Categorical Predictor Variables

- Not all potential predictors in a regression model need to be values measured on a continuous numeric scale. In fact, in addition to numeric predictors we have looked at variables that could better be described as ordinal (ordered numeric categories) than continuous.

  Variables such as geographic region or strain of mouse, are examples of purely qualitative categorical variables.

- Non-continuous scale variables can be divided (for the most part) into
  - ordinal variables: ordering between categories, e.g., youth, young adult, middle-aged and elderly, level of education
  - nominal variables: no ordering whatsoever, e.g., male/female; Hawaii/non-Hawaii, race/ethnicity groups.

Categorical Predictor Variables

SPRM 3.16

- **Such variables are easily accommodated in linear regression, In fact, the** *general linear model* **subsumes all types of predictor variables.** The familiar ANOVA method is a model to compare means between several groups. **Membership to these groups can be considered categorical predictors**

- When we approach categorical variables from our linear regression model framework, such variables can be represented by one or more **indicator variables**. Such variables 'indicate" the presence or absence of a given feature, and are encoded using the two values, usually 1 (feature present) and 0 (feature absent), respectively. Equivalently, this indicates membership to discrete groups (male vs. female)

Categorical Predictor Variables

- **Why is this necessary?**
  - If we have just two categories, say male and female, then labeling (0,1) or (1,2) works (although we have to pay careful attention to the coding in the latter case).
  - If there are three categories or more, we cannot just plug values (1,2,3,4) for single/married/widowed/divorced into the model, as it will be assumed that these are ordered numeric categories. What is the correct ordering? Is it meaningful?
- Before we go further, let's look at how to use SLR to analyze categorical data. This will motivate the necessity for indicator variables.

SLR for Categorical Predictor Variables

- We want to 'predict' some quantity Y with sex (male, female), As we will see, this is tantamount to comparing the means between the groups.

- We could therefore write the modell as

$$\mathrm{E}(Y \mid \text{sex}) = \beta_0 + \beta_1 \times \text{sex}$$

where

$$\text{sex} = \begin{cases} 0 \text{ if female} \\ 1 \text{ if male} \end{cases}$$

- Recall that the meaning of β_0 is is $\mathrm{E}(Y \mid \text{X=0})$. So, β_0 is the mean of Y for females. What is β_1?

SLR for Categorical Predictor Variables

- For men, note that $\mathrm{E}(Y \mid \text{sex}) = \beta_0 + \beta_1$. Thus, β_1 is the **shift or change** in the mean due to sex=1 (male) vs. sex=0 (female), or the change in Y for a one unit change in X, as always in SLR. $\beta_0 + \beta_1$ is the mean of Y in males.

• What then does the test of $\beta_1 = 0$ represent? There really is no 'slope' to consider.

- The test of $\beta_1 = 0$ is equivalent to testing 'no shift in the mean due to membership in group(male) vs. group(female)'. Other ways to write are

$$H_0 : \mu_{female} = \mu_{male}$$

or

$$H_0 : \mu_{female} - \mu_{male} = 0$$

In other words, a hypothesis typically evaluated with a two-sample t-test

SLR for Categorical Predictor Variables

- This two-sample t-test has DF $= n_1 - 1 + n_2 - 1 = n - 2$, same as in testing β_1 in SLR

- **Why not model as**

$$\mathrm{E}(Y \mid \textsf{sex}) = \beta_0 + \beta_1 \times \textsf{sexfemale} + \beta_2 \times \textsf{sexmale}$$

  with two indicators sexfemale and sexmale (again, taking on value 0,1), so that we have βs for each sex effect separately?

- **Two problems with this approach:**
  - What is the meaning of β_0 in this model? In regression, $\beta_0 = E(Y|X = 0)$. Not meaningful (assuming 2 groups)
  - These variables are linear functions of each other, a violation of regression assumptions: sexfemale $=$ (1 - sexmale). They are completely collinear, meaning redundant in information (for matrix afficianatos, $X^T X$ has no unique inverse - can't get β vector)

Back to Categorical Predictor Variables in General

- Any variable with k possible values (feature types, groups, etc) requires a set of $k-1$ indicator variables, which answer the sequence of questions: "Is feature 1 present? Is feature 2 present? ... Is feature k-1 present?" Note that if all answers are "no" then it must be true that feature k is present; likewise, if one of the answers is "yes", then k-th answer is "no". There can be at most one "yes" answer among the k-1 questions.

- For example, consider the variable "smoking status": We can categorize it as *never smoker, occasional smoker, or heavy smoker* , so $k = 3$.
  If you want to be more specific, can also categorize it as *never, ex-, light, moderate, or heavy smoker*, 5 categories.

  Note:

  The different categories must be defined as mutually exclusive.

Categorical Predictor Variables in General

For the smoking status variable with 3 categories, we can code this variable using $3 - 1 = 2$ dummy (indicator) variables.

$$\text{occasSmoker} = \begin{cases} 1 \text{ if occasionally smoked, 0 otherwise} \end{cases}$$

$$\text{heavySmoker} = \begin{cases} 1 \text{ if smoked above some threshold frequency, 0 otherwise} \end{cases}$$

The omitted category, we might call neverSmoker (corresponding to 0's for the other two indicators above) is called the **base, control, or reference category**.

