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

Tim Farkas

January 28, 2022

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`.

```
library(erikmisc)
library(tidyverse)

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

```
spec_tbl_df [32 x 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)
```

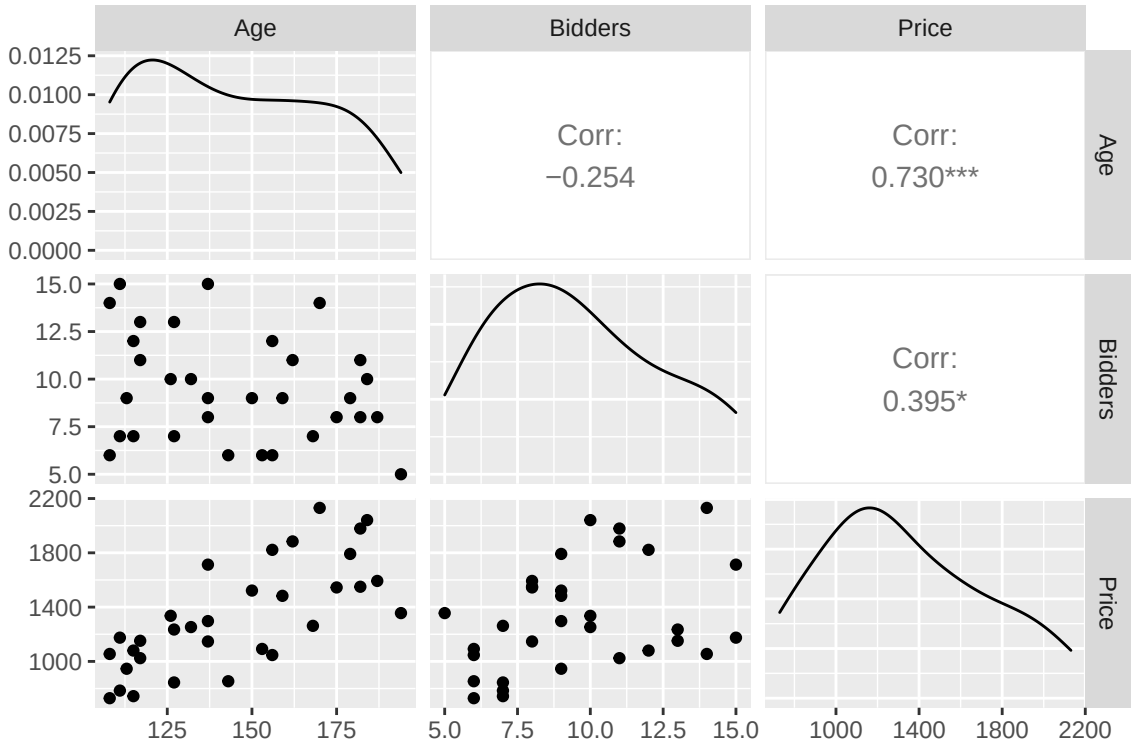
```
      Age           Bidders          Price     
 Min.   :108.0   Min.   : 5.000   Min.   : 729  
 1st Qu.:117.0   1st Qu.: 7.000   1st Qu.:1053  
 Median :140.0   Median : 9.000   Median :1258  
 Mean   :144.9   Mean   : 9.531   Mean   :1327  
 3rd Qu.:168.5   3rd Qu.:11.250   3rd Qu.:1561  
 Max.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

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



Solution

I don't see much of a need for transformation here ...

(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))
```

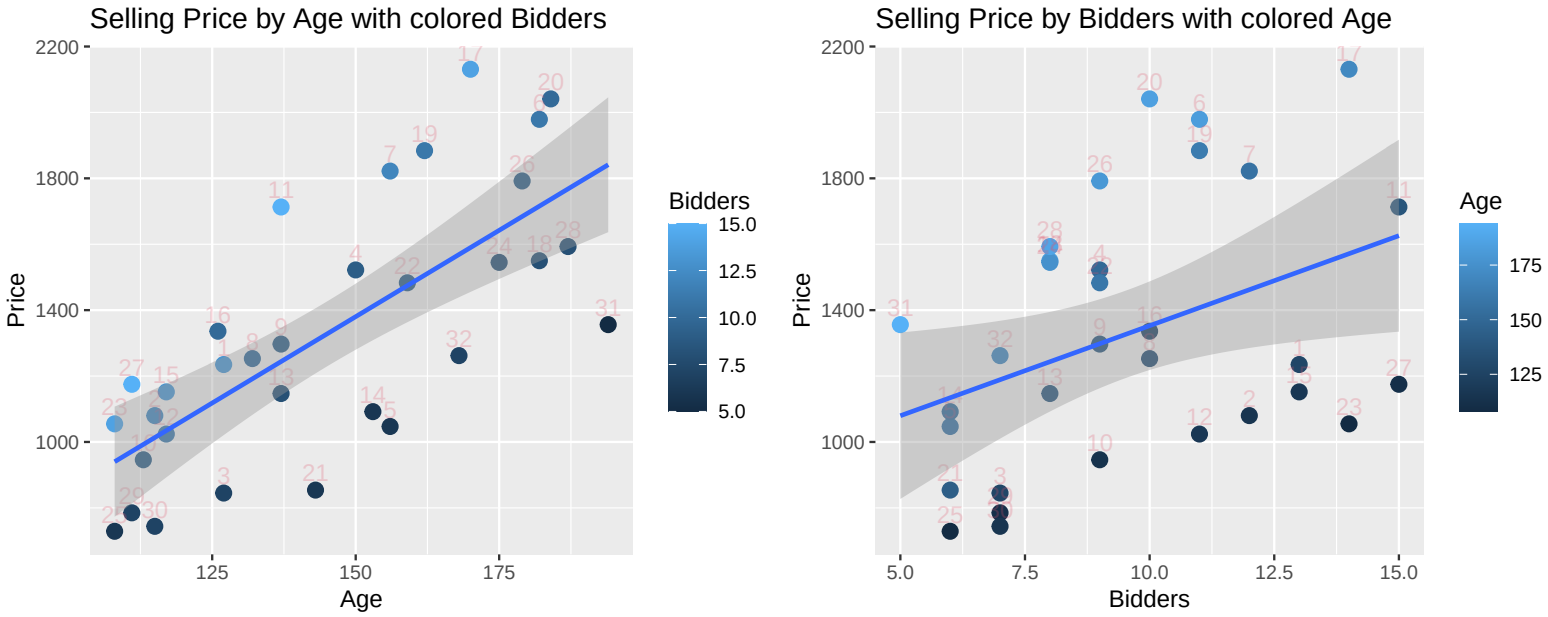
```
          Age Bidders Price
Age      1.00   -0.25  0.73
Bidders -0.25    1.00  0.39
Price    0.73    0.39  1.00

n= 32


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

Solution

The three hypothesis tests are for pairwise correlations. The results show significant correlations between Price and Age, and Price and Bidders, but no between Age and Bidders.

(1 p) Plot interpretation

Below are two plots. The first has y = Price, x = Age, and colour = Bidders, and the second has y = Price, x = 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.



Solution

These plots show pairwise relationships between price and age, and price and bidders, complementing the correlation matrix. It's hard to interpret anything beyond this – the coloration is just hard to parse.

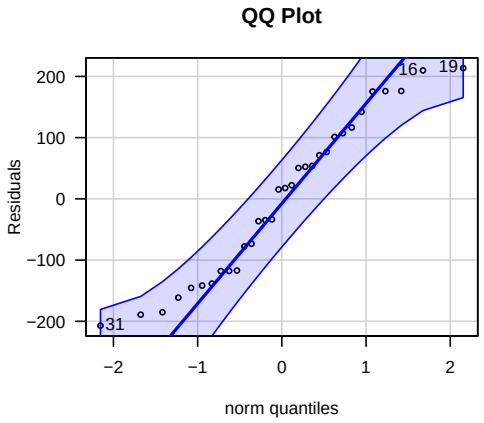
(2 p) Multiple regression assumptions (assessing model fit)

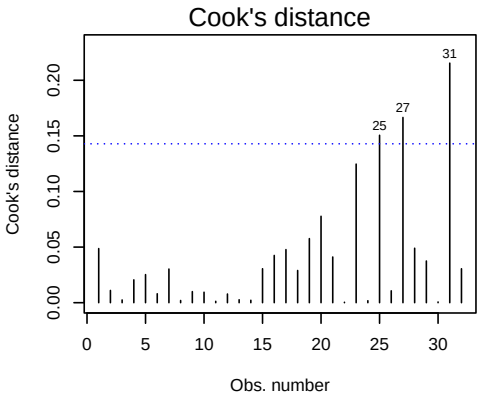
Below the multiple regression is fit. Start by assessing the model assumptions by interpretting what you learn from the first six plots (save the added variable plots for the next question). If assumptions are not met, attempt to address by transforming a variable and restart at the beginning using the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

