Course on

Statistical Learning

University of Rome Tor Vergata

Linear Regression Model

Simple Regression wtih R

Given the relation between `testscore` and `student teaching ratio` let's compute the OLS to estimate the regression coefficients: β_0 and β_1.

Command lines

```
mydata <- read.csv2(``caschool.csv")
View(mydata)
str(mydata)
```

Score variable and descriptive statistics

Create `Score` variable from `read score` and `math score`

Command line

```
mydata$score = (mydata$read_scr+mydata$math_scr)/2
```

Summary

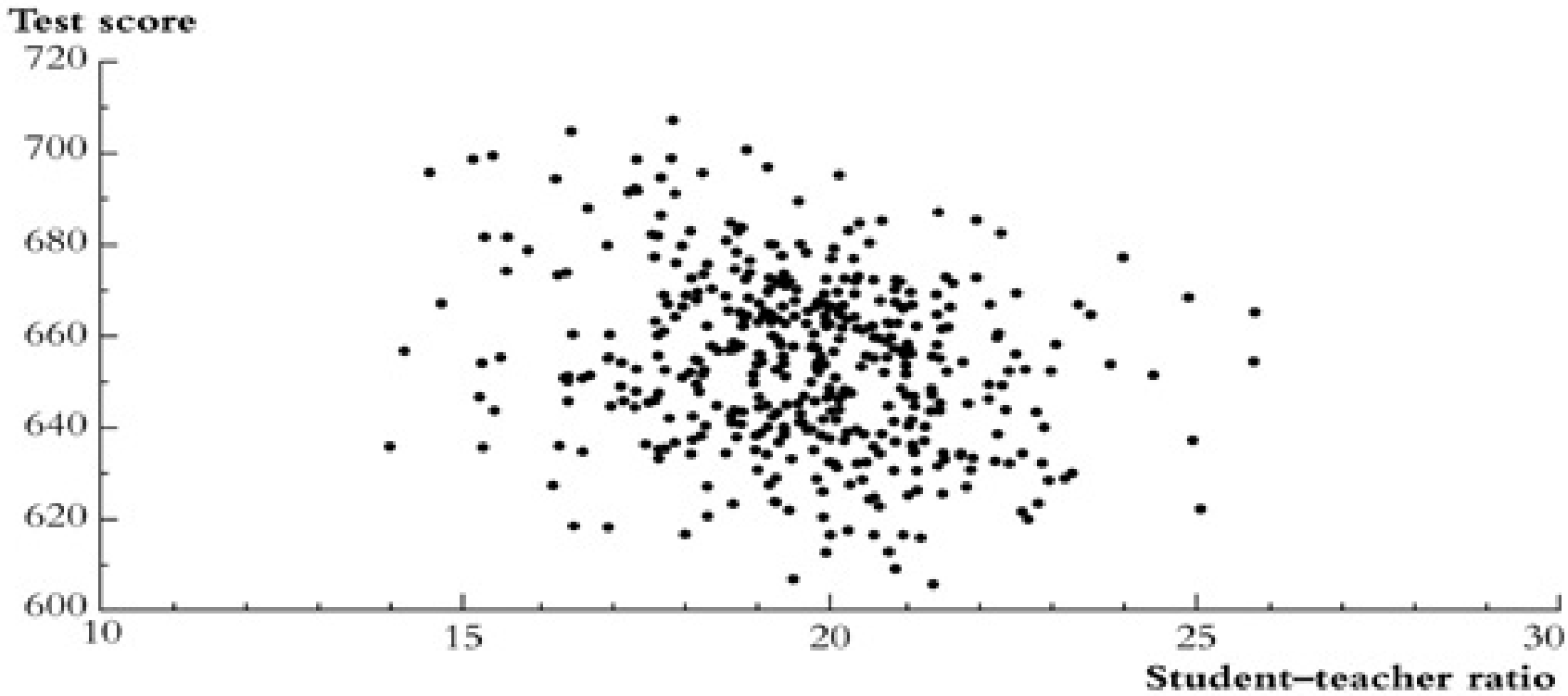
Command lines

```
summary(mydata$str)
sd(mydata$str)
var(mydata$str)
summary(mydata$score)
sd(mydata$score)
var(mydata$score)
```

Scatter plot

Command line

```
plot(mydata$str, mydata$score, xlab=‘‘STR", ylab=‘‘SCORE")
```



Mean values and correlation coefficient

Let's show in the plot the value of the two means

Command line

```
segments (mean(mydata$str), mean(mydata$score),
mean(mydata$str), min(mydata$score) - 5, lty = ‘‘dashed")
segments (min(mydata$str) - 5, mean(mydata$score),
mean(mydata$str), mean(mydata$score), lty = ‘‘dashed")
```

