Project #1

Gavin Valenzuela

2024-07-04

```
knitr::opts_chunk$set(echo = TRUE, warning = FALSE, message = FALSE)
options(width = 80)
load("C:/Users/valen/Downloads/hp.RData")
library(magrittr)
library(MASS)
library(car)
```

```
## Loading required package: carData
```

```
library(RRPP)
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
model <- lm(sell ~ . , data = hp)
```

```
viftest <- vif(model)
viftest
```

```
##    lotsize   bedrooms   fullbath    stories   driveway recreation   basement
##   1.321632   1.365633   1.282494   1.478584   1.163091   1.210501   1.316543
##        gas centralair     garage  neighbour
##   1.038246   1.201397   1.200839   1.147639
```

```
knitr::opts_chunk$set(echo = TRUE, warning = FALSE, message = FALSE)
options(width = 80)
cat("Removing the stories and bedrooms variables because they have the highest VIFs and it also\n")
```

```
## Removing the stories and bedrooms variables because they have the highest VIFs and it also
```

```
cat("makes logical sense since the larger the lot size, the more likely more bedrooms are, so\n")
```

```
## makes logical sense since the larger the lot size, the more likely more bedrooms are, so
```

```
cat("lot size covers that aspect.\n")
```

```
## lot size covers that aspect.
```

```
model1 <- lm(sell ~ . - stories - bedrooms, data = hp)
#Found a command for generating best variables to use via AIC test which turns out to be all the variab
lm(sell ~ . , data = hp) %>%
```

```
stepAIC(direction = "backward", trace = 0) %>%
summary()
```

```
##
## Call:
## lm(formula = sell ~ lotsize + bedrooms + fullbath + stories +
##     driveway + recreation + basement + gas + centralair + garage +
##     neighbour, data = hp)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -41389  -9307   -591   7353  74875
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4038.3504  3409.4713  -1.184 0.236762
## lotsize         3.5463     0.3503  10.124  < 2e-16 ***
## bedrooms     1832.0035  1047.0002   1.750 0.080733 .
## fullbath    14335.5585  1489.9209   9.622  < 2e-16 ***
## stories      6556.9457   925.2899   7.086 4.37e-12 ***
## driveway     6687.7789  2045.2458   3.270 0.001145 **
## recreation   4511.2838  1899.9577   2.374 0.017929 *
## basement     5452.3855  1588.0239   3.433 0.000642 ***
## gas         12831.4063  3217.5971   3.988 7.60e-05 ***
## centralair  12632.8904  1555.0211   8.124 3.15e-15 ***
## garage       4244.8290   840.5442   5.050 6.07e-07 ***
## neighbour    9369.5132  1669.0907   5.614 3.19e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15420 on 534 degrees of freedom
## Multiple R-squared:  0.6731, Adjusted R-squared:  0.6664
## F-statistic: 99.97 on 11 and 534 DF,  p-value: < 2.2e-16
```

```
AIC(model)
```

```
## [1] 12094.19
```

```
AIC(model1)
```

```
## [1] 12158.24
```

```
# Assuming model is already defined and fitted to your data

# Generate fitted values from the model
fittedmodel <- fitted.values(model)

# Additional regressor like squared and cubed terms
fittedsquared <- fittedmodel^2
fittedcubed <- fittedmodel^3

# Make a new model with the new variables added
newmodel <- lm(sell ~ fittedsquared + fittedcubed, data = hp)

# Check the summary of the new model
summary(newmodel)
```

```
##
## Call:
## lm(formula = sell ~ fittedsquared + fittedcubed, data = hp)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -40377  -8397   -907   7592  68508
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.151e+04  2.250e+03  14.004  < 2e-16 ***
## fittedsquared  8.045e-06  1.141e-06   7.053 5.34e-12 ***
## fittedcubed   -1.091e-11  9.153e-12  -1.192    0.234
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15000 on 543 degrees of freedom
## Multiple R-squared:  0.6857, Adjusted R-squared:  0.6845
## F-statistic: 592.3 on 2 and 543 DF,  p-value: < 2.2e-16
```

```
#performing reset test on the new model
Rtest <- resettest(model, power = 2:3,type = "regressor")
print(Rtest)
```

```
##
##  RESET test
##
## data:  model
## RESET = 0.85583, df1 = 22, df2 = 512, p-value = 0.6546
```

#3

```
# Fit the baseline model
baseline_model <- lm(sell ~ ., data = hp)

# Create a summary of the baseline model
baseline_summary <- summary(baseline_model)

# Print the summary
print(baseline_summary)
```

```
##
## Call:
## lm(formula = sell ~ ., data = hp)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -41389  -9307   -591   7353  74875
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4038.3504  3409.4713  -1.184 0.236762
