Intro to Computational Statistics - Final Project

Vandana Thannir (worked alone)

April 14, 2021

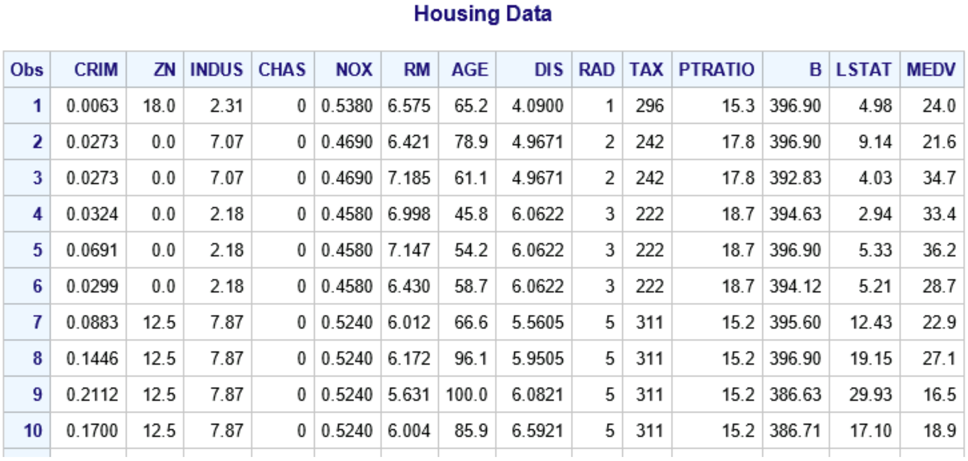
1 Problem 1

Downloaded data set

2 Problem 2

2.1 Part A

Naming the 14 variables results in the following table:
Output of Table 2.1 (only first 10 obs to make it simpler)

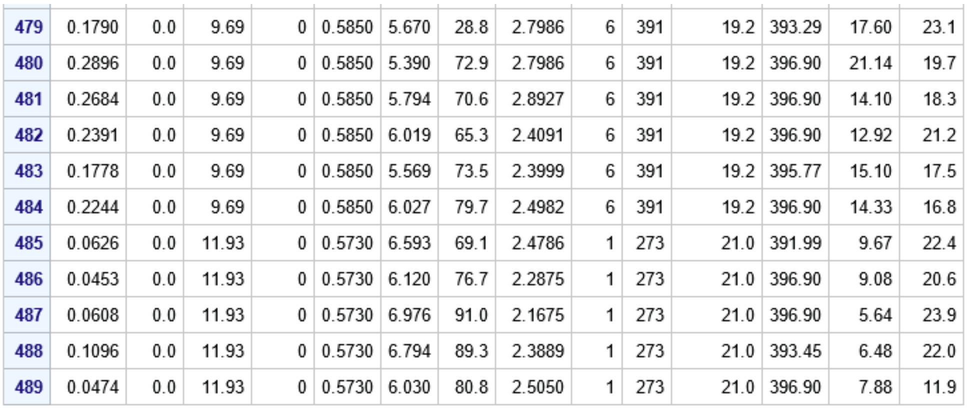


Housing Data

Obs	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
1	0.0063	18.0	2.31	0	0.5380	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	24.0
2	0.0273	0.0	7.07	0	0.4690	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6
3	0.0273	0.0	7.07	0	0.4690	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
4	0.0324	0.0	2.18	0	0.4580	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
5	0.0691	0.0	2.18	0	0.4580	7.147	54.2	6.0622	3	222	18.7	396.90	5.33	36.2
6	0.0299	0.0	2.18	0	0.4580	6.430	58.7	6.0622	3	222	18.7	394.12	5.21	28.7
7	0.0883	12.5	7.87	0	0.5240	6.012	66.6	5.5605	5	311	15.2	395.60	12.43	22.9
8	0.1446	12.5	7.87	0	0.5240	6.172	96.1	5.9505	5	311	15.2	396.90	19.15	27.1
9	0.2112	12.5	7.87	0	0.5240	5.631	100.0	6.0821	5	311	15.2	386.63	29.93	16.5
10	0.1700	12.5	7.87	0	0.5240	6.004	85.9	6.5921	5	311	15.2	386.71	17.10	18.9

2.2 Part B

Removing the outliers results in the following table:
Output of Table 2.2 (only last 10 obs to make it simpler)



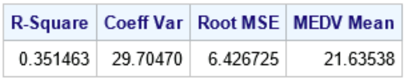
479	0.1790	0.0	9.69	0	0.5850	5.670	28.8	2.7986	6	391	19.2	393.29	17.60	23.1
480	0.2896	0.0	9.69	0	0.5850	5.390	72.9	2.7986	6	391	19.2	396.90	21.14	19.7
481	0.2684	0.0	9.69	0	0.5850	5.794	70.6	2.8927	6	391	19.2	396.90	14.10	18.3
482	0.2391	0.0	9.69	0	0.5850	6.019	65.3	2.4091	6	391	19.2	396.90	12.92	21.2
483	0.1778	0.0	9.69	0	0.5850	5.569	73.5	2.3999	6	391	19.2	395.77	15.10	17.5
484	0.2244	0.0	9.69	0	0.5850	6.027	79.7	2.4982	6	391	19.2	396.90	14.33	16.8
485	0.0626	0.0	11.93	0	0.5730	6.593	69.1	2.4786	1	273	21.0	391.99	9.67	22.4
486	0.0453	0.0	11.93	0	0.5730	6.120	76.7	2.2875	1	273	21.0	396.90	9.08	20.6
487	0.0608	0.0	11.93	0	0.5730	6.976	91.0	2.1675	1	273	21.0	396.90	5.64	23.9
488	0.1096	0.0	11.93	0	0.5730	6.794	89.3	2.3889	1	273	21.0	393.45	6.48	22.0
489	0.0474	0.0	11.93	0	0.5730	6.030	80.8	2.5050	1	273	21.0	396.90	7.88	11.9

