Chapter 7 Code

Ryan Martin

February 28, 2018

Chapter 7 - Dummy Variables

Dummies on the RHS (Independent Variables)

Note, stata calculates the marginal effects of a binary variable (Utown, for example) on the output variable (price, for example) as

$$E(Price|Utown = 1) - E(Price|Utown = 0)$$

If the binary is only an indicator, this is just it's coefficient. If there are interaction terms too, then these also matter. Note that a binary {0,1 } variable to any power is still just that binary variable.

Hedonic Models - Characteristic space way before IO people thought of characteristic spaces. http://www.jstor.org/stable/1830899?seq=1#page_scan_tab_contents.

Note, there is a "dummy variable trap". If have an indicator for all cases and an intercept, get exact collinearity. Have to omit one. Doesn't matter which (but people almost always keep the coefficient and all but 1 of the dummies)

Dummy interacted with a continuous term is called a slope-indicator or a slope dummy variable

Indicator Vars - a.k.a. dummy, binary or dichotomous

Chow Test

Chow test is an F-test for the equivalence of two regressions. It's a really simple idea. Stick an indicator on the whole regression and do an F-test for the original vs the original plus an indicator times the original.

More concretely, suppose D is your indicator variable. Suppose you have

$$y = reg_1 + e$$

where $reg_1 = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_m x_m$. Note that reg_1 does not include the dummy variable. Then the Chow test also runs the regression

$$y = reg_1 + D \times reg_1 + e$$

. Call the first model the restricted and second model the unrestricted. Note that the number of terms that are forced to be 0 between the two models is m, not 1. So there are $J = m$ restrictions to the model. Look at SSE_R and SSE_U. $K = 2 \times m$. Do the F-test. This is a chow test. The null says we don't need to separte any part of reg 1's terms between the group D = 1 and the group with D = 0. The alternative says that there is at least one part of

Note, this still assumes homoskedasticity between groups D = 1 and D = 0. Allowing for heteroskedasticity in the test done in chapter 8

7.3 Log-Linear Models with Indicators

Suppose $logy = \beta_1 + \delta D + \beta_2 x_2$. Then, before we learned the interpretation of δ is that $100 \times \delta$ is the percent change in y for a 1 unit change in x. This was an approximation, through the fact that $\%\Delta y/\Delta x \approx \frac{dy}{dx} \times \frac{1}{y}$. The derivative approximation is less good when the change in x is large. The change in an indicator variable is usually a large change, so the percent change approximation can often be poor for indicator variable coefficients. Instead, we get the exact percent change in wage in this example is simply

$$100 \times \frac{Wage|_{D=1} - Wage|_{D=0}}{Wage|_{D=0}} = 100 \times (e^{\delta} - 1)$$

7.4 Linear Probability Model (Dummies on the LHS (Dependent Var))

Suppose y is either 1 or 0. Typically used to model choice between two (or more) products. Suppose each person the same, probability choose $y = 1$ is p and probability choose $y = 0$ is $1 - p$. Then, $E(y) = P(y = 1) = p$. We can model $E(y) = p = \beta_1 + \beta_2 x_2 + \cdots + \beta_K x_K$. This is called linear probability model. Two problems

1. for some x's, p < 0 or p>1 is possible.
2. var(e) = var(y) = p(1 - p)$ from econ 41. but

$$p(1-p) = (\beta_1 + \beta_2 x_2 + \cdots + \beta_K x_K)(1 - \beta_1 - \beta_2 x_2 - \cdots - \beta_K x_K)$$

   which depends on x. That is, variance of e depends on x, so this modeling choice requires heteroskedasticity, with variance largest when x values such that $\beta_1 + \beta_2 x_2 + \cdots + \beta_K x_K \approx 1/2$. Skew distribution when x are such that $p \approx 1, 0$

This model works best when x are s.t. p close to 1/2.

7.5 Treatment Effects

post hoc, ergo propter hoc - just because A happens before B does not mean A causes B.

If treatment is not random, people may self-select in, causes bias in treated effect. e.g. if mostly people who are working hard at losing weight join an spinning class, the weight-loss effects of the spinning exercise class will likely be biased upward. vs if most people who go out surfing only pretend to want exercise but really just want to sit in the ocean chatting with their buddies away from ear shot of their spouses, surfings effect on weight loss will be underestimated (because most of these people aren't putting in the effort). No way to know the bias without knowing characteristics of the people in group. Have to tell a story. Story is circumstancial.

