Statistics and Econometrics: Problem Set 3

Siow Meng Low

29 October 2016

Question 1

Question 1a: Usual Standard Errors and Robust Standard Errors

The equation is given as:

$$math4 = \beta_0 + \beta_1 lunch + \beta_2 log(enroll) + \beta_3 log(exppp) + u$$

Table 1 below displays the model information of regressing *math*4 (percentage of students passing fourth grade passing math test) on *lunch* (percentage of students eligible for free or reduced lunch), *lenroll* (school enrollment, in natural log form), and *lexppp* (expenditures per pupil, in natural log form). The left column contains the information of usual standard errors (in parentheses) whereas the right column shows the robust standard errors (in parentheses).

Table 1: Question 1a - Standard Errors vs Robust Standard Errors

	math4: Math Performance of 4th Graders	
	Usual Standard Errors	Robust Standard Errors
	(1)	(2)
lunch	-0.449^{***}	-0.449^{***}
	(0.015)	(0.017)
lenroll	-5.399^{***}	-5.399^{***}
	(0.940)	(1.131)
lexppp	3.525^{*}	3.525
	(2.098)	(2.354)
Constant	91.932^{***}	91.932^{***}
	(19.962)	(23.087)
N	1,692	1,692
R^2	0.373	0.373
Adjusted R^2	0.372	0.372
Residual Std. Error (df = 1688)	15.302	15.302
F Statistic (df = 3; 1688)	334.567^{***}	334.567^{***}

Notes:
***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

From the above table, we can observe that the robust standard errors are larger than the usual standard errors. As a result, the statistical significance of some coefficients have been lowered. For instance, *lexppp*, which is significant at 10% level when usual standard errors are used, becomes not significant when we use robust standard errors.

Question 1b: White Test for Heteroskedasticity

The test statistic of the White Test is 229.78 and p-value is 1.269e-50. The p-Value is extremely small. At 5% significance level, we reject the null hypothesis of homoskedasticity and conclude that heteroskedasticity is present in this model.

Question 1c: Weighted Least Squares Estimation

Table 2 below compares the OLS with WLS estimates. The left column contains the information of the OLS estimates (with usual standard errors) whereas the right column displays the information of the WLS estimates (with usual standard errors).

Table 2: Question 1c - OLS vs WLS Estimates

	math4: Math Performance of 4th Graders	
	OLS Estimates	WLS Estimates
	(1)	(2)
lunch	-0.449^{***}	-0.449^{***}
	(0.015)	(0.015)
lenroll	-5.399^{***}	-2.647^{***}
	(0.940)	(0.836)
lexppp	3.525^{*}	6.474^{***}
	(2.098)	(1.686)
Constant	91.932^{***}	50.478^{***}
	(19.962)	(16.510)
N	1,692	1,692
R^2	0.373	0.360
Adjusted R^2	0.372	0.359
Residual Std. Error (df = 1688)	15.302	1.900
F Statistic (df = 3; 1688)	334.567^{***}	316.475^{***}
Notes:	***Significant at the 1 percent level.	
	**Significant at the 5 percent level.	
	*Significant at the 10 percent level.	

We can observe from the table that coefficients of WLS estimates are very different from OLS estimates. While the coefficient of *lunch* remains unchanged, the coefficients of *lenroll*, *lexppp*, and the constant (i.e. intercept) term have all changed. The usual standard errors of WLS estimates are also smaller than that of OLS estimates. Consequently, the independent variable *lexppp* has become much more statistically significant (significant at 1% level for WLS estimates).

Question 1d: Robust Standard Errors of WLS

To allow misspecification of the variance function, we shall use the robust standard errors of the WLS estimates. Table 3 below compares the usual standard errors and robust standard errors of WLS. The left column contains the information of usual standard errors (in parentheses) whereas the right column shows the robust standard errors (in parentheses).