Let's compute the correlation coefficient

Command line

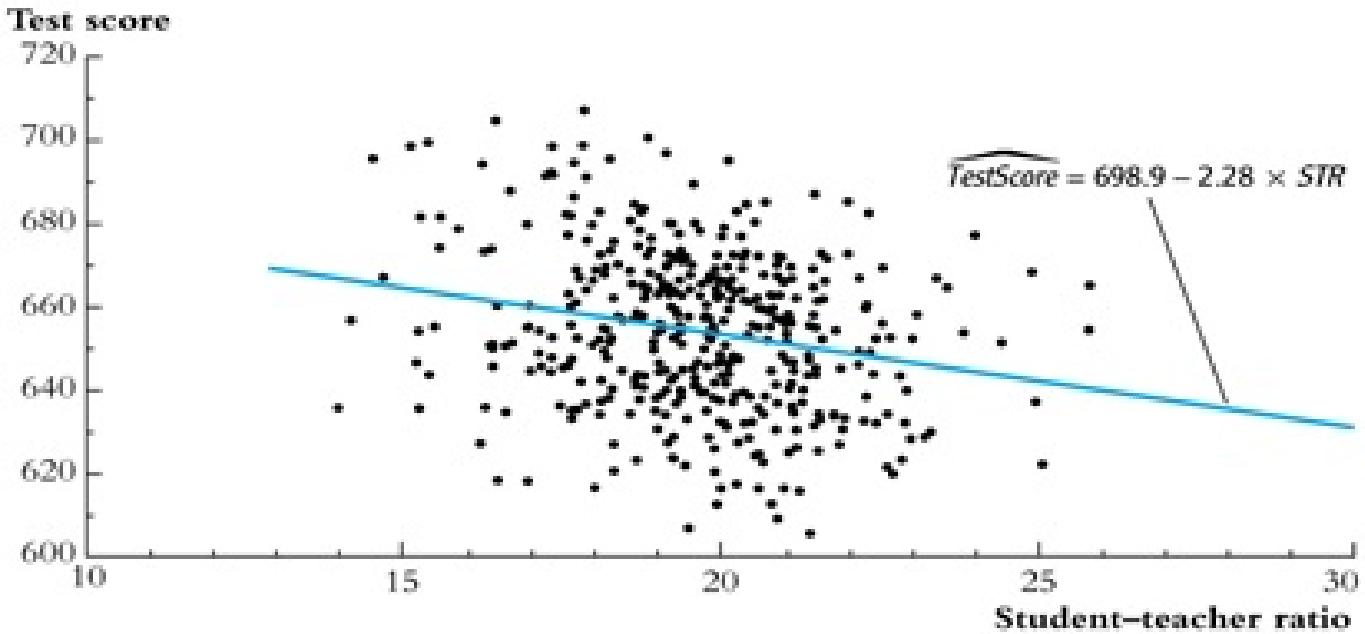
```
cor(mydata$str,mydata$score)
```

Linear regression

Let's show in the plot the regression line

Command line

```
score.lm <- lm(mydata$score   mydata$str, data=mydata)
score.lm
```



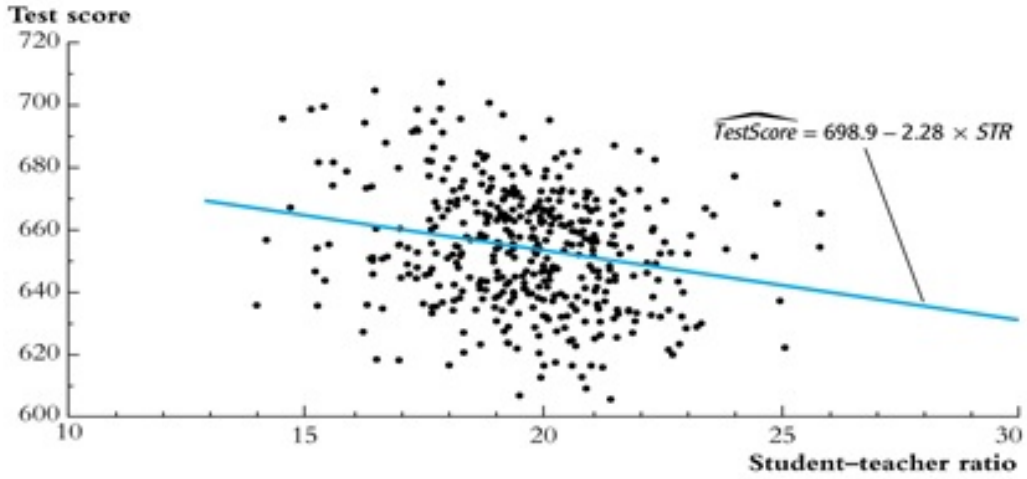
Results Interpretation

$$\texttt{Test Score} = 698.9 - 2.28 \ \texttt{STR}$$

- Districts with one more student per teacher on average have 2.28 points less than the other students ($\hat{\beta} = -2.28$)
- The intercept means that, according to this estimated line, districts with zero students per teacher would have a (predicted) test score of 698.9.

Note: This interpretation of the intercept makes no sense – it extrapolates the line outside the range of the data – here, the intercept is not economically meaningful.

Predicted values and residuals



One of the districts in the dataset is Antelope, CA, for which $\texttt{STR} = 19.33$ and $\texttt{Test Score} = 657.8$.

Given the regression line results we get:

$$\hat{Y}_{Antelope} = 698.9 - 2.28 \cdot 19.33 = 654.8 \text{ (Predicted value)}$$
$$\hat{u}_{Antelope} = 657.8 - 654.8 = 3 \text{ (Residual)}$$

OLS regression: R output

Command lines

```
score.lm <- lm(mydata$score   mydata$str, data=mydata)
score.lm
summary(score.lm)
confint(score.lm)
```

```
Regression                                      Number of obs =     420
                                                F(  1,   418) =   22.58
                                                R-squared     =  0.0489
                                                Root MSE      =  18.581

------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cons |     698.933     9.4675    73.825   0.000     680.3231    717.5428
         str |     -2.2798      .4798    -4.751   0.000     -3.22298   -1.336636
------------------------------------------------------------------------------
```

OLS regression: R output

Standard deviation

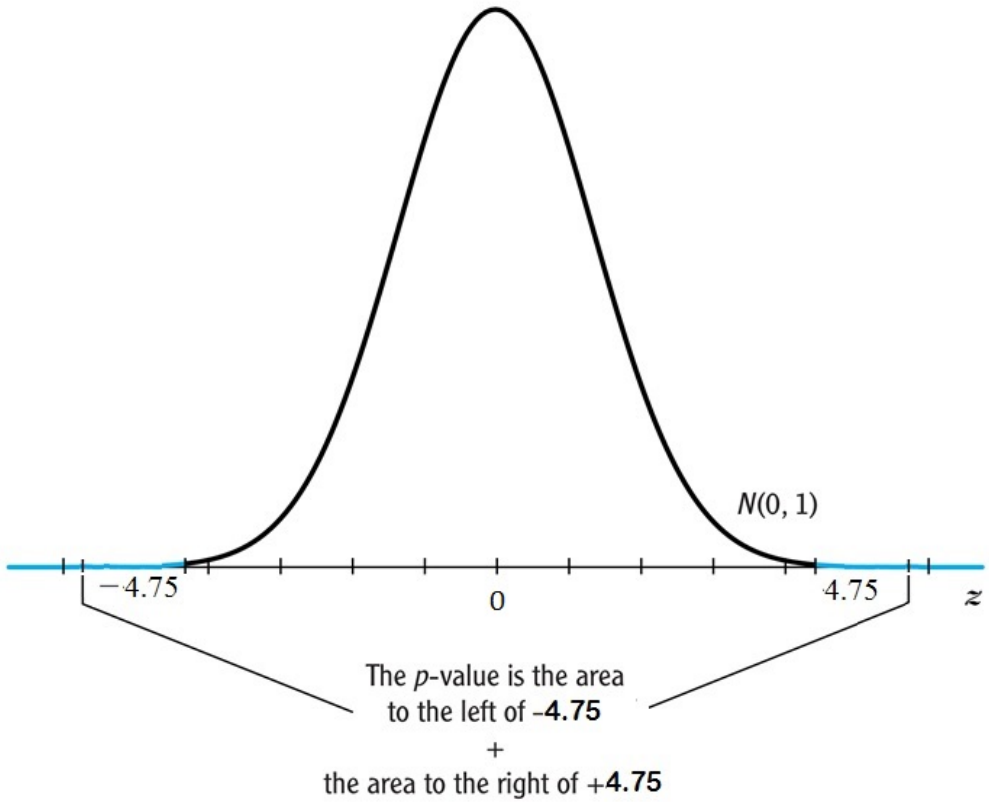
$$SE_{\beta_0} = 9.4675 \qquad SE_{\beta_1} = 0.4798$$

