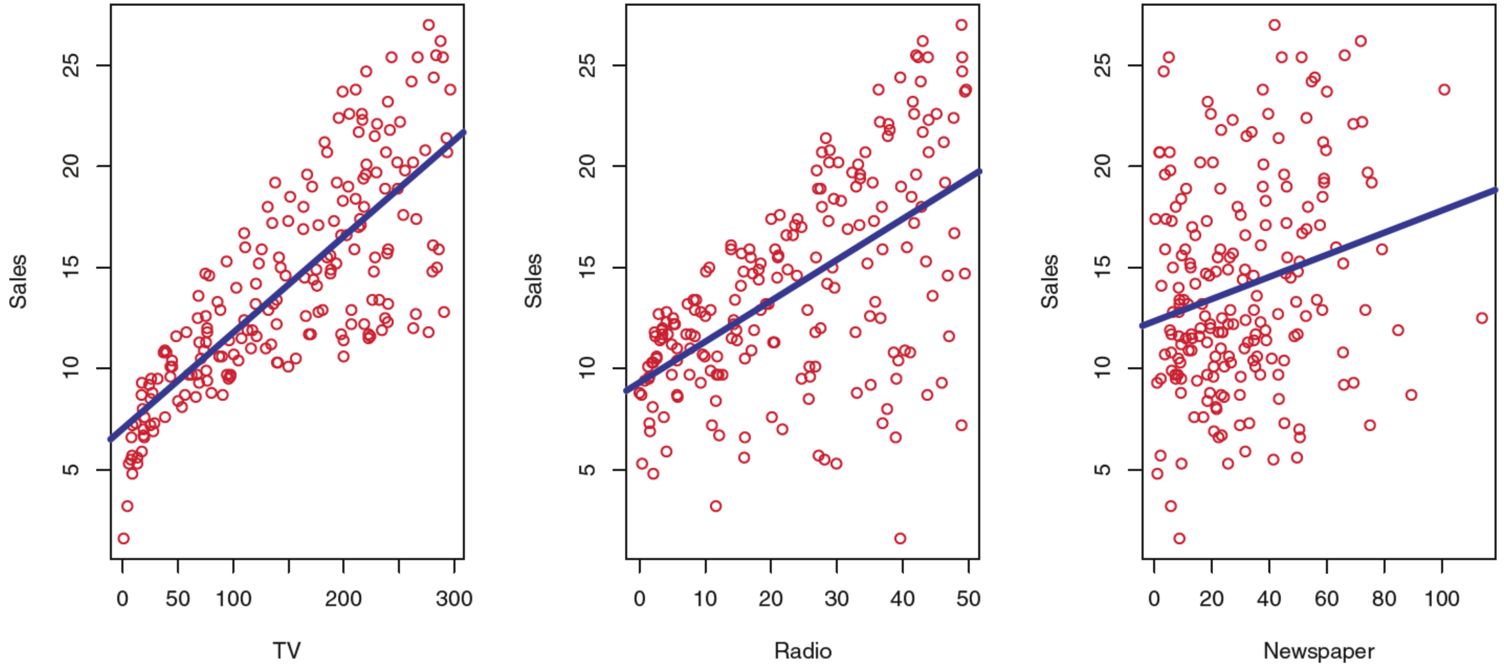
- **The Simple Economics of Machine Intelligence** by A. Agrawal, J. Gans and A. Goldfarb, November 17, 2016, Harvard Business Review
- Authors' conclusions:
  - Machine learning is in essence a **prediction technology**.
  - As cost of prediction plummets, two things will happen:
    - Prediction will be used in previously unapplied areas (since it is cheap)
    - The value of things that **complement prediction** will rise
  - The value goes up for complements and down for substitutes.
  - All human activities have 5 high-level components: data, prediction, **judgment**, action and outcomes
  - The value of human judgment skills will increase.
    - There will be greater demand for the application of ethics, and for emotional support.
    - **Safe jobs**: CEOs, caregivers, artists, counselors, social workers, beauty consultant, PR/Marketing directors, elderly companions, etc.

