Problem set 2: understanding ordinary least squares regressions

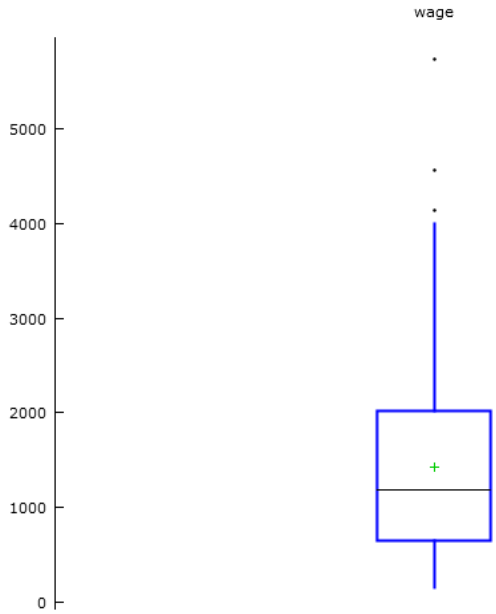
September 12, 2013

1 Introduction

This problem set is meant to accompany the undergraduate econometrics video series on youtube; covering roughly the 30th video through to the 85th. These are the answers to this problem set.

2 NBA Wages - practical

1. The results of the boxplot are shown below. The wage data is positively skewed.



2. The correlation matrix of all the variables in the dataset is shown below. (Forgive the format which this matrix is outputted in - I know it's not the neatest table!)

Correlation coefficients, using the observations 1–269
(missing values were skipped)
5% critical value (two-tailed) = 0.1196 for n = 269

marr	wage	exper	age	coll	
1.0000	0.1581	0.3283	0.3673	−0.0435	marr
	1.0000	0.4092	0.3424	−0.1056	wage
		1.0000	0.9412	0.0873	exper
			1.0000	0.0743	age
				1.0000	coll

games	minutes	guard	forward	center	
0.0691	0.1051	0.0305	−0.0253	−0.0069	marr
0.3038	0.5634	−0.1247	0.0511	0.0967	wage
0.1482	0.2143	−0.0520	−0.0002	0.0684	exper
0.1264	0.1471	−0.0657	0.0059	0.0784	age
−0.0149	−0.0681	0.0693	−0.0494	−0.0263	coll
1.0000	0.7878	0.1181	0.0186	−0.1791	games
	1.0000	0.1184	0.0326	−0.1978	minutes
		1.0000	−0.7079	−0.3865	guard
			1.0000	−0.3778	forward
				1.0000	center

points	rebounds	assists	draft	allstar	
0.1237	−0.0330	0.1607	0.0267	0.0536	marr
0.6570	0.5409	0.3282	−0.3625	0.3973	wage
0.1908	0.1635	0.1499	0.0246	0.0800	exper
0.1048	0.1165	0.0783	0.1164	0.0027	age
−0.1204	−0.1152	−0.0335	0.1016	−0.0810	coll
0.5006	0.3328	0.3562	−0.1155	0.1847	games
0.8392	0.5852	0.5997	−0.2326	0.4247	minutes
0.1008	−0.4797	0.5085	0.1487	0.0231	guard
0.0005	0.3678	−0.2923	−0.1532	−0.0160	forward
−0.1328	0.1486	−0.2849	0.0049	−0.0093	center
1.0000	0.5633	0.5393	−0.3228	0.6086	points
	1.0000	0.0600	−0.2995	0.3325	rebounds
		1.0000	−0.0631	0.3798	assists
			1.0000	−0.2314	draft
				1.0000	allstar

