7 Multiple Regression Analysis with Qualitative Information

1. Which variables appear to reveal qualitative information? What are their features?

2. How do we understand regression with such variables and how do we study interactions of quantitative with such variables? What is the partial effect?

3. How binary dependent variables can be treated separately using our information on Binomial distributions?

4. Introducing natural experiments

7.1 Describing Qualitative Information

1. *Binary or zero-one or dummy variables*: describe experiments (or events) which have only two possible outcomes. Which variable is assigned as zero is a choice.

2. *Why only zero-one?* provides ease of interpretation.

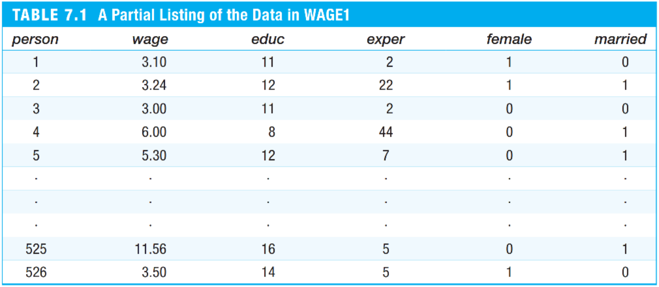


TABLE 7.1 A Partial Listing of the Data in WAGE1

person	*wage*	*educ*	*exper*	*female*	*married*
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
525	11.56	16	5	0	1
526	3.50	14	5	1	0

7.2 Single Dummy Variable

1. Example: Consider the effect of education on wages.

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + u$$

Now we want to understand whether there are systematic differences wages across gender i.e. we want to study discrimination. We modify the model as

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 female + u$$

2. **Base group**: is the group which is denoted with 0. Here female = 1 when the observation refers to females and 0 if males, hence the base group is that of male. Your comparison would be with a male.

3. **Dummy variable trap**: we could instead code the variable male, when male = 1 when the observation refers to male, and 0 if females.

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 female + \beta_4 male + u$$

But including both in one model would include perfect collinearity and hence wrong.

4. **Discrimination on females in the labor market**: If we suspect this to be the case, we use this model with the null hypothesis that $H_0 : \beta_3 = 0$ i.e. there is no discrimination with alternative as $H_a : \beta_3 < 0$. Let us assume we estimated this regression:

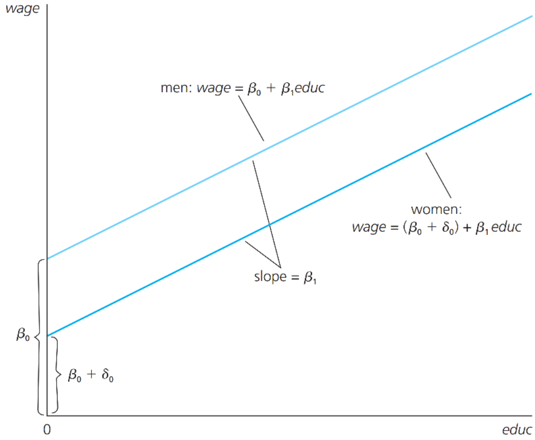
$$\hat{wage} = \hat{\beta}_0 + \hat{\beta}_1 educ + \hat{\beta}_2 exper + \hat{\beta}_3 female$$

5. **Interpretation**: magnitude of β_3 tells us the difference in hourly wage between females and males, given the same amount of education and experience. If the estimates are causal then:

$$\begin{aligned}\beta_3 =& E(wage|female = 1, educ, exper) \\ & - E(wage|female = 0, educ, exper)\end{aligned}$$

This is intercept shift between the fitted lines for male and female when the variables are coded as zero-one.

FIGURE 7.1 Graph of $wage = \beta_0 + \delta_0\, female + \beta_1\, educ$ for $\delta_0 < 0$.



Lecture 6: 17 Feb 2021

24 students present

(a) Practical computational exercises.

Lecture 7: 22 Feb 2021, Monday

24 students present

(a) Dummy variables (section 7.2)

6. Estimation results:

$$\begin{aligned}\hat{wage} =& -1.57(0.72) + 0.572(0.049)educ \\ &+ 0.25(0.012)exper - 1.81(0.26)female\end{aligned}$$

(a) Females are paid 1.81$ less than men for every hour worked and none of these differences are explained by education or experience difference.

(b) **Nested model**: tells us that the average wage for men is 7.10$ in the sample while women earn only 4.59 without controlling for any other factors. This merely describes the sample and we do not claim causality.

$$\hat{wage} = 7.10(0.21) - 2.51(0.30)female$$

(c) **Inference**: for the t-stat on female to be valid, we assume that population variance in male is identical to females i.e. homoskedasticity.

7. Computer ownership and GPA: could be evaluated by examining a dummy where owns = 1 implies that student has a personal computer and owns = 0 implies student does not. *Here owns is a choice and not an attribute* and dummy variables could be used in this too.

8. **Program evaluation and Binary variables**: Programs are special cases of policy where there are likely to be two groups (a) control group: those who are not directly affected by policy and (b) treatment group i.e. those who are directly affected by policy.

   (a) Comes from natural sciences and experiments on rats etc.

   (b) Vaccine efficacy studies are a type of program evaluation where you choose people of similar characteristic and give vaccine to half (called the treated) and give the other half a placebo (called the control)

7.2.1 When Dependent Variable is $Log(y)$

1. Recall: when dependent variable is $log(y)$, coefficient on the dummy variable multiplied by 100 is interpreted as percentage difference in y (for the category dummied)

2. If the estimated $\hat{\beta}$ is large, then the we need a more accurate correction for computing the β associated with y when estimating a $log(y)$ model. This is again as previously undertaken in section 6.2

$$100\Big[exp(\hat{\beta}_1) - 1\Big]$$

Lecture 8: 24 Feb 2021, Wednesday

15 students present

1. Marriage and gender premium (section 7.3)
2. Many categories and recoding our categories (7.3.1)
3. Slope variations and dummy variables with interactions (7.4)

7.3 Using Dummy Variable for Multiple Categories

1. Example on Marriage Premium: The question is whether there is discrimination of women and even more discrimination against married women. This is an example where there are 4 groups:

   (a) $Marrmale$, $marrfem$, $singfem$ with base group as single male

   (b) You could have taken any other base group too.

2. Estimates and interpretation:

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

   **: tells us that all variables are significant at 5% level. Married men earn about 21.3% more than single male (base group)

3. **Discrimination against married women**: is of the order of 19.8% in comparison to single males of the same education and experience.

4. Difference between groups with correct standard errors:

    - We can approximate the difference between single and married women as (-0.110 - (-0.198) = 0.088) i.e. 9% difference. But we dont know the correct SD on this for testing. We can re-estimate the model with married females as base to directly obtain SD.

7.3.1 Incorporating Ordinal Information by Using Dummy Variables

1. *Example: Scale of Physical Attractiveness*

   Consider five categories i.e. homely, quite plain, average, good looking, and strikingly beautiful (or handsome). **A survey was conducted asking independent experts in fashion about employees physical appearance** (withholding their name, designation etc.)

   Since there were very few extreme observations i.e. homely and strikingly handsome, **the categories were renamed** as : average, below and above.

2. Estimated equation is:

$$\text{Men}: \quad log(\hat{wage}) = \hat{\beta}_0 - 0.164^{**} belavg + 0.016 abvavg + ...$$

$$\text{Women}: \quad log(\hat{wage}) = \hat{\beta}_0 - 0.124^{*} belavg + 0.035 abvavg + ...$$

3. **Interpretation**: a man who appears not good looking obtains a 16% less wage and a similar wage for a below average women is 12%. This is discrimination on physical appearance against men.

   Hamermesh, D. S., & Biddle, J. E. (1993). Beauty and the labor market (No. w4518). National Bureau of Economic Research.

4. Too many ordinal variables: could be reduced into “arbitrary” categories and the used with similar interpretation, albeit with caution.

   ▷ Law school example 7.8

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.

