Assignment 2: Multiple Regression Analysis

Keith G. Williams

800690755

Tuesday, June 1, 2015

Completed as part of DSBA 6201, SUM 1, 2015

The Data Set

The “Boston Housing” dataset recorded properties of 506 housing zones in the Greater Boston area. Typically, one is interested in predicting MEDV (median home value) based on other attributes.

Here is a list of attribute information:

1. CRIM: per capita crime rate by town

2. ZN: proportion of residential land zoned for lots over 25,000 ft2

3. INDUS: proportion of non-retail business acres per town

4. CHAS: Charles River dummy variable (=1 if tract bounds river; 0 otherwise)

5.NOX: nitric oxides concentration (parts per 10 million)

6. RM: average number of rooms per dwelling

7. AGE: proportion of owner-occupied units built prior to 1940

8. DIS: weighted distances to five Boston employment centers

9. RAD: index of accessibility to radial highways

10. TAX: full-value property-tax rate per $10,000

11. PTRATIO: pupil-teacher ratio by town

12. LSTAT: % lower status of the population

13. MEDV: median value of owner-occupied homes in $1000’s

Pre-processing

MEDV has somewhat longish tail and is not so normally distributed, so we will take the log transform, and then predict LMEDV instead.

|  |
| --- |
| **Histogram of MEDV** |

|  |
| --- |
|  |

**PROC** **SQL**;

CREATE TABLE WORK.QUERY\_FOR\_BOSTON\_HOUSING AS

SELECT t1.CRIM,

t1.ZN,

t1.INDUS,

t1.CHAS,

t1.NOX,

t1.RM,

t1.AGE,

t1.DIS,

t1.RAD,

t1.TAX,

t1.PTRATIO,

t1.LSTAT,

t1.MEDV,

/\* LMEDV \*/

(LOG(t1.MEDV)) AS LMEDV

FROM WORK.'BOSTON HOUSING'n t1;

**QUIT**;

|  |
| --- |
| **Histogram of log(MEDV)** |

|  |
| --- |
|  |

Questions

1. **Please perform the multicollinearity diagnosis based on the VIF calculation results. Do we need to drop any variables that might have multicollinearity concerns?**

|  |
| --- |
|  |
| **The REG Procedure Model: Linear\_Regression\_Model Dependent Variable: LMEDV** |
| | Number of Observations Read | 506 | | --- | --- | | Number of Observations Used | 506 | |
| | **Analysis of Variance** | | | | | | | --- | --- | --- | --- | --- | --- | | **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** | | Model | 12 | 66.09326 | 5.50777 | 148.51 | <.0001 | | Error | 493 | 18.28323 | 0.03709 |  |  | | Corrected Total | 505 | 84.37649 |  |  |  | |
| | Root MSE | 0.19258 | R-Square | 0.7833 | | --- | --- | --- | --- | | Dependent Mean | 3.03451 | Adj R-Sq | 0.7780 | | Coeff Var | 6.34620 |  |  | |
| | **Parameter Estimates** | | | | | | | | --- | --- | --- | --- | --- | --- | --- | | **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** | **Variance Inflation** | | Intercept | 1 | 4.33112 | 0.19812 | 21.86 | <.0001 | 0 | | CRIM | 1 | -0.01087 | 0.00132 | -8.20 | <.0001 | 1.76749 | | ZN | 1 | 0.00120 | 0.00055706 | 2.15 | 0.0322 | 2.29846 | | INDUS | 1 | 0.00215 | 0.00249 | 0.86 | 0.3887 | 3.98718 | | CHAS | 1 | 0.10769 | 0.03492 | 3.08 | 0.0022 | 1.07117 | | NOX | 1 | -0.82243 | 0.15458 | -5.32 | <.0001 | 4.36909 | | RM | 1 | 0.08409 | 0.01687 | 4.99 | <.0001 | 1.91253 | | AGE | 1 | 0.00034023 | 0.00053500 | 0.64 | 0.5251 | 3.08823 | | DIS | 1 | -0.04976 | 0.00809 | -6.15 | <.0001 | 3.95404 | | RAD | 1 | 0.01353 | 0.00269 | 5.04 | <.0001 | 7.44530 | | TAX | 1 | -0.00064120 | 0.00015256 | -4.20 | <.0001 | 9.00216 | | PTRATIO | 1 | -0.03760 | 0.00531 | -7.09 | <.0001 | 1.79706 | | LSTAT | 1 | -0.03025 | 0.00203 | -14.88 | <.0001 | 2.87078 | |

The TAX variable has a high variance inflation factor, and can be seen to be highly correlated with RAD; since this violates the assumption of no multicollinearity, it will be removed from the model.

| **Pearson Correlation Coefficients, N = 506  Prob > |r| under H0: Rho=0** | | |
| --- | --- | --- |
|  | **RAD** | **TAX** |
| RAD | |  | | --- | | 1.00000 | |  | | |  | | --- | | 0.91023 | | <.0001 | |
| TAX | |  | | --- | | 0.91023 | | <.0001 | | |  | | --- | | 1.00000 | |  | |

1. **Please run the linear regression analyses.**

* **Use the stepwise model selection approach to determine the final model. Drop variables based on their significance.**
* **Report summary output for *each step*, including ANOVA, R2, and parameter estimates.**

|  |
| --- |
| **Linear Regression Results**  **The REG Procedure Model: Linear\_Regression\_Model Dependent Variable: LMEDV** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| | Number of Observations Read | 506 | | --- | --- | | Number of Observations Used | 506 | |
| | **Analysis of Variance** | | | | | | | --- | --- | --- | --- | --- | --- | | **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** | | Model | 11 | 65.43814 | 5.94892 | 155.18 | <.0001 | | Error | 494 | 18.93835 | 0.03834 |  |  | | Corrected Total | 505 | 84.37649 |  |  |  | |
| | Root MSE | 0.19580 | R-Square | 0.7755 | | --- | --- | --- | --- | | Dependent Mean | 3.03451 | Adj R-Sq | 0.7706 | | Coeff Var | 6.45236 |  |  | |
| | **Parameter Estimates** | | | | | | | | --- | --- | --- | --- | --- | --- | --- | | **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** | **Variance Inflation** | | Intercept | 1 | 4.24264 | 0.20029 | 21.18 | <.0001 | 0 | | CRIM | 1 | -0.01082 | 0.00135 | -8.03 | <.0001 | 1.76735 | | ZN | 1 | 0.00067450 | 0.00055212 | 1.22 | 0.2224 | 2.18417 | | INDUS | 1 | -0.00245 | 0.00228 | -1.08 | 0.2822 | 3.21795 | | CHAS | 1 | 0.12571 | 0.03523 | 3.57 | 0.0004 | 1.05502 | | NOX | 1 | -0.87235 | 0.15670 | -5.57 | <.0001 | 4.34330 | | RM | 1 | 0.08919 | 0.01710 | 5.21 | <.0001 | 1.90264 | | AGE | 1 | 0.00027655 | 0.00054373 | 0.51 | 0.6112 | 3.08576 | | DIS | 1 | -0.05044 | 0.00823 | -6.13 | <.0001 | 3.95244 | | RAD | 1 | 0.00459 | 0.00167 | 2.75 | 0.0061 | 2.77221 | | PTRATIO | 1 | -0.03926 | 0.00538 | -7.30 | <.0001 | 1.78705 | | LSTAT | 1 | -0.03015 | 0.00207 | -14.59 | <.0001 | 2.87041 | |

AGE has the highest p-value at p = 0.61, so it will be dropped.

