Problem Set 2

Philip Nye

2/26/2022

Contents

NBA Wages - Practical

Problem 1 . . . .
Problem 2 . . . .
Problem 3 . . . .
Problem 4 . . . .
Problem 5 . . . .
Problem 6 . . . .
Problem 7 . . . .
Problem 8 . . . .
Problem 9 . . . .
Problem 10 . . . .
Problem 11 . . . .
Problem 12 . . . .
Problem 13 . . . .
Problem 14 . . . .
Problem 15 . . . .
Problem 16 . . . .
Problem 17 . . . .
Problem 18 . . . .

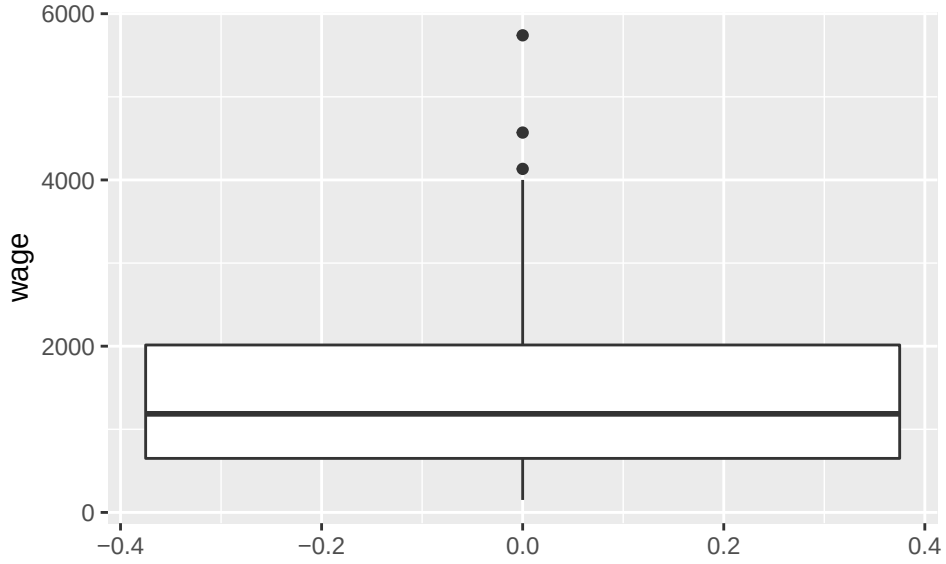
Theory

Problem 1 . . . .
Problem 2 . . . .
Problem 3 . . . .