avgmin	lwage	black	children	expersq	
0.1098	0.1478	−0.0947	0.2811	0.2883	marr
0.6218	0.8938	0.0768	0.1657	0.3458	wage
0.2234	0.4056	−0.0050	0.2048	0.9503	exper
0.1411	0.3178	−0.0543	0.1925	0.9046	age
−0.0632	−0.0559	0.0289	−0.0802	0.0489	coll
0.5809	0.3697	0.1039	0.0812	0.0935	games
0.9353	0.5915	0.1402	0.1742	0.1278	minutes
0.1050	−0.0859	0.0542	−0.0011	−0.0657	guard
0.0377	0.0984	0.1008	0.0313	−0.0118	forward
−0.1869	−0.0158	−0.2027	−0.0395	0.1016	center
0.8870	0.6194	0.1163	0.1755	0.1226	points
0.6351	0.4882	0.1218	0.1419	0.1164	rebounds
0.6327	0.3614	0.0245	0.1853	0.0677	assists
−0.2676	−0.4022	−0.0818	−0.0518	0.0076	draft
0.4536	0.2954	0.0587	0.0804	0.0588	allstar
1.0000	0.6407	0.1360	0.1946	0.1293	avgmin
	1.0000	0.1203	0.1888	0.3175	lwage
		1.0000	0.0194	−0.0058	black
			1.0000	0.1583	children
				1.0000	expersq

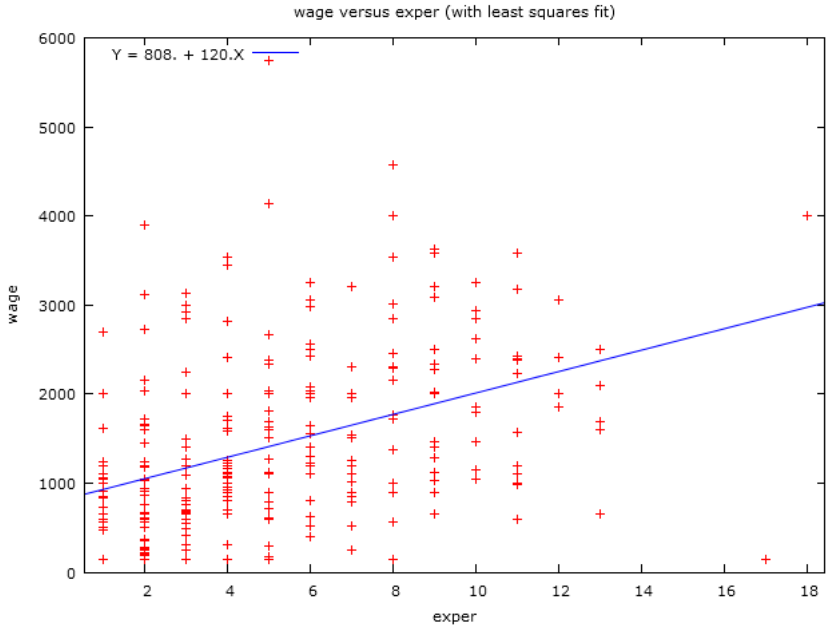
agesq	marrblck	
0.3574	0.8028	marr
0.3393	0.1514	wage
0.9421	0.2319	exper
0.9968	0.2441	age
0.0690	0.0074	coll
0.1181	0.0827	games
0.1351	0.1151	minutes
−0.0704	0.0601	guard
0.0020	0.0126	forward
0.0897	−0.0952	center
0.0978	0.1342	points
0.1118	−0.0071	rebounds
0.0646	0.1236	assists
0.1084	−0.0317	draft
0.0052	0.0372	allstar
0.1286	0.1215	avgmin
0.3108	0.1293	lwage
−0.0495	0.3500	black
0.1839	0.2237	children
0.9269	0.2036	expersq
1.0000	0.2380	agesq
	1.0000	marrblck

From the various experience bivariate correlations, it is clear that age (as you might expect) is highly correlated with experience. The issue with including both of these

variables in an OLS regression is due to the high level of multicollinearity amongst them. Intuitively, OLS is going to struggle to disentangle the effect of experience from age on players' wages. This will be realised by a large estimated standard error for both coefficients, and perhaps a lack of individual significance.

3. A number of variables are quite highly correlated with wages, having a correlation over 0.3: experience, age, games, minutes, points, rebounds, assists, draft, allstar, avgmin, and trivially lwage, expersq, agesq.

4. The result of an X-Y scatter is shown below. There appears to be quite a strong positive correlation between these two variables.