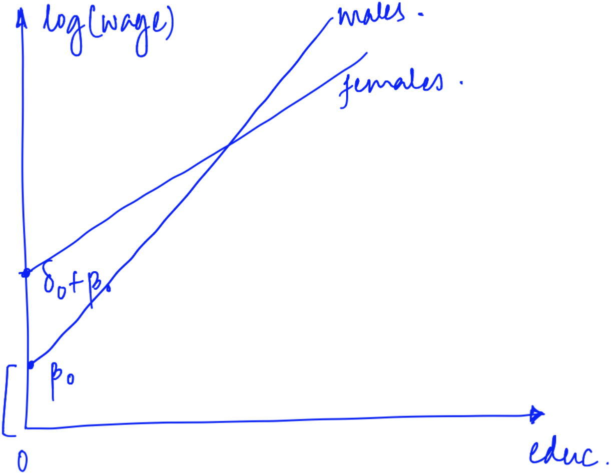
7.4.1 Allowing for Different Slopes

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

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

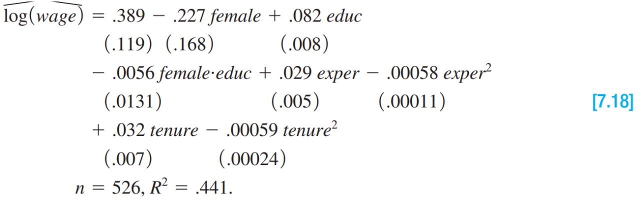
2. Interpretation: female $= 0$ gives us β_0 as intercept for males and β_1 as the partial effect of education for males.
   female $= 1$ gives us $\beta_0 + \delta_0$ as the intercept for females and $\beta_1 + \delta_1$ as slope for females.

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:



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

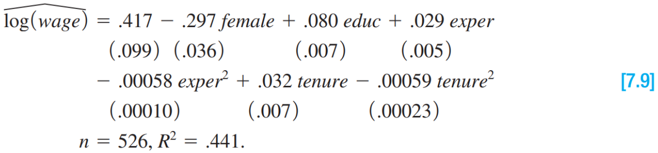
4. **Testing difference in wages**: If we claim that wages are identical for men and women with identical education, then we are saying that δ_0 and δ_1 are both zero i.e. joint test of hypothesis.



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

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:



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

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.

Lecture 9: 1 Mar 2021, Monday

Missed class to be scheduled later...

1. Testing for Differences in Regression Functions across Groups (7.4.1)
2. Linear probability model (7.5)
3. Introducing program evaluations (7.6)
4. Program evaluations and unrestricted regression adjustment (7.6.1) TBA
5. Interpreting Regression results with discrete dependent variables (7.7) TBA

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:

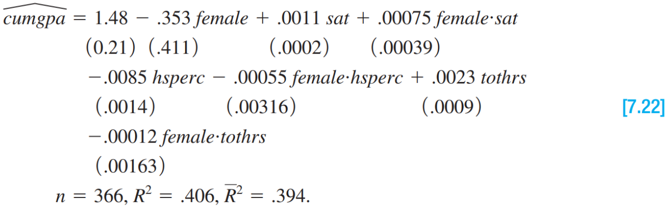
$$cumpga = \beta_0 + \beta_1 sat + \beta_2 hsperc + \beta_3 tothrs + u$$

   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.

7.5 The Linear Probability Model

1. **Motivation**: What happens if we want to use multiple regression to explain a qualitative event? Till now we said the following:

   (a) What if the intercepts differ across the two categories

   $$y = \beta_0 + \delta_0 D_1 + \beta_1 x_1 + u$$

   When $D_1 = 0$, intercept is β_0, and when $D_1 = 1$ intercept is $\beta_0 + \delta_0$.

   (b) What if the slopes differ across two categories

   $$y = \beta_0 + (\beta_1 + \delta_0 D_1) x_1 + u$$

   when $D_1 = 0$, the slope is β_1 and when $D_1 = 1$ slope is $\beta_1 + \delta_0$.

   (c) What happens if slope and intercepts differ across the two categories:

   $$y = (\beta_0 + \delta_0 D_1) + (\beta_1 + \delta_1 D_1) x_1 + u$$

2. **Binary dependent variable**: $y = 1$ when an individual voted for the NDA and $y = 0$ otherwise. Under such assumption if we have the following model under MLR assumptions

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

then we have the probability of $y = 1$, called the "Response probability" is defined as:

$$P(y = 1|\mathbf{x}) = E(y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

3. **Linear probability models**: These belong to a class of linear probability models where the response probability is linear in β_j. In order to get the probability of $y = 0$ we use probability theory as

$$P(y = 0|\mathbf{x}) = 1 - P(y = 1|\mathbf{x})$$

4. Estimated equation where $\hat{y}$ represents the probability of success i.e. $y = 1$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$$

5. *Explaining Labor Force Participation Rates*: Consider the following model:

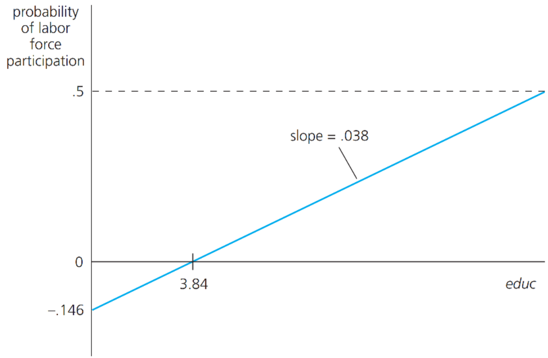
$$\begin{aligned}\widehat{inlf} =& 0.586^{**} - 0.0034^{**} nwifeinc + 0.038^{**} educ + 0.039^{**} exper \\ & - 0.00060^{**} exper^2 - 0.016^{**} age - 0.262^{**} kidslt6 + 0.013 kidsge6\end{aligned}$$

$inlf = 1$ if individual reports to be in labor force is success for this example. $nwifeinc$ is the income of partner, etc.

   (a) **Educ**: cetrius paribus, an additional year of education increases the probability of being in the labor force by 0.03 times, or a 10 year increase increases the probability of participating by 1/3rd.

(b) We could plot the probability of participation based on educ for arbitrary values of remaining variables as follows:

FIGURE 7.3 Estimated relationship between the probability of being in the labor force and years of education, with other explanatory variables fixed.



6. Limitations: (a) we cannot explain negative or greater than 1 probability (b) possible that probability is not linearly related and independent to other explanatory variables.

7. Correcting prediction errors in LPM models: $\hat{y}$ should have been $\in [0, 1]$. Re-define a new variable $\tilde{y} = 1$ if $\hat{y} \geq 0.5$ and 0 otherwise. This gives us a binary predicted variable which is a widely used goodness of fit measure.

8. Problem of heteroskedasticity: arises because the dependent variable is binary with a conditional variance as follows:

$$Var(p|\mathbf{x}) = p(\mathbf{x})(1 - p(\mathbf{x}))$$

   This would not be a problem if $p(\cdot)$ is not a function of $\mathbf{x}$. The problem is that standard errors are incorrect and inference is limited.

9. Example 7.12: Crime and arrests data.

7.6 Program Evaluation

1. Selection bias and Programs: consider the following example where those firms who received for job training grants by government are considered treated while those who did not are controls

   ▷ The argument is that firms apply on a first come, first serve basis. The estimated regression from this data (JTRAIN for 1988) is

   $$\widehat{log(scrap)} = 4.99 - 0.052 grant - 0.455 log(sales) + 0.639^{*} log(employ)$$

   ▷ *grant*: firms recieving grants have a scrap rate of 0.052 less than the ones not receiving grant, but small t prevents us from trusting the results completely.

- Selection bias: What if the firms who applied for the grants had more resources and were otherwise more proactive to seek alternative finances?

- We must be careful to include factors that might be systematically related to the binary independent variable of interest in order to provide correct inference.

- The bias induced by omiting out key control factors which explain systematic difference between the control and treated induce what we call a “Selection bias”.

7.6.1 Program Evaluation and Unrestricted Regression Adjustment

1. General model for program evaluation when w is the policy or program indicator, $x_1, x_2 \ldots, x_k$ are control variables,

$$E(y|w, \mathbf{x}) = \alpha + \tau w + \gamma_1 x_1 + \cdots + \gamma_k x_k$$

   where $y = (1-w)\hat{y}(0) + w\hat{y}(1)$ is the observed average y.

2. *Controls and self selection*: The problem of participation decisions differing systematically by individual characteristics is often referred to as the self-selection problem.
   Once we control for all the variables (**co-variates**) we could possibly control for, we could capture the effect of the program.

3. **Example**: We want to capture the effect of drug consumption during teens on labor force participation. However, those who consume drugs are likely to be from the poor family, subjected to lower levels of education and less skilled by the time they enter labor force. These other characteristics interact with their ability to join labor force.

4. Regression adjusted estimate: is the estimate of $\hat{\tau}$ for the program in the presence of co-variates. This implies conditional independence of the treatment.

