

6 – Multiple Regression

More than one “X” variable.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \epsilon_i \quad (6.1)$$

Why?

- Might be interested in more than one marginal effect
- Omitted Variable Bias (OVB)

6.1 and 6.2 – House prices and OVB

Should I build a fireplace?

The following empirical example uses data on house prices, in the New York area in 2002-2003 (the data are from Richard De Veaux of Williams College).

Let's try to determine the value of a fireplace. First, load the data and take a look at it.

```
houses <-  
read.csv("http://rtgodwin.com/data/houseprice.csv")  
  
head(houses)
```

The “head” command prints out the first 6 observations from each variable. You should see something like:

Price	Lot.Size	Waterfront	Age	Land.Value	New.Construct
132500	0.09	0	42	50000	0
181115	0.92	0	0	22300	0
109000	0.19	0	133	7300	0
155000	0.41	0	13	18700	0
86060	0.11	0	0	15000	1
120000	0.68	0	31	14000	0
Central.Air	Fuel.Type	Heat.Type	Sewer.Type	Living.Area	Pct.College
0	3	4	2	906	35
0	2	3	2	1953	51
0	2	3	3	1944	51
0	2	2	2	1944	51
1	2	2	3	840	51
0	2	2	2	1152	22
Bedrooms	Fireplaces	Bathrooms	Rooms		
2	1	1.0	5		
3	0	2.5	6		
4	1	1.0	8		
3	1	1.0	5		
2	0	1.0	3		
4	1	1.0	8		

We are interested in the effect of the variable *Fireplaces* on *Price*.

Is *Fireplaces* a dummy variable?

```
summary(house$Fireplaces)
```

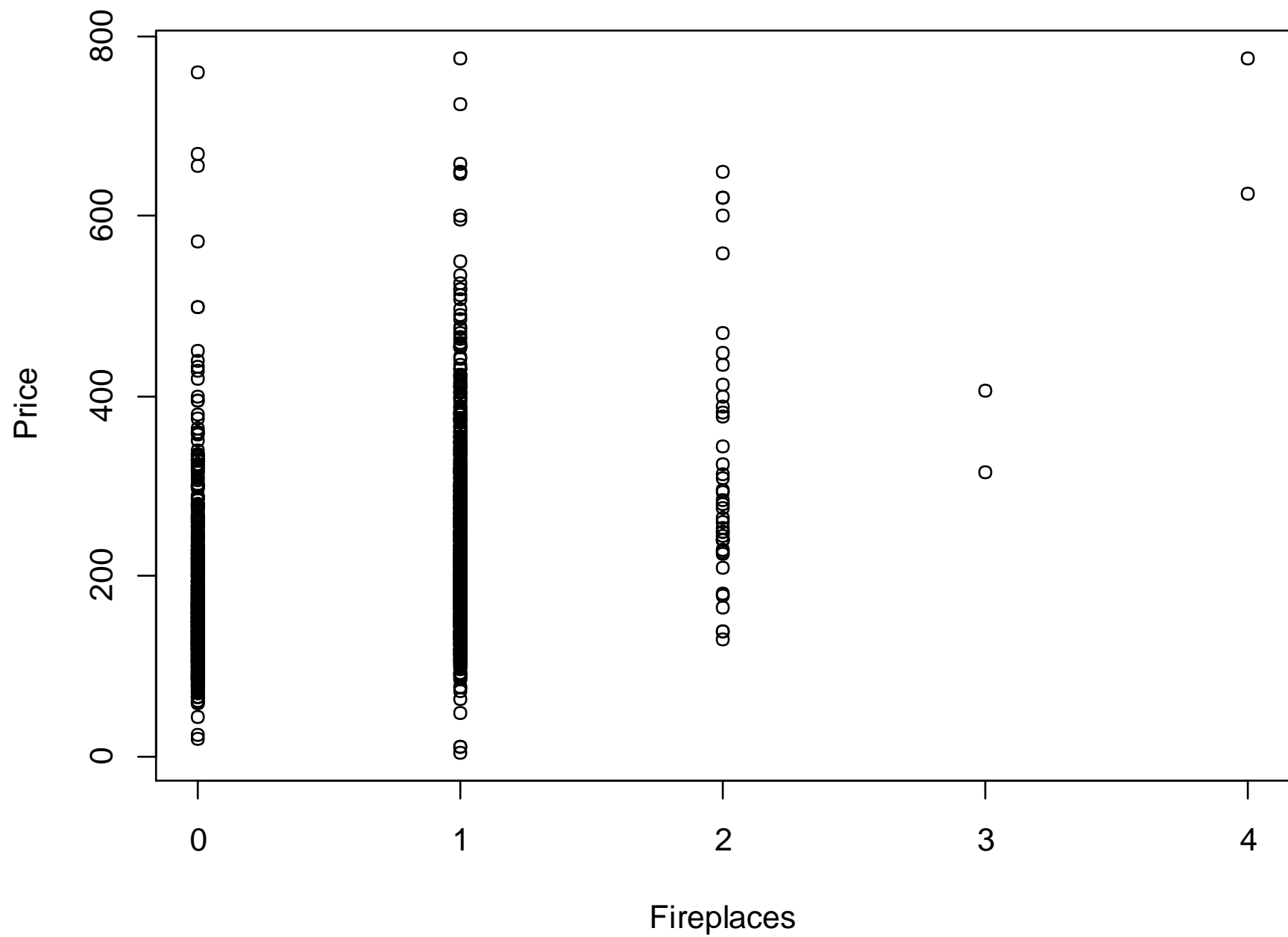
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0000	0.0000	1.0000	0.6019	1.0000	4.0000

