Economics 6400: Econometrics

Lecture 7: Multiple Regression Analysis with Qualitative Information – Binary (or Dummy) Variables

CSU, East Bay

November 14, 2017

In the past few weeks...

- Our left- and right-hand side variables have been **quantitative**
  - Hourly wage rate, years of education, college GPA, air pollution, firm sales, birth weight
  - Magnitude of the variable conveys interpretable information
- Often we want to include **qualitative** factors
  - Gender or race, industry of a firm (financial versus consumer product), state in the U.S. (CA, IL, etc.)
  - We will use binary (0/1) variables to incorporate qualitative factors

Describing qualitative information

- Qualitative factors are usually described by a binary (yes/no or 0/1) relationship
  - A person is male or female; a person does or does not smoke; a state administers capital punishment or not
- Relevant information can be coded as a 1 (if true) or 0 if otherwise
  - E.g. the variable $female = 1$ if the person is female and $female = 0$ if the person is not female (i.e. male)
- It would not be incorrect to define a dummy variable with values other than 0 and 1 but it would make the interpretation more difficult with no apparent benefit

female and *married* are dummy variables

```
. list wage educ exper female married
```

	wage	educ	exper	female	married
1.	3.1	11	2	1	0
2.	3.2	12	22	1	1
3.	3	11	2	0	0
4.	6	8	44	0	1
5.	5.3	12	7	0	1
6.	8.8	16	9	0	1
7.	11	18	15	0	0
8.	5	12	5	1	0
9.	3.6	12	26	1	0
10.	18	17	22	0	1

Single dummy independent variable

- Consider the hourly wage equation:

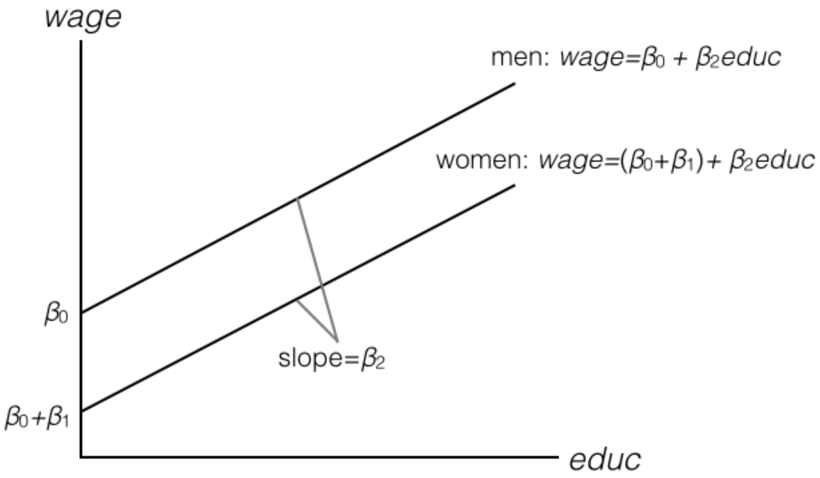
$$wage = \beta_0 + \beta_1 female + \beta_2 educ + u$$

- β_1 is the difference in hourly wage between females and males, *given* the same amount of education and same error term u
- If $\beta_1 < 0$ then there is evidence for discrimination
- Given the zero conditional mean assumption and same level of education:

$$\beta_1 = E(wage|female = 1, educ) - E(wage|female = 0, educ)$$

- Situation can be illustrated graphically by an intercept shift equal to β_1

$wage = \beta_0 + \beta_1 female + \beta_2 educ + u$ assuming $\beta_1 < 0$



Why no dummy variable for male?