Note: we might also reasonably consider this variable on an ordered numeric scale, meaning ordered categories over limited range

Categorical Predictor Variables in General

Another simple example, combining categorical and continuous predictors: predicting salary with qualification and gender:

If the variable *gender* is coded as 1 (males) and 0 (females) then we have:

$$\text{Wage}_i = \beta_0 + \beta_1 \text{Qualification}_i + \beta_2 \text{Gender}_i + \epsilon_i$$

This really means:

$$\text{Wage}_i = (\beta_0 + \beta_2) + \beta_1 \text{Qualification}_i + \epsilon_i, \text{ if } i\text{-th person is male}$$

$$\text{Wage}_i = \beta_0 + \beta_1 \text{Qualification}_i + \epsilon_i, \text{ if } i\text{-th person is female.}$$

Using Categorical Predictor Variables

Example: salary survey data (from Chatterjee and Hadi textbook)

The data for this example are from a survey of computer professionals' salaries in a large corporation.

Response variable:

S – *salary*: **thousands of dollars**

Predictor variables:

X – *experience*: Experience in years

E – *education*: Highest level of education completed (coded as: $1 =$ HS, $2 =$ Bachelor's , $3 =$ higher (advanced degree))

M – *management*: Management responsibility ($1=$yes, $0 =$ no)

Using Categorical Predictor Variables

How to represent education in the model?:

- Education is a categorical variable with three categories. Currently, this variable is coded as ordinal, meaning ordered categories
  – Should we treat it as ordinal? While the categories have quantitative meaning and likely effect on salary simple ordinal coding imposes a linearity constraint which may or may not make sense
  – We can ask, Does it make sense to talk about an "average increase corresponding to 1 unit change in education"?
  – Does it make sense to expect the same average increase going from HS to B, and from B to advanced degree?
- Initially, we choose to treat it as a nominal (categorical) variable for now and see if we can justify a linear effect later.

Using Categorical Predictor Variables

To treat education as a nominal variable, we need to generate indicators for its categories. Stata will create a set of indicator variables for us using the *tabulate* command. This creates three indicators here which correspond to the three variables: E1, E2, E3 in C&H.

```
. tab e, generate (E)

          e |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |         14       30.43       30.43
          2 |         19       41.30       71.74
          3 |         13       28.26      100.00
------------+-----------------------------------
      Total |         46      100.00


.* Check frequency of indicator
. tab E1
```

```
        e== |
     1.0000 |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         32       69.57       69.57
          1 |         14       30.43      100.00
------------+-----------------------------------
      Total |         46      100.00
```

```
. list in 1/4
     +-----------------------------------+
     |     s   x   e   m   E1   E2   E3 |
     |-----------------------------------|
  1. | 13876   1   1   1    1    0    0 |
  2. | 11608   1   3   0    0    0    1 |
  3. | 18701   1   3   1    0    0    1 |
  4. | 11283   1   2   0    0    1    0 |
     +-----------------------------------+
```

NOTE: We only need any two of these to define categorical eduction as a predictor in the model- as mentioned earlier

Using Categorical Predictor Variables

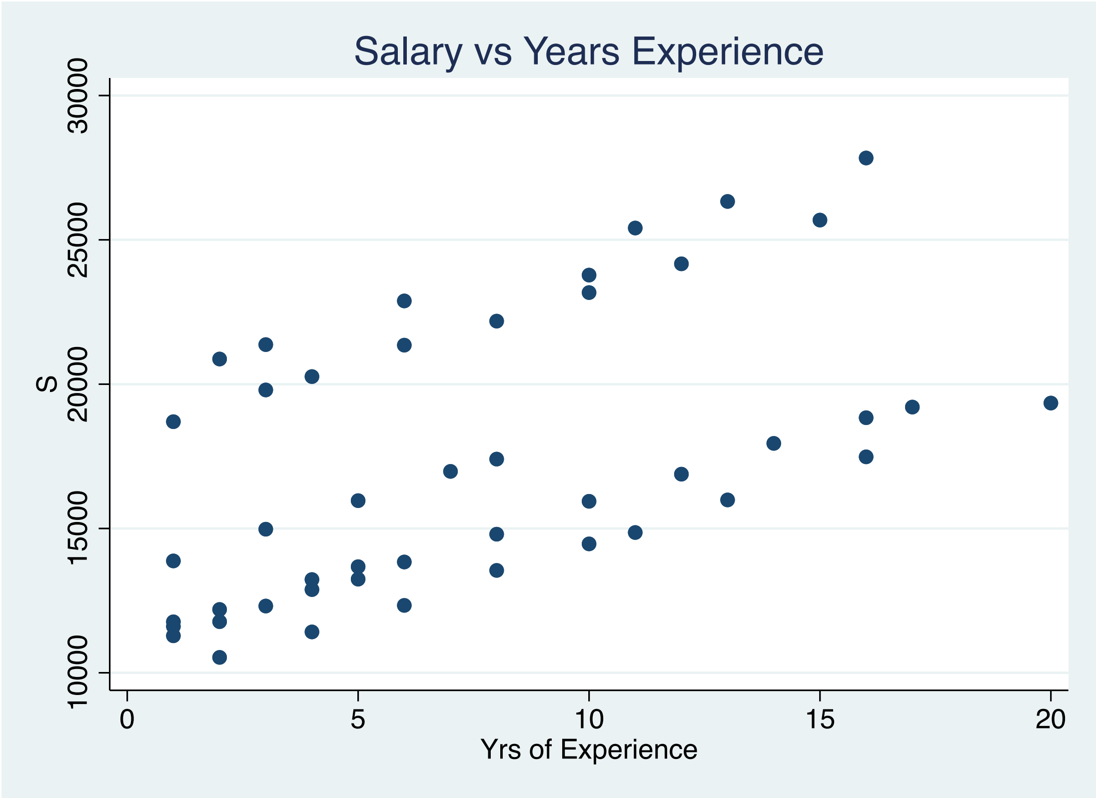
The model we would like to fit is then:

$$\text{Salary} = \beta_0 + \beta_1 X + \beta_2 E2 + \beta_3 E3 + \beta_4 M + \epsilon$$

