Regression Analysis

BIOL 5081 INTRO TO BIOSTATISTICS

OCTAVIA MAHDIYAN, ELENI FEGARAS & KARAM DAHYALEH

Scatter plots

- Scatterplots are used to depict the relationship between 2 variables
    - Linear relationships
    - Curve linear relationships
    - Strong or weak relationships
    - No relationships

Covariance and correlation

- Covariance measures how 2 variables vary with respect to one another
  - Measures the direction of the linear relationship but does not measure the strength
- Correlation coefficient
  - Population correlation coefficient (ρ)
  - Sample correlation coefficient (r)
  - Measures the strength and direction of a linear relationship
  - Unit free and ranges from -1 to +1

$$r = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{n}(x_i - \mu_x)^2 . \sum_{i=1}^{n}(y_i - \mu_y)^2}}$$

- r = sample correlation coefficient
- n = sample size
- x = value of the predictor variable
- y = value of the response variable

Examples of *r* values

- Stronger linear relationships (r = -1, r = +1)
- Weaker linear relationships (r = -0.6, r = + 0.3)
- No linear relationship (r = 0)

Significance

Hypotheses

Null hypothesis → H_0: $\rho = 0$ (no correlation)

Alternate hypothesis → H_A: $\rho \neq 0$ (correlation)

t-value to test significance of a correlation

$$t = \frac{r}{\sqrt{\frac{(1-r^2)}{(n-2)}}}$$

- r = correlation coefficient
- df = n-2

Linear regression analysis

- Statistical analysis to describe the relationship between 2 or more continuous variables

response variable = model + error

- Simple linear regression is part of bivariate statistics
- Working with 2 variables
  - y variable = **response**, dependant, outcome
  - x variable = **predictor**, independent, explanatory

Linear model for regression

Slope intercept form a line

$\mathbf{y = m\,x + b + \varepsilon}$

x = random variable

m = slope of the line

b = y-intercept

Population linear regression model

$\mathbf{y_i = \beta_0 + \beta_1\,x_1 + \varepsilon_i}$

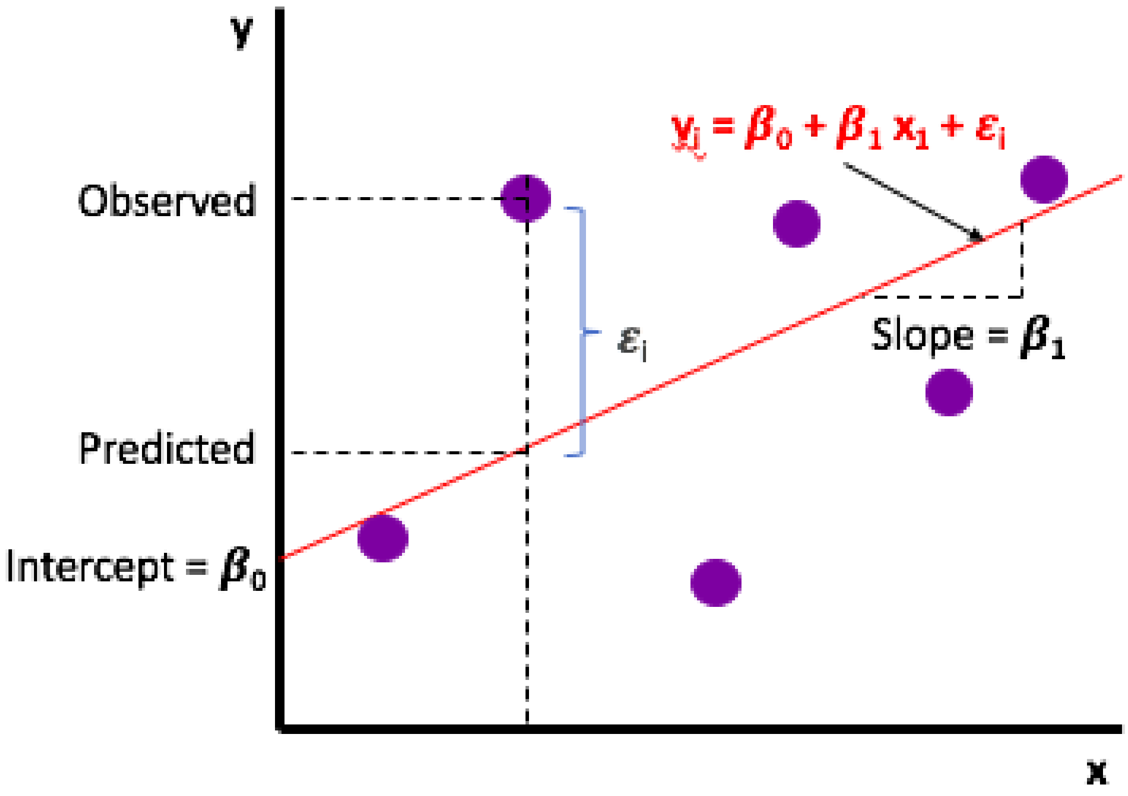
β_0 = population y-intercept

β_1 = population slope

x = predictor variable

ε = error term, unexplained variation in y

Linear regression model



Linear regression assumptions

- Individual variables are normally distributed
- The relationship between the x and y variable is linear
- Random sampling
- Independence of observations
- The probability distribution of the errors has a constant variance

Estimating model parameters

Sample regression line

$\hat{y}_i = b_0 + b_1 x_i$

$\hat{y}_i$ = value of the y_i predicted by the fitted regression line for each x

b_0 = estimate of the regression intercept

b_1 = estimate of the regression slope

x = predictor variable

The main aim of regression analysis is to estimate the parameters (β_0, β_1) of the linear regression model

Sample regression line provides an estimate of the population regression line using sample data

Sample regression line

- Model of the least squares regression line and residual values

- The difference between each observed Y-value and each predicted value $\hat{y}_i$ value is called a residual

Quinn, G. P., & Keough, M. J. (2002). *Experimental design and data analysis for biologists*. Cambridge University Press.

Analysis of variance

$$SST = SSR + SSE$$

Total sum of squares → measures the variation of the y_i values around their mean

Sum of squares regression → explained variation attributable to the relationship between x and y

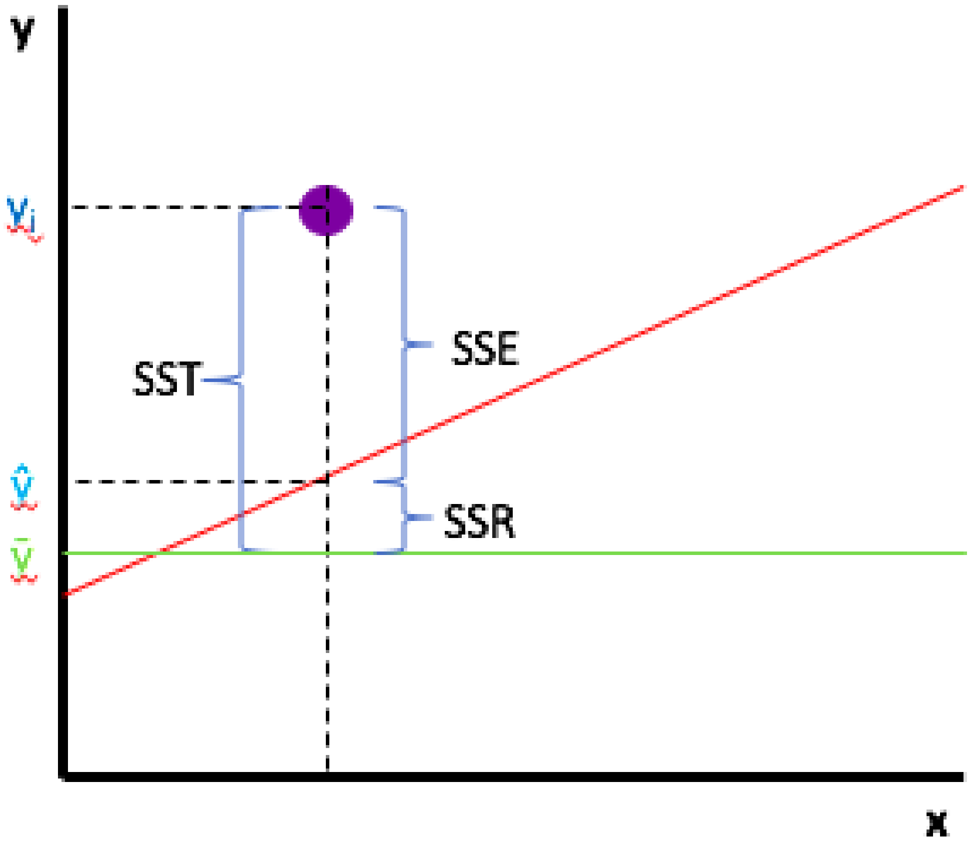
Sum of squares error → variation attributed to factors other than the relationship between x and y

Analysis of variance

$$SST = SSR + SSE$$

$SST = \Sigma(y-\bar{y})^2$ $SSR = \Sigma(\hat{y}-\bar{y})^2$ $SSE = \Sigma(y-\hat{y})^2$

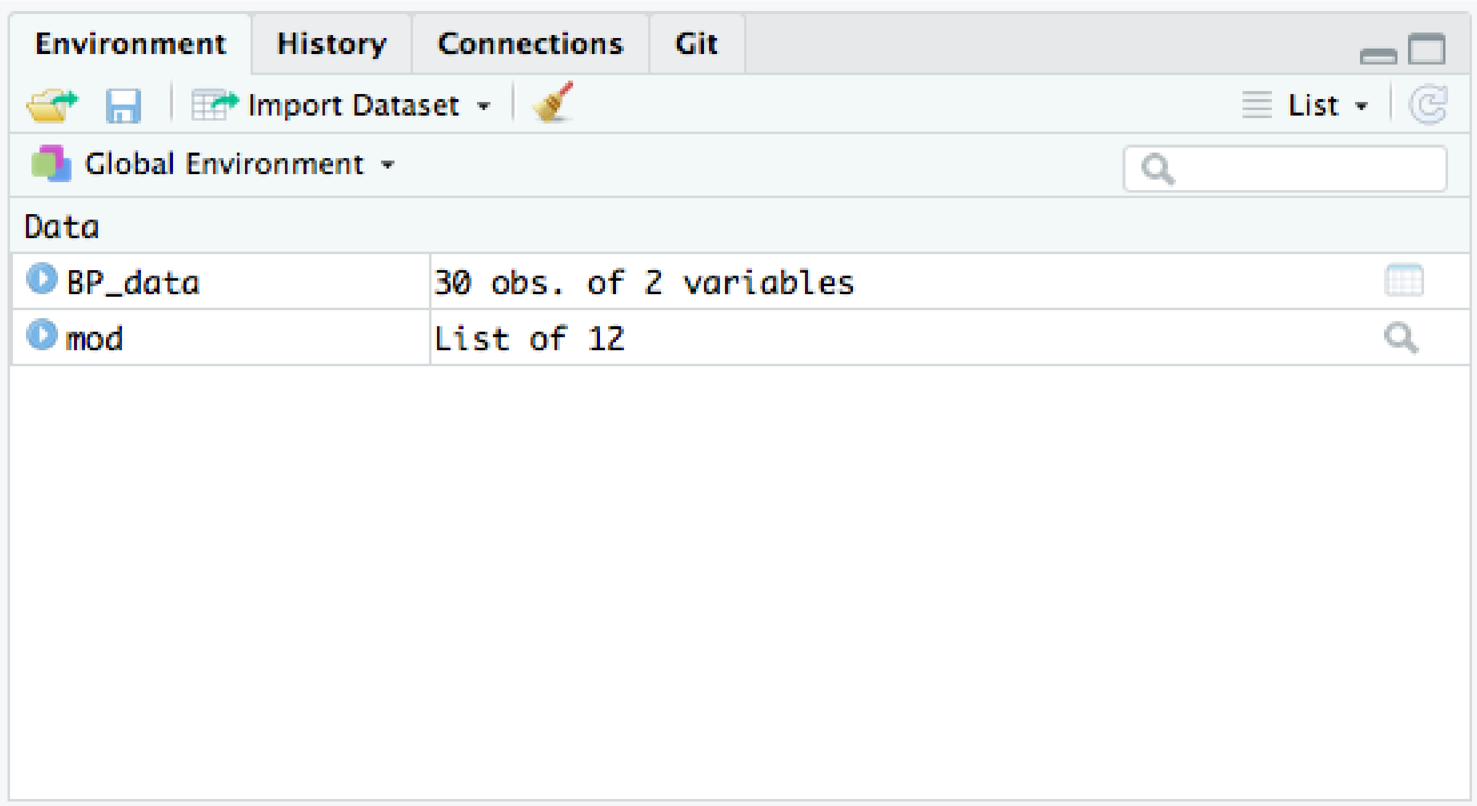
Explained and unexplained variation



Coefficient of Determination (R^2)

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = \frac{\text{sum of squares explained by regression}}{\text{total sum of squares}}$$

Linear regression in R



BP_data × | Linear regression script .R ×

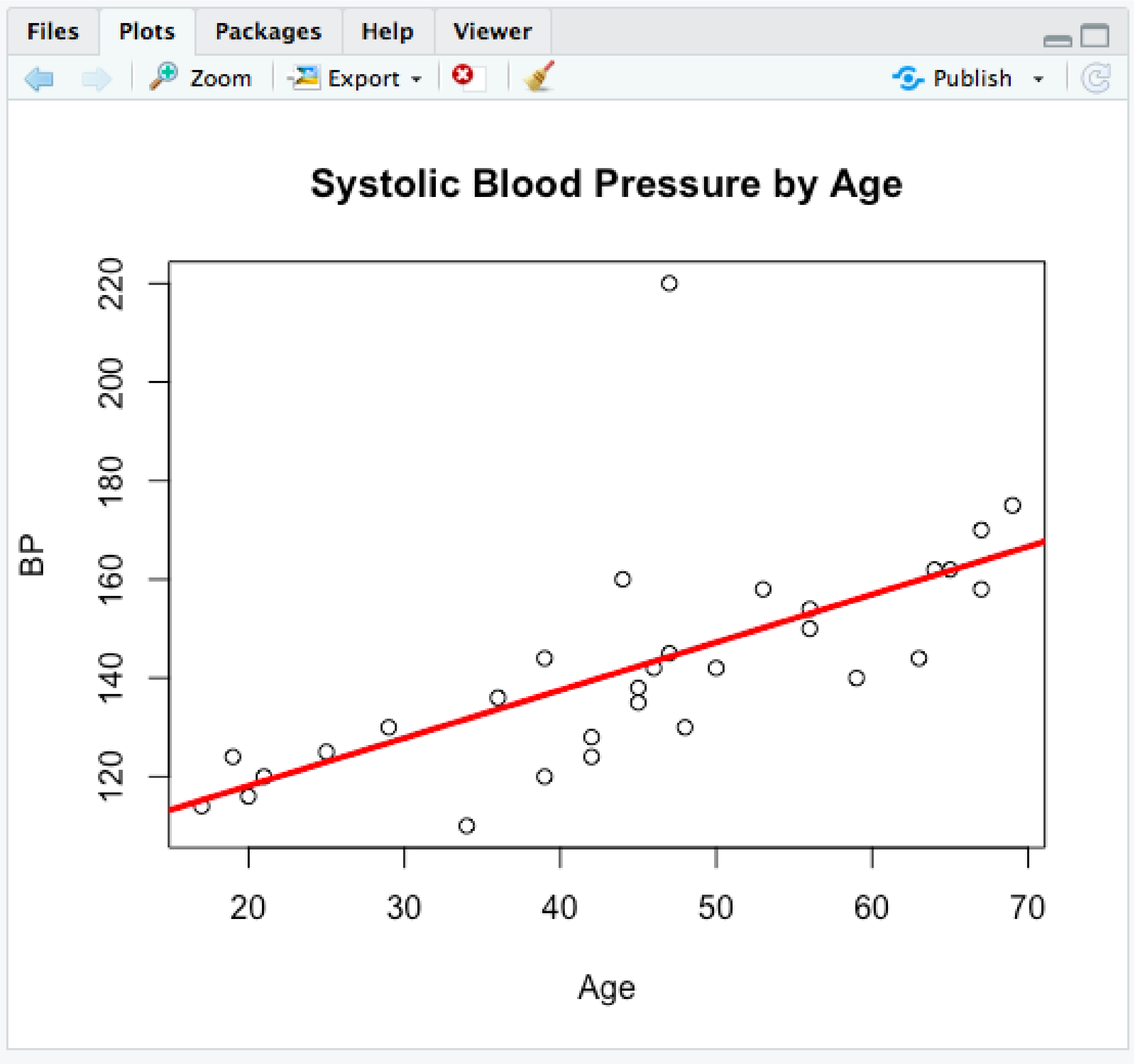
Filter

	Age	BP
1	39	144
2	47	220
3	45	138
4	47	145
5	65	162
6	46	142
7	67	170
8	42	124
9	67	158
10	56	154
11	64	162
12	56	150
13	59	140
14	34	110
15	42	128
16	48	130
17	45	135
18	17	114
19	20	116
20	19	124
21	36	136
22	50	142
23	39	120
24	21	120
25	44	160
26	53	158
27	63	144
28	29	130
29	25	125
30	69	175

Showing 1 to 30 of 30 entries

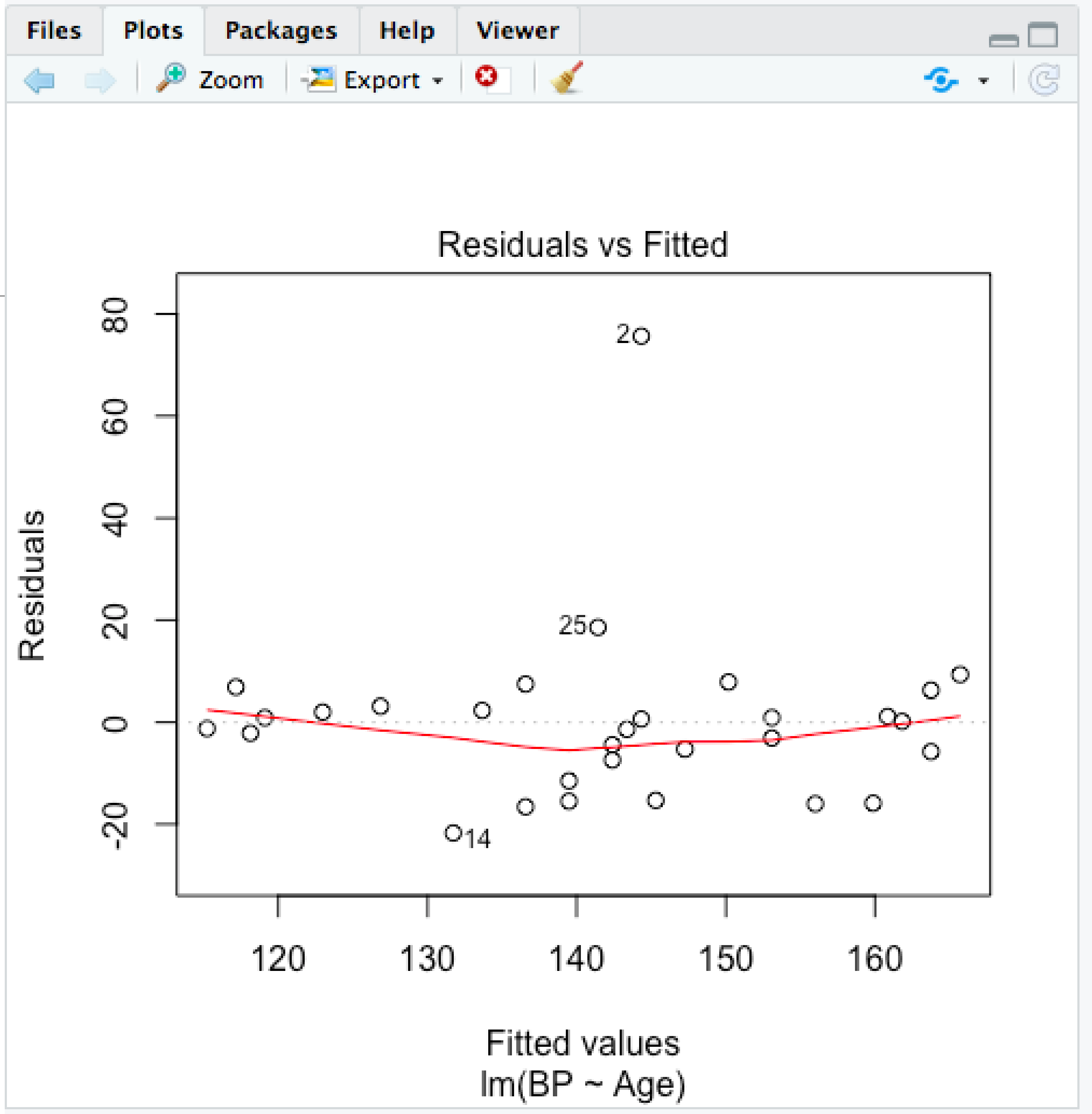
BP_data | Linear regression script .R

```r
#Import your dataset into your environment
#Preview the dimensions of the BP_data
dim(BP_data)

#See the names of the BP_data
names(BP_data)

#Check the type of variable for Age and BP
class(Age)
class(BP)

#Plot your data in a scatter plot
plot(Age, BP, main="Systolic Blood Pressure by Age")

#Fit a linear regression model with the lm fnx + summary
mod <- lm(BP ~ Age)
summary(mod)

#Adding the regression line to your model (can also change colour and line width)
abline(mod, col=2, lw=3)

summary(mod)
anova(mod)

#Regression disgnostic plots
plot(mod)
```


Files
Plots
Packages
Help
Viewer
Zoom
Export
Publish
Systolic Blood Pressure by Age
BP
220
200
180
160
140
120
20
30
40
50
60
70
Age