If the errors fall into two classes, e_1 and e_0, the 1 type occuring when D = 1 and the 0 type occuring when D = 0, then (in treatment case)

$$b_2 = \beta_2 + \frac{\sum(d_i - \bar{d})(e_i - \bar{e})}{\sum(d_i - \bar{d})^2} = \beta_2 + (\bar{e_1} - \bar{e_0})$$

where the first equality comes from the formula for b_2 from the formula sheet with d_i plugged in for x_i and the second from some tricky algebra shown in the appendix. The tricks include the fact that $d_i^2 = d_i$ since d_i is 0 or 1. Then b_2 is consistent for β_2 if $E(\bar{e_1} - \bar{e_0}) = 0$.

Can regress treatment indicator on the other covariates to see if there is any significant linear dependence between treatment and other vars.

7.5.5 Diff-in-Diff - The most popular regression technique

Use this when have two groups over time. Assume the control groups change is exactly what change the treated group would undergo if they weren't treated. The treatment effect is the difference between their actual post-treatment outcome and their assumed outcome if they had the same changes as the untreated/control group. Called Diff in diff, because it's the difference of two differences.

$$\begin{aligned}\delta_{DD} &:= \bar{y}_{treatment,after} - \bar{y}_{control,after} - (\bar{y}_{treatment,before} - \bar{y}_{control,before}) \\ &= \bar{y}_{treatment,after} - \bar{y}_{treatment,before} - (\bar{y}_{control,after} - \bar{y}_{control,before})\end{aligned}$$

The first formula can be interpreted as "how much did treatment change the gap between treated and controlled." The second line can be interpreted as "How much more did the treated group change as compared to the untreated group". They are mathematically equivalent because of the distributive property.

Can show, if do the regression

$$y_{it} = \beta_1 + \beta_2 Treat_i + \beta_3 AFTER_t + \delta(Treat_i \times AFTER_t) + e_{it}$$

then δ is exactly the diff-in-diff estimator.

Could just calculate diff-in-diff using sample means, but regression is probably easier.

7.5.7 Panel Data

Time series is usually observation of one thing or a small collection of things over time. Panel is usually observing large collection of individuals over smaller time.

Since we observe each individual repeatedly in panel data, we can estimate some individual fixed effects.

7.9

In the STAR experiment (Section 7.5.3), children were randomly assiged within schools into three types of c lasses: small classes with 13-17 students, regular-sized classes with 22-25 students, and regular-sized classes with a full-time teacher aide to assist the teacher. Student scores on acheievement tests were recorded, as was some information about the students, teachers and schools. Data e kindergarten classes is contained in the data file `star.dat`

a

Calculate the average of TOTALSCORE for (i) students in regular-sized classrooms with full time teachers, but no aide; (ii) students in regular-sized classrooms with full time teachers and an aide; and (iii) students in small classrooms. What do you observe about test scores in these three types of learning environments?

```
my_wd <- "C:/Users/ryanj/Dropbox/TA/Econ 103/Winter 2018/Data/s4poe_statadata"
my_file <- paste(my_wd, "star.dta", sep = "/")
library(haven)
dat <- read_stata(my_file)
attach(dat)


#averages
mean( totalscore[  (small == 0)&(aide == 0) ])
```

```
## [1] 918.0429
```

```
mean( totalscore[ (small==0)&(aide==1) ] )
```

```
## [1] 918.3568
```

```
mean( totalscore[ (small==1) ])
```

```
## [1] 931.9419
```

b

Estimate the regression model

$$TOTALSCORE_i = \beta_1 + \beta_2 SMALL_i + \beta_3 AIDE_i + e_i$$

```
reg_out <- lm(totalscore ~ small + aide) #note, no interaction
    #because never have small and an aide
summary(reg_out)
```

```
##
## Call:
## lm(formula = totalscore ~ small + aide)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -283.04  -50.94   -7.04   42.64  334.64
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 918.0429     1.6412 559.363  < 2e-16 ***
## small        13.8990     2.4085   5.771  8.3e-09 ***
## aide          0.3139     2.3102   0.136    0.892
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 73.49 on 5783 degrees of freedom
## Multiple R-squared:  0.0073, Adjusted R-squared:  0.006957
## F-statistic: 21.26 on 2 and 5783 DF,  p-value: 6.3e-10
```

```
#note same results here from regression

reg_out$coefficients[1] + reg_out$coefficients[2] #small classroom
```

```
## (Intercept)
##    931.9419
```

```
    #no aide

reg_out$coefficients[1]    #regular no aid classroom
```

```
## (Intercept)
##    918.0429
```

```
reg_out$coefficients[1] + reg_out$coefficients[3]
```

```
## (Intercept)
##    918.3568
```

```
    #regular with aid classroom
```

