7.4 Interactions with dummy variables

1. Recall the marriage and gender model which captures two levels of wage discrimination. We recall:

$$\begin{aligned}\widehat{log(wage)} =& 0.321^{**} + 0.213^{**} marrmale \\ &- 0.198^{**} marrfem - 0.110^{**} singfem \\ &+ 0.79^{**} educ + 0.027^{**} exper - 0.00054^{**} exper^2\end{aligned}$$

2. We can write this as interaction between marriage and female dummy variables. Which one is better and why?

$$\begin{aligned}\widehat{log(wage)} =& 0.321^{**} - 0.110^{**} female + 0.231^{**} married \\ &- 0.301^{**} female \cdot married\end{aligned}$$

3. Interpretation: setting $female = 0$ and $married = 0$ corresponds to the base group. The above example tells us that a married female is likely to earn 30% less than unmarried male controlling for education, experience.

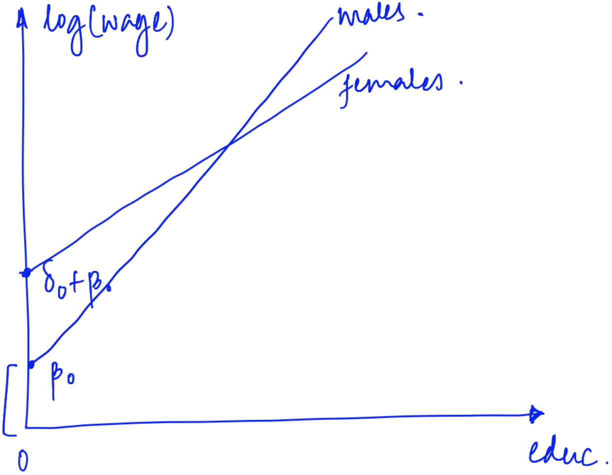
7.4.1 Allowing for Different Slopes

1. What happens when dummy variables interact with regular variables?

$$log(wage) = (\beta_0 + \delta_0 female) + (\beta_1 + \delta_1 female)educ + u$$

2. Interpretation: female = 0 gives us β_0 as intercept for males and β_1 as the partial effect of education for males.
   female = 1 gives us $\beta_0 + \delta_0$ as the intercept for females and $\beta_1 + \delta_1$ as slope for females.

3. Consider the case when average wage in the sample for females is higher than that of males, but the marginal effect of education for males is still higher than females. This would look as follows:



$\delta_0 > 0 \quad \delta_1 < 0.$

4. **Testing difference in wages**: If we claim that wages are identical for men and women with identical education, then we are saying that δ_0 and δ_1 are both zero i.e. joint test of hypothesis.

$$
\begin{aligned}
\widehat{\log(wage)} = {} & \underset{(.119)}{.389} - \underset{(.168)}{.227}\, female + \underset{(.008)}{.082}\, educ \\
& - \underset{(.0131)}{.0056}\, female \cdot educ + \underset{(.005)}{.029}\, exper - \underset{(.00011)}{.00058}\, exper^2 \\
& + \underset{(.007)}{.032}\, tenure - \underset{(.00024)}{.00059}\, tenure^2 \\
n = 526,\ & R^2 = .441.
\end{aligned}
\tag{7.18}
$$

5. Interpretation: No evidence against discrimination since t-stat on females is less than 2 on female. Are we sure about that?

6. Look at another specification we used earlier:

$$\widehat{\log(wage)} = .417 - .297\, female + .080\, educ + .029\, exper \qquad [7.9]$$
$$\phantom{\widehat{\log(wage)} =}\ (.099)\ \ (.036) \qquad\quad (.007) \qquad\quad (.005)$$
$$- .00058\, exper^2 + .032\, tenure - .00059\, tenure^2$$
$$(.00010) \qquad\quad (.007) \qquad\quad (.00023)$$
$$n = 526,\ R^2 = .441.$$

Because we have added the interaction $female \cdot educ$ to the equation, the coefficient on female is now estimated much less precisely than it was in equation (7.9)

7. Why?: cause female and education are correlated i.e. the likelihood of attaining education if you are a female is likely to be smaller than if you are male. This is true across countries.

7.4.2 Testing for Differences in Regression Functions across Groups

1. Example: We want to understand whether there are systematic differences between *cumgpa* (grade point averages) for college men and women athletes. We suspect that there are no gender differences so we use the following model:

$$cumpga = \beta_0 + \beta_1 sat + \beta_2 hsperc + \beta_3 tothrs + u$$

where *sat* are SAT scores, *hsperc* is high school rank percentile and *tothrs* is total hours of college courses.

2. **Different perspectives**: after a talk with your colleague from another college town you think that there is a possibility for gender to interact with all independent variables such that the correct model might be:

$$
\begin{aligned}
cumpga = &\beta_0 + \delta_0 female + \beta_1 sat + \delta_1 female \cdot sat + \beta_2 hsperc \\
&+ \delta_2 female \cdot hsperc + \beta_3 tothrs + \delta_3 female \cdot tothrs + u
\end{aligned}
$$

You want to test your conjecture and your estimated model is:

$$
\begin{aligned}
\widehat{cumgpa} = \underset{(0.21)}{1.48} &- \underset{(.411)}{.353}\, female + \underset{(.0002)}{.0011}\, sat + \underset{(.00039)}{.00075}\, female \cdot sat \\
&- \underset{(.0014)}{.0085}\, hsperc - \underset{(.00316)}{.00055}\, female \cdot hsperc + \underset{(.0009)}{.0023}\, tothrs \\
&- \underset{(.00163)}{.00012}\, female \cdot tothrs
\end{aligned} \tag{7.22}
$$

$$n = 366,\ R^2 = .406,\ \bar{R}^2 = .394.$$

3. **Model comparison**: our null hypothesis on account of our colleague from another college town is

$$H_0 : \delta_0 = 0, \delta_1 = 0, \delta_2 = 0, \delta_3 = 0$$

You run your original model and obtain an R^2. You compute F-stat from the two restricted and un-restricted model as follows

$$F \equiv \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} \geq 0$$

where SSR_r is the sum of squared residuals from the restricted model and SSR_{ur} is the sum of squared residuals from the unrestricted model.

- ▷ We find that and find that the F-stat is 8.14 or the p-value is 0.00000.
- ▷ It is pretty clear from the definition of F that we will reject H_0 in favor of H_a when F is sufficiently "large". How large depends on our chosen significance level (Section 4.5, Wooldridge)

4. What if we have a larger model with k independent variables. We could follow the same process or

   ▷ We could run the restricted model for each group.

   ▷ Let $n_1 = 90$ are the number of females in our bigger sample and $n_2 = 76$ are the number of males in the bigger sample.

   ▷ To obtain $SSR_{ur} = SSR_1 + SSR_2$ or the unrestricted model is merely a pooling estimator (*later*) and the F-stat is

$$F = \frac{[SSR_{ur} - (SSR_1 + SSR_2)]}{[SSR_1 + SSR_2]} \cdot \frac{[n - 2(k+1)]}{k+1}$$

   This is called the **Chow test** and is a form of F-test valid under homoskedasticity assumption. Note you have to use SSR and not R^2

5. *Overcoming limitations of Chow Test*: Say we want to test the intercept on interaction terms only and leave out the intercept.

   (a) Include dummy interactions like [7.22] and test joint significance of interaction terms only.

   (b) Estimate an SSR_r on a model with just the intercept term and no interactions i.e. for the college GPA model as follows:

   $$cumpga = \beta_0 + \delta_0 female + \beta_1 sat + \beta_2 hsperc + \beta_3 tothrs + u$$

   (c) and obtain an F for k restrictions i.e. $H_0 : \delta_1 = 0, \delta_2 = 0, \ldots, \delta_k = 0$

   $$F = \frac{[SSR_{ur} - (SSR_1 + SSR_2)]}{[SSR_1 + SSR_2]} \cdot \frac{[n - 2(k+1)]}{k}$$

   We obtain a p-value of 0.205 i.e. we fail to reject the null at 5% which tells us that this might be the best model and not the one with interactions.