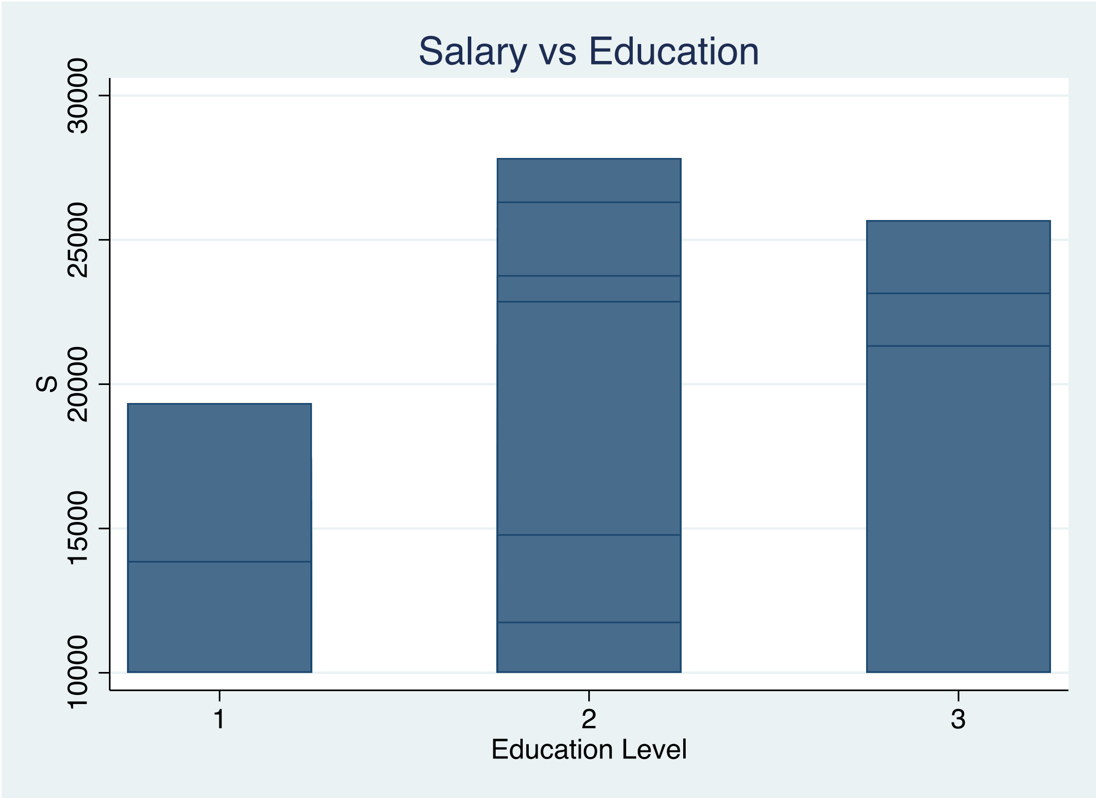
Note that here we have chosen to use a different reference group for education than the one in C&H. This changes the βs and their interpretation only.

Choice of reference/baseline group is flexible, usually chosen based on questions of interest or context. Usually lowest or highest group

Examining the Relationships



Examining the Relationships



Examining the Relationships

Using Categorical Predictor Variables

The model $\text{Salary} = \beta_0 + \beta_1 X + \beta_2 E2 + \beta_3 E3 + \beta_4 M + \epsilon$ predicts the following:

– For people with HS only (note no education term - it is the reference category)

$$\text{Salary} = \beta_0 + \beta_1 X + \beta_4 M + \epsilon$$

– For people with a Bachelor's level degree (E2 $=$ 1)

$$\text{Salary} = \beta_0 + \beta_1 X + \beta_2 + \beta_4 M + \epsilon$$

– For people with higher degree (E3 $=$ 1)

$$\text{Salary} = \beta_0 + \beta_1 X + \beta_3 + \beta_4 M + \epsilon$$

Note that β_2 measures the increment in mean salary for those with a college degree, relative to those with a high-school diploma only.

Using Categorical Predictors

$$\text{Salary} = \beta_0 + \beta_1 X + \beta_2 E2 + \beta_3 E3 + \beta_4 M + \epsilon$$

What does the sum $\beta_0 + \beta_2$ represent in the model?

A: This is the mean salary for individuals with college level education, X=0 (zero years experience) and no management responsibility (M=0)

What does the parameter β_3 represent in the model?

A: This term represents the *change* or *increment* in mean salary when going from high-school (reference level) to advanced college education, at any given level of the other variables.

What does β_0 measure?

