INSH5301 Intro Computational Statistics - Final Exam

Ali Banijamali

04/21/2020

Loading Libraries and Data :

```
library(Ecdat) # For housing data set
library(AER)   # For health insurance data set
library(ggplot2)
library(glmnet)
library(e1071)
library(stargazer)
library(knitr)
library(kableExtra)
library(car)
library(psych)

data(Housing)
kable( head(Housing, 2) , booktabs=T, caption='Housing Dataset') %>%
  kable_styling( latex_options=c('scale_down', 'hold_position'))
```

Table 1: Housing Dataset

price	lotsize	bedrooms	bathrms	stories	driveway	recroom	fullbase	gashw	airco	garagepl	prefarea
42000	5850	3	1	2	yes	no	yes	no	no	1	no
38500	4000	2	1	1	yes	no	no	no	no	0	no

```
data(HealthInsurance)
kable( head(HealthInsurance, 2) , booktabs=T, caption='Health Insurance Dataset') %>%
  kable_styling( latex_options=c('scale_down', 'hold_position'))
```

Table 2: Health Insurance Dataset

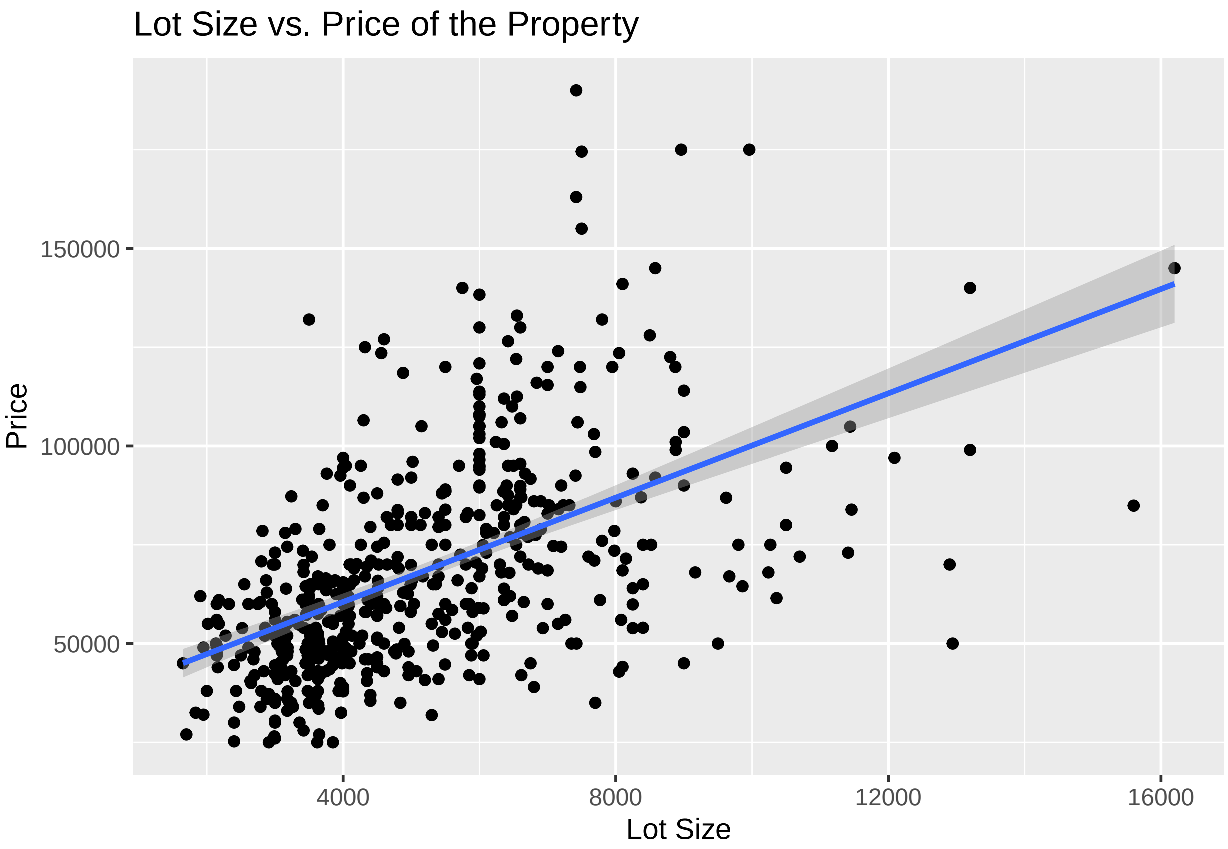
health	age	limit	gender	insurance	married	selfemp	family	region	ethnicity	education
yes	31	no	male	yes	yes	yes	4	south	cauc	bachelor
yes	31	no	female	yes	yes	no	4	south	cauc	highschool

