PADERBORN UNIVERSITY
The University for the Information Society

W4479 – Econometrics Summer 2021

Applied Project No. 3

Further Topics in Regression, Ch09 – Ch13

Lecturer:	Prof. Yuanhua Feng
Project beginning:	25.08.2021
Project due:	27.09.2021

Question No.	1	2	3	4	Total
Points awarded					
Points max.	25	25	25	25	100

Submitted by Group No. 10:

Name	**Matriculation Number**
Manizha Hakimova	**6874923**
Ratul Chowdhury	**6899992**
Mohammad Sirajul Islam	**6904392**
Kerstin Zemlianski	**7246469**
Alexander Szawiolo	**7214404**

Exercise 1 – Application of Dummy Variables in Regression

a)

In this project the CPSSW9204 data from the AER package in R has been used to show the different applications of dummy variables in regression. We have filtered the dataset to represent data for the year 1992 only to simplify our analysis on the effects of dummy variable. The use of two dummy variables will also help us determine whether there is a level of education and gender dimension to the earnings during this period. The filtered dataset consists of 7602 observations and 5 variables.

b)

The categorical variable degree is in factor format and has been converted into a dummy variable to simplify our regression.

Categorical variable values	Dummy variable values ($\boldsymbol{degree}$)
highschool	1
bachelor	0

The dummy variable ***degree*** is represented by D_1 in the regression model, where the categorical value **highschool** has been encoded to 1, while the value of **bachelor** has been encoded to 0.

- Regression model:

$$\widehat{Y_i} = \beta_0 + \beta_1 X_i + \beta_2 D_{1i} + u_i$$

Dummy Variable ($\boldsymbol{D_1}$)	Model
1	$\widehat{Y_i} = (\beta_0 + \beta_2) + \beta_1 X_i + u_i$
0	$\widehat{Y_i} = \beta_0 + \beta_1 X_i + u_i$

As the dummy variable degree has been regressed in the levels, the value of y-intercept changes but the value of slope remains unchanged.
Therefore, the plot of the different regression lines at different values of D_1 should be parallel to each other.

- Estimated model:

$$\widehat{earnings_i} = 3.974 + 0.345\, age_i - 4.266\, degree_i + u_i$$

Dummy Variable ($\boldsymbol{Degree}$)	Model
1	$\widehat{earnings_i} = -0.292 + 0.345\ age_i$
0	$\widehat{earnings_i} = 3.974 + 0.345\ age_i$

From the estimated regression model, we can see that the value of slope remains unchanged with different values of dummy variable degree. However, the y-intercept has changed. We can conclude that a person with only a highschool degree has an adverse effect on their earnings, whereas a person with a bachelor's degree an additional 3.974 units. Additionally, all the coefficients for this model are significant.

c)

To analyze the effect of regressing earnings on **age** and **degree** in the levels and slope, we include an interaction term i.e., the product of **age** and **degree** into the model.

- Regression model:

$$\widehat{Y_i} = \beta_0 + \beta_1 X_i + \beta_2 D_{1i} + \beta_3 (X_i * D_{1i})$$

Dummy Variable ($\boldsymbol{D_1}$)	Model
1	$\widehat{Y_i} = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)\ X_i + u_i$
0	$\widehat{Y_i} = \beta_0 + \beta_1 X_i + u_i$

In this regression model, we can see that there is change in both the y-intercept and the slope with different values of the dummy variable.

- Estimated model:

$$\widehat{earnings_i} = -1.154 + 0.519\ age_i + 4.170\ degree_i - 0.285(age_i * degree_i)$$

Dummy Variable ($\boldsymbol{D_1}$)	Model
1	$\widehat{earnings_i} = 3.016 + 0.234\ age_i$
0	$\widehat{earnings_i} = -1.154 + 0.519\ age_i$

Like the regression model, we can observe that both the y-intercept and slope changes after introduction of the interaction term. From the estimated model, we can conclude that people with highschool degree can start work early compared to those pursing a bachelor's degree and earn in the initial years of work. However, the disparity in earning works to the favor of those pursuing a bachelor's degree as the age variable increases. Among people holding a highschool degree, 1 year increase in age only increases the earning by 0.234 units. On the other hand, people with a bachelor's degree can increase their earnings by 0.519 units with 1 year increase in age.

d)

In this segment, we introduce the categorical variable gender into the regression model. Like the degree variable, the variable is in factor format and has been converted into a dummy variable to simplify our regression.

