The Salary Survey

马学俊(主讲) 杜悦(助教)

苏州大学
数学科学学院

https://xuejunma.github.io/

Outline

1. The Salary Survey

2. Interaction effect

The Salary Survey

- The objective of the survey was to identify and quantify those variables that determine salary differentials
- The response variable: salary (S)
- The predictors:
  - experience (X), measured in years
  - education (E),
    - 1: high school (H.S.)
    - 2: bachelor degree (B.S.)
    - 3: advanced degree
  - management (M),
    - 1: management
    - 0: otherwise

```
> rm(list=ls())
> dat <- read.table("p130.txt", head=TRUE)
> attach(dat)
> head(dat)
      S X E M
1 13876 1 1 1
2 11608 1 3 0
3 18701 1 3 1
```

- Note that when using indicator variables to represent a set of categories, the number of these variables required is one less than the number of categories.
- For example, in the case of the education categories above, we create two indicator variables $E1$ and $E2$, where

$$E_{i1} = \begin{cases} 1 = & \text{if } i\text{th person is in the H.S. category} \\ 0 = & otherwise \end{cases}$$

and

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

- stated above, these two variables taken together uniquely represent the three groups.
  - H.S., E1 = 1, E2 = 0;
  - B.S., E1 = 0, E2 = 1;
  - advanced degree, E1 = 0, E2 = 0.

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

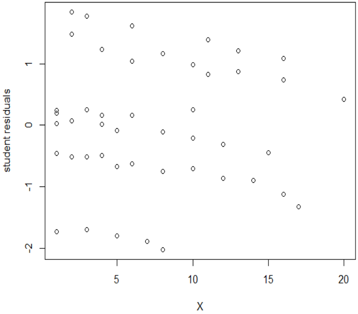
Table: Regression Equations for the Six Categories of Education and Management

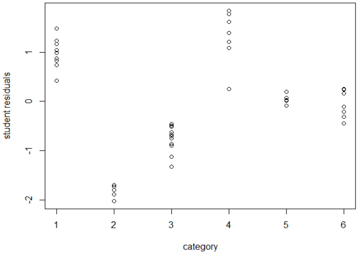
Category	E	M	Regression Equation
1	1	0	$S = \beta_1 X + \varepsilon + (\beta_0 + \gamma_1)$
2	1	1	$S = \beta_1 X + \varepsilon + (\beta_0 + \gamma_1 + \delta_1)$
3	2	0	$S = \beta_1 X + \varepsilon + (\beta_0 + \gamma_2)$
4	2	1	$S = \beta_1 X + \varepsilon + (\beta_0 + \gamma_2 + \delta_1)$
5	3	0	$S =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

```
> E1 <- as.numeric(dat$E == 1)
> E2 <- as.numeric(dat$E == 2)
> fit <- lm(S~X+E1+E2+M, data=dat)
> summary(fit)
            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 ***
```

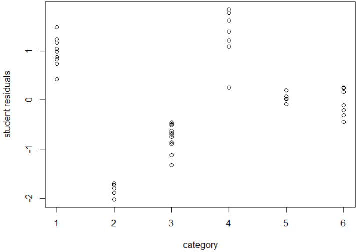
$$S = \beta_0 + \beta_1 X + \gamma_1 E_1 + \gamma_2 E_2 + \delta_1 M + \varepsilon$$

- The coefficient of X is 546.16. That is, each additional year of experience is estimated to be worth an annual salary increment of 546.
- The coefficient of the `management` indicator variable, δ_1, is estimated to be 6883.50. Interpret this amount to be the average incremental value in annual salary associated with a management position.
- For the education variables,
  - γ_1:measures the salary differential for the H.S. category relative to the advanced degree category
  - γ_2:measures the differential for the B.S. category relative to the advanced degree category.
  - $\gamma_2 - \gamma_1$: measures the differential salary for the H.S. category relative to the B.S. category.
  - An advanced degree is worth 2996 more than a high school diploma,
  - A B.S. is worth 148 more than an advanced degree (this differential is not statistically significant, $t = 0.38$),
  - A B.S. is worth about 3144 more than a high school diploma.


