Introduction to Statistical Learning and Applications
CC2: Multiple Linear Regression

Razan MHANNA
STATIFY team, Inria centre at the University Grenoble Alpes
LEMASSON/CHRISTEN team, Grenoble Institute of Neurosciences GIN

13 February 2024

Contents

Current Section

Simple Linear Regression

- The simple linear regression model is defined by

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

- β_1 and β_0 are the parameters to be estimated by Least Squarres method
- The total sum of squarres is equal to the regression sum of squarres plus the error sum of squarres

$$SST = SSR + SSE$$

- The best case scenario when SSE=0 meaning that the variation of Y is totally explained by X perfectly linear,
- The worst case scenario is when SSR=0 where X gives no inforamtoion about y,

Simple Linear Regression

- The coefficient of determination indicates the proportion of variation of Y explained by the model.

$$r^2 = \frac{SSR}{SST}$$

  Assumptions on the error:

  1. E[ε_i]=0
  2. V(ε_i)=constant
  3. cov(x_i,ε_i)=0
  4. cov(ε_i,ε_j)=0
  5. $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon)$
     Ir is not crucial, if this condition holds the model falls into Gaussian Simple Linear regression category, and the estimatots are by then Gaussian too.

- The standard deviation of the variation of observations around the regression line is estimated by $s_\varepsilon = \sqrt{\frac{SSE}{n-k-1}}$ where n is the sample size and k the number of independent variables in the model.

Current Section

Multiple Linear regression

$$Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + ... + \beta_p X_p + \varepsilon$$

X= $\begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$

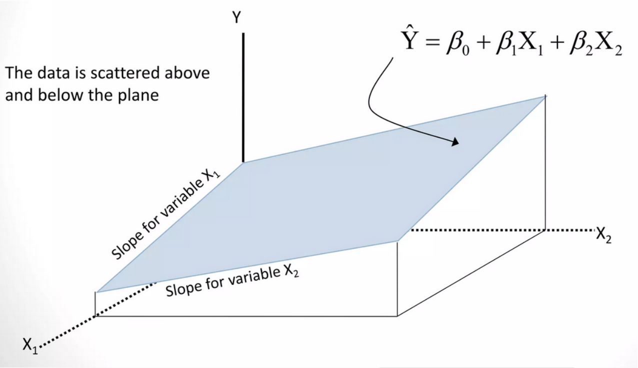
Y=$(y_1\, y_2 \dots \; \dots y_n)^T$

Matrix notation

$$\text{Y=}X\beta + \varepsilon$$
$$Y \in \mathbb{R}^n;\; X \in \mathbb{R}^{n\times(p+1)};\; \varepsilon \in \mathbb{R}^n;\; \beta \in \mathbb{R}^{p+1}$$

n is the sample size (number of observations) and p is the number of predictors.

Multiple Linear regression



Multiple Linear regression

- We want to predict $\hat{Y} = X\hat{\beta}$, by following the same method as in simple linear regression, our estimator will minimize $\mathbb{E}_{\mathbb{XY}}\left[(Y - f(X))^2\right]$.
- This means finding the plane or hyper-plane that minimizes the error between the y values in the observations set and the y values that the plane or hyper-plane passes through.
- In other words, we want the plane or hyper-plane that "best fits" the training samples.

$\hat{\beta} = (X^T X)^{-1} X^T y$

If $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon)$ *then* $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (X^T X)^{-1})$

Multicollinearity: IVs shouldn't be overly correlated (e.g >.7), if so- consider removing one.

Current Section

Null Hyptothesis

The null hypothesis H_0 states that all β_i are equal to 0(no linear relationship):

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$$

The alternative hypothesis H_1 is the opposite(linear relationship):

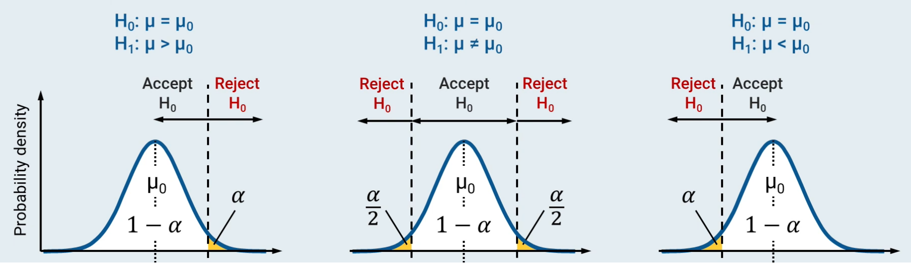
$$H_1 : \text{At least one } \beta_i \text{ is not equal to 0}$$

The rejection region R is of the form $R = \{x : T(x) \geq c\}$, where T is a test statistic and c is a critical value.
$\Rightarrow$ Hypothesis testing $\Leftarrow$ find appropriate T and c.

