

Lecture 15: Collinearity in multiple regression

Stat 102

- **Collinearity**

is a disease that infects many multiple regression data sets.

- With proper care the worst symptoms of the disease can usually be avoided or lessened.
- But sometimes there is no desirable cure.
- We'll use our **Baseball salaries** data to illustrate the symptoms of the disease.
- Then we'll discuss some of the features of the disease and how to cure it, when that's possible.

Baseball Salaries '87

- **Data** is each ML player's salary for 1987, along with performance indicators for 1986 and for their entire career (including Years played until 1987).
- A possible goal in analyzing this data is to use performance indicators to predict salary
 - In the next lecture we'll pursue that goal in detail
- For now, we'll just see what can happen when we build a particular multiple regression model containing 3 predictive factors.

- Previous investigation showed (or should have shown)
 - $Y = \text{Log}(\text{Salary})$ is a better choice than $Y = \text{Salary}$.
 - Performance indicators over the entire career such as **$\text{Runs}_{\text{cr}} = \text{Runs}_{\text{cr}}/\text{yr} \times \text{Years}$** seem to work better in simple regressions than do the per-yr variables. We'll use the full career versions in the following.
 - The data for **Mike Schmidt** and **Terry Kennedy** is mis-recorded. They should be excluded.
 - Baseball aficionados will note that **Pete Rose** was a player/manager. For that reason we'll exclude him too. *[Optional choice]*
 - Most of the simple regressions worked better with Logged x -variables. For the following analyses we'll use that form without further comment;

except that in the next lecture we'll use "putouts '86" and "Years" without taking their Log.

{Log(Putouts'86) is awkward because some players have 0. Previous research with other types of salaries suggests "Years" may work as well in a big model as "Log(Years)", even though in our data Log(Years) turns out to be clearly preferable in simple regressions of Y on X .}

Correlations with Log(Salary)

- The Correlation Matrix is a convenient tool for finding the best candidate(s) for a simple regression predictor of Log(Salary). Here's its first row, relative to three possible predictors of interest:

	LogSal	Logrns,cr	Logrbi,cr	Loghits,cr
LogSal	1.000	0.823	0.813	0.824

- The best single predictor here is Loghits,cr (*by a whisker*) but it's in a virtual tie with Logrns,cr, & Logrbi,cr, [*& also with Logab,cr if we had included that variable*].
- For LogSal on Loghits,cr we have $R^2 = 0.677$. [*With Pete Rose one still gets for LogSal on Loghits: $R^2 = 0.676$.*]

Suppose for now that –

[in order to improve the prediction and still have a fairly simple model]

We want (at most) a **3-factor** model

- We might look at the correlation matrix for all variables and choose the three factors: **Loghits,cr**, **Logrns,cr**, & **Logrbi,cr**.
- We get $R^2 = 0.685$. This is a modest improvement over the previous 1-factor value of $R^2 = 0.677$.
- Whenever you add more factors you must increase R^2 . [**Why?**]
Hence the increase from 0.677 to 0.685 really isn't very much.
- Here's the 3-factor ANOVA Table:

3-Factor ANOVA Table

Summary of Fit

RSquare	0.6853
Root Mean Square Error	0.499
Observations	260

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	138.99	46.33	185.8
Error	256	63.82	0.2493	Prob > F
C. Total	259	202.81		<.0001

- Note that this 3-Factor model is statistically highly significant, with F-Ratio = 185.8 and P-value $\ll .0001$, as well as $R^2 = .685$
- Now, here's the 3-factor Parameter Estimates Table

3-Factor Parameter Estimates Table

Parameter Estimates				
Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	1.943	0.258	7.53	<.0001
Logrns,cr	0.277	0.198	1.40	0.1618
Loghits,cr	0.245	0.220	1.11	0.2664
Logrbi,cr	0.181	0.113	1.61	0.1088

- **None** of the 3 factors is statistically significant at 0.05.
- This is a curious effect of **Collinearity**.
- **What does it mean?**
- **Why does it occur?**
- **What can we do to cure it?**

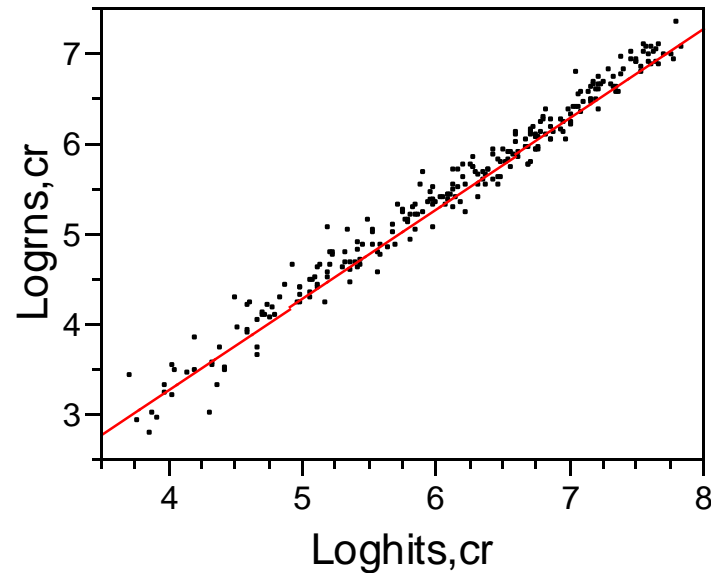
None of the 3 factors is statistically significant at 0.05.