Categorical variable values	Dummy variable values ($\boldsymbol{gender}$)
female	1
male	0

In the dummy variable ***gender***, the categorical value ***female*** has been encoded to 1, while the value of ***male*** has been encoded to 0.

Estimated model:

$$\widehat{earnings_i} = -2.5527 + 0.5947\ age_i + 3.5355\ degree_i + \ 4.4438\ gender_i - 0.2642\ \mathrm{I}(age_i * degree_i) - 0.2133\ \mathrm{I}(age_i * gender_i) - 0.1285\ \mathrm{I}(gender_i * degree_i)$$

degree	gender	Estimated Models
1	1	$\widehat{earnings_i} = 5.2981 + 0.1172\ age_i$
1	0	$\widehat{earnings_i} = 0.9828 + 0.3305\ age_i$
0	1	$\widehat{earnings_i} = 1.8911 + 0.3814\ age_i$
0	0	$\widehat{earnings_i} = -2.5527 + 0.5947\ age_i$

In this regression model, all the coefficients are significant. From the different regression outcomes for different combinations of dummy variables degree and gender, we observe that a male opting for a bachelor's degree can raise their earning the highest with 1 year increase in age. However, they must forgo the initial income they would have earned if they chose to work instead of pursuing a bachelor's degree. Because of introducing the gender dummy variable, we can also see a disparity in earnings between male and female who have only obtained a highschool degree. Males who have obtained only a highschool degree can earn more in the long run with increase in age compared to females.

Exercise 2 – Discussion on Multicollinearity Problem

a)

The CASchool dataset contains data on test performance school characteristics and student demographic backgrounds for school districts in California. This data frame contains 420 observations on 14 variables. We are using 10 variables containing 420 observations. Grades variable contains factor indicating grade span of district, students contain total enrolment students, teachers contain number of teachers, calworks contain percent qualifying for calworks (income assistance), lunch contains percent qualifying for reduced-price lunch, computer contains total number of computers, expenditure contains total expenditure per student, income indicates district average income (in USD 1,000), english contains percent of english learners whose 2nd language is English, read contains average reading score, math contains average math score. The data is for two-year 1998 and 1999. Test scores are on the Stanford 9 standardized test administered to 5th grade students. School characteristics include enrolment, number of teachers, number of computers per classroom, and expenditures per student. Demographic variables for the students are averaged across the district. The demographic variables include the percentage of students in the public assistance program CalWorks, the percentage of students that qualify for a reduced-price lunch, and the percentage of students that are English learners whom English is a second language. Coefficients of expected model:

	Estimate	***Standard Error***	***T-value***	***Pr (>|t|)***
(Intercept)	-1.043e+02	1.714e+01	-6.089	2.63e-09
Grades	5.252e-01	6.427e-01	0.817	0.41426
Students	-1.682e-03	7.640e-04	-2.202	0.02823
Teachers	3.810e-02	1.682e-02	2.265	0.02401
Calworks	4.240e-03	3.046e-02	0.139	0.88936
Lunch	-1.294e-01	2.094e-02	-6.181	1.54e-09
Computer	1.177e-04	1.460e-03	0.081	0.93581
Expenditure	2.497e-03	3.805e-04	6.561	1.62e-10
English	1.094e-01	1.923e-02	5.688	2.44e-08
Read	7.248e-02	3.835e-02	1.890	0.05948
Math	9.480e-02	3.148e-02	3.012	0.00276
$\boldsymbol{R^2}$	0.6389			
$\boldsymbol{\bar{R}^2}$	0.6301			

Expected model:

$\widehat{Income}_i$ = -104.3+0.5252 Grades$_{\mathrm{i}}$-0.0017 Students$_{\mathrm{i}}$+0.038 Teachers$_{\mathrm{i}}$+0.00424 Calworks$_{\mathrm{i}}$-0.129 Lunch_i+0.00012 Computer_i+0.002497 Expenditure_i+0.1094 English_i+0.0725 Read_i+0.0948 Math_i

b)

VIF and TOL of independent variables:

Independent Variables	***VIF***	***TOL***
Grades	1.114986	0.896872141
Students	193.905738	0.005157145
Teachers	216.691866	0.004614848
Calworks	2.640542	0.378710164
Lunch	6.996165	0.142935452
Computer	9.009284	0.110996608
Expenditure	1.262595	0.792019332
English	2.682700	0.372758733
Read	12.900539	0.077516142
Math	7.561026	0.132257176

From this chart of VIF and TOL values, we can clearly see that independent value Students, Teachers and Read has the VIF values more than 10 and TOL values less than .1. So, we can easily eliminate these three variables from our expected model to reduce the multicollinearity of our expected model. So, our expected model will be –

$\widehat{Income}_i$ = -87.769+0.414 Grades$_{\mathrm{i}}$+0.0115 Calworks$_{\mathrm{i}}$-0.147 Lunch_i+0.00139 Computer_i+0.002852 Expenditure_i+0.099 English_i+0.141 Math_i

c)

After removing these three variables which have values in VIF more than 10 and Values in TOL less than .1, we get a model which has less multicollinearity in data. Without these three independent variables, all other variables have stable coefficient. We have four linear regression models in total. On one regression, we regress income on all other variables including students, teachers and read. We have another three regressions where we regress income on other variables excluding students, teachers and read. Coefficients of students, teachers and read differs a lot in these four linear regression models, but other independent variable's coefficients are a bit more stable in these four models. Coefficients of grades, calworks, lunch, computers, expenditure, English and math are more stable than the coefficients of independent variables of students, teachers and read. In comparison to models, “grades”, “calworks”, “lunch”, “computer”, “expenditure”, “English” and “math” stay significant and the estimated coefficients but also the standard error just changed marginally. So, the impact of these parameters on “income” is not really changed. But, “students”, “teachers” and “read” do not stay significant and standard error and coefficients are changed marginally. After removing multicollinearity variables, we find this expected model of expected income:

$\widehat{Income}_i$ = -87.769+0.414 Grades_i+0.0115 Calworks_i-0.147 Lunch_i+0.00139 Computer_i+0.002852 Expenditure_i+0.099 English_i+0.141 Math_i

d)

Selecting models according to AIC or BIC:

MODEL NO.	*MODELS*	*AIC*	*BIC*
1	**Model regressed with all variables**	2448.295	2496.778
2	**Model regressed without students**	2451.244	2495.687
3	**Model regressed without teachers**	2451.533	2495.975
4	**Model regressed without read**	2449.947	2494.39
5	**Model regressed without multicollinearity**	2451.427	2487.79

Among the five regression models, the model which has the lowest AIC value, making it the best model when assessing based on AIC. Same criteria will apply for BIC value, which model has the lowest value will be selected. If we consider the AIC values from the chart, we will select model 1 which regressed with all variables. This model has the lowest AIC value. If we consider the BIC values, then we will select model 5 which regressed without multicollinearity. This model also has the lowest BIC value among all other models.

Exercise 3 – Linear Regression with Heteroscedasticity

a)

The dataset CPSSWEducation contains data of the Current Population Survey of the Bureau of Labor Statistics. The dataset *CPSSWEducation* is provided by Stock and Watson (2007) and obtained from the package *AER* in R. The data frame contains 2,950 observations on the earnings distribution in the United States in 2004 for full-time workers aged 29-30 with a college degree. The variables age and gender are not taken into consideration to build a simple linear regression model for the purpose of this exercise. The simple linear regression model used for investigating the problem of heteroscedasticity represents the relationship between average hourly earnings (dependent variable) and the number of years of education (independent variable). A scatterplot of this regression is illustrated below to see if heteroscedasticity is clearly visible.

