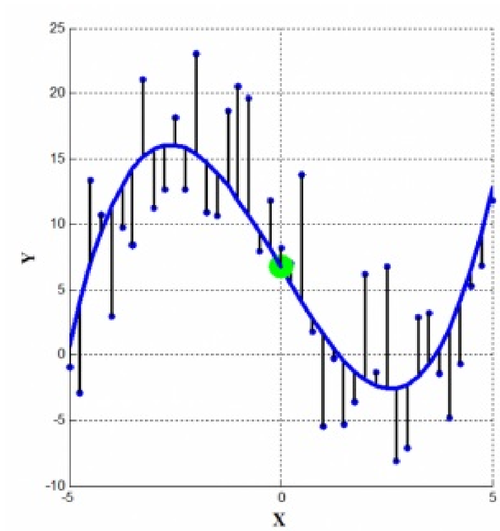
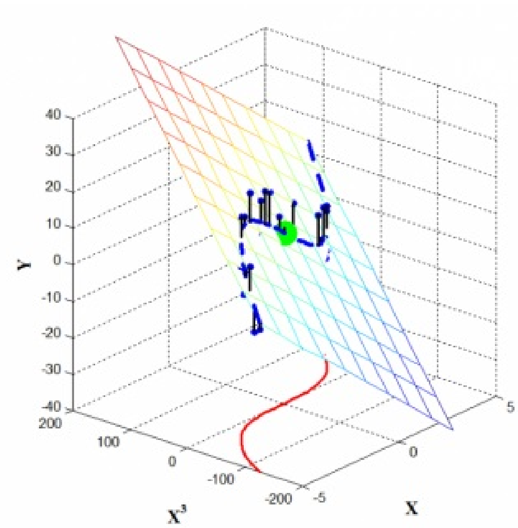
MAT 3375 - Regression Analysis
Polynomial Regression Models and the Indicator function

Professor: Termeh Kousha

Fall 2017





1 Polynomial Regression Models

Polynomials are widely used when the response is curvilinear. For example, consider the second-order model with interaction

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

It can be written as a multiple linear regression model as below

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \epsilon$$

where $\beta_3 = \beta_{11}$,$\beta_4 = \beta_{22}$,$\beta_5 = \beta_{12}$,$x_3 = x_1^2$,$x_4 = x_2^2$ and $x_5 = x_1 x_2$.

2 Polynomial Models in One Variable

2.1 Basic Properties

In general, the kth-order polynomial model in one variable is

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \epsilon \tag{1}$$

if we set $x_j = x^j, j = 1, \cdots, k$, then the equation (1) becomes a multiple linear regression model with k regressors.

The expected value of y is

$$E(y) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k$$

which is a kth-order function. We call β_1 the linear effect parameter, β_2 the quadratic effect parameter and so on. If the range of the data includes $x = 0$, then the parameter β_0 is the mean of y when $x = 0$, otherwise it has no physical interpretation.

Polynomial models are practical in many situations because an arbitrary function can be approached by a polynomial function in small neighborhood, which can be considered as the Taylor series expansion of the true function.

When fitting a polynomial model in one variable, we should take several points into consideration.

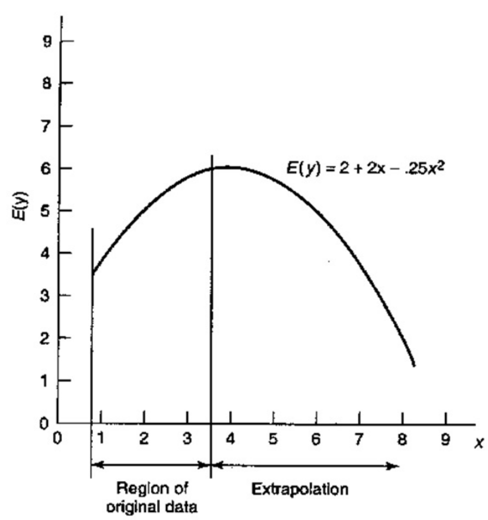
1. Order of the Model
   The order of the model should be as low as possible. When the response appears to be curvilinear, transformations should be considered

first to keep the model first order. If this method fails, then we can try a second-order polynomial. It is important that the higher-order polynomials ($k > 2$) should be avoided unless they can be justified for reasons outside the data.

2. Model-Building Strategies

   - **Forward Selection** Fitting models by increasing order until the t test for the highest order term is insignificant.
   - **Back Elimination** Fitting models with highest order and then delete terms from the highest order term until the t test for the highest order remaining term is significant.

3. Extrapolation
   Extrapolation is used to predict the future based on the trend of the model. However it should satisfy two assumptions that the trend should be continuous without jump and the change of the response in the future is similar to that at present. If the curve of the polynomial model is different in the range of extrapolation and in the range of the original data, it will be extremely hazardous. Generally speaking, polynomial function can just be used to describe the relationship between response and regressors in the data range and the extrapolation is not reliable.



4. Ill-Condition I
   We have known that the normal equations in the least square can be written as
   $$X'y = X'X\hat{\beta}$$
   and the estimated least-squares estimator β is
   $$\hat{\beta} = (X'X)^{-1}X'y$$
   the condition of obtaining the estimated value is the existence of $(X'X)^{-1}$. If the $X'X$ is nearly singular or ill-conditioned, in the other word, the determinant of $X'X$ is nearly to be zero $(\det(X'X) \approx 0)$, then the $X'X$ matrix is called as ill-conditioned matrix because a small change in one of the values of the coefficient matrix will cause large change in the solution. For example,

   In addition, if the first order variables are not centered, then the cross terms and the power polynomial terms are highly correlated with the variables of which they are comprised (R^2 between X and X^2 can approach 1.0). The problems of multicollinearity produced by noncentered variables are referred as nonessential ill-conditioning.

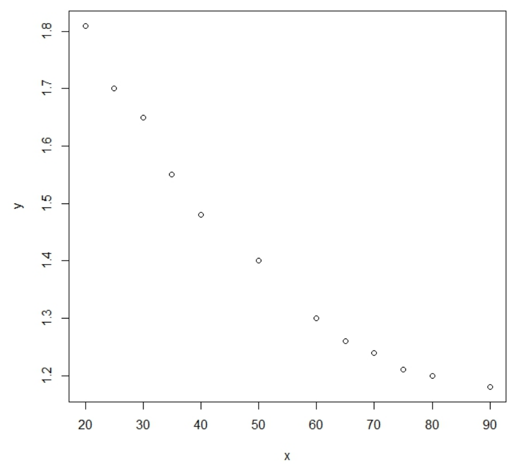
   To lessen the multicollinearity, we can first centralize the regressor variable $(x - \bar{x})$ to remove nonessential ill-conditioning. It may still result in correlations between certain regression coefficients, and then we can deal with the problem with constructing orthogonal polynomial.

5. Ill-Conditioning II
   If the values of x are limited to a narrow range, these columns in X may cause significant ill conditioning or multicollinearity.

6. Hierarchy
   A regression model is said to be hierarchial when it concludes all terms of the highest order and lower. Peixoto[1987,1990] points out that only hierarchical models are invariant under transformation. However, some terms of model may be unnecessary so checking the significant of all the terms are more important than just follow the rules.

Example 1. Sidewall panels for interior of an airplane are formed in 1500-ton press. The unit manufacturing cost varies with the production lot size. The data shown below give the average cost per unit (in hundreds of dollars) for this product (y) and the production lot size (x). The scatter diagram, indicates that a second order polynomial may be appropriate.

