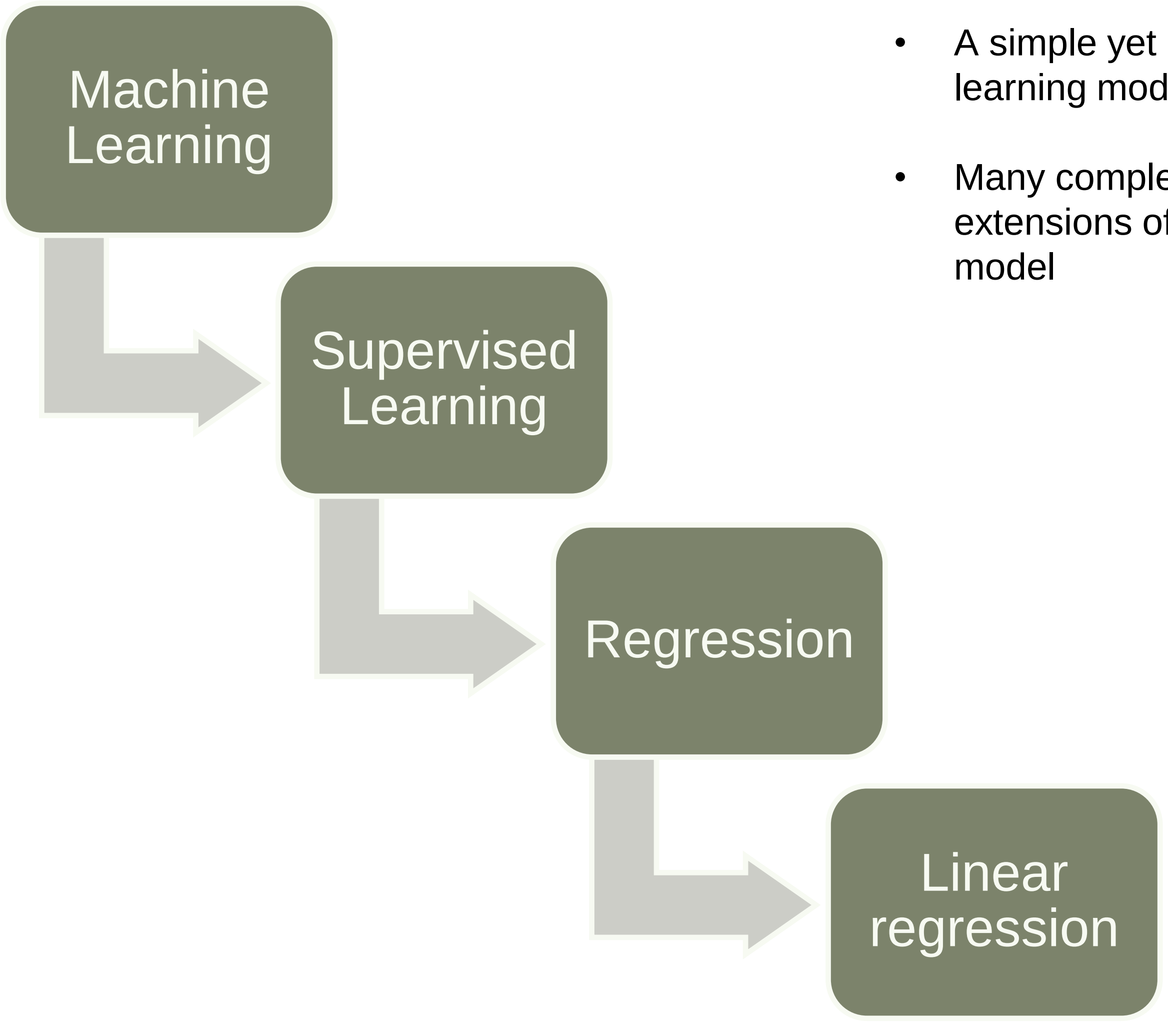
X2. Linear Regression



- A simple yet important supervised learning model
- Many complex models are extensions of linear regression model

The Mathematical Mapping between Y and Xs

- We have a series of $\mathrm{X} = (X_1, X_2, \ldots, X_p)$ and a variable Y to predict
- We intend to find a good
  $Y = f(X) + \epsilon$,
  ϵ is measurement errors or other random discrepancies between true Y and mapped Y - $f(X)$.

