BWH - Biostatistics

Intermediate Biostatistics for Medical Researchers

Robert Goldman
Professor of Statistics
Simmons College

Topics in Multiple Regression

Tuesday, March 28, 2017

Topics

1. Multicollinearity

2. Qualitative Explanatory Variables

3. Interaction in Multiple Regression

4. Obtaining the ‘Best’ Regression Model

5. Checking the Conditions for Inference in Regression

1. Multicollinearity

Multicollinearity refers to the problems that occur in multiple regression when the explanatory (X) variables are themselves highly correlated.

Two (related) problems occur when two highly correlated explanatory variables (X_1 and X_2) are included in the model:

- The regression coefficients (slopes) for X_1 and X_2 change substantially from the corresponding coefficients when each is in the model alone. [Occasionally, the coefficients change sign.]

- The standard errors of the slopes are greatly inflated. This makes tests about the β's less sensitive and confidence intervals for the β's wider.

```
round(cor(d),3)
          Age    BMI Height Smoke
Age     1.000  0.802 -0.015 0.245
BMI     0.802  1.000 -0.095 0.287
Height -0.015 -0.095  1.000 0.179
Smoke   0.245  0.287  0.179 1.000
```

Age alone

```
The regression equation is
SBP = 59.1 + 1.60 Age

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  59.0916    12.8163   4.611 6.98e-05
Age           1.6045     0.2387   6.721 1.89e-07

Multiple R-squared:  0.6009
```

Age and BMI

```
The regression equation is
SBP = 41.8 + 1.05 Age + 1.82 BMI

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  41.7736    15.6842   2.663   0.0125
Age           1.0496     0.3855   2.723   0.0108
BMI           1.8221     1.0150   1.795   0.0831

Multiple R-squared:  0.6409
```

```
round(cor(d),3)
          Age    BMI Height Smoke
Age     1.000  0.802 -0.015 0.245
BMI     0.802  1.000 -0.095 0.287
Height -0.015 -0.095  1.000 0.179
Smoke   0.245  0.287  0.179 1.000
```

Age alone

The regression equation is
SBP = 59.1 + 1.60 Age

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  59.0916    12.8163   4.611 6.98e-05
Age           1.6045     0.2387   6.721 1.89e-07


Multiple R-squared:  0.6009
```

Age and Smoke

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  62.3892    11.8647   5.258 1.24e-05
Age           1.4639     0.2265   6.462 4.52e-07
Smoke         7.8818     3.1082   2.536   0.0169

Multiple R-squared:  0.6734
```

With two X variables in a multiple regression model, we acknowledge collinearity as a problem when:

$r^2_{X1, X2} > 0.8$

$1 - r^2_{X1, X2} = \text{Tolerance} < 0.2$

$$\frac{1}{1 - r^2_{X1, X2}} = \text{Variance Inflation Factor (VIF)} > 5$$

For Age and BMI:

$r^2_{X1, X2} = 0.802^2 = 0.6432$

$1 - r^2_{X1, X2} = \text{Tolerance} = 1 - 0.643 = 0.3586$

$$\frac{1}{1 - r^2_{X1, X2}} = \text{VIF} = \frac{1}{0.3586} = 2.802$$

The VIF is the factor by which the variance of a regression coefficient increases when the explanatory variables are linearly correlated (collinear) relative to the situation in which they were independent.

There is a VIF for each explanatory variable.

The basic R package will not compute VIF values; You need to install the *car* package. Once this is done:

```
model <- lm(SBP ~ Age + BMI, hd)
library(car)
vif(model)
     Age      BMI
2.801016 2.801016
```

If you wanted to compute the VIF for Age from the formula:

```
X <- lm(Age ~ BMI, hd)
summary(X)

Multiple R-squared:  0.643

vif<- 1/(1-0.643)
vif
[1] 2.80112
```

Since $r^2_{X1, X2} = r^2_{X2, X1}$, when there are only two X variables, they have the same VIF values.

With three or more X variables, the VIF values will be different.

```
model <- lm(SBP ~ Age + BMI + Smoke, hd)
vif(model)
     Age      BMI    Smoke
2.802773 2.871977 1.090623
```

$$VIF_{Age} = \frac{1}{1 - r^2_{Age}} = 2.803$$

where r^2_{Age} is the coefficient of determination when Age is regressed on BMI and Smoke.

$$VIF_{BMI} = \frac{1}{1 - r^2_{BMI}} = 2.872$$

where r^2_{BMI} is the coefficient of determination when BMI is regressed on Age and Smoke.

$$VIF_{Smoke} = \frac{1}{1 - r^2_{Smoke}} = 1.091$$

where r^2_{Smoke} is the coefficient of determination when Smoke is regressed on Age and BMI.

The extreme case of multicollinearity occurs where one of the X variables can be expressed as a linear function of other X variables, for example,

$X_1 = aX_2 + bX_3$

In this case the method of least squares will not produce a unique set of regression coefficients and so software will throw out one of the variables.

Options for Dealing with Multicollinearity

1. Drop one or more of the variables.

2. Combine two or more of the variables by substituting the mean or the sum, for example.

3. Use a method other than least squares (i.e. Ridge Regression).

2. Including Qualitative Explanatory Variables

(a) **A Dichotomous Explanatory Variable**

Predicting SBP from Age and Smoke

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  62.3892    11.8647   5.258 1.24e-05
Age           1.4639     0.2265   6.462 4.52e-07
Smoke         7.8818     3.1082   2.536   0.0169
```

$$\widehat{SBP} = 62.389 + 1.464Age + 7.882Smoke$$

How should we interpret 1.464 and 7.882?

After adjusting for smoking status, for every additional year of age, predicted SBP increases by 1.46 mm.

After adjusting for age, the predicted SBP for smokers (1) is 7.9 mm higher than that for non-smokers (0).

You can gain considerable visual insight into the meaning of these ‘slopes’ by breaking down the least-squares line into two separate lines.

(i) For smokers, Smoke = 1 and the model becomes

$$\widehat{\text{SBP}} = 63.389 + 1.464\text{Age} + 7.882(1)$$

$$= 71.271 + 1.464\text{Age}$$

(ii) For non-smokers, Smoke = 0 and the model becomes

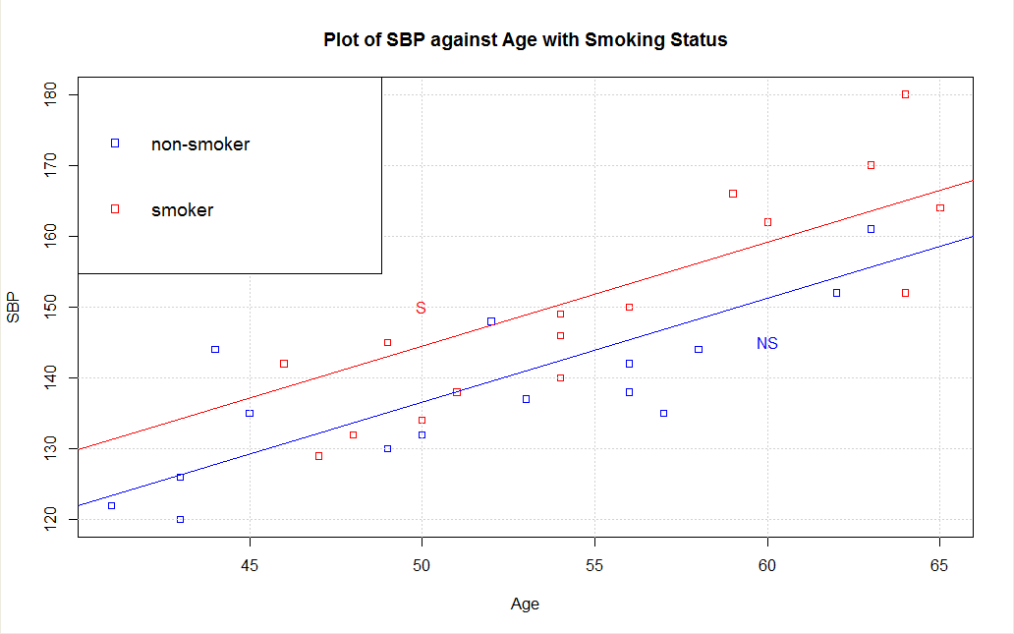
$$\widehat{\text{SBP}} = 63.389 + 1.464\text{Age} + 7.882(0)$$

$$= 63.389 + 1.464\text{Age}$$

$\widehat{SBP} = 62.389 + 1.464Age + 7.882Smoke$

For smokers $\widehat{SBP} = 71.271 + 1.464Age$

For non-smokers $\widehat{SBP} = 63.389 + 1.464Age$



```
hd$smk <- as.factor(hd$Smoke)

plot(SBP ~ Age, hd, col =
      c("blue", "red")[hd$smk], pch = 22,
      main = "Plot of SBP against Age
      with Smoking Status")

grid()

legend("topleft", levels(hd$smk),
      col = c("blue", "red"),
      pch = 22, cex = 1.1)

abline(71.27,1.464, col = "red")
abline(63.389,1.464, col = "blue")

text(50, 150, "S", col = "red")
text(60, 145, "NS", col = "blue")
```

(b) Adding a Qualitative Predictor with three or more Categories to a Regression Model

Suppose the categorical variable has three or more categories? Here is the general rule for representing such categorical variables as predictors in a regression model.

Suppose the variable has K categories:

(i) Create an indicator variable for each of the categories. This is a variable that takes the value 1 if that category is present and zero otherwise.

(ii) The categorical variable can be represented in the regression model by any K – 1 of these indicator variables.

(iii) The category omitted from the regression is called the **reference** group/category.