1. *Bivariate Regression*

1. Using the Housing dataset, create a scatter plot of sale price of a house (y-axis) and the lot size of the property (x-axis). Use the ggplot function and include a regression line. Using the graph, describe the relation between the two variables.

ANS.

```r
ggplot(data=Housing, aes(x=lotsize, y=price)) + geom_point() + geom_smooth(method=lm) +
  xlab('Lot Size') + ylab('Price') +
  ggtitle('Lot Size vs. Price of the Property')
```



According to the plot, there is a positive relationship between Lot Size and the price of the property. Meaning, as the size of the lot increases, the price of the property goes up.

2. Estimate a bivariate regression of the sale price of a house on the lot size of the property. Interpret the estimated β parameters, the statistical significance and R^2.

ANS.

```r
bivar.model <- lm(price ~ lotsize, data=Housing)

stargazer(bivar.model, align=T, no.space=T, header=F,
          title="Price of the Property vs. Lot Size",
          dep.var.labels=c("Price"),
          covariate.labels=c("Lot Size"),
          omit.stat=c("LL", "ser", "f"),
          table.placement = "H" # TO hold the table at it's place (Not floating)
          # Floting tables get out of order in the printed output
         )
```

Table 3: Price of the Property vs. Lot Size

	Dependent variable:
	Price
Lot Size	6.599***
	(0.446)
Constant	34, 136.190***
	(2, 491.064)
Observations	546
R^2	0.287
Adjusted R^2	0.286
Note:	*p<0.1; **p<0.05; ***p<0.01

$\beta = +$ 6.599 This shows a positive correlation, i.e. as the size of houses' lots increase, their price goes up. The estimated β is statistically significant (p-value<2.2e-16) and the adjusted R^2 which almost equals to the R^2 in our case, because we only have one independent variable, is 0.286. meaning the independent variable (lot size), explains ~%27 of the variations in the house's price. β also tells us that each square foot increase in the lotsize, increases the price of the house by ~6.6.

3. Is there any reason to believe that the estimated slope parameter in the previous regression is biased? (Explain)

ANS.
I think the lot size of the house is a completely reasonable variable for explainging the price of the property. Let's look at other variables in our data:

```
colnames(Housing)
```

```
##  [1] "price"    "lotsize"  "bedrooms" "bathrms"  "stories"  "driveway"
##  [7] "recroom"  "fullbase" "gashw"    "airco"    "garagepl" "prefarea"
```

Let's think of exogenity. When exogenity is violated, our model will suffer from the Omitted Variable Bias (OVB). Although I think it is not there, we cn examine different scenarios. First, let's check for Spurious asscociations. I will get a regression with all variables and see if the effect of lotsize goes away or not:

```
# 1. Spurious association check:
spurious.check <- lm(price ~ ., data=Housing)
stargazer(spurious.check, align=T, no.space=T, header=F,
          title="Price of the Property vs. all vars",
          dep.var.labels=c("Price"),
          omit.stat=c("LL", "ser", "f"),
          table.placement = "H"
         )
```

