BWH - Biostatistics

Intermediate Biostatistics for Medical Researchers

Robert Goldman
Professor of Statistics
Simmons College

Thursday, April 5, 2018

Multiple Regression

1. Introduction

A multiple linear regression model is one in which there is more than one explanatory (X) variable. In the examples that follow I will use the *hd* data.

```
head(hd)
  SBP Age  BMI Height Smoke  Race
1 135  45 22.8     70     0 Black
2 122  41 24.7     67     0 White
3 130  49 23.9     69     0 Black
4 148  52 27.4     70     0 White
5 146  54 23.3     71     1 White
6 129  47 22.3     76     1 Black
```

Here are some possible sample, multiple linear regression models:

1. $\widehat{SBP} = b_0 + b_1 Age + b_2 BMI$

2. $\widehat{SBP} = b_0 + b_1 Age + b_2 Smoke$

3. $\widehat{SBP} = b_0 + b_1 Age + b_2 Age^2$

4. $\widehat{SBP} = b_0 + b_1 Age + b_2 BMI + b_3 Smoke$

5 $\widehat{SBP} = b_0 + b_1 Age + b_2 BMI + b_3 Age*BMI$

Why bother with multiple regression models?

1. Our interest in fitting a regression model is often:

(a) Accounting for variability in Y

(b) Using the fitted model for prediction

Either way, we can generally do better by including two or more predictor/explanatory variables in the model.

2. You can use multiple regression to remove the effect of confounding variables, particularly in observational studies.

3. Finally, even if we have only a single explanatory variable, multiple regression allows to consider a wide range of models such as quadratic models and cubic models. For example:

$$\hat{Y} = b_0 + b_1 X + b_2 X^2 + b_3 \frac{1}{X}$$

Some of these models don't look very linear!

$$\widehat{SBP} = b_0 + b_1 Age + b_2 Age^2$$

$$\hat{Y} = b_0 + b_1 X + b_2 X^2 + b_3 \frac{1}{X}$$

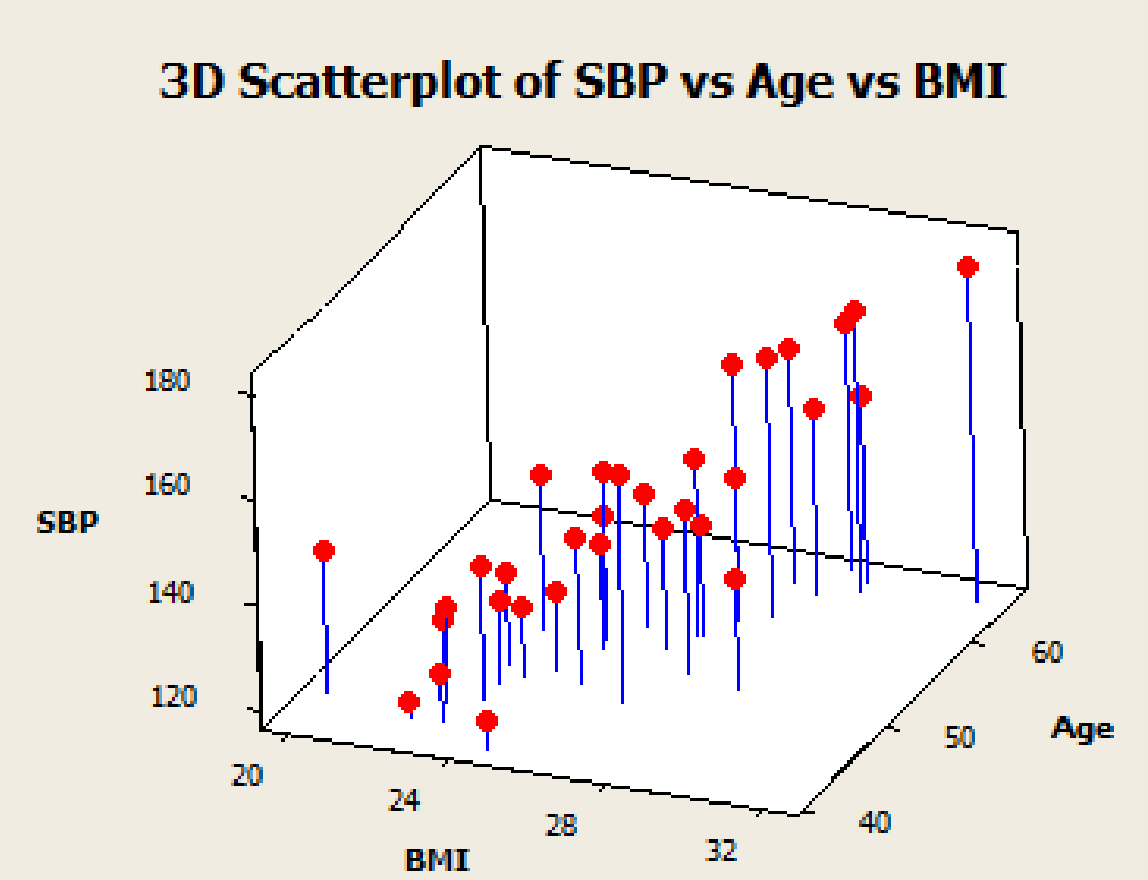
When we talk about 'linear' models we mean linear in the coefficients (the b's) not in the X's.

Here is an example of a non-linear model

$$\widehat{P(Y = 1)} = \frac{e^{b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3}}{1 + e^{b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3}}$$

The coefficients/parameters to be estimated are located in the exponent of the expression. We discuss this non-linear model in week 4.

2. Begin by fitting a model predicting SBP from both Age and BMI.



$$\widehat{SBP} = b_0 + b_1 Age + b_2 BMI$$

Our task is to find the particular values b_0, b_1, and b_2 associated with the 'plane'

$$y = b_0 + b_1 Age + b_2 BMI$$

which minimizes the sum of squared differences:

$$\sum d^2 = \sum [\, y - (b_0 + b_1 Age + b_2 BMI)]^2$$

```
model2 <- lm(SBP ~ Age + BMI, hd)
model2
```

```
Coefficients:
(Intercept)          Age          BMI
     41.774        1.050        1.822
```

The fitted model with Age and BMI is

$$\widehat{SBP} = 41.774 + 1.050 Age + 1.822 BMI$$

The fitted model with Age alone is

$$\widehat{SBP} = 59.114 + 1.604\ Age$$

(a) Regression Coefficients (b_0, b_1, and b_2)

$\widehat{SBP} = 41.774 + 1.050Age + 1.822BMI$

The intercept, 41.774 does not have a meaningful interpretation. It is the predicted SBP for patients aged 0 with a BMI of 0.

