STAC67: Regression Analysis

Lecture 16

Sohee Kang

Mar. 11, 2021

Interaction

- However,for example, One's initial salary depends on eduction level (categorical) and that increases are with year of experiene (continuous) and also depends on education level
- Interaction Term: can be used to expand the linear model to deal with these siuations : product of variables in the regression model

Interaction of 2 categorical variables

X: continuous var , Z_1 Z_2: binary variables $z_i \begin{cases} 0 & \dots \\ 1 & \dots \end{cases}$

γ : how much the intercept in the model deviates from the additive model for the two categorical variables

$$y = \beta_0 + \beta_1 x_1 + \alpha_1 z_1 + \alpha_2 z_2 + \gamma z_1 z_2 + \varepsilon$$

z_1	z_2	$E(Y \mid X, z_1, z_2)$
0	0	$\beta_0 + \beta_1 X$
1	0	$\beta_0 + \alpha_1 + \beta_1 X$
0	1	$(\beta_0 + \alpha_2) + \beta_1 X$
1	1	$(\beta_0 + \alpha_1 + \alpha_2 + \gamma) + \beta_1 X$

Interaction

$k-1$ dummy vars

2 interaction terms for each

- If both categorical variables have 3 categories: $(2) \times 2 = 4$
- Including interactions increases the number of parameters to etimate.
- In many cases, the additive structure fits the data well.

We fit model the model with and without interaction and use the F-test to compare to compare the interaction term is sig or not.

Interaction between a categorical and a continous variable

Y, X : continuous $\qquad Z$: Binary

$$Y = B_0 + B_1 X + \alpha_1 Z + \alpha_2 (XZ)$$

$$E(Y \mid Z=0) = B_0 + B_1 X$$

$$E(Y \mid Z=1) = (B_0 + \alpha_1) + (B_1 + \alpha_2) X$$

Example: Insurance Innovation Example

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 D_i + \beta_3 D_i X_{i1} + \epsilon_i \qquad D_i = \begin{cases} 1 \\ 0 \end{cases}$$

- The estimated model:

- Test whether the effect of firm size changes with the firm type.

$H_0: \beta_3 = 0$

$$t = \frac{\hat{\beta}_3}{SE(\hat{\beta}_3)} = \frac{-0.0004171}{0.0183312} = -0.02$$

① t val $t(0.975, 16) = 2.12$ $|t| <$ critical val

② $2 * pt(-0.02, 12) = 0.9843$ $\therefore$ Fail to reject H_0

- Conclusion: Fail to reject, there is evidence that $\beta_3 = 0$ with 95% confidence.

R codes:

```
Innovation = read.table("Table8-2.txt", header=F,  col.names=c("Y","X1","X2"))
fit = lm(Y~X1*X2, data=Innovation)
summary(fit)
```

or lm(Y ~ X1 + X2 + X1:X2 , data=Innovation)

```
##
## Call:
## lm(formula = Y ~ X1 * X2, data = Innovation)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7144 -1.7064 -0.4557  1.9311  6.3259
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.8383695  2.4406498  13.864 2.47e-10 ***
## X1          -0.1015306  0.0130525  -7.779 7.97e-07 ***
## X2           8.1312501  3.6540517   2.225   0.0408 *
## X1:X2       -0.0004171  0.0183312  -0.023   0.9821
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.32 on 16 degrees of freedom
## Multiple R-squared:  0.8951, Adjusted R-squared:  0.8754
## F-statistic: 45.49 on 3 and 16 DF,  p-value: 4.675e-08
```

Example

- Samples of male atheletes from the National ~~Basketball~~ EPL Association (NBA), National Hockey League (NHL), and English premier (Football) League are obtained, and the relationship between players' Weight (Y) and Height (X) is measured.
- NBA $= \begin{cases} 1 & \text{From NBA} \\ 0 & \text{ow} \end{cases}$ NHL $= \begin{cases} 1 & \text{From NHL} \\ 0 & \text{ow} \end{cases}$

Baseline Cat = EPL

- Full Model:

$$Y = B_0 + B_1 X_1 + B_2 NBA + B_3 NHL + \varepsilon$$

1. Test the identitiy of three regression functions

$$H_0 : B_2 = B_3 = 0$$

$$F^* = \frac{\frac{SSR(NHL, NBA, X_1) - SSR(X_1)}{4-2}}{MSE_F} = \frac{\frac{5587 + 2157}{2}}{161}$$

