BS2280 - Econometrics 1

Lecture 8 - Part 2: Dummy Variables

Dr. Yichen Zhu

Structure of today's lecture

1. Dummies With More Than Two Categories
2. Dummy Variable Trap
3. Slope Dummy Variables
4. Hypothesis Testing of Dummy Variables

Intended Learning Outcomes

- Understanding dummies with more than 2 categories
- Explaining dummy variable trap
- Understanding slope dummy variables
- Testing joint explanatory power of dummy variables

Background

- Often we will be confronted with qualitative variables that have more than 2 categories
- Example: Regional dummies, industry dummies, etc.

- **Shanghai Secondary School Example**:
- 2 types of regular schools: General and Vocational
- 2 types of occupational schools: Technical and Skilled Workers
- We have now in total four qualitative categories – General, Vocational, Technical and Skilled Workers

- How can we include them into our model?

Background

- Standard procedure: choose one category as the reference/base category and define dummy variables for each of the others
- Good practice: select the most normal or basic category as the reference/base category

- **Shanghai Secondary School Example**:
- We have now in total four qualitative categories – General, Vocational, Technical and Skilled Workers
- Reference/Base category: General schools
- We want to estimate the following model:

$$COST_i = \beta_1 + \beta_2 N_i + \delta_V VOC_i + \delta_T TECH_i + \delta_W WORKER_i + u_i$$

- **We do not include a dummy variable for Reference/Base category: General schools!!!! (will discuss why later)**

Background

- Reference/Base category: General schools

$$COST_i = \beta_1 + \beta_2 N_i + \delta_V VOC_i + \delta_T TECH_i + \delta_W WORKER_i + u_i$$

- Accordingly we will define dummy variables for the other three types: Vocational, Technical and Skilled Workers
- VOC: dummy for the vocational schools:
  $VOC = 1$: vocational school, $VOC = 0$ otherwise
- $TECH$: dummy for the technical schools:
  $TECH = 1$: technical school, $TECH = 0$ otherwise
- $WORKER$: dummy for the Skilled Workers schools:
  $WORKER = 1$: Skilled Workers school, $WORKER = 0$ otherwise

Background

- Reference/Base category: General schools

$$COST_i = \beta_1 + \beta_2 N_i + \delta_V VOC_i + \delta_T TECH_i + \delta_W WORKER_i + u_i$$

- Accordingly we will define dummy variables for the other three types: Vocational, Technical and Skilled Workers
- VOC: dummy for the vocational schools:
  $VOC = 1$: vocational school, $VOC = 0$ otherwise
- $TECH$: dummy for the technical schools:
  $TECH = 1$: technical school, $TECH = 0$ otherwise
- $WORKER$: dummy for the Skilled Workers schools:
  $WORKER = 1$: Skilled Workers school, $WORKER = 0$ otherwise

Background

- Reference/Base category: General schools

$$COST_i = \beta_1 + \beta_2 N_i + \delta_V VOC_i + \delta_T TECH_i + \delta_W WORKER_i + u_i$$

- Accordingly we will define dummy variables for the other three types: Vocational, Technical and Skilled Workers
- VOC: dummy for the vocational schools:
  $VOC = 1$: vocational school, $VOC = 0$ otherwise
- $TECH$: dummy for the technical schools:
  $TECH = 1$: technical school, $TECH = 0$ otherwise
- $WORKER$: dummy for the Skilled Workers schools:
  $WORKER = 1$: Skilled Workers school, $WORKER = 0$ otherwise

Background

- Reference/Base category: General schools

$$COST_i = \beta_1 + \beta_2 N_i + \delta_V VOC_i + \delta_T TECH_i + \delta_W WORKER_i + u_i$$

- Accordingly we will define dummy variables for the other three types: Vocational, Technical and Skilled Workers
- VOC: dummy for the vocational schools:
  $VOC = 1$: vocational school, $VOC = 0$ otherwise
- $TECH$: dummy for the technical schools:
  $TECH = 1$: technical school, $TECH = 0$ otherwise
- $WORKER$: dummy for the Skilled Workers schools:
  $WORKER = 1$: Skilled Workers school, $WORKER = 0$ otherwise

Background

- Reference/Base category: General schools

$$COST_i = \beta_1 + \beta_2 N_i + \delta_V VOC_i + \delta_T TECH_i + \delta_W WORKER_i + u_i$$

- Accordingly we will define dummy variables for the other three types: Vocational, Technical and Skilled Workers
- VOC: dummy for the vocational schools:
  $VOC = 1$: vocational school, $VOC = 0$ otherwise
- $TECH$: dummy for the technical schools:
  $TECH = 1$: technical school, $TECH = 0$ otherwise
- $WORKER$: dummy for the Skilled Workers schools:
  $WORKER = 1$: Skilled Workers school, $WORKER = 0$ otherwise

Interpretations

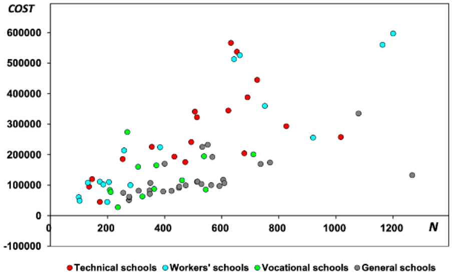
	$COST_i = \beta_1 + \beta_2 N_i + \delta_V VOC_i + \delta_T TECH_i + \delta_W WORKER_i + u_i$	
General school	$VOC = TECH = WORKER = 0$	$COST_i = \beta_1 + \beta_2 N_i + u_i$
Vocational school	$VOC = 1$; $TECH = WORKER = 0$	$COST_i = (\beta_1 + \delta_V) + \beta_2 N_i + u_i$
Technical school	$TECH = 1$; $VOC = WORKER = 0$	$COST_i = (\beta_1 + \delta_T) + \beta_2 N_i + u_i$
Skilled Worker school	$WORKER = 1$; $VOC = TECH = 0$	$COST_i = (\beta_1 + \delta_W) + \beta_2 N_i + u_i$

- Each dummy will have a coefficient which represents the extra overhead costs of the schools relative to the reference/base category.
- Example:
  δ_V represents the costs differences between general school (Reference/Base category) and vocational school
  δ_T represents the costs differences between general school (Reference/Base category) and technical school
  δ_W represents the costs differences between general school (Reference/Base category) and skilled worker school

Example

- The table shows the data for the first 10 schools in the sample
- The scatter diagram shows the data for the entire sample, differentiating by type of school.

School	Type	*COST*	*N*	*TECH*	*WORKER*	*VOC*
1	Technical	345,000	623	1	0	0
2	Technical	537,000	653	1	0	0
3	General	170,000	400	0	0	0
4	Workers'	526.000	663	0	1	0
5	General	100,000	563	0	0	0
6	Vocational	28,000	236	0	0	1
7	Vocational	160,000	307	0	0	1
8	Technical	45,000	173	1	0	0
9	Technical	120,000	146	1	0	0
10	Workers'	61,000	99	0	1	0



Example

```
Call:
lm(formula = COST ~ N + TECH + WORKER + VOC, data = schools)

Residuals:
    Min      1Q  Median      3Q     Max
-246690  -46624   -6272   38957  250374

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -54893.09   26673.08  -2.058   0.0434 *
N              342.63      40.22   8.519 2.25e-12 ***
TECH        154110.89   26760.41   5.759 2.15e-07 ***
WORKER      143362.38   27852.80   5.147 2.38e-06 ***
VOC          53228.64   31061.65   1.714   0.0911 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 88580 on 69 degrees of freedom
Multiple R-squared:  0.632,   Adjusted R-squared:  0.6107
F-statistic: 29.63 on 4 and 69 DF,  p-value: 2.387e-14
```

$$\widehat{COST}_i = -54893.09 + 342 N_i + 53228.64\, VOC_i + 154110.89\, TECH_i + 143362.38\, WORKER_i$$

Example

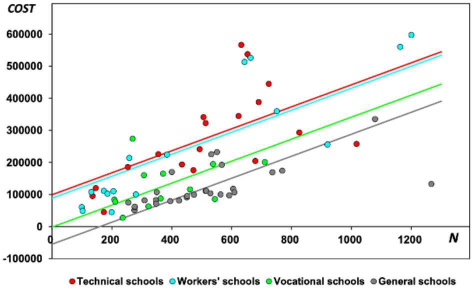
$$\widehat{COST}_i = -54893.09 + 342N_i + 53228.64\,VOC_i + 154110.89\,TECH_i + 143362.38\,WORKER_i$$