NBA Wages - Practical

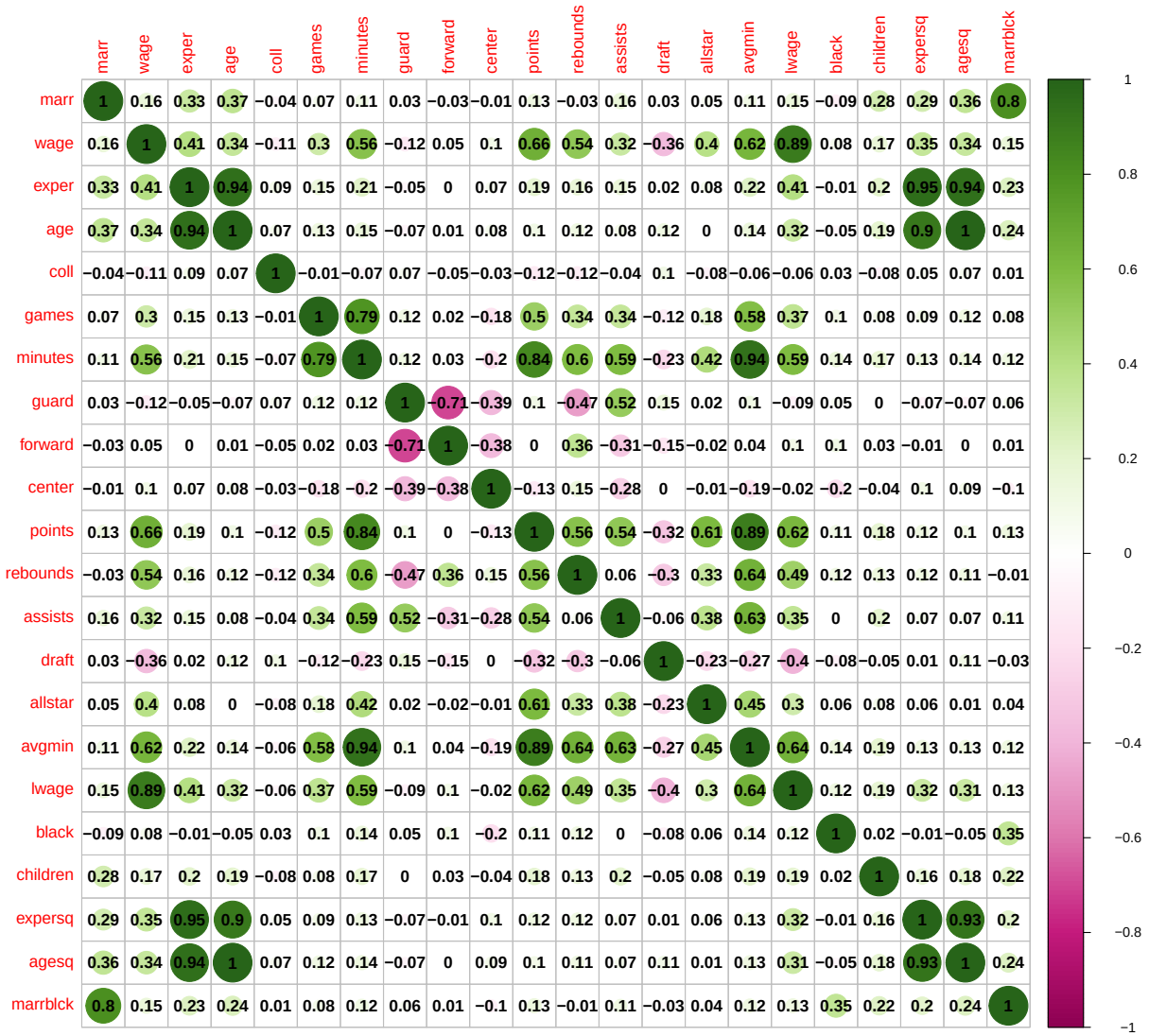
Problem 1

1. Draw a boxplot for the players' wages. (If you don't know how to do a given plot etc. then consult the user manual by clicking the 'Help' menu). Which way are the players' wages skewed? Towards infinity or zero?



Problem 2

2. Let's investigate the relationships between variables in our dataset. In practice if two variables are highly correlated with one another, then we may run into the problems caused by multicollinearity. This will make it hard for ordinary least squares to decipher the effect of one variable from another in a regression model. One way of investigating the relationships is via their correlation. If you select the option of 'Correlation matrix' from the 'View' menu, and include all the variables in the dataset, this will output the bivariate correlations between all variables in the NBA dataset. This can be a useful tool to allow one to get a quick handle on the data by seeing how strong the relationship is between different variables in your dataset. Are there any variables that are particularly highly correlated with experience? What would be the issue of including both of these measures in a regression with wages as the dependent variable?



`exper` is highly correlated with `age`, `expersq`, and `agesq`. Including more than one of these variables as regressors in our model will lead to a high level of multicollinearity. 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.

