Multicollinearity in Regression

Paul E. Johnson[1] [2]

[1]Department of Political Science

[2]Center for Research Methods and Data Analysis, University of Kansas

2014

Multicollinearity

- There's a small R program persp-multicollinearity-1.R. It is available with this document and can be used to experiment with 3-D illustrations.

Outline

What is Multi Collinearity?

- Definition: Multicollineary exists if it is possible to calculate the value of one IV using a linear formula that combines the other IV's.
- Note. It is not the same as "bivariate correlation" among x's.
  The word has "multi" for a reason! It is not *bi-collinearity*!
- Pearson correlation matrix not best way to check for multicollinearity.

Perfect Multicollinearity

- Example: If a model has IV's $X1_i$ and $X2_i$ and $X3_i$, perfect MC would exist if one could find constants k_0, k_1 and k_2 such that

$$X3_i = k_0 + k_1 X1_i + k_2 X2_i$$

- If you put the same variable in a regression with two different names, what do you get? The estimation process for the model should crash and complain to you. Perfect multicollinearity!
- Silly mistakes: Put in "gender" (=Male Female) and "sex" (=Man Woman) and "biologically capable of giving birth" (=Yes No) in the same model.
- Those variables cannot be differentiated from one another.

(ImPerfect?) Multicollinearity

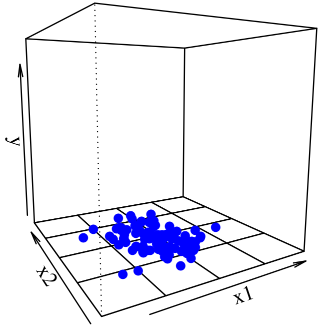
- If you put variables in a model that are similar, but not identical, then you have multicollinearity.
- It is not "perfect" (perfectly bad), but it is still (imperfectly) bad.
- In a model with 10 IV, that means it is possible to predict (at least) one IV from others with some weighted formula

$$X10_i = k_0 + k_1 X1_i + k_2 X2_i + k_3 X3_i + \ldots + k_9 X9_i$$

- If $R^2_{x10.x1\ldots x9}$ is high, then $x10_i$ has "not much separate variation", it cannot be distinguished from the others. AND the others cannot be distinguished from it.

Illustration in 3 Dimensions

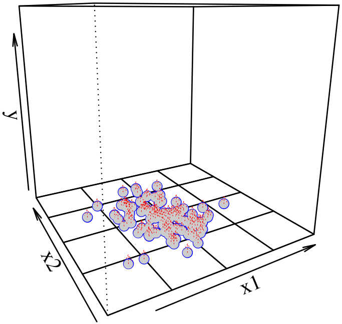
- No values drawn yet for dependent variable
- Please notice dispersion in the x1-x2 plane



Watch Data Cloud Develop

- The true relationship is

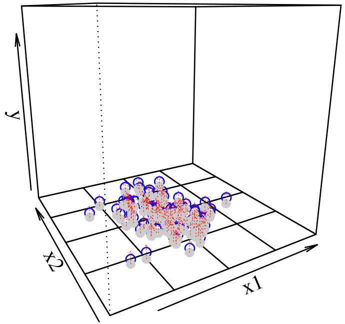
$$y_i = .2\,x1_i + .2\,x2_i + e_i,\ e_i \sim N(0, 7^2)$$



Watch Data Cloud Develop

- The true relationship is

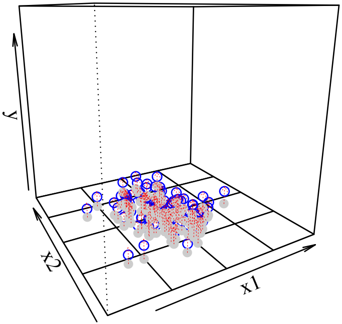
$$y_i = .2\,x1_i + .2\,x2_i + e_i,\; e_i \sim N(0, 7^2)$$



Watch Data Cloud Develop

- The true relationship is

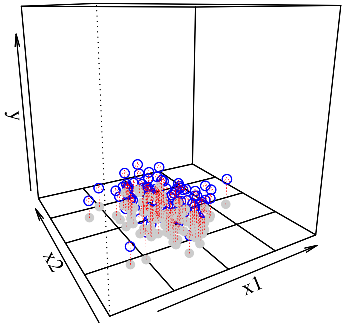
$$y_i = .2\,x1_i + .2\,x2_i + e_i,\ e_i \sim N(0, 7^2)$$



Watch Data Cloud Develop

- The true relationship is

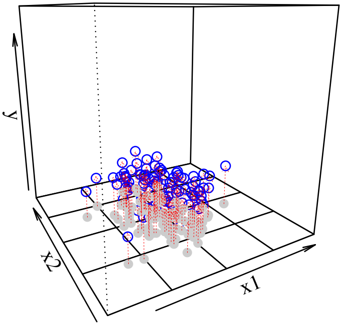
  $$y_i = .2\,x1_i + .2\,x2_i + e_i,\ e_i \sim N(0, 7^2)$$



Watch Data Cloud Develop

- The true relationship is

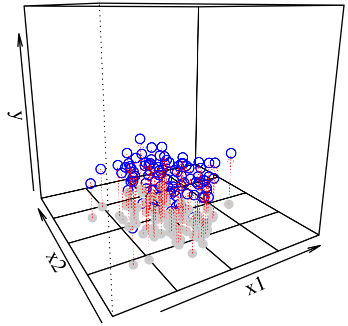
$$y_i = .2\,x1_i + .2\,x2_i + e_i,\ e_i \sim N(0, 7^2)$$



Watch Data Cloud Develop

- The true relationship is

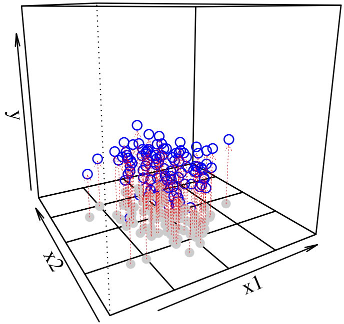
$$y_i = .2\,x1_i + .2\,x2_i + e_i,\ e_i \sim N(0, 7^2)$$



Watch Data Cloud Develop

- The true relationship is

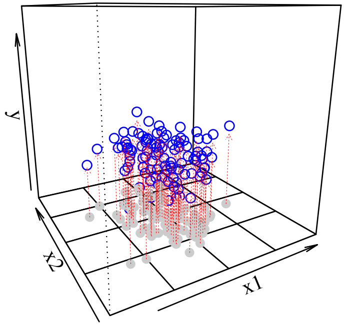
$$y_i = .2\,x1_i + .2\,x2_i + e_i,\ e_i \sim N(0, 7^2)$$



Watch Data Cloud Develop

- The true relationship
  is

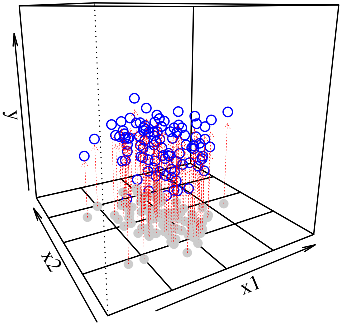
  $$y_i = .2\,x1_i + .2\,x2_i + e_i,\ e_i \sim N(0, 7^2)$$



Watch Data Cloud Develop

- The true relationship is

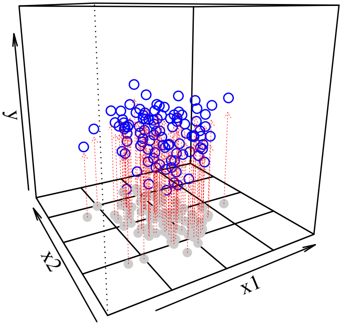
$$y_i = .2\,x1_i + .2\,x2_i + e_i,\ e_i \sim N(0, 7^2)$$



Watch Data Cloud Develop

- The true relationship is

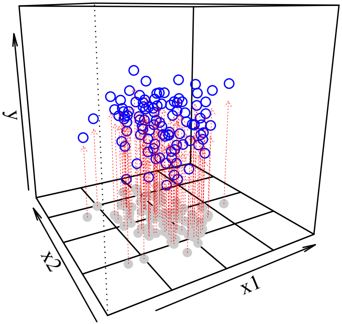
$$y_i = .2\,x1_i + .2\,x2_i + e_i,\ e_i \sim N(0, 7^2)$$



Watch Data Cloud Develop

- The true relationship is

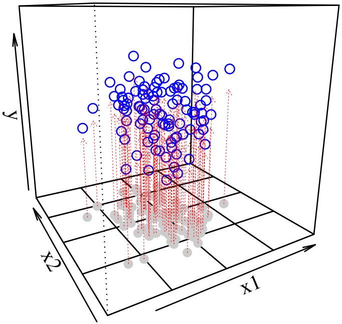
  $$y_i = .2\,x1_i + .2\,x2_i + e_i,\ e_i \sim N(0, 7^2)$$