A: This is the mean salary for no experience (X=0), high school education level (E2=0, E3=0), and no management responsibility (M=0)

The model is fit in Stata as follows:

```
. regress  s x E2 E3 m

      Source |       SS       df       MS              Number of obs =      46
-------------+------------------------------           F(  4,    41) =  226.84
       Model |   957816858     4   239454214           Prob > F      =  0.0000
    Residual |  43280719.5    41   1055627.3           R-squared     =  0.9568
-------------+------------------------------           Adj R-squared =  0.9525
       Total |  1.0011e+09    45  22246612.8           Root MSE      =  1027.4

------------------------------------------------------------------------------
           s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |    546.184   30.51919    17.90   0.000     484.5493    607.8188
          E2 |   3144.035   361.9683     8.69   0.000     2413.025    3875.045
          E3 |    2996.21   411.7527     7.28   0.000     2164.659    3827.762
           m |   6883.531    313.919    21.93   0.000     6249.559    7517.503
       _cons |   8035.598   386.6893    20.78   0.000     7254.663    8816.532
------------------------------------------------------------------------------
```

Using Categorical Predictors

Q: Tests are given for βs associated with E2 and E3. What is being tested here?

A: Whether mean salary shifts significantly when going from HS to B (E2) or HS to advanced degree (E3)

- From here, we can test several hypotheses. For example, test whether education level is needed in the model at all

```
. test E2 E3

 ( 1)  E2 = 0
 ( 2)  E3 = 0

       F(  2,    41) =   43.35
            Prob > F =    0.0000
```

- We can also formally test if a higher degree worth more in this context than a BS

```
.  test  E2 = E3
 ( 1)  E2 - E3 = 0.0

       F(  1,    41) =    0.15
            Prob > F =    0.7049
```

Note that we could use the model

$$\text{Salary} = \beta_0 + \beta_1 X + \beta_2 E1 + \beta_3 E2 + \beta_4 M + \epsilon$$

Here we have changed the reference category by omitting the E3 indicator

```
. regress  s x E1 E2 m

      Source |       SS       df       MS              Number of obs =      46
-------------+------------------------------           F(  4,    41) =  226.84
       Model |   957816858     4   239454214           Prob > F      =  0.0000
    Residual |  43280719.5    41   1055627.3           R-squared     =  0.9568
-------------+------------------------------           Adj R-squared =  0.9525
       Total |  1.0011e+09    45  22246612.8           Root MSE      =  1027.4
```

s	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x	546.184	30.51919	17.90	0.000	484.5493	607.8188
E1	-2996.21	411.7527	-7.28	0.000	-3827.762	-2164.659
E2	147.8249	387.6593	0.38	0.705	-635.0689	930.7188
m	6883.531	313.919	21.93	0.000	6249.559	7517.503
_cons	11031.81	383.2171	28.79	0.000	10257.89	11805.73

Some Observations

- The p-value of the F-test for $H_0 : \beta_{E2} = \beta_{E3}$ is identical to the p-value for the coefficient of E2 in the above regression. Why?
- ANOVA table output is identical - **we have not changed the overall model** - just the baseline level for education is changed
- Note that the intercept $cons$ term *is* different: This is now the mean salary for zero experience, no management experience, and advanced degree education (now baseline level)
- What if we tried to use all three indicator variables in the model?

Using Categorical Predictors: what about this model?

```
. regress  s x E1 E2 E3 m
note: E3 omitted because of collinearity

      Source |       SS       df       MS              Number of obs =      46
-------------+------------------------------           F(  4,    41) =  226.84
       Model |   957816858     4   239454214           Prob > F      =  0.0000
    Residual |  43280719.5    41   1055627.3           R-squared     =  0.9568
-------------+------------------------------           Adj R-squared =  0.9525
       Total |  1.0011e+09    45  22246612.8           Root MSE      =  1027.4

------------------------------------------------------------------------------
           s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |    546.184   30.51919    17.90   0.000     484.5493    607.8188
          E1 |   -2996.21   411.7527    -7.28   0.000    -3827.762   -2164.659
          E2 |   147.8249   387.6593     0.38   0.705    -635.0689    930.7188
          E3 |          0  (omitted)
           m |   6883.531    313.919    21.93   0.000     6249.559    7517.503
       _cons |   11031.81   383.2171    28.79   0.000     10257.89    11805.73
------------------------------------------------------------------------------
```

- **Oops.** Model cannot be executed as specified. Instead, the reference category is set as E3 - Stata chooses this to fix problem

Ordinal Predictor Variables

What if we had used education as the variable "e" (coded as 1,2,3) instead of subsets of dummy vars E1, E2, E3?

Then the regression would look like:

$$\text{Salary} = \beta_0 + \beta_1 X + \beta_2 e + \beta_3 M + \epsilon$$

and specifies that:

– For people with HS only (e=1)

$$\text{Salary} = \beta_0 + \beta_1 X + \beta_2 + \beta_3 M + \epsilon$$

– For people with a B-level degree (e=2)

$$\text{Salary} = \beta_0 + \beta_1 X + \beta_2 \times 2 + \beta_3 M + \epsilon$$

– For people with a higher degree (e=3)

$$\text{Salary} = \beta_0 + \beta_1 X + \beta_2 \times 3 + \beta_3 M + \epsilon$$

Ordinal Predictor Variables

```
. regress  s x e m

      Source |       SS       df       MS              Number of obs =      46
-------------+------------------------------           F(  3,    42) =  179.63
       Model |   928714168     3   309571389           Prob > F      =  0.0000
    Residual |  72383409.5    42  1723414.51           R-squared     =  0.9277
-------------+------------------------------           Adj R-squared =  0.9225
       Total |  1.0011e+09    45  22246612.8           Root MSE      =  1312.8

------------------------------------------------------------------------------
           s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   570.0874   38.55905    14.78   0.000     492.2721    647.9027
           e |    1578.75   262.3216     6.02   0.000     1049.364    2108.137
           m |    6688.13   398.2756    16.79   0.000     5884.377    7491.883
       _cons |   6963.478   665.6947    10.46   0.000     5620.051    8306.904
------------------------------------------------------------------------------
```

The estimated coefficient for education is $\hat{\beta}_2 = 1578.75$ (this is the increment in dollars), i.e., we would expect the difference between HS vs BS holders to be the same as the difference between BS vs higher-degree holders. Does that seem reasonable?

