Assignment #10: Simple Linear Regression

Four datasets are displayed to the right.

You'll notice that for each dataset:

- We have 11 observations
- The mean of X = 9
- The mean of Y = 7.5
- The std. dev. of X = 3.32
- The std. dev. of Y = 2.03
- The correlation of X and Y = 0.816.

You can download this data at:
bradthiessen.com/html5/data/anscombe.csv

	x1	y1	x2	y2	x3	y3	x4	y4
	10	8.04	10	9.14	10	7.46	8	6.58
	8	6.95	8	8.14	8	6.77	8	5.76
	13	7.58	13	8.74	13	12.74	8	7.71
	9	8.81	9	8.77	9	7.11	8	8.84
	11	8.33	11	9.26	11	7.81	8	8.47
	14	9.96	14	8.10	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.10	4	5.39	19	12.5
	12	10.84	12	9.13	12	8.15	8	5.56
	7	4.82	7	7.26	7	6.42	8	7.91
	5	5.68	5	4.74	5	5.73	8	6.89
n	11	11	11	11	11	11	11	11
Mean	9	7.5	9	7.5	9	7.5	9	7.5
Std. Dev	3.32	2.03	3.32	2.03	3.32	2.03	3.32	2.03
Correlation	0.816		0.816		0.816		0.816	

1. Let's focus on the first dataset (with x1 and y1). Using the formulas we derived in Lesson #10, calculate the slope and y-intercept for the least-squares regression line.

y1 = ____________________ (x1) + ____________________

2. Calculate and interpret R^2 for this first dataset.

R^2 = ____________________. Interpretation: ____

____

3. Using the sample size and sample standard deviation for Y1 provided in the table, calculate SSY for this first dataset.

SSY = ______________________________

4. Using the R^2 value you calculated in question #2, calculate and interpret SSE and SSreg for this first dataset.

SSE = ______________________________ SSreg = ______________________________

5. Just this one time, let's calculate SSE the conceptual (long) way. In the table below, calculate the values of y1 you would predict from the regression line you calculated in question #1. Then, calculate the squared error for each prediction. 8 of the 11 rows have been completed for you, so you can check that you're doing it correctly.

Finally, calculate the sum of those squared error terms. Verify that this value is equal to the value you calculated in the previous question.

x1	y1	Predicted Y (from regression line calculated in #1)	Squared Error = (y1 - predicted)^2
10	8.04	_________________	_________________
8	6.95	_________________	_________________
13	7.58	_________________	_________________
9	8.81	7.5	1.716
11	8.33	8.5	0.029
14	9.96	10	0.002
6	7.24	6	1.538
4	4.26	5	0.548
12	10.84	9	3.386
7	4.82	6.5	2.822
5	5.68	5.5	0.032
		Sum of squared errors =	13.763

6. Complete the following ANOVA summary table using the SS values you calculated in the previous questions. What conclusion can you make from this?

Source	**SS**	**df**	**MS**	**MSR (F)**
Regression ($b_1 \mid b_0$)	____________	________	____________	____________
Error	____________	________	____________	p = ________
Total	____________	________	MS_{total}	R^2 = ________

Conclusion:

7. Conduct a t-test to determine if the slope of your regression line (the one you calculated in question #1) significantly differs from zero (use the formula we found in question #27 in Lesson #10). Notice that this also tests if the correlation between x1 and y1 significantly differs from zero.

t =

8. Use the omnibus F-statistic to determine if x1 is a significant predictor of y1. This formula appears in question 30 in Lesson #10.

F =

9. Verify that your answer to question #8 matches the F-statistic (MSR) you calculated in the ANOVA summary table for question #6. Also, verify that the square root of this F-statistic agrees with the t-test you calculated in question #7. Note that they might not all match exactly due to rounding.