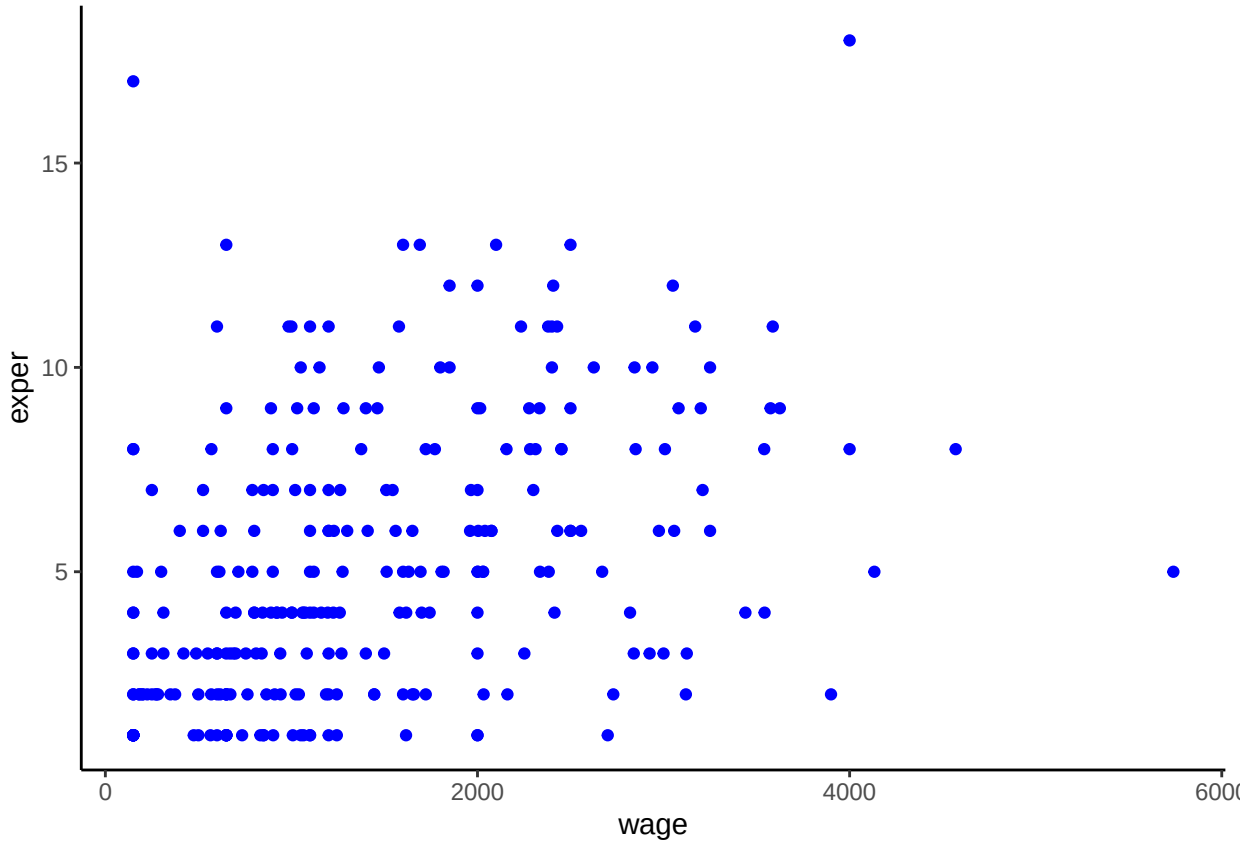
Problem 3

3. Another useful aspect of a correlation matrix is that it can give you a feel for which variables are correlated with your dependent variable. Which variables (other than the log of wages), show the highest correlation with wages?

`wage` has a relatively high correlation with `exper`, `minutes`, `points`, `rebounds`, and `avgmin`.

Problem 4

4. Graphically investigate whether players who are more experienced earn more. How strong is the correlation between these two variables?



The graph suggests players with more experience tend to be paid higher wages. The correlation between `exper` and `wage` is 0.4091764.

Problem 5

5. Create an ordinary least squares model which investigates how experience affects a player's wages. (Tip: You can save models for future use/viewing by clicking 'save as icon and close' in the model window. To access your model go to View → Icon view then click on the model. If you right click on a model you can change its name.)

Table 1: Bivariate Experience Model

	Dependent variable:
	wage
exper	120.317***
	(16.420)
	t = 7.327
	p = 0.000
Constant	807.932***
	(100.847)
	t = 8.011
	p = 0.000
Observations	269
R^2	0.167
Adjusted R^2	0.164
Residual Std. Error	913.956 (df = 267)
F Statistic	53.692*** (df = 1; 267) (p = 0.000)
Note:	*p<0.1; **p<0.05; ***p<0.01

Problem 6

6. What is the average wage increase associated with an increase in experience by one year implied by your model?

In this simple model, an increase in experience by one year is associated with an approximate $120K increase in average wage.

Problem 7

7. Do you think that the estimates of the effect of experience on wages is likely too big or too small? Which Gauss-Markov assumption is being violated, and why?

The zero conditional mean of the error term assumption is being violated such that $E[u|exper] \neq 0$. This is due to omitted variable bias in which independent variables that are correlated with both `exper` and `wage` are not included in the model. In this instance, it's likely the coefficient on `exper` is upwardly biased as both `exper` and `wage` are positively correlated with `points`, `rebounds`, and `assists`. In other words, experience is taking credit for these other omitted variables in our model.

Problem 8

8. Create another regression with wages as a dependent variable, and age as the independent variable (along with a constant). Does this imply that the effect of age is positive or negative?

Table 2: Bivariate Age Model

	Dependent variable:
	wage
age	100.955***
	(16.951)
	t = 5.956
	p = 0.000
Constant	−1,341.728***
	(467.888)
	t = −2.868
	p = 0.005
Observations	269
R^2	0.117
Adjusted R^2	0.114
Residual Std. Error	941.083 (df = 267)
F Statistic	35.470*** (df = 1; 267) (p = 0.000)
Note:	*p<0.1; **p<0.05; ***p<0.01

This model implies that the effect of age is positive such that an increase in age by one year is associated with an approximate $101K increase in wage.

Problem 9

9. What would would be the average wage implied by your model for an individual of 30? What about for a 90 year old? What is the problem with the latter estimate?

From our model, the average wage for an individual of 30 would be about $1687K while an individual of 90 would earn about $7744K. Clearly, it's ridiculous to think an NBA team would pay more for a player who's 90 in comparison with an individual who is 30. This emphasizes the need for caution when predicting out of sample results. (Out of sample here means that we currently do not have any data for the wages of 90 year old basketball players.)

