1. Example. Data showing nonlinearity.

```r
dxmp13.07
days yield
1 16 2508
2 18 2518
...
15 44 3103
16 46 2776

> polyfit <- lm(yield ~ days + I(days^2))
> beta <- coef(polyfit)
> beta
   (Intercept) days I(days^2)
 -1070.397689 293.482948 -4.535802

x <- 15:45
lines(x, y)
```

```r
to.r
> summary(polyfit)
Call: lm(formula = yield ~ days + I(days^2))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1070.398   617.253  -1.734   0.107
      days  293.483    42.178   6.958  9.94e-06 ***
I(days^2)  -4.536     0.674  -6.726  1.41e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 203.9 on 13 degrees of freedom Multiple
R-Squared: 0.7942,    Adjusted R-squared: 0.7625 F-statistic: 25.08
on 2 and 13 DF,  p-value: 3.452e-05
```

2. We will consider a general model that includes a linear model as a special case. For given \( n \) pured measurements \((x_j, y_j)\), we will fit the following model:

\[
Y_i = \phi(x_i) + \epsilon_i,
\]

where \( \epsilon_i \sim N(0, \sigma^2) \). Some possible model choice would be \( \phi(x) = \beta_0 + \beta_1 x \) (linear), \( \phi(x) = \beta_0 + \beta_1 x + \beta_2 x^2 \) (quadratic). The \( k \)-th degree polynomial regression model is given by

\[
\phi(x_j) = \beta_0 + \beta_1 x_j + \cdots + \beta_k x_j^k.
\]
Figure 1: Grain yield vs. the number of days after flowering. Left: quadratic model. Right: cubic model. Since do not see much difference, the cubic model is not suitable for this data.

In general $\phi_i(x)$ is expressed as

$$\phi(x) = \sum_{i=0}^{m} \beta_i \phi_i(x),$$

where $\phi_i$ are basis functions such as $\phi_i(x) = x^i$ and we estimate $\beta_i$ via the least squares estimation. We need to solve the system of the linear equation

$$
\begin{bmatrix}
\phi_1(x_1) & \cdots & \phi_m(x_1) \\
\phi_1(x_2) & \cdots & \phi_m(x_2) \\
\vdots & \ddots & \vdots \\
\phi_1(x_n) & \cdots & \phi_m(x_n)
\end{bmatrix}
\begin{bmatrix}
\beta_1 \\
\beta_2 \\
\vdots \\
\beta_m
\end{bmatrix} =
\begin{bmatrix}
y_1 \\
y_2 \\
\vdots \\
y_n
\end{bmatrix}.
$$

It can be written as $y = X\beta$ where $X$ is rectangular. Then

$$X'X\beta = X'y$$

If there are more data then the number of basis functions, i.e. $n \geq m + 1$, $X'X$ is invertible. So we get

$$\hat{\beta} = (X'X)^{-1}X'y.$$  

Popular choice for $\phi_i$ is Hermite polynomials $H_i(x)$ which forms orthