



THE UNIVERSITY OF TEXAS AT AUSTIN  
McCOMBS SCHOOL OF BUSINESS

# Interactions 2

---

## Lecture 16

STA 371G

# Project

- It's time to start thinking about our final project!
- You will work in groups of 4—sign on Canvas by Thursday. You can create your own group or add to an existing group. Please fill partial groups before starting a new group.
- Your task will be to find or gather a regression data set, build an appropriate regression model, and write and present a report containing your findings.
- You must send me a proposal by Sunday (details on handout available on Canvas)

# Project

- Your data set must include
  - At least 100 data points.
  - At least 8 explanatory variables, at least one quantitative and one categorical.
- Your data set must not be:
  - A data set for which an existing analysis is published online.
  - A data set that is built in to R or from the R dataset package.
  - A data set that is more than 10 years old, or a data set for which more current data is readily available.

# Project

Some examples from past years:

- Predicting **NBA player points-per-game**, with predictors including player height, position, and years in the NBA.
- Predicting **GPA**, with predictors including gender, number of classes, and hours of sleep.
- Predicting **grocery expenditure**, with predictors including age, gender, amount of exercise, and income.
- Predicting **high school graduation rates**, with predictors including presence of AP program, SAT/ACT scores, and spending per capita.
- Predicting **flight prices**, with predictors including mileage, days in advance, and weekday of flight.

## NBA data

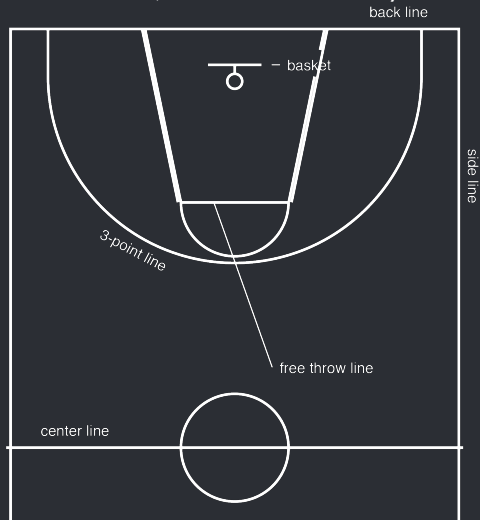
Basketball-Reference.com provides detailed data on NBA teams and players. We'll look at team data for 4 seasons ending in 2016; each of these metrics is the average across the season:

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

There are 30 NBA teams  $\times$  4 seasons = 120 cases in this file.

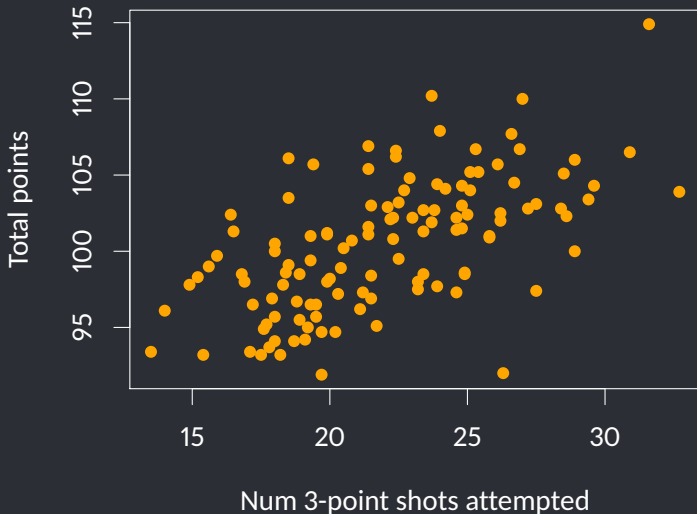
## NBA data

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

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



```
modell1 <- lm(PTS ~ N3PA, data=nba)
summary(modell1)
```

Call:

```
lm(formula = PTS ~ N3PA, data = nba)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-11.2454	-2.5114	0.0549	2.2252	8.6405

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	86.19204	1.77464	48.569	< 2e-16 ***
N3PA	0.64842	0.07935	8.171	3.89e-13 ***

---

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

Residual standard error: 3.496 on 118 degrees of freedom

Multiple R-squared: 0.3614, Adjusted R-squared: 0.356

F-statistic: 66.77 on 1 and 118 DF, p-value: 3.889e-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
	-8.3487	-2.1392	-0.0791	1.8691	9.1904

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	62.00493	5.61396	11.045	< 2e-16 ***
N3PA	0.56467	0.07587	7.442	1.82e-11 ***
PCT3P	0.73415	0.16292	4.506	1.57e-05 ***

---

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

Residual standard error: 3.241 on 117 degrees of freedom

Multiple R-squared: 0.4558, Adjusted R-squared: 0.4465

F-statistic: 49 on 2 and 117 DF, p-value: 3.478e-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!

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

Call:

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

Residuals:

	Min	1Q	Median	3Q	Max
	-7.2629	-2.2757	0.1148	1.9698	9.3756

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	122.84903	30.58937	4.016	0.000105	***
N3PA	-2.11904	1.32903	-1.594	0.113561	
PCT3P	-0.98410	0.86465	-1.138	0.257400	
N3PA:PCT3P	0.07561	0.03739	2.023	0.045423	*

---

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

Residual standard error: 3.199 on 116 degrees of freedom

Multiple R-squared: 0.4743, Adjusted R-squared: 0.4608

F-statistic: 34.89 on 3 and 116 DF, p-value: 3.798e-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  $\cdot$  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  $\cdot$  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  $\cdot$  PCT3P** (0.08) can be interpreted in two ways:
  - the increase in the *slope coefficient* for N3PA for each 1-unit increase of PCT3P.



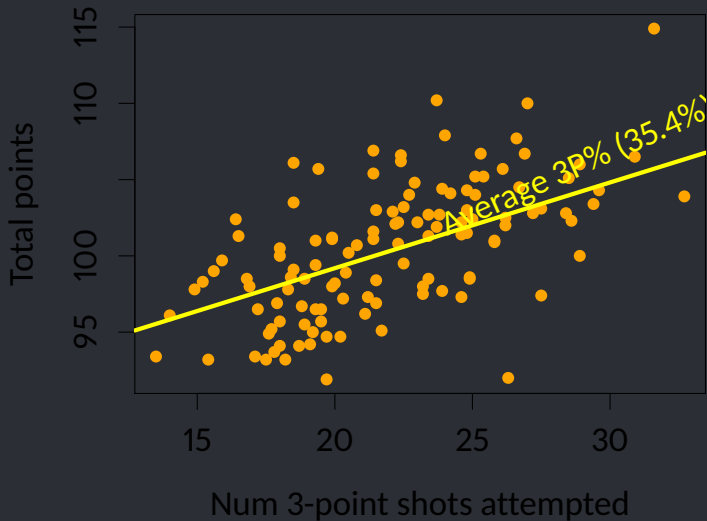
Model 3 corresponds to the regression equation

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

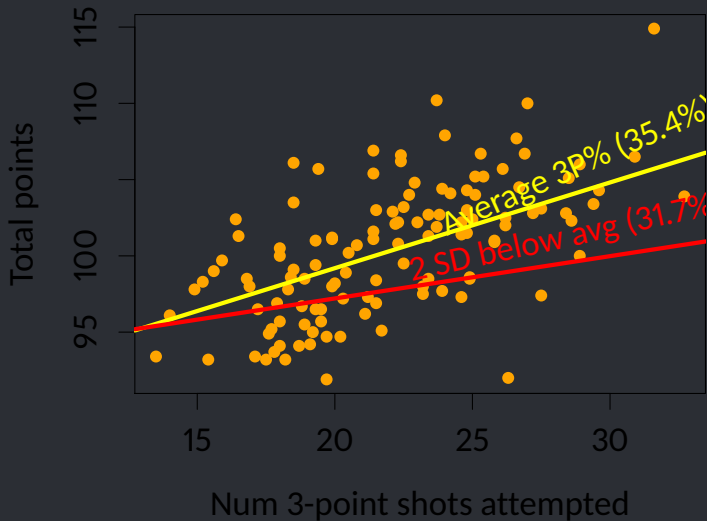
We interpret the coefficients as follows:

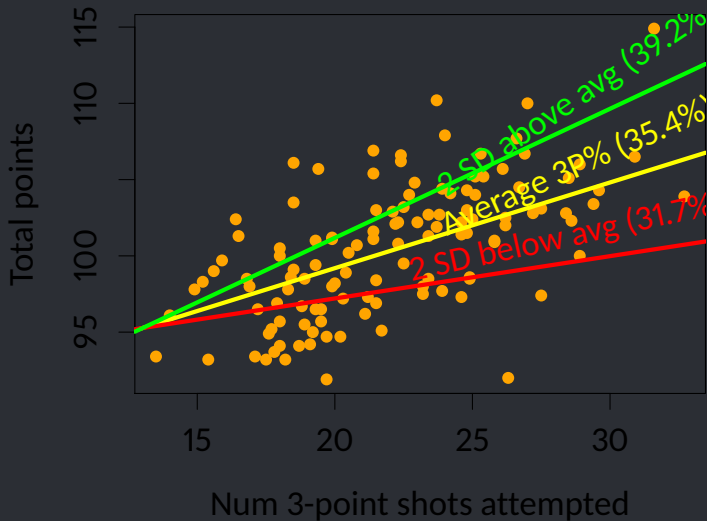
- **Intercept** (122.85) is our prediction of total points when  $\text{N3PA} = \text{PCT3P} = 0$ . (Meaningless in this context!)
- **N3PA** ( $-2.12$ ) is the predicted increase in total points for each additional 3-pointer taken, when  $\text{PCT3P} = 0$ .
- **PCT3P** ( $-0.98$ ) is the predicted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











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

- 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?