General school	$VOC = TECH = WORKER = 0$	$\widehat{COST}_i = -54893.09 + 342N_i$
Vocational school	$VOC = 1$; $TECH = WORKER = 0$	$\widehat{COST}_i = -54893.09 + 53228.64 + 342N_i$ $\widehat{COST}_i = -1664.45 + 342N_i$
Technical school	$TECH = 1$; $VOC = WORKER = 0$	$\widehat{COST}_i = -54893.09 + 154110.89 + 342N_i$ $\widehat{COST}_i = 99217.8 + 342N_i$
Skilled Worker school	$WORKER = 1$; $VOC = TECH = 0$	$\widehat{COST}_i = -54893.09 + 143362.38 + 342N_i$ $\widehat{COST}_i = 88469.29 + 342N_i$

Example

$$\widehat{COST}_i = -54893.09 + 342N_i + 53228.64\,VOC_i + 154110.89\,TECH_i + 143362.38\,WORKER_i$$

General school	$VOC = TECH = WORKER = 0$	$\widehat{COST}_i = -54893.09 + 342N_i$
Vocational school	$VOC = 1$; $TECH = WORKER = 0$	$\widehat{COST}_i = -1664.45 + 342N_i$
Technical school	$TECH = 1$; $VOC = WORKER = 0$	$\widehat{COST}_i = 99217.8 + 342N_i$
Skilled Worker school	$WORKER = 1$; $VOC = TECH = 0$	$\widehat{COST}_i = 88469.29 + 342N_i$



Dummy Variable Trap

- **Dummy variable trap**: When the number of dummy variables created is equal to the number of values the categorical value can take on.
- When this happens, at least two of the dummy variables will suffer from perfect multicollinearity. That is, they'll be perfectly correlated.

Age	Gender
23	Female
25	Male
22	Male
21	Female

→

Age	Male	Female
23	0	1
25	1	0
22	1	0
21	0	1

- In this case, Female and Male are perfectly correlated and have a correlation coefficient of -1.
- We create $k - 1$ dummy variables to avoid falling into what is called the dummy variable trap

Dummy Variable Trap

- When creating dummy variables, we have to follow the rule:
  If our qualitative variable has k categories, we always just create $k - 1$ dummy variables
- **Shanghai Secondary School Example:**
  - Dummy "type of school" in total has 4 qualitative categories: General, Vocational, Technical and Skilled Workers
  - We only need to create $4 - 1 = 3$ dummies: VOC, $TECH$ and $WORKER$

$$COST_i = \beta_1 + \beta_2 N_i + \delta_V VOC_i + \delta_T TECH_i + \delta_W WORKER_i + u_i$$

  - General school is the base/benchmark/reference
- The category that is omitted is called the base/benchmark/reference
- All other categories (which would have a dummy variable created for them) are compared against the base category
- If we have no base/reference category we cannot make any comparisons!!!!

Dummy Variable Trap

- When creating dummy variables, we have to follow the rule:
  If our qualitative variable has k categories, we always just create $k-1$ dummy variables
- **Shanghai Secondary School Example:**
  - Dummy "type of school" in total has 4 qualitative categories: General, Vocational, Technical and Skilled Workers
  - We only need to create $4-1=3$ dummies: VOC, $TECH$ and $WORKER$

$$COST_i = \beta_1 + \beta_2 N_i + \delta_V VOC_i + \delta_T TECH_i + \delta_W WORKER_i + u_i$$

  - General school is the base/benchmark/reference
- The category that is omitted is called the base/benchmark/reference
- All other categories (which would have a dummy variable created for them) are compared against the base category
- If we have no base/reference category we cannot make any comparisons!!!!

Dummy Variable Trap

- When creating dummy variables, we have to follow the rule: If our qualitative variable has k categories, we always just create $k - 1$ dummy variables
- **Shanghai Secondary School Example:**
  - Dummy "type of school" in total has 4 qualitative categories: General, Vocational, Technical and Skilled Workers
  - We only need to create $4 - 1 = 3$ dummies: *VOC*, *TECH* and *WORKER*

$$COST_i = \beta_1 + \beta_2 N_i + \delta_V VOC_i + \delta_T TECH_i + \delta_W WORKER_i + u_i$$

  - General school is the base/benchmark/reference
- The category that is omitted is called the base/benchmark/reference
- All other categories (which would have a dummy variable created for them) are compared against the base category
- If we have no base/reference category we cannot make any comparisons!!!!

