Problem Set 1

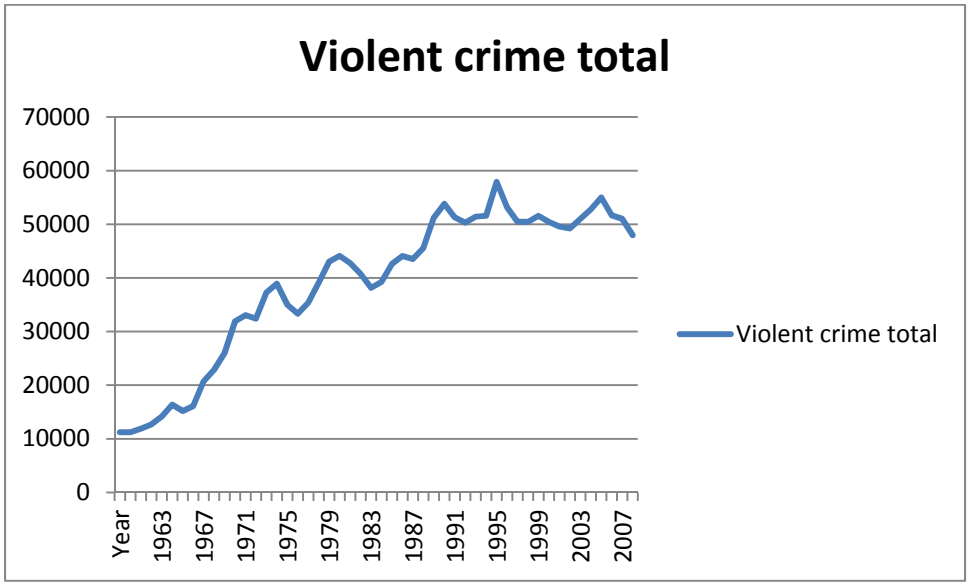
Note: this problem set is primarily intended to get you used to manipulating and presenting data using a spreadsheet program. While subsequent problem sets will be useful indicators of the difficulty of exam questions, this one is not.

Descriptive Statistics

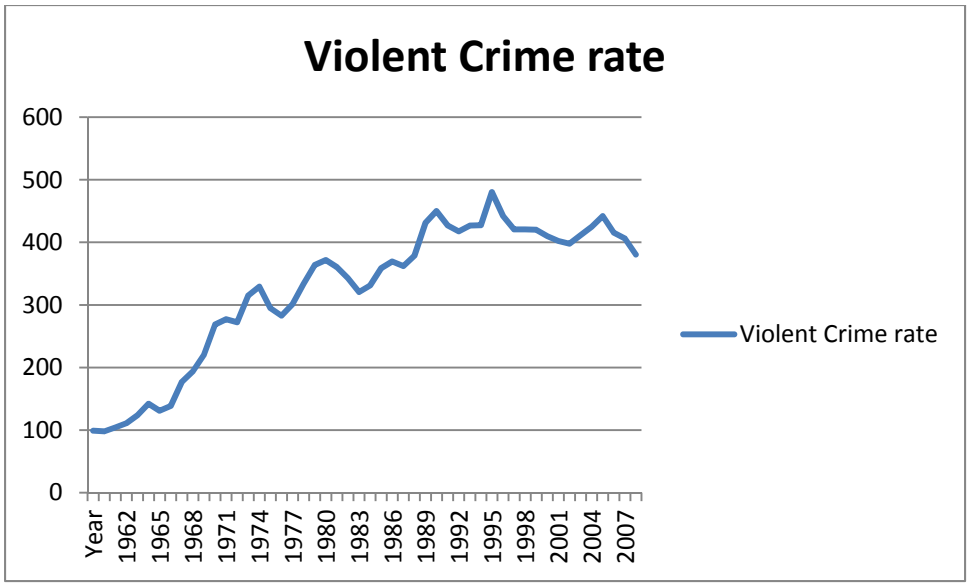
1. Go to the Bureau of Justice Statistics. Find the data for reported crime in Pennsylvania for 1960-2009. Put these data in a spreadsheet.

These data are available in convenient formats at
http://www.ucrdatatool.gov/Search/Crime/State/StateCrime.cfm

2. Graph the total violent crime by year.



3. Graph the total violent crime rate by year.



4. If you were trying to give someone a sense of the evolution of crime over time, is the graph from #2 or #3 more useful? Why?

It turns out that population growth in PA is more or less linear, so the total and rate series give, more or less, the same perspective (especially since inter-Census years are linearly interpolated for population). In terms of which is better, it depends on the questions of interest. If one is interested in getting some sense of how likely an individual is to be a victim of violent crime, rates are more useful. If, instead, someone was interested in budgetary issues, totals may be more useful.

5. Calculate the standard deviation of violent crime rates in PA over this period.

In excel, the standard deviation macro is "=STDEV.P(cell range)" [for the purposes of this class, ignore the difference between a standard deviation for a population and for a sample].

