Problem set 3: hypothesis testing and model selection

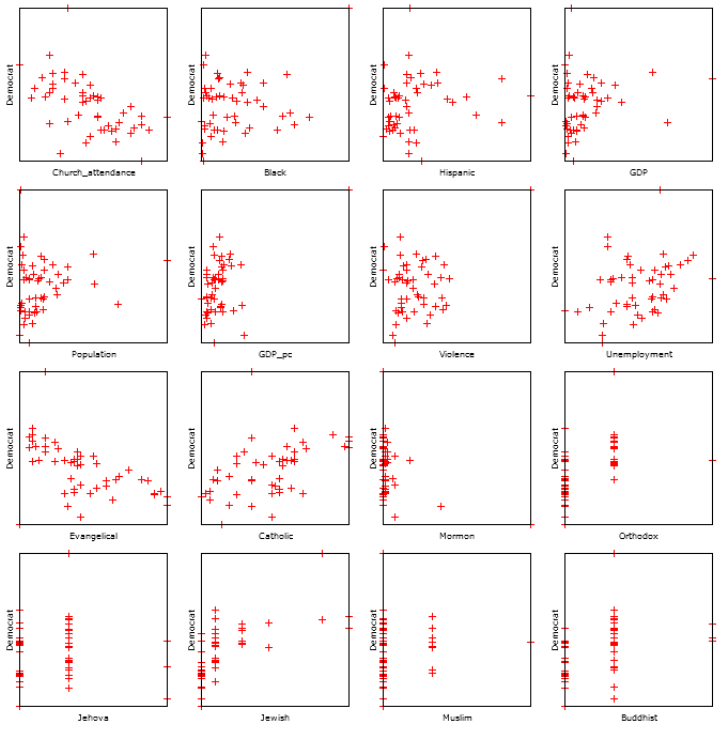
September 16, 2013

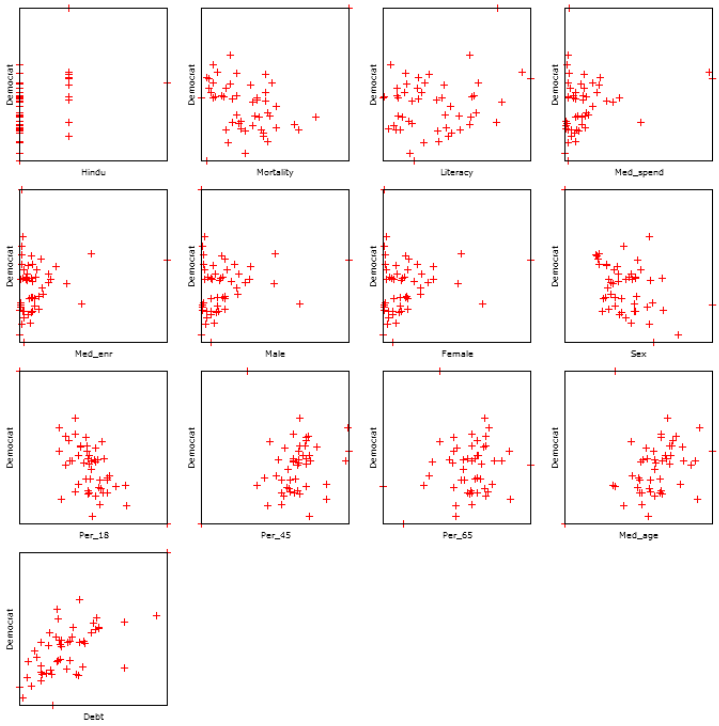
1 Introduction

These are the answers to the third econometrics problem set.

2 2012 Presidential Election Results - practical

1. (a) The results of this exercise are shown in the two plots below.





In my view the variables which show the most robust, (and convincing) relationships with the Democratic share of the vote, *Democrat*, are:

- Church attendance
- GDP_pc
- Unemployment
- Evangelical
- Catholic
- Debt

These are very much only on first glances, so I am sure you can argue that others similarly fit the bill.

(b) The results of the regression are shown below. All the variables apart from *Med_enr* are significant at the 5% level. This is because the p value for this variable is above 0.05.

Model X: OLS, using observations 1–51 (n = 49)
Missing or incomplete observations dropped: 2
Dependent variable: Democrat

	Coefficient	Std. Error	t-ratio	p-value
const	0.332479	0.0360133	9.2321	0.0000
GDP_pc	2.93950e–006	6.02674e–007	4.8774	0.0000
Mormon	−0.00400827	0.00147351	−2.7202	0.0092
Med_enr	1.17382e–008	6.92476e–009	1.6951	0.0970

Mean dependent var	0.487653	S.D. dependent var	0.115906
Sum squared resid	0.354681	S.E. of regression	0.088779
R^2	0.449974	Adjusted R^2	0.413305
$F(3, 45)$	12.27143	P-value(F)	5.39e–06
Log-likelihood	51.21678	Akaike criterion	−94.43356
Schwarz criterion	−86.86628	Hannan–Quinn	−91.56254

(c) The interpretation of each coefficient is that it is the increase in Democratic share of the vote from a 1 unit change in that variable. Note the importance of units here. Since each variable has different units it is very hard/impossible to compare these.