t-statistic testing

$$\beta_{1,0} = 0 \Rightarrow \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)} = \frac{-2.2798 - 0}{0.4798} = -4.751$$

- The 1% 2 – sided significance level is 2.58, so we reject the null at the 1% significance level. (***)
- Alternatively, we can see the p-value as well that is lower than 0.01

Test Results



The p-value based on the large-n standard normal approximation to the t-statistic is 0.000001 (10^{-6})

Confidence Interval

Given the standard deviation

$$SE_{\beta_0} = 9.4675 \qquad SE_{\beta_1} = 0.4798$$

The 95% confidence interval for $\hat{\beta}_1$ is equal to:

$$\{\hat{\beta}_1 \pm 1.96 \cdot SE(\hat{\beta}_1)\} = \{-2.28 \pm 1.96 \cdot 0.4798\} = (-3.22, -1.33)$$

- The 95% confidence interval does not include zero
- The hypothesis $\beta_1 = 0$ is rejected at the 5% level

The R^2 and the SER

```
Regression                                      Number of obs =     420
                                                F(  1,   418) =   22.58
                                                R-squared     =  0.0489
                                                Root MSE      =  18.581

------------------------------------------------------------------------------
             |               Robust
    testscr  |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cons |     698.933     9.4675    73.825   0.000     680.3231    717.5428
         str |     -2.2798      .4798    -4.751   0.000     -3.22298   -1.336636
------------------------------------------------------------------------------
```

From the results we can notice that: $R^2 = .05$, $SER = 18.6$

STR explains only a small fraction of the variation in test scores.

Does this make sense? Does this mean the STR is unimportant in a policy sense?

Graphic Analysis

To ass whether the linear model selected is appropriate. We define four plots that are also important diagnostic tools:

1. *Residuals vs Fitted:* When a linear model is appropriate, we expect
   - the residuals will have constant variance when plotted against fitted values
   - the residuals and fitted values will be uncorrelated. If there are clear trends in the residual plot, or the plot looks like a funnel, these are clear indicators that the given linear model is inappropriate

   If there are clear trends in the residual plot, or the plot looks like a funnel, these are clear indicators that the given linear model is inappropriate.

Graphic Analysis

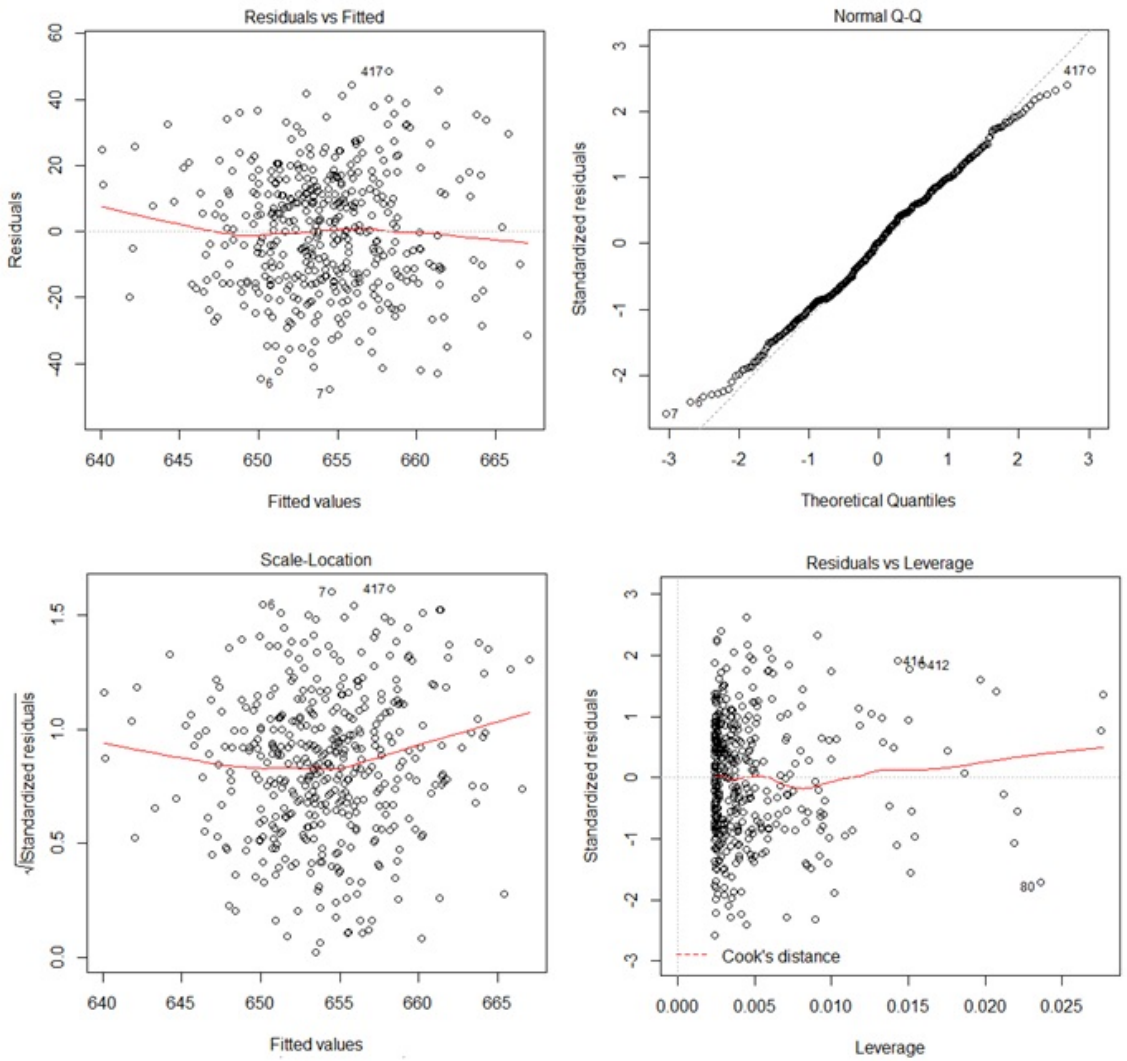
To ass whether the linear model selected is appropriate. We define four plots that are also important diagnostic tools:

2. *Normal Q-Q plot:* You can use a linear model for prediction even if the underlying normality assumptions don't hold. However, in order to let the p-values be believable, the residuals from the regression must look approximately normally distributed.
3. *Scale-location plot:* this is another version of the residuals vs fitted plot. There should be no discernible trends in this plot.

Graphic Analysis

4 *Residuals vs Leverage:* Leverage is a measure of how much an observation influenced the model fit. It's a one-number summary of how different the model fit would be if the given observation was excluded, compared to the model fit where the observation is included. Points with high residual (poorly described by the model) and high leverage (high influence on model fit) are outliers. They're skewing the model fit away from the rest of the data, and don't really seem to fit with the rest of the data.

Results


Residuals vs Fitted
Residuals
Fitted values
417
6
7
Normal Q-Q
Standardized residuals
Theoretical Quantiles
417
7
Scale-Location
√|Standardized residuals|
Fitted values
7
6
417
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
416
412
80

Comments

1. The residuals do not seem to have an obvious trend based on the `X` used (`STR`). Indeed, as we see, there is no real correlation between them.
   The fitted values vs residuals appear quite random. Based on this we would say that we have no great problems of heteroskedasticity as the residuals appear to have almost the same variance everywhere. However we will go deeper on this point with the Scale – Location plot.
2. The Q-Q plot reveal a good level of normality. Indeed the points are arranged with a certain regularity along the bisector line.

Comments

3 The Scale Location plot shows that there is not a clear homeskedasticity in the data. Indeed the red line is not flat and at the beginning and at the end its slope changes.

4 This last plot is helpful for checking the existence of outliers by looking at the existence of some points out from the Cook's distance. In our case the plot identifies three outliers (414, 412 and 80).

Application with dummy variable

Instead of using the STR variable, in this new exercise we consider as independent variable a dummy which define the district with a $\texttt{STR} > 20$ to all the others.
To construct this variable we use the following command line:

```
mydata$d_str<-as.numeric(mydata$str>=20)
```

Moreover to make some statistics we separate the DB in large a small in the following way:

```
large <- mydata[mydata$str>20,]
small <- mydata[mydata$str<=20,]
```

Descriptive statistics

We compute some descriptive statistics for the two variable of interest:

```
summary(large$score)
sd(large$score)
summary(small$score)
sd(small$score)
```

		Testscore	
D	N	Mean	SD
0	177	650.00	17.97
1	243	657,19	19.29
All	420	654.16	19.05

Regression

Let's estimate the linear regression

Command lines

```
lm(score ~ d_str)
```

```
Regression                                      Number of obs =     420
                                                F(  1,   418) =   15.05
                                                R-squared     =  0.0348
                                                Root MSE      =   18.74

------------------------------------------------------------------------------
             |               Robust
    testscr  |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cons |     657.185      1.202   546.62   0.000     654.8213    659.5478
         str |      -7.185      1.852    -3.88   0.000     -10.8255   -3.544716
------------------------------------------------------------------------------
```

Difference in mean: $\bar{Y}_{small} - \bar{Y}_{large} = 657.19 - 650.0 = 7.19$

Standard Error (SE): $\sqrt{\frac{S_s^2}{n_s} + \frac{S_l^2}{n_l}} = \sqrt{\frac{19.29^2}{243} + \frac{17.97^2}{177}} = 1.83$

Comments

- The regression coefficient associated to `STR` is negative and significantly different from 0.
- The value of -7.185 indicates the following: compared to classes with a lower student teacher ratio, classes with an higher `STR` present -7.185 points.
- It means that a class with a lower `STR` provide a better result in terms of student score.
- Clearly there will be only two predicted values:
  1. For `STR` $> 20 \Rightarrow 657.185 - 7.185 = 650.00$
  2. For `STR` $\leq 20 \Rightarrow 657.185$

Example

Let us consider a sample of 400 credit card titles, on which, among others, the following variables have been observed [1]:

- `Income`: in thousand of dollars;
- `Cards`: number of credit cards;
- `Student`: (dummy variabl: `No`, `Yes`);
- `Eathnicity`: (categorical variable: `African American`, `Asian`, `Caucasian`);
- `Balance`: total charge on credit cards held, in dollars.

[1] Data source: "An Introduction to Statistical Learning, with applications in R" (Springer, 2013).

Example: continuous explanatory variable

Let us analyse the relation between `Balance` (dependent variable) and `Income` (explanatory variable):

$$\texttt{Balance} = \beta_0 + \beta_1 \ \texttt{Income} + u$$

- β_0 is the total charge on credit card observed when `Income` is equal to zero.
- β_1 is the variation in the average charge when `Income` increases of 1000\$.

Example: explanatory dummy variable

Let's check if to be a student (explanatory variable) gets changing the total charge on credit cards held (dependent variable):

$$\texttt{Balance} = \beta_0 + \beta_1 \texttt{ Student} + u$$

The variable `Student` is transformed in a **variabile dummy** assuming value 0 when **the guy observed** ("`is not a student (No)`") and 1 **in the opposite case** ("`is a student (Yes)`") [2].