111.11

6. Calculate the standard deviation of homicide rates in PA over this period.

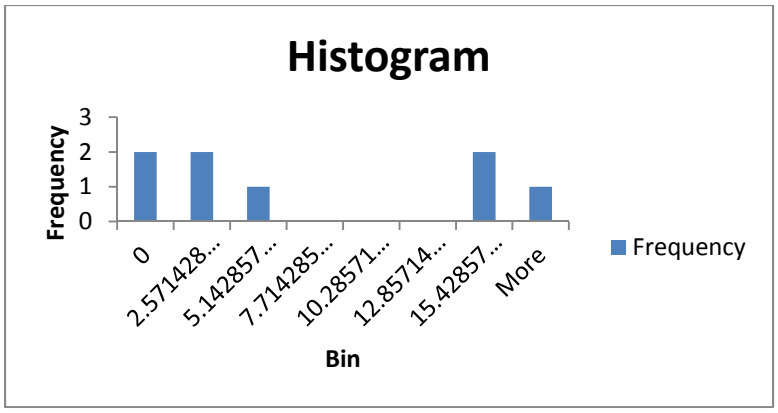
1.18

7. From the Bureau of Justice Statistics, pull crime totals and rates by all available categories of crime for the year 2009 into a spreadsheet.

http://www.ucrdatatool.gov/Search/Crime/State/StateCrime.cfm

8. Create a histogram for violent crime rates.

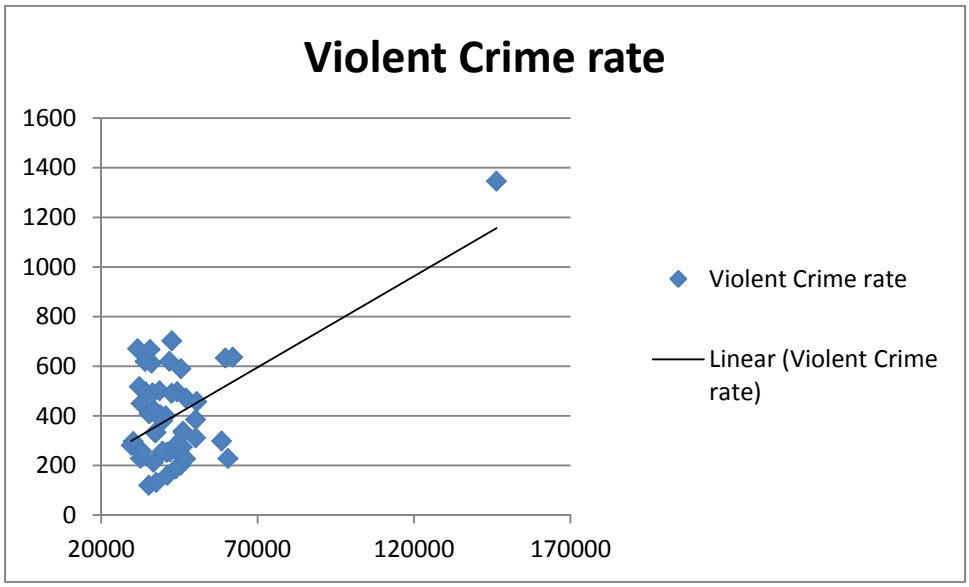
You need to load the "Analysis Tool Pack." To see how to do that, search in the help.



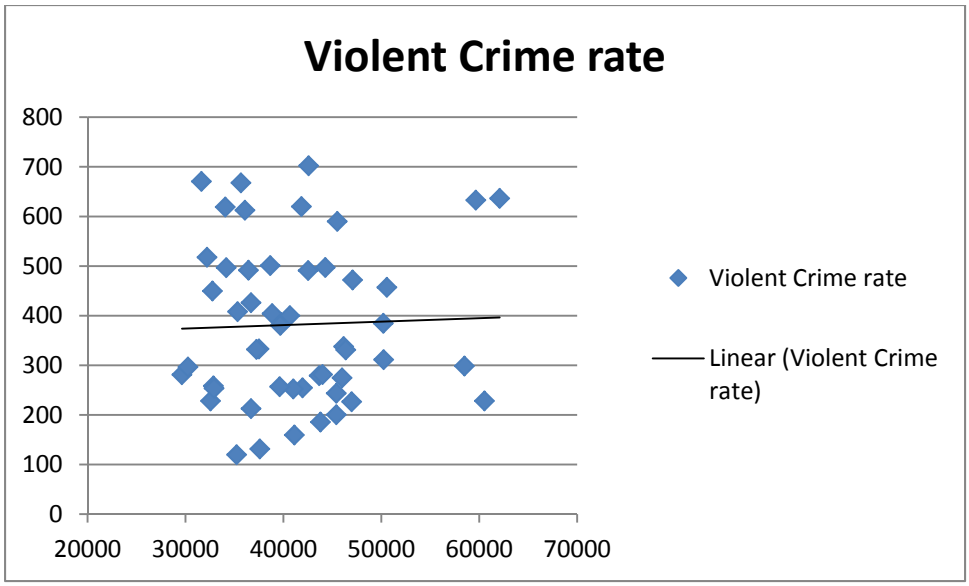
9. Get state per capita GDP data from the Bureau of Economic Analysis for 2009. Create a scatterplot of violent crime rates on the vertical axis and per capita GDP on the horizontal axis. What does the relationship appear to be? Does this relationship change if DC is eliminated from the graph?