- A good $f(X)$ helps to:
  - Predict Y for any given new x : $f(x) = E(Y|X = x)$, where $E(Y|X = x)$ is the expected value (mean) of Y when $\mathrm{X} = x$
  - Understand which factors are important to predict Y, which are irrelevant
  - Depending on the complexity of *f()*, we may be able to tell *how* X affect Y

A simple $f(X)$ - Linear model

- The linear model is an important example of a parametric model. It is specified in terms of p + 1 parameters, $\beta_0, \beta_1, \ldots \beta_p$.

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_p X_p$$

- Although it is almost never correct, linear regression serves as a good and interpretable approximation to the true function $f(X)$
- Although it seems overly simplistic, linear regression is extremely useful conceptually and practically

With one predictor X

- We assume a model

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

where β_0 and β_1 are two unknown constants that represent the *intercept* and *slope*, also known as *coefficients* or *parameters*, and ϵ is the error term.

- Given some estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model coefficients, we predict future sales using

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

where $\hat{y}$ indicates a prediction of Y on the basis of $X = x$. The *hat* symbol denotes an estimated value.

An Example

- Problem: Can we predict sales based on advertising spendings?

- Data: the sales of a product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: TV, radio, and newspaper.

- Model: Linear model

TV	radio	newspaper	sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5
180.8	10.8	58.4	12.9
8.7	48.9	75	7.2
57.5	32.8	23.5	11.8
120.2	19.6	11.6	13.2
8.6	2.1	1	4.8
199.8	2.6	21.2	10.6
66.1	5.8	24.2	8.6
214.7	24	4	17.4
23.8	35.1	65.9	9.2
97.5	7.6	7.2	9.7
204.1	32.9	46	19
195.4	47.7	52.9	22.4
67.8	36.6	114	12.5
281.4	39.6	55.8	24.4
69.2	20.5	18.3	11.3
147.3	23.9	19.1	14.6
218.4	27.7	53.4	18
237.4	5.1	23.5	12.5
13.2	15.9	49.6	5.6
228.3	16.9	26.2	15.5
62.3	12.6	18.3	9.7
262.9	3.5	19.5	12
142.9	29.3	12.6	15
240.1	16.7	22.9	15.9
248.8	27.1	22.9	18.9
70.6	16	40.8	10.5
292.9	28.3	43.2	21.4

How to find parameter values? – Least squares

The error / residual: the difference between true y and predicted y

- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the ith value of X. Then $e_i = y_i - \hat{y}_i$ represents the ith *residual*
- We define the *residual sum of squares* (RSS) as

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

or equivalently as

$$\text{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$$

Minimize the sum of error squares by choosing betas

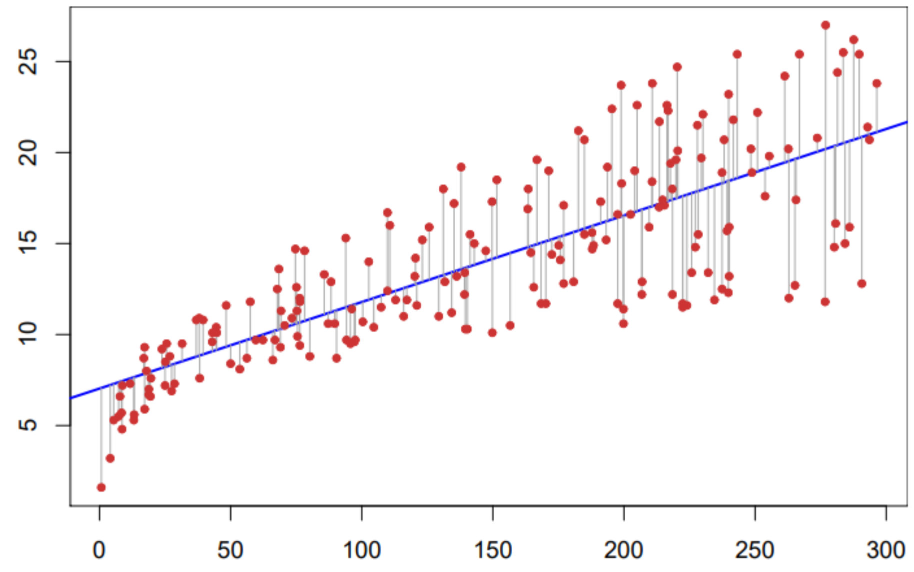
- The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS. The minimizing values can be shown to be

