Midterm Part 2

Langchen Liu

2024-02-20

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(pubtheme)
```

```
## Loading required package: plotly
##
## Attaching package: 'plotly'
##
## The following object is masked from 'package:ggplot2':
##
##     last_plot
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following object is masked from 'package:graphics':
##
##     layout
##
## Loading required package: scales
##
## Attaching package: 'scales'
##
## The following object is masked from 'package:purrr':
##
##     discard
##
## The following object is masked from 'package:readr':
##
##     col_factor
##
```

```
## Loading required package: ggrepel
```

```r
d = read.csv('data/branford.csv')
d = d %>%
  select(pid, value, land, living, beds, baths,
         good, style, grade, ac, miles_to_coastline)
head(d,2)
```

```
##   pid  value land living beds baths good       style grade      ac
## 1   1 247400 0.51   2194    3     2   87 Split-Level   B - Central
## 2 100 177200 1.30   1200    3     1   78   Old Style     C    None
##   miles_to_coastline
## 1          0.6500547
## 2          0.4848638
```

There is one row per property, each with one single family home, and the columns have the following meanings:

- `value`: the assessed value of the proper

```r
## build several linear models to predict the log(value) using the other variables, compare the models
## and choose the best one

## first refacroring the grade variable
d$grade = factor(d$grade, levels = c('A ++', 'A +', 'A', 'B +', 'B', 'B -', 'C +', 'C', 'C -', 'D +', 'I
# let's see what styles the dataset have
d$style %>% table()
```

```
## .
##     Bungalow     Cape Cod     Colonial      Cottage       Custom  Mobile Home
##          116          957         1485          120          264          136
##    Old Style Raised Ranch  Split-Level        Tudor
##          921          460          333            1
```

```r
# exculde the mobile homes
d = d %>% filter(style != 'Mobile Home')
d$style = factor(d$style)
```

```r
# do the same for ac
d$ac = factor(d$ac)
```

We start building linear models to predict the log of the value of the property using the other variables. We will compare the models and choose the best one.

```r
# first check all numerical variables
lm1 = lm(log(value) ~ land, data = d)
lm2 = lm(log(value) ~ land + living, data = d)
lm3 = lm(log(value) ~ land + living + beds + baths, data = d)
lm4 = lm(log(value) ~ land + living + beds + baths + good, data = d)
lm5 = lm(log(value) ~ land + living + beds + baths + good + miles_to_coastline, data = d)
summary(lm1)
```

```
##
## Call:
## lm(formula = log(value) ~ land, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0070 -0.3195 -0.0974  0.1960  2.1281
```

```
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 12.455990   0.008439 1475.969  < 2e-16 ***
## land         0.049424   0.006616    7.471 9.46e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.509 on 4655 degrees of freedom
## Multiple R-squared:  0.01185,    Adjusted R-squared:  0.01164
## F-statistic: 55.81 on 1 and 4655 DF,  p-value: 9.461e-14
```

```r
summary(lm2)
```

```
##
## Call:
## lm(formula = log(value) ~ land + living, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.87777 -0.19885 -0.07153  0.09254  2.41288
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.167e+01  1.183e-02 986.958  < 2e-16 ***
## land        -1.938e-02  4.526e-03  -4.282 1.89e-05 ***
## living       4.000e-04  5.293e-06  75.571  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3411 on 4654 degrees of freedom
## Multiple R-squared:  0.5563, Adjusted R-squared:  0.5561
## F-statistic:  2918 on 2 and 4654 DF,  p-value: < 2.2e-16
```

```r
summary(lm3)
```

```
##
## Call:
## lm(formula = log(value) ~ land + living + beds + baths, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.58300 -0.19878 -0.07029  0.09464  2.42307
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.160e+01  1.801e-02 644.418  < 2e-16 ***
## land        -1.779e-02  4.493e-03  -3.960 7.60e-05 ***
## living       3.513e-04  8.184e-06  42.927  < 2e-16 ***
## beds         1.037e-02  6.486e-03   1.599     0.11
## baths        6.818e-02  8.894e-03   7.666 2.15e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3383 on 4650 degrees of freedom
```

```
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.5638, Adjusted R-squared:  0.5634
## F-statistic:  1502 on 4 and 4650 DF,  p-value: < 2.2e-16
```

```r
summary(lm4)
```

```
##
## Call:
## lm(formula = log(value) ~ land + living + beds + baths + good,
##     data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.31827 -0.19696 -0.07518  0.08633  2.51191
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.118e+01  5.220e-02 214.122  < 2e-16 ***
## land        -1.502e-02  4.469e-03  -3.361 0.000783 ***
## living       3.374e-04  8.274e-06  40.782  < 2e-16 ***
## beds         1.505e-02  6.457e-03   2.331 0.019779 *
## baths        5.811e-02  8.899e-03   6.530 7.26e-11 ***
## good         5.773e-03  6.633e-04   8.704  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3356 on 4649 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.5707, Adjusted R-squared:  0.5703
## F-statistic:  1236 on 5 and 4649 DF,  p-value: < 2.2e-16
```