- The intercept for males is β_0, and the intercept for females is $\beta_0 + \beta_1$
- Since there are only two groups, we need only need two different intercepts
- Using two dummy variables would introduce perfect collinearity as $male = 1 - female$
  - When using dummy variables, one category has to be omitted
- If we had chosen females to be the base group or benchmark group instead then the model would be:

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

Testing for discrimination

- Consider the hourly wage equation:

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

- Estimating this equation using data from 1976:

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	-1.810852	.2648252	-6.84	0.000	-2.331109	-1.290596
educ	.5715048	.0493373	11.58	0.000	.4745802	.6684293
exper	.0253959	.0115694	2.20	0.029	.0026674	.0481243
tenure	.1410051	.0211617	6.66	0.000	.0994323	.1825778
_cons	-1.567939	.7245511	-2.16	0.031	-2.991339	-.144538

- All else equal, women earn $1.81 less per hour than men
- Coefficient is very statistically significant ($|t_{female}| \approx 7$)

Effect of training grants on hours of training

- A special case of policy analysis is **program evaluation**, in which we seek to determine the effect of economic or social programs
- In the simplest case there are two groups:
  1. Control group: does not participate in the program
  2. Experimental (or treatment) group: does take part in the program
- Consider the effect of a training grant:

$$hrsemp = \beta_0 + \beta_1 grant + \beta_2 \log(sales) + \beta_3 \log(employ) + u$$

where *hrsmp* is hours of training per employee and $grant = 1$ if the firm received a grant

Effect of training grants on hours of training

- Estimating this equation using data from 1988:

hrsemp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
grant	26.2545	5.591765	4.70	0.000	15.16194	37.34705
lsales	-.9845809	3.539903	-0.28	0.781	-8.006797	6.037635
lemploy	-6.069871	3.882893	-1.56	0.121	-13.77249	1.632744
_cons	46.66508	43.4121	1.07	0.285	-39.45284	132.783

- The variable *grant* is very statistically significant ($t_{grant} \approx 5$)
- A firm that received a grant trained each worker 26.25 hours on average more than firms who did not receive a grant, controlling for sales and employment

Interpreting coefficients on dummy right-hand side variables when dependent variable is log(y)

- When the dependent variable appears in logarithmic form, such as in the house price equation:

$$\widehat{\log(price)} = -1.35 + 0.168\log(lotsize) + 0.707\log(sqrft) + 0.027bdrms + 0.054colonial$$

  then the coefficient on the dummy has a % interpretation
- In the above example, the dummy variable *colonial* (=1 if the house has a colonial style) implies that a house with a colonial style is predicted to sell for about 5.4% more, holding other factors fixed
- For larger coefficients, the exact percentage difference is: $100 \cdot [exp(\hat{\beta}_j) - 1]$

Using dummy variables for multiple categories

- Suppose we wanted to add a dummy variable for *married* to the log(*wage*) equation
  - The coefficient would indicate the percentage change in salary from being married all else equal, including gender
- If we wanted to allow the effect from being married to differ for females and males we could add three dummy variables:
  1. *marrmale*
  2. *marrfem*
  3. *singfem*
- The excluded/base group is single men
- General rule: if the regression model includes g groups, then we need to include $g - 1$ dummy variables in the model along with an intercept

Estimating the model

```
. gen marrmale=(1-female)*married

. gen marrfem=female*married

. gen singfem=female*(1-married)

. reg lwage marrmale marrfem singfem educ exper expersq tenure tenursq
```

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
marrmale	.2126757	.0553572	3.84	0.000	.103923	.3214284
marrfem	-.1982676	.0578355	-3.43	0.001	-.311889	-.0846462
singfem	-.1103502	.0557421	-1.98	0.048	-.219859	-.0008414
educ	.0789103	.0066945	11.79	0.000	.0657585	.092062
exper	.0268006	.0052428	5.11	0.000	.0165007	.0371005
expersq	-.0005352	.0001104	-4.85	0.000	-.0007522	-.0003183
tenure	.0290875	.006762	4.30	0.000	.0158031	.0423719
tenursq	-.0005331	.0002312	-2.31	0.022	-.0009874	-.0000789
_cons	.3213781	.100009	3.21	0.001	.1249041	.5178521

Using dummy variables for multiple categories

- All of the coefficients are statistically significant
- To interpret the coefficients on the dummy variable, note that the base group is single men
  - Married men, all else equal, earn 21.3% more than single men
  - Married women, all else equal, earn 19.8% less than single men
  - Single women, all else equal, earn 11.0% less than single men
- Single women earn $-11.0 - (-19.8) = 8.8\%$ more than married women
  - To check if this is a statistically significant difference, it is easiest to re-run the regression with married women as the excluded/base group and see if the coefficient on *singfem* is statistically significant

