IFI 9000 Analytics Methods
Linear and Logistic Regression

by **Houping Xiao**

January 12^{th}, 2021

Some Basic Probability Overview

- For a continuous random variable X we often define a probability density distribution $f_X(x)$ where $\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)dx$
- A random variable X is normally distributed with mean μ and variance σ^2 (denoted as $\mathrm{N}(\mu, \sigma^2)$), when

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{1}$$

- Sum of normals: if $x_1, x_2, \cdots, x_n$ are normal (**not necessarily independent**), the weighted sum $\alpha_1 x_1 + \cdots + \alpha_n x_n$ is also normal
- Expectation of the weight sum: if $x_1, x_2, \cdots, x_n$ are random variables with mean $\mathbb{E}(x_i) = \mu_i$, then for constants α_i:

$$\mathbb{E}(\alpha_1 x_1 + \alpha_n x_n) = \alpha 1 \mu_1 + \cdots + \alpha_n \mu_n \tag{2}$$

- Variance of weighted sum: if $x_1, x_2, \cdots, x_n$ are **independent** random variables with variance $Var(x_i) = \sigma_i^2$, then

$$Var(\alpha_1 x_1 + \cdots + \alpha_n x_n) = \alpha_1^2 \sigma_1^2 + \cdots + \alpha_n^2 \sigma_n^2 \tag{3}$$

A Brief Review of Hypothesis Testing

Hypothesis Testing

- A hypothesis is often a conjecture about one ore more populations
- To prove a hypothesis is true we need to examine all the population which is often not practical, **instead we take a random sample** and assess if we have enough evidence to support a hypothesis in probability
- **Example**: All human beings respond well to a specific treatment
    - Instead of testing it on all people on the planet, we look into a fraction of people who are infected by the decease and see if the treatment works
- Since we are focused on a limited sample, we can only state our confidence about the conjecture in probability

Hypothesis Testing

- We call H_0 as the **null** hypothesis and refer to H_1 as the **alternative** hypothesis
- The hypothesis we want to test is if H_1 is likely to be true
- Usually, the equality hypothesis is chosen to be the null hypothesis
  - For many problems that we encounter in this course, Hypothesis Testing is simply testing the chances of a random variable to be in region defined by H_0 or H_1

Hypothesis Testing

- Hypothesis testing is often formulated in terms of two hypotheses
  - H_0: the null hypothesis
  - H_1**: the alternative hypothesis**
- You decide to make a claim about the null hypothesis H_0 in probability at least $1 - \alpha$
- **Example**: We are 90% confident that this drug works on patients with xxx decease
- α: measure how confident you want to make the claim
  - determined by you

Hypothesis Testing

Usually, one of the following two cases happens:

- **Reject** H_0 and **accept** H_1, as we have enough evidence to support H_1
- **Fail to reject** H_0, as we don't have enough evidence to support H_1
  - H_0 may be false, but the data is not enough to reject it

Hypothesis Testing: Types of Error

	Reject H_0 (accept H_1)	Fail to reject H_0
H_0 is True	Type I error: α	Correct: $1 - \alpha$
H_1 is True	Correct: $1 - \beta$	Type II error: β

- α is a small number (e.g. 0.01, 0.05) that we determine and is called the significance level
  - the probability to make type I error
- We decide on how confident we want to make a claim in favor of H_0 and $1 - \alpha$ is our confidence about this
- *We won't focus on type II error here*

Hypothesis Testing Example

In the context of our linear regression problem, we are interested in hypothesis testing problems on the basis of samples

Example: There is a normal distribution with variance 1 and unknown μ. Below list 10 independent samples x_i of this distribution:

$-0.04450595, -0.48165, 0.09475972, 0.83689004, -1.4314154$,
-1.12870336, 0.68414548, 0.54675891, -0.2334923 , -0.5824023

The sample mean is

$$\frac{x_1 + \cdots + x_{10}}{10} = -0.17396 \tag{4}$$

A hypothesis testing: $H_0 : \mu = 0$ vs $H_1 : \mu \neq 0$ (two sided test)

Hypothesis Testing Example

Solution: Check the random variable

$$\bar{x} = \frac{x_1 + \cdots + x_{10}}{10} \tag{5}$$

Then $\bar{x} \sim \mathcal{N}(\mu, 0.1)$ Why?. Further, $z = \frac{\bar{x}-\mu}{\sqrt{0.1}} \sim \mathcal{N}(0,1)$ Why?.