After adjusting for BMI, for every additional year of age, predicted SBP increases by 1.050 mm.

After adjusting for Age, for every additional point in BMI, predicted SBP increases by 1.822 mm.

So, each of the two regression slopes/coefficients are interpreted *conditional on adjusting for the other variable(s).*

Be careful not to interpret these regression coefficients as implying causation.

Here is a three –predictor model

```
model3 <- lm(SBP ~ Age + BMI + Smoke, hd)
model3


Coefficients:
(Intercept)      Age      BMI       Smoke
     48.08    1.029    1.473       7.139
```

$\widehat{SBP}$ = 48.08 + 1.029 Age + 1.47 BMI

+ 7.14 Smoke

After adjusting for BMI and Smoking Status, for each additional year of age, predicted SBP increases by 1.029 mm.

After adjusting for Age and Smoking Status, for each extra unit of BMI, predicted SBP increases by 1.47 mm.

After adjusting for Age and BMI, …..

A closer look at interpreting slopes in multiple regression.

$\widehat{SBP} = 41.774 + 1.050\text{Age} + 1.822\text{BMI}$

After adjusting for BMI, for every additional year of age, predicted SBP increases by 1.050 mm.

What does *adjusting for BMI* mean?

Regress SBP on BMI and save the residuals.

```
SBP.WO.BMI <- resid(lm(SBP ~ BMI, hd))
```

Now, regress Age on BMI and save the residuals.

```
Age.WO.BMI <- resid(lm(Age ~ BMI, hd))
```

Finally, regress SBP.WO.BMI on Age.WO.BMI

```
model <- lm(SBP.WO.BMI ~ Age.WO.BMI)
model

 (Intercept)   Age.WO.BMI
  4.042e-16    1.050e+00
```

SBP.WO.BMI = 1.05 Age.WO.BMI

$\widehat{SBP}$ = 48.08 + 1.029 Age + 1.47 BMI

+ 7.14 Smoke

After adjusting for BMI and Smoking Status, for each additional year of age, predicted SBP increases by 1.029 mm.

What does adjusting for BMI and Smoking Status mean?

Regress SBP on BMI and Smoking Status and save the residuals.

```
SBP.WO.BMI_SMK <-
          resid(lm(SBP ~ BMI + Smoke, hd))
```

Now, regress Age on BMI and Smoking Status and save the residuals.

```
Age.WO.BMI_SMK <-
          resid(lm(Age ~ BMI + Smoke, hd))
```

Finally, regress SBP.WO.BMI_SMK on Age.WO.BMI_SMK

```
model <- lm(SBP.WO.BMI_SMK ~
                                Age.WO.BMI_SMK)
model

Coefficients:
   (Intercept)  Age.WO.BMI_SMK
     5.563e-16       1.029e+00
```

SBP.WO.BMI_SMK = 1.029 Age.WO.BMI_SMK

Avoid phrases such as those italicized in the first half of the interpretations below.

Holding BMI and Smoking Status constant, for each additional year of age, predicted SBP increases by 1.029 mm.

Among patients with the same BMI and Smoking Status, for each additional year of age,predicted SBP increases by 1.029 mm.

(a) In complex models with many explanatory (X) variables not every combination of X variables is possible.

(b) When some of the explanatory (X) variables are moderately (or, even highly) correlated, it makes no sense to think of some being held constant as another increases by one unit.

(b) Predicted and Residual Values

$\widehat{SBP} = 41.774 + \mathbf{1.050}Age + 1.822BMI$

We define predicted values and residuals for each individual in multiple regression just as we did in simple regression. For example, the first person in the data set has a SBP of Y = 135 mm, is 45 years old with a BMI of 22.8.

The predicted SBP for this individual is

$\widehat{SBP} = 41.774 + 1.050(45) + 1.822(22.8)$

$= 130.55$ mm.

The residual for this individual is:

residual = e = SBP - $\widehat{SBP}$ = 135 - 130.55

= 4.45 mm.

```
fit2<- round(fit2, 2)
res2<- round(res2,2)
a <- data.frame(fit2, res2)
a
     fit2   res2
1  130.55   4.45
2  129.81  -7.81
3  136.75  -6.75
4  146.28   1.72
5  140.91   5.09
6  131.74  -2.74
7  153.76   8.24
8  140.62  19.38
9  124.40  19.60
10 167.44  12.56
11 154.72  11.28
12 147.96  -9.96
13 162.33 -10.33
14 149.56 -11.56
15 146.55  -6.55
16 136.71  -2.71
17 139.30   5.70
18 132.87   9.13
19 145.69 -10.69
20 147.01  -5.01
21 149.38   0.62
22 152.57  -8.57
23 142.95  -5.95
24 138.89  -6.89
25 143.82   5.18
26 134.97  -2.97
27 127.54  -7.54
28 129.00  -3.00
29 158.37   2.63
30 161.47   8.53
31 158.78  -6.78
32 162.29   1.71
```

What is the sum of the residuals?

(c) The Coefficient of Determination in Multiple Regression

```
Model2 <- lm(SBP ~ Age + BMI, hd)
summary(model2)

Multiple R-squared:  0.6409,

Adjusted R-squared:  0.6161
```

In multiple regression, the coefficient of determination has an interpretation similar that that in simple regression; it is the fraction of the variability in Y that can be associated with the linear relationship between Y and the X variables in the model.

Here, 64.1% of the variability in SBP can be associated with the linear relationship between SBP and both Age and BMI.

```
Multiple R-squared:  0.6409, Adjusted R-squared:
0.6161
```

The value 0.6409 looks like the square of a correlation coefficient but is it the correlation between SBP and Age, between SBP and BMI, or between Age and BMI?

It is none of these correlations! It is, in fact, the correlation between the actual SBP and the predicted SBP (i.e. between Y and $\hat{Y}$). The output below shows this.

```
cor(hd$SBP, fit)
[1] 0.8006015
```

$0.8006^2 = 0.6409$

In multiple regression the coefficient of determination, R^2, is the square of the correlation between Y and $\hat{Y}$.

Age Alone

```
Multiple R-squared:  0.6009
```

Age and BMI

```
Multiple R-squared:  0.6409
```

Why does adding the variable BMI add such little explanatory power to the model?