R codes

```
Player = read.csv("sample.csv", header=T)
fit = lm(Weight ~ Height + NBA + NHL, data=Player)
fit2 = lm(Weight~Height, data=Player)
```

```
\begin{verbatim}
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -272.7683    42.6965  -6.389 3.52e-08 ***
Height         6.1373     0.6017  10.199 2.22e-14 ***
NBA            5.7244     6.3590   0.900    0.372
NHL           26.5333     4.5029   5.892 2.27e-07 ***
---
\end{verbatim}
```

```
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: Weight
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Height     1  45416   45416 282.254 < 2.2e-16 ***
## NBA        1   2157    2157  13.404 0.0005576 ***
## NHL        1   5587    5587  34.721 2.272e-07 ***
## Residuals 56   9011     161
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example: Interaction Model

- Full Model: $n = 60$

$$Y_i = \beta_0 + \beta_1 Height + \beta_2 NBA + \beta_3 NHL + \beta_4 Height * NBA + \beta_5 Height * NHL + \epsilon_i$$

2) Test the equality of slopes of three regression models:

$H_0: \beta_4 = \beta_5 = 0$

NBA : $E(Y) = (\beta_0 + \beta_2) + (\beta_1 + \beta_4) X_1$

NHL : $E(Y) = (\beta_0 + \beta_3) + (\beta_1 + \beta_5) X_1$

EPL : $E(Y) = \beta_0 + \beta_1 X_1$

$$F^* = \frac{\frac{SSE_R - SSE_F}{2}}{MSE_F} = \frac{\frac{9011809}{2}}{148}$$

R codes

```
fit2 = lm(Weight ~ Height*NBA + Height*NHL , data=Player)
anova(fit2)
```

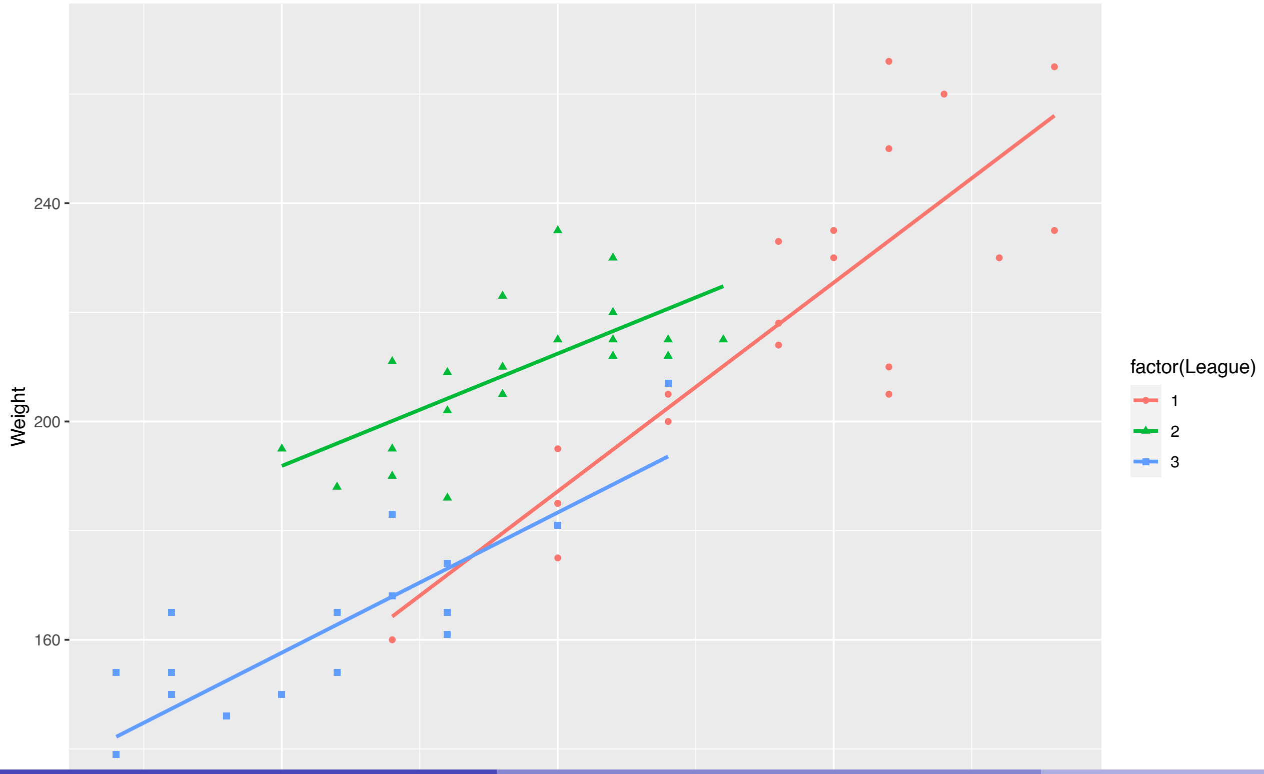
```
## Analysis of Variance Table
##
## Response: Weight
##            Df Sum Sq Mean Sq  F value    Pr(>F)
## Height      1  45416   45416 306.2009 < 2.2e-16 ***
## NBA         1   2157    2157  14.5413 0.0003541 ***
## NHL         1   5587    5587  37.6672 1.026e-07 ***
## Height:NBA  1    944     944   6.3644 0.0146215 *
## Height:NHL  1     57      57   0.3867 0.5366565
## Residuals  54   8009     148
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
\begin{verbatim}
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -201.914     73.134  -2.761  0.00786 **
Height         5.136      1.032   4.976 6.98e-06 ***
NBA         -184.599     98.395  -1.876  0.06605 .
NHL          105.640    119.624   0.883  0.38110
Height:NBA     2.513      1.326   1.895  0.06345 .
Height:NHL    -1.020      1.641  -0.622  0.53666
\end{verbatim}
```

R codes

```
library(ggplot2)
ggplot(data=Player, aes(x=Height, y=Weight, color= factor(League), shape=factor(League))) + geom_point() + geom_
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Case Study (SENIC)

- The primary objective of the Study on the Efficacy of Nosocomial Infection Control (SENIC Project) was to determine whether infection surveillance and control programs have reduced the rates of nosocomial (hospital-acquired) infection in United States hospitals. This data set consists of a random sample of 113 hospitals selected from the original 338 hospitals surveyed. Each line of the data set has an identification number and provides information on 11 other variables for a single hospital. The data presented here are for the 1975-76 study period.
- Consider a model of regressing infectious risk Y against age X_1, routine culturing ratio X_2, average daily census X_3, available facilities and service X_4, Medical school affiliation X_5. For each region, we can find a model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon$$

- Var9 is the **Region**: Geographic region, where: 1=NE, 2=NC, 3=S, 4=W Are the estimated regression functions similar for the four regions? Discuss.

Case study

Let D1 = $\begin{cases} 1 \\ 0 \end{cases}$ NE

Baseline = W

Let D2 = $\begin{cases} 1 \\ 0 \end{cases}$ NC

Let D3 = $\begin{cases} 1 \\ 0 \end{cases}$ S

- Full model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$
  $+ \alpha_1 D_1 + \alpha_2 D_2 + \alpha_3 D_3 + \gamma_1 D_1 X_1 + \gamma_2 D_1 X_2 + \cdots$
  $+ \gamma D_3 X_5$
- Reduced model: $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$

```
Data = read.table("senic.txt")
Y = Data[,4]
X1 = Data[,3]
X2 = Data[,5]
X3 = Data[,10]
X4 = Data[,12]
X5 = Data[,8]
Z = Data[,9]
```

R codes

```r
###Now consider a big model with dummy variable
D1 = as.numeric(Z==1)
D2 = as.numeric(Z==2)
D3 = as.numeric(Z==3)

regFULL = lm(Y~X1+X2+X3+X4+X5+D1+D2+D3
        +D1:X1+D1:X2+D1:X3+D1:X4+D1:X5
        +D2:X1+D2:X2+D2:X3+D2:X4+D2:X5
        +D3:X1+D3:X2+D3:X3+D3:X4+D3:X5)
#summary(regFULL)
anova(regFULL)
```

$H_0: \alpha_1 = \alpha_2 = \cdots \gamma = 0$

H_a: at least one not equal to zero

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X1         1  0.000   0.000  0.0002   0.98768
## X2         1 66.406  66.406 66.2007 2.280e-12 ***
## X3         1 19.072  19.072 19.0128 3.478e-05 ***
## X4         1  2.924   2.924  2.9153   0.09123 .
## X5         1  2.044   2.044  2.0381   0.15690
## D1         1  0.016   0.016  0.0157   0.90051
## D2         1  0.041   0.041  0.0412   0.83966
## D3         1  4.658   4.658  4.6433   0.03388 *
## X1:D1      1  2.256   2.256  2.2490   0.13724
## X2:D1      1  0.168   0.168  0.1678   0.68305
## X3:D1      1  0.555   0.555  0.5533   0.45892
## X4:D1      1  0.014   0.014  0.0143   0.90501
## X5:D1      1  0.006   0.006  0.0062   0.93722
## X1:D2      1  0.160   0.160  0.1596   0.69051
## X2:D2      1  2.798   2.798  2.7897   0.09838
```

R codes

```r
regR = lm(Y~X1+X2+X3+X4+X5)
#summary(regR)
anova(regR)
```

```
## Analysis of Variance Table
##
## Response: Y
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## X1          1   0.000   0.000  0.0002   0.98787
## X2          1  66.406  66.406 64.0515 1.559e-12 ***
## X3          1  19.072  19.072 18.3956 3.947e-05 ***
## X4          1   2.924   2.924  2.8207   0.09598 .
## X5          1   2.044   2.044  1.9720   0.16314
## Residuals 107 110.933   1.037
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

op of code for f-test

```r
anova(regR, regFULL)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X1 + X2 + X3 + X4 + X5
## Model 2: Y ~ X1 + X2 + X3 + X4 + X5 + D1 + D2 + D3 + D1:X1 + D1:X2 + D1:X3 +
##     D1:X4 + D1:X5 + D2:X1 + D2:X2 + D2:X3 + D2:X4 + D2:X5 + D3:X1 +
##     D3:X2 + D3:X3 + D3:X4 + D3:X5
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1    107 110.933
## 2     89  89.276 18    21.657 1.1995 0.2791
```

cannot reject, no correlation between region and infection rate.

8.1 Polynomial Regression Models

- Polynomial regression models have two basic types of uses:

1. When the true curvilinear response function is indeed a polynomial function.
2. When the true curvilinear response function is unknown (or complex) but a polynomial function is a good approximation to the true function. [More Common]

- Danger of polynomial regression models:

Polynomial regression models may provide good fits for the data at hand, but may turn in unexpected directions when extrapolated beyond the range of the data.

One Predictor - Second order

only for quantative predictors

- A polynomial regression model with one predictor variable raised to the first and second powers:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i$$

where, $X_i = X_i - \bar{X}$ — centering

- This polynomial regression model is called a second-order model with one predictor variable because the single predictor variable is expressed in the model to the first and second powers. reduce multicolinearity
- The predictor variable is centered.
- The reason for centering is: to reduce high correlation between X and X^2
- This model is frequently rewritten:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_{11} X_i^2 + \epsilon_i$$

One Predictor Variable - higher order

- Third order model:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_{11} X_i^2 + \beta_{111} X_i^3 + \epsilon_i$$

- ℓ-th order model with one predictor variable

$$Y_i = \beta_0 + \sum_{k=1}^{\ell} \beta_k X_i^k + \epsilon_i$$

- To be used with special caution because
  - Difficult to interpret of regression coefs
  - Uncertain behaviour of the model for extrapolation
  - We can almost always find sufficiently large order to fit data (Overfitting)

Polynomial regression models - several predictor variables

- Second order model with two predictor variables:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_{11} X_{i1}^2 + \beta_{22} X_{i2}^2 + \beta_{12} X_{i1} X_{i2} + \epsilon_i$$

- Second order model with three predictor variables

$$Y = B_0 + B_1 X_1 + B_2 X_2 + B_3 X_3 + B_4 X_1^2 + B_5 X_2^2 + B_6 X_3^2 + B_7 X_1 X_2 + B_8 X_1 X_3$$
$$+ B_9 X_2 X_3 + \varepsilon$$

- And so on...
- A polynomial regression model is a particular case of multiple regression model.

Polynomial Models Fitting

- The second order model is equivalent to:

$$\underset{\sim}{Y} = \boldsymbol{X}\underset{\sim}{\beta} + \underset{\sim}{\epsilon}$$

with

$$\boldsymbol{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & X_{11}^2 & X_{12}^2 & X_{11}X_{12} \\ 1 & X_{21} & X_{22} & X_{21}^2 & X_{22}^2 & X_{21}X_{22} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & X_{n1}^2 & X_{n2}^2 & X_{n1}X_{n2} \end{bmatrix}$$

- All earlier theory and results on general linear regression model apply.
- Polynomial models fitting is not a new problem!

