THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

# Indicator Variables and Interactions

## Lecture 18

STA 371G

1. Using categorical variables with 2 categories in a regression model

2. Using categorical variables with 3+ categories in a regression model

3. Interactions between a categorical and a quantitative variable

4. Interactions between two quantitative variables

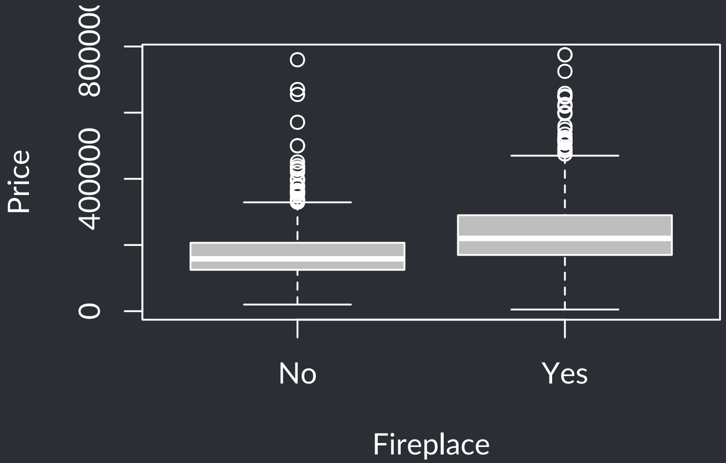
## Housing price data

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')
```



## Using a categorical variable as a predictor

- We can't just throw a categorical variable into a regression—all $X$'s have to be quantitative

# Using a categorical variable as a predictor

- We can't just throw a categorical variable into a regression—all $X$'s have to be quantitative
- Idea: Recode a binary (yes/no) categorical variable as 0/1

# Using a categorical variable as a predictor

- We can't just throw a categorical variable into a regression—all $X$'s have to be quantitative
- Idea: Recode a binary (yes/no) categorical variable as 0/1
- This is called a indicator variable or dummy variable

# Using a categorical variable as a predictor

- We can't just throw a categorical variable into a regression—all $X$'s have to be quantitative
- Idea: Recode a binary (yes/no) categorical variable as 0/1
- This is called a indicator variable or dummy variable
- Let's set 1 = Yes and 0 = No (doesn't matter; just need to be consistent)

# How much is a fireplace worth?

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

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

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
```

1. Using categorical variables with 2 categories in a regression model

2. Using categorical variables with 3+ categories in a regression model

3. Interactions between a categorical and a quantitative variable

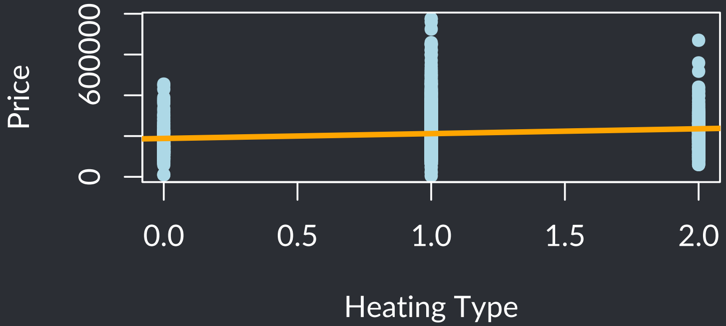
4. Interactions between two quantitative variables

# Using a categorical variable with 3+ categories as a predictor

- Let's say we want to predict price from type of heat (electric, hot air, hot water)
- We CANNOT set 0 = electric, 1 = hot air, 2 = hot water and throw that into the model!

## Using a categorical variable with 3+ categories as a predictor

- Let's say we want to predict price from type of heat (electric, hot air, hot water)
- We CANNOT set 0 = electric, 1 = hot air, 2 = hot water and throw that into the model!