	SBP	Age	BMI	Height	Smoke
SBP	1.000	**0.775**	**0.741**	0.143	0.451
Age	0.775	1.000	**0.802**	-0.015	0.245
BMI	0.741	0.802	1.000	-0.095	0.287
Height	0.143	-0.015	-0.095	1.000	0.179
Smoke	0.451	0.245	0.287	0.179	1.000

The impact that an explanatory (X) variable has on the response variable (Y) depends critically on what point it is added to the model.

(d) ANOVA in Multiple Regression

Age

Source of Variation	Sum of Squares	df	Mean Square	F	p
Regression	3861.6	1	3861.6	45.177	0.000
Residual	2564.3	30	85. 5		
Total	6426.0	31			

Age and BMI

Source of Variation	Sum of Squares	df	Mean Square	F	p
Regression	4118.1	2	2059.05	25.87	0.000
Residual	2307.9	29	79.6		
Total	6426.0	31			

Source	df	Seq SS
Age	1	3861.6
BMI	1	256.5

Notice that the SSTOT (6426.0) is exactly the same as it was in the simple regression model with just Age. This is true because SSTOT = $\sum (Y_i - \bar{Y})^2$ depends only on the Ys; it will not change regardless of how many and which X variables are in the model.

By contrast, the SSRES = $\sum (Y - \hat{Y})^2$ has declined (by 256.5, from 2564.3 to 2307.9) and SSREG = $\sum (\hat{Y} - \bar{Y})^2$ has increased by the same amount (from 3861.6 to 4118.1).

As before, the coefficient of determination is the ratio of the SSREG and SSTOT:

$$R^2 = 0.641 = \frac{4118.1}{6426.0} = 0.641$$

As you can see, much of the descriptive structure of simple regression is maintained when we go to multiple regression.

ANOVA in Regression with R

```
model1 <- lm(SBP ~ Age, hd)
anova(model1)
Analysis of Variance Table

Response: SBP
          Df Sum Sq Mean Sq F value    Pr(>F)
Age        1 3861.6  3861.6  45.177 1.894e-07
Residuals 30 2564.3    85.5
```

```
model2 <- lm(SBP ~ Age + BMI, hd)
anova(model2)
Analysis of Variance Table

Response: SBP
          Df Sum Sq Mean Sq F value    Pr(>F)
Age        1 3861.6  3861.6 48.5239 1.171e-07
BMI        1  256.5   256.5  3.2226   0.08306 .
Residuals 29 2307.9    79.6
```

```
anova(model1, model2)
Analysis of Variance Table

  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     30 2564.3
2     29 2307.9  1    256.46 3.2226 0.08306 .
```

(e) Adjusted R^2

```
summary(model2)

Multiple R-squared:  0.6409,
Adjusted R-squared:  0.6161
```

$$\text{Adj } R^2 = 1 - (1 - R^2)\frac{n - 1}{n - K - 1}$$

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

For our data above, n = 32, K = 2 and R^2 = 0.6409.

$$\text{Here, Adj } R^2 = 1 - (1 - 0.6409)\frac{32 - 1}{(32 - 2 - 1)} = 0.6161$$

Suppose n = 12, K = 5 and $R^2 = 0.75$.

Then, Adj $R^2 = 1 - (1 - 0.75)\frac{12 - 1}{(12 - 5 - 1)} = 0.542$

Suppose n = 1200, K = 5 and $R^2 = 0.75$.

Then, Adj $R^2 = 1 - (1 - 0.75)\frac{1200 - 1}{(1200 - 5 - 1)} = 0.749$

In the first example above, with n = 12 and K = 5 the value for Adj R^2 (0.542) is substantially less than the value for R^2 (0.75). This is an indication of *overfitting*. This occurs when the complexity of the statistical model is too great for the amount of data that you have.

(f) Capturing the 'Importance' of Variables

$\widehat{\text{SBP}} = 48.08 + 1.029\ \text{Age} + 1.47\ \text{BMI} + 7.14\ \text{Smoke}$

In this model, which of the three predictors, Age, BMI, or Smoke, is the most 'important' or 'influential'?

Because these three variables are not linearly independent of one another, it is difficult to answer this question (or even to understand exactly what the question means).

You certainly cannot judge the importance of variables by their slope. This is because the the units for the three variables are quite different.

So, one approach to getting at this problem is to perform the regression on *standardized variables*.

Standardizing a variable is generally taken to mean subtracting the mean and dividing the difference by the standard deviation.

If we have a column/vector/variable of values:

$x = (x_1, x_2, x_3, \ldots, x_n)$

with mean, $\bar{x}$ and standard deviation, S_x, the standardized form of x will be:

$$\left(\frac{x_1 - \bar{x}}{S_x}, \frac{x_2 - \bar{x}}{S_x}, \frac{x_3 - \bar{x}}{S_x}, \ldots, \frac{x_n - \bar{x}}{S_x}\right)$$

In R we standardize a variable with the **scale** function

```
S_sbp <- scale(hd$SBP)
S_age<- scale(hd$Age)
S_bmi <- scale(hd$BMI)
S_smoke <- scale(hd$Smoke)
```

```
S_model <- lm(s_sbp ~ s_age + s_bmi + s_smoke)
S_model
```

```
Coefficients:
(Intercept)     s_age       s_bmi      s_smoke
  5.560e-17    0.4970      0.2702       0.2514
```

The intercept is 0.

The three slopes can be compared (sort of)

```
Coefficients:
(Intercept)     s_age        s_bmi       s_smoke
  5.560e-17    0.4970       0.2702        0.2514
```

After adjusting for BMI and Smoke, as Age increases by one standard deviation, the predicted SBP goes up by 0.497 standard deviations.

After adjusting for Age and Smoke, as BMI increases by one standard deviation, the predicted SBP goes up by 0.2702 standard deviations.

After adjusting for Age and BMI, as Smoke increases by one standard deviation, the predicted SBP goes up by 0.2514 standard deviations.

This last interpretation makes no sense, but I hope you get the idea!

3. The Population Model for Multiple Regression

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

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

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

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

3. **The equal standard deviation condition**: for any particular combination of Age and BMI, 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 estimate the four parameters of the population model (β_0, β_1, β_2, and σ) from the sample.

$\hat{\beta}_0 = b_0 = 41.774$

$\hat{\beta}_1 = b_1 = 1.050$

$\hat{\beta}_2 = b_2 = 1.822$

We estimate σ from the residuals.

$$\hat{\sigma}^2 = \frac{\Sigma(Y - \hat{Y})^2}{n - 3} = SSRES/(n - 3) = MSRES = 79.6$$

So, an estimate for σ is $S_e = \sqrt{79.6} = 8.921$ mm.

```
Analysis of Variance Table

Response: SBP
          Df Sum Sq Mean Sq F value    Pr(>F)
Age        1 3861.6  3861.6 48.5239 1.171e-07
BMI        1  256.5   256.5  3.2226   0.08306
Residuals 29 2307.9    79.6


Residual standard error: 8.921 on 29 degrees of
freedom
```

4. The t-Test in Multiple Regression

```
summary(model2)

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
```

In multiple linear regression there is a t test for each variable in the model! Each test is used to check whether it is worth-while to add that explanatory variable to a model based on the other variables.

For example, we would use the t test to answer the question; given that we have a model with Age alone, is it worth adding the variable BMI?

H_0: For the population model:

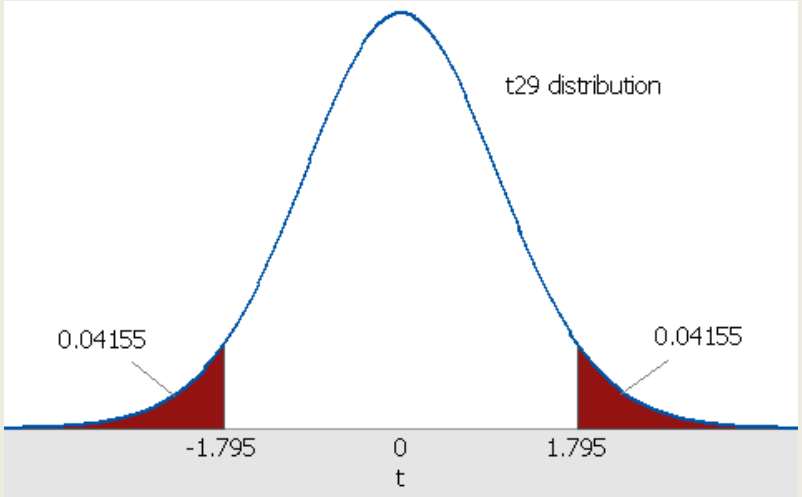
$$\mu_{SBP} = \beta_0 + \beta_1 Age + \beta_2 BMI \qquad \beta_2 = 0$$

H_A: For the population model:

$$\mu_{SBP} = \beta_0 + \beta_1 Age + \beta_2 BMI \qquad \beta_2 \neq 0$$

```
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
```

$$t = \frac{b_2 - 0}{SE(b_2)} = \frac{1.822}{1.015} = 1.795.$$



The p-value is 0.0831 which suggests that (at the 5% level of significance) the population slope associated with BMI in a model with Age is not (quite) significantly different from 0.

The data suggest that it is probably not worth adding BMI to a model with Age.

What about the t test associated with Age?

```
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
```

The t value, 2.723 (= 1.0496/0.3855) for Age, allows you to test for the value of adding Age to a pre-existing model with just BMI.

The p-value is 0.0108 which suggests that the population slope associated with Age in a model with BMI is significantly different from 0 (actually, greater than 0 because $b_1 = 1.0496$).

The data suggest that it is certainly worth adding Age to a model with BMI.

Continuing with the population model:

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

R will give you confidence intervals for β_1 and β_2.

```
round(confint(model2),3)
              2.5 % 97.5 %
(Intercept)  9.696 73.851
Age          0.261  1.838
BMI         -0.254  3.898
```

We can be 95% confident that β_1 lies between 0.261 and 1.838.

We can be 95% confident that β_2 lies between -0.254 and 3.898.

The fact that our t-test for adding BMI after Age was not significant (i.e., β_2 was not significantly different from 0) is reflected in the fact that our 95% confidence interval contains 0.

```
Model3 <- lm(SBP ~ Age + BMI + Smoke, hd)
summary(model3)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  48.0815    14.8681   3.234  0.00313 **
Age           1.0287     0.3594   2.862  0.00787 **
BMI           1.4726     0.9579   1.537  0.13542
Smoke         7.1385     3.0758   2.321  0.02780 *
```

The partial t-tests tell us that:

It is worth adding the variable Age to an existing model predicting SBP from BMI and smoking status (p = 0.00313)

It is not worth adding the variable BMI to an existing model predicting SBP from Age and smoking status (p = 0.13542)

It is worth adding the variable smoking status to an existing model predicting SBP from Age and BMI (p = 0.029)

5. The General F-Test in Multiple Regression

In simple regression the F test and the (two-sided) t test are equivalent. In multiple regression they are generally not. The F-test is a general tool for comparing models.

We have a regression model predicting Y from p predictors. We are interested in checking whether it is worth adding q more variables **as a block** (as opposed to one at a time).

There is an F-test for this!

The original model with p variables is called the **reduced** model. The model with p + q variables is called the **full** model.

R makes it really straightforward to conduct such F-tests.

The null hypothesis is that in the full model, the coefficients associated with the q variables are all 0.

The alternative hypothesis is that not all these coefficients are 0.

The test statistic is

$$F = \frac{[SSRES(p) - SSRES(p + q)]/q}{SSRES(p + q)/[n - (p + q) - 1]}$$

$$= \frac{\text{Decline in SSRES/(number of variables added)}}{MSRES(p + q)}$$

If the null hypothesis is true, F will follow an

$F_{q,\, n - (p+q) - 1}$ distribution

Example 1

We have a response variable, SBP, and are quite happy with a model predicting SBP from Age. Now we ask the question: Can we improve our model by adding the three variables BMI, Height and Smoke, *as a block*, to this model?

modelR	Age
modelF	Age, BMI, Height, Smoke

H_0: In the modelF

$$\mu_{SBP} = \beta_0 + \beta_1 Age + \beta_2 BMI + \beta_3 Height + \beta_4 Smoke$$

$$\beta_2 = \beta_3 = \beta_4 = 0$$

H_A: In the modelF

$$\mu_{SBP} = \beta_0 + \beta_1 Age + \beta_2 BMI + \beta_3 Height + \beta_4 Smoke$$

