# STAT 408
# Applied Regression Analysis

Nan Miles Xi

Department of Mathematics and Statistics

Loyola University Chicago

Fall 2022

# Transformation

# Transformation

- In linear regression, the “linear” refers to parameter
- The predictors themselves do not have to be linear

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 \log X_2 + \beta_3 X_1 X_2 + \varepsilon$$

$$Y = \beta_0 + \beta_1 X_1^{\beta_2} + \varepsilon$$

- Therefore, we can transform both response Y and predictors and model is still linear
- Let’s first look at transforming response Y

# Transforming the Response

- Reasons we may consider transforming response Y in linear model
    1. Reduce the impact of outliers and increase the normality of error distribution
    2. Improve the model fit
    3. Some real questions require us to transform the response Y

- We already see log and square root transformation for 1 and 2
- Let’s see one example for 3

# Transforming the Response

- Production function in microeconomics shows

$$Y = AL^{\beta}K^{\alpha}$$

where $Y$ = total production, $L$ = labor input, $K$ = capital input, $A$ = productivity

$\alpha$ and $\beta$ are “elasticity” of labor and capital, which are our interests

- Log-transformation gives us a linear model to estimate $\alpha$ and $\beta$:

$$\log(Y) = \log(A) + \beta \log(L) + \alpha \log(K)$$

# Transforming the Response

- When we use log-transformation, the regression parameters have a particular interpretation

$$\begin{aligned}\hat{y} &= e^{\hat{\beta}_0} e^{\hat{\beta}_1 x_1} \dots e^{\hat{\beta}_p x_p} \\ \log \hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_p x_p\end{aligned}$$

- If $\hat{\beta}_1 = 0.01$, then one unit increase of $X_1$ will increase $\log(\hat{y})$ by 0.01

- Because $\log(1 + \hat{y} - 1)$ = 0.01 and log (1+x) ≈ x for small x, $\hat{y} - 1 \approx$ 0.01
  - $\hat{y}$ will increase from 1 to 1.01, that is, 1%

- After log-transformation, <u>small</u> $\hat{\beta}$ is the percentage increase of Y if X increases by one unit

# Segmented Regression

- Now let's focus on the transformation of predictors – segmented regression

- The saving dataset contains five variables for 50 countries
  - sr: saveing rate
  - pop15: percent population under age of 15
  - pop75: percent population over age of 75
  - dpi: per-capita disposable income in dollars
  - ddpi: percent growth rate of dpi

- The data is over the period 1960-1970

| | sr | pop15 | pop75 | dpi | ddpi |
|---|---|---|---|---|---|
| Australia | 11.43 | 29.35 | 2.87 | 2329.68 | 2.87 |
| Austria | 12.07 | 23.32 | 4.41 | 1507.99 | 3.93 |
| Belgium | 13.17 | 23.80 | 4.43 | 2108.47 | 3.82 |
| Bolivia | 5.75 | 41.89 | 1.67 | 189.13 | 0.22 |
| Brazil | 12.88 | 42.19 | 0.83 | 728.47 | 4.56 |
| Canada | 8.79 | 31.72 | 2.85 | 2982.88 | 2.43 |

# Segmented Regression

- The motivation of segmented regression is that different linear regression models may apply in different regions of the data