- β_0 represents the average charge for individuals that are not students.
- β_1 represents the expected change in average charge for students with respect to no students.

[2] **R** by default it sets the basic mode to the first mode in alphanumeric order

Model Estimation

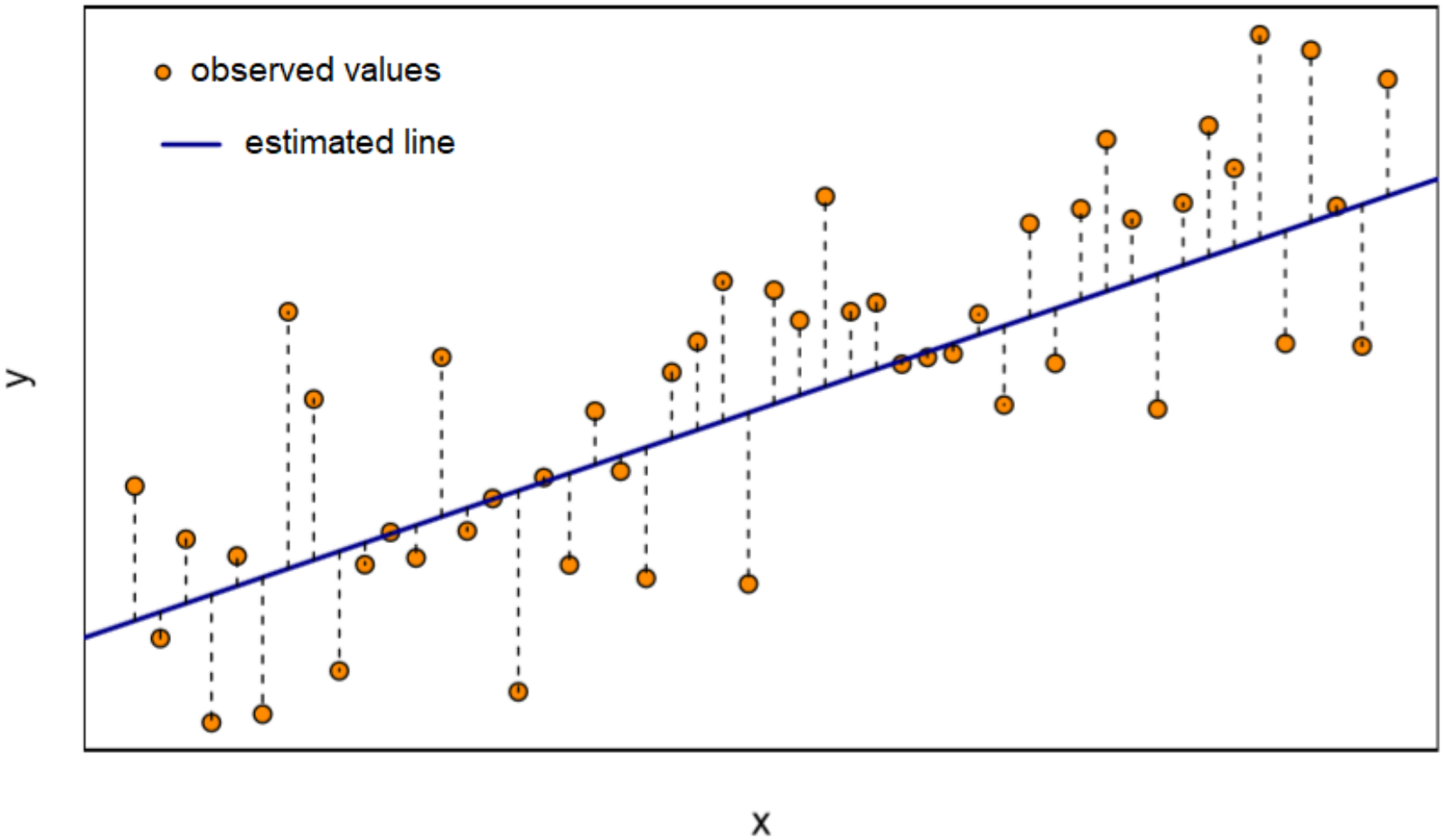
To estimate the model coefficients (**unknown**) β_0 e β_1 we need a sample of n observations both for explanatory and dependent variable.

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n))$$

Our aim is to estimate **the coifficients**, $\hat{\beta}_0$ and $\hat{\beta}_1$, so that:

$$\underbrace{y_i}_{\substack{\text{observed} \\ \text{value}}} \approx \underbrace{\hat{\beta}_0 + \hat{\beta}_0 \; x_i}_{\text{estimated value}} \qquad (i = 1, \ldots, n)$$

in such way we estimate a **line as near as possible** to observed data.


observed values
estimated line
y
x

Least Squares Estimation

Estimation method: **Ordinary Least Squares**, (OLS), or, **least squares**.

- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \; x_i$ be the **estimated value** of the variable Y associated with the i^{th} observation of the variable X.
- Let $\hat{u}_i = y_i - \hat{y}_i = y_i - \left(\hat{\beta}_0 - \hat{\beta}_1 \; x_i\right)$ be the i^{th} **residual**.
- Let $RSS = \hat{u}_1^2 + \hat{u}_2^2 + \cdots + \hat{u}_i^2 + \cdots + \hat{u}_n^2 = \sum_{i=1}^{n} \hat{u}_i^2$ be the **residual sum of squares** (RSS).

The least squares approach aims to determine $\hat{\beta}_0$ and $\hat{\beta}_1$ in order to **minimize** RSS.

Example

Coefficients estimated for the relation between `Balance` and `income` are:

$$\hat{\beta}_0 = 246.51 \qquad \hat{\beta}_1 = 6.05$$

- For `income`$= 0$, the expected value of `Balance` is equal to 246.51\$.
- For each 1000\$ increasing of `income`, `Balance` increases on average of 6.05\$.

In the model with the dummy variable `Student` we get:

$$\hat{\beta}_0 = 480.37 \qquad \hat{\beta}_1 = 396.46$$

It means that:

- For no student `Balance` is on average equal to 480.37\$.
- If the credit card owner is a student `Balance` increases on average by 396.46\$, with respect to no students.

