ADA2: Class 03, Ch 02 Introduction to Multiple Linear Regression

[Advanced Data Analysis 2](https://StatAcumen.com/teach/ada12, Stat 428/528, Spring 2023, Prof. Erik Erhardt, UNM

AUTHOR
Ryan Riner

PUBLISHED
January 24, 2023

Auction selling price of antique grandfather clocks

The data include the selling price in pounds sterling at auction of 32 antique grandfather clocks, the age of the clock in years, and the number of people who made a bid. In the sections below, describe the relationship between variables and develop a model for predicting selling `Price` given `Age` and `Bidders`.

```r
library(erikmisc)
```

```
── Attaching packages ─────────────────────────────────────── erikmisc 0.1.18 ──

✔ tibble 3.1.8      ✔ dplyr  1.0.10

── Conflicts ──────────────────────────────────────────── erikmisc_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()

erikmisc, solving common complex data analysis workflows
  by Dr. Erik Barry Erhardt <erik@StatAcumen.com>
```

```r
library(tidyverse)
```

```
── Attaching packages
───────────────────────────────────────
tidyverse 1.3.2 ──

✔ ggplot2 3.4.0     ✔ purrr   1.0.1
✔ tidyr   1.2.1     ✔ stringr 1.5.0
✔ readr   2.1.3     ✔ forcats 0.5.2
── Conflicts ─────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
```

```r
dat_auction <- read_csv("ADA2_CL_03_auction.csv")
```

```
Rows: 32 Columns: 3
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
dbl (3): Age, Bidders, Price

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
str(dat_auction)
```

```
spc_tbl_ [32 × 3] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ Age    : num [1:32] 127 115 127 150 156 182 156 132 137 113 ...
 $ Bidders: num [1:32] 13 12 7 9 6 11 12 10 9 9 ...
 $ Price  : num [1:32] 1235 1080 845 1522 1047 ...
 - attr(*, "spec")=
  .. cols(
  ..   Age = col_double(),
  ..   Bidders = col_double(),
  ..   Price = col_double()
  .. )
 - attr(*, "problems")=<externalptr>
```

```
summary(dat_auction)
```

```
      Age           Bidders           Price     
 Min.   :108.0   Min.   : 5.000   Min.   : 729  
 1st Qu.:117.0   1st Qu.: 7.000   1st Qu.:1053  
 Median :140.0   Median : 9.000   Median :1258  
 Mean   :144.9   Mean   : 9.531   Mean   :1327  
 3rd Qu.:168.5   3rd Qu.:11.250   3rd Qu.:1561  
 Max.   :194.0   Max.   :15.000   Max.   :2131  
```

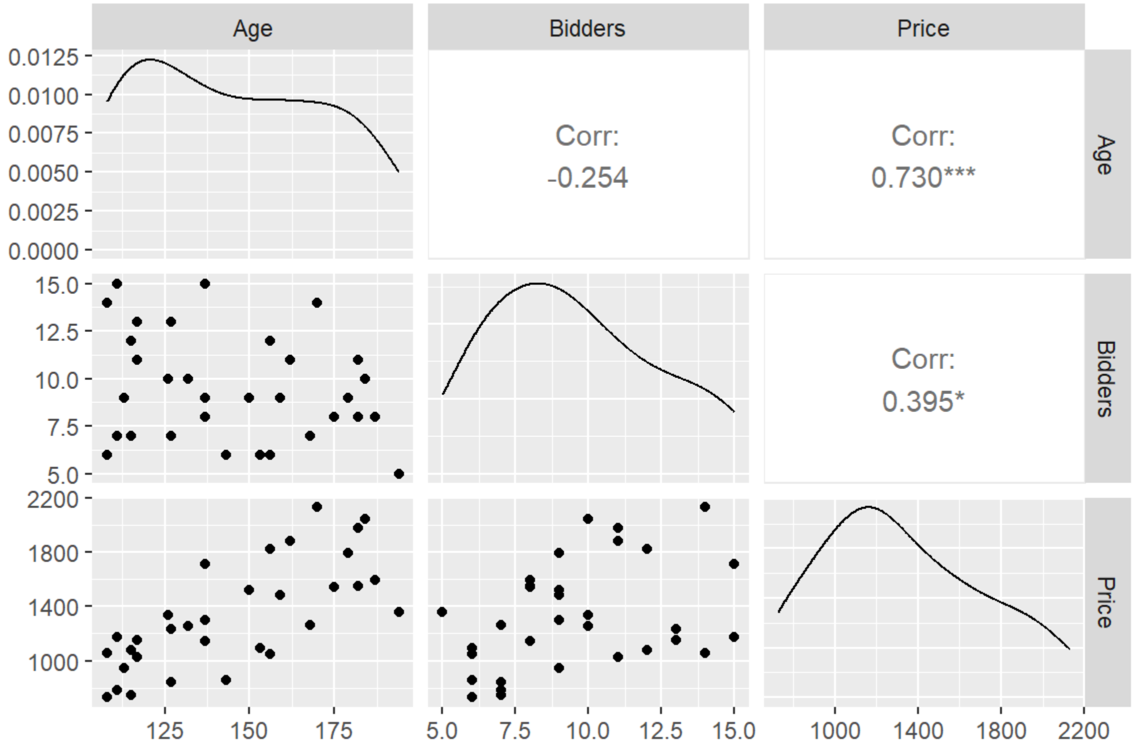
(1 p) Scatterplot matrix

In a scatterplot matrix below interpret the relationship between each pair of variables. If a transformation is suggested by the plot (that is, because there is a curved relationship), also plot the data on the transformed scale and perform the following analysis on the transformed scale. Otherwise indicate that no transformation is necessary.

```
library(ggplot2)
library(GGally)
```

```
Registered S3 method overwritten by 'GGally':
  method from
  +.gg   ggplot2
```

```
p <- ggpairs(dat_auction)
print(p)
```



Solution

There is a positive relationship between age and price with quite a strong correlation of 0.73.

There is also a positive relationship between bidders and price with a fairly strong correlation of 0.395.

There is no discernible relationship between bidders and age, though there is a weakly negative correlation of -0.254.

Because of the generally linear nature of these relationships, no transformation is necessary.

(1 p) Correlation matrix

Below is the correlation matrix and tests for the hypothesis that each correlation is equal to zero. Interpret the hypothesis tests and relate this to the plot that you produced above.

```
# correlation matrix and associated p-values testing "H0: rho == 0"
#library(Hmisc)
Hmisc::rcorr(as.matrix(dat_auction))
```

```
          Age Bidders Price
Age      1.00   -0.25  0.73
Bidders -0.25    1.00  0.39
Price    0.73    0.39  1.00

n= 32


P
        Age    Bidders Price 
Age            0.1611  0.0000
Bidders 0.1611         0.0254
Price   0.0000 0.0254        
```

Solution

Using Bonferoni's adjustment for three variables (0.05/3) we will use a significance level of 0.017 for these correlations.

The p-value for the correlation between `Age` and `Bidders` is 0.1611 > 0.017, so their negative correlation observed above is insignificant.

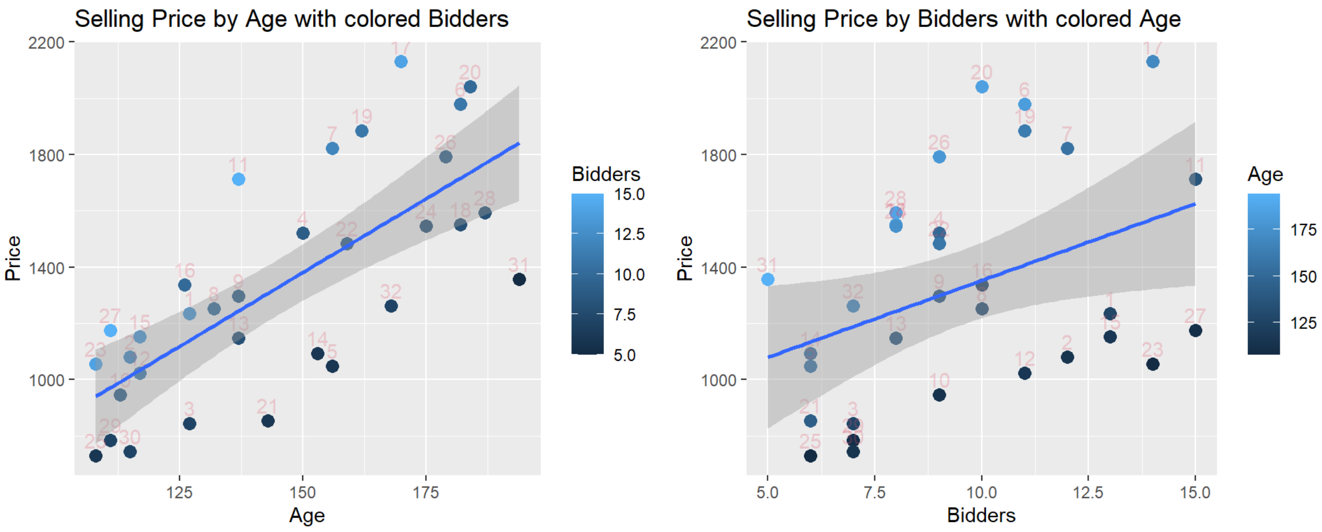
The p-value for the correlation between `Age` and `Price` is effectively 0 < 0.017, so their positive correlation observed above is significant.

The p-value for the correlation between `Bidders` and `Price` is 0.0254 > 0.017, so their positive correlation observed above is insignificant.

(1 p) Plot interpretation

Below are two plots. The first has $y = \mathit{Price}$, $x = \mathit{Age}$, and colour = Bidders, and the second has $y = \mathit{Price}$, $x = \mathit{Bidders}$, and colour = Age. Interpret the relationships between all three variables, simultaneously. For example, say how Price relates to Age, then also how Price relates to Bidders conditional on Age being a specific value.

```

Attaching package: 'gridExtra'

The following object is masked from 'package:dplyr':
    combine

`geom_smooth()` using formula = 'y ~ x'

Warning: The following aesthetics were dropped during statistical transformation:
label
ℹ This can happen when ggplot fails to infer the correct grouping structure in
  the data.
ℹ Did you forget to specify a `group` aesthetic or to convert a numerical
  variable into a factor?

`geom_smooth()` using formula = 'y ~ x'

Warning: The following aesthetics were dropped during statistical transformation:
label
ℹ This can happen when ggplot fails to infer the correct grouping structure in
  the data.
ℹ Did you forget to specify a `group` aesthetic or to convert a numerical
  variable into a factor?
```



Solution

`Price` has a positive relationship to `Age`, although there are a lot of samples outside the limits of the confidence band. For each individual `Age`, an increase in `Bidders` corresponds to an increase in `Price`.

`Price` also has a positive relationship to `Bidders`, though a much weaker one. For each individual number of `Bidders`, an increase in `Age` corresponds to an increase in `Price`.

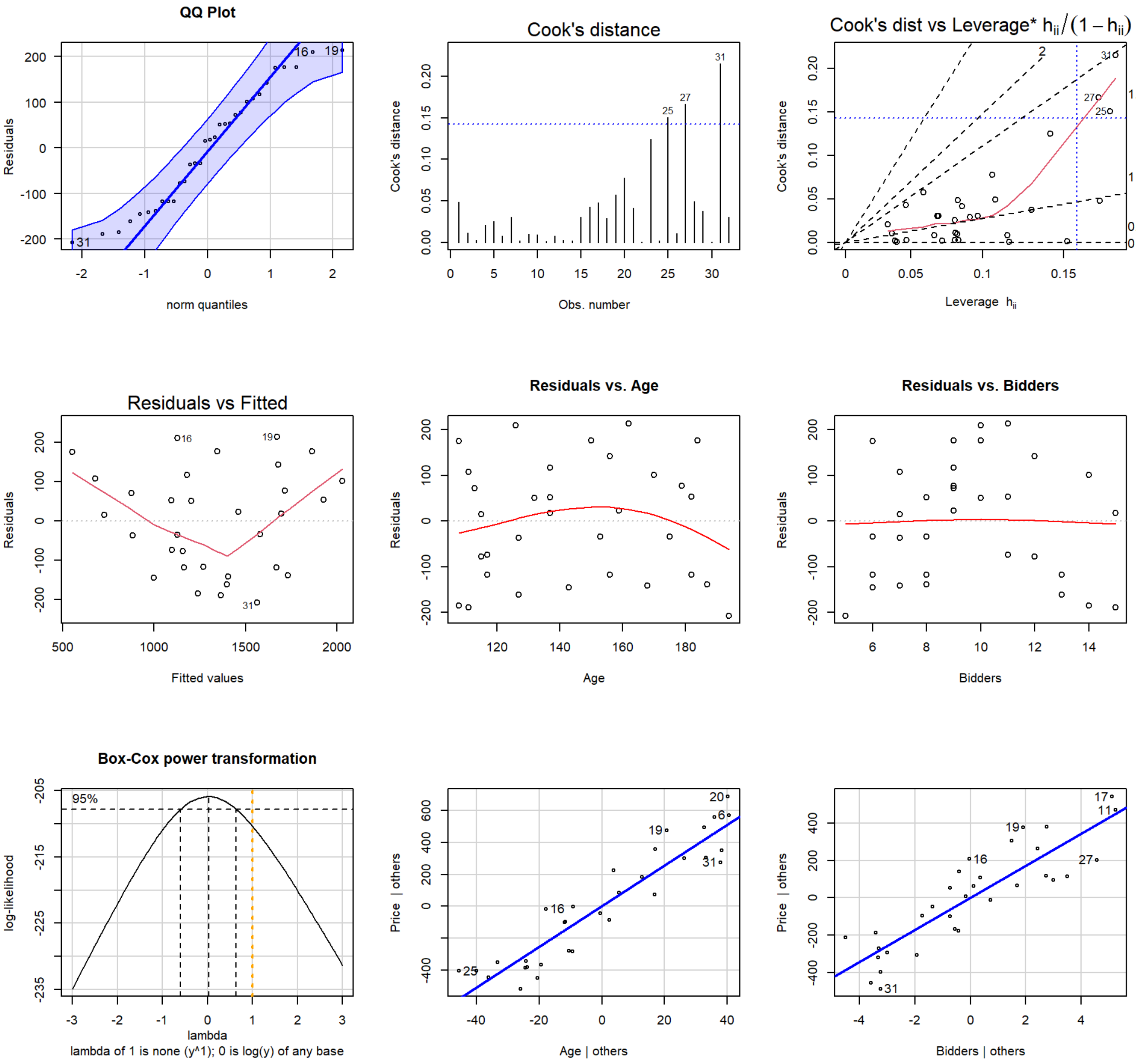
(2 p) Multiple regression assumptions (assessing model fit)

Below the multiple regression is fit. Start by assessing the model assumptions by interpreting what you learn from the first six plots (save the added variable plots for the next question). If assumptions are not met, attempt to address by transforming a variable and restart at the beginning using the new transformed variable.

```r
# fit the simple linear regression model
lm_p_a_b <- lm(Price ~ Age + Bidders, data = dat_auction)
```

Plot diagnostics.

```r
# plot diagnostics
e_plot_lm_diagostics(lm_p_a_b, sw_plot_set = "simpleAV")
```



