Multiple Regression Models

Statistics 104

Autumn 2004

Multiple Regression Models

The multiple regression model can be used to describe a wide range of models including

- Curved response functions
- Categorical predictors
- Interactions - the effect of one predictor variable depends on the level of another predictor variables
- Non-normal responses, multiplicative errors, etc

These models can be fit into the general framework by adding adding functions of the predictor variables to the model, such as

$$x^2, x^3, \log x, x_1 x_2, I(x = man), I(x = woman)$$

or by looking at transformations of the response variable, such as $\frac{100}{MPG}$ (as in the fuel use example) or $\log y$.

Polynomial Regression

One approach for describing curved relationships

Include terms like x^2 and x^3 in the model

General Polynomial Regression Model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_p x_i^p + \epsilon_i$$

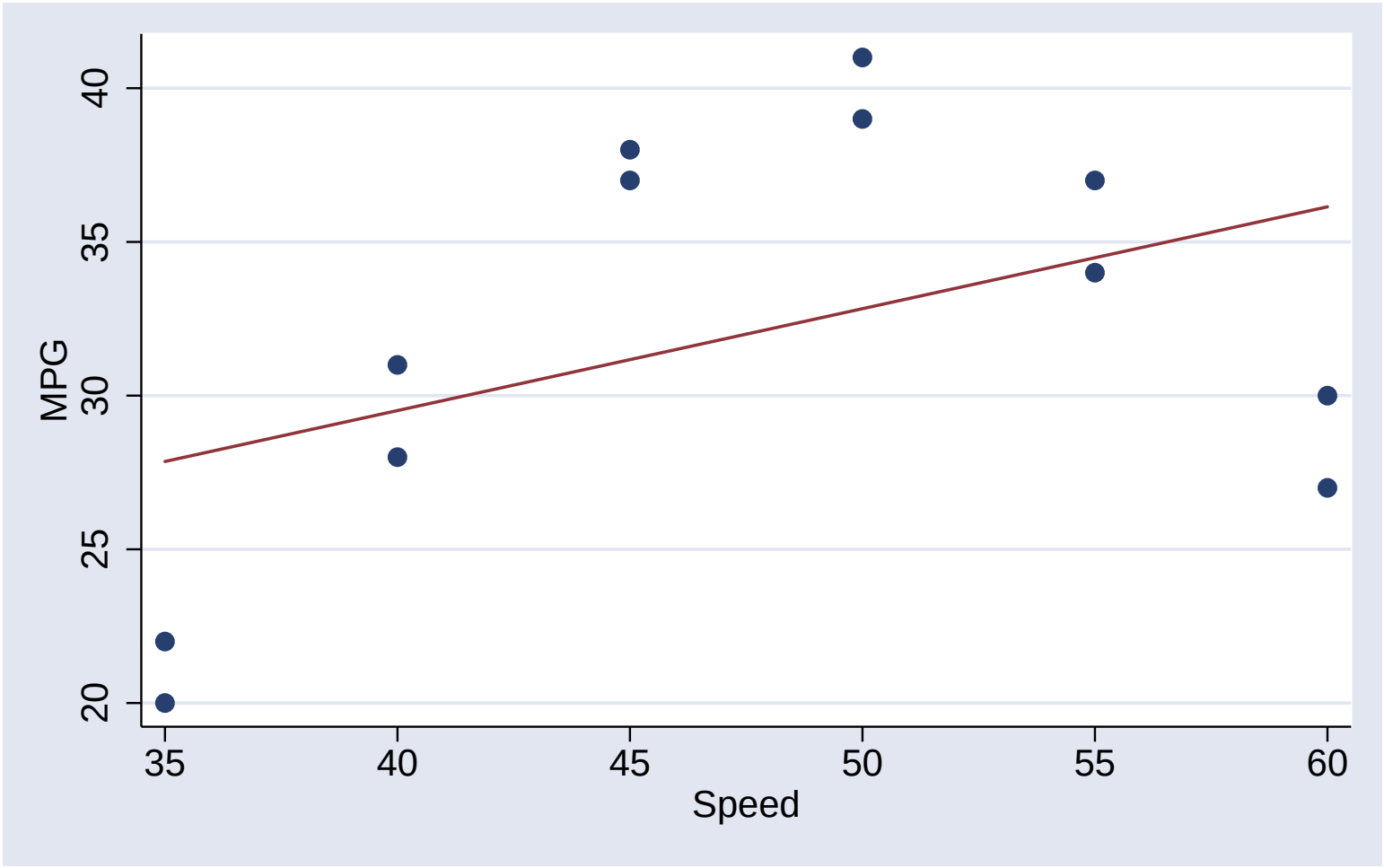
where $\epsilon_i \sim N(0, \sigma)$.

Example: Speed vs MPG

The effectiveness of a new experimental overdrive gear in reducing gasoline consumption was studied in 12 trials with a light truck equipped with this gear.

$y =$ MPG

$x =$ Truck Speed



Fairly obvious that a straight line model is a poor description of the relationship between Speed and MPG.

