Question 1(a)

Regression modelling uses statistics to ascertain the relationship between a dependent variable and one or many independent variables. A regression analysis performed on the home selling price (*sprice*) in a town in the late 1900s against its living area (*livarea*), age as well as the number of bedrooms (*beds*) and bathrooms (*baths*)[1] is shown in Figure 1, while the analysis of variance, parameters estimates and effect tests for the model can be found in Figures 2 to 4 respectively.

Figure 1

Regression analysis of home selling price against living area, age as well as the number of bedrooms and bathrooms

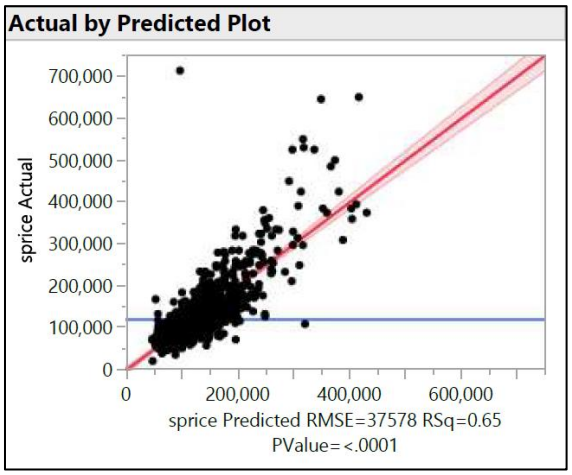


Figure 2

Analysis of Variance

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	3.8859e+12	9.715e+11	687.9524
Error	1495	2.1111e+12	1.4121e+9	Prob > F
C. Total	1499	5.997e+12		<.0001*

[1] The number of bedrooms and bathrooms are assumed to be continuous variables for purpose of the regression model.

Figure 3

Parameters Estimates

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	11154.289	6555.113	1.70	0.0890
livarea	10680.002	273.1491	39.10	<.0001*
baths	-7019.296	2903.816	-2.42	0.0158*
beds	-15552.44	1970.006	-7.89	<.0001*
age	-11.33396	80.50204	-0.14	0.8881

Figure 4

Effect Tests

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
livarea	1	1	2.1588e+12	1528.773	<.0001*
beds	1	1	8.801e+10	62.3250	<.0001*
baths	1	1	8251276702	5.8432	0.0158*
age	1	1	27991221.4	0.0198	0.8881

From the parameters estimates, the regression formula is given by:

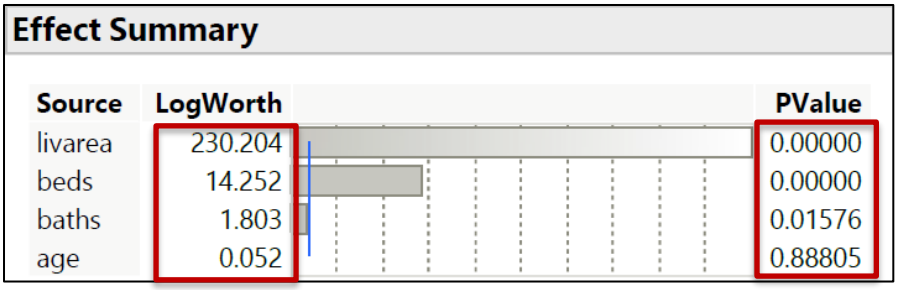
$$sprice = 11{,}154.289 + 10{,}680.002\ \text{livarea} - 7{,}019.296\ \text{baths} - 15{,}552.44\ \text{beds} - 11.33396\ \text{age}$$

The analysis of variance shows that the model has a small Prob>F which implies that there is minimally one explanatory independent variable in the model. Based on the parameter estimates and effects test, the significant independent variables (i.e. with low Prob>|t| and Prob>F) are living area as well as the number of bedrooms and bathrooms, while the age is not a significant variable. Specifically, the coefficient for living area is positive which signifies that as living area increases, ceteris paribus, home selling price also increases. Conversely, the coefficients for the number of bedrooms and bathrooms are negative which implies that ceteris paribus, the home selling price decreases when the number of bedrooms and bathrooms increases. Furthermore, these significant independent variables are also identified through the effect summary (Figure 5) which summarises

the effect of different independent variables on the dependent variables i.e. independent variables with low p-values or high LogWorth[2] values are explanatory variables.