Dummy Variable Trap

- When creating dummy variables, we have to follow the rule:
  If our qualitative variable has k categories, we always just create $k-1$ dummy variables
- **Shanghai Secondary School Example:**
  - Dummy "type of school" in total has 4 qualitative categories: General, Vocational, Technical and Skilled Workers
  - We only need to create $4-1=3$ dummies: VOC, $TECH$ and $WORKER$

$$COST_i = \beta_1 + \beta_2 N_i + \delta_V VOC_i + \delta_T TECH_i + \delta_W WORKER_i + u_i$$

  - General school is the base/benchmark/reference
- The category that is omitted is called the base/benchmark/reference
- All other categories (which would have a dummy variable created for them) are compared against the base category
- If we have no base/reference category we cannot make any comparisons!!!!

Dummy Variable Trap

- When creating dummy variables, we have to follow the rule:
  If our qualitative variable has k categories, we always just create $k-1$ dummy variables
- **Shanghai Secondary School Example:**
  - Dummy "type of school" in total has 4 qualitative categories: General, Vocational, Technical and Skilled Workers
  - We only need to create $4-1=3$ dummies: VOC, $TECH$ and $WORKER$

$$COST_i = \beta_1 + \beta_2 N_i + \delta_V VOC_i + \delta_T TECH_i + \delta_W WORKER_i + u_i$$

  - General school is the base/benchmark/reference
- The category that is omitted is called the base/benchmark/reference
- All other categories (which would have a dummy variable created for them) are compared against the base category
- If we have no base/reference category we cannot make any comparisons!!!!

Dummy Variable Trap

- When creating dummy variables, we have to follow the rule:
  If our qualitative variable has k categories, we always just create $k-1$ dummy variables
- **Shanghai Secondary School Example:**
  - Dummy "type of school" in total has 4 qualitative categories: General, Vocational, Technical and Skilled Workers
  - We only need to create $4-1=3$ dummies: *VOC*, *TECH* and *WORKER*

$$COST_i = \beta_1 + \beta_2 N_i + \delta_V VOC_i + \delta_T TECH_i + \delta_W WORKER_i + u_i$$

  - General school is the base/benchmark/reference
- The category that is omitted is called the base/benchmark/reference
- All other categories (which would have a dummy variable created for them) are compared against the base category
- If we have no base/reference category we cannot make any comparisons!!!!

Dummy Variable Trap

- When creating dummy variables, we have to follow the rule:
  If our qualitative variable has k categories, we always just create $k-1$ dummy variables
- **Shanghai Secondary School Example:**
  - Dummy "type of school" in total has 4 qualitative categories: General, Vocational, Technical and Skilled Workers
  - We only need to create $4-1=3$ dummies: VOC, $TECH$ and $WORKER$

$$COST_i = \beta_1 + \beta_2 N_i + \delta_V VOC_i + \delta_T TECH_i + \delta_W WORKER_i + u_i$$

  - General school is the base/benchmark/reference
- The category that is omitted is called the base/benchmark/reference
- All other categories (which would have a dummy variable created for them) are compared against the base category
- If we have no base/reference category we cannot make any comparisons!!!!

Dummy Variable Trap

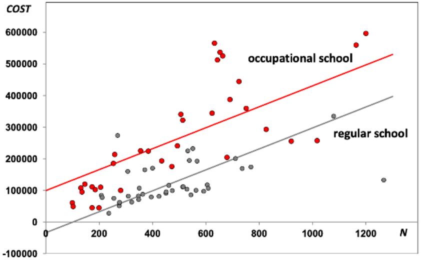
- When creating dummy variables, we have to follow the rule:
  If our qualitative variable has k categories, we always just create $k - 1$ dummy variables
- **Shanghai Secondary School Example:**
  - Dummy "type of school" in total has 4 qualitative categories: General, Vocational, Technical and Skilled Workers
  - We only need to create $4 - 1 = 3$ dummies: VOC, $TECH$ and $WORKER$

$$COST_i = \beta_1 + \beta_2 N_i + \delta_V VOC_i + \delta_T TECH_i + \delta_W WORKER_i + u_i$$

  - General school is the base/benchmark/reference
- The category that is omitted is called the base/benchmark/reference
- All other categories (which would have a dummy variable created for them) are compared against the base category
- If we have no base/reference category we cannot make any comparisons!!!!

