



THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

Model building: interactions 1

Lecture 12

STA 371G

Announcements

- Reminder to submit your team evaluations for Part 1 of the project by Friday at 11:59 PM.
- Starting next Tuesday night, the R help session will move permanently to GSB 3.130 (lets us produce better video recordings of the session for people that cannot attend).

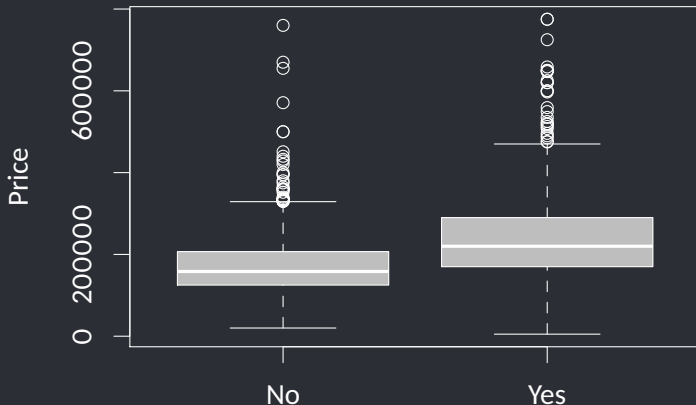
Housing price data

Today we'll consider a 2007 housing price data set from Saratoga County, NY.

- **Price:** price of house (\$)
- **Living.Area:** amount of living space (sq ft)
- **Fireplace:** whether house has a fireplace (yes/no)

How much is a fireplace worth?

```
boxplot(Price ~ Fireplace, data=houses,  
        col='gray', ylab='Price')
```



How much is a fireplace worth?

If we regress Price on Fireplace, we get the regression equation

$$\widehat{\text{Price}} = 174653 + 65261 \cdot (\text{Fireplace} = \text{Yes})$$

The average difference between houses with and without a fireplace is \$65261.



How much is a fireplace worth?

Note that the coefficient represents the difference between the means, and the intercept in the mean price when Fireplace is “No”:

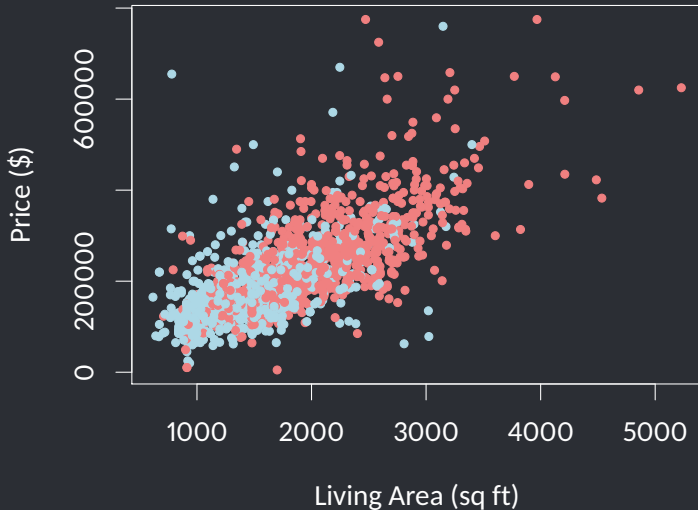
```
tapply(houses$Price, houses$Fireplace, mean)
```

No	Yes
174653	239914

$239914 - 174653$

```
[1] 65261
```

What is the relationship between price and size?



Predicting price from living area

Let's start by creating a simple regression predicting price from living area (in sq ft).


```
modell1 <- lm(Price ~ Living.Area, data=houses)
summary(modell1)
```

Call:

```
lm(formula = Price ~ Living.Area, data = houses)
```

Residuals:

Min	1Q	Median	3Q	Max
-277022	-39371	-7726	28350	553325

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	13439.39	4992.35	2.69	0.0072	**
Living.Area	113.12	2.68	42.17	<2e-16	***

Signif. codes: 0 '***' 0 '**' 0 '*' 0 '.' 0 ' ' 1

Residual standard error: 69100 on 1726 degrees of freedom

Multiple R-squared: 0.507, Adjusted R-squared: 0.507

F-statistic: 1.78e+03 on 1 and 1726 DF, p-value: <2e-16

Can we do better by adding a dummy variable for fireplace to the model?

```
model2 <- lm(Price ~ Living.Area + Fireplace, data=houses)
summary(model2)
```

Call:

```
lm(formula = Price ~ Living.Area + Fireplace, data = houses)
```

Residuals:

Min	1Q	Median	3Q	Max
-271421	-39935	-7887	28215	554651

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	13599.16	4991.70	2.72	0.0065	**
Living.Area	111.22	2.97	37.48	<2e-16	***
FireplaceYes	5567.38	3716.95	1.50	0.1344	

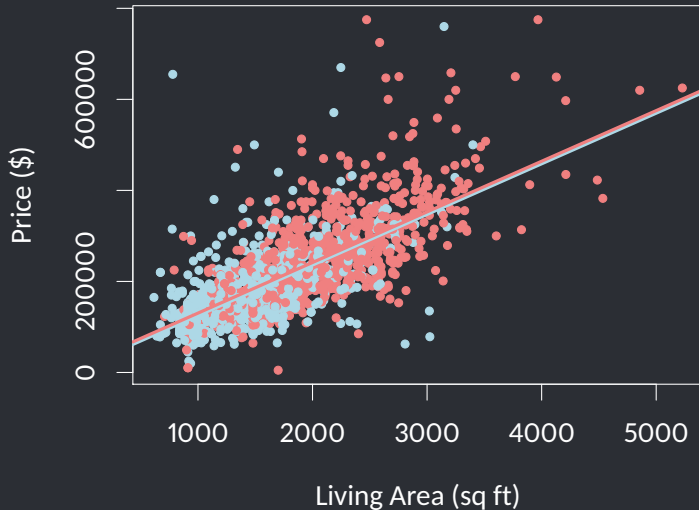
Signif. codes: 0 '***' 0 '**' 0 '*' 0 '.' 0 ' ' 1

Residual standard error: 69100 on 1725 degrees of freedom

Multiple R-squared: 0.508, Adjusted R-squared: 0.508

F-statistic: 891 on 2 and 1725 DF, p-value: <2e-16

By adding the dummy variable, we are essentially fitting two regression lines:



They have the same slope, but different intercepts

Interactions

Our regression equation is

$$\widehat{\text{Price}} = 13599 + 111 \cdot \text{Living.Area} + 5567 \cdot \text{FireplaceYes.}$$

Interactions

Our regression equation is

$$\widehat{\text{Price}} = 13599 + 111 \cdot \text{Living.Area} + 5567 \cdot \text{FireplaceYes.}$$

What if the *slope* of the best-fit line is different for houses with a fireplace than for houses without?

Interactions

Our regression equation is

$$\widehat{\text{Price}} = 13599 + 111 \cdot \text{Living.Area} + 5567 \cdot \text{FireplaceYes.}$$

What if the *slope* of the best-fit line is different for houses with a fireplace than for houses without?

Equivalently, what if the *effect* of having a bigger house is different for houses with fireplaces than for houses without fireplaces?

Interactions

To model this, we can add an **interaction term** that consists of the product of the two predictors:

$$\begin{aligned}\text{Price} = & \beta_0 + \beta_1 \cdot \text{Living.Area} + \beta_2 \cdot \text{FireplaceYes} \\ & + \beta_3 \cdot \text{Living.Area} \cdot \text{FireplaceYes} + \epsilon_j.\end{aligned}$$

Interactions

To model this, we can add an **interaction term** that consists of the product of the two predictors:

$$\begin{aligned}\text{Price} = & \beta_0 + \beta_1 \cdot \text{Living.Area} + \beta_2 \cdot \text{FireplaceYes} \\ & + \beta_3 \cdot \text{Living.Area} \cdot \text{FireplaceYes} + \epsilon_j.\end{aligned}$$

Now, the *slope* of Living.Area depends on the *value* of Fireplace!

Houses with a fireplace have a slope of $\beta_1 + \beta_3$, houses without have a slope of β_1 .

```
model3 <- lm(Price ~ Living.Area * Fireplace, data=houses)
summary(model3)
```

Call:

```
lm(formula = Price ~ Living.Area * Fireplace, data = houses)
```

Residuals:

Min	1Q	Median	3Q	Max
-241710	-39588	-7821	28480	542055

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	40901.29	8234.66	4.97	0.00000075	***
Living.Area	92.36	5.41	17.07	< 2e-16	***
FireplaceYes	-37610.41	11024.85	-3.41	0.00066	***
Living.Area:FireplaceYes	26.85	6.46	4.16	0.00003376	***

Signif. codes: 0 '***' 0 '**' 0 '*' 0 '.' 0 ' ' 1

Residual standard error: 68800 on 1724 degrees of freedom

Multiple R-squared: 0.513, Adjusted R-squared: 0.512

F-statistic: 605 on 3 and 1724 DF, p-value: <2e-16

This corresponds to the regression equation:

$$\widehat{\text{Price}} = 40901 + 92 \cdot \text{Living.Area} - 37610 \cdot \text{FireplaceYes} \\ + 27 \cdot \text{Living.Area} \cdot \text{FireplaceYes}$$

