psyc30815 2019/20 Assessment

Jane Doe

03/12/2019

Question 1: Linear Regression

Data set

For this analysis, I chose to use a data set based on sales prices of houses in the city of Windsor during 1987. 547 observations were collected. This is available from Anglin and Gencay (1996).

In the data, the *outcome* variable is:

price The sale price of the house.

From the same data set, the following *predictor* variables were obtained: These predictor variables were chosen as they present physical features of a house.

lotsize The lot size of the property in square feet.

stories The number of stories exlcluding the basement.

bedrooms The number of bedrooms in the house.

bathrms The number of bathrooms in the house.

driveway Binary variable, does the house have a driveway; yes or no.

recroom Binary variable, does the house have a recreational room; yes or no.

fullbase Binary variable, does the house have a finnished basement; yes or no.

garagepl The number of garage places the house has.

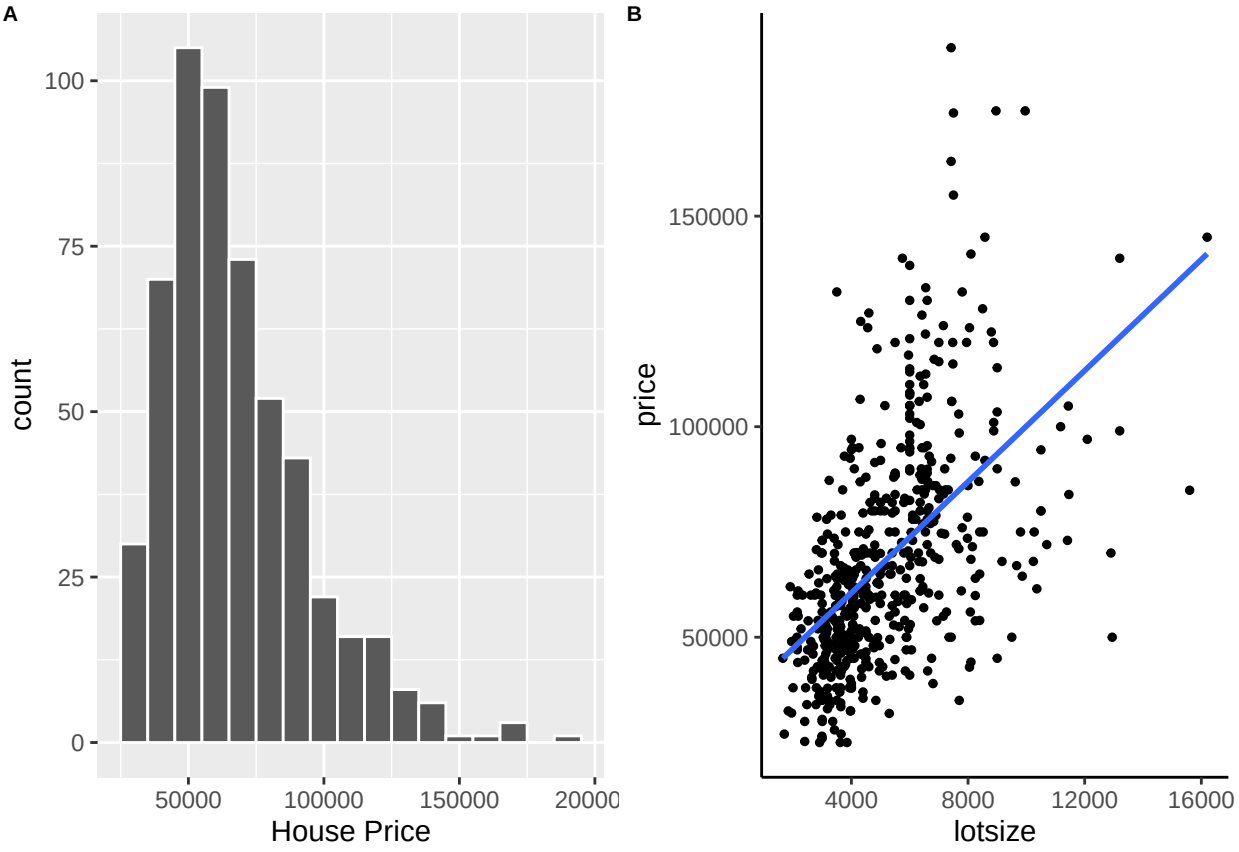
Objective of analysis

This data set is interesting as it shows potential variables predicting house price,and therefore will answer the theoretical question: what variables predict house price. Lotsize could be considered an explanatory predictor variable. It is arguable that the *bigger* the lotsize, the *larger* the house price. The bigger the lot size, the more likely that the house will have space for more *stories*, more *bedrooms*, more *bathrooms*, more *garage places* and will be more likely to have a *full basement*, a *recreation room* and a *driveway*. Therefore, the focus of this analysis will be on whether increased sale price of a house is associated with these factors.