Incorporating ordinal information by using dummy variables

- We can also use dummy variables to deal with **ordinal variables** such as rankings or ratings, in which one unit increases are difficult to interpret quantitatively
- Example: Credit ratings for local government debt, which can be one of 5 ratings: $CR \in \{0, 1, 2, 3, 4\}$
  - Denote 0 credit rating as the base group and create four dummy variables, e.g. $CR_1 = 1$ if rating is 1, and 0 if otherwise
- Consider the following equation for the municipal bond rate:

$$MBR = \beta_0 + \delta_1 CR_1 + \delta_2 CR_2 + \delta_3 CR_3 + \delta_4 CR_4 + other\ \ factors$$

- Movement between different ratings is allowed to have a different effect
  - Difference between 3 and 2 credit rating is $\delta_3 - \delta_2$

Grouping ordinal values

- In some cases the ordinal values take on too many values (e.g. law school rankings)
- One option is to categorize the ordinal values (e.g. top 10 law schools, schools ranked 11-25, schools ranked 26-40 etc.)
- Example: Effect of physical attractiveness on wage
  - Each person is ranked for physical attractiveness (homely, quite plain, average, good looking, strikingly beautiful or handsome)
  - Group bottom and top two categories
  - Results from Hamermesh and Biddle (1994) for men:

$$\widehat{\log(wage)} = \hat{\beta}_0 - 0.164 belavg + 0.016 abvavg + other\ factors$$

  - Below average looking men earn 16.4% less than average looking men
  - Above average men earn 1.6% more than average looking men (but not statistically significant)

Interactions involving dummy variables

- We can recast the wage model using interactions between *female* and *married*:

$$\widehat{\log(wage)} = 0.321 - 0.110 female + 0.213 married - 0.301 female \cdot married + \dots$$

- Interpreting the coefficients:
  - $\log(wage)$ is 0.321+0.213=0.534 for married men
  - Estimated wage is 11% less for single women compared to single men
  - Estimated wage is -0.11+0.213-0.301=19.8% less for married women compared to single men

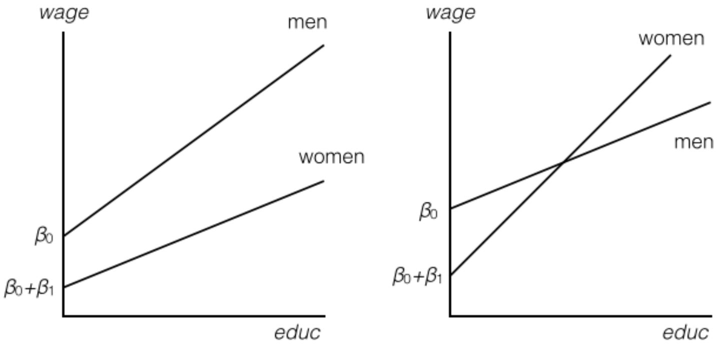
Allowing for different slopes: Interactions between dummy with non-dummy explanatory variables

- Suppose we seek to test whether the returns to education are different from women and men:
- We can accomplish this with an interaction between *female* and *education*:

$$\log(wage) = \beta_0 + \beta_1 female + \beta_2 educ + \beta_3 female \cdot educ + u$$

  - If *female* $= 0$ then intercept for males is β_0 and the slope on education is β_2
  - If *female* $= 1$ then intercept for females is $\beta_0 + \beta_1$ and the slope is $\beta_2 + \beta_3$

Different slopes:
Left panel: $\beta_1 < 0$, $\beta_3 < 0$, Right panel: $\beta_1 < 0$, $\beta_3 > 0$



- In the right panel, women earn less than men at low levels of education but the gap narrows as both men and women obtain more education

Testing for differential returns

- Two interesting hypotheses to test:
  1. H_0: $\beta_3 = 0$ (return to education same for both sexes)
  2. H_0: $\beta_1 = 0$ and $\beta_3 = 0$ (whole equation same for both sexes)