Data available at: http://www.bea.gov/regional/gsp/

Scatter with DC



With DC Excluded


Violent Crime rate
800
700
600
500
400
300
200
100
0
20000
30000
40000
50000
60000
70000
Violent Crime rate
Linear (Violent Crime rate)

Problem Set 2
Klick

You are the plaintiff's lawyer in a torts suit where the defendant's negligent driving led to a head on crash that killed your client's 33 year old husband. There is no dispute over the negligence or causation. However, the defense claims the defense of contributory negligence because the deceased was not wearing a seatbelt. The defense lawyer cites data from the National Highway Traffic Safety Administration saying, "Fatality data show that in the age range 25-34, 66 percent of fatal accident victims do not wear a seatbelt. Because this demonstrates that it is more likely than not that the deceased would not have died had he been wearing a seatbelt, no liability is justified. In this same age group, the data are quite clear that most people, 70 percent in fact, do wear their seatbelt and in the average head on crash, only 10 percent of people die. This man would almost surely be alive today if he had worn his seatbelt." What is your response to this argument?

This is a Bayes Theorem problem (drawn verbatim from last semester's test).

Recall:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

In this case, the relevant question is what is the likelihood of death if a seatbelt is not worn.

$$P(death|noseatbelt) = \frac{P(noseatbelt|death) * P(death)}{P(noseatbelt)}$$

We know from the above that P(seatbelt) = 0.70 so the P(noseatbelt)=0.30 and P(death) = 0.10 so we just need to determine P(noseatbelt|death) to use Bayes Theorem. We're told that 66% of people in fatal accidents of this kind do not wear their seatbelt.

$$P(death|noseatbelt) = \frac{0.66*0.10}{0.30} = 0.22$$

While it is true that it is more likely than not that a person will survive this crash if he wears his seatbelt, it is also true that it is more likely than not that a person not wearing a seatbelt will also survive this crash. This suggests that the probability evidence alone is not enough to carry the contributory negligence claim.

In response, a sophisticated plaintiff might suggest that if one does the calculation

$$P(death|seatbelt) = \frac{P(seatbelt|death) * P(death)}{P(seatbelt)} = \frac{0.34*0.10}{0.70} = 5\%$$

So if we implement something like the Hand rule, the benefit of the precaution (i.e., seatbelt) = (0.22-0.05)*value of life which is likely to be a non-trivial amount relative to the minimal cost of wearing a seatbelt and therefore wearing the seatbelt is required.

Problem Set 3
Klick

As part of an investigation of discriminatory hiring practices and a potentially hostile work environment at a firm, you perform a survey of firm employees finding that out of 170 employees, 3 self-identify as being homosexual. Studies of the nationwide workforce suggest that about 5 percent of workers self-identify as being homosexual. The Supreme Court has suggested in the discrimination context, that outcomes that are so rare such that they would be expected to occur by random chance less than 5% of the time are suspect and likely the result of some non-random process (see Castaneda v. Partida 430 U.S. 482 (1977) at footnote 17).

1. Using only the information given above, is there reason to suspect that discrimination or a hostile work environment adversely affect the employment prospects of homosexuals at the firm in question?

This is a problem that can be addressed using the binomial distribution. In this case, if the firm were employing people randomly (with respect to sexual orientation), we can determine the likelihood of the firm ending up with 3 or fewer homosexuals out of a workforce of 170 employees. To do this, we make use of the binomial probability formula:

$$P(x) = \frac{n!}{(n-x)!x!} p^x (1-p)^{n-x}$$

To determine the likelihood of 3 or fewer, we need to figure out P(3)+P(2)+P(1)+P(0)

$$P(3) = \frac{170!}{(170-3)!3!} 0.05^3 (0.95)^{167} = \frac{170*169*168*167!}{167!3!} * 0.000125 * 0.00019 =$$

$$\frac{170*169*168}{3*2*1} * 0.0000000238 = 0.019$$

$$P(2) = \frac{170!}{168!2!} 0.05^2 (0.95)^{168} = 14365 * 0.0025 * 0.000181 = 0.0065$$

$$P(1) = \frac{170!}{169!1!} 0.05^1 (0.95)^{169} = 170 * 0.05 * 0.00017 = 0.0015$$

$$P(0) = \frac{170!}{170!0!} 0.05^0 (0.95)^{170} = 1*1*0.00016 = 0.00016$$

$$P(x \le 3) = 1.9\% + 0.65\% + 0.15\% + 0.016\% = 2.7\%$$

If this firm's hiring and employment practices were unrelated to sexual orientation of its employees/applicants, the likelihood that this firm of 170 people would have 3 or fewer homosexual employees, based on a labor market wide rate of homosexuality of 5%, is about 2.7%. Given that this falls below the threshold the Supreme Court suggests, there is a decent argument to be made that this firm is engaging in practices that have a disparate impact on homosexual employees/applicants.