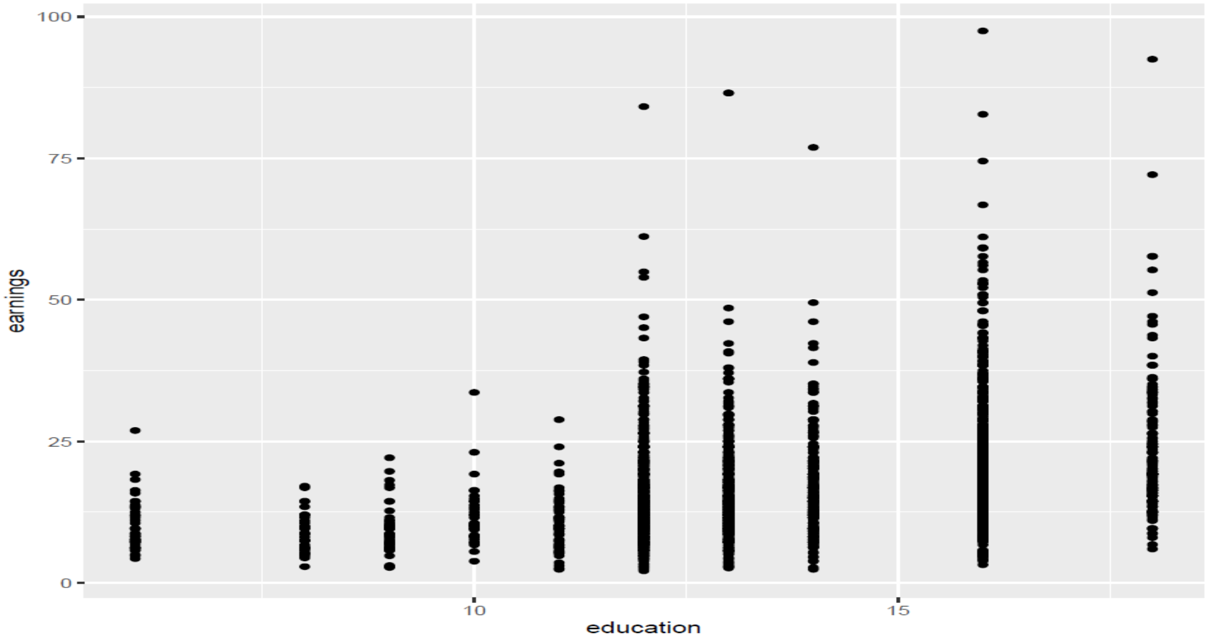

Figure 1: Visibility of Heteroscedasticity in the model

Taking a first look on the scatterplot It can be clearly noted that a higher number of years of education leads to a higher variance in contrast to a low number of years of education. Thus, we come to the result that there is visible heteroscedasticity in the dataset. A suitable test in form of the park test will be applied in subpart c) to further emphasize the existence of heteroscedasticity.

b)

The simple linear regression model (OLS estimators) is as follows: $\widehat{earnings_i} = -3.13437 + 1.46693 * education_i$

	Estimate	Standard Error	t-value
(Intercept)	-3.13437	0.95925	- 3.268
education	1.46693	0.06978	21.021
$\bar{R}^2$	0.1301		

The 95% confidence interval for education	[1.330098, 1.603753]

The standard errors as well as the 95% confidence interval are needed to do a comparison between the model with OLS estimators and the two models with WLS estimators which will be reported and compared in detail in subpart d).

Furthermore, the residuals of this regression are obtained. Details regarding the residuals and the calculation can be obtained from the attached r-file.