Y	1.18	1.7	1.65	1.55	1.48	1.4	1.3	1.26	1.24	1.21	1.2	1.18
X	20	25	30	35	40	50	60	65	70	75	80	90

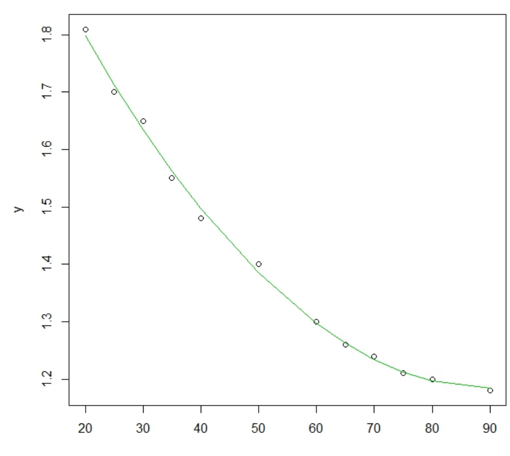


We will fit the model

$$Y = \beta_0 + \beta_1 x + \beta_{11} x^2 + \epsilon$$

The y vector, X matirx and the β vector are as follows:

```
Call:
lm(formula = y ~ x + I(x^2))

Residuals:
       Min         1Q     Median         3Q        Max
-0.0174763 -0.0065087  0.0001297  0.0071482  0.0151887

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.198e+00  2.255e-02   97.48 6.38e-15 ***
x           -2.252e-02  9.424e-04  -23.90 1.88e-09 ***
I(x^2)       1.251e-04  8.658e-06   14.45 1.56e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.01219 on 9 degrees of freedom
Multiple R-squared:  0.9975,    Adjusted R-squared:  0.9969
F-statistic:  1767 on 2 and 9 DF,  p-value: 2.096e-12
```



```
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x          1 0.49416 0.49416 3326.12 7.134e-13 ***
I(x^2)     1 0.03100 0.03100  208.67 1.564e-07 ***
Residuals  9 0.00134 0.00015
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Find the regression model. Can we drop the quadratic term form from the model?

2.2 Multicollinearity

In multiple regression problems, we expect to find dependencies between the response variable Y and the regressors x_j, In most regression problems, however, we find that there are also dependencies among the regressor variables x_j. In situations where the dependencies are strong, we say that multicollinearity exists. Multicollinearity can have serious effects on the estimates of the regression coefficients and on the general applicability of the estimated model. The effect of multicollinearity may be easily demonstrated. The diagonal elements of the matrix $C = (X'X)^{-1}$ can be written as

$$C_{jj} = \frac{1}{(1 - R_j^2)} \qquad j = 1, \cdots, k$$

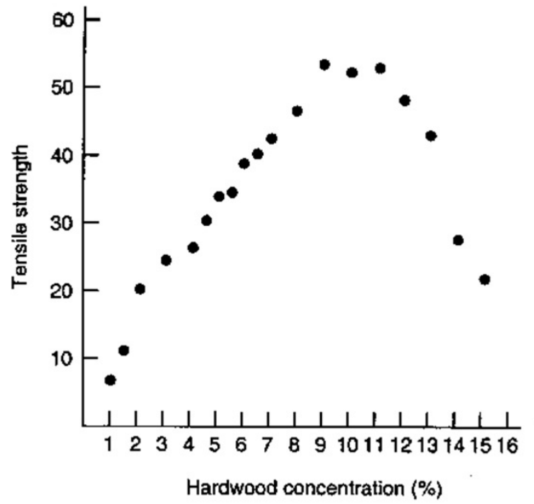
where R_j^2 is the coefficient of multiple determination resulting from regression x_j on the other k-1 regressor variables. Clearly the stronger the linear dependency of x_j on the remaining regressor variables, and hence the stronger the multicollinearity, the larger the value of R_j^2 will be. Recall that $Var(\hat{\beta}_j) = \sigma^2 C_{jj}$. Therefore, we say that the variance of $\hat{\beta}_j$ is inflated by the quantity $(1 - R_j^2)^{-1}$. Consequently we define the variance inflation factor for B_j as

$$\boxed{VIF(\beta_j) = \frac{1}{(1 - R_j^2)} \qquad j = 1, \cdots, k}$$

Theses factors are important measure of the extend to which multicollinearity is present. The larger the VIF, the more severe the multicollinearity . If VIF exceeds 10, multicollinearity is a problem. Some authors suggest VIF should not exceed 4 or 5.
The vif function is in package and calculate VIFs for the terms in the linear model:

```
library(car)
vif(lm)
```

Example 2 (The Hardwood data). Data concerning the strength of kraft paper and the percentage of hardwood in the batch of pulp from which the paper was produced. A scatter diagram of these data is shown below.



This display and knowledge of the production process suggests that a quadratic model may adequately describe the relationship.

```
lm<-lm(y~hardwood$x+I((hardwood$x)^2),data=hardwood)
summary(lm)

Call:

lm<-lm(formula = y ~ hardwood$x + I((hardwood$x)^2), data = hardwood)

Residuals:
Min 1Q Median 3Q Max
-5.8503 -3.2482 -0.7267 4.1350 6.5506
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.67419 3.39971 -1.963 0.0673
hardwood$x 11.76401 1.00278 11.731 2.85E-09
I((hardwood$x)^2) -0.63455 0.06179 -10.27 1.89E-08
---
```

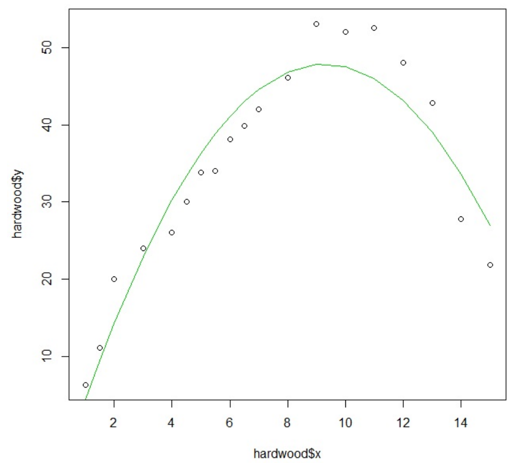
```
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 4.42 on 16 degrees of freedom
Multiple R-squared: 0.9085,Adjusted R-squared: 0.8971
F-statistic: 79.43 on 2 and 16 DF,p-value: 4.912e-09

anova(lm)

Analysis of Variance Table

Response: y
Df Sum Sq Mean Sq F value Pr(>F)
x1 1 1043.43 1043.43 5.34E+01 1.76E-06
x2 1 2060.82 2060.82 1.05E+02 1.89E-08
Residuals 16 312.64 19.54

---

Signif. codes:   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

library(car)

vif(lm)

hardwood$x I((hardwood$x)^2)

17.12077 17.12077
```

To sketch the model:

```
x<-hardwood$x y<-hardwood$y mean(x)
lines(x,predict(lm),col=3)
```

Variance Inflation Factor (VIF) is used to determine whether a model has serious multicollinearity. We can find the VIFs of uncentered models are larger than 10. It indicates that the model is ill-conditioned and exists multicollinearity. We centralize the data like

```
mean(x)
```




```
[1] 7.263158
 x1<-x-mean(x)
 x2<-(x-mean(x))^2
lm2<-lm(y~x1+x2,data=hardwood)
summary(lm2)


Call:

lm(formula = y ~ x1 + x2, data = hardwood)

Residuals:
Min 1Q Median 3Q Max
-5.8503 -3.2482 -0.7267 4.135 6.55E+00
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 45.29497 1.48287 30.55 1.29E-15
x1 2.54634 0.25384 10.03 2.63E-08
x2 -0.63455 0.06179 -10.27 1.89E-08

---
```

```
Signif. codes:   0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 4.42 on 16 degrees of freedom