5. The results of this regression are shown below. Since the p value on experience is less than 0.05 we can conclude that the effect of experience on wages is significant (in this current model setting).

Model 1: OLS, using observations 1–269
Dependent variable: wage

	Coefficient	Std. Error	t-ratio	p-value
const	807.932	100.847	8.0114	0.0000
exper	120.317	16.4199	7.3275	0.0000

Mean dependent var	1423.828	S.D. dependent var	999.7741
Sum squared resid	2.23e+08	S.E. of regression	913.9559
R^2	0.167425	Adjusted R^2	0.164307
$F(1, 267)$	53.69196	P-value(F)	2.79e–12
Log-likelihood	−2214.674	Akaike criterion	4433.348
Schwarz criterion	4440.538	Hannan–Quinn	4436.236

6. One more year of experience is associated with an average increase in salary of around $120K.

7. In my view this coefficient likely overstates the effect of experience on wages, since those who are better tend to be employed as professional basketball players for longer, and hence are paid more. In other words there is another third variable 'quality' which is causing both wages and experience to be higher. A measure of players' quality is contained within the variables: 'points', 'minutes', 'rebounds' etc. variables. So their inclusion is likely to bring the estimate of the effect of experience down.

8. The results of this regression are shown below. Unsurprisingly this model suggests that the effect of age on wages is positive, although, at first glance, the effect appears that it might be smaller than experience.

Model 2: OLS, using observations 1–269
Dependent variable: wage

	Coefficient	Std. Error	t-ratio	p-value
const	−1341.73	467.888	−2.8676	0.0045
age	100.955	16.9510	5.9557	0.0000

Mean dependent var	1423.828	S.D. dependent var	999.7741
Sum squared resid	2.36e+08	S.E. of regression	941.0834
R^2	0.117268	Adjusted R^2	0.113962
$F(1, 267)$	35.47000	P-value(F)	8.14e–09
Log-likelihood	−2222.542	Akaike criterion	4449.085
Schwarz criterion	4456.274	Hannan–Quinn	4451.972

9. The average wage for a 30 year old would be 100.955 x age - 1341.73 which is around $1.7m - this seems reasonable. For a 90 year old, our model predicts that their wage will be close to $8m! This latter prediction is completely out of sample, and also very unrealistic. One has to be very careful when extending the results of a regression to make out of sample predictions. (Out of sample here means that we currently do not have any data for the wages of 90 year old basketball players.)

10. One way might be to include the square of the age in the regression, as this would suggest that there might be diminishing marginal returns to age. I include the results of this regression below.

Model 3: OLS, using observations 1–269
Dependent variable: wage

	Coefficient	Std. Error	t-ratio	p-value
const	−2701.01	3045.89	−0.8868	0.3760
age	197.275	213.941	0.9221	0.3573
agesq	−1.67914	3.71785	−0.4516	0.6519

Mean dependent var	1423.828	S.D. dependent var	999.7741
Sum squared resid	2.36e+08	S.E. of regression	942.4894
R^2	0.117944	Adjusted R^2	0.111312
$F(2, 266)$	17.78412	P-value(F)	5.64e–08
Log-likelihood	−2222.439	Akaike criterion	4450.878
Schwarz criterion	4461.662	Hannan–Quinn	4455.209

Note that the results of this regression aren't that suggestive of diminishing marginal returns to age. Perhaps a better way to deal with this issue would be not to make predictions on out of sample data!

11. The results of this regression are shown below. Note that the coefficient on age is now negative, and the coefficient on experience has nearly doubled. This unstable change in the coefficient values is due to high multicollinearity between age and experience.