Console ~/Desktop/SWC/

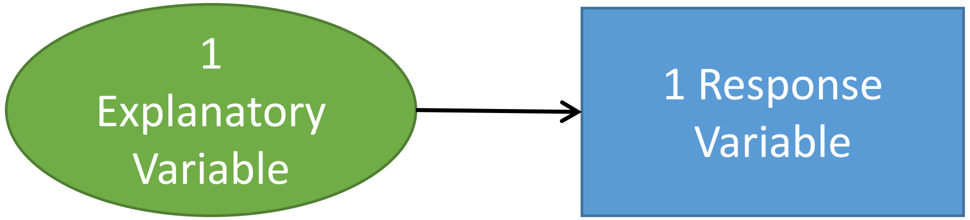
```
> summary(mod)

Call:
lm(formula = BP ~ Age)

Residuals:
    Min      1Q  Median      3Q     Max 
-21.724  -6.994  -0.520   2.931  75.654 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  98.7147    10.0005   9.871 1.28e-10 ***
Age           0.9709     0.2102   4.618 7.87e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.31 on 28 degrees of freedom
Multiple R-squared:  0.4324,	Adjusted R-squared:  0.4121 
F-statistic: 21.33 on 1 and 28 DF,  p-value: 7.867e-05

>
>
>
> anova(mod)
Analysis of Variance Table

Response: BP
          Df Sum Sq Mean Sq F value    Pr(>F)    
Age        1 6394.0  6394.0   21.33 7.867e-05 ***
Residuals 28 8393.4   299.8                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```


Files
Plots
Packages
Help
Viewer
Zoom
Export
Residuals vs Fitted
Residuals
80
60
40
20
0
-20
2
25
14
120
130
140
150
160
Fitted values
lm(BP ~ Age)

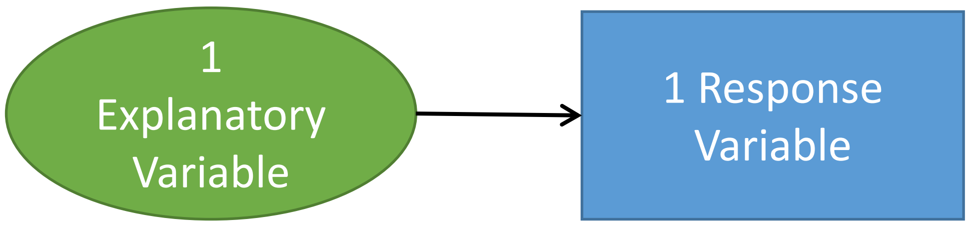
Multiple Regression

Eleni Fegaras

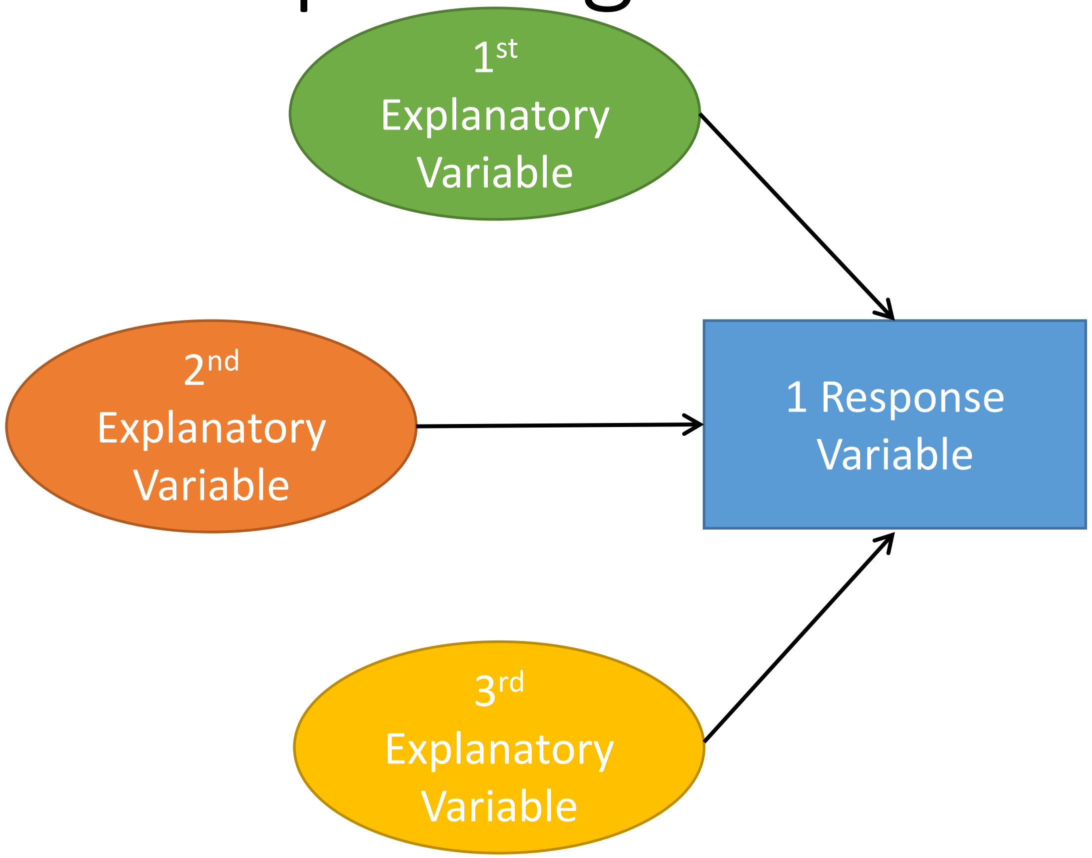
Linear Regression



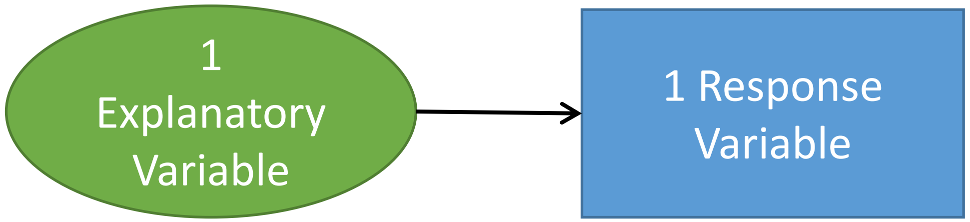
Linear Regression



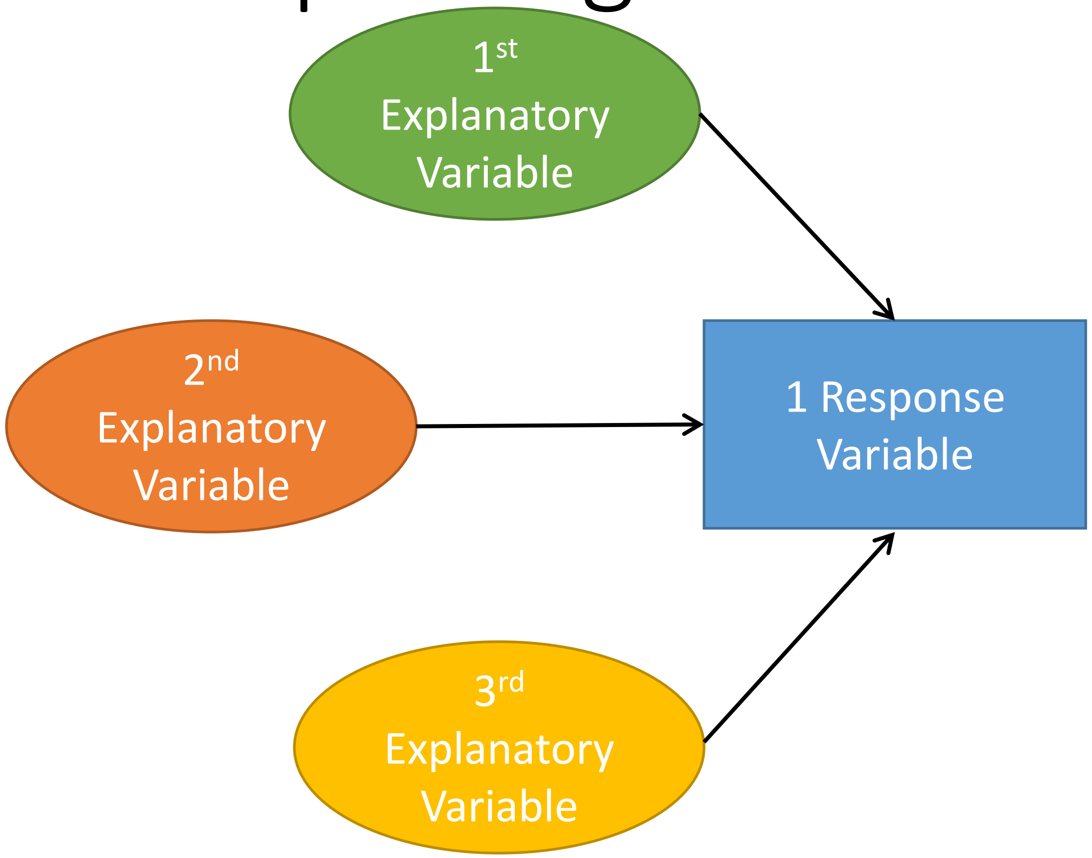
Multiple Regression



Linear Regression



Multiple Regression



- Prediction: the value of a variable based on the value of 2+ other variables
- Causal: You can determine the overall fit of the model and the relative contribution of each explanatory variable to the response

Linear Regression

- Population model
- $y_i = \beta_0 + \beta_1 x_{i\,1} + \varepsilon_i$
- β_0 = population y-intercept
- β_1 = population slope
- x = predictor variable
- ε = error term, unexplained variation in y

Multiple Regression

- Population model
- $y_i = \beta_0 + \beta_1 x_{i\,1} + \beta_2 x_{i\,2} + \cdots + \beta_k x_{i\,k} + \varepsilon_i$
- β_0 = population y-intercept
- $\beta_{1,2\ldots k}$ = population slope for that predictor variable, holding other variables constant
- $x_{1,2\ldots k}$ = predictor variable
- ε = error term, unexplained variation in y

Linear Regression

- Population model
- $y_i = \beta_0 + \beta_1 x_{i\,1} + \varepsilon_i$
- β_0 = population y-intercept
- β_1 = population slope
- x = predictor variable
- ε = error term, unexplained variation in y

Predicted regression line

$\hat{y}_i = b_0 + b_1 x_{i1}$

Multiple Regression

- Population model
- $y_i = \beta_0 + \beta_1 x_{i\,1} + \beta_2 x_{i\,2} + \cdots + \beta_k x_{i\,k} + \varepsilon_i$
- β_0 = population y-intercept
- $\beta_{1,2\ldots k}$ = population slope for that predictor variable, holding other variables constant
- $x_{1,2\ldots k}$ = predictor variable
- ε = error term, unexplained variation in y

Predicted regression line

$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_k x_{ik}$

Our example case study:

- Q: Are a person's brain size and body size predictive of his or her intelligence? Willerman *et al.*, 1991
- <u>Response variable </u>($\mathbf{y}_i$): Performance IQ (PIQ) from the Wechsler Adult Intelligence Scale
- <u>Explanatory variables</u>: ($\mathbf{x}_{i\,1}$) Brain size in MRI ($\mathbf{x}_{i\,2}$) Height in inches ($\mathbf{x}_{i\,3}$) Weight in pounds

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

$$(PIQ)_I = \beta_0 + \beta_1 \text{(brain size)} + \beta_2 \text{(height inches)} + \beta_3 \text{(weight pounds)}$$

Some Additional Assumptions

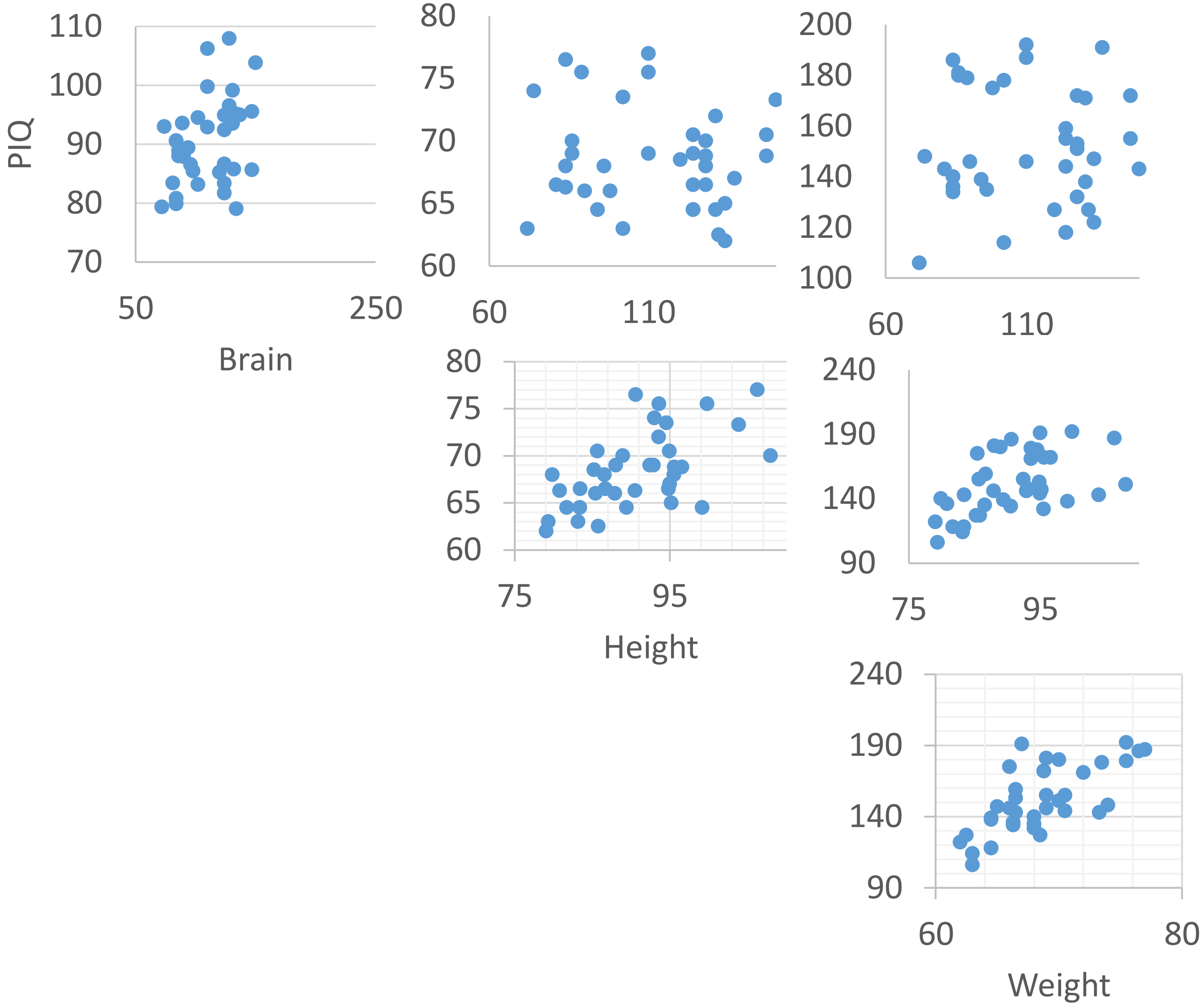
1. Linear relationship between the response variable and each of the explanatory variables, and the response variable and the explanatory variables collectively

2. Try to eliminate multicollinearity

3. Minimum number of observations

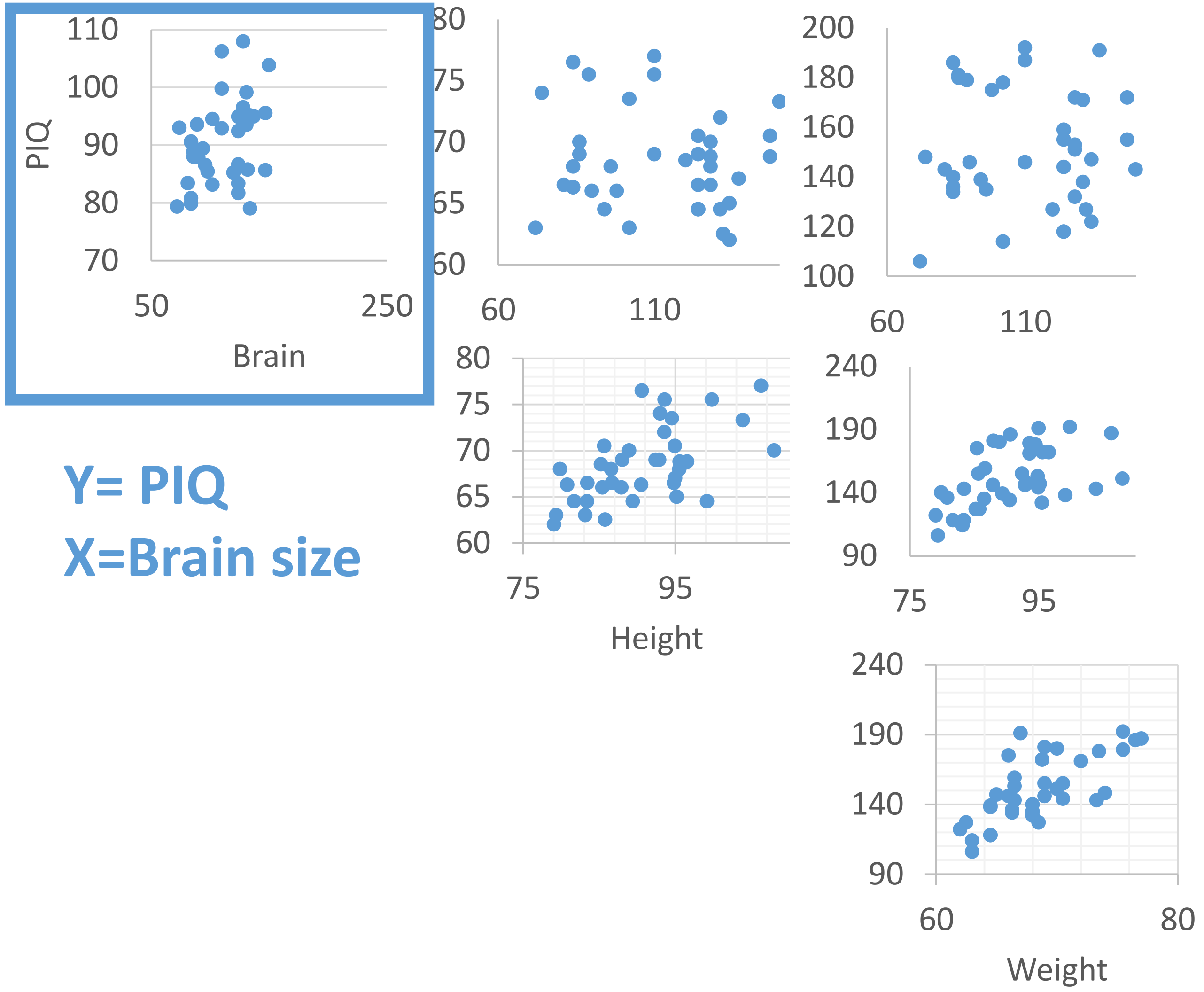
1) Linear Relationship

- **Scatter plot matrixes**
- Investigate the relationships among all the variables
- Illustrates “marginal relationships”; no regard to other variables



1) Linear Relationship

- **Scatter plot matrixes**
- Investigate the relationships among all the variables
- Illustrates "marginal relationships"; no regard to other variables

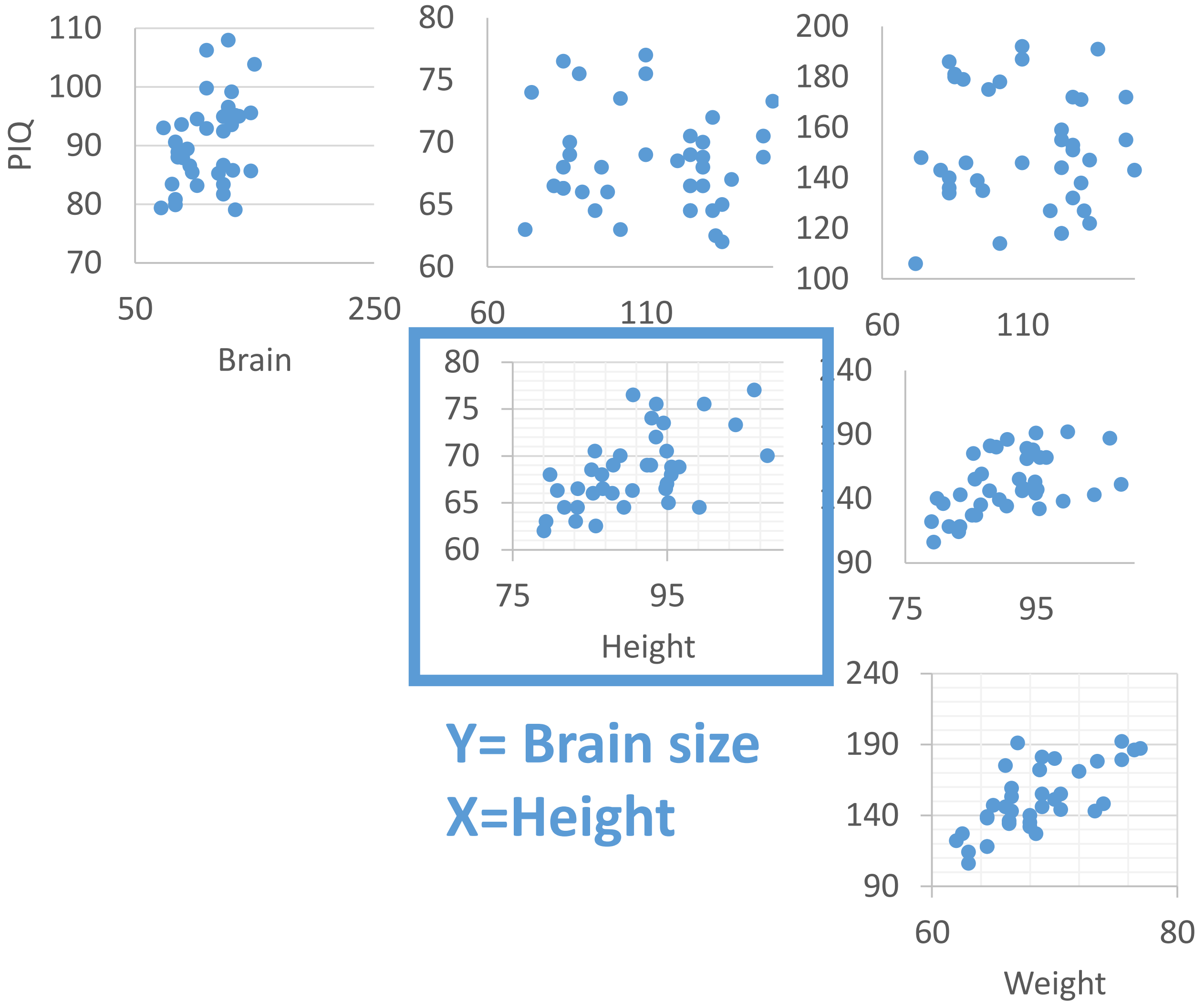


Y= PIQ

X=Brain size

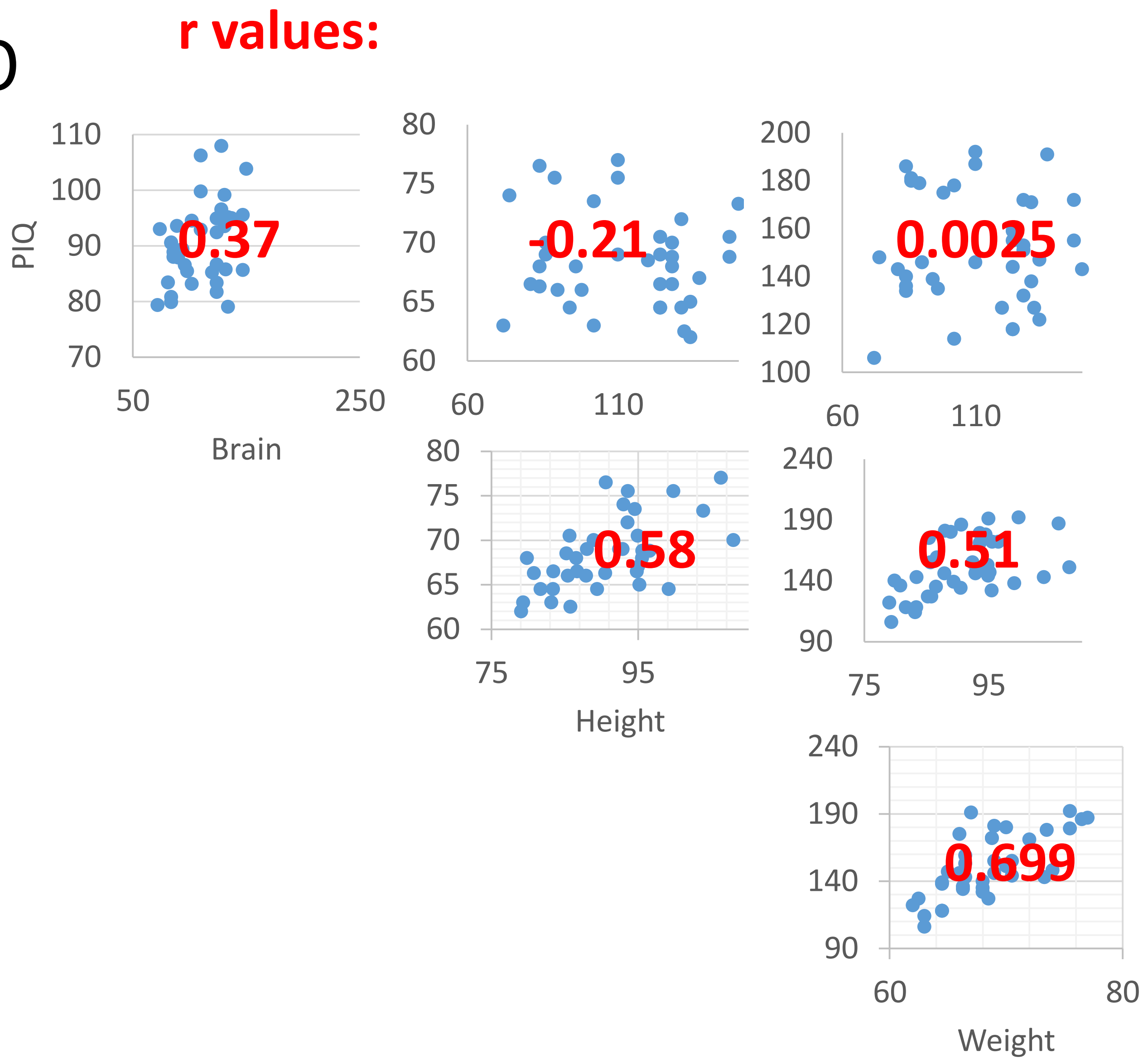
1) Linear Relationship

- **Scatter plot matrixes**
- Investigate the relationships among all the variables
- Illustrates "marginal relationships"; no regard to other variables

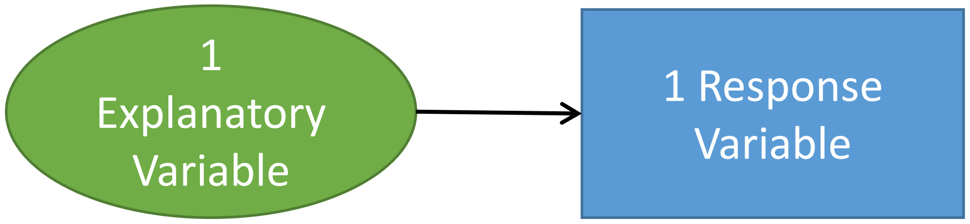


1) Linear Relationship

- **Scatter plot matrixes**
- Investigate the relationships among all the variables
- Illustrates "marginal relationships"; no regard to other variables

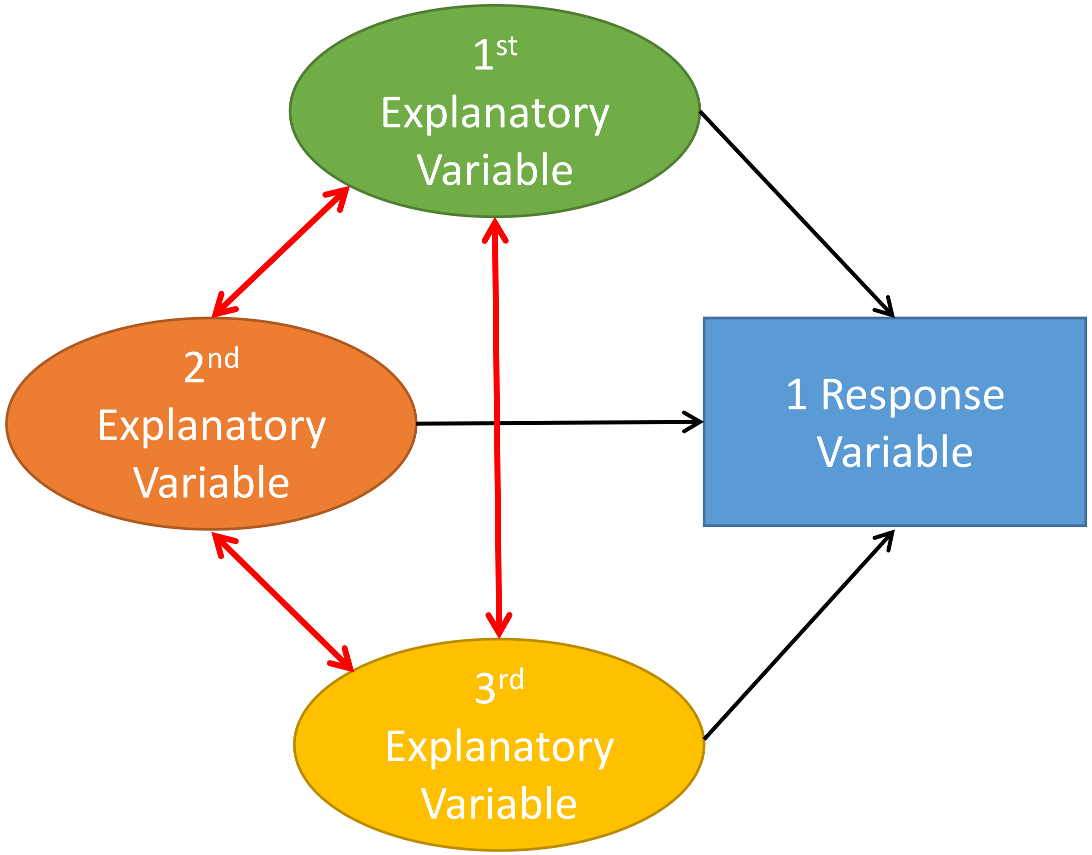


Linear Regression



Multiple Regression

2) Multicollinearity



2) The Issues of Multicollinearity

Multicollinearity is the most often faced issue

- 1) small changes to data (adding or deleting data) can greatly change the estimated regression coefficients
- 2) standard errors of the estimated regression slopes are inflated

Basically: different sample, different population may yield very different results

3) Minimum number of observations

- Green (1991) ratio of # of predictors + 104 : observations
- Neter et al (1996) ratio of 6-10(# of predictors) : observations
- Maximize your number of observations
- If you must, reduce the number of variables you're testing

- For example: Our study 38 volunteers, 3 predictors:

  (3) + 104 > 38 ☒

  6(3) to 10(3) = 18 to 30 < 38 ☑

Setup in R statistics

1. Estimated model coefficients and regression equation
2. Determine how well the model fits (r-squared)
3. Which explanatory variables contributes the most (ANOVA)
4. Choosing the best model (AICc and Partial F-test)