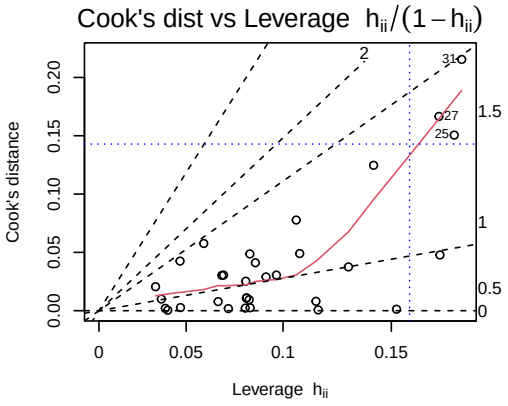
```
# fit the simple linear regression model
lm_p_a_b <- lm(Price ~ Age + Bidders, data = dat_auction)
```

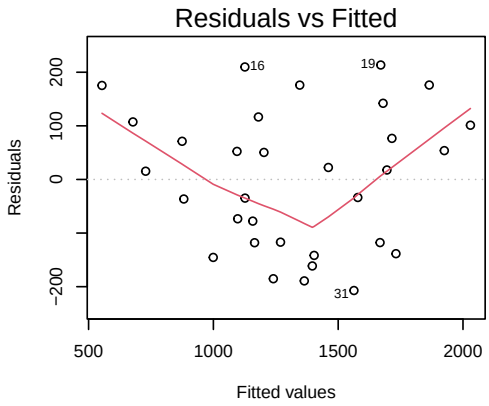
Plot diagnostics.

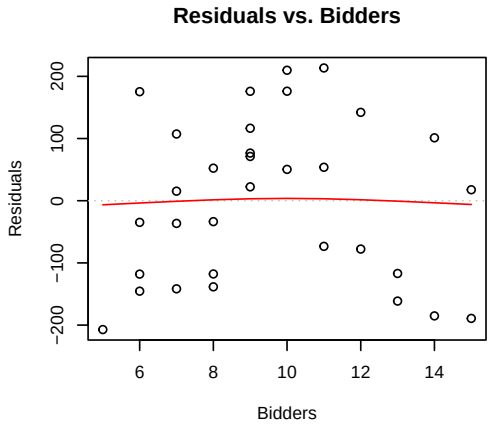
```
# plot diagnostics
e_plot_lm_diagostics(lm_p_a_b, sw_plot_set = "simpleAV")
```

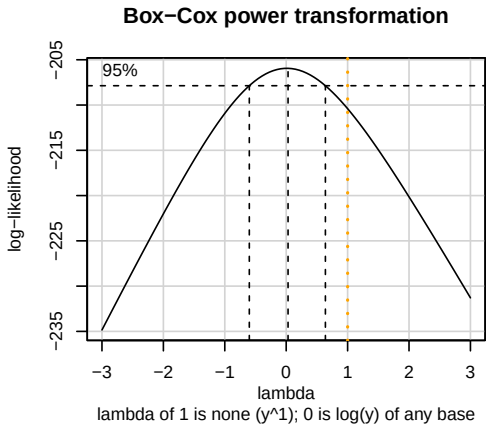


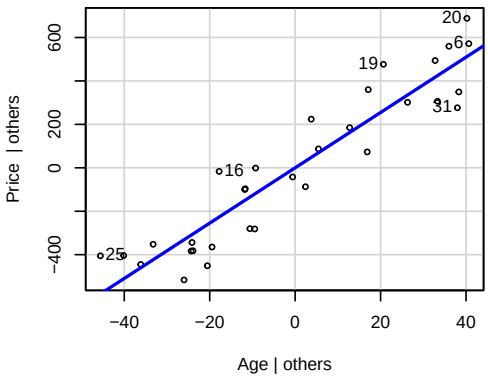


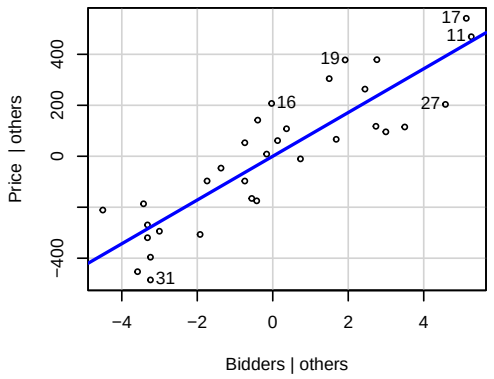












Solution

From the diagnostic plots above,

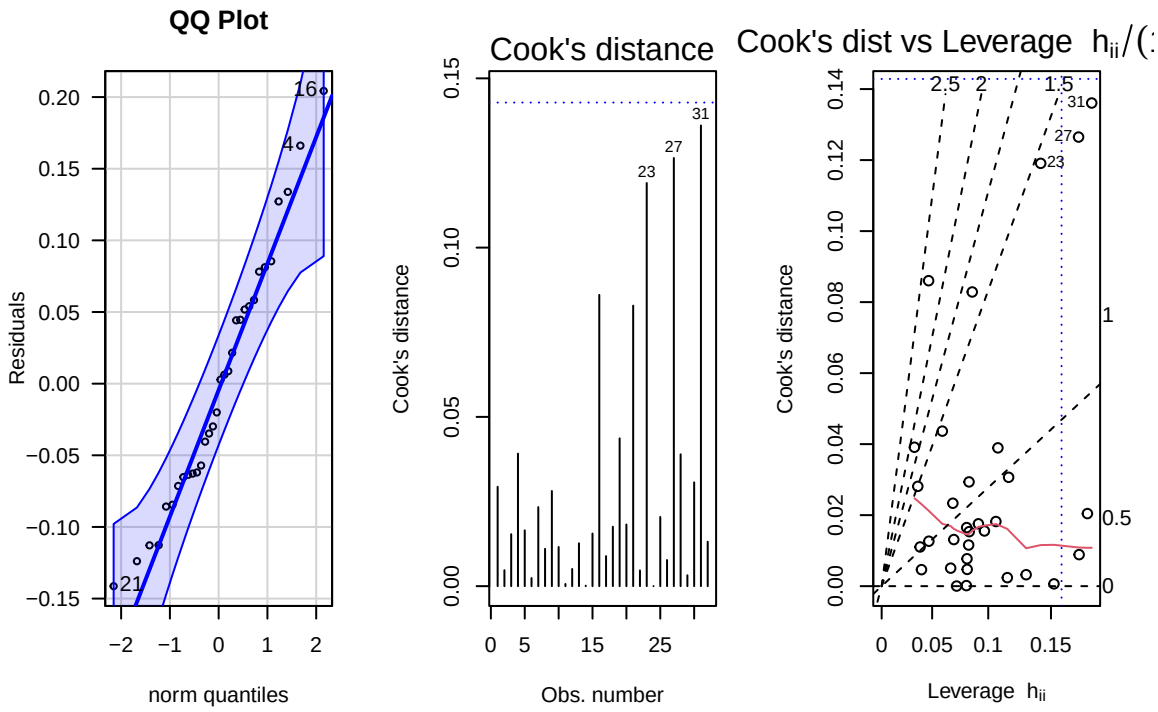
(1) The residuals look pretty darn normal!
(2) Cook's distance suggests there are a few potential outliers influencing the fit of this model!
(3) Cook's distance vs. leverage shows that the points with high leverage are those exerting undue influence on the model fit. Except for one point with high leverage that does not influence the fit much.
(4) Residuals vs. fitted shows a pretty cloud-like pattern, though middle fitted values seem to show low residuals on average. Not sure this pattern looks too strong.
(5) Residuals vs. Age shows no particular pattern.
(6) Residuals vs. Bidders shows no particular pattern.

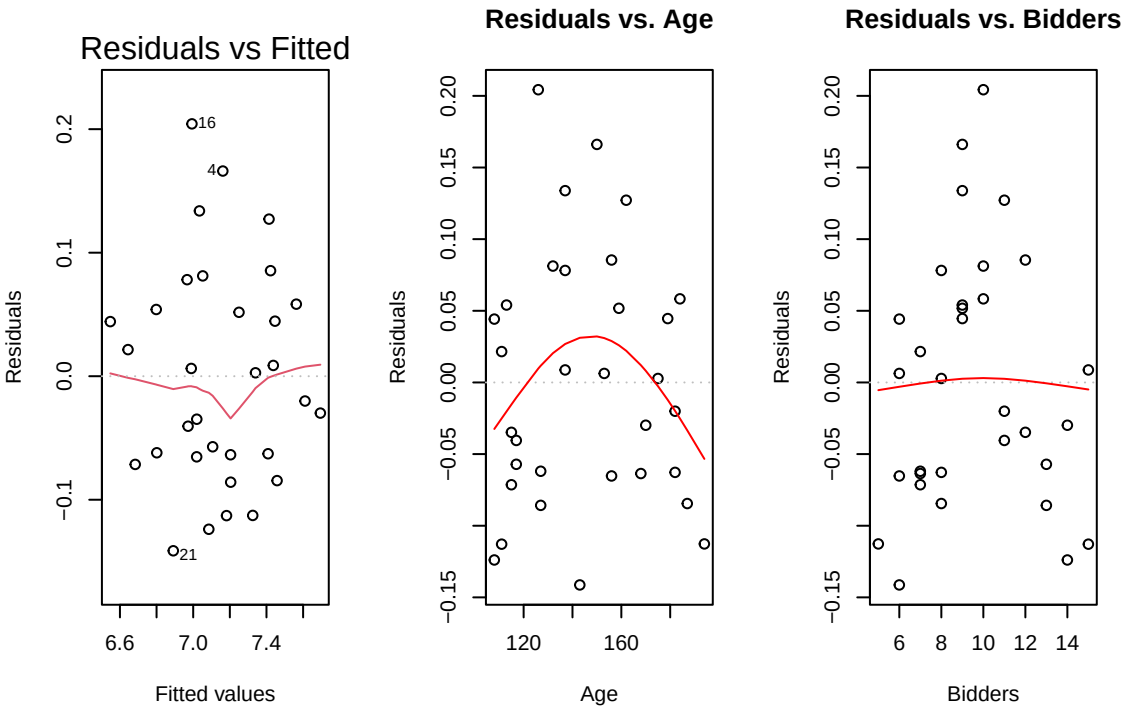
Based on the Box-Cox profile, I'm going to try a log transformation:

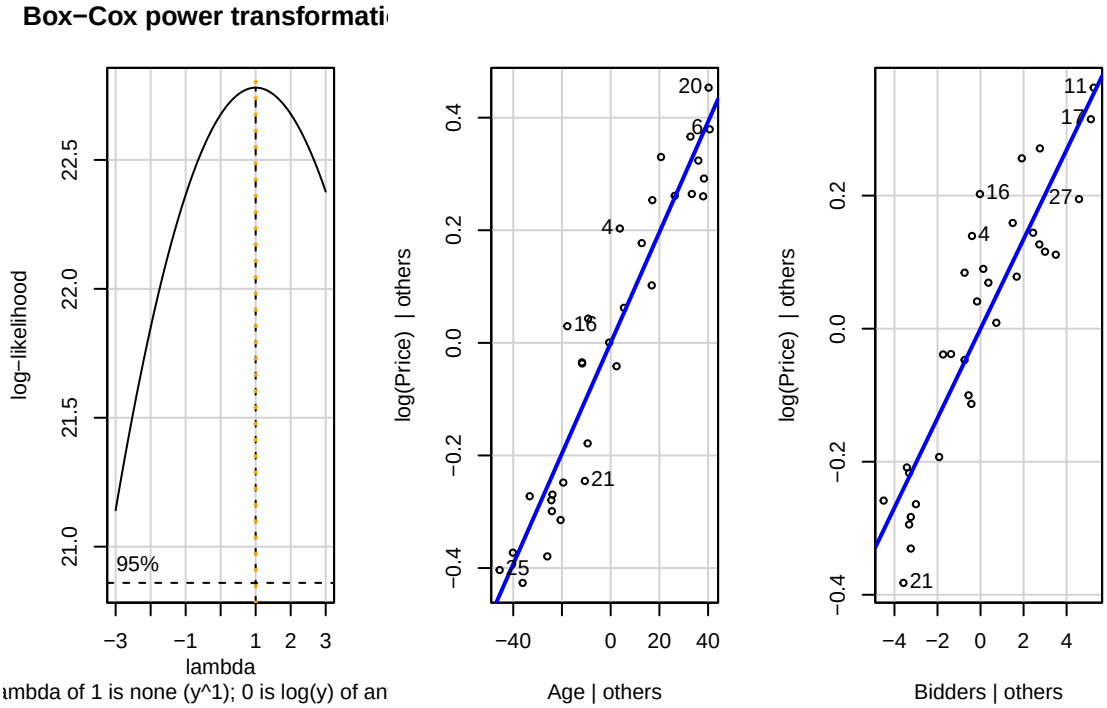
```
lm_lnp_a_b <- lm(log(Price) ~ Age + Bidders, data = dat_auction)
summary(lm_lnp_a_b)
```

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

Residuals:
      Min        1Q    Median        3Q       Max
-0.141325 -0.064007 -0.008658  0.055097  0.204297

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.0833094  0.1186522   42.84  < 2e-16 ***
Age         0.0098227  0.0006176   15.90 7.36e-16 ***
Bidders     0.0672172  0.0059586   11.28 4.01e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09112 on 29 degrees of freedom
Multiple R-squared:  0.9142,    Adjusted R-squared:  0.9083
F-statistic: 154.5 on 2 and 29 DF,  p-value: 3.426e-16
```

```
e_plot_lm_diagostics(lm_lnp_a_b, sw_plot_set = "simpleAV")
```





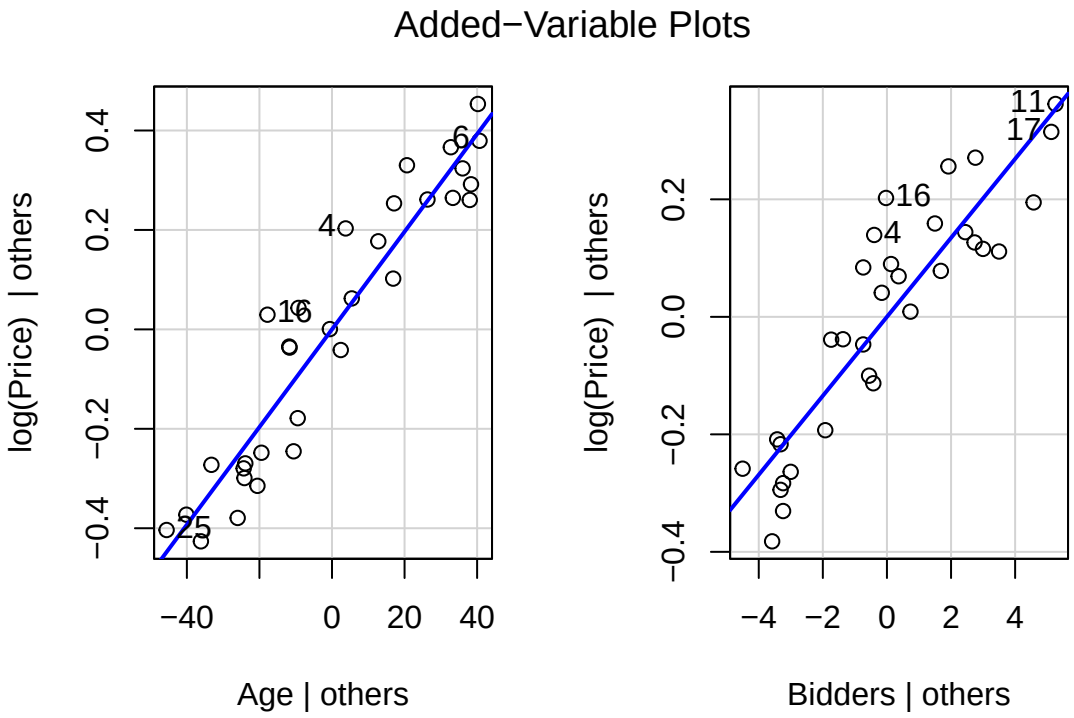


Oh yeah, that looks better. The outliers no longer seem extreme, judging by cooks distance, and the data look even more normal!

(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.

```
car::avPlots(lm_lnp_a_b)
```



No transformation other than log needed. Those are tight, linear relationships!

Solution

(1 p) Multiple regression hypothesis tests

State the hypothesis test and conclusion for each regression coefficient.

```
# 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

Intercept: H_0: The mean of log(Price) is 0 when Age and Bidders is 0. H_α: Otherwise.

Age: H_0: The effect of Age on ln(Price) is 0, holding Bidders constant. H_α: Otherwise.

Bidders: H_0: The effect of Bidderes on ln(Price) is 0, holding Age constant. H_α: Otherwise.

(1 p) Multiple regression interpret coefficients

Interpret the coefficients of the multiple regression model.

Solution

Intercept: The value of ln(Price) for a new clock with no bidders is 5.08. Age: An increase in clock age of one year increase the ln(Price) of the clock by 0.0098. Bidders: An increase in Bidders by one increases the ln(Price) of the clock by 0.067.

(1 p) Multiple regression R^2

Interpret the Multiple R-squared value.

Solution

The R^2 is 0.91, indicating that 91% of variation in ln(Price) is explained by this model.

(1 p) Summary

Summarize your findings in one sentence.

Solution

Both the age of a clock and the number of bidders on the clock have a strong, positive relationship with the price of clock, together accounting for most of the variation in clock 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 3D.
## 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)
```