San José State University

Math 261A: Regression Theory & Methods

Indicator Variables

Dr. Guangliang Chen

This lecture is based on the following part of the textbook:

- Sections 8.1 – 8.2

Outline of the presentation:

- A single categorical variable with exactly two levels
- A single categorical variable with more than two levels
- Models with two or more indicator variables
- From quantitative to qualitative

Introduction

So far we have only used quantitative variables in our regression analysis (i.e., variables with numerical values), such as height, weight, temperature, distance, and pressure.

Sometimes it is necessary to use qualitative/categorical variables as predictors, such as sex, employment status, educational level, etc.

These are variables whose values are nonnumerical, and thus you cannot add them or multiply them by a scalar.

We need a way to convert categorical variables to quantitative ones in order to carry out the regression analysis.

A single categorical variable with exactly 2 levels

Suppose that a mechanical engineer wishes to relate the **effective life of a cutting tool** (y) used on a lathe to the **lathe speed** in revolutions per minute (x_1) and the **type of cutting tool** used.

The second regressor variable, "tool type", is qualitative and has two levels (e.g., tool types A and B). We can use an **indicator variable** x_2 that takes on the values 0 and 1 to identify the classes of the regressor variable "tool type" as follows:

$$x_2 = \begin{cases} 0, & \text{if the observation is of tool type A} \leftarrow \text{baseline} \\ 1, & \text{if the observation is of tool type B} \end{cases}$$

With the quantitative variable x_2 we can formulate a linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

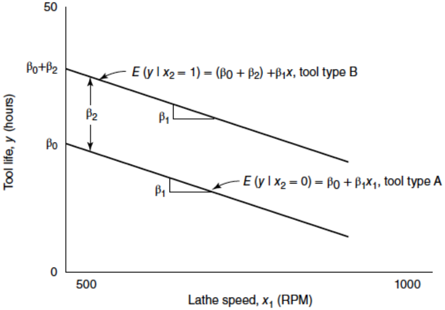
This model can be rewritten as follows

$$y = \begin{cases} \beta_0 + \beta_1 x_1 + \epsilon, & x_2 = 0 \, (\text{type A}) \\ (\beta_0 + \beta_2) + \beta_1 x_1 + \epsilon, & x_2 = 1 \, (\text{type B}) \end{cases}$$

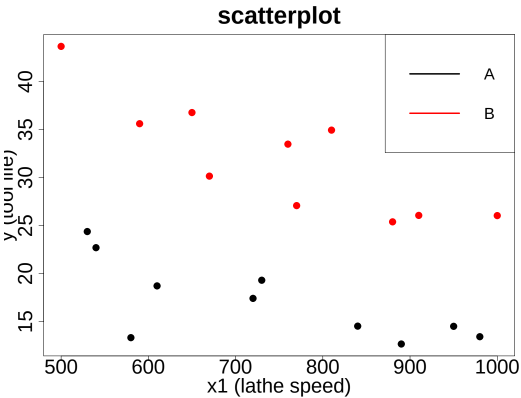
Interpretation

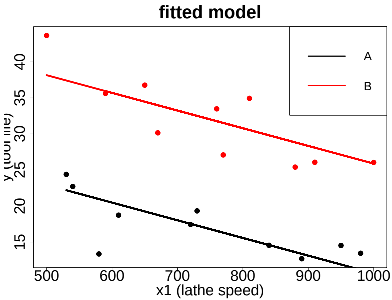
- Mathematically, **there is a separate response function for each tool type**, but they only differ in the intercept.

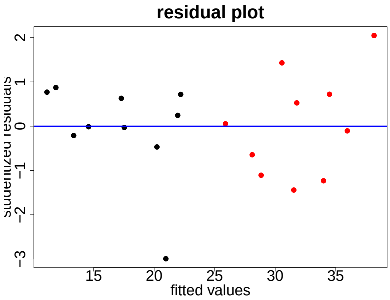
- Geometrically, the equation (with the indicator variable) defines **two parallel regression lines** (with a common slope but different intercepts), one for each tool type.