# Using a categorical variable with 3+ categories as a predictor

1. Pick an (arbitrary) reference category, say Electric
2. For the other categories, create an indicator variable that is 1 if the value is that category, and 0 otherwise

| Value | Heat Type is Hot Air | Heat Type is Hot Water |
|---|---|---|
| Electric | 0 | 0 |
| Hot Air | 1 | 0 |
| Hot Water | 0 | 1 |

If you add a categorical variable to a model, R will pick a reference category and create indicator variables for you:

```
model <- lm(Price ~ Heat.Type, data=houses)
summary(model)


Call:
lm(formula = Price ~ Heat.Type, data = houses)

Residuals:
    Min      1Q  Median      3Q     Max
-221355  -63355  -17644   43895  548645

Coefficients:
                   Estimate Std. Error t value     Pr(>|t|)
(Intercept)          161889       5469   29.60      < 2e-16 ***
Heat.TypeHot Air      64467       6168   10.45      < 2e-16 ***
Heat.TypeHot Water    47244       7754    6.09 0.0000000014 ***
---
Signif. codes:  0 '***' 0 '**' 0 '*' 0 '.' 0 ' ' 1

Residual standard error: 95500 on 1725 degrees of freedom
Multiple R-squared:  0.0597,Adjusted R-squared:  0.0586
F-statistic: 54.8 on 2 and 1725 DF,  p-value: <2e-16
```

# Interpreting indicator variable slopes

- The slope of an indicator variable represents the predicted difference in $Y$ between the corresponding category and the reference category
- Example: The "`Heat.TypeHot Air`" slope of 64467 represents the predicted difference in prices between houses with hot air heat and houses with electric heat

Regression equation:

$$
\begin{aligned}
\widehat{\text{Price}} = 161889 &+ 64467 \cdot (\text{Heat Type = Hot Air}) \\
&+ 47244 \cdot (\text{Heat Type = Hot Water})
\end{aligned}
$$

Regression equation:

$$\widehat{\text{Price}} = 161889 + 64467 \cdot (\text{Heat Type = Hot Air}) \\ + 47244 \cdot (\text{Heat Type = Hot Water})$$

Let's write out the equations:

- Electric $\longrightarrow \widehat{\text{Price}} = 161889 + 64467 \cdot 0 + 47244 \cdot 0 = 161889$

Regression equation:

$$\begin{aligned}\widehat{\text{Price}} = 161889 + 64467 \cdot (\text{Heat Type = Hot Air}) \\ +47244 \cdot (\text{Heat Type = Hot Water})\end{aligned}$$

Let's write out the equations:

- Electric $\longrightarrow \widehat{\text{Price}} = 161889 + 64467 \cdot 0 + 47244 \cdot 0 = 161889$
- Hot Air $\longrightarrow \widehat{\text{Price}} = 161889 + 64467 \cdot 1 + 47244 \cdot 0 = 226355$

Regression equation:

$$\widehat{\text{Price}} = 161889 + 64467 \cdot (\text{Heat Type = Hot Air}) \\ + 47244 \cdot (\text{Heat Type = Hot Water})$$

Let's write out the equations:

- Electric $\longrightarrow \widehat{\text{Price}} = 161889 + 64467 \cdot 0 + 47244 \cdot 0 = 161889$
- Hot Air $\longrightarrow \widehat{\text{Price}} = 161889 + 64467 \cdot 1 + 47244 \cdot 0 = 226355$
- Hot Water $\longrightarrow$
  $\widehat{\text{Price}} = 161889 + 64467 \cdot 0 + 47244 \cdot 1 = 209132$

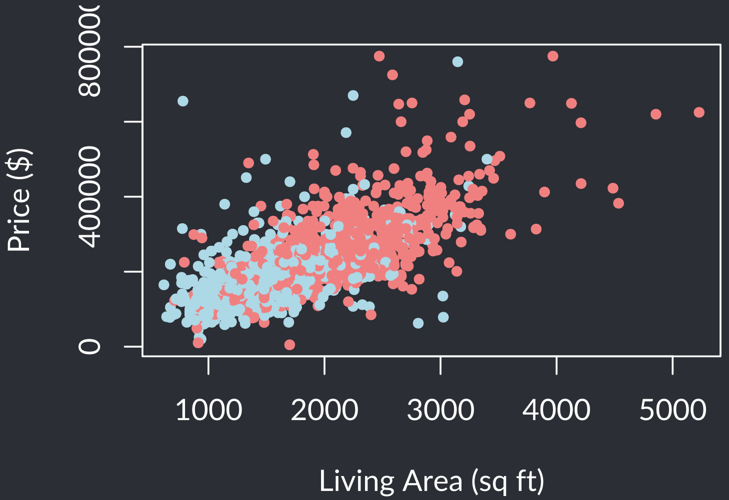
1. Using categorical variables with 2 categories in a regression model