Table 3: Question 1d - WLS Standard Errors vs Robust Standard Errors

	math4: Math Performance of 4th Graders	
	Usual Standard Errors	Robust Standard Errors
	(1)	(2)
lunch	-0.449^{***}	-0.449^{***}
	(0.015)	(0.014)
lenroll	-2.647^{***}	-2.647^{**}
	(0.836)	(1.055)
lexppp	6.474^{***}	6.474^{***}
	(1.686)	(1.813)
Constant	50.478^{***}	50.478^{***}
	(16.510)	(18.925)
N	1,692	1,692
R^2	0.360	0.360
Adjusted R^2	0.359	0.359
Residual Std. Error (df = 1688)	1.900	1.900
F Statistic (df = 3; 1688)	316.475^{***}	316.475^{***}

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

With the exception of *lunch*, the robust WLS standard errors of *lenroll*, *lexppp*, and constant (i.e. intercept) term are all larger than the usual WLS standard errors. As a result, the statistical significance of some coefficients have been lowered. For instance, *lenroll*, which is significant at 1% level with usual WLS standard errors, becomes significant at 5% level when we use robust WLS standard errors.

Question 1e: Estimation of Effect of Spending

Table 4 below compares the OLS with WLS estimates (with robust standard errors for both types of estimates). The left column contains the information of the OLS estimates whereas the right column displays the information of the WLS estimates.

Table 4: Question 1e - OLS vs WLS Estimating Effect of Spending (with Robust Standard Errors)

	math4: Math Performance of 4th Graders	
	OLS Estimates	WLS Estimates
	(1)	(2)
lunch	-0.449^{***}	-0.449^{***}
	(0.017)	(0.014)
lenroll	-5.399^{***}	-2.647^{**}
	(1.131)	(1.055)
lexppp	3.525	6.474^{***}
	(2.354)	(1.813)
Constant	91.932^{***}	50.478^{***}
	(23.087)	(18.925)
N	1,692	1,692
R^2	0.373	0.360
Adjusted R^2	0.372	0.359
Residual Std. Error (df = 1688)	15.302	1.900
F Statistic (df = 3; 1688)	334.567^{***}	316.475^{***}
Notes:	***Significant at the 1 percent level.	
	**Significant at the 5 percent level.	
	*Significant at the 10 percent level.	

Since we are interested in estimating the effect of spending, we shall focus on the independent variable, *lexppp*. Its WLS coefficient, β_{lexppp} is much larger than OLS coefficient. Further more, the WLS estimate has a smaller robust standard error than OLS estimate. Consequently, the WLS coefficient is more precise and statistically more significant (WLS estimate of β_{lexppp} is significant at 1% level whereas OLS estimate is not statistically significant).

As we can observe from this example, WLS addresses heteroskedasticity and provide a more efficient estimates (with lower robust standard errors than OLS estimates).

Question 2

Question 2a: Identify Multicollinearity

If we choose to include all the mentioned independent variables, the equation of this "Full Model" is in the following form (*center* is used as the base case, hence not shown in the equation):

$$log(wage) = \beta_0 + \beta_1 exper + \beta_2 coll + \beta_3 games + \beta_4 avgmin + \beta_5 guard + \beta_6 forward + \beta_7 points + \beta_8 rebounds + \beta_9 assists + \beta_{10} draft + \beta_{11} allstar + \beta_{12} black + \beta_{13} children + \beta_{14} marr + u$$

Table 5 displays the Variance Inflation Factors for all the independent variables. As a rule of thumb, a VIF value greater than 10 indicates multicollinearity problem. As we can see from the table, only *avgmin* has a VIF value greater than 10. Therefore, it indicates that the presence of *avgmin* causes multicollinerity problem.

Table 5: Variance Inflation Factors of All Independent Variables

	VIF Value
coll	1.083
black	1.108
children	1.162
draft	1.186
exper	1.210
marr	1.268
games	1.489
allstar	1.900
forward	2.202
assists	3.244
guard	4.019
rebounds	4.024
points	6.517
avgmin	10.532

Question 2b: Forward and Backward Stepwise Selections

Table 6 tabulates the models with lowest AIC using three stepwise selection methods:

- Leftmost Column: Forward-stepwise selection
- Middle Column: Backward-stepwise selection
- Rightmost Column: Both directions (Initial model contains no independent variable)

