Fachhochschule Südwestfalen
University of Applied Sciences

Statistical Modeling

CH.4 - Qualitative Variables

SS 2021 || Prof. Dr. Buchwitz

Wir geben Impulse

Outline

Course Contents

Session	Topic
1	Simple Linear Regression
2	Multiple Linear Regression
3	Regression Diagnostics
4	Qualitative Variables as Predictors
5	Transformation of Variables
6	Weighted Least Squares
7	Correlated Errors
8	Analysis of Collinear Data
9	Working with Collinear Data
10	Variable Selection Procedures
11	Logistic Regression
12	Further Topics

Outline

Introduction

- Qualitative or categorical variables (such as gender, marital status, etc) are useful predictors and are usually called **indicator** or **dummy variables**.
- Those variables usually only take two values, 0 and 1, which signify that the observation belongs to one of two possible categories.
- The numerical values of indicator variables**do not reflect quantitative ordering**.
- **Example Variable:** Gender, coded as 1 for *female* and 0 for *male*.
- Indicator variables can also be used in a regression equation to distinguish between three or more groups.
- The response variable is stil a quantiative continuous in all discussed cases.

Example: Salary Survey Data

```
P130
```

```
##          S  X E M
## 1  13876  1 1 1
## 2  11608  1 3 0
## 3  18701  1 3 1
## 4  11283  1 2 0
## 5  11767  1 3 0
## 6  20872  2 2 1
## 7  11772  2 2 0
## 8  10535  2 1 0
## 9  12195  2 3 0
## 10 12313  3 2 0
## 11 14975  3 1 1
## 12 21371  3 2 1
## 13 19800  3 3 1
## 14 11417  4 1 0
## 15 20263  4 3 1
## 16 13231  4 3 0
## 17 12884  4 2 0
## 18 13245  5 2 0
## 19 13677  5 3 0
## 20 15965  5 1 1
## 21 12336  6 1 0
## 22 21352  6 3 1
## 23 13839  6 2 0
## 24 22884  6 2 1
## 25 16978  7 1 1
## 26 14803  8 2 0
```

Your turn

Salary survey of computer professionals with objective to identify and quantify variables that determine salary differentials.

S Salary (Response)

X Experience, measured in years

E Education, 1 (High School/HS), 2 (Bachelor/BS), 3 (Advanced Degree/AD)

M Management 1 (is Manager), 0 (no Management Responsibility)

Example: Salary Survey Data

- **Experience:** We assume linearity, which means that each additional year is worth a fixed salary increment.
- **Education:** Can be used in a linear or categorial form.
  - Using the the variable in its raw form would assume that each step up in education is worth a fixed increment in salary. This may be too restrictive.
  - Using education as categorical variable can be done by defining **two indicator variables**. This allows to pick up the effect of education wether it is linear or not.
- **Management:** Is also an indicator variable, that allows to distinguish between management (1) an regular staff positions (0).

Indicator Variables

When using indicator variables to represent a set of categroies, the number of these variables required is **one less than the number of categories**. For *education* we can create two indicators variables:

$$E_{i1} = \begin{cases} 1, \text{if the i-th person is in the HS category} \\ 0, \text{otherwise.} \end{cases}$$

$$E_{i2} = \begin{cases} 1, & \text{if the i-th person is in the BS category} \\ 0, & \text{otherwise.} \end{cases}$$

These two variables allow representing the three groups (HS, BS, AD).

$$\text{HS: } E_1 = 1, E_2 = 0, \text{ BS: } E_1 = 0, E_2 = 1, \text{ AD: } E_1 = 0, E_2 = 0$$

Indicator Variables

- The regression equation from the Salary Survey Data is:

$$S = \beta_0 + \beta_1 X + \gamma_1 E_1 + \gamma_2 E_2 + \delta_1 M + \epsilon$$

Indicator Variables

- The regression equation from the Salary Survey Data is:

$$S = \beta_0 + \beta_1 X + \gamma_1 E_1 + \gamma_2 E_2 + \delta_1 M + \epsilon$$

- There is a different valid regression equation for each of the six (three education and two managmeent) categories.

