清华大学统计学辅修课程

Linear Regression Analysis

Lecture 9- Extra Sum of Squares & General Linear Tests & Multicollinearity & Polynomial Regression

周在莹

清华大学统计学研究中心

http://www.stat.tsinghua.edu.cn

Topic 1:
Extra Sum of Squares
&
General Linear Tests

Outline

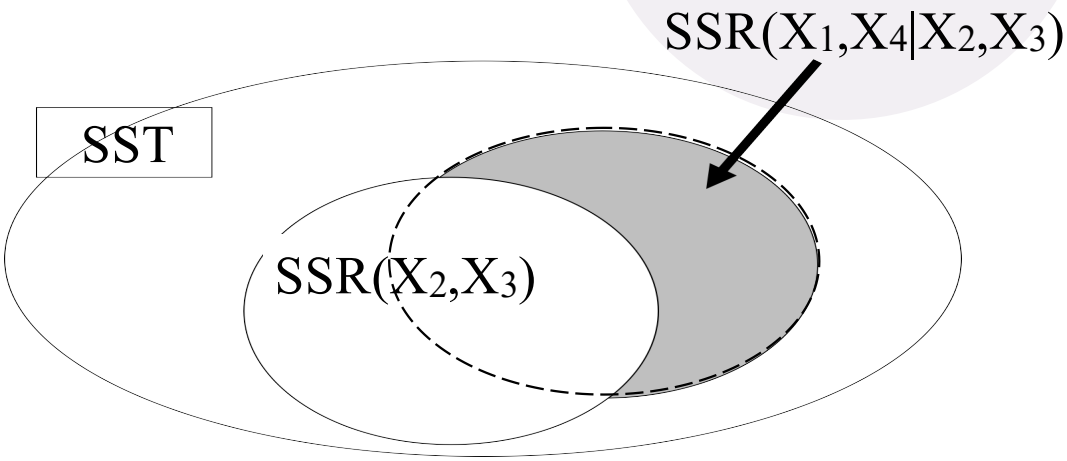
- Extra Sums of Squares with Applications
- General Linear Test (Review Section 2.8)
  - Testing single $\beta_k = 0$
  - Testing several $\beta_k = 0$
  - Other general linear tests
- Using and Interpreting R^2 and Partial- R^2, Partial Correlations
- Standardized Regression and Interpretation of Coefficients

Extra Sums of Squares

- Basic Ideas
  - *Extra SS* measure the marginal reduction in the error sum of squares (*SSE*) from the addition of a group of predictor variables to the model
- Examples
  - $SSR(X_1, X_2, X_3)$ is the total variation explained by X_1, X_2 and X_3 in a model
  - $SSR(X_1 | X_2)$ is the additional variation explained by X_1 when added to a model already containing X_2
  - $SSR(X_1, X_4 | X_2, X_3)$ is the additional variation explained by X_1 and X_4 when added to a model already containing X_2 and X_3

Extra SS Illustration and Calculation

- *Extra SS* represents the part of the *SSE* that is explained by an added group of variables that was not previously explained by the rest
- Calculation
  - $SSR(X_1 \mid X_2) = SSE(X_2) - SSE(X_1, X_2)$
  - $SSR(X_1, X_4 \mid X_2, X_3) = SSE(X_2, X_3) - SSE(X_1, X_2, X_3, X_4)$



Add One Variable at a Time: Type I SS

- Regression SS can be partitioned into pieces (in any order):

$$SSR(X_1,X_2,X_3,X_4) = SSR(X_1) + SSR(X_2 \mid X_1) + SSR(X_3 \mid X_1, X_2) + SSR(X_4 \mid X_1, X_2, X_3)$$

- This particular breakdown is called Type I sums of squares (variables added in order), also called "sequential" SS
- 'Extended' AVONA table of R: Row for "Model" or "Regression" becomes $p - 1$ rows, in terms of Type I SS and MS
- Numerator df is 1 for each of these F tests

Types of Sums of Squares

- Suppose we have a model with two factors and the terms appear in the order *A, B, AB*
- *SSE*(*A,B,AB*) is the residual sum of squares fitting the whole model, *SSE*(*A*) is the residual sum of squares fitting just the main effect of *A*, and *SSE*(1) is the residual sum of squares fitting just the mean
- The three types of SS are defined as follows:

Term	Type I SS	Type II SS	Type III SS
A	SSR(A) = SSE(1)-SSE(A)	SSR(A|B) = SSE(B)-SSE(A,B)	SSR(A|B,AB) = SSE(B,AB)-SSE(A,B,AB)
B	SSR(B|A) = SSE(A)-SSE(A,B)	SSR(B|A) = SSE(A)-SSE(A,B)	SSR(B|A,AB) = SSE(A,AB)-SSE(A,B,AB)
AB	SSR(AB|A,B) = SSE(A,B)-SSE(A,B,AB)	Yates: no significant interaction is assumed →	SSR(AB|A,B) = SSE(A,B)-SSE(A,B,AB)

Note on Types of SS

- The notation of Type I, II and III SS seems to have been introduced into statistics from the *SAS* package but is now widespread
- *SAS* and *SPSS* use Type III SS as their default, while functions that ship with *R* use Type I SS. This can lead to different results when analyzing the same data with different statistical packages
- KNNL uses Type I SS, where decomposition works:

$$SSR(\text{A,B,AB}) = SSR(\text{A}) + SSR(\text{B}|\text{A}) + SSR(\text{AB}|\text{A,B})$$

- Type III sums of squares refers to variables added last. These do NOT add to the *SSR*

General Linear Tests

- Different ways to look at the comparison of two models
- Recall: Look at the difference in
  - *SSE* (reduce unexplained SS)
  - *SSR* (increase explained SS)

  between a full and reduced model
- Because $SSR+SSE=SST$, these two comparisons are equivalent

General Linear Tests

- Models we compare are hierarchical/nested in the sense that one (the full model) includes all of the explanatory variables of the other (the reduced model)

- We can compare models with different explanatory variables such as
  - 1. X_1, X_2 *vs* X_1
  - 2. X_1, X_2, X_3, X_4, X_5 *vs* X_1, X_2, X_3

  (Note the first model includes all X's of the second)

General Linear Tests

- We will get an F test that compares the two models
  - Full Model: All variables / parameters
  - Reduced Model: Apply NULL hypothesis to full model
- We are testing the null hypothesis that the regression coefficients for the *extra* variables are all zero
- For X_1, X_2, X_3, X_4, X_5 *vs* X_1, X_2, X_3
  - H_0: $\beta_4 = \beta_5 = 0$
  - H_1: at least one of $\beta_4, \beta_5 \neq 0$

General Linear Tests

$$F^* = \frac{(\text{SSE(R)} - \text{SSE(F)})/(df_E(\text{R}) - df_E(\text{F}))}{\text{SSE(F)}/df_E(\text{F})}$$

- Under H_0:

$$F^* \sim F_{df_E(R)-df_E(F),df_E(F)}$$

- Degrees of freedom for the F statistic are the number of *extra* variables and the df_E for the larger model

- Suppose n = 100 and we compare models with

$$X_1, X_2, X_3, X_4, X_5 \ \ vs \ \ X_1, X_2, X_3$$

  - Numerator df is 2
  - Denominator df is n-6 = 94

Application of the Extra SS

- $SSE(X_1, X_2, X_3, X_4, X_5)$ is the *SSE* for the *full* model
- $SSE(X_1, X_2, X_3)$ is the *SSE* for the *reduced* model
- $SSR(X_4, X_5 \mid X_1, X_2, X_3)$ is the difference in the *SSE*s (reduced minus full)

$$SSE(X_1, X_2, X_3) - SSE(X_1, X_2, X_3, X_4, X_5)$$

- In terms of either *SSR* or *SSE*

$$SSR(X_4, X_5 \mid X_1, X_2, X_3)$$

$$= SSR(X_1, X_2, X_3, X_4, X_5) - SSR(X_1, X_2, X_3)$$

$$= SSE(X_1, X_2, X_3) - SSE(X_1, X_2, X_3, X_4, X_5)$$

Perform the F Test

$$F^* = \frac{(\text{SSE(R)} - \text{SSE(F)})/(df_E(\text{R}) - df_E(\text{F}))}{\text{SSE(F)}/df_E(\text{F})}$$

- Numerator : $SSE(X_4, X_5 \mid X_1, X_2, X_3)/2$
- Denominator : $SSE(X_1, X_2, X_3, X_4, X_5)/(n\text{-}6)$
- Under H_0, $F \sim F(2, n\text{-}6)$
- Reject if the P-value ≤ 0.05 and conclude that either X_4 or X_5 or both contain additional information useful for predicting Y in a linear model that also includes X_1, X_2 and X_3

Examples

- Predict bone density using age, weight and height; does diet (sugars, protein) add any useful information?
- Predict GPA using 3 HS grade variables; do SAT scores add any useful information?
- Predict yield of an industrial process using temperature and pH; does the supplier of the raw material (categorical) add any useful information?