   - ▷ Thus an individual who took drugs conditional on all other controls will have the following predicted value

   $$y(1) = \psi_1 + (\mathbf{x} - \boldsymbol{\eta})\boldsymbol{\gamma_1} + u(1)$$

   - ▷ If the same individual did not take drugs, then his predicted labor market outcome would be

   $$y(0) = \psi_0 + (\mathbf{x} - \boldsymbol{\eta})\boldsymbol{\gamma_0} + u(0)$$

   Read: section 2.7 again from JW Chapter 2 to distinguish between ATE and TE (a) under $\tau = y_i(1) - y_i(0)$ i.e. constant treatment effect and (b) different linear effect functions for treatment and control i.e. $\psi_0 \neq \psi_1$

5. Average Treatment effect: without restrictions (a) and (b) i.e. identical linear effect $\psi_0 = \psi_1$, we can write the ATE as:

$$E(te_i) = (\psi_1 - \psi_0) + E\{(\mathbf{x}_i - \boldsymbol{\eta})(\boldsymbol{\gamma_1} - \boldsymbol{\gamma_0}) + u_i(1) - u_i(0)\} = \tau_i$$

where $(\mathbf{x}_i - \boldsymbol{\eta})$ has a zero mean by construction and $u_i(1)$ and $u_i(0)$ has zero mean because they are the errors obtained from conditional expectations.

Unconfoundedness: implies conditional independence where the dummy variable for treatment does not depend on individual level co-variates i.e. $\mathbf{x}_i$

$$w \text{ is independent of } [y(0), y(1)] \text{ conditional on } \mathbf{x}$$

$$\begin{aligned} E(u_i|w_i, \mathbf{x}_i) =& E(u_i(0)|w_i, \mathbf{x}_i) + w_i E\{[u_i(1) - u_i(0)]|w_i, \mathbf{x}_i\} \\ =& E(u_i(0)|\mathbf{x}_i) + E\{[u_i(1) - u_i(0)]|\mathbf{x}_i\} \end{aligned}$$

6. Directly estimating average treatment effect: by regressing y_i on $w_i, x_{i1}, x_{i2}, \ldots$ and on demeaned controls $x_{ik} - \bar{x}_{ik} \cdot w_i$. The estimate of $\hat{\tau}$ on w_i is out ATE from a *restricted regression adjustment*.

   - It is critical to demean the x_j before constructing the interactions in order to obtain the average treatment effect as the coefficient on w_i.

   ▷ The two methods give us exact same estimated of $\hat{\tau}$

7. Consider the following example

Job Training program example: uses data JTRAIN98 and measures the effect of job training in 1997 on wages in 1998 for workers who got the training in comparison with those who didnt. We control for past earnings, education, age and maritial status.

- Simple difference of means estimate (3.74) of -2.05

$$\widehat{earn98} = 10.61 - 2.05 train$$

- Examine the regression with controls (3.75) called the restricted regression estimate of +2.41

$$\begin{aligned}\widehat{earn98} =& 4.67 + 2.41 train + 0.373 earn96 + 0.363 educ \\ & - 0.181 age + 2.48 married\end{aligned}$$

8. Now consider our estimate without any restriction

- Examine the regression with the un-restricted regression estimate of 3.11

$$
\begin{aligned}
\widehat{earn98} =& 5.08 + 3.11 train + 0.353 earn96 + 0.378 educ \\
& - 0.196 age + 2.76 married \\
& + 0.133 train \cdot (earn96 - \bar{earn}96) \\
& - 0.035 train \cdot (educ - \bar{educ}) \\
& + 0.58 train \cdot (age - \bar{age}) \\
& - 0.993 train \cdot (married - \bar{married})
\end{aligned}
$$

  These estimates prove to provide the most significant effect of the job training program. The t-stat on train is significant. Which model would you choose?

9. *Average Treatment Effect*: demean married and replace the final interaction term with $unmarried \cdot train$ to obtain an estimate of 3.79 as $\hat{\beta}$ of train. Coefficient on married is the same as 0.993. The ATE is 2.797.

10. Spurious effect of treatment is possible unless there are claims to causality.

7.7 Interpreting Regression Results with Discrete Dependent Variables

1. When the dependent variable follows a multi-nominal distribution, then the interpretation follows our usual multiple linear regression, i.e. for a given set of x_js we interpret the predicted value as $\hat{\beta}_0+\hat{\beta}_1 x_1+\ldots\hat{\beta}_k x_k$ as an estimate of the *conditional expectation function* i.e. $E(y|x_1, x_2, \ldots, x_k)$.

2. The estimates of $\hat{\beta}$ then tell us how the average of y changes over the sample when y takes many discreete values.