Statistics for Lawyers

Problem Set

1. Compare the expected return and the variance of the following three portfolios:

Portfolio 1: 100% of assets are invested in stock A
Portfolio 2: 50% of assets are invested in stock A and 50% of assets are invested in stock B
Portfolio 3: 10% of assets are invested in each of stocks A, B, C, D, E, F, G, H, I, and J

Assume the following: i) each stock A through J has an expected return of 10%
ii) the returns of stocks A through J are all statistically independent
iii) the standard deviation of each stock A through J's return is σ

Expected return is simply the share of the portfolio in a stock multiplied by the stock's expected return. Thus, the return for Portfolio 1 is 100% * 10% = 10%; Portfolio 2 is 50%*10% + 50% * 10% = 10%; and Portfolio 3 is 10%*10% + 10%*10% . . . = 10*10%*10%= 10%.

Each portfolio has the same expected return.

For variance, for portfolio 1, since it contains just stock A, the variance of the portfolio will be the variance of the stock which is $\sigma * \sigma = \sigma^2$

For portfolio 2, recall two things: the variance of the sum of two independent random variables is just the sum of the individual variances and recall that a constant times a random variable has a variance of the constant squared times the variance of the random variable. Since portfolio 2 is composed of .5*stock A and .5*stock B, the portfolio variance will be the sum of variance of the two subportfolios. The variance of each subportfolio will be $(.5*.5)*\sigma^2$ **so the sum of the subportfolio variances will be** $(.5*.5)*\sigma^2 + (.5*.5)*\sigma^2 = 2*.25*\sigma^2 = 0.5\,\sigma^2$

So the variance of portfolio 2 is half as large as the variance of portfolio 1.

Portfolio 3 involves the sum of the variances of 10 subportfolios each of which has an individual variance of $(.1*.1)*\sigma^2 = .01\,\sigma^2$ **which when added with the other 9 subportfolios =** $10*0.01\,\sigma^2 = 0.1\,\sigma^2$

This implies that portfolio three, which has the same expected value as portfolio 1 has a variance that is 1/10 as large. This is the idea of diversification.

2. The plaintiff's lawyer in a wrongful death case against a defendant who crashed his car into the deceased defendant argues the following: according to 2013 FARS data, 40% of all fatal crashes occurring on a weekend night involved drivers whose blood alcohol content (BAC) equaled or exceeded 0.08, just as the defendant's did on the Saturday night when the crash occurred, even though government statistics suggest that just 1 in 12 weekend night drivers has a BAC greater than or equal to 0.08. This, the plaintiff lawyer contends, demonstrates that it is more likely than not that had the defendant not been drunk, the plaintiff would still be alive. The defense lawyer argues that since more fatal crashes involve drivers who are not drunk than drivers who

are drunk, such a claim is not justified. The evidence rules in the state allow the judge to hire an impartial expert so the judge hires you to use the data contained in the plaintiff's argument to determine whether it is more likely than not that the defendant's drinking led to the plaintiff's death. There is no other evidence in the case beyond the fact that the defendant's BAC was above 0.08, the defendant crashed into the plaintiff's car, there is no claim of contributory negligence, and the nighttime driving conditions were not abnormal. Assume that the total odds of dying in a car accident on a weekend night is 15 out of 1 million.

In this contrived problem, we are ignoring all sorts of stuff (what kinds of cars are involved, what are the ages of the parties, what are the specific conditions, how far above 0.08 is the BAC, etc) that might be relevant in a real case. With that caveat in mind, one might think of this as a Bayes Theorem Problem. You might want to compare the probability of a fatal accident given that the driver was drinking with the probability of a fatal accident given that the driver was not drinking.