Table 6: Question 2b - Lowest AIC by Forward and Backward Selections

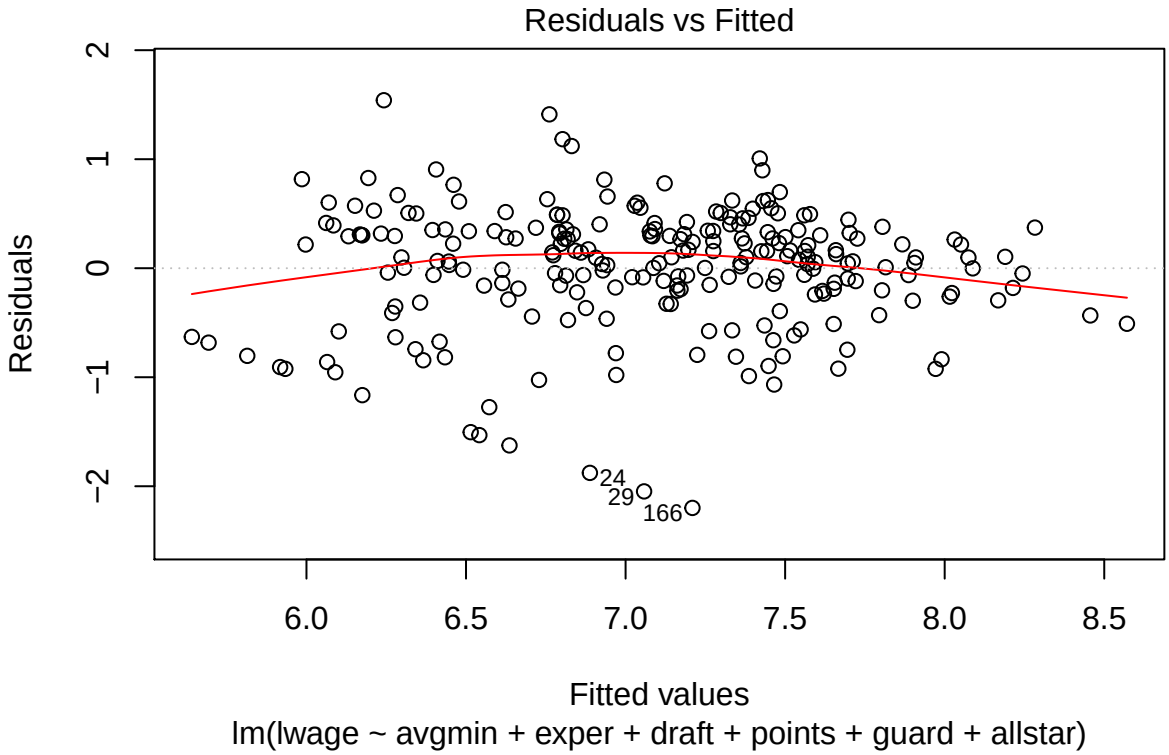
	log(wage)		
	Forward Selection	Backward Selection	Both Directions
	(1)	(2)	(3)
avgmin	0.023***	0.023***	0.023***
	(0.008)	(0.008)	(0.008)
exper	0.065***	0.065***	0.065***
	(0.011)	(0.011)	(0.011)
draft	−0.011***	−0.011***	−0.011***
	(0.002)	(0.002)	(0.002)
points	0.044***	0.044***	0.044***
	(0.015)	(0.015)	(0.015)
guard	−0.134*	−0.134*	−0.134*
	(0.077)	(0.077)	(0.077)
allstar	−0.242*	−0.242*	−0.242*
	(0.142)	(0.142)	(0.142)
Constant	5.994***	5.994***	5.994***
	(0.133)	(0.133)	(0.133)
N	240	240	240
R^2	0.520	0.520	0.520
Adjusted R^2	0.508	0.508	0.508
Residual Std. Error (df = 233)	0.568	0.568	0.568
F Statistic (df = 6; 233)	42.125***	42.125***	42.125***
Notes:			***Significant at the 1 percent level.
			**Significant at the 5 percent level.
			*Significant at the 10 percent level.

As we can see from the table, all the three selection methods produce the exact same model.

