1 Polynomial regression with a single predictor - section 7.1

- The following example simulates from a second order relationship \( y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon \), and fits an incorrect, first order relationship \( y = \beta_0 + \beta_1 x + \epsilon \).

- the QQ plot of the residuals indicates a non-normal short tailed distribution of the residuals, but is otherwise not indicative of the true relationship.
> x=seq(1,10,length.out=100)
> y=1+2*x+.75*x^2+rnorm(100,0,.3)
> lm.out=lm(y~x)
> lm.resid=residuals(lm.out)
> lm.fits=fitted(lm.out)
> qqnorm(lm.resid,main="normal quantile plot of residuals")
> qqline(lm.resid)
• the plot of residuals vs fitted values suggests the addition of a quadratic term

```r
> plot(lm.fits, lm.resid, main="plot of residuals vs fitted values",
+     xlab="fitted values", ylab="residuals")
```

![Plot of residuals vs fitted values](image)
> lm.out2=lm(y~x+I(x^2))
> lm.resid2=residuals(lm.out2)
> lm.fits2=fitted(lm.out2)
> plot(lm.fits2,lm.resid2,main="Quadratic model",
+ xlab="fitted values",ylab="residuals")
• note the use of the $I()$ operator in the R model statement to get the polynomial term.
• there is no indication of a problem in the residual plot
• the multiple linear regression model can incorporate higher order polynomial terms. For example

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_k x^k + \epsilon$$
1.1 Polynomial models in two variables

- suppose we have observations on a dependent variable $y$ and two independent variables $x_1$ and $x_2$.
- in the following model the mean of $y$ is quadratic in the two variables $x_1$ and $x_2$.

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \epsilon \]

- Following are some plots of the mean of $y$ for a couple of different choices of $\beta_0, \ldots, \beta_5$. 
- the following plots the surface

\[ E(y) = 800 + 10x_1 + 7x_2 - 8.5x_1^2 + 5x_2^2 + 4x_1x_2 + \epsilon \]

for \( x_1 \) and \( x_2 \) both taking values on \((-10,10)\)

\[
\text{Ey} = \text{function}(x1,x2,beta=c(800,10,7,-8.5,5,4))\{
  \text{+ return}(beta[1]+beta[2]*x1+beta[3]*x2+beta[4]*x1^2+
  \text{+ beta[5]*x2^2+beta[6]*x1*x2})
\}
\]

\[
\text{nrow}=60; \text{ncol}=60
\]
\[
\text{x1}=\text{seq}(-10,10,\text{length.out= nrow})
\]
\[
\text{x2}=\text{seq}(-10,10,\text{length.out= ncol})
\]
\[
\text{y}=\text{matrix}(\text{rep}(0,\text{nrow*ncol}), \text{byrow=T, nrow=nrow})
\]
\[
\text{for (i in 1:nrow)}{
  \text{for (j in 1:ncol)}{
    y[i,j]=Ey(x1[i],x2[j],beta=c(800,10,7,-8.5,5,4))}
}\]
\[
\text{persp}(x1,x2,y,xlab="X1",ylab="X2",zlab="y",ticktype="detailed",
\text{+ phi=30,theta=135})
\]
- the next plot just changes the sign on the coefficient of $x_2^2$.

$$E(y) = 800 + 10x_1 + 7x_2 - 8.5x_1^2 - 5x_2^2 + 4x_1x_2 + \epsilon$$

```r
> y2=matrix(rep(0,nrow*ncol),byrow=T,nrow=nrow)
> for (i in 1:nrow)
+    for (j in 1:ncol)
+        y2[i,j]=Ey(x1[i],x2[j],beta=c(800,10,7,-8.5,-5,4))
> persp(x1,x2,y2,xlab="X1",ylab="X2",zlab="y",ticktype="detailed",
+    phi=30,theta=135)
```
• Because the model is quadratic, it can accommodate at most one extreme point (as in the second figure), or a saddle point (as in the first figure).

• In general, as indicated in chapter 7,
  – higher order polynomials can fit surfaces with several local maxima or minima
  – they can approximate most nonlinear functions, as they are essentially Taylor approximations to the true underlying function
  – high order polynomial models rarely provide an understanding of a true unknown nonlinear function
  – the estimated coefficients are often imprecise, as the $X^TX$ matrix is typically ill conditioned for a high degree polynomial.
Problem 7.18 provides some data on solubility.

The variables are:

- The response variable $y$ is the negative logarithm of mole fraction solubility.
- $x_1 =$ dispersion partial solubility
- $x_2 =$ dipolar partial solubility
- $x_3 =$ hydrogen bonding Hansen partial solubility

The problem asks to fit a complete quadratic model, and to test for the contribution of all second order terms. The reduced model retains only the linear terms in $x_1$, $x_2$ and $x_3$. 
> data=read.csv("http://bsmith.mathstat.dal.ca/stat3340/Data/data-prob-7-18.csv",header=T)
> data.2ndorder=lm(y~x1+x2+x3+I(x1^2)+I(x2^2)+I(x3^2)+I(x1*x2)+I(x1*x3)+I(x2*x3),data=data)
> data.1storder=lm(y~x1+x2+x3,data=data)
> anova(data.2ndorder,data.1storder)

Analysis of Variance Table

Model 1: y ~ x1 + x2 + x3 + I(x1^2) + I(x2^2) + I(x3^2) + I(x1 * x2) + I(x1 * x3) + I(x2 * x3)
Model 2: y ~ x1 + x2 + x3

Res.Df RSS Df Sum of Sq F Pr(>F)
1  16 0.059386
2  22 0.095294 -6 -0.035908 1.6124 0.2076

• The F test for the second order terms are not significant.
> resids.1storder = residuals(data.1storder)
> predict.1storder = predict(data.1storder)
> qqnorm(resids.1storder)
> qqline(resids.1storder)

The QQ plot of residuals appears to show some deviation from normality in the tails.
The plot of residuals vs fitted values shows no obvious trend, and no suggestion that variance of the residuals is non-constant.