After removing any observation with MEDV = 50 or RM = 8.780, we find that the data set now contains 489 observations, as opposed to the previous 506. Thus, we can conclude that there are 17 observations where either MEDV = 50 or RM = 8.780. These outliers could be harmful to our modeling because they can distort the results of the data, and cause us to make observations or analysis that are false. These observations are not reflective of the population as a whole, so they should be removed.

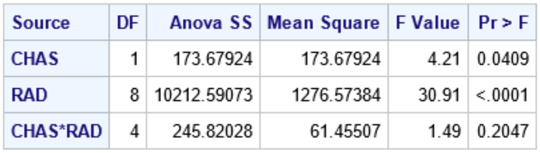
3 Problem 3

3.1 Part A

Output of Table 3.1



R-Square	Coeff Var	Root MSE	MEDV Mean
0.351463	29.70470	6.426725	21.63538



Source	DF	Anova SS	Mean Square	F Value	Pr > F
CHAS	1	173.67924	173.67924	4.21	0.0409
RAD	8	10212.59073	1276.57384	30.91	<.0001
CHAS*RAD	4	245.82028	61.45507	1.49	0.2047

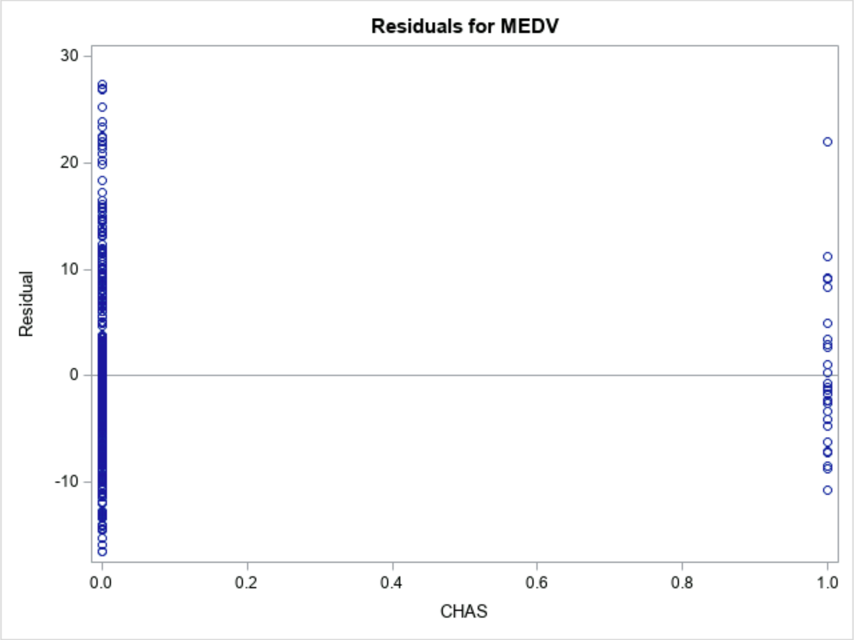
Null Hypothesis:
μ_0 = CHAS and RAD do not have an interaction affect on MEDV
Alternate Hypothesis:
μ_0 = CHAS and RAD do have an interaction affect on MEDV

Two-Way ANOVA Test
Interaction term (CHAS * RAD) is not significant (p-value=0.2047). CHAS is significant (p-value=0.0409). RAD is significant (p-value=0.0001). Therefore, we can not reject the null, and can assume that CHAS and RAD do not have an interaction affect on MEDV. However, CHAS and RAD do independently have an effect on MEDV.

3.2 Part B

The model assumptions are that there is a linear relationship between the variables, and that the residuals are independent, normal, and have constant variance.

We can check the model assumptions by checking the residuals. For instance, after performing a residual test for res vs. chas, I obtained the following plot.



As we can see, the error randomly scatters about zero, and this is the same result that we obtain for all of the residuals. Performing the normality test also further proves validity.

4 Appendix with Code

```
FILENAME housing 'downloads/housing.data';
*Problem 2A of the homework;
DATA housing;
        INFILE housing;
        *Name the variables;
        INPUT CRIM ZN INDUS CHAS NOX RM AGE DIS RAD TAX PTRATIO B LSTAT MEDV;
RUN;
PROC PRINT data=housing;
TITLE 'Housing_Data';
RUN;

*Problem 2B of the homework;
DATA housing2;
        SET housing;
        *Removes outliers;
    IF NOT (MEDV=50 or RM = 8.780)  THEN OUTPUT housing2;
RUN;
PROC PRINT data = housing2;
RUN;

*Problem 3A of the homework;
```

```
PROC ANOVA DATA=housing2;
CLASS CHAS RAD;
*MODEL MEDV = CHAS RAD CHAS*RAD;
MODEL MEDV = CHAS|RAD;
*MEANS MEDV/alpha=.01;
RUN;
*Problem 3B of the homework;
PROC REG DATA=housing2;
MODEL MEDV= CHAS;
RUN;
PROC REG DATA=housing2;
MODEL MEDV= RAD;
RUN;
```