Before we proceed, let's instead measure *Price* in thousands of dollars:

```
house$Price = house$Price / 1000
```

Now, let's see the relationship between *Fireplaces* and *Price*.

```
plot(house$Fireplaces, house$Price)
```



Let's see the average Price conditional on different number of Fireplaces:

```
mean(house$Price[house$Fireplaces == 0])  
[1] 174.6533  
mean(house$Price[house$Fireplaces == 1])  
[1] 235.1629  
mean(house$Price[house$Fireplaces == 2])  
[1] 318.8214  
mean(house$Price[house$Fireplaces == 3])  
[1] 360.5  
mean(house$Price[house$Fireplaces == 4])  
[1] 700
```

Correlation?

```
cor(house$Price, house$Fireplaces)  
[1] 0.3767862
```

It appears that the more Fireplaces, the higher the Price.

Let's try estimating the population model:

$$Price_i = \beta_0 + \beta_1 Fireplaces_i + \epsilon_i$$

```
summary(lm(Price ~ Fireplaces, data = house))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	171.824	3.234	53.13	<2e-16	***
Fireplaces	<u>66.699</u>	3.947	<u>16.90</u>	<2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

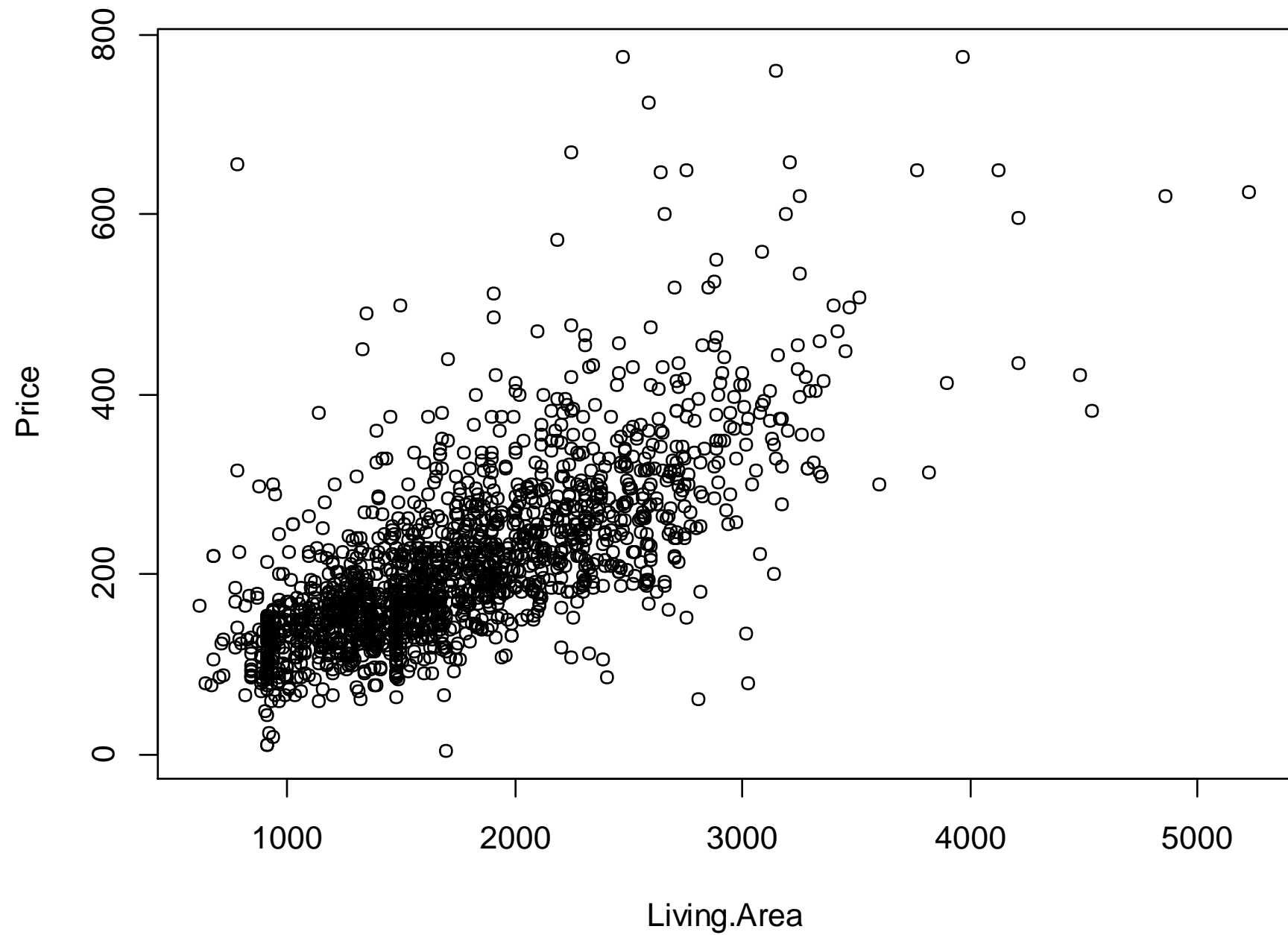
Residual standard error: 91.21 on 1726 degrees of freedom

Multiple R-squared: 0.142, Adjusted R-squared: 0.1415

F-statistic: 285.6 on 1 and 1726 DF, p-value: < 2.2e-16

Questions:

- What is the marginal effect of *Fireplaces* on *Price*?
- How much does it cost to install a fireplace?
- Should I install a fireplace in my home?
- What the ? is going on here?
- What do you think the main determinant of *Price* should be?



The above plot was generated using the code:

```
plot(house$Living.Area, house$Price)
```

Is there a positive relationship between *Living.Area* and *Price*?

Now, estimate the model:

$$Price_i = \beta_0 + \beta_1 Living.Area_i + \epsilon_i$$

```
summary(lm(Price ~ Living.Area))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	13.439394	4.992353	2.692	0.00717	**
Living.Area	<u>0.113123</u>	0.002682	<u>42.173</u>	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 69.1 on 1726 degrees of freedom

Multiple R-squared: 0.5075, Adjusted R-squared: 0.5072

F-statistic: 1779 on 1 and 1726 DF, p-value: < 2.2e-16

- What is the marginal effect?
- Note the R^2 between the two regressions.
- What might be a problem with determining these two marginal effects separately?

```
cor(Living.Area, Fireplaces)
```

```
[1] 0.4737878
```

- If the variable *Living.Area* is excluded from the original regression, then it goes into the error term, u_i .
- If *Living.Area* and *Fireplaces* are positively correlated, then more fireplaces are just indicating a bigger house!
- That is, the error term is correlated with the “X” variable, and L.S.A. #1 is violated! The OLS estimator for β_1 in the first regression will be biased.

How can we take care of this problem? Include both variables in the model!

$$Price_i = \beta_0 + \beta_1 Fireplaces_i + \beta_2 Living.Area_i + \epsilon_i$$

```
summary(lm(Price ~ Fireplaces + Living.Area, data = house))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	14.730146	5.007563	2.942	0.00331	**
Fireplaces	<u>8.962440</u>	3.389656	2.644	<u>0.00827</u>	**
Living.Area	<u>0.109313</u>	0.003041	35.951	< 2e-16	***