c

Use a t-test to test the hypothesis $H_0 : \beta_2 \geq 1$ against the alternative $H_1 : \beta_2 < 1$

```
s <- summary(reg_out)
t.stat <- (coef(s)[2,1]- 1 )/coef(s)[2,2]
t.c <- qt(.95, s$df[2])
t.c
```

```
## [1] 1.645117
```

```
t.stat #reject null
```

```
## [1] 5.355524
```

c

To the regression in (b) add the additional explanatory variable TCHEXPER. Is this variable statistically significant? Does its addition to the model affect the estimates of β_2 and β_3?

```
reg_out2 <- lm(totalscore ~ small + aide + tchexper)
summary(reg_out2)
```

```
##
## Call:
## lm(formula = totalscore ~ small + aide + tchexper)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -274.13  -51.41   -7.44   42.57  341.53
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 904.7212     2.2280 406.071  < 2e-16 ***
## small        14.0061     2.3953   5.847 5.27e-09 ***
## aide         -0.6006     2.3065  -0.260    0.795
## tchexper      1.4690     0.1672   8.784  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 73.08 on 5762 degrees of freedom
##   (20 observations deleted due to missingness)
## Multiple R-squared:  0.02033,    Adjusted R-squared:  0.01982
## F-statistic: 39.86 on 3 and 5762 DF,  p-value: < 2.2e-16
```

Note, aide remains statistically insignificant, but now has switched signs to be a negative impact on scores! Teacher experience is positive and statistically significant, but practically insignificant (1.5 point impact only). The strength of small has now also increased, suggesting that many of the smaller classrooms are taught by inexperienced teachers. A quick check of the correlation between small and tchexper should verify this.

```
cor(small, tchexper, use = "pairwise.complete.obs")
```

```
## [1] -0.03542852
```

```
  #as anticipated, slightly negative
```

d

To the regression in (c) add the additional explanatory variables BOY, FREELUNCH and WHITE_ASIAN. Are any of these variables statistically significant? Does their addition to the model affect the estimates of β_2 and β_3?

```
reg_out3 <- lm(totalscore ~ small + aide + tchexper + boy + freelunch
               + white_asian) #note, no interaction
    #because never have small and an aide
summary(reg_out3)
```

```
##
## Call:
## lm(formula = totalscore ~ small + aide + tchexper + boy + freelunch +
##     white_asian)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -255.27  -47.92   -8.07   40.29  323.48
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 923.2498     3.1210 295.822  < 2e-16 ***
## small        13.8960     2.2936   6.059 1.46e-09 ***
## aide          0.6983     2.2093   0.316    0.752
## tchexper      1.1140     0.1613   6.908 5.43e-12 ***
## boy         -14.0452     1.8457  -7.610 3.19e-14 ***
## freelunch   -34.1170     2.0639 -16.531  < 2e-16 ***
## white_asian  11.8373     2.2108   5.354 8.92e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 69.98 on 5759 degrees of freedom
##   (20 observations deleted due to missingness)
## Multiple R-squared:  0.1023, Adjusted R-squared:  0.1013
## F-statistic: 109.3 on 6 and 5759 DF,  p-value: < 2.2e-16
```

All these new variables are statistically significant. The (still) statistically insignificant "aide" has swithced sign again (but is still very small). Small gets slightly reduced by its inclusion, suggesting the smaller classrooms either had more white/asian students, slightly less boys or slightly less freelunchs (or some combination)

```
lm(small ~ freelunch + boy + white_asian + tchexper + aide)
```

```
##
## Call:
## lm(formula = small ~ freelunch + boy + white_asian + tchexper +
##     aide)
##
## Coefficients:
## (Intercept)    freelunch          boy  white_asian     tchexper
##   0.4655965   -0.0004265    0.0009103    0.0027876   -0.0003800
##        aide
```

```
##  -0.4639696
```

e

To the regression in (d), add the additional explanatory variables TCHWHITE, TCHMASTERS, SCHURBAN and SCHRURAL. Are any of these variables statistically significant? Does their addition to the model affect the estimates of β_2 and β_3?

```
reg_out4 <- lm(totalscore ~ small + aide + tchexper +
                 boy + freelunch + white_asian +
                 tchwhite + tchmasters + schurban + schrural)
summary(reg_out4)
```

```
##
## Call:
## lm(formula = totalscore ~ small + aide + tchexper + boy + freelunch +
##     white_asian + tchwhite + tchmasters + schurban + schrural)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -262.91  -47.12   -8.43   39.97  314.46
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 931.7553     3.9401 236.482  < 2e-16 ***
## small        13.9803     2.3023   6.072 1.34e-09 ***
## aide          1.0023     2.2166   0.452  0.65114
## tchexper      1.1562     0.1657   6.979 3.31e-12 ***
## boy         -14.0083     1.8426  -7.602 3.37e-14 ***
## freelunch   -32.5325     2.1260 -15.302  < 2e-16 ***
## white_asian  16.2326     2.7804   5.838 5.57e-09 ***
## tchwhite     -7.6683     2.8420  -2.698  0.00699 **
## tchmasters   -3.5598     2.0193  -1.763  0.07798 .
## schurban     -5.7499     2.8580  -2.012  0.04428 *
## schrural     -7.0061     2.5585  -2.738  0.00619 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 69.86 on 5755 degrees of freedom
##   (20 observations deleted due to missingness)
## Multiple R-squared:  0.1059, Adjusted R-squared:  0.1043
## F-statistic: 68.14 on 10 and 5755 DF,  p-value: < 2.2e-16
```