R demonstration (Tool Life Data)




fitted model
A
B
y (tool life)
40
35
30
25
20
15
500
600
700
800
900
1000
x1 (lathe speed)


residual plot
studentized residuals
2
1
0
−1
−2
−3
15
20
25
30
35
fitted values

Indicator Variables

```
> mymodel <- lm(y~x1+ToolType, data=mydata)
> summary(mymodel)

Call:
lm(formula = y ~ x1 + ToolType, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-7.6255 -1.6308  0.0612  2.2218  5.5044

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 35.208726   3.738882   9.417 3.71e-08 ***
x1          -0.024557   0.004865  -5.048 9.92e-05 ***
ToolTypeB   15.235474   1.501220  10.149 1.25e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.352 on 17 degrees of freedom
Multiple R-squared:  0.8787,    Adjusted R-squared:  0.8645
F-statistic:  61.6 on 2 and 17 DF,  p-value: 1.627e-08
```

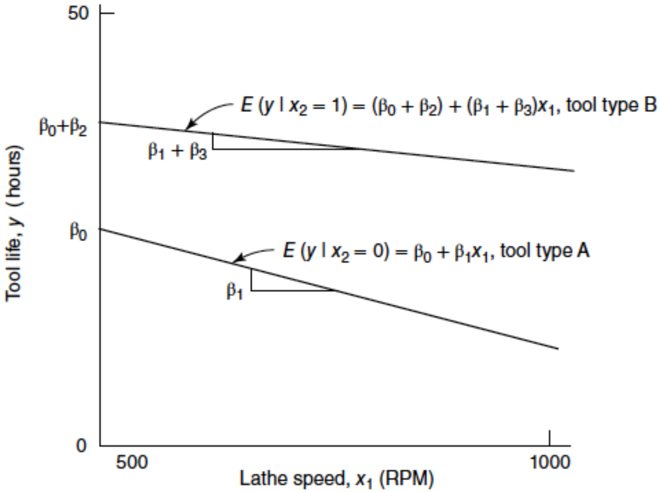
The categorical predictor is significant! What does it mean?

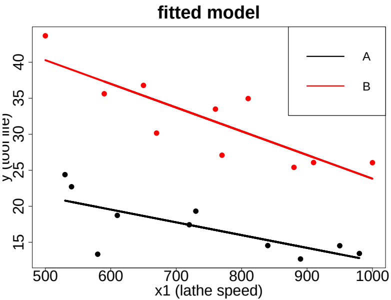
To fit a model consisting of two regression lines that differ in both intercept and slope, we can add an interaction term:

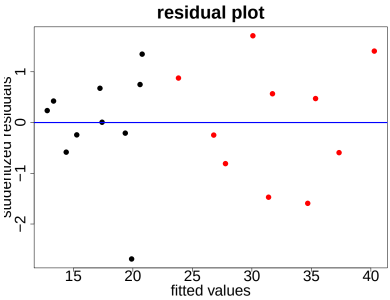
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$$

This new model is equivalent to

$$y = \begin{cases} \beta_0 + \beta_1 x_1 + \epsilon, & x_2 = 0 \,(\text{type A}) \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1 + \epsilon, & x_2 = 1 \,(\text{type B}) \end{cases}$$


50
$\beta_0+\beta_2$
$E(y \mid x_2 = 1) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x_1$, tool type B
$\beta_1 + \beta_3$
β_0
$E(y \mid x_2 = 0) = \beta_0 + \beta_1 x_1$, tool type A
β_1
Tool life, y (hours)
0
500
1000
Lathe speed, x_1 (RPM)


fitted model
A
B
y (tool life)
x1 (lathe speed)
500
600
700
800
900
1000
15
20
25
30
35
40


residual plot
studentized residuals
fitted values
15
20
25
30
35
40
−2
−1
0
1

Indicator Variables

```
> mymodel <- lm(y~x1*ToolType, data=mydata)
> summary(mymodel)

Call:
lm(formula = y ~ x1 * ToolType, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.5534 -1.7088  0.3283  2.0913  4.8652 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  30.176013   4.724895   6.387 9.01e-06 ***
x1           -0.017729   0.006262  -2.831  0.01204 *  
ToolTypeB    26.569340   7.115681   3.734  0.00181 ** 
x1:ToolTypeB -0.015186   0.009338  -1.626  0.12345    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.201 on 16 degrees of freedom
Multiple R-squared:  0.8959,	Adjusted R-squared:  0.8764 
F-statistic: 45.92 on 3 and 16 DF,  p-value: 4.37e-08
```

The interaction term is not significant. What does it imply?