2. What criticism(s) might the defendant firm offer in response to the analysis in question 1?

You could argue that some of the firm's employees who are homosexual simply chose not to self-identify leading one to overstate the unlikeliness of observing the firm's particular employee pool. This is tricky though for at least two reasons: 1) if employees are hesitant to self-identify, this could provide ancillary evidence of the hostile work environment and 2) it seems likely that some people will have hidden their homosexuality in the nationwide studies as well suggesting that p should have been higher leading these two effects to cancel each other out to some extent.

More sensibly, you might argue that the nationwide studies are not representative of the relevant applicant pool which may be different than the nationwide labor force. If p in the local labor pool is lower than 5%, the calculation above will be under-stated.

Problem Set 4
Klick

Assume you are investigating a market timing case as discussed in class. However, in this case, assume that the total number of across market trades made by a suspect customer is 1,000. Retain the assumption that market outcomes are a random walk such that the likelihood that a trade on any given day will be "successful" by random chance is 50%. Determine the likelihood that this customer would have at least x successful trades out of the 1,000 by mere coincidence where x =:

1) 518
2) 525
3) 540
4) 565
5) 600

For this problem, you need to recognize that this is a binomial distribution AND that 1,000 trials is surely enough for the normal distribution to serve as a reasonable approximation. To analyze this problem then, we need to determine how far away each x is from the mean expectation normalized in standard deviation terms. For this, we need to know the mean expectation (which is n*p = 500) and the standard deviation (which for the binomial is $\sqrt{(n*p*[1-p])} = \sqrt{(1000*.5*.5)} = 15.8$).

Thus, the normalized scores are

1) **(518-500)/15.8=1.14**
2) **(525-500)/15.8=1.58**
3) **(540-500)/15.8=2.53**
4) **(565-500)/15.8=4.11**
5) **(600-500)/15.8=6.33**

We now need to consult a table of z scores for the normal distribution wherein we find the probability of having a standardized score of that listed above or greater is

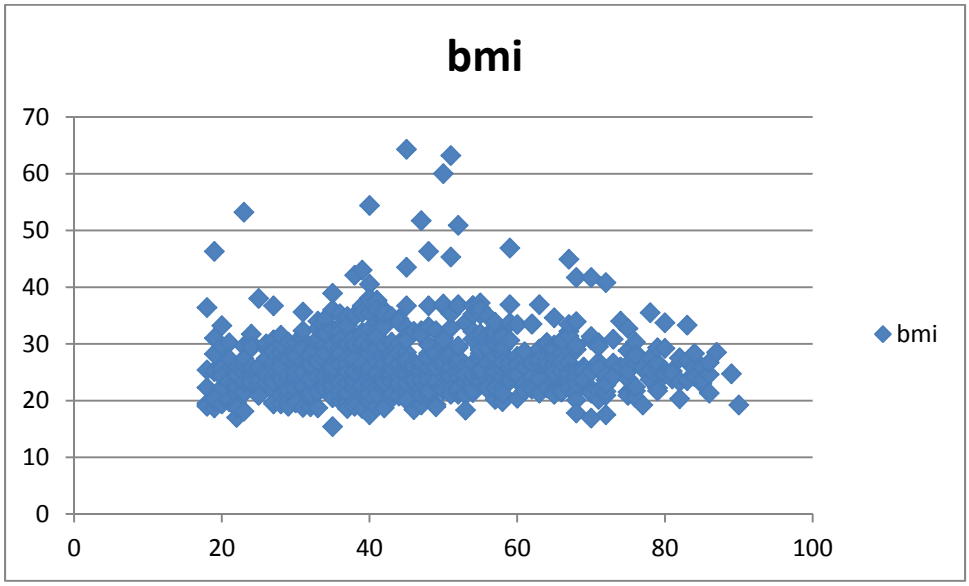
1) **1-0.8729 = 12.7%**
2) **1-0.9429 = 5.7%**
3) **1-0.9943 = 0.5%**
4) **1 – 0.9999 <0.01% (actually 0.002%)**
5) **1-0.9999 <0.01% (actually 0.0000000099%)**

Problem Set 5

Use the dataset PS5.csv for this problem set. It contains individual level data on the characteristics of 716 Pennsylvania residents in the year 2000. The dataset includes:

BMI: Body Mass Index
Income: Measured categorically (higher value = higher income category)
Age: Measured in years
Female: Dummy variable taking value of 1 if individual is a woman
Educa: Education measured categorically (higher value = more education received)
Smoke: Dummy variable taking value of 1 if individual is a smoker
Married: Dummy variable taking the value of 1 if individual is married
Sepdiv: Dummy variable taking value of 1 if individual is separated or divorced
Children: Number of children the individual has
Unemployed: Dummy variable taking value of 1 if individual is unemployed
Race: =1 if individual is white; =2 if individual is black

1. Create a scatterplot of bmi on age. Do you see any relationship?



There doesn't seem to be much of a relationship.