Multiple R-squared:   0.9085,Adjusted R-squared: 0.8971
F-statistic: 79.43 on 2 and 16 DF,p-value: 4.912e-09
anova(lm2)
Analysis of Variance Table
Response: y
Df Sum Sq Mean Sq F value Pr(>F)
x1 1 1043.43 1043.43 5.34E+01 1.76E-06
x2 1 2060.82 2060.82 1.05E+02 1.89E-08
Residuals 16 312.64 19.54
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
vif(lm2)
x1 x2
1.09705 1.09705
```

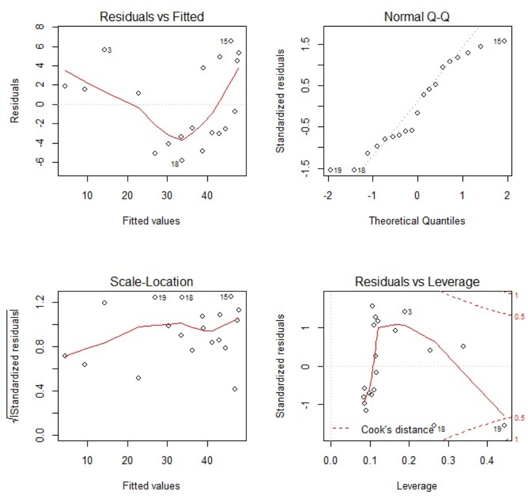
From the coefficients and variance tables, the value of estimated parameters from uncentered and centered models are same but the VIFs of centered model are very small. We can believe the problem of multicollinearity has been solved.

Now supposed that we wish to investigate the contribution of the quadratic term to the model:

The plot of residuals is shown below:

```
par(mfrow=c(2,2))
```

```
plot(lm2)
```



The plot for the residuals looks like a curve, which indicates that the model may not be linear so that we should add variable into the model. The qq-plot in right figure indicates that the errors distribution may has heavier tails than the normal assumption. Here we add the cross term to see if we can resolve this problem:

```
lm3<-lm(y~x1*x2)
summary(lm3)

Call:
lm(formula = y ~ x1 * x2)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6250 -1.6109  0.0413  1.5892  5.0216

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 44.975562   0.869032  51.754  < 2e-16 ***
```

```
x1           4.339394   0.350978  12.364 2.87e-09 ***
x2          -0.548873   0.039199 -14.002 5.11e-10 ***
x1:x2       -0.055188   0.009789  -5.638 4.72e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.585 on 15 degrees of freedom
Multiple R-squared:  0.9707,    Adjusted R-squared:  0.9648
F-statistic: 165.4 on 3 and 15 DF,  p-value: 1.025e-11
lines(x,predict(lm3),col=3)
```

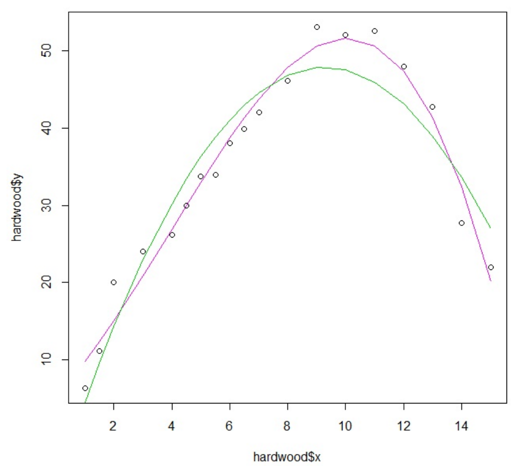


Figure 1: lm2 (green) versus lm3 (purple)

The new regression model can be written as

$$\hat{y} = 44.976 + 4.339(x - 7.2632) - 0.549(x - 7.2632)^2 - 0.055(x - 7.2632)^3.$$

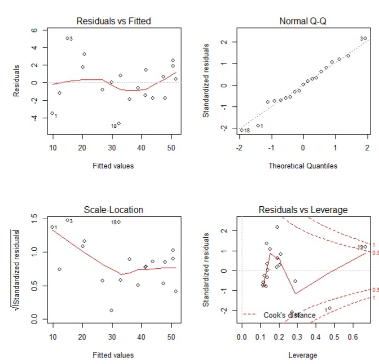
Note that here $x_1 = (x - mean(x))$ and $x_2 = (x - mean(x))^2$. therefore the cross term $x_1 x_2$ will be $x_3 = (x - mean(x))^3$.