```r
Black <- rep(0, 32) # Initialize Black
Black[hd$Race == "Black"]<- 1

Hispanic <- rep(0, 32)
Hispanic[hd$Race =="Hispanic"] <- 1

White <- rep(0, 32)
White[hd$Race == "White"] <- 1

hd <- data.frame(hd, Black, Hispanic, White)

hd
```

	SBP	Age	BMI	Height	Smoke	Race	Black	Hispanic	White
1	135	45	22.8	70	0	Black	1	0	0
2	122	41	24.7	67	0	White	0	0	1
3	130	49	23.9	69	0	Black	1	0	0
4	148	52	27.4	70	0	White	0	0	1
5	146	54	23.3	71	1	White	0	0	1
6	129	47	22.3	76	1	Black	1	0	0
7	162	60	26.9	79	1	White	0	0	1
8	160	48	26.6	67	1	White	0	0	1
9	144	44	20.0	75	0	White	0	0	1
10	180	64	32.1	74	1	Hispanic	0	1	0
11	166	59	28.0	70	1	White	0	0	1
12	138	51	28.9	73	1	White	0	0	1
13	152	64	29.3	64	1	White	0	0	1
14	138	56	26.9	71	0	White	0	0	1
15	140	54	26.4	72	1	White	0	0	1
16	134	50	23.3	67	1	Hispanic	0	1	0
17	145	49	25.3	74	1	Hispanic	0	1	0
18	142	46	23.5	72	1	Hispanic	0	1	0
19	135	57	24.2	75	0	Black	1	0	0
20	142	56	25.5	70	0	White	0	0	1
21	150	56	26.8	65	1	White	0	0	1
22	144	58	27.4	66	0	White	0	0	1
23	137	53	25.0	70	0	Black	1	0	0
24	132	50	24.5	70	0	White	0	0	1
25	149	54	24.9	73	1	White	0	0	1
26	132	48	23.5	74	1	White	0	0	1
27	120	43	22.3	68	0	Black	1	0	0
28	126	43	23.1	74	0	White	0	0	1
29	161	63	27.7	70	0	White	0	0	1
30	170	63	29.4	81	1	Black	1	0	0
31	152	62	28.5	69	0	White	0	0	1
32	164	65	28.7	66	1	Hispanic	0	1	0

Predicting SBP from Age and Race

We must represent the variable Race by any two of the indicator variables, Black, Hispanic, and White. But, which two?

Variable	Race	N	Mean	StDev
SBP	Black	7	136.57	15.80
	Hispanic	5	153.00	18.68
	White	20	145.20	11.97

I will omit the variable Black. The other two variables White and Hispanic are sufficient to represent the three classes as you can see from the table below

	White	Hispanic	
White	1	0	
Hispanic	0	1	
Black	0	0	Reference Class

```
model4 <- lm(SBP ~ Age + Hispanic + White,
hd)
summary(model4)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  58.7915    12.4604   4.718 5.99e-05
Age           1.5251     0.2351   6.486 5.00e-07
Hispanic     10.6332     5.3173   2.000   0.0553
White         4.5871     3.9804   1.152   0.2589
```

$$\widehat{SBP} = 58.79 + 1.5251Age + 10.633Hispanic$$

$$+ 4.587White$$

After adjusting for age an Hispanic patient can be expected to have a SBP 10.63 mm *greater than a Black patient.*

After adjusting for age a White patient can be expected to have a SBP 4.587 mm *greater than a Black patient.*

Notice that the slopes/coefficients associated with Hispanic and White are interpreted relative to the reference race, Black.

For Hispanic patients:

$$\widehat{SBP} = 58.79 + 1.5251Age + 10.633(1)$$

$$+ 4.587(0)$$

$$\widehat{SBP} = 69.423 + 1.5251Age$$

For White patients:

$$\widehat{SBP} = 58.79 + 1.5251Age + 10.633(0)$$

$$+ 4.587(1)$$

$$\widehat{SBP} = 63.377 + 1.5251Age$$

For Black patients:

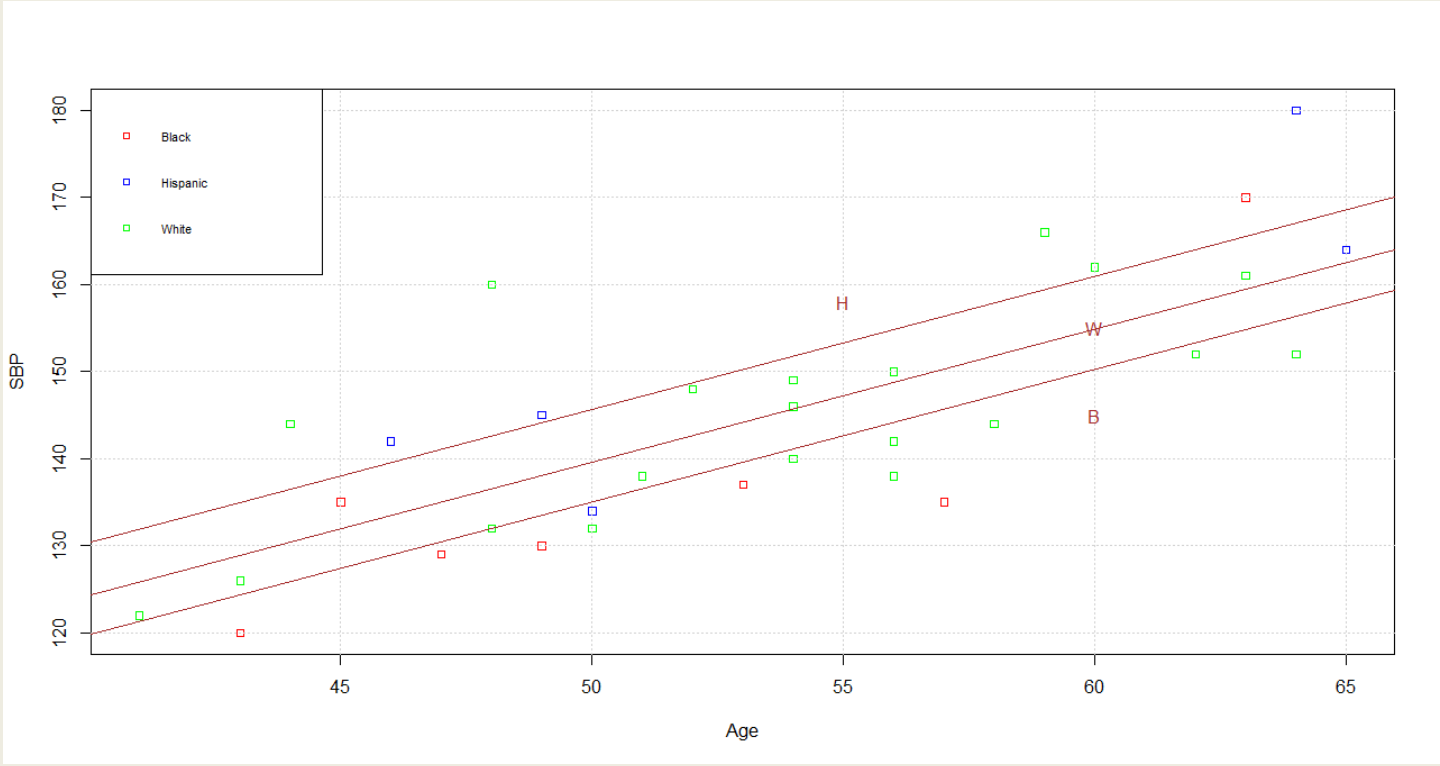
$$\widehat{SBP} = 58.79 + 1.5251Age + 10.633(0)$$

$$+ 4.587(0)$$

$$\widehat{SBP} = 58.79 + 1.5251Age$$

$$\widehat{SBP} = 58.79 + 1.5251Age + 10.633Hispanic$$

$$+ 4.587White$$



In fact, R will include a categorical predictor without explicitly having to create the indicator variables.