https://www.youtube.com/watch?v=wWvXVehccjw

https://www.youtube.com/watch?v=ajGgd9Ld-Wc

Linear Regression - A motivating example

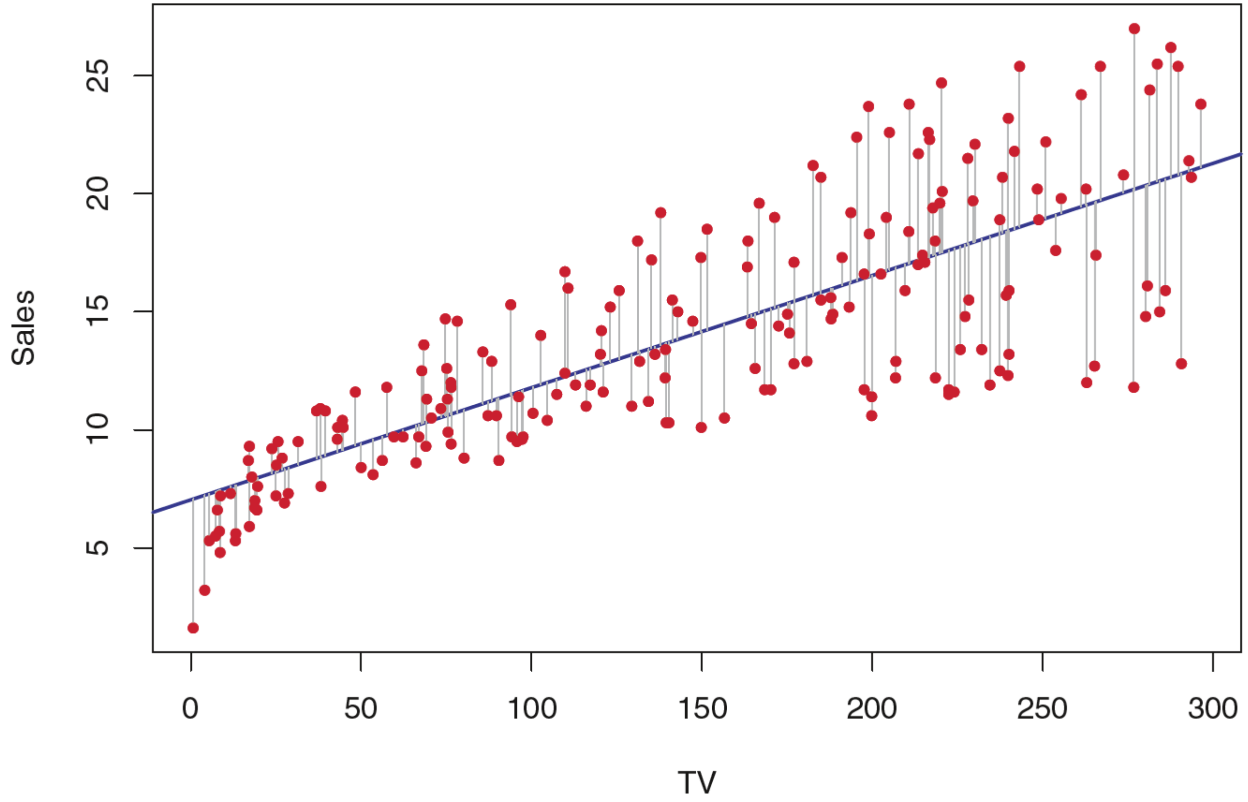
- **Input** variables are advertising budgets in TV, Radio and Newspaper
- **Output** variable is Sales



- Is there a relationship between advertising budget and sales? How strong is the relationship?
- How accurately can we estimate the effect of each medium on sales?
- How accurately can we predict future sales?
- Is there synergy among the advertising media?

Simple Linear Regression

- Model: $Y = \beta_0 + \beta_1 X + \epsilon$
- β_0 is the intercept, the expected value of Y without knowing X
- β_1 is the slope, the average increase in Y associated with a one-unit increase in X



$$Sales = \beta_0 + \beta_1\, TV + \epsilon$$

Multiple Linear Regression

- Model: $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$ for $i = 1, \dots, n$
- Assumptions: linear relation between E(Y_i) and x_i, E(ϵ_i) = 0, Var(ϵ_i) is constant, and ϵ_i uncorrelated with each other
- $\beta_j,\ j = 0,1, \dots, p$ are regression parameters or coefficients, whose values are (assumed to be) fixed but unknown
- We estimate β_j's using sample observations $(X_1, y_1), (X_2, y_2), \dots, (X_n, y_n)$
- Denote the estimators by $\hat{\beta}_j$
- **Important:** Understand the properties of $\hat{\beta}_j$