Problem 10

10. How might we rectify the issue of the unrealistic estimates from the previous model?

First off, one should refrain from trying to predict out of sample results. We didn't have any (nor do they exist) entries for 90 year olds in the NBA. Thus, it's unwise to try to predict a 90 year old NBA player's salary. Additionally, to rectify the previous model we should add `agesq` as a regressor so that a diminishing marginal return to age can be expressed in our model.

Problem 11

11. Now create a regression with both experience and age in the model. What has happened to the sign of the coefficient on age? Why has this happened?

Table 3: Experience and Age Multiple Linear Regression Model

	Dependent variable:
	wage
age	-110.115^{**}
	(48.335)
	t = −2.278
	p = 0.024
exper	223.686^{***}
	(48.210)
	t = 4.640
	p = 0.00001
Constant	$3{,}295.293^{***}$
	(1,096.404)
	t = 3.006
	p = 0.003
Observations	269
R^2	0.183
Adjusted R^2	0.177
Residual Std. Error	906.868 (df = 266)
F Statistic	29.862^{***} (df = 2; 266) (p = 0.000)
Note:	*p<0.1; **p<0.05; ***p<0.01

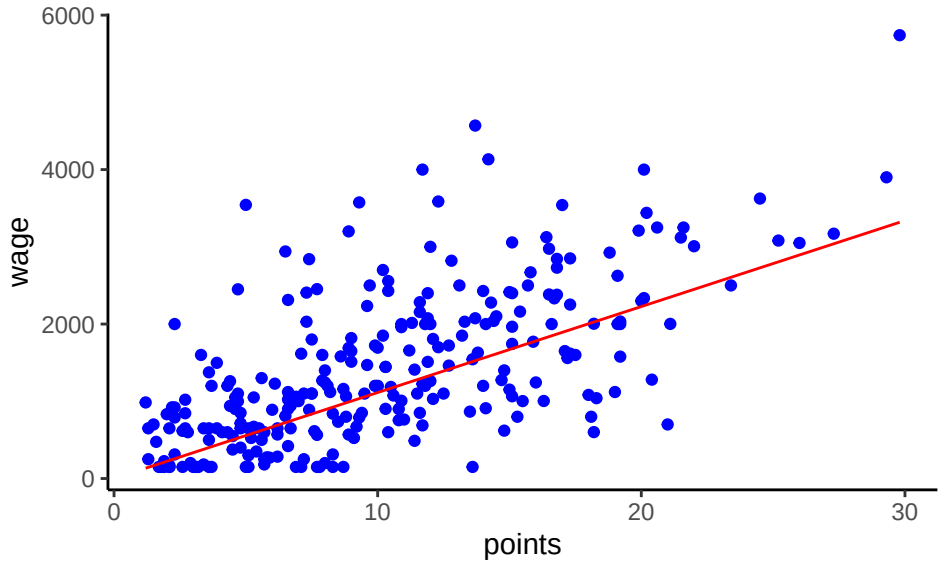
The coefficient on age changed from 100.9546125 in the bivariate model to -110.1153323 in the new model. This has occurred because of the high level of multicollinearity in this model. The correlation between `age` and `exper` is 0.9411652. Clearly, OLS is having difficulty deciphering between the effects of `age` and `exper` on `wage` in this model.

Problem 12

12. Let's now try to examine whether individuals who score more points tend to earn more by creating a regression of wages on points per game (and a constant). You can view a graph of the fitted regression line by navigating to Graphs → Fitted, actual plot from within the model window. The option you should select is 'against points'. From this you can see a graph of actual vs predicted wages vs points. What does your model suggest would be the increase in wages for an increase in 10 points per game?

Table 4: Bivariate Points Model

	Dependent variable:
	wage
points	111.330***
	(7.817)
	t = 14.243
	p = 0.000
Constant	287.101***
	(92.137)
	t = 3.116
	p = 0.003
Observations	269
R^2	0.432
Adjusted R^2	0.430
Residual Std. Error	755.069 (df = 267)
F Statistic	202.857*** (df = 1; 267) (p = 0.000)
Note:	*p<0.1; **p<0.05; ***p<0.01