By Bayes Theorem:

$$P(fatality \mid drinking) = \frac{P(drinking \mid fatalaccident) * P(fatalaccident)}{P(drinking)} = \frac{0.4 * 0.000015}{1/12} = 0.0072\%$$

$$P(fatality \mid notdrinking) = \frac{P(notdrinking \mid fatalaccident) * P(fatalaccident)}{P(notdrinking)} = \frac{0.6 * 0.000015}{11/12} = 0.001\%$$

Without any other evidence, it would seem as though drinking increases the risk of death by a factor of 7. There are all sorts of causality concerns here (maybe those who drink and drive just tend to be bad/risky drivers in general, etc), so it would be very hard to simply say the evidence demonstrates that the drinking led to the plaintiff's death. Putting those causality concerns aside, there are still some difficulties here. The probability of a death occurring even when drinking is still very low, even though it is 7 times higher than the probability of death occurring when not drinking. These probabilities may be more probative with respect to negligence determination (e.g., multiply the change in probability .006% times some value of life and then compare that to the cost of refraining from drinking), but they are not very useful (strictly speaking) in demonstrating causation. More evidence would likely be needed.

3. If individual jurors each have a 50% probability of voting guilty, and there is no discussion among the jurors, and there is secret ballot voting, what is the probability of a 12 person jury:
    a. Reaching a unanimous guilty verdict
    b. Reaching a majority guilty verdict
    c. Being evenly split between guilty and not guilty determinations

This problem involves the binomial distribution (guilty/not guilty). Although you could do it by hand, online calculators (such as this one http://stattrek.com/online-calculator/binomial.aspx) are handy.

a. P(X=12; n=12) = 0.000244 = 0.0244%
b. P(X>6; n=12) = 0.387 = 38.7%
c. P(x=6; n=12) = 0.2256 = 22.56%

4. Professor Klick just bought $10 worth of Powerball tickets (each play costs $2). The advertised jackpot is $235 million. Klick had the machine randomly pick his numbers and none of his 5 plays duplicates the set of numbers on any of his other plays. The Powerball number is determined by drawing a number between 1 and 59 from each of 5 separate/independent containers and then choosing a powerball separately from a container including powerballs labeled 1-35. For the first 5 numbers, order does not matter; the winner must simply match all 5 numbers but s/he must also separately match the powerball (i.e., there is a separate powerball number listed for each of the gambler's plays). What is the expected value of Klick's investment in the Powerball assuming that only one player will win the jackpot and assuming that there are no non-jackpot prizes. How would you go about determining the probability that the assumption that only a single person hits if Klick hits is a valid assumption (i.e., what other info would you need and how would you use that info)?

To determine the probability of matching the powerball with a single number, you calculate the number of ways to choose 5 different numbers from a set of 59. Here to you could simply use your calculator or an online calculator. It is a combination problem since order doesn't matter. nCr = $\frac{59!}{(59-5)!5!} = 5{,}006{,}386$ **this represents the number of possible plays for the first set of numbers drawn, so the probability of any randomly chosen set of 5 numbers matching is 1/5006386 but to win the jackpot, you also need to match the powerball and the chance of doing that is 1/35, so the probability of matching the five numbers plus the powerball is the product of these or 0.0000000057 or 1 chance in about 175 million.**

However, Professor Klick played 5 sets of numbers so his probability of winning is 5*0.0000000057. To find the expected value of his investment, you multiply this probability times the jackpot of $235,000,000 to get about $6.70. Since he spent $10, his expected value is -$3.30.

To think about the validity of the assumption that, if he wins, he wins by himself, you need to know how many numbers are likely to be played and you need to assume that any given number is just as likely to be played as any other number. If you think that n plays will be made, then the probability that none matches Klick's winning number is (1-0.0000000057)^n= 0.9999999943^n. So, for example, if there were 100 million sets of numbers played (and each was chosen randomly), the probability that Klick wins alone if he wins is around 50%. This probability should be included in the expected value calculation as well.

5. Use excel (or calculate by hand) the covariance and correlation coefficient for the variables income and education using the data below:

Observation	Income	Education
1	48312	12
2	34149	7
3	33553	8
4	82619	15
5	62355	10
6	32277	3

7	33089	5
8	42148	11
9	39930	5
10	51706	15

Covariance (using covariance.P in excel)= 46393.42
Correlation (using correl in excel)= 0.759464

6. Bonus: Figure out how to run a regression of income on education using excel. Hint: you will need to go to your options menu then your add-ins menu, and then you need to add the analysis tool pack). What is the interpretation of the parameters you estimated?

Problem Set #2
Regression Exercises

Use the excel sheet PS2 for this problem set.

Income is calculated as 5000+(2500*education)+(750*IQ points above 50)-(500*female)+(500*age)-(6*age*age)+ a random error (which takes a value between -10,000 and +10,000)

1) Run the "correct" regression and qualitatively interpret the output (for this step, use just the IQ variable to control for IQ i.e., don't create an IQ points above 50 variable).

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-30145.62154	3012.698617	-10.00618561	1.38123E-21
iq	736.1444025	9.710924788	75.80579797	3.1896E-274
education	2537.483177	52.23693113	48.57642135	2.8962E-190
female	-130.406281	492.2508387	-0.264918352	0.791182861
age	404.0086778	132.5391724	3.048220917	0.002425379
age2	-4.696905572	1.419601298	-3.308608959	0.001006032

Everything is statistically significant at better than the 1% type 1 error level (as shown by a p value<0.01) except the female control.

The coefficients imply that the baseline income rate is -$30,146, each additional IQ point increases income by $736, each additional year of education increases income by $2537, women make $130 less than men, and the effect of age is nonlinear (income goes up with age but at some point starts to come down).

2) Discuss/provide intuition for any departures between your estimates and the "true" coefficients revealed above.

The big differences are the female coefficient and the constant. For female, the effect of being a woman is presumably pretty small relative to the error term in the model, and, what's more, there are relatively few women in the dataset, so there is lots of noise and perhaps too few observations for LLN/CLT to have kicked in.

The constant term is more puzzling (not only is it not anywhere close to the $5000 in the true model, but it is large and negative). This makes more sense when you recognize that I had you control for IQ, but the true model was a function of IQ-50. This means for low IQ people (given the large positive effect of IQ), there essentially must be a correction in the baseline income.

3) For any departures between the truth and the estimates, predict what would happen if you had 10X as many data points (as calculated above). Provide your intuition.

Presumably, with more data, all of the coefficients will be estimated with less noise. Further, the female coefficient will likely start to converge to the true coefficient (since you'll have a couple thousand female observations) so it is likely that the female coefficient will get closer to 500 and will be statistically significant since the signal (which is still small relative to the noise) will be estimated more precisely. The constant issue, however, is not one of too little data, so it is unlikely to be fixed with more observations.

4) Verify your intuition by simply copying the 500 observations and pasting them 9 more times and re-running the regression.

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-30145.62154	947.5341985	-31.81481111	2.0159E-202
iq	736.1444025	3.054216337	241.0256253	0
education	2537.483177	16.42921678	154.4494306	0
female	-130.406281	154.8195034	-0.842311712	0.399653832
age	404.0086778	41.68535074	9.691862263	5.09787E-22
age2	-4.696905572	0.446483684	-10.51976979	1.29569E-25

The female intuition is not borne out entirely. While the estimate did get more precise, there was no convergence to -500. This is not a failure of LLN/CLT, rather since you don't actually have "real" additional datapoints (just a repetition of the original ones), there isn't the normal balancing out of large and small errors that you would expect to get with 5000 independent datapoints.

5) Using the original 500 data points, calculate the correlation coefficient between education and IQ, as well as IQ and the other control variables. Based on these calculations, make a prediction of what will happen to your coefficients if you re-run the regression (on the original 500 data points) leaving IQ out of the control variables. Provide the intuition for your prediction. Re-run the regression to verify your intuition.

Correlation between IQ and education = 0.481, IQ and female = 0.024, IQ and age = 0.005 (for completeness, the education and female and age correlations are all zero too).

Only IQ and education are correlated; I predict that the estimated education effect will increase, but the other effects will not change. This is omitted variable bias since IQ matters for income and is correlated with education. The other variables are not correlated at all, so there will not be an effect there.

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	34421.38148	10260.6314	3.354704027	0.000855
education	4443.564383	162.5740924	27.33254922	6E-101
female	257.9317805	1747.715353	0.147582259	0.882733
age	79.16690187	470.3540659	0.168313421	0.866405
age2	-1.087718815	5.037670468	-0.215917024	0.829141

Was right about the education coefficient, but other stuff changed too; goes to show you that it is very hard on simple intuition to make predictions in the presence of omitted variable bias (one needs to look at residual correlations after other stuff is controlled for, but that is beyond what we want to do in this

class; instead the take home is that omitted variable bias causes problems across the regression estimates as a general matter).

6) Would adding more data points (say, the 10X increase noted above) affect the outcome from #5 with respect to estimating the "causal" effect of education? Provide your intuition. Re-run the regression with the larger dataset to verify your intuition.

No, bias is not a small observations problem. Here you would get a comparable result if you actually created 4500 new datapoints as described by the true model, since bias isn't about sample size.

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	34421.38148	3230.047728	10.65661699	3.10164E-26
education	4443.564383	51.17833956	86.8250987	0
female	257.9317805	550.1809571	0.468812628	0.639224032
age	79.16690187	148.0675041	0.534667633	0.592903505
age2	-1.087718815	1.585859136	-0.685886149	0.492816738

7) Rerun the regression from #1 (using the original 500 datapoints) but this time don't use IQ; instead use IQ-50 as the IQ control variable. What differences do you find between your estimates from #1?

This solves the constant problem more or less; no longer is the constant adjusting for the very low IQ folks.

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	6661.598582	2912.900202	2.286929905	0.022622
education	2537.483177	52.23693113	48.57642135	2.9E-190
female	-130.406281	492.2508387	-0.264918352	0.791183
age	404.0086778	132.5391724	3.048220917	0.002425
age2	-4.696905572	1.419601298	-3.308608959	0.001006
netiq	736.1444025	9.710924788	75.80579797	3.2E-274

8) If you were to rerun the regression from #7 using just the female observations, would anything change in terms of the outputs? Provide your intuition. Re-run the regression from #7 to verify your intuition.

Since female isn't correlated with anything, you wouldn't expect the coefficients to change. The intercept would change some (something like 6700-130), but it is not likely that the other estimates will change (though since the sample size is smaller, standard errors will be bigger in general.

	Coefficients	*Standard Error*	*t Stat*	*P-value*

Intercept	6143.587783	4439.541926	1.383833712	0.167717
education	2477.589145	79.53257704	31.15187809	1.41E-85
female	0	0	65535	#NUM!
age	442.6503929	204.7809958	2.161579453	#NUM!
age2	-5.243691051	2.194411059	-2.389566453	0.017654
netiq	743.7418661	14.10599204	52.72524356	1.5E-132

Problem Set 3

The file PS3.xlsx contains stock price data for Exxon and the S&P 500 since the beginning of the year. The ExxonAdjClose variable is the price of exxon stock adjusted for splits and dividends.

On November 6, President Obama announced that he would reject an application of for the completion of the Keystone XL oil pipeline. Arguably, this rejection hurts oil companies who will have to pay higher prices to transport oil from Canada to refiners on the Gulf Coast.

1) Use an event study analysis to examine the effect of the Obama rejection on the price of Exxon stock. Note that the spreadsheet has price information, while an event study focuses on returns.

Regressing the Exxon return on the S&P500 return yields a model of -0.0004+1.0498*S&P (this is if you use all of the data provided. If you use just 100 days before, the intercept is still essentially 0 but with a positive sign this time and the slope is a little bigger). Thus, on November 6, you would have predicted something like a -0.0007 return and on November 9, you would have predicted something like a -0.0107 return. The standard deviation of the abnormal returns is something like 0.009. Thus, in either case, the abnormal return is not statistically significant, though it is substantially larger on the 9th.

2) Is there a difference between using November 6 as the event date vs using November 9 as the event date. Note that Obama made his announcement in the middle of the day on November 6 (i.e., before markets closed).

See above for the differences. Because Obama made his announcement before the close of trading on the 6th, we might have expected that any effect would have been capitalized in the closing price that day. However, maybe it took the weekend for the full effect to be appreciated by the market.

3) Instead of using the multi step approach to event studies used in class, you can also implement an event study by running a single regression that includes an "event" variable that takes the value of 1 on the event date and the value of 0 on the non-event dates. Verify that the two approaches yield similar results.

If you create a variable called event that has the value of 1 for the November 9 observation and the value of 0 for the rest of the observations and then run a regression (I used all of the data; results will be comparable if you use just the event date and the 100 preceding days), you get:

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-0.000377916	0.000617309	-0.61219907	0.541062099
SPr	1.049835999	0.06366938	16.48886786	7.43046E-40
event	-0.010685552	0.009072786	-1.17775866	0.240213261

In which case the event coefficient is comparable to what you calculated above, and the standard error is pretty close to the standard deviation for the non-event date abnormal returns.