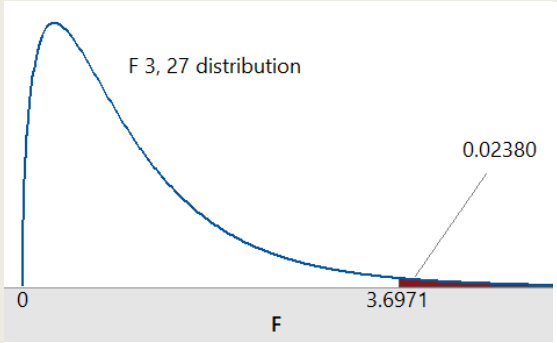
some of β_2, β_3, β_4 are not 0

```
modelR <- lm(SBP ~ Age, data = hd)
modelF <- lm(SBP ~ Age + BMI + Height + Smoke,
  data = hd)

anova(modelR, modelF)
Analysis of Variance Table

Model 1: SBP ~ Age
Model 2: SBP ~ Age + BMI + Height + Smoke
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     30 2564.3
2     27 1817.7  3    746.68 3.6971 0.0238 *
```

SSRES(p) SSRES(p + q) q SSRES(p)-SSRES(p + q) F

n – (p + q) – 1 = 32 – 4 – 1 = 27

The p-value for H_0 is the area under the $F_{3, 27}$ distribution to the right of 3.6971.



We may reject the null hypothesis at the 5% level of significance. The data suggest that it is worth adding this block of three variables.

Example 2 (The Overall F test)

Suppose we begin with a model predicting SBP from Age and BMI. Is this model statistically significant?

In this case p = 0 and q = 2.

H_0: In the modelF

$\mu_{SBP} = \beta_0 + \beta_1 Age + \beta_2 BMI \qquad \beta_1 = \beta_2 = 0$

H_A: In the modelF

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

at least one of $\beta_1, \beta_2 \neq 0$

The reduced model is

$\mu_{SBP} = \beta_0$

I could not get R to fit this reduced model but we can still perform the F test

```
modelF <- lm(SBP ~ Age + BMI, data = hd)
anova(modelF)
Analysis of Variance Table

Response: SBP
          Df Sum Sq Mean Sq F value    Pr(>F)
Age        1 3861.6  3861.6 48.5239 1.171e-07
BMI        1  256.5   256.5  3.2226   0.08306 .
Residuals 29 2307.9    79.6
```

Age and BMI

Source of Variation	Sum of Squares	df	Mean Square	F	p
Regression	4118.1	2	2059.05	25.87	0.000
Residual	2307.9	29	79.6		
Total	6426.0	31			

This test is a special case of the general F test.

For the reduced model

$SSRES(0) = \sum (Y - \bar{Y})^2 = SSTOT = 6426$

Example 3

There is one situation in multiple regression where the F test and the t test are identical.

Is it worth adding the variable Smoke to a model with Age alone?

```
modelR <- lm(SBP ~ Age, data = hd)
modelF <- lm(SBP ~ Age + Smoke, data = hd)
summary(modelF)

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(modelF)
Analysis of Variance Table
```

$F = 6.4304 = 2.536^2 = t^2$

```
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
```

```
anova(modelR, modelF)
Analysis of Variance Table

Model 1: SBP ~ Age
Model 2: SBP ~ Age + Smoke
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     30 2564.3
2     29 2098.9  1    465.41 6.4304 0.01687
```

6. Prediction in Multiple Regression

Suppose we want to obtain a 90% CI for the mean SBP for the following combination of Age and BMI.

	Age	BMI
1.	40	25
2.	40	30
3.	60	25
4.	60	30

```
a <- c(40, 40, 60, 60)  # new values for Age
b <- c(25, 30, 25, 30)  # new values for BMI
k <- data.frame(Age = a, BMI = b)
p <- predict(model2, newdata = k, interval =
     "confidence", level = 0.9)
round(p, 2)
     fit    lwr    upr
1 129.31 121.14 137.48
2 138.42 122.94 153.90
3 150.30 144.21 156.39
4 159.41 154.04 164.78
```

More elegantly:

```
  Age BMI    fit    lwr    upr
1 40  25  129.31 121.14 137.48
2 40  30  138.42 122.94 153.90
3 60  25  150.30 144.21 156.39
4 60  30  159.41 154.04 164.78
```

Model	$100R^2$
Age	60.1
Age, BMI	64.1
Age, Smoking Status	67.3
Age, Smoking Status, Height	68.5

7. Checking Inference Conditions in Regression

For the SBP data, a good model includes 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.

Random Sample?

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.

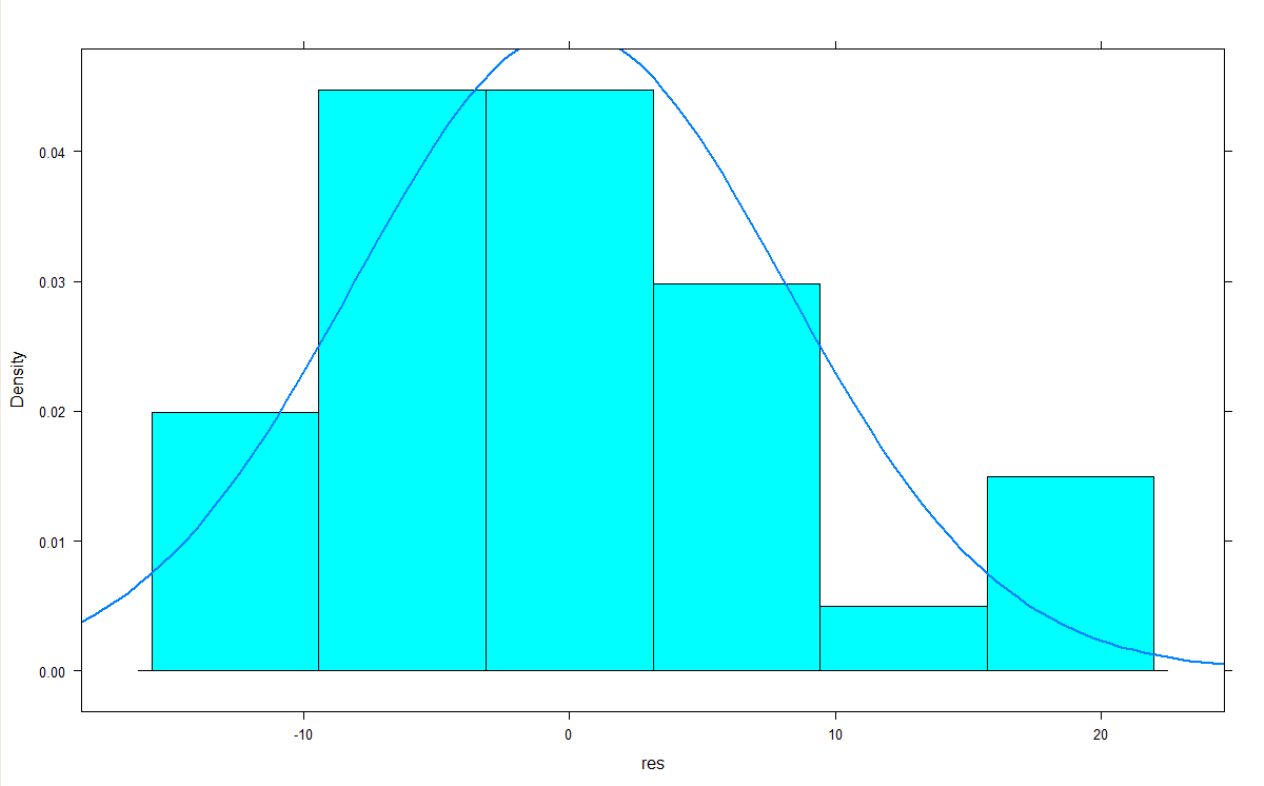
2. We check each of the conditions by focusing on the residuals, $Y - \hat{Y}$, or the *standardized residuals*,