Special Cases of Extra SS or GLT

- Compare models that differ by one explanatory variable:

$$F_{1,n-p} = t^2_{n-p}$$

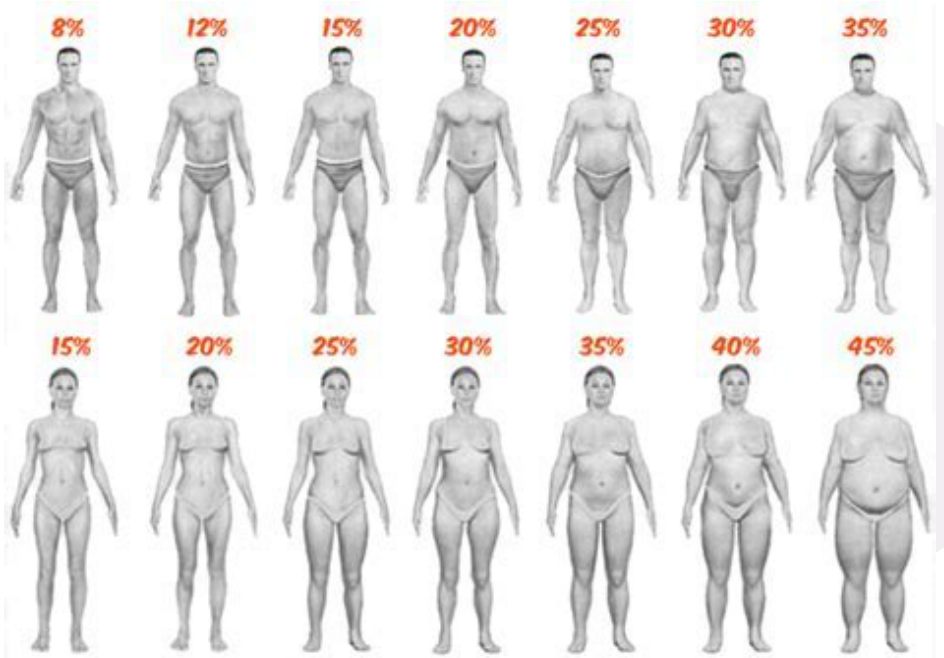
- Individual parameter *t*-tests are equivalent to the general linear test based on

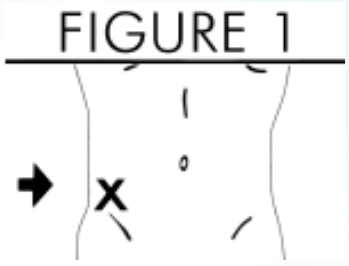
$$SSR(X_i \,|X_1, \ldots, X_{i\text{-}1}, X_{i+1}, \ldots, X_{p\text{-}1})$$

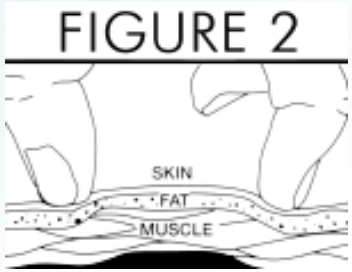
- We will come back later

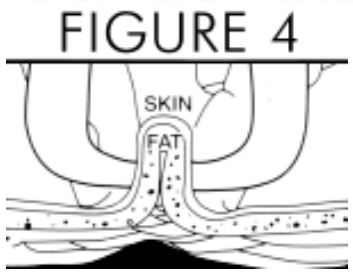
Body Fat Example (KNNL p256)

- 20 healthy female subjects
- Y is body fat
- X_1 is triceps skin fold thickness
- X_2 is thigh circumference
- X_3 is midarm circumference
- Underwater weighing is the "gold standard" used to obtain Y







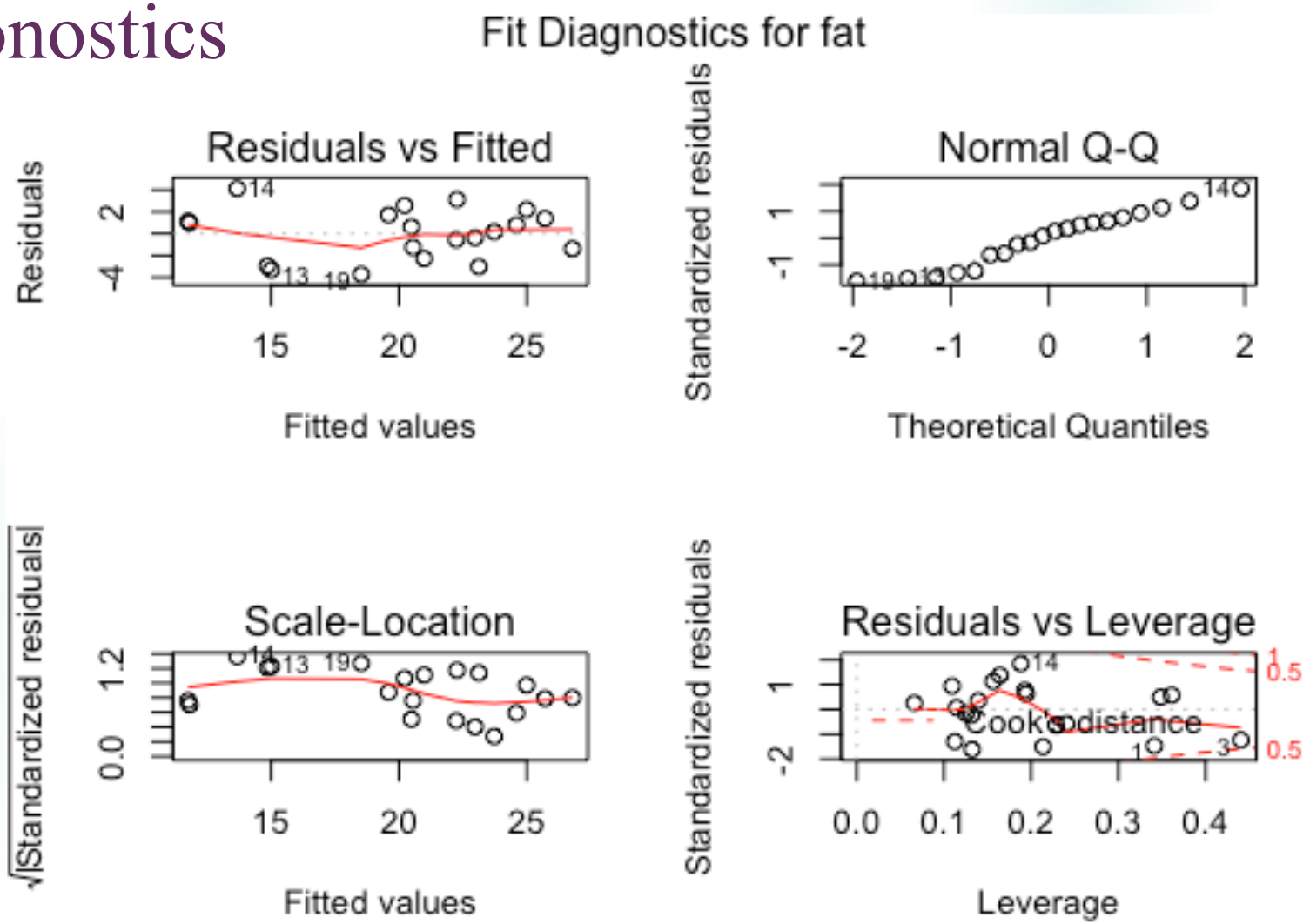


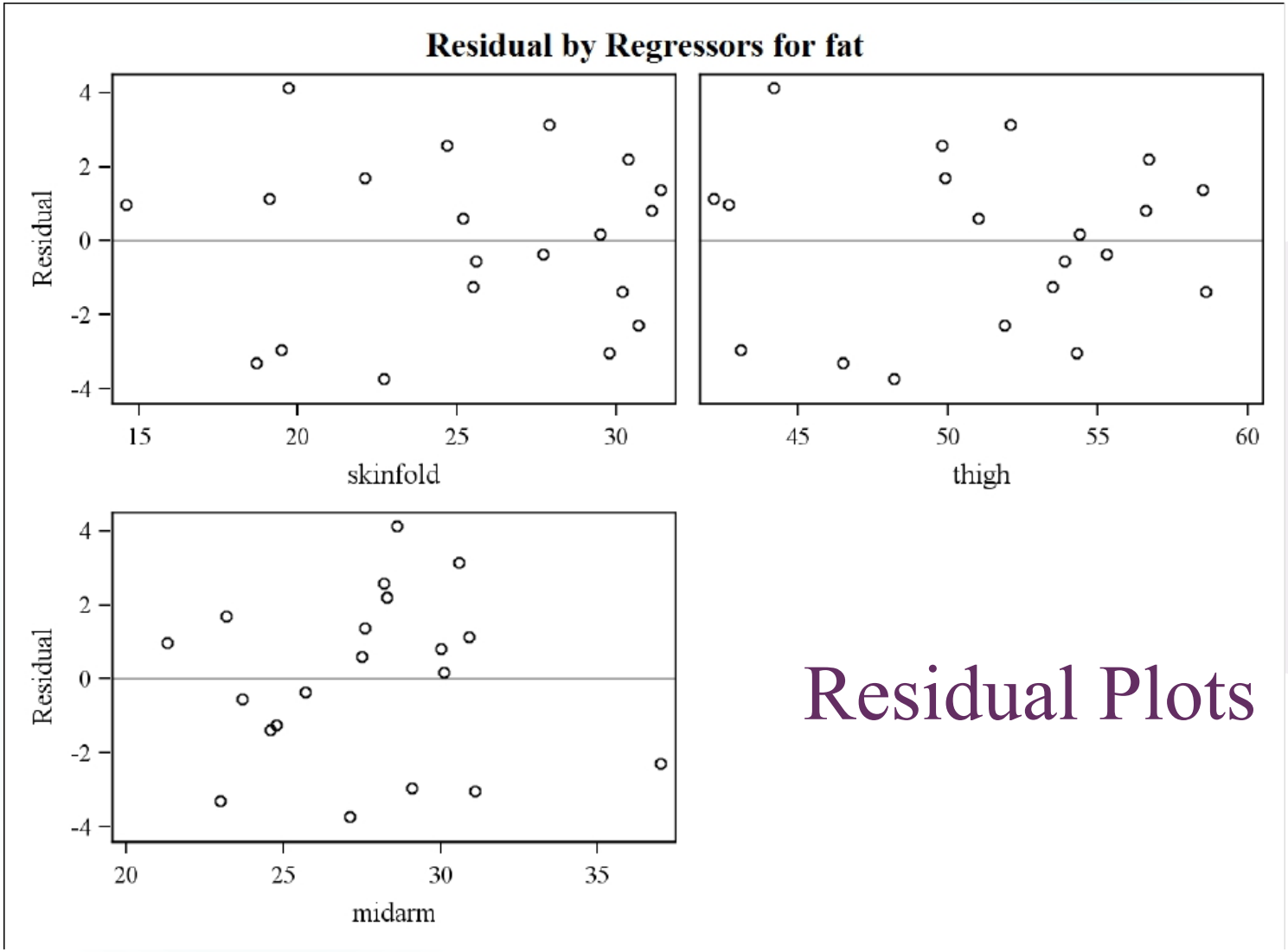
R Code


```
# Input and data check
> a1 = read.table(....) # "CH07TA01.txt"
> colnames(a1) = c("skinfold", "thigh", "midarm", "fat")
> View(a1)

# Fit model
> reg1 <- lm(fat ~ skinfold + thigh + midarm, data=a1)
> summary(reg1)
> anova(reg1)
> plot(reg1)
```

Diagonostics





Residual Plots

Output

None of the individual t-tests are significant?!!

Analysis of Variance Table

Response: fat

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
skinfold	1	352.27	352.27	57.2768	1.131e-06	***
thigh	1	33.17	33.17	5.3931	0.03373	*
midarm	1	11.55	11.55	1.8773	0.18956	
Residuals	16	98.40	6.15			

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	117.085	99.782	1.173	0.258
skinfold	4.334	3.016	1.437	0.170
thigh	-2.857	2.582	-1.106	0.285
midarm	-2.186	1.595	-1.370	0.190

Residual standard error: 2.48 on 16 degrees of freedom
Multiple R-squared: 0.8014, Adjusted R-squared: 0.7641
F-statistic: 21.52 on 3 and 16 DF, p-value: 7.343e-06

Group of predictors helpful in predicting percent body fat

It seems that this set of variables as a whole are helpful but each individual is not!

Look at This Using Extra SS


```
> library("car")
> Anova(reg1, type="II")   # Type II tests
> Anova(reg1, type="III")  # Type III tests
```

Anova Table (Type II tests)

Response: fat

	Sum Sq	Df	F value	Pr(>F)
skinfold	12.705	1	2.0657	0.1699
thigh	7.529	1	1.2242	0.2849
midarm	11.546	1	1.8773	0.1896
Residuals	98.405	16		

Anova Table (Type III tests)

Response: fat

	Sum Sq	Df	F value	Pr(>F)
(Intercept)	8.468	1	1.3769	0.2578
skinfold	12.705	1	2.0657	0.1699
thigh	7.529	1	1.2242	0.2849
midarm	11.546	1	1.8773	0.1896
Residuals	98.405	16		

Analysis of Variance Table

Response: fat

	Df	Sum Sq	Pr(>F)
skinfold	1	352.27	1.131e-06 ***
thigh	1	33.17	0.03373 *
midarm	1	11.55	0.18956
Residuals	16	98.40	6.15

- Notice how different these SS are for skinfold and thigh

Interpretation

- Fact: Type I and Type III SS are calculated in very different ways
  - Type I SS – Fit in order specified in model
  - Type III SS – extra SS for variable fitted last
- Fact: If we reorder the variables in the model statement we will get
  - Different Type I SS
  - The same Type III SS
- Could variables be explaining same SS and canceling each other out?

Run Additional Models

- Rerun with skinfold as the explanatory variable

```
> reg2 <- lm(fat ~ skinfold, data=a1)
> summary(reg2)
> anova(reg2)
```

```
Coefficients:
             Estimate  Std. Error t value Pr(>|t|)
(Intercept)  -1.4961    3.3192   -0.451    0.658
skinfold      0.8572    0.1288    6.656    3.02e-06 ***

Residual standard error: 2.82 on 18 degrees of freedom
Multiple R-squared:  0.7111, Adjusted R-squared:  0.695
F-statistic:  44.3 on 1 and 18 DF,  p-value: 3.024e-06
```

```
Analysis of Variance Table

Response: fat
           Df Sum Sq Mean Sq F value    Pr(>F)
skinfold    1  352.27  352.27  44.305 3.024e-06 ***
Residuals 18 143.12    7.95
```

skinfold by itself is a highly significant linear predictor

Use GLT to See if Other Predictors Contribute beyond skinfold


```
> anova(reg2, reg1)
```

```
Analysis of Variance Table

Model 1: fat ~ skinfold
Model 2: fat ~ skinfold + thigh + midarm
  Res.Df    RSS  Df Sum of Sq      F     Pr(>F)
1     18 143.120
2     16  98.405  2    44.715 3.6352 0.04995 *
```

```
Analysis of Variance Table

Response: fat
           Df Sum Sq Pr(>F)
skinfold    1 352.27 1.131e-06 ***
thigh       1  33.17  0.03373 *
midarm      1  11.55  0.18956
Residuals 16  98.40  6.15
```

- Yes, they do help after skinfold is in the model
- Perhaps the best model includes only two predictors
- Use GLT to assess midarm

Use GLT to Assess midarm

- Test whether the variable 'midarm' can be dropped from the model

```
> reg3 <- lm(fat ~ skinfold + thigh,  data=a1)
> anova(reg3, reg1)
```

Analysis of Variance Table

Model 1: fat ~ skinfold + thigh
Model 2: fat ~ skinfold + thigh + midarm

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	17	109.951				
2	16	98.405	1	11.546	1.8773	0.1896

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	117.085	99.782	1.173	0.258
skinfold	4.334	3.016	1.437	0.170
thigh	-2.857	2.582	-1.106	0.285
midarm	-2.186	1.595	-1.370	0.190

Residual standard error: 2.48 on 16 degrees of freedom
Multiple R-squared: 0.8014, Adjusted R-squared: 0.7641
F-statistic: 21.52 on 3 and 16 DF, p-value: 7.343e-06

- With skinfold and thigh in the model, midarm is not a significant predictor
- It is just the *t*-test for this coefficient in the full model

Be Careful with Our Interpretations!

- The equivalence of the *t*-test to the *F*-test implies

when X_k is orthogonal to $R(X_{-k})$

$$SSR(X_k|X_{-k})=\frac{b_k^2}{[(X'X)^{-1}]_{kk}}=||X_k||^2 b_k^2$$

$$\left(F=\frac{SSR(X_k|X_{-k})}{MSE}=t^2=\left(\frac{b_k}{s(b_k)}\right)^2, s^2(b_k)=[(X'X)^{-1}]_{kk}MSE\right)$$

  - The *t*-test is a test for the *marginal* significance of the X_3 predictor *after* the other predictors X_1 and X_2 have been taken into account
  - It does NOT test for the significance of the relationship between the response Y and the predictor X_3 alone
- Recall that β_i is called *partial* regression coefficient. It represents the change in $E(Y)$ associated with a one-unit increase in X_i when all other predictors are held constant
- Partial *F* test and overall *F* test

Other Uses of GLT: Test Linear Hypothesis

- The test statement can be used to perform a significance test for any hypothesis involving a linear combination of the regression coefficients, i.e. a contrast

- Examples
  - H_0: $\beta_2 = \beta_3$
  - H_0: $\beta_2 = 2$
  - H_0: $\beta_2 - 3\beta_3 = 12$

```
> glt1 <- lm(fat ~ skinfold + I(thigh + midarm), data=a1)
> anova(glt1, reg1)
```

```
> glt2 <- lm(fat ~ skinfold + offset(2*thigh) + midarm , data=a1)
#with known coefficient 1 rather than an estimated coefficient
> anova(glt2, reg1)
```

```
> library(car)
> linearHypothesis(reg1, "thigh - midarm = 0")
> linearHypothesis(reg1, " thigh  = 2")
> linearHypothesis(reg1, "thigh - 3*midarm = 12")
```

Note on GLT

- $H_0: C\beta = t$

$$F = \frac{(C\hat{\beta}-t)^T\ (C(X^TX)^{-1}C^T)^{-1}(C\hat{\beta}-t)}{qs^2}\ \sim F_{q,n-p}$$

where q is the row of contrast matrix C, the number of contrasts

- Most of the time, the alternative will be that at least one of the variables in the null group is important
- Often looking to *"fail to reject"* when performing a test like this – our goal is to eliminate unnecessary variables
- This means POWER / sample size must be a consideration! If our sample size is too small, we may incorrectly remove variables

Discussion: CS Example Revisited

- Recall that testing whether HSS, SATM, SATV as a group are important when added to model containing HSM and HSE
  - F < 1 so no need to even look up the P value; fail to reject. With 224 data points, we likely have the power required to conclude that the three variables are not useful in the model that already contains HSM and HSE
  - P-value is 0.4361 (as long as its > 0.1 and the sample size is reasonably large, we can discard the additional variables)
- How would we test...
  - Importance of HSS in addition to the rest
  - Importance of SAT's added to HS's
  - Importance of HSE after HSM
- Can obtain the P-value you need for any partial *F* test by arranging the variables correctly

Coefficients of Partial Determination

- Recall: The *coefficient of multiple determination* R^2
  - may be interpreted as the percentage of the total variation that has been explained by the model
  - measures the proportionate reduction in the variation of Y achieved by the introduction of the entire set of X variables in the model
- A *coefficient of partial determination*, in contrast, measures
  - the marginal contribution of one X variable when all others are already included in the model
  - Or, the amount of *remaining variation* explained by a variable *given* other variables already in the model

$$R^2_{Yk|1,\ldots,k-1,k+1,\ldots,p} = \frac{SSR\,(X_k|X_1,\ldots,X_{k-1},X_{k+1},\ldots,X_p)}{SSE(X_1,\ldots,X_{k-1},X_{k+1},\ldots,X_p)}$$

- Subscripts after bar (|) or period(.) represent variables already in model;
- The bar or period separates the correlated variables and the controlled for variables

Variable Importance: Illustration

- $A+B+C+D$ represents all the variability in the dependent variable(DV) Y to be explained
  - $A+B+C = R^2$ for the model
- The partial determination is the amount a variable explains relative to the amount in the DV that is left to explain after the contributions of the other predictors have been removed from both the predictor and criterion
  - For Predictor 1 it is $A/(A+D)$
  - For Predictor 2 it would be $B/(B+D)$

Figure 1

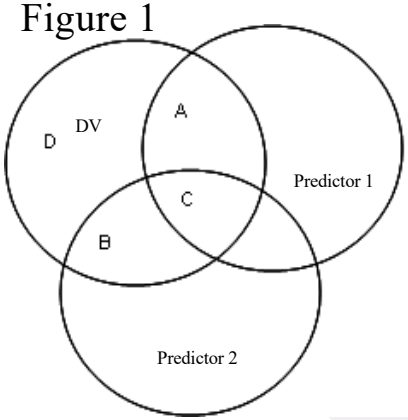


Figure 2

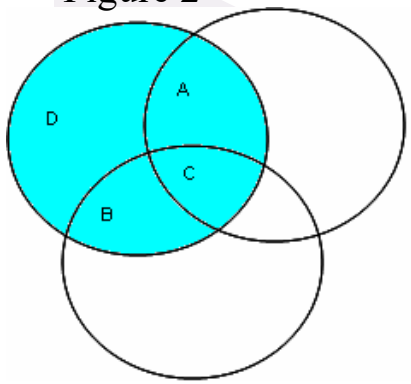


Figure 3

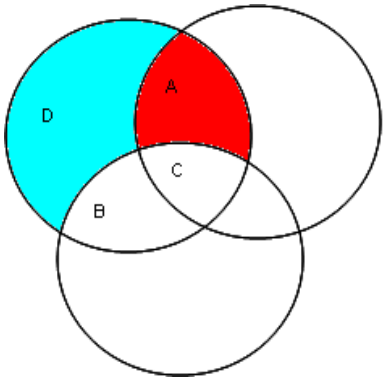
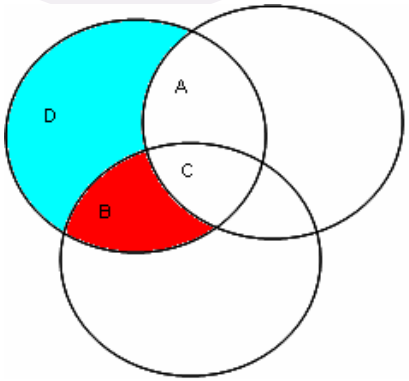


Figure 4



Examples

- $R^2_{Y1|23}$ represents the percentage of the leftover variation in Y (after regressing on X_2 and X_3) that is explained by X_1

$$R^2_{Y1|23} = \frac{SSR\ (X_1|X_2, X_3)}{SSE(X_2, X_3)} = \frac{SSE(X_2, X_3) - SSE(X_1, X_2, X_3)}{SSE(X_2, X_3)}$$

- Suppose that total sums of squares is 100, and X_1 explains 60; Of the remaining 40, X_2 then explains 20, and of the remaining 20, X_3 explains 5. Then
  - $R^2_{Y2|1} = \frac{20}{40} = 0.50$
  - $R^2_{Y3|12} = \frac{5}{20} = 0.25$

Model #1's Analysis of Variance Table

Response: gpa

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
hsm	1	25.810	25.8099	52.6975	6.621e-12 ***
hss	1	1.237	1.2371	2.5258	0.1134
hse	1	0.665	0.6654	1.3585	0.2451
Residuals	220	107.750	0.4898		

- HSE explains 0.6% of remaining variation after HSM and HSS in model

$$R^2_{GPA\ HSE|HSM\ HSS} = \frac{0.665}{0.665 + 107.750} = 0.00613$$

Regression Results for Several Fitted Models

(a) Regression of Y on X_1
$\hat{Y} = -1.496 + .8572X_1$

Source of Variation	SS	df	MS
Regression	352.27	1	352.27
Error	143.12	18	7.95
Total	495.39	19	

Variable	Estimated Regression Coefficient	Estimated Standard Deviation	t^*
X_1	$b_1 = .8572$	$s\{b_1\} = .1288$	6.66

(b) Regression of Y on X_2
$\hat{Y} = -23.634 + .8565X_2$

Source of Variation	SS	df	MS
Regression	381.97	1	381.97
Error	113.42	18	6.30
Total	495.39	19	

Variable	Estimated Regression Coefficient	Estimated Standard Deviation	t^*
X_2	$b_2 = .8565$	$s\{b_2\} = .1100$	7.79

(c) Regression of Y on X_1 and X_2
$\hat{Y} = -19.174 + .2224X_1 + .6594X_2$

Source of Variation	SS	df	MS
Regression	385.44	2	192.72
Error	109.95	17	6.47
Total	495.39	19	

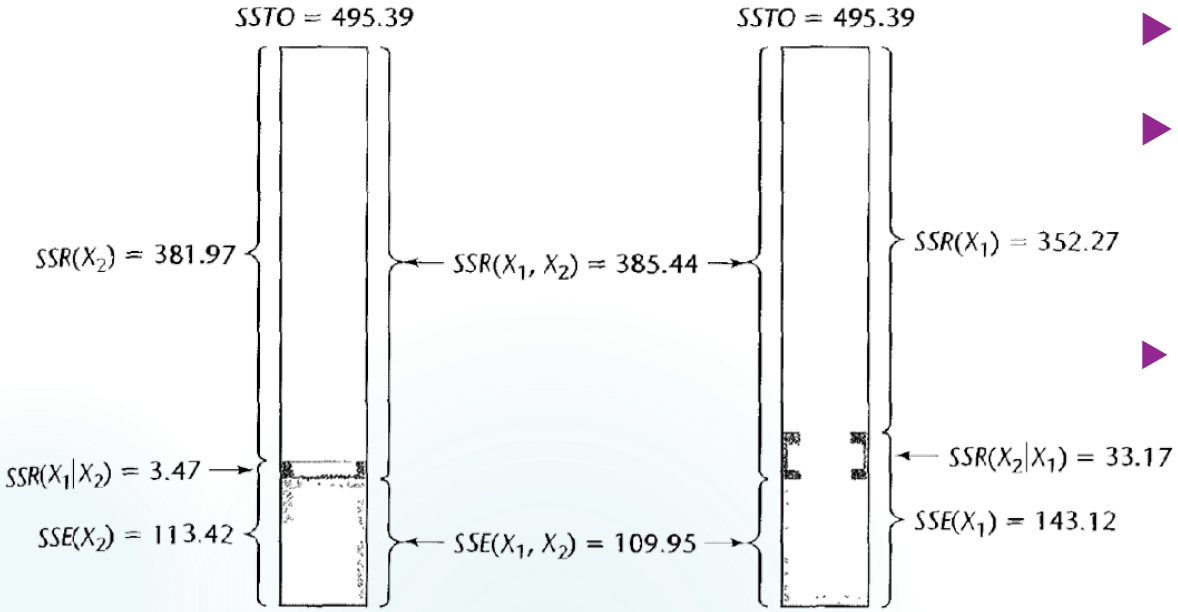
Variable	Estimated Regression Coefficient	Estimated Standard Deviation	t^*
X_1	$b_1 = .2224$	$s\{b_1\} = .3034$	.73
X_2	$b_2 = .6594$	$s\{b_2\} = .2912$	2.26

(d) Regression of Y on X_1, X_2, and X_3
$\hat{Y} = 117.08 + 4.334X_1 - 2.857X_2 - 2.186X_3$

Source of Variation	SS	df	MS
Regression	396.98	3	132.33
Error	98.41	16	6.15
Total	495.39	19	

Variable	Estimated Regression Coefficient	Estimated Standard Deviation	t^*
X_1	$b_1 = 4.334$	$s\{b_1\} = 3.016$	1.44
X_2	$b_2 = -2.857$	$s\{b_2\} = 2.582$	−1.11
X_3	$b_3 = -2.186$	$s\{b_3\} = 1.596$	−1.37

Schematic Representation of Extra SS



- Notice that *SST* never changes
- $F^* = \dfrac{(SSE(R)-SSE(F))/(df_E(R)-df_E(F))}{SSE(F)/df_E(F)}$
  $= \dfrac{(R^2(F)-R^2(R))/(df_E(R)-df_E(F))}{(1-R^2(F))/df_E(F)}$
- $R^2_{Yk|1,\ldots,k-1,k+1,\ldots,p} = \dfrac{SSR\,(X_k|X_1,\ldots,X_{k-1},X_{k+1},\ldots,X_p)}{SSE(X_1,\ldots,X_{k-1},X_{k+1},\ldots,X_p)}$
  $= \dfrac{SSE(R)-SSE(F)}{SSE(R)} \in [0,1]$
  $= R^2\ (Y|1,\ldots,k-1,k+1,\ldots,p,\ k|1,\ldots,k-1,k+1,\ldots,p)$

- Example of ANOVA Table with Decomposition of SSR for Three *X* Variables

Source of Variation	*SS*	*df*	*MS*
Regression	$SSR(X_1, X_2, X_3)$	3	$MSR(X_1, X_2, X_3)$
X_1	$SSR(X_1)$	1	$MSR(X_1)$
$X_2\|X_1$	$SSR(X_2\|X_1)$	1	$MSR(X_2\|X_1)$
$X_3\|X_1, X_2$	$SSR(X_3\|X_1, X_2)$	1	$MSR(X_3\|X_1, X_2)$
Error	$SSE(X_1, X_2, X_3)$	$n-4$	$MSE(X_1, X_2, X_3)$
Total	*SSTO*	$n-1$	

η^2 in R

- Partial η^2 appears to be another name for partial R^2
- Package 'heplots' :Visualizing Hypothesis Tests in Multivariate Linear Models
- > library(heplots)
- > etasq(reg1)

'skinfold' and 'midarm' explain the most remaining variation when added last

	Partial
eta^2	
skinfold	0.11434540
thigh	0.07107507
midarm	0.10500972
Residuals	NA

Note on Partial Determinations

- Can be useful in model selection (Chapter 9)
- Can get any partial coefficient of determination that we want, but may have to rearrange model to do it
  - Example: If we want HSE given HSM, we would need to list variables HSM and HSE as the first and second in the model
- Can get any desired Type I SS in the same way

(Coefficient of) Partial Correlation

- Square root of the coefficient of partial determination
- Given plus/minus sign according to the corresponding regression coefficient
- Measures the strength of a linear relation between two variables taking into account other variables
- Interpreting the result:
  - If the partial correlation, $r_{12.3}$, is smaller than the simple (two-variable) correlation r_{12}, but greater than 0, then variable 3 partly explains the correlation between variable 1&2
- Procedure to find partial correlation Y, X_k
  - Predict Y using other X's
  - Predict X_k using other X's
  - Find correlation between the two sets of residuals

Standardized Regression Model

- Dwaine Studios Example

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

  - Y is the total sale in a city (unit: thousands)
  - X_1 : population aged 16 and under (thousands)
  - X_2 : per capita disposable income (thousands)
- The units of β_1 and β_2 are different, and the magnitudes of β_1 and β_2 depend on the units of X_1 and X_2, respectively. β_1 and β_2 can NOT be directly compared
- When X_1 and X_2 are dramatically different in magnitudes, there will be numerical problems in computing the estimates of β_1
- We typically prefer this because most of the measures (such as those used in psychology) are on arbitrary scales

Correlation Transformation

- Let $\bar{Y}$, $\bar{X}_1, \bar{X}_2$ be the means of $\{Y_i\}$, $\{X_{i1}\}$, $\{X_{i2}\}$, respectively
- Let s_Y, s_{X1}, s_{X2} be the standard deviations of $\{Y_i\}$, $\{X_{i1}\}$, $\{X_{i2}\}$, respectively

Standardization = Centering + Scaling

= subtracting mean + dividing standard deviation

$$s_Y = \sqrt{\frac{\sum_i (Y_i - \bar{Y})^2}{n-1}}, \; s_{X_k} = \sqrt{\frac{\sum_i (X_{ik} - \bar{X}_k)^2}{n-1}}, \quad k = 1,2$$

$$\frac{1}{\sqrt{n-1}} \frac{Y_i - \bar{Y}}{s_Y} = \frac{1}{\sqrt{n-1}} \frac{(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i) - (\beta_0 + \beta_1 \bar{X}_1 + \beta_2 \bar{X}_2 + \bar{\varepsilon})}{s_Y}$$

$$= \frac{1}{\sqrt{n-1}} \frac{\beta_1 (X_{i1} - \bar{X}_1) + \beta_2 (X_{i2} - \bar{X}_2) + (\varepsilon_i - \bar{\varepsilon})}{s_Y}$$

$$= \frac{\beta_1 s_{X1}}{s_Y} \frac{(X_{i1} - \bar{X}_1)}{\sqrt{n-1} s_{X1}} + \frac{\beta_2 s_{X2}}{s_Y} \frac{(X_{i2} - \bar{X}_2)}{\sqrt{n-1} s_{X2}} + \frac{\varepsilon_i - \bar{\varepsilon}}{\sqrt{n-1} s_Y}$$

$$\boxed{Y_i^* = 0 + \beta_1^* X_{i1}^* + \beta_2^* X_{i2}^* + \ldots + \beta_{p-1}^* X_{i,p-1}^* + \varepsilon_i^*}$$ Correlation transformation

Standardized Regression Model

$$Y_i^* = 0 + \beta_1^* X_{i1}^* + \beta_2^* X_{i2}^* + \ldots + \beta_{p-1}^* X_{i,p-1}^* + \varepsilon_i^*$$

where

$$\beta_k^* = \frac{s_{X_k}}{s_Y}\beta_k, \quad \text{for } k = 1,2,\ldots,p-1$$

- Standardized regression coefficients:
  - $\beta_1^*, \beta_2^*, \ldots, \beta_{p-1}^*$ are scale/unit free, admit same interpretation, and can be directly compared with each other in an intuitive sense
- Standardized design matrix:
  - $X^{*T}X^*$ is the sample correlation matrix r_{XX} for $X_1, X_2, \ldots, X_{p-1}$. Bounded between -1 and 1, numerical problems in computing the inverse of $(X^{*T}X^*)^{-1}$ can be avoided or mitigated
  - $X^{*T}Y^* = r_{YX}$

$$X^*_{n\times(p-1)} = \begin{pmatrix} X^*_{11} & \cdots & X^*_{1,p-1} \\ \vdots & \ddots & \vdots \\ X^*_{n1} & \cdots & X^*_{n,p-1} \end{pmatrix}$$

$$X^{*T}X^* = r_{XX} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1,p-1} \\ r_{21} & 1 & & r_{2,p-1} \\ \vdots & & \ddots & \vdots \\ r_{p-1,1} & r_{p-1,2} & \cdots & 1 \end{pmatrix}$$

$$X^{*T}Y^* = r_{YX} = \begin{pmatrix} r_{Y1} \\ r_{Y2} \\ \vdots \\ r_{Y,p-1} \end{pmatrix}$$

Note on Standardization

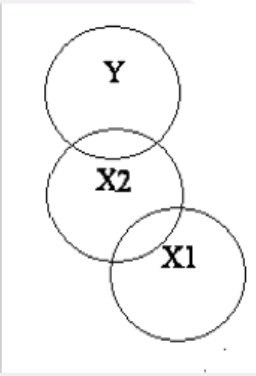
- Refer to KNNL p276, since

$$r_{XX}=\begin{bmatrix}1 & r_{12}\\ r_{12} & 1\end{bmatrix},\ r_{YX}=\begin{bmatrix}r_{Y1}\\ r_{Y2}\end{bmatrix},\ b_1^*=\frac{r_{Y1}-r_{12}r_{Y2}}{1-r_{12}^2},\ b_2^*=\frac{r_{Y2}-r_{12}r_{Y1}}{1-r_{12}^2},\ b_k^*=\frac{s_{X_k}}{s_Y}b_k$$

Transformed back to the original scale, we have

$$b_1=\frac{\frac{\sum(X_{i1}-\bar{X}_1)(Y_i-\bar{Y})}{\sum(X_{i1}-\bar{X}_1)^2}-\left[\frac{\sum(Y_i-\bar{Y})^2}{\sum(X_{i1}-\bar{X}_1)^2}\right]^{1/2}r_{12}r_{Y2}}{1-r_{12}^2}$$

(7.56) on KNNL p281. We need this to prove (7.41) on p271



$$R^2_{Y2|1}=\left(r_{Y2|1}\right)^2=\frac{(r_{Y2}-r_{12}r_{Y1})^2}{(1-r_{12}^2)(1-r_{Y1}^2)}$$

When $r_{Y1}\approx 0$,

$$R^2_{Y2|1}\approx\frac{r_{Y2}^2}{1-r_{12}^2}>r_{Y2}^2=R^2_{Y2}$$

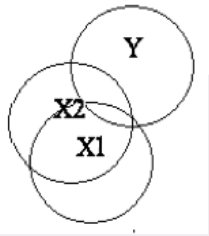
- Similarly,

$$b_2=\frac{\frac{\sum(X_{i2}-\bar{X}_2)(Y_i-\bar{Y})}{\sum(X_{i2}-\bar{X}_2)^2}-\left[\frac{\sum(Y_i-\bar{Y})^2}{\sum(X_{i2}-\bar{X}_2)^2}\right]^{1/2}r_{12}r_{Y1}}{1-r_{12}^2},\ R^2_{Y2|13}=\left(r_{Y2|13}\right)^2=\frac{\left(r_{Y2|3}-r_{12|3}r_{Y1|3}\right)^2}{(1-r^2_{12|3})(1-r^2_{Y1|3})}$$

Even though X_1 is not correlated with Y, having it in the equation raises the R^2 from what it would have been with just X_2

Suppressor Variables

- If SSR($X_2|X_1$) > SSR(X_2) then X_1 is called a suppressor variable
  - SSR(skinfold) = 352.27, SSR(skinfold|midarm) = 379.40
  - SSR(midarm) = 10.05, SSR(midarm|skinfold) = 37.19



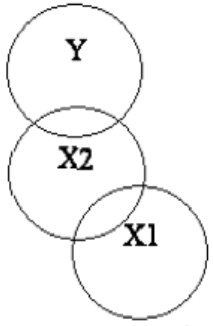
Response: fat

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
skinfold	1	352.27	352.27	56.5312	8.406e-07	***
midarm	1	37.19	37.19	5.9674	0.02579	*
Residuals	17	105.93	6.23			

Response: fat

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
midarm	1	10.05	10.05	1.6131	0.2212	
skinfold	1	379.40	379.40	60.8856	5.117e-07	***
Residuals	17	105.93	6.23			

- The general idea is that there is some kind of noise (error) in X_2 that is not correlated with Y, but is correlated with X_1. By including X_1we suppress (account for) this noise, and leave X_2 as an improved predictor of Y
- It will suppress irrelevant variance of other independent variables



R Code

- Get the coefficients after using standardized variables
  - library(QuantPsyc)
  - lm.beta(reg1)

```
skinfold     thigh      midarm
4.263705 -2.928701 -1.561417
```

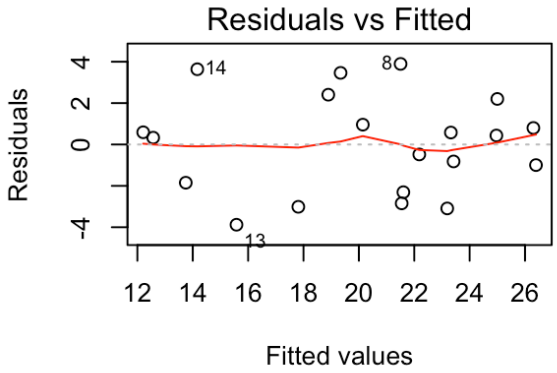
- Or, we could do it without above package
  - sa <- scale(a1) #standardized, notice now sa is a matrix, not a data frame
  - n <- nrow(sa); sa <- data.frame(sa/sqrt(n-1)) #correlation transformation
  - srm <- lm(fat ~ skinfold + thigh + midarm, sa)
  - summary(srm)
- Different ANOVA; Exactly the same η^2

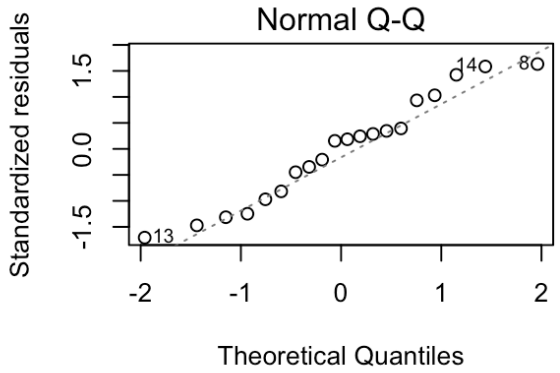
Skinfold and thigh suggest largest standardized change

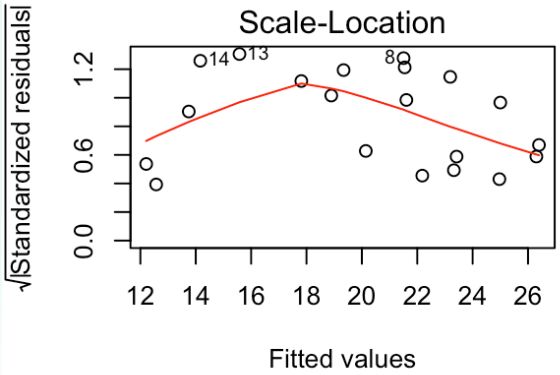
Don’t Forget the Diagnostics

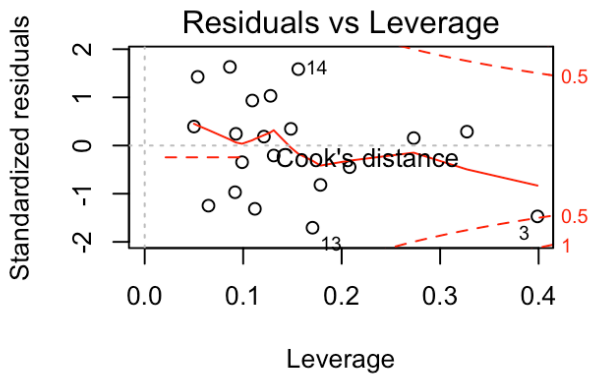
- Finally,

fat ~ skinfold + thigh









Background Reading

- We went over 7.1 – 7.5
- We used program lec9_1. R to generate the output

Topic 2: Multicollinearity & Polynomial Regression

Outline

- Multicollinearity
  - Zero Collinearity
  - Linearly Dependent
  - General Case
  - Effects of Multicollinearity
  - Pairwise Correlations
- Polynomial regression

Body Fat Example, Revisited

- The P-value for ANOVA F-test is $<.0001$
- The P values for the individual regression coefficients are 0.170, 0.285, and 0.190
  - None of these are near our standard significance level of 0.05
- What is the explanation/cause?
- **Multicollinearity!!!**

	skinfold	thigh	midarm	fat
skinfold	1.0000000	0.9238425	0.4577772	0.8432654
thigh	0.9238425	1.0000000	0.0846675	0.8780896
midarm	0.4577772	0.0846675	1.0000000	0.1424440
fat	0.8432654	0.8780896	0.1424440	1.0000000

Zero Collinearity

- Recall

$$b = (X'X)^{-1}X'Y \sim N(\beta, \sigma^2 (X'X)^{-1})$$

- Extreme Case 1: **Orthogonal Design** X- zero collinearity
- Explanatory variables are not correlated to each other, or their corresponding columns in X are orthogonal to each other, that is,

$$X'X = \text{diag}(\|X_0\|^2, \|X_1\|^2, \cdots, \|X_{p-1}\|^2)$$

$$b_j = \frac{X_j'Y}{\|X_j\|^2}; \quad \text{Var}(b_j) = \frac{\sigma^2}{\|X_j\|^2}$$

- Recall that when $r_{12} = 0$,

$$b_1 = \frac{\frac{\sum(X_{i1}-\bar{X}_1)(Y_i-\bar{Y})}{\sum(X_{i1}-\bar{X}_1)^2} - \left[\frac{\sum(Y_i-\bar{Y})^2}{\sum(X_{i1}-\bar{X}_1)^2}\right]^{1/2} r_{12}r_{Y2}}{1-r_{12}^2} = \frac{\sum(X_{i1}-\bar{X}_1)(Y_i-\bar{Y})}{\sum(X_{i1}-\bar{X}_1)^2}$$

Properties of Zero Collinearity

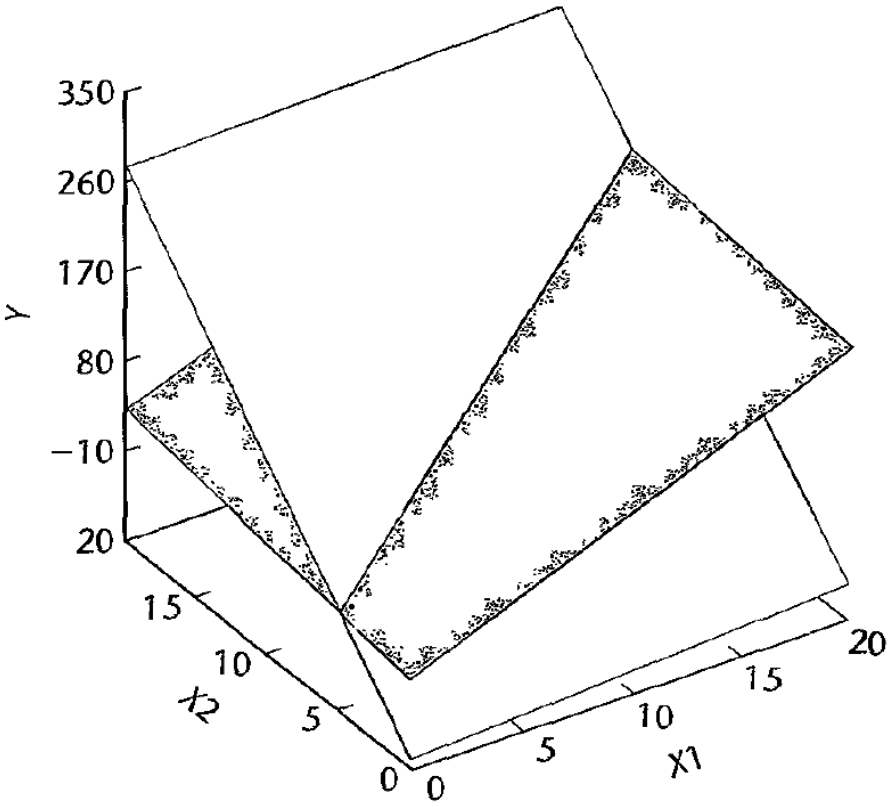
- The estimate of β_j, i.e. b_j, does not depend on the other explanatory variables, that is, it does not change whether other explanatory variables are included in the model or not
- The contribution of explanatory variable X_j to the Total Sum of Squares (*SST*) is clear-cut, that is, it does not depend on whether other explanatory variables are in the model or not, and Type I and III (or II) *SS* of X_j will be the same
- Under the assumption that the linear regression model is true, orthogonal design (X with zero collinearity) is optimal: No ambiguity between the explanatory variables, and the variances are the smallest possible (high power in testing)
- P-values for testing β_j will change, because they depend on *MSE*, which further depends which explanatory variables are included in the model

Multicollinearity: Linearly Dependent

- Extreme Case 2: **Degenerate Design** X- some explanatory variables (or their columns) are linearly dependent
  - $X_1 = X_2$
  - $X_2 - 3X_3 + X_4 = 5$
  - $c_1X_{j1} + c_2X_{j2} + \ldots + c_kX_{jk} =$ constant
- Consequences of a degenerate design
  - $rank(X) < p,\ rank(X'X) = rank(X) < p$
  - $(X'X)^{-1}$ does not exist, $b = (X'X)^{-1}X'Y$ can not be calculated
- The full model including all the explanatory variables cannot be fitted, due to linear dependence between the variables

Nature of Problem

- When X_1 and X_2 follow a linear relation, i.e., the predictor variables are perfectly correlated
- The two response surfaces have the same fitted values only when they intersect
- Since many different response functions provide the same good fit, we cannot interpret anyone set of regression coefficients as reflecting the effects of the different predictor variables