Model 4: OLS, using observations 1–269
Dependent variable: wage

	Coefficient	Std. Error	t-ratio	p-value
const	3295.29	1096.40	3.0055	0.0029
exper	223.686	48.2105	4.6398	0.0000
age	−110.115	48.3352	−2.2782	0.0235

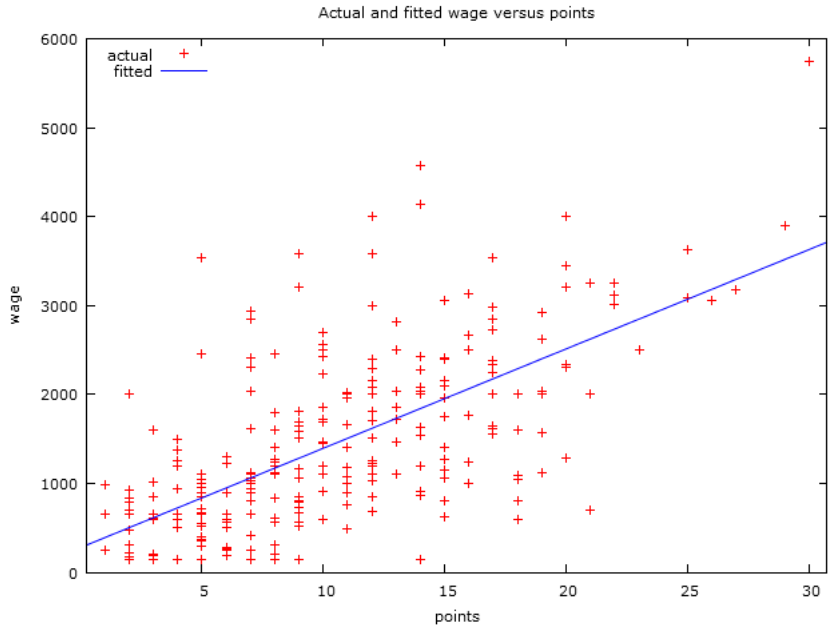
Mean dependent var	1423.828	S.D. dependent var	999.7741
Sum squared resid	2.19e+08	S.E. of regression	906.8679
R^2	0.183359	Adjusted R^2	0.177219
$F(2, 266)$	29.86228	P-value(F)	2.00e–12
Log-likelihood	−2212.075	Akaike criterion	4430.150
Schwarz criterion	4440.934	Hannan–Quinn	4434.481

12. The results of this regression are shown below along with the graph of actual vs fitted wages. The results of this model are suggestive that an increase in 10 points per game is associated with an increase in wages on average by $1.1m.

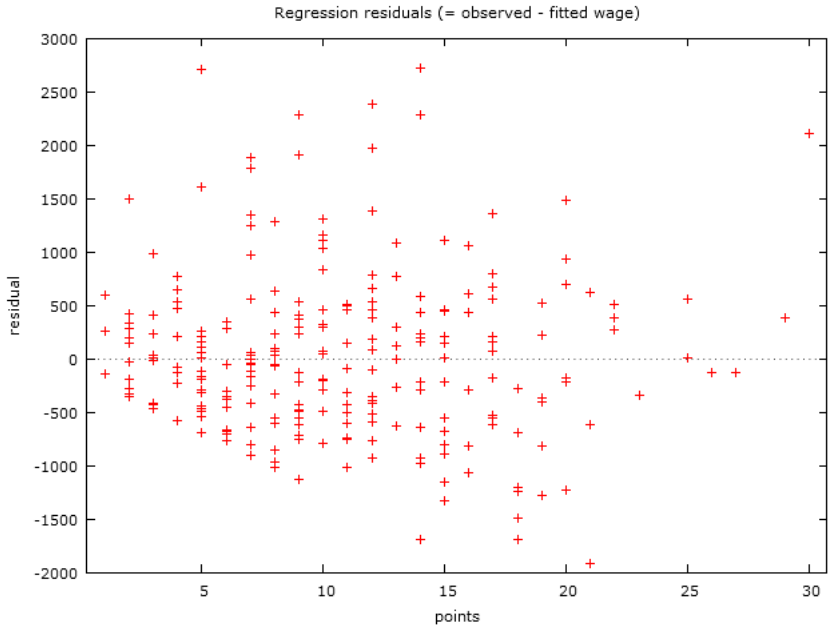
Model 5: OLS, using observations 1–269
Dependent variable: wage