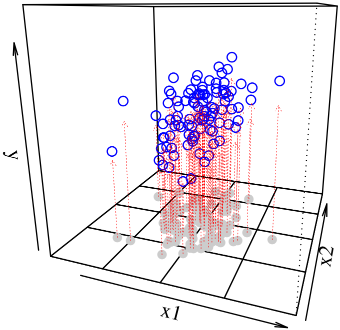
Watch Data Cloud Develop

- The true relationship is

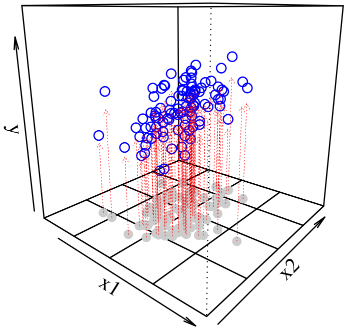
$$y_i = .2\,x1_i + .2\,x2_i + e_i,\ e_i \sim N(0, 7^2)$$



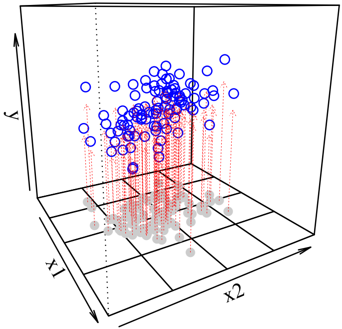
Can Spin the Cloud (Just Showing Off)



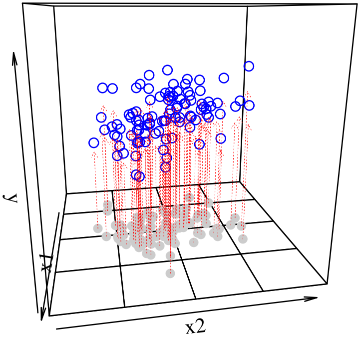
Can Spin the Cloud (Just Showing Off)



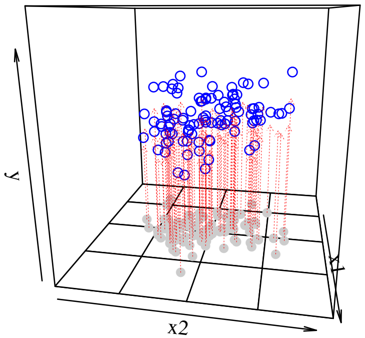
Can Spin the Cloud (Just Showing Off)



Can Spin the Cloud (Just Showing Off)



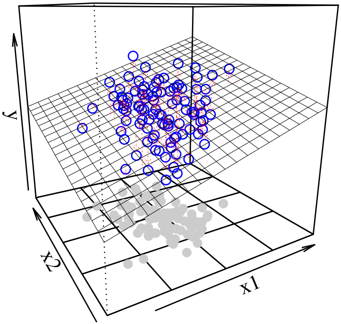
Can Spin the Cloud (Just Showing Off)



Regression Plane Sits Nicely in the Data Cloud

	M1	
	Estimate	(S.E.)
(Intercept)	-0.522	(4.037)
x1	0.193***	(0.051)
x2	0.216***	(0.057)
N	100	
RMSE	5.661	
R^2	0.212	
adj R^2	0.196	

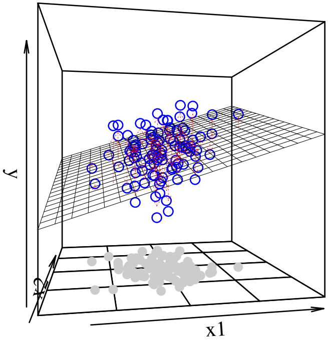
$*p \leq 0.05 ** \ p \leq 0.01 *** p \leq 0.001$



Regression Plane Sits Nicely in the Data Cloud

	M1	
	Estimate	(S.E.)
(Intercept)	-0.522	(4.037)
x1	0.193***	(0.051)
x2	0.216***	(0.057)
N	100	
RMSE	5.661	
R^2	0.212	
adj R^2	0.196	

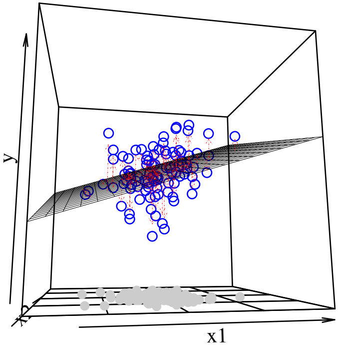
$*p \leq 0.05 ** \ p \leq 0.01 *** p \leq 0.001$



Regression Plane Sits Nicely in the Data Cloud

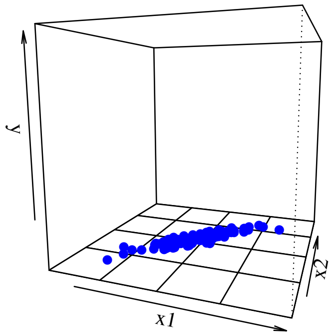
	M1	
	Estimate	(S.E.)
(Intercept)	-0.522	(4.037)
x1	0.193***	(0.051)
x2	0.216***	(0.057)
N	100	
RMSE	5.661	
R^2	0.212	
adj R^2	0.196	

$*p \le 0.05 ** \ p \le 0.01 *** p \le 0.001$

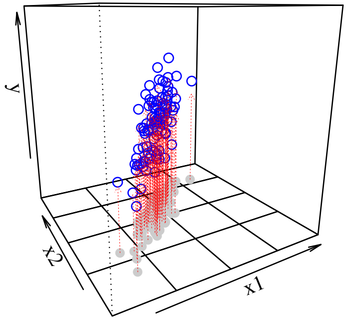


Severe Collinearity: r(x1,x2)=0.9

- Nearly linear dispersion in the x1-x2 plane



Cloud Is More like Data Tube

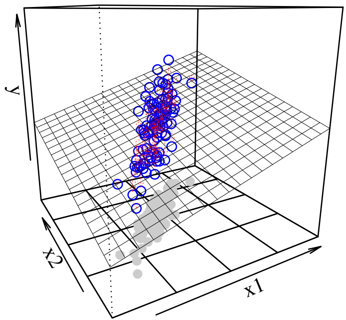


Cloud Is More like Data Tube

	M1	
	Estimate	(S.E.)
(Intercept)	-4.120	(3.054)
x1	0.267	(0.157)
x2	0.207	(0.153)
N	100	
RMSE	6.827	
R^2	0.394	
adj R^2	0.381	

$*p \le 0.05 ** \ p \le 0.01 *** p \le 0.001$

- plane does not sit as "comfortably"
- greater standard errors



Outline

Symptom #1: Weak Hypothesis Tests.

- MC inflates the variance $Var(\hat{b}_1)$ and the estimated variance, $\widehat{Var(\hat{\beta}_1)}$, and its square root, the $std.err.(\hat{\beta}_1)$.

Suppose $H_0 : b = 0$.

- MC makes t-statistics smaller, since $t = \frac{\widehat{b}}{std.err(\widehat{b})}$
- Find a book that gives the formula for the $s.e.(\hat{\beta})$ for a model with a few independent variables. It should be easy to see that as the variables become more similar, then the $s.e(\hat{\beta})$ gets bigger.

Warning Sign: Mismatch of F and t tests

- Standard output: "no significant(ly different from zero) t statistics"
- But the F statisticis is significant, or there is a really big R^2

You Did Not Do Something Wrong!

- Suppose Nature used this formula:

$$y_i = 1.1 + 4.4\,x1_i + 2.1\,x2_i + e_i,\;\; e_i \sim N(0, \sigma_e^2) \tag{1}$$

- You estimate the correct formula, with the right variables:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x1_i + \hat{\beta}_2 x2_i \tag{2}$$

- If it "just so happens" that high levels of $x1_i$ are observed in the same cases in which we also observe $x2_i$, then we have trouble estimating β_1 and β_2.
- The "other things equal" theory cannot be explored with this data.
- Do they love you because you are beautiful? Or because you are clever? Or modest? Or because you are a good listener?

Multicollinearity Causes High Variance in Slope Estimates

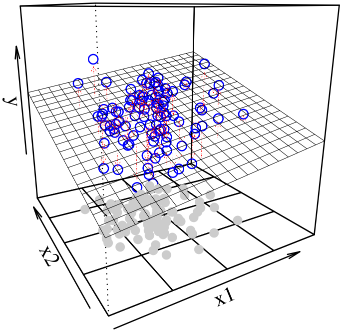
- Here's my demonstration plan.
- First, I'll draw data from not-correlated independent variables
  - fit regressions (remember the "true" slopes are 0.2
  - draw the 3d regression planes
- I'll do that, say, 20 times. In class, I might run the script that does this 500 times so we can "really see" it. But that would make this PDF too large.
- After that, I will repeat the process, but with data that is multi-correlated. If the demonstration works properly, the reader should see that the fitted models are more stable when there is no collinearity than when there is collinearity.