## lotsize         3.5463     0.3503  10.124  < 2e-16 ***
## bedrooms     1832.0035  1047.0002   1.750 0.080733 .
## fullbath    14335.5585  1489.9209   9.622  < 2e-16 ***
## stories      6556.9457   925.2899   7.086 4.37e-12 ***
```

```
## driveway     6687.7789  2045.2458   3.270 0.001145 **
## recreation   4511.2838  1899.9577   2.374 0.017929 *
## basement     5452.3855  1588.0239   3.433 0.000642 ***
## gas         12831.4063  3217.5971   3.988 7.60e-05 ***
## centralair  12632.8904  1555.0211   8.124 3.15e-15 ***
## garage       4244.8290   840.5442   5.050 6.07e-07 ***
## neighbour    9369.5132  1669.0907   5.614 3.19e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15420 on 534 degrees of freedom
## Multiple R-squared:  0.6731, Adjusted R-squared:  0.6664
## F-statistic: 99.97 on 11 and 534 DF,  p-value: < 2.2e-16
```

```
knitr::opts_chunk$set(echo = TRUE, warning = FALSE, message = FALSE)
options(width = 80)
# Interpretation
#(Comment on the Statistical and
#Economic Significance of Your Individual Estimates
#From the summary output of the baseline model
#*Statistically Significant Variables:
#Variables with p-values less than 0.05 are statistically significant
#Significant variables include lotsize, fullbath, stories,
#driveway, recreation, basement, gas, centralair, garage,
#and neighbour
#*Insignificant Variables
#The variable bedrooms has a p-value of 0.0807,
#which is greater than 0.05, indicating it
#is not statistically significant at the 5% level
```

```
# Identify statistically significant variables
coef_summary <- baseline_summary$coefficients
significant_vars <-
coef_summary[coef_summary[, 4] < 0.05, ]

# Print significant variables
print(significant_vars)
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## lotsize        3.546303     0.3503 10.123618 3.732442e-22
## fullbath   14335.558468  1489.9209  9.621691 2.570369e-20
## stories     6556.945711   925.2899  7.086369 4.374046e-12
## driveway    6687.778890  2045.2458  3.269914 1.145151e-03
## recreation  4511.283826  1899.9577  2.374413 1.792936e-02
## basement    5452.385539  1588.0239  3.433440 6.422381e-04
## gas        12831.406266  3217.5971  3.987885 7.595575e-05
## centralair 12632.890405  1555.0211  8.123935 3.150681e-15
## garage      4244.829004   840.5442  5.050096 6.069790e-07
## neighbour   9369.513239  1669.0907  5.613544 3.189602e-08
```

```
knitr::opts_chunk$set(echo = TRUE, warning = FALSE, message = FALSE)
options(width = 80)
# Economic Significance
# Interpretation of Coefficients:
# - **Lotsize:** The coefficient is 3.5463, meaning a one-unit increase in lot size is associated
```

```
#with an increase of approximately 3.55 units in sell, holding other factors constant.
#This is economically significant as it shows a positive impact on property value.
# - **Fullbath:** The coefficient is 14335.5585,
#meaning an additional full bathroom is
#associated with an increase of approximately
#14335.56 units in sell.
#This is a substantial impact and economically significant.
# - **Stories:** The coefficient is 6556.9457,
#meaning an additional story in the building is associated
#with an increase of approximately 6556.95
#units in sell.
# - **Driveway:** The coefficient is 6687.7789,
#meaning having a driveway is associated
#with an increase of approximately
#6687.78 units in sell.
# - **Recreation:** The coefficient is
#4511.2884, meaning the presence of recreation
#facilities is associated with
#an increase of approximately 4511.29 units in sell.
# - **Basement:** The coefficient is 5452.3835,
#meaning having a basement is associated
#with an increase of approximately 5452.38
#units in sell.
# - **Gas:** The coefficient is 12381.4066,
#meaning the presence of a gas system
#is associated with an increase
#of approximately 12381.41 units in sell.