Of the newly added variables, only `tchmasters` is not statistically significant (but its almost). Given the rest of the model, a teacher with a masters degree (surprisingly?) lowers test scores. suburb schools do the best, with rural and urban schools below (unsurprisingly). The teacher being white seems to significantly lower scores by nearly 8 points, on average. Still, the most (practically) significant term is freelunch. small's effect remains close to all of its previous estimates, near 14.

f

Discuss the importantce of parts (c), (d) and (e) to our estimation of the "treatment" effects in part(b).

Solution The inclusion of our controls (a.k.a. covariates or other dependent variables that aren't treatment

effects) didn't impact our estimate of the "treatment" effect much at all. This is because all of our controls/covariates are relatively uncorrelated with treatment. If there were a control that were strongly colinear with small, we would expect it to take away from small's estimate.

g

Add to the models in (b) through (e) indicator variables for each school.

$$SCHOOL_j = \begin{cases} 1 \text{ if student is in school } j \\ 0 \text{ otherwise} \end{cases}$$

Test the significance of these school "fixed effects." Does the inclusion of these fixed effect indicator variables substantially alter the estimate of β_2 and β_3

```
reg_out_sfe <- lm(totalscore ~ small + aide + schid)
summary(reg_out_sfe)
```

```
##
## Call:
## lm(formula = totalscore ~ small + aide + schid)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -283.07  -51.04   -7.08   42.74  334.78
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 9.174e+02  5.576e+00 164.508  < 2e-16 ***
## small       1.390e+01  2.409e+00   5.771 8.27e-09 ***
## aide        3.162e-01  2.310e+00   0.137    0.891
## schid       3.194e-06  2.518e-05   0.127    0.899
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 73.5 on 5782 degrees of freedom
## Multiple R-squared:  0.007303,   Adjusted R-squared:  0.006788
## F-statistic: 14.18 on 3 and 5782 DF,  p-value: 3.327e-09
```

```
reg_out2_sfe <- lm(totalscore ~ small + aide + tchexper +
                  schid)
summary(reg_out2_sfe)
```

```
##
## Call:
## lm(formula = totalscore ~ small + aide + tchexper + schid)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -274.24  -51.07   -7.26   42.56  342.18
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.015e+02  5.829e+00 154.661  < 2e-16 ***
## small        1.403e+01  2.396e+00   5.855 5.03e-09 ***
## aide        -5.936e-01  2.307e+00  -0.257    0.797
```

```
## tchexper     1.474e+00  1.675e-01   8.803  < 2e-16 ***
## schid        1.479e-05  2.507e-05   0.590    0.555
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 73.09 on 5761 degrees of freedom
##   (20 observations deleted due to missingness)
## Multiple R-squared:  0.02039,    Adjusted R-squared:  0.01971
## F-statistic: 29.98 on 4 and 5761 DF,  p-value: < 2.2e-16
```

```
reg_out3_sfe <- lm(totalscore ~ small + aide + tchexper +
                 boy + freelunch + white_asian+schid)
summary(reg_out3_sfe)
```

```
##
## Call:
## lm(formula = totalscore ~ small + aide + tchexper + boy + freelunch +
##     white_asian + schid)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -258.76  -47.60   -7.91   40.23  333.59
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.878e+02  6.983e+00 127.136  < 2e-16 ***
## small        1.408e+01  2.288e+00   6.156 7.95e-10 ***
## aide         8.831e-01  2.204e+00   0.401    0.689
## tchexper     1.124e+00  1.608e-01   6.990 3.07e-12 ***
## boy         -1.401e+01  1.841e+00  -7.609 3.21e-14 ***
## freelunch   -3.336e+01  2.063e+00 -16.172  < 2e-16 ***
## white_asian  1.708e+01  2.391e+00   7.144 1.02e-12 ***
## schid        1.484e-04  2.616e-05   5.671 1.49e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 69.79 on 5758 degrees of freedom
##   (20 observations deleted due to missingness)
## Multiple R-squared:  0.1072, Adjusted R-squared:  0.1062
## F-statistic: 98.81 on 7 and 5758 DF,  p-value: < 2.2e-16
```

```
reg_out4_sfe <- lm(totalscore ~ small + aide + tchexper +
                 boy + freelunch + white_asian +
                 tchwhite + tchmasters + schurban + schrural+
                   schid)
summary(reg_out4_sfe)
```

```
##
## Call:
## lm(formula = totalscore ~ small + aide + tchexper + boy + freelunch +
##     white_asian + tchwhite + tchmasters + schurban + schrural +
##     schid)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -264.50  -47.13   -8.51   40.26  324.84
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.972e+02  7.549e+00 118.844  < 2e-16 ***
## small        1.401e+01  2.297e+00   6.099 1.14e-09 ***
## aide         1.040e+00  2.211e+00   0.471  0.63799
## tchexper     1.170e+00  1.653e-01   7.081 1.60e-12 ***
## boy         -1.397e+01  1.838e+00  -7.601 3.40e-14 ***
## freelunch   -3.173e+01  2.126e+00 -14.924  < 2e-16 ***
## white_asian  1.997e+01  2.860e+00   6.982 3.25e-12 ***
## tchwhite    -5.408e+00  2.866e+00  -1.887  0.05924 .
## tchmasters  -3.902e+00  2.016e+00  -1.936  0.05294 .
## schurban    -6.782e+00  2.858e+00  -2.373  0.01766 *
## schrural    -6.984e+00  2.552e+00  -2.736  0.00623 **
## schid        1.422e-04  2.653e-05   5.361 8.59e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 69.7 on 5754 degrees of freedom
##   (20 observations deleted due to missingness)
## Multiple R-squared:  0.1103, Adjusted R-squared:  0.1086
## F-statistic: 64.86 on 11 and 5754 DF,  p-value: < 2.2e-16
```

```
length(unique(schid)) #79 schools
```

```
## [1] 79
```

```
nrow(dat) #5786 classroom observations
```

```
## [1] 5786
```