Omitted Variable Bias

Consider the linear regression:

$$Y = \beta_0 + \beta_1 X + u$$

- The error u arises because of factors, or variables, that influence Y but are not included in the regression function. There are always omitted variables
- Sometimes, the omission of those variables can lead to bias in the OLS estimator
- The bias in OLS estimator that occurs as a result of an omitted factor, or variable, is called **omitted variable bias**.

Definition

For omitted variable bias to occur, two conditions must be fulfilled:

- The omitted variable, let's call it Z, is a determinant of the dependent variable Y ($\beta_2 \neq 0$)
- X is correlated with the omitted variable. (i.e. $\sigma_{xz} \neq 0$)

Example

Let's go back to the class size example:

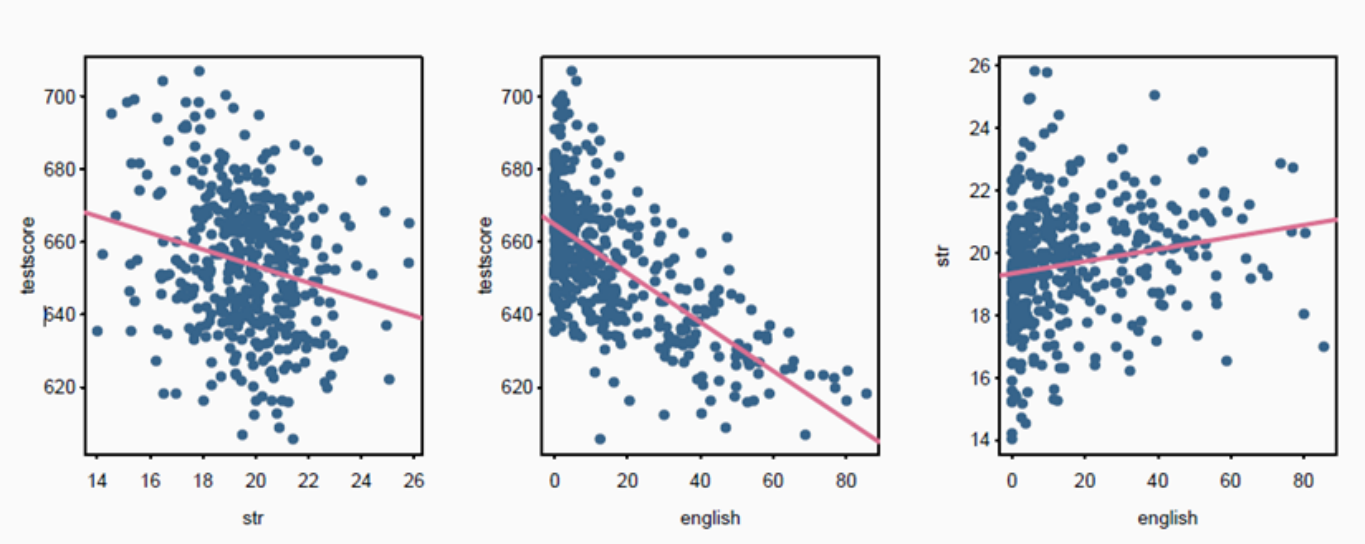
$$testscore = \beta_0 + \beta_1 STR + u$$

- Consider a new variable: `English language ability` (Z) measuring whether the student has English as a second language which plausibly affects standardized test scores: Z is a determinant of Y.
- Immigrant communities tend to be less affluent and thus have smaller school budgets and higher STR: Z is correlated with X.

  Thus β_1 is bias, i.e. $\hat{\beta}_1 \xrightarrow{p} \beta_1 + bias$

Example

The California School Dataset has data on the fraction of English learning in a district, named 'English'



- Districts with fewer English Learners have higher test scores
- Districts with lower percent EL (PctEL) have smaller classes

Focus on the theoretical problem

If this new variable is significant it means that the correct model is given by two X's variables

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

And we should estimate

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{u}$$

We know that $\mathbb{E}\hat{\beta}_1 = \beta_1$ and $\mathbb{E}\hat{\beta}_2 = \beta_2$ i.e. the regression coefficients are unbiased estimators of the population parameters.
Now suppose that we estimate,

$$Y = \hat{\beta}_0^* + \hat{\beta}_1^* X_1 + \hat{\epsilon}$$

And mistakenly we omitted X_2

Questions

1. How does $\hat{\beta}_1$ (the regression estimate from the correctly specified model) compare to $\hat{\beta}_1^*$ (the regression estimate from the miss-specified model)?
2. What is $\mathbb{E}\hat{\beta}_1^*$?
3. Is it a biased or unbiased estimator of $\hat{\beta}_1$? ù
4. If it is biased, how is it biased?

Solution

1. Formula for bivariate regression coefficient:

$$\beta_1 = \frac{\hat{Cov}(X_1, Y)}{\hat{Var}(X_1)}$$

2. Substituting in Y its value from the correctly specified model:

$$\begin{aligned}\beta_1 &= \frac{\widehat{Cov}(X_1, \hat{\beta}_0 + \widehat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{u})}{\hat{Var}(X_1)} \\ &= \frac{\widehat{Cov}(X_1, \hat{\beta}_0) + \hat{\beta}_1 \widehat{Cov}(X_1, X_1) + \hat{\beta}_2 \widehat{Cov}(X_1, X_2) + \widehat{Cov}(X_1, \hat{u})}{\widehat{Var}(X_1)}\end{aligned}$$

3. Recalling that each variable is uncorrelated with the constant and the residuals:

$$\beta_1 = \frac{0 + \hat{\beta}_1 \widehat{Cov}(X_1, X_1) + \hat{\beta}_2 \widehat{Cov}(X_1, X_2) + 0}{\widehat{Var}(X_1)}$$

Results

If X_2 has mistakenly been omitted from the model, then, taking expectation we get:

$$E(\hat{\beta}_1^*) = \hat{\beta}_1 + \hat{\beta}_2 \frac{\sigma_{12}}{\sigma_1^2}$$

It means that:

- $\hat{\beta}_1^*$ is a biased estimator of β_1
- This bias will not disappear as the sample size gets larger, so the omission of a variable from a model also leads to inconsistent estimator
- To let$\hat{\beta}_1^*$ be not biased two condition should be respected:
  1. $\beta_2 = 0$ Of course if $\beta_2 = 0$, this means that the model is not miss-specified (X_2 does not belong to the model because it has no effect on Y)
  2. $\sigma_{12} = 0$. That is, if the 2 $X's$ are uncorrelated, then omitting one does not result in biased estimates of the effect of the other

Example Results

Considering the class size example. It is very likely that:

- $\hat{\beta}_2 < 0$ it is reasonable to assume that districts with more English learners have lower $testscore$ (the sample analysis also seems to suggest so)
- $\sigma_{xz} > 0$ the covariance between `STR` and `English` is probably positive (the sample analysis also seems to suggest so)

Thus, the bias is probably negative in which case we say that β_1 is downward biased, that is, it is smaller than the true β_1.