Lets try the quadratic model

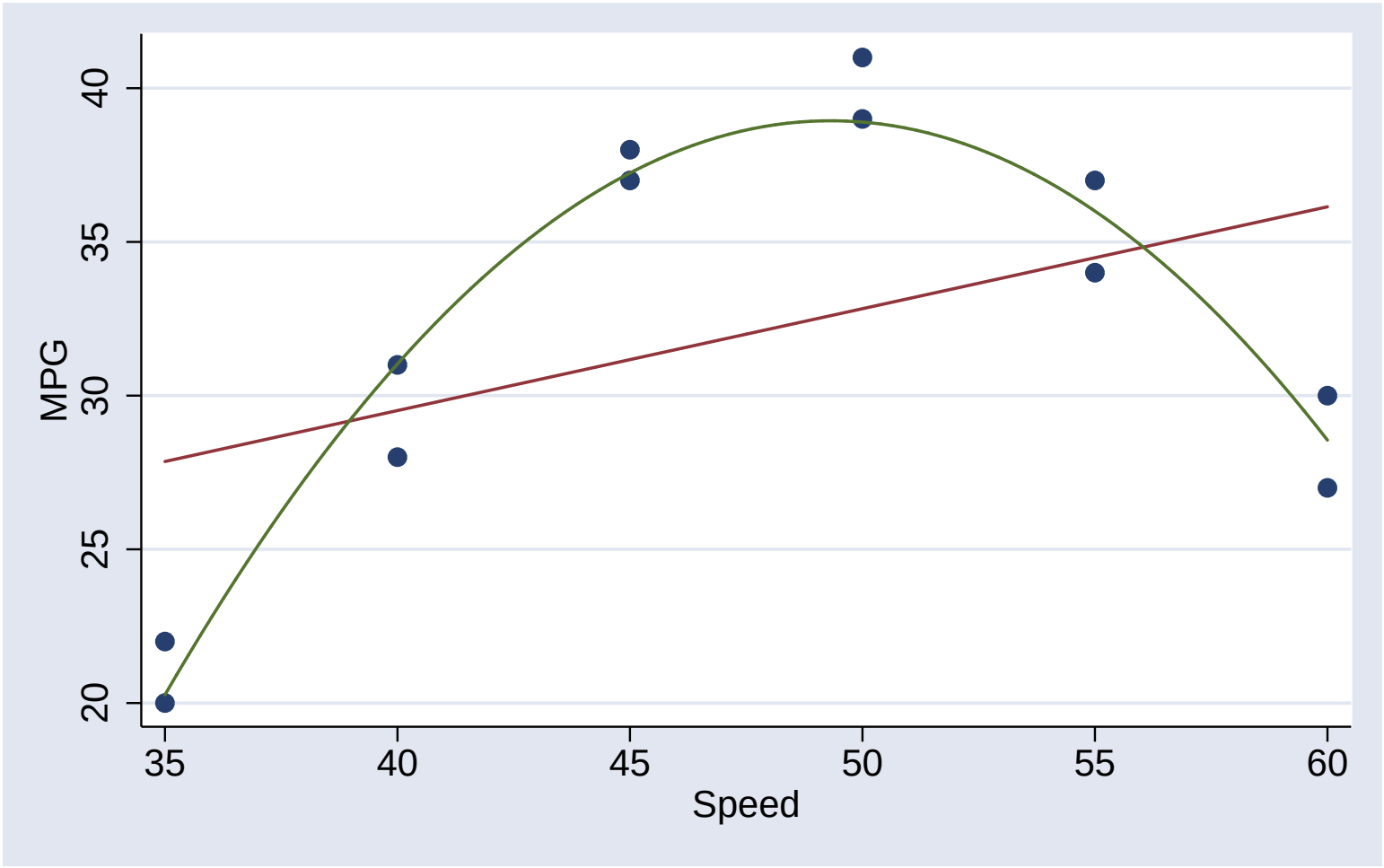
$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$$

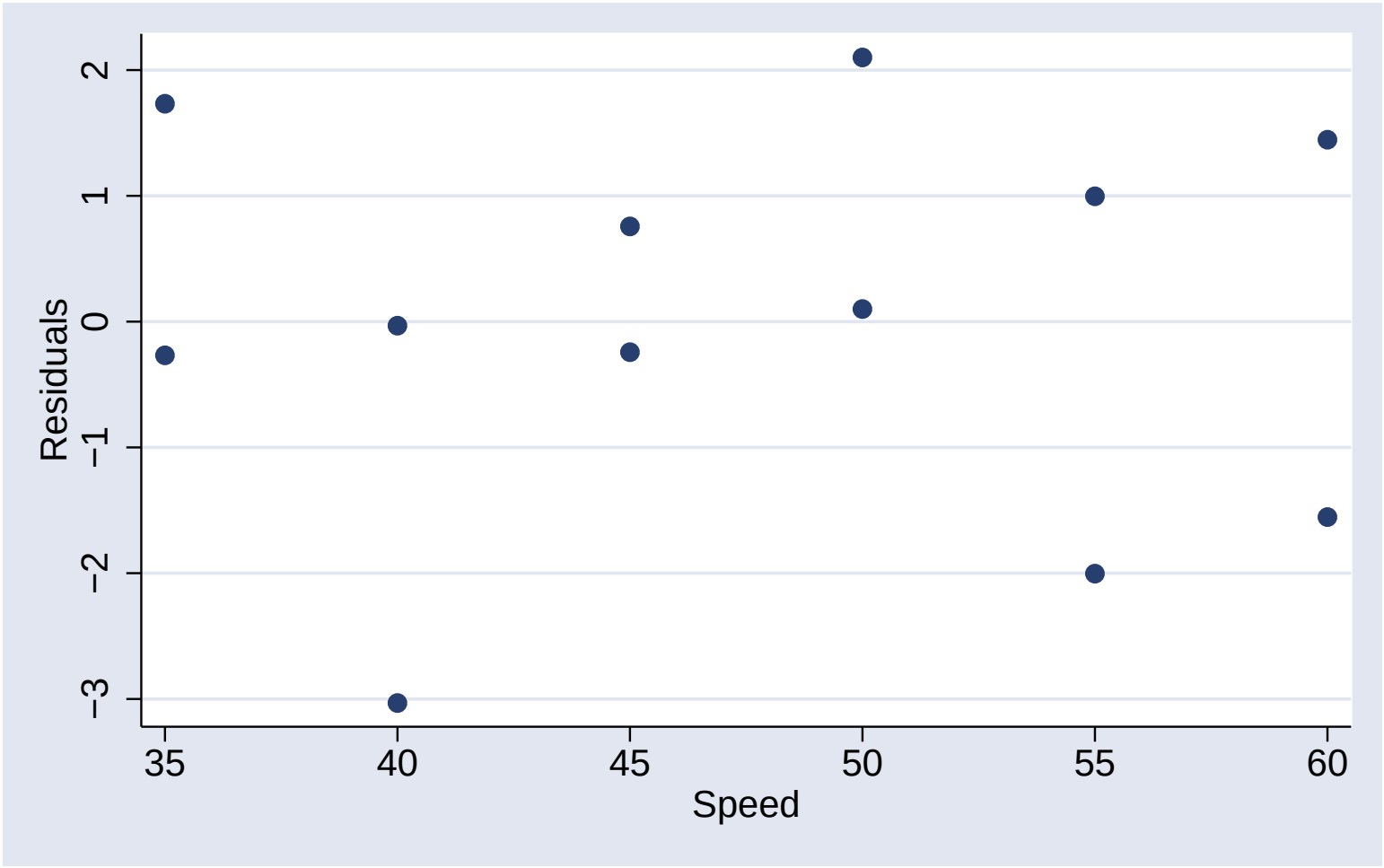
For polynomial models, the interpretation of the βs isn't as nice.