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

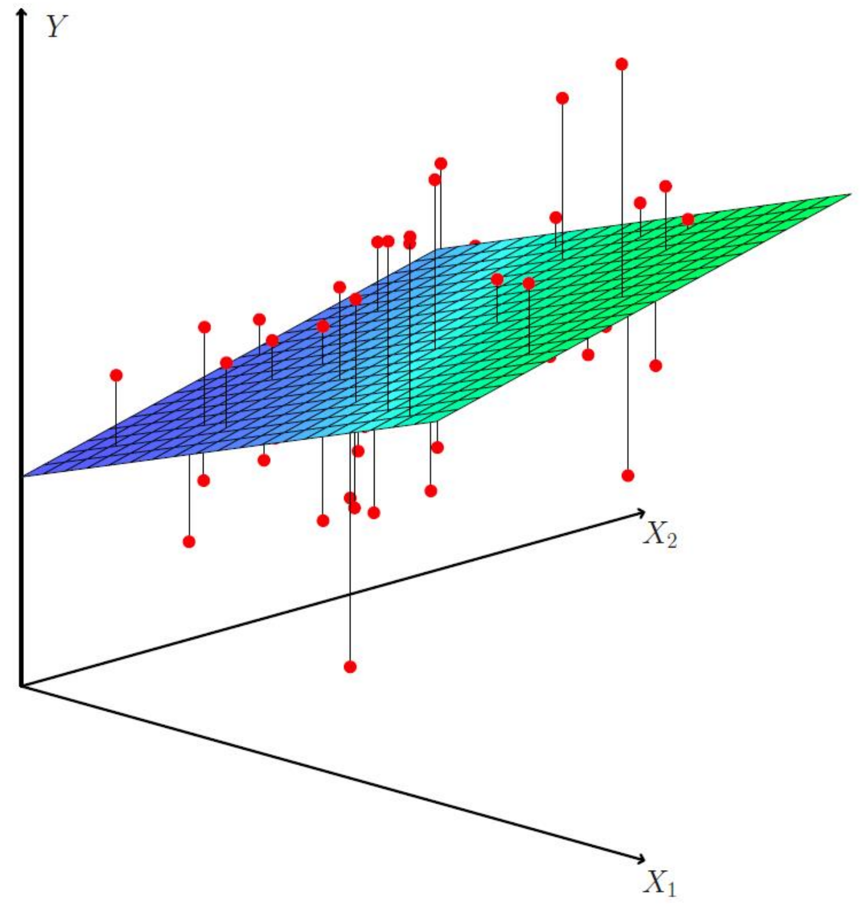
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where $\bar{y} \equiv \frac{1}{n} \sum_{i=1}^n y_i$ and $\bar{x} \equiv \frac{1}{n} \sum_{i=1}^n x_i$ are the sample means.

As shown in the graph



- Minimize the sum of squared errors (vertical distances between each observation, as shown in red, and the line).



Minimize the sum of squared errors (vertical distances between each observation, as shown in red, and the plane).

The output for a single variable

		Coefficient	Std. error	t-statistic	p-value
β_0	Intercept	7.0325	0.4578	15.36	< 0.0001
β_1	TV	0.0475	0.0027	17.67	< 0.0001

The coefficient of TV advertising: An increase of 1 unit in the TV advertising budget is associated with an increase in sales by 0.0475 units.