Background

Combined regression model	$\widehat{COST}_i = -33612.6 + 331.4N_i + 133259.1\,TYPE_i$
Regular school, $TYPE = 0$	$\widehat{COST}_i = -33612.6 + 331.4N_i$
Occupational school, $TYPE = 1$	$\widehat{COST}_i = 99646 + 331.4N_i$



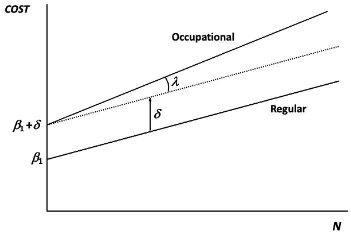
- This model assumes that the marginal cost per student (slope) is the same for different types of schools.
- Hence the cost functions are parallel.
- Is this assumption realistic?
- Occupational schools incur higher costs that are related to the number of students.

Slope Dummy Variables

- Let's relax this assumption by introducing a slope dummy variable $NTYPE$.
- $NTYPE$: defined as the $N \times TYPE$

Combined regression model	$COST_i = \beta_1 + \delta TYPE_i + \beta_2 N_i + \lambda NTYPE_i + u_i$
Regular school, $TYPE = 0$**,** $NTYPE = 0$	$COST_i = \beta_1 + \beta_2 N_i + u_i$
Occupational school, $TYPE = 1$**,** $NTYPE = N$	$COST_i = (\beta_1 + \delta) + (\beta_2 + \lambda) N_i + u_i$

- The model now allows:
- marginal cost per student to be an amount λ greater than that in regular schools
- the overhead costs to be different



Slope Dummy Variables

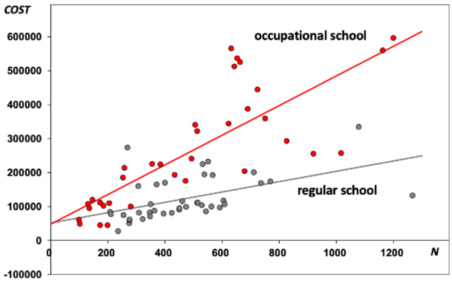
School	Type	*COST*	*N*	*TYPE*	*NTYPE*
1	Occupational	345,000	623	1	623
2	Occupational	537,000	653	1	653
3	Regular	170,000	400	0	0
4	Occupational	526.000	663	1	663
5	Regular	100,000	563	0	0
6	Regular	28,000	236	0	0
7	Regular	160,000	307	0	0
8	Occupational	45,000	173	1	173
9	Occupational	120,000	146	1	146
10	Occupational	61,000	99	1	99

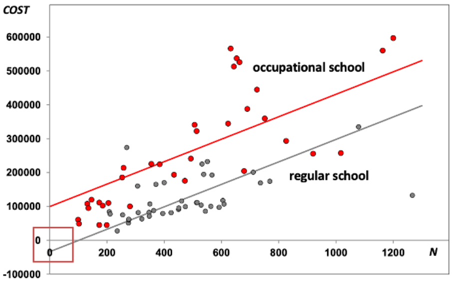
```
Call:
lm(formula = COST ~ N + TYPE + NTYPE, data = schools)

Residuals:
    Min      1Q  Median      3Q     Max
-234588  -34362  -16561   35663  242119

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  51475.25   31314.84   1.644 0.104703
N              152.30      60.02   2.537 0.013395 *
TYPE         -3501.18   41085.46  -0.085 0.932332
NTYPE          284.48      75.63   3.761 0.000348 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 81980 on 70 degrees of freedom
Multiple R-squared:  0.6803,  Adjusted R-squared:  0.6666
F-statistic: 49.64 on 3 and 70 DF,  p-value: < 2.2e-16
```

Combined regression model	$\widehat{COST}_i = 51475.25 - 3501.18 TYPE_i + 152.30 N_i + 284.48 NTYPE_i$
Regular school, $TYPE = 0$**,** $NTYPE = 0$	$\widehat{COST}_i = 51475.25 + 152.30 N_i$
Occupational school, $TYPE = 1$**,** $NTYPE = N$	$\widehat{COST}_i = (51475.25 - 3501.18) + (152.30 + 284.48) N_i$
	$\widehat{COST}_i = 47974.07 + 436.78 N_i$

Slope Dummy Variables

- Marginal costs (slope) for regular schools: 152 ¥
- Marginal costs (slope) for occupational schools: 436 ¥
- Cost functions fit the data much better than before and that the real difference is in the marginal cost
- The assumption of the same marginal cost led to an estimate of the marginal cost that was a compromise between the marginal costs of occupational and regular schools.
- The cost function for regular schools was too steep and as a consequence the intercept was underestimated, actually becoming negative and indicating that something must be wrong with the specification of the model.