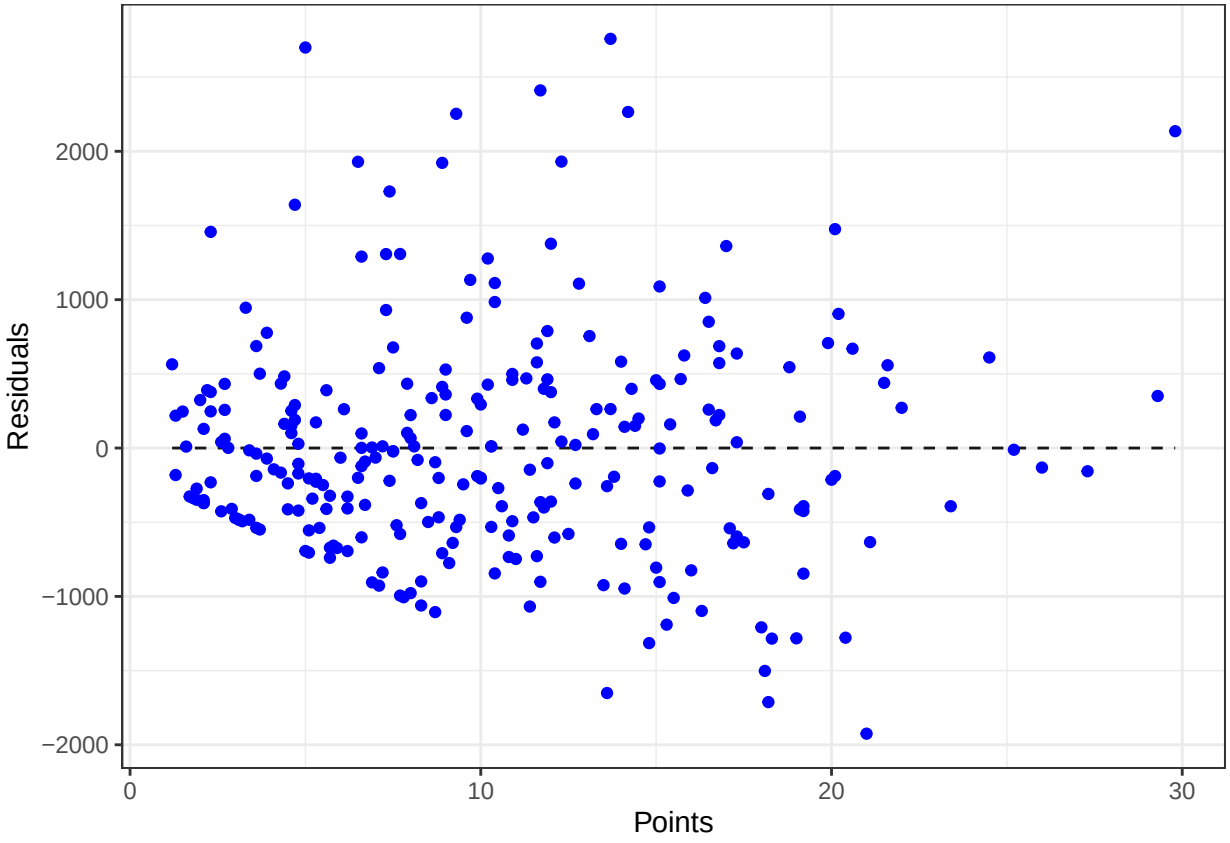


This model suggests an increase of $1113K in wages for an increase in 10 points per game.

Problem 13

13. You can look at a graph of the residuals (the estimated errors) from the regression, by clicking into the model (if you are not already in the model window), and navigating to Graphs → Residual plot. There are then a number of different options available, from seeing a plot of residuals against observation number to seeing a plot against the values of experience. Which of these plots should you use to graphically inspect for heteroscedasticity?

One should look at a plot of the residuals against the independent variable in the model, `points`.



Problem 14

14. Does there appear to be heteroscedasticity? What might be causing it? How might we rectify it?

There definitely appears to be some larger variance in the residuals for point values between 10 and 20 and smaller variance for point values greater than 20. This may be a result of model misspecification. We might rectify this by adding additional regressors to our model.

Problem 15

15. Do you think that the effect of points on wages predicted by your model is too high or low? Why might this be the case?

I think that the effect of points on wages predicted by your model is too high. I doubt a team is willing to pay an additional \$111330 for each additional point a player scores on average. Looking other factors that may play an important role in wage such as `rebounds` and `assists`, they too are positively correlated with `points` and `wage` meaning our coefficient on `points` is likely crediting `points` too much for additions to `wage`.