Question 2c: Residual Plots

Residual vs Fitted Plot

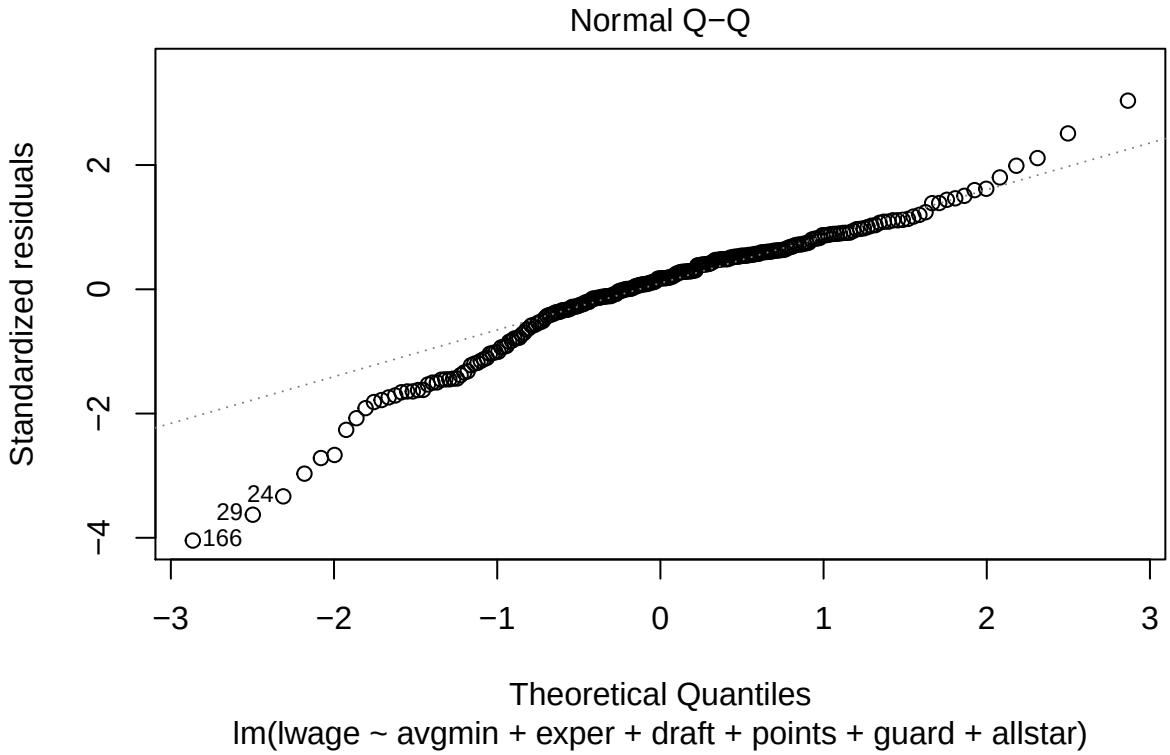
The first residual plot we are going to investigate is the "Residuals vs Fitted" plot.



From the plot above, we can see that there are no extreme outliers. The spread of the residuals at each fitted value are roughly similar hence the linear model is adequate.

Normal Q-Q Plot

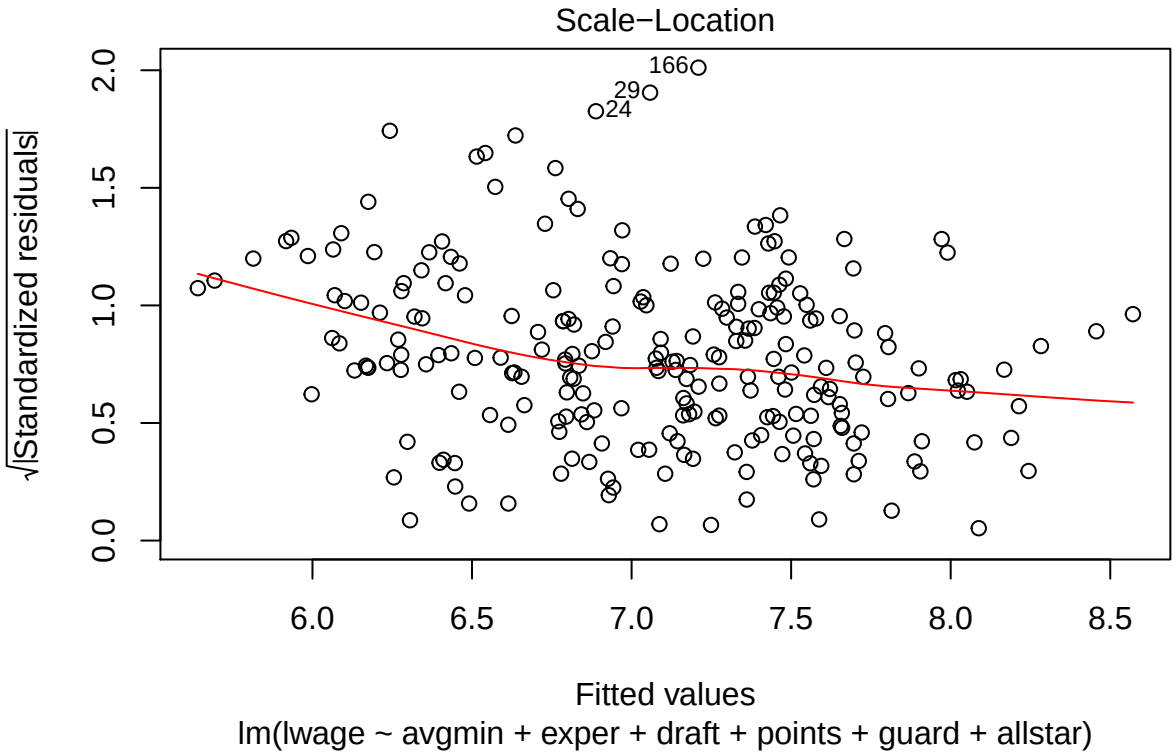
The second plot to investigate is the normal QQ Plot.