	Coefficient	Std. Error	t-ratio	p-value
const	278.102	92.6940	3.0002	0.0030
points	111.667	7.84116	14.2411	0.0000

Mean dependent var	1423.828	S.D. dependent var	999.7741
Sum squared resid	1.52e+08	S.E. of regression	755.1072
R^2	0.431684	Adjusted R^2	0.429555
$F(1, 267)$	202.8089	P-value(F)	1.28e–34
Log-likelihood	−2163.316	Akaike criterion	4330.632
Schwarz criterion	4337.821	Hannan–Quinn	4333.519



13. Observation number here is arbitrary and does not reflect any mechanism of interest. Hence a residual plot against points is most appropriate. This plot is shown below.



14. I would say that there is definitely evidence of systematic increases in the variance of our estimates as points increases from 0-15. There is then a decline in variance towards the latter end of the points spectrum. This makes intuitive sense. When players score few points, they are not paid much. When they score a reasonable number of points they tend to be paid more, but there is a higher variance. This could be because of the fact that players on a basketball team occupy different positions - some are more focussed on scoring, others on defending. This would mean that once a threshold number of points is reached the players are paid more

or less dependent on their respective abilities in their positions. When a player scores a high number of points, they are in short supply, and can hence command a higher wage.

15. I suspect it is too high. There is probably some reverse causality happening, whereby players that are paid more tend to score more points. Also, points likely also captures some of the effects of other variables that are important in determining wages. For example, players from better teams have better team mates, and hence tend to have more chances to score. At the same time the players on the better teams tend to be paid more.

16. No real answer here - the two variables should now be in your Gretl variable selection.

17. Now create two new regression models (keeping your current regression of wages on points):

$$wage_i = \alpha + \beta_1 points_i + \beta_2 pointsq_i$$

$$wage_i = \alpha + \beta_1 points_i + \beta_2 pointsq_i + \beta_3 pointsc_i$$

The results of these two regressions are shown below. Model 6: OLS, using observations 1–269Dependent variable: wage

	Coefficient	Std. Error	t-ratio	p-value
const	396.641	148.060	2.6789	0.0078
points	85.4355	26.7260	3.1967	0.0016
pointsq	1.07765	1.04967	1.0267	0.3055

Mean dependent var	1423.828	S.D. dependent var	999.7741
Sum squared resid	1.52e+08	S.E. of regression	755.0309
R^2	0.433927	Adjusted R^2	0.429671
$F(2, 266)$	101.9520	P-value(F)	1.36e–33
Log-likelihood	−2162.784	Akaike criterion	4331.568
Schwarz criterion	4342.352	Hannan–Quinn	4335.899

Model 7: OLS, using observations 1–269
Dependent variable: wage

	Coefficient	Std. Error	t-ratio	p-value
const	59.2577	221.555	0.2675	0.7893
points	207.232	65.4185	3.1678	0.0017
pointsq	−9.68923	5.38662	−1.7988	0.0732
pointsc	0.260608	0.127911	2.0374	0.0426