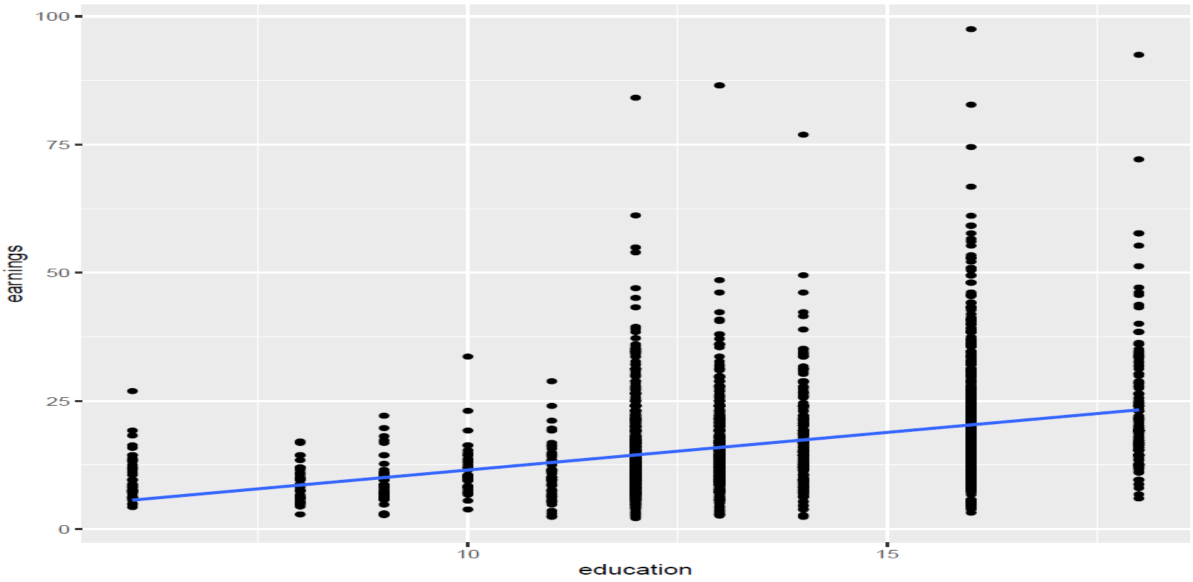

Figure 2: Simple Linear Regression with OLS estimators

The plotted simple linear regression shows that the regression is strongly affected by the heteroscedasticity. This is especially visible for larger number of years of education and can be detected by a widen out area in which the error terms are scattering around the regression line. Due to the increased variance for higher educated people, the OLS estimators cannot be minimized variance estimators anymore and are therefore no more the best linear unbiased estimators. The assumption for homoscedasticity therefore does not hold any more. Regarding the consequences of heteroscedasticity, it has no impact on the point estimation of the parameters itself but causes problems regarding the efficiency of the estimation of the standard errors of the regression coefficients.

c)

In the subparts a) and b) we did observe the existence of heteroscedasticity in a graphic used for illustrating the simple linear regression. Now, a formal method in form of the park test will be applied to test the existence of heteroscedasticity in a different way. This test is suitable choice as the variances are usually not known. Therefore, a double-log model including the relationship between the squared residuals from the linear regression of subpart b) and the independent variable education is taken into consideration and the following regression will be run:

$$ln\ \hat{u}_i^2 = \ln \sigma^2 + \beta \ln X_i + v_i$$

This leads to the following regression model:

$$\widehat{\ln \hat{u}_i^2} = = -3.0875 + 2.1954 \ln(\text{education}_i)$$

	Estimate	Standard Error	t-value
(Intercept)	-3.0875	0.6010	-5.138
ln(education)	2.1954	0.2314	9.487

The $\bar{R}^2$ is 0.0293 and the p-value takes a value of < 2.2e-16.

To test for heteroscedasticity, we test for the hypothesis H_0: Homoscedasticity and H_1: Heteroscedasticity setting $\alpha = 0.05$, using the park test. As the p-value of $< 2.2e - 16 \ < \ 0.05$, H_0 can be rejected and therefore, a problem of heteroscedasticity does exist. Furthermore, β is highly significant which further strengthens the existence of heteroscedasticity.

d)

To increase the efficiency of the estimation of the standard errors and to minimize the variance, a weighted linear regression with weights 1/X and $1/X^2$ are used. To stabilize the regression model by using WLS estimators, smaller weights are used for larger variances and vice versa.

