Ve406 Lecture 8

Jing Liu

UM-SJTU Joint Institute

June 6, 2018

- In multiple linear regression, we explore the relation between the response

$$Y_i \qquad i = 1, 2, \ldots, n$$

and two or more predictors, k of them in general,

$$X_{i1}, \quad X_{i2}, \qquad \ldots \qquad X_{ij}, \qquad \ldots \qquad X_{ik} \qquad i = 1, 2, \ldots, n$$

where we are assuming that we observed on n cases, and

1. The conditional mean of the response is given by

$$\mathbb{E}\left[Y_i \mid X_{i1}, X_{i2}, \ldots, X_{ik}\right] = \mathbb{E}\left[Y_i \mid \mathbf{X}_i\right] = \beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + \cdots + \beta_k x_{ik}$$

2. The errors have zero mean and constant variance

$$\mathbb{E}\left[\varepsilon_i \mid \mathbf{X}_i\right] = 0 \quad \text{and} \quad \mathrm{Var}\left[\varepsilon_i \mid \mathbf{X}_i\right] = \sigma^2 \qquad \text{where} \quad \varepsilon_i = Y_i - \mathbb{E}\left[Y_i \mid \mathbf{X}_i\right]$$

3. The errors are independent of $\mathbf{X}_i$, and of each other.
4. The errors follow the normal distribution of $\mathrm{N}\left(0, \sigma^2\right)$.

- In matrix form, we have

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \text{ and } \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

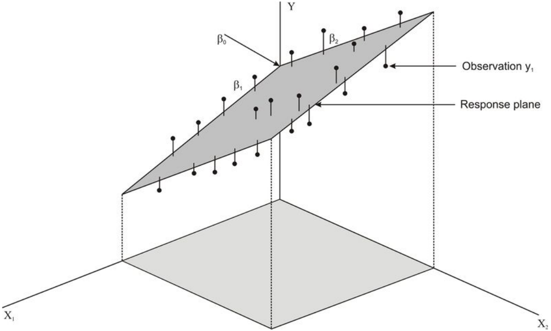
Q: How to interpret the regression coefficients β_0, β_1, …, β_n now?

- Suppose there are only two predictors

$$m(x_{i1}, x_{i2}) = \mathbb{E}\left[Y_i \mid X_{i1} = x_{i1}, X_{i2} = x_{i2}\right] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

$$\implies \frac{\partial m}{\partial x_{i1}} = \beta_1; \qquad \frac{\partial m}{\partial x_{i2}} = \beta_2; \qquad m(0,0) = \beta_0$$

- The regression coefficients in multiple regression are sometimes called partial regression coefficients to emphasise that their interpretation requires that the other variables should be held fixed.



- Notice the essential idea is the same, but we are fitting a plane/hyperplane instead of a line to the data by estimating the parameters in

$$\mathbb{E}\left[Y \mid \mathbf{X}\right]$$

- Most of what we have done can be extended with a bit of linear algebra

$$\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\mathbf{e}}$$

we choose $\hat{\boldsymbol{\beta}} = \mathbf{b}$ that minimises RSS, which is given by

$$\begin{aligned} f(b_0, b_1, \ldots, b_k) = \hat{\mathbf{e}}^{\mathrm{T}}\hat{\mathbf{e}} &= (\mathbf{y} - \mathbf{X}\mathbf{b})^{\mathrm{T}} (\mathbf{y} - \mathbf{X}\mathbf{b}) \\ &= \left(\mathbf{y}^{\mathrm{T}} - \mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\right)(\mathbf{y} - \mathbf{X}\mathbf{b}) \\ &= \mathbf{y}^{\mathrm{T}}\mathbf{y} - \mathbf{y}^{\mathrm{T}}\mathbf{X}\mathbf{b} - \mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} + \mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{b} \end{aligned}$$

- Notice the all terms are scalars, thus equal

$$\mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \left(\mathbf{y}^{\mathrm{T}}\mathbf{X}\mathbf{b}\right)^{\mathrm{T}} = \text{scalar} = \mathbf{y}^{\mathrm{T}}\mathbf{X}\mathbf{b}$$