Significance tests

We have seen that fitting the model (with a quantitative predictor x_1 and a categorical predictor x_2)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$$

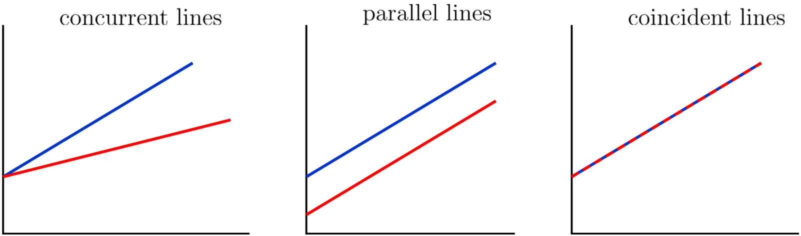
is equivalent to fitting two separate regression lines (with different intercepts and/or slopes).

$$y = \begin{cases} \beta_0 + \beta_1 x_1 + \epsilon, & x_2 = 0 \,(\text{type A}) \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1 + \epsilon, & x_2 = 1 \,(\text{type B}) \end{cases}$$

We can test any of the following

- **Concurrent lines**: the intercepts are identical

$$H_0 : \beta_2 = 0 \text{ (same intercept but slope could be different)}$$
$$H_1 : \beta_2 \neq 0 \text{ (separate intercepts needed)}$$



- **Parallel lines**: the slopes are identical

$$H_0 : \beta_3 = 0 \text{ (same slope but intercept could be different)}$$
$$H_1 : \beta_3 \neq 0 \text{ (separate slopes)}$$

- **Coincident lines**: the two regression models are identical

$$H_0 : \beta_2 = \beta_3 = 0 \text{ (same intercept and slope)}$$
$$H_1 : \beta_2 \neq 0 \text{ (separate intercepts needed),}$$
$$\text{or } \beta_3 \neq 0 \text{ (separate slopes needed)}$$

by fitting the two models in each case

- a reduced model (H_0 true): $SS_{Res}(RM), df_{RM}$
- the full model: $SS_{Res}(FM), df_{FM} = n - 4$

and comparing them using an extra-sum-of-squares F-test:

$$F_0 = \frac{\frac{SS_{Res}(RM) - SS_{Res}(FM)}{df_{RM} - df_{FM}}}{\frac{SS_{Res}(FM)}{df_{FM}}} \sim F_{df_{RM} - df_{FM}, df_{FM}}$$