```
anova(lm3)
Analysis of Variance Table
```

```
Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x1         1 1043.43 1043.43 156.144 2.481e-09 ***
x2         1 2060.82 2060.82 308.392 2.058e-11 ***
x1:x2      1  212.40  212.40  31.785 4.722e-05 ***
Residuals 15  100.24    6.68
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
vif(lm3)
      x1       x2    x1:x2
6.132746 1.291093 6.817759
```

We still use extra-sum-of-squares method to test the significance of the cross term. We can find $SSR(x1 * x2|x1, x2) = 212.4$, the p-value is $2.06E - 11$, which is less than 0.05. The null hypothesis $H0 : \beta_3 = 0$ should be rejected, so the cubic term contributes significantly to the model. VIF increases, which means the multicollinearity may exit.
For the plot of residuals for the third model: The plot of residuals is shown below:

```
par(mfrow=c(2,2))
plot(lm3)
```



The residuals are more random and the points in the qq-plot are closer to the red line. We can consider that the problem of nonlinearity has been solved.

3 Indicator Function

In some research of practical issues, besides the quantitative variables with well-defined scale of measurement, there are also some qualitative variables such as sex (male or female), data status (censored or not), times (before or after a war), type of products and so on. In statistics and econometrics, particularly in regression analysis, an indicator variable is one that takes the value 0 or 1 to indicate the absence or presence of some categorical effect that may be expected to shift the outcome.

Regression analysis treats all regressor variables (x_i) in the analysis as numerical. An indicator variable which also be called a dummy variable has no natural scale of measurement so it must be defined by us through assigning a set of levels, that is to say that it is an artificial variable created to represent an attribute with two or more distinct categories or levels. Unlike the quantitative variables that have concrete meaning, the qualitative variables are just used to indicate or identify the levels of the variable and do not have intrinsic meaning of their own. If there is a categorial explanatory variable in a data set, we can use it as a variable in the regression model following below steps. Usually, we let 0 or 1 (can be any other two arbitrary number) to identify the two levels of a qualitative variable. To illustrate, we want to build a model including the survival state of patients.

1. Choose one of the levels as the reference level, here we regard the state of dead as the reference level.

2. Make a dummy variable by using indicator for the other level,

$$x_i = \begin{cases} 0 & observation_i = reference(dead) \\ 1 & other(survival) \end{cases}$$

3. Run a simple linear regression using this indicator function as the explanatory variable

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

   here β_0 is the mean of the reference group(dead group), β_1 is the difference between the dead group and the survival group

For the variable has more than 2 states, we take another example that there are 3 brands of product and we want to use the levels of them as qualitative variables. The 3 levels are numbered $i, 1$, $i, 2$ and $i, 3$, here the numbers do not have concrete meaning but just the signs of products.

1. Choose brand 1 as the reference level for variable x_1, and brand 2 as the reference level for variable x_2.

2. $$x_{i,1} = \begin{cases} 1 & brand_i = 2 \\ 0 & others, \end{cases} \quad , \quad x_{i,2} = \begin{cases} 1 & brand_i = 3 \\ 0 & others, \end{cases}$$

3. $y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \epsilon_i$,

For indicator variables, we should keep the following points in mind.

- The number of indicator variables necessary to represent a single attribute variable is equal to the number of levels in that variable minus one.

- For a given attribute variable, none of the indicator variables constructed can be redundant. That is, one indicator variable can not be a constant multiple or a simple linear relation of another.

- The interaction of two attribute variables (e.g. Gender and Marital Status) is represented by a third indicator variable which is simply the product of the two individual indicator variables.

If a data set has one quantitative variable x_1 and a qualitative variable which has more than one levels (suppose to be a levels), we can represent them by $a-1$ indicator variables $x_2 \sim x_a$, each taking on the values of 0 and 1. The levels of the indicator variables are

x_2	x_3	$\cdots$	x_a	
0	0	$\cdots$	0	if the observation is from the 1st level
1	0	$\cdots$	0	if the observation is from the 2nd level
0	1	$\cdots$	0	if the observation is from the 3th level
$\vdots$	$\vdots$		$\vdots$	
0	0	$\cdots$	1	if the observation is from the ath level

and the regression model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_a x_a + \epsilon$$

Example 3 (The tool life data). The tool life data (y) present twenty observations on tool life and lathe speed (x_1) and the type of the cutting tool (x_2). The second regressor variable, tool type, is qualitative and has two levels:

$$x_2 \begin{cases} 0, & \text{type A;} \\ 1, & \text{type B .} \end{cases}$$

Assuming that a first-order model is appropriate, we have

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

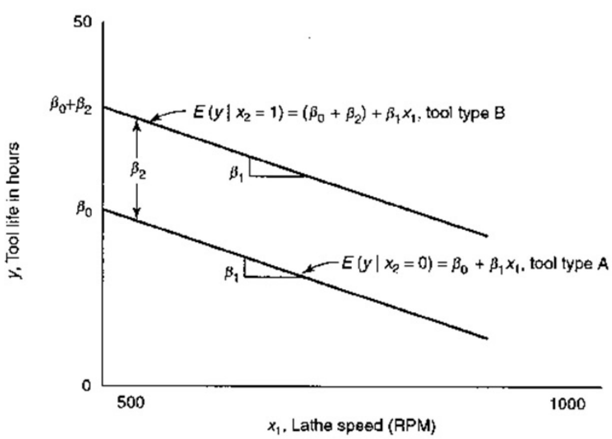
For the first tool type A, we have $x_2 = 0$ and the model become:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 0 + \epsilon = \beta_0 + \beta_1 x_1 + \epsilon$$

For the first tool type B, we have $x_2 = 1$ and the model become:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 1 + \epsilon = (\beta_0 + \beta_2) + \beta_1 x_1 + \epsilon$$

The two models are parallel.



tool