2. Regress bmi on age. Interpret the coefficient and determine the smallest type 1 error level at which the age effect would be statistically significant (relative to a hypothesis of 0 effect).

Using the regression function in the data analysis tool pack, pointing to the bmi column for your y range and the age column as your x range, you get:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.034601078
R Square	0.001197235
Adjusted R Square	-0.000201649
Standard Error	5.84290756
Observations	716

ANOVA

	df	*SS*	*MS*	*F*	*Significance*
Regression	1	29.21835472	29.21835	0.855850141	0.35
Residual	714	24375.65209	34.13957		
Total	715	24404.87045			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	26.22348456	0.666884823	39.32236	7.8278E-181	24.
age	0.012352325	0.013352108	0.925122	0.355215173	-0.01

This suggests that as you get 1 year older, on average, your bmi increases by 0.01 points. We can look to the pvalue to answer the type 1 error question (we would reject 0 effect at any type 1 error larger than 36%).

3. Regress bmi on age as a quadratic function. Interpret the age coefficients. Is the age effect statistically significant at a 5% type 1 error?

To include age as a quadratic, we need to create a variable for age squared. To do this, add a column next to age and type in the first data cell (i.e., c2) "=b2*b2" then copy that throughout the column.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.159451322
R Square	0.025424724
Adjusted R Square	0.022690993
Standard Error	5.775654198
Observations	716

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	620.4870978	310.2435	9.300373574	0.000102964
Residual	713	23784.38335	33.35818		
Total	715	24404.87045			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	19.45357464	1.737897165	11.19374	6.59334E-27	16.04156685	22.86558
age	0.314511537	0.072973759	4.309926	1.86267E-05	0.171242395	0.457781
agesq	-0.003002402	0.000713145	-4.21009	2.87933E-05	-0.004402516	-0.0016

This suggests that there is a non-linear effect of age. Specifically, we can see that as age increases, bmi goes up, but at some point the effect becomes negative. Precisely, as age increases by 1 year, the effect on bmi is 0.3145 + (2*-.003002)*age = 0.3145-0.006age implying that the shift from a positive effect to a negative effect occurs around age 52 (don't worry if you don't know how to get this precisely since it requires some calculus). From the p values, we can see this age effect is statistically significant.

4. Regress bmi on married allowing for women to have a separate intercept. Interpret the female coefficient.

Change your x range to include female and married

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.136648604
R Square	0.018672841
Adjusted R Square	0.01592017
Standard Error	5.795626607
Observations	716

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	455.7082646	227.8541	6.783536	0.001207
Residual	713	23949.16218	33.58929		
Total	715	24404.87045			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*
Intercept	27.35245831	0.431017056	63.46027	1.6E-295	26.50624	28.19867	26.50624
female	-1.492747901	0.439940442	-3.39307	0.000729	-2.35648	-0.62901	-2.35648
married	0.556106141	0.440154274	1.263435	0.206846	-0.30805	1.42026	-0.30805

The interpretation of the female coefficient is that women, on average, have bmi's that are 1.49 points lower than men or you could say that the baseline female bmi is 25.86 (27.35-1.49). The female effect is statistically significant.

5. Regress bmi on married allowing women to have a seprate intercept and a differential slope with respect to the married variable. What is the average effect on bmi of being a married woman (relative to an unmarried man)? What is the average effect on bmi of being a married woman (relative to a married man)?

To do this, we need to create a female X married interaction and then include female, married, and the interaction in the x range

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.137886626
R Square	0.019012722
Adjusted R Square	0.014879348
Standard Error	5.798690699
Observations	716

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	464.0030092	154.6677	4.599807	0.00338
Residual	712	23940.86744	33.62481		
Total	715	24404.87045			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Up 95*
Intercept	27.19043478	0.540730696	50.28461	1.7E-236	26.12882	28.25205	26.12882	28.
female	-1.228346871	0.690754232	-1.77827	0.075786	-2.58451	0.127812	-2.58451	0.1
married	0.819947731	0.69002307	1.18829	0.235115	-0.53478	2.174671	-0.53478	2.1
femalemar	-0.445171236	0.896304405	-0.49667	0.619572	-2.20489	1.314544	-2.20489	1.3

The effect of being a married woman relative to an unmarried man (which is represented by the intercept given that we control for female and married, therefore the reference group is neither female

nor married i.e., an unmarried man) is a 1.23 point reduction for being a woman, a .82 increase for being married, and a .45 reduction for being a married woman, so the total effect of being a married woman is .82-1.23-.45 = -.86 points lower than an unmarried man.

To determine the effect of being a married woman relative to a married man, we need to change the reference baseline. A married man starts off with the intercept in bmi, 27.19 and has a .82 increase for being married, leading to a comparison baseline of 28.01. A married woman is 1.23 points less than the married man because she is a woman and her marriage effect is .45 lower, so the married woman is 1.68 points lower than a married man.