We see that adding the school fixed effects doesn't change the estimated treatment effect either

7.2

In September 1998, a local TV station contacted an econometrician to analyze some data for them. They were going to do a Halloween story on the legend of full mons' affecting behavior in strange ways. They collected data from a local hospital on emergency room cases for the period from January 1, 1998 until mid-August. There were 229 observations (days). During this time, there were eight full moons and seven new moons (a related myth concerns new moons) and three holidays (NYD, Memorial Day and Easter). If there is a full-moon effect, then hospital administrators will adjust numbers of emergency room doctors and nurses and local police may change the number of officers on duty. (This sounds like a very 90s era question.)

`Using the data in the file \texttt{fullmoon.dat} we obtain the regression results in the following tabl`

a

Interpret these regression results. When should emergency rooms expect more calls?

Solution from this data, holidays seem to have the largest effect on emergency rooms. (As an aside, note that the most dangerous days to drive are new years day - which includes the 3am drives home from new year's eve parties - and Memorial Day). The weekend also drives up incidences. Fullmoon and newmoon have positive, but statistically insignificant coefficients. So, the evidence isn't strong enough to confirm that fullmoons or newmoons cause problems, separately.

Emergency Room Cases Regression—Model 1

Variable	Coefficient	Std. Error	*t*-Statistic	Prob.
C	93.6958	1.5592	60.0938	0.0000
T	0.0338	0.0111	3.0580	0.0025
HOLIDAY	13.8629	6.4452	2.1509	0.0326
FRIDAY	6.9098	2.1113	3.2727	0.0012
SATURDAY	10.5894	2.1184	4.9987	0.0000
FULLMOON	2.4545	3.9809	0.6166	0.5382
NEWMOON	6.4059	4.2569	1.5048	0.1338

$R^2 = 0.1736$ $SSE = 27108.82$

Figure 1: New Moon and Full Moon Included

```
my_wd <- "C:/Users/ryanj/Dropbox/TA/Econ 103/Winter 2018/Data/s4poe_statadata"
my_file <- paste(my_wd, "fullmoon.dta", sep = "/")
library(haven)
dat <- read_stata(my_file)
attach(dat)
library(Hmisc)
```

```
## Loading required package: lattice
## Loading required package: survival
## Loading required package: Formula
## Loading required package: ggplot2
##
## Attaching package: 'Hmisc'
## The following objects are masked from 'package:base':
##
##     format.pval, round.POSIXt, trunc.POSIXt, units
```

```
describe(dat) #closer to Stata's tabulate
```

```
## dat
##
##  7  Variables      229  Observations
## ---------------------------------------------------------------------------
## cases : number of emergency room cases  Format:%8.0g
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      229        0       52    0.999    100.6    13.54       82       86
##      .25      .50      .75      .90      .95
##       92      100      109      117      120
##
## lowest :  69  75  77  79  81, highest: 125 127 129 131 141
## ---------------------------------------------------------------------------
## t : time indicator, t = 1, ..., 229  Format:%8.0g
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      229        0      229        1      115    76.67     12.4     23.8
##      .25      .50      .75      .90      .95
##     58.0    115.0    172.0    206.2    217.6
##
## lowest :   1   2   3   4   5, highest: 225 226 227 228 229
## ---------------------------------------------------------------------------
```

