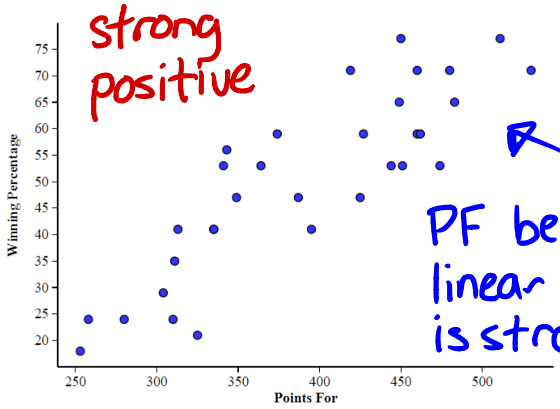
Offense or Defense?

Let's look at offensive and defensive statistics for National Football League teams from the 2021 season, shown in the table below. What variable does a better job at predicting a team's winning percentage (PCT): the number of points an offense scores (PF = points for) or the number of points a defense allows (PA = points against)?

Team	49ers	Bears	Bengals	Bills	Broncos	Browns	Buccaneers	Cardinals	Chargers	Chiefs	Colts	Cowboys	Dolphins	Eagles	Falcons	Football Team	Giants
PF	427	311	460	483	335	349	511	449	474	480	451	530	341	444	313	335	258
PA	365	407	376	289	322	371	353	366	459	364	365	358	373	385	459	434	416
PCT	59	35	59	65	41	47	77	65	53	71	53	71	53	53	41	41	24

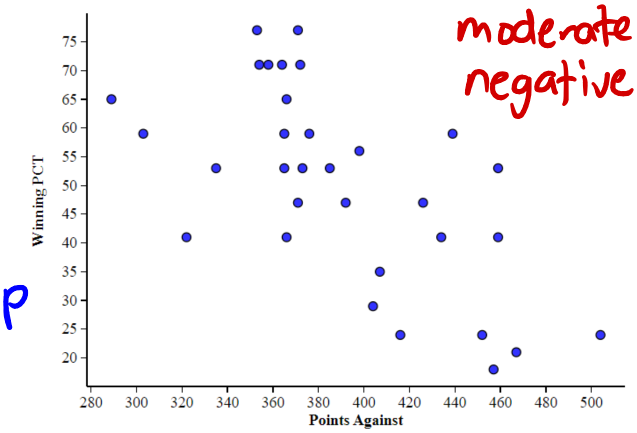
Team	Jaguars	Jets	Lions	Packers	Panthers	Patriots	Raiders	Rams	Ravens	Saints	Seahawks	Steelers	Texans	Titans	Vikings
PF	253	310	325	450	304	462	374	460	387	364	395	343	280	419	425
PA	457	504	467	371	404	303	439	372	392	335	366	398	452	354	426
PCT	18	24	21	77	29	59	59	71	47	53	41	56	24	71	47

1. The scatterplots below show the association between a team's winning percentage with either points for (PF) or points against (PA). Based on the scatterplots, which explanatory variable – PF or PA – would you guess will do a better job at predicting a team's winning percentage?



strong positive

PF because the linear relationship is stronger



moderate negative

2. On stapplet.com, select the *Multiple Regression* applet. Input PF as the first explanatory variable, PA as the second explanatory variable, and PCT as the response variable. **Be sure that the only box selected with "included in model" is PF**. Write the equation of the LSRL using PF and record the value of R^2 and S.

LSRL: $\hat{PCT} = -24.835 + 0.192(PF)$ R^2: 0.746 S: 8.722

typical prediction error

3. Using the LSRL, calculate the residual for the San Francisco 49ers, with 427 points for (PF) and a winning percentage (PCT) of 59 percent.

residual = $y - \hat{y}$

$\hat{PCT} = -24.835 + 0.192(427) = 57.149$ residual = $59 - 57.149 = 1.851$ percent

4. Go to "edit inputs" and deselect the box next to PF; select the box next to PA (now only PA is "included in model"). Write the equation of the LSRL using PA and state the value of R^2 and S.

LSRL: $\hat{PCT} = 135.548 - 0.219(PA)$ R^2: 0.411 S: 13.279

5. Using this new LSRL, calculate the residual for the San Francisco 49ers, with 365 points against (PA) and a winning percentage (PCT) of 59 percent.

$\hat{PCT} = 135.548 - 0.219(365) = 55.613$

residual = $59 - 55.613 = 3.387$ percent

Rather than using just one explanatory variable at a time, what if we used both PF and PA *in the same model*? Would this improve our predictions? Select "Edit inputs" and click both PF and PA to be included in the model. Begin analysis!

6. You should see regression output like the table to the right. Fill in the coefficient boxes, and write the equation of the multiple regression model, in the form:

Predicted PCT = Constant + (coef) PF + (coef) PA

$$\widehat{PCT} = 16.470 + 0.164(PF) - 0.077(PA)$$

Predictor	Coef	SE Coef	T	P
Constant	16.470	20.618	0.799	0.431
PF	0.164	0.023	7.002	<0.001
PA	-0.077	0.036	-2.159	0.039

7. Using this new multiple regression model, calculate the residual for the 49ers, with a winning percentage 59 percent, 427 points for, and 365 points against.

$$\widehat{PCT} = 16.470 + 0.164(427) - 0.077(365) = 58.393$$

residual = 59 - 58.393 = 0.607 percent ← lower residual

8. What was the value of R^2 and S for this multiple regression model? R^2: 0.781 ↑ S: 8.234 ↓

9. Which of the three models did the best at predicting winning percentage among these NFL teams? Explain.

Using both PF & PA.
- R^2 was highest $(0.781 > 0.746 > 0.411)$
- S was lowest $(8.234 < 8.722 < 13.279)$
- residual for 49ers lowest $(0.607 < 1.851 < 3.387)$

10. What is a variable that may increase the value of R^2 in our model? Why do you think so?

- time of possession
- presence of all-star QB
- number of turnovers
- number of sacks

11. What is a variable that would not increase the value of R^2 in our model? Why do you think so?

- color of the jersey
- age of a punter

Multiple Regression

QuickNotes

Multiple Regression - uses 2+ explanatory variables to predict a response variable

$\hat{y} = a + b_1x_1 + b_2x_2 + b_3x_3 + \ldots$

A strong model has a large r^2 and a low s.

Check Your Understanding

Here is a multiple regression model for predicting y = long jump distance (in inches) using x_1 = 40-yard dash time (in seconds) and x_2 = grade level (input 1 for junior or senior; input 0 for freshmen or sophomore) for a sample of students:

$$\hat{y} = 293.56 - 31.05x_1 + 42.02x_2$$

a) Predict the long-jump distance for a senior student who had a dash time of 5.41 seconds.

$\widehat{\text{distance}} = 293.56 - 31.05(5.41) + 42.02(1) = 167.6$ inches

b) The student in part (a) had a long jump distance of 171 inches. Calculate and interpret the residual.

residual = 171 - 167.6 = 3.4 inches.

The senior student with a dash time of 5.41 seconds jumped 3.4 inches farther than predicted.