Category	E	M	Regression Equation
1	1	0	$S = (\beta_0 + \gamma_1) + \beta_1 X + \epsilon$
2	1	1	$S = (\beta_0 + \gamma_1 + \delta_1) + \beta_1 X + \epsilon$
3	2	0	$S = (\beta_0 + \gamma_2) + \beta_1 X + \epsilon$
4	2	1	$S = (\beta_0 + \gamma_2 + \delta_1) + \beta_1 X + \epsilon$
5	3	0	$S = \beta_0 + \beta_1 X + \epsilon$
6	3	1	$S = (\beta_0 + \delta_1) + \beta_1 X + \epsilon$

Indicator Variables

```r
d <- P130
d$E1 <- as.numeric(d$E == 1)
d$E2 <- as.numeric(d$E == 2)
mod <- lm(S ~ 1 + X + E1 + E2 + M, data=d)
summary(mod)
```

```
##
## Call:
## lm(formula = S ~ 1 + X + E1 + E2 + M, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1884.60  -653.60    22.23   844.85  1716.47
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11031.81     383.22  28.787  < 2e-16 ***
## X             546.18      30.52  17.896  < 2e-16 ***
## E1          -2996.21     411.75  -7.277 6.72e-09 ***
## E2            147.82     387.66   0.381    0.705
## M            6883.53     313.92  21.928  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1027 on 41 degrees of freedom
## Multiple R-squared:  0.9568, Adjusted R-squared:  0.9525
## F-statistic: 226.8 on 4 and 41 DF,  p-value: < 2.2e-16
```

Your turn

Interpret the regression coefficients. Assume that the residual patterns are satisfactory.

Model Comparison

Table 3

	Dependent variable:	
	S	
	(1)	(2)
X	546.184*** (30.519)	570.087*** (38.559)
E1	−2,996.210*** (411.753)	
E2	147.825 (387.659)	
E		1,578.750*** (262.322)
M	6,883.531*** (313.919)	6,688.130*** (398.276)
Constant	11,031.810*** (383.217)	6,963.478*** (665.695)
Observations	46	46
R^2	0.957	0.928
Adjusted R^2	0.953	0.923
Residual Std. Error	1,027.437 (df = 41)	1,312.789 (df = 42)
F Statistic	226.836*** (df = 4; 41)	179.627*** (df = 3; 42)
Note:		*p<0.1; **p<0.05; ***p<0.01

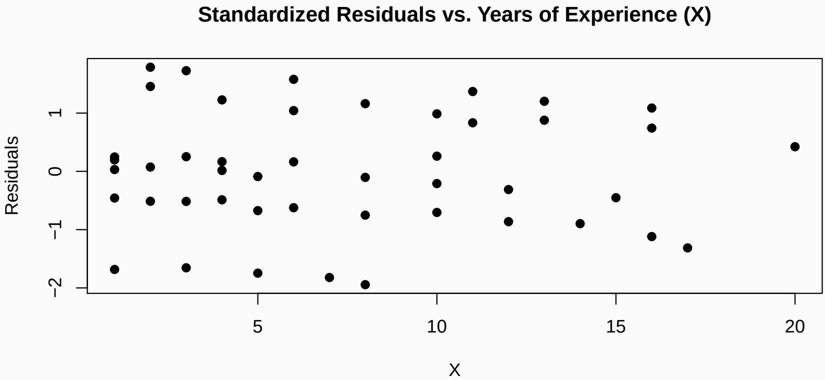
Regression Diagnostics

Before we continue we check the residuals

1) Residuals vs. Years of Experience
2) Residuals vs. Categories from Dummys

Regression Diagnostics

```r
plot(x = d$X, y = rstandard(mod), pch=19,
     ylab="Residuals", xlab = "X",
     main = "Standardized Residuals vs. Years of Experience (X)")
```



Regression Diagnostics

```r
d$cat <- factor((paste0("E=",d$E,"&M=",d$M)))
plot(x = as.numeric(d$cat), y = rstandard(mod), pch=19, xaxt="n",
     ylab="Residuals", xlab = "Category",
     main = "Standardized Residuals vs. Education-Management Category")
axis(1,at=1:6,labels=levels(d$cat))
```