- Hence the function we need to minimise is given by

$$f(\mathbf{b}) = \mathbf{y}^{\mathrm{T}}\mathbf{y} - 2\mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} + \mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{b}$$

- Recall to minimise a function,

$$f(\mathbf{b}) = \mathbf{y}^{\mathrm{T}}\mathbf{y} - 2\mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} + \mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{b}$$

  we set the gradient to zero,

$$\nabla f = 0 - 2\mathbf{X}^{\mathrm{T}}\mathbf{y} + 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{b}$$

  Setting this to zero, we have

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

- Hence, the fitted value is given by

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{P}\mathbf{y}$$

  and the residual can be found using

$$\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{P})\,\mathbf{y}$$

- With more linear algebra, we have

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\left(\mathbf{X}\boldsymbol{\beta} + \mathbf{e}\right) = \boldsymbol{\beta} + \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{e}$$

which means it is unbiased as expected,

$$\mathbb{E}\left[\hat{\boldsymbol{\beta}} \mid \mathbf{X}\right] = \boldsymbol{\beta} + \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbb{E}\left[\boldsymbol{\varepsilon} \mid \mathbf{X}\right] = \boldsymbol{\beta}$$

- The variance is given by

$$\begin{aligned}\mathrm{Var}\left[\hat{\boldsymbol{\beta}} \mid \mathbf{X}\right] &= \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathrm{Var}\left[\boldsymbol{\varepsilon} \mid \mathbf{X}\right]\left(\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\right)^{\mathrm{T}} \\ &= \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\sigma^2\mathbf{I}\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} = \sigma^2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\end{aligned}$$

- With the normal assumption, we see

$$\hat{\boldsymbol{\beta}} \sim \text{Normal}\left(\boldsymbol{\beta}, \sigma^2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\right)$$

- For the estimation σ^2, we only need adjust the scalar a bit

$$\hat{\sigma}^2 = \frac{1}{n-k-1}\hat{\mathbf{e}}^{\mathrm{T}}\hat{\mathbf{e}} \qquad \text{where} \quad \hat{\mathbf{e}} = (\mathbf{I}-\mathbf{P})\,\mathbf{y}$$

  so that it is unbiased as well as being consistent as before.

- It can be shown the residual

$$\hat{\mathbf{e}} = (\mathbf{I}-\mathbf{P})\,\mathbf{y} = (\mathbf{I}-\mathbf{P})\,(\mathbf{X}\boldsymbol{\beta}+\mathbf{e})$$

  is an unbiased and consistent estimator of the error $\mathbf{e}$, and the variance is

$$\begin{aligned}\operatorname{Var}[\hat{\mathbf{e}} \mid \mathbf{X}] &= \operatorname{Var}[(\mathbf{I}-\mathbf{P})\,(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}) \mid \mathbf{X}] \\ &= (\mathbf{I}-\mathbf{P})\operatorname{Var}[\boldsymbol{\varepsilon} \mid \mathbf{X}]\,(\mathbf{I}-\mathbf{P})^{\mathrm{T}} \\ &= (\mathbf{I}-\mathbf{P})\,\sigma^2\mathbf{I}\,(\mathbf{I}-\mathbf{P})^{\mathrm{T}} = \sigma^2\,(\mathbf{I}-\mathbf{P})\end{aligned}$$

- Thus with the normal assumption, we have

$$\hat{\mathbf{e}} \sim \text{Normal}\left(\mathbf{0}, \sigma^2\,(\mathbf{I}-\mathbf{P})\right)$$

An unusual study on Blood Pressure

- A random sample of Peruvian Indians born in the Andes mountains who had since migrated to lower altitudes was selected. The subjects were all male, age 21 years or over, and whose parents were born at high altitudes as well. Previous research suggested that migration of this kind might cause higher blood pressure at first, but over time, blood pressure would decrease.
- The variables measured were:

`age`	the age of the subject
`years`	the number of years since migration down from the mountains
`weight`	the weight of the subject (in kg)
`height`	the height of the subject (in mm)
`BP`	the subject's systolic blood pressure (in mm Hg)

- After loading the data

```
> bp.df = read.table("~/Desktop/peru.txt",
+                              header = TRUE)
```

we should check whether the data is loaded correctly, i.e. the correct type

```
> str(bp.df)
```

```
'data.frame':   39 obs. of  5 variables:
 $ age    : int  21 22 24 24 25 27 28 28 31 32 ...
 $ years  : int  1 6 5 1 1 19 5 25 6 13 ...
 $ weight: num  71 56.5 56 61 65 62 53 53 65 57 ...
 $ height: int  1629 1569 1561 1619 1566 1639 1494 1568 1540 1530 ...
 $ BP     : int  170 120 125 148 140 106 120 108 124 134 ...
```

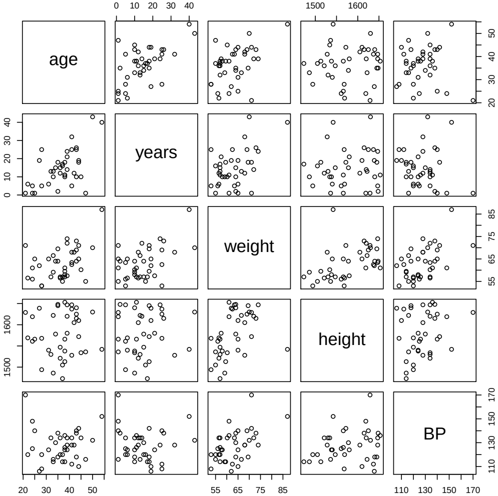
and whether there is any unusual value in the data

```
> summary(bp.df)
```

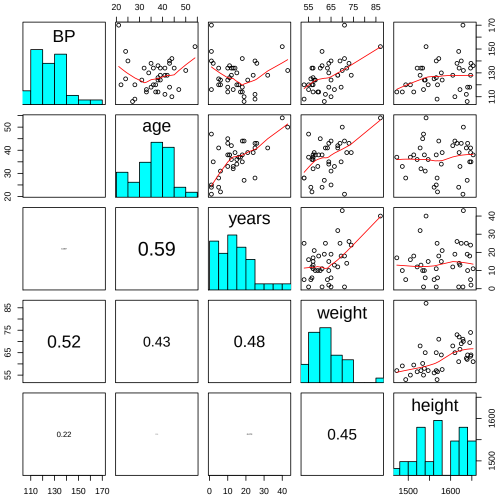
```
      age             years            weight           height          BP       
 Min.   :21.00   Min.   : 1.00   Min.   :53.00   Min.   :1473   Min.   :106.0  
 1st Qu.:32.50   1st Qu.: 8.00   1st Qu.:57.00   1st Qu.:1537   1st Qu.:118.0  
 Median :38.00   Median :13.00   Median :62.50   Median :1572   Median :126.0  
 Mean   :36.54   Mean   :14.74   Mean   :63.16   Mean   :1579   Mean   :127.4  
 3rd Qu.:41.50   3rd Qu.:19.00   3rd Qu.:68.00   3rd Qu.:1627   3rd Qu.:134.0  
 Max.   :54.00   Max.   :43.00   Max.   :87.00   Max.   :1653   Max.   :170.0  
```

- Pairs plot is also useful to identify pattens or unusual values in the data

```
> pairs(bp.df)
```



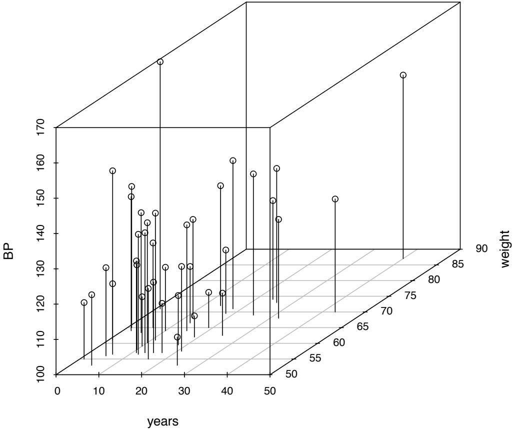
- We can include histograms, correlation coefficients, and trend lines as well