Table 4: Price of the Property vs. all vars

	Dependent variable:
	Price
lotsize	3.546***
	(0.350)
bedrooms	1,832.003*
	(1,047.000)
bathrms	14,335.560***
	(1,489.921)
stories	6,556.946***
	(925.290)
drivewayyes	6,687.779***
	(2,045.246)
recroomyes	4,511.284**
	(1,899.958)
fullbaseyes	5,452.386***
	(1,588.024)
gashwyes	12,831.410***
	(3,217.597)
aircoyes	12,632.890***
	(1,555.021)
garagepl	4,244.829***
	(840.544)
prefareayes	9,369.513***
	(1,669.091)
Constant	−4,038.350
	(3,409.471)
Observations	546
R^2	0.673
Adjusted R^2	0.666
Note:	*p<0.1; **p<0.05; ***p<0.01

As we can see, lotsize still has a sigificant role, so this can't be the case. Now let's look at the correlations. Note that there are factor variables in some columns which I don't see any of them affecting the lot size (the other way might be true in a few cases), therefore, I am just removing them from the correlation test:

```
kable(cor(Housing[c(1:5, 11)]), booktabs=T,
      caption='Correlation between Numerical Variables') %>%
  kable_styling( latex_options=c('striped', 'hold_position'))
```

We can see that lotsize has a correlation with all of the variables to some extent (including price), but we have to also keep in mind that those variables are not affecting lotsize! For example a bigger lotsize, gives the option of having more bedroom or more garage places but the relationship is not the other way. So I conclude that the relationship is not biased.

Table 5: Correlation between Numerical Variables

	price	lotsize	bedrooms	bathrms	stories	garagepl
price	1.0000000	0.5357957	0.3664474	0.5167193	0.4211902	0.3833020
lotsize	0.5357957	1.0000000	0.1518515	0.1938335	0.0836750	0.3528717
bedrooms	0.3664474	0.1518515	1.0000000	0.3737688	0.4079737	0.1391172
bathrms	0.5167193	0.1938335	0.3737688	1.0000000	0.3240656	0.1781783
stories	0.4211902	0.0836750	0.4079737	0.3240656	1.0000000	0.0434119
garagepl	0.3833020	0.3528717	0.1391172	0.1781783	0.0434119	1.0000000

2. *Multivariate Regression*

4. Using the rest of the variables in the dataset, construct a correlation matrix and use it to check if the assumption of exogeneity is valid in the estimated model in question (2). (Explain) Hint: See module 13's Memo 3 for dummies and then see Module 13's Memo 1 to get familiar with OV B, explain accordingly.

ANS.
lm model recognizes factors as different categories and can handle dummy variables, but cor() doesn't understand non-numeric values, so I have to define dummy variables manually for it:

```
Housing.d <- cbind(Housing[1:5],
                   driveway =ifelse(Housing$driveway=='yes', 1, 0),
                   recroom  =ifelse(Housing$recroom =='yes', 1, 0),
                   fullbase =ifelse(Housing$fullbase=='yes', 1, 0),
                   gashw    =ifelse(Housing$gashw   =='yes', 1, 0),
                   airco    =ifelse(Housing$airco   =='yes', 1, 0),
                   Housing[11],
                   prefarea =ifelse(Housing$prefarea=='yes', 1, 0))


kable(cor(Housing.d), booktabs=T, caption='Correlation between All Variables') %>%
  kable_styling( latex_options=c('scale_down', 'striped', 'hold_position'))
```