```
model5 <- lm(SBP ~ Age + Race, hd)

summary(model5)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   58.7915    12.4604   4.718 5.99e-05
Age            1.5251     0.2351   6.486 5.00e-07
RaceHispanic  10.6332     5.3173   2.000   0.0553
RaceWhite      4.5871     3.9804   1.152   0.2589
```

```
anova(model5)
```

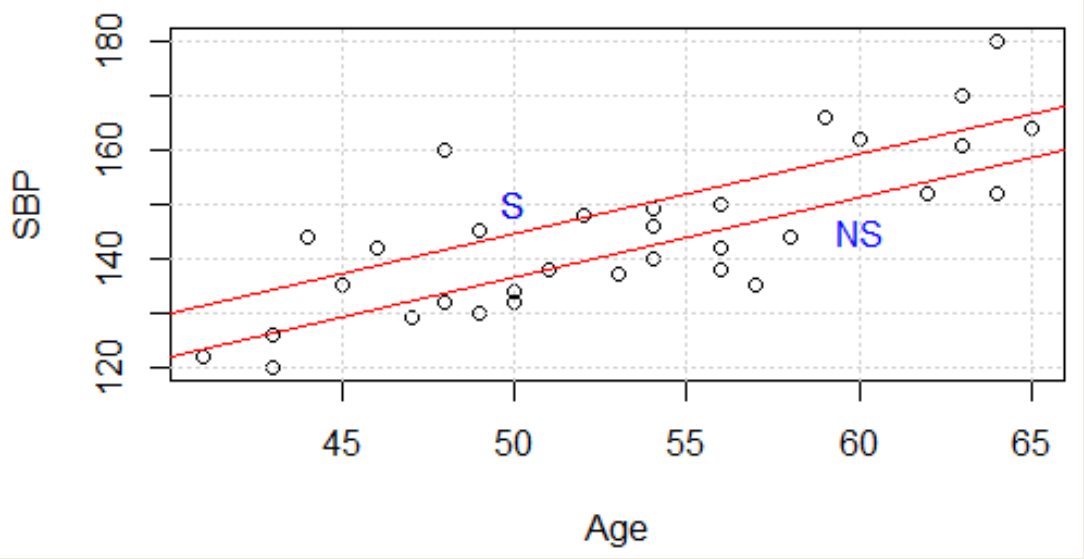
```
Analysis of Variance Table

