Lecture 9: Dummy-Variable Regression

Ailin Zhang

2023-05-31

Agenda

- Dummy-Variable Regression
- Qualitative (Categorical) Predictors
- Interactions Between Qualitative Variables

Dichotomous Factor

- In a general linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

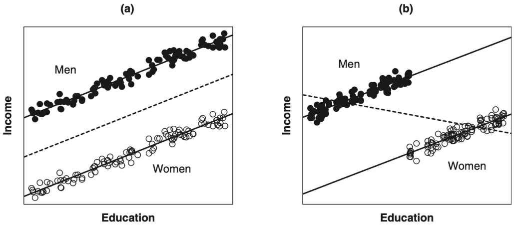
  - Y: A continuous dependent variable
  - X_j: Continuous independent variables.

- What if X is no longer continuous?

  gender, blood type, marital status, education level, etc ...

- Categorical variables can provide useful information in predicting the response variable Y.

Example: Relationship between Education and Income among Women and Men



The additive dummy-variable regression model: $Y_i = \beta_0 + \beta_1 X_i + \gamma D_i + \epsilon_i$

Where D, called a dummy-variable regressor or an indicator variable is coded 1 for men and 0 for women.

$$D_i = \begin{cases} 1 & \text{for men} \\ 0 & \text{for women} \end{cases}$$

Example: Relationship between Education and Income among Women and Men

- For women the model becomes $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
- For men, $Y_i = \beta_0 + \beta_1 X_i + \gamma + \epsilon_i$
- The coefficient γ for the dummy regressor gives the difference in intercepts for the two regression lines.
- The within-gender regression lines are parallel, γ also represents the constant vertical separation between the lines.
- It may interpreted as the expected income advantage accruing to men when education is held constant.
- If men were disadvantaged relative to women with the same level of education, then γ would be negative.

Dichotomous Factor

- Dummy variable coded with 0 or 1
- A model incorporating a dummy regressor represents parallel regression surfaces
- The constant vertical separation between the surfaces given by the coefficient of the dummy regressor.

Statistical Inference

- To determine whether gender affects income, controlling for education: $H_0 : \gamma = 0$ v.s. $H_A : \gamma \neq 0$
  - t-test
  - F-test with reduced and full models
  - Statistical-inference procedures of the previous lectures apply

Polytomous Factors: General Qualititative/ Categorical Predictors

```
salary = read.table("salary.txt", header=TRUE)
```

$$
\begin{array}{rcc}
S & = & \text{Salary} \\
X & = & \text{Experience, in years} \\
E & = & \text{Education} \\
 & & \text{1 if H.S. only,} \\
 & & \text{2 if Bachelor's only,} \\
 & & \text{3 if Advanced degree} \\
M & = & \text{Management Status} \\
 & & \text{1 if manager, 0 if non-manager}
\end{array}
$$

Indicator Varibales (aka. Dummy Variables)

- For Education (E), it contains 3 categories, so we can start from 3 indicator variables:

$$E_{i1} = \begin{cases} 1 & \text{if sample i has a high school diploma only} \\ 0 & \text{otherwise} \end{cases}$$

$$E_{i2} = \begin{cases} 1 & \text{if sample i has a B.S. only} \\ 0 & \text{otherwise} \end{cases}$$

$$E_{i3} = \begin{cases} 1 & \text{if sample i has an advanced degress only} \\ 0 & \text{otherwise} \end{cases}$$

Additional constraint on indicator variables

- Education (E) only has 3 categories.
- Each sample must fall in exactly one of the 3 categories: only one of E_1, E_2, and E_3 can be 1 and the other 2 must be 0.
- Identity holds: $E_1 + E_2 + E_3 = 1$
- One of E_1, E_2 and E_3 is redundant. The last one is known once the remainin are known.
- In general, a categorical predictor with c categories needs only c-1 indicator variables
- The command `as.factor()` or `factor()` tells R that E is categorical and the indicator variables E1, E2, E3 are created automatically. By default, R drops the indicator E1 for the lowest level