Figure 5

Effect Summary



Effect Summary

Source	LogWorth	PValue
livarea	230.204	0.00000
beds	14.252	0.00000
baths	1.803	0.01576
age	0.052	0.88805

The RSquare value in the Summary of Fit shown in Figure 6 provides an indication of the % of dependent variable variation explained by the model i.e. the closer to 1, the better the model. The model has a RSquare value of 0.648 which signifies an average fit. This is also echoed in the Lack of Fit where the small Prob>F value denotes a significant lack of fit of the model (Figure 7).

Figure 6

Summary of Fit

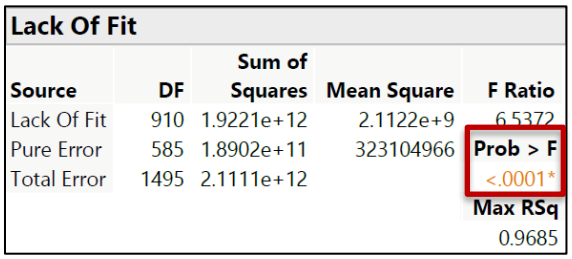
Summary of Fit

RSquare	0.647971
RSquare Adj	0.647029
Root Mean Square Error	37578.21
Mean of Response	123693.9
Observations (or Sum Wgts)	1500

[2] LogWorth = $-\log_{10}$(p-value). Independent variables with LogWorth values more than 2 are significant at the 0.01 level to explain the dependent variable.

Figure 7

Lack of Fit



Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	910	1.9221e+12	2.1122e+9	6.5372
Pure Error	585	1.8902e+11	323104966	Prob > F
Total Error	1495	2.1111e+12		<.0001*
				Max RSq
				0.9685

Question 1(b)

Using the regression formula in part (a), the expected price difference between the two houses with 8 years age difference is:

Expected price difference

$= 11.33396 \times (10 - 2)$

$= \$90.67168$

The value is calculated using the coefficient of the regressor *age* in the regression formula, multiplied by the difference in age of the two houses. The exclusion of the other regressors when calculating the expected price difference is due to the values of the other regressors remaining the same across both houses. Thus, they are excluded to calculate the marginal effect of a change in regressor *age*.Accordingly, the 95% interval estimate for the expected price difference (with t-value of 1.96 for 95% confidence interval and ∞ i.e. >120 degree of freedom) is:

Interval estimate

$= \bar{x} \pm 1.96 \frac{S}{\sqrt{n}}$

$= 90.67168 \pm 1.96 \frac{80.50204}{\sqrt{1{,}500}}$

$= [86.60, \quad 94.75]$

Question 1(c)

Using the regression formula in part (a), the expected price increase for the living room extension of 200 square feet is:

Expected price increase

$= 10{,}680.002 \times 2$

$= \$21{,}360.004$

To test if the increase in price will be at least \$20,000, the null (H_0) and alternate (H_1) hypothesis are set as per below respectively:

H_0: Expected price increase (μ) < \$20,000

H_1: Expected price increase (μ) ≥ \$20,000

The test statistics is then computed as follows:

Test statistics

$$= \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$$= \frac{21{,}360.004 - 20{,}000}{273.1491/\sqrt{1500}}$$

$$= 192.84$$

Comparing the computed test statistics to the t-value of 1.645 for 95% confidence interval and ∞ i.e. >120 degree of freedom), the computed test statistics is higher and hence, the null hypothesis is rejected. Accordingly, the price increase will be at least \$20,000 at the 5% significance level.

Question 1(d)

Using the regression formula in part (a), the expected price increase for the additional bedroom of 200 square feet is:

Total expected price difference

= Expected price difference due to increase in living area + Expected price difference due to additional bedroom

$= 10{,}680.002 \times 2 - 15{,}552.44 \times 1$

$= \$5{,}807.564$

Accordingly, the 95% interval estimate for the price increase (with t-value of 1.96 for 95% confidence interval and ∞ i.e. >120 degree of freedom) is:

Interval estimate

$$= \bar{x} \pm 1.96 \frac{S}{\sqrt{n}}$$

$$= 5{,}807.564 \pm 1.96 \frac{273.1491 + 1{,}970.006}{\sqrt{1{,}500}}$$

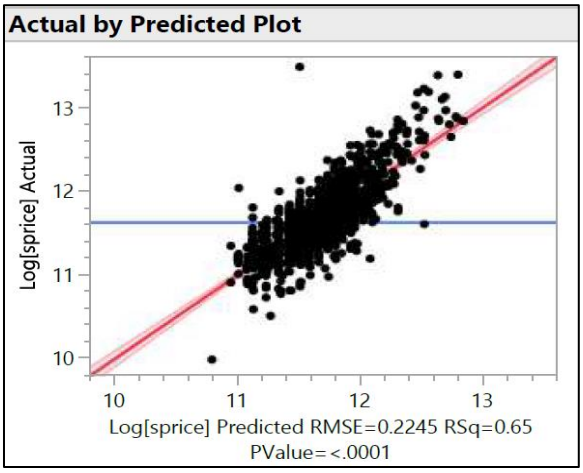
$= [5{,}694.04, \quad 5{,}921.08]$

Question 1(e)

Another regression analysis performed on the log of the home selling price against the log of its living area, while controlling for its age as well as the number of bedrooms and bathrooms[3] (in levels) is shown in Figure 8, while the analysis of variance and parameters estimates for the model can be found in Figures 9 and 10 respectively.

Figure 8

Regression analysis of log home selling price against log living area, while controlling for age as well as the number of bedrooms and bathrooms



[3] The number of bedrooms and bathrooms are assumed to be continuous variables for purpose of the regression model.

Figure 9

Analysis of Variance

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	141.61932	35.4048	702.5096
Error	1495	75.34448	0.0504	Prob > F
C. Total	1499	216.96380		<.0001*

Figure 10

Parameters Estimates

Parameter Estimates				
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	8.7171963	0.062302	139.92	<.0001*
Log[livarea]	1.1438678	0.030759	37.19	<.0001*
baths	-0.005438	0.017145	-0.32	0.7511
beds	-0.071998	0.011903	-6.05	<.0001*
age	3.2029e-5	0.000479	0.07	0.9467

From the parameters estimates, the regression formula is given by:

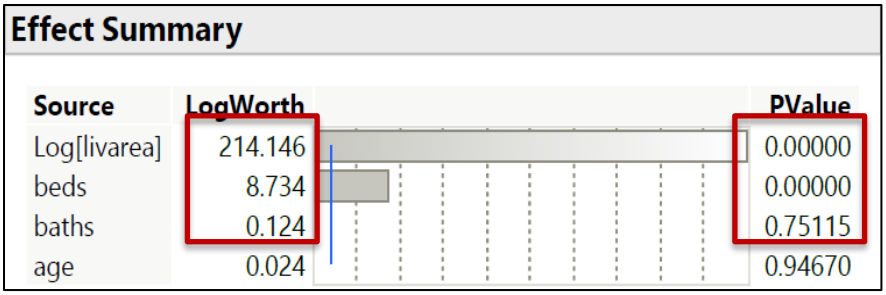
$$logsprice = 8.7171963 + 1.1438678\ \text{loglivarea} - 0.005438\ \text{baths} - 0.071998\ \text{beds} - 3.2029\text{e}^{-5}\ \text{age}$$

The analysis of variance shows that the model has a small Prob>F which implies that there is minimally one explanatory independent variable in the model. Based on the parameter estimates, the significant independent variables (i.e. with low Prob>|t|) are the log of the living area as well as the number of bedrooms. Specifically, the coefficient for the log of the living area is positive which signifies that as living area increases, ceteris paribus, home selling price also increases. Conversely, the coefficient for the number of bedrooms is negative which implies that ceteris paribus, the home selling price decreases when the number of bedrooms increases. Furthermore, these significant independent variables are also identified through the effect summary (Figure 11)

which summarises the effect of different independent variables on the dependent variables i.e. independent variables with low p-values or high LogWorth[4] values are explanatory variables.

Figure 11

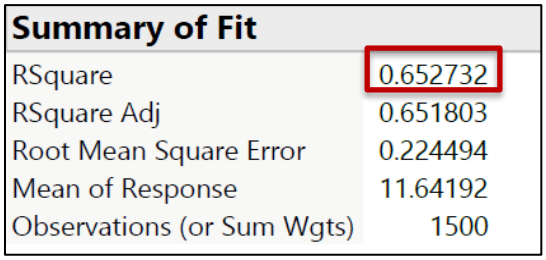
Effect Summary



The RSquare value in the Summary of Fit shown in Figure 12 provides an indication of the % of dependent variable variation explained by the model i.e. the closer to 1, the better the model. The model has a RSquare value of 0.653 which signifies an average fit. This is also echoed in the Lack of Fit where the small Prob>F value denotes a significant lack of fit of the model (Figure 13). In addition, comparing the adjusted RSquare value of this model (0.652) against that of the model in part (a) (0.647), this revised model provides a slightly better fit for the data.

Figure 12

Summary of Fit



[4] LogWorth = $-\log_{10}$(p-value). Independent variables with LogWorth values more than 2 are significant at the 0.01 level to explain the dependent variable.

Figure 13

Lack of Fit

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	910	60.041159	0.065979	2.5222
Pure Error	585	15.303317	0.026160	**Prob > F**
Total Error	1495	75.344476		<.0001*
				Max RSq
				0.9295

A further deep dive into the regression analysis is performed by specifying the measurement type for the number of bedrooms and bathrooms as nominal fields. From the parameters estimates in Figure 14, it appears that the home selling price is also mainly driven by the presence of 2 to 3.5 bathrooms and 1 to 3 bedrooms.

Figure 14

Regression analysis with number of bedrooms and bathrooms as nominal fields

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	8.4177802	0.1107	76.04	<.0001*
Log[livarea]	1.1420694	0.031873	35.83	<.0001*
baths[1]	0.0048528	0.040501	0.12	0.9046
baths[1.5]	-0.06462	0.043897	-1.47	0.1412
baths[2]	-0.08559	0.029177	-2.93	0.0034*
baths[2.5]	-0.079598	0.029302	-2.72	0.0067*
baths[3]	-0.086341	0.029807	-2.90	0.0038*
baths[3.5]	0.173897	0.057869	3.01	0.0027*
baths[4]	0.0046825	0.084541	0.06	0.9558
baths[4.5]	0.1212694	0.102248	1.19	0.2358
age	-0.00048	0.000505	-0.95	0.3421
beds[1]	-0.420174	0.189157	-2.22	0.0265*
beds[2]	0.1724182	0.058203	2.96	0.0031*
beds[3]	0.1776954	0.054267	3.27	0.0011*
beds[4]	0.0703778	0.054713	1.29	0.1985
beds[5]	-0.001319	0.059956	-0.02	0.9825

<u>Question 1(f)</u>

A further regression analysis performed on the log of the home selling price against the log of both its living area and the square of its living area, while controlling for its age as well as the

number of bedrooms and bathrooms (in levels) is shown in Figure 15, while the parameter estimates for the model can be found in Figure 16.

Figure 15

Regression analysis of log home selling price against log living area and log living area square, while controlling for age as well as the number of bedrooms and bathrooms

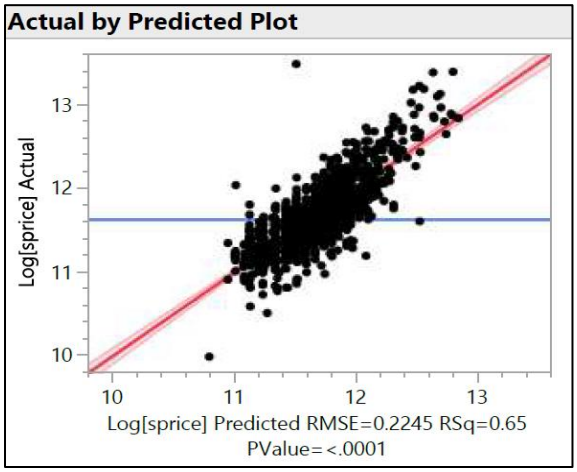


Figure 16

Parameters Estimates

Parameter Estimates

Term		Estimate	Std Error	t Ratio	Prob>|t|
Intercept		8.7171963	0.062302	139.92	<.0001*
Log[livarea]	Biased	1.1438678	0.030759	37.19	<.0001*
baths		-0.005438	0.017145	-0.32	0.7511
beds		-0.071998	0.011903	-6.05	<.0001*
age		3.2029e-5	0.000479	0.07	0.9467
Log[livarea^2]	Zeroed	0	0	.	.

