Course Materials May Not Be Distributed or Posted Electronically

These course materials are the sole property of Dr. Todd M. Gross. They are strictly for use by students enrolled in a course taught by Dr. Gross. They may not be altered, excerpted, distributed, or posted to any website or other document-sharing service without express permission.

PSTAT 126
Regression Analysis

Dr. Todd Gross

Department of Statistics and Applied Probability

UCSB

Lecture 8

Multiple Linear Regression – Part II

Lecture Outline

- Worked Example of Multiple Linear Regression in R
- Extra Sum of Squares Principle
- Testing Extra Sum of Squares

A Worked Example of Multiple Linear Regression in R

Worked Example of Multiple Regression

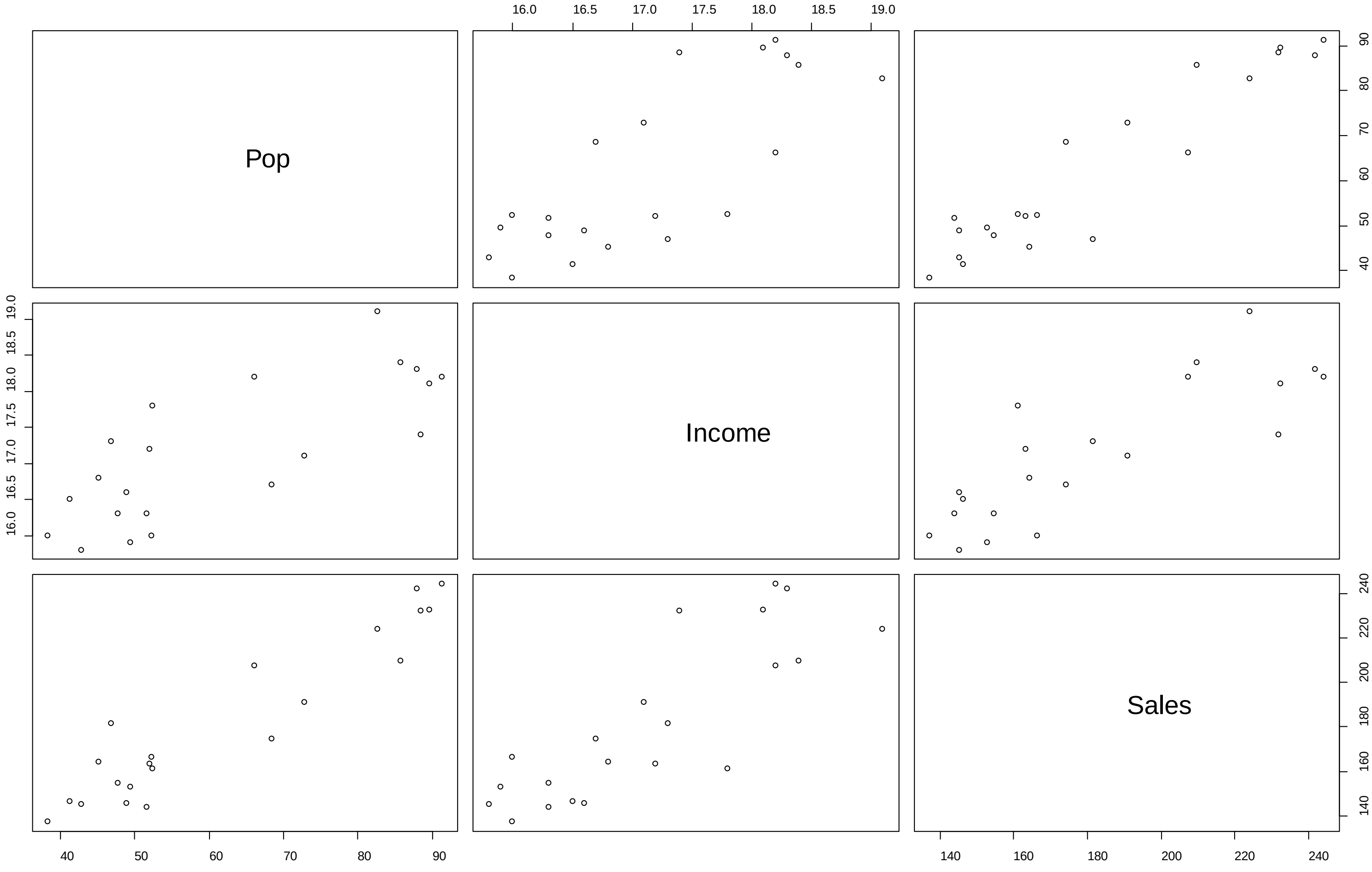
- Photography Studio Example

Dwaine Studios, Inc specialize in portraits of children. The company wishes to investigate whether sales (Y, thousand dollars) in a community can be predicted from the number of persons aged 16 or younger in the community (X_1, thousand persons) and the per capita disposable personal income in the community (X_2, thousand dollars).