Solution

From the diagnostic plots above,

1. QQ Plot: Although there is some minor deviation, there are no true outliers and the residuals do follow a normal distribution. Assumption met.
2. Cook's Distance: There are no extreme points compared to the bulk of the data, so none are necessarily overly influential. Assumption met.
3. Cook's Distance vs Leverage: There are no extreme points in Cook's distance with excessive leverage shown. Assumption met.
4. Residuals vs Fitted: There is little to no structure to the plot of residuals, meaning the value of residuals is fairly evenly distributed. Assumption met.
5. Residuals vs Age: Similarly, there is no obvious structure to the distribution of residuals. Assumption met.
6. Residuals vs Bidders: Again, we can see an even distribution of residuals across this plot. Assumption met.

(1 p) Added variable plots

Use partial regression residual plots (added variable plots) to check for the need for transformations. If linearity is not supported, address and restart at the beginning.

Solution

Given the linear and correlative qualities of the distribution of residuals in `Price` vs `Age` after adjusting for `Bidders`, and likewise for `Price` vs `Bidders` after adjusting for `Age`, there is no need for transformation and we may continue with the current linear plots.

(1 p) Multiple regression hypothesis tests

State the hypothesis test and conclusion for each regression coefficient.

```r
# fit the simple linear regression model
lm_p_a_b <- lm(Price ~ Age + Bidders, data = dat_auction)
# use summary() to get t-tests of parameters (slope, intercept)
summary(lm_p_a_b)
```

```
Call:
lm(formula = Price ~ Age + Bidders, data = dat_auction)

Residuals:
   Min     1Q Median     3Q    Max 
-207.2 -117.8   16.5  102.7  213.5 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1336.7221   173.3561  -7.711 1.67e-08 ***
Age            12.7362     0.9024  14.114 1.60e-14 ***
Bidders        85.8151     8.7058   9.857 9.14e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 133.1 on 29 degrees of freedom
Multiple R-squared:  0.8927,	Adjusted R-squared:  0.8853 
F-statistic: 120.7 on 2 and 29 DF,  p-value: 8.769e-15
```

Solution

Age Coefficient: $H_0 : coefficient = 0$ $H_A : coefficient \neq 0$ $p - value = 1.60e^{-}14 < 0.05$ With a p-value less than $\alpha = 0.05$ we must reject the null hypothesis and conclude that the `Age` coefficient is not equal to 0.

Bidders Coefficient: $H_0: $ coefficient $= 0$ $H_A: $ coefficient $\neq 0$ $p - value = 9.14e^{-}11 < 0.05$ With a p-value less than $\alpha = 0.05$ we must reject the null hypothesis and conclude that the `Age` coefficient is not equal to 0.

Intercept: We cannot in this case interpret the negative intercept.

(1 p) Multiple regression interpret coefficients

Interpret the coefficients of the multiple regression model.

Solution

The coefficient of `Age` is estimated to be 12.7. Thus, for every unit increase in `Age` and holding the value of `Bidders` constant, we would expect an increase of 12.7 in `Price`.

The coefficient of `Bidders` is estimated to be 85.8. Thus, for every unit increase in `Bidders` and holding the value of `Age` constant, we would expect an increase of 85.8 in `Price`.

(1 p) Multiple regression R^2

Interpret the Multiple R-squared value.

Solution

The regression model explains 89.27% of the variability of `Price`.

(1 p) Summary

Summarize your findings in one sentence.

Solution

In examining the data (which met model assumptions and did not need any transformation), we found significant positive linear relationships between `Price` and both `Age` and `Bidders`, which explains about 90% of the variability of the response variable `Price`.

```
## Aside: I generally recommend against 3D plots for a variety of reasons.
## However, here's a 3D version of the plot so you can visualize the surface fit in 3
## I will point out a feature in this plot that we wouldn't see in other plots
## and it would typically only be detected by careful consideration
## of a "more complicated" second-order model that includes curvature.

# library(rgl)
# library(car)
# scatter3d(Price ~ Age + Bidders, data = dat_auction)
```