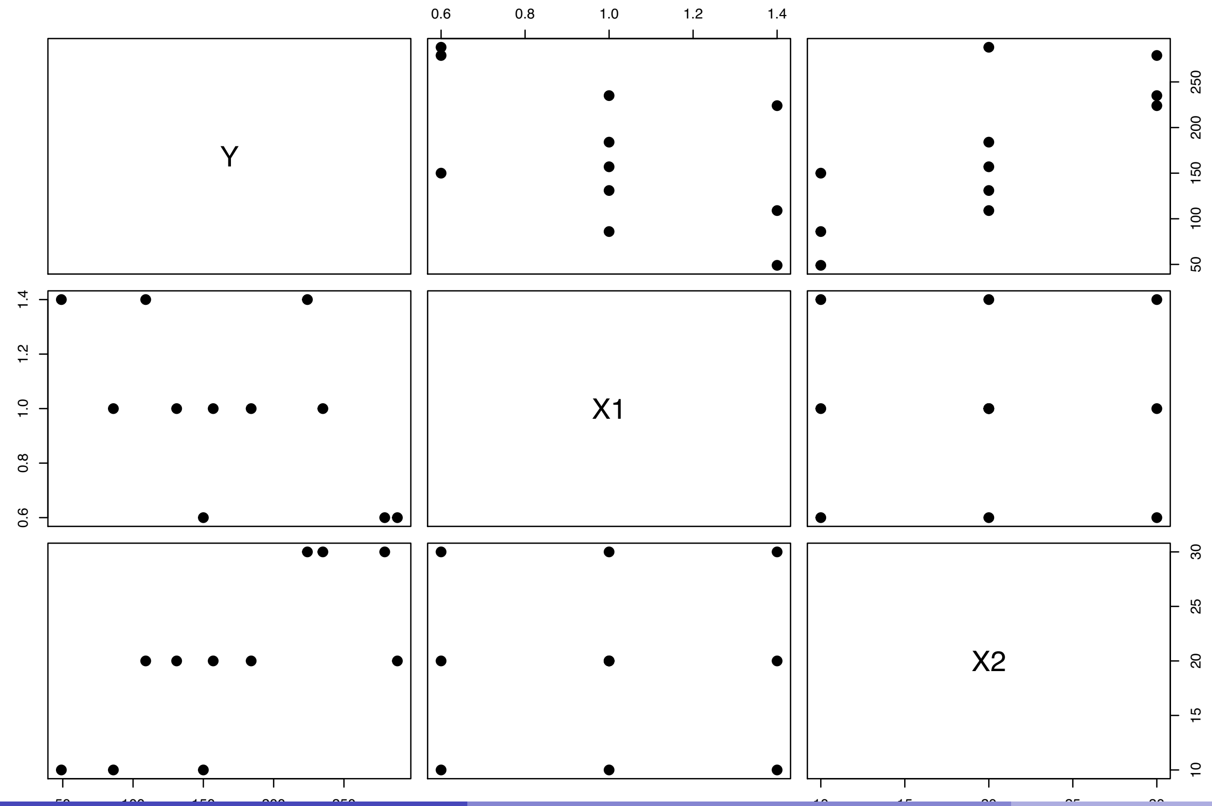
Power cells example

- Researcher studies the effects of the charge rate (amperes) and temperature (degrees Celsius) of a new type of power cell in a preliminary small-scale experiment.
- Three levels of charge rate and of temperature
- Life of the power cell in terms of the number of discharge-charge cycles before the cell failed

Cell i	Number of cycles Y_i	Charge rate X_{i1}	Temperature X_{i2}
1	150	0.6	10
2	86	1.0	10
3	49	1.4	10
$\vdots$	$\vdots$	$\vdots$	$\vdots$
11	224	1.4	30
Mean		1.0	20

Power cells example

```
Powercell= read.table("Table8-1.txt", header=T)
par(mfrow=c(2,2))
pairs(Powercell, pch=19, cex=1.5)
```