Parameter, Estimator and Estimate

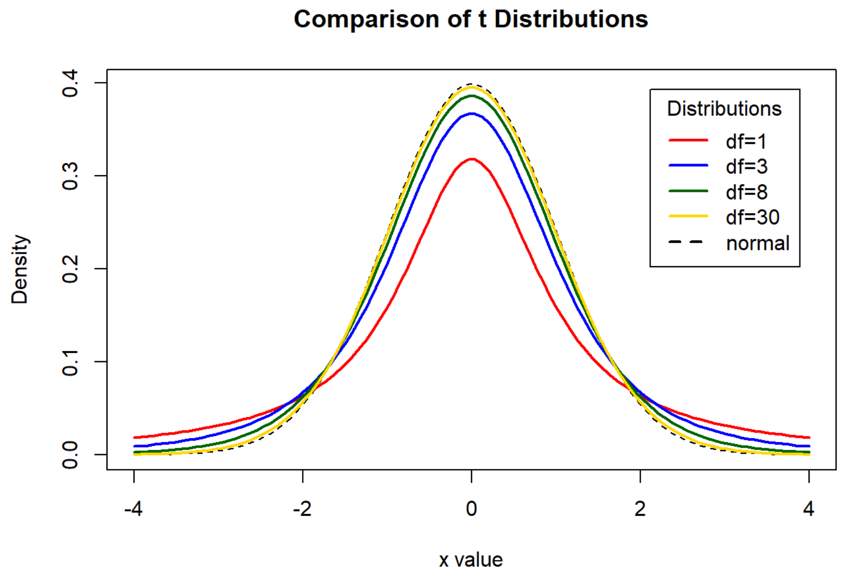
- $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$
- For each $j = 0,1, \dots, p$
  - β_j is a model **parameter**, an unknown deterministic quantity
  - $\hat{\beta}_j$ is (before a sample is drawn) a random variable, called an **estimator**
  - The value of $\hat{\beta}_j$ can be calculated from data (random sample)
  - Each set of sample data gives a specific realization of $\hat{\beta}_j$, called an **estimate**. Different samples will generally produce different estimates.
  - An estimate is a realization of an estimator at a given sample.

Estimating β_j

- Given a training sample, $(X_1, y_1), (X_2, y_2), \dots, (X_n, y_n)$, how do we estimate the coefficients β_j?
- Least Squares method: $\underset{\hat{\beta}}{\text{minimize}} \sum_{i=1}^{n} (\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{ij} - y_i)^2$
- Properties of least squares estimators
  - Unbiased: $E(\hat{\beta}_j) = \beta_j$
  - Minimum variance: Least squares estimator has the minimum variance among all unbiased estimators of β_j

Properties of the estimator $\hat{\beta}_j$

- The estimator $\hat{\beta}_j$ for each j has a Normal distribution with mean β_j and standard deviation σ_{β_j}
- Both β_j and σ_{β_j} are unknown, fixed numbers
- While we use $\hat{\beta}_j$ to estimate β_j, we use the *Standard Error* $\mathrm{SE}(\hat{\beta}_j)$, a statistic calculated from sample data, to estimate σ_{β_j}
- The statistic $t = \dfrac{\hat{\beta}_j - \beta_j}{\mathrm{SE}(\hat{\beta}_j)}$ will have a *t distribution*



Concept of the Confidence Interval

- Let θ be a parameter and $\hat{\theta}$ be its unbiased estimator with a Normal distribution
- Let $\mathrm{SE}(\hat{\theta})$ be the standard error of $\hat{\theta}$
- Then the statistic

$$\hat{I}_{1-\alpha} = [\hat{\theta} - t_{\alpha/2,v}\mathrm{SE}(\hat{\theta}), \hat{\theta} + t_{\alpha/2,v}\mathrm{SE}(\hat{\theta})]$$

  is the *confidence interval* of θ at the $(1-\alpha) \times 100\%$ *confidence level*, where v is the degree of freedom, equal to $n-2$.
- Each time we calculate the statistic with a different sample, we will get a different value for $\hat{\theta}$ and a different interval $\hat{I}_{1-\alpha}$
- $(1-\alpha) \times 100\%$ of such intervals will encompass θ, the true parameter value

Talk intelligently about CI

- Suppose we choose a confidence level of 95% and calculated the confidence interval of θ from a given sample
- “With 95% chance the true value of θ will fall in the confidence interval”
  - This statement is **Wrong**. θ is a fixed number, not a RV. It is where it is, its whereabouts is not random.
  - It is the confidence interval that is a random variable -- its center location and width depend on the sample by which it is calculated.
- “If we constructed 100 such confidence intervals each with a different random sample, we would expect 95 of them to cover the true value of θ”
  - **Right**

