Assignment 2 - Omitted and Redundant Variables Bias

Aliaksandr Panko, Nadzeya Yakimchyk, Nikolai Agafonov, Ilia Kurennoi

May 29, 2018

Contents

1 Objectives

1. Compute the omitted variable bias.
2. Discuss the trade-off between the omitted variable bias and multi-collinearity.

2 Choose the model

Let's start with the regression where all factors are statistically significant

2.1 R-code

```
train <- read.csv2(file.choose(), header=TRUE, sep=",", na.strings="")

train["NbrCrawfor"] <- train$Neighborhood == "Crawfor"
train["NbrStoneBr"] <- train$Neighborhood == "StoneBr"

regr <- lm(log(SalePrice) ~  log(YearBuilt) +
log(LotArea) +
log(LivAreaSF) +
OverallQual +
OverallCond +
NbrCrawfor +
NbrStoneBr +
```

```
Fireplaces +
log(X1stFlrSF),
train)

summary(regr)
```

2.2 Regression Result

```
Call:
lm(formula = log(SalePrice) ~ log(YearBuilt) + log(LotArea) +
    log(LivAreaSF) + OverallQual + OverallCond + NbrCrawfor +
    NbrStoneBr + Fireplaces + log(X1stFlrSF), data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-1.48978 -0.06348  0.01079  0.08101  0.41643

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -62.288004   3.774280 -16.503  < 2e-16 ***
log(YearBuilt)   9.083924   0.495465  18.334  < 2e-16 ***
log(LotArea)     0.106959   0.012929   8.273 6.33e-16 ***
log(LivAreaSF)   0.350201   0.026099  13.418  < 2e-16 ***
OverallQual      0.095764   0.006741  14.206  < 2e-16 ***
OverallCond      0.054689   0.005907   9.259  < 2e-16 ***
NbrCrawforTRUE   0.143743   0.032664   4.401 1.24e-05 ***
NbrStoneBrTRUE   0.128496   0.044761   2.871  0.00422 **
Fireplaces       0.026418   0.011039   2.393  0.01696 *
log(X1stFlrSF)   0.137169   0.024835   5.523 4.65e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1554 on 720 degrees of freedom
Multiple R-squared:  0.8407,    Adjusted R-squared:  0.8387
F-statistic: 422.2 on 9 and 720 DF,  p-value: < 2.2e-16
```

3 Omitted Variables Bias

Let’s take as omitted variables log(LivAreaSF), OverallQual and Fireplaces So the reduced model is:

3.1 R-code

```
regr_r <- lm(log(SalePrice) ~  log(YearBuilt) +
```

```
log(LotArea) +
#log(LivAreaSF) +
#OverallQual +
OverallCond +
NbrCrawfor +
NbrStoneBr +
#Fireplaces +
log(X1stFlrSF),
train)

summary(regr_r)

AIC(regr)
AIC(regr_r)
BIC(regr)
BIC(regr_r)
```

3.2 Regression Result

```
Call:
lm(formula = log(SalePrice) ~ log(YearBuilt) + log(LotArea) +
    OverallCond + NbrCrawfor + NbrStoneBr + log(X1stFlrSF), data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-1.16706 -0.14752 -0.02032  0.15021  0.84108

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -98.466469   4.788457 -20.563  < 2e-16 ***
log(YearBuilt)  13.933375   0.635221  21.935  < 2e-16 ***
log(LotArea)     0.158458   0.018662   8.491  < 2e-16 ***
OverallCond      0.061812   0.008655   7.142 2.25e-12 ***
NbrCrawforTRUE   0.295951   0.047812   6.190 1.01e-09 ***
NbrStoneBrTRUE   0.217598   0.066358   3.279  0.00109 **
log(X1stFlrSF)   0.424935   0.033904  12.534  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2326 on 723 degrees of freedom
Multiple R-squared:  0.6419,	Adjusted R-squared:  0.639
F-statistic:   216 on 6 and 723 DF,  p-value: < 2.2e-16

> AIC(regr)
[1] -634.1398
> AIC(regr_r)
[1] -48.87182
> BIC(regr)
[1] -583.6163
> BIC(regr_r)
[1] -12.12746
>
```

Both AIC and BIC drastically increase when we omit these variables, and R^2 with adjusted R^2 heavily drops. We clearly lose information by omitting these variables.

3.3 Bias calculation

Let's now calculate the bias:

3.3.1 R-code

```
X1 <- cbind(log(train$YearBuilt),
log(train$LotArea),
train$OverallCond,
train$NbrCrawfor,
train$NbrStoneBr,
log(train$X1stFlrSF))
```

```
X2 <- cbind(log(train$LivAreaSF),
train$OverallQual,
train$Fireplaces)

bias1 <- qr.solve(t(X1) %*% X1) %*% t(X1) %*% X2 %*% regr$coefficients[c(4,5,9)]
print(bias1)
```

```
            [,1]
[1,]  0.05488600
[2,]  0.03309558
[3,] -0.01649894
[4,]  0.10324056
[5,]  0.11178711
[6,]  0.35831293
```

4 Redundant Variables. Multicollinearity problem

Let's add to our basic regression GarageArea, GarageCars and both of them.

4.0.1 R-code

```
regrGA <- lm(log(SalePrice) ~  log(YearBuilt) +
log(LotArea) +
log(LivAreaSF) +
OverallQual +
OverallCond +
NbrCrawfor +
NbrStoneBr +
Fireplaces +
log(X1stFlrSF) +
GarageArea,
train)

summary(regrGA)
```

```
regrGC <- lm(log(SalePrice) ~  log(YearBuilt) +
log(LotArea) +
log(LivAreaSF) +
OverallQual +
OverallCond +
NbrCrawfor +
NbrStoneBr +
Fireplaces +
log(X1stFlrSF) +
GarageCars,
train)

summary(regrGC)

regrGBoth <- lm(log(SalePrice) ~  log(YearBuilt) +
log(LotArea) +
log(LivAreaSF) +
OverallQual +
OverallCond +
NbrCrawfor +
NbrStoneBr +
Fireplaces +
log(X1stFlrSF) +
GarageArea +
GarageCars,
train)

summary(regrGBoth)
```

4.0.2 Regression Result

```
Call:
lm(formula = log(SalePrice) ~ log(YearBuilt) + log(LotArea) +
    log(LivAreaSF) + OverallQual + OverallCond + NbrCrawfor +
    NbrStoneBr + Fireplaces + log(X1stFlrSF) + GarageArea, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-1.54942 -0.05935  0.01205  0.07583  0.39844

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -5.715e+01  3.828e+00 -14.927  < 2e-16 ***
log(YearBuilt)  8.457e+00  5.004e-01  16.902  < 2e-16 ***
log(LotArea)    9.522e-02  1.288e-02   7.393 4.00e-13 ***
log(LivAreaSF)  3.269e-01  2.599e-02  12.578  < 2e-16 ***
OverallQual     8.961e-02  6.717e-03  13.341  < 2e-16 ***
OverallCond     5.444e-02  5.798e-03   9.390  < 2e-16 ***
NbrCrawforTRUE  1.510e-01  3.209e-02   4.705 3.04e-06 ***
NbrStoneBrTRUE  1.322e-01  4.394e-02   3.008  0.00272 **
Fireplaces      2.982e-02  1.085e-02   2.747  0.00616 **
log(X1stFlrSF)  1.143e-01  2.475e-02   4.617 4.61e-06 ***
GarageArea      1.854e-04  3.481e-05   5.326 1.35e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1526 on 719 degrees of freedom
Multiple R-squared:  0.8467,    Adjusted R-squared:  0.8446
F-statistic: 397.3 on 10 and 719 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = log(SalePrice) ~ log(YearBuilt) + log(LotArea) +
    log(LivAreaSF) + OverallQual + OverallCond + NbrCrawfor +
    NbrStoneBr + Fireplaces + log(X1stFlrSF) + GarageCars, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-1.37782 -0.06162  0.01282  0.07937  0.40623

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -54.684262   3.875922 -14.109  < 2e-16 ***
log(YearBuilt)   8.127074   0.506759  16.037  < 2e-16 ***
log(LotArea)     0.097846   0.012687   7.712 4.12e-14 ***
log(LivAreaSF)   0.318069   0.025956  12.254  < 2e-16 ***
OverallQual      0.088522   0.006673  13.266  < 2e-16 ***
OverallCond      0.055443   0.005758   9.628  < 2e-16 ***
NbrCrawforTRUE   0.147006   0.031842   4.617 4.62e-06 ***
NbrStoneBrTRUE   0.130959   0.043631   3.001  0.00278 **
Fireplaces       0.027019   0.010761   2.511  0.01226 *
log(X1stFlrSF)   0.122936   0.024315   5.056 5.44e-07 ***
GarageCars       0.063300   0.010158   6.232 7.85e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1515 on 719 degrees of freedom
Multiple R-squared:  0.8489,	Adjusted R-squared:  0.8468
F-statistic: 403.8 on 10 and 719 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = log(SalePrice) ~ log(YearBuilt) + log(LotArea) +
    log(LivAreaSF) + OverallQual + OverallCond + NbrCrawfor +
    NbrStoneBr + Fireplaces + log(X1stFlrSF) + GarageArea + GarageCars,
    data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-1.40484 -0.06141  0.01215  0.07881  0.40551

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -5.470e+01  3.878e+00 -14.107  < 2e-16 ***
log(YearBuilt)  8.133e+00  5.071e-01  16.040  < 2e-16 ***
log(LotArea)    9.676e-02  1.281e-02   7.557 1.26e-13 ***
log(LivAreaSF)  3.178e-01  2.597e-02  12.238  < 2e-16 ***
OverallQual     8.829e-02  6.685e-03  13.206  < 2e-16 ***
OverallCond     5.529e-02  5.766e-03   9.590  < 2e-16 ***
NbrCrawforTRUE  1.480e-01  3.189e-02   4.640 4.13e-06 ***
NbrStoneBrTRUE  1.314e-01  4.365e-02   3.009  0.00271 **
Fireplaces      2.761e-02  1.080e-02   2.555  0.01081 *
log(X1stFlrSF)  1.203e-01  2.466e-02   4.880 1.31e-06 ***
GarageArea      3.667e-05  5.750e-05   0.638  0.52391
GarageCars      5.469e-02  1.690e-02   3.236  0.00127 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1516 on 718 degrees of freedom
Multiple R-squared:  0.849,     Adjusted R-squared:  0.8466
F-statistic: 366.9 on 11 and 718 DF,  p-value: < 2.2e-16
```