Response: SBP
           Df Sum Sq Mean Sq F value    Pr(>F)
Age         1 3861.6  3861.6 48.1878 1.513e-07
Race        2  320.5   160.2  1.9997    0.1543
Residuals 28 2243.8    80.1
```

3. Interaction in Multiple Regression

In the model

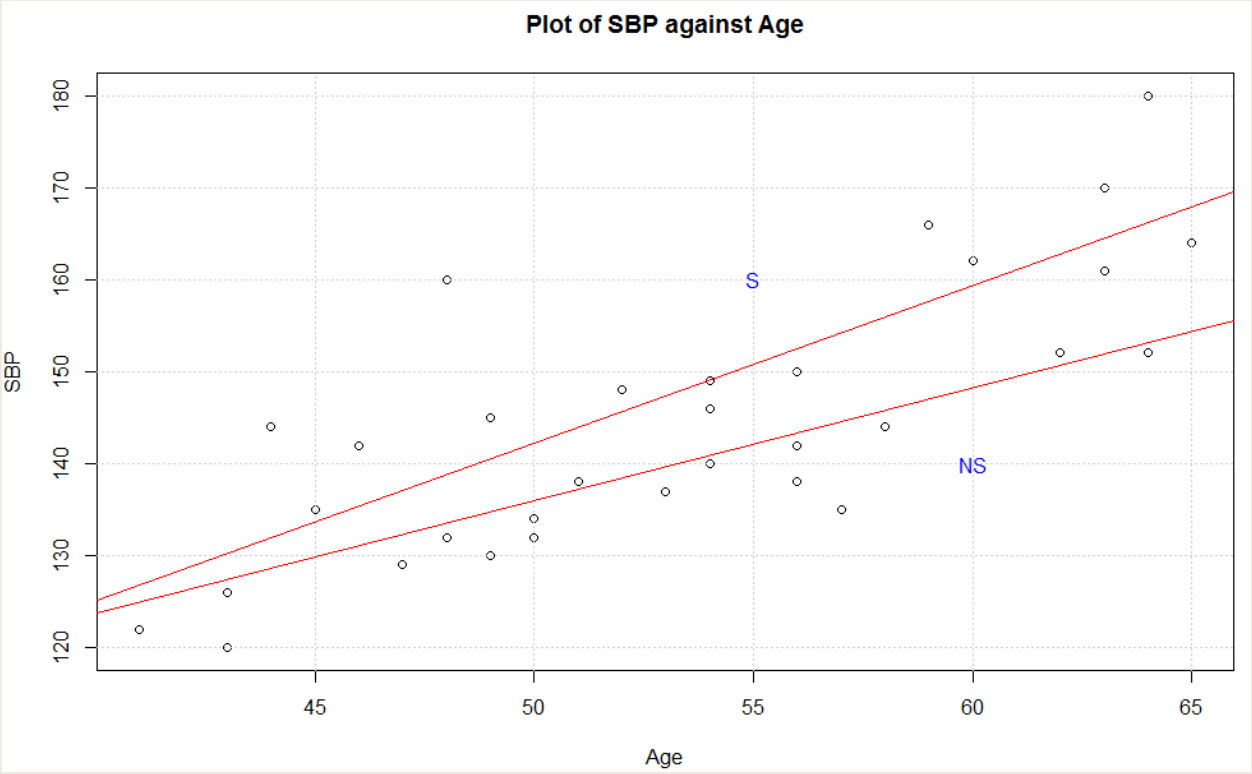
$$\widehat{\text{SBP}} = 62.389 + 1.464\text{Age} + 7.882\text{Smoke}$$

the impact on SBP of a one-year increase in age is the same regardless of smoking status. Similarly, the impact of being a smoker (rather than a non-smoker) is independent of age. Does this make sense?

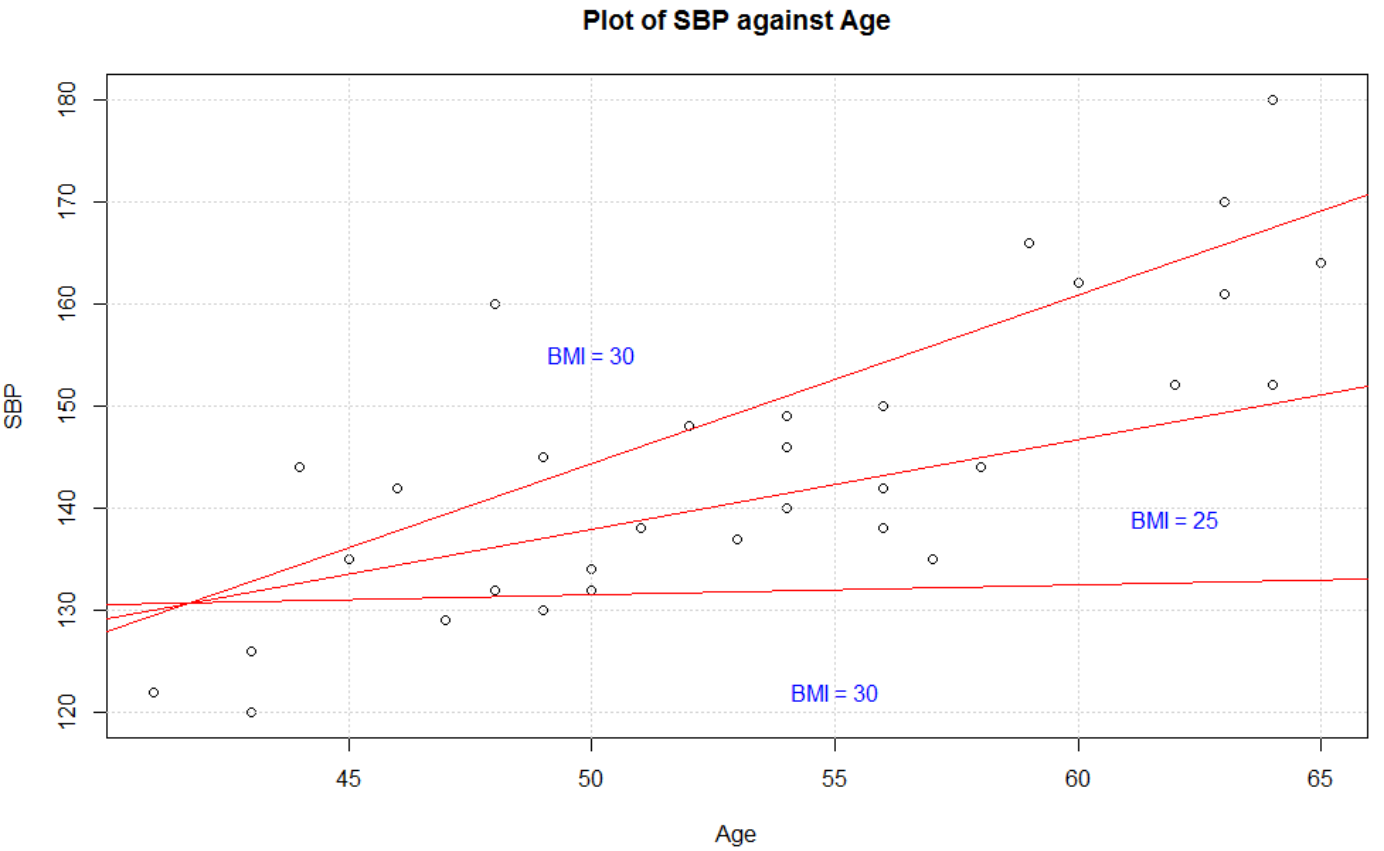


We introduce an interaction term by creating a new variable, Age*Smoke.

$$\widehat{\text{SBP}} = 74.673 + 1.225\text{Age} - 18.041\text{Smoke} + 0.487\text{Age*Smoke}$$



Exploring Interaction is not limited to the situation where one of the two X variables is 0/1. Here is the result of adding an interaction term when we regress SBP on Age and BMI.



4. Obtaining the ‘Best’ Regression Model

We have a response variable, Y and K potential explanatory (X) variables. We want to construct a model based on a subset of the K variables that provide useful information about Y

Almost always, the selection of the ‘best’ model involves a compromise between a model that explains a high percentage of the variability in Y and a model that is parsimonious—that is, simple and intuitive.

Preliminary Task 1

Explore the relationships the data thoroughly both numerically and graphically before beginning the task of model building.

Identify the k potential variables. An early analysis should identify higher-order powers, (X^2), other transformations such as ln(X) or X^2, and interaction terms (X1*X2).

Preliminary Task 2

Specify the criteria you will use to compare models:

1. High value for R^2

2. High value for R^2_Adj $= 1 - (1 - R^2)\frac{n - 1}{n - K - 1}$

Here, n is the number of observations and K the number of explanatory (X) variables in the model.

3. Low value for S_e^2 = MSRES

4. Akaike Information Criterion (AIC).

$$AIC = n*[\ln(2\pi) + 1 + \ln(SSRes/n)] + 2*(p + 2)$$

The magnitude of this measure of the quality of the fit for a model varies from context to context. Generally, the lower the value for AIC, the better the fit of the model to the data.

5. Mallow's $C_p = \frac{SSRES(p)}{MSRES(K)} - [n - 2(p + 1)]$

SSRES(p) is the SSRES with p variables (p < K) variables in the model and MSRES(K) is the MSRES with the *maximum model*—the model with all K variables included. It has been widely suggested that a good model is one in which C_p is close to p + 1. But, be aware that the full model (if you insert all K variables will always have Cp = K +1.

Preliminary Task 3

Select the procedure for obtaining a ‘best’ model

1. **Forward selection**—begin with one variable and build up.

2. **Backward selection**—start with the maximum model and work down by throwing out variables

3. **All-possible subsets**—computes every possible one of the $2^K - 1$ possible models and presents the best performing ones

Preliminary Task 4

Choose between an automated procedure and a procedure that you control.

Goldman's Forward Selection Procedure

Potential predictors $X_1, X_2, X_3, X_4, \ldots X_K$

Step 1

Begin with the variable (say X_3) that is the best single predictor of Y. Check that the t test produces a p-value < α.

Step 2

(a) Examine every pair of variables that include X_3, that is $X_3X_1, X_3X_2, \ldots X_3X_K$

(b) Select the pair with the highest R^2 (or R^2_Adj, or smallest S_e^2 = MSRES), say X_3X_1.

(c) Check for collinearity.

(d) If the p-value for the partial t test for X_1 < α move to step 3; otherwise, stop with X_3.

Step 3

(a) Examine every triple of variables that include X_3 and X_1 that is $X_3X_1X_2$, $X_3X_1X_4$, …$X_3X_1X_K$.

(b) Select the triple with the highest R^2 (or R^2_Adj, or smallest S_e^2 = MSRES), say $X_3X_1X_7$.

(c) Check for multicollinearity.

(d) If the p-value for the partial t test for $X_7 < \alpha$ move to step 4; otherwise, stop with X_3 X_1.

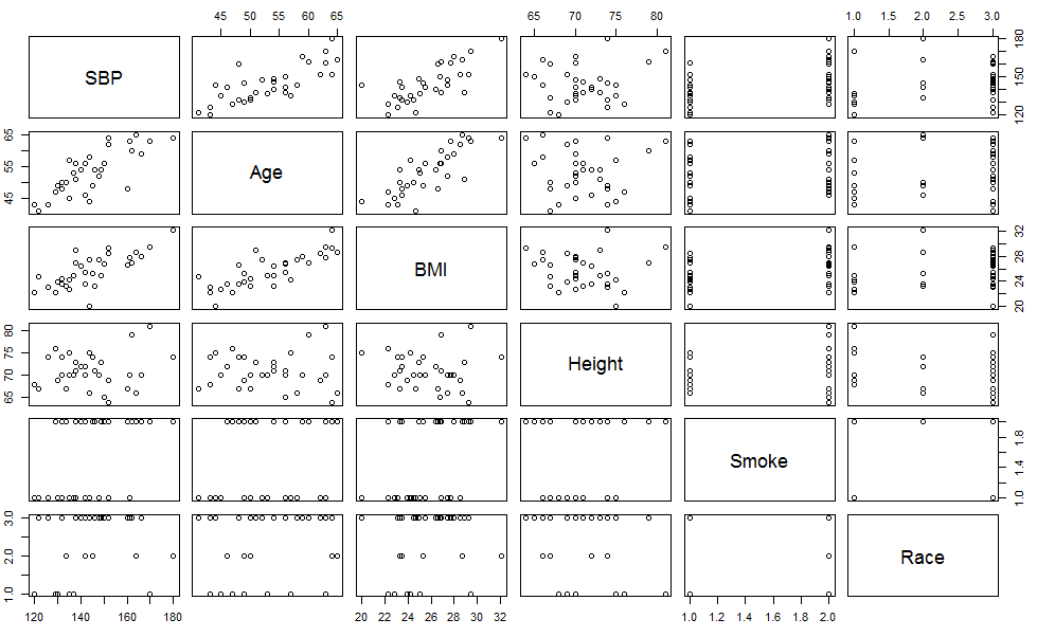
Continue this procedure until the p-value for adding the best new variable has a p-value > α.

At every stage, check whether it is worth throwing out a variable that, at an earlier stage, was promising.

Y SBP X_1 Age, X_2 BMI X_3 Height

X_4 Smoke X_5 Race

```
pairs(hd)
```




```
d <- hd[,-c(5,6)]
round(cor(d),3)
```

```
          SBP    Age    BMI Height
SBP     1.000  0.775  0.741  0.143
Age     0.775  1.000  0.802 -0.015
BMI     0.741  0.802  1.000 -0.095
Height  0.143 -0.015 -0.095  1.000
```

Step 1

	Variable(s)	R^2	R^2 - Adj	S_e	AIC	p-value
√	Age	0.6009	0.5876	9.245	237.1	0
	BMI	0.5490	0.5340	9.83	241.0	0
	Height	0.0204	-0.0123	14.49	265.8	0.436
	Smoke	0.2031	0.1765	13.07	259.2	0.0096
	Race	0.1262	0.066	13.91	264.2	0.1414

The p-value for Race is based on the overall F test; the other p-values are for t tests.

Step 2

	Variable(s)	R^2	R^2 - Adj	S_e	AIC	p-value
	Age + BMI	0.6409	0.6161	8.921	235.7	0.0831
	Age + Height	0.6250	0.5990	9.117	237.1	0.184
√	Age + Smoke	0.6734	0.6508	8.507	232.7	0.0169
	Age + Race	0.6508	0.6134	8.952	232.7	0.1543*

The p-value for Race is based on the partial F test; the other p-values are for t tests for adding a variable.

Step 3

Variable(s)	R^2	R^2 - Adj	S_e	AIC	p-value
Age + Smoke + BMI	0.6988	0.6665	8.314	232.1	0.135
Age + Smoke + Height	0.6846	0.6508	8.508	233.6	0.327
Age + Smoke + Race	0.6894	0.6434	8.597	235.1	0.506

The p-value for Race is based on the partial F test; the other p-values are for t tests for adding a variable.

The 'best' model involves Age and Smoke

$\widehat{SBP} = 62.4 + 1.46\ Age + 7.88\ smokers$

Note: At this point you will want to check if it is worth adding an Age*Smoke interaction term.

Note: At this point you will want to check for collinearity

Note: There may be some reason for including BMI.

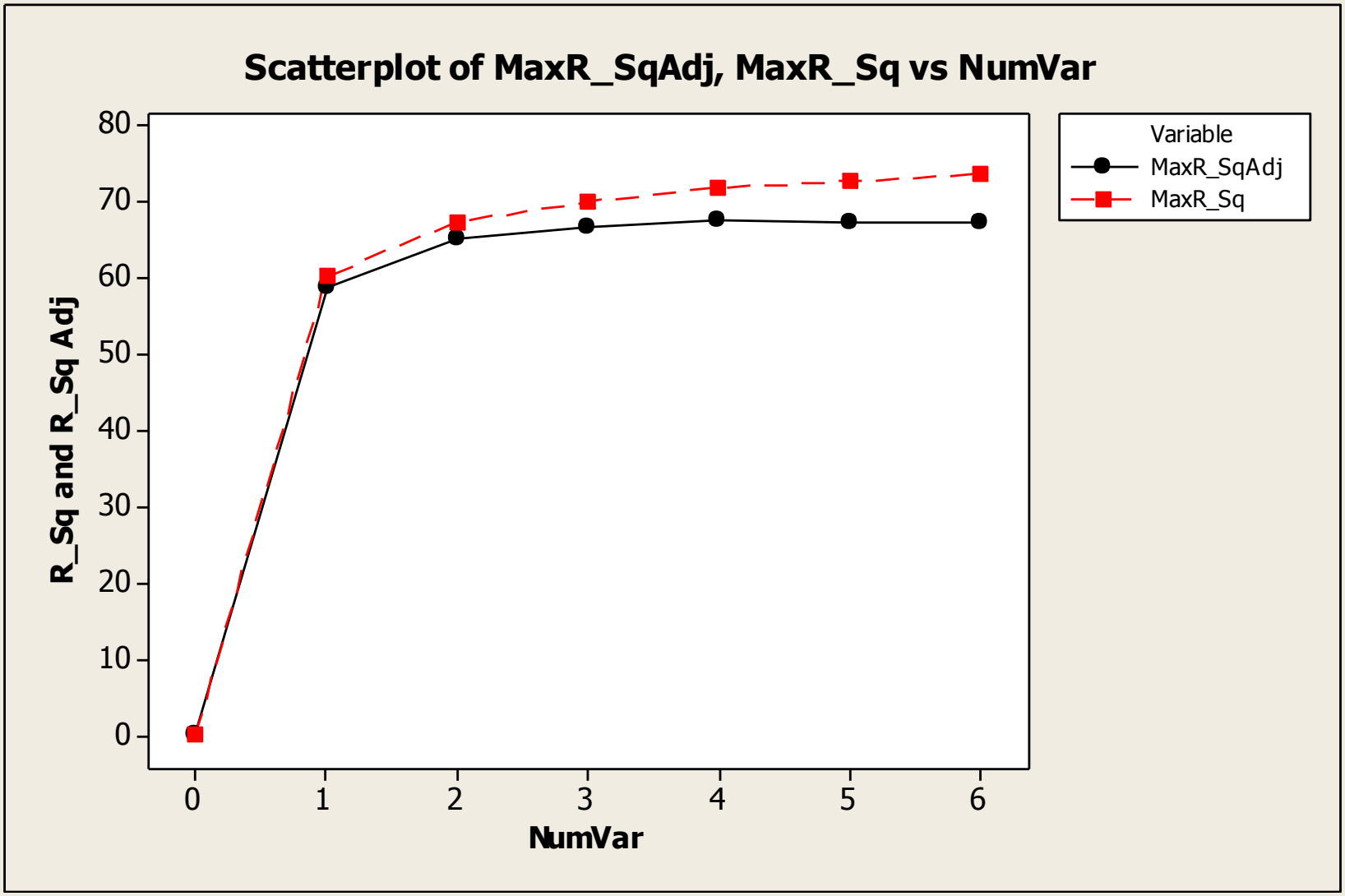
All Possible Subsets

```
library(leaps)
leaps <- regsubsets(SBP ~ Age + BMI + Height +
          Smoke + Race, hd, nbest = 2)


Selection Algorithm: exhaustive
          Age BMI Height Smokesmoker RaceHispanic RaceWhite
1  ( 1 ) "*" " " " "    " "          " "           " "
1  ( 2 ) " " "*" " "    " "          " "           " "
2  ( 1 ) "*" " " " "    "*"          " "           " "
2  ( 2 ) "*" "*" " "    " "          " "           " "
3  ( 1 ) "*" "*" " "    "*"          " "           " "
3  ( 2 ) "*" " " "*"    "*"          " "           " "
4  ( 1 ) "*" "*" "*"    "*"          " "           " "
4  ( 2 ) "*" "*" " "    "*"          "*"           " "
5  ( 1 ) "*" "*" "*"    "*"          "*"           " "
5  ( 2 ) "*" "*" "*"    " "          "*"           "*"
6  ( 1 ) "*" "*" "*"    "*"          "*"           "*"
```

The regsubsets functions uses Mallows Cp as the criterion for finding the best models.

Minitab's Best Subsets

Vars	R-Sq	R-Sq(adj)	Mallows Cp	S	Age	BMI	Height	Smoke	Hispanic	White
1	60.1	58.8	9.8	9.2454	X					
1	54.9	53.4	14.7	9.8282		X				
2	**67.3**	**65.1**	**4.9**	**8.5075**	**X**			**X**		
2	64.1	61.6	8.0	8.9209	X	X				
3	**69.9**	**66.7**	**4.5**	**8.3142**	**X**	**X**		**X**		
3	68.5	65.1	5.9	8.5084	X		X	X		
4	71.7	67.5	4.8	8.2049	X	X	X	X		
4	70.4	66.1	6.0	8.3868	X	X		X	X	
5	72.6	67.3	6.0	8.2331	X	X	X	X	X	
5	71.9	66.5	6.6	8.3304	X	X	X		X	X
6	73.6	67.3	7.0	8.2392	X	X	X	X	X	X


Scatterplot of MaxR_SqAdj, MaxR_Sq vs NumVar
R_Sq and R_Sq Adj
NumVar
0
10
20
30
40
50
60
70
80
0
1
2
3
4
5
6
Variable
MaxR_SqAdj
MaxR_Sq

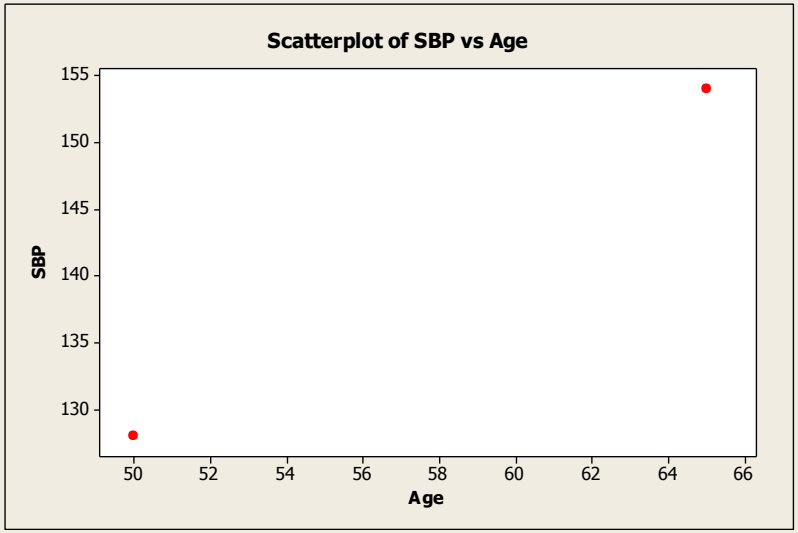
Overfitting

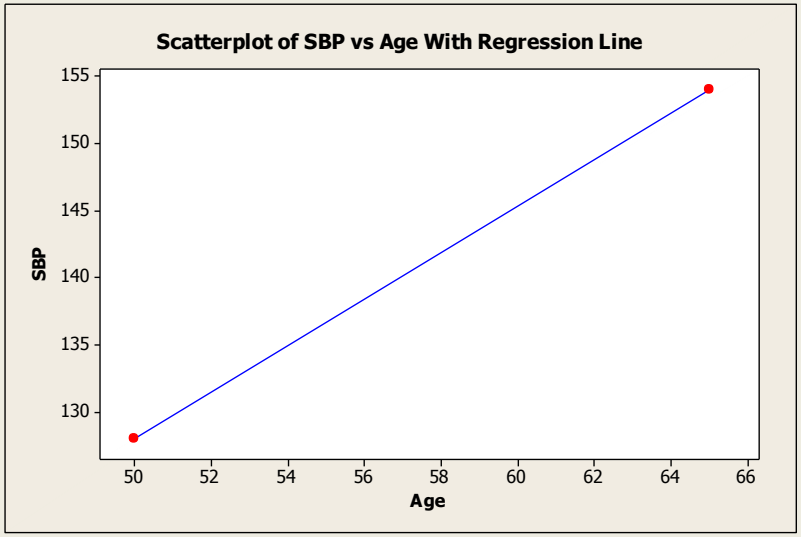
"Overfitting means that some of the relationships that appear statistically significant are actually just noise. It occurs when the complexity of the statistical model is too great for the amount of data that you have. You will find that a model with overfitting does not replicate well and does a lousy job of predicting future responses. In simple words, the model stinks!"

What matters in fitting models to data is not so much the sample size (n) itself but the sample size relative to the number of variables (K) in the model.

When the sample size (n) is small relative to the number of variables (K) in regression, the researcher risks finding significant slopes/coefficients just by chance. One rule of thumb for t tests in multiple regression is that there must be at least 20 times as many cases as independent variables. I like the compromise that suggests at least 10 times as many cases as variables.

```
x <- c(50, 65)
y <- c(128, 154)
model <- lm(y ~ x)
```



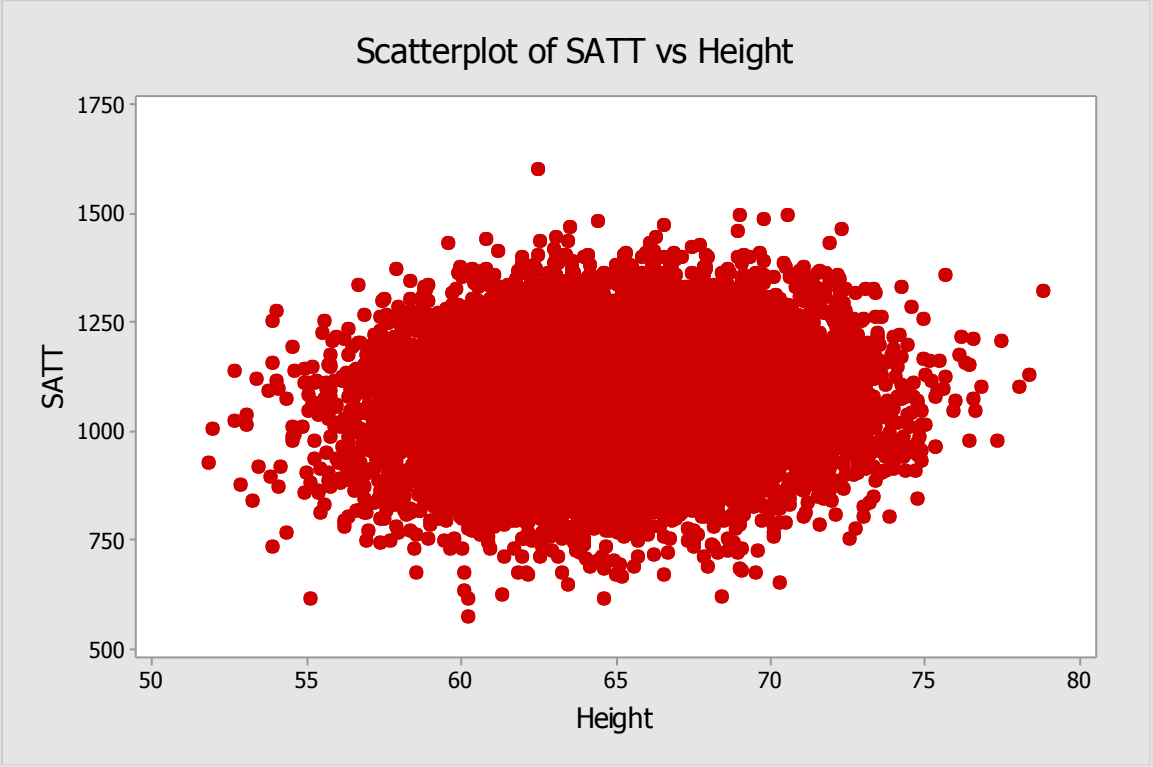


```
summary(model)

Residuals:
ALL 2 residuals are 0: no residual degrees of
freedom!

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   41.333         NA      NA       NA
x              1.733         NA      NA       NA

Residual standard error: NaN on 0 degrees of
freedom

Multiple R-squared:       1

F-statistic:   NaN on 1 and 0 DF,  p-value: NA
```

```
anova(model)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value Pr(>F)
x          1    338     338
Residuals  0      0
Warning message:
In anova.lm(model) :
  ANOVA F-tests on an essentially perfect fit
are unreliable
```

Can The Sample Size be too large?

In large samples classical hypothesis tests become meaningless because even tiny departures from the null hypothesis are statistically significant. Here is a scatterplot of SAT scores for about 15,000 female 12th graders against height.



```
model <- lm(SATT ~ Height)
model


Coefficients:
(Intercept)       Height
    920.974        2.208

summary(model)


Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 920.9742    17.9442  51.324  < 2e-16
Height        2.2080     0.2758   8.006 1.27e-15

Multiple R-squared:  0.004256
```

This situation illustrates one of the major problems with using classical statistical methods with ‘Big Data”.

5. Checking Inference Conditions in Regression

For the SBP data, our “best” model included the two explanatory variables, Age and Smoke.

```
model <- lm(SBP ~ Age + Smoke, hd)
summary(model)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  62.3892    11.8647   5.258 1.24e-05
Age           1.4639     0.2265   6.462 4.52e-07
Smoke         7.8818     3.1082   2.536   0.0169


anova(model)
Response: SBP
          Df Sum Sq Mean Sq F value  Pr(>F)
Age        1 3861.6  3861.6 53.3545 4.8e-08
Smoke      1  465.4   465.4  6.4304 0.01687
Residuals 29 2098.9    72.4