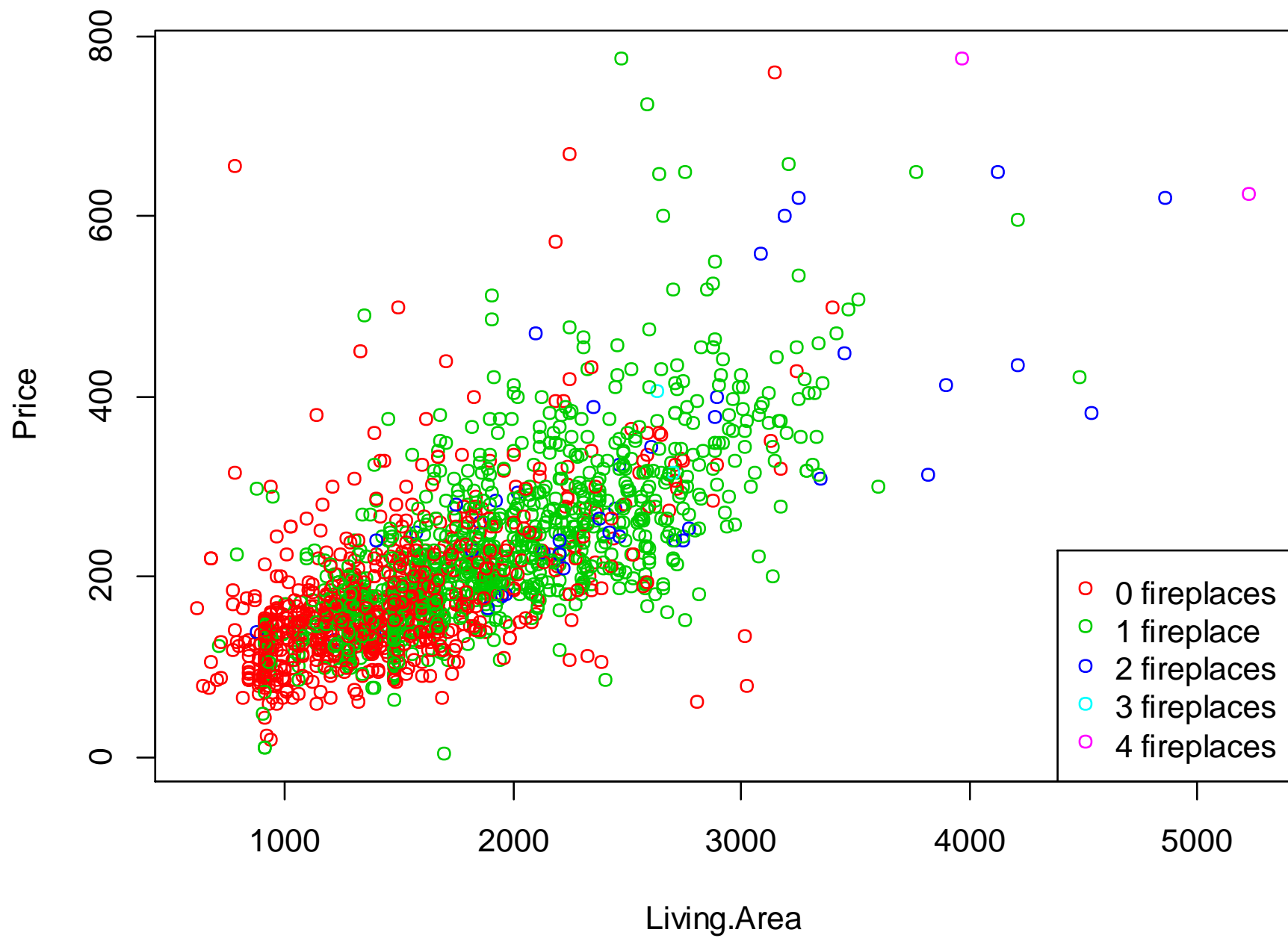
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 68.98 on 1725 degrees of freedom

Multiple R-squared: 0.5095, Adjusted R-squared: 0.5089

F-statistic: 895.9 on 2 and 1725 DF, p-value: < 2.2e-16

- Notice how the estimated marginal effects have changed.
- Notice that *Fireplaces* is now a lot less significant.
- This is an example of omitted variable bias (OVB).



Omitted Variable Bias

$$\hat{Price} = 171.82 + 66.70 \times Fireplaces, \quad R^2 = 0.142$$

(3.23) (3.95)

$$\hat{Price} = 14.73 + 8.96 \times Fireplaces + 0.11 \times Living.Area, \quad R^2 = 0.511$$

(5.01) (3.39) (0.003)

Several results have changed with the addition of the `Living.Area` variable:

- The estimated value of an additional fireplace has dropped from \$66,699 to \$8,962.
- The R^2 has increased from 0.142 to 0.5095.
- The estimated intercept has changed by a lot (but this is unimportant).
- There is a new estimated β : $b_2 = 0.11$. This means that, it is estimated that an additional square-foot of house size increases price by \$110.

Omitted Variable Bias

- Omitted variable bias (OVB) occurs when one or more of the variables in the random error term ϵ are related to one or more of the X variables
- A.5: X and ϵ are independent. OVB is a violation of this assumption, resulting in bias and inconsistency of OLS
- Suppose that X and Z both cause Y
- Suppose X and Z are correlated
- What happens when X changes?
- What is the problem with attributing changes in X to changes in Y ?

Solution: include the omitted variable if possible