For example, in the quadratic model, increasing x by 1, leads to an expected change in y of

$$\beta_1 + \beta_2(2x + 1)$$

So the change depends on the level of x. Can't keep x^2 fixed when changing x.


MPG
20
25
30
35
40
35
40
45
50
55
60
Speed



The fit of the model seems reasonable. The fitted curve seems to follow the basic pattern in the data and nothing stands out strongly in the residual plot.

Fitting polynomial regression models in Stata

As a polynomial regression model is a special case of the multiple linear regression model, we can do it basically the same way. However you need to need to create the transformed variables. For example,

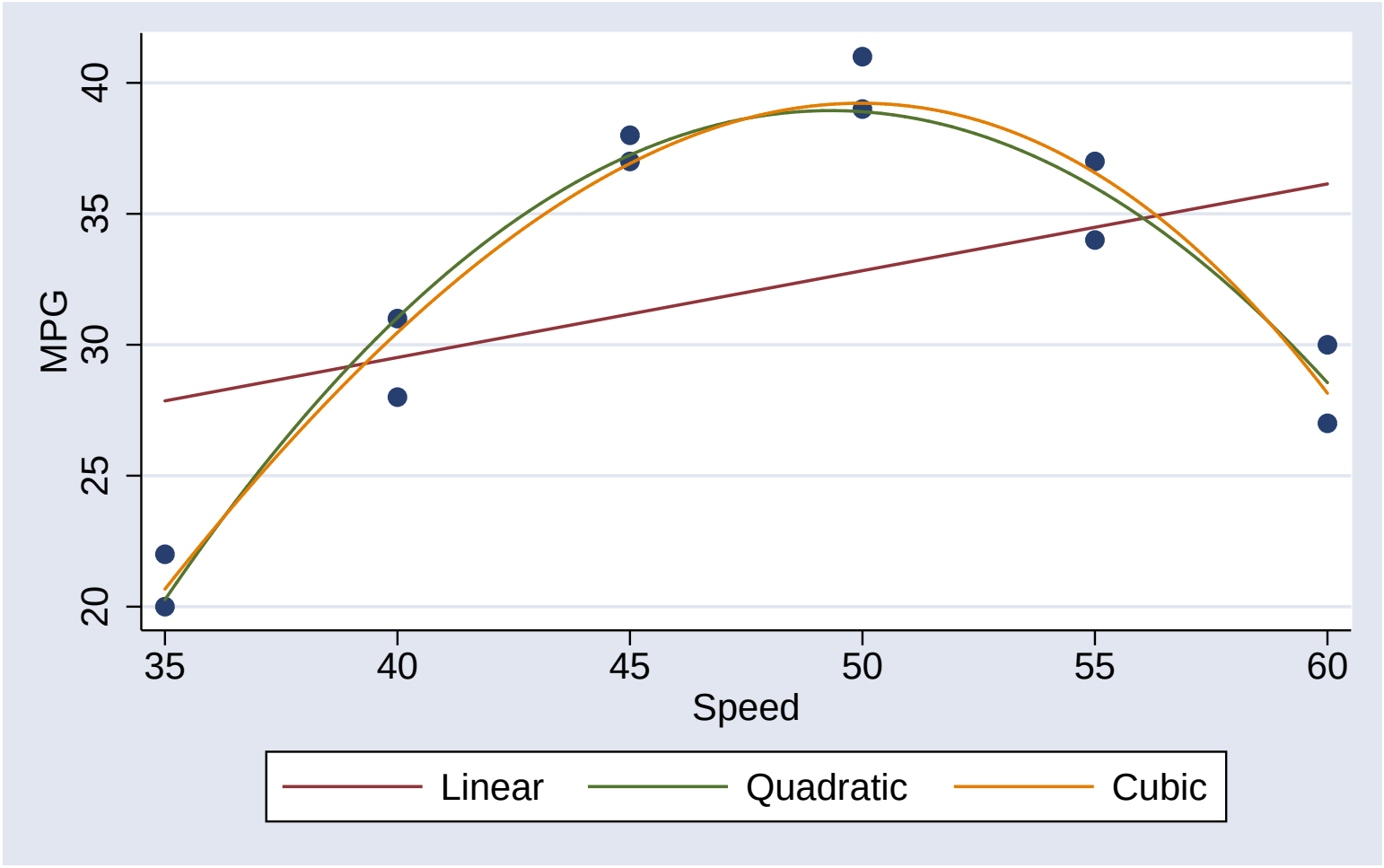
```
. gen speed2 = speed*speed
. regress mpg speed speed2
```

```
  Source |       SS       df       MS              Number of obs =      12
---------+------------------------------           F(  2,     9) =   81.03
   Model |  483.167857     2  241.583929           Prob > F      =  0.0000
Residual |  26.8321429     9  2.98134921           R-squared     =  0.9474
---------+------------------------------           Adj R-squared =  0.9357
   Total |         510    11  46.3636364           Root MSE      =  1.7267

------------------------------------------------------------------------------
     mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
   speed |   8.983214   .7615635    11.80   0.000     7.260438    10.70599
  speed2 |  -.0910714   .0079929   -11.39   0.000    -.1091526   -.0729903
   _cons |  -182.5821   17.67703   -10.33   0.000    -222.5704   -142.5939
------------------------------------------------------------------------------
```

The t test for the Speed^2 term is significant, implying that adding this term improves the fit. Lets see if the cubic model does even better.

```
. regress mpg speed speed2 speed3

  Source |       SS       df       MS              Number of obs =      12
---------+------------------------------           F(  3,     8) =   52.85
   Model |  485.503968     3  161.834656           Prob > F      =  0.0000
Residual |  24.4960317     8  3.06200397           R-squared     =  0.9520
---------+------------------------------           Adj R-squared =  0.9340
   Total |         510    11  46.3636364           Root MSE      =  1.7499

------------------------------------------------------------------------------
     mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
   speed |   1.848677   8.204499     0.23   0.827    -17.07093    20.76828
  speed2 |   .0619841   .1754158     0.35   0.733    -.3425255    .4664937
  speed3 |  -.0010741   .0012297    -0.87   0.408    -.0039097    .0017616
   _cons |   -73.9127   125.6955    -0.59   0.573    -363.7671    215.9417
------------------------------------------------------------------------------
```

The t test for the Speed^3 term is not significant, implying that adding this term doesn't improve the fit.



Also notice that R^2 barely increases when the cubic term is added (0.9520 vs 0.9474). However going from linear to quadratic, the increase is from 0.1885 to 0.9474.

When building polynomial models, usually you want to keep the order low (i.e. $p = 2$ or 3). Letting p get too large will often give very squiggly fitted curves.

Also you usually don't want to skip terms (i.e. don't use $y_i = \beta_0 + \beta_2 x_i^2 + \epsilon_i$). This can lead to poor fits and poor predictions.

Indicator Variables

When dealing with categorical factors, you can't directly put them into a regression model - what does $2 \times$ man or $4 \times$ woman mean?

Categorical factors can be useful in describing relationships.

An approach for dealing with them is through the use of indicator variables (sometimes called dummy variables).

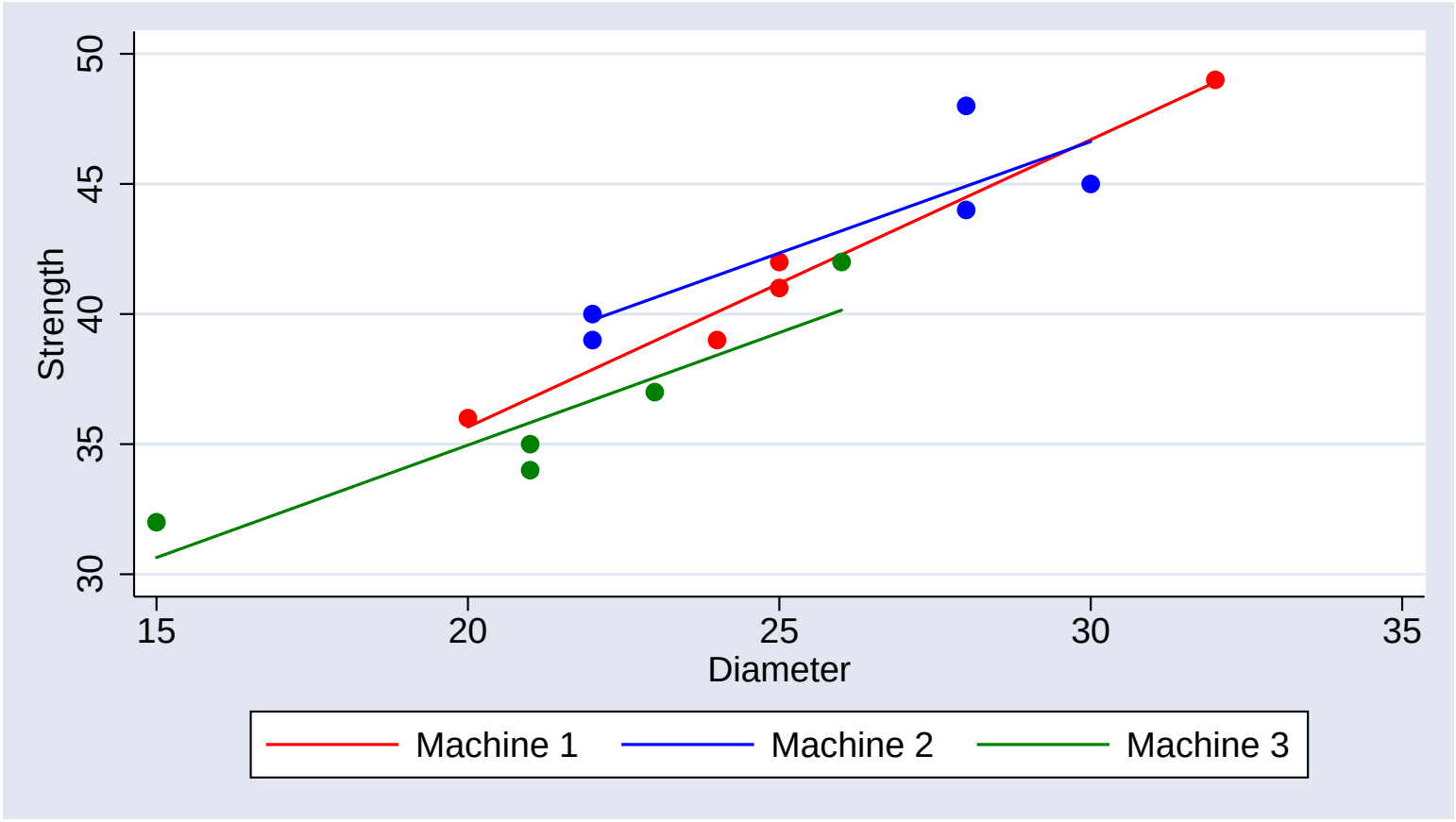
Example: Fiber Strength

Three different machines produce a monofilament fiber for a textile company. The process engineer is interested in determining if their is a difference in the breaking strength of the fiber produced by the three machines. However the strength of the fiber is related to its diameter, with thicker fibers being generally stronger than thinner ones. A random sample of five fiber specimens is selected from each machine.

