Polynomial Regression

- **One Predictor Variable - Second Order.** $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i$.

  where, ϵ_i are independent, identically distributed, $N(0, \sigma^2)$

  However, X_i and X_i^2 are usually highly correlated. This leads to the multicollinearity problem. To reduce the this problem, we transform the predictor and use $X_i^* = X_i - \bar{X}$.

- **One Centered Predictor Variable - Second Order.** $Y_i = \beta_0^* + \beta_1^* X_i^* + \beta_{11}^* (X_i^*)^2 + \epsilon_i$.

  where:

  1. X_i^* is the ith centered observation for predictor X.
  2. ϵ_i are independent, identically distributed, $N(0, \sigma^2)$

  To get the estimates for the parameters using the original variable, we can use the following equations:

  1. $b_{11} = b_{11}^*$
  2. $b_1 = b_1^* - 2 * b_{11}^* \bar{X}$
  3. $b_0 = b_0^* - b_1^* * \bar{X} + b_{11}^* \bar{X}^2$

- **Simulation Example**

```
n=50
x=sample(20:40,n,replace=T)
y=100-30*x+0.5*x^2 + rnorm(n,mean=0,sd=10)
plot(x,y)

x.2=x^2
cor(x,x.2)
summary(lm(y~x+x.2))

xc=x-mean(x)
xc.2=xc^2
cor(xc,xc.2)
results=(lm(y~xc+xc.2))
summary(results)
b0=coef(results)[1]
b1=coef(results)[2]
b2=coef(results)[3]

b0.orig=b0-b1*mean(x)+b2*(mean(x))^2
b1.orig=b1-2*b2*mean(x)
b2.orig=b2
curve(b0.orig+b1.orig*x+b2.orig*x^2,20,40,add=T)

y.hat=b0.orig+b1.orig*x+b2.orig*x^2 # Or use, yhat=results$fitted
residuals=y-y.hat                   # Or use, residuals=results$res

# Testing lack of fit of the quadratic regression
red=lm(y~xc+xc.2)
full=lm(y~factor(xc))
anova(red,full)
```

- **Power Cell Example.** A researcher studied the effects of the charge rate and temperature on the life of a new type of power cell in a preliminary small-scale experiment. The rate (X_1) was controlled at three levels (0.6, 1.0, and 1.4 amperes) and the ambient temperature (X_2) was controlled at three levels (10, 20, and 30^oC). Factors pertaining to the discharge of the power cell were held at fixed levels. The life of the power cell (Y) was measured in terms of the number of discharge-charge cycles that a power cell underwent before it failed. The data are stored in the file "PowerCells.csv".

```
data=read.csv("PowerCells.csv",header=T)
attach(data)
cor(charge,charge^2)   # Note how correlated they are.
[1] 0.9910312
cor(temperature,temperature^2)
[1] 0.9860911

y=cycles
x1=(charge-mean(charge))/.4 # Transforming will reduce the correlation
x2=(temperature-mean(temperature))/10
x1.2=x1^2 # Check: cor(x1,x1.2) = -4.042173e-16
x2.2=x2^2 # Check: cor(x2,x2.2) = 0
x1x2=x1*x2

summary(lm(y~x1+x1.2+x2+x2.2+x1x2)) # Note that the interaction effect is not significant.
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   162.84      16.61   9.805 0.000188 ***
x1            -55.83      13.22  -4.224 0.008292 **
x1.2           27.39      20.34   1.347 0.235856
x2             75.50      13.22   5.712 0.002297 **
x2.2          -10.61      20.34  -0.521 0.624352
x1x2           11.50      16.19   0.710 0.509184        <-- Not significant

summary(lm(y~x1+x1.2+x2+x2.2)) # Note that x2^2 is not significant.
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   162.84      15.91  10.237 5.07e-05 ***
x1            -55.83      12.66  -4.410 0.004517 **
x1.2           27.39      19.48   1.406 0.209306
x2             75.50      12.66   5.964 0.000996 ***
x2.2          -10.61      19.48  -0.544 0.605825        <-- Not significant

summary(lm(y~x1+x1.2+x2)) # Note that x1^2 is not significant.
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   158.60      13.15  12.059 6.15e-06 ***
x1            -55.83      12.01  -4.650 0.002341 **
x1.2           24.57      17.81   1.380 0.210194        <-- Not significant
x2             75.50      12.01   6.288 0.000409 ***

summary(lm(y~x1+x2)) # Note how much R^2 changed as you remove variables from the model.
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  172.000      9.354  18.387 7.88e-08 ***
x1           -55.833     12.666  -4.408 0.002262 **
x2            75.500     12.666   5.961 0.000338 ***

# Or you can test Ho:beta_11=beta_22=beta_12=0
red=lm(y~x1+x2)
full=lm(y~x1+x1.2+x2+x2.2+x1x2)
anova(red,full)
Model 1: y ~ x1 + x2
Model 2: y ~ x1 + x1.2 + x2 + x2.2 + x1x2
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1      8 7700.3
2      5 5240.4  3    2459.9 0.7823 0.5527
# Since the p-value=0.5527, we can conclude Ho,
# that no curvature and interaction effects are needed.

# Changing back to the original data:
b0.orig=172+55.833*mean(charge)/.4-75.5*mean(temperature)/10  # b0.orig=160.5825
b1.orig=-55.833/.4  # b1.orig=-139.5825
b2.orig=75.5/10     # b0.orig=7.55
y.hat=160.58-139.58*charge+7.55*temperature
```