- z is referred to as the test statistic

p-value: is a useful quantity in the analysis of the test and is the probability of obtaining a result equal or "more extreme" than what we have observed, given that the null hypothesis is true. In the case of this example: $z^* = \frac{\bar{x}-\mu}{\sqrt{0.1}} = -0.5501$

$$\text{p-value} = \mathbb{P}\left(z > |z^*|\right) + \mathbb{P}\left(z < -|z^*|\right) \tag{6}$$

$$= \mathbb{P}\left(z > 0.5501\right) + \mathbb{P}\left(z < -0.5501\right) = 0.5823 \tag{7}$$

Hypothesis Testing Example

Assume that the significance level is $\alpha = 0.05$

- If p-value $\leq \alpha$: reject H_0 and accept H_1
- If p-value $> \alpha$: do not reject H_0

In this example, p-value$= 0.5823 > 0.05$, so we cannot reject the hypothesis $\mu = 0$

- If the value of $\bar{x} = -0.17396$ was calculated based on 200 samples (sample size increased), then $z^* = -1.7396$ and

$$\text{p-value} = \mathbb{P}\left(z > 1.7396\right) + \mathbb{P}\left(z < -1.7396\right) = 0.0139 < 0.05 \qquad (8)$$

  then we were able to reject H_0

- In other words we are more than 95% confident that it is not possible to take the sample mean over 200 random number of mean $\mu = 0$ and variance 1, and get a value as far from 0 as -0.17396

(See the code)

Hypothesis Testing Example: Unknown Variance

Suppose in the previous example, we did not know σ and instead of working with the standard random normal variable $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ we work with the random variable

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \quad \text{where} \quad s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{9}$$

- s an unbiased estimate of σ
- t Student's t distribution
- similar procedure for p-value calculation

Take Away from Hypothesis Testing

- Some independent samples from a distribution
- Make some guesses about the data
- Formulate a hypothesis test, H_0 and H_1
- Determine the confidence level, $1 - \alpha$
- Find a test statistic
- Calculate the p-value:
  - If p-value $\leq \alpha$: reject H_0 and accept H_1
  - If p-value $> \alpha$: do not reject H_0

Now Lets Start Linear Regression!

Introduction to Linear Regression

- An ideal regression function $f(\boldsymbol{x})$
  - this regression function was the actual function behind our data generation
  - observations y were in the form

$$y = f(\boldsymbol{x}) + \epsilon \tag{10}$$

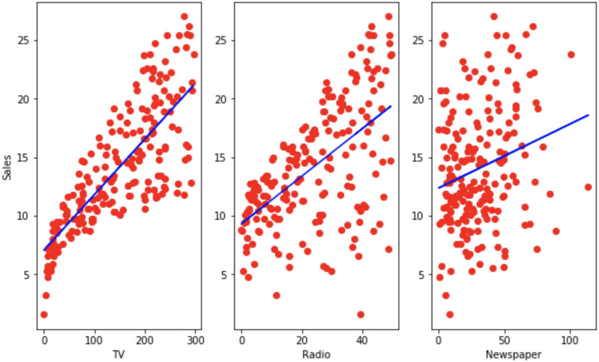
where $\boldsymbol{x} = (x_1, \cdots, x_p)^\top$

- estimate $\hat{f}$ to approximate the **unknown** $f(\boldsymbol{x})$
- **Linear regression**: estimate f in the following form

$$\hat{f}(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{11}$$

Introduction to Linear Regression

- A simple model with $p = 1$



$$\text{Sales} = \beta_0 + \beta_1 \text{TV/Radio/Newspaper} + \epsilon$$

or

$$y = \beta_0 + \beta_1 x + \epsilon$$

where x is the one feature

More on Simple Linear Regression

Consider

$$y = \beta_0 + \beta_1 x + \epsilon$$

- β_0 and β_1 are unknown coefficients that represent the intercept and slope, respectively
- Based on the available $(x_1, y_1), \cdots, (x_n, y_n)$, estimate $\hat{\beta}_0$ and $\hat{\beta}_1$
- using $\hat{\beta}_0$ and $\hat{\beta}_1$ to make prediction

$$\hat{y}(x^*) = \hat{\beta}_0 + \hat{\beta}_1 x^*$$

Determining the Model Coefficients

- Remember that we had samples $(x_1, y_1), \cdots, (x_n, y_n)$, the $\hat{f}$ are derived from solving the following optimization problem

$$\min \sum_{i=1}^{n} (y_i - \hat{f}(x_i))^2$$

- Using the simple model β_0 and β_1 are determined such that the **Residual Sum of Squares** (RSS)