The weighted simple linear regression model with 1/X is as follows:

$\widehat{earnings_i} = -3.13437 + \ 1.46693 * education_i$

WLS estimation with 1/X			
	Estimate	Standard Error	t-value
(Intercept)	- 1.82265	0.83978	-2.17
education	1.37012	0.06307	21.72
$\bar{R}^2$	0.1377		

The 95% confidence interval for education	[1.246456, 1.493785]

The weighted simple linear regression model with $1/X^2$ is as follows:

$\widehat{earnings} = -0.16956 + \ 1.24378 education_i$

WLS estimation with $1/X^2$			
	Estimate	Standard Error	t-value
(Intercept)	- 0.16956	0.70268	- 0.241
education	1.24378	0.05495	22.634
$\bar{R}^2$	0.1478		

The 95% confidence interval for education	[1.136031, 1.351531]

Comparing the models with WLS estimators to the one with OLS estimators, it can be noted that the standard errors for the intercept and slope coefficients are lower for the weighted linear regression models. The ranges of the confidence intervals are also smaller for the weighted linear regression models. Therefore, we can conclude that the WLS estimators are more accurate or more efficient in estimating the standard errors of the regression coefficients. Furthermore, in terms of model selection criterion, the WLS estimation with weights $1/X^2$ has the lowest standard errors, smallest ranges and highest $\bar{R}^2$. In this case, the WLS with weights $1/X^2$ are variance minimizing as well as it is stabilizing the regression model the most.

Question 4 – Omitted variable bias

a)

The fitted model for regression 1 ran will be:

$$\widehat{Price}_i = 67745.47 \quad + 12135.27 * LotSize_i - \ 495.21 * Age_i + \ 21575.31 * \ Rooms_i$$

	Estimate	Standard Error	t-value
(Intercept)	67745.47	6747.5	10.039
LotSize	12135.27	2837.66	4.277
Age	-495.21	66.01	−7.502
Rooms	21575.31	858.42	25.134
$\bar{R}^2$	0.3128		

Variable $LotSize_i$ for example is relevant and significant for the regression analysis because, the correlation $(Price_i + LotSize_i) \neq 0$, correlation $(LotSize_i + Rooms_i\,) \neq 0$ and correlation $(LotSize_i + Age_i) \neq 0$. Therefore, the omission of $LotSize_i$ from the regression analysis will make the parameter estimation of $\alpha_2\ and\ \alpha_3$ to bebiased. The omission of any of the three variables will make the results of the regression to be biased because the true correlation is not equal to zero (0).

The 95% confidence intervals for $\alpha_1, \alpha_2\ and\ \alpha_3$ are given below:

95% confidence interval for $\alpha_1 LotSize_i$	[6569.672, 17700.87]
95% confidence interval for $\alpha_2 Age_i$	[-624.6861, -365.7378]
95% confidence interval for $\alpha_3 Rooms_i$	[19891.67, 23258.96].

b)

To test if there is a misspecification causing an omitted variable bias in the regression model when omitting the variable LivingArea, the Durbin-Watson statistic d is calculated.

The Durbin Watson test represents a test for autocorrelation setting the null hypothesis H_0 that there is no autocorrelation among regression residuals and that the model is without specification error. Critical values d_L and d_U must be calculated to determine if the value for the d-statistic will lead to a rejection of the null hypothesis or not. A clear specification error exists if d is smaller than both d_L and d_U. Critical values can be obtained from a Durbin-Watson table, indicating n = 1734, k = 3 and α = 0.05. The corresponding critical values are $d_L = approx. 1.918\ and\ d_U = approx. 1.925$.