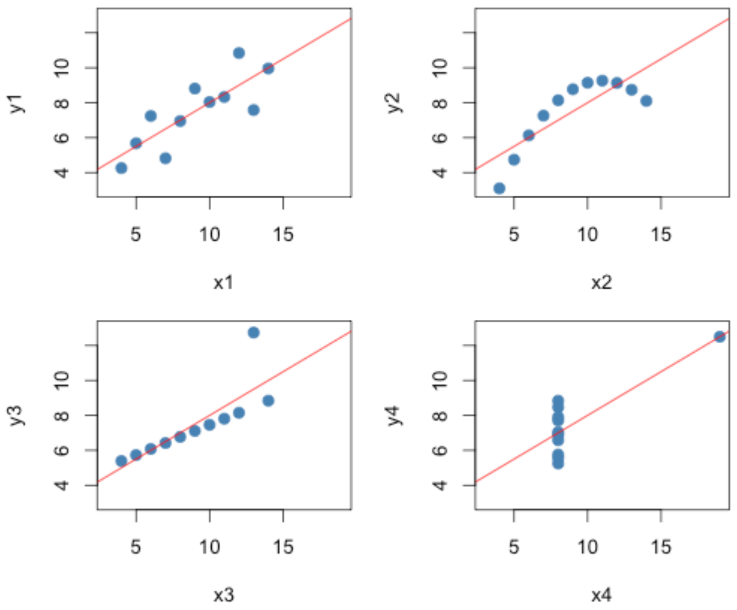
10. All four datasets on the first page of this assignment have the same means, standard deviations, and correlations. Thus, if we were to calculate regression lines for each dataset, each regression line would have the same values for the slope and y-intercept.

Furthermore, if we constructed ANOVA summary tables or calculated t-tests or omnibus F-ratios, we would get the same results for each dataset. From these, we would make the same conclusion for each dataset: x is a significant predictor of y (or, alternatively, that x and y have a linear relationship).

Now take a look to the right at scatterplots for each dataset. Would we still conclude that x and y have a linear relationship in each dataset?

This, hopefully, demonstrates the importance of plotting your data prior to running any regression analysis!

Anscombe's 4 Regression data sets



11. Now it's your turn to try to analyze data using R.
Download the template at http://www.bradthiessen.com/html5/stats/m301/yourturn10.Rmd and open it in RStudio. Then, do the following:

a) Open an **nba** data.frame using the command:

```
nba <- read.csv("http://www.bradthiessen.com/html5/data/nba.csv")
```

This dataset contains data from the 30 professional basketball teams in the NBA during the 2010-11 season.
Specifically, it shows: • **Wins** (the number of games each team won that season), and
• **PtsAgainst** (the average number of points scored per game against each team)
In a simple sense, this data shows the relationship between a team's defense and their winning percentage.

b) Find the least-squares regression line to predict wins as a function of points against. Record those coefficients here:

The least squares regression line is: Wins = ________________ (points against) + ________________

c) Interpret the slope and y-intercept you recorded above (question #11b). What does each value represent in terms of the points scored against, and the number of wins for, each team in 2010-11?

Slope: ____

Y-int: ____

d) Summarize the model to record the values for R-squared and residual standard error. Interpret those values.

R-squared = ___________ Interpretation: ____

____

Residual std. error = ___________ Interpretation: ____

____

e) Compare the full model with one predictor (points against) to a reduced model with no predictors. Fill-in the following ANOVA summary table:

Source	**SS**	**df**	**MS**	**MSR (F)**
Regression $(b_1 \mid b_0)$	____________	________	____________	____________
Error	____________	________	____________	p = ________
Total	____________	________	MS_{total}	R^2 = ________

f) Calculate the standard error of estimate ($Sy|x = \sqrt{MS_E}$). Compare that to what you recorded in question (d) and interpret.

Sy|x = ____________________

Interpretation = ____

g) Use your regression model to predict the number of wins for a team that averages 100 points against. Record that value plus a 95% confidence interval for that prediction.

Predicted number of wins = ________________. Confidence interval: ________________________

h) This time, construct a 95% prediction interval for the number of wins for a team that averages 100 points against. Record and interpret that prediction interval.

Prediction interval = __________________. Interpretation: ____________________________

____

i) Use plots to evaluate the conditions necessary for a linear regression model. Comment on whether you think the conditions are satisfied in this example.

Evaluation of the necessary conditions: ____

____

j) Using the R^2 value you recorded in question (d), calculate the omnibus F-statistic. Then, verify this F-statistic is the same as what you recorded in the ANOVA summary table in question (e).

Omnibus F-statistic:

k) Use a randomization-based methods to test the slope of the regression line. Estimate the p-value and compare it to the p-value obtained from the ANOA table in question (e).

p-value = ________________________. Comparison to p-value from ANOVA table: __________________________

l) Construct a 95% bootstrap confidence interval for the slope of our regression model. Record that interval.

Confidence interval for slope parameter = ____

m) Calculate the log-likelihood and AIC values for our full model (with one predictor). Record those values.

Log-likelihood = ________________________.   AIC = _________________________.