Ordinal Predictor Variables

How would we check if this model makes sense? How should we decide if we want to treat a categorical variable as ordinal or nominal?

- **First**, it has to make sense to be treated as ordered categories. We would expect some sort of monotone increasing effect here, but must consider carefully whether this is correct. We could look for monotone change in means by (education) levels
- **Then**, . . .
  For k categories, we could plot k separate regression lines (using the other predictors), for each level, and check if the fitted lines are roughly evenly spaced.
  Or we could also use the STATA command `test 2E2 = E3` in a regression treating E as nominal first and see if linear increments in effect are appropriate

Ordinal Predictor Variables

- IF an ordinal representation is adequate to capture the effect, then there are some advantages with respect to model properties. What are the degrees of freedom for the overall F-tests (and the DF (i.e., denominator) for the error term MSE) if we treat education as ordinal or nominal?

  Nominal Parameters are two education levels, experience, management responsibility = 4. Then DF = 46 - 4 - 1(intercept) = 41.

  Ordinal: Parameters are **one** education levels, experience, management responsibility = 3. Then DF = 46 - 3 - 1(intercept) = 42.

- This 'savings' in DF <u>if the model form fits</u> the data is more efficient (higher statistical power per n observations)

Expanding the Flexibility of the Model

- Thus far we have considered models where, say, the effect of years of experience on salary is the same regardless of which education level one is in. This may or may not be reasonable.
  - What if those with lower education are subject to 'capped' salary according to their job description/qualifications? This is a realistic scenario, and could alter the salary trajectory over years of experience.
- This possibility can be investigated and added to the model if warranted to improve the fit and/or explain the observations on salary

Interactions (AKA Effect Modification)

- All our models thus far specify that any given level/value of one predictor, the relative increment in (mean of) Y due to another predictor is the same. There are many situations that the effects of a predictor on the response are NOT the same in, say, different categories of some other predictor.

  For example, how much do oral contraceptives (OC) increase the risk of blood clots?

  - For women who don't smoke, the increase in risk is modest.
  - For women who smoke, the increase in risk is very large.

  This is also called *effect modification* in epidemiology

Interactions (Effect Modification)

(SPRM 3.20, covered extensively in other texts)

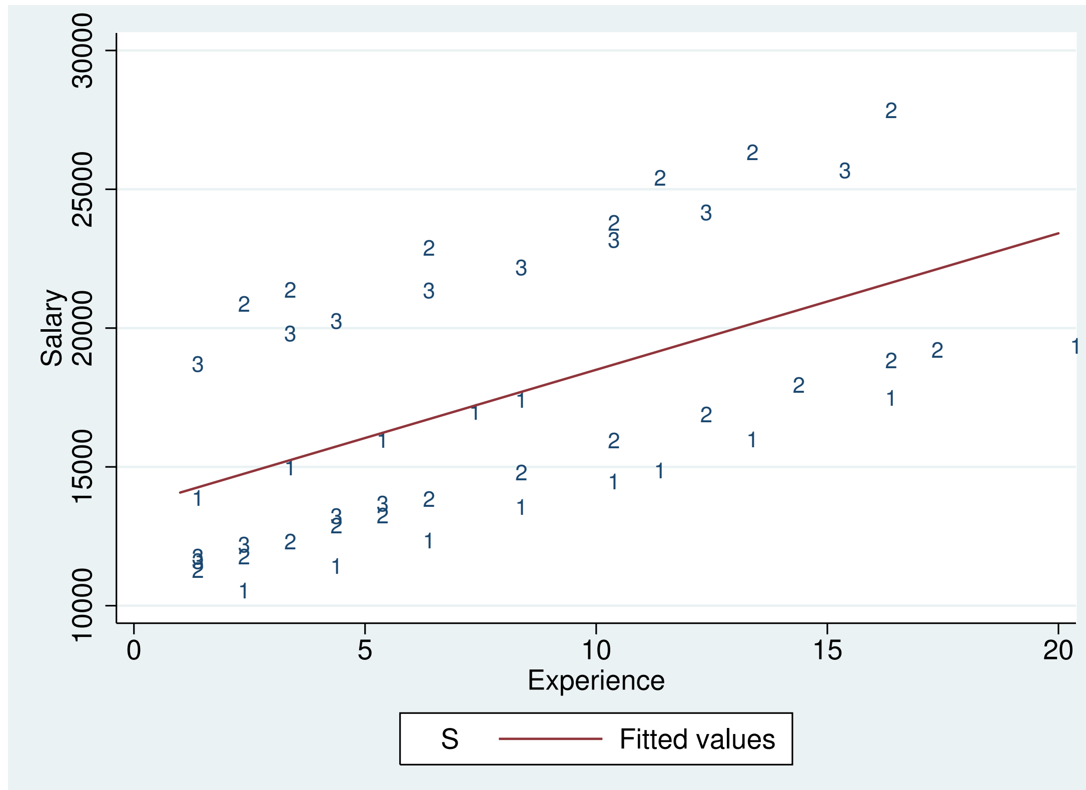
- That is, the relationship of OC and blood clot risks is different for smokers and non-smokers. If we plotted separate regression lines for clot risk versus time-on-OC we would see two lines with different slopes for smokers and non-smokers.