6. Controlling for all of the variables (income, age as a quadratic, female, educa, smoke, married, sepdiv, children, and unemployed) in a bmi regression, is the effect of being black positive, negative, or zero? Is it statistically significant at a 1 percent type 1 error?

We need to create a black dummy variable from the race variable (dummy =1 if race =2). You can figure out how to do if then statements in excel or, in this case, you could simply subtract 1 from the race variable, leaving whites =0 and blacks=1.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.326472906
R Square	0.106584558
Adjusted R Square	0.092624942
Standard Error	5.565172763
Observations	716

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	11	2601.182335	236.4711	7.635206877	1.58E-12
Residual	704	21803.68811	30.97115		
Total	715	24404.87045			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*
Intercept	23.93773368	2.046752681	11.69547	5.31944E-29	19.91926	27.9562	19.91926
income	-0.243299231	0.129827223	-1.87402	0.061341191	-0.49819	0.011596	-0.49819
age	0.38158626	0.078375474	4.868695	1.38791E-06	0.227709	0.535464	0.227709
agesq	-0.003814366	0.000770566	-4.95008	9.29211E-07	-0.00533	-0.0023	-0.00533
educa	-0.647899295	0.228899295	-3.05739	0.002317247	-1.14924	-0.25043	-1.14924

	0.699835545						
	-						
smoke	0.377767638	0.42512079	-0.88861	0.374514894	-1.21242	0.456889	-1.21242
	-						
female	1.743098388	0.429079962	-4.06241	5.40313E-05	-2.58553	-0.90067	-2.58553
married	0.308198176	0.569958961	0.540737	0.588859577	-0.81082	1.427221	-0.81082
	-						
sepdiv	1.192934185	0.743725399	-1.604	0.109162954	-2.65312	0.267251	-2.65312
	-						
children	0.149181026	0.221579639	-0.67326	0.501001935	-0.58422	0.285855	-0.58422
unemployed	0.988948923	1.132687189	0.8731	0.382906401	-1.2349	3.212798	-1.2349
black	3.134220839	0.74571253	4.202988	2.97339E-05	1.670134	4.598308	1.670134

We can see that the effect of being black increases BMI and the effect is statistically significant at the 1 percent level (as seen from the pvalue).

Problem Set 6
Klick

The provided dataset comes from http://www.economics.harvard.edu/faculty/shleifer/dataset and relate to Rafael LaPorta, Florencio Lopez-de-Silanes, and Andrei Shleifer, "The Economic Consequences of Legal Origins," Journal of Economic Literature, 46(2): 285-332(2008) which is available at http://www.economics.harvard.edu/faculty/shleifer/files/consequences_JEL_final.pdf .

1. In Table 4 of the paper, they argue that their legal origin effects on creditor rights are robust to the inclusion of various cultural control variables. Assess this claim by doing your own robustness checks using the alternate controls they considered. Provide some indication of how stable the legal origins variables are.

In order to perform this analysis, you need to run various permutations of the regression provided in Table 4. In excel, you would need to do this in a fairly ad hoc way, but in fancier stats programs, this could be automated to run every possible combination of control variables.

Here are (more or less) the results from the original Table 4 (there are some minor differences because the authors updated the dataset after publication:

	(1)	(2)	(3)	(4)	(5)
Dependent Variable: Creditors' Rights in 2002					
% catholic	-0.25 (0.22)				
Power Distance Index		-0.00 (0.01)			
Individualism			-0.01 (0.01)		
Uncertainty Avoidance Index				-0.01 (0.01)	
Masculinity					-0.02* (0.01)
French Legal Origin	-0.76*** (0.24)	-0.84** (0.35)	-0.92** (0.34)	-0.48 (0.40)	-1.02*** (0.37)
German Legal Origin	-0.25 (0.27)	-0.52 (0.45)	-0.56 (0.43)	-0.25 (0.45)	-0.28 (0.43)
Scandinavian Legal Origin	-1.10** (0.48)	-0.87 (0.58)	-0.87 (0.58)	-0.92 (0.55)	-1.69** (0.70)
Log(GDP per capita in 2002)	0.28*** (0.08)	0.19 (0.24)	0.32 (0.21)	0.22 (0.19)	0.20 (0.20)
Constant	0.00 (0.73)	0.65 (2.50)	-0.27 (1.80)	0.74 (1.65)	1.70 (1.88)
Observations	136	52	52	52	52
R-squared	20%	14%	15%	16%	19%

Note: Data come from <http://www.economics.harvard.edu/faculty/shleifer/files/JEL_%20web.xls>.
Heteroskedasticity robust standard errors presented in parentheses.
*** $p < 0.01$ (two sided test of zero null hypothesis)
** $p < 0.05$ (two sided test of zero null hypothesis)
* $p < 0.10$ (two sided test of zero null hypothesis)

If we include all of those first 5 culture variables, we already see some important changes in the legal origins coefficients:

Dependent Variable: Creditors' Rights in 2002	
Catholic	-0.79** (0.36)
Power Distance Index	-0.00 (0.01)
Individualism	-0.00 (0.01)
Uncertainty Avoidance Index	-0.01* (0.01)
Masculinity	-0.02 (0.01)
French Legal Origin	-0.00 (0.39)
German Legal Origin	0.19 (0.39)
Scandinavian Legal Origin	-1.86*** (0.68)
Log(GDP per capita in 2002)	0.33 (0.22)
Constant	1.10 (2.44)
Observations	52
R-squared	29%
Note: Data come from http://www.economics.harvard.edu/faculty/shleifer/files/JEL_%20web.xls. Heteroskedasticity robust standard errors presented in parentheses. *** p < 0.01 (two sided test of zero null hypothesis) ** p < 0.05 (two sided test of zero null hypothesis) * p < 0.10 (two sided test of zero null hypothesis)	

Specifically, the effect of French legal origin goes to zero (and is not statistically significant), the German legal origins indicator coefficient flips signs, and the Scandinavian one becomes much larger in magnitude.

If we run every permutation of the regression, always including the income and origins effects, and then allowing for all the various subsets of the cultural indicators (including the others tried in the original article), we find the following:

	Baseline	Mean	Max	Min	SE
French Legal Origin	-0.85***	-0.05	1.12	-1.04	0.59
German Legal Origin	-0.30	0.07	1.14	-0.82	0.70
Scandinavian Legal Origin	-1.03**	-1.63	-0.56	-2.57	0.82
Log(GDP per capita in 2002)	0.25***	-0.04	0.41	-0.36	0.31
Catholic		-0.86	-0.25	-1.25	0.49
Power Distance Index		-0.02	0.00	-0.03	0.01
Individualism		0.00	0.01	-0.01	0.01
Uncertainty Avoidance Index		-0.01	0.00	-0.02	0.01
Masculinity		-0.02	-0.02	-0.03	0.01

% agree child obedience is important		0.20	1.68	-1.80	1.85
% agree child independence is important		1.93	2.70	1.10	1.16
% agree parents must do their best for children		0.10	2.78	-2.74	2.76
% agree that parents must be respected regardless		-0.93	0.43	-2.17	1.34
% agree family life is very important		1.90	3.47	0.27	2.77
% agree strangers can generally be trusted		0.12	1.97	-1.29	1.47

This suggests that only the Scandinavian legal origin effect is stable in terms of the sign it generates. If we go a bit further, we find that for both French and German, we are as likely to find a positive as a negative sign, and the effects are generally not statistically significant:

	%(Significant)	**%(Positive)**	**%(Negative)**	**Average T**	**Number of Specifications**
French Legal Origin	0.05	0.45	0.55	0.75	2048
German Legal Origin	0	0.50	0.50	0.50	2048
Scandinavian Legal Origin	0.51	0	1	1.96	2048
Log(GDP per capita in 2002)	0.00	0.40	0.60	0.45	2048
Catholic	0.24	0	1	1.771	1024
Power Distance Index	0.21	0.00	1.00	1.67	1024
Individualism	0	0.63	0.37	0.29	1024
Uncertainty Avoidance Index	0	0.00	1.00	0.91	1024
Masculinity	0.05	0	1	1.69	1024
% agree child obedience is important	0	0.61	0.39	0.33	1024
% agree child independence is important	0.23	1	0	1.69	1024
% agree parents must do their best for children	0	0.50	0.50	0.50	1024
% agree that parents must be respected regardless	0.00	0.05	0.95	0.74	1024
% agree family life is very important	0	1	0	0.68	1024
% agree	0	0.55	0.45	0.36	1024

	%(Significant)	%(Positive)	%(Negative)	Average T	Number of Specifications
strangers can generally be trusted					

2. In what sense is robustness a useful criterion for assessing an empirical finding?

This kind of robustness checking can show us pretty conclusively that there is an omitted variable bias at least in some of the specifications; since we can't know which are free of the bias, we can confidently conclude that the results are unreliable. However, if we had found stability, that would not give us any confidence that there was no bias. It may just be the case that we did not examine the effects of any variables related to the important omitted ones. Robustness tests are asymmetric. They can cause us to lose confidence, but they do not ever provide confidence.

Problem Set 7

For background see Jonathan Klick and Robert Sitkoff, "Agency Costs, Charitable Trusts, and Corporate Control: Evidence from Hershey's Kiss Off," Columbia Law Review 108(4): 749-838 available at http://www.law.upenn.edu/fac/jklick/108CLR749.pdf .

The provided dataset includes HSY which is the closing price of the class A shares of the Hershey company adjusted for stock splits and dividends, as well as similar data for three of Hershey's competitors in the US chocolate market (CSG, RMCF, and TR). It also includes three proxies for the daily market return (CRSP which is the value weighted return for the Center for Research on Securities Prices which measures a mean return for all securities in the CRSP database weighted by market capitalization, SP500 which is the daily return for the Standard and Poor's 500, and DJIA which is the daily return for the Dow Jones Industrial Average).