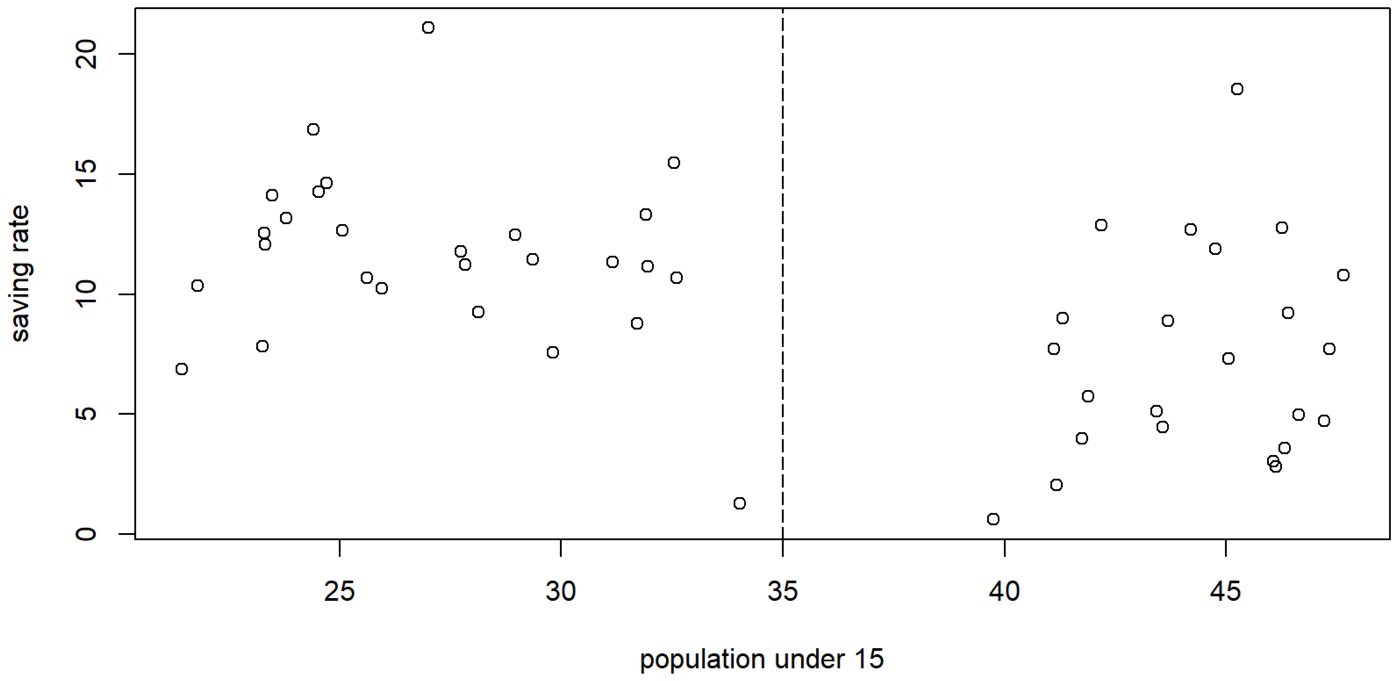
- In the saving dataset, we suspect the relations between saving rate and population age are different in younger countries and older countries

- We use pop15=35 as the cutoff

```
saving <- read.csv("saving.csv")
plot(sr ~ pop15, data = saving, xlab='population under 15', ylab='saving rate')
abline(v=35, lty=5)
```

# Segmented Regression

- We need to fit two separate models to capture the two relationships



# Segmented Regression

- Fit two models in two regions

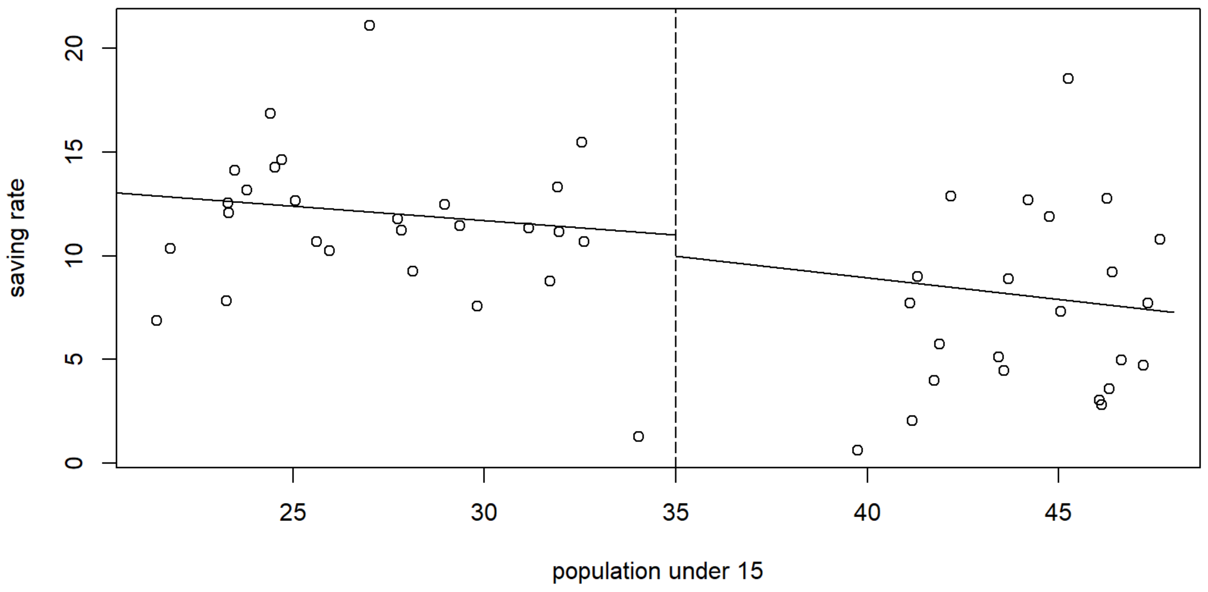
```
lm1 <- lm(sr~pop15, data = saving, subset = (pop15<35))
lm2 <- lm(sr~pop15, data = saving, subset = (pop15>35))
```