```
>
> data1 <-read.delim(file.choose(), header=T)
> data1
   PIQ  Brain Height Weight
1  124  81.69   64.5    118
2  150 103.84   73.3    143
3  128  96.54   68.8    172
4  134  95.15   65.0    147
5  110  92.88   69.0    146
6  131  99.13   64.5    138
7   98  85.43   66.0    175
8   84  90.49   66.3    134
9  147  95.55   68.8    172
10 124  83.39   64.5    118
11 128 107.95   70.0    151
12 124  92.41   69.0    155
13 147  85.65   70.5    155
14  90  87.89   66.0    146
15  96  86.54   68.0    135
16 120  85.22   68.5    127
17 102  94.51   73.5    178
18  84  80.80   66.3    136
19  86  88.91   70.0    180
20  84  90.59   76.5    186
21 134  79.06   62.0    122
22 128  95.50   68.0    132
23 102  83.18   63.0    114
24 131  93.55   72.0    171
25  84  79.86   68.0    140
26 110 106.25   77.0    187
27  72  79.35   63.0    106
28 124  86.67   66.5    159
29 132  85.78   62.5    127
30 137  94.96   67.0    191
31 110  99.79   75.5    192
32  86  88.00   69.0    181
33  81  83.43   66.5    143
34 128  94.81   66.5    153
35 124  94.94   70.5    144
36  94  89.40   64.5    139
37  74  93.00   74.0    148
38  89  93.59   75.5    179
```

What about non-numeric data in R?

Ordinal scale represent use “dummy variables”

Or more simpler categories you assign male – 1 female – 0