a <- c(40, 40, 60, 60)
b <- c(1, 0, 1, 0)
k <- data.frame(Age = a, Smoke = b)
p <- predict(model, newdata = k, interval =
"confidence")
round(p, 2)
 Age Smoke  fit    lwr    upr
1 40   1   128.83 120.77 136.89
2 40   0   120.95 113.99 127.90
3 60   1   158.11 153.25 162.96
4 60   0   150.23 144.24 156.21
```

Validity conditions??

1. **The linearity condition**: there is a straight line relationship of the form

$\mu_{SBP} = \beta_0 + \beta_1 Age + \beta_2 Smoke$

between the age of patient (X_1), Smoking Status (X_2) and mean SBP (μ_Y).

2. **The Normality condition**: for any particular combination of Age and Smoke, the distribution of SBP is Normal.

3. **The equal standard deviation condition**: for any particular combination of Age and Smoke, the standard deviation (σ) of SBP is the same.

For inference, we require the condition that the n = 32 patients are a random sample from the population to which we wish to make the inference.

We cannot confirm the validity of these conditions because we don't see the entire population; but we can use the information in the sample to get an idea of how valid they are.

1. The methods we use to check the validity of our regression model are the same regardless of the complexity of the model.

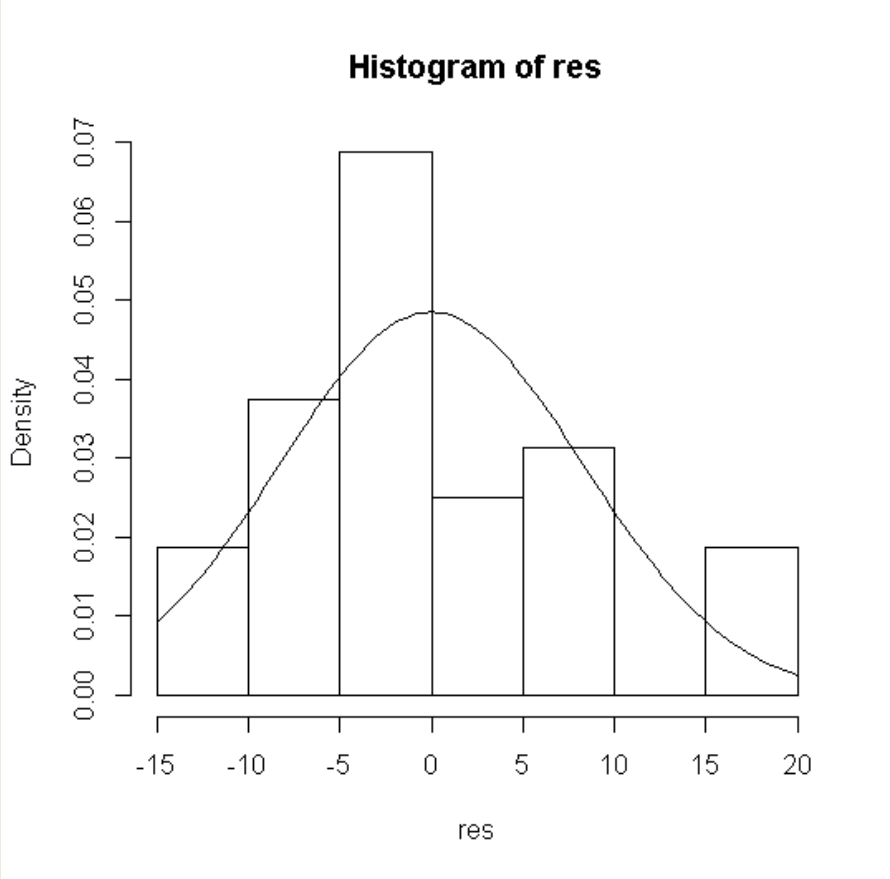
2. We check each of the conditions by focusing on the residuals, $Y - \hat{Y}$, or the *standardized residuals*, $\frac{Y - \hat{Y}}{S_e}$.

We use the residuals from the model predicting SBP from Age and Smoke to illustrate the procedures.

Checking the Normality Condition

1. Histogram of Residuals with Overlaid Normal

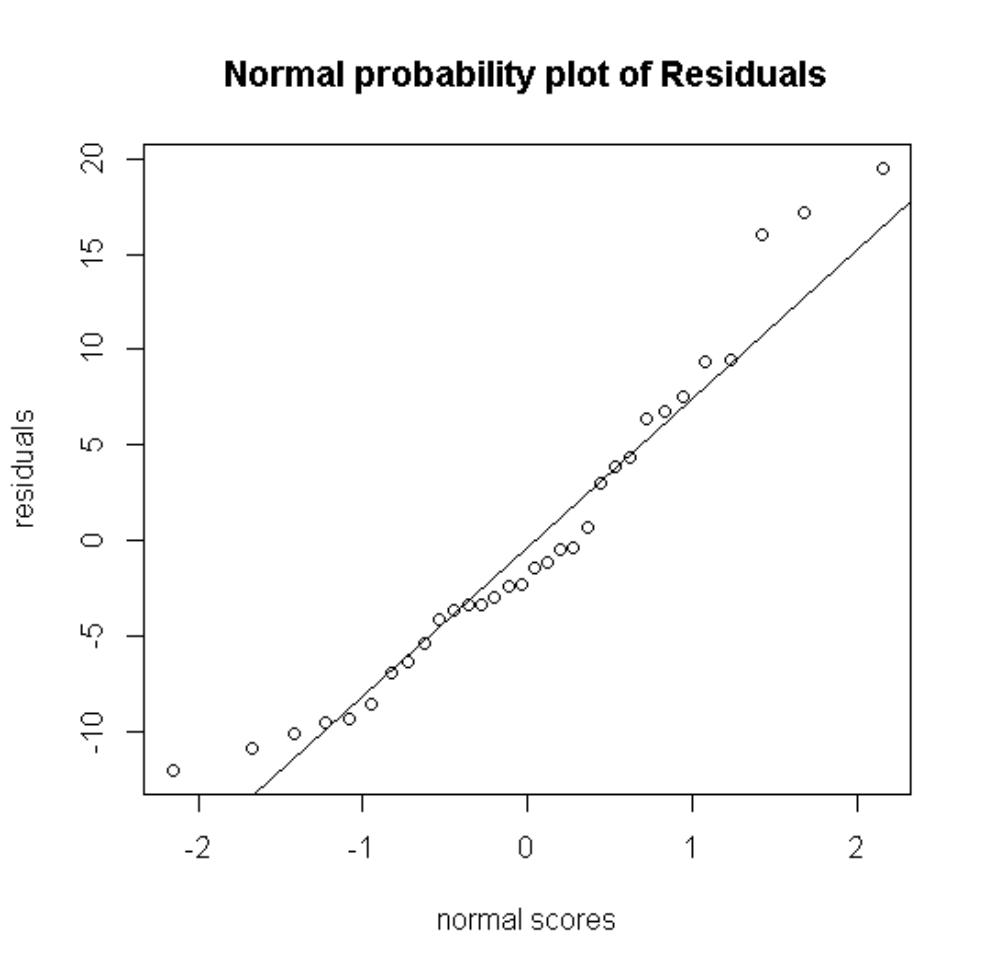
```
res <- resid(model)
fit <- fitted(model)
res<- resid(model)
hist(res,prob = T)
curve(dnorm(x, mean=mean(res), sd =
sd(res)), add = T)
```



2. Normal Probability Plot of Residuals

.

```
qqnorm(res, ylab = "residuals", xlab = "normal scores", main = "Normal probability plot of Residuals")

qqline(res)
```



3. Formal Test for Normal Residuals

.

H_0: The sample of residuals is drawn from a Normal distribution

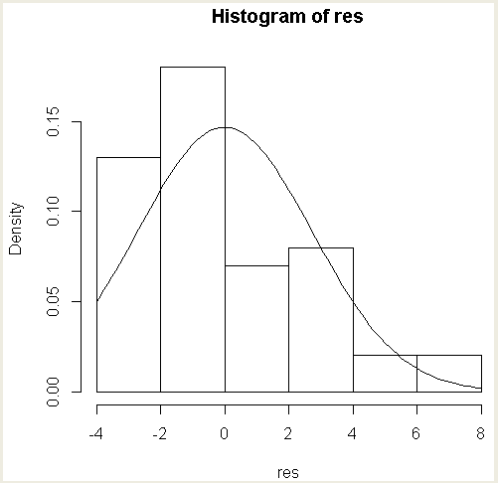
H_A: The sample of residuals is drawn from a non-Normal distribution

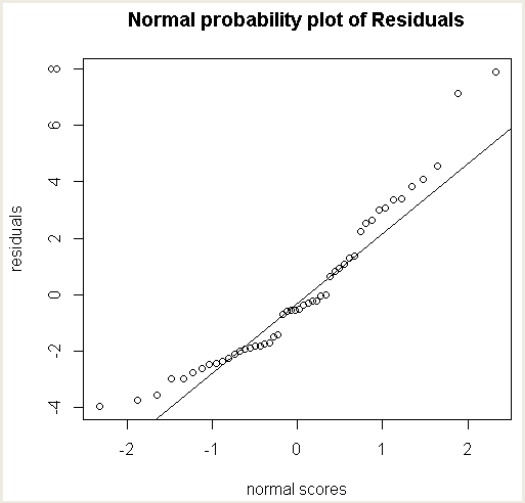
The Shapiro-Wilks Test

```
shapiro.test(res)

    Shapiro-Wilk normality test

data:  res
W = 0.9457, p-value = 0.1085
```

Tests for Normality is the one situation where researchers do not want to reject the null hypothesis. We want a high p-value!

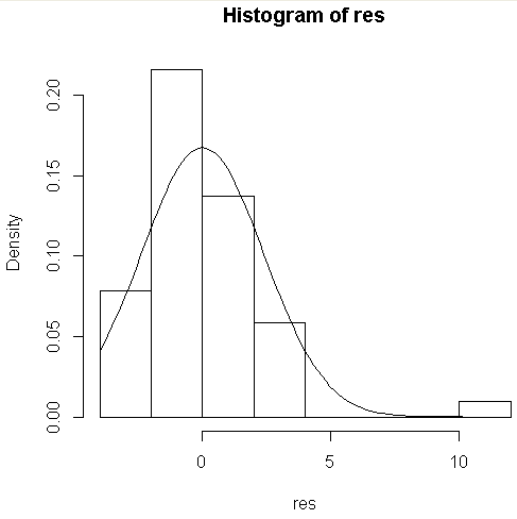
Bad News Residuals (1)

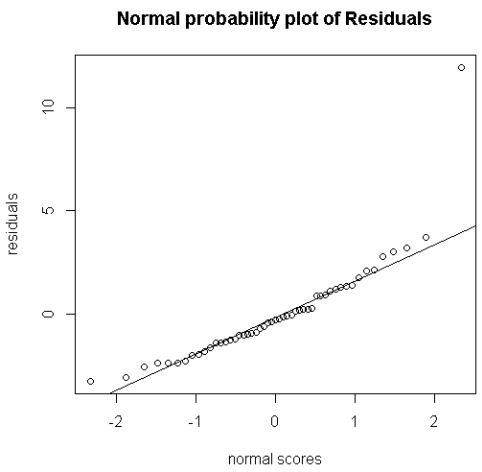