1. *A test statistic is a numerical value computed from sample data used to determine whether to accept or reject a null hypothesis in hypothesis testing.*
2. *p-value* $= \inf\{\alpha : T(x) \in R_\alpha\}$ *is the smallest level at which we can reject* H_0*.*

Statistical Hypothesis Test

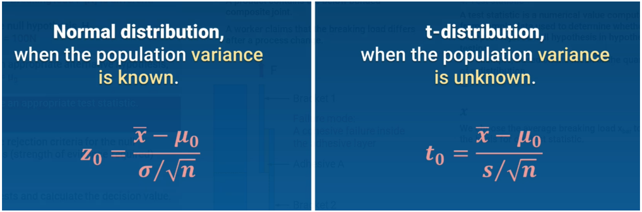
Significance mean the percentage risk to reject a null hypothesis, and its donated by α.[1]
(1-α) is the confidence interval in which the null hypothesis will exist xhen it's true.



[1] What α should we use? It is become conventional to set $\alpha = 0.05$.

Outline

Confidence Interval



- For linear regression, the 95% confidence interval for β_1 approximately takes the form $\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1)$
- Similarly, a confidence interval for ${}_0$ approximately takes the form $\hat{\beta}_0 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_0)$
- The t-statistic for the following problem

  $H_0 : \beta_1 = 0 \quad \text{versus} \quad H_a : \beta_1 \neq 0$ is $t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$

Current Section

Interpretation in R

- **lm()** command to fit a linear regression model in R,
- **summary()** command to view the output of the regression model.

Example the mtcars dataset using hp, drat, and wt as predictor variables and mpg as the response variable.
Tutorial here.

This section displays the estimated coefficients of the regression model. We can use these coefficients to form the following estimated regression equation: $mpg = 29.39 - 0.03 \times hp + 1.62 \times drat - 3.23 \times wt$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 29.394934   6.156303   4.775 5.13e-05 ***
hp          -0.032230   0.008925  -3.611 0.001178 **
drat         1.615049   1.226983   1.316 0.198755
wt          -3.227954   0.796398  -4.053 0.000364 ***
```

Coefficients

- Estimate: The estimated coefficient. This tells us the average increase in the response variable associated with a one unit increase in the predictor variable, assuming all other predictor variables are held constant.
- Std. Error: This is the standard error of the coefficient. This is a measure of the uncertainty in our estimate of the coefficient.
- t value: This is the t-statistic for the predictor variable, calculated as (Estimate) / (Std. Error).
- Pr(>|t|): This is the p-value that corresponds to the t-statistic. If this value is less than some alpha level (e.g. 0.05) than the predictor variable is said to be statistically significant.[2]

[2] If we used an alpha level of α = .05 to determine which predictors were significant in this regression model, we would say that hp and wt are statistically significant predictors while drat is not.

Residuals

This section displays a summary of the distribution of residuals from the regression model. Recall that a residual is the difference between the observed value and the predicted value from the regression model.

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.3598 -1.8374 -0.5099  0.9681  5.7078
```

The minimum residual was -3.3598, the median residual was -0.5099 and the max residual was 5.7078.

Assessing Model Fit

This last section displays various numbers that help us assess how well the regression model fits our dataset.

```
Residual standard error: 2.561 on 28 degrees of freedom
Multiple R-squared:  0.8369,    Adjusted R-squared:  0.8194
F-statistic: 47.88 on 3 and 28 DF,  p-value: 3.768e-11
```

Residual standard error: This tells us the average distance that the observed values fall from the regression line. The smaller the value, the better the regression model is able to fit the data.

The degrees of freedom is calculated as n-k-1 where n = total observations and k = number of predictors. In this example, mtcars has 32 observations and we used 3 predictors in the regression model, thus the degrees of freedom is $32 - 3 - 1 = 28$.

Assessing Model Fit

Multiple R-Squared: This is known as the coefficient of determination. It tells us the proportion of the variance in the response variable that can be explained by the predictor variables.
This value ranges from 0 to 1. The closer it is to 1, the better the predictor variables are able to predict the value of the response variable.

Adjusted R-squared: This is a modified version of R-squared that has been adjusted for the number of predictors in the model. It is always lower than the R-squared.
$R^2_{\text{adj}} = 1 - \frac{(1-R^2)(n-1)}{n-k-1}$

The adjusted R-squared can be useful for comparing the fit of different regression models that use different numbers of predictor variables.

Assessing Model Fit

F-statistic: This indicates whether the regression model provides a better fit to the data than a model that contains no independent variables. In essence, it tests if the regression model as a whole is useful.

p-value: This is the p-value that corresponds to the F-statistic. If this value is less than some significance level (e.g. 0.05), then the regression model fits the data better than a model with no predictors. When building regression models, we hope that this p-value is less than some significance level because it indicates that the predictor variables are actually useful for predicting the value of the response variable.

Current Section

Categorical Predictors

- When some inputs are categories (e.g. gender) rather than numbers (e.g. age), we need to represent the category values as numbers so they can be used in the linear regression equations.
- In one-hot encoding, we allocate each category value its own dimension in the inputs. So, for example, having three car types as predictors we allocate
  1. For Audi X_1=(1,0,0)
  2. For BMW X_2=(0,1,0)
  3. For Mercedes X_3=(0,0,1)
- y= $\beta + \beta_B X_B + \beta X_1 + \cdots + \beta_p X_p + \varepsilon$
- For one dimensional dummy varuable (binary case), we will be having two sub models with different intercepts
  1. y= $\beta + \beta X_1 + \ldots$
  2. y= $\beta + \beta_B + \beta X_1 + \ldots$