```
        y   x1 x2
1  18.73  610  a
2  14.52  950  a
3  17.43  720  a
4  14.54  840  a
5  13.44  980  a
6  24.39  530  a
7  13.34  680  a
8  22.71  540  a
9  12.68  890  a
10 19.32  730  a
11 30.16  670  b
12 27.09  770  b
13 25.40  880  b
14 26.05 1000  b
15 33.49  760  b
16 35.62  590  b
17 26.07  910  b
18 36.78  650  b
19 34.95  810  b
20 43.67  500  b
> x1<-tool$x1
> x2<-tool$x2
> y<-tool$y
> lm<-lm(y~x1+factor(x2))
> summary(lm)

Call:
lm(formula = y ~ x1 + factor(x2))

Residuals:
    Min      1Q  Median      3Q     Max
-5.5527 -1.7868 -0.0016  1.8395  4.9838

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.98560    3.51038  10.536 7.16e-09 ***
x1          -0.02661    0.00452  -5.887 1.79e-05 ***
```

```
factor(x2)b 15.00425    1.35967  11.035 3.59e-09 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3.039 on 17 degrees of freedom
Multiple R-squared:  0.9003,    Adjusted R-squared:  0.8886
F-statistic: 76.75 on 2 and 17 DF,  p-value: 3.086e-09
```

In the summary table the t statistics for β_1 and β_2 have small p-value, therefore we conclude both x_1 and x_2 contribute to the model. In the codes above factor(x2) changes the indicator variable into binomial variable. The regression model that we find is:

$$\hat{y} = 36.986 - 0.027x_1 + 15.004x_2$$

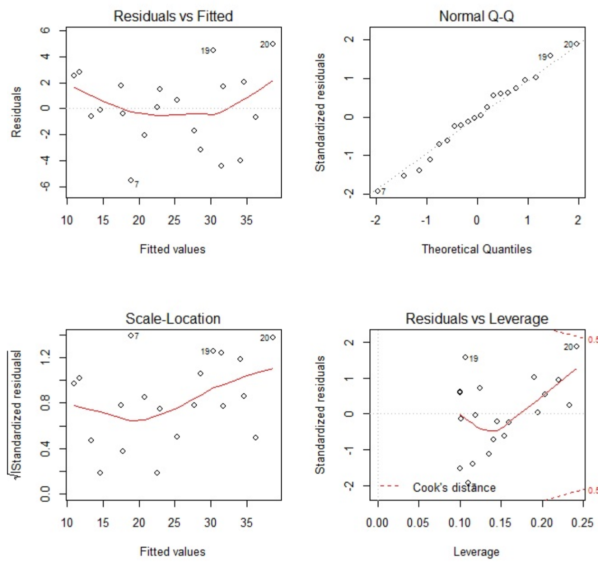
Here is the ANOVA table for this model. Since the p-values are very small, the hypothesis of significance of regression is rejected.

```
Analysis of Variance Table

Response: y
           Df  Sum Sq Mean Sq F value    Pr(>F)
x1          1  293.01  293.01  31.716 2.990e-05 ***
factor(x2)  1 1125.03 1125.03 121.776 3.587e-09 ***
Residuals  17  157.05    9.24
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The parameter β_2 is the change in mean tool life resulting from a change from tool type A to tool type B. The 95% confidence interval on β_2 is

```
beta.range<-confint(lm,level=0.95)
> beta.range
                  2.5 %      97.5 %
(Intercept) 29.57934160 44.39186079
x1          -0.03614301 -0.01707145
factor(x2)b 12.13559830 17.87290293
> beta.range[3,]
  2.5 %  97.5 %
12.1356 17.8729
```

The plots of residuals imply that there maybe a mild inequality-of-variance problem. The normal probability plot of residuals shows no indication of serious model inadequacies. Now Consider the following model (cross product)