- We say that OC and smoking **interact** in their effect on blood-clotting risk.

- **Back to the salary example**, does the effect of experience on salary same for people with different levels of education?

  We first fit Salary on Experience (only). Model is $\text{Salary} = \beta_0 + \beta_1 X + \epsilon$

```
. twoway (scatter s x, mlabel(e) msymbol(none))
  (lfit s x), xtitle("Experience") ytitle("Salary")
```

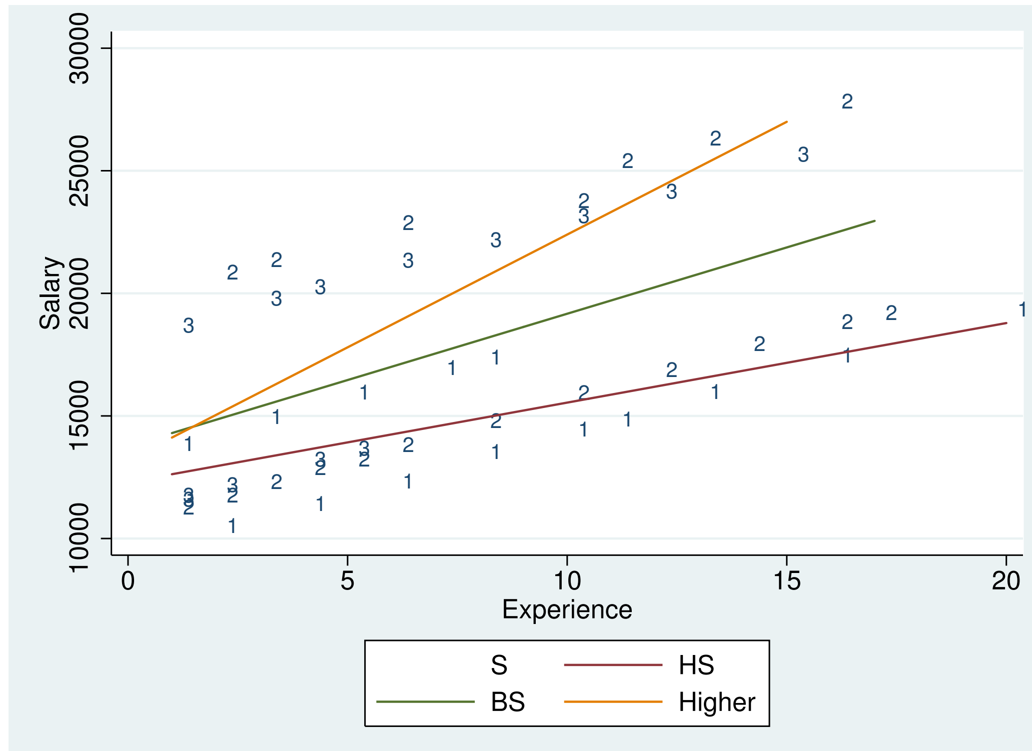


- Recall this model ignores (can be thought of averaging over) education level. The line 'fits' the data slope well, but there is a clear education effect (horizontal shift). Also, it's not clear that lines through the three education levels would be parallel, which is required for truly same effect at each education level.

Interactions - empirical assessment

To explore whether there might be different effects (slopes) for experience according to education level, we *stratify* the regression analysis for different levels of education: This means we fit separate models within each education category, still using only experience as the predictor.

```
. twoway (scatter s x, mlabel(e) msymbol(none))
 (lfit s x if (E1==1))  (lfit s x if (E2==1))  (lfit s x if (E3==1)),
  xtitle("Experience") ytitle("Salary") legend(order(1 "S" 2 "HS" 3 "BS" 4 "Higher"))
```



- The lines appear non-parallel, suggesting a differential experience effect by education level
- How non-parallel in order to declare slopes different? To some extent this requires considerations beyond statistical, but we can formally test for the necessity of **interaction effects** in the model

Linear Model with Interactions

- To accommodate the interaction between experience (X) and education on salary, the appropriate regression model will look like:

  $$\text{Salary} = \beta_0+\beta_1 X+\beta_2 E2+\beta_3 E3+\beta_4(E2\times X)+\beta_5(E3\times X)+\beta_6 M+\epsilon$$

  The **product of experience and education variables** are now predictors in the model, meaning that:

  – For people with HS only

  $$\text{Salary} = \beta_0 + \beta_1 X + \beta_6 M + \epsilon$$

  – For people with a B-level degree

  $$\text{Salary} = \beta_0 + \beta_2 + (\beta_1 + \beta_4)X + \beta_6 M + \epsilon$$

  – For people with higher degree

  $$\text{Salary} = \beta_0 + \beta_3 + (\beta_1 + \beta_5)X + \beta_6 M + \epsilon$$