$$RSS = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

  is minimized

Determining the Model Coefficients

- Taking the derivative w.r.t. β_0 and β_1 and letting them to be zero, we have

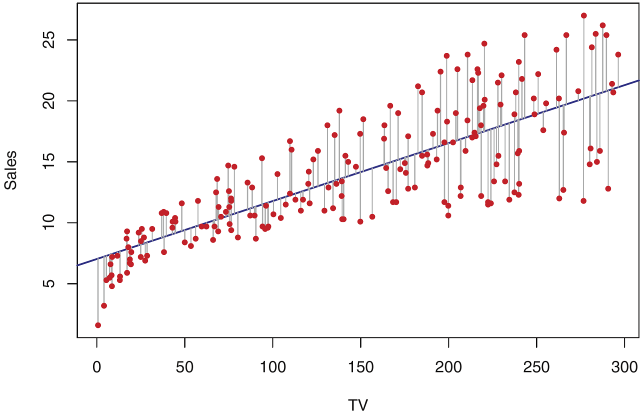
$$\hat{\beta}_0 = \hat{y} - \hat{\beta}_1 \bar{x} \quad \text{and} \quad \hat{\beta} = \frac{\sum_{i=1}^n x_i y_i - n x \bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$

- Using the following equations:

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

How Well Did We Do the Fit?



Now that we have our fit we would like to address few questions about it!

(See the code)

What is the Confidence Interval for the Coefficients?

- Note that $\hat{\beta}_0$ and $\hat{\beta}_1$ are both normally distributed when the noise is normally distributed
- Consider the 95% confidence intervals that cover the true β_0 and β_1
- $$SE(\hat{\beta}_0)^2 = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right), \quad SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$
- The 95% confidence intervals for β_0 and β_1 are

$$\left[\beta_0 - 2SE(\hat{\beta}_0)^2, \beta_0 + 2SE(\hat{\beta}_0)^2\right], \quad \left[\beta_1 - 2SE(\hat{\beta}_1)^2, \beta_1 + 2SE(\hat{\beta}_1)^2\right]$$

(See the code)

Is There a Relationship Between x and y?

- We want to know whether there is really a relationship between x and y or if the fit is useless?
- **Hypothesis testing for** β_1:
  - H_0: $\beta_1 = 0$
  - H_1: $\beta_1 \neq 0$
- Test statistic: $t = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$
- t follows t-distribution, and get the p-value
  - If p-value $\leq \alpha$: reject H_0 and accept H_1
  - If p-value $> \alpha$: do not reject H_0
- For the example provided p-value$= 2 \times 10^{-}16$ and we reject H_0 and accept H_1

How Well does the Model Explain the Data?

- Check
$$R^2 = 1 - \frac{RSS}{TSS}$$
where TSS is the Total Sum of Squares
$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2, \quad TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$$
- For general regression problems (not just the simple one with only one feature) R^2 measures the proportion of variability in y that can be explained by x
- R^2 close to 1 indicates that our model explains a large proportion of the response variability, and R^2 close to zero indicates that our model cannot explain much of the variability in response

(See the code)

Multiple Linear Regression

Multiple Linear Regression

- If we have multiple features $x_1, \cdots, x_p$, the fit could be

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

For instance, $Sales = \beta_0 + \beta_1 \cdot TV + \beta_2 \cdot Radio + \beta_3 \cdot Newspaper + \epsilon$

- Suppose that we have n training samples $(\boldsymbol{x}^{(1)}, y_1), \cdots, (\boldsymbol{x}^{(n)}, y_n)$ and

$$\boldsymbol{x}^{(1)} = \begin{pmatrix} x_1^{(1)} \\ x_2^{(1)} \\ \vdots \\ x_p^{(1)} \end{pmatrix}, \quad \boldsymbol{x}^{(2)} = \begin{pmatrix} x_1^{(2)} \\ x_2^{(2)} \\ \vdots \\ x_p^{(2)} \end{pmatrix}, \quad \cdots, \quad \boldsymbol{x}^{(n)} = \begin{pmatrix} x_1^{(n)} \\ x_2^{(n)} \\ \vdots \\ x_p^{(n)} \end{pmatrix}$$

- The objective function to minimize is the following squared error

$$RSS = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_1^{(i)} - \beta_2 x_2^{(i)} \cdots - \beta_p x_p^{(i)} \right)^2$$

- In matrix manner

$$RSS = (\boldsymbol{y} - \boldsymbol{X}\beta)^\top (\boldsymbol{y} - \boldsymbol{X}\beta)$$

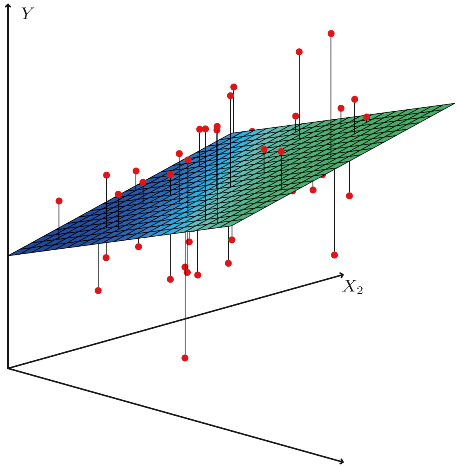
where

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \boldsymbol{X}_{n \times (p+1)} = \begin{pmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \cdots & x_p^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \cdots & x_p^{(2)} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_1^{(n)} & x_2^{(n)} & \cdots & x_p^{(n)} \end{pmatrix}$$

Multiple Linear Regression

- Similar to what we did before we can set $\frac{\partial RSS}{\partial \boldsymbol{\beta}} = 0$ and get

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$



Are the Features and Response Related?