Confidence Interval and Prediction Interval

- Suppose $p = 1$, i.e., simple linear regression
- Given an particular predictor value X_i
- The *Confidence Interval* is intended for the mean response E(Y_i), or equivalently $\beta_0 + \beta_1 x_i$, or $f(X_i)$.
  - "The 95% confidence interval at x = 35 is [10.985, 11.528]" means that 95% of intervals so obtained will contain the true value of E(Y) given x = 35.
- The *Prediction Interval* is intended for the actual response Y_i, or equivalently $\beta_0 + \beta_1 x_i + \epsilon_i$, which is a R.V.
  - "95% prediction interval at x = 35 is [7.930, 14.580]" means that 95% of intervals of this form will contain the actual value of Y corresponding to x = 35
- A prediction interval is wider than a confidence interval at the same predictor value x, since it accounts for more uncertainty

t-statistic
p-value

Model: $Y = \beta_0 + \beta_1 X + \epsilon$

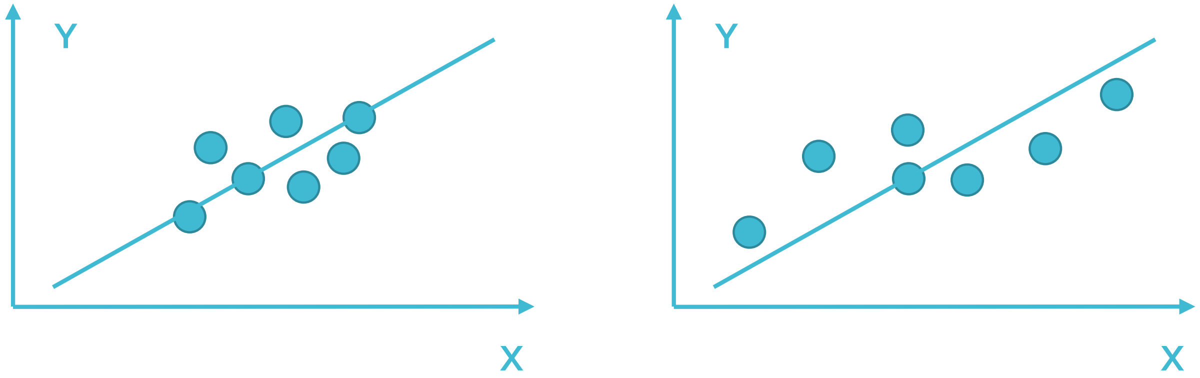
- The question: Is there a relationship between X and Y?
- Is equivalent to: Is the true value of β_1 equal to zero or not equal to zero?
- Null hypothesis H_0: $\beta_1 = 0$
- We can compute the estimate $\hat{\beta}_1$ and SE($\hat{\beta}_1$) from data
- If H_0 were true, we would expect $t = (\hat{\beta}_1 - 0)/\mathrm{SE}(\hat{\beta}_1)$ to follow a t distribution
  - If we get a t-statistic of 15.36, does it seem to come from t distribution? Do you believe H_0 is true?
- *p-value* is the probability of "seeing a sample data that produces this t value or a more extreme one (>|t|)" under the null hypothesis
- A small p-value gives strong evidence to reject the null hypothesis – there *is* a significant relationship between X and Y
- Failing to reject H_0 does not give a strong evidence to conclude X and Y are unrelated – they still may be related but we simply haven't yet observed a strong evidence from data

Leverage

The standard error of an estimator reflects how it varies under repeated sampling. We have

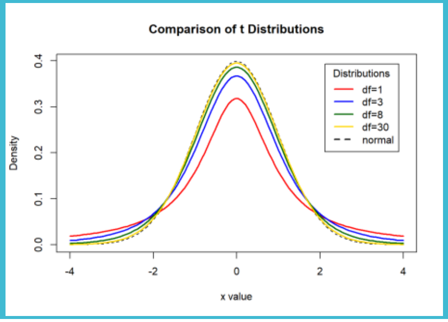
$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

where $\sigma^2 = \mathrm{Var}(\epsilon)$



- $SE(\hat{\beta}_1)$ is smaller when x_i are more spread out
- For each training point, the leverage statistic: $h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2}$
- If an observation has a leverage statistic way greater than $\frac{p+1}{n}$, we should be alerted.