Linear Model with Interactions

What is the effect of experience on salary? With one more year of experience, how much do we expect the salary to change? **Now it depends on the level of education**

– For HS, it is β_1 (slope for HS only)

– For B-level, it is $\beta_1 + \beta_4$ (a different slope for college degree)

– For Higher degree, it is $\beta_1 + \beta_5$ (another different slope for advanced college degree)

- **Fitting this model:** We need to create the additional model predictors needed. We can use partial F-tests to evaluate the need for these extra predictors being in the model (i.e., test whether slopes are different)

Fitting Interaction Models

```
. gen xE2 = x*E2
. gen xE3 = x*E3
. regress s x m E2 E3 xE2 xE3

      Source |       SS       df       MS              Number of obs =      46
-------------+------------------------------           F(  6,    39) =  158.61
       Model |   961686897     6   160281150           Prob > F      =  0.0000
    Residual |  39410679.8    39  1010530.25           R-squared     =  0.9606
-------------+------------------------------           Adj R-squared =  0.9546
       Total |  1.0011e+09    45  22246612.8           Root MSE      =  1005.3


           s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   632.2878   53.18525    11.89   0.000     524.7105    739.8651
           m |   7102.454   333.4416    21.30   0.000     6428.005    7776.903
          E2 |   4172.504   674.9662     6.18   0.000     2807.256    5537.753
          E3 |   3946.365   686.6934     5.75   0.000     2557.396    5335.333
         xE2 |  -125.5147   69.86292    -1.80   0.080    -266.8258    15.79635
         xE3 |  -141.2741   89.28056    -1.58   0.122    -321.8611    39.31286
       _cons |    7256.28   549.4936    13.21   0.000     6144.824    8367.736
------------------------------------------------------------------------------
```

Fitting Interaction Models

- Just examining the coefficients table, the interaction terms xE2 and xE3 do not quite satisfy conventional statistical significance, meaning that the βs are not different from zero. **A more appropriate test, evaluating the set of interaction effects as a group, is:**

```
. test xE2 xE3

 ( 1)  xE2 = 0
 ( 2)  xE3 = 0
       F(  2,    39) =    1.91
            Prob > F =    0.1610
```

- **This result suggest that there is weak evidence at best of different slopes per education level (because βs for these terms not different from zero)** We may choose to omit the interaction terms, opting for the simpler **main effects model**. For some purposes, we might choose to include (R^2 larger by 4%)

Interactions Among Continuous/Ordinal Variables

If we had used education as the variable "e" (coded as 1,2,3) instead of dummies E1, E2, and E3, then the regression would look like:

$$\text{Salary} = \beta_0 + \beta_1 X + \beta_2 e + \beta_3 (e \times X) + \beta_4 M + \epsilon$$

- **interpretation:** Effect of experience on salary is $\beta_1 + \beta_3 e$, i.e, the effect of experience changes linearly with increasing level of education.

  For each additional year of experience increase, the salary increase will depend on education, specifically:

  - For HS, it is $\beta_1 + \beta_3$
  - For B-level, it is $\beta_1 + 2\beta_3$
  - For higher degree, it is $\beta_1 + 3\beta_3$

  Two continuous covariates that interact are interpreted similarly.

Linear Model with Interactions - Comments

- When interaction effects exist, we call the effects of predictor as the **main effects**. In general, **interactions alone should not be kept in a model without corresponding main effects in the model**. There are exceptions as we will illustrate
- In the presence of interaction effects, **we can no longer talk about effects of one predictor variable without considering the value taken on by another.**
- A critical issue with interaction effects is statistical power. These effects may be present, but as detecting them depends on adequate observations for **predictor variable value combinations**, there is often low statistical power.

Linear Model with Interaction Effects - Reviewing Model Parameters

- **For a model with continuous predictor X and categorical predictor C (that has $k = 2$ levels: 0 and 1), the full model is**

$$\mathrm{Y} = \beta_0 + \beta_1 X + \beta_2 C + \beta_3 (C \times X) + \epsilon$$

The meaning of the parameters (βs) in the model:

- β_0 - value of Y when $X = 0, C = 0$
- β_1 - effect on Y of incrementing X one unit, <u>when $C = 0$</u>
- β_2 - effect on Y of $C = 1$, <u>when $X = 0$</u>
- β_3 - additional effect on Y of one unit increment in X <u>when $C = 1$</u> (also, additional effect on Y of $C = 1$ <u>when $X = x$, a specific value)</u>

Linear Model with Interaction Effects - Reviewing Model Parameters

$$\mathrm{Y} = \beta_0 + \beta_1 X + \beta_2 C + \beta_3 (C \times X) + \epsilon$$