- Draw two models in two regions

```
segments(x0 = 20, y0 = lm1$coefficients[1]+lm1$coefficients[2]*20,
         x1 = 35, y1 = lm1$coefficients[1]+lm1$coefficients[2]*30)
segments(x0 = 35, y0 = lm1$coefficients[1]+lm1$coefficients[2]*35,
         x1 = 48, y1 = lm1$coefficients[1]+lm1$coefficients[2]*48)
```

# Segmented Regression

- One issue is that the two models are disconnected at x=35
- X=35 is a break point, which seems unrealistic
- Segmented regression solves this issue by smoothing the break points



# Segmented Regression

- We transform the predictor X into two predictors by two basis functions

$$B_l(x) = \begin{cases} c - x & \texttt{if } x < c \\ 0 & \texttt{otherwise} \end{cases}$$

$$B_r(x) = \begin{cases} x - c & \texttt{if } x > c \\ 0 & \texttt{otherwise} \end{cases}$$

where c is the cutoff between the two groups and c = 35 in this model

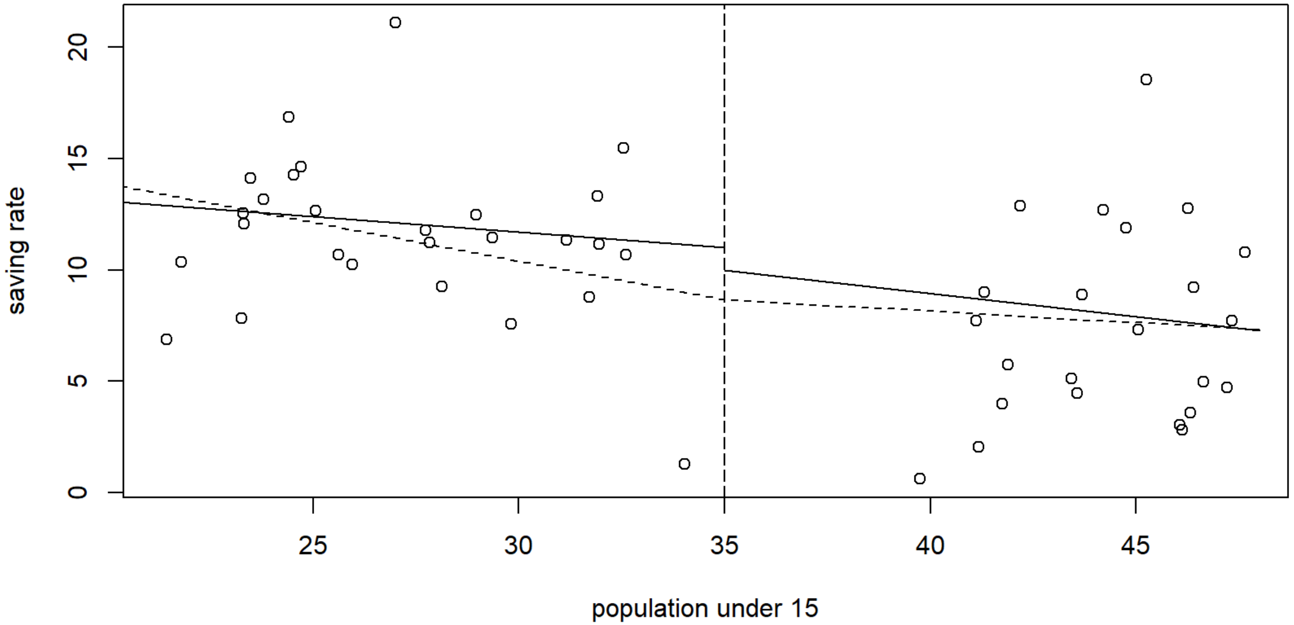
- Under this transformation, the linear model is