2. Using categorical variables with 3+ categories in a regression model

3. Interactions between a categorical and a quantitative variable

4. Interactions between two quantitative variables

# 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).

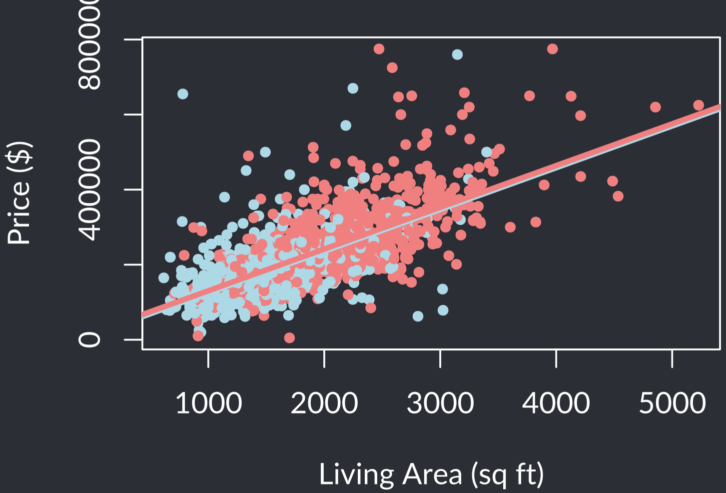
```
model1 <- lm(Price ~ Living.Area, data=houses)
summary(model1)


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?

```r
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:

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

# Interactions

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

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

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

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

```r
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

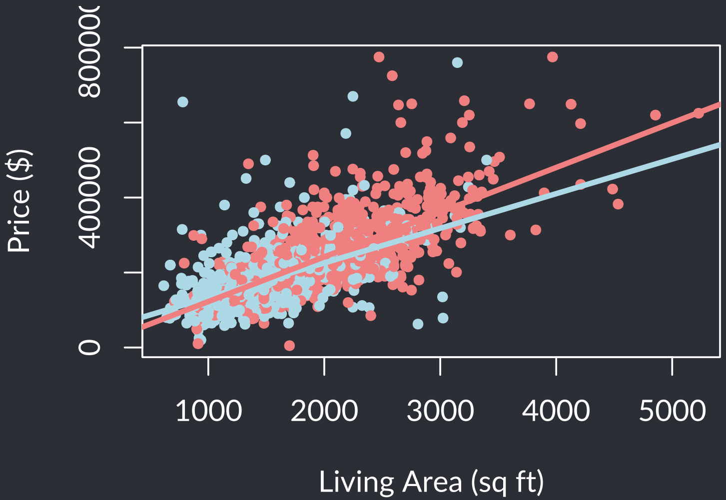
Let's make predictions for the price of a 2500 sq ft house, both with and without a fireplace:

```
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
```



Price ($)
0
400000
800000
1000
2000
3000
4000
5000
Living Area (sq ft)

# 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!

1. Using categorical variables with 2 categories in a regression model

2. Using categorical variables with 3+ categories in a regression model

3. Interactions between a categorical and a quantitative variable

4. Interactions between two quantitative variables

## NBA data

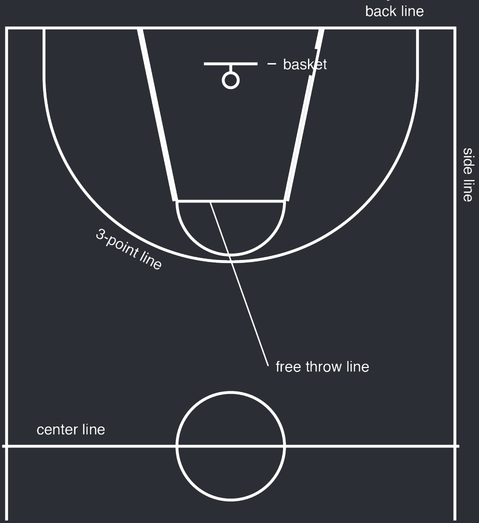
Basketball-Reference.com provides detailed data on NBA teams and players. We'll look at team data for 4 seasons ending in 2016; each of these metrics is the average across the season:

- **PTS**: Total points
- **PCT3P**: Percentage of 3-point shots made
- **N3PA**: Number of 3-point shots attempted

There are 30 NBA teams × 4 seasons = 120 cases in this file.

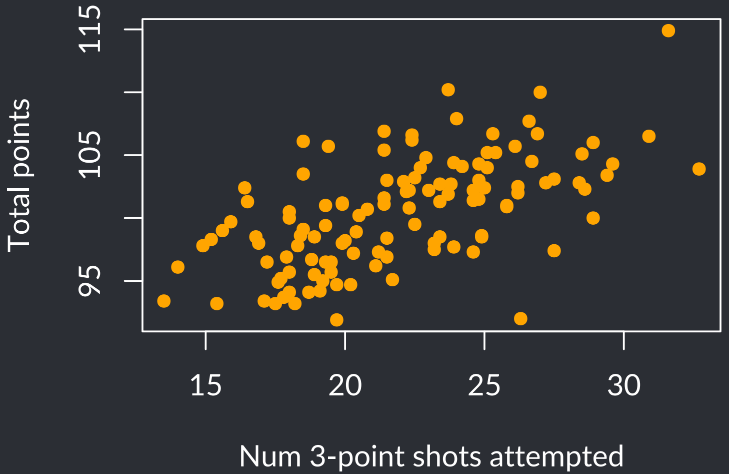
# NBA data

In basketball, there are three ways to score:




- **1 point** for free throws made after a foul by the other team
- **2 points** for shots made inside the 3-point line
- **3 points** for shots made outside the 3-point line

