

# Lecture 2: Multiple Predictors in OLS Regression

Paul E. Johnson<sup>1</sup>

<sup>1</sup>Departments of Political Science & Psychology, KU

# Outline

- 1 Choosing Variables: An Unsolved Problem
- 2 Non-Orthogonal Predictor Columns
- 3 Hat matrix
- 4 Multiple  $R^2$  and adjusted  $R^2$
- 5 Which Variables are “Important”?
  - The Famous Ballantine Graph of Cohen, et al
  - Partial Correlation
  - Semi-partial Correlation Coefficients
  - relaimpo package
- 6 R Interactions (W&R notation).
- 7 Practice Problems

## What the heck is this lecture about?

- Cautions about variable selection
- partial and semi-partial correlations, and “importance” conundrum
- W&R notation for regression

# Outline

- 1 Choosing Variables: An Unsolved Problem
- 2 Non-Orthogonal Predictor Columns
- 3 Hat matrix
- 4 Multiple  $R^2$  and adjusted  $R^2$
- 5 Which Variables are “Important”?
  - The Famous Ballantine Graph of Cohen, et al
  - Partial Correlation
  - Semi-partial Correlation Coefficients
  - relaimpo package
- 6 R Interactions (W&R notation).
- 7 Practice Problems

## How to decide which variables belong in a model?

- At its core, multiple regression has NOT MUCH to say about this.
- Regression models are fit with the assumption that the correct variables were selected.
- Leaving a variable out means you assume  $\beta_j = 0$ .
- “Automatic” variable selection tools are inadequate/misleading.

## The "Oracle" Model, in a Perfect World

- Suppose you have 100s of  $x$  variables floating about. The general model would be too huge:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \dots + \beta_{100} x_{100i} + \dots + e_i$$

- Estimation would likely fail. Even if we could get estimates, they would have huge standard errors.
- Suppose the Oracle (a voice from God) tells you the "true slope" is zero for most coefficients.
- The Oracle says "set all of the irrelevant  $\beta_j$ 's set to 0", so you fit

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + e_i$$

## If You Stop There, OK!

- The statistical derivations we emphasize pre-suppose you stop at this point. Standard errors, p-values, etc, are derived with that assumption
- Danger arises from building a model “stepwise”, especially “building-up” by adding variables one-at-a-time if they have “good” p-values.
- More commonly, we fit a model with a large number of coefficients, and then wrestle with the question of “which ones should we throw out?”
- You risk misleading yourself and others.
  - Edward Leamer. 1983. Lets Take the Con Out of Econometrics. *American Economic Review*, 31-43.

# The Big Picture: Selection and Regularization

- 1 Variable “selection”. Exclude variables, assuming their true coefficient is actually 0
  - 1 stepwise methods add and remove variables in a sequence
  - 2 “best subset” methods: systematic search over all possible regressions for the combination of predictors that reduces prediction error
- 2 Parameter regularization. Try to estimate a giant vector  $\beta = (\beta_0, \beta_1, \dots, \beta_{100})$  but use some criteria to “stabilize” the vector, to “penalize” the estimates that are unstable and shrink them toward 0
  - 1 ridge regression.
  - 2 The “lasso” (described well in James et al, *An Introduction to Statistical Learning*,(2013))

## Don't Use Stepwise Variable Selection

- Frank Harrell. 2001. *Regression Modeling Strategies*. Stepwise regression is bad because it biases parameter estimates and often chooses the “wrong” variables.
- Flom and Cassell. 2007. “Stopping Stepwise: Why stepwise and similar selection methods are bad, and what you should use.”
- Stata website: Problems with stepwise regression:  
<http://www.stata.com/support/faqs/stat/stepwise.html>
- Forward stepwise regression seems most dangerous to me. Many of us are “closet” practitioners of a subject “backwards stepwise” approach that throws out variables

## I suggest a hierarchical approach

- Lets avoid all variants of automatic “stepwise regression.”
- Don't try the “lasso” or other fancy tools unless you really understand them
- Lets
  - follow our substantive knowledge as far as it will guide us
  - When uncertain, fit several models and compare side by side.
- Basic idea:
  - If  $\hat{\beta}_j$  is similar across various regressions, then we probably have a good estimate
  - Otherwise, we have a problem that requires some hard thought.

## Including Extra Variables: What Danger?

- If extra  $X$ 's are
  - uncorrelated with your predictors that do belong,
    - then you are doing no "damage" to your estimates.
- But, if those extra variables are correlated with your predictors that do belong, then
  - you are diluting the estimates of the ones that do belong, and
  - causing "inefficiency" (higher variance) and
  - the estimated standard errors are inflated

## Excluding Variables: What Danger?

- If you remove variables that really do belong, and those variables are
  - uncorrelated with your predictors that are still included, then
    - your estimated slopes for the included variables are still unbiased
- But, if they are correlated with your other predictors, then
  - your estimated slopes for the included variables are biased. (omitted variable bias)
- Called the “omitted variable bias”.

# Outline

- 1 Choosing Variables: An Unsolved Problem
- 2 Non-Orthogonal Predictor Columns**
- 3 Hat matrix
- 4 Multiple  $R^2$  and adjusted  $R^2$
- 5 Which Variables are “Important”?
  - The Famous Ballantine Graph of Cohen, et al
  - Partial Correlation
  - Semi-partial Correlation Coefficients
  - relaimpo package
- 6 R Interactions (W&R notation).
- 7 Practice Problems

## The Virtues Orthogonal data

- 2 columns are orthogonal if the deviations about the means are not related (Interpretation: Knowing a respondent's score on one variable implies nothing about the other one).
- The numerator of the covariance is zero:

$$\sum_{i=1}^N (x1_i - \bar{x1}) \times (x2_i - \bar{x2}) = (x1_1 - \bar{x1})(x2_1 - \bar{x2}) \\ + (x1_2 - \bar{x1})(x2_2 - \bar{x2}) + (\dots N \text{ times}) = 0$$

- Since  $r = \frac{\text{Cov}(x1, x2)}{\sqrt{\text{Var}(x1) \cdot \text{Var}(x2)}}$ , Pearson correlation is 0.
- Geometrically,  $x1$  and  $x2$  are “perpendicular”, at right angles to one another

## Here is the important thing I want to convey

- 1 If the columns of  $X$  are not correlated with each other, THEN regression is easy!
  - 1  $\hat{\beta}$  estimates are stable, not affected by inserting or removing other variables!
- 2 If the columns of  $X$  are correlated with each other, THEN regression is difficult!
  - 1  $\hat{\beta}$  fluctuate as variables go in and out.
  - 2 it is difficult to know which variables “truly belong” in the model (“specification” is difficult)

## For Interpreting Regression Formulas

- Variance  $\iff$  sums of squares

$$\sum (x_i - \bar{x})^2 \text{ is the same as } (N - 1) \widehat{Var}[x_i] \quad (1)$$

- CoVariance  $\iff$  sums of “cross products”

$$\sum (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) \text{ is the same as } (N - 1) \widehat{Cov}[x_1, x_2] \quad (2)$$

- Covariance  $\iff$  Correlation

$$\widehat{Cov}[x_1, x_2] = \widehat{Std.Dev.}(x_1) \cdot \widehat{Std.Dev.}(x_2) \cdot r_{x_1, x_2} \quad (3)$$

- Cohen et al, for example, show  $\hat{\beta}_j$  in terms of correlation coefficients.

## Cohen offers estimation formula for 2 predictors

- For 2 predictors  $x_1$  and  $x_2$ .

$$\hat{\beta}_1 = \frac{(\sum(x_{1i} - \bar{x}_1)(y_i - \bar{y})) (\sum(x_{2i} - \bar{x}_2)^2) - (\sum(x_{2i} - \bar{x}_2)(y_i - \bar{y})) (\sum(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2))}{(\sum(x_{1i} - \bar{x}_1)^2) (\sum(x_{2i} - \bar{x}_2)^2) - (\sum(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2))^2} \quad (4)$$

- With “mean centered” data, formula simplifies:  $\bar{x}_1 = 0$ ,  $\bar{x}_2 = 0$ ,  $\bar{y} = 0$

$$\hat{\beta}_1 = \frac{(\sum x_{1i}y_i) (\sum x_{2i}^2) - (\sum x_{2i}y_i) (\sum x_{1i}x_{2i})}{(\sum x_{1i}^2) (\sum x_{2i}^2) - (\sum x_{1i}x_{2i})^2} \quad (5)$$

## Can interpret beta coefficient as a ratio of Pearson correlations

- With Centered data, we replace  $\bar{x}_1$  and  $\bar{x}_2$  with 0
- With Standardized data:

$$\widehat{Var}[x_1] = \frac{\sum (x_{1i} - \bar{x}_1)^2}{N-1} = \widehat{Var}[x_2] = \frac{\sum (x_{2i} - \bar{x}_2)^2}{N-1} = 1 \quad (6)$$

- Recalling  $\widehat{Cov}[x_1, y] = \frac{\sum (x_{1i} - \bar{x}_1)(y_i - \bar{y})}{N-1}$  and  $r_{y,x_1} = \frac{\widehat{Cov}[x_1, y]}{\sqrt{\widehat{Var}[x_1]}\sqrt{\widehat{Var}[y]}}$ :

$$\hat{\beta}_1 = \frac{r_{y,x_1} - r_{y,x_2}r_{x_1,x_2}}{1 - r_{x_1,x_2}^2} \quad (7)$$

- Note, if  $r_{x_1,x_2} = 0$ , it all collapses to  $\hat{\beta}_1 = r_{y,x_1}$

## Another way to describe it

- With orthogonal data, the regression coefficient is the ratio of covariance to variance.

$$\hat{\beta}_1 = \frac{\sum(x1_i - \bar{x1})(y_i - \bar{y})}{\sum(x1_i - \bar{x1})^2} = \frac{\widehat{\text{Cov}}[y, x1]}{\widehat{\text{Var}}[x1]} \quad (8)$$

- Some experimental designs can give us orthogonal data.
- We rarely find in “real life samples.”
- You see now why political scientists who collect data in the field are jealous of psychologists who design experiments so that their predictors are orthogonal.

# Outline

- 1 Choosing Variables: An Unsolved Problem
- 2 Non-Orthogonal Predictor Columns
- 3 Hat matrix**
- 4 Multiple  $R^2$  and adjusted  $R^2$
- 5 Which Variables are "Important"?
  - The Famous Ballantine Graph of Cohen, et al
  - Partial Correlation
  - Semi-partial Correlation Coefficients
  - relaimpo package
- 6 R Interactions (W&R notation).
- 7 Practice Problems

## Foreshadowing: The Hat Matrix

Recall the OLS estimator

- $\hat{\beta} = (X^T X)^{-1} X^T y$
- The part before the  $y$  is a critical part of the “hat matrix”.
  - $H = H(X^T X)^{-1} X^T$  is a  $N \times N$  weighting matrix
  - called hat because

$$\begin{aligned}\hat{y} &= X\hat{\beta} \\ \hat{y} &= X(X^T X)^{-1} X^T y \\ &= Hy\end{aligned}\tag{9}$$

- **The Hat Matrix** turns observed  $y$  into predictions  $\hat{y}$
- The Hat matrix is  $N \times N$ .
  - $(X^T X)$  is a  $p \times p$  square. The covariance of the  $X$ 's.
  - $X^T$  is  $p \times N$ .
  - $X(X^T X)^{-1} X^T$  is a matrix:  $N \times N$  (Because matrices of sizes:  $(N \times p)(p \times p)(p \times N)$ )

## Foreshadowing: The Hat Matrix ...

- Most regression diagnostics and analysis of predictive power focus on the Hat matrix.
- See Simon Wood, *Generalized Additive Models: An Introduction with R*.

# Outline

- 1 Choosing Variables: An Unsolved Problem
- 2 Non-Orthogonal Predictor Columns
- 3 Hat matrix
- 4 Multiple  $R^2$  and adjusted  $R^2$**
- 5 Which Variables are “Important”?
  - The Famous Ballantine Graph of Cohen, et al
  - Partial Correlation
  - Semi-partial Correlation Coefficients
  - relaimpo package
- 6 R Interactions (W&R notation).
- 7 Practice Problems

## We Need to Revise $R^2$

- If you were enthusiastic about  $R^2$  before, wait till you hear this awesome news:

### Ever Expanding R-Square

$R^2$  always gets bigger when more variables are added. Never gets smaller.

- The Sum of Squared errors can't get worse, so we are misled into adding more and more variables into a regression model.
- Various “corrections” have been suggested, either
  - a revision of  $R^2$  to include a penalty for more predictors, or
  - alternative “information criteria” that help us to gauge the predictive fit of the model.

## Adjusted $R^2$

- Proposal. Replace  $R^2 = 1 - \frac{ESS}{TSS}$  with

$$\text{adjusted } R^2 = 1 - \frac{ESS/(N - k - 1)}{TSS/(N - 1)} = \frac{MSE}{\widehat{\text{Var}}[y]} \quad (10)$$

$$\text{adjusted } R^2 = 1 - (1 - R^2) \frac{N - 1}{N - k - 1}$$

- *adjusted*  $R^2$  can go down as more variables are added
- Sometimes called “corrected  $R^2$ ” or “shrunk  $R^2$ ”.

## Other Suggestions: Information Criteria

- Akaike's Information Criterion (AIC)
- Bayesian Information Criterion (BIC)
- Mallows's  $C_p$  statistic

These are decisions that are similar to  $Adj R^2$  in a linear model, but they are applicable to maximum likelihood models (as opposed to ordinary least squares)

# Outline

- 1 Choosing Variables: An Unsolved Problem
- 2 Non-Orthogonal Predictor Columns
- 3 Hat matrix
- 4 Multiple  $R^2$  and adjusted  $R^2$
- 5 Which Variables are "Important"?**
  - The Famous Ballantine Graph of Cohen, et al
  - Partial Correlation
  - Semi-partial Correlation Coefficients
  - relaimpo package
- 6 R Interactions (W&R notation).
- 7 Practice Problems

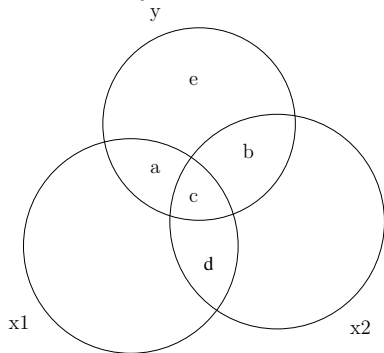
## Struggle to Assess Separate Effect of Predictors

- We fit a model, and we want to report to our sponsors “variable X1 is very influential”, and “variable X2 doesn't have a very substantial effect.”
- We want to gauge the relative “predictive power” of the predictors.
- We want these things; I think they don't exist, aren't very meaningful
- But, nevertheless, I have to introduce
  - “partial correlation” coefficients.
  - semi-partial correlations

## Ballantine Graph=Venn Diagram

Think of "variance accounted for" as areas in a Venn diagram (Cohen, et al, 3ed, p. 72)

Variance of  $y$ , Subdivided

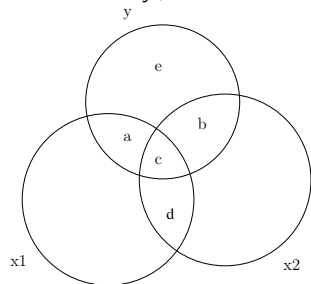


$$\text{Var}(y) = a + b + c + e$$

- $e$  is error term's variance
- $R^2$  is represented by area  $(a + b + c) / (a + b + c + e)$   
(separate plus shared parts = "accounted for")

## Areas c and d are troublemakers

Variance of  $y$ , Subdivided

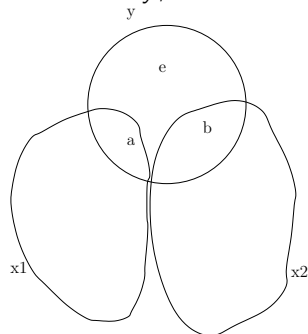


$$\text{Var}(y) = a + b + c + e$$

- $a$  and  $b$  are "uniquely" accounted for by predictors
- " $c$ " is overlapped explanatory power—we can't tell which causes what
- $c+d$  is "multicollinearity", overlap of 2 predictors

## All is Well if there's no Overlap

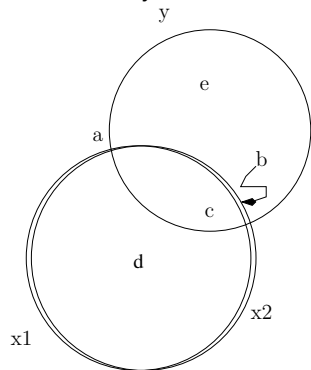
### Variance of $y$ , Subdivided



- Here,  $x_1$  and  $x_2$  are "orthogonal", completely uncorrelated.
- Separate regressions  $y \sim x_1$  and  $y \sim x_2$  would yield the same slope estimates as  $y \sim x_1 + x_2$
- Clearly, the separate effect of  $x_1$  is "a"
- However, that is not true if  $x_1$  and  $x_2$  overlap (Multicollinearity)

# Big Overlap=Bad Problem

Variance of  $y$ , Subdivided



- When 2 predictors “coincide”, it means we can’t distinguish them.
- We can’t say meaningfully that  $x_1$  has an effect that is separate from  $x_2$

# Partial Correlation

**partial correlation coefficient:**  $r_{y \cdot x_1 \cdot x_2}$  Correlation between  $y$  and  $x_1$ , after "accounting for" or "partialing out"  $x_2$

Can be described as a sequence of regressions:

- Fit:  $\hat{y}_i = \hat{c}_0 + \hat{c}_1 x_{2i}$  and  $\hat{x}_{1i} = \hat{d}_0 + \hat{d}_1 x_{2i}$
- Create "residuals"  $y_i^* = y_i - \hat{y}_i$  and  $x_{1i}^* = x_{1i} - \hat{x}_{1i}$
- We have "partialed out" the effect of  $x_2$ , so
- Correlation b/t  $y^*$  and  $x_{1i}^*$  equals the "partial correlation between  $x_1$  and  $y$ ."

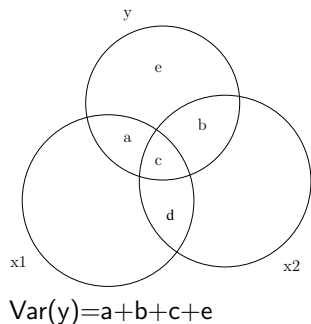
## Partial Correlation

- Lots of ways to write and re-organize this, Cohen, et al (p. 74):

$$pr_1 = r_{y \cdot x_1 \cdot x_2} = \frac{r_{y,x_1} - r_{y,x_2}r_{x_1,x_2}}{\sqrt{1 - r_{x_1,x_2}^2}\sqrt{1 - r_{y,x_2}^2}} \quad (11)$$

- Valid only if there are 2 predictors  $x_1$  and  $x_2$ .
- Look what happens if  $r_{x_1,x_2} = 0$  (orthogonal variables!)
- I think this is "harsh" toward  $x_1$ .  $x_2$  is given "all the influence it can account for" and  $x_1$  gets only "what it can account for after  $x_2$  takes everything it can grab."
- If  $x_1$  and  $x_2$  overlap substantially, meaning both have small partial correlations  $y$ .

## Partial in the Ballantine Graph



- Pearson correlations referred to as “zero order” correlations
- $r_{yx1}$  is correlation between  $y$  and  $x1$ , ignoring other variables
- Partial for  $X1 = pr_1^2 = r_{yx1.x2}^2 = (R^2 - r_{yx2}^2) / (1 - r_{yx2}^2)$
- In the diagram  $r_{yx1.x2}^2 = a / (a + e)$
- Partial for  $x2$  is same, just replace  $x2$  with  $x1$
- $r_{yx2.x1}^2 = b / (b + e)$

## I'd Never Calculated These Before 2010

- They were not emphasized in Econometrics or political science when I was a student
- More popular with behavioral researchers.
- I found various R packages, students ran into trouble with them. In 2013, the rockchalk package introduced a function "getPartialCor"

## Get Public Spending Regression

	M1	
	Estimate	(S.E.)
(Intercept)	356.182	(306.486)
ECAB	1.419**	( 0.430)
MET	-0.660	( 0.353)
GROW	0.572	( 0.425)
OLD	-1.855	( 7.137)
YOUNG	-6.675	( 7.481)
WEST	35.472*	( 13.771)
N	48	
RMSE	39.844	
$R^2$	0.599	
adj $R^2$	0.541	

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

## Partial Regression Matrix from `rockchalk::getPartialCor`

```
mod1pcor <- getPartialCor(mod1, dvonly = FALSE)  
round(mod1pcor, 4)
```

	EX	ECAB	MET	GROW	OLD	YOUNG	WEST
EX	-1.0000	0.4579	-0.2806	0.2055	-0.0406	-0.1380	0.3732
ECAB	0.4579	-1.0000	-0.1138	-0.0364	-0.4397	-0.5440	0.1137
MET	-0.2806	-0.1138	-1.0000	0.1404	-0.4716	-0.6449	0.0584
GROW	0.2055	-0.0364	0.1404	-1.0000	-0.4355	-0.2175	0.1699
OLD	-0.0406	-0.4397	-0.4716	-0.4355	-1.0000	-0.8268	0.3311
YOUNG	-0.1380	-0.5440	-0.6449	-0.2175	-0.8268	-1.0000	0.4053
WEST	0.3732	0.1137	0.0584	0.1699	0.3311	0.4053	-1.0000

## And We really want $pr^2$

```
round(mod1pcor^2, 4)
```

	EX	ECAB	MET	GROW	OLD	YOUNG	WEST
EX	1.0000	0.2097	0.0788	0.0422	0.0016	0.0190	0.1393
ECAB	0.2097	1.0000	0.0129	0.0013	0.1934	0.2960	0.0129
MET	0.0788	0.0129	1.0000	0.0197	0.2224	0.4159	0.0034
GROW	0.0422	0.0013	0.0197	1.0000	0.1896	0.0473	0.0289
OLD	0.0016	0.1934	0.2224	0.1896	1.0000	0.6836	0.1096
YOUNG	0.0190	0.2960	0.4159	0.0473	0.6836	1.0000	0.1642
WEST	0.1393	0.0129	0.0034	0.0289	0.1096	0.1642	1.0000

Compare effect of Young and West.

## Conduct "regression on residuals" to verify

```
m1 <- lm(EX ~ MET + GROW + YOUNG + OLD + WEST, data=dat)
EXStar <- resid(m1)
m2 <- lm(ECAB ~ MET + GROW + YOUNG + OLD + WEST, data=dat)
ECABStar <- resid(m2)
m3 <- lm(EXStar ~ ECABStar)
hopePcor <- summary(m3)
hopePcor$r.squared
```

```
[1] 0.2097133
```

- Oh, Sweet mystery of life at last I've found you! My regression result partial correlation matrix

## It seems like we made progress, but we really didn't

- Seems like we have a way to summarize each variable's separate "explanatory power"
- But estimates not robust against changes in columns of the dataset
- Compare  $pr^2$  with fewer variables

```
mod2 <- lm(EX ~ YOUNG + WEST, data = dat)
mod2pcor <- getPartialCor(mod2, dvonly = FALSE)
round(mod2pcor^2, 4)
```

	EX	YOUNG	WEST
EX	1.0000	0.2041	0.2507
YOUNG	0.2041	1.0000	0.2020
WEST	0.2507	0.2020	1.0000

Compare effect of Young and West.

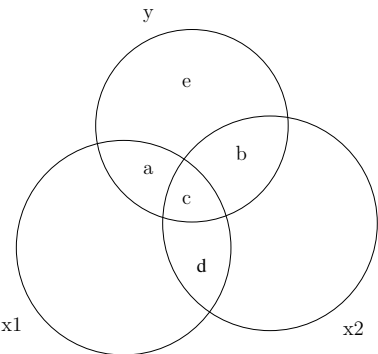
## My conclusion about partial correlation estimates

- I avoid partial correlations. Because:
- The conclusions depend VITALLY on which columns from  $X$  are included in the calculations
- If I knew the "Oracle model," these might be useful.
- These don't help in choosing variables for a regression (because they are calculated on the initial assumption of which variables need to be included in the calculation of  $pr^2$ )
- So I'll file this in the category of "something some people know, but I don't use".

## Semi-partial Correlation Coefficients

- I never heard term "semi-partial correlation" before reading Cohen, et al, but have heard other names like delta- $R^2$ .
- Simplest explanation I've found
  - Fit the whole model, with all the predictors, get  $R^2$
  - Remove one predictor, re-estimate, calculate  $\Delta R_j^2$  (change in  $R^2$  from dropping  $j'$ th variable).
  - The  $sr_{x_j}^2$  is the change in the  $R^2$  that results when  $x_j$  is removed from the full model.
- My "gut" reaction: still seems "harsh".  $x_1$  gets credit only for "what it can account for by itself", the part  $c$  is "lost" as neither  $x_1$  nor  $x_2$  is given "credit" for it.

## Semi-Partial in the Ballantine Graph



$$\text{Var}(y) = a + b + c + \text{effect}$$

- Pearson correlations referred to as “zero order” correlations
- zero order correlation  $r_{yx1}$  is correlation between  $y$  and  $x1$ , ignoring other variables
- Notice the semipartial can be seen as
  - $sr_1^2 = R^2 - r_{yx2}^2 = a / (a + b + c + e)$
  - $sr_2^2 = R^2 - r_{yx1}^2 = b / (a + b + c + e)$

## Recently added to rockchalk: getDeltaRsquare for Semi-partial $R^2$

```
EXfull <- lm(EX ~ ECAB + MET + GROW + YOUNG + OLD + WEST, data=dat)
getDeltaRsquare(EXfull)
```

The deltaR-square values: the change in the R-square observed when a single term is removed.  
Same as the square of the 'semi-partial correlation coefficient'

	deltaRsquare
ECAB	0.106307818
MET	0.034249030
GROW	0.017664838
YOUNG	0.007779037
OLD	0.000660166
WEST	0.064831204

- Which Variables are "Important"?
- Semi-partial Correlation Coefficients

## Calculate $\Delta R^2$ Explicitly To verify

$sr_{ECAB}^2$ . Check the change in R-square that results when ECAB is removed

```
EXfull <- lm(EX ~ ECAB + MET + GROW + YOUNG + OLD + WEST, data=dat)
EXnoECAB <- lm(EX ~ MET + GROW + YOUNG + OLD + WEST, data=dat)
(deltaRecab <- summary(EXfull)$r.square - summary(EXnoECAB)$
  r.square)
```

```
[1] 0.1063078
```

## Semi-Partial $r^2$ is Unstable (Dependent on inclusion of columns in data frame)

- semi-partial  $r^2$  jumps around depending on what other variables are in the model.
- Since we don't know for sure which ones belong, hard to hold much faith in it (that's where I stand now, despite recent changes).
- Others are seeking ways to improve results.

## A "New-ish" Proposal

*Grömping, U. (2006). Relative Importance for Linear Regression in R: The Package relaimpo. Journal of Statistical Software, 17(1), 1–27.*

*Grömping, U. (2007). Estimators of Relative Importance in Linear Regression Based on Variance Decomposition. The American Statistician, 61, 139-147. doi:10.1198/000313007X188252*  
*Discusses a number of criteria to decide how important each variable might be, how much variance it accounts for.*

## Grömping Proposes

- Consider possible range of “variance accounted for” estimates, semi-partial correlations ( $R^2$  changes) when a variable is put into a model.
- $sr^2$  biggest when the variable is by itself in the model
- $sr^2$  smallest when competing with other predictors in the model
- “lmg” procedure enters regression variables in all possible orders, and then averages the changes in the  $R^2$ .
- Implemented in the “relaimpo” package (see also package hier.part)

- Which Variables are "Important"?

- relaimpo package

## Load relaimpo

```
library(relaimpo)  
EXFull <- lm(EX ~ ECAB+ MET + GROW + YOUNG + OLD + WEST, data=dat)
```

# Test out relaimp

```
(EXFull.relimp <- calc.relimp(EXFull, type=c("lmg")))
```

Response variable: EX

Total response variance: 3456.829

Analysis based on 48 observations

6 Regressors:

ECAB MET GROW YOUNG OLD WEST

Proportion of variance explained by model: 59.94%

Metrics are not normalized (rela=FALSE).

Relative importance metrics:

	lmg
ECAB	0.28416663
MET	0.03492833
GROW	0.08033756
YOUNG	0.06748339
OLD	0.01429318
WEST	0.11817898

Average coefficients for different model sizes:

1X

2Xs

3Xs

4Xs

5Xs

6Xs

## Test out relaimp ...

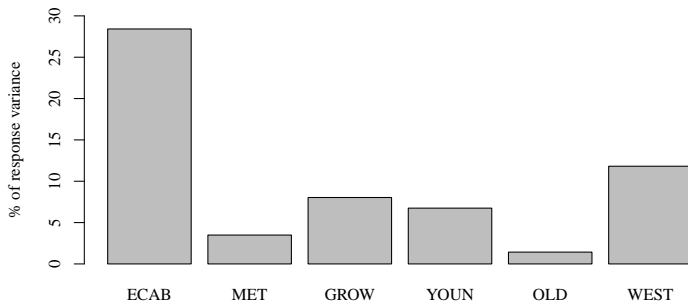
```
ECAB  1.73287208  1.7961836  1.8130491  1.7476127  1.6128010  1
      .4185020
MET   0.09872711 -0.1759874 -0.4565307 -0.6267283 -0.6783790
      -0.6601532
GROW  1.26246694  1.1260393  0.9792009  0.8118053  0.6649301  0
      .5715895
YOUNG -8.02338567 -7.5065678 -7.4332358 -7.2917564 -6.9958279
      -6.6746586
OLD   -0.83908956 -0.6832795 -0.7204064 -1.1259130 -1.5630752
      -1.8550734
WEST  43.45833333 46.5175733 45.8928772 42.4128710 38.6907909 35
      .4723358
```

```
plot(EXFull.relimp)
```

# lm model averages together the $sr^2$

## Relative importances for EX

### Method LMG



$R^2 = 59.94\%$ , metrics are not normalized.

# Outline

- 1 Choosing Variables: An Unsolved Problem
- 2 Non-Orthogonal Predictor Columns
- 3 Hat matrix
- 4 Multiple  $R^2$  and adjusted  $R^2$
- 5 Which Variables are “Important”?
  - The Famous Ballantine Graph of Cohen, et al
  - Partial Correlation
  - Semi-partial Correlation Coefficients
  - relaimpo package
- 6 R Interactions (W&R notation).**
- 7 Practice Problems

## R Uses W&R Notation

- Wilkinson & Rogers introduced a system of notation, which R uses
- A simple additive model:

```
mod <- lm(y ~ x1 + x2 + x3)
```

asks R to estimate the theoretical model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + e_i$$

- Note you get an intercept “for free” (to suppress that, insert -1 or 0 as a predictor)
- W&R notation uses “:”, “\*” and “/” in special ways. .

## In W&R notation, "\*" Means Multiplicative Interaction

- Suppose  $x_1$  and  $x_2$  are numeric predictors. This notation

```
mod <- lm(y ~ x1*x2, data = dat)
```

- Is the same as

```
mod <- lm(y ~ x1 + x2 + x1:x2, data = dat)
```

- R always inserts the "main effects" that sit underneath an interaction. You ask for  $x_1*x_2$ , get estimates for  $x_1$ ,  $x_2$ , and  $x_1:x_2$ .
- In R output, the product term " $x_{1i} \times x_{2i}$ " will appear as " $x_1 : x_2$ "
- Interactions can fit together with other predictors as well, of course:

```
mod <- lm(y ~ x1 + x2*x3)
```

Asks R to estimate  $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{2i} \times x_{3i} + e_i$

It is considered gauche to write the request explicitly:

```
mod <- lm(y ~ x1 + x2 + x3 + x2:x3)
```

## In W&R notation, "\*" Means Multiplicative Interaction ...

In the design matrix,  $x_2 : x_3$  is a new constructed column, the product of  $x_2$  and  $x_3$ .

- Combine several:

```
mod <- lm(y ~ x1*x2*x3)
```

Asks R to fit:  $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{1i} \times x_{2i} + \beta_5 x_{1i} \times x_{3i} + \beta_6 x_{2i} \times x_{3i} + \beta_7 x_{1i} \times x_{2i} \times x_{3i} + e_i$

## Interact a numeric with a categorical predictor

```
mod2 <- lm(y ~ xnum1 * z, data = dat)
mod2sum <- summary(mod2)
coef(mod2sum)
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.01095327	0.1581058	-0.0692781	0.94491862
xnum1	0.28981053	0.1681968	1.7230442	0.08824016
zB	0.27670253	0.2417953	1.1443668	0.25544018
zC	0.37603376	0.2569725	1.4633227	0.14678733
zD	-0.27113912	0.2088686	-1.2981327	0.19748670
xnum1:zB	0.39229326	0.2801136	1.4004794	0.16473350
xnum1:zC	0.11537877	0.3911294	0.2949887	0.76866715
xnum1:zD	0.27251832	0.2047313	1.3311026	0.18644453

That estimated all of the parallel lines, plus 4 additional terms.

## Now let's see what happens with "\*"

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 zB_i + \beta_3 zC_i + \beta_4 zD_i + \beta_5(x \cdot zA_i) + \beta_6(x \cdot zB_i) + \beta_7(x \cdot zC_i) + \beta_8(x \cdot zD_i) + e_i \quad (12)$$

- Think of  $\beta_0$  and  $\beta_1$  as the “baseline” estimates, for everybody.
- Then we have to look specifically at groups  $A$ ,  $B$ ,  $C$  and  $D$  to figure out what their predicted values need to be.
- Group  $A$ , which implies  $zB_i = zC_i = zD_i = 0$ , so

$$\begin{aligned} \text{Group } A : \hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_5 x_i \\ &= \hat{\beta}_0 + (\hat{\beta}_1 + \hat{\beta}_5) x_i \end{aligned} \quad (13)$$

- Group  $B$ , which implies  $zB_i = 1$  and  $zC_i = zD_i = 0$ . Predicted values would be:

Now let's see what happens with "\*" ...

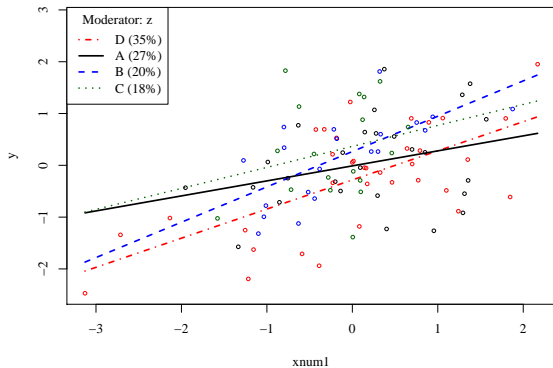
$$\begin{aligned} \text{Group } B : \hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 z B_i + \hat{\beta}_6 (x \cdot z B_i) \\ &= (\hat{\beta}_0 + \hat{\beta}_2) + (\hat{\beta}_1 + \hat{\beta}_6) x_i \end{aligned} \quad (14)$$

- $\beta_2$  is the "intercept shift" due to membership in group  $B$
- $\beta_6$  is the "slope shift" due to membership in group  $B$
- Lets check, and then plot, the predicted values for the 4 levels of  $z$ . Here's the predictOMatic() output. We don't really need this now, but we will later.

```
mod4pmo <- predictOMatic(mod2, predVals = list(z = levels(dat$z),
  xnum1 = range(dat$xnum1)))
mod4pmo
```

	xnum1	z	fit
1	-3.128437	A	-0.9176073
2	-3.128437	B	-1.8681696
3	-3.128437	C	-0.9025288
4	-3.128437	D	-2.0413029
5	2.168475	A	0.6174936
6	2.168475	B	1.7448742
7	2.168475	C	1.2437233
8	2.168475	D	0.9373036

## rockchalk::plotSlopes output



```
plotSlopes(mod2, plotx = "xnum1", modx = "z")
```

# Outline

- 1 Choosing Variables: An Unsolved Problem
- 2 Non-Orthogonal Predictor Columns
- 3 Hat matrix
- 4 Multiple  $R^2$  and adjusted  $R^2$
- 5 Which Variables are “Important”?
  - The Famous Ballantine Graph of Cohen, et al
  - Partial Correlation
  - Semi-partial Correlation Coefficients
  - relaimpo package
- 6 R Interactions (W&R notation).
- 7 Practice Problems**

# Problems

- 1 If you want to see an example of really bad “overlapping” variables, download the cystic fibrosis data set from the DataSets folder. Try to predict the DV “pemax” as a function of these variables age + sex + height + weight. Wow. That’s discouraging.
  - 1 Make a regression table to summarize the results and try to write a paragraph about the effect of height in contrast with that of weight.
  - 2 Write out the formula for the theory you are fitting, and the equation for the predicted values.
  - 3 Conduct a “fancy” t test to find out if the effect of height is different from that of weight. Computing hint. If the fitted model is “mod1”, run “vcov(mod1)” to review the var/covar matrix, then pick the ones you need for the denominator. You can do this calculation with a calculator, but you can make R do it if you learn how to grab rows and columns from a matrix. I can help you with that if you like.

## Problems ...

- 2 Continue with `cystfibr`. Run a regression with each independent variable separately entered.  
Compare the findings of these one-at-a-time models, try to make sense out of the “statistical significance” of the coefficients.  
You have to stare at this one a while, it really puzzled me at first. It helped me to run `termplot` so I could see the problem graphically.
- 3 Test out R’s “`anova`” function to to a test of the null hypothesis that several coefficients are actually 0. Any data set with several predictors will do. Suppose the “unrestricted” model is `m1`, and the restricted model is `m2`

```
m1 <- lm(y ~ x1 + x2 + x3 + x4 + x5 + x6, data=dat)
m2 <- lm(y ~ x1 + x5 + x6, data=dat)
anova(m2, m1, test="F")
```

Your idea here is that variables `x3` and `x4` do not belong, so you need to test the null that  $\beta_3 = \beta_4 = 0$ . Can you interpret the results?

## Problems ...

- 4 Practice making professional looking regression tables. Make a table that summarizes the “full” model from problem 1 and also includes columns for one or two of the “partial” models in problem 2.

To do that work, I’ve used some R functions that generate  $\text{\LaTeX}$  tables. The `oureg` function in `new-ish` `rockchalk` can create html files, which Word can import. If you like word.

## Problems ...

- 5 While you have that `cystfibr` data open, here are a few fun things to do.
  - 1 Create a correlation matrix showing the bivariate (one-on-one) correlations of all the predictors. (code hint: Run `cor(dat)` if your data set is named `dat`. To round numbers after 3 decimal values, do `round(cor(dat), 3)`. Then try to think of some way to wrestle that into a document. (For me, that's a hassle).
  - 2 Regress each IVs on the other IVs. Do the  $R_j^2$  of these “auxiliary” regressions differ much from the bivariate correlations?
  - 3 Get the squared partial and semi-partial correlation coefficients. I implemented those in `rockchalk`, `getPartialCor` and `getDeltaRsquare`, but I'm sure you'll find other packages with similar
  - 4 If you have `relaimpo`, that's interesting. Run `calc.relimp(mod1)`. It gives a somewhat different conclusion than the correlation coefficients. `calc.relimp` has a lot of options I've not tested yet.