Problem 16

16. We are now going to create two new variables in Gretl: 'pointsq' and 'pointsc' equal to the square of points and its cube respectively. To do this navigate to Add → Define new variable... This allows a user to enter a formula for the construction of a new variable from an old one. For example to create 'pointsq', you can enter the formula: 'pointsq = points^2'. Here the '^' means 'raise that variable to the power'. Go ahead and create 'pointsq' and 'pointsc'.

```
nbasal <- nbasal %>%
  mutate(
    pointsq = points ** 2,
    pointsc = points ** 3
  )
head(nbasal)
```

```
##   marr   wage exper age coll games minutes guard forward center points rebounds
## 1    1 1002.5     4  27    4    77    2867     1       0      0   15.5      3.9
## 2    1 2030.0     5  28    4    78    2789     1       0      0   13.3      2.5
## 3    0  650.0     1  25    4    74    1149     0       0      1    5.5      3.3
## 4    0 2030.0     5  28    4    47    1178     0       1      0    7.3      5.1
## 5    0  755.0     3  24    4    82    2096     1       0      0   10.8      4.3
## 6    0 2014.5     9  31    4    82    1971     0       1      0   11.3      4.9
##   assists draft allstar   avgmin    lwage black children expersq agesq marrblck
## 1     4.5    19       0 37.23376 6.910252     1        0      16   729        1
## 2     8.8    28       0 35.75641 7.615791     1        1      25   784        1
## 3     0.2    19       0 15.52703 6.476973     1        0       1   625        0
## 4     1.5     1       0 25.06383 7.615791     1        0      25   784        0
## 5     2.6    24       0 25.56098 6.626718     1        0       9   576        0
## 6     1.5     4       0 24.03658 7.608126     1        0      81   961        0
##   pointsq  pointsc
## 1  240.25 3723.875
## 2  176.89 2352.637
## 3   30.25  166.375
## 4   53.29  389.017
## 5  116.64 1259.712
## 6  127.69 1442.897
```

Problem 17

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

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

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

(a) Which of these regressions has the highest value of R-squared? What does this mean?

(b) What is the interpretation of the coefficient on 'pointsc' in the last regression model?

(c) Which of these regressions has the highest value of adjusted R-squared?

(d) Out of the three specifications, which would you prefer?

Table 5: Comparing Points Models

	Dependent variable:		
	wage		
	(1)	(2)	(3)
points	111.330***	90.119***	211.174***
	(7.817)	(26.648)	(65.716)
	t = 14.243	t = 3.382	t = 3.213
	p = 0.000	p = 0.001	p = 0.002
pointsq		0.872	−9.814*
		(1.047)	(5.409)
		t = 0.833	t = −1.814
		p = 0.406	p = 0.071
pointsc			0.258**
			(0.128)
			t = 2.013
			p = 0.046
Constant	287.101***	382.550***	47.203
	(92.137)	(147.102)	(221.690)
	t = 3.116	t = 2.601	t = 0.213
	p = 0.003	p = 0.010	p = 0.832
Observations	269	269	269
R^2	0.432	0.433	0.442
Adjusted R^2	0.430	0.429	0.435
Residual Std. Error	755.069 (df = 267)	755.503 (df = 266)	751.205 (df = 265)
Note:			*p<0.1; **p<0.05; ***p<0.01

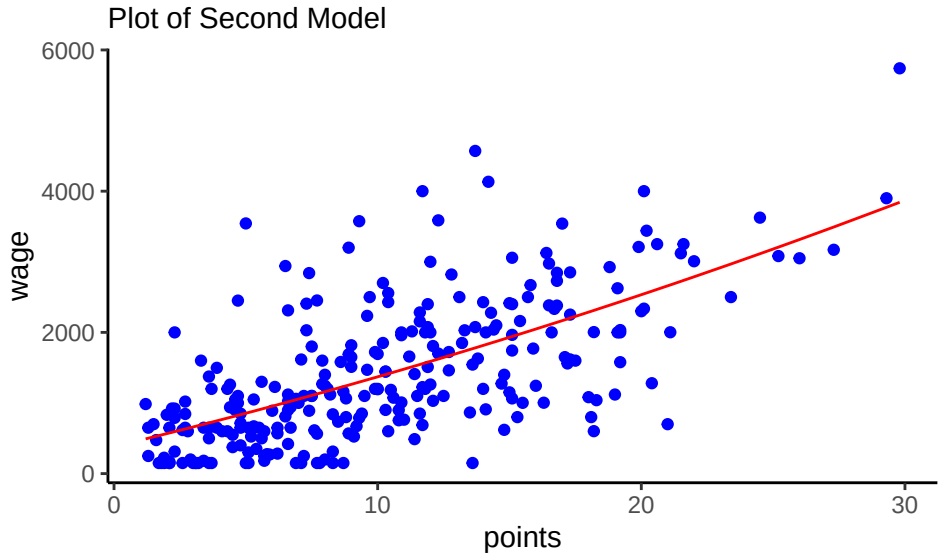
17. (a) The third model, $wage_i = \alpha + \beta_1 points_i + \beta_2 pointsq_i + \beta_3 pointsc_i$, has the greatest coefficient of determination (as expected because R^2 doesn't decline as regressors are added to a model). This mean the independent variables in the model explain about 44.2

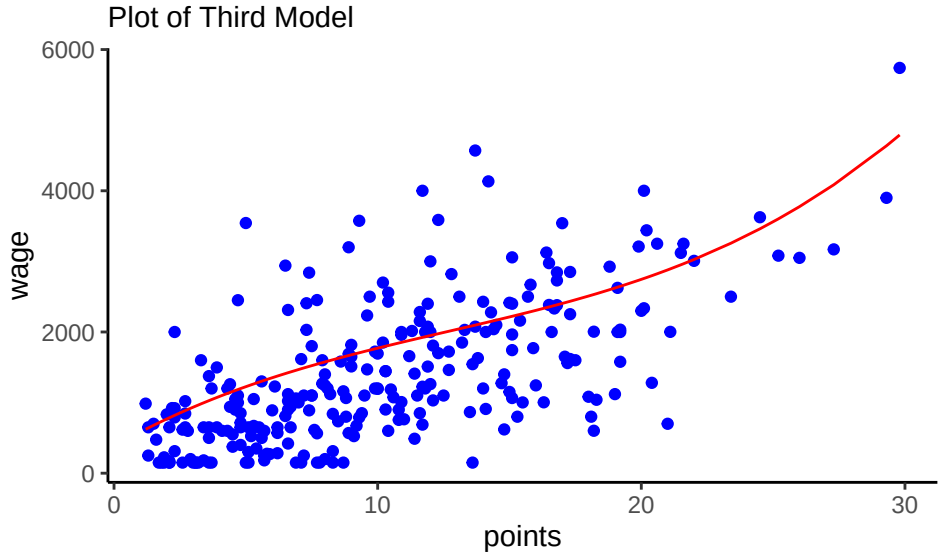
(b) For each additional cubed point a player scores on average, we expect to see an increase of $258 to

a player's wage. It also implies a diminishing diminishing return of points on wage.

(c) The third model, $wage_i = \alpha + \beta_1 points_i + \beta_2 pointsq_i + \beta_3 pointsc_i$, also has the greatest $\bar{R}^2$, value.

(d) Solely based on $\bar{R}^2$, I would prefer the third model; however, based on my intuition I would've suggested the second model because I believe there's an increasing return in wage to points.





Problem 18

18. Freestyle: try to create a model which you believe captures explains players wages best in terms of other attributes.

Table 6: Comparing Models

	Dependent variable:			
	wage			
	(1)	(2)	(3)	(4)
exper	99.021**		102.551**	
	(41.012)		(41.073)	
	t = 2.414		t = 2.497	
	p = 0.017		p = 0.014	
expersq	−1.245		−1.508	
	(2.952)		(2.957)	
	t = −0.422		t = −0.510	
	p = 0.674		p = 0.611	
age		−53.060		−43.924
		(157.268)		(157.470)
		t = −0.337		t = −0.279
		p = 0.737		p = 0.781
agesq		2.271		2.110
		(2.729)		(2.733)
		t = 0.832		t = 0.772
		p = 0.407		p = 0.441
points	78.238***	81.151***	46.977*	54.105*
	(10.758)	(10.833)	(27.624)	(27.782)
	t = 7.272	t = 7.491	t = 1.701	t = 1.947
	p = 0.000	p = 0.000	p = 0.091	p = 0.053
pointsq			1.196	1.036
			(0.973)	(0.980)
			t = 1.228	t = 1.057
			p = 0.221	p = 0.292
rebounds	79.661***	82.562***	81.488***	84.216***
	(18.645)	(18.747)	(18.687)	(18.808)
	t = 4.272	t = 4.404	t = 4.361	t = 4.478
	p = 0.00003	p = 0.00002	p = 0.00002	p = 0.00002
assists	7.648	17.398	15.657	24.471
	(25.540)	(25.592)	(26.335)	(26.446)
	t = 0.299	t = 0.680	t = 0.595	t = 0.925
	p = 0.765	p = 0.498	p = 0.553	p = 0.356
Constant	−203.966*	−86.794	−86.371	−106.236
	(122.349)	(2,231.721)	(155.255)	(2,231.299)
	t = −1.667	t = −0.039	t = −0.556	t = −0.048
	p = 0.097	p = 0.970	p = 0.579	p = 0.963
Observations	269	269	269	269
R^2	0.551	0.544	0.553	0.546
Adjusted R^2	0.542	0.536	0.543	0.536
Residual Std. Error	676.431 (df = 263)	681.328 (df = 263)	675.777 (df = 262)	681.176 (df = 262)

Note: *p<0.1; **p<0.05; ***p<0.01

All of the models I tested have a similar coefficient of determination and adjusted R^2 with the third model marginally taking the cake in each. Instantly, solely based on intuition, I would drop the second and fourth model from consideration because the desire for age is negative but has increasing returns. This is likely due to better players staying in the NBA longer, but it doesn't make sense intuitively. The other two models are as I expected where experience is valued but at a decreasing rate. Personally, I would choose the third model because I believe there's an increasing return to points. Additionally, while I could go searching for a model that provides a greater $\bar{R}^2$, I believe this model and the coefficients makes sense intuitively.