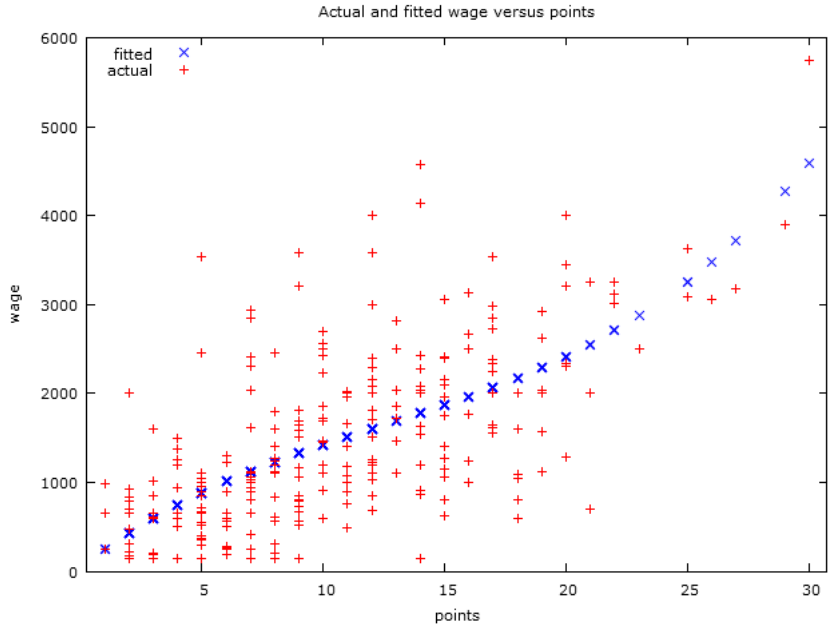
Mean dependent var	1423.828	S.D. dependent var	999.7741
Sum squared resid	1.49e+08	S.E. of regression	750.5981
R^2	0.442657	Adjusted R^2	0.436348
$F(3, 265)$	70.15684	P-value(F)	2.00e–33
Log-likelihood	−2160.694	Akaike criterion	4329.387
Schwarz criterion	4343.766	Hannan–Quinn	4335.162

(a) The latter model has a higher R-squared of 0.44. This means that these independent variables explain 44% of the variation in the dependent variable.

(b) This would suggest that there is diminishing diminishing (yep two diminishings!) marginal returns to points. In other words it is pretty nonsensical.

(c) The last regression still has the highest adjusted R-squared.

(d) The first regression (with only wages regressed on points), since it is parsimonious, and has a clear interpretation. The latter of the three especially is overfitting the data. I include the last regression fitted residuals below so you can see for yourself. Don't let apophenia (seeing meaningless patterns in data) get the better of you - a linear line is still better than this curve.



18. Freestyle: this is the best I /a few of my colleagues could do. I'm not suggesting by any means this is the gold standard, but in my view it represents a reasonably good first stab at a reasonable model for wages. The variable 'centerpoint' is equal to center times points, and ldraft is the log of the draft variable. All the variables have the expected signs, and there is not a great deal of movement in coefficient values when the specification is changed slightly. There is a logic to the inclusion of the multiplication of center and points, since it is suggested that centers that score lots of points will be valued disproportionately more (perhaps this makes sense, my knowledge of basketball could be better!) Also, by logging the draft variable this allows it to have a nonlinear effect on wages, which is to be expected.

Model 49: OLS, using observations 1–269 (n = 240)
Missing or incomplete observations dropped: 29
Dependent variable: wage

	Coefficient	Std. Error	t-ratio	p-value
const	1003.72	185.293	5.4169	0.0000
points	68.6397	8.14700	8.4252	0.0000
centerpoint	36.1528	9.28749	3.8926	0.0001
ldraft	−281.611	45.2199	−6.2276	0.0000
exper	83.9601	12.0779	6.9515	0.0000

Mean dependent var	1532.652	S.D. dependent var	996.5671
Sum squared resid	95068638	S.E. of regression	636.0404
R^2	0.599478	Adjusted R^2	0.592661
$F(4, 235)$	87.93361	P-value(F)	1.45e–45
Log-likelihood	−1887.282	Akaike criterion	3784.563
Schwarz criterion	3801.967	Hannan–Quinn	3791.576

3 Theory

1. A researcher is interested in quantifying the effect of the number of broken windows in a block on property prices, and the results of a preliminary regression are:

$$Hprice_i = 100 - 10windows_i$$

Where $windows_i$ represents the number of broken windows counted on a block, i, and $Hprice_i$ is the average property value (in thousands of $) on that same block.

   (a) Each additional broken window on a block is associated with a decline in house prices of around $10K.

   (b) Broken windows are likely associated with other omitted variables which are correlated with house prices. An example might be the level of crime. There is also possibly an argument that there is a degree of reverse causation happening here, whereby lower-priced houses are more likely to suffer broken windows.

   (c) Due to its correlation with crime, it most likely overstates the effect which broken windows have on house prices.