TV advertising: 1 unit = 1000 dollars;
Sales: 1 unit = 1000

The estimated standard deviation of the slope: Roughly, there is a 95% chance that the true slope lies within 2 standard deviations, i.e., between 0.0421 and 0.0526

The probability of observing such t-statistic given β_1 to be 0

Q1. Which factors are important in predicting Y?

H_0 : There is no relationship between X and Y

versus the *alternative hypothesis*

H_A : There is some relationship between X and Y.

- Mathematically, this corresponds to testing

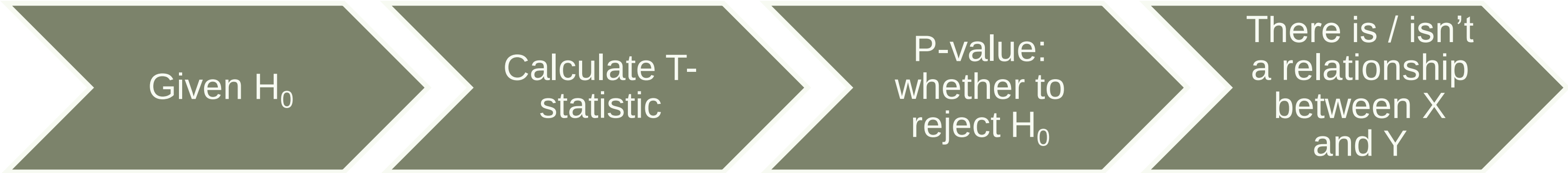
$$H_0 : \beta_1 = 0$$

versus

$$H_A : \beta_1 \neq 0,$$

since if $\beta_1 = 0$ then the model reduces to $Y = \beta_0 + \epsilon$, and X is not associated with Y.

- The outcome of the test exhibits in p-values: if p-value is less than 5%, we are confident to reject the null hypothesis.
- That means, the alternative hypothesis is correct.



Q2. How does each factor affect Y?

- We interpret β_1 as the average effect on Y of a one unit increase in X
- For multiple linear regression, we interpret β_j as the average effect on Y of a one unit increase in X_j, holding all other predictors fixed.
- For instance, the impact of radio advertising,

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

For a given amount of TV and newspaper advertising, spending an additional $1,000 on radio advertising is associated with approximately 189 units of additional sales.

Q3. How well does the model fit the data?

- We compute the *Residual Standard Error*

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

where the *residual sum-of-squares* is $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

- *R-squared* or fraction of variance explained is

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

where $\mathrm{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the *total sum of squares*.

- R-squared (R^2) measures the proportion of variability in Y that can be explained using X
- The larger, the better

Comparisons of R^2

Model	R^2	RSE
TV	0.61	3.26
TV + Radio	0.89719	1.681
TV + Radio + Newspaper	0.8972	1.686

- Newspaper provides no real improvement in model fit but increases RSE (residual standard error).
- R^2 is always larger with more variables
- Adjusted R^2 takes that into consideration and should be a better indicator for variable selection.
- So, the best model is TV + Radio

Q4. How to make predictions of Y?

- It is straightforward to predict Y using the estimated coefficients and the model

- F $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p.$ l:

    - Estimated sales: 2.939 + 0.046*44.5 + 0.189 * 39.3 + (-0.001) * 45.1 = **12.37**

    - Observed sales: **10.4**
- Errors could come from

    - The estimated coefficients have errors

    - The linear model only approximately captures the relationship between Xs and Y

    - Even with a perfect model, there are random errors we cannot control

Intervals to quantify the uncertainty

- For given values of Xs, we can estimate two types of Y values:
  - E.g., $100,000 TV, $20 radio
- Confidence interval (95%)
  - Estimated average Y value over many markets
  - [10,985, 11,528]
  - 95% of such intervals will contain the true value
- Prediction interval (95%)
  - Estimated Y value for a single market
  - [7,930, 14,580]
  - 95% of such intervals will contain the true value
- As prediction interval is more restricted, the interval is always wider than the confidence interval
  - Reflecting the uncertainty is larger for predicting sales of a single market than the average of many markets