Table 6: Correlation between All Variables

	price	lotsize	bedrooms	bathrms	stories	driveway	recroom	fullbase	gashw	airco	garagepl	prefarea
price	1.0000000	0.5357957	0.3664474	0.5167193	0.4211902	0.2971668	0.2549595	0.1862177	0.0928365	0.4533466	0.3833020	0.3290743
lotsize	0.5357957	1.0000000	0.1518515	0.1938335	0.0836750	0.2887778	0.1403273	0.0474867	-0.0092009	0.2217649	0.3528717	0.2347822
bedrooms	0.3664474	0.1518515	1.0000000	0.3737688	0.4079737	-0.0119963	0.0804923	0.0972006	0.0460278	0.1604122	0.1391172	0.0789527
bathrms	0.5167193	0.1938335	0.3737688	1.0000000	0.3240656	0.0419552	0.1268918	0.1027914	0.0673651	0.1849551	0.1781783	0.0640133
stories	0.4211902	0.0836750	0.4079737	0.3240656	1.0000000	0.1224986	0.0422812	-0.1738596	0.0182605	0.2962157	0.0434119	0.0429397
driveway	0.2971668	0.2887778	-0.0119963	0.0419552	0.1224986	1.0000000	0.0919591	0.0434284	-0.0119421	0.1062901	0.2036822	0.1993781
recroom	0.2549595	0.1403273	0.0804923	0.1268918	0.0422812	0.0919591	1.0000000	0.3724339	-0.0101186	0.1366256	0.0381222	0.1612918
fullbase	0.1862177	0.0474867	0.0972006	0.1027914	-0.1738596	0.0434284	0.3724339	1.0000000	0.0046773	0.0452479	0.0525240	0.2286510
gashw	0.0928365	-0.0092009	0.0460278	0.0673651	0.0182605	-0.0119421	-0.0101186	0.0046773	1.0000000	-0.1303499	0.0681440	-0.0591697
airco	0.4533466	0.2217649	0.1604122	0.1849551	0.2962157	0.1062901	0.1366256	0.0452479	-0.1303499	1.0000000	0.1565956	0.1156259
garagepl	0.3833020	0.3528717	0.1391172	0.1781783	0.0434119	0.2036822	0.0381222	0.0525240	0.0681440	0.1565956	1.0000000	0.0923638
prefarea	0.3290743	0.2347822	0.0789527	0.0640133	0.0429397	0.1993781	0.1612918	0.2286510	-0.0591697	0.1156259	0.0923638	1.0000000

Many of these variables seem to have relatively high correlation with both price and lotsize. In particular look at bedrooms, bathrms, driveway, recroom, airco, garagepl and prefarea. So exogeneity was probably violated in the bivariate model.

5. Estimate a set of multivariate models to address the potential issue of OVB, adding at most one additional variable each time. (See Memo 2: Multivariate Models under Real World Example) Display all the estimated models side-by-side (you may need two or more stargazer tables here). Using the multivariate models, do you think there is evidence that the estimated parameter in (2) was biased? which of the estimated models you consider the least bias (from now on, we'll call this model the best model)?

Hint: See Module 13's Memo 1 to get familiar with OV B, and follow Memo 2's Multivariate Models to add one variable each time.

ANS.

```r
mv.model.1 <- lm(price ~ lotsize + bedrooms, data=Housing.d)
mv.model.2 <- lm(price ~ lotsize + bedrooms + bathrms, data=Housing.d)
mv.model.3 <- lm(price ~ lotsize + bedrooms + bathrms + driveway,
                 data=Housing.d)
mv.model.4 <- lm(price ~ lotsize + bedrooms + bathrms + driveway +
                 recroom, data=Housing.d)
mv.model.5 <- lm(price ~ lotsize + bedrooms + bathrms + driveway +
                 recroom + airco, data=Housing.d)
mv.model.6 <- lm(price ~ lotsize + bedrooms + bathrms + driveway +
                 recroom + airco + garagepl, data=Housing.d)
mv.model.7 <- lm(price ~ lotsize + bedrooms + bathrms + driveway +
                 recroom + airco + garagepl + prefarea, data=Housing.d)

# Note: The command for the second table is hidden in order to fit both tables in 1 page.
stargazer(list(bivar.model, mv.model.1, mv.model.2, mv.model.3),
          align=T, no.space=T, header=F,
          title="Comparison of Models Part 1",
          column.sep.width = "1pt",
          dep.var.labels=c("Price"),
          omit.stat=c("LL", "ser", "f", "n"),
          table.placement = "H"
         )
```