Example

```
salary$Edu = factor(salary$E,
                    labels=c("High School","College","Advanced"))
lmsalary = lm(S ~ X+Edu, salary)
summary(lmsalary)
```

```
##
## Call:
## lm(formula = S ~ X + Edu, data = salary)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -4320  -3182  -1372   2812   6079
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10474.3     1305.4   8.024 5.19e-10 ***
## X              548.6      107.6   5.100 7.69e-06 ***
## EduCollege    3221.1     1275.8   2.525  0.01544 *
## EduAdvanced   4780.1     1422.7   3.360  0.00167 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3622 on 42 degrees of freedom
## Multiple R-squared:  0.4498, Adjusted R-squared:  0.4104
## F-statistic: 11.44 on 3 and 42 DF,  p-value: 1.291e-05
```

Treat as E_1 is removed from the model

When E_1 is removed from the model:

$$Y = \beta_0 + \beta_1 X + \gamma_2 E_2 + \gamma_3 E_3 + \epsilon$$

And we can look at the E(Y) for different education levels:

Education(E)	Indicator	E(Y)
1 (HS)	$E_2 = E_3 = 0$	$\beta_0 + \beta_1 X$
2 (B.S.)	$E_2 = 1, E_3 = 0$	$\beta_0 + \beta_1 X + \gamma_2$
3 (Advanced)	$E_2 = 0, E_3 = 1$	$\beta_0 + \beta_1 X + \gamma_3$

Based on the model above, for people w/ the same years of experience (X), the difference in their mean salary are

- γ_2 : B.S.-HS
- γ_3 : Advanced-HS
- $\gamma_3 - \gamma_2$: Advanced-B.S.

Hypothesis Testing of Parameters

- To test whether those w/ a Bachelor's degree had a higher mean salary than those w/ only a HS diploma, after accounting for experience, which parameter should we test?

  $H_0 : \gamma_2 = 0$ v.s. $H_A : \gamma_2 > 0$

- To test whether those w/ an advanced degree had a higher mean salary than those w/ only a HS diploma, after accounting for experience, which parameter should we test?

  $H_0 : \gamma_3 = 0$ v.s. $H_A : \gamma_3 > 0$

- To test whether an advanced degree increases mean salary than a Bachelor's degree after accounting for experience, which parameter should we test?

  $H_0 : \gamma_3 = \gamma_2$ v.s. $H_A : \gamma_3 > \gamma_2$

Right or Wrong?

```r
lm0 = lm(S ~ X + E, data=salary)
summary(lm0)$coef
```

```
##             Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 8279.8631  1814.5882 4.562943 4.175777e-05
## X            560.8138   105.8320 5.299095 3.780641e-06
## E           2418.4016   706.8593 3.421334 1.377546e-03
```

- Issue: R treats E (education) as numerical values 1, 2, and 3, not a categorical one.

Confidence Intervals

```
confint(lmsalary, level=0.95)
```

```
##                  2.5 %     97.5 %
## (Intercept) 7839.8114 13108.7185
## X            331.5143   765.7014
## EduCollege   646.4217  5795.8228
## EduAdvanced 1908.9201  7651.3558
```

- Each extra year of experience worths \$548 more in salary, with a 95% CI of \$331.5 to \$765.1.
- Completing college increases salary by \$3221, with a 95%CI of \$646.4 to \$5795.8.
- Completing college + advanced degree increases salary by \$4780, with a 95% CI of \$1908.9 to \$7651.4.

How to Drop a Different Indicator Variable in R?

If not happy with R's choice of which indicator to drop, one can manually create the indicator variables E1 and E3

```
salary$E1=ifelse(salary$E==1, 1, 0)
salary$E3=ifelse(salary$E==3, 1, 0)
lm1 = lm(S ~ X + E1 + E3, data = salary)
summary(lm1)$coef
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 13695.3873  1225.0304 11.179631 3.626080e-14
## X             548.6078   107.5742  5.099808 7.694601e-06
## E1          -3221.1223  1275.8158 -2.524755 1.544273e-02
## E3           1559.0156  1338.6033  1.164658 2.507301e-01
```

```
# For one tail test:
pt(1.165,df=42, lower.tail=F)
```

```
## [1] 0.1252967
```

The large P-value indicate an advanced degree did not increase more salary than B.S significantly

It Doesn't Matter Which Indicator is Dropped

The 2 models have identical fitted values $\hat{y}_i$, residuals e_i, SSE, SSR and hence $\hat{\sigma}^2 = \text{MSE}$, multiple and adjust R^2, and many others, despite they drop different indicators.

```r
summary(lmsalary)
```

```
##
## Call:
## lm(formula = S ~ X + Edu, data = salary)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -4320  -3182  -1372   2812   6079
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10474.3     1305.4   8.024 5.19e-10 ***
## X              548.6      107.6   5.100 7.69e-06 ***
## EduCollege    3221.1     1275.8   2.525  0.01544 *
## EduAdvanced   4780.1     1422.7   3.360  0.00167 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3622 on 42 degrees of freedom
## Multiple R-squared:  0.4498, Adjusted R-squared:  0.4104
## F-statistic: 11.44 on 3 and 42 DF,  p-value: 1.291e-05
```

It Doesn't Matter Which Indicator is Dropped

```r
summary(lm1)
```

```
##
## Call:
## lm(formula = S ~ X + E1 + E3, data = salary)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -4320  -3182  -1372   2812   6079
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  13695.4     1225.0  11.180 3.63e-14 ***
## X              548.6      107.6   5.100 7.69e-06 ***
## E1           -3221.1     1275.8  -2.525   0.0154 *
## E3            1559.0     1338.6   1.165   0.2507
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3622 on 42 degrees of freedom
## Multiple R-squared:  0.4498, Adjusted R-squared:  0.4104
## F-statistic: 11.44 on 3 and 42 DF,  p-value: 1.291e-05
```

Model with Two Categorical Predictors

Now let's take another categorical predictor, management status (M), into account.

$$M = \begin{cases} 1 & \text{if manager} \\ 0 & \text{if not manager} \end{cases}$$

Since M is categorical, just like E, we should create indicator variables M_0 and M_1 for the two categories. The model now becomes:

$$Y = \beta_0 + \beta_1 X + \alpha_0 M_0 + \alpha_1 M_1 + \gamma_1 E_1 + \gamma_2 E_2 + \gamma_3 E_3 + \epsilon$$

After droping one column from each categorical group:

$$Y = \beta_0 + \beta_1 X + \alpha_1 M_1 + \gamma_2 E_2 + \gamma_3 E_3 + \epsilon$$

Model with Two Categorical Predictors

Education(E)	E(Y) Not Manager: $M_1 = 0$	E(Y) Manager: $M_1 = 1$
1 (HS,$E_2 = E_3 = 0$)	$\beta_0 + \beta_1 X$	$\beta_0 + \beta_1 X + \alpha_1$
2 (B.S., $E_2 = 1, E_3 = 0$)	$\beta_0 + \beta_1 X + \gamma_2$	$\beta_0 + \beta_1 X + \gamma_2 + \alpha_1$
3 (Adv, $E_2 = 0, E_3 = 1$)	$\beta_0 + \beta_1 X + \gamma_3$	$\beta_0 + \beta_1 X + \gamma_3 + \alpha_1$

The model implies that on average:

- Managers earn α_1 more than non-managers, regardless of E and X
- Completing college increases salary by γ_2, regardless of M and X
- Advanced degree earn γ_3 more than HS, regardless of M and X

Model Construction

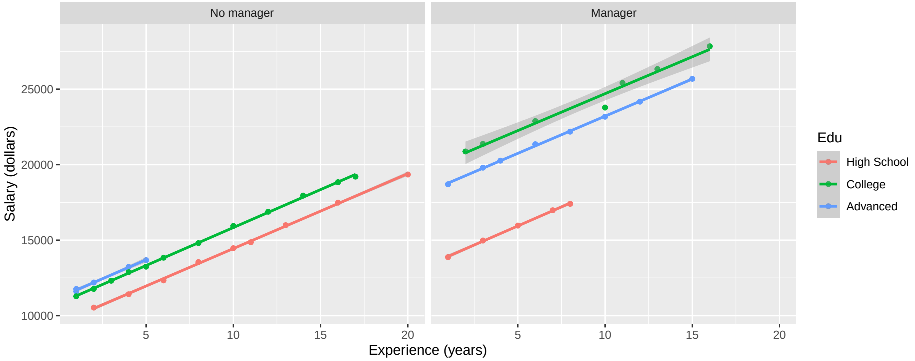
```r
salary$Man = factor(salary$M,
                    labels=c("No manager","Manager"))
lmsalm = lm(S ~ X+Edu+Man, salary)
summary(lmsalm)
```

```
##
## Call:
## lm(formula = S ~ X + Edu + Man, data = salary)
##
## 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)  8035.60     386.69  20.781  < 2e-16 ***
## X             546.18      30.52  17.896  < 2e-16 ***
## EduCollege   3144.04     361.97   8.686 7.73e-11 ***
## EduAdvanced  2996.21     411.75   7.277 6.72e-09 ***
## ManManager   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
```

Model Construction

The previous model assumes the effect of management status (M) on salary (S) does not change with education levels E. However, from the plot below

```
library(ggplot2)
ggplot(salary, aes(x = X, y = S, color=Edu)) + geom_point() +
  facet_grid(~Man) + geom_smooth(method="lm", formula='y~x') +
  xlab("Experience (years)") + ylab("Salary (dollars)")
```



Adding Interaction Terms

We may consider the model below with M*E interactions.

$$Y = \beta_0 + \beta_1 X + \alpha_1 M_1 + \gamma_2 E_2 + \gamma_3 E_3 + \theta_2 (M_1 \cdot E_2) + \theta_3 (M_1 \cdot E_3) + \epsilon$$

Here $(M_1 \cdot E_2)$ means the product of the variables M_1 and E_2.

Education(E)	E(Y) Not Manager: $M_1 = 0$	E(Y) Manager: $M_1 = 1$
1 (HS, $E_2 = E_3 = 0$)	$\beta_0 + \beta_1 X$	$\beta_0 + \beta_1 X + \alpha_1$
2 (B.S., $E_2 = 1, E_3 = 0$)	$\beta_0 + \beta_1 X + \gamma_2$	$\beta_0 + \beta_1 X + \gamma_2 + \alpha_1 + \theta_2$
3 (Adv, $E_2 = 0, E_3 = 1$)	$\beta_0 + \beta_1 X + \gamma_3$	$\beta_0 + \beta_1 X + \gamma_3 + \alpha_1 + \theta_3$

- For HS, managers earns α_1 more than others with the same X
- For B.S, managers earns $\alpha_1 + \theta_2$ more than others with the same X
- For advanced degree, managers earns $\alpha_1 + \theta_3$ more than others with the same X

Model Construction (with Interaction)

```r
lmsalm_int = lm(S ~ X+Edu+Man+Edu*Man, salary)
summary(lmsalm_int)
```

```
##
## Call:
## lm(formula = S ~ X + Edu + Man + Edu * Man, data = salary)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -928.13  -46.21   24.33   65.88  204.89
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            9472.685     80.344  117.90   <2e-16 ***
## X                       496.987      5.566   89.28   <2e-16 ***
## EduCollege             1381.671     77.319   17.87   <2e-16 ***
## EduAdvanced            1730.748    105.334   16.43   <2e-16 ***
## ManManager             3981.377    101.175   39.35   <2e-16 ***
## EduCollege:ManManager  4902.523    131.359   37.32   <2e-16 ***
## EduAdvanced:ManManager 3066.035    149.330   20.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 173.8 on 39 degrees of freedom
## Multiple R-squared:  0.9988, Adjusted R-squared:  0.9986
## F-statistic:  5517 on 6 and 39 DF,  p-value: < 2.2e-16
```