```r
summary(lm5)
```

```
##
## Call:
## lm(formula = log(value) ~ land + living + beds + baths + good +
##     miles_to_coastline, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.32584 -0.18814 -0.05154  0.12237  2.44960
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.111e+01  4.877e-02 227.838  < 2e-16 ***
## land                1.121e-02  4.288e-03   2.614  0.00898 **
## living              3.313e-04  7.725e-06  42.882  < 2e-16 ***
## beds                3.557e-02  6.076e-03   5.854 5.13e-09 ***
## baths               4.393e-02  8.322e-03   5.278 1.36e-07 ***
## good                7.455e-03  6.223e-04  11.978  < 2e-16 ***
## miles_to_coastline -1.773e-01  6.748e-03 -26.267  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3132 on 4648 degrees of freedom
```

```
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.6262, Adjusted R-squared:  0.6257
## F-statistic:  1298 on 6 and 4648 DF,  p-value: < 2.2e-16
```

From the very first observation, `beds` and `land` seems to be less important. Let's discard land and create a variable bed+bath

```
d$bed_bath = d$beds + d$baths
lm6 = lm(log(value) ~ living + good + bed_bath + miles_to_coastline, data = d)
summary(lm6)
```

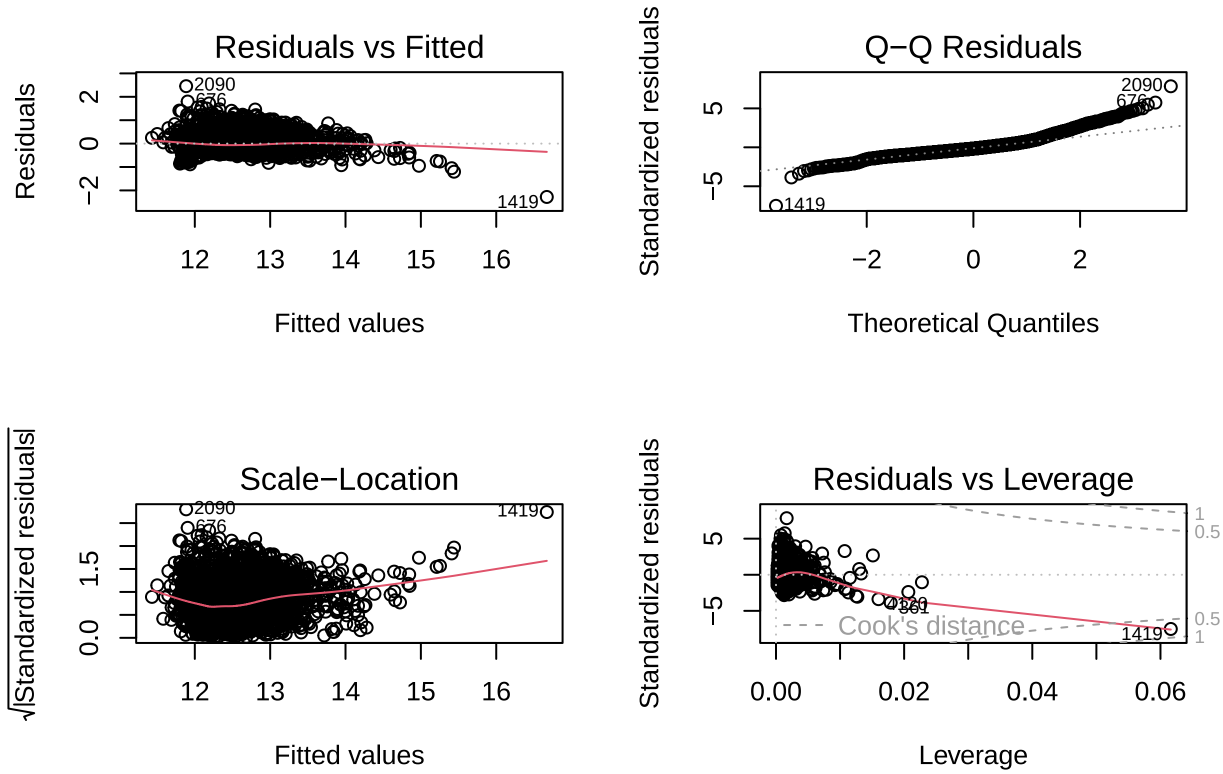
```
##
## Call:
## lm(formula = log(value) ~ living + good + bed_bath + miles_to_coastline,
##     data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.28007 -0.18816 -0.05198  0.12278  2.45418
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        11.1162398  0.0474939 234.056   <2e-16 ***
## living              0.0003361  0.0000074  45.421   <2e-16 ***
## good                0.0073776  0.0006126  12.043   <2e-16 ***
## bed_bath            0.0381304  0.0044927   8.487   <2e-16 ***
## miles_to_coastline -0.1737624  0.0065190 -26.655   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3134 on 4650 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.6256, Adjusted R-squared:  0.6253
## F-statistic:  1943 on 4 and 4650 DF,  p-value: < 2.2e-16
```

Note that lm6 has the highest R-squared value, so we will use it as our model.

As a conclusion, I choose `log(value)` to be the dependent variable, and `living`, `good`, `bed_bath`, and `miles_to_coastline` to be the independent variables. The reason I choose them is because they are all numerical values and are all significant in the linear model.

Let's check on the linear model assumptions

```
par(mfrow=c(2,2))
plot(lm6)
```



From the plot, the Q-Q residuals plot is not perfectly normal, but it is close enough. So we can say that the normality assumption is satisfied. The residuals vs fitted plot looks not good, the relationship may not be linear. So the linearity assumption might not be perfectly satisfied. But we are okay with this. The indepedence assumption is fine. The constant variance assumption fails as I cannot tell the residuals have constant variance.