Due to the fact, that the d-statistic takes a value of 1.6308, H_0 needs to be rejected at the 5% significant level in favor of the alternative hypothesis that the autocorrelation among residuals is greater than zero. This is due to d < d_L as well d < d_U. Also the p-value of 6.295e-15 < 0.05. To draw a conclusion, model 1 is with specification error and the variable *LivingArea* should therefore be included in the regression model to avoid an omitted variable bias.

c)

The fitted model regression 2 will be:

$$\widehat{Price}_i = 19987{,}351 \quad + 6040.360 * LotSize_i - \; 241.377 * Age_i + \; 1038.985 * \; Rooms_i \\ + 107.296 * \; LivingArea_i$$

	Estimate	Standard Error	t-value
(Intercept)	19987.351	5941.251	3.364
LotSize	6040.360	2394.774	2.522
Age	-241.377	56.259	−4.290
Rooms	1038.985	1050.811	0.989
LivingArea	107.296	3.993	26.871
$\bar{R}^2$	0.515		

95% confidence interval for $\beta_1 LotSize_i$	[1343.401, 10737.32]
95% confidence interval for $\beta_2 Age_i$	[-351.719, -131.0346]
95% confidence interval for $\beta_3 Rooms_i$	[-1.022.009, 3099.979].

Evidently, the confidence intervals significantly differ from each other. The 95% confidence intervals show that the ranges in model 2 are smaller for β_1 and β_2 but larger for β_3. This leads to the fact that the estimation of the standard errors of β_1 and β_2 is more efficient after adding *LivingArea* into the model but becomes less accurate for β_3. Once the variable *LivingArea* is included which is a relevant one in this case, the already included variable *Rooms* may become unnecessary. Comparing the variable *Rooms* in both regression models it can be noted that the t-value falls below the critical value of 2, according to the rule of thumb for a two-tailed test with α = 0.05, in the second model. But as the standard errors are smaller for β_1 and β_2 in model 2 as well as the $\bar{R}^2$ of 0.5161 is higher compared to a $\bar{R}^2$ of 0.314 for model 1, the latter is to be preferred.

d).

As the model (2) still not considers all variables included in the dataset, there may be a possibility that at least one of the four variables *NewConstruct, Beedrooms, LandValue and CentralAir* is important and should be included. The Durbin-Watson test is applied again to find out if at least one of the mentioned variables needs to be included to avoid misspecifications.

The hypothesis is set H_0: $model\ is\ correclty\ specified,$

H_1: $true\ autocorrelation\ is\ greater\ than\ 0.$

Critical values can be obtained from a Durbin-Watson table, indicating n = 1734, k = 4 and α = 0.05. The corresponding critical values are $d_L = approx. 1.917\ and\ d_U = approx. 1.926.$

Variable	d-statistic	p-value	H_0
NewConstruct	1.9756	0.3056	cannot be rejected
Bedrooms	1.9642	0.2237	cannot be rejected
LandValue	1.4917	2.2e-16	can be rejected
CentralAir	1.9491	0.1442	cannot be rejected

The table shows that only the null hypothesis for the variable *LandValue* can be rejected as its d-statistic is below the critical values as well as its p-vakue is below 0.05. Only In this case, there is a misspecification in the model. In all the three other cases, the model is assumed to be correctly specified. This is due to the d-statistic being higher than $d_L = 1.917$ and $d_U = 1.926$ in all three cases. Besides, the individual p-values are greater than 0.05. Therefore, only the variable *LandValues* should be added to the model.

e)

Including the variable *LandValue,* the fitted regression model (3) is as follows:

$$\widehat{Price}_i = 28742.6488 + 6598.2308 LotSize_i - 304.7849 Age_i + 1491.2280\ Rooms_i + 82.1015 LivingArea_i + 0.9768 LandValue_i$$

	Estimate	Standard Error	t-value
(Intercept)	28742.6488	5321	5.401
LotSize	6598.2308	2139	3.085
Age	-304.7849	50.33	−6.056
Rooms	1491.2280	938.5	1.589
LivingArea	82.1015	3.762	21.823
LandValue	0.9768	0.04654	20.988
$\bar{R}^2$	0.6133		

The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) for all 3 models are stated in the table below:

MODELS	AIC	BIC
1	44152.72	44180.01
2	43549.61	43582.36
3	43157.86	43196.06

Model 3 will be preferable because it has the lowest The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) when compared to other models as well as the $\bar{R}^2$ of model 3 is the highest.