On July 25, 2002 (before the market opened for the day), the trustees who manage the Milton Hershey School Trust, whose portfolio contains all of Hershey's Class B shares which controls 80 percent of the votes in the Hershey company, asked the company to find a buyer for the trust's controlling interest. On September 18, 2002, the trust abandoned the sale due to pressure from the state Attorney General's office (which didn't want the sale to occur for political reasons).

1. Standard practice in applied finance performs an event study by estimating a market model (where firm returns are taken to be a linear function of a market index) in some period before the event of interest (generally using 100 trading days). This estimated model is then used to predict the return on the event day. Abnormal returns are then calculated as the difference between the observed return for a given day and the predicted return for that day. For the test statistic in the event study, this approach takes the abnormal return for the event day and standardizes it by dividing by the standard deviation of abnormal returns during the estimation period. Use this "standard" approach to determine whether the movement on the day the sale was announced was statistically significant.

The first thing we need to do is to calculate the return of HSY based on the prices. A return is a rate of change, so the return is equal to (P(t)-P(t-1))/P(t-1) the change in price between today and yesterday divided by yesterday's price. We then need to regress this on the market index (which is already in return terms). If we do this (using the 100 days before July 25, 2002 and using CRSP as the market), we get:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.359096252
R Square	0.128950118
Adjusted R Square	0.120061854
Standard Error	0.012429018
Observations	100

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0.002241	0.002241	14.50791	0.000244
Residual	98	0.015139	0.000154		
Total	99	0.01738			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Up 95.*
Intercept	-0.00033244	0.001263	-0.26313	0.793003	-0.00284	0.002175	-0.00284	0.00
X Variable 1	0.335238197	0.088014	3.808925	0.000244	0.160578	0.509899	0.160578	0.50

This is our market model then. This tells us that on July 25, 2002, when the CRSP return was -0.0057, we should have expected a HSY return of -0.00033244 +0.335238197*(-0.0057) = -0.002243. Instead, we observe an actual HSY return on that day of 0.2528. Thus, our abnormal return is 0.2528 – (-0.002243) = 0.255. To standardize this, we need a measure of the volatility of abnormal returns. To do this, we need to calculate the abnormal return for each of the 100 days before July 25, 2002 based on the model above and then calculate the standard deviation of that series. Using =stdev.p as the standard deviation macro in excel, we find a standard deviation of abnormal returns for those 100 days equal to 0.0123. Standardizing the abnormal return for July 25, 2002, we have 0.255/.0123 = 20.73 which is statistically significant at the 5% (and the 1%or even the 0.1%) level.

2. Re-do #1 using the one step dummy variable regression approach and confirm they lead to the same outcome. Intuitively explain why the two methods are the same.

To do this, we need to create a dummy variable for the event which takes the value of 0 for all days but July 25, 2002 which takes the value of 1. Doing that yields:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.902045029
R Square	0.813685235
Adjusted R Square	0.809882893
Standard Error	0.012429018
Observations	101

ANOVA

	df	*SS*	*MS*	*F*	*Significance*

					F
Regression	2	0.066116345	0.033058	213.9958	1.75E-36
Residual	98	0.015139087	0.000154		
Total	100	0.081255432			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*
Intercept	-0.00033244	0.001263405	-0.26313	0.793003	-0.00284	0.002175	-0.0028
X Variable 1	0.335238197	0.088013866	3.808925	0.000244	0.160578	0.509899	0.16057
X Variable 2	0.255043767	0.012494037	20.41324	4.51E-37	0.23025	0.279838	0.2302

We see that the X variable 2 coefficient is exactly the abnormal return we calculated for the event date. The standard error is very close (the differences arise because of choices regarding how to calculate the standard deviation but will be slight in any case). From this approach, we can see that the event effect is statistically significant at a type 1 error level of 0.00000000000000000000000000000000000000451.

As for intuition, the cookbook approach calculates a predicted value and then nets it off of the true value for the event day which is equivalent to the residual for the event day if you had included it in the model estimation. In the regression approach, the dummy variable takes the value of the residual; thus the two approaches are effectively the same.

3. Note that between July 25, 2002 and September 18, 2002, there were multiple court actions regarding the proposed sale. Given that, should these days be used to estimate the effect of abandonment of the sale? Why or why not?

No. For your model, you want to use data that is informative for the question of what would the return have been on this day given the market return if nothing else was happening. The court decisions are clearly something else happening. To estimate the abandonment effect, one would probably want to use pre-sale announcement data.

4. Is there any basis for preferring any of the market indexes available in the dataset for your analysis? Re-do the sale announcement event study with each of the indexes and observe whether your results are robust to the choice of market index.

Theoretically, the broader index is probably better since most theoretical models (e.g., the CAPM) use the market of all possible assets as their market value. Clearly the DJIA (30 stocks) is not the whole market. However, in practice, the choice doesn't matter much.