Emergency Room Cases Regression—Model 2

Variable	Coefficient	Std. Error	*t*-Statistic	Prob.
C	94.0215	1.5458	60.8219	0.0000
T	0.0338	0.0111	3.0568	0.0025
HOLIDAY	13.6168	6.4511	2.1108	0.0359
FRIDAY	6.8491	2.1137	3.2404	0.0014
SATURDAY	10.3421	2.1153	4.8891	0.0000

$R^2 = 0.1640$ $SSE = 27424.19$

Figure 2: New Moon and Fullmoon omitted

```
## holiday : =1 if day was a holiday  Format:%8.0g
##        n  missing distinct     Info      Sum     Mean      Gmd
##      229        0        2    0.039        3   0.0131  0.02597
##
## ------------------------------------------------------------------------
## friday : =1 if friday  Format:%8.0g
##        n  missing distinct     Info      Sum     Mean      Gmd
##      229        0        2     0.37       33   0.1441   0.2478
##
## ------------------------------------------------------------------------
## saturday : =1 if saturday  Format:%8.0g
##        n  missing distinct     Info      Sum     Mean      Gmd
##      229        0        2     0.37       33   0.1441   0.2478
##
## ------------------------------------------------------------------------
## fullmoon : =1 if moon was full  Format:%8.0g
##        n  missing distinct     Info      Sum     Mean      Gmd
##      229        0        2    0.101        8  0.03493  0.06772
##
## ------------------------------------------------------------------------
## newmoon : =1 if moon was new  Format:%8.0g
##        n  missing distinct     Info      Sum     Mean      Gmd
##      229        0        2    0.089        7  0.03057  0.05953
##
## ------------------------------------------------------------------------
```

```
  #note, only 3 holidays in 229 observations. should be wary of that
  #8 full moons, 33 friday and saturdays ,7 newmoons
```

b

The model was reestimated omitting the variables FULLMOON and NEWMOON, as shown in Figure 2. Comment on any changes you observe.

Solution

Regression results seem pretty much unchanged after dropping fullmoon and newmoon.

c

Test the joint significance of FULLMOON and NEWMOON. State the null and alternative hypotheses and indicate the test statistic you use. What do you conclude?

Solution The null is that Fullmoon and Newmoon both have 0 coefficients (or do not influence emergency room incidents). The alternative is that at least one of them does.

```
SSE_R = 27424.19
J = 2
N = nrow(dat)
K = 7
SSE_U = 27108.82
F.stat = (SSE_R - SSE_U)*(N - K)/ J / SSE_U
F.stat
```

```
## [1] 1.291317
```

```
F.c = qf(.95, J, N-K)
F.c
```

```
## [1] 3.036524
```

Fail to reject the null. Given the data, neither is likely to have an influence.

7.14

Professor Ray C. Fair's voting model was introduced in Exercise 2.14. He builds models that explain and predict the U.S. presidential elections. See his website at http://fairmodel.econ.yale.edu/vote2008/index2.htm. The basic premise of the model is that the incumbent party's share of the two-party popular vote is affected by a number of factors relating to the economy and variables relating to the politics, such as how long the incumbent party has been in power and whether the president is running for reelection. Fair's data, 33 observations for the election years from 1880 to 2008, are in the file `fair4.dat`. The dependent variable is VOTE = percentage share of the popular vote won by the incumbent party. The explanatory variables include:

- PARTY: Binary variable that's either 1 or -1. 1 if Democratic incumbent, -1 if Republican.
- PERSON: binary (0-1), 1 if incumbent is runing for (re)election.
- DURATION: discrete variable. 0 if the incumbent party has been in power for one term, 1 if the incumbent party has been in power for two consecutive terms, 1.25 if the incumbent has been in power for three consecutive terms, 1.50 for four, etc.
- WAR. binary (0-1). 1 in 1920, 1944, 1948. 0 otherwise
- GROWTH: growth rate of real per capita GDP in the first 3 quarters of the election year (annual rate).
- INFLATION: absolute value of the growth rate of GDP deflator in first 15 quarters of the administration (annual rate) except for 1920, 1944 and 1948 (where values are 0).
- GOODNEWS: The number of quarters in the first 15 quarters of the administration in which the growth rate of real per capita GDP is greater than 3.2% at an annual rate, except for 1920, 1944 and 1948, where the values are 0.

