STATS 500 Test 1

Martin Zanaj.

Academic Honor Code

"*I have **not** used any resources from outside the class or discussed the exam with anyone.*" -Martin Zanaj

1. The dataset **hprice** is data representing housing prices in 36 US metropolitan statistical areas (MSAs) over 9 years from 1986-1994. It is made up of 8 different variables. The variables are: narsp, ypc, perypc, regtest, rcdum, ajwtr, msa, time from these ones four are quantitative & four are categorical.

Variable	Type	Description
narsp	Quantitative	natural log average sale price in thousands of dollars
ypc	Quantitative	average per capita income
perypc	Quantitative	percentage growth in per capita income
regtest	Quantitative	Regulatory environment index
rcdum	Categorical	Rent control (0=no, 1=yes)
ajwtr	Categorical	Adjacent to a coastline (0=no, 1=yes)
msa	Categorical	indicator for the MSA (1-36)
time	Categorical?	Year 1=1986 to 9= 1994

From the summary below we can see the overall distribution & numerical frequency. Perhaps, some 'interesting' numerical facts that can immediately grab the eye are: negative score in perypc (does negative data make sense in this case), no rent control is by far more popular than rent control, more houses seem to be away from coastline (make sense given limited coastline). Further analysis will reveal their importance, or lack thereof.

```
##      narsp            ypc            perypc          regtest       rcdum
##  Min.   :3.920   Min.   :12535   Min.   :-2.054   Min.   :13.00   0:279
##  1st Qu.:4.264   1st Qu.:16609   1st Qu.: 3.535   1st Qu.:18.00   1: 45
##  Median :4.412   Median :18454   Median : 3.964   Median :20.00
##  Mean   :4.484   Mean   :18769   Mean   : 4.268   Mean   :20.42
##  3rd Qu.:4.575   3rd Qu.:20323   3rd Qu.: 5.711   3rd Qu.:22.00
##  Max.   :5.563   Max.   :33383   Max.   : 8.788   Max.   :29.00
##
##  ajwtr        msa           time
##  0:189   1      :  9   Min.   :1
##  1:135   2      :  9   1st Qu.:3
##          3      :  9   Median :5
##          4      :  9   Mean   :5
##          5      :  9   3rd Qu.:7
##          6      :  9   Max.   :9
##          (Other):270
```

2. Linear regression model with narsp as the response and ypc, perypc, regtest, rcdum, and time as

predictors.

```
##
## Call:
## lm(formula = narsp ~ ypc + perypc + regtest + rcdum + time, data = hprice)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.31051 -0.11653 -0.01862  0.07919  0.57618
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.661e+00  8.458e-02  31.460  < 2e-16 ***
## ypc          7.180e-05  4.291e-06  16.735  < 2e-16 ***
## perypc      -1.387e-02  5.091e-03  -2.725 0.006794 **
## regtest      2.973e-02  3.112e-03   9.555  < 2e-16 ***
## rcdum1       1.587e-01  3.199e-02   4.960 1.15e-06 ***
## time        -1.886e-02  5.103e-03  -3.695 0.000258 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1656 on 318 degrees of freedom
## Multiple R-squared:  0.7547, Adjusted R-squared:  0.7508
## F-statistic: 195.7 on 5 and 318 DF,  p-value: < 2.2e-16
```

3. The **RSS** is equal to **8.72**. Now refering to Ch.2 of "Linear Models with R" on page 16, there is a clear formula for calculating tha variance of e. The esimtate for such variance is represented by simga_hat^2 which is equal to RSS/df.residuals, where df.residuals is eqaul to (n-p) [total observations - (#predictors +1)]{324-6}.This formula can be rearranged, so as to solve for RSS by RSS= sigma_hat^2 * df.residuals. Both of these values can be found in the regression summary output where **sigma_hat = 0.1656** and **df. residulas = 318**. Hence, our calculations can be carried as **0.1656^2 * 318 = 8.72**. The results can be checked through the deviance() command which gives the RSS for our model.

```r
#RSS calculartion
(0.1656**2)*318
```

```
## [1] 8.720628
```

```r
#RSS
deviance(lmod)
```

```
## [1] 8.723506
```

4. The **total sum of sqaures** is **35.55**. Again, by making use of the previous results, summary, and our dear book (page 23) the original formula is R^2 = 1- RSS/TSS, where R^2 is the *multiple R-squared* from the linear regression output;the RSS is the residual sum of squares, and the TSS is the total sum of squares. This formula can be rearranged, so as to solve for TSS by TSS = RSS/ (1-R^2). Hence, TSS = **8.72/(1-0.7547) = 35.54**. The results can be checked by taking the sqaured deviance of response - mean(response) [sum((yi-ybar)**2)].

```r
#TSS=RSS/(1-R^2)
(8.72)/(1-0.7547)
```

```
## [1] 35.54831
```

```r
#Check
sum((hprice$narsp- mean(hprice$narsp))**2)
```

```
## [1] 35.5627
```

5. Dropping the variable **rcdum** gives the bigger reduction in R2 (i.e. 0.7547-0.7357=0.019). This is a way to infer about variable importance- having a bigger reduction in R^2 when a particular variable is dropped is a sign that the variable is indeed important in explaining the variance in the response.

   Summary of fit of the model without *perypc* variable. Total reduction from original fit **0.0057**.

```
##
## Call:
## lm(formula = narsp ~ ypc + regtest + rcdum + time, data = hprice)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.31188 -0.11171 -0.01991  0.07752  0.55794
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.593e+00  8.168e-02  31.750  < 2e-16 ***
## ypc          7.117e-05  4.327e-06  16.447  < 2e-16 ***
## regtest      2.960e-02  3.143e-03   9.418  < 2e-16 ***
## rcdum1       1.587e-01  3.231e-02   4.911 1.45e-06 ***
## time        -1.425e-02  4.863e-03  -2.931  0.00363 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1673 on 319 degrees of freedom
## Multiple R-squared:  0.749,  Adjusted R-squared:  0.7458
## F-statistic: 237.9 on 4 and 319 DF,  p-value: < 2.2e-16
```

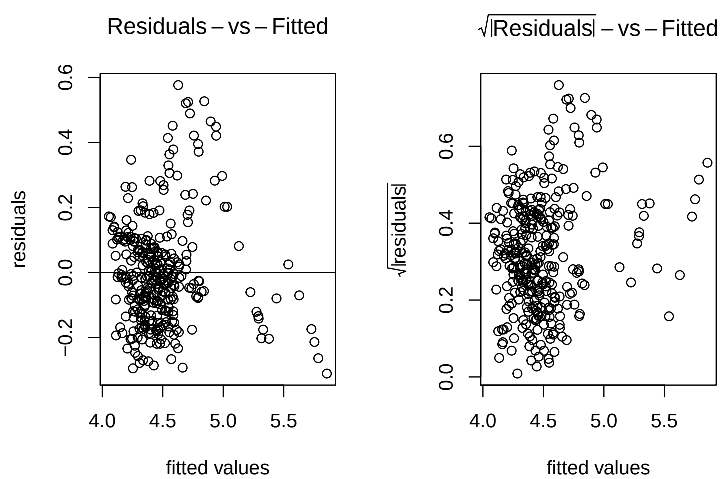
   Summary of fit of the model without *rcdum* variable. Total reduction from original fit **0.019**.

```
##
## Call:
## lm(formula = narsp ~ ypc + perypc + regtest + time, data = hprice)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32096 -0.11125 -0.02201  0.08062  0.62596
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.461e+00  7.711e-02  31.920  < 2e-16 ***
```

```
## ypc          8.006e-05  4.098e-06  19.539  < 2e-16 ***
## perypc      -1.387e-02  5.276e-03  -2.629  0.00898 **
## regtest      3.452e-02  3.066e-03  11.257  < 2e-16 ***
## time        -2.510e-02  5.124e-03  -4.898 1.54e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1716 on 319 degrees of freedom
## Multiple R-squared:  0.7357, Adjusted R-squared:  0.7324
## F-statistic:   222 on 4 and 319 DF,  p-value: < 2.2e-16
```