   (d) Another variable is included in the regression, $emerg_i$, which is a measure of the number of emergency services calls which were made from each block over a period of time. And the result of the regression is:

$$Hprice_i = 100 - 3windows_i - 5emerg_i$$

   The effect of broken windows is diluted due to its correlation with emergency calls - a measure of crime. They both may be highly correlated, meaning that multicollinearity is at play, and can cause individual significances which are relatively low compared to their joint significance.

2. The zero conditional mean assumption of the Gauss-Markov conditions is often stated as:

$$\mathbb{E}[\varepsilon_i | X_i] = 0 \tag{1}$$

(a) In order to prove this we need to be aware of the law of iterated expectations. This relates the unconditional expectation to the conditional expectation by the following:

$$\mathbb{E}(X_i\varepsilon_i) = \mathbb{E}[\mathbb{E}(X_i\varepsilon_i|X_i)]$$

Since we are given X_i the expectations operator passes through it. QED.

$$\mathbb{E}(X_i\varepsilon_i) = \mathbb{E}[X_i\mathbb{E}(\varepsilon_i|X_i)] = 0$$

(b) Yes, it does. The covariance between X_i and ε_i can be written:

$$Cov(X_i, \varepsilon_i) = \mathbb{E}(X_i - \mathbb{E}(X_i))(\varepsilon_i - \mathbb{E}(\varepsilon_i)) = \mathbb{E}(X_i\varepsilon_i) - \mathbb{E}(X_i)\mathbb{E}(\varepsilon_i)$$

Since we know that $\mathbb{E}(\varepsilon_i) = 0$ by definition, we have that $Cov(X_i, \varepsilon_i) = 0$.

(c) No. Independence implies covariance being zero, but not necessarily vice versa. The reason independence implies covariance being zero, is that the definition of independence of X_i and Y_i is:

$$\mathbb{E}(X_iY_i) = \mathbb{E}(X_i)\mathbb{E}(Y_i)$$

Hence, we have that:

$$Cov(X_i, Y_i) = \mathbb{E}(X_iY_i) - \mathbb{E}(X_i)\mathbb{E}(Y_i) = 0$$

3. A researcher is interested in measuring what the effect of an individual's innate 'language intelligence' is on their ability to learn a language. She finds 100 volunteers for the study who have all not learned French before, nor have they learned any other languages to any serious fluency. Her theory is that those individuals who have higher innate measures of 'language intelligence' will take less time to reach of level of proficiency in French.

Each volunteer is enrolled in a day course in basic French, and is tested at the end of the day in French. At the end of the day each participant also takes a standardised IQ test. The researcher then carries out the following regression:

$$score_i = \alpha + \beta IQ_i + u_i$$

(a) There are a range of variables which could help explain their ability to learn a language which are correlated with IQ. Examples include tiredness (because of the contemporaneous testing of both abilities), parental education, etc.

(b) It likely underestimates the effect of 'language intelligence' on their language-learning abilities. See this Youtube video for an explanation of why:
`http://tinyurl.com/qz8syjk`

(c) Using IQ in the regression is equivalent to having measurement error in the independent variable. Hence we have that:

$$IQ_i = LI_i + \varepsilon_i$$

Where we have that by definition $Cov(LI_i, \varepsilon_i) = 0$. Hence we know that we have a covariance between IQ and the error term:

$$Cov(IQ_i, \varepsilon_i) = Cov(LI_i + \varepsilon_i, \varepsilon_i) = \sigma_\varepsilon^2$$

Writing out the expression for the True relationship proposed we have that:

$$score_i = \gamma_0 + \gamma_1 LI_i + \eta_i$$

However, we don't observe LI_i hence we substitute in for it using our relation between IQ_i and LI_i above.

$$score_i = \gamma_0 + \gamma_1 IQ_i + (\eta_i - \gamma_1 \varepsilon_i)$$

Where the term in parenthesis is the observed error. Since we know that $Cov(IQ_i, \varepsilon_i) = \sigma_\varepsilon^2$, we can see straight away that there is endogeneity. Thus meaning that OLS estimators are both biased and inconsistent. Furthermore the bias will be downwards due to the negative sign in the expression above.