# - **Centralair:** The coefficient
#is 12632.8905, meaning the presence
#of central air conditioning is associated
#with an increase of approximately 12632.89 units in sell.
# - **Garage:** The coefficient is 4244.8290,
#meaning the presence of a garage is
#associated with an increase of
#approximately 4244.83 units in sell.
# - **Neighbour:** The coefficient is 9369.5132,
#meaning being in a certain neighborhood
#is associated with an increase of
#approximately 9369.51 units in sell.
```

```
knitr::opts_chunk$set(echo = TRUE, warning = FALSE, message = FALSE)
options(width = 80)
# - The intercept has a negative coefficient (-4038.3540),
#which might be unexpected in some contexts
#but can be interpreted as the base value
#when all other variables are zero.
# - The large coefficients for fullbath, gas, and centralair
#suggest these variables have a
#substantial economic impact, which should be verified for practical realism.
```

```
knitr::opts_chunk$set(echo = TRUE, warning = FALSE, message = FALSE)
options(width = 80)
# Overall fit metrics
r_squared <- baseline_summary$r.squared
```

```r
adjusted_r_squared <- baseline_summary$adj.r.squared
f_statistic <- baseline_summary$fstatistic
```

```r
knitr::opts_chunk$set(echo = TRUE, warning = FALSE, message = FALSE)
options(width = 80)

#R-squared: 0.6731, indicating that approximately
#67.31% of the variance in sell
#is explained by the mode
#Adjusted R-squared: 0.6664
#slightly lower than R-squared, accounting for the number of predictors.
#This is still a good fit, showing the model
#explains a substantial portion of the variance.",
#The F-statistic is 99.97 with a p-value
#< 2.2e-16, indicating that the model is statistically
#significant overall.
#This suggests that the predictors,
#taken together, significantly explain the variability in sell."
```

```r
# Print the overall fit metrics
cat("R-squared:", r_squared, "\n")
```

```
## R-squared: 0.6731236
```

```r
cat("Adjusted R-squared:", adjusted_r_squared, "\n")
```

```
## Adjusted R-squared: 0.6663902
```

```r
cat("F-statistic:", f_statistic[1], "on", f_statistic[2], "and", f_statistic[3], "DF, p-value:", pf(f_s
```

```
## F-statistic: 99.96774 on 11 and 534 DF, p-value: 6.177731e-122
```

```r
# Perform RESET test on the new model
Rtest <- resettest(baseline_model, power = 2:3, type = "regressor")
print(Rtest)
```

```
##
##  RESET test
##
## data:  baseline_model
## RESET = 0.85583, df1 = 22, df2 = 512, p-value = 0.6546
```

```r
knitr::opts_chunk$set(echo = TRUE, warning = FALSE, message = FALSE)
options(width = 80)
#The RESET test has a p-value of 0.6546,
#which is not significant. This indicates
#that there is no strong evidence against
#the model specification. In other words,
#the model does not suffer from omitted
#variable bias or incorrect functional form.
# Summary
#Statistical Significance:**
#Most variables are statistically significant, #indicating they have a meaningful relationship
#with the dependent variable.
#Economic Significance:** Variables like lotsize,
#fullbath, and centralair show substantial economic
#impacts, which align with practical expectations.",
```

```
#"- **Model Fit:** The model has a good fit with
#an R-squared of 67.31% and an
#adjusted R-squared of 66.64%. The model is also
#statistically significant
#as indicated by the F-statistic.",
# **Model Specification** The RESET test suggests
# that the model does not have specification errors.",
#"Overall, the baseline model appears to be robust
#and provides meaningful insights into the factors
#affecting property values (sell)."
```

#8

```
library(lmtest)
library(sandwich)
```

```
# Fit the baseline model
baseline_model <- lm(sell ~ ., data = hp)

# Perform the Breusch-Pagan test for heteroskedasticity
bp_test <- bptest(baseline_model)

# Print the results of the Breusch-Pagan test
print(bp_test)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  baseline_model
## BP = 61.953, df = 11, p-value = 4.014e-09
```

#The Breusch-Pagan test results show #a BP = 61.953 with a p-value of 4.014e-09. #Since the p-value is significantly #less than 0.05, we reject the #null hypothesis of homoscedasticity. #This indicates the presence of #heteroskedasticity in the model.