To solve this problem we move to **multiple linear regression**

Multiple Linear Regression

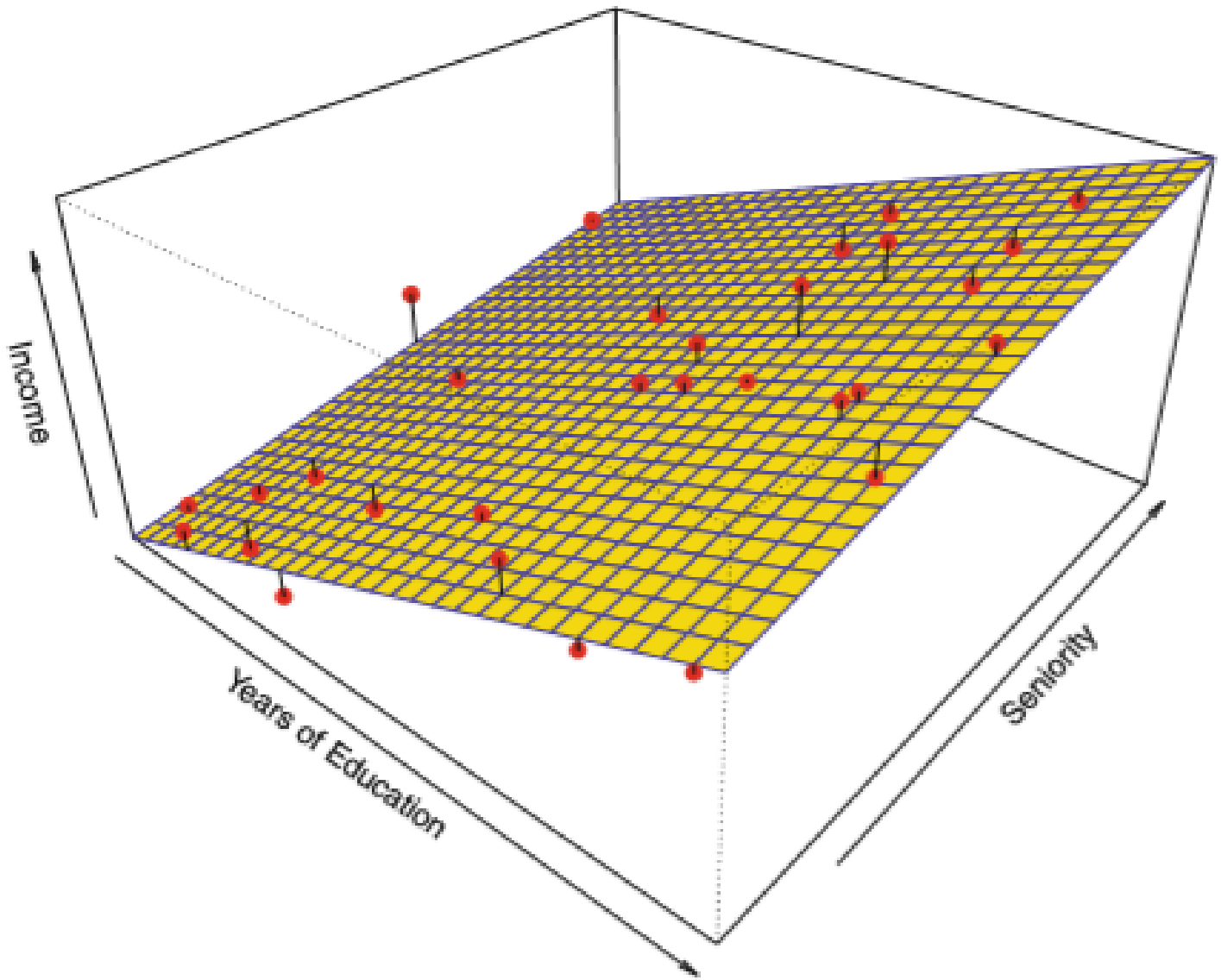
Multiple Linear Regression Model allows to study the relationship between a set of p explanatory variables, $X_1, X_2, \ldots, X_j, \ldots, X_p$ (quantitative or qualitative) and a dependent variable, taking care to the effect of **each explanatory variable**:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_j X_j + \cdots + +\beta_p X_p + u$$

All the **assumptions** considered for the **simple** linear regression model are easily **generalized** for **multiple linear regression**.
In this case, the **deterministic component** of the model:

$$\mathbb{E}\left(Y | X_1 = x_1, \ldots, X_p = x_p\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

is called **regression hyperplane**.


Income
Years of Education
Seniority

Coefficients Interpretation

For the multiple linear regress regression model:

- β_0 represents the expected value of Y when **all the explanatory variables** are equal to 0;
- β_j represents the expected increasing (decreasing) that occurs to Y, when the variable X_j **increases (decreases) of one unit**, while all the others p_1 explanatory variables **do not change** (**marginal effect of X_j on Y**).

In other words, β_j represents the effect of X_j on Y, **given all the other conditions**.