```
>
> data1 <-read.delim(file.choose(), header=T)
> data1
   PIQ  Brain Height Weight
1  124  81.69   64.5    118
2  150 103.84   73.3    143
3  128  96.54   68.8    172
4  134  95.15   65.0    147
5  110  92.88   69.0    146
6  131  99.13   64.5    138
7   98  85.43   66.0    175
8   84  90.49   66.3    134
9  147  95.55   68.8    172
10 124  83.39   64.5    118
11 128 107.95   70.0    151
12 124  92.41   69.0    155
13 147  85.65   70.5    155
14  90  87.89   66.0    146
15  96  86.54   68.0    135
16 120  85.22   68.5    127
17 102  94.51   73.5    178
18  84  80.80   66.3    136
19  86  88.91   70.0    180
20  84  90.59   76.5    186
21 134  79.06   62.0    122
22 128  95.50   68.0    132
23 102  83.18   63.0    114
24 131  93.55   72.0    171
25  84  79.86   68.0    140
26 110 106.25   77.0    187
27  72  79.35   63.0    106
28 124  86.67   66.5    159
29 132  85.78   62.5    127
30 137  94.96   67.0    191
31 110  99.79   75.5    192
32  86  88.00   69.0    181
33  81  83.43   66.5    143
34 128  94.81   66.5    153
35 124  94.94   70.5    144
36  94  89.40   64.5    139
37  74  93.00   74.0    148
38  89  93.59   75.5    179
```

```
> model <- lm(PIQ ~ Brain + Height + Weight)
> summary(model)

Call:
lm(formula = PIQ ~ Brain + Height + Weight)

Residuals:
    Min      1Q  Median      3Q     Max 
-32.74 -12.09  -3.84  14.17  51.69 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.114e+02  6.297e+01   1.768 0.085979 .  
Brain        2.060e+00  5.634e-01   3.657 0.000856 ***
Height      -2.732e+00  1.229e+00  -2.222 0.033034 *  
Weight       5.599e-04  1.971e-01   0.003 0.997750    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.79 on 34 degrees of freedom
Multiple R-squared:  0.2949,	Adjusted R-squared:  0.2327 
F-statistic: 4.741 on 3 and 34 DF,  p-value: 0.007215
```

```
> confint(model1, conf.level=0.95)
                  2.5 %      97.5 %
(Intercept) -16.6190567 239.3262733
Brain         0.9153051   3.2054285
Height       -5.2304287  -0.2334296
Weight       -0.3999266   0.4010465
```

```
>
> data1 <-read.delim(file.choose(), header=T)
> data1
   PIQ  Brain Height Weight
1  124  81.69   64.5    118
2  150 103.84   73.3    143
3  128  96.54   68.8    172
4  134  95.15   65.0    147
5  110  92.88   69.0    146
6  131  99.13   64.5    138
7   98  85.43   66.0    175
8   84  90.49   66.3    134
9  147  95.55   68.8    172
10 124  83.39   64.5    118
11 128 107.95   70.0    151
12 124  92.41   69.0    155
13 147  85.65   70.5    155
14  90  87.89   66.0    146
15  96  86.54   68.0    135
16 120  85.22   68.5    127
17 102  94.51   73.5    178
18  84  80.80   66.3    136
19  86  88.91   70.0    180
20  84  90.59   76.5    186
21 134  79.06   62.0    122
22 128  95.50   68.0    132
23 102  83.18   63.0    114
24 131  93.55   72.0    171
25  84  79.86   68.0    140
26 110 106.25   77.0    187
27  72  79.35   63.0    106
28 124  86.67   66.5    159
29 132  85.78   62.5    127
30 137  94.96   67.0    191
31 110  99.79   75.5    192
32  86  88.00   69.0    181
33  81  83.43   66.5    143
34 128  94.81   66.5    153
35 124  94.94   70.5    144
36  94  89.40   64.5    139
37  74  93.00   74.0    148
38  89  93.59   75.5    179
```

```
> model <- lm(PIQ ~ Brain + Height + Weight)
> summary(model)

Call:
lm(formula = PIQ ~ Brain + Height + Weight)

Residuals:
    Min      1Q  Median      3Q     Max 
-32.74  -12.09   -3.84   14.17   51.69 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.114e+02  6.297e+01   1.768 0.085979 .  
Brain        2.060e+00  5.634e-01   3.657 0.000856 ***
Height      -2.732e+00  1.229e+00  -2.222 0.033034 *  
Weight       5.599e-04  1.971e-01   0.003 0.997750    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.79 on 34 degrees of freedom
Multiple R-squared:  0.2949,    Adjusted R-squared:  0.2327 
F-statistic: 4.741 on 3 and 34 DF,  p-value: 0.007215
```

```
> confint(model1, conf.level=0.95)
                   2.5 %      97.5 %
(Intercept) -16.6190567 239.3262733
Brain         0.9153051   3.2054285
Height       -5.2304287  -0.2334296
Weight       -0.3999266   0.4010465
```

Regression Equation:

$(PIQ)_I = \beta_0 + \beta_1$ (brain size) $+ \beta_2$(height inches) $+ \beta_3$ (weight pounds)

$(PIQ)_I = 111.4 + 2.06$ (brain size) $- 2.73$ (height inches) $+ 0.001$ (weight pounds)

```
>
> data1 <-read.delim(file.choose(), header=T)
> data1
   PIQ  Brain Height Weight
1  124  81.69   64.5    118
2  150 103.84   73.3    143
3  128  96.54   68.8    172
4  134  95.15   65.0    147
5  110  92.88   69.0    146
6  131  99.13   64.5    138
7   98  85.43   66.0    175
8   84  90.49   66.3    134
9  147  95.55   68.8    172
10 124  83.39   64.5    118
11 128 107.95   70.0    151
12 124  92.41   69.0    155
13 147  85.65   70.5    155
14  90  87.89   66.0    146
15  96  86.54   68.0    135
16 120  85.22   68.5    127
17 102  94.51   73.5    178
18  84  80.80   66.3    136
19  86  88.91   70.0    180
20  84  90.59   76.5    186
21 134  79.06   62.0    122
22 128  95.50   68.0    132
23 102  83.18   63.0    114
24 131  93.55   72.0    171
25  84  79.86   68.0    140
26 110 106.25   77.0    187
27  72  79.35   63.0    106
28 124  86.67   66.5    159
29 132  85.78   62.5    127
30 137  94.96   67.0    191
31 110  99.79   75.5    192
32  86  88.00   69.0    181
33  81  83.43   66.5    143
34 128  94.81   66.5    153
35 124  94.94   70.5    144
36  94  89.40   64.5    139
37  74  93.00   74.0    148
38  89  93.59   75.5    179
```

```
> model <- lm(PIQ ~ Brain + Height + Weight)
> summary(model)

Call:
lm(formula = PIQ ~ Brain + Height + Weight)

Residuals:
    Min      1Q  Median      3Q     Max
-32.74  -12.09   -3.84   14.17   51.69

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.114e+02  6.297e+01   1.768 0.085979 .
Brain        2.060e+00  5.634e-01   3.657 0.000856 ***
Height      -2.732e+00  1.229e+00  -2.222 0.033034 *
Weight       5.599e-04  1.971e-01   0.003 0.997750
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.79 on 34 degrees of freedom
Multiple R-squared:  0.2949,    Adjusted R-squared:  0.2327
F-statistic: 4.741 on 3 and 34 DF,  p-value: 0.007215
```

```
> confint(model1, conf.level=0.95)
                   2.5 %      97.5 %
(Intercept) -16.6190567 239.3262733
Brain         0.9153051   3.2054285
Height       -5.2304287  -0.2334296
Weight       -0.3999266   0.4010465
```

Regression Equation:

$$(PIQ)_i = \beta_0 + \beta_1 \text{ (brain size)} + \beta_2 \text{(height inches)} + \beta_3 \text{ (weight pounds)}$$

↓

$$(PIQ)_i = 111.4 + 2.06 \text{ (brain size)} - 2.73 \text{ (height inches)} + 0.001 \text{ (weight pounds)}$$

```
>
> data1 <-read.delim(file.choose(), header=T)
> data1
   PIQ  Brain Height Weight
1  124  81.69   64.5    118
2  150 103.84   73.3    143
3  128  96.54   68.8    172
4  134  95.15   65.0    147
5  110  92.88   69.0    146
6  131  99.13   64.5    138
7   98  85.43   66.0    175
8   84  90.49   66.3    134
9  147  95.55   68.8    172
10 124  83.39   64.5    118
11 128 107.95   70.0    151
12 124  92.41   69.0    155
13 147  85.65   70.5    155
14  90  87.89   66.0    146
15  96  86.54   68.0    135
16 120  85.22   68.5    127
17 102  94.51   73.5    178
18  84  80.80   66.3    136
19  86  88.91   70.0    180
20  84  90.59   76.5    186
21 134  79.06   62.0    122
22 128  95.50   68.0    132
23 102  83.18   63.0    114
24 131  93.55   72.0    171
25  84  79.86   68.0    140
26 110 106.25   77.0    187
27  72  79.35   63.0    106
28 124  86.67   66.5    159
29 132  85.78   62.5    127
30 137  94.96   67.0    191
31 110  99.79   75.5    192
32  86  88.00   69.0    181
33  81  83.43   66.5    143
34 128  94.81   66.5    153
35 124  94.94   70.5    144
36  94  89.40   64.5    139
37  74  93.00   74.0    148
38  89  93.59   75.5    179
```

```
> model <- lm(PIQ ~ Brain + Height + Weight)
> summary(model)

Call:
lm(formula = PIQ ~ Brain + Height + Weight)

Residuals:
    Min      1Q  Median      3Q     Max
-32.74  -12.09   -3.84   14.17   51.69

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.114e+02  6.297e+01   1.768 0.085979 .
Brain        2.060e+00  5.634e-01   3.657 0.000856 ***
Height      -2.732e+00  1.229e+00  -2.222 0.033034 *
Weight       5.599e-04  1.971e-01   0.003 0.997750
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.79 on 34 degrees of freedom
Multiple R-squared:  0.2949,	Adjusted R-squared:  0.2327
F-statistic: 4.741 on 3 and 34 DF,  p-value: 0.007215
```

```
> confint(model1, conf.level=0.95)
                   2.5 %      97.5 %
(Intercept) -16.6190567 239.3262733
Brain         0.9153051   3.2054285
Height       -5.2304287  -0.2334296
Weight       -0.3999266   0.4010465
```

Regression Equation:

$$(PIQ)_I = \beta_0 + \beta_1 \text{(brain size)} + \beta_2 \text{(height inches)} + \beta_3 \text{(weight pounds)}$$

↓

$$(PIQ)_I = 111.4 + 2.06 \text{(brain size)} - 2.73 \text{(height inches)} + 0.001 \text{(weight pounds)}$$

```
>
> data1 <-read.delim(file.choose(), header=T)
> data1
   PIQ  Brain Height Weight
1  124  81.69   64.5    118
2  150 103.84   73.3    143
3  128  96.54   68.8    172
4  134  95.15   65.0    147
5  110  92.88   69.0    146
6  131  99.13   64.5    138
7   98  85.43   66.0    175
8   84  90.49   66.3    134
9  147  95.55   68.8    172
10 124  83.39   64.5    118
11 128 107.95   70.0    151
12 124  92.41   69.0    155
13 147  85.65   70.5    155
14  90  87.89   66.0    146
15  96  86.54   68.0    135
16 120  85.22   68.5    127
17 102  94.51   73.5    178
18  84  80.80   66.3    136
19  86  88.91   70.0    180
20  84  90.59   76.5    186
21 134  79.06   62.0    122
22 128  95.50   68.0    132
23 102  83.18   63.0    114
24 131  93.55   72.0    171
25  84  79.86   68.0    140
26 110 106.25   77.0    187
27  72  79.35   63.0    106
28 124  86.67   66.5    159
29 132  85.78   62.5    127
30 137  94.96   67.0    191
31 110  99.79   75.5    192
32  86  88.00   69.0    181
33  81  83.43   66.5    143
34 128  94.81   66.5    153
35 124  94.94   70.5    144
36  94  89.40   64.5    139
37  74  93.00   74.0    148
38  89  93.59   75.5    179
```

```
> model <- lm(PIQ ~ Brain + Height + Weight)
> summary(model)

Call:
lm(formula = PIQ ~ Brain + Height + Weight)

Residuals:
    Min      1Q  Median      3Q     Max
-32.74  -12.09   -3.84   14.17   51.69

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.114e+02  6.297e+01   1.768 0.085979 .
Brain        2.060e+00  5.634e-01   3.657 0.000856 ***
Height      -2.732e+00  1.229e+00  -2.222 0.033034 *
Weight       5.599e-04  1.971e-01   0.003 0.997750
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.79 on 34 degrees of freedom
Multiple R-squared:  0.2949,    Adjusted R-squared:  0.2327
F-statistic: 4.741 on 3 and 34 DF,  p-value: 0.007215
```

```
> confint(model1, conf.level=0.95)
                   2.5 %       97.5 %
(Intercept) -16.6190567  239.3262733
Brain         0.9153051    3.2054285
Height       -5.2304287   -0.2334296
Weight       -0.3999266    0.4010465
```

Regression Equation:

$$(PIQ)_I = \beta_0 + \beta_1 \text{ (brain size)} + \beta_2 \text{(height inches)} + \beta_3 \text{ (weight pounds)}$$

↓

$$(PIQ)_I = 111.4 + 2.06 \text{ (brain size)} - 2.73 \text{ (height inches)} + 0.001 \text{ (weight pounds)}$$

```
>
> data1 <-read.delim(file.choose(), header=T)
> data1
   PIQ  Brain Height Weight
1  124  81.69   64.5    118
2  150 103.84   73.3    143
3  128  96.54   68.8    172
4  134  95.15   65.0    147
5  110  92.88   69.0    146
6  131  99.13   64.5    138
7   98  85.43   66.0    175
8   84  90.49   66.3    134
9  147  95.55   68.8    172
10 124  83.39   64.5    118
11 128 107.95   70.0    151
12 124  92.41   69.0    155
13 147  85.65   70.5    155
14  90  87.89   66.0    146
15  96  86.54   68.0    135
16 120  85.22   68.5    127
17 102  94.51   73.5    178
18  84  80.80   66.3    136
19  86  88.91   70.0    180
20  84  90.59   76.5    186
21 134  79.06   62.0    122
22 128  95.50   68.0    132
23 102  83.18   63.0    114
24 131  93.55   72.0    171
25  84  79.86   68.0    140
26 110 106.25   77.0    187
27  72  79.35   63.0    106
28 124  86.67   66.5    159
29 132  85.78   62.5    127
30 137  94.96   67.0    191
31 110  99.79   75.5    192
32  86  88.00   69.0    181
33  81  83.43   66.5    143
34 128  94.81   66.5    153
35 124  94.94   70.5    144
36  94  89.40   64.5    139
37  74  93.00   74.0    148
38  89  93.59   75.5    179
```

```
> model <- lm(PIQ ~ Brain + Height + Weight)
> summary(model)

Call:
lm(formula = PIQ ~ Brain + Height + Weight)

Residuals:
    Min      1Q  Median      3Q     Max 
-32.74  -12.09   -3.84   14.17   51.69 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.114e+02  6.297e+01   1.768 0.085979 .  
Brain        2.060e+00  5.634e-01   3.657 0.000856 ***
Height      -2.732e+00  1.229e+00  -2.222 0.033034 *  
Weight       5.599e-04  1.971e-01   0.003 0.997750    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.79 on 34 degrees of freedom
Multiple R-squared:  0.2949,	Adjusted R-squared:  0.2327 
F-statistic: 4.741 on 3 and 34 DF,  p-value: 0.007215
```

```
> confint(model1, conf.level=0.95)
                  2.5 %      97.5 %
(Intercept) -16.6190567 239.3262733
Brain         0.9153051   3.2054285
Height       -5.2304287  -0.2334296
Weight       -0.3999266   0.4010465
```

Regression Equation:

$$(PIQ)_I = \beta_0 + \beta_1 \text{ (brain size)} + \beta_2 \text{(height inches)} + \beta_3 \text{ (weight pounds)}$$

$$(PIQ)_I = 111.4 + 2.06 \text{ (brain size)} - 2.73 \text{ (height inches)} + 0.001 \text{ (weight pounds)}$$

Which explanatory variables contribute the most

```
> library(car)
> Anova(model, type="III") # Adjusted (type III)
Anova Table (Type III tests)

Response: PIQ
             Sum Sq Df F value    Pr(>F)
(Intercept)  1225.2  1  3.1270 0.0859785 .
Brain        5239.2  1 13.3716 0.0008556 ***
Height       1934.7  1  4.9378 0.0330338 *
Weight          0.0  1  0.0000 0.9977495
Residuals   13321.8 34
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

***So far we don’t know if the model with these three explanatory variables is the *best* model!**

look at AICc and Partial F-Test

Choosing the Best Model

```
> model.1  = lm(PIQ ~ Brain,                  data=data1)
> model.2  = lm(PIQ ~ Height,                 data=data1)
> model.3  = lm(PIQ ~ Weight,                 data=data1)
> model.4  = lm(PIQ ~ Brain  + Height,        data=data1)
> model.5  = lm(PIQ ~ Brain  + Weight,        data=data1)
> model.6  = lm(PIQ ~ Brain  + Height + Weight, data=data1)
> library(rcompanion)
Error in library(rcompanion) : there is no package called ‘rcompanion’
> install.packages("rcompanion")
> library(rcompanion)
>
> compareLM(model.1, model.2, model.3, model.4, model.5, model.6)
$Models
  Formula
1 "PIQ ~ Brain"
2 "PIQ ~ Height"
3 "PIQ ~ Weight"
4 "PIQ ~ Brain + Height"
5 "PIQ ~ Brain + Weight"
6 "PIQ ~ Brain + Height + Weight"

$Fit.criteria
  Rank Df.res   AIC  AICc   BIC R.squared Adj.R.sq  p.value Shapiro.W
1    2     36 343.9 344.6 348.8 1.427e-01  0.11890 0.019350    0.9574
2    2     36 349.5 350.2 354.4 8.678e-03 -0.01886 0.578000    0.9415
3    2     36 349.8 350.5 354.7 6.311e-06 -0.02777 0.988100    0.9313
4    3     35 338.5 339.7 345.1 2.949e-01  0.25460 0.002208    0.9760
5    3     35 343.7 344.9 350.2 1.925e-01  0.14640 0.023690    0.9771
6    4     34 340.5 342.4 348.7 2.949e-01  0.23270 0.007215    0.9760
  Shapiro.p
1   0.15620
2   0.04687
3   0.02211
4   0.57640
5   0.61510
6   0.57580
```

Akaike Information Criterion (AIC)
Schwarz Bayesian Information Criterion (BIC)

- BIC is more harsh
- AICc is used for smaller sample size
- Smaller values indicate better models

Reduced Model (Model 4) seems to be a better fit in comparison to the *Full Model* (Model 5)

Partial F Test

$$F_{stat} = \frac{\left(SSE(Reduced.Model) - SSE(Full.Model)\right)/(Change\ in\ \#\ of\ Parameters)}{MSE(Full)}$$

If F_{stat} is large and significant, there is a large difference between the two models -> use full model

If F_{stat} is small or not significant, models do not differ greatly -> use reduced model

Partial F Test

$$F_{stat} = \frac{(SSE(Reduced.Model) - SSE(Full.Model))/(Change\ in\ \#\ of\ Parameters)}{MSE(Full)}$$

If F_{stat} is large and significant, there is a large difference between the two models -> use full model

If F_{stat} is small or not significant, models do not differ greatly -> use reduced model

Partial F Test

```
> anova(reduced.model1, model1)
Analysis of Variance Table

Model 1: PIQ ~ Brain + Height
Model 2: PIQ ~ Brain + Height + Weight
  Res.Df   RSS Df Sum of Sq  F Pr(>F)
1     35 13322
2     34 13322  1 0.0031633  0 0.9977
> 
```

RSS is identical, F=0, p>0.1

Partial F Test

$$F_{stat} = \frac{(SSE(Reduced.Model) - SSE(Full.Model))/(\text{Change in \# of Parameters})}{MSE(Full)}$$

If F_{stat} is large and significant, there is a large difference between the two models -> use full model

If F_{stat} is small or not significant, models do not differ greatly -> use reduced model

Partial F Test

```
> anova(reduced.model1, model1)
Analysis of Variance Table

Model 1: PIQ ~ Brain + Height
Model 2: PIQ ~ Brain + Height + Weight
  Res.Df   RSS Df Sum of Sq  F Pr(>F)
1     35 13322
2     34 13322  1 0.0031633  0 0.9977
>
```

RSS is identical, F=0, p>0.1

Reduced Model *without* Weight

```
> reduced.model1 <- lm(PIQ ~ Brain + Height)
> summary(reduced.model1)

Call:
lm(formula = PIQ ~ Brain + Height)

Residuals:
    Min      1Q  Median      3Q     Max 
-32.750 -12.090  -3.841  14.174  51.690 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 111.2757    55.8673   1.992 0.054243 .  
Brain         2.0606     0.5466   3.770 0.000604 ***
Height       -2.7299     0.9932  -2.749 0.009399 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.51 on 35 degrees of freedom
Multiple R-squared:  0.2949,	Adjusted R-squared:  0.2546 
F-statistic: 7.321 on 2 and 35 DF,  p-value: 0.002208
```

Partial F Test

$$F_{stat} = \frac{(SSE(Reduced.Model) - SSE(Full.Model))/(Change\ in\ \#\ of\ Parameters)}{MSE(Full)}$$

If F_{stat} is large and significant, there is a large difference between the two models -> use full model

If F_{stat} is small or not significant, models do not differ greatly -> use reduced model

Partial F Test

```
> anova(reduced.model1, model1)
Analysis of Variance Table

Model 1: PIQ ~ Brain + Height
Model 2: PIQ ~ Brain + Height + Weight
  Res.Df   RSS Df Sum of Sq  F Pr(>F)
1     35 13322
2     34 13322  1 0.0031633  0 0.9977
> |
```

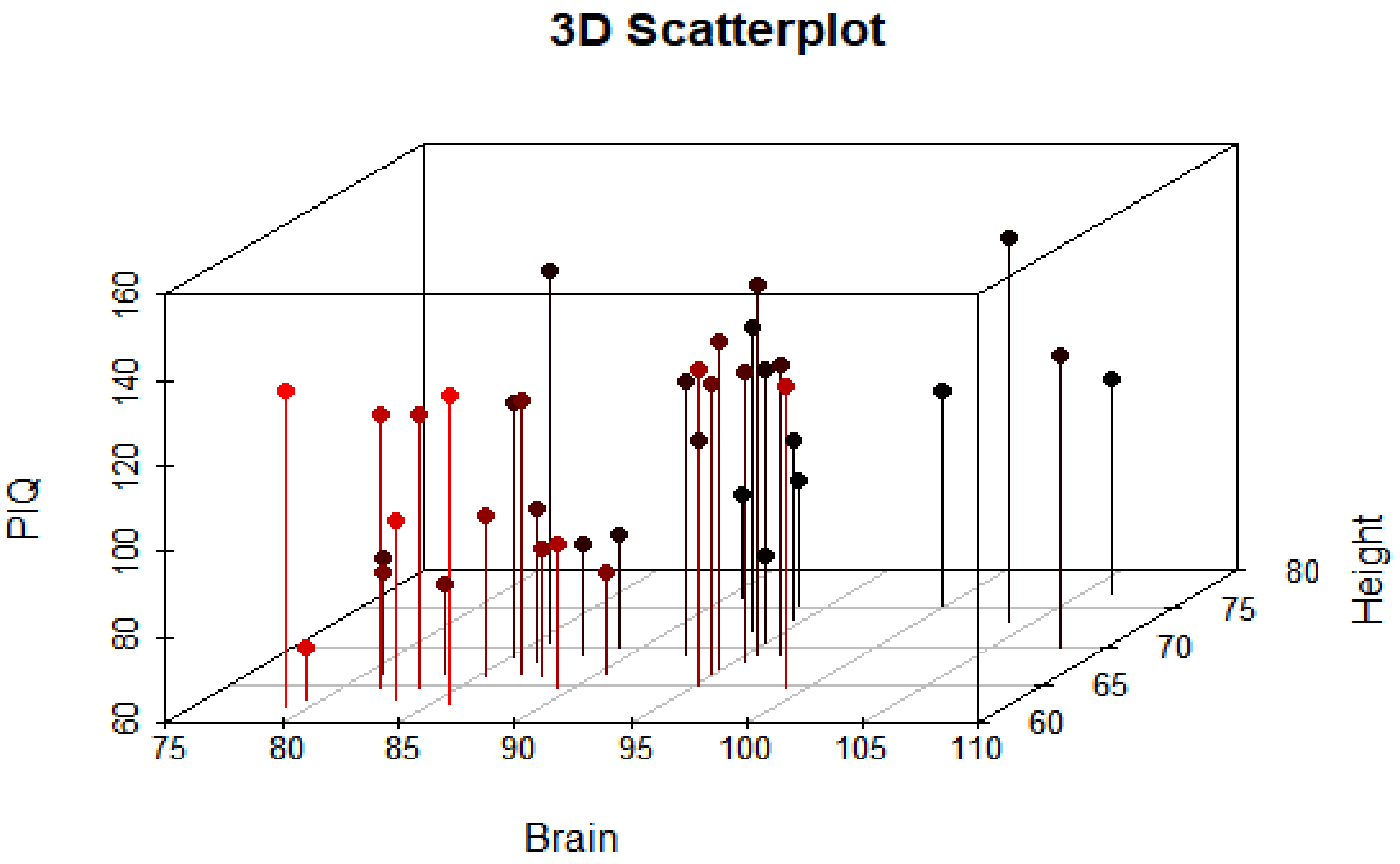
RSS is identical, F=0, p>0.1

$(PIQ)_I = 111.3 + 2.06$ (brain size) -2.73 (height inches)

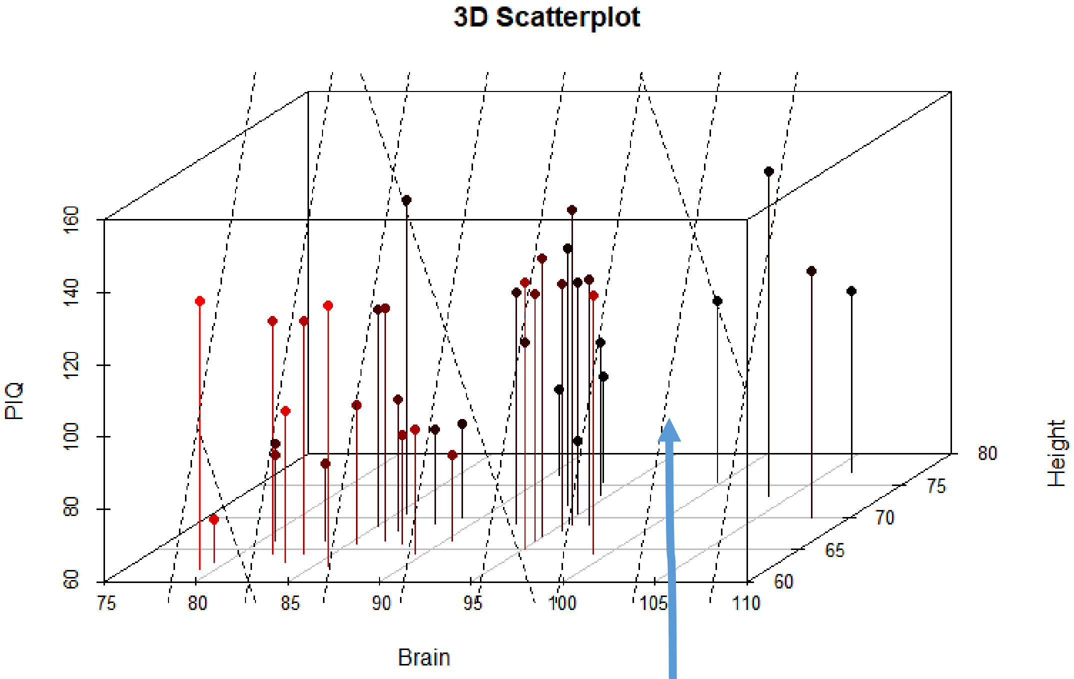
Reduced Model *without* Weight

```
> reduced.model1 <- lm(PIQ ~ Brain + Height)
> summary(reduced.model1)

Call:
lm(formula = PIQ ~ Brain + Height)

Residuals:
    Min      1Q  Median      3Q     Max
-32.750 -12.090  -3.841  14.174  51.690

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 111.2757    55.8673   1.992 0.054243 .
Brain         2.0606     0.5466   3.770 0.000604 ***
Height       -2.7299     0.9932  -2.749 0.009399 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.51 on 35 degrees of freedom
Multiple R-squared:  0.2949,    Adjusted R-squared:  0.2546
F-statistic: 7.321 on 2 and 35 DF,  p-value: 0.002208
```

Visualize Data with 2 Explanatory Variables



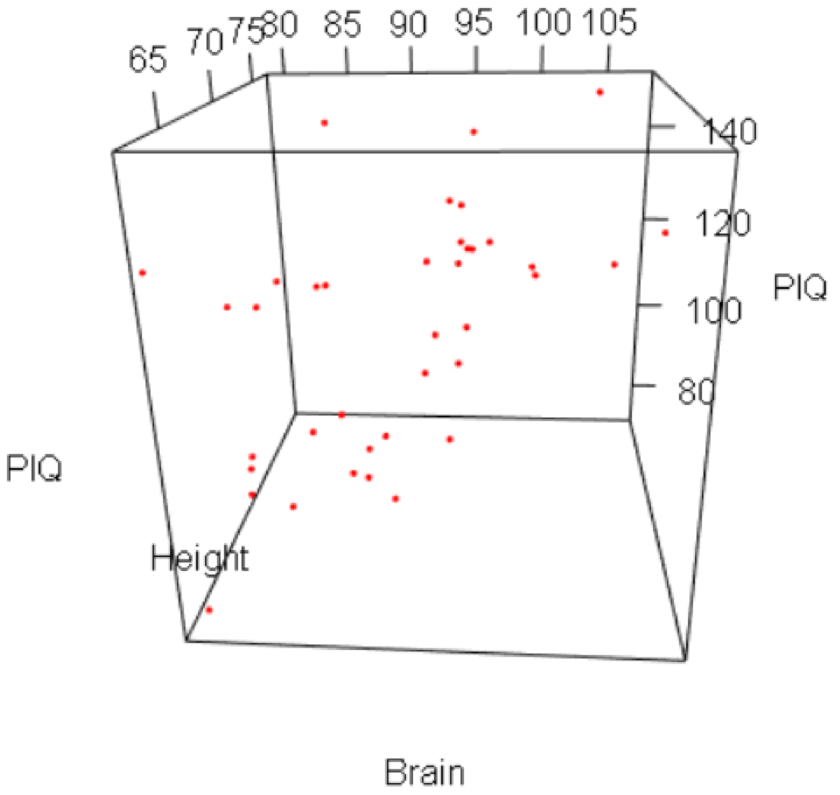

```
> scatterplot3d(Brain, Height, PIQ, main="3D Scatterplot")
> scatterplot3d(Brain, Height, PIQ,
+               pch = 16,
+               highlight.3d = TRUE,
+               type = "h",
+               main = "3D Scatterplot")
> fit <- lm(PIQ ~ Brain+Height)
> s3d$plane3d(fit)
> |
```

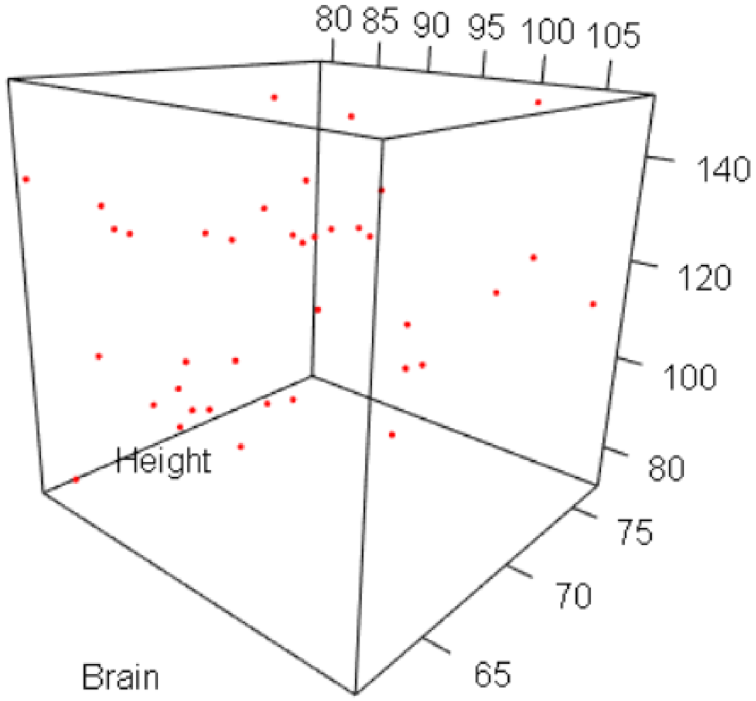


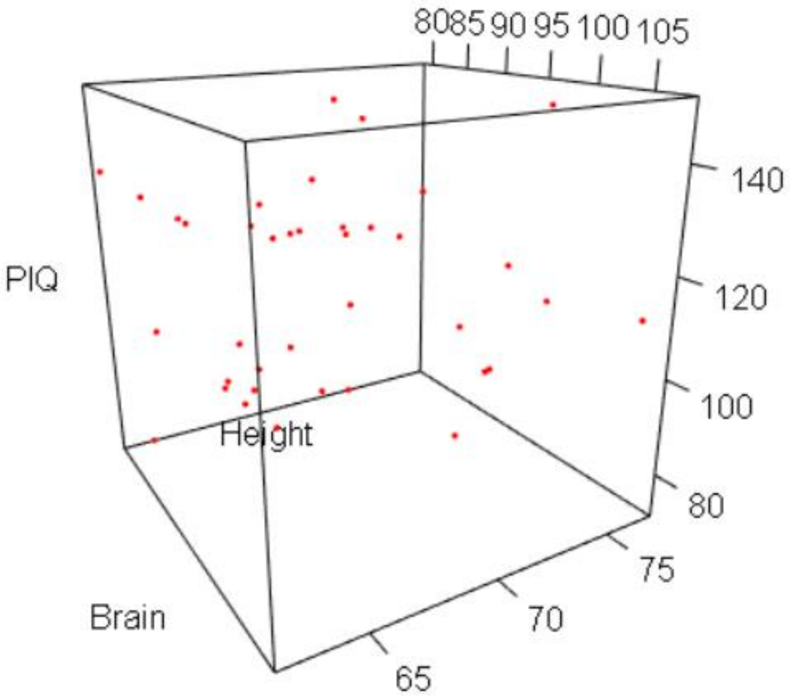
$(PIQ)_I$ = $\mathbf{111.3}$ + $\mathbf{2.06}$ (brain size) $\mathbf{-2.73}$ (height inches)