Regression with Uncorrelated Predictors

	M1	
	Estimate	(S.E.)
(Intercept)	-3.901	(4.321)
x1	0.159*	(0.065)
x2	0.297***	(0.065)
N	100	
RMSE	6.564	
R^2	0.237	
adj R^2	0.221	

$*p \leq 0.05 ** \ p \leq 0.01 *** p \leq 0.001$

- $r_{x1,x2} =$ 0.0
- Sample 1

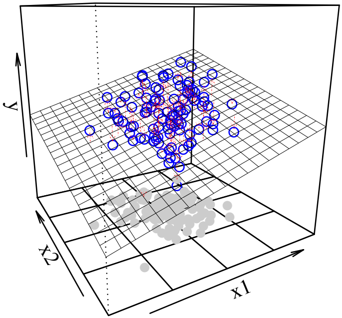


Regression with Uncorrelated Predictors

	M1	
	Estimate	(S.E.)
(Intercept)	-4.550	(4.741)
x1	0.267***	(0.066)
x2	0.220***	(0.063)
N	100	
RMSE	6.821	
R^2	0.218	
adj R^2	0.202	

$*p \le 0.05 ** p \le 0.01 *** p \le 0.001$

- $r_{x1,x2} =$ 0.0
- Sample 2

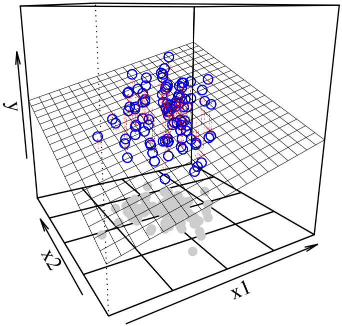


Regression with Uncorrelated Predictors

	M1	
	Estimate	(S.E.)
(Intercept)	-1.803	(4.998)
x1	0.170**	(0.064)
x2	0.260***	(0.068)
N	100	
RMSE	7.228	
R^2	0.174	
adj R^2	0.157	

$*p \leq 0.05 ** \; p \leq 0.01 *** p \leq 0.001$

- $r_{x1,x2} =$ 0.0
- Sample 3

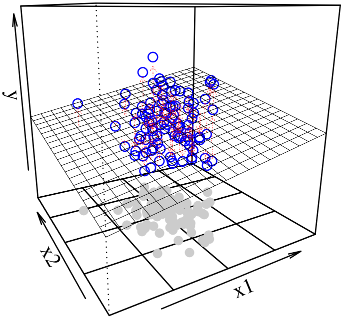


Regression with Uncorrelated Predictors

	M1	
	Estimate	(S.E.)
(Intercept)	6.996	(4.592)
x1	0.133	(0.072)
x2	0.122*	(0.060)
N	100	
RMSE	6.616	
R^2	0.073	
adj R^2	0.054	

$*p \leq 0.05 ** \ p \leq 0.01 *** p \leq 0.001$

- $r_{x1,x2} = 0.0$
- Sample 4

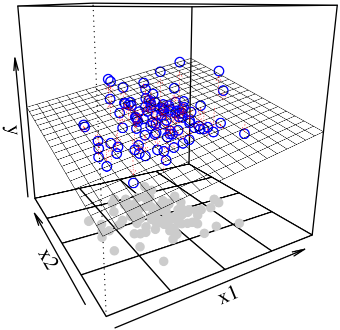


Regression with Uncorrelated Predictors

	M1	
	Estimate	(S.E.)
(Intercept)	-2.752	(4.794)
x1	0.227***	(0.066)
x2	0.211***	(0.061)
N	100	
RMSE	6.151	
R^2	0.184	
adj R^2	0.167	

$*p \leq 0.05 ** \ p \leq 0.01 *** p \leq 0.001$

- $r_{x1,x2} =$ 0.0
- Sample 5

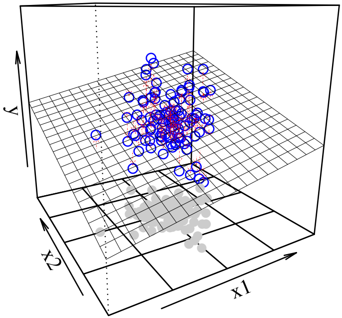


Regression with Uncorrelated Predictors

	M1	
	Estimate	(S.E.)
(Intercept)	1.109	(4.501)
x1	0.155*	(0.069)
x2	0.222***	(0.061)
N	100	
RMSE	6.328	
R^2	0.169	
adj R^2	0.152	

$*p \leq 0.05 ** \; p \leq 0.01 *** p \leq 0.001$

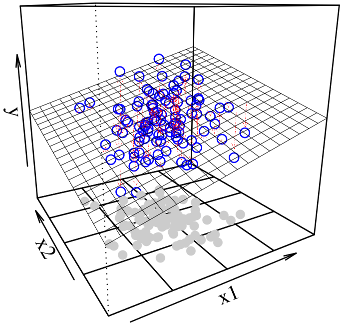
- $r_{x1,x2} =$ 0.0
- Sample 6



Regression with Uncorrelated Predictors

	M1	
	Estimate	(S.E.)
(Intercept)	-3.312	(5.076)
x1	0.250***	(0.064)
x2	0.239**	(0.072)
N	100	
RMSE	6.290	
R^2	0.203	
adj R^2	0.186	

$*p \leq 0.05 ** \ p \leq 0.01 *** p \leq 0.001$



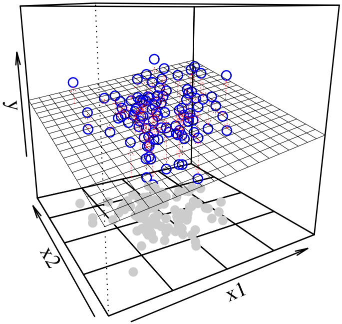
- $r_{x1,x2} =$ 0.0
- Sample 7

Regression with Uncorrelated Predictors

	M1	
	Estimate	(S.E.)
(Intercept)	0.841	(5.115)
x1	0.148*	(0.072)
x2	0.238**	(0.076)
N	100	
RMSE	7.230	
R^2	0.134	
adj R^2	0.116	

$*p \leq 0.05 ** \; p \leq 0.01 *** p \leq 0.001$

- $r_{x1,x2} =$ 0.0
- Sample 8

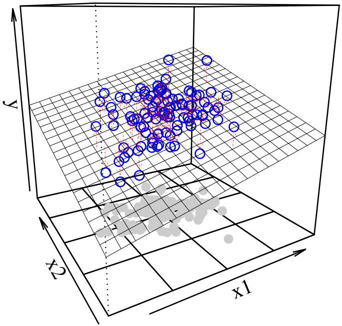


Regression with Uncorrelated Predictors

	M1	
	Estimate	(S.E.)
(Intercept)	-6.454	(4.913)
x1	0.268***	(0.062)
x2	0.252***	(0.072)
N	100	
RMSE	7.014	
R^2	0.230	
adj R^2	0.215	

$*p \leq 0.05 ** \ p \leq 0.01 *** p \leq 0.001$

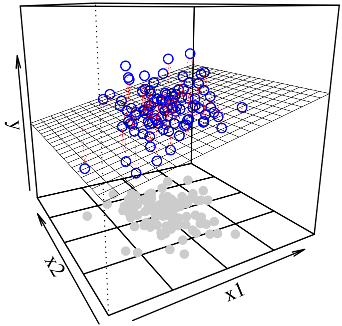
- $r_{x1,x2} =$ 0.0
- Sample 9



Regression with Uncorrelated Predictors

	M1	
	Estimate	(S.E.)
(Intercept)	7.048	(5.563)
x1	0.270***	(0.078)
x2	0.007	(0.082)
N	100	
RMSE	7.222	
R^2	0.112	
adj R^2	0.094	

$*p \leq 0.05 ** \ p \leq 0.01 *** p \leq 0.001$



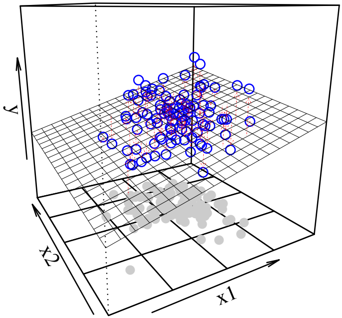
- $r_{x1,x2} =$ 0.0
- Sample 10

Regression with Uncorrelated Predictors

	M1	
	Estimate	(S.E.)
(Intercept)	-2.513	(5.690)
x1	0.311***	(0.080)
x2	0.133	(0.073)
N	100	
RMSE	6.925	
R^2	0.152	
adj R^2	0.135	

$*p \le 0.05 ** \ p \le 0.01 *** p \le 0.001$

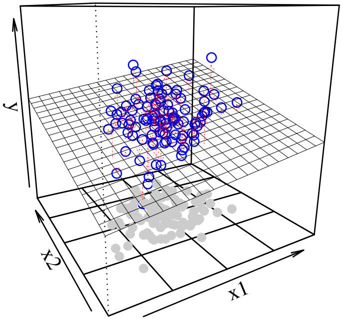
- $r_{x1,x2} =$ 0.0
- Sample 11



Regression with Uncorrelated Predictors

	M1	
	Estimate	(S.E.)
(Intercept)	-4.626	(4.953)
x1	0.168*	(0.074)
x2	0.316***	(0.086)
N	100	
RMSE	7.753	
R^2	0.201	
adj R^2	0.185	

$*p \leq 0.05 ** \ p \leq 0.01 *** p \leq 0.001$

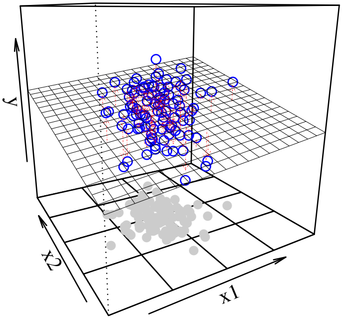


- $r_{x1,x2} =$ 0.0
- Sample 12

Regression with Uncorrelated Predictors

	M1	
	Estimate	(S.E.)
(Intercept)	8.310	(4.354)
x1	0.076	(0.064)
x2	0.165**	(0.059)
N	100	
RMSE	6.101	
R^2	0.088	
adj R^2	0.069	

$*p \le 0.05 ** \ p \le 0.01 *** p \le 0.001$



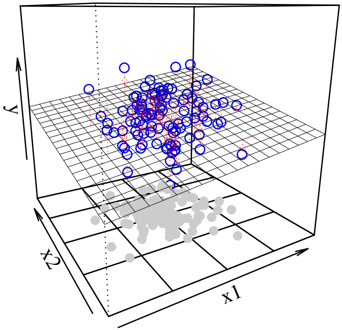
- $r_{x1,x2} =$ 0.0
- Sample 13

Regression with Uncorrelated Predictors

	M1	
	Estimate	(S.E.)
(Intercept)	3.447	(5.601)
x1	0.154	(0.078)
x2	0.183*	(0.074)
N	100	
RMSE	7.032	
R^2	0.093	
adj R^2	0.075	

$*p \le 0.05 ** \ p \le 0.01 *** p \le 0.001$

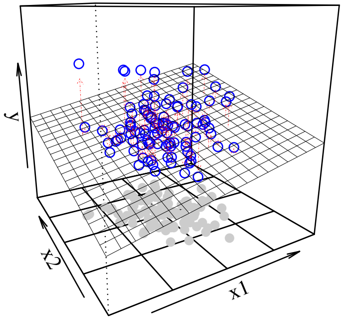
- $r_{x1,x2} =$ 0.0
- Sample 14



Regression with Uncorrelated Predictors

	M1	
	Estimate	(S.E.)
(Intercept)	-3.310	(6.130)
x1	0.185**	(0.068)
x2	0.279**	(0.089)
N	100	
RMSE	7.402	
R^2	0.133	
adj R^2	0.115	

$*p \leq 0.05 ** \ p \leq 0.01 *** p \leq 0.001$



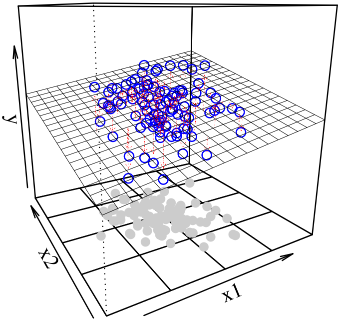
- $r_{x1,x2} =$ 0.0
- Sample 15

Regression with Uncorrelated Predictors

	M1	
	Estimate	(S.E.)
(Intercept)	-0.763	(5.980)
x1	0.160*	(0.072)
x2	0.249**	(0.084)
N	100	
RMSE	7.344	
R^2	0.108	
adj R^2	0.089	

$*p \leq 0.05 ** \ p \leq 0.01 *** p \leq 0.001$

- $r_{x1,x2} =$ 0.0
- Sample 16

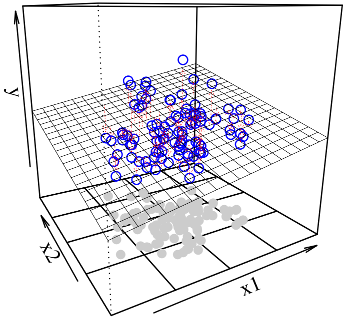


Regression with Uncorrelated Predictors

	M1	
	Estimate	(S.E.)
(Intercept)	-1.623	(5.331)
x1	0.219**	(0.071)
x2	0.223**	(0.071)
N	100	
RMSE	6.586	
R^2	0.159	
adj R^2	0.141	

$*p \leq 0.05 ** \ p \leq 0.01 *** p \leq 0.001$

- $r_{x1,x2} =$ 0.0
- Sample 17

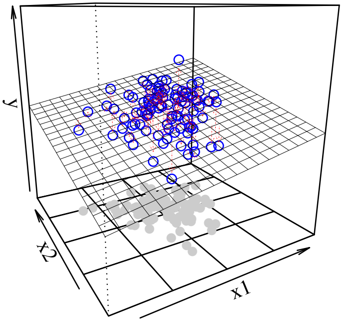


Regression with Uncorrelated Predictors

	M1	
	Estimate	(S.E.)
(Intercept)	-1.453	(4.921)
x1	0.213**	(0.071)
x2	0.212***	(0.060)
N	100	
RMSE	6.451	
R^2	0.171	
adj R^2	0.154	

$*p \leq 0.05 ** \ p \leq 0.01 *** p \leq 0.001$

- $r_{x1,x2} =$ 0.0
- Sample 18

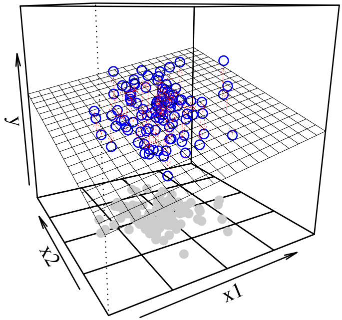


Regression with Uncorrelated Predictors

	M1	
	Estimate	(S.E.)
(Intercept)	-2.770	(5.263)
x1	0.202*	(0.077)
x2	0.245***	(0.067)
N	100	
RMSE	7.631	
R^2	0.165	
adj R^2	0.148	

$*p \leq 0.05 ** \ p \leq 0.01 *** p \leq 0.001$

- $r_{x1,x2} =$ 0.0
- Sample 19

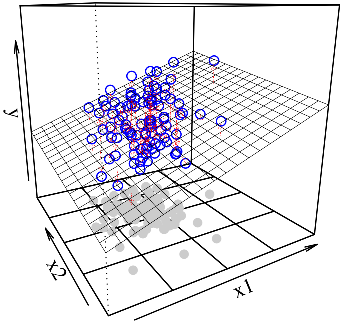


Regression with Uncorrelated Predictors

	M1	
	Estimate	(S.E.)
(Intercept)	3.445	(4.171)
x1	0.223***	(0.058)
x2	0.109	(0.063)
N	100	
RMSE	5.764	
R^2	0.164	
adj R^2	0.147	

$*p \leq 0.05 ** \ p \leq 0.01 *** p \leq 0.001$

- $r_{x1,x2} = 0.0$
- Sample 20



Change Gears Now!

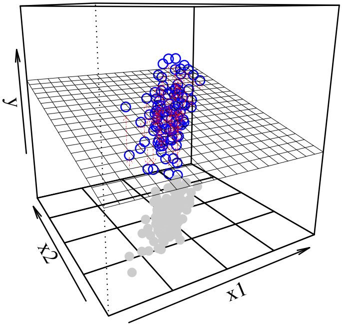
- You should see that the planes are more-or-less “the same”, the sampling process is not causing wild fluctuations.
- Those had no correlation between $x1_i$ and $x2_i$.
- Now, repeat with high correlation between those variables. The plane will wobble a lot more.

High Collinearity Causes Unstable Estimates

	M1	
	Estimate	(S.E.)
(Intercept)	4.016	(4.532)
x1	0.001	(0.174)
x2	0.303	(0.181)
N	100	
RMSE	7.369	
R^2	0.116	
adj R^2	0.098	

$*p \leq 0.05 ** \; p \leq 0.01 *** p \leq 0.001$

- $r_{x1,x2} = 0.9$
- Sample 1

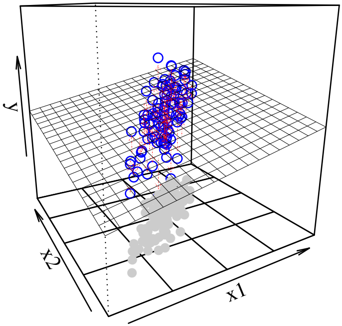


High Collinearity Causes Unstable Estimates

	M1	
	Estimate	(S.E.)
(Intercept)	-1.779	(3.806)
x1	0.233	(0.163)
x2	0.201	(0.150)
N	100	
RMSE	7.017	
R^2	0.271	
adj R^2	0.256	

$*p \leq 0.05** \; p \leq 0.01***p \leq 0.001$

- $r_{x1,x2} =$ 0.9
- Sample 2

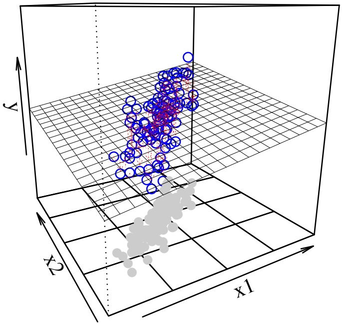


High Collinearity Causes Unstable Estimates

	M1	
	Estimate	(S.E.)
(Intercept)	-0.032	(4.362)
x1	0.232	(0.187)
x2	0.165	(0.175)
N	100	
RMSE	7.262	
R^2	0.188	
adj R^2	0.172	

$*p \leq 0.05 ** \ p \leq 0.01 *** p \leq 0.001$

- $r_{x1,x2} =$ 0.9
- Sample 3

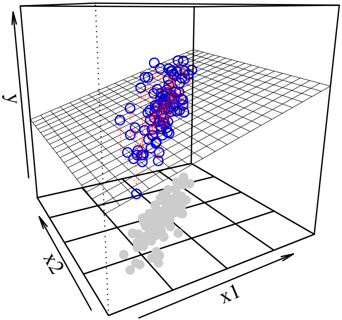


High Collinearity Causes Unstable Estimates

	M1	
	Estimate	(S.E.)
(Intercept)	-1.076	(3.722)
x1	0.354*	(0.157)
x2	0.077	(0.163)
N	100	
RMSE	7.568	
R^2	0.283	
adj R^2	0.269	

$*p \leq 0.05 ** \; p \leq 0.01 *** p \leq 0.001$

- $r_{x1,x2} =$ 0.9
- Sample 4

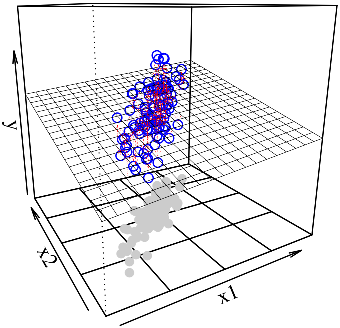


High Collinearity Causes Unstable Estimates

	M1	
	Estimate	(S.E.)
(Intercept)	0.285	(4.298)
x1	0.122	(0.187)
x2	0.252	(0.183)
N	100	
RMSE	7.591	
R^2	0.178	
adj R^2	0.161	

$*p \leq 0.05$ $**\ p \leq 0.01$ $***p \leq 0.001$

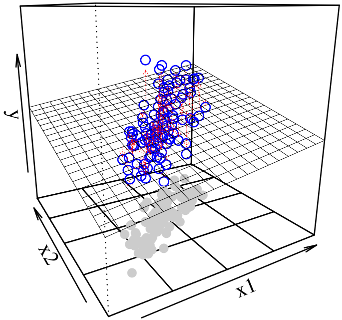
- $r_{x1,x2} =$ 0.9
- Sample 5



High Collinearity Causes Unstable Estimates

	M1	
	Estimate	(S.E.)
(Intercept)	0.553	(4.059)
x1	0.173	(0.134)
x2	0.222	(0.143)
N	100	
RMSE	6.355	
R^2	0.212	
adj R^2	0.196	

$*p \leq 0.05 ** \ p \leq 0.01 *** p \leq 0.001$

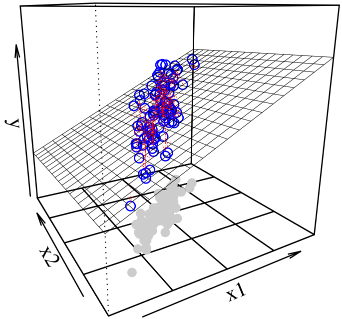


- $r_{x1,x2} =$ 0.9
- Sample 6

High Collinearity Causes Unstable Estimates

	M1	
	Estimate	(S.E.)
(Intercept)	-2.770	(3.985)
x1	0.468*	(0.188)
x2	-0.044	(0.182)
N	100	
RMSE	7.434	
R^2	0.238	
adj R^2	0.223	

$*p \leq 0.05 ** \ p \leq 0.01 *** p \leq 0.001$

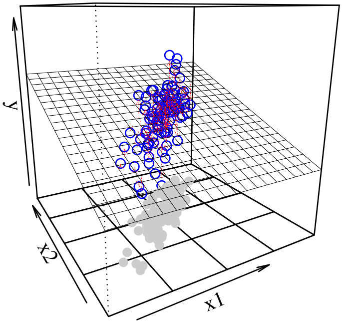


- $r_{x1,x2} =$ 0.9
- Sample 7

High Collinearity Causes Unstable Estimates

	M1	
	Estimate	(S.E.)
(Intercept)	-2.193	(3.835)
x1	-0.003	(0.181)
x2	0.463**	(0.171)
N	100	
RMSE	7.096	
R^2	0.286	
adj R^2	0.272	

$*p \leq 0.05 ** \ p \leq 0.01 *** p \leq 0.001$



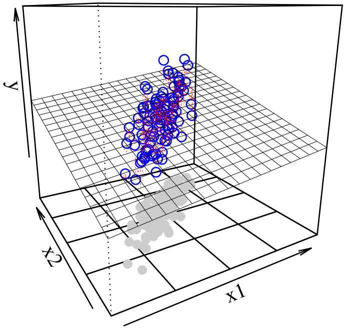
- $r_{x1,x2} = 0.9$
- Sample 8

High Collinearity Causes Unstable Estimates

	M1	
	Estimate	(S.E.)
(Intercept)	-1.883	(3.750)
x1	0.151	(0.156)
x2	0.283	(0.156)
N	100	
RMSE	6.655	
R^2	0.275	
adj R^2	0.260	

$*p \leq 0.05 ** \ p \leq 0.01 *** p \leq 0.001$

- $r_{x1,x2} =$ 0.9
- Sample 9

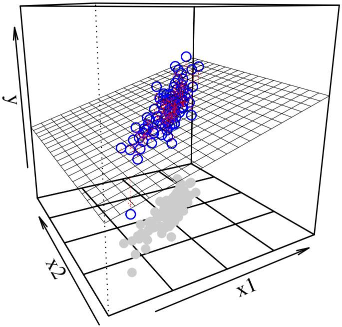


High Collinearity Causes Unstable Estimates

	M1	
	Estimate	(S.E.)
(Intercept)	-4.236	(3.089)
x1	0.421**	(0.146)
x2	0.062	(0.140)
N	100	
RMSE	6.557	
R^2	0.389	
adj R^2	0.376	

$*p \leq 0.05 ** \ p \leq 0.01 *** p \leq 0.001$

- $r_{x1,x2} =$ 0.9
- Sample 10

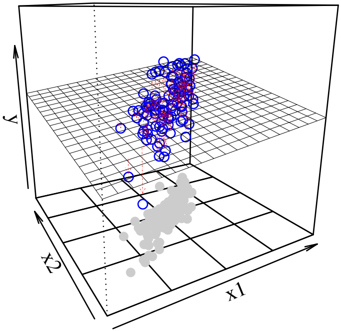


High Collinearity Causes Unstable Estimates

	M1	
	Estimate	(S.E.)
(Intercept)	3.325	(4.127)
x1	0.160	(0.165)
x2	0.173	(0.161)
N	100	
RMSE	6.999	
R^2	0.154	
adj R^2	0.136	

$*p \le 0.05$ $**\ p \le 0.01$ $***p \le 0.001$

- $r_{x1,x2} =$ 0.9
- Sample 11

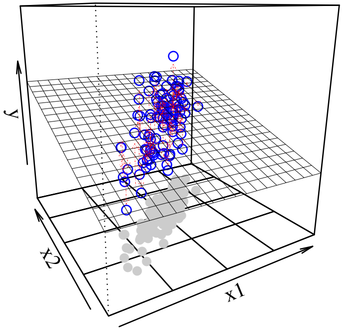


High Collinearity Causes Unstable Estimates

	M1	
	Estimate	(S.E.)
(Intercept)	4.513	(3.833)
x1	-0.011	(0.176)
x2	0.335*	(0.167)
N	100	
RMSE	7.282	
R^2	0.182	
adj R^2	0.165	

$*p \leq 0.05 ** \ p \leq 0.01 *** p \leq 0.001$

- $r_{x1,x2} =$ 0.9
- Sample 12

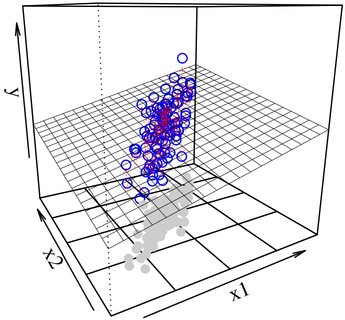


High Collinearity Causes Unstable Estimates

	M1	
	Estimate	(S.E.)
(Intercept)	-3.865	(3.776)
x1	0.276	(0.157)
x2	0.196	(0.171)
N	100	
RMSE	6.985	
R^2	0.300	
adj R^2	0.285	

$*p \le 0.05 ** \ p \le 0.01 *** p \le 0.001$

- $r_{x1,x2} = 0.9$
- Sample 13

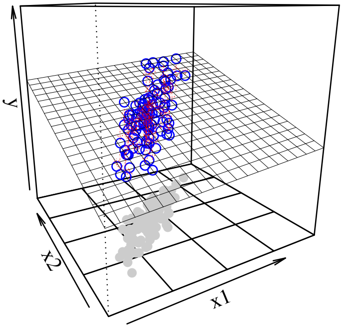


High Collinearity Causes Unstable Estimates

	M1	
	Estimate	(S.E.)
(Intercept)	-1.898	(3.463)
x1	0.093	(0.148)
x2	0.345*	(0.157)
N	100	
RMSE	6.805	
R^2	0.297	
adj R^2	0.283	

$*p \leq 0.05 ** \ p \leq 0.01 *** p \leq 0.001$

- $r_{x1,x2} = 0.9$
- Sample 14

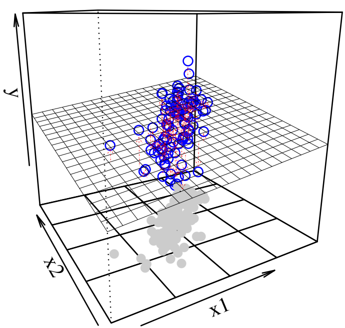


High Collinearity Causes Unstable Estimates

	M1	
	Estimate	(S.E.)
(Intercept)	3.555	(4.459)
x1	0.148	(0.160)
x2	0.195	(0.155)
N	100	
RMSE	7.815	
R^2	0.135	
adj R^2	0.117	

$*p \leq 0.05 ** \; p \leq 0.01 *** p \leq 0.001$

- $r_{x1,x2} =$ 0.9
- Sample 15

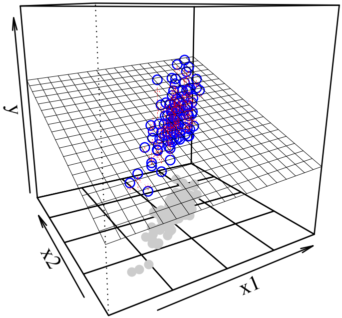


High Collinearity Causes Unstable Estimates

	M1	
	Estimate	(S.E.)
(Intercept)	-1.632	(3.451)
x1	0.070	(0.161)
x2	0.376*	(0.151)
N	100	
RMSE	6.550	
R^2	0.327	
adj R^2	0.313	

$*p \leq 0.05 ** \ p \leq 0.01 *** p \leq 0.001$

- $r_{x1,x2} =$ 0.9
- Sample 16

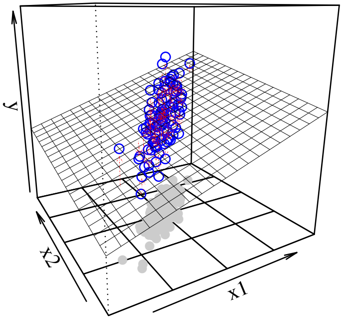


High Collinearity Causes Unstable Estimates

	M1	
	Estimate	(S.E.)
(Intercept)	-7.464*	(3.676)
x1	0.344*	(0.154)
x2	0.177	(0.159)
N	100	
RMSE	6.938	
R^2	0.360	
adj R^2	0.347	

$*p \leq 0.05 ** \ p \leq 0.01 *** p \leq 0.001$

- $r_{x1,x2} = 0.9$
- Sample 17

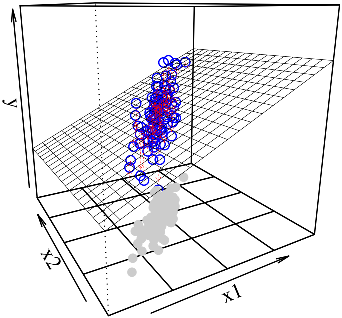


High Collinearity Causes Unstable Estimates

	M1	
	Estimate	(S.E.)
(Intercept)	-0.470	(3.560)
x1	0.449**	(0.170)
x2	-0.035	(0.166)
N	100	
RMSE	7.132	
R^2	0.263	
adj R^2	0.248	

$*p \leq 0.05 ** \ p \leq 0.01 *** p \leq 0.001$

- $r_{x1,x2} =$ 0.9
- Sample 18

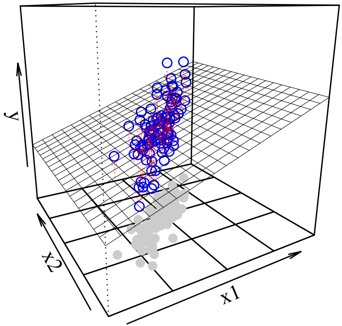


High Collinearity Causes Unstable Estimates

	M1	
	Estimate	(S.E.)
(Intercept)	-0.239	(3.381)
x1	0.374*	(0.150)
x2	0.037	(0.144)
N	100	
RMSE	6.169	
R^2	0.284	
adj R^2	0.270	

$*p \leq 0.05 ** \ p \leq 0.01 *** p \leq 0.001$

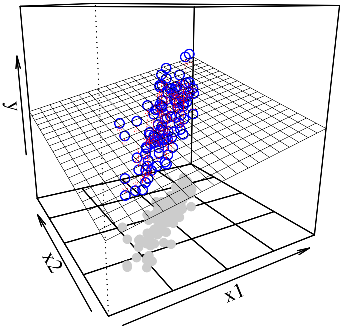
- $r_{x1,x2} =$ 0.9
- Sample 19



High Collinearity Causes Unstable Estimates

	M1	
	Estimate	(S.E.)
(Intercept)	-4.415	(3.572)
x1	0.254	(0.171)
x2	0.205	(0.164)
N	100	
RMSE	7.080	
R^2	0.317	
adj R^2	0.303	

$*p \leq 0.05 ** \ p \leq 0.01 *** p \leq 0.001$



- $r_{x1,x2} =$ 0.9
- Sample 20

Symptom 2: "Bouncing B's"

- If there is NO COLLINEARITY, estimates of slopes do not change when variables are put in and removed from the model.
- If there IS COLLINEARITY, the estimate of each $\hat{\beta}_j$ depends on all of the data for all of the variables.
- Slope estimates "jump around" when variables are inserted and removed from the model.

Omitted Variable Bias

- If the Right fitted model is

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x1_i + \hat{\beta}_2 x2_i \tag{3}$$

- But you fit

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x1_i \tag{4}$$

- Then the estimate of $\hat{\beta}_1$is “biased” because $x1_i$ “gets credit” for the effect of $x2_i$.

Formula to Demonstrate Effect of Collinearity with 2 IVs

- The simple one-input regression

$$y_i = c_0 + c_1 x1_i + u_i \tag{5}$$

- The two-input regression

$$y_i = \beta_0 + \beta_1 x1_i + \beta_2 x2_i + e_i \tag{6}$$

- The auxiliary regression

$$x2_i = d_0 + d_1 x1_i + v_i \tag{7}$$

- The following summarizes the effect of excluding $x2_i$

$$\hat{c}_1 = \hat{\beta}_1 + \hat{\beta}_2 \cdot \hat{d}_1 \tag{8}$$

- If you leave out $x2_i$, the estimate $\hat{c}_1$ is a "biased" estimate of the slope $\hat{\beta}_1$.
- Equivalently, Here's how the $\hat{\beta}_1$ "jumps" when $x2_i$ is added to the model

$$\hat{\beta}_1 = \hat{c}_1 - \hat{\beta}_2 \cdot \hat{d}_1 \tag{9}$$

Suppressor Variables

- In practice, it usually seems that, leaving a variable out makes the b's (and t's) of the included variables "bigger".
- Not logically necessary, however. A "Suppressor" variable is one that makes $\hat{\beta}$ from another variable become greater when the suppressor is included in the model. (Leaving out the other "suppresses" $\hat{\beta}$).
- Including a variable may make the estimated coefficients bigger for both variables.

Example: Heaven and Hell

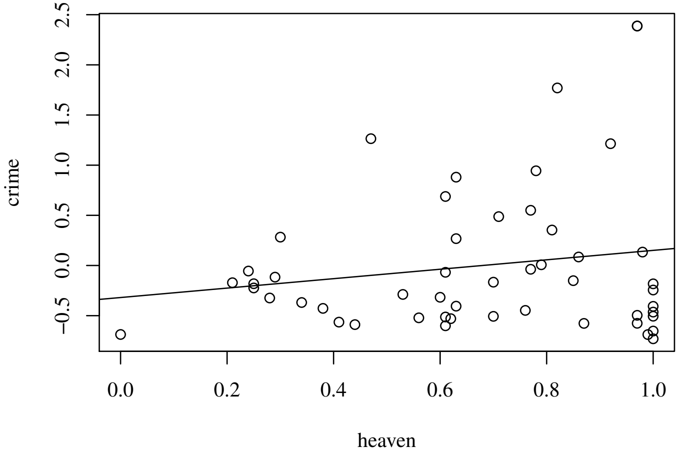
- in rockchalk, dataset religion crime is described
  - "The data national-level summary indicators of public opinion about the existence of heaven and hell as well as the national rate of violent crime."
- Special thanks to the anonymous data donor

Crime is not a function of the belief in Heaven

	M1 Estimate (S.E.)
(Intercept)	-0.319
	(0.281)
heaven	0.470
	(0.386)
N	51
RMSE	0.737
R^2	0.029

$*p \leq 0.05 ** \ p \leq 0.01 *** p \leq 0.001$

See?

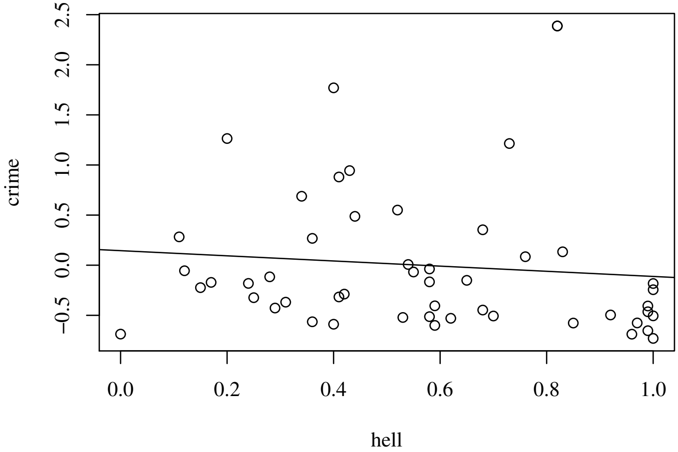


Crime is not a function of the belief in Hell

	M1 Estimate (S.E.)
(Intercept)	0.145
	(0.235)
hell	-0.257
	(0.369)
N	51
RMSE	0.744
R^2	0.010

$*p \leq 0.05 ** \ p \leq 0.01 *** p \leq 0.001$

See?

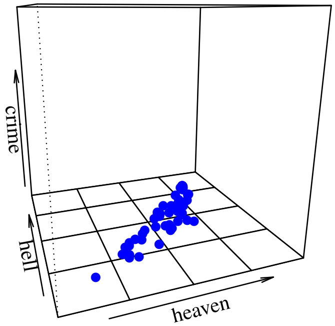


But Heaven and Hell Both Affect Crime

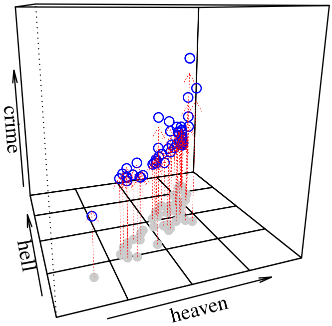
	M1 Estimate (S.E.)
(Intercept)	-0.760***
	(0.212)
heaven	5.187***
	(0.746)
hell	-4.813***
	(0.706)
N	51
RMSE	0.531
R^2	0.507
adj R^2	0.486

$*p \leq 0.05 ** \ p \leq 0.01 *** p \leq 0.001$

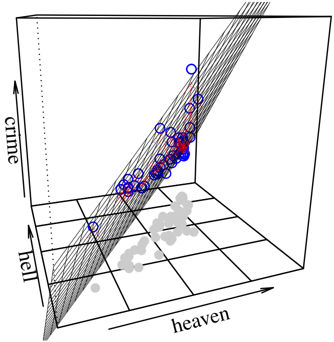
Visualize that ...



Visualize that ...



Visualize that ...



Outline

The Bivariate Correlation Matrix

- A simple, but not completely informative approach
- cor(x) shows pearson correlations
- does not demonstrate the multi in multi-collinearity

Putting the "Multi" in Multicollinearity

- Regress Each Predictor on all of the others (creates k fitted models)

$$Auxiliary\ Regression\, j:\ \widehat{xj_i} = \hat{d}_0 + \hat{d}_1 x1_i + \ldots exclude\, j'th \ldots + xk_i \quad (10)$$

- Cohen's notation for the R^2 from that fit is $R^2_{xj.x2,x3,\ldots,(j),\ldots,k}$
- I write R^2_j: R^2 from the $j'th$ auxiliary regression
- Intuition: $1 - R^2_j$ indicates magnitude of xj's separate effect.
  - if $1 - R^2_j$ is almost 0, it means the other variables can predict xj almost perfectly
- "Tolerance" is a name for $1 - R^2_j$ (according to Cohen, et al).

Variance Inflation Factor

- Weird but true. The true variance of $\hat{\beta}_j$ can be re-organized thusly:

$$Var(\hat{\beta}_j) = \frac{\sigma_e^2}{(1 - R_j^2)\sum(xj_i - \overline{xj})^2} \tag{11}$$

  - R_j^2 is R-square from regressing xj on all other predictors in auxiliary regression.
  - Note denominator: Product of
    - tolerance, $(1 - R_j^2)$
    - the sum of squares for the $j'th$ variable
  - Test question: If your $\widehat{Var(\hat{\beta}_j)}$ is huge, what changes would you like to make in your data so as to make it smaller?

Variance inflation factor (page 2)

- Re-write the variance formula like so

$$Var(\hat{\beta}_j) = \frac{\sigma_e^2}{(1-R_j^2)\sum(xj_i - \overline{xj})^2} = \frac{1}{(1-R_j^2)} \times \frac{\sigma_e^2}{\sum(xj_i - \overline{xj})^2} \quad (12)$$

- See why the first term is called a “variance inflation factor”?

$$VIF_j = \frac{1}{1-R_j^2} \quad (13)$$

mcDiagnose in rockchalk package: Hell!

```
mcDiagnose(mod3)
```

```
The following auxiliary models are being estimated
   and returned in a list:
heaven ~ hell
<environment: 0x755f8b8>
hell ~ heaven
<environment: 0x755f8b8>
Drum roll please!

And your R_j Squareds are (auxiliary Rsq)
   heaven      hell
0.8607671 0.8607671
The Corresponding VIF, 1/(1-R_j^2)
 heaven     hell
7.18221 7.18221
Bivariate Correlations for design matrix
```

mcDiagnose in rockchalk package: Hell! ...

```
          heaven hell
heaven      1.00 0.93
hell        0.93 1.00
```

Big Take-Away Points (So Far)

Fit your model with all of the variables your theory leads you to include

- $\hat{\beta}$'s are still unbiased.
- If $std.err.(\hat{\beta})$ is small, and t's are "good", don't worry about it.
- MC between two variables (or within a block of variables) need not affect estimates of other coefficients.

Outline

I: Do nothing:

- acknowledge problem
- Can do F test for groups of variables

II: Get More Diverse Data!

- Gather more data, so the X's are not so intercorrelated.
- This is the best and only truly meaningful solution.
- In research planning, be conscious of MC dangers

III: Combine Variables Into An Index

- Begin with a set of variables that are almost the same, and then they combine them by adding them or calculating an average.
- Best when variables are "conceptually related".
- Better still if they are thought of as multiple measures of the same thing.
- Poor person's "structural equation model"

More Sophisticated way to Create an Index: Principal Components

- Definition: principal component is an "underlying variable" (unmeasured variable) that is related to the observed $X1$, $X2$,$X3$, $X4$.
- PCs can be "extracted" from the data and used as predictors. 2 PCs might effectively summarize 4 X's.
- Some authors very enthusiastic about it, some find components difficult to understand.

PC's Require Matrix Notation

- Center variables so they are in "deviations form"(So $X1_i = x1_i - \bar{x}$, if $x1_i$ was the "original" data.)
- Let's reproduce 4 X's with 2 PCs, called $Z1$and $Z2$. The theory is that:

$$\begin{bmatrix} X1 & X2 & X3 & X4 \end{bmatrix} = \begin{bmatrix} Z1 & Z2 \end{bmatrix} A + \begin{bmatrix} u1 & u2 & u3 & u4 \end{bmatrix}$$

  - $u1$,$u2$,$u3$, $u4$ are columns of random errors, $E[u_j] = 0$
  - a is a matrix of weights (actually, "eigenvalues" of X).

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \end{bmatrix} \qquad (14)$$

- Two PCs might effectively "reproduce" the X's like this:

$$X1_i = a_{11}\,Z1_i + a_{21}\,Z2_i + u1_i \qquad X2_i = a_{12}\,Z1_i + a_{22}\,Z2_i + u2_i$$
$$X3_i = a_{13}\,Z1_i + a_{23}\,Z2_i + u3_i \qquad X4_i = a_{14}\,Z1_i + a_{24}\,Z2_i + u1_i$$

Substantive Interpretation of PCs

- Suppose the “city” level IVs are like this:
  - $X1$ number of children in public school
  - $X2$ number of teachers in public schools
  - $X3$ number of employees in city government
  - $X4$ number of desks owned by city
- The first 2 go together, the last 2 go together

How Would that Look In PC Output?

- Search for pattern in matrix of a's.

$$\begin{bmatrix} X1 & X2 & X3 & X4 \end{bmatrix} = \begin{bmatrix} Z1 & Z2 \end{bmatrix} \begin{bmatrix} .5 & .5 & 0.0 & 0.0 \\ 0.0 & 0.0 & .5 & .5 \end{bmatrix}$$

- This example makes it very clear.
  - PC $Z1$ is driving the predictions for $X1$ and $X2$,
  - PC $Z2$ is driving $X3$ and $X4$.

PCA: Math Facts

1. By design, the columns $Z1$ and $Z2$ give the "best possible" linear prediction of the values of the X's.
2. Can add more PCs if desired.
3. The PCs $Z1$and $Z2$ are uncorrelated w/each other (orthogonal). So if we remove the $X's$ from the regression model, and we use the $Z's$ instead, then our "inputs" are not intercorrelated any more.

Greene Does Not Endorse PC

The leading econometrics text, William Greene, *Econometric Analysis*, 5th ed (p. 58)

> *The problem here is that if the original model in the form $y = X\beta + \epsilon$ were correct, then it is unclear what one is estimating when one regresses y on some set of linear combinations of the columns of X. Algebraically, it is simple; at least for the principal components case, in which we regress y on $Z = XC_L$ to obtain d, it follows that $E(d) = \delta = C_L C_L' \beta$. In an economic context, if β has an interpretation, then it is unlikely that δ will. (How do we interpret the price elasticity plus minus twice the income elasticity?)*

Cohen, et al., also Reluctant

A leading text, Cohen, et al, Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, 3rd ed (p. 429)

> *Unfortunately, however, these $\tilde{\beta}_i$ are only rarely interpretable. The component scores are linear combinations of the original IVs and will not typically have a clear meaning.... On the positive side, dropping components that account for small proportions of variance eliminates major sources of multicollinearity. The result is that the back transformed regression coefficients, β_i,for the original IVs will be biased, but will be more robust to small changes in the data set than are the original OLS estimates.*

IV: Ridge Regression

- Adding information adds efficiency.
- See Practical Regression and Anova using R by Julian Faraway (in R contributed documentation on `http://www.r-project.org`)
- Instead of using the OLS estimator

$$\hat{\beta}^{OLS} = (X'X)^{-1}X'y \quad Var(\hat{\beta}^{OLS}) = \sigma_e^2(X'X)^{-1}$$

  Insert a scalar value, λ, known as the "ridge constant," to create an adjusted estimator:

$$\hat{\beta}^{ridge} = (X'X + \lambda I)^{-1}X'y \quad Var(\hat{\beta}^{ridge}) = \sigma_e^2(X'X + \lambda I)^{-1}$$

- Ameliorates multicollinearity. If a small value of λ is used, then, of course, the estimates are not far from the $\hat{\beta}^{ols}$.

This estimator is known to be biased, but it is also known to havelower variance than the OLS estimator.

Evaluating Biased Estimators

- Suppose we want the smallest squared-error

$$E[(\hat{\beta} - b)^2]$$

- Assert: That is

$$E[(\hat{\beta} - b)^2] = (E[\hat{\beta} - b])^2 + E[(\hat{\beta} - E(\hat{\beta}))^2]$$

which is

$$E[(\hat{\beta} - b)^2] = \textit{bias of estimator}^2 + \textit{variance of estimator}$$

Outline

OLS Estimators

- The OLS estimator is

$$\hat{\beta} = (X'X)^{-1}X'Y \qquad \widehat{V(\hat{\beta})} = \widehat{\sigma_e^2} * (X'X)^{-1}$$

  $\widehat{\sigma_e^2}$ estimated variance of the error term, also known as the MeanSquareError.

- The slope and variance estimates require us to calculate:

$$(X'X)^{-1}$$

- Perfect multicollinearity: $(X'X)^{-1}$ cannot be calculated– $(X'X)$ cannot be "inverted."
- In practice, multicollinearity is not severe enough to prevent calculations. But it does make the estimated variances larger.

Here is an analogy with ordinary numbers.

- Take $X = 0$. Then the inverse, $X^{-1}$is undefined. X cannot be inverted.
- Suppose instead $X = 0.0000000001$. Now the inverse of X does exist, but it is some HUGE number, $X^{-1} = \frac{1}{0.0000000001} = 10^9$.

Inverse of a Matrix

- Recall that $(X'X)^{-1}$ is a matrix defined in the following way:

$$(X'X) * (X'X)^{-1} = I = \begin{matrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & ... & 0 & 0 & 0 \\ 0 & 0 & ... & ... & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{matrix}$$

- If X is an nxp matrix, then X' is pxn, and so the product (X'X) is pxp. (X'X) is square.
- If $(X'X)$ cannot be inverted, it means that there are 2 or more redundant rows in $(X'X)$. Some computer programs will give the error "the model is not full rank." If "rank" of $(X'X)$ is smaller than p, then it means there are redundant rows.
- Usually, in practice, the cross product matrix $(X'X)$ can still be inverted, however, the values in $(X'X)^{-1}$ are HUGE.

Envision $(X'X)$ and $(X'X)^{-1}$

- I kept wondering what the matrix $(X'X)^{-1}$ would look like, so here's an example.
- Suppose

$$X = \begin{bmatrix} 1 & X1_1 & X2_1 \\ 1 & X1_2 & X2_2 \\ 1 & X1_3 & X2_3 \\ 1 & X1_4 & X2_4 \\ \dots & \dots & \dots \\ 1 & X1_N & X2_N \end{bmatrix}$$

- The first column is the "y intercept" and there are 2 variables.
- $X'X$. The product of X transpose and X

Envision $(X'X)$ and $(X'X)^{-1}$ ...

$$X'X = \begin{bmatrix} 1 & 1 & 1 & 1 & \vdots & 1 \\ X1_1 & X1_2 & X1_3 & X1_4 & \vdots & X1_N \\ X2_1 & X2_2 & X2_3 & X2_4 & & X2_N \end{bmatrix} \begin{bmatrix} 1 & X1_1 & X2_1 \\ 1 & X1_2 & X2_2 \\ 1 & X1_3 & X2_3 \\ 1 & X1_4 & X2_4 \\ \cdots & \cdots & \cdots \\ 1 & X1_N & X2_N \end{bmatrix}$$

$$X'X = \begin{bmatrix} N & \sum X1_i & \sum X2_i \\ \sum X1_i & \sum X1_i^2 & \sum X1_i \cdot X2_i \\ \sum X2_i & \sum X1_i \cdot X2_i & \sum X2_i^2 \end{bmatrix} \tag{15}$$

- We want to know $(X'X)^{-1}$ in order to calculate $\hat{\beta}$ and $Var(\hat{\beta})$.

Envision $(X'X)$ and $(X'X)^{-1}$ ...

- From matrix algebra: the inverse of a matrix can be written as the product of the determinant, $det(X'X)$, and a matrix called the adjoint $adj(X'X)$. (Consult any linear algebra textbook, such as Howard Anton, *Elementary Linear Algebra, 3ed*, New York, John Wiley, 1980, p. 80). The formula is:

$$(X'X)^{-1} = \frac{1}{det(X'X)} adj(X'X)$$

- The determinant of the 3x3 matrix (15) is

$$\begin{aligned} det(X'X) \quad = \quad & N\left(\sum X1_i^2\right)\cdot\left(\sum X2_i^2\right) + 2\left(\sum X1_i\right)\cdot\left(\sum X2_i\right)\cdot\left(\sum X1_i \cdot X2_i\right) \\ & -\left(\sum X2_i\right)^2\left(\sum X1_i^2\right) - N\cdot\left(\sum X1_i\cdot X2_i\right)^2 \\ & -\left(\sum X1_i\right)^2\cdot\left(\sum X2_i^2\right) \end{aligned} \tag{16}$$

If that determinant is equal to 0, then the inverse is not defined.

Envision $(X'X)$ and $(X'X)^{-1}$ ...

- If you mistakenly put in two identical columns, so $X1_i = X2_i$, the determinant is 0. Replace $X2_i$ by $X1_i$:

$$
\begin{aligned}
det(X'X) \quad = \quad & N\left(\sum X1_i^2\right)\cdot\left(\sum X1_i^2\right) \\
& +2\left(\sum X1_i\right)\cdot\left(\sum X1\right)\cdot\left(\sum X1_i\cdot X1_i\right) \\
& -\left(\sum X1_i\right)^2\left(\sum X1_i^2\right) - N\cdot\left(\sum X1_i\cdot X1_i\right)^2 \\
& -\left(\sum X1_i\right)^2\cdot\left(\sum X1_i^2\right)
\end{aligned}
$$

$$
\begin{aligned}
& N\left(\sum X1_i^2\right)^2 + 2\left(\sum X1_i\right)^2\cdot\left(\sum X1_i^2\right) \\
= \quad & -N\cdot\left(\sum X1_i^2\right)^2 - 2\left(\sum X1_i\right)^2\left(\sum X1_i^2\right)
\end{aligned}
$$

The terms with plus signs are exactly counterbalanced by negative signs, and so $det(X'X) = 0$ if two redundant variables are included. When there are 2 redundant columns, there can be no linear regression analysis. What if 2 columns are not exactly the same, but instead just "similar" or "correlated".

$$X2_i = X1_i + \gamma_i$$

Envision $(X'X)$ and $(X'X)^{-1}$ ...

$$
\begin{aligned}
det(X'X) \quad = \quad & N\left(\sum X1_i^2\right)\cdot\left(\sum (X1_i+\gamma_i)^2\right) \\
& +2\left(\sum X1_i\right)\cdot\left(\sum (X1_i+\gamma_i)\right)\cdot\left(\sum X1_i\cdot(X1_i+\gamma_i)\right) \\
& -\left(\sum (X1_i+\gamma_i)\right)^2\left(\sum X1_i^2\right) - N\cdot\left(\sum X1_i\cdot(X1_i+\gamma_i)\right)^2 \\
& -\left(\sum X1_i\right)^2\cdot\left(\sum (X1_i+\gamma_i)^2\right)
\end{aligned} \tag{17}
$$

The intuition: if $\gamma_i = 0$, $det(X'X) = 0$.
If γ_i is small–$X2_i$ is redundant with $X1_i$– then this determinant will be small.

Outline

Problems

1. Get the cystfibr example dataset. You should have already seen that data and conducted part of this exercise. Pick several of the predictors and run a regression. The standard errors will likely be large, the t's small. Calculate the auxiliary regressions. List out the R_j^2 and calculate the VIF

   Many programs have routines to calculate VIF for you, but I think it really does help if you work this out at least once the "old fashioned way." Then use your software's vif calculator to double check. In the "car" package in R, a vif function is available.

   I've not checked this myself, but you can let me know. Do the results based on the vif differ from the conclusions you would draw from the bivariate correlation coefficients by themselves? Sometimes the 2 methods lead to the same simple answer, but not always.

   Can you think of a situation in which the VIF analysis would be richer than the analysis based on the bivariate correlations alone?

Problems ...

2. The procedure known as "stepwise regression" is hated by political scientists and many sociologists. In fact, we thought it was dead! But recently I found out that some psychologists like it! Actually, quite a few of them do. It is apparently something of a "culture war."

   Do a quick Google to find out what stepwise regression is.

   Once you have just a basic grasp of it, let me ask you this. What effect do you think multicollinearity will have on the step-by-step decisions made during stepwise regression?

3. I sent a paper into a journal. The column of $\hat{b}$'s was so awesome you wouldn't believe it. I could tell a good story about every one. And, furthermore, my R^2 was huge and all my t statistics were bigger than 2. The paper was unceremoniously rejected because one of the reviewers said that variables $x1$, $x2$, and $x3$ were strongly inter-correlated. In fact, he/she said they were multicorrelated. What should I say in response to the editor?

Problems ...

4. I once told a student, "you have an obvious multicollinearity problem. Go look into that." He kept coming back and showing me the Pearson product-moment correlation coefficient matrix. None of the correlations between predictors were bigger than 0.3. But, after a glance at his regression table, I could say for sure I was correct. Why do you think I was so sure?