Testing Joint Explanatory Power of Dummy Variables

- We can also perform an F test of the joint explanatory power of the dummy variables, comparing RSS when the dummy variables are included with RSS when they are not.
- Original model specification (Model 1): $COST_i = \beta_1 + \beta_2 N_i + u_i \quad RSS_1$
  Modified model specification (Model 2): $COST_i = \beta_1 + \delta TYPE_i + \beta_2 N_i + \lambda NTYPE_i + u_i \quad RSS_2$

Step 1. State the null and alternative hypotheses

Null Hypothesis	$H_0 : \delta = \lambda = 0$
Alternative Hypothesis	$H_1 : \delta \neq 0 \; or \; \lambda \neq 0 \; or \; both \; \delta \; and \; \lambda \neq 0$

Step 2. Select the significance level. Significance level $\alpha = 5\%$
Step 3. Select and calculate the test statistics

$$F(cost\ in\ dof, dof\ remaining) = \frac{reduction\ in\ RSS / cost\ in\ dof}{RSS\ remaining / dof\ remaining} = \frac{(RSS_1 - RSS_2)/cost\ in\ dof}{RSS_2 / dof\ remaining} \tag{1}$$

Testing Joint Explanatory Power of Dummy Variables

Step 3. Select and calculate the test statistics

$$F(\textit{cost in dof}, \textit{dof remaining}) = \frac{\textit{reduction in RSS}/\textit{cost in dof}}{\textit{RSS remaining}/\textit{dof remaining}} = \frac{(RSS_1 - RSS_2)/\textit{cost in dof}}{RSS_2/\textit{dof remaining}} \qquad (2)$$

```
> nobs(costfit4)
[1] 74
```

```
Analysis of Variance Table # with dummies

Response: COST
          Df       Sum Sq      Mean Sq F value    Pr(>F)
N          1 579744371821 579744371821  86.263 7.863e-14 ***
TYPE       1 326072074645 326072074645  48.518 1.444e-09 ***
NTYPE      1  95082497498  95082497498  14.148 0.0003475 ***
Residuals 70 470448182865   6720688327
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Analysis of Variance Table # without dummies

Response: COST
          Df       Sum Sq      Mean Sq F value    Pr(>F)
N          1 579744371821 579744371821  46.816 2.157e-09 ***
Residuals 72 891602755008  12383371597
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$RSS_1 = 891602755008$; $RSS_2 = 470448182865$; cost in dof $= 2$ dof remaining $n - k = 74 - 4 = 70$

$$F(\textit{cost in dof}, \textit{dof remaining}) = \frac{(RSS_1 - RSS_2)/\textit{cost in dof}}{RSS_2/\textit{dof remaining}} = \frac{(891602755008 - 470448182865)/2}{470448182865/70} = 31.4$$

Testing Joint Explanatory Power of Dummy Variables

Step 4. Set the decision rule.

$cost\ in\ dof = number\ of\ new\ variables\ added = 2$

$k = 4, n = 74, dof\ remaining = n - k = 74 - 4 = 70$

$F_{crit,5\%}(cost\ in\ dof, dof\ remaining) = F_{crit,5\%}(2, 70) = 3.12$

Step 5. Make statistical decisions.

$F = 31.4 > F_{crit,5\%}(2, 70) = 3.12$

We can reject the null $H_0 : \delta = \lambda = 0$.

We conclude that adding *TYPE* and *NTYPE* improves the overall fit of the model-at least one of the coefficients is statistically significant.

Student Task

- We regress hourly wages in USD on a gender dummy (1...female, 0...male), an education variable (measured in years), and their product

$$wage_i = \beta_1 + \delta female_i + \beta_2 educ_i + \lambda female_i \times educ_i + u_i$$

```
Call:
lm(formula = wage ~ educ + female + femaleeduc, data = wage1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.20050    0.84356   0.238    0.812
educ         0.53948    0.06422   8.400 4.24e-16 ***
female      -1.19852    1.32504  -0.905    0.366
femaleeduc  -0.08600    0.10364  -0.830    0.407
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.186 on 522 degrees of freedom
Multiple R-squared:  0.2598,  Adjusted R-squared:  0.2555
F-statistic: 61.07 on 3 and 522 DF,  p-value: < 2.2e-16
```

- Using the R output, write down the wage functions for men and women separately and interpret the coefficients.
- Provide a regression line diagram to illustrate the differences between the wages of men and women based on education

What to do next:

- Attempt homework 7
- Read chapter 5 of Dougherty