$$y = \beta_0 + \beta_1 B_l(x) + \beta_2 B_r(x) + \varepsilon$$

- We can use regular linear regression to fit this model (coding 6.r)

# Segmented Regression

- The segmented model will have different slopes in the two groups but connects at X = 35
- Segmented model also reduces the four parameters in the two linear models to three
- Since we change the X value in both groups, the two slopes are different from the two separate regressions



# Polynomial Regression

- Polynomial regression includes the higher order of predictors into a linear model

$$y = \beta_0 + \beta_1 x + \cdots + \beta_d x^d + \varepsilon$$

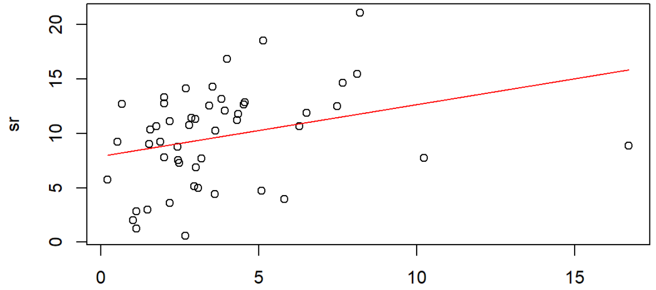
- It introduces non-linearity into the model and provides more flexibility and complexity

- There are three ways to determine the higher order $d$
  1. Forward: adding terms until the added term is not statistically significant
  2. Backward: starting with a large $d$ and eliminate non-statistically significant terms
  3. Choose $d$ based on prior knowledge

# Polynomial Regression

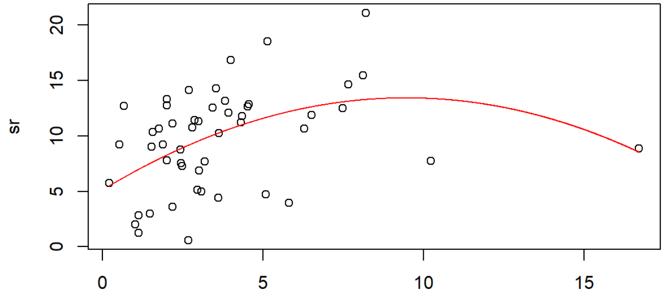
- We use <u>forward</u> method to examine the relation between saving and the increase of disposable income

```
> summary(lm(sr ~ ddpi,savings))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    7.883      1.011    7.80  4.5e-10
ddpi           0.476      0.215    2.22    0.031
```



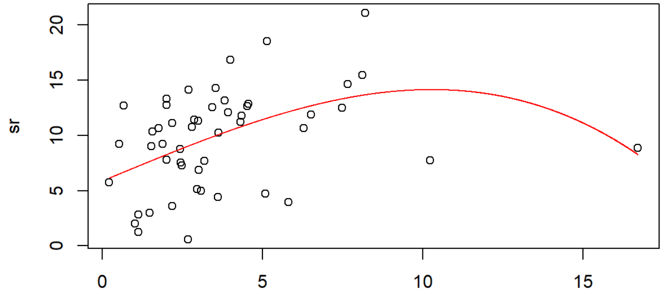

Linear

```
> summary(lm(sr ~ ddpi+I(ddpi^2),savings))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.1304     1.4347    3.58  0.00082
ddpi          1.7575     0.5377    3.27  0.00203
I(ddpi^2)    -0.0930     0.0361   -2.57  0.01326
```