From the parameters estimates, coefficients of the independent variables that are also used in the model in part (e) yielded similar results. In addition, the coefficient for the additional independent variable i.e. log of the square of living area is zero, and there is a biased indicator on the coefficient for the log of the living area.

A plausible explanation on the zero coefficient for the log of the square of living area could be that this additional independent variable does not have an impact on the model, given that it is

highly correlated to the log of the living area variable i.e. multicollinearity exists. This could have also led to the biased indicator for the log of the living area variable.

Question 2

From 1980 to 2009, regulation on assault weapons and concealed carry weapons (CCW) in the United States underwent gradual changes. Prior to 1989, there were no bans on assault weapons. Thereafter, state-level bans were enacted by several states until 1994, when a nationwide Federal ban was enacted till 2004. After which, several states continued with the enaction of state-level bans. On CCW laws, there were broadly four restriction categories imposed by states at different junctures as follows:

- Unrestricted - No permit required
- Shall issue - Permit must be granted to all qualified individuals
- May issue - Permit may be granted to qualified individuals
- Restricted - No carrying allowed

In the study, the author sought to find out the impact of assault weapon bans/CCW laws (independent variables) on gun-related murder rate (dependent variable). This was performed using a fixed effects regression model[5] that controlled for state-level regulation and time. For modelling purposes, states with assault weapon bans (including the Federal ban)/CCW laws in place for specific time periods were denoted by one, while the absence of regulations were denoted by zero. Specifically, the value of one was used for CCW laws falling under the 'restricted' and 'may issue' categories. In addition, other demographic and socioeconomic variables were also included in the model.

The result showed that CCW laws were significant in explaining gun-related murder rate while assault weapon ban was not. Specifically, CCW laws and gun-related murder rate shared a positive relationship i.e. rate increased with more restrictive laws. Separately, the Federal ban on assault weapon from 1994 to 2004 was also significant in explaining the gun-related murder rate in a positive direction i.e. rate also increased when the Federal ban was in force.

[5] Used to analyse variables that change very slowly over time e.g. regulations in a country.

The findings appeared to be counter intuitive as gun-related murder rate was expected to decrease with more restrictive CCW laws/assault weapon bans. However, other explanatory variables with high coefficients and low p-values as shown in Figure 17 could also have contributed to the results. These include the real per capita median income, % of population that is rural or between the age of 18 to 25 and per capita alcohol consumption. Furthermore, there could also be potential loopholes/exemptions in the enforcement process or that regulations could be set tighter in states where the violence rate was higher. Easy access to weapons in black markets could also be another contributing factor.

Figure 17

Regression results

Table 1. Fixed effects regression gun-related murder rate

Constant	−3.02 (−3.20)***
Assault weapons ban	−0.29 (−1.57)
Federal assault weapons ban	0.66 (2.42)**
Restrictive concealed carry laws	0.365 (3.74)***
Proportion of population that is white	0.172 (1.76)*
Proportion of population that is rural	1.93 (3.97)***
Real per capita median income	0.00021 (6.03)***
Proportion of population with college degree	−1.367 (−1.20)
Unemployment rate	3.397 (1.34)
Proportion of population >18 and <25	11.45 (2.27)**
Proportion of population >24 and <35	−2.876 (−0.91)
Per capita alcohol consumption	0.688 (4.05)***

Notes: $R^2 = 0.797$.
Test statistics in parentheses.
* 5% < *p*-value < 10%; ** 1% < *p*-value < 5%; *** *p*-value < 1%.

The use of fixed effects regression models has its shortcomings such as low statistical power (Hill et al., 2020) when applied on variables with little variation - as with regulation which could be longer term in nature and unlikely to vary much over time. Fixed effects regression models are also prone to reverse causality (Collischon & Eberl, 2020) i.e. whether the effect of gun-related murder rate stemmed from regulation or that an increase in the gun-related murder rate could impact the CCW laws/assault weapon ban decisions.

(Word count for Q2: 500)

References

Collischon, M., & Eberl, A. (2020, August 5). Let's Talk About Fixed Effects: Let's Talk About All the Good Things and the Bad Things. https://link.springer.com/article/10.1007/s11577-020-00699-8

Hill, T. D., Davis, A. P., Roos, J. M., & French, M. T. (2020). Limitations of Fixed-Effects Models for Panel Data. *Sociological Perspectives*, 63(3), 357–369. https://doi.org/10.1177/0731121419863785