From the plot above, the residuals seem to be approximately normally distributed since most of the points lie along the straight line. However, the distribution of residuals seem to have heavy tails at both ends (since both tails of Q-Q plot twist counterclockwise). This is especially the case for the left tail, where the (counterclockwise) twist is much more severe.

Scaled Location Plot

Next plot to investigate is the Scaled Location plot.

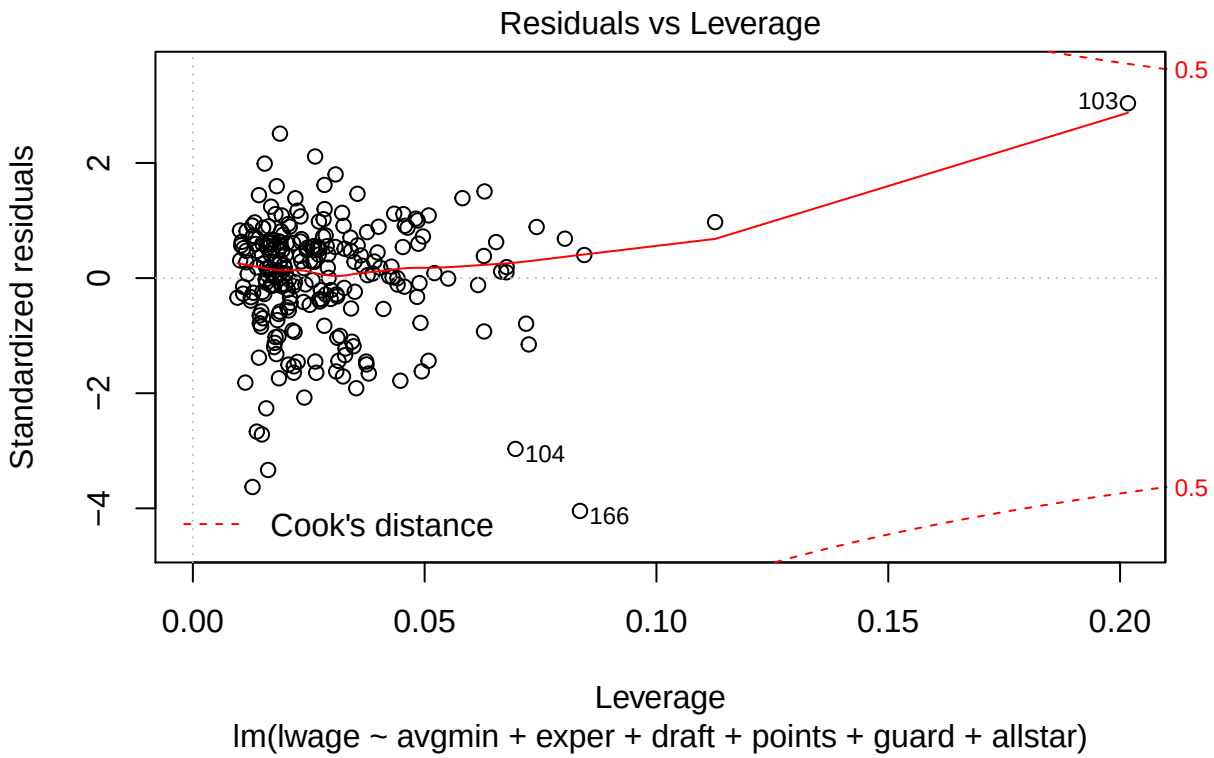


Scaled Location plot shows us the spread of the residuals across different fitted values. From the plot, we can see that the residuals are roughly spreaded equally across all of the fitted values. This implies that our assumption of homoskedasticity should hold.