```r
plot(nba$N3PA, nba$PTS, pch=16, col='orange',
  xlab='Num 3-point shots attempted', ylab='Total points')
```



```r
model1 <- lm(PTS ~ N3PA, data=nba)
summary(model1)
```

```

Call:
lm(formula = PTS ~ N3PA, data = nba)

Residuals:
    Min      1Q  Median      3Q     Max 
-11.245  -2.511   0.055   2.225   8.640 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  86.1920     1.7746   48.57  < 2e-16 ***
N3PA          0.6484     0.0794    8.17  3.9e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.5 on 118 degrees of freedom
Multiple R-squared:  0.361,Adjusted R-squared:  0.356 
F-statistic: 66.8 on 1 and 118 DF,  p-value: 3.89e-13
```

# Can we do better?

$R^2 = 36\%$, so we can explain 36% of the variance in total points based only on knowing the number of 3-point attempts.

## Can we do better?

$R^2 = 36\%$, so we can explain 36% of the variance in total points based only on knowing the number of 3-point attempts.

This means that **most** of the variance (64%) in total points is **not** explained by the number of 3-point attempts.

# Can we do better?

$R^2 = 36\%$, so we can explain 36% of the variance in total points based only on knowing the number of 3-point attempts.

This means that **most** of the variance (64%) in total points is **not** explained by the number of 3-point attempts.

Let's add another variable to our model — why might 3-point percentage be useful as another predictor?

# Can we do better?

```
model2 <- lm(PTS ~ N3PA + PCT3P, data=nba)
summary(model2)


Call:
lm(formula = PTS ~ N3PA + PCT3P, data = nba)

Residuals:
   Min     1Q Median     3Q    Max 
-8.349 -2.139 -0.079  1.869  9.190 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  62.0049     5.6140   11.04  < 2e-16 ***
N3PA          0.5647     0.0759    7.44  1.8e-11 ***
PCT3P         0.7342     0.1629    4.51  1.6e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.2 on 117 degrees of freedom
Multiple R-squared:  0.456,Adjusted R-squared:  0.447 
F-statistic:   49 on 2 and 117 DF,  p-value: 3.48e-16
```

# Can we do even better?

It would make sense that the **impact** of the number of 3-pointers taken on total points would **depend on** how well the team shoots the 3!

# Can we do even better?

It would make sense that the **impact** of the number of 3-pointers taken on total points would **depend on** how well the team shoots the 3!

This sounds like an interaction — let's make a model with an interaction between the two predictors!