Table 7: Comparison of Models Part 1

	Dependent variable:			
	Price			
	(1)	(2)	(3)	(4)
lotsize	6.599^{***}	6.053^{***}	5.411^{***}	4.785^{***}
	(0.446)	(0.424)	(0.388)	(0.395)
bedrooms		$10,567.350^{***}$	$5,826.802^{***}$	$6,196.547^{***}$
		(1,247.676)	(1,206.571)	(1,177.634)
bathrms			$19,750.210^{***}$	$19,688.790^{***}$
			(1,785.083)	(1,739.429)
driveway				$13,144.810^{***}$
				(2,405.969)
Constant	$34,136.190^{***}$	5,612.600	$-2,418.293$	$-11,501.410^{***}$
	(2,491.064)	(4,102.819)	(3,779.412)	(4,040.559)
R^2	0.287	0.370	0.486	0.513
Adjusted R^2	0.286	0.368	0.483	0.510
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 8: Comparison of Models Part 2

	Dependent variable:			
	Price			
	(1)	(2)	(3)	(4)
lotsize	4.626^{***}	4.101^{***}	3.639^{***}	3.316^{***}
	(0.391)	(0.368)	(0.376)	(0.370)
bedrooms	$6,052.177^{***}$	$5,194.116^{***}$	$4,924.599^{***}$	$4,671.830^{***}$
	(1,159.817)	(1,082.922)	(1,066.079)	(1,037.370)
bathrms	$19,052.260^{***}$	$17,543.270^{***}$	$16,912.280^{***}$	$17,013.920^{***}$
	(1,718.864)	(1,607.321)	(1,586.089)	(1,542.058)
driveway	$12,560.470^{***}$	$11,581.140^{***}$	$10,422.160^{***}$	$8,759.434^{***}$
	(2,372.514)	(2,209.577)	(2,187.170)	(2,146.379)
recroom	$8,952.255^{***}$	$7,308.760^{***}$	$7,645.728^{***}$	$6,364.557^{***}$
	(2,097.103)	(1,958.949)	(1,926.867)	(1,886.790)
airco		$15,202.320^{***}$	$14,748.280^{***}$	$14,349.720^{***}$
		(1,648.742)	(1,623.683)	(1,580.061)
garagepl			$4,070.540^{***}$	$4,121.579^{***}$
			(911.278)	(885.966)
prefarea				$9,854.542^{***}$
				(1,735.582)
Constant	$-10,526.990^{***}$	$-7,017.655^{*}$	$-4,769.932$	$-3,049.459$
	(3,984.282)	(3,725.855)	(3,696.439)	(3,606.332)
R^2	0.529	0.593	0.608	0.630
Adjusted R^2	0.525	0.589	0.603	0.624
Note:				*p<0.1; **p<0.05; ***p<0.01

Now let's talk about the models:

1. As we add more variables, the amplitude of β for lotsize decreases, this tells us that there was OVB. However, this variable (lotsize), remains a strong and significant predictor even after adding new independent variables.

2. Let's look at how adjusted R^2 increased after addition of variables:

- mv.1: bedrooms: ~ + %8.2
- mv.2: bathrms: ~ + %11.5
- mv.3: driveway: ~ + %2.7
- mv.4: recroom: ~ + %1.5
- mv.5: airco: ~ + %6.4
- mv.6: garagepl: ~ + %1.4
- mv.7: prefarea: ~ + %2.1
  As we can see, all of the additional variables are significant and increased the adjusted R^2, however, the most important variables were bathrms, bedrooms, airco, driveway and prefarea. The other 2 variables, garagepl and recroom didn't increase adjusted R^2 very much, so we can probably exclude them from our model but including them won't hurt either.