Power cells example

- seems to be a good idea.

```
fit = lm(Y~X1 + X2 + I(X1^2)+ I(X1*X2) + I(X2^2), data=Powercell)
summary(fit)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + I(X1^2) + I(X1 * X2) + I(X2^2), data = Powercell)
##
## Residuals:
##       1       2       3       4       5       6       7       8       9      10
## -21.465   9.263  12.202  41.930  -5.842 -31.842  21.158 -25.404 -20.465   7.263
##      11
##  13.202
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  337.7215   149.9616   2.252   0.0741 .
## X1          -539.5175   268.8603  -2.007   0.1011
## X2             8.9171     9.1825   0.971   0.3761
## I(X1^2)      171.2171   127.1255   1.347   0.2359
## I(X1 * X2)     2.8750     4.0468   0.710   0.5092
## I(X2^2)       -0.1061     0.2034  -0.521   0.6244
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.37 on 5 degrees of freedom
## Multiple R-squared:  0.9135, Adjusted R-squared:  0.8271
## F-statistic: 10.57 on 5 and 5 DF,  p-value: 0.01086
```

R output

- The correlation matrix of the variables included in the model is:

	Y	X_1	X_2	X_1^2	X_2^2	X_1X_2
Y	1.000	-0.556	0.751	-0.529	0.737	0.255
X_1	-0.556	1.000	0.000	0.991	0.000	0.605
X_2	0.751	0.000	1.000	0.000	0.986	0.757
X_1^2	-0.529	0.991	0.000	1.000	0.006	0.600
X_2^2	0.737	0.000	0.986	0.006	1.000	0.746
X_1X_2	0.255	0.605	0.757	0.600	0.746	1.000

- Based on the R output on this slide and the previous one, would you say that the model considered is appropriate? Justify.

Recording of the variables

- Let's center the variables around the mean:

- The correlation matrix of the recoded variables is:

	Y	x_1	x_2	x_1^2	x_2^2	x_1x_2
Y	1.000	-0.556	0.751	0.165	-0.022	0.093
x_1	-0.556	1.000	0.000	0.000	0.000	0.000
x_2	0.751	0.000	1.000	0.000	0.000	0.000
x_1^2	0.165	0.000	0.000	1.000	0.267	0.000
x_2^2	-0.022	0.000	0.000	0.267	1.000	0.000
x_1x_2	0.093	0.000	0.000	0.000	0.000	1.000

R codes

```r
attach(Powercell)
x1 = X1 - mean(X1)
x2 = X2 - mean(X2)
fit2 = lm(Y ~ x1 + x2 + I(x1^2) + I(x2^2) + I(x1*x2))
summary(fit2)
```

```
##
## Call:
## lm(formula = Y ~ x1 + x2 + I(x1^2) + I(x2^2) + I(x1 * x2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.05486 -0.78425  0.06687  0.82103  2.58254
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.4551446  0.1303992  34.165  < 2e-16 ***
## x1           0.0371199  0.0236446   1.570   0.1194
## x2           0.1014715  0.0135585   7.484 2.14e-11 ***
## I(x1^2)      0.0030020  0.0031124   0.965   0.3370
## I(x2^2)     -0.0014825  0.0006857  -2.162   0.0328 *
## I(x1 * x2)   0.0005464  0.0024179   0.226   0.8216
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.082 on 107 degrees of freedom
## Multiple R-squared:  0.3779, Adjusted R-squared:  0.3488
## F-statistic:    13 on 5 and 107 DF,  p-value: 6.94e-10
```

Polynomial regression model and centered data

- Reason of centering:

Hierarchical approach to fitting

- First fit a second-order or third-order model and then explore whether a lower-order model is adequate
- **Exercise 1**:

Consider the third order model

$$Y_i = \beta_0 + \beta_1 X_i + \beta_{11} X_i^2 + \beta_{111} X_i^3 + \epsilon_i$$

How can we test whether the cubic term can be dropped? And how can we test whether both the cubic term and quadratic term can be dropped?

Regression function in terms of the initial variables

- We often wish to express the final model in terms of the original variables (rather than the centered variables).
- Example: we consider the fitted model

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_{11} x_i^2 + \hat{\beta}_{111} x_i^3$$

  which we want to express in terms of Xi rather than $x_i = X_i - \bar{X}$.
- **Exercise 2**: Show that the fitted model

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_{11} x_i^2$$

  can be express in terms of X_i as

$$\hat{Y}_i = \hat{\beta}_0' + \hat{\beta}_1' X_i + \hat{\beta}_{11}' X_i^2$$