Theory

Problem 1

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) What is the interpretation of the coefficient on $windows_i$?

(b) What are the causes of endogeneity in this model?

(c) Do you think that this coefficient over or under estimates the effect of broken windows 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:

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

Why has the coefficient on $windows_i$ fallen relative to what it was before in this new model? Not only this, but neither coefficients on $windows_i$ or $emerg_i$ are statistically significant. However, both coefficients jointly appear to be significant in determining house prices. Why might this be?

1. (a) It means, for each additional broken window on a block, the average property value drops $10K.

(b) The causes of endogeneity in this model are the omitted variables. There are several factors that are likely correlated with the number of broken windows on a block and housing prices. Such variables may include the number of crimes in the area, the level of wealth in the area, and many other factors. This model may also suffer from reverse causality in which cheaper (or possibly more expensive) homes are more likely to suffer from broken windows.

(c) It likely overstates (the actual value is less negative) the effect of broken windows on house prices because the coefficient likely captures some of the effect of the omitted variables discussed in the previous part.

(d) The new model now attempts to capture some of the effect of crime in an area, which is reducing some (or possibly all) of the bias on the broken windows coefficient. Separately, neither of the coefficients are statistically significant because there's likely a high level of multicollinearity in this model.

Problem 2

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) Can you prove that this implies that $\mathbb{E}[\varepsilon_i X_i] = 0$?

   (b) Does this imply that X_i and ε_i are uncorrelated? Prove it.

   (c) Does the covariance between X_i and ε_i being zero imply independence of these variables?

2. (a) The law of iterated expectations states $E[X_i\varepsilon_i] = E[E[X_i\varepsilon_i|X_i]] = E[X_i E[\varepsilon_i|X_i]] = E[X_i \cdot 0] = 0$

   (b)

$$\begin{aligned} corr(X_i, \varepsilon_i) &= \frac{cov(X_i, \varepsilon_i)}{\sigma_X \sigma_\varepsilon} \\ &= \frac{E[X_i\varepsilon_i] - E[X_i]E[\varepsilon_i]}{\sigma_X \sigma_\varepsilon} \\ &= \frac{0 - E[X_i] \cdot 0}{\sigma_X \sigma_\varepsilon} = 0 \end{aligned}$$

   (c) No, a covariance of zero doesn't necessarily imply independence. It's possible for two random variables to have a zero covariance but still be dependent on one another.

Problem 3

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 their progress in the language. 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) **Do you think that β fairly represents the effect of an incremental point of IQ on an individual's performance in the end of day test? Why/why not?**

   (b) **The above equation is amended to include any other relevant explanatory variables. The researcher is aware that IQ is not a perfect measure of an individual's 'language intelligence'. However, she supposes that it is an adequate proxy - meaning that it is not a biased estimate of 'language intelligence'. Will the least squares estimator $\hat{\beta}$ be unbiased?**

   (c) **Prove either way your answer for the last question.**

3. (a) No, I don't think β fairly represents the effect of an incremental point of IQ on an individual's performance in the end of day test. There are several other factors that may be correlated to both IQ and an individual's test score. Wealth, age, and GPA may be omitted variables from the true population model. Additionally, it may be the case that some individuals already know a language similar to French while others may not.

   (b) This new model likely underestimates the effect of 'language intelligence' on their language learning abilities. An individual with a higher IQ doesn't necessarily signify a 1:1 ratio of return on IQ to score.

   (c) Suppose $IQ_i = lang\_abil_i + \varepsilon_i$. By definition, we have

$$cov(lang\_abil_i, \varepsilon_i) = 0$$

   Thus, we know the covariance between IQ and the error term is

$$cov(IQ_i, \varepsilon_i) = cov(lang\_abil_i + \varepsilon_i, \varepsilon_i) = 0 + \sigma_\varepsilon^2 = \sigma_\varepsilon^2$$

   The true population model is

$$score_i = \gamma_0 + \gamma_1 lang\_abil + \eta_i$$

   , which can be rewritten as

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

   or

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

   This violates the exogeneity assumption because $cov(IQ_i, \varepsilon_i) = \sigma_\varepsilon^2 \neq 0$ and thus β is likely biased.