$$\frac{Y - \hat{Y}}{S_e}.$$

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

1. Histogram of Residuals with Overlaid Normal

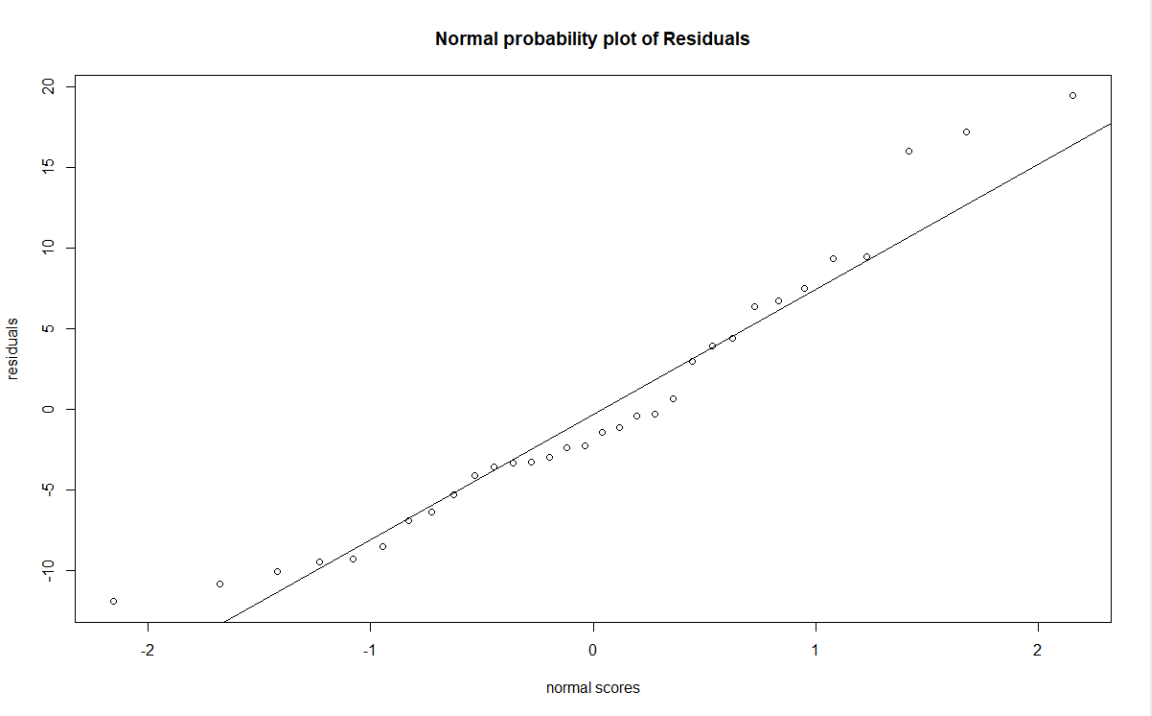
```
model <- lm(SBP ~ Age + Smoke, data = hd)
res <- resid(model)
fit <- fitted(model)
histogram(~res, fit = "normal")
```



2. Normal Probability Plot/Q-Q 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

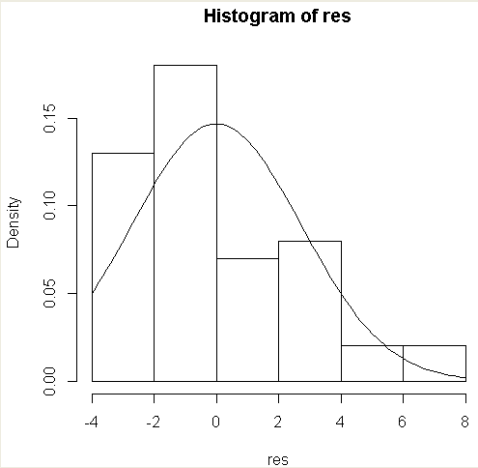
H_A: The sample of residuals is drawn from a non-Normal distribution