- Whether features $x_1, \cdots, x_p$ is useful in predicting the response
- **Hypothesis testing**:
  - $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$
  - H_1 : at least one $\beta_i \neq 0$
- Test statistic:

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}$$

where $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ and $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$

  - F follows F distribution, and calculate the p-value
    - If p-value $\leq \alpha$: reject H_0 and accept H_1
    - If p-value $> \alpha$: do not reject H_0
  - Or check the value of F
    - If F is much larger than 1, reject H_0
    - If F is very close to 1, do not reject H_0

(See the code)

Multiple Linear Regression

Now that we have our fit, again we would like to address few questions about it!

(See the code)

Assessing the P-values and Correlations Among Features

- Sometimes features are correlated and the contribution of one feature can be taken care of by the others

	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

	`TV`	`radio`	`newspaper`	`sales`
`TV`	1.0000	0.0548	0.0567	0.7822
`radio`		1.0000	0.3541	0.5762
`newspaper`			1.0000	0.2283
`sales`				1.0000

(See the code)

What are the Best Selection of Features?

- As noted sometimes some features can be redundant and we would like to find the best subset of features that predicts well and is not redundant
- In general this problem is "NP-hard" (computationally very hard) and we need to assess 2^p models
- Some heuristics to do this that we will see later:
  - Forward selection: model p regressions each with only one feature, pick the one with least RSS, repeat it with selected feature and combination of others, etc
  - Backward selection: Start with all features and remove variable with largest p-value, run a new regression, remove variable of largest p-value, etc

How to Handle Categorical Features?

- Sometimes our features do not take numerical values, instead they take categorical values
- **Example**: In a regression problem we have a feature called ethnicity, which takes possible values of Asian, Caucasian, African-American
- We can introduce 2 dummy variables (features) eA; eC
    - eA $= 1$; eC $= 0$ if Asian
    - eA $= 0$; eC $= 1$ if Caucasian
    - eA $= 0$; eC $= 0$ if African-American
- Basically, for every categorical feature that has L levels, we need to define $L - 1$ dummy variables

Can We Only Fit Flat Curves with Linear Regression

Multiple linear regression

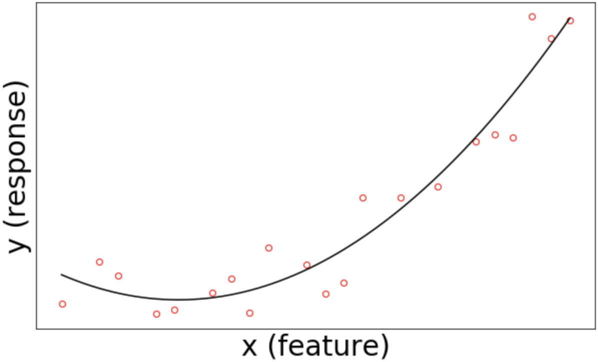
$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

- one might get the impression that linear regression is only good for fitting at surfaces (linear manifolds)
- **Handle nonlinearity**: introduce powers of a feature, e.g., $x_1, x_2, \cdots$ or cross termsbetween the features, e.g., $x_1 x_2, x_1 x_2 x_3$, etc
- Hard to determine the degree

Can We Only Fit Flat Curves with Linear Regression

- **Example**: For a problem with only one feature shown in the figure, we can use the regression

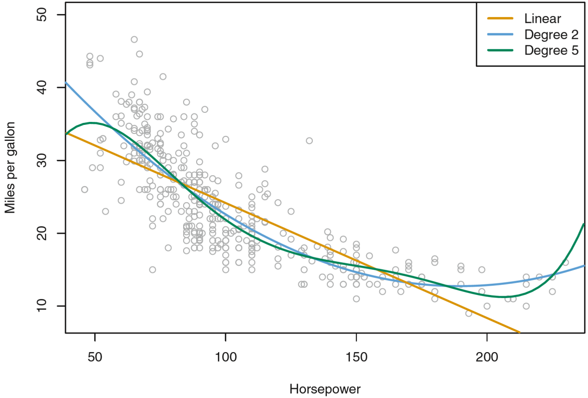
$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$



Can We Only Fit Flat Curves with Linear Regression

- **Example**: Regressing Mile per Gallon in terms of the Horse Power

$$\text{Miles per gallon} = \beta_0 + \sum_{i=1}^{p} \beta_i (\text{horse power})^i$$



The End