```r
model3 <- lm(PTS ~ N3PA * PCT3P, data=nba)
summary(model3)
```

```
Call:
lm(formula = PTS ~ N3PA * PCT3P, data = nba)

Residuals:
   Min     1Q Median     3Q    Max 
-7.263 -2.276  0.115  1.970  9.376 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 122.8490    30.5894    4.02  0.00011 ***
N3PA         -2.1190     1.3290   -1.59  0.11356    
PCT3P        -0.9841     0.8646   -1.14  0.25740    
N3PA:PCT3P    0.0756     0.0374    2.02  0.04542 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.2 on 116 degrees of freedom
Multiple R-squared:  0.474,Adjusted R-squared:  0.461 
F-statistic: 34.9 on 3 and 116 DF,  p-value: 3.8e-16
```

Model 3 corresponds to the regression equation

$$\widehat{\text{PTS}} = 122.85 - 2.12 \cdot \text{N3PA} - 0.98 \cdot \text{PCT3P} + 0.08 \cdot \text{N3PA} \cdot \text{PCT3P}.$$

Model 3 corresponds to the regression equation

$$\widehat{\text{PTS}} = 122.85 - 2.12 \cdot \text{N3PA} - 0.98 \cdot \text{PCT3P} + 0.08 \cdot \text{N3PA} \cdot \text{PCT3P}.$$

We interpret the coefficients as follows:

- **Intercept** (122.85) is our prediction of total points when $\text{N3PA} = \text{PCT3P} = 0$. (Meaningless in this context!)

Model 3 corresponds to the regression equation

$$\widehat{\text{PTS}} = 122.85 - 2.12 \cdot \text{N3PA} - 0.98 \cdot \text{PCT3P} + 0.08 \cdot \text{N3PA} \cdot \text{PCT3P}.$$

We interpret the coefficients as follows:

- **Intercept** (122.85) is our prediction of total points when $\text{N3PA} = \text{PCT3P} = 0$. (Meaningless in this context!)
- **N3PA** ($-2.12$) is the predicted increase in total points for each additional 3-pointer taken, when $\text{PCT3P} = 0$.

Model 3 corresponds to the regression equation

$$\widehat{\text{PTS}} = 122.85 - 2.12 \cdot \text{N3PA} - 0.98 \cdot \text{PCT3P} + 0.08 \cdot \text{N3PA} \cdot \text{PCT3P}.$$

We interpret the coefficients as follows:

- **Intercept** (122.85) is our prediction of total points when $\text{N3PA} = \text{PCT3P} = 0$. (Meaningless in this context!)
- **N3PA** ($-2.12$) is the predicted increase in total points for each additional 3-pointer taken, when $\text{PCT3P} = 0$.