(d) The results of the new regression are shown below.

Model X: OLS, using observations 1–51 ($n = 49$)
Missing or incomplete observations dropped: 2
Dependent variable: Democrat

	Coefficient	Std. Error	t-ratio	p-value
const	0.352458	0.0347161	10.1526	0.0000
GDP_pc	2.87681e–006	6.13664e–007	4.6879	0.0000
Mormon	−0.00432290	0.00149124	−2.8989	0.0057

Mean dependent var	0.487653	S.D. dependent var	0.115906
Sum squared resid	0.377328	S.E. of regression	0.090569
R^2	0.414853	Adjusted R^2	0.389412
$F(2, 46)$	16.30638	P-value(F)	4.44e–06
Log-likelihood	49.70031	Akaike criterion	−93.40062
Schwarz criterion	−87.72516	Hannan–Quinn	−91.24736

We can use one of the three tests we discussed in lectures: Goldfeld-Quandt, Breusch-Pagan or White. The Goldfeld-Quandt looks for heteroscedasticity along one variable, the Breusch-Pagan across all variables, and the White looks for heteroscedasticity across the squares/cubes etc. of variables and their cross-products. I have opted for the Breusch-Pagan test here. The reason I have not chosen to use the White is because the sample size is quite small. The LM statistic for this test is found to be 0.989452, with a p value which is 0.609738. Hence we conclude that there is no serial correlation at the 5% level.

(e) To test for functional misspecification I use the Ramsey RESET test option on Gretl. I use the 'squares and cubes' option for the test, and get an F statistic of 0.268, with an associated p value of 0.766. Hence we conclude that the model is not functionally misspecified at the 5 % level.

(f) No need for an answer.

(g) The results of this regression are shown below.

Model X: OLS, using observations 1–51 ($n = 49$)
Missing or incomplete observations dropped: 2
Dependent variable: Democrat

	Coefficient	Std. Error	t-ratio	p-value
const	−2.35960	0.639633	−3.6890	0.0006
lgdp_pc	0.234963	0.0527394	4.4552	0.0001
lmormon	−0.0228561	0.0135063	−1.6923	0.0975
lmed_enr	0.0229046	0.0133206	1.7195	0.0924

Mean dependent var	0.487653	S.D. dependent var	0.115906
Sum squared resid	0.390733	S.E. of regression	0.093182
R^2	0.394065	Adjusted R^2	0.353670
$F(3, 45)$	9.755141	P-value(F)	0.000045
Log-likelihood	48.84502	Akaike criterion	−89.69005
Schwarz criterion	−82.12277	Hannan–Quinn	−86.81903

(h) The interpretation of the coefficient is the % change in Democratic share of the vote with respect to a 1 % increase in GDP per capita in that State.

(i) A 10% increase in GDP per capita would result in a 2.35% increase in the Democratic share of the vote.

(j) We could create a new variable, equal to the difference between the two variables, and regress this on *Democratic*, along with logged GDP per capita and logged Mormon % in the State. If we then conducted a t test on the Mormon variable, this would be a test of whether these two variables have the same magnitude, but opposite sign, effects. This is slightly different, (but the essential idea remains), to the way in which Gretl tests coefficient restrictions.

(k) The results from this test are shown below. We cannot reject the null hypothesis that these two coefficients are the same in magnitude, but opposite in sign.

Restriction:

b[lmormon] + b[lmed_enr] = 0

Test statistic: F(1, 45) = 4.90311e-006, with p-value = 0.9982

Restricted estimates:

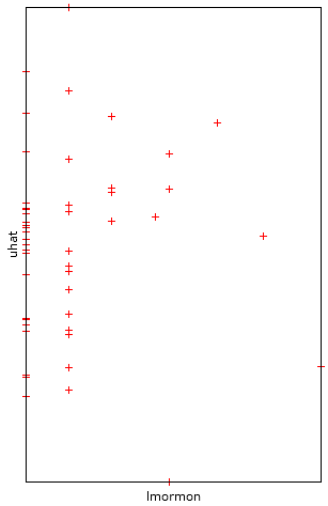
	coefficient	**std. error**	**t ratio**	**p value**
const	2.35892	0.554575	4.254	0.0001
lgdp_pc	0.23493	0.050051	4.694	2.44e-05 ***
lmormon	0.0228806	0.007659	2.988	0.0045***
lmed_enr	0.0228806	0.007659	2.988	0.0045***

(l) No answer required.