F-test of Interactions:

H_0: Model without interactions is true v.s. H_A: Model with interactions is true

```
anova(lmsalm,lmsalm_int)
```

```
## Analysis of Variance Table
##
## Model 1: S ~ X + Edu + Man
## Model 2: S ~ X + Edu + Man + Edu * Man
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1     41 43280719
## 2     39  1178168  2  42102552 696.84 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Conclusion: There are significant E*M interactions!

Extra: Interaction between Qualitative and Quantitative variables

We may consider the model below with M*E and M*X interactions.

$$Y = \beta_0 + \beta_1 X + \beta_2 (X \cdot M_1) + \alpha_1 M_1 + \gamma_2 E_2 + \gamma_3 E_3 + \theta_2 (M_1 \cdot E_2) + \theta_3 (M_1 \cdot E_3) + \epsilon$$

Education(E)	E(Y) Not Manager: $M_1 = 0$	E(Y) Manager: $M_1 = 1$
1 (HS,$E_2 = E_3 = 0$)	$\beta_0 + \beta_1 X$	$\beta_0 + (\beta_1 + \beta_2)X + \alpha_1$
2 (B.S., $E_2 = 1, E_3 = 0$)	$\beta_0 + \beta_1 X + \gamma_2$	$\beta_0 + (\beta_1 + \beta_2)X + \gamma_2 + \alpha_1 + \theta_2$
3 (Adv, $E_2 = 0, E_3 = 1$)	$\beta_0 + \beta_1 X + \gamma_3$	$\beta_0 + (\beta_1 + \beta_2)X + \gamma_3 + \alpha_1 + \theta_3$

- For HS, managers earns $\alpha_1 + \beta_2 X$ more than others with the same X
- For B.S, managers earns $\alpha_1 + \theta_2 \beta_2 X$ more than others with the same X
- For advanced degree, managers earns $\alpha_1 + \theta_3 \beta_2 X$ more than others with the same X

Model Construction

```
lmsalm_int2 = lm(S ~ X+Edu+Man+Edu*Man+X*Man, salary)
summary(lmsalm_int2)
```

```
##
## Call:
## lm(formula = S ~ X + Edu + Man + Edu * Man + X * Man, data = salary)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -921.81  -38.21   13.26   67.29  216.12
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            9444.419     91.420 103.308   <2e-16 ***
## X                       499.814      7.037  71.024   <2e-16 ***
## EduCollege             1386.853     78.268  17.719   <2e-16 ***
## EduAdvanced            1751.666    110.666  15.828   <2e-16 ***
## ManManager             4033.223    128.336  31.427   <2e-16 ***
## EduCollege:ManManager  4916.570    133.987  36.694   <2e-16 ***
## EduAdvanced:ManManager 3057.767    150.924  20.260   <2e-16 ***
## X:ManManager             -7.739     11.644  -0.665     0.51
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 175.1 on 38 degrees of freedom
## Multiple R-squared:  0.9988, Adjusted R-squared:  0.9986
## F-statistic:  4661 on 7 and 38 DF,  p-value: < 2.2e-16
```

F-test of Interactions (X*Man):

H_0: Model without interactions is true v.s. H_A: Model with interactions is true

```
anova(lmsalm_int,lmsalm_int2)
```

```
## Analysis of Variance Table
##
## Model 1: S ~ X + Edu + Man + Edu * Man
## Model 2: S ~ X + Edu + Man + Edu * Man + X * Man
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1     39 1178168
## 2     38 1164630  1     13538 0.4417 0.5103
```

Conclusion: There is no significant X*M interactions!