y = Fiber Strength (in pounds)

x = Fiber Diameter (in 1/1000 of an inch)

z = Machine (1, 2, or 3)



It appears that the 3 machines might be different. The fitted line for machine 3 is lower than the rest and the fitted line for machine 1 is steeper than the rest.

However there isn't much data, so the differences might be consistent with random variation.

Indicator variables

Take values either 0 or 1, depending on the level of the factor.

There is one indicator variable for each level of the factor.

For the example

$$m_1 = \begin{cases} 1 & \text{if machine 1} \\ 0 & \text{if machine 2 or 3} \end{cases}$$

$$m_2 = \begin{cases} 1 & \text{if machine 2} \\ 0 & \text{if machine 1 or 3} \end{cases}$$

$$m_3 = \begin{cases} 1 & \text{if machine 3} \\ 0 & \text{if machine 1 or 1} \end{cases}$$

When using indicator variables, you need to omit one of them. You could fit a model like

$$y_i = \beta_0 + \beta_1 m_{i1} + \beta_2 m_{i2} + \epsilon_i$$

What do we get for each machine?

- Machine 1

$$y_i = \beta_0 + \beta_1 1 + \beta_2 0 + \epsilon_i = (\beta_0 + \beta_1) + \epsilon_i$$

- Machine 2

$$y_i = \beta_0 + \beta_1 0 + \beta_2 1 + \epsilon_i = (\beta_0 + \beta_2) + \epsilon_i$$

- Machine 3

$$y_i = \beta_0 + \beta_1 0 + \beta_2 0 + \epsilon_i = \beta_0 + \epsilon_i$$

This gives a different mean level for each machine with the same standard deviation.

(Note: This is a different way of describing the one-way ANOVA model, which will be discussed in the future.)

Note that this is probably not a reasonable description of the data, as it doesn't include the diameter of the fiber, which from the plot appears to be important.

A possibly better model might be

$$y_i = \beta_0 + \beta_1 m_{i1} + \beta_2 m_{i2} + \beta_3 x_i + \epsilon_i$$

For each machine we get

- Machine 1

$$y_i = \beta_0 + \beta_1 1 + \beta_2 0 + \beta_3 x_i + \epsilon_i = (\beta_0 + \beta_1) + \beta_3 x_i + \epsilon_i$$

- Machine 2

$$y_i = \beta_0 + \beta_1 0 + \beta_2 1 + \beta_3 x_i + \epsilon_i = (\beta_0 + \beta_2) + \beta_3 x_i + \epsilon_i$$

- Machine 3

$$y_i = \beta_0 + \beta_1 0 + \beta_2 0 + \beta_3 x_i + \epsilon_i = \beta_0 + \beta_3 x_i + \epsilon_i$$

This model gives 3 different lines, but all with the same slope (parallel lines).

Interaction Models

We may want to use a model with possibly different slopes as well. This can be done with interaction terms.

Can use the product of two variables as terms in the model like

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$$

For the fiber example, could try a model something like

$$y_i = \beta_0 + \beta_1 m_{i1} + \beta_2 m_{i2} + \beta_3 x_i + \beta_4 x_i m_{i1} + \beta_5 x_i m_{i2} + \epsilon_i$$

For each machine we get

- Machine 1

$$y_i = \beta_0 + \beta_1 1 + \beta_2 0 + \beta_3 x_i + \beta_4 x_i 1 + \beta_5 x_i 0 + \epsilon_i = (\beta_0 + \beta_1) + (\beta_3 + \beta_4) x_i + \epsilon_i$$

- Machine 2

$$y_i = \beta_0 + \beta_1 0 + \beta_2 1 + \beta_3 x_i + \beta_4 x_i 0 + \beta_5 x_i 1 + \epsilon_i = (\beta_0 + \beta_2) + (\beta_3 + \beta_5) x_i + \epsilon_i$$

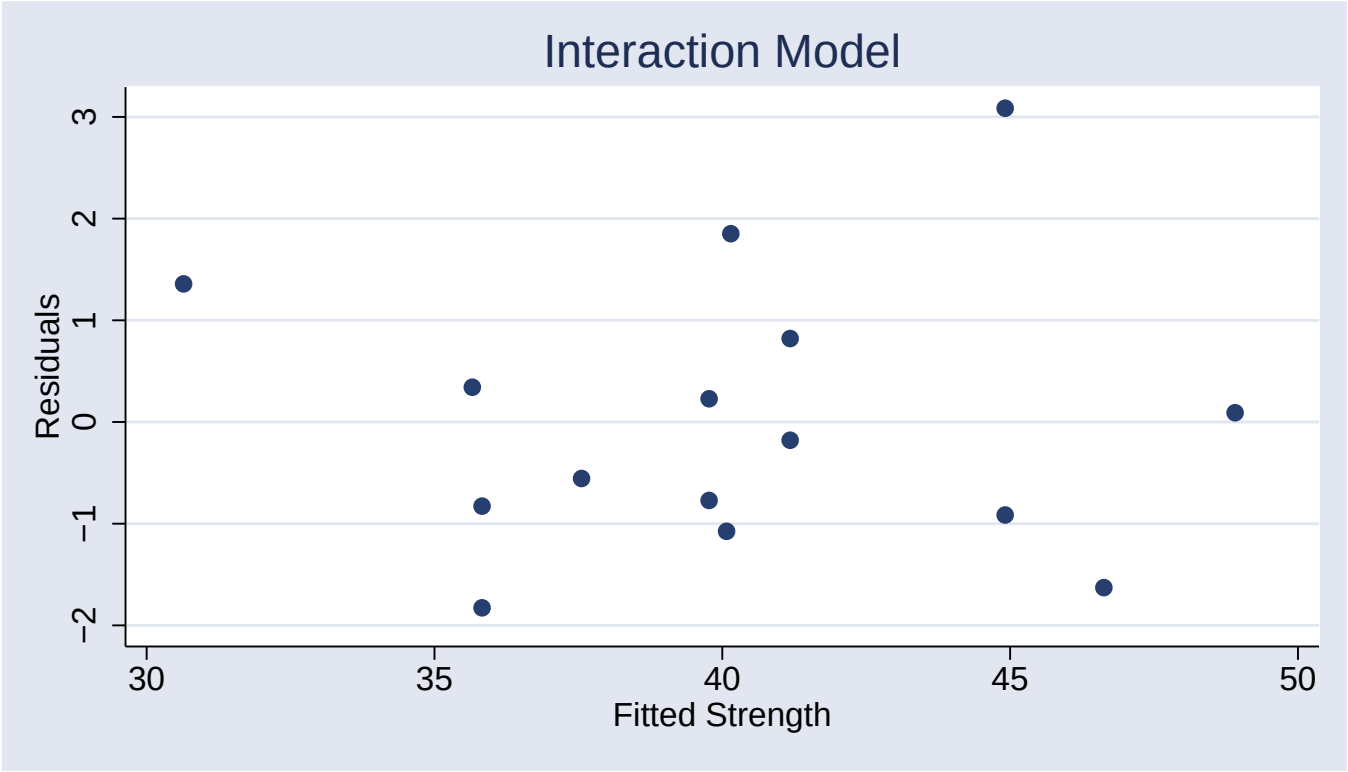
- Machine 3

$$y_i = \beta_0 + \beta_1 0 + \beta_2 0 + \beta_3 x_i + \beta_4 x_i 0 + \beta_5 x_i 0 + \epsilon_i = \beta_0 + \beta_3 x_i + \epsilon_i$$

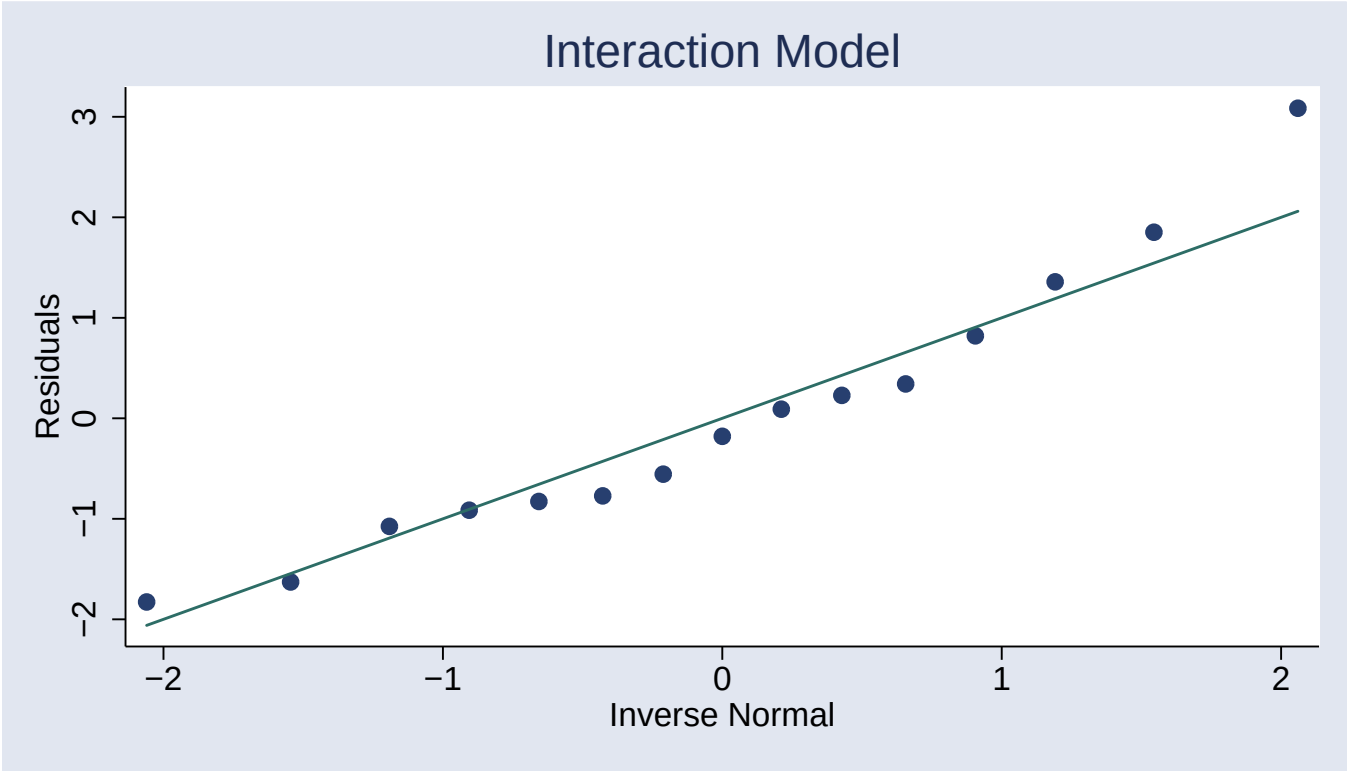
This gives us want we want, three different lines, one for each machine.

This is an example of an interaction, where the effect of one variable depends on (is influenced by) another variable. In this case, the effect of diameter on fiber strength depends on which machine the fiber is sampled from.

```
. regress strength diameter mach1 mach2 dmach1 dmach2

  Source |       SS       df       MS              Number of obs =      15
---------+------------------------------           F(  5,     9) =   22.90
   Model |  321.151288     5  64.2302576           Prob > F      =  0.0001
Residual |  25.2487121     9  2.80541246           R-squared     =  0.9271
---------+------------------------------           Adj R-squared =  0.8866
   Total |       346.4    14  24.7428571           Root MSE      =  1.6749

------------------------------------------------------------------------------
strength |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
diameter |   .8641975   .2080707     4.15   0.002      .393509    1.334886
   mach1 |   -4.10682   6.663141    -0.62   0.553    -19.17989    10.96625
   mach2 |   3.235273   7.378705     0.44   0.671    -13.45652    19.92706
  dmach1 |   .2400805   .2842515     0.84   0.420     -.402941    .8831021
  dmach2 |  -.0070547   .3055979    -0.02   0.982    -.6983651    .6842557
   _cons |   17.67901   4.474245     3.95   0.003     7.557567    27.80046
------------------------------------------------------------------------------
```



There doesn't seem to be a pattern that stands out here, though there is one fairly large residual, though it is less than $2MSE$.



Fairly straight, so the normality assumption doesn't seem to be violated.

Testing whether a set of parameters are all zero simultaneously

For example it would be nice to test whether both interaction parameters are both 0 (which corresponds to the parallel line model).

$$H_0 : \beta_4 = \beta_5 = 0 \qquad \text{vs} \qquad H_A : \beta_4 \neq 0 \text{ or } \beta_5 \neq 0 \text{ or both } \neq 0$$

- Full model (H_A):

$$y_i = \beta_0 + \beta_1 m_{i1} + \beta_2 m_{i2} + \beta_3 x_i + \beta_4 x_i m_{i1} + \beta_5 x_i m_{i2} + \epsilon_i$$

- Reduced model (for example $H_0 : \beta_4 = \beta_5 = 0$):

$$y_i = \beta_0 + \beta_1 m_{i1} + \beta_2 m_{i2} + \beta_3 x_i + \epsilon_i$$

These two models can be compared with an F test which measures how much the fit of the Full model is better than the fit of the Reduced model. This can be done in Stata with the `testparm` command.

```
. testparm dmach1 dmach2

( 1)  dmach1 = 0
( 2)  dmach2 = 0

      F(  2,     9) =    0.49
           Prob > F =    0.6293
```

Since the p-value is large (F is small), we don't want to reject the null hypothesis. The data appears consistent with a constant slope model.

In this F test, the numerator degrees of freedom is the number of βs being examined and the denominator degrees of freedom is the error df for the Full model.

```
. regress strength diameter mach1 mach2

  Source |       SS       df       MS              Number of obs =      15
---------+------------------------------           F(  3,    11) =   41.72
   Model |   318.41411     3  106.138037           Prob > F      =  0.0000
Residual |  27.9858896    11  2.54417178           R-squared     =  0.9192
---------+------------------------------           Adj R-squared =  0.8972
   Total |       346.4    14  24.7428571           Root MSE      =   1.595

------------------------------------------------------------------------------
strength |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
diameter |   .9539877   .1140483     8.36   0.000     .7029691    1.205006
   mach1 |   1.584049    1.10715     1.43   0.180    -.8527714     4.02087
   mach2 |   2.620859   1.147759     2.28   0.043     .0946588    5.147059
   _cons |   15.77546   2.520854     6.26   0.000      10.2271    21.32382
------------------------------------------------------------------------------
```

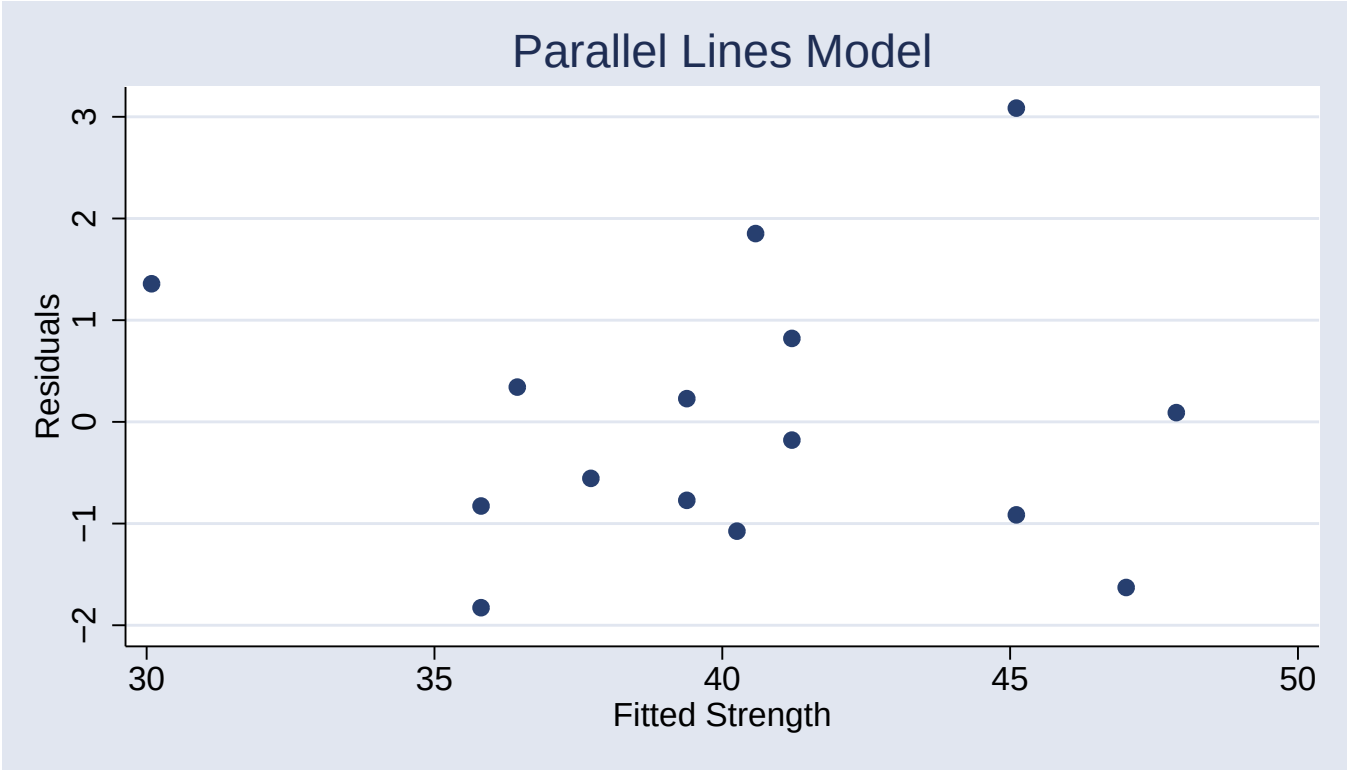
```
. testparm mach1 mach2

( 1)  mach1 = 0
( 2)  mach2 = 0

      F(  2,    11) =    2.61
           Prob > F =    0.1181
```