We will reject H_0 at level α if

$$F_0 > F_{\alpha, df_{RM} - df_{FM}, df_{FM}}$$

or equivalently,

$$p\text{-value} < \alpha$$

For example, in the coincident lines test, $df_{RM} = n - 2$. It is conducted in R as follows:

```
> myReducedModel <- lm(y~x1, data=mydata)
> myFullModel <- lm(y~x1*ToolType, data=mydata)
> anova(myReducedModel, myFullModel)
Analysis of Variance Table

Model 1: y ~ x1
Model 2: y ~ x1 * ToolType
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     18 1348.06
2     16  163.89  2    1184.2 57.802 4.773e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Conclusion: Since the p-value is significant, we reject $H_0 : \beta_2 = \beta_3 = 0$ and correspondingly conclude that two separate regression models are needed (that are different in slope or intercept or both).

A single categorical variable with > 2 levels

An electric utility is investigating the effect of the **size of a single-family house** x_1 (square feet of floor space) and the **type of air conditioning** used in the house on the **total electricity consumption** y (in kilowatt-hours) during the period of June through September.

There are four types of air conditioning systems: (1) no air conditioning, (2) window units, (3) heat pump, and (4) central air conditioning.

Type of air conditioning is a categorical variable with 4 levels. We need to convert it to numerical in order to carry out a regression analysis.

Option 1: Regression with **allocated codes**

Type of Air Conditioning System	x_2
No air conditioning	0
Window units	1
Heat pumps	2
Central air conditioning	3

We then fit a model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

Option 2: The 4 levels of this factor can be modeled by 3 **indicator/ dummy** variables, x_2, x_3, and x_4, defined as follows:

Type of Air Conditioning	x_2	x_3	x_4
No air conditioning (baseline)	0	0	0
Window units	1	0	0
Heat pump	0	1	0
Central air conditioning	0	0	1

The corresponding regression model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$$

Option 1 implies that

$$y = \begin{cases} \beta_0 + \beta_1 x_1 + \epsilon & \text{(no air conditioning)} \\ \beta_0 + \beta_1 x_1 + \beta_2 + \epsilon & \text{(window units)} \\ \beta_0 + \beta_1 x_1 + 2\beta_2 + \epsilon & \text{(heat pump)} \\ \beta_0 + \beta_1 x_1 + 3\beta_2 + \epsilon & \text{(central air conditioning)} \end{cases}$$

This may be unrealistic or even very wrong.

Option 2 implies that

$$y = \begin{cases} \beta_0 + \beta_1 x_1 + \epsilon & \text{(no air conditioning)} \\ (\beta_0 + \beta_2) + \beta_1 x_1 + \epsilon & \text{(window units)} \\ (\beta_0 + \beta_3) + \beta_1 x_1 + \epsilon & \text{(heat pump)} \\ (\beta_0 + \beta_4) + \beta_1 x_1 + \epsilon & \text{(central air conditioning)} \end{cases}$$

where $\beta_2, \beta_3, \beta_4$ represent the separate effects of the three air conditioning systems, all relative to the baseline (no air conditioning).

This is the correct way to handle a categorical predictor with >2 levels.

Geometrically, this defines four parallel lines, one for each level of the categorical variable.