```
 # Refit the model using robust standard errors
  robust_model <- coeftest(baseline_model, vcov = vcovHC(baseline_model, type = "HC1"))

  # Print the summary of the model with robust standard errors
  print(robust_model)
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept) -4038.35043  3182.32917 -1.2690  0.204997
## lotsize         3.54630     0.39384  9.0044 < 2.2e-16 ***
## bedrooms     1832.00347  1038.15817  1.7647  0.078191 .
## fullbath    14335.55847  1899.66388  7.5464 1.944e-13 ***
## stories      6556.94571   869.60670  7.5401 2.030e-13 ***
## driveway     6687.77889  1657.45708  4.0350 6.259e-05 ***
## recreation   4511.28383  2144.41628  2.1037  0.035869 *
## basement     5452.38554  1769.05413  3.0821  0.002162 **
## gas         12831.40627  4242.97928  3.0242  0.002613 **
## centralair  12632.89040  1666.22480  7.5817 1.520e-13 ***
## garage       4244.82900   946.28548  4.4858 8.901e-06 ***
```

```
## neighbour    9369.51324  1870.88352  5.0081 7.482e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

#lotsize:

#Coefficient: 3.5463 #Statistically significant with p-value < 2.2e-16 #conomically significant: A one-unit increase in lot #size is associated with an increase of approximately #3.55 units in property value (sell). #Bedrooms:

#Coefficient: 1832.00347 #Statistically insignificant with p-value = 0.07819 #This indicates that the number of bedrooms does not #have a significant effect on property value at the 5% #significance level. #Fullbath:

#Coefficient: 14335.55847 #Statistically significant with p-value < 2e-16 #Economically significant: An additional full bathroom #is associated with an increase of approximately #14335.56 units in property value. #Stories:

#Coefficient: 6556.94571 #Statistically significant with p-value = 2.03e-13 #Economically significant: An additional story in the #building is associated with an increase of #approximately 6556.95 units in property value. #Driveway:

#Coefficient: 6687.77889 #Statistically significant with p-value = 6.259e-05 #Economically significant: Having a driveway is #associated with an increase of approximately 6687.78 #units in property value. #Recreation:

#Coefficient: 4511.28834 #Statistically significant with p-value = 0.035869 #Economically significant: The presence of recreation #facilities is associated with an increase of #approximately 4511.29 units in property value. #Basement:

#Coefficient: 5452.38346 #Statistically significant with p-value = 0.002001 #Economically significant: Having a basement is #associated with an increase of approximately 5452.38 #units in property value. #Gas:

#Coefficient: 12381.40627 #Statistically significant with p-value = 0.002613 #Economically significant: The presence of a gas system #is associated with an increase of approximately #12381.41 units in property value. #Centralair:

#Coefficient: 12632.89004 #Statistically significant with p-value = 1.52e-13 #Economically significant: The presence of central air #conditioning is associated with an increase of #approximately 12632.89 units in property value. #Garage:

#Coefficient: 4244.82900 #Statistically significant with p-value = 8.901e-06 #Economically significant: The presence of a garage is #associated with an increase of approximately 4244.83 #units in property value. #Neighbour:

#Coefficient: 9369.51324 #Statistically significant with p-value = 7.482e-07 #Economically significant: Being in a certain #neighborhood is associated with an increase of #approximately 9369.51 units in property value. #Overall Model Fit #R-squared: 0.6731 #Indicates that approximately 67.31% of the variance in #sell is explained by the model. #Adjusted R-squared: 0.6664 #Slightly lower than R-squared, accounting for the #number #of predictors. Still a good fit, showing the #model explains a substantial portion of the variance. #F-statistic: 99.97 with a p-value < 2.2e-16 #Indicates that the model is statistically significant #overall, suggesting that the predictors, taken #together, significantly explain the variability in #sell. #Summary #Statistical Significance: Most variables are #statistically significant, indicating they have a #meaningful relationship with the dependent variable. #Economic Significance: Variables like lotsize, #fullbath, and centralair show substantial economic #impacts, which align with practical expectations. #Model Fit: The model has a good fit with an R-squared #of 67.31% and an adjusted R-squared of 66.64%. The #model is also statistically significant as indicated by #the F-statistic. #Model Specification: The RESET test suggests that the #model does not have specification errors. #Heteroskedasticity: The Breusch-Pagan test indicates #the presence of heteroskedasticity, which has been #corrected using robust standard errors. The corrected #model provides more reliable standard errors and #significance tests.