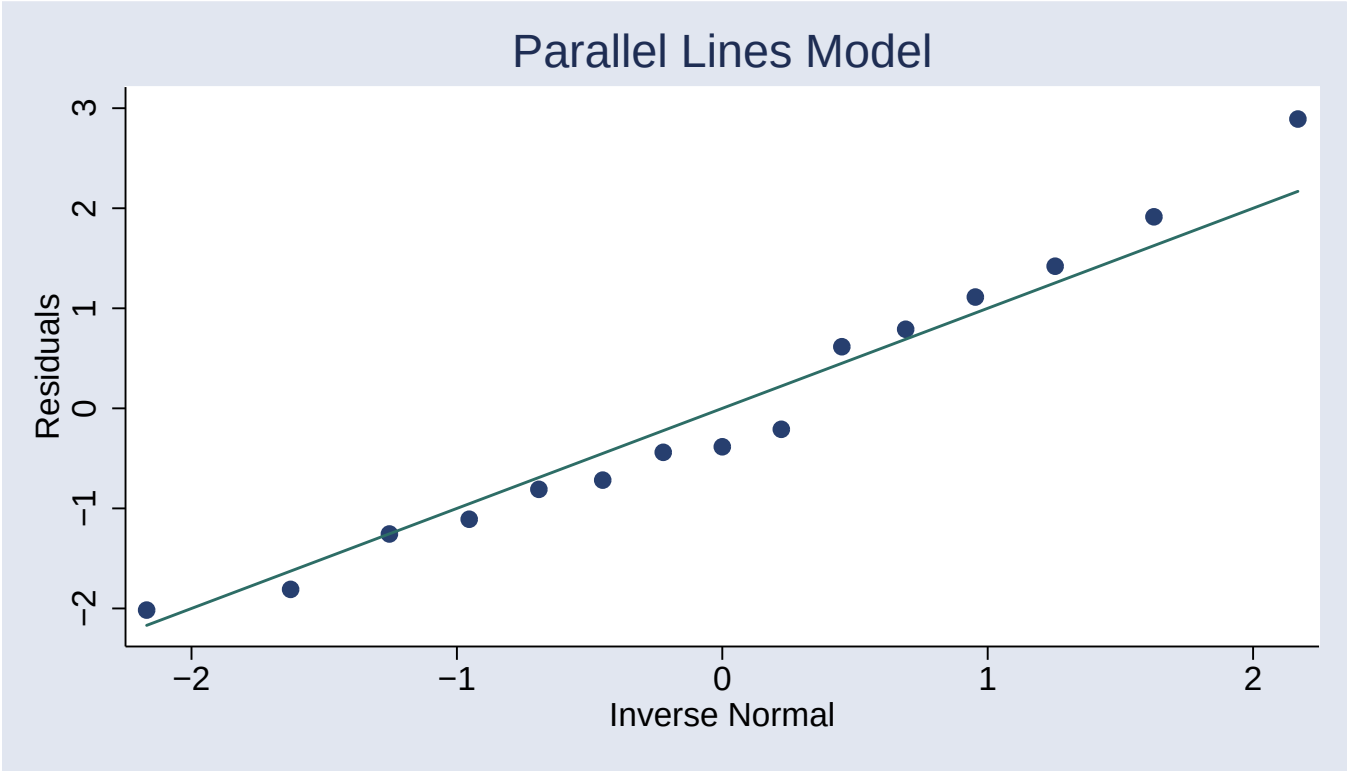
Based on this F test, it appears that the data is also consistent with all the intercepts being the same, i.e. the relationship between strength and diameter is the same for all three machines.

However this isn't completely clear as the t test for β_2 has a p-value of 0.043 which is suggestive. This suggests that the intercept for machine 2 is different from the intercept for machine 3. β_2 can also be thought of as the expected difference in response for machine 2 vs machine 3 for any given diameter.

Though we need to adjust for multiple comparisons since we have two βs for describing the machine effects. The Bonferroni correction states we should compare the p-value with $\frac{\alpha}{2}$ in this case.

The graphical summary on the residuals suggests that there are no serious deviations from the parallel lines model.




Parallel Lines Model
Residuals
−2
−1
0
1
2
3
−2
−1
0
1
2
Inverse Normal