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	-.2267886	.1675394	-1.35	0.176	-.5559289	.1023517
educ	.0823692	.0084699	9.72	0.000	.0657296	.0990088
female_educ	-.0055645	.0130618	-0.43	0.670	-.0312252	.0200962
exper	.0293366	.0049842	5.89	0.000	.019545	.0391283
expersq	-.0005804	.0001075	-5.40	0.000	-.0007916	-.0003691
tenure	.0318967	.006864	4.65	0.000	.018412	.0453814
tenursq	-.00059	.0002352	-2.51	0.012	-.001052	-.000128
_cons	.388806	.1186871	3.28	0.001	.1556388	.6219732

- Coefficient on *female_educ* is not statistically significant
- Coefficient on *female* now insignificant due to multicollinearity

Testing for differences in regression functions across groups

- Often we want to test whether two populations follow the same regression function
- Consider the model of GPAs for college athletes:

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

  where *hsperc* is high school rank percentile, and *tothrs* is total hours on college courses
- We want to see if there are *any* differences between sexes
- One option is to create a series of interactions:

$$\begin{aligned} cumgpa = & \beta_0 + \beta_1 female + \beta_2 sat + \beta_3 female \cdot sat \\ & + \beta_4 hsperc + \beta_5 female \cdot hsperc + \beta_6 tothrs \\ & + \beta_7 female \cdot tothrs + u \end{aligned}$$

Unrestricted model results

Source	SS	df	MS
Model	53.5391808	7	7.6484544
Residual	78.3545052	358	.218867333
Total	131.893686	365	.361352564

Number of obs	=	366
F(7, 358)	=	34.95
Prob > F	=	0.0000
R-squared	=	0.4059
Adj R-squared	=	0.3943
Root MSE	=	.46783

cumgpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	-.3534862	.4105293	-0.86	0.390	-1.160838	.4538659
sat	.0010516	.0001811	5.81	0.000	.0006955	.0014078
female_sat	.0007506	.0003852	1.95	0.052	-6.88e-06	.0015081
hsperc	-.0084516	.0013704	-6.17	0.000	-.0111465	-.0057566
female_hsperc	-.0005498	.0031617	-0.17	0.862	-.0067676	.0056681
tothrs	.0023441	.0008624	2.72	0.007	.0006482	.0040401
female_tothrs	-.0001158	.0016277	-0.07	0.943	-.0033169	.0030852
_cons	1.480812	.2073336	7.14	0.000	1.073067	1.888557

Testing for differences in regression functions across groups

- Null hypothesis that *cumgpa* follows the same model for males and females:

$$H_0 : \beta_1 = 0, \beta_3 = 0, \beta_5 = 0, \beta_7 = 0$$

- t statistics for *female* and the interactions are not large
- But these variables are highly correlated so an F test is required
- $SSR_{UR} = 78.354$, $SSR_R = 85.515$ (next slide), $q = 4$
- F statistic is:

$$F = \frac{(SSR_R - SSR_{UR})/q}{SSR_{UR}/(n-k-1)} = \frac{(85.515 - 78.354)/4}{78.354/(366-7-1)} = 8.175$$

- Critical value at 5% level is 2.37 so we can reject the null hypothesis

Restricted model results

Source	SS	df	MS
Model	46.3786194	3	15.4595398
Residual	85.5150666	362	.236229466
Total	131.893686	365	.361352564

Number of obs	=	366
F(3, 362)	=	65.44
Prob > F	=	0.0000
R-squared	=	0.3516
Adj R-squared	=	0.3463
Root MSE	=	.48603

cumgpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
sat	.001185	.0001648	7.19	0.000	.0008609	.001509
hsperc	-.0099569	.0012446	-8.00	0.000	-.0124044	-.0075094
tothrs	.0023429	.0007554	3.10	0.002	.0008574	.0038285
_cons	1.49085	.1836782	8.12	0.000	1.12964	1.85206

When there are too many independent variables?