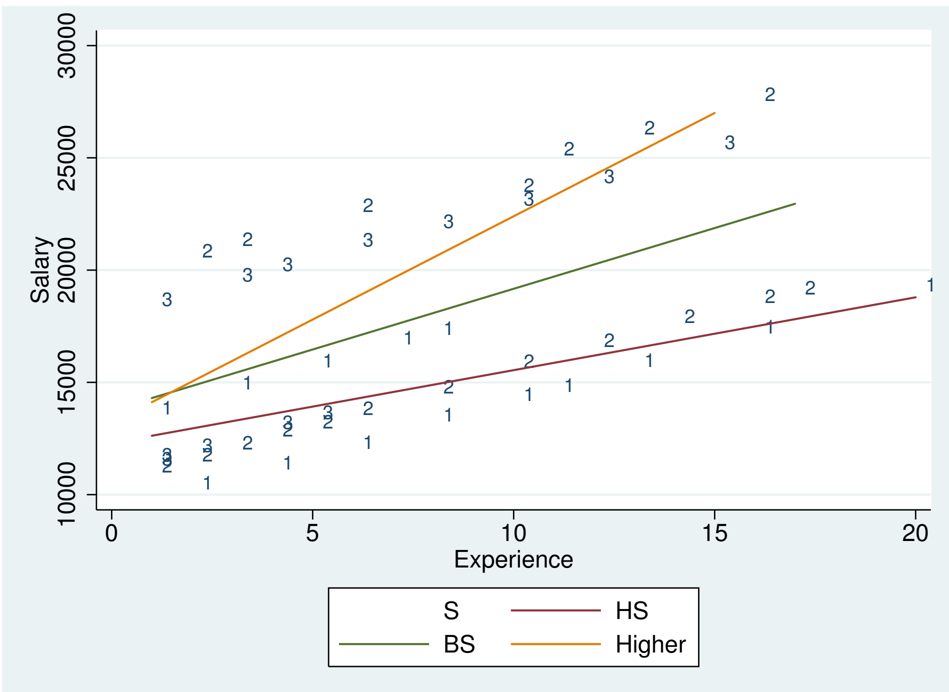
The result on outcome Y produced by the (full) model:

- when $x = 0, C = 0$: $\quad Y = \beta_0 + \epsilon$
- when $x = 1, C = 0$: $\quad Y = \beta_0 + \beta_1 + \epsilon$
- when $x = 0, C = 1$: $\quad Y = \beta_0 + \beta_2 + \epsilon$
- when $x = 1, C = 1$: $\quad Y = \beta_0 + \beta_1 + \beta_2 + \beta_3 + \epsilon$

Interactions and Stratification

Question: Will slopes from interaction model look like the empirical 'stratified' estimates (separate regression by Educ level)?

Original Plot -



Estimates from interaction model

s	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x	324.5148	179.6417	1.81	0.078	-38.55451	687.5841
E2	1461.181	2326.397	0.63	0.534	-3240.642	6163.005
E3	898.2396	2357.149	0.38	0.705	-3865.737	5662.216
xE2	216.3477	238.6374	0.91	0.370	-265.9565	698.6518
xE3	595.5212	288.8711	2.06	0.046	11.6909	1179.352
_cons	12299.02	1740.366	7.07	0.000	8781.61	15816.43

This model allows a separate intercept and slope for the combined effects of years of experience and each education level

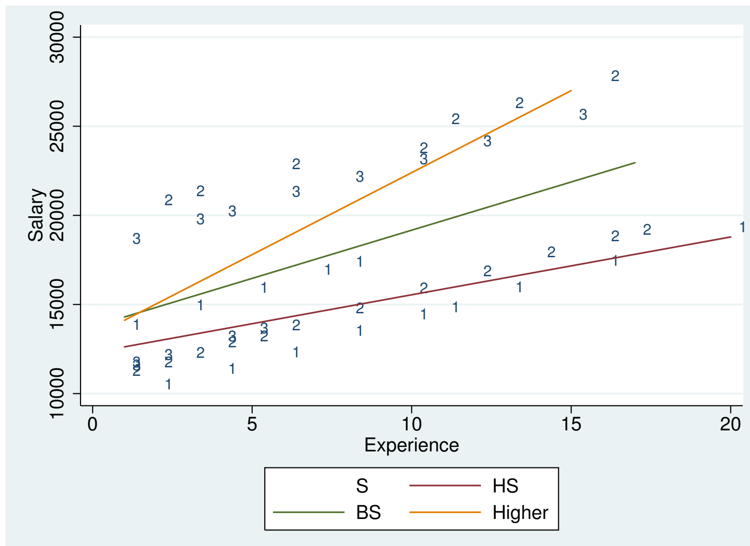
Interactions and Stratification

From the interaction model above:

- Intercept for HS, (no experience): $\$12,299$
- Intercept for College: $\$12299 + 1461 = \$13,760$
- Slope for HS $\$324$ per year
- slope for College: $\$324 + 216 = \$540$ per year
- slope for Advanced: $\$324 + 595 = \$919$ per year

  These are identical to what one would get with separate regression models in each education category

Interactions and Stratification



Example, for the HS group analyzed by itself:

```
. regress s x if e==1
------------------------------------------------------------------------------
           s | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
           x |   324.5148   92.42068     3.51   0.004     123.1474    525.8822
       _cons |   12299.02   895.3706    13.74   0.000     10348.18    14249.87
------------------------------------------------------------------------------
```

These are same estimates as the model with interaction terms

Linear Model with Interaction Effects - Summary

- We discussed models supported by the 'main effects' model, that is, without the interaction effect. For this model, three **new** (non-null) models can result :

  1. Different intercepts, different slopes

  $$\mathrm{Y} = \beta_0 + \beta_1 X + \beta_2 C + \beta_3 (C \times X) + \epsilon$$

  2. Different intercepts, same slope (this model really isn't new)

  $$\mathrm{Y} = \beta_0 + \beta_1 X + \beta_2 C + \epsilon$$

  3. Same intercept, different slopes

  $$\mathrm{Y} = \beta_0 + \beta_1 X + \beta_3 (C \times X) + \epsilon$$