6. Check if the best model suffers from multicollinearity (if it does, don't try to fix it, just explain β what problems it may cause).

Hint:Use vif() in car package to easily calculate VIF and lecture 13's Memo 2 of multicollinearity to explain.

ANS.

```r
# Best model:
# I took out garagepl & recroom in the best model.
best.mv.model <- lm(price ~ lotsize + bedrooms + bathrms + driveway +
                   airco + prefarea, data=Housing.d)

# Looking at multicollinearity:
kable(t( vif(best.mv.model)), booktabs=T, caption='Best Model Multicollinearity Test') %>%
  kable_styling( latex_options=c('hold_position'))
```

Table 9: Best Model Multicollinearity Test

lotsize	bedrooms	bathrms	driveway	airco	prefarea
1.207809	1.186777	1.203482	1.120413	1.089609	1.086345

As we can see, all values are below 5. This means means weak imperfect multicollinearity and is not an issue at all. We can also test the model with all of the variabes:

```r
kable(t( vif(mv.model.7)), booktabs=T, caption='All Vars Model Multicollinearity Test') %>%
  kable_styling( latex_options=c('hold_position'))
```

Table 10: All Vars Model Multicollinearity Test

lotsize	bedrooms	bathrms	driveway	recroom	airco	garagepl	prefarea
1.309573	1.190941	1.22043	1.137937	1.060491	1.101906	1.185171	1.102347

Ase we can see, here too, the multicollinearity isn't an issue at all. All of the values are below 5, which means weak imperfect multicollinearity.

3. *Non − linear Functional Forms*

7. Take a look at the graph from part (1), do you think there is any reason to believe that the effect of lot size on price is not the same for all the domain of lot size? if yes, is the effect increasing or decreasing?

ANS.
Looking closer at the data, we can see that the effect of the lotsize is fairly linearly increasing up to 4000~5000 sq ft, but it suddenly increasing at a very fast rate afterwards and specifically around 8000 square ft. After that we really don't have a lot of data points to have a fair judgement about the traend of the data but it seems like it is tracking the same linear increase in the beginning (smaller lot sizes).

8. Estimate the best model again, but this time transform the lot size variable to natural logarithms. Interpret the estimated parameter for log of the lot size.

ANS.

```r
best.mv.v2 <- lm(price ~ I(log(lotsize)) + bedrooms + bathrms + driveway +
                 airco + prefarea, data=Housing.d)

stargazer(best.mv.v2, align=T, no.space=T, header=F,
          title="Best MV model + log term",
          column.sep.width = "1pt",
          dep.var.labels=c("Price"),
          omit.stat=c("LL", "ser"),
          table.placement = "H"
         )
```

Table 11: Best MV model + log term

	Dependent variable:
	Price
I(log(lotsize))	$22,221.930^{***}$
	$(2,013.864)$
bedrooms	$5,025.226^{***}$
	$(1,055.136)$
bathrms	$17,853.870^{***}$
	$(1,562.003)$
driveway	$8,737.133^{***}$
	$(2,201.341)$
airco	$14,341.690^{***}$
	$(1,614.173)$
prefarea	$10,849.550^{***}$
	$(1,748.892)$
Constant	$-172,471.700^{***}$
	$(16,256.980)$
Observations	546
R^2	0.614
Adjusted R^2	0.610
F Statistic	142.980^{***} (df = 6; 539)
Note:	*p<0.1; **p<0.05; ***p<0.01

9. Estimate the best model twice: (a) first, adding a quadratic term for lot size, and, (b) second, adding a quadratic and cubic terms. Using the change in lot size as a one standard deviation change from the mean, compare the effect of lot size in the original model, model (a), and, model (b). Can you reject the hypothesis that the relation between lot size and price is linear? quadratic? cubic? (Explain)

ANS.

```
# a) lotsize quadratic:
best.mv.v3 <- lm(price ~ I(lotsize^2) + bedrooms + bathrms + driveway +
                   airco + prefarea, data=Housing.d)
```

10. Using the best model as the nested model, test the hypothesis that the effect of lot size on price is moderated by prefarea.