Visualize Data with 2 Explanatory Variables

```
> Library(rgl)
> plot3d(Brain, Height, PIQ, col="red", size=3)
```



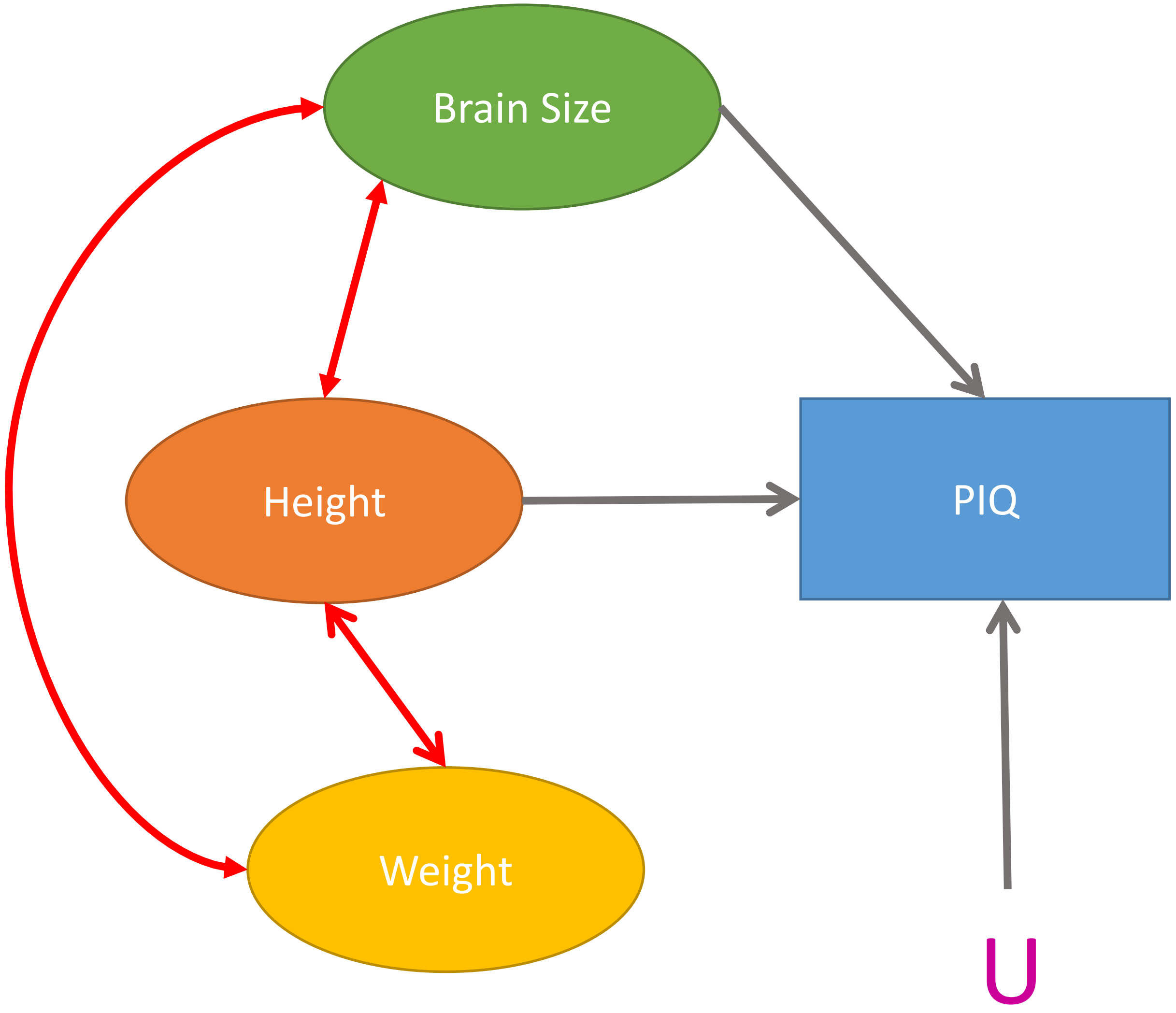




Path Analysis

Includes all correlations and all supposed causal links

Can account for unexplained causes that might affect the response variable, variables we have not yet measured (U)



Regression Model Analysis

Tests for linear association in a simple regression model

- Two primary methods:
  - t-test for the slope
    - Used to test whether a slope is positive or negative.
  - Analysis of Variance test (ANOVA) F-test
    - Useful for testing whether or not the slope = 0

t-test for slope

$$t^* = \frac{b_1 - \beta}{\left(\frac{\sqrt{MSE}}{\sqrt{\sum (x_i - \bar{x})^2}}\right)} = \frac{b_1 - \beta}{se(b_1)}$$

Null hypothesis $H_0 : \beta_1 = \text{some number } \beta$

Alternative hypothesis $H_A : \beta_1 \neq \text{some number } \beta$

- The resulting t-statistic obtained from the above formula is used to calculate the P-value. The P-value is determined by referring to a t-distribution with n-2 degrees of freedom.

ANOVA F-test

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

SSTO	*SSR*	*SSE*
Total sum of squares	Regression sum of squares	Error sum of squares

$$SSTO = SSR + SSE$$

ANOVA F-test

the null hypothesis H_0: $\beta_1 = 0$

against the alternative hypothesis H_A: $\beta_1 \neq 0$.

These values help test the null and alternative hypotheses:

$$MSR = \frac{\sum(\hat{y}_i - \bar{y})^2}{1} = \frac{SSR}{1}$$

$$MSE = \frac{\sum(y_i - \hat{y}_i)^2}{n-2} = \frac{SSE}{n-2}$$

$$F^* = \frac{MSR}{MSE}$$

Simple Linear Regression assumptions - LINE

- Linearity (L): The mean of the response of a sample population at each value of the predictor value Xi is a linear function of Xi
- Independence (I): The errors at each predictor value are independent
- Normally distributed (N): The errors at each predictor value are normally distributed
- Equal variance (E): The errors at each predictor value have equal variances

Assessing Linearity (L)

- Visual inspection
- Residuals vs Fit (estimated values) plot
  - This can also be a good check for equal variances and outliers
  - Residuals vs Predictor is a similar plot, but can help assess whether a new, additional predictor can make the model better

Residuals: $e_i = y_i - \hat{y}_i$

Assessing Linearity - Example: alcohol consumption vs muscle strength
(Marquez et al, 1989)

Source: https://onlinecourses.science.psu.edu/stat501/node/277

Assessing Linearity - What a non-linear plot looks like

Source: https://onlinecourses.science.psu.edu/stat501/node/279

Assessing Independence (I)

- Residuals vs Order plot
  - NB: This test can only be performed for data collected in an ordered or numbered fashion.
  - A scatter plot with the residuals on the y axis and order in which the data were collected on the x axis.

Assessing Independence - What to look for when error shows no independence

Positive serial correlation:

Negative serial correlation:

Source: https://onlinecourses.science.psu.edu/stat501/node/280

Assessing Normal Distribution

- Normal probability plot of residuals is used where a plot of the theoretical percentiles of the normal distribution vs the the observed sample percentiles is plotted.

  - This resulting plot should be linear.

Assessing Error Variance - what an unequal variance looks like on a residual vs fits plot

Example of a fanning scatter plot:

Source: https://onlinecourses.science.psu.edu/stat501/node/279

Data Transformation

- If the data presented does not adhere to the SLR model, a number of approaches can be considered:
  - Omitting predictor variables to improve the model.
  - If the mean of the response is not a linear function of the predictors, a different function can be used. Eg: Polynomial regression or Log transformation
  - If there are unequal variances, use the “weighted least squares regression” to transform response and/or predictor variables
  - If an outlier exists, use “robust estimation procedure”
  - If error terms are not independent, try a “time series model”.

Data Transformation: Transforming predictor values (X) only

- Transforming Predictor values is usually performed when nonlinearity is the ONLY problem; All other assumptions must hold true after transformation

Eg: Proportion of words recalled vs time:

Regression Model:

Residual vs Fit:

Source: https://onlinecourses.science.psu.edu/stat501/node/319

Data Transformation: Transforming predictor values (X) only

- Transforming Predictor values is usually performed when nonlinearity is the ONLY problem; All other assumptions must hold true after transformation

Taking the natural log of predictor value (time)

time	*prop*	*lntime*
1	0.84	0.00000
5	0.71	1.60944
15	0.61	2.70805
30	0.56	3.40120
60	0.54	4.09434
120	0.47	4.78749
240	0.45	5.48064
480	0.38	6.17379
720	0.36	6.57925
1440	0.26	7.27240
2880	0.20	7.96555
5760	0.16	8.65869
10080	0.08	9.21831

Data Transformation: Transforming response values (Y) only

- Transforming response values is usually performed when non-normality and/or unequal variances are the problem; All other assumptions must hold true after transformation

Eg: Gestation length vs birthweight:

Source: https://onlinecourses.science.psu.edu/stat501/node/320

Data Transformation: Transforming response values (Y) only

- Transforming response values is usually performed when non-normality and/or unequal variances are the problem; All other assumptions must hold true after transformation

Take the natural log of response value (gestation time):

Mammal	Birthwgt	Gestation	lnGest
Goat	2.75	155	5.04343
Sheep	4.00	175	5.16479
Deer	0.48	190	5.24702
Porcupine	1.50	210	5.34711
Bear	0.37	213	5.36129
Hippo	50.00	243	5.49306
Horse	30.00	340	5.82895
Camel	40.00	380	5.94017
Zebra	40.00	390	5.96615
Giraffe	98.00	457	6.12468
Elephant	113.00	670	6.50728

Source: https://onlinecourses.science.psu.edu/stat501/node/320

Data Transformation: Transforming both predictor and response values

- Transforming response values is usually performed when non-normality and/or unequal variances as well as non-linearity are the problem.

Eg: Tree volume vs diameter (Schumacher et al, 1935):

Regression model:

Residuals vs fit:

Source: https://onlinecourses.science.psu.edu/stat501/node/321

Eg: Tree volume vs diameter (Schumacher et al, 1935):

Source: https://onlinecourses.science.psu.edu/stat501/node/321

Eg: Tree volume vs diameter (Schumacher et al, 1935):

Transforming predictor values only:

Diameter	*Volume*	*lnDiam*
4.4	2.0	1.48160
4.6	2.2	1.52606
5.0	3.0	1.60944
5.1	4.3	1.62924
5.1	3.0	1.62924
5.2	2.9	1.64866
5.2	3.5	1.64866
5.5	3.4	1.70475
5.5	5.0	1.70475
5.6	7.2	1.72277
5.9	6.4	1.77495
5.9	5.6	1.77495
7.5	7.7	2.01490
7.6	10.3	2.02815

Eg: Tree volume vs diameter (Schumacher et al, 1935):

Transforming predictor values only:

Source: https://onlinecourses.science.psu.edu/stat501/node/321

Eg: Tree volume vs diameter (Schumacher et al, 1935):

Transforming both predictor and response values:

Diameter	*Volume*	*lnDiam*	*lnVol*
4.4	2.0	1.48160	0.69315
4.6	2.2	1.52606	0.78846
5.0	3.0	1.60944	1.09861
5.1	4.3	1.62924	1.45862
5.1	3.0	1.62924	1.09861
5.2	2.9	1.64866	1.06471
5.2	3.5	1.64866	1.25276
5.5	3.4	1.70475	1.22378
5.5	5.0	1.70475	1.60944
5.6	7.2	1.72277	1.97408
5.9	6.4	1.77495	1.85630
5.9	5.6	1.77495	1.72277
7.5	7.7	2.01490	2.04122
7.6	10.3	2.02815	2.33214

Source: https://onlinecourses.science.psu.edu/stat501/node/321

Eg: Tree volume vs diameter (Schumacher et al, 1935):

Transforming both predictor and response values:

Source: https://onlinecourses.science.psu.edu/stat501/node/321

Polynomial Regression

- The scatter plot of residuals vs predictor may suggest a non-linear relationship. Polynomial regression may be a more suitable model for the data.

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_h X^h + \epsilon$$

h = degree of the polynomial

Polynomial regression guidelines:

1. The fitted model is more reliable when the sample size is large
2. Do not extrapolate beyond the limit of the observed values
3. Be aware of statistical overflow when trying to incorporate higher degree terms
4. Use practical significance vs statistical significance

Polynomial Regression - Example

- How is the length of a bluegill fish related to its age? (*Cook and Weisberg, 1999)*

$$y_i = (\beta_0 + \beta_1 x_i + \beta_{11} x_i^2) + \epsilon_i$$

Source: https://onlinecourses.science.psu.edu/stat501/node/325