Категориальные переменные

Категориальные переменные

- Принимают конечное, но большее двух множество значений
- **Например, переменная education – принимает значение:**
- 1 - для индивидов с незаконченным средним образованием;
- 2 – для индивидов с законченным средним образованием;
- 3 – для индивидов с незаконченным высшим образованием;
- 4 – для индивидов с законченным высшим образованием;
- 5 – для закончивших аспирантуру.

- **Например, переменная trustgovernment – принимает значение:**
- 1 – если индивид полностью доверяет правительству;
- 2 – если скорее доверяет;
- 3 – если относится нейтрально;
- 4 – если скорее не доверяет;
- 5 – если совсем не доверяет.

- **Например, переменная fedokrug– федеральный округ, в котором проживает индивид, принимает значение:**
- 1 – для Северо-Западного ФО ;
- 2 – для Центрального ФО;
- 3 – для Южного ФО;
- 4 – для Сибирского ФО;
- 5 – для Уральского ФО
- 6 – для Приволжского ФО
- 7 – для Дальневлсточного ФО
- 8 – для Северо-Кавказского ФО
- 9 – для Крымского ФО.

- Категориальные переменные не рекомендуется включать в уравнение регрессии в первоначальном виде.
- Вместо одной категориальной в уравнение регрессии включается набор фиктивных переменных
- При этом (важно!!!) **фиктивных переменных** в уравнение регрессии следует включать **на одну меньше, чем выделено категорий.**
- Невключенная категория называется базовой и все остальные категории сравниваются с ней.

- Например, при моделировании зависимости спроса на некоторый товар **Y** от его цены **P** и среднего дохода покупателей **I** нередко возникает необходимость учитывать сезонность. Пусть данные являются квартальными, тогда можно создать 4 дополнительные дамми-переменные:
- D1, которая =1 если период наблюдения первый квартал, и =0, если период наблюдения 2, 3 или 4 кварталы;
- D2, которая =1 если период наблюдения второй квартал, и =0, если период наблюдения 1, 3 или 4 кварталы;
- D3, которая =1 если период наблюдения третий квартал, и =0, если период наблюдения 1, 2 или 4 кварталы;
- D4, которая =1 если период наблюдения четвертый квартал, и =0, если период наблюдения 1, 2 или 3 кварталы;

- Но в уравнение регрессии следует включать **не все 4, а только 3** квартальные дамми-переменные.
- Это объясняется тем, что дамми-переменные D1, D2,D3 и D4 в сумме дают единичный столбец, и тогда условие теоремы Гаусса-Маркова о независимости столбцов матрицы X будет нарушено (возникнет мультиколлинеарность).

- Если в примере с сезонностью в качестве базового выбран первый квартал, то уравнение регрессии имеет вид

$$Y = \beta_0 + \beta_p P + \beta_I I + \beta_2 D_2 + \beta_3 D_3 + \beta_4 D_4 + \varepsilon$$

- Оцененное уравнение регрессии

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_p P + \hat{\beta}_I I,$$

- Для 1-го квартала

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_2 + \hat{\beta}_p P + \hat{\beta}_I I$$

- Для 2-го квартала

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_3 + \hat{\beta}_p P + \hat{\beta}_I I$$

- Для 3-го квартала

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_4 + \hat{\beta}_p P + \hat{\beta}_I I$$

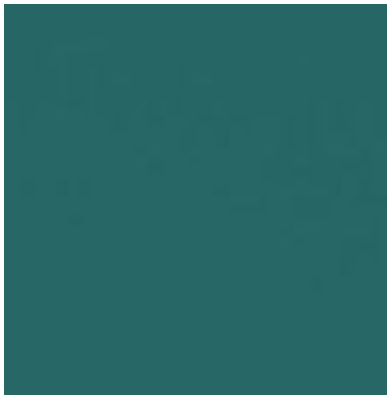
- Для 4-го квартала

Интерпретация коэффициентов:

Если коэффициент β_2 значим, то разница в спросе в первом и втором кварталах составляет β_2. Аналогично значимость $\beta_3 (\beta_4)$ отражает разницу в спросе в первом и третьем (четвертом) квартале

Пример

Имеются данные о цвете (Color), длине (Length), ширине (Width) лепестков и показателе роста цветков (Rate).



```
Color,Length,Width,Rate
blue, 5.4, 1.8, 0.9
blue, 4.8, 1.5, 0.7
blue, 4.9, 1.6, 0.8
pink, 5.0, 1.9, 0.4
pink, 5.2, 1.5, 0.3
pink, 4.7, 1.9, 0.4
teal, 3.7, 2.2, 1.4
teal, 4.2, 1.9, 1.2
```

```
> model <- lm(data$Rate ~ (data$Color + data$Length + data$Width))
> summary(model)
```

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -0.14758    0.48286  -0.306  0.77986
data$Colorpink -0.49083    0.04507 -10.891  0.00166 **
data$Colorteal  0.35672    0.09990   3.571  0.03754 *
data$Length     0.04159    0.07876   0.528  0.63406
data$Width      0.45200    0.11973   3.775  0.03255 *
```