Consequences of Linear Dependency

- For each (perfect) linear relation, the involved explanatory variables are confounded, and at least one explanatory variable is redundant
- It is possible to mathematically remove the redundancy by deleting one of the explanatory variables, so that the reduced model can be fitted, and coefficients of the remaining variables can be estimated
- However, the confounding issue still remains and cannot be resolved, and the resulting model cannot be interpreted
- Consider three variables: X_1, X_2, X_3, with $X_1 + X_2 + X_3 = 0$
  - Model (X_1, X_2), Model (X_1, X_3), and Model (X_2, X_3) cannot be distinguished

Multicollinearity: General Case

- Between the two extreme cases, not entirely orthogonal design, not degenerate either, some correlation exists among explanatory variables
- Although not optimal, it is fine when the amount of correlation is small. When the amount of correlation increases and becomes large, it leads to both statistical and numerical problems
  - Statistically, ambiguity between the involved variables increases, it is therefore difficult to determine the regression coefficients. Type I *SS* and Type II *SS* becomes different, $Var(b) = \sigma^2(X'X)^{-1}$ increases; Particularly, $Var(b_j)$ increases (inflates), leading to inconsistency between ANOVA F test and t tests (recall Body Fat Example)
  - Numerically (i.e. computationally), $X'X$ is close to singular and is therefore difficult to invert accurately, $(X'X)^{-1}$ becomes unstable; The same happens to b
- Solve the statistical problem and the numerical problem will also be solved

CS Example, Simulation I

- Consider the extreme case where one predictor is a linear combination of other predictors

```
> csdata$hs <- (csdata$hsm + csdata$hss + csdata$hse)/3
> reg1 <- lm(gpa ~ hsm + hss + hse + hs, data = csdata)
> summary(reg1)
> anova(reg1)
```

Coefficients: (1 not defined because of singularities)

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.58988	0.29424	2.005	0.0462	*
hsm	0.16857	0.03549	4.749	3.68e-06	***
hss	0.03432	0.03756	0.914	0.3619	
hse	0.04510	0.03870	1.166	0.2451	
hs	NA	NA	NA	NA	

Analysis of Variance Table

Response: gpa

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
hsm	1	25.810	25.8099	52.6975	6.621e-12	***
hss	1	1.237	1.2371	2.5258	0.1134	
hse	1	0.665	0.6654	1.3585	0.2451	
Residuals	220	107.750	0.4898			

Something is wrong!

Explanation


```
Coefficients: (1 not defined because of singularities)
             Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)  0.58988    0.29424    2.005    0.0462 *
...
hse          0.04510    0.03870    1.166    0.2451
hs               NA         NA         NA       NA
```

- NOTE: The parameters of 'hs' have been set to NA, since the variable is a linear combination of other variables as shown

$$hs = 0.33333*hsm + 0.33333*hss + 0.33333*hse$$

- Now the model is not full rank. Least-squares solutions for the parameters are not unique. Some statistics will be misleading

Type II &III SS Do Not Work


```
> library("car")
> Anova(reg1, type="II")   # Type II tests
> Anova(reg1, type="III")  # Type III tests
```

- In this extreme case, R does not consider aliases for the Type II SS, and simply refuse to work for Type III SS

```
Note: model has aliased coefficients
        sums of squares computed by
model comparison
Anova Table (Type II tests)

Response: gpa
          Sum Sq  Df F value Pr(>F)
hsm            0
hss            0
hse            0
hs             0
Residuals 107.75 220
```

```
Error in Anova.III.lm(mod, error, singular.ok = singular.ok, ...) :
  there are aliased coefficients in the model
```

Extent of multicollinearity

- This example had one explanatory variable equal to a linear combination of other explanatory variables
- This is the most extreme case of multicollinearity and is detected by statistical software because $(X'X)$ does not have an inverse
- We are concerned with cases less extreme

CS Example, Simulation II

- Now we add a little noise to break up the perfect linear association

```
> n <- nrow(csdata)
> csdata$hs1 <- csdata$hs + rnorm(n)*0.05
> reg2 <- lm(gpa ~ hsm + hss + hse + hs1, data = csdata)
> summary(reg2)
> anova(reg2)
> library("car")
> Anova(reg2, type="II")
```

Output

- Model seems to be good here

- None of the predictors significant
- Much larger SEs
- Look at the differences in Type I and II SS
- Sign of each coefficient make sense?

```
Coefficients:
             Estimate Std. Error  t value  Pr(>|t|)
(Intercept)  0.58988    0.29424   2.005    0.0462 *
hsm          0.16857    0.03549   4.749 3.68e-06 ***
hss          0.03432    0.03756   0.914    0.3619
hse          0.04510    0.03870   1.166    0.2451

Residual standard error: 0.6998 on 220 degrees of freedom
Multiple R-squared:  0.2046,  Adjusted R-squared:  0.1937
F-statistic: 18.86 on 3 and 220 DF,  p-value: 6.359e-11
```

	Estimate	Std. Error	t value	Pr(>|t|)	Type I SS	Type II SS
(Intercept)	0.5919	0.2943	2.011	0.0456 *		
hsm	-0.1396	0.3306	-0.422	0.6733	25.810	0.087
hss	-0.2776	0.3349	-0.829	0.4079	1.237	0.337
hse	-0.2639	0.3318	-0.795	0.4274	0.665	0.310
hs1	0.9285	0.9904	0.937	0.3495	0.431	0.431

Effects of Multicollinearity

- Regression coefficients are not well estimated and may be meaningless
- Similarly for standard errors of these estimates
- Type I *SS* and Type II *SS* will differ
- R^2 and predicted values are usually ok in these situations

- We want to refine a model that currently has redundancy in the explanatory variables
- Do this regardless if $X'X$ can be inverted without difficulty

Pairwise Correlations

- Pairwise correlations can be used to check for "pairwise" collinearity

  > cor(a1[, c("skinfold", "thigh", "midarm", "fat")])

  - Cor(skinfold, thigh) = 0.9238
- Multicollinearity may involve multiple *X*'s
  - Cor(midarm, skinfold+thigh) = 0.9952!!!

	skinfold	thigh	midarm
skinfold	1.0000000	0.9238425	0.4577772
thigh	0.9238425	1.0000000	0.0846675
midarm	0.4577772	0.0846675	1.0000000

- Change in coeff values of skinfold and thigh depending on what variables are in the model->

Variables in Model	b_1	b_2
X_1	.8572	—
X_2	—	.8565
X_1, X_2	.2224	.6594
X_1, X_2, X_3	4.334	−2.857

‘midarm’ or ‘thigh’


```
Call:
lm(formula = fat ~ skinfold + midarm, data = a1)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.7916     4.4883   1.513   0.1486
skinfold      1.0006     0.1282   7.803   5.12e-07 ***
midarm       -0.4314     0.1766  -2.443   0.0258 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.496 on 17 degrees of freedom
Multiple R-squared:  0.7862,  Adjusted R-squared:  0.761
F-statistic: 31.25 on 2 and 17 DF,  p-value: 2.022e-06
```

```
Call:
lm(formula = fat ~ skinfold + thigh, data = a1)

Coefficients:
             Estimate  Std. Error t value Pr(>|t|)
(Intercept) -19.1742     8.3606  -2.293   0.0348 *
skinfold      0.2224     0.3034   0.733   0.4737
thigh         0.6594     0.2912   2.265   0.0369 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.543 on 17 degrees of freedom
Multiple R-squared:  0.7781,  Adjusted R-squared:  0.7519
F-statistic:  29.8 on 2 and 17 DF,  p-value: 2.774e-06
```

- fat ~ skinfold + midarm?
- cor(reg2$residuals, cbind(midarm, thigh))= c(-0.4531811, 0.184271)

Polynomial Regression

- We can fit a quadratic, cubic, etc. relationship by defining squares, cubes, etc., of a single *X* in a data step and using them as additional explanatory variables
- We can do this with more than one explanatory variable if needed

- Issue:
  - When we do this we generally create a multicollinearity problem

KNNL Example p300

- Response variable is the life (in cycles) of a power cell
- Explanatory variables are
  - Charge rate (3 levels)
  - Temperature (3 levels)
- This is a designed experiment!