```
    Shapiro-Wilk normality test

data:  res
W = 0.9276, p-value = 0.004486
```

Bad News Residuals (2)



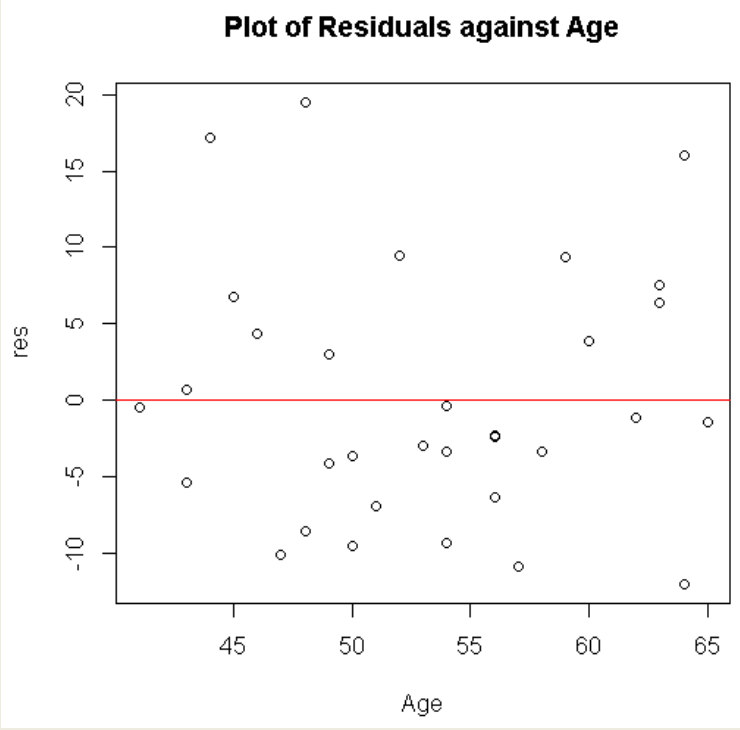



```
    Shapiro-Wilk normality test

data:  res
W = 0.8019, p-value = 7.99e-07
```

Checking the Linearity and the Equal Standard Deviation Condition

Both conditions can be checked visually with a scatterplot of the residuals plotted against either X or $\hat{Y}$.

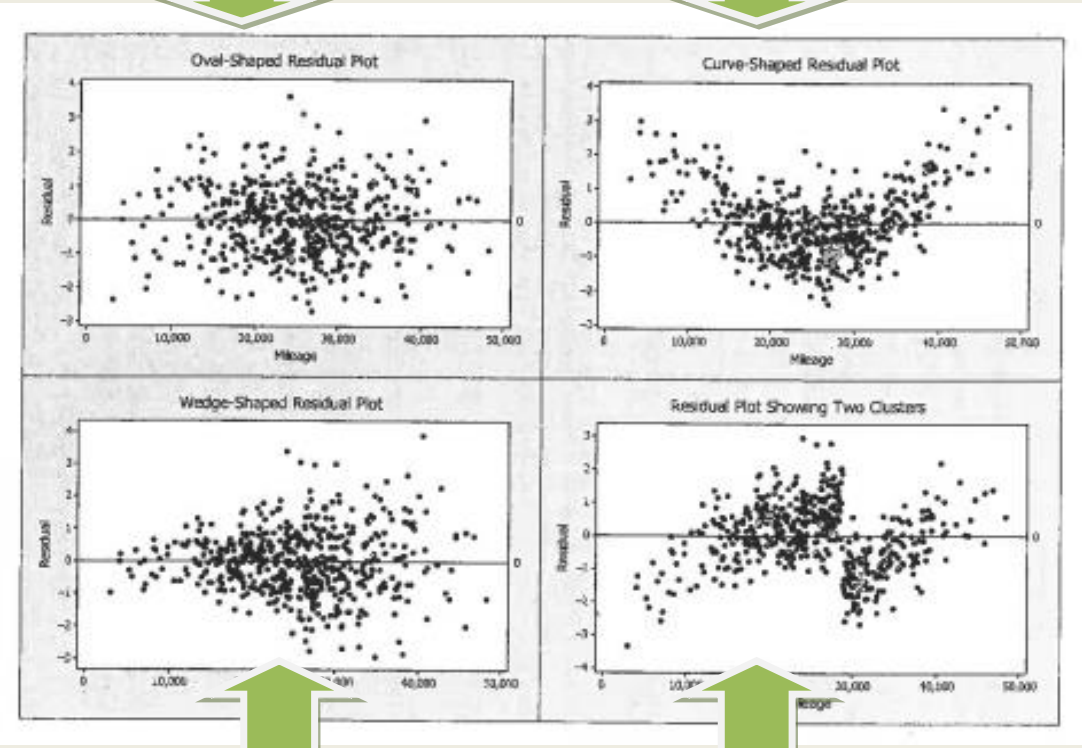
```
plot(res ~ Age, main = "Plot of Residuals
against Age")
abline(0,0, col = "red")
```



Some Residual Plots

'Ideal' Plot . The plot suggests random noise around 0; the variability of the residuals remains roughly constant

This residual plot suggests a non-linear trend in the original data



This plot suggests the linearity condition is valid but the condition that the spread of the residuals remain constant is clearly violated.

In this case neither the linearity nor the equal standard deviation condition is valid.

Multiple Linear Regression Issues not Discussed

- Regression Diagnostics—Dealing with outliers, influential values, transformations

- Techniques other than Least squares—Ridge Regression, Robust Regression, Computer-intensive methods in multiple regression

- Polynomial Regression—fitting models like

$$Y = b_0 + b_1X + b_2X^2 + b_3X^3$$

- Cross-Validation Techniques

- One-Way and Two-Way ANOVA as Regression Models

Appendix: The Letter 'e' and Natural Logarithms

The letter 'e' plays an important role in mathematics and statistics and in logistic regression, in particular. In fact, e is a constant whose value is approximately 2.71828.

Think of e as the base amount of growth shared by all continually growing processes**.** It shows up whenever systems grow exponentially and continuously: populations, radioactive decay, interest calculations, and so on.

Suppose $1000 is invested at 5% per year compounded continuously. After two years the $1000 will have grown to

$$1000e^{0.05(2)} = 1000e^{0.1} = 1000(1.10517) = \$1105.17$$

In this module you will have to compute quantities like $e^{1.42}$ or $e^{0.062}$. Use the $\mathbf{e^x}$ button on your calculator to do this.

Verify that (a) $e^{1.42} = 4.137$ and (b) $e^{0.062} = 1.064$

It is useful to recall that $e^{-X} = \frac{1}{e^X}$.

So, (a) $e^{-1.42} = \frac{1}{e^{1.42}} = \frac{1}{4.137} = 0.2417$

and (b) $e^{0.062} = \frac{1}{e^{0.062}} = \frac{1}{1.064} = 0.9399$

A standard alternative to the notation e^X is exp(X). This is particularly helpful when the exponent is a nasty expression. For example, it is much easier to write $e^{-1.23 + 4.567X}$ as exp(-1.23 + 4.567X).

It is tough typing expressions like $\frac{e^{b0 + b1X}}{1+e^{b0 + b1X}}$. You can write this in the form $\exp(b_0 + b_1X) / [1 + \exp(b_0 + b_1X)]$.

You will have encountered logarithms to the base 10 in the past. For example:

$10^3 = 1000$ The log to the base 10 of 1000 is 3 $\log_{10}1000 = 3$

$10^2 = 100$ The log to the base 10 of 100 is 2 $\log_{10}10 = 2$

$10^1 = 10$ The log to the base 10 of 10 is 1 $\log_{10}1 = 1$

$10^0 = 1$ The log to the base 10 of 1 is 1 $\log_{10}1 = 0$

In general, the log to the base 10 of a number X ($\log_{10}X$) is the power of 10 that yields X.

In the distant past (before calculators) logs to the base 10 were handy for doing calculations. There were tables of logarithms.

Today, we use 'natural' logarithms which are logarithms to the base e. We use the notation $\log_eX = \ln(X)$ to denote the logarithm to the base e of X. Here are some examples

$\log_e100 = \ln(100) = 4.605$ $e^{4.605} = \exp(4.605) = 100$

$\log_e2.5 = \ln(2.5) = 0.9163$ $e^{0.9163} = \exp(0.9163) = 2.5$

$\log_e8.1662 = \ln(8.1662) = 2.1$ $e^{2.1} = \exp(2.1) = 8.1662$

Practice finding ln(100), ln(2.5), and ln(8.1662) on your calculator.

(On most scientific calculators, $\log_e$ is the **ln** button.)