- **PCT3P** ($-0.98$) is the predicted increase in total points for each additional percentage point of 3-point shooting accuracy, when $\text{N3PA} = 0$.

Model 3 corresponds to the regression equation

$$\widehat{\text{PTS}} = 122.85 - 2.12 \cdot \text{N3PA} - 0.98 \cdot \text{PCT3P} + 0.08 \cdot \text{N3PA} \cdot \text{PCT3P}.$$

We interpret the coefficients as follows:

- **Intercept** (122.85) is our prediction of total points when $\text{N3PA} = \text{PCT3P} = 0$. (Meaningless in this context!)
- **N3PA** ($-2.12$) is the predicted increase in total points for each additional 3-pointer taken, when $\text{PCT3P} = 0$.
- **PCT3P** ($-0.98$) is the predicted increase in total points for each additional percentage point of 3-point shooting accuracy, when $\text{N3PA} = 0$.
- **N3PA · PCT3P** ($0.08$) can be interpreted in two ways:

Model 3 corresponds to the regression equation

$$\widehat{\text{PTS}} = 122.85 - 2.12 \cdot \text{N3PA} - 0.98 \cdot \text{PCT3P} + 0.08 \cdot \text{N3PA} \cdot \text{PCT3P}.$$

We interpret the coefficients as follows:

- **Intercept** (122.85) is our prediction of total points when $\text{N3PA} = \text{PCT3P} = 0$. (Meaningless in this context!)
- **N3PA** ($-2.12$) is the predicted increase in total points for each additional 3-pointer taken, when $\text{PCT3P} = 0$.
- **PCT3P** ($-0.98$) is the predicted increase in total points for each additional percentage point of 3-point shooting accuracy, when $\text{N3PA} = 0$.
- **N3PA · PCT3P** ($0.08$) can be interpreted in two ways:

Model 3 corresponds to the regression equation

$$\widehat{\text{PTS}} = 122.85 - 2.12 \cdot \text{N3PA} - 0.98 \cdot \text{PCT3P} + 0.08 \cdot \text{N3PA} \cdot \text{PCT3P}.$$

We interpret the coefficients as follows:

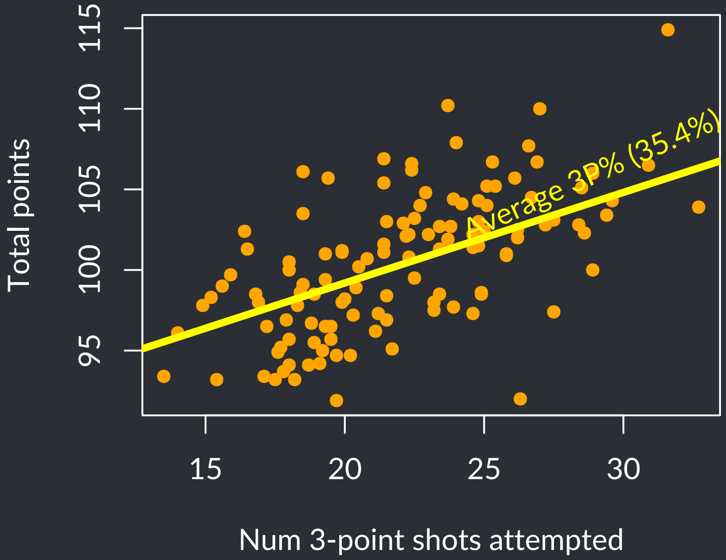
- **Intercept** (122.85) is our prediction of total points when N3PA = PCT3P = 0. (Meaningless in this context!)
- **N3PA** ($-2.12$) is the predicted increase in total points for each additional 3-pointer taken, when PCT3P = 0.
- **PCT3P** ($-0.98$) is the predicted increase in total points for each additional percentage point of 3-point shooting accuracy, when N3PA = 0.
- **N3PA · PCT3P** (0.08) can be interpreted in two ways:
  - the increase in the *slope coefficient* for N3PA for each 1-unit increase of PCT3P.

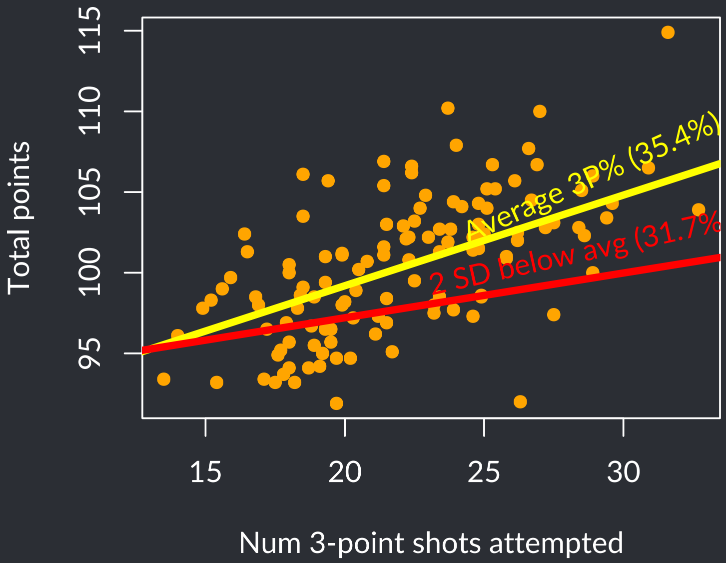
Model 3 corresponds to the regression equation

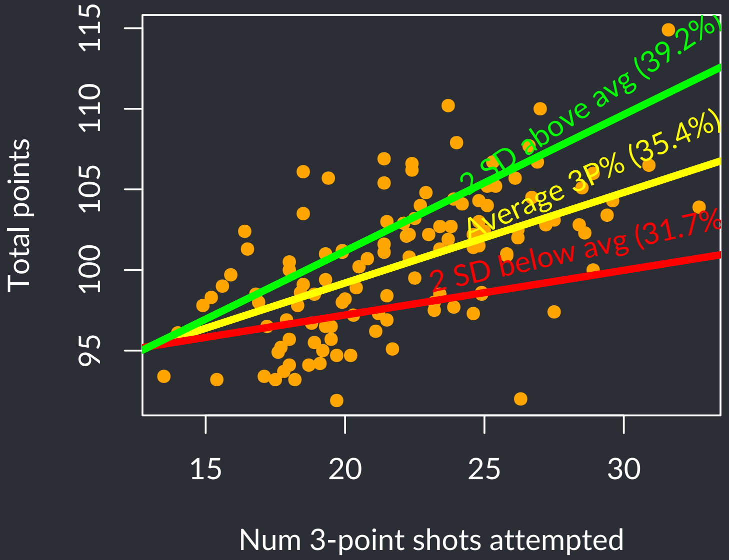
$$\widehat{\text{PTS}} = 122.85 - 2.12 \cdot \text{N3PA} - 0.98 \cdot \text{PCT3P} + 0.08 \cdot \text{N3PA} \cdot \text{PCT3P}.$$

We interpret the coefficients as follows:

- **Intercept** (122.85) is our prediction of total points when N3PA $=$ PCT3P $= 0$. (Meaningless in this context!)
- **N3PA** ($-2.12$) is the predicted increase in total points for each additional 3-pointer taken, when PCT3P $= 0$.
- **PCT3P** ($-0.98$) is the predicted increase in total points for each additional percentage point of 3-point shooting accuracy, when N3PA $= 0$.
- **N3PA · PCT3P** ($0.08$) can be interpreted in two ways:
  - the increase in the *slope coefficient* for N3PA for each 1-unit increase of PCT3P.
  - the increase in the *slope coefficient* for PCT3P for each 1-unit increase of N3PA.


Average 3P% (35.4%)
115
110
105
100
95
Total points
15
20
25
30
Num 3-point shots attempted


Average 3P% (35.4%)
2 SD below avg (31.7%)
Total points
95
100
105
110
115
15
20
25
30
Num 3-point shots attempted



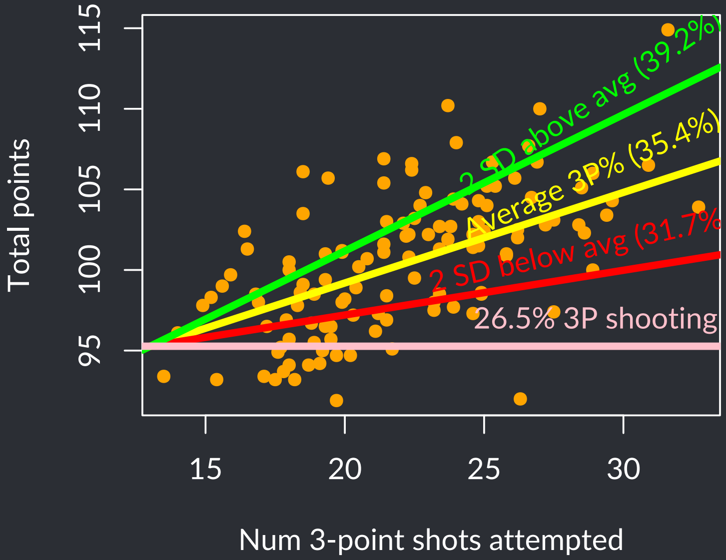
Total points
95
100
105
110
115
15
20
25
30
Num 3-point shots attempted
2 SD above avg (39.2%)
Average 3P% (35.4%)
2 SD below avg (31.7%)

$$\widehat{\text{PTS}} = 122.85 - 2.12 \cdot \text{N3PA} - 0.98 \cdot \text{PCT3P} + 0.08 \cdot \text{N3PA} \cdot \text{PCT3P}.$$

- How many points per game do you predict for a team that shoots 3-pointers at the NBA average rate (35.4) and that takes 30 3-pointers per game?

$$\widehat{\text{PTS}} = 122.85 - 2.12 \cdot \text{N3PA} - 0.98 \cdot \text{PCT3P} + 0.08 \cdot \text{N3PA} \cdot \text{PCT3P}.$$

- How many points per game do you predict for a team that shoots 3-pointers at the NBA average rate (35.4) and that takes 30 3-pointers per game?
- How bad would a team have to shoot the 3 before taking 3-point shots start to have a negative impact on total points?



Total points
115
110
105
100
95
15
20
25
30
Num 3-point shots attempted
2 SD above avg (39.2%)
Average 3P% (35.4%)
2 SD below avg (31.7%)
26.5% 3P shooting