Analysis

The predictor variables are the nine variables aforementioned: *lotsize*, *stories* , *bedrooms*, *bathrms*, *recroom*, *driveway*, *fullbase* and *garagepl*. These will be used without being modified. The binary variables *recroom*, *driveway*and *fullbase* will be coded with 1 indicating yes and 0 indicating no. Figure A shows a histogram for the outcome variable (price), the mean sale price of the house was $68121.60 with a standard deviation of $26702.67. The large SD of houseprice can be explained by the amount house price varies based on the

predictors used in this sample. Figure B shows a scatterplot of the correlation between price and *lotsize* (the explanatory predictor variable).



For this analysis, the assumed linear model can be written as follows: For $i \in \{1 \dots 547\}$

$$y_i \sim N(\mu_i, \sigma^2), \quad y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots \ \beta_8 x_{8i} + \epsilon_i$$

Here, y_i is the sale price of a house in observation i (outcome variable) and $x_{1i} \dots x_{9i}$ are the values of the 9 predictor variables in observation i. Here, it is assumed that y_i is normally distributed with a mean of μ_i and a variance of σ^2. Expected values of y_i should be a linear function of the predictors. The parameters to be inferred of this model are: the intercept term β_0, the coefficients for each of the 9 predictor variables when all other predictors are held constant ($\beta_1 \dots \beta_8$) and the variance of the outcome variable σ^2. ϵ_i is the error term accounting for random variance independent of predictors. The parameters of this model are inferred using maximum likelihood estimation and are shown below. The values of the coefficients are determined based on the data.

$$\hat{\boldsymbol{\beta}} = \hat{\beta}_0, \hat{\beta}_1 \dots \hat{\beta}_8.$$

The predicted values of the outcome variable (sale price of a house), assuming maximum likelihood estimation of the coefficients will be as follows.

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \dots \ \hat{\beta}_8 x_{8i}$$

The overall fit to the data using this is quantified using R^2. However, any predictor variable will increase the R^2 value despite its predictive value. Therefore, the *adjusted* R^2 value provides a more reliable measure of model fit as it takes in to account the number of predictors and their predictive value.

Under the null hypothesis that,

$$\hat{\beta}_1 = \hat{\beta}_2 \dots = \hat{\beta}_8 = 0$$

$$\frac{R^2}{(1-R^2)}\frac{N-K-1}{K}$$

is distributed as an F distribution with K and $N-K-1$ degrees of freedom. The F-test of overall significance indicates whether the linear regression model provides a better fit to the data than a model that contains no independent variables.

When focusing on the explanatory predictor variable (*lotsize*) alone, the price outcome variable had an $R^2 = 0.287$, meaning that $\approx 29\%$ of the variance in the price variable can be explained by *lotsize*. The adjusted R^2 values is $\text{AdjR}^2 = 0.286$. The F statistic corresponding to the global null hypothesis that all predictor variables are simultaneously of no predictive value is $\text{F}(1, 544) = 219.056$. The p-value corresponding to this is $< .000$, highlighting high significance of *lotsize* as an explanatory predictor variable. However, despite significance, the R^2 and adjusted R^2 values are relatively low, this could be explained by a lack of covariates which impact house price.

The maximum likelihood estimate, standard error, t-statisitc and p-value for *lotsize* are provided in the following table:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	34136	2491	13.7	6.275e-37
lotsize	6.599	0.4458	14.8	6.77e-42

The overall model fit of the linear model using the $K = 8$ predictors of the price outcome variable had an $R^2 = 0.608$, meaning that $\approx 61\%$ of the variance in the price variable can be explained by these predictors. The adjusted R^2 values is $\text{AdjR}^2 = 0.602$. The F statistic corresponding to the global null hypothesis that all predictor variables are simultaneously of no predictive value is $\text{F}(8, 537) = 104.089$. The p-value corresponding to this is $< .000$, highlighting high significance of the overall model.