It can be seen that if either GarageArea or GarageCars is used, it is statistically significant. It fits common sense that some factor reflecting the presense of a garage should influence on the price of a house, especially in Ames, Iowa, since small American towns are known for not having extensive public transport system (if any). However, if we use both factors, GarageArea becomes statistically insignificant. Therefore, we should use only one of them. Both AIC and BIC criteria clearly demonstrate that we should use the model with GarageCars

```
AIC(regrGA)
AIC(regrGC)
AIC(regrGBoth)
BIC(regrGA)
BIC(regrGC)
BIC(regrGBoth)

> AIC(regrGC)
[1] -670.541
> AIC(regrGBoth)
```

```
[1] -668.9543
> BIC(regrGA)
[1] -605.2656
> BIC(regrGC)
[1] -615.4245
> BIC(regrGBoth)
[1] -609.2447
```

It corresponds to some common sense that the capacity of garage in the terms of the number of cars which could be parked there better reflects the utility of the garage than its area. In principle, if the garage is badly planned, larger area may not increase its utility. On a relatively small data set it is hardly possible to distinguish the effects of car capacity and area. Anyway, let's compare the biases which arise if we remove from the full model either GarageCars or GarageArea:

4.0.3 Bias calculation

```
X1GC <- cbind(log(train$YearBuilt),
log(train$LotArea),
log(train$LivAreaSF),
train$OverallQual,
train$OverallCond,
train$NbrCrawfor,
train$NbrStoneBr,
log(train$X1stFlrSF),
train$Fireplaces,
train$GarageArea)

X2GC <- cbind(train$GarageCars)

biasGC <- qr.solve(t(X1GC) %*% X1GC) %*% t(X1GC) %*% X2GC %*% regrGBoth$coeffici
print(biasGC)
```

```
                [,1]
[1,]  4.691614e-03
[2,] -2.385473e-03
[3,]  4.808618e-03
[4,]  3.257110e-03
[5,] -2.271647e-03
[6,]  1.594275e-04
```

```
[7,]  5.430548e-05
[8,] -4.940581e-03
[9,]  2.516873e-03
[10,]  1.542976e-04

X1GA <- cbind(log(train$YearBuilt),
log(train$LotArea),
log(train$LivAreaSF),
train$OverallQual,
train$OverallCond,
train$NbrCrawfor,
train$NbrStoneBr,
log(train$X1stFlrSF),
train$Fireplaces,
train$GarageCars)

X2GA <- cbind(train$GarageArea)

biasGA <- qr.solve(t(X1GC) %*% X1GC) %*% t(X1GC) %*% X2GC %*% regrGBoth$coeffici
print(biasGA)


               [,1]
[1,]  3.145373e-06
[2,] -1.599279e-06
[3,]  3.223815e-06
[4,]  2.183647e-06
[5,] -1.522968e-06
[6,]  1.068841e-07
[7,]  3.640773e-08
[8,] -3.312287e-06
[9,]  1.687374e-06
[10,]  1.034449e-07
```

It is clear that bias for all coefficients is considerably larger if we omit GarageCars than if we omit GarageArea. It also confirms our conclusions that we should use the model with GarageCars.