|  |
| --- |
| **Linear Regression Results** |

|  |
| --- |
| **The REG Procedure Model: Linear\_Regression\_Model Dependent Variable: LMEDV** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| | Number of Observations Read | 506 | | --- | --- | | Number of Observations Used | 506 | |
| | **Analysis of Variance** | | | | | | | --- | --- | --- | --- | --- | --- | | **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** | | Model | 10 | 65.42823 | 6.54282 | 170.92 | <.0001 | | Error | 495 | 18.94827 | 0.03828 |  |  | | Corrected Total | 505 | 84.37649 |  |  |  | |
| | Root MSE | 0.19565 | R-Square | 0.7754 | | --- | --- | --- | --- | | Dependent Mean | 3.03451 | Adj R-Sq | 0.7709 | | Coeff Var | 6.44753 |  |  | |
| | **Parameter Estimates** | | | | | | | | --- | --- | --- | --- | --- | --- | --- | | **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** | **Variance Inflation** | | Intercept | 1 | 4.23589 | 0.19970 | 21.21 | <.0001 | 0 | | CRIM | 1 | -0.01082 | 0.00135 | -8.04 | <.0001 | 1.76731 | | ZN | 1 | 0.00064168 | 0.00054792 | 1.17 | 0.2421 | 2.15434 | | INDUS | 1 | -0.00244 | 0.00228 | -1.07 | 0.2847 | 3.21743 | | CHAS | 1 | 0.12663 | 0.03516 | 3.60 | 0.0003 | 1.05226 | | NOX | 1 | -0.85125 | 0.15100 | -5.64 | <.0001 | 4.03899 | | RM | 1 | 0.09093 | 0.01675 | 5.43 | <.0001 | 1.82640 | | DIS | 1 | -0.05167 | 0.00786 | -6.57 | <.0001 | 3.61317 | | RAD | 1 | 0.00450 | 0.00166 | 2.72 | 0.0068 | 2.74684 | | PTRATIO | 1 | -0.03903 | 0.00536 | -7.29 | <.0001 | 1.77476 | | LSTAT | 1 | -0.02980 | 0.00195 | -15.31 | <.0001 | 2.54673 | |

INDUS has the next highest p-value at p = 0.28, so it will be dropped.

|  |
| --- |
| **Linear Regression Results** |

|  |
| --- |
| **The REG Procedure Model: Linear\_Regression\_Model Dependent Variable: LMEDV** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| | Number of Observations Read | 506 | | --- | --- | | Number of Observations Used | 506 | |
| | **Analysis of Variance** | | | | | | | --- | --- | --- | --- | --- | --- | | **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** | | Model | 9 | 65.38432 | 7.26492 | 189.73 | <.0001 | | Error | 496 | 18.99218 | 0.03829 |  |  | | Corrected Total | 505 | 84.37649 |  |  |  | |
| | Root MSE | 0.19568 | R-Square | 0.7749 | | --- | --- | --- | --- | | Dependent Mean | 3.03451 | Adj R-Sq | 0.7708 | | Coeff Var | 6.44848 |  |  | |
| | **Parameter Estimates** | | | | | | | | --- | --- | --- | --- | --- | --- | --- | | **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** | **Variance Inflation** | | Intercept | 1 | 4.24137 | 0.19966 | 21.24 | <.0001 | 0 | | CRIM | 1 | -0.01076 | 0.00134 | -8.00 | <.0001 | 1.76426 | | ZN | 1 | 0.00064853 | 0.00054797 | 1.18 | 0.2372 | 2.15405 | | CHAS | 1 | 0.12459 | 0.03512 | 3.55 | 0.0004 | 1.04919 | | NOX | 1 | -0.90820 | 0.14135 | -6.43 | <.0001 | 3.53822 | | RM | 1 | 0.09335 | 0.01660 | 5.62 | <.0001 | 1.79324 | | DIS | 1 | -0.04966 | 0.00763 | -6.51 | <.0001 | 3.40711 | | RAD | 1 | 0.00428 | 0.00164 | 2.60 | 0.0096 | 2.70103 | | PTRATIO | 1 | -0.04008 | 0.00527 | -7.61 | <.0001 | 1.71584 | | LSTAT | 1 | -0.03000 | 0.00194 | -15.49 | <.0001 | 2.52325 | |

ZN has the next highest p-value at p = 0.24, and is the final predictor above the α = 0.05 threshold, so it will be dropped to produce the final model.