Quadratic

```
> summary(lm(sr ~ ddpi+I(ddpi^2)+I(ddpi^3),savings))
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.145360   2.198606    2.34    0.024
ddpi         1.746017   1.380455    1.26    0.212
I(ddpi^2)   -0.090967   0.225598   -0.40    0.689
I(ddpi^3)   -0.000085   0.009374   -0.01    0.993
```




Cubic

# Polynomial Regression

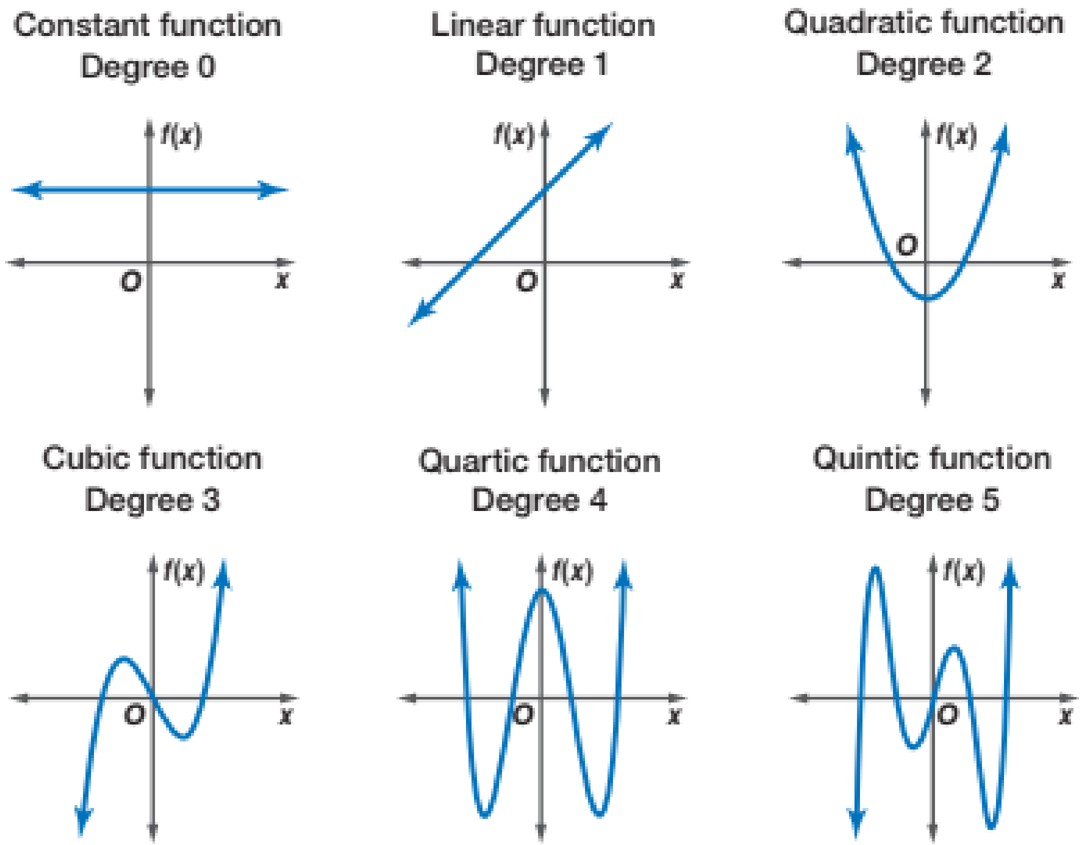
- Any term higher than the second order is insignificant, so the final model is quadratic of dppi
- It is a bad idea to eliminate lower order terms from the model before the higher order terms, even if they are not significant

- We can also define polynomials in more than one variable, also called response surface model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2$$

```
lmod <- lm(sr ~ pop15 + ddpi + I(pop15^2) + I(ddpi^2) + I(pop15*ddpi),savings)
summary(lmod)
```