Interpreting the LR output for an individual predictor



Model: Sales = $\beta_0 + \beta_1 TV + \epsilon$ In R: `m1 <- lm(sales ~ TV, data = ad)`

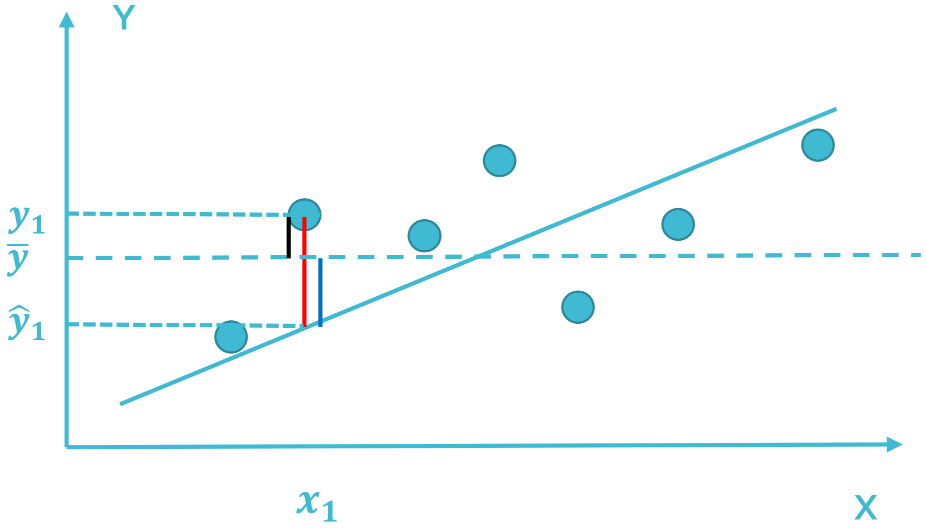
	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

- $\hat{\beta}_0$ = 7.0325, SE($\hat{\beta}_0$) = 0.4578, $\hat{\beta}_1$ = 0.0475, SE($\hat{\beta}_0$) = 0.0027
- H_0: There is no relationship between Ad budget on TV and Sales, $\beta_1 = 0$
- H_1: There is some relationship between Ad budget on TV and Sales, $\beta_1 \neq 0$
- Regardless of the hypothesis, the statistic $t = \frac{\hat{\beta}_1 - \beta_1}{\text{SE}(\hat{\beta}_1)}$ has a t distribution.
- To test H_0, we compute the t-statistic **under the assumption that H_0 is true**:
  - $t = \frac{\hat{\beta}_1 - 0}{\text{SE}(\hat{\beta}_1)} = \frac{0.0475}{0.0027} = 17.67$
- It is extremely unlikely to encounter a random number drawn from the t-distribution to have a value as high as (or higher than) 17.67.
- The probability of seeing such a thing is the **p-value**, $P(> |t|) < 0.0001$
- Therefore, the assumption that H_0 is true is most likely wrong, we reject H_0
- We declare that a relationship exists between Ad budget on TV and Sales
- The same argument applied to the analysis of each individual predictor in a multiple linear regression.

Assessing the Accuracy of the Model

- Residual Sum of Squares (RSS): $\sum_{i=1}^{n}\left(y_i - \hat{f}(X_i)\right)^2$, measures the amount of variability in Y that is left unexplained after performing the regression.
- Residual Standard Error (RSE) is an estimate of the standard deviation of ϵ.
  - RSE = $\sqrt{\left(\frac{1}{n-p-1}\right) RSS}$
  - Note that RSE depends on p, so adding a useless predictor to the model increases $\left(\frac{1}{n-p-1}\right)$, overall RSE might also increase
- RSE represents the average amount that the response will deviate from the true regression line. It is a measure of the *lack of fit* of the model to the data, in the units of Y.
- R^2 statistic: the proportion of variance in Y that is explained by the model.
  - $R^2 = \frac{TSS-RSS}{TSS}$, where $TSS = \sum(y_i - \bar{y})^2$ is the total sum of squares.
  - Adding an extra predictor will always increase R^2
  - Adjusted R^2 accounts for the model complexity

Decompose the TSS



- Suppose the linear model is fit by the OLS method, then
- Total variation in Y is decomposed into two parts:
  - Variation explained by the model
  - Variation left unexplained by the model
- $Total\ SS = Explained\ SS + Residual\ SS$
- $\sum_{i=1}^{n}(y_i - \bar{y}_i)^2 = \sum_{i=1}^{n}(\bar{y}_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

About the F-statistic

- F-statistic is used for testing whether at least one of the predictors has a significant effect on the response variable.
- $H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$
- H_1: at least one β_j is non-zero
- A large F-statistic value will lead to rejection of H_0. The rejection threshold depends on both n and p.
- Why use F test since we already have the t test?
- For a model with many predictors (i.e., large p), it can happen that the p-value for some individual predictor(s) is small (e.g., < 0.05), but the model as a whole fails the F test (i.e., fail to reject H_0).
  - For instance, if there are 100 variables, all unrelated to Y, the p-values for about 5% of the variables will be below 0.05 **by chance**. We would expect to see about 5 small p-values even in the absence of any true association between the predictors and the response.
  - F-statistic is immune to this type of fallacy.