ANS.

4. *Unsupervised Machine Learning*

11. Run a factor analysis or PCA on the Housing dataset, examine the loadings of the factors on the variables. Sort the variables by their loadings, and try to interpret what the first one "mean".

ANS.
I am using PCA:

```
# Scaling some columns to have all of them relatively
# on the same order using psych::rescale
Housing.d.sc <- cbind(rescale(Housing.d$lotsize, mean=0, sd=1),
                      Housing.d[3:12])
colnames(Housing.d.sc) <- colnames(Housing.d[2:12])

# PCA using singular value decomposition (SVD):
pca <- prcomp(Housing.d.sc)
# 1st component:
kable(t(pca$rotation[ ,1][order(pca$rotation[ ,1])]), booktabs=T, caption='Fisrt Principal Component')
  kable_styling( latex_options=c('hold_position', 'scale_down')) # order() for sorting
```

Table 12: Fisrt Principal Component

lotsize	garagepl	stories	bedrooms	bathrms	airco	driveway	prefarea	recroom	fullbase	gashw
-0.7146536	-0.4681746	-0.3202026	-0.2987256	-0.1835323	-0.1502877	-0.0963425	-0.0941105	-0.0570081	-0.0251744	-0.004824

The result is very interesting. Here we can see the two poles of the 1st PC. On one side, we have lot size, number of bedrooms and bathrooms and generally variables which are related to the size of the property. On the other side, we can see variables about amenities of the building like is it heated by gas, or does it has a full finished basement or does the house has recreational room?

12. Use k-means algorithm and examine the centers of each cluster using only two centroids. How are they similar to and different from the factor loadings of the first factor?

ANS.

```
set.seed(1)
# Running kmeans w/ 2 centers and 25 random start:
kmeans <- kmeans(Housing.d.sc, centers=2, nstart=25)

kmeans.centroids <- kmeans$centers

kmeans.topvars_centroid1 <- kmeans.centroids[1, order(kmeans.centroids[1, ])]
kable(t(tail(kmeans.topvars_centroid1)), booktabs=T, caption='Fisrt Centroid') %>%
  kable_styling( latex_options=c('hold_position', 'scale_down'))
```

Table 13: Fisrt Centroid

fullbase	garagepl	driveway	bathrms	stories	bedrooms
0.3213213	0.3213213	0.7867868	1.126126	1.573574	2.750751

```
kmeans.topvars_centroid2 <- kmeans.centroids[2, order(kmeans.centroids[2, ])]
kable(t(tail(kmeans.topvars_centroid2)), booktabs=T, caption='Second Centroid') %>%
  kable_styling( latex_options=c('hold_position', 'scale_down'))
```

Table 14: Second Centroid

lotsize	driveway	garagepl	bathrms	stories	bedrooms
0.836792	0.971831	1.272301	1.535211	2.173709	3.30047

5. *Supervised Machine Learning*

13. Divide the Housing data into two equally sized samples (one for training and one for testing). The dependent variable is price. Using the training sample, estimate a ridge model using the Housing dataset and find the optimal value of λ.

ANS.

14. How does the model performs in the testing sample? Compare the results of the ridge model with a linear regression. Which model performs best?

ANS.

15. Using the HealthInsurance dataset. Divide the data into two equally sized samples (one for training and one for testing). The dependent variable is health. Using the training sample; and a radial kernel and the following two values for cost C = (1e - 05, 1e + 01), estimate a support vector machine model and choose the optimal cost parameter using the function tune.(In this part, feel free to reduce the size of each sample to improve the speed of the calculations.)

ANS.

16. How does the svm model performs in the testing sample? How does the model compares to a logit in terms of accuracy?

ANS.