Chapter 7
Regression with Qualitative Information

Jim Bang

Dummy (Binary) Independent Variables

Qualitative Information (Categorical/Factor Variables)

- Gender, Race, Industry, Occupation
- Create separate dummy variables for each category.
- Avoid the *dummy variable trap*!
- R automatically does this for you.

Single Dummy Independent Variable

Estimate the wage differential for women, controlling for education, experience, and experience squared. Call this `wage.lm9` and run a `summary()` of the results.

```
wage.lm9 <- lm(wage ~ educ + exper + I(exper^2) + female, data = wage1)
summary(wage.lm9)
```

```
##
## Call:
## lm(formula = wage ~ educ + exper + I(exper^2) + female, data = wage1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.8380 -1.8998 -0.3573  1.2691 14.2932
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.3192037  0.7388254  -3.139  0.00179 **
```

```
## educ         0.5562848  0.0502875  11.062  < 2e-16 ***
## exper        0.2551276  0.0348671   7.317 9.64e-13 ***
## I(exper^2)  -0.0044396  0.0007762  -5.720 1.80e-08 ***
## female      -2.1140347  0.2625501  -8.052 5.57e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.989 on 521 degrees of freedom
## Multiple R-squared:  0.3501, Adjusted R-squared:  0.3451
## F-statistic: 70.17 on 4 and 521 DF,  p-value: < 2.2e-16
```

Dummy Variables for Multiple Categories

1. Use the `factor()` function to create a factor, `occ`, as a variable in `wage1` from the occupational dummy variables in `wage1` (`profocc`, `clerocc`, and `servocc`) that takes integer values from one to four. Use the `labels` option to assign the labels `Manufacturing`, `Professional`,`Clerical`, and `Services` to the values.
2. Replicate `wage.lm7` using the new factor variable you created. Call this `wage.lm10` and summarize both regressions using a text stargazer() output.

```
wage.lm7 <- lm(wage ~ educ + exper + I(exper^2) + tenure + profocc + clerocc + servocc, data = wage1)
wage1$occ <- factor(1 + wage1$profocc + 2*wage1$clerocc + 3*wage1$servocc,
                    labels = c("Manufacturing", "Professional", "Clerical", "Services"))
wage.lm10 <- lm(wage ~ educ + exper + I(exper^2) + tenure + occ, data = wage1)
stargazer(wage.lm7, wage.lm10, type = "text")
```

```
##
## ===========================================================
##                                    Dependent variable:
##                                -----------------------------
##                                            wage
##                                     (1)            (2)
## -----------------------------------------------------------
## educ                              0.394***      0.394***
##                                   (0.057)        (0.057)
##
## exper                             0.181***      0.181***
##                                   (0.035)        (0.035)
##
## I(exper2)                        -0.004***      -0.004***
##                                   (0.001)        (0.001)
```

```
##
## tenure                           0.150***      0.150***
##                                  (0.020)        (0.020)
##
## profocc                          1.512***
##                                  (0.355)
##
## clerocc                          -0.675*
##                                  (0.388)
##
## servocc                          -0.947**
##                                  (0.408)
##
## occProfessional                                1.512***
##                                                 (0.355)
##
## occClerical                                     -0.675*
##                                                 (0.388)
##
## occServices                                    -0.947**
##                                                 (0.408)
##
## Constant                         -1.378*        -1.378*
##                                  (0.757)        (0.757)
##
## -----------------------------------------------------------
## Observations                       526            526
## R2                                0.397          0.397
## Adjusted R2                       0.389          0.389
## Residual Std. Error (df = 518)    2.887          2.887
## F Statistic (df = 7; 518)       48.709***      48.709***
## ===========================================================
## Note:                          *p<0.1; **p<0.05; ***p<0.01
```

Interactions

Interactions among Dummy Variables

Estimate the wage effect of marriage differ across gender by interacting `female` with `married`, controlling for education, experience, and job tenure. Call this `wage.lm11` and use a pipe to print a `summary()` of the results.

```
wage.lm11 <- lm(wage ~ educ + exper + tenure + female*married, data = wage1) |>
  summary()
```

Different Slopes

Estimate the effect of gender on wages *and the returns to education*, controlling for experience and job tenure. Call this `wage.lm12` and print a `summary()` of your result.

```
wage.lm12 <- lm(wage ~ educ + exper + tenure + female*educ, data = wage1)
summary(wage.lm12)
```

```
##
## Call:
## lm(formula = wage ~ educ + exper + tenure + female * educ, data = wage1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8607 -1.7730 -0.4345  1.0240 14.0358
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.27461    0.87227  -2.608  0.00938 **
## educ         0.62577    0.06186  10.116  < 2e-16 ***
## exper        0.02563    0.01156   2.217  0.02702 *
## tenure       0.14233    0.02116   6.727 4.58e-11 ***
## female      -0.06001    1.23480  -0.049  0.96125
## educ:female -0.13974    0.09626  -1.452  0.14721
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.954 on 520 degrees of freedom
## Multiple R-squared:  0.3661, Adjusted R-squared:   0.36
## F-statistic: 60.07 on 5 and 520 DF,  p-value: < 2.2e-16
```

Binary Dependent Variables

Probability of Arrest

Estimate the effect of prior convictions (`pcnv`) on *whether* a person was arrested in 1986, controlling for the average sentence of prior convictions (`avesen`), total time spent in prison prior to 1986 (`tottime`), total number of months spent in prison in 1986 (`ptime86`), and total number of quarters officially employed in 1986 (`qemp86`). Call this `crime.lm1` and print a `summary()` of your result.

```
crime.lm1 <- lm((narr86 > 0) ~ pcnv + avgsen + tottime + ptime86 + qemp86, data = crime1)
summary(crime.lm1)
```

```
##
## Call:
## lm(formula = (narr86 > 0) ~ pcnv + avgsen + tottime + ptime86 +
##     qemp86, data = crime1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5577 -0.2889 -0.2157  0.5734  0.8931
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.440615   0.017233  25.568  < 2e-16 ***
## pcnv        -0.162445   0.021237  -7.649 2.79e-14 ***
## avgsen       0.006113   0.006452   0.947    0.344
## tottime     -0.002262   0.004978  -0.454    0.650
## ptime86     -0.021966   0.004635  -4.739 2.25e-06 ***
## qemp86      -0.042829   0.005405  -7.925 3.31e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4373 on 2719 degrees of freedom
## Multiple R-squared:  0.04735,    Adjusted R-squared:  0.0456
## F-statistic: 27.03 on 5 and 2719 DF,  p-value: < 2.2e-16
```

What does the coefficient on prior convictions represent?

Linear Probability Model

$$Arrest_i^{86} = \ beta_0 + \beta_1 Priors_i + \beta_2 Sentence_i + \beta_3 PriorTime_i + \\ \beta_4 PrisonTime_i^{86} + \beta_5 QuaartersEmployed_i^{86} + u_i$$

$$Arrest_i^{86} = \begin{cases} 1 \text{ if number of arrests} > 0; \\ 0 \text{ otherwise} \end{cases}$$

$$\hat{y}_i = \hat{P}(Arrest_i^{86} = 1|x) = X\beta$$

$$\beta_j = \frac{\Delta \hat{P}(Arrest_i^{86} = 1|x)}{\Delta x_j}$$

Problems with the Linear Probability Model

Estimate the model including the respondent's income in 1986 (`inc86`) and name it `crime.lm2`. Summarize the `fitted.values` stored in the estimation results.

```
crime.lm2 <- lm((narr86 > 0) ~ pcnv + avgsen + tottime + ptime86 + qemp86, data = crime1)
summary(crime.lm2$fitted.values)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.006643 0.215691 0.269298 0.277064 0.354957 0.557690
```

Notice a problem?