# Nonlinearity in Linear Regression




The visualization for polynomial models with different orders

# Collinearity

# Collinearity

- Recall that we estimate linear model by

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

- If $X$ is singular (perfect linear relation of predictors), then $X^T X$ is not inversible, $\hat{\beta}$ does not have unique solution

- In this case, we need to drop certain predictors to break prefect linear relation

- This is called <u>exact collinearity</u>

# Collinearity

- A more challenging problem is $X$ close to singular but not exactly (collinearity)

- Recall $\hat{\beta}$ is a random variable with a normal distribution:

$$\hat{\beta} \sim N(\beta, (X^T X)^{-1} \sigma^2)$$

- Close to singular will cause $X^T X$ “small” and $(X^T X)^{-1}$ “large”, then the variance of $\hat{\beta}$ will be large
    1. The model estimation is unstable, small measurement errors leads to large changes in $\hat{\beta}$
    2. t statistic is small, t-test may fail to find significant predictors
    3. The signs of the coefficients could be the opposite of the truth

# Collinearity Detection

1. Examine the correlation matrix of the predictors
   - Matrix entry close to −1 or +1 indicates large pairwise collinearities

2. Regress predictor $X_i$ on all other predictors, then check R-square of this regression
   - large R-square (close to one) indicates collinearity

# Collinearity Detection

- Let's see one real-data example of dealing with collinearity

- Dataset seatpos contains 8 predictors of 38 driver's body size, weight, age and response variable hipcenter (seating position)

| | Age | Weight | HtShoes | Ht | Seated | Arm | Thigh | Leg | hipcenter |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 46 | 180 | 187.2 | 184.9 | 95.2 | 36.1 | 45.3 | 41.3 | -206.300 |
| 2 | 31 | 175 | 167.5 | 165.5 | 83.8 | 32.9 | 36.5 | 35.9 | -178.210 |
| 3 | 23 | 100 | 153.6 | 152.2 | 82.9 | 26.0 | 36.6 | 31.0 | -71.673 |
| 4 | 19 | 185 | 190.3 | 187.4 | 97.3 | 37.4 | 44.1 | 41.0 | -257.720 |
| 5 | 23 | 159 | 178.0 | 174.1 | 93.9 | 29.5 | 40.1 | 36.9 | -173.230 |
| 6 | 47 | 170 | 178.7 | 177.0 | 92.4 | 36.0 | 43.2 | 37.4 | -185.150 |
| 7 | 30 | 137 | 165.7 | 164.6 | 87.7 | 32.5 | 35.6 | 36.2 | -164.750 |
| 8 | 28 | 192 | 185.3 | 182.7 | 96.9 | 35.8 | 39.9 | 43.1 | -270.920 |

# Collinearity Detection

```
> lmod <- lm(hipcenter ~ ., seatpos)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 436.43213  166.57162   2.620   0.0138 *
Age           0.77572    0.57033   1.360   0.1843
Weight        0.02631    0.33097   0.080   0.9372
HtShoes      -2.69241    9.75304  -0.276   0.7845
Ht            0.60134   10.12987   0.059   0.9531
Seated        0.53375    3.76189   0.142   0.8882
Arm          -1.32807    3.90020  -0.341   0.7359
Thigh        -1.14312    2.66002  -0.430   0.6706
Leg          -6.43905    4.71386  -1.366   0.1824
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 37.72 on 29 degrees of freedom
Multiple R-squared:  0.6866,    Adjusted R-squared:  0.6001
F-statistic:  7.94 on 8 and 29 DF,  p-value: 1.306e-05
```

<u>Issue of this full model</u>

- No single predictor is significant (t-test), but they are jointly significant (F-test)
- Standard deviations of estimated parameters are large
- The R-square is reasonable
- Multiple predictors measure driver's body size
- Those are evidence of collinearity: highly correlated predictors

# Collinearity Detection

- Let’s check the predictor correlation matrix

```
> round(cor(seatpos[,-9]),2)
          Age Weight HtShoes    Ht Seated  Arm Thigh   Leg
Age      1.00   0.08   -0.08 -0.09  -0.17 0.36  0.09 -0.04
Weight   0.08   1.00    0.83  0.83   0.78 0.70  0.57  0.78
HtShoes -0.08   0.83    1.00  1.00   0.93 0.75  0.72  0.91
Ht      -0.09   0.83    1.00  1.00   0.93 0.75  0.73  0.91
Seated  -0.17   0.78    0.93  0.93   1.00 0.63  0.61  0.81
Arm      0.36   0.70    0.75  0.75   0.63 1.00  0.67  0.75
Thigh    0.09   0.57    0.72  0.73   0.61 0.67  1.00  0.65
Leg     -0.04   0.78    0.91  0.91   0.81 0.75  0.65  1.00
```

- There are some large pairwise correlations between predictors, mainly those predictors that measure height/length

# Collinearity Detection

- Let's regress each predictor on others and check their R squares

```r
x <- model.matrix(lmod)[,-1]

for(i in 1:8){
  r2 <- summary(lm(x[,i] ~ x[,-i]))$r.squared
  cat(colnames(x)[i], '\t', r2, '\n')
}
```

```
Age      0.4994823
Weight   0.7258043
HtShoes          0.9967472
Ht       0.9969982
Seated   0.8882813
Arm      0.7775983
Thigh    0.6380596
Leg      0.850619
```

- There are some large R-squares indicating collinearity

# Instable estimation

- We simulate a new dataset by adding random noise (std = 10) to response variable hipcenter

```
lm.model <- lm(hipcenter+rnorm(n=38,mean=0,sd=10)~., data = seatpos)
summary(lm.model)
```

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | 420.9929 | 172.4276 | 2.442 | 0.021 | * |
| Age | 0.7916 | 0.5904 | 1.341 | 0.190 | |
| Weight | -0.1128 | 0.3426 | -0.329 | 0.744 | |
| HtShoes | -6.2007 | 10.0959 | -0.614 | 0.544 | |
| Ht | 4.1962 | 10.4860 | 0.400 | 0.692 | |
| Seated | 0.3776 | 3.8941 | 0.097 | 0.923 | |
| Arm | -0.6793 | 4.0373 | -0.168 | 0.868 | |
| Thigh | -1.1333 | 2.7535 | -0.412 | 0.684 | |
| Leg | -5.8780 | 4.8796 | -1.205 | 0.238 | |

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | 436.43213 | 166.57162 | 2.620 | 0.0138 | * |
| Age | 0.77572 | 0.57033 | 1.360 | 0.1843 | |
| Weight | 0.02631 | 0.33097 | 0.080 | 0.9372 | |
| HtShoes | -2.69241 | 9.75304 | -0.276 | 0.7845 | |
| Ht | 0.60134 | 10.12987 | 0.059 | 0.9531 | |
| Seated | 0.53375 | 3.76189 | 0.142 | 0.8882 | |
| Arm | -1.32807 | 3.90020 | -0.341 | 0.7359 | |
| Thigh | -1.14312 | 2.66002 | -0.430 | 0.6706 | |
| Leg | -6.43905 | 4.71386 | -1.366 | 0.1824 | |

- Many “length-related” predictors have very different parameter estimations, some even change signs

# Mitigation of Collinearity

- Too many variables try do the same job of explaining the response and there is redundant information in predictors

- When we have a new dataset from the same population, the model “randomly” reassign importance to similar predictors and causes instable parameter estimation

- The high degree of instability inflates the variance of estimation and hides the significance

- The solution is simple: remove highly correlated predictors, leave remain only one of them

# Mitigation of Collinearity

- Let’s remove all “length-related” predictors except driver’s height

```
lm.model <- lm(hipcenter~Age+Weight+Ht, data = seatpos)
summary(lm.model)
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 528.297729 135.312947   3.904 0.000426 ***
Age           0.519504   0.408039   1.273 0.211593    
Weight        0.004271   0.311720   0.014 0.989149    
Ht           -4.211905   0.999056  -4.216 0.000174 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 36.49 on 34 degrees of freedom
Multiple R-squared:  0.6562,	Adjusted R-squared:  0.6258 
F-statistic: 21.63 on 3 and 34 DF,  p-value: 5.125e-08
```

- In the new model, the standard deviations of estimated parameters is much smaller
- The height predictor now is significant
- Adjusted $R^2$ is improved