4) Estimate (using the single step regression approach described in #3) the event study using 25 days, 100 days, and all of the data in the spreadsheet. Does anything change with respect to

your results? Conceptually, explain how you should think about how many observations to use in the event study.

25 days:

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	**0.000927775**	**0.00240526**	**0.385727**	**0.703243**
SPr	**1.173721061**	**0.303276207**	**3.870139**	**0.000776**
event	**-0.010774353**	**0.012024281**	**-0.89605**	**0.379509**

100 days:

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	**0.00012541**	**0.00101126**	**0.124013**	**0.901559**
SPr	**1.090035725**	**0.086996517**	**12.52965**	**4.55E-22**
event	**-0.010794006**	**0.010199207**	**-1.05832**	**0.292512**

All data:

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	**-0.000377916**	**0.000617309**	**-0.6122**	**0.541062**
SPr	**1.049835999**	**0.06366938**	**16.48887**	**7.43E-40**
event	**-0.010685552**	**0.009072786**	**-1.17776**	**0.240213**

As you can see, the answers are not exactly the same, but they are close. As you add more data, the standard errors decline, which will generally be the case. Also, in principle, as you increase the sample size, you can be more confident that the results will be unbiased (LLN/CLT kick in) assuming you do not think there is omitted variable bias. Generally, in event studies, there will be a trade-off between wanting to use more data and worrying that data from more distant periods will not be informative regarding the relationship between the asset and the market as it exists at the time of the event.

5) Is there any reason to believe that the event study may not be fully capturing the effect of the Obama decision on the value of Exxon stock? Explain why.

Many observers believed Obama would do this long before it happened. Presumably, some of that effect was already capitalized into the Exxon price before the official announcement was made.

Problem Set 4

Read the article "The Effect of Abortion Legalization on Sexual Behavior: Evidence from Sexually Transmitted Diseases" 32 Journal of Legal Studies 407 (2003) available at https://www.law.upenn.edu/fac/jklick/32JLS407.pdf .

1) What is the hypothesis tested in the article?

The specific hypothesis is whether increasing access to abortion led to an increase in risky sex. A more general hypothesis is whether individuals' risky sex decisions are influenced by incentives (costs and benefits)

2) What is the research design used to investigate the hypothesis?

Using STD rates (specifically gonorrhea and syphilis) as a proxy for risky sex, the article compares STD rates before and after abortion was legalized in the early legalizing states (CA, NY, etc) relative to what was going on in the same time period in other states, and then examines the effect of Roe v. Wade on the non-(early) legalizing states relative to the states that had already legalized.

3) What are some alternate hypotheses that are consistent with the primary finding of the article?

Perhaps people's views of sexuality were changing during this period (e.g., summer of love, etc) which led both to changes in abortion law and changes in risky sex without there being a causal link between those two things. Also, perhaps, the underlying risky sex and STD rates were not changing at all; instead, perhaps the STDs were under-diagnosed and once people gained access to abortion, they got tested more frequently (as many abortion providers also provide STD testing), making it appear as though STD rates had risen.

4) How does the article address these alternate hypotheses?

The changing views of sexuality is very difficult to address. The article provides some analysis where differential pre-existing trends are controlled for (so, if you think such changes occur slowly, this might account for them) and the result does not change. However, if these changes are occurring in discrete or more complicated ways (e.g., some non-linear trend), this will not do much to address the concern. Also, to some extent, the fact that the jump also occurs with Roe v. Wade (which presumably wasn't decided because of social changes in, for instance, Iowa) in all sorts of states, gives some confidence that it's not simply changing views. The testing possibility is likewise difficult to rule out. The paper analyzes males and females separately (on the assumption that any testing effect would be concentrated among females) and does not find a significant difference. That said, if tested females then inform their untested partners, this approach does not do much to rule out the testing hypothesis.

5) Are there any policy implications of the article's findings?

It's not clear that there are. Abortion law is largely unrelated to outcomes such as those examined in this paper. However, the general idea that risky sex is influenced by costs and benefits may be helpful

in general terms regarding the design of policies aimed at reducing risky sex (e.g., condom distribution policies, etc).