The maximum likelihood estimates, standard errors, t-statisitc and p-values for the coefficients are provided in the following table:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-8408	3689	-2.279	0.02306
lotsize	4.235	0.3735	11.34	7.117e-27
stories	8617	978.2	8.809	1.746e-17
bedrooms	1932	1143	1.69	0.09152
bathrms	14942	1623	9.208	7.394e-19
recroomyes	6094	2067	2.948	0.003335
drivewayyes	8102	2214	3.659	0.0002784
fullbaseyes	7776	1701	4.573	5.982e-06
garagepl	4965	912.1	5.443	7.962e-08

All of the predictors except *bedrooms* are significant. Therefore, if all the other predictors are held constant, each predictor significantly predicts a change in the outcome variable. Contrastingly *bedrooms* provides no predicitive value in this model. We can test this through using a nested model comparision. *Bedrooms* is removed from the model, an F test is perfomed to determine whether there is a significant difference in model fit. The p-value for the model fit and the new model without *bedrooms* is 0.09. Here we can see that removing *bedrooms* from the model leads to no significant decrease in the model fit. Therefore, it can be removed completely. Having removed *bedrooms* the model fit is $R^2 = 0.606$ ($\text{AdjR}^2 = 0.601$), with the F statistic corresponding null hypothesis that all predictor variables are simultaneously of no predictive value is $\text{F}(7, 538) = 118.142$. The p-value corresponding to this is $< .000$, highlighting high significance of the overall model. The R^2 of the new model is lower as a *bedrooms* has been removed. However, the ajusted R^2 is the same for both models, again highlighting that *bedrooms* did not add any predictive value to the model. The

maximum likelihood estimates, standard errors, t-statistic, and p-values for the coefficients in the $K = 7$ model are provided in the following table:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-4676	2961	-1.579	0.1149
lotsize	4.289	0.3728	11.5	1.545e-27
stories	9217	912.9	10.1	4.578e-22
bathrms	15564	1583	9.831	4.361e-21
recroomyes	6019	2070	2.908	0.003793
drivewayyes	7660	2203	3.477	0.0005472
fullbaseyes	8203	1685	4.869	1.475e-06
garagepl	5082	911	5.579	3.847e-08

All of the predictors have a positive slope. This means that when all else is held constant, each predictor increases the estimated house price. For every square foot increase in *lotsize* increases house price by 4.29 dollars. For every increase in number of stories, house price increases by 9217.48 dollars. For every aditional bathroom, house price increases by 15563.89 dollars. If the house has a recreational room, house price increases by 6018.87 dollars. If the house has a driveway and a full basement, house price increases by 7659.78 dollars and 8202.68 dollars respectively. Finally, for each increase in gargage place, house price increases by 5081.94 dollars.

Interestingly, when the number of *bedrooms* were included within the model, it was not significant when controlling for all the other variabes. However, when running bedrooms in a model with just lotsize, *bedrooms* was a significant predictor of house price ($p < .000$). Significance here could be explained by the idea that number of *bedrooms* is highly correlated with other variables that also predict house price. The overall model fit of the linear model using the $K = 2$ predictors of the price outcome variable had an $R^2 = 0.37$, meaning that $\approx 37\%$ of the variance in the price variable can be explained by *lotsize* and *bedrooms*. The adjusted R^2 values is $\text{AdjR}^2 = 0.368$. The F statistic corresponding to the global null hypothesis that all predictor variables are simultaneously of no predictive value is $\text{F}(8, 537) = 104.089$. The p-value corresponding to this is $< .000$, highlighting high significance of the overall model. However the low R^2 and R^2 adjusted values can be explained by the lack of predictors in the model.

Prediction

As *lotsize* can be considered an explanatory predictor, house price varies as a function of the lotsize. It can be argued that the larger the lot size, the more components the house will have or the more potential it will have. For example a larger lot size is more likely to have a recreational room or more bathrooms or more space for these components. Predicted house prices when lot size varies were calculated whilst all the other predictors remained constant. The predictions are shown in the table below.

lotsize	stories	bathrms	garagepl	recroom	driveway	fullbase	pred
3000	2	1	0	yes	no	yes	56411
4000	2	1	0	yes	no	yes	60699
5000	2	1	0	yes	no	yes	64988

Here it can be seen that, as lotsize increases from 3000 square feet to 5000 square feet, house price increases from 56,411 dollars to 64,988 dollars.

Conclusion

The results of this linear regression indicate that a bigger sale price of a house is correlated with a larger lot size, a larger number of stories and bathrooms and garage places and the inclusion of a recreation room, a

driveway and a full basement. It can be argued that lot size is an explanatory factor as the larger the lot, the more space there is to have multiple stories/bathrooms for example. However, it can be argued that other factors also need to be considered with regards to house price. For example, Adam, Kuang, and Marcet (2012) mentioned interest rate changes and housing preferences, this can also vary across countries. Therefore it would be beneficial to run this analysis again, adding variables such as countries, housing preferances and interest rates.

References

Adam, Klaus, Pei Kuang, and Albert Marcet. 2012. "House Price Booms and the Current Account." *NBER Macroeconomics Annual* 26 (1): 77–122.

Anglin, Paul M, and Ramazan Gencay. 1996. "Semiparametric Estimation of a Hedonic Price Function." *Journal of Applied Econometrics* 11 (6): 633–48.