a

Consider the model

$$VOTE = \beta_1 + \beta_2 GROWTH + \beta_3 INFLATION + \beta_4 GOODNEWS + \beta_5 PERSON + \beta_6 DURATION + \beta_7 PARTY + \beta_8 WAR + e$$

Discuss the anticipated effects of the dummy variables PERSON and WAR

Solution We expect the effect of person to be positive. It is well known that incumbents have an advantage in elections. We know that, of the three election years with war in this data, the incumbent party lost once (1920) and won twice (1940s). So, the program will estimate war's coefficient to be positive (and probably statistically insignificant since so few observations where war = 1)

b

The binary variable PARTY is somewhat different from the dummy variables we have considered. Write out the regression function E(VOTE) for the two values of PARTY. Discuss the effects of this specification.

Solution Party is -1 or 1 rather than 0 or 1. So, this makes it so that the difference between the two is $2 \times \beta_7$ rather than just β_7 in the original specification. This just affects our interpretation.

More formally,

$$\begin{aligned} E(VOTE|PARTY = 1) - E(VOTE|PARTY = 0) = \\ = \beta_7 \times 1 - (\beta_7 \times -1) \\ = 2\beta_7 \end{aligned}$$

and all the other terms cancel

c

Use the data for the period 1916-2004 to estimate the proposed model. Discuss the estimation results. Are the signs as expected? Are the estimates statistically significant? How well does the model fit the data?

```
my_wd <- "C:/Users/ryanj/Dropbox/TA/Econ 103/Winter 2018/Data/s4poe_statadata"
my_file <- paste(my_wd, "fair4.dta", sep = "/")
library(haven)
dat <- read_stata(my_file)
attach(dat)
library(Hmisc)
describe(dat) #closer to Stata's tabulate
```

```
## dat
##
##  9  Variables      33  Observations
## ---------------------------------------------------------------------------
## year  Format:%9.0g
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##       33        0       33        1     1944    45.33     1886     1893
##      .25      .50      .75      .90      .95
##     1912     1944     1976     1995     2002
##
## lowest : 1880 1884 1888 1892 1896, highest: 1992 1996 2000 2004 2008
## ---------------------------------------------------------------------------
## vote : Incumbent share of the two-party presidential vote  Format:%9.0g
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##       33        0       33        1     52.1     6.87    43.09    45.07
##      .25      .50      .75      .90      .95
##    48.95    51.68    55.00    59.84    61.52
```

```
##
## lowest : 36.119 40.841 44.595 44.697 46.545, highest: 59.170 60.006 61.344 61.789 62.458
## ---------------------------------------------------------------------------
## party : = 1 if Democratic incumbent at election time; -1 if a Republican incumbent  Format:%9.0g
##        n  missing distinct     Info     Mean      Gmd
##       33        0        2    0.733  -0.1515    1.008
##
## Value         -1     1
## Frequency     19    14
## Proportion 0.576 0.424
## ---------------------------------------------------------------------------
## person : = 1 if incumbent is running for election and 0 otherwise  Format:%9.0g
##        n  missing distinct     Info      Sum     Mean      Gmd
##       33        0        2    0.733       19   0.5758   0.5038
##
## ---------------------------------------------------------------------------
## duration : number of terms incumbent administration in power  Format:%9.0g
##        n  missing distinct     Info     Mean      Gmd
##       33        0        6    0.894   0.7121   0.7301
##
## Value       0.00  1.00  1.25  1.50  1.75  2.00
## Frequency     14    10     4     2     2     1
## Proportion 0.424 0.303 0.121 0.061 0.061 0.030
## ---------------------------------------------------------------------------
## war : = 1 for elections of 1920, 1944, and 1948 and 0 otherwise.  Format:%9.0g
##        n  missing distinct     Info      Sum     Mean      Gmd
##       33        0        2    0.248        3  0.09091   0.1705
##
## ---------------------------------------------------------------------------
## growth : growth rate GDP in first three quarters of the election year  Format:%9.0g
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##       33        0       33        1   0.6243    5.827  -10.600   -6.135
##      .25      .50      .75      .90      .95
##   -1.451    2.229    3.902    5.096    5.630
##
## lowest : -14.499 -11.463 -10.024  -6.281  -5.553
## highest:   5.043   5.109   5.440   5.914  11.765
## ---------------------------------------------------------------------------
## inflation : growth rate of GDP deflator during first 15 quarters of admin  Format:%9.0g
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##       33        0       31    0.999    2.666    2.296   0.0000   0.1014
##      .25      .50      .75      .90      .95
##   1.4420   2.2740   3.2800   5.2394   7.3720
##
## lowest : 0.000 0.081 0.183 0.604 1.055, highest: 5.161 5.259 7.200 7.630 7.831
## ---------------------------------------------------------------------------
## goodnews : number of quarters in first 15 with real GDP per capita growth > 3.2  Format:%9.0g
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##       33        0       11    0.986    5.273     3.36      0.0      1.2
##      .25      .50      .75      .90      .95
##      3.0      5.0      8.0      8.8      9.4
##
## Value          0     1     2     3     4     5     6     7     8     9
## Frequency      3     1     2     3     4     5     1     5     5     2
```

```
## Proportion 0.091 0.030 0.061 0.091 0.121 0.152 0.030 0.152 0.152 0.061
##
## Value         10
## Frequency      2
## Proportion 0.061
## ---------------------------------------------------------------------------
```

```
  #note, only 3 holidays in 229 observations. should be wary of that
  #8 full moons, 33 friday and saturdays ,7 newmoons

reg_out <- lm(vote ~ growth + inflation + goodnews + person +
                duration + party + war)
s <- summary(reg_out)
s
```

```
##
## Call:
## lm(formula = vote ~ growth + inflation + goodnews + person +
##     duration + party + war)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.7018  -2.1252   0.7512   1.9468   8.9780
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  51.4352     2.9767  17.279 2.06e-15 ***
## growth        0.5132     0.1444   3.555  0.00154 **
## inflation    -0.8111     0.4042  -2.007  0.05572 .
## goodnews      0.7182     0.3171   2.265  0.03245 *
## person        1.1734     1.8475   0.635  0.53113
## duration     -3.3463     1.4542  -2.301  0.03000 *
## party        -1.7603     0.8263  -2.130  0.04316 *
## war           1.7635     3.7206   0.474  0.63963
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.095 on 25 degrees of freedom
## Multiple R-squared:  0.6429, Adjusted R-squared:  0.543
## F-statistic: 6.431 on 7 and 25 DF,  p-value: 0.000217
```

As expected, war and person both have a positive sign. Also as expected, war is not significant (probably due to such few observations). Surprisingly, person is not significant. Parties negative coefficient says that republicans have a relatively stronger predicted probability of staying in power than democrats over the time period

d

Predict the outcome of the 2008 election using the given 2008 data for values of the explanatory variables. Based on the prediction, would you have picked the outcome of the election correctly?

Solution Prediction just has us plugging in 2008 election values to our regression output. Since there are many, I will let the computer do it automatically. We predict the incumbent will get less than .50 percent of the popular vote (which was correct in 2008, GWB lost and Obama won, switching parties.). (Note that we did use this data to estimate it too! It might be more honest to drop the 2008 data and fit it without, then

try to see if we can still get it right!).

```
predict(reg_out, newdata = dat[dat$year=="2008",],
        interval = "confidence")
```

```
##        fit      lwr      upr
## 1 49.78077 46.70148 52.86006
```

```
predict(reg_out, newdata = dat[dat$year=="2008",],
        interval = "prediction")
```

```
##        fit      lwr      upr
## 1 49.78077 40.80321 58.75833
```

e

Construct a 95% prediction interval for the outcome of the 2008 election.

Solution Prediciton interval was calculated above.

f

Using data values of your choice (you must explain them), predict the outcome of the 2012 election.

Solution This is a bonus one. You can look up the data if you want. a lot is online through government websites, but it's usually time-consuming to find, download and clean.

Chapter 7 New Stata Code

```
!missing
areg
  - regression that automatically absorbs/suppresses
  - some of the outputs. do when have a large dummy variable
  - collection where not concerned about their coefficient estimates
areg varlist, absorb var
  - reg on varlist, subpresses output of vars
esttab
  - creates table, see below
tabulate
  - summary table
tabul varname, gen()
pwcorr
  - calculates pairwise correlations
  - almost same as corr, difference in
  - how treats missing variables.
global
  - create a variable that can be inserted into
  - multiple data collections, with $. p.231



// when in doubt, can always use, e.g.
// help areg
```

```
//examples
esttab model1 model2 model3 model4, se(%12, 3f) b(%12, 3f) ///
star(*.10 ** .05 *** .0) gaps ar2 bic scalars(rss) ///
  title("Project Star: Kindergarden")
// creates table ith 4 models with coefficients and standard errors
// sets signifiacnace stars as described

// see also
//help estout
```