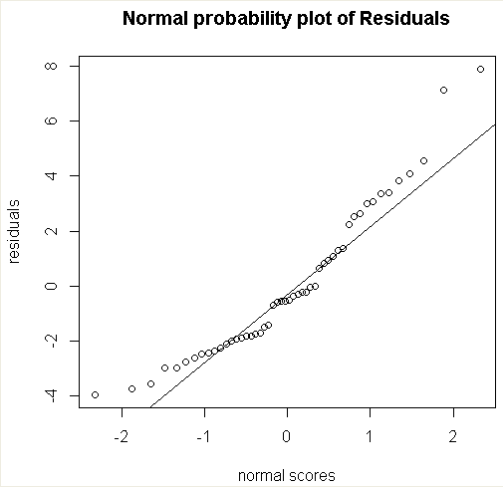
The Shapiro-Wilks Test

```
shapiro.test(res)

     Shapiro-Wilk normality test

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

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

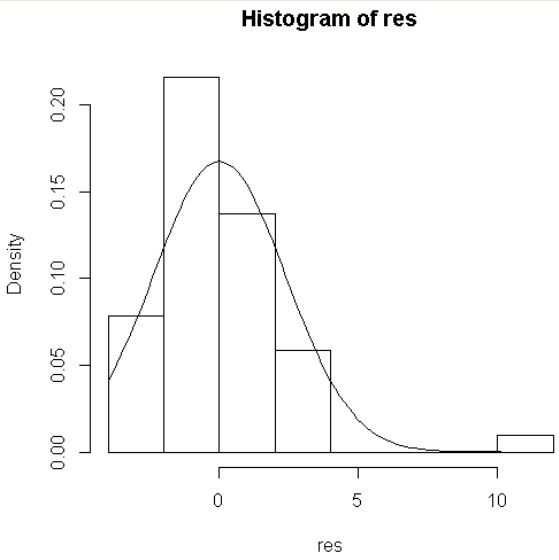
Bad News Residuals (1)

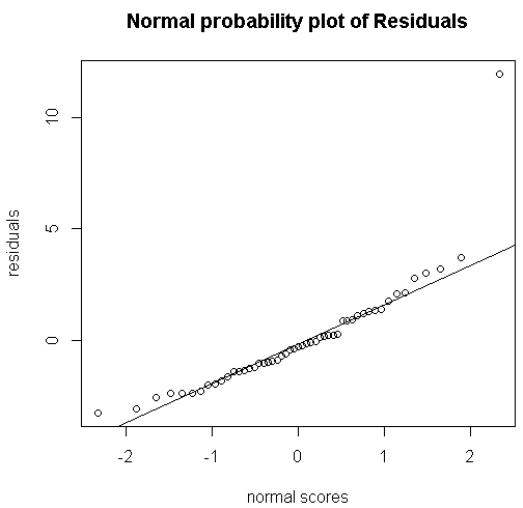





```
    Shapiro-Wilk normality test

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

Bad News Residuals (2)



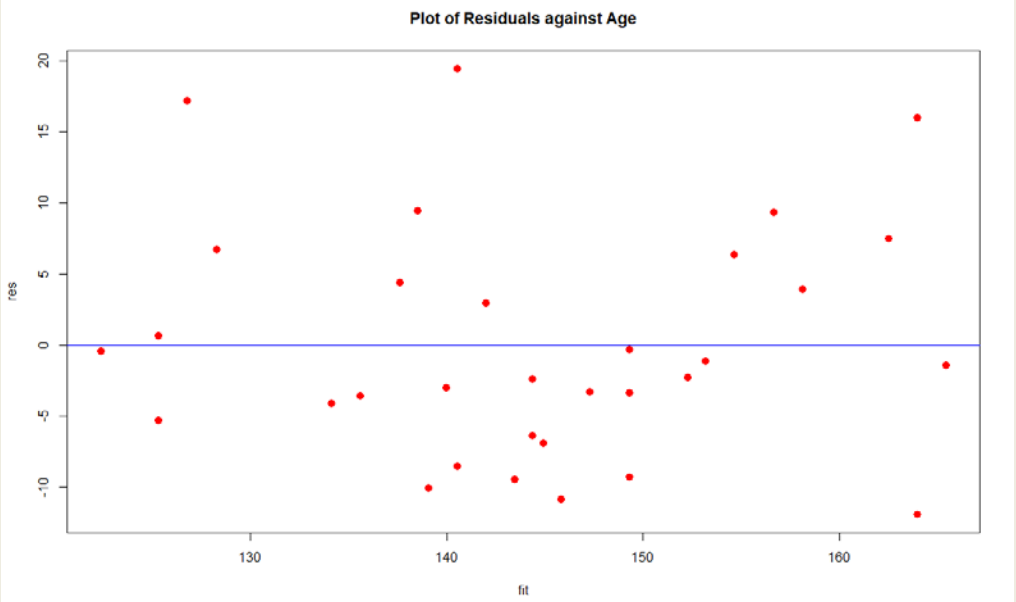



```
    Shapiro-Wilk normality test
data:  res

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 $\hat{Y}$.

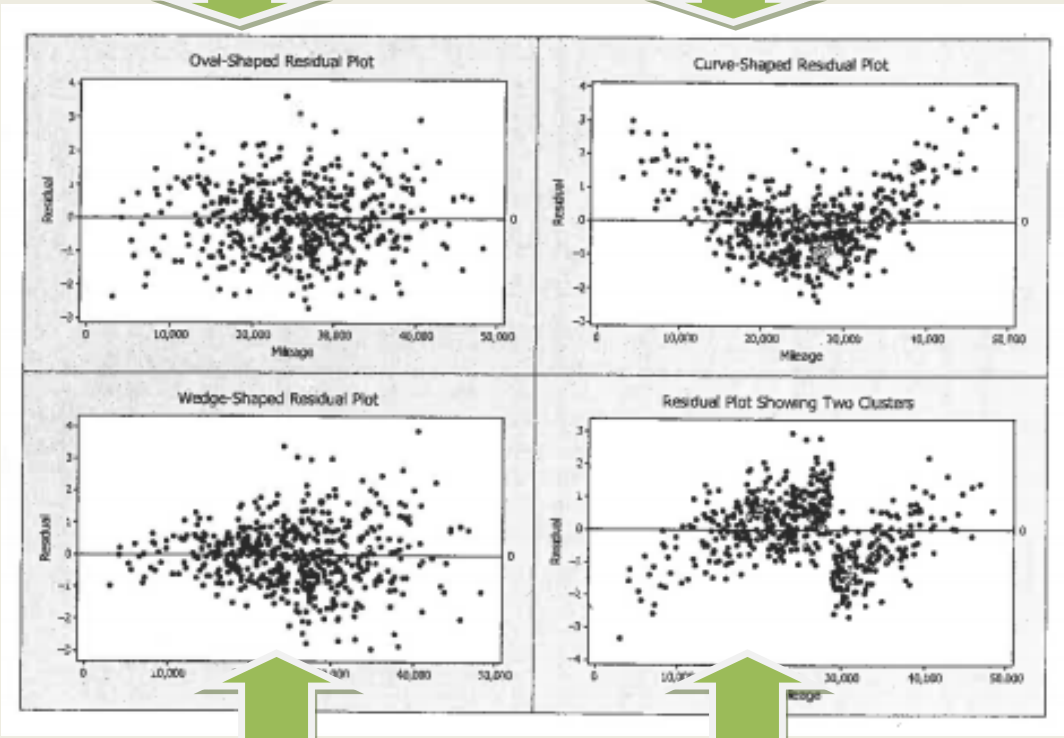
```
plot(res ~ fit,
 main = "Plot of Residuals against Age",
 col = "red",
 pch = 19,
 cex = 1.3)
abline(h = 0, col = "blue")
```



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.

8. Multiple Regression as a Tool for Adjusting for the Effect of Confounding Variables: Alzheimer's Example

Column		Count	Name	
	C1	509	Survival	Survival Time (in months)
	C2	509	AAO	Age at onset
	C3	509	Education	Years of Education
	C4	509	Sex	F = 1, M = 0
T	C5	509	Sex	F,M

Pat	Survival	AAO	Education	Sex	
1	55.8	81.9	12	1	F
2	137.1	60.1	10	0	M
3	31.1	86.9	2	1	F
4	84.3	83.7	5	0	M
5	52.1	86.4	15	0	M
6	62.9	73.0	15	1	F
7	47.2	77.4	5	0	M
8	56.6	96.8	11	0	M
9	56.8	93.2	10	1	F
10	64.3	75.3	8	0	M
11	11.0	77.2	7	1	F
12	233.1	65.6	8	0	M
13	72.6	69.4	3	1	F
14	4.1	88.4	13	0	M
15	102.4	82.3	12	1	F
16	31.7	90.7	9	0	M
:	:	:	:	:	
:	:	:	:	:	
503	69.3	75.4	6	1	F
504	75.4	67.8	8	0	M
505	23.8	92.4	10	1	F
506	148.1	63.8	16	0	M
507	85.1	84.8	5	0	M
508	47.0	82.4	8	0	M
509	14.6	82.6	5	1	F

How does Survival time vary with Sex?

The unadjusted effect of sex on survival time is 7.61 months.

(i)

Variable	Sex	N	Mean	StDev
Survival	F (1)	164	70.36	52.36
	M (0)	345	77.97	48.43

			-7.61	

(ii)

The regression equation is

Survival = 77.97 **- 7.61** Sex