1. Qualitative variables

- Examples
  - Ownership (own/not own a house)
  - Education levels (High school, undergraduate, graduate, professional)
  - Product categories (Food, Electronics, Appliances, etc.)
  - …
- Solution:
  - Dummy variables: indicators we generate based on the values of the qualitative variable

A qualitative variable with two levels

- Ownership: a person owns a house vs. a person does not owns a house
- We can create a dummy variable takes two values: 1 – owns a house; 0 – does not own a house

$$x_i = \begin{cases} 1 & \text{if } i\text{th person owns a house} \\ 0 & \text{if } i\text{th person does not own a house,} \end{cases}$$

- The estimation function turns into

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person owns a house} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person does not.} \end{cases}$$

- The coefficient β_1 indicates the difference between people who own a house vs. people who do not own a house

	Coefficient	Std. error	t-statistic	p-value
Intercept	509.80	33.13	15.389	< 0.0001
own[Yes]	19.73	46.05	0.429	0.6690

  - Average credit balance (the dependent variable) for people who do not own a house: 509.80
  - Average credit balance for people who own a house will be β_1 more, 19.73 more

A qualitative variable with more than levels

- We can create additional dummy variables
- For example, variable ‘region’ has three levels: East, West and South
- We can create the following dummy variables

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is from the South} \\ 0 & \text{if } i\text{th person is not from the South,} \end{cases} \qquad x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is from the West} \\ 0 & \text{if } i\text{th person is not from the West.} \end{cases}$$

- There will always be one dummy variable fewer than the number of levels (n – 1). The level with no dummy variable - East in this example - is known as the baseline.
- The level selected as the baseline category is arbitrary, and the final predictions for each group will be the same regardless of this choice.
- The estimation function will be

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is from the South} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is from the West} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is from the East.} \end{cases}$$

More than two levels - Continued

- The results

	Coefficient	Std. error	t-statistic	p-value
Intercept	531.00	46.32	11.464	< 0.0001
region[South]	−18.69	65.02	−0.287	0.7740
region[West]	−12.50	56.68	−0.221	0.8260

- The average credit card balance for individuals from the East (baseline) is β_0
- The difference in the average balance between people from the South versus the East is β_1. That is
  - People from the South have averagely $18.69 less in credit balance than people from the East
- The difference in the average balance between those from the West versus the East is β_2
- The choice of dummy variables does not affect the final predictions, but does affect the coefficient estimations
- To determine whether a qualitative variable has a relationship with the dependent variable, we do not look at the p-value of a single dummy variable, but an F-test to examine all dummy variables

2. Interaction

- Example: synergy effect in marketing
  - The linear model states that the average increase in sales associated with a one-unit increase in TV is always β_1, regardless of the amount spent on radio.
  - In reality: spending money on radio advertising actually increases the effectiveness of TV advertising.
  - For instance, the impact of $100,000 spending on TV will be greater if we also spend on radio – the synergy effect.
- Statisticall $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon.$ eraction terms.
  - Previous model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \underbrace{\beta_3 X_1 X_2}_{\text{Interaction term}} + \epsilon.$$

  - After considering the synergy

Interaction - continued

- The coefficient of TV changes to $\tilde{\beta}_1 = \beta_1 + \beta_3 X_2$

$$
\begin{aligned}
Y &= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon \\
&= \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon
\end{aligned}
$$

- When the spending on radio (X_2) increases, the effectiveness of TV spending ($\tilde{\beta}_1 = \beta_1 + \beta_3 X_2$ becomes larger (given β_3 is positive).
- We can interpret β_3 as the increase in the effectiveness of TV advertising associated with a one-unit increase in radio advertising (or vice-versa)

	Coefficient	Std. error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

The R^2 for this model is 96.8 %, compared to only 89.7% for the model that predicts sales using TV and radio without an interaction term.

Interaction – Hierarchy principle

- The hierarchical principle states that if we include an interaction($X_1 X_2$) in a model, we should also include the main effects (X_1, and X_2) even if the p-values associated with their coefficients are not significant.