It is also possible to use different slopes by adding interaction terms between the quantitative variable x_1 and each of the three indicator variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_1 x_2 + \beta_6 x_1 x_3 + \beta_7 x_1 x_4 + \epsilon$$

This is equivalent to

$$y = \begin{cases} \beta_0 + \beta_1 x_1 + \epsilon & \text{(no air conditioning)} \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_5) x_1 + \epsilon & \text{(window units)} \\ (\beta_0 + \beta_3) + (\beta_1 + \beta_6) x_1 + \epsilon & \text{(heat pump)} \\ (\beta_0 + \beta_4) + (\beta_1 + \beta_7) x_1 + \epsilon & \text{(central air conditioning)} \end{cases}$$

Models with two or more indicator variables

Consider again the setting where a mechanical engineer wishes to relate the **effective life of a cutting tool** (y) used on a lathe to the **lathe speed** in revolutions per minute (x_1) and the **type of cutting tool** used.

The regressor variable, tool type, is qualitative and has two levels (e.g., tool types A and B). We used an **indicator variable** x_2 that takes on the values 0 and 1 to identify the classes of the regressor variable "tool type" as follows:

$$x_2 = \begin{cases} 0, & \text{if the observation is of tool type A} \leftarrow \text{baseline} \\ 1, & \text{if the observation is of tool type B} \end{cases}$$

Suppose that a second qualitative factor, **the type of cutting oil used**, must be considered. Assuming that this factor has two levels, "low-viscosity oil" and "medium-viscosity oil", we may define a second indicator as follows;

$$x_3 = \begin{cases} 0, & \text{if low-viscosity oil is used} \leftarrow \text{baseline} \\ 1, & \text{if medium-viscosity oil is used} \end{cases}$$

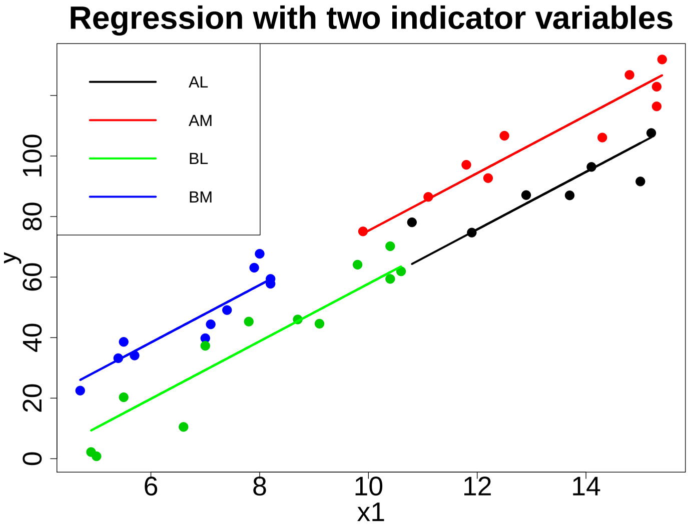
A regression model relating tool life (y) to cutting speed (x_1), tool type (x_2), and oil type (x_3) is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

This model can be rewritten as follows:

$$y = \begin{cases} \beta_0 + \beta_1 x_1 + \epsilon & \text{(type A tool, low viscosity oil)} \\ (\beta_0 + \beta_2) + \beta_1 x_1 + \epsilon & \text{(type B tool, low viscosity oil)} \\ (\beta_0 + \beta_3) + \beta_1 x_1 + \epsilon & \text{(type A tool, medium viscosity oil)} \\ (\beta_0 + \beta_2 + \beta_3) + \beta_1 x_1 + \epsilon & \text{(type B tool, medium viscosity oil)} \end{cases}$$

This defines **four parallel regression lines** corresponding to the four pairs of levels of the two categorical variables.

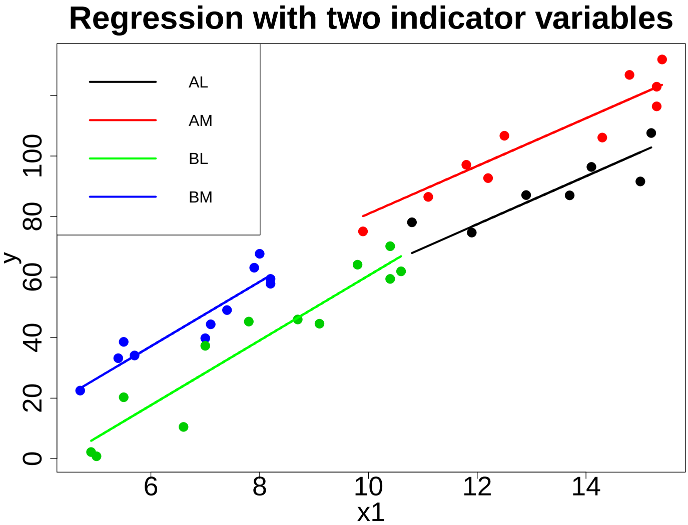

Regression with two indicator variables
AL
AM
BL
BM
y
0
20
40
60
80
100
6
8
10
12
14
x1

To allow the regression lines to have different slopes, we can add all the **interaction effects** (between the quantitative variable and the two categorical variables):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \epsilon$$

which can be rewritten as

$$y = \begin{cases} \beta_0 + \beta_1 x_1 + \epsilon & \text{(type A tool, low viscosity oil)} \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_4) x_1 + \epsilon & \text{(type B tool, low viscosity oil)} \\ (\beta_0 + \beta_3) + (\beta_1 + \beta_5) x_1 + \epsilon & \text{(type A tool, medium viscosity oi} \\ (\beta_0 + \beta_2 + \beta_3) + (\beta_1 + \beta_4 + \beta_5) x_1 + \epsilon & \text{(type B tool, medium viscosity oi} \end{cases}$$