Coefficients Interpretation *in formula*

Given the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_j X_j + \cdots + \beta_p X_p + u$$

Let us suppose that X_1 increases by ΔX_1, holding $X_2, \ldots, X_p$ constant. The deterministic component of the model becomes:

$$\mathbb{E}\left(Y | X_1 = x_1 + \Delta x_1, \ldots, X_p = x_p\right) = \beta_0 + \beta_1 (x_1 + \Delta x) + \cdots + \beta_p x_p$$

Estimated Effect

$$\begin{aligned} &\mathbb{E}(Y | X_1 = x_1 + \Delta x_1, \ldots, X_p) \\ &\qquad - \mathbb{E}\left(Y | X_1 = x_1, \ldots, X_p\right) = \beta_1 \mathbf{\Delta x_1} \end{aligned}$$

Perfect multicollinearity (the 4^{th} assumption)

In multiple linear regression we have a 4^{th} assumption to take care: **perfect multicollinearity**

- Perfect multicollinearity is when one of the regressors is an exact linear function of the other regressors.
- When there is multicollinearity we cannot identify the marginal effect of each dependent viariable.

Example

Let us consider the following regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

Suppose that $X_1 = \phi X_2$ (the elements of the second X_1 are proportional to those of the X_2).
So that we have:

$$Y = \beta_0 + \beta_1 \phi X_2 + \beta_2 X_2 + u = \beta_0 + (\beta_1 \phi + \beta_2) X_2 + u$$

In this case we cannot separate the effect of X_1 by that of X_2. The two parameters are not β_1 and β_2 cannot be distinguished separately while we can capture their linear combination $(\beta_1 \phi + \beta_2)$.

Comments

- Perfect multicollinearity usually reflects a mistake in the definitions of the regressors, or an oddity in the data
- If you have perfect multicollinearity, your statistical software will let you know either by crashing or giving an error message or by "dropping" one of the variables arbitrarily
- The solution to perfect multicollinearity is modifying the list of regressors so that you no longer have perfect multicollinearity.

Imperfect multicollinearity

Imperfect multicollinearity occurs when two or more regressors are very highly correlated.

Why the term multicollinearity? If two regressors are very highly correlated, then their scatterplot will pretty much look like a straight line they are "co-linear" but unless the correlation is exactly ± 1, that collinearity is imperfect.

One of the **consequences** of imperfect multicollinearity is that the standard errors of the coefficients tend to be large. In that case, the test of the hypothesis that the coefficient is equal to zero may lead to a failure to reject a false null hypothesis of no effect of the explanatory.

Continuous explanatory variables

Let us consider as **dependent variable** `Balance`, while, as explanatory variable `Income` and `Cards`:

$$\texttt{Balance} = \beta_0 + \beta_1 \ \texttt{Income} + \beta_2 \ \texttt{Cards} + u$$

- β_0 is the average value of `Balance` for an individual who has no `income` and no `credit cards`.
- β_1 is the expected change in `Balance` if the income increases by 1000\$ **given the** number of credit cards.
- β_2 is the expected change in `Balance` for a student with respect to a no student, **given** the same amount of `income`.

Continuous and dummy explanatory variables

Let us consider as explanatory variables `Income` and `Student`

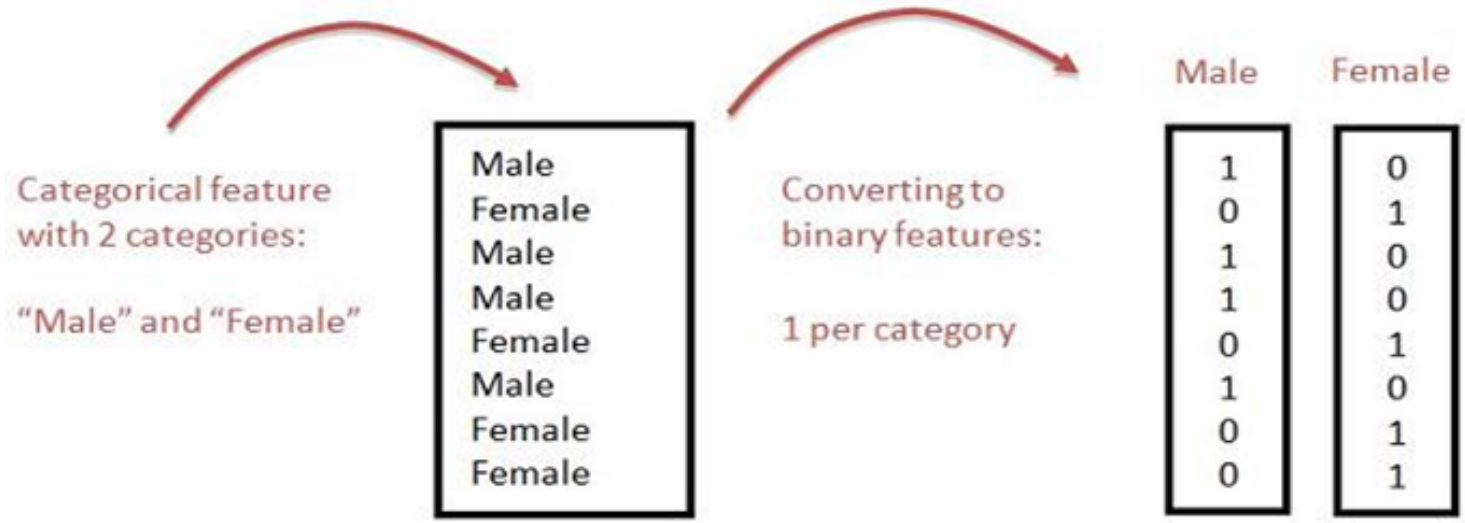
$$\texttt{Balance} = \beta_0 + \beta_1 \ \texttt{Income} + \beta_2 \ \texttt{Student} + u$$

- β_0 is the average value of `balance` for a no student without income.
- β_1 is the expected change on `balance` if `income` increases by 1000\$ regardless of whether the credit card holder is a student or not.
- β_2 is the average change on `balance` for a student, given the same amount of `income`.

Categorical independent variable

Suppose you have a categorical variable, i.e. multiple categories and every observation falls in one and only one category (like region of residence: Sicily, Lazio, Tuscany,...)

To deal with them you have to transform the categorical variable with m attribute into a set of m binary variables, which are mutually exclusive and exhaustive. The simplest example: dicothomic variable...



Dummy Variable Trap

If you include all these m dummy variables and a constant, you will have perfect multicollinearity this is sometimes called the dummy variable trap.

Solutions to the dummy variable trap:

1. Omit one of the groups (e.g. Lazio)
2. Omit the intercept