BT1101 Introduction to Business Analytics

Hands-on Tutorial of Regression

Dependent Variable : Y Independent Variables : $X_1, X_2, \cdots, X_d$.

- DV is discrete : Logistic Regression.
- DV is continuous : **Linear Regression**.

Linear Regression

Given Dataset $D = \{(\vec{X_t}, y_t)\}_{t=1}^{n}$:

random variable!

$$Y = \underbrace{b_0 + b_1X_1 + b_2X_2 + \cdots + b_dX_d}_{\text{non-random}} + \varepsilon \quad \text{for some } b_1, b_2 \cdots, b_d.$$

noise $\sim N(0, \delta^2)$ (random variable!)

Question: What's the distribution of the random variable Y? $N(\underbrace{b_1X_1+b_2X_2+\cdots+b_dX_d+b_0}_{\text{mean}}, \underbrace{\delta^2}_{\text{variance}})$

Steps :

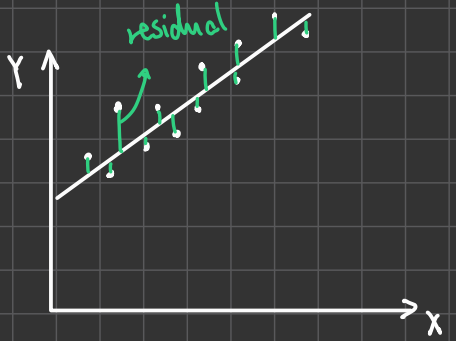
① plotting the points (X_t, y_t), seems like a linear relation btw X and Y

if not, some techniques should be applied : log(variable)

② We want to find a linear function (= to find $(b_0, b_1, \cdots, b_d)$) s.t. $\sum_{t=1}^{n} (y_t - \hat{y_t})^2$ is minimized.

your model prediction : $\hat{y_t} = b_1X_1 + b_2X_2 + \cdots + b_dX_d + b_0$

the real value : y_t



③ Using R to calculate $(b_0, b_1, \cdots, b_d)$

```
summary(lm(formula = y ~ x1+x2+... + xd,
           data = dataframe you have ))
```

④ Get the summary table and interpret the meanings

```
Call:
lm(formula = y ~ x, data = df1)

Residuals:
     Min       1Q   Median       3Q      Max
-1.0781 -0.5736  0.1260  0.3071  1.5452

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.26117    0.46171  -4.897 0.000851 ***
x            2.10376    0.07804  26.956 6.44e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8185 on 9 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.9878,    Adjusted R-squared:  0.9864
F-statistic: 726.6 on 1 and 9 DF,  p-value: 6.442e-10
```

Residual Statistics

$\hat{b}_0$ $\hat{b}_1$

hypothesis testing: what is the Ho?

Goodness-of-fit

Coefficients:

① Interpretation of $\hat{b}_0$ and $\hat{b}_1$

② null hypothesis: The parameter is equal to zero.

p-value < 0.05 means:

Residuals:

each sample has a residual (unexplained variance).

We are interested in: the residuals are big or not?

⇒ Compare the Residual standard error with the value of Y. It's relative.

Goodness of fit: Is linear model an appropriate model for this D?

① R^2 (coefficient of determination): the proportion of the variance of Y can be explained by X

- $R^2 \in [0,1]$ the bigger, the better.
- Ajusted R-squared in the summary table.

NOTE: Total variance $\sum(y_i-\bar{y})^2$

explainable variance (the percentage larger, the better)

unexplained variance (Residuals)

② F test:

- Ho: the model has no predictive power (all the b's are zero)
- When F is large and p-value is very small (<.05), we should reject Ho. ⇒ the model is useful.

Categorical Independent Variable: $\text{UmbrellaSales} = b_0 + b_1 \cdot \text{Weather}$. $\text{Weather} \in \{\text{Sunny}, \text{Rainy}\}$

We need a dummy variable "Rainy" in our regression model $\quad \text{rainy} = \begin{cases} 0, & \text{weather is Sunny} \\ 1, & \text{weather is Rainy.} \end{cases}$

if sunny: $\text{umbre.} = b_0 + 0 = b_0$

if rainy: $\text{umbre.} = b_0 + b_1 = b_0 + b_1$ interpretation!

Rmk: When weather has n possible values, we need $(n-1)$ dummy variables.

Logistic Regression: Y is discrete / Categorical (Good / Bad Client)

$$\log(\text{odds}) = \log \frac{p}{1-p} = b_0 + b_1 X_1 + \cdots + b_d X_d \quad \text{with } p = \text{the probability of success.}$$

- Interpretation: b_0, b_1, b_2
- summary (glm ($Y \sim X_1 + X_2$, family = "binomial", data = ...))