- Pairwise relationship between the response and the predictors
- Correlation coefficient is particularly relevant for a linear relation
- Watch out for really strong correlations between two predictors
- Watch out for unusual observations
- Watch out for small variabilities, i.e. categorical variables

```
> panel.hist = function(x, ...) {
+   usr = par("usr"); on.exit(par(usr))
+   par(usr = c(usr[1:2], 0, 1.5) )
+   h = hist(x, plot = FALSE)
+   breaks = h$breaks; nB = length(breaks)
+   y = h$counts; y = y/max(y)
+   rect(breaks[-nB],0,breaks[-1],y,col="cyan",...)
+ }
> panel.cor = function(x, y, digits = 2,
+                       prefix = "", cex.cor, ...)
+   usr = par("usr"); on.exit(par(usr))
+   par(usr = c(0, 1, 0, 1)); r = abs(cor(x, y))
+   txt = format(c(r,0.123456789),digits=digits)[1]
+   txt = paste0(prefix, txt)
+   if(missing(cex.cor)) cex.cor = 0.8/strwidth(txt)
+   text(0.5, 0.5, txt, cex = cex.cor * r)
+ }
> pairs(bp.df[, c(5,1:4)], upper.panel = panel.smooth,
+       lower.panel = panel.cor, diag.panel = panel.hist)
```

- 3D plot can be used to see unusual values that may not present in pairs plots



```
> library(scatterplot3d)
>
> vname = names(bp.df)
> k = length(vname) - 1
> m = combn(k, 2)
>
> pdf()
> for (j in 1:ncol(m)){
+   scatterplot3d(bp.df[, m[1, j]],
+                 bp.df[, m[2, j]],
+                 bp.df$BP, type="h",
+                 xlab = vname[m[1, j]],
+                 ylab = vname[m[2, j]],
+                 zlab = "BP")
+ }
> dev.off()
```

- Running the model with all the predictors,

```
> bp.LM = lm(BP~., data = bp.df)
```

- Checking assumptions by looking at residuals.
- All of the plots we have for simple linear regression remain valuable.

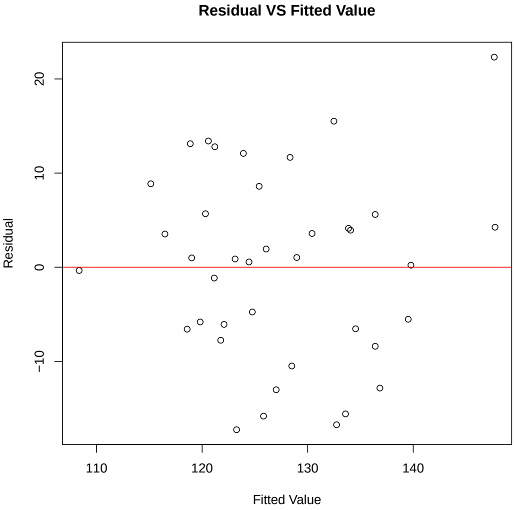
1. Residual Vs Fitted Value
2. Residual Vs Previous Residual
3. QQ normal

- We also include the following

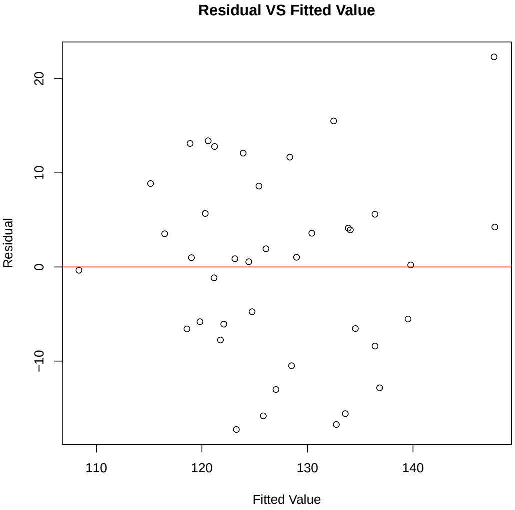
4. Residual Vs predictor

```
# Residual Vs Fitted Value
> res = bp.LM$residuals
> fit = bp.LM$fitted.values
```

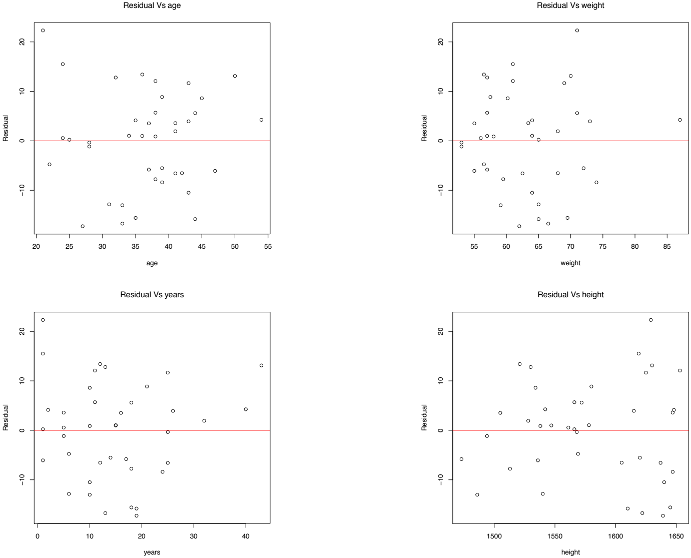
- There is no indication of nonlinearity, correlation or non-constant variance.



- There is no indication of nonlinearity, correlation or non-constant variance.

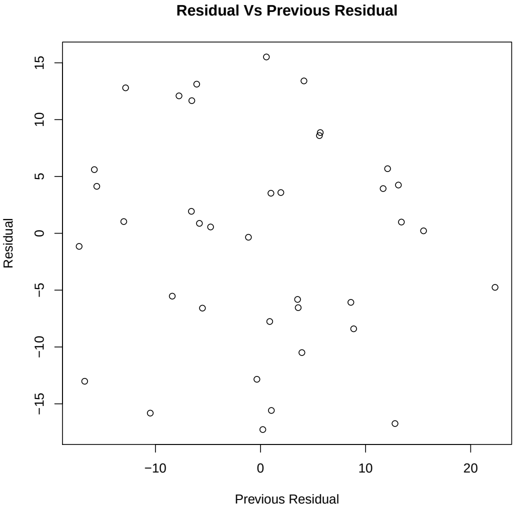


- There is no indication of nonlinearity, correlation or non-constant variance.

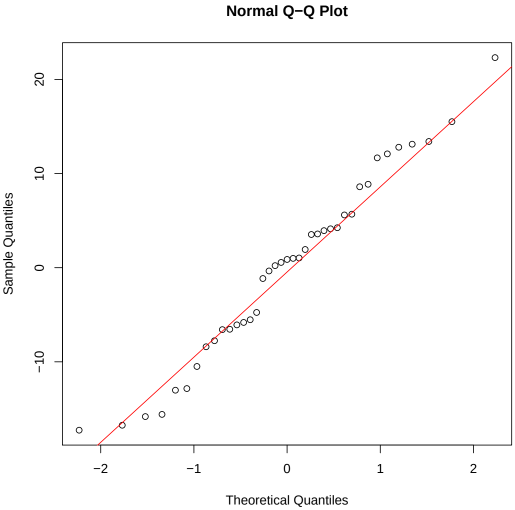


```
> plot(fit, res,
+       xlab = "Fitted Value", ylab = "Residual",
+       main = "Residual VS Fitted Value")
> abline(h = 0, col = "red")

> # Residual Vs Predictor
> pdf()
> for (i in 1:k){
+ plot(bp.df[[vname[i]]], res,
+       xlab = vname[i], ylab = "Residual",
+       main = bquote("Residual Vs"~.(vname[i])))
+ abline(h = 0, col = "red")
+ }
> dev.off()

> plot(res[-nrow(bp.df)], res[-1],
+       main = "Residual Vs Previous Residual",
+       ylab = "Residual", xlab = "Previous Residual")
```

- There is no indication that the residuals are correlated.



- There is no clear indication that normality is violated.



```
> qqnorm(res); qqline(res, col = "red")
> shapiro.test(res)
```

```
        Shapiro-Wilk normality test

data:  res
W = 0.97609, p-value = 0.5633
```

- Given there is no indication of any violation of the assumptions, we proceed

```
> summary(bp.LM)
```

```
Call:
lm(formula = BP ~ ., data = bp.df)

Residuals:
    Min      1Q  Median      3Q     Max 
-17.263  -6.561   0.875   5.644  22.320 

Coefficients:
            Estimate Std. Error t value Pr(>t)    
(Intercept) 90.17743   52.47507   1.718   0.0948 .  
age         -0.16133    0.27948  -0.577   0.5676    
years       -0.53803    0.21951  -2.451   0.0195 *  
weight       1.49865    0.31726   4.724 3.91e-05 ***
height      -0.02761    0.03674  -0.752   0.4575    
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 10.42 on 34 degrees of freedom
Multiple R-squared:  0.4344,    Adjusted R-squared:  0.3679 
F-statistic: 6.529 on 4 and 34 DF,  p-value: 0.0005195
```

- Notice the R^2 and the adjusted R^2 are relatively low,

```
Multiple R-squared:  0.4344,
Adjusted R-squared:  0.3679
```

so the model does not explain the variability in the response very well.

Q: What is the difference between R^2 and the adjusted R^2?

```
> n = nrow(bp.df); k = ncol(bp.df) - 1
>
> y = bp.df$BP; ybar = 1/nrow(bp.df) * sum(y)
>
> (R2 = sum((fit - ybar)^2) / sum((y - ybar)^2))
```

```
[1] 0.4344447
```

```
> (adjR2 = 1 - (1 - R2) * (n - 1) / (n - k - 1))
```

```
[1] 0.3679088
```

- R^2 is a measure of goodness of fit when all the assumptions are satisfied.
- However, large R^2 does not mean:

1. the assumptions are satisfied.
2. a more precise prediction for a single set of predictors.
3. a better model across different datasets.
4. a causal relationship between the response and the predictors.

- The adjusted R^2 is a relative measure, thus cannot be interpreted along.

```
> summary(lm(BP~., data = bp.df))$adj.r.squared
```

```
[1] 0.3679088
```

```
> summary(lm(BP~.-age, data = bp.df))$adj.r.squared
```

```
[1] 0.3799508
```

Q: Any one remember how F-test works?

```
> summary(bp.LM)
```

```
Call:
lm(formula = BP ~ ., data = bp.df)

Residuals:
    Min      1Q  Median      3Q     Max
-17.263  -6.561   0.875   5.644  22.320

Coefficients:
            Estimate Std. Error t value Pr(>t)
(Intercept) 90.17743   52.47507   1.718   0.0948 .
age         -0.16133    0.27948  -0.577   0.5676
years       -0.53803    0.21951  -2.451   0.0195 *
weight       1.49865    0.31726   4.724 3.91e-05 ***
height      -0.02761    0.03674  -0.752   0.4575
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 10.42 on 34 degrees of freedom
Multiple R-squared:  0.4344,    Adjusted R-squared:  0.3679
F-statistic: 6.529 on 4 and 34 DF,  p-value: 0.0005195
```

- The underlying assumption of F-statistics is that all slopes are zero

$$\beta_1 = \beta_2 = \cdots = \beta_k = 0$$

- Under this assumption and model assumptions, F-statistics follows

$$F(d_1 = k, d_2 = n - k - 1)$$

where k is the number of predictors, and n is the number of observations.

```
> bp.sm$fstatistic
```

```
   value      numdf      dendf
6.529476   4.000000  34.000000
```

```
> 1 - pf(bp.sm$fstatistic[1], bp.sm$fstatistic[2],
+         bp.sm$fstatistic[3])
```

```
       value
0.0005195276
```

- The p-value is very small, thus we have strong evidence from our data, at least one explanatory variable explains the variability in the response.

```
> null.LM = lm(BP~ 1, data = bp.df)
>
> (d = null.LM$df - bp.LM$df) # Essentiall k
```

```
[1] 4
```

```
> RSS.null = sum(null.LM$residuals^2)
>
> RSS.full = sum(bp.LM$residuals^2)
>
> RMS.full = RSS.full / bp.LM$df
>
> (f0 = (RSS.null - RSS.full) / (d * RMS.full))
```

```
[1] 6.529476
```

```
> bp.sm$fstatistic
```

```
    value     numdf     dendf
 6.529476  4.000000 34.000000
```