This corresponds to the regression equation:

$$\widehat{\text{Price}} = 40901 + 92 \cdot \text{Living.Area} - 37610 \cdot \text{FireplaceYes} \\ + 27 \cdot \text{Living.Area} \cdot \text{FireplaceYes}$$

In other words, for houses without a fireplace:

$$\widehat{\text{Price}} = 40901 + 92 \cdot \text{Living.Area}$$

This corresponds to the regression equation:

$$\widehat{\text{Price}} = 40901 + 92 \cdot \text{Living.Area} - 37610 \cdot \text{FireplaceYes} \\ + 27 \cdot \text{Living.Area} \cdot \text{FireplaceYes}$$

In other words, for houses without a fireplace:

$$\widehat{\text{Price}} = 40901 + 92 \cdot \text{Living.Area}$$

And for houses with a fireplace:

$$\widehat{\text{Price}} = (40901 - 37610) + (92 + 27) \cdot \text{Living.Area}$$

Making predictions

Let's make predictions for the price of a 2500 sq ft house, both with and without a fireplace:

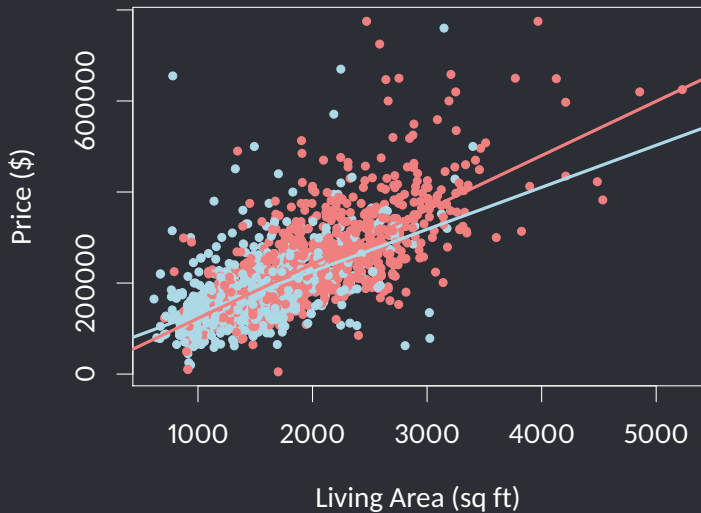
```
predict(model3, list(Living.Area=2500, Fireplace="Yes"),  
        interval="prediction")
```

	fit	lwr	upr
1	301331	166362	436300

```
predict(model3, list(Living.Area=2500, Fireplace="No"),  
        interval="prediction")
```

	fit	lwr	upr
1	271811	136405	407217





Main effects and interaction effects

In the output, the coefficients for Living.Space and Fireplace are **main effects**, and the coefficient for Living.Space · Fireplace is an **interaction effect**.

```
summary(model3)$coefficients
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	40901	8234.7	5.0	7.5e-07
Living.Area	92	5.4	17.1	1.8e-60
FireplaceYes	-37610	11024.9	-3.4	6.6e-04
Living.Area:FireplaceYes	27	6.5	4.2	3.4e-05

Main effects and interaction effects

In the output, the coefficients for Living.Space and Fireplace are **main effects**, and the coefficient for Living.Space · Fireplace is an **interaction effect**.

```
summary(model3)$coefficients
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	40901	8234.7	5.0	7.5e-07
Living.Area	92	5.4	17.1	1.8e-60
FireplaceYes	-37610	11024.9	-3.4	6.6e-04
Living.Area:FireplaceYes	27	6.5	4.2	3.4e-05

The main effect for Living.Area (92.36) represents the predicted incremental effect of each additional square foot of living space, when there is no fireplace present.

Main effects and interaction effects

In the output, the coefficients for Living.Space and Fireplace are **main effects**, and the coefficient for Living.Space · Fireplace is an **interaction effect**.

```
summary(model3)$coefficients
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	40901	8234.7	5.0	7.5e-07
Living.Area	92	5.4	17.1	1.8e-60
FireplaceYes	-37610	11024.9	-3.4	6.6e-04
Living.Area:FireplaceYes	27	6.5	4.2	3.4e-05

The main effect for Living.Area (92.36) represents the predicted incremental effect of each additional square foot of living space, when there is no fireplace present.

When we have an interaction term in the model, we *must* include the main effect as well!