```
Model5 <- lm(Survival ~ Sex, Alzheimers)
Model5

Call:
lm(formula = Survival ~ Sex)

Coefficients:
(Intercept)          Sex
     77.973       -7.611
```

But are we comparing like with like? This is an observational study and there is every likelihood that the characteristics of male and female Alzheimer's patients differ. Do they differ by Education level and/or Age at Onset (AAO)?

A confounding variable is one that is related to to both the explanatory variable (X, Sex) and the response variable (Y, Survival time).

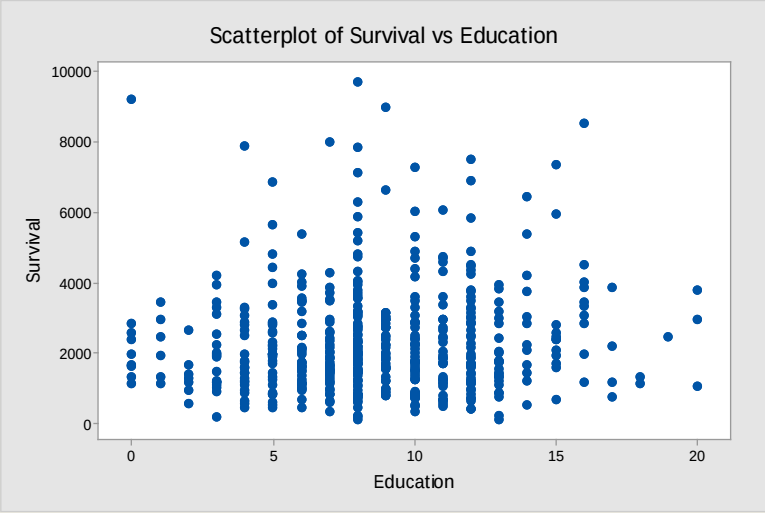
Education level

(a) Is Education level related to Sex?

Variable	Sex	N	Mean	StDev	Median
Education	M	345	8.899	3.446	8.000
	F	164	8.091	4.155	8.000

0.808 years

(b) Is Survival Status related to Education level?



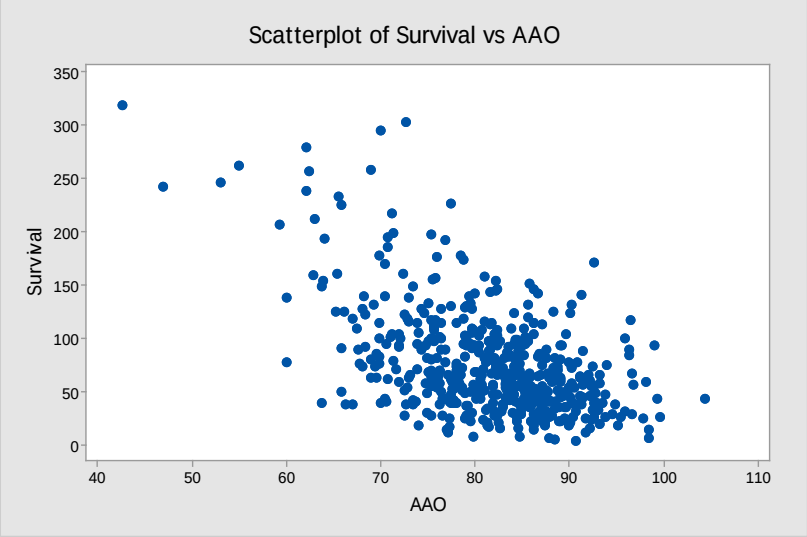

```
cor(Survival, Education)
[1] 0.08089806
```

(a) Is Age at Onset related to Sex?

Variable	Sex	N	Mean	StDev	Median
AAO	M	345	82.564	7.839	83.500
	F	164	79.453	8.948	80.200

3.111 years

(b) Is Survival Status related to Age at Onset?




```
cor(Survival, AAO)
[1] -0.5364428
```

We can remove/adjust for the effect of Age at Onset by obtaining a Regression of Survival Time on Sex and Age at Onset

$\widehat{\text{Survival}}$ = 357.4 **- 18.14** Sex - 3.385 AAO

After adjusting for Age at Onset, the impact of being a female (rather than a male) on survival time is -18.1 months.

We can remove/adjust for the effect of both Age at Onset and Education by obtaining a Regression of Survival Time on Sex, Age at Onset, and Education.

$\widehat{\text{Survival}}$ = 351.6 **- 17.68** Sex - 3.370 AAO + 0.514 Education

After adjusting for Age at Onset and Education, the impact of being a female (rather than a male) on survival time is -17.7 months.