Hw3

2025-09-18

Consider the house price data in Table B.4. a. Fit a multiple regression model relating selling price to all nine regressors. b. Test for significance of regression. What conclusions can you draw? c. Use t tests to assess the contribution of each regressor to the model. Discuss your findings. d. What is the contribution of lot size and living space to the model given that all of the other regressors are included? e. Is multicollinearity a potential problem in this model?

y : Sale price of the house/1000 x1 : Taxes (local, school, county)/1000 x2 : Number of baths x3 : Lot size (sq ft × 1000) x4 : Living space (sq ft × 1000) x5 : Number of garage stalls x6 : Number of rooms x7 : Number of bedrooms x8 : Age of the home (years) x9 : Number of fi replaces

```
library(ggplot2)
library(moderndive)
```

```
## Warning: package 'moderndive' was built under R version 4.3.3
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
df<- read.csv("B4.csv")
head(df)
```

```
##      y     x1 x2    x3    x4 x5 x6 x7 x8 x9
## 1 29.5 5.0208  1 3.531 1.500  2  7  4 62  0
## 2 27.9 4.5429  1 2.275 1.175  1  6  3 40  0
## 3 25.9 4.5573  1 4.050 1.232  1  6  3 54  0
## 4 29.9 5.0597  1 4.455 1.121  1  6  3 42  0
## 5 29.9 3.8910  1 4.455 0.988  1  6  3 56  0
## 6 30.9 5.8980  1 5.850 1.240  1  7  3 51  1
```

```
#1 regression using all 9 regressors

model <- lm(y~x1+x2+x3+x4+x5+x6+x7+x8+x9,data=df)
get_regression_table(model)
```

```
## # A tibble: 10 x 7
##    term      estimate std_error statistic p_value lower_ci upper_ci
##    <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
##  1 intercept   14.9       5.91      2.52    0.024    2.25    27.6
##  2 x1           1.92      1.03      1.87    0.083   -0.284    4.13
##  3 x2           7.00      4.3       1.63    0.126   -2.22    16.2
##  4 x3           0.149     0.49      0.304   0.765   -0.903    1.20
##  5 x4           2.72      4.36      0.625   0.542   -6.63    12.1
##  6 x5           2.01      1.37      1.46    0.166   -0.939    4.95
##  7 x6          -0.41      2.38     -0.172   0.866   -5.51     4.69
##  8 x7          -1.40      3.40     -0.413   0.686   -8.69     5.88
##  9 x8          -0.037     0.067    -0.557   0.586   -0.18     0.106
## 10 x9           1.56      1.94      0.805   0.434   -2.60     5.72
```

#2 From the regression table above. the intercept so when all other values x1-x9 are 0 the selling price or the Y value is estimated to be 14.98. This value for the intercept is statistically significant because the p-value is 0.024 so we can reject the null hypothesis.

On the other hand for the sake of time instead of analyzing each regression individually from the graph we can see all the other regressors x1-x9 have much lower estimates. They also all have P-values greater than 0.05 so we can not reject the null hypothesis so from the regresson above we can not determine if the regressors are significant or not yet.

```
#3
summary(model)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9,
##     data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.720 -1.956 -0.045  1.627  4.253
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.92765    5.91285   2.525   0.0243 *
## x1           1.92472    1.02990   1.869   0.0827 .
## x2           7.00053    4.30037   1.628   0.1258
## x3           0.14918    0.49039   0.304   0.7654
## x4           2.72281    4.35955   0.625   0.5423
## x5           2.00668    1.37351   1.461   0.1661
## x6          -0.41012    2.37854  -0.172   0.8656
## x7          -1.40324    3.39554  -0.413   0.6857
## x8          -0.03715    0.06672  -0.557   0.5865
## x9           1.55945    1.93750   0.805   0.4343
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.949 on 14 degrees of freedom
## Multiple R-squared:  0.8531, Adjusted R-squared:  0.7587
## F-statistic: 9.037 on 9 and 14 DF,  p-value: 0.000185
```

Using the summary function we can see the t-values of all the Regressors. When comparing our t-test value to 2 we see all of the t values x1-x9 are less then 2. So we do not have strong evidence that any of them are statistically significant contributers. But we some stil contribute more then others even though evidence is weak. For example x1, x2, and x5 all contribute even though there is weak evidence. there values are much closer to 2 then the rest. Combining this with the fact they all have high p-values and low t values means they may not have any effect on the y intercept.

```
model2 <-lm(y~x1+x2+x5+x6+x7+x8+x9, data=df)
anova(model2, model)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x1 + x2 + x5 + x6 + x7 + x8 + x9
## Model 2: y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     16 127.36
## 2     14 121.75  2    5.6083 0.3225 0.7296
```

4. To test the joint contribution of lot size (x3) and living space (x4), we compared the full model (all nine regressors) to a reduced model excluding x3 and x4. The partial F-test gave F(2,14) = 0.32 with p = 0.73. Since the p-value is much greater than 0.05, we fail to reject the null hypothesis that = = 0. This indicates that, after accounting for the other variables, lot size and living space do not significantly improve the fit of the model.

5. In this housing dataset, multicollinearity is a potential problem because variables such as living space, number of rooms, and number of bedrooms likely overlap in the information they provide. This can inflate standard errors and explain why many t-tests were insignificant despite the model potentially fitting the data reasonably well overall.

But All the Vifs are less than 10 so its not a problem.