for our example:

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

For type A :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 0 + \beta_3 0 + \epsilon = \beta_0 + \beta_1 x_1 + \epsilon$$

For type B :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 1 + \beta_3 x_1 + \epsilon = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1 + \epsilon$$

If we want to check whether or not the two models are identical we would test $H_0 : \beta_2 = \beta_3 = 0$ against $H_1 : \beta_2 \neq 0$ or/and $\beta_3 \neq 0$. If the null hypothesis is not rejected (by using extra sum of squares), this would imply a single regression model can explain between tool life and lathe speed.

```
lm2<-lm(y~x1*x2)
```

```
> summary(lm2)

Call:
lm(formula = y ~ x1 * x2)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.1750 -1.4999  0.4849  1.7830  4.8652 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 32.774760   4.633472   7.073 2.63e-06 ***
x1          -0.020970   0.006074  -3.452  0.00328 ** 
x2b         23.970593   6.768973   3.541  0.00272 ** 
x1:x2b      -0.011944   0.008842  -1.351  0.19553    
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1 

Residual standard error: 2.968 on 16 degrees of freedom
Multiple R-squared:  0.9105,    Adjusted R-squared:  0.8937 
F-statistic: 54.25 on 3 and 16 DF,  p-value: 1.319e-08 

> anova(lm2)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq  F value    Pr(>F)    
x1         1  293.01  293.01  33.2545 2.889e-05 ***
x2         1 1125.03 1125.03 127.6847 4.891e-09 ***
x1:x2      1   16.08   16.08   1.8248    0.1955    
Residuals 16  140.98    8.81                       
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1 
```

Here we get model

$$\hat{y} = 32.775 - 0.021x_1 + 23.971x_2 - 0.021x_1x_2.$$

We can see that the p-value for x_1x_2 is 0.19553 which indicates that we can't reject the null hypothesis ($\beta_3 = 0$) and this variable does not contribute significant;y to the model. To test the hypothesis that the two regression lines are identical:

To test the hypothesis that two lines have different intercepts and a common slope ($H_0 : \beta_3 = 0$)