### **Final model**

|  |
| --- |
| **Linear Regression Results** |

|  |
| --- |
| **The REG Procedure Model: Linear\_Regression\_Model Dependent Variable: LMEDV** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| | Number of Observations Read | 506 | | --- | --- | | Number of Observations Used | 506 | |
| | **Analysis of Variance** | | | | | | | --- | --- | --- | --- | --- | --- | | **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** | | Model | 8 | 65.33068 | 8.16634 | 213.10 | <.0001 | | Error | 497 | 19.04581 | 0.03832 |  |  | | Corrected Total | 505 | 84.37649 |  |  |  | |
| | Root MSE | 0.19576 | R-Square | 0.7743 | | --- | --- | --- | --- | | Dependent Mean | 3.03451 | Adj R-Sq | 0.7706 | | Coeff Var | 6.45108 |  |  | |
| | **Parameter Estimates** | | | | | | | --- | --- | --- | --- | --- | --- | | **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** | | Intercept | 1 | 4.25599 | 0.19936 | 21.35 | <.0001 | | CRIM | 1 | -0.01062 | 0.00134 | -7.93 | <.0001 | | CHAS | 1 | 0.12379 | 0.03512 | 3.52 | 0.0005 | | NOX | 1 | -0.91847 | 0.14114 | -6.51 | <.0001 | | RM | 1 | 0.09583 | 0.01647 | 5.82 | <.0001 | | DIS | 1 | -0.04521 | 0.00664 | -6.80 | <.0001 | | RAD | 1 | 0.00453 | 0.00163 | 2.78 | 0.0057 | | PTRATIO | 1 | -0.04209 | 0.00499 | -8.44 | <.0001 | | LSTAT | 1 | -0.02997 | 0.00194 | -15.47 | <.0001 | |

LMEDV = 4.26 − 0.011 \* CRIM + 0.124 \* CHAS − 0.918 \* NOX + 0.096 \* RM − 0.045 \* DIS + 0.004 \* RAD − 0.042 \* PTRATIO − 0.030 \* LSTAT + e

### **Interpretation of Coefficients:**

The natural log of median home value ($1000) changes by the estimated coefficient for each predictor for each unit change in that predictor, while all other predictors are held constant.

CRIM:

The log median home value ($1000) decreases on average by 0.011 ± 0.003 for each unit increase in the per capita crime rate, while all other predictors are held constant.

CHAS:

The log median home value ($1000) is on average 0.124 ± 0.069 higher for homes that bound the Charles River versus those that don’t, while all other predictors are held constant.

NOX:

The log median home value ($1000) decreases on average by 0.918 ± 0.277 for each parts per 10 million increase in nitric oxide concentration, while all other predictors are held constant.

RM:

The log median home value ($1000) increases on average by 0.096 ± 0.032 for each 1 room increase in the average number of rooms per dwelling, while all other predictors are held constant.

DIS:

The log median home value ($1000) decreases on average by 0.045 ± 0.013 for each 1 unit increase in the weighted distances to five Boston employment centers, while all other predictors are held constant.

RAD:

The log median home value ($1000) increases on average by 0.004 ± 0.003 for each 1 unit increase in the index of accessibility to radial highways, while all other predictors are held constant.

PTRATIO:

The log median home value ($1000) decreases on average by 0.042 ± 0.010 for each 1 pupil increase in the pupil-teacher ratio, while all other predictors are held constant.

LSTAT:

The log median home value ($1000) decreases on average by 0.030 ± 0.004 for each 1% increase in the percent lower status of the population, while all other predictors are held constant.

The 95% confidence interval for each coefficient is calculated by obtaining the 0.975 t-quantile on 497 degrees of freedom multiplied by the standard error for that coefficient. This value is added and subtracted from the fitted value at each Xi to get the upper and lower bounds.

Finally the R2 value of 0.77 means that 77% of the variance in the data is explained by the model.