- **What does this mean?**
- **It doesn't say** that “*none of the 3-factors is useful*”.
- **It does say:**
- **After you have two of the 3 factors**, adding the third factor does not provide a statistically significant improvement (at level $\alpha = 0.05$). (*Actually, in this data even more is true: After you have any one of the factors, the other two aren't of very much use.*)
- **ALSO**, the individual coefficient estimates aren't very stable.
- **SO**, although the prediction provided by the overall 3-Factor linear equation is quite good,
- **BUT** the individual coefficient values aren't very trustworthy.

What are the conditions that allow this to occur?

- A *necessary* condition is that one of the predictors is nearly a linear function of the others.
 - ie, $x_K \approx \beta'_0 + \beta'_1 x_1 + \dots + \beta'_{K-1} x_{K-1}$.
- This is called **Collinearity** of the predictor variables.
- In particular, if one of the x_i is nearly a linear function of another one, there is collinearity. That is the case here. For example, here is the plot of Logruns,cr on Loghits,cr:

Logrns,cr on Loghits,cr



- Here $R^2 = .976!$ These two variables are **extremely** collinear.
- Even though these variables are individually significant (as predictors of $\text{Log}(\text{sal})$), because of collinearity we *may* find only one – or none – of these variables to have significant coefficients in a multiple regression that involves both of them.

- In summary, the “**disease**” is that when we put together predictor variables that are individually significant they may turn out to not be significant in the multiple regression (or to only be weakly significant).
- Collinearity is a necessary contributing factor to this “disease”.
- It won’t happen unless this factor is collinear with some other factors that are included in the multiple regression.
- Even if you have collinear variables this problem may not occur.
- With just logruns and loghits, we have a moderate case of the disease:

Parameter Estimates				
Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	1.777	0.237	7.49	<.0001
Logrns,cr	0.315	0.197	1.60	0.111
Loghits,cr	0.396	0.200	1.98	0.049

What is the “cure”?

If collinearity causes this problem, what can we do about it?

A. In your final model don't include as factors two or more variables that are highly collinear. Thus, if $\log_{\text{hits,cr}}$ are in your model, don't include $\log_{\text{rns,cr}}$ or $\log_{\text{rbi,cr}}$.

- But doing this may fail to tell the whole story, and may not produce as good a prediction or prediction CI.

B. Gather additional data that isn't so collinear.

- We don't have that option here. Even with more years of data hits and runs will be highly collinear,

- We can't conduct an experiment to find out how much a player with many hits and few runs scored would be payed. etc.

Summary of 3 Factor Analysis

- We investigated 3 factors – Loghits,cr, Logrns,cr and Logrbi,cr – as possible predictors of Log(Salary).
- Each one individually is a good (linear) predictor; each one has an R^2 with Log(Salary) of about 0.67.
- Every pair of these predictors is **very** highly collinear. (For each pair $R^2 > 0.96$.)
- Among the 3 predictors Loghits,cr is very slightly the best individual predictor of Log(Salary). But the difference is very small and not statistically significant. [There is a formal statistical test to investigate such a difference in R^2 values, but we won't pursue that here.]
- All 3 predictors together do not do a much better job than any single one of them. For example, a test in the 3-factor model of $H_0: \beta_2 = \beta_3 = 0$ is as follows: For the ordinary regression using Loghits,cr we have $SSE=65.107$; for the 3-factor model $SSE=63.822$ and $MSE=.249$. Hence the statistic for testing H_0 is $F = [(65.107 - 63.822) / 2] \div 0.249 = 2.58$. This has P-value = 0.078 (from F with 2&256 DF).

Summary of “Collinearity”

- Collinearity describes a situation where (at least) one of the predictive factors is close to being a linear combination of (some of) the others.
- The simplest case is where the R^2 of a pair of factors is near 1.
- When collinearity is present it may happen that one factor is significant by itself, but is not significant in combination with the other factors.
- If this happens you can do one of the following
 - A. Leave some factors out of your analysis.
 - B. Try to collect more (and less collinear) data.
 - C. Keep all factors in the larger analysis, but explain that individually they are important even though in combination it may not be significant **after controlling for the others.**