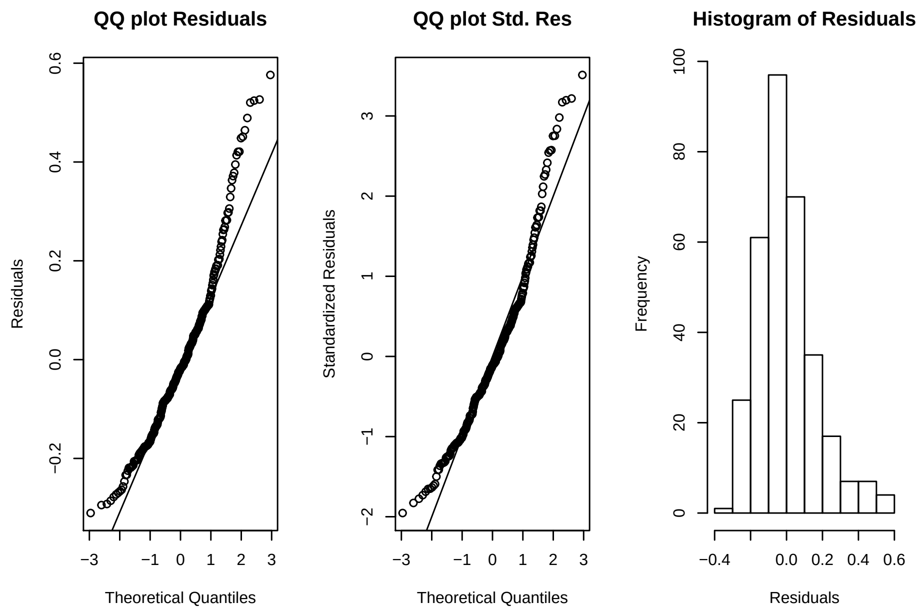
6. **Yes**, one can infer about the importance of the predictors by looking at the **p-values** in the original regression output (first fit). Through these p-values one can learn the respective satistical significance of each predictor. In our case, comparing the p-value for perypc & rcdum does indeed confirm the results in question 5- rcdum has a lower p-value (more significant) than perypc. Hence, rcdum is more importan than perypc. As a result, taking away an important predictor from the model will yield in a bigger reduction in Rˆ2 than an unimportant one.

7. The residuals against the fitted values is a common diagnostics graphical tool to check for the *constant variance* assumption that we make when performing regression. In order to check if this assumption holds, we plot residuals agains fitted values. For a better resolution one can take the square root of the abs(residuals). It is not super clear, but the constant variance assumption might not be valid because the errors seem to follow particular trends for different regions in the fitted values space (e.g. errors for 4 to 4.5 seem to be smaller than errors from 4.5 to 5). There should be no pattern in order for the constant variance assumption to hold. We can verify the fact that there might be a relationship between residuals & fitted by running a linear regression. The Rˆ2 from this model is 0.036 (should be zero if no relationship) which does confirm our doubts about the possible non-constant variance.




```
## [1] 0.03597842
```

8. The QQ plot of residuals is a common diagnostics graphical tool to check for the *normality* assumption that we make about residuals when performing linear regression. To confirm the normality result, we can use other methods such as QQ plot (standardized residuals), histogram , & Shapiro-Wilk (H0: residuals are normal) test. All graphical and numerical methods suggest that normaility is **NOT** satisfied.




```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(lmod)
## W = 0.94851, p-value = 3.243e-09
```

9. Observation **40** has the largest residual.

```
#Residuls
res=lmod$residuals
lmod$residuals[which.max(abs(res))]
```

```
##        40
## 0.5761788
```

10. The point is **not** an outlier given that the p-value **0.5649** is bigger than the alpha level **0.0001**. Hence, we fail to reject H0, and conclude that the point is not an outlier.

11. Leverage points are extreme values in the X space. In our case, the observation that have the highest leverage is **54**. One can confirm for such point through a half plot.

```
#Leverages are extreme values in the X spacv
hatv <- hatvalues(lmod)
#Max
hatv[which.max((hatv))]
```

```
##         54
## 0.07916241
```

```r
#Confirm results throguh half normal plot
halfnorm(hatv,ylab="Leverages", main='Half-normal plot for Leverages')
```