- F test can be adapted for cases with two groups but too many independent variables to construct interactions for
- Key insight: the SSR from the unrestricted model can be obtained from two separate regressions, one for each group giving SSR_1 and SSR_2
- $SSR_1 + SSR_2$ can be compared to SSR_P, which is obtained by running a regression on the pooled/combined sample
- Unrestricted model, with a group dummy variable and k interaction terms has $n - 2(k+1)$ degrees of freedom
- F or **Chow statistic** is:

$$F = \frac{[SSR_P - (SSR_1 + SSR_2)]/(k+1)}{(SSR_1 + SSR_2)/[n - 2(k+1)]}$$

Computing the Chow statistic

```
. reg cumgpa sat hsperc tothrs if spring==1 & female==0
```

Source	SS	df	MS
Model	27.2497343	3	9.08324475
Residual	58.7517192	272	.215998968
Total	86.0014535	275	.312732558

Number of obs	=	276
F(3, 272)	=	42.05
Prob > F	=	0.0000
R-squared	=	0.3169
Adj R-squared	=	0.3093
Root MSE	=	.46476

```
. reg cumgpa sat hsperc tothrs if spring==1 & female==1
```

Source	SS	df	MS
Model	13.1465734	3	4.38219113
Residual	19.602786	86	.227939372
Total	32.7493594	89	.36797033

Number of obs	=	90
F(3, 86)	=	19.23
Prob > F	=	0.0000
R-squared	=	0.4014
Adj R-squared	=	0.3805
Root MSE	=	.47743

$$F = \frac{[85.515 - (58.752 + 19.603)]/4}{(58.752 + 19.603)/358} = 8.175$$

Testing for <u>different</u> slopes but <u>same</u> intercept

- Chow test tests for no differences at all between groups
- A similar F statistic can be calculated for this test
  - Replace SSR_P with SSR from a regression with an intercept shift but no interaction terms
  - F statistic becomes:

$$F = \frac{[SSR_P - (SSR_1 + SSR_2)]/k}{(SSR_1 + SSR_2)/[n - 2(k+1)]}$$

  - For wage example $SSR_{UR} = 79.362$ so F statistic is:

$$F = \frac{[79.362 - (58.752 + 19.603)]/3}{(58.752 + 19.603)/358} = 1.533$$

  - p value ≈ 0.205 so cannot reject null that slopes are the same
- This result combined with the Chow test result suggest the best model allows for a different intercept but no interaction terms (i.e. no differential slopes)

Pooled model with intercept shift but no interaction terms

```
. reg cumgpa female sat hsperc tothrs if spring==1
```

Source	SS	df	MS
Model	52.5320205	4	13.1330051
Residual	79.3616656	361	.219838409
Total	131.893686	365	.361352564

Number of obs =	366
F(4, 361) =	59.74
Prob > F =	0.0000
R-squared =	0.3983
Adj R-squared =	0.3916
Root MSE =	.46887

cumgpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	.3100975	.0586128	5.29	0.000	.1948321	.4253629
sat	.0012144	.0001591	7.63	0.000	.0009016	.0015272
hsperc	-.0084413	.0012343	-6.84	0.000	-.0108687	-.0060139
tothrs	.0024638	.0007291	3.38	0.001	.00103	.0038976
_cons	1.328541	.1798275	7.39	0.000	.9748996	1.682182

A binary dependent variable: The Linear Probability Model

- We can also use binary/dummy variables as left-hand side variables
- Predicted or expected value is the probability of "success:"

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

- Multiple linear regression model with a binary dependent variable is called the **linear probability model (LPM)**
- β_j measures the change in the probability of success when x_j changes, holding other factors fixed:

$$\Delta P(y = 1|\mathbf{x}) = \beta_j \Delta x_j$$

Linear probability model of arrests

- Let $arr86$ be a binary variable equal to 1 if a man was arrested during 1986, and zero if otherwise
- Population is group of young men in CA born 1960 or 1961 who have at least one prior arrest
- LPM for $arr86$ is:

$$
\begin{aligned}
arr86 = & \beta_0 + \beta_1 pcnv + \beta_2 avgsen + \beta_3 tottime + \beta_4 ptime86 \\
& + \beta_5 qemp86 + u
\end{aligned}
$$

where

- $pcnv$ = proportion of prior arrests that led to a conviction
- $avgsen$ = average sentence served from prior convictions
- $tottime$ = months spent in prison since age 18 prior to 1986
- $ptime86$ = months spent in prison in 1986
- $qemp86$ = number of quarters (0 to 4) that the man was legally employed in 1986

Linear probability model of arrests

```
. reg arr86 pcnv avgsen tottime ptime qemp86
```

Source	SS	df	MS
Model	25.8452455	5	5.16904909
Residual	519.971268	2719	.191236215
Total	545.816514	2724	.20037317

Number of obs =	2725
F(5, 2719) =	27.03
Prob > F =	0.0000
R-squared =	0.0474
Adj R-squared =	0.0456
Root MSE =	.43731

arr86	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
pcnv	-.1624448	.0212368	-7.65	0.000	-.2040866	-.120803
avgsen	.0061127	.006452	0.95	0.344	-.0065385	.018764
tottime	-.0022616	.0049781	-0.45	0.650	-.0120229	.0074997
ptime86	-.0219664	.0046349	-4.74	0.000	-.0310547	-.0128781
qemp86	-.0428294	.0054046	-7.92	0.000	-.0534268	-.0322319
_cons	.4406154	.0172329	25.57	0.000	.4068246	.4744063

Linear probability model of arrests

- Interpreting coefficients:
  - *avgsen* and *tottime* are insignificant
  - Intercept of 0.441 implies that someone who has not been convicted, spent no time in prison, and was unemployed in 1986, has a 44% predicted probability of being arrested
  - Coefficient on *pcnv* implies that a 50% increase in proportion of convictions reduces probability by $0.5 \cdot 0.162 = 8.1\%$
  - Coefficient on *ptime*86 implies that six more months in prison reduces probability of arrest by $0.022 \cdot 6 = 13.2\%$
  - Coefficient on *qemp*86 implies a man employed all four quarters is $4 \cdot -0.043 = 17.2\%$ less likely to be arrested than a man who is not employed at all, all else equal

Advantages and disadvantages of LPM

- Disadvantages
  - Predicted probabilities can be larger than 1 and smaller than 0
  - Marginal probability effects sometimes logically impossible
  - LPM is necessarily heteroskedastic:

$$
\begin{aligned}
Var(y|\mathbf{x}) =& E(y^2|\mathbf{x}) - E(y|\mathbf{x})^2 \\
=& P(y=1|\mathbf{x}) \cdot 1^2 + (1 - P(y=1|\mathbf{x})) \cdot 0^2 \\
& - \Big(P(y=1|\mathbf{x}) \cdot 1 + (1 - P(y=1|\mathbf{x}) \cdot 0 \Big)^2 \\
=& P(y=1|\mathbf{x})[1 - P(y=1|\mathbf{x})]
\end{aligned}
$$

  - Need to estimate heteroskedasticity consistent standard errors
- Advantages of the LPM
  - Easy estimation and interpretation
  - Estimated effects and predictions often reasonably good in practice

Policy analysis and program evaluation

- Example: Effect of job training grants on worker productivity

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

where *scrap* is the firm's scrap rate (% of failed assemblies or material that cannot be repaired or restored), and *grant* is a dummy variable indicating whether the firm received a grant in 1988 for job training

- Firms receiving the grant have scrap rates 5.2% lower than firms without grants, all else equal

Self-selection into treatment as a source for endogeneity

- Treatment group: grant receivers, Control group: firms that received no grant
- Concern: grants were not assigned randomly but were given out on a first-come, first-served basis
  - Might be that firms with less productive workers saw an opportunity to improve productivity and applied first
  - Would imply a large effect from the grant since those firms stood the most to gain
  - Unobserved factors affecting productivity such as education, ability, experience, tenure etc. are correlated with the *grant*
  - In experiments, assignment to treatment is random so causal effects can be inferred using a simple regression:

$$y = \beta_0 + \beta_1 partic + u$$

where *partic* indicates participation (1) or not (0)

Further example of an endogenous dummy regressor

- Are nonwhite customers discriminated against?

$$\begin{aligned} \textit{approved} = & \beta_0 + \beta_1 \textit{nonwhite} + \beta_2 \textit{income} + \beta_3 \textit{wealth} \\ & + \beta_4 \textit{credrate} + u \end{aligned}$$

  where *approved* is a dummy indicating whether a mortgage application approved, and *nonwhite* is a dummy for minorities
- It is important to control for other characteristics that may be important for loan approval (e.g. profession, unemployment)
- Omitting important characteristics that are correlated with the *nonwhite* dummy will produce spurious evidence for discrimination

Next lecture

- Further Issues and Heteroskedasticity (Chapter 8)