One thing to note is that towards the right end of the x-axis (with fitted value greater than 8), the residuals seem to be less spreaded compared with other fitted values. There might be some weak indications of heteroskedasticity, homoskedasticity tests might need to be performed to affirm this.

Residuals vs Leverage Plot

The last plot to investigate is the "Residuals vs Leverage" plot.



From the plot, we can see that there are a few high leverage points. However, their Cook's distances are relatively small (smaller than 0.5). This means that although they have high leverage values, they are not particularly influential.

In other words, these data points would not be extreme outliers. However, it is still good to examine these points with high leverage (especially the point marked as '103').

Question 2d: AIC Without *avgmin*

Table 7 tabulates the models with lowest AIC using three stepwise selection methods (without *avgmin* in the full model):

- Leftmost Column: Forward-stepwise selection
- Middle Column: Backward-stepwise selection
- Rightmost Column: Both directions (Initial model contains no independent variable)

Table 7: Question 2d - Lowest AIC by Forward and Backward Selections (without avgmin)

	log(wage)		
	Forward Selection	Backward Selection	Both Directions
	(1)	(2)	(3)
points	0.059***	0.070***	0.059***
	(0.011)	(0.009)	(0.011)
exper	0.066***	0.065***	0.066***
	(0.011)	(0.011)	(0.011)
guard		−0.255***	
		(0.094)	
draft	−0.011***	−0.011***	−0.011***
	(0.002)	(0.002)	(0.002)
allstar	−0.325**	−0.369***	−0.325**
	(0.140)	(0.141)	(0.140)
rebounds	0.042**		0.042**
	(0.016)		(0.016)
black		0.158*	
		(0.094)	
assists	0.036*	0.057**	0.036*
	(0.022)	(0.025)	(0.022)
Constant	6.064***	6.072***	6.064***
	(0.120)	(0.134)	(0.120)
N	240	240	240
R^2	0.515	0.520	0.515
Adjusted R^2	0.503	0.506	0.503
Residual Std. Error	0.571 (df = 233)	0.569 (df = 232)	0.571 (df = 233)
F Statistic	41.241*** (df = 6; 233)	35.945*** (df = 7; 232)	41.241*** (df = 6; 233)

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

As observed from the above table, "Forward Selection" and "Both Direction" produced the same linear model whereas "Backward Selection" produced another linear model which uses a different set of independent variables. The differences between these models are that:

- Forward Selection has *rebounds* in final model, which is not present in the final model of Backward Selection
- Backward Selection has *guard* and *black* in the final model, which are not present in the final model of Forward Selection

Note the the R^2 and adjusted R^2 of both models are very close. Although Forward Selection and Backward Selection picked different set of independent variables, the final models are equally adequate.

As we noted from Question 2a, the VIF value of *avgmin* is greater than 10 and it might cause multicollinearity

as it is highly correlated with some of the independent variables in the "Full Model". The three variables, which are most highly correlated with *avgmin* in the "Full Model", are listed in Table 8 below.

Table 8: Independent Variables Highly Correlated with avgmin

	Correlation
points	0.874
rebounds	0.626
assists	0.608

We can see that *points* is highly correlated with *avgmin*. By ensuring *avgmin* is not in the final model, the multicollinearity effect is lessened and the standard error for *points* in Table 7 is now lower than in Table 6 (Question 2b).

One last observation to note is that the R^2 and adjusted R^2 values in Table 7 are very close to the values in Table 6 (Question 2b). This is because AIC aims to seek a balance between fit and simplicity. Since *avgmin* is no longer in consideration, AIC will now pick other highly correlated independent variables, such as *rebounds* and *assists*, to explain the variations in *lwage*.