(m) There is heteroscedasticity across the Mormon variable, as well as the cross product of (log) GDP per capita and the Mormon variable.

(n) A scatter plot of the residuals (or squared residuals) against the variable in question might be a good way to check visually for heteroscedasticity. You

first need to create the residuals series, as indicated in the problem set. Then by typing 'scatters uhat; lmormon' you can draw a scatter plot like the one shown below. From inspection, one can see that the average error is decreasing as the % population who are Mormon increases. The logic behind this could be that, there are many different types of State who have low Mormon populations, these could be composed of both democratic and right wing states. However, as the State's Mormon population increases this tends to increase the probability that the State will vote Republican (due in part to Mitt Romney's religious status); reducing the variability for these states. In theory if enough variables were controlled for, then this heteroscedasticity might go away.



(o) Drawing a histogram of the residuals and testing for normality is easy using Gretl. Just type 'freq uhat –normal' into the console. Gretl then draws the frequency plot of your residuals, overlaying a normal distribution. It then does a Doornik-Hansen chi-squared test for normality. (The details of this test are beyond the scope of this course, but it will suffice for our purposes.) We conclude that we cannot reject the null hypothesis that our residuals are normal at the 5% level.

(p) Inference should be fine based on the results of the previous test. There is no need to use non-normal distributions for inference.

(q) An example (by no means best!) model which I came up with for the non-logged Democratic share of the vote is shown below.

Model Example output: OLS, using observations 1–51 ($n = 49$)
Missing or incomplete observations dropped: 2
Dependent variable: Democrat

	Coefficient	Std. Error	t-ratio	p-value
const	0.446800	0.0535356	8.3458	0.0000
Evangelical	−0.00530384	0.000913823	−5.8040	0.0000
Debt	1.00062e–005	3.98749e–006	2.5094	0.0159
Mormon	−0.00512913	0.00102022	−5.0275	0.0000
Black	0.295791	0.0873684	3.3856	0.0015
Unemployment	0.0105263	0.00534880	1.9680	0.0555

Mean dependent var	0.487653	S.D. dependent var	0.115906
Sum squared resid	0.138512	S.E. of regression	0.056756
R^2	0.785200	Adjusted R^2	0.760223
$F(5, 43)$	31.43725	P-value(F)	2.52e–13
Log-likelihood	74.25312	Akaike criterion	−136.5062
Schwarz criterion	−125.1553	Hannan–Quinn	−132.1997

3 Theory

2. (a) The coefficient on *books* is elasticity of wages with respect to books read. Equivalently, it is the % increase in wages associated with a 1% increase in the number of books read.

   (b) No. The number of books read is not significant. This is because the t stat on this variable is 1.11; less than the 95% critical value of a t distribution with 96 degrees of freedom.

   (c) The test for this part is based upon the constructed t statistic:

   $$t = \frac{\hat{\beta}-1}{se(\hat{\beta})}$$

   Which in this case is just:

   $$t = \frac{1.24-1}{0.23} = 1.04$$

   Since this t statistic is less than the critical value for a t with 96 degrees of freedom, we cannot conclude that the value of this coefficient is statistically different from one.

   (d) The way this works is by investigating how much of the variance of each independent variable can be explained by the variance in linear combinations of the other variables. If it turns out that a lot of variance can be explained, then R-squared is close to one, meaning that 1 minus R-squared is close to zero, causing the *VIF* to be high; indicating a large degree of multicollinearity.

   (e) Books is very correlated with each of the other factors. This means that inclusion of this variable along with the other two may cause large standard errors, and cause instability in the regression results.

   (f) Since *fatheduc* is likely positively correlated with *books*, we would expect that the removal of *books* from the regression results will cause an increase in the coefficient on *fatheduc*. The intuition here is that *books* is taking some of the credit away from *fatheduc*.

   (g) The R-squared of 0.37 means that 37% of the variance in the dependent variable is explained by the model.

(h) Since the coefficient on *iq* is not statistically different from zero, we cannot conclude that there is statistically any difference between father's education and iq (in logged form), on the log of wage. See this video for further explanation.

3. (a) The interpretation is the $ return in sales, on each extra $ spent on advertising. This likely overstates the effect of advertising since there is likely reverse causation occuring, where companies which spend more have more money to spend on advertising.

   (b) The interpretation of the coefficient on *Consumer* is the difference in average sales of a SME for selling a consumer-facing product, opposed to a B2B business.

   (c) It looks like there might be heteroscedasticity along the *Years* variable, since it is the only variable which has a t statistic above the critical value. (This auxillary regression is essentially a Breusch-Pagan test).

   (d) OLS estimators will be inefficient, but still unbiased. GLS estimators could be used, and will be BLUE. Alternatively, corrected standard errors could be used.

   (e) Taking the log of the dependent variable acts to suppress variation in it. This can often have the desirable effect of removing heteroscedasticity from a model.