Regression with two indicator variables
AL
AM
BL
BM
y
0
20
40
60
80
100
x1
6
8
10
12
14

The model on the preceding slide is still additive (despite the interaction terms).

If we further add the interaction between the two indicator variables

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3 + \epsilon$$

$$= \begin{cases} \beta_0 + \beta_1 x_1 + \epsilon & \text{(type A tool, low viscosity oi} \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_4) x_1 + \epsilon & \text{(type B tool, low viscosity oi} \\ (\beta_0 + \beta_3) + (\beta_1 + \beta_5) x_1 + \epsilon & \text{(type A tool, medium viscos} \\ (\beta_0 + \beta_2 + \beta_3 + \beta_6) + (\beta_1 + \beta_4 + \beta_5) x_1 + \epsilon & \text{(type B tool, medium viscosi} \end{cases}$$

then it indicates that

the effect of one indicator variable on the intercept depends on the level of the other indicator variable.

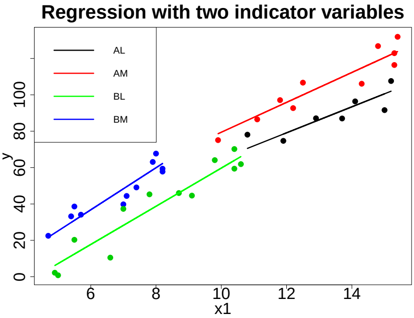
Similarly, adding the interaction term among all three variables (1 quantitative and 2 qualitative) to the model

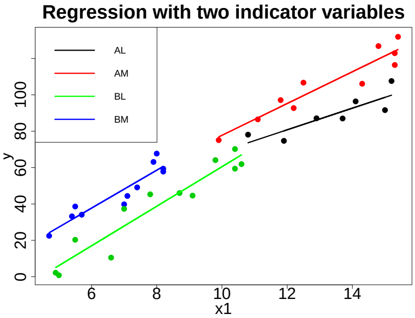
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3 + \beta_7 x_1 x_2 x_3 + \epsilon$$

$$= \begin{cases} \beta_0 + \beta_1 x_1 + \epsilon & \text{(type A tool, low viscos} \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_4) x_1 + \epsilon & \text{(type B tool, low viscos} \\ (\beta_0 + \beta_3) + (\beta_1 + \beta_5) x_1 + \epsilon & \text{(type A tool, medium v} \\ (\beta_0 + \beta_2 + \beta_3 + \beta_6) + (\beta_1 + \beta_4 + \beta_5 + \beta_7) x_1 + \epsilon & \text{(type B tool, medium v} \end{cases}$$

results in that

the effect of one indicator variable on the slope depends on the level of the other indicator variable.


Regression with two indicator variables
AL
AM
BL
BM
y
0
20
40
60
80
100
6
8
10
12
14
x1


Regression with two indicator variables
AL
AM
BL
BM
y
0
20
40
60
80
100
6
8
10
12
14
x1

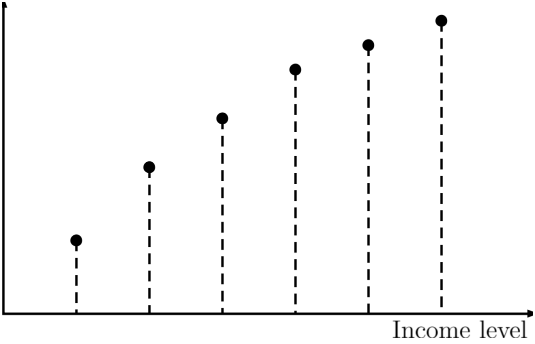
From quantitative to qualitative

Values of a quantitative variable may be grouped into classes to produce a qualitative variable, e.g., income $\rightarrow$ income level.

Household Income Range	Millions of Households	% of Total	Income class
Less than $20,000	19.7	15%	Below or near poverty level
$20,000 - $44,999	28.7	23%	Low income
$45,000 - $139,999	57.7	45%	Middle class
$140,000 - $149,999	2.6	2%	Upper middle class
$150,000 - $199,999	9.0	7%	High income
$200,000 and over	9.9	8%	Highest tax brackets

Source: "Table HINC-01, 2018 Household Income Survey", U.S. Census Bureau

Advantage: It does not require any prior assumption about the functional form of the relationship between the response and the regressor variable.



Disadvantages:

- It may lead to degraded accuracy
- It requires more regression parameters (as $\ell - 1$ indicator variables need to be introduced for ℓ levels)
- It reduces the degree of freedom for the error sum of squares (SS_{Res}) $\longleftarrow$ more data may be needed