Dwaine Studio Example – data set

```
> a=read.table("Dwaine.txt",header=T)
> a
   Pop  Income Sales
1  68.5   16.7 174.4
2  45.2   16.8 164.4
3  91.3   18.2 244.2
4  47.8   16.3 154.6
5  46.9   17.3 181.6
6  66.1   18.2 207.5
7  49.5   15.9 152.8
8  52.0   17.2 163.2
9  48.9   16.6 145.4
10 38.4   16.0 137.2
11 87.9   18.3 241.9
12 72.8   17.1 191.1
13 88.4   17.4 232.0
14 42.9   15.8 145.3
15 52.5   17.8 161.1
16 85.7   18.4 209.7
17 41.3   16.5 146.4
18 51.7   16.3 144.0
19 89.6   18.1 232.6
20 82.7   19.1 224.1
21 52.3   16.0 166.5
```

Scatterplot Matrix



Fitting the Linear Model

We fit the model

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

where $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

- We use the lm function in R to fit the model:

```
> fit1=lm(income~pop+sales)
```

- Does this linear model fit the data?
  - We need to look at the summary output

Summary of the Linear Model

```
Call:
lm(formula = Sales ~ Income + Pop)

Residuals:
     Min       1Q   Median       3Q      Max 
-18.4239  -6.2161   0.7449   9.4356  20.2151 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -68.8571    60.0170  -1.147   0.2663    
Income        9.3655     4.0640   2.305   0.0333 *  
Pop           1.4546     0.2118   6.868    2e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.01 on 18 degrees of freedom
Multiple R-squared:  0.9167, Adjusted R-squared:  0.9075 
F-statistic:  99.1 on 2 and 18 DF,  p-value: 1.921e-10
```

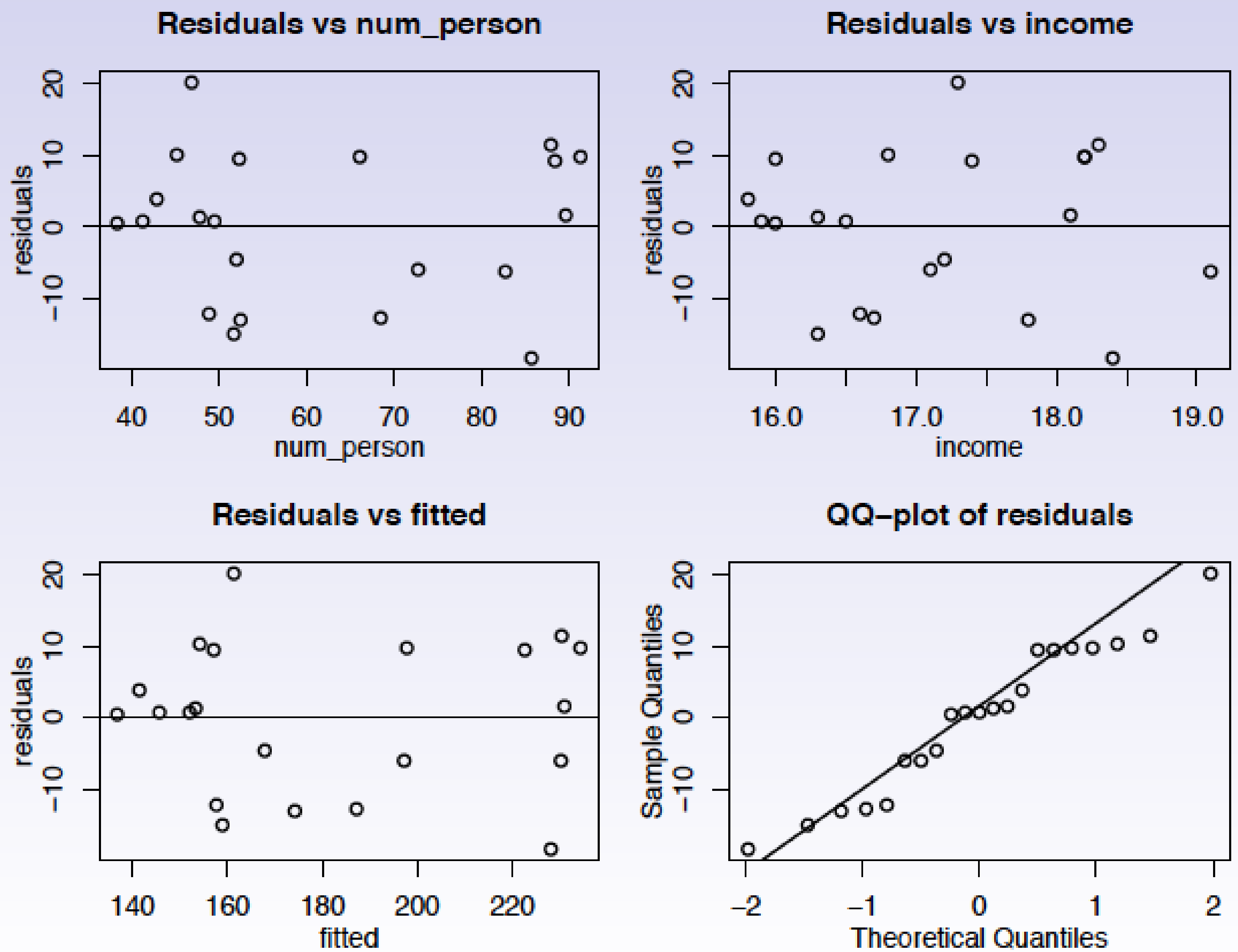
Overall fit

- $R^2 = 0.9167$
- For $H_0 : \beta_1 = \beta_2 = 0$, H_1 : not both β_1 and β_2 equal zero, $F^* = 99.1$ with p-value=1.921e-10. Strong evidence that sales are related to size of the targeted population and per capita disposable income

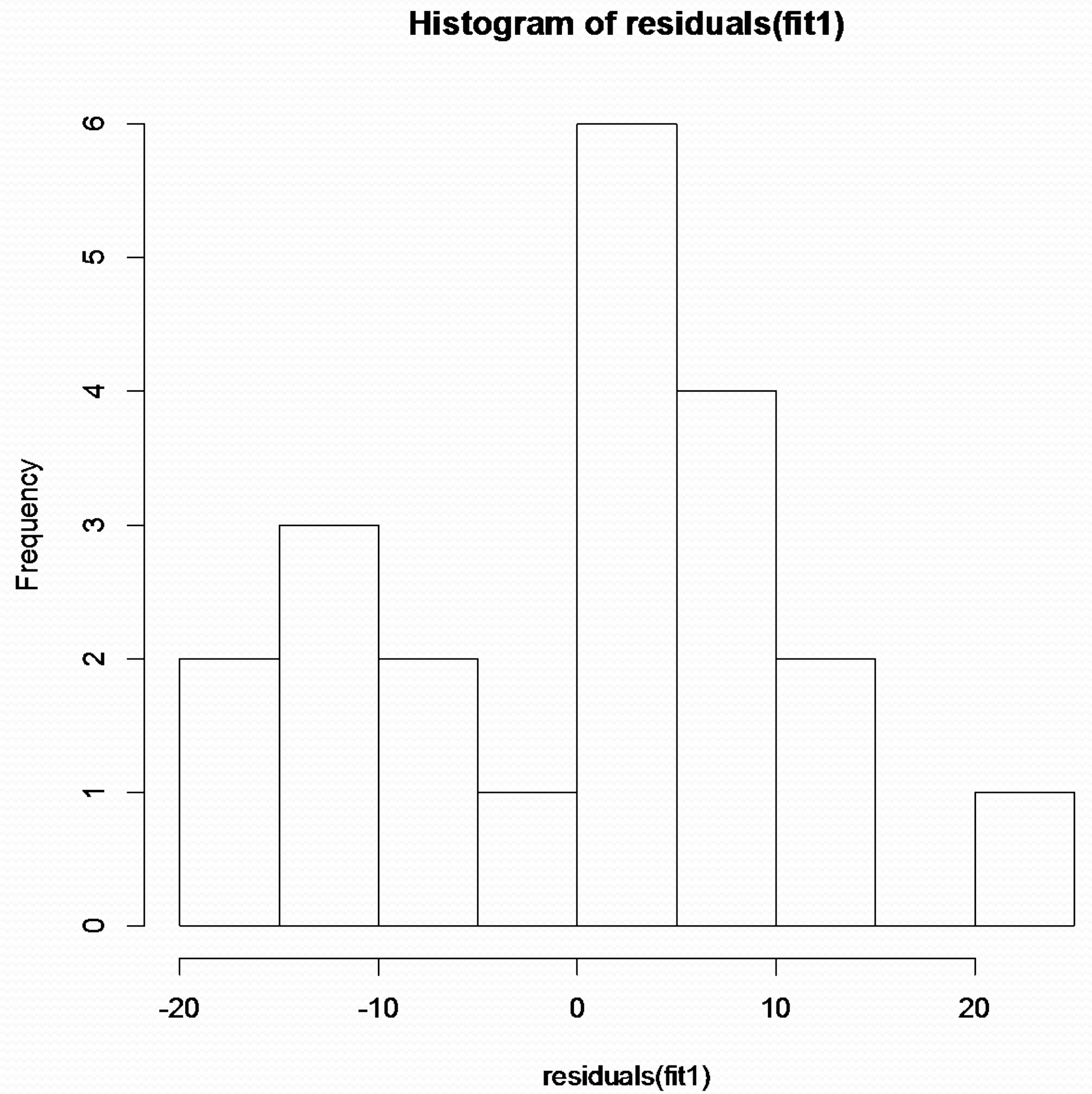
Residual plots

```
plot(x1, residuals(fit1),
     xlab=''num_person'',ylab=''residuals'')
abline(h=0)
title(``Residuals vs num_person'')

plot(x2, residuals(fit1),
     xlab=''income'', ylab=''residuals'')
abline(h=0)
title(``Residuals vs income'')

plot(fitted(fit1), residuals(fit1),
     xlab=''fitted'',ylab=''residuals'')
abline(h=0)
title(``Residuals vs fitted'')

qqnorm(residuals(fit1), main="")
qqline(residuals(fit1))
title(``QQ-plot of residuals'')
```

Residual plots



Histogram of Residuals



Confidence intervals of parameters

```
> confint(fit1)
                   2.5 %     97.5 %
(Intercept) -194.9480130 57.233867
x1             1.0096226  1.899497
x2             0.8274411 17.903560
```

Estimates and inference of parameters

- $b_0 = -68.8571$ with $s(b_0) = 60.0170$, fail to reject $H_0 : \beta_0 = 0$ at 5% level since p-value=0.2663. 95% confidence interval for β_0 is $(-194.9480130, 57.233867)$
- $b_1 = 1.4546$ with $s(b_1) = 0.2118$, reject $H_0 : \beta_1 = 0$ at 5% level since p-value=2e-06. 95% confidence interval for β_1 is $(1.0096226, 1.899497)$. With number of persons aged 16 or younger in a community increase by 1000, the expected sales increases by \$1454.6 with 95% confidence interval (\$1009.6, \$1899.5)
- $b_2 = 9.3655$ with $s(b_2) = 4.0640$, reject $H_0 : \beta_2 = 0$ at 5% level since p-value=0.0333. 95% confidence interval for β_1 is $(0.8274411, 17.903560)$. With per capita disposable personal income in the community increase by \$1000, the expected sales increases by \$9365.5 with 95% confidence interval (\$827.4, \$17903.6)

Estimation of Mean Response

- The company would like to estimate the <u>mean</u> sales for cities that have
  - a target population of 65,000 children, and
  - Per-capita disposable income of $17,000

```
> predict(fit1,data.frame(x1=65,x2=17),interval="confidence")
       fit      lwr      upr
1 184.9028 179.3134 190.4922
```

- "We predict annual sales of $185K for all cities with 65K children and $17K disposable per-capita income, and are 95% confident that sales will be between $179K and $190K"

Prediction of a Future Observation

- The company would like to estimate the sales for an individual city that has:
  - a target population of 65,000 children, and
  - Per-capita disposable income of $17,000

```
> predict(fit1,data.frame(x1=65,x2=17),interval="prediction")
       fit      lwr      upr
1 184.9028 161.1113 208.6944
```

- "We predict annual sales of $185K for an individual city with 65K children and $17K disposable per-capita income, and are 95% confident that sales will be between $161K and $209K"

The Extra Sum of Squares Principle

Extra Sum of Squares Principle

- When we add more predictors to a regression model, SSR always <u>increases</u> and SSE always <u>decreases</u>, while SSTO remains <u>unchanged</u>.

- The amount of increase in SSR (or reduction in SSE) is the <u>extra sum of squares</u>.
  - It measures the contribution of the <u>added</u> terms to the regression model <u>given</u> the other terms that are already in the model.

- Question: Does the model with more predictors fit the data <u>significantly</u> better than the model with less predictors?

The R^2 Interpretation of Extra SS

- Extra Sum of Squares involves comparing two models, one with more predictors than the other

- Consider two models
  - Model 1: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$
  - Model 2: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$

- Recall that R^2 describes the amount of variance in Y that is explained by the regression model

- R^2 for Model 2 will always be bigger than R^2 for Model 1

- Question: Is the increase in R^2 enough to conclude that the additional predictor makes a statistically significant contribution to the regression model?

- We perform an F-test to test the hypothesis that adding an additional predictor (or predictors) to the regression model produces a better fit.

Extra SS – Sequential and Partial

- We will consider two types of Extra SS: Sequential and Partial
  - Note that there are other SS types (SAS offers 4 types)

- Sequential SS – Each predictor is tested sequentially against any predictors that appear earlier in the model

- Partial SS – Each predictor is tested against all other predictors in the full regression model, whether they appear earlier or later

Extra Sum of Squares Principle (Overview)

- Consider a FULL model and a RESTRICTED (or REDUCED) model.
- Under a specific H_0, the Full Model includes additional predictors which have slopes of zero ($\beta_i = 0$)
- Fit the FULL model and obtain the error sum of squares - SSE(Full)
- Fit the RESTRICTED model and obtain the error sum of squares - SSE(Restricted)
- SSE(R) will always be larger than SSE(F). The difference SSE(R) – SSE(F) is the Extra SS.
- Under H_0, this difference should be small (i.e., when H_0 is true, the difference will be negligible).

The F-test for Extra SS

- We can test H_0 using an F-test that compares the error under the Restricted Model to the Error under the Full Model.

The F test statistic

$$F^* = \frac{(SSE(R) - SSE(F))/(df_R - df_F)}{SSE(F)/df_F} \overset{H_0}{\sim} F_{df_F - df_R, df_F}$$

where df_F and df_R are degrees of freedoms associated with $SSE(F)$ and $SSE(R)$ respectively

Reject H_0 if $F^* > F(1-\alpha;\ df_R - df_F, df_F)$

Sequential Sum of Squares

To introduce the concepts, consider the following simple linear model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i$$

We first introduce the sequential (type I) SS. Consider fitting three nested models

Model 1: $Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i$
Model 2: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$
Model 3: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i$

- Which of these models has the largest SSRegression?
- Which has the largest SSError?
- Which has the largest SSTotal?

Sequential Sum of Squares

Let $SSR(X_1)$, $SSR(X_1, X_2)$ and $SSR(X_1, X_2, X_3)$ be regression sum of squares of models 1, 2, and 3 respectively, and $SSE(X_1)$, $SSE(X_1, X_2)$ and $SSE(X_1, X_2, X_3)$ be error sum of squares of models 1, 2, and 3 respectively. Define

$$
\begin{aligned}
SSR(X_2|X_1) &= SSR(X_1, X_2) - SSR(X_1) \\
&= SSE(X_1) - SSE(X_1, X_2) \\
SSR(X_3|X_1, X_2) &= SSR(X_1, X_2, X_3) - SSR(X_1, X_2) \\
&= SSE(X_1, X_2) - SSE(X_1, X_2, X_3)
\end{aligned}
$$

Sequential Sum of Squares

The $SSTO$ may be decomposed into

$$\begin{aligned} SSTO &= SSR(X_1) + SSE(X_1) \\ &= SSR(X_1) + \overbrace{SSR(X_2|X_1) + SSE(X_1, X_2)} \\ &= SSR(X_1) + SSR(X_2|X_1) + \overbrace{SSR(X_3|X_1, X_2) + SSE(X_1, X_2, X_3)} \end{aligned}$$

- The first equality corresponds to fitting Model 1
- The first equality corresponds to adding X_2 to Model 1, ie Model 2
- The first equality corresponds to adding X_3 to Model 2, ie Model 3

Sequential sum of squares

When we fit the full model (Model 3), we have

$$SSTO = SSR(X_1, X_2, X_3) + SSE(X_1, X_2, X_3)$$

we see that $SSR(X_1, X_2, X_3)$ has been split into $SSR(X_1)$, $SSR(X_2|X_1)$ and $SSR(X_3|X_1, X_2)$.

- $SSR(X_1)$: SS explained by X_1
- $SSR(X_2|X_1)$: extra SS due to the addition of X_2 to the model that already includes X_1
- $SSR(X_3|X_1, X_2)$: extra SS due to the addition of X_3 to the model that already includes X_1 and X_2
- One can calculate $SSR(X_1)$, $SSR(X_2|X_1)$ and $SSR(X_3|X_1, X_2)$ by fitting three models. Sequential (type I) SS computes them simultaneously

Sequential Sum of Squares

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

- The Sequential or Type I SS add up to the SSRegression
- Type I SS depends on the order in which the variables are entered into the model, e.g., lm(Y~X1+X2+X3)
- In R, the anova function gives sequential, or Type I SS

Testing Extra SS for each predictor

- The ANOVA Table in R provides the sequential test of each predictor, given the preceding predictors in the model:

```
Analysis of Variance Table

Response: bodyfat
          Df Sum Sq Mean Sq F value    Pr(>F)
tricep     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
```

Test of X_1 alone

Test of X_2 given X_1

Test of X_3 given X_1, X_2

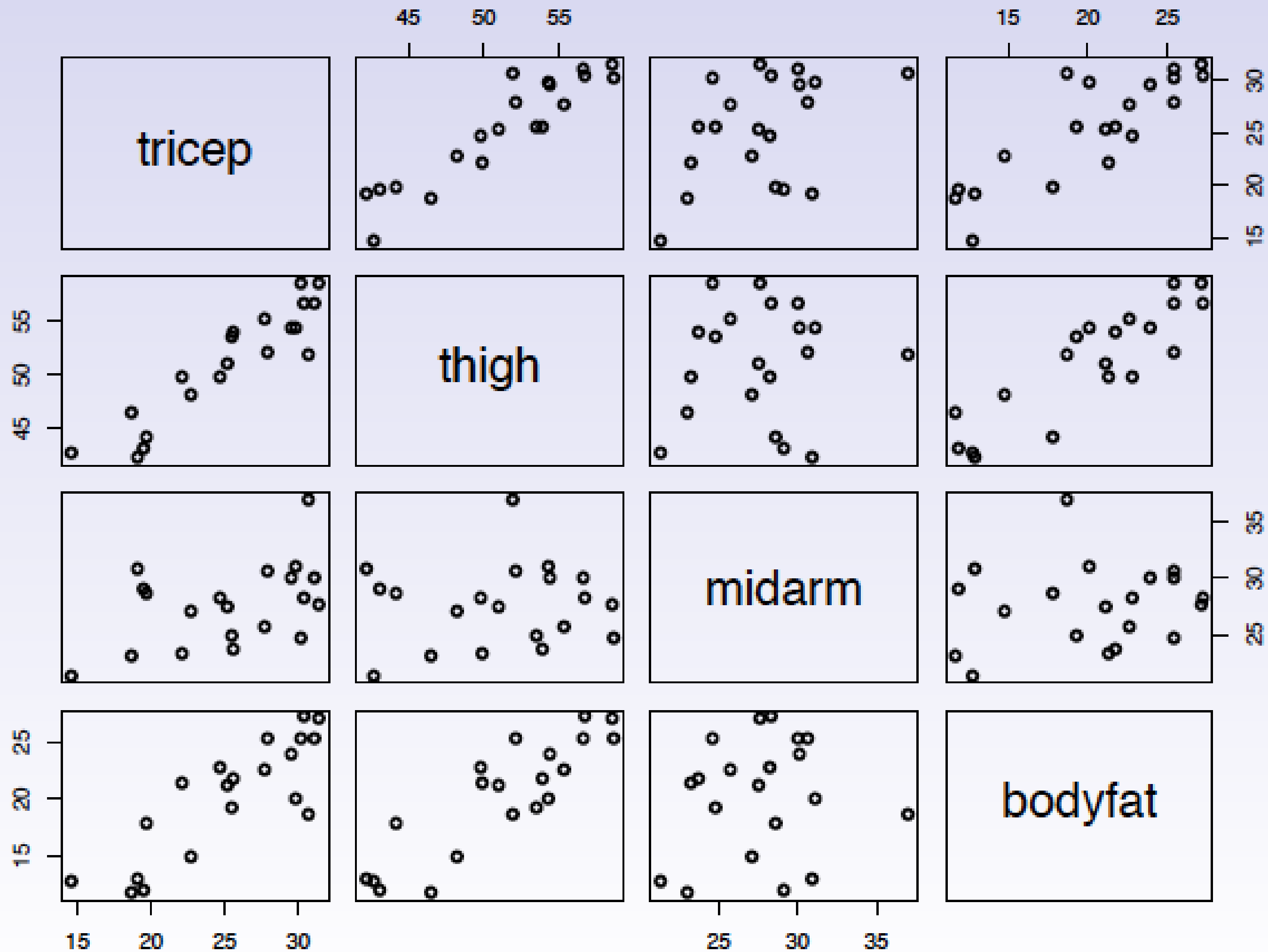
Body fat example

We want to build a regression model to predict body fat (Y) using triceps skinfold thickness (X_1), thigh circumference (X_2) and midarm circumference (X_3) (Example in Section 7.1).

```
> a <- matrix(scan("CH07TA01.DAT"),ncol=4,byrow=T)
> a
>          [,1] [,2] [,3] [,4]
  [1,] 19.5 43.1 29.1 11.9
  [2,] 24.7 49.8 28.2 22.8
  [3,] 30.7 51.9 37.0 18.7
  [4,] 29.8 54.3 31.1 20.1
  [5,] 19.1 42.2 30.9 12.9
  [6,] 25.6 53.9 23.7 21.7
  [7,] 31.4 58.5 27.6 27.1
  [8,] 27.9 52.1 30.6 25.4
  [9,] 22.1 49.9 23.2 21.3
 [10,] 25.5 53.5 24.8 19.3
 [11,] 31.1 56.6 30.0 25.4
 [12,] 30.4 56.7 28.3 27.2
 [13,] 18.7 46.5 23.0 11.7
```

Scatter plots

```
pairs(a,labels=c("tricep","thigh","midarm","bodyfat"))
```



Sequential sum of squares

We can fit a sequence of nested models to get the sequential sum of squares:

```
> tricep <- a[,1]; thigh <- a[,2]
> midarm <- a[,3]; bodyfat <- a[,4]
> fit0 <- lm(bodyfat ~ 1)
> fit1 <- lm(bodyfat ~ tricep)
> fit2 <- lm(bodyfat ~ tricep+thigh)
> fit3 <- lm(bodyfat ~ tricep+thigh+midarm)
> anova(fit0, fit1, fit2, fit3)
Analysis of Variance Table

Model 1: bodyfat ~ 1
Model 2: bodyfat ~ tricep
Model 3: bodyfat ~ tricep + thigh
Model 4: bodyfat ~ tricep + thigh + midarm
  Res.Df    RSS Df Sum of Sq       F    Pr(>F)
1     19 495.39
2     18 143.12  1    352.27 57.2768 1.131e-06 ***
3     17 109.95  1     33.17  5.3931   0.03373 *
4     16  98.40  1     11.55  1.8773   0.18956
```

Sequential sum of squares

We can get the sequential sum of squares directly using the `anova` function:

```
> anova(fit3)
Analysis of Variance Table

Response: bodyfat
          Df Sum Sq Mean Sq F value    Pr(>F)    
tricep     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                      
---
Signif. codes:  0 .***. 0.001 .**. 0.01 .*. 0.05 ... 0.1 . . 1
```

Sequential sum of squares with a different order

```
> fit4 <- lm(bodyfat ~ thigh+tricep+midarm)
> anova(fit4)
Analysis of Variance Table

Response: bodyfat
          Df Sum Sq Mean Sq F value    Pr(>F)
thigh      1 381.97  381.97 62.1052 6.735e-07 ***
tricep     1   3.47    3.47  0.5647    0.4633
midarm     1  11.55   11.55  1.8773    0.1896
Residuals 16  98.40    6.15
---
Signif. codes:  0 .***. 0.001 .**. 0.01 .*. 0.05 ... 0.1 . . 1
```

Note that the extra SS for `midarm` in two fits are the same. Why?

Changing the Order of Predictors in the Model

- Place a predictor in the model last in order to test the sequential effect of that predictor, given all other predictors in the model

```
Analysis of Variance Table

Response: bodyfat
          Df Sum Sq Mean Sq F value    Pr(>F)    
thigh      1 381.97  381.97 62.1052 6.735e-07 ***
midarm     1   2.31    2.31  0.3762    0.5483    
tricep     1  12.70   12.70  2.0657    0.1699    
Residuals 16  98.40    6.15                      
```

Test of tricep given thigh, midarm

Adding more than one predictor to the Model

- You can test the effect of adding more than one predictor to the model.
  - Model 1: $Y = \beta_0 + \beta_1 X_1 + \varepsilon$
  - Model 2: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$

```
> model1=lm(bodyfat~tricep)
> model2=lm(bodyfat~tricep+thigh+midarm)
> anova(model1,model2)
Analysis of Variance Table

Model 1: bodyfat ~ tricep
Model 2: bodyfat ~ tricep + 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 *
```

Test of thigh, midarm <u>given</u> tricep

Partial (type III) sum of squares

- To test the hypothesis $H_0: \beta_1 = 0$, we can fit the full model and the reduced model without X_1, and apply the extra sum of squares principal. That is, we need to compute $SSR(X_1|X_2, X_3)$ which is the extra sum of squares explained by X_1 when X_2 and X_3 are included in the model
- Similarly, to test $H_0: \beta_2 = 0$ and $H_0: \beta_3 = 0$, we need to compute $SSR(X_2|X_1, X_3)$ and $SSR(X_3|X_1, X_2)$
- One can calculate these SS's by fitting multiple models. Partial (type III) sum of squares computes them simultaneously

Partial (type III) sum of squares

Source	SS	H_0
X_1	$SSR(X_1 \| X_2, X_3)$	$\beta_1 = 0$
X_2	$SSR(X_2 \| X_1, X_3)$	$\beta_2 = 0$
X_3	$SSR(X_3 \| X_1, X_2)$	$\beta_3 = 0$

- SS for a given effect is adjusted for all other effects
- ANOVA table does not really make sense here since the SS's do not add up to $SSR(X_1, X_2, X_3)$
- Use type III or simply t-tests

Partial (type III) sum of squares for body fat data

We can get the sequential sum of squares from the `Anova` function in the library `car`:

```
> library(car)
> Anova(fit3, type="III")
Anova Table (Type III tests)

Response: bodyfat
            Sum Sq Df F value Pr(>F)
(Intercept)  8.468  1  1.3769 0.2578
tricep      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
```

Another Way to Get Partial p-values

- The Summary function in R will give you the partial p-values, but not the partial SS (these are the same p-values as the ANOVA on the prior slide)

```
> summary(model2)
Call:
lm(formula = bodyfat ~ tricep + thigh + midarm)
Residuals:
    Min      1Q  Median      3Q     Max
-3.7263 -1.6111  0.3923  1.4656  4.1277

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  117.085     99.782   1.173    0.258
tricep         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
```

Test of tricep given thigh, midarm

Test of thigh given tricep, midarm

Test of midarm given tricep, thigh

Extra sum of squares principle

Use the extra sum of squares principle to test general hypothesis. For example, to test $H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$:

```
> fit5 <- lm(bodyfat ~ thigh+midarm) # restricted model
> anova(fit5,fit3)
Analysis of Variance Table

Model 1: bodyfat ~ thigh + midarm
Model 2: bodyfat ~ tricep + thigh + midarm
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     17 111.110
2     16  98.405  1    12.705 2.0657 0.1699
```

Testing Addition of More Than 1 Predictor

To test $H_0: \beta_1 = \beta_2 = 0$ vs $H_1:$ not both β_1 and β_2 equal 0:

```
> fit6 <- lm(bodyfat ~ midarm)  # restricted model
> anova(fit6,fit3)
Analysis of Variance Table

Model 1: bodyfat ~ midarm
Model 2: bodyfat ~ tricep + thigh + midarm
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     18 485.34
2     16  98.40  2    386.93 31.456 2.856e-06 ***
---
Signif. codes:  0 .***. 0.001 .**. 0.01 .*. 0.05 ... 0.1 . . 1
```

Testing a Specific Value of β

To test $H_0: \beta_1 = 3$ vs $H_1: \beta_1 \neq 3$:

```
> fit7 <- lm(bodyfat ~ offset(3*tricep)+thigh+midarm)
> anova(fit7,fit3)
Analysis of Variance Table

Model 1: bodyfat ~ offset(3 * tricep) + thigh + midarm
Model 2: bodyfat ~ tricep + thigh + midarm
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     17 99.609
2     16 98.405  1     1.204 0.1957 0.6641

> confint(fit3)
                 2.5 %     97.5 %
(Intercept) -94.444550 328.613940
tricep       -2.058507  10.726691
thigh        -8.330476   2.616780
midarm       -5.568367   1.196247
```

Extra SS Summary

- Extra SS allow us to test the addition of predictor(s) to an existing model
- The ANOVA function provides a flexible method for comparing two models
  - Fit two different models

    ```
    fit1 = lm(Y=X1)
    fit2 = lm(Y~X1+X2+X3)
    ```

  - Compare using ANOVA (note: smaller model goes first)

    ```
    ANOVA(fit1,fit2)
    ```

  - A significant p-value tells you that the larger model fits better than the smaller model

Overview of Extra Sum of Squares

- The estimated regression coefficients are the same regardless of the order of predictors in the model

- The Summary function in R provides p-values based on Partial effects. Each predictor is tested GIVEN that all other predictors are in the model:
  - $X_1 | X_2, X_3$
  - $X_2 | X_1, X_3$
  - $X_3 | X_1, X_2$

- The ANOVA function in R provides p-values based on Sequential effects. Each predictor is tested GIVEN that earlier predictors are in the model:
  - X_1 alone
  - $X_2 | X_1$
  - $X_3 | X_1, X_2$
  - Significance depends on the order of predictors in the model (e.g., lm<-(Y~X1+X2+X3)

Summary of R functions

Type of Test	Purpose	R Function
Extra Sum of Squares	Test the addition of any number of predictors to the model	`anova(model1,model2)`
Sequential or Type I SS	Test the effect of predictor GIVEN previous predictors	`anova(model2)`
Partial or Type III SS	Test the effect of each predictor GIVEN all other predictors in the model	`summary(model2)`