student residuals
-2
-1
0
1
5
10
15
20
X

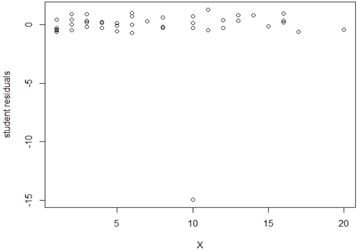


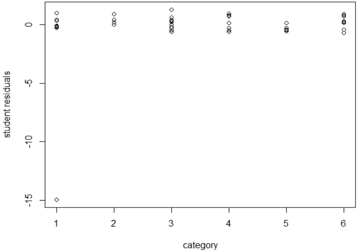
- The graph that the residuals cluster by size according to their education-management category.
- The combinations of education and management have not been satisfactorily treated in the model.
- Within each of the six groups, the residuals are either almost totally positive or totally negative.
- This behavior implies that the model does not adequately explain the relationship between salary and experience, education, and management variables.
- The graph points to some hidden structure in the data that has not been explored.
- The graphs strongly suggest that the effects of education and management status on salary determination are not additive.


student residuals
-2
-1
0
1
category
1
2
3
4
5
6

$$S = \beta_0 + \beta_1 X + \gamma_1 E_1 + \gamma_2 E_2 + \delta_1 M + \alpha_1 E_1 \bullet M + \alpha_2 E_2 \bullet M + \varepsilon$$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 11203.434     79.065 141.698  < 2e-16 ***
X             496.987      5.566  89.283  < 2e-16 ***
E1          -1730.748    105.334 -16.431  < 2e-16 ***
E2           -349.078     97.568  -3.578 0.000945 ***
M            7047.412    102.589  68.695  < 2e-16 ***
E1:M        -3066.035    149.330 -20.532  < 2e-16 ***
E2:M         1836.488    131.167  14.001  < 2e-16 ***
```


student residuals
0
-5
-10
-15
5
10
15
20
X


student residuals
0
-5
-10
-15
1
2
3
4
5
6
category

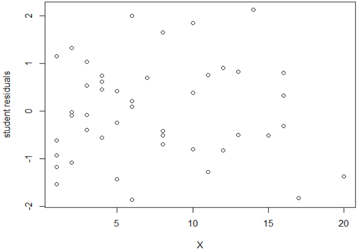
Observation 33 Deleted

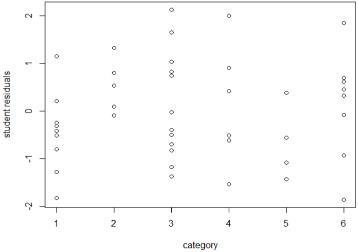
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 11199.714     30.533 366.802  < 2e-16 ***
X             498.418      2.152 231.640  < 2e-16 ***
E1n         -1741.336     40.683 -42.803  < 2e-16 ***
E2n          -357.042     37.681  -9.475 1.49e-11 ***
M            7040.580     39.619 177.707  < 2e-16 ***
E1n:M       -3051.763     57.674 -52.914  < 2e-16 ***
E2n:M        1997.531     51.785  38.574  < 2e-16 ***
```

- increments of approximately 500 are added to a starting salary that is specified for each of the six education-management groups.
- Since the final regression model is not additive, it is rather difficult to directly interpret the coefficients of the indicator variables.
- To see how the qualitative variables affect salary differentials, we use the coefficients to form estimates of the base salary for each of the six categories. These

Table: Regression Equations for the Six Categories of Education and Management

Category	E	M	Coefficients	Estimation
1	1	0	$\beta_0 + \gamma_1$	9459
2	1	1	$\beta_0 + \gamma_1 + \delta_1 + \alpha_1$	13448
3	2	0	$\beta_0 + \gamma_2$	10843
4	2	1	$\beta_0 + \gamma_2 + \delta_1 + \alpha_2$	19880
5	3	0	β_0	11200
6	3	1	$\beta_0 + \delta_1$	18240


student residuals
-2
-1
0
1
2
5
10
15
20
X


student residuals
-2
-1
0
1
2
1
2
3
4
5
6
category