

Ken's Comments Based on a Light Reading of Labs #6

1. The lab asked for plots of residuals vs. each predictor in the final model. Most people included plots of leverage residuals vs. each predictor in the final model. Leverage residuals and residuals are not the same thing.

2. The lab asked for a discussion of how prediction intervals for salary based on the final model compared to actual salaries. A good idea would be to report the percentage of PI's that contained their corresponding salary. A low percentage would indicate that a model does not provide a good fit to the data. (Presumably, about 95% of 95% prediction intervals should bracket their own y 's.)

3. Correlation r is a measure of association between EXACTLY two variables. R^2 measures association between a response variable and a set (i.e., 1 or more) predictor variables. Some interpreted the data to have a "low correlation" when in fact they were talking about a multiple regression model with a low R^2 value.

4. When searching for a model, some groups began by looking at pairwise correlations r between each of the potential predictors and the response (salary or $\log(\text{salary})$ in most cases). This is a good idea. However, some groups came to the conclusion that since a predictor had a correlation with the response that was close to zero, it would not be useful in their multiple regression model. This is not true. The following data set illustrates:

X1	X2	Y
1	1	3
2	1	2
3	1	1
4	2	4
5	2	3
6	2	2
7	3	5
8	3	4
9	3	3
10	4	6
11	4	5
12	4	4

	1	2	3	4
	1	2	3	4
Y	1	2	3	4

	X1			

	*	*	*	*
	*	*	*	*
Y	*	*	*	*

	X2			

Regression Models for predicting Y based on $X1$ and $X2$

model	predictors	R^2
1	$X1$	0.0559
2	$X2$	0.0000
3	$X1$ and $X2$	1.0000

In the plot of Y vs. $X1$, the value of $X2$ is the plotting symbol. Note that although Y and $X2$ have correlation $r = 0$, it is a good idea to include both $X1$ and $X2$ in a multiple regression model to predict Y , as R^2 increases from 0.0559 to 1.0.

To assess whether or not a predictor is useful in a multiple regression model, use partial t -tests, not the pairwise correlation of that predictor with the response. Partial t -tests investigate whether a predictor is useful after accounting for all other predictors in a model. Although $X2$ is not a useful predictor of Y by itself, $X2$ is a useful predictor of Y after you account for $X1$.

5. Many were disappointed with the "low" R^2 values associated with their models. Depending which model was selected and what data points were omitted from the analysis, R^2 values for the class ranged from about 0.5 to 0.7. If for any fixed combination of the x 's, y is noisy, then R^2 will be "low." In some settings, researchers would be happy to see $R^2 = 0.2$. Once one sees that he or she isn't going to get a model with an R^2 of more than 0.7, he or she should look for models with an R^2 close to 0.7 that satisfies regression assumptions (i.e., ones with residual plots that look as expected).

6. Very few calculated the C_p or $PRESS$ statistics to assess model fit. To provide some kind of justification for a chosen model

- calculate C_p and note that it is in fact close to p as it is for a "good" model, and
- calculate $PRESS$ for the candidate model and note that it is not too much larger than SSE , as it should be if the model is "good."