Input and Check the Data

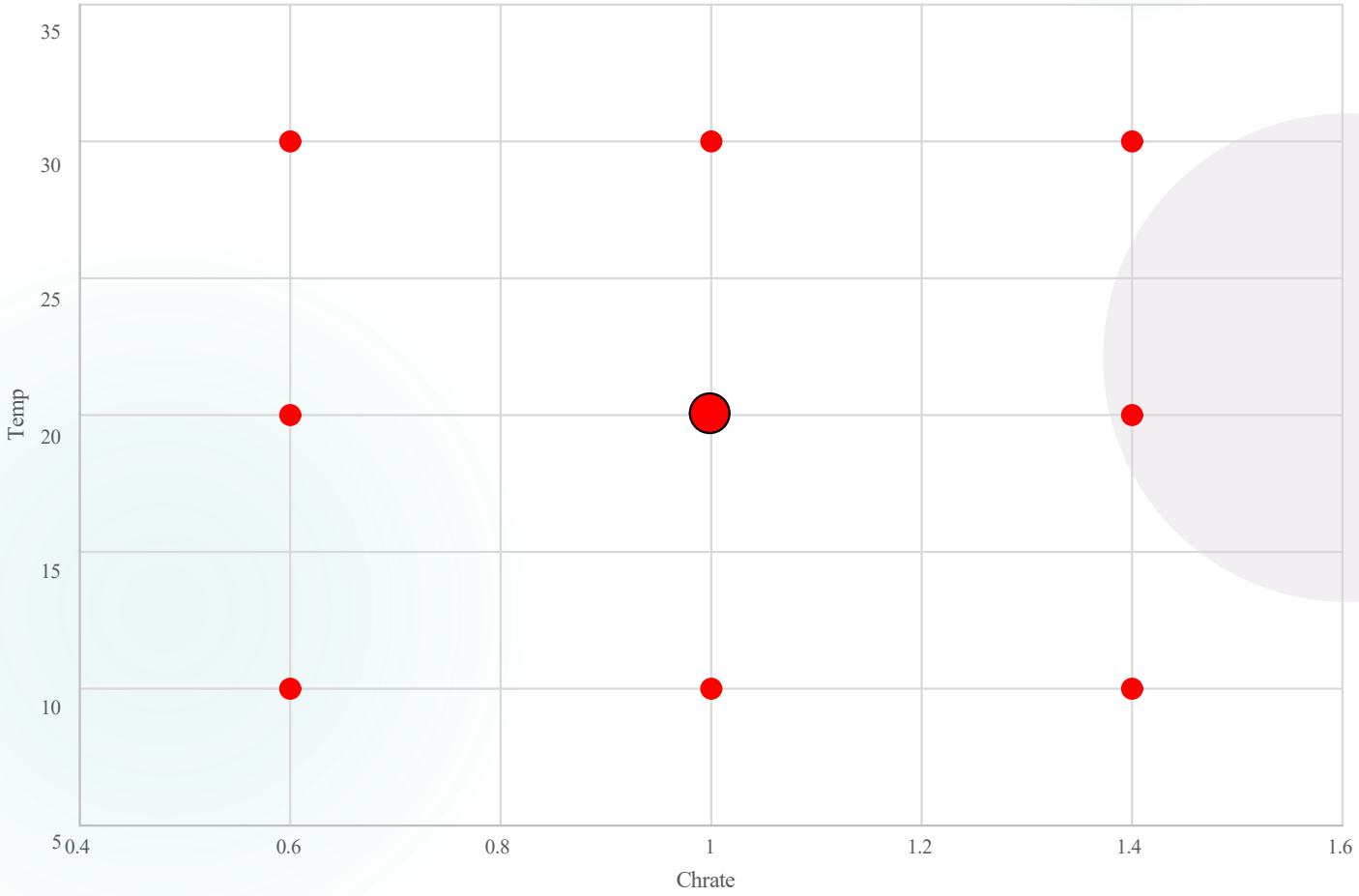
```
> b1 = read.table ("CH08TA01.txt")
> colnames(b1) = c("cycles",  "chrate", "temp")
> View(b1)
```

- Design of Experiments (DOE) is the perfect tool to efficiently determine if key inputs are related to key outputs
- Behind the scenes, DOE is simply a regression analysis
- What's not simple, however, is all of the choices you have to make when planning your experiment
  - What *X*'s should you test?
  - What ranges should you select for your *X*'s? How many replicates should you use?
  - Do you need center points? Etc.

	cycles	chrate	temp
1	150	0.6	10
2	86	1.0	10
3	49	1.4	10
4	288	0.6	20
5	157	1.0	20
6	131	1.0	20
7	184	1.0	20
8	109	1.4	20
9	279	0.6	30
10	235	1.0	30
11	224	1.4	30

Known as center points

Design Layout



Create New Variables and Run the Regression

- b1$chrate2 = b1$chrate * b1$chrate
- b1$temp2 = b1$temp * b1$temp
- b1$ct = b1$temp * b1$chrate
- reg5 <- lm(cycles ~ chrate + temp + chrate2 + temp2 + ct, data=b1)
- summary(reg5)
- anova(reg5)

Analysis of Variance Table

Response: cycles

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
chrate	1	18704	18704	17.8460	0.008292 **
temp	1	34201	34201	32.6323	0.002297 **
chrate2	1	1646	1646	1.5704	0.265552
temp2	1	285	285	0.2719	0.624352
ct	1	529	529	0.5047	0.509184
Residuals	5	5240	1048		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Conclusion

- Overall *F* significant, individual *t*'s not significant
  → multicollinearity problem
- Look at the correlations
- There are some very high correlations
  - r(chrate,chrate2) = 0.99103
  - r(temp,temp2) = 0.98609
- Common to have correlation between powers of a variable

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	337.7215	149.9616	2.252	0.0741	.
chrate	-539.5175	268.8603	-2.007	0.1011	
temp	8.9171	9.1825	0.971	0.3761	
chrate2	171.2171	127.1255	1.347	0.2359	
temp2	-0.1061	0.2034	-0.521	0.6244	
ct	2.8750	4.0468	0.710	0.5092	

Residual standard error: 32.37 on 5 degrees of freedom
Multiple R-squared: 0.9135, Adjusted R-squared: 0.8271
F-statistic: 10.57 on 5 and 5 DF, p-value: 0.01086

A Remedy

- We can often remove the correlation between explanatory variables and their powers by centering
- Centering means that you subtract off the mean before squaring etc.
- KNNL rescaled by standardizing (subtract the mean and divide by the standard deviation) but subtracting the mean is key here because you get positive and negative values of X

- Use scale() to center the explanatory variables
- Recompute the squares, cubes, etc., using the centered variables
- Rerun the regression analysis

Rerun


```
> b1$schrate = scale(b1$chrate)
> b1$stemp = scale(b1$temp)
```

- Recompute squares and cross product

```
> b1$schrate2 = b1$schrate * b1$schrate
> b1$stemp2 = b1$stemp * b1$stemp
> b1$sct = b1$stemp * b1$schrate
```

- Rerun regression

```
> reg6 <- lm(cycles ~ schrate + stemp  + schrate2 + stemp2 + sct, data=b1)
> summary(reg6)
> anova(reg6)
```

	cycles	schrate	stemp
1	150	-1.29099	-1.29099
2	86	0.00000	-1.29099
3	49	1.29099	-1.29099
4	288	-1.29099	0.00000
5	157	0.00000	0.00000
6	131	0.00000	0.00000
7	184	0.00000	0.00000
8	109	1.29099	0.00000
9	279	-1.29099	1.29099
10	235	0.00000	1.29099
11	224	1.29099	1.29099

Output

Analysis of Variance Table

Response: cycles

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
chrate	1	18704	18704	17.8460	0.008292	**
temp	1	34201	34201	32.6323	0.002297	**
chrate2	1	1646	1646	1.5704	0.265552	
temp2	1	285	285	0.2719	0.624352	
ct	1	529	529	0.5047	0.509184	
Residuals	5	5240	1048			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	162.842	16.608	9.805	0.000188	***
schrate	-43.248	10.238	-4.224	0.008292	**
stemp	58.482	10.238	5.712	0.002297	**
schrate2	16.437	12.204	1.347	0.235856	
stemp2	-6.363	12.204	-0.521	0.624352	
sct	6.900	9.712	0.710	0.509184	

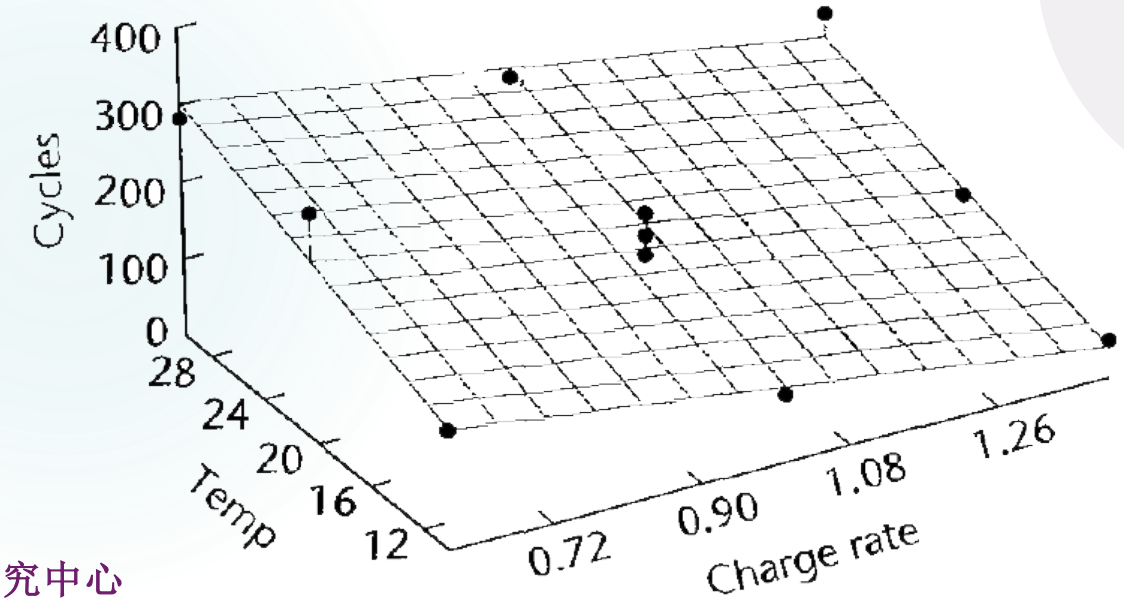
Residual standard error: 32.37 on 5 degrees of freedom
Multiple R-squared: 0.9135, Adjusted R-squared: 0.8271
F-statistic: 10.57 on 5 and 5 DF, p-value: 0.01086

- Exact same ANOVA table as before!!

- Overall *F* significant
- Individual *t*'s significant for chrate and temp
- Appears linear model will suffice
- Could do formal general linear test to assess this (*P*-value is 0.5527)

Conclusion

- Overall F significant
- Individual t's significant for chrate and temp
- Appears linear model will suffice
- Could do formal general linear test to assess this. (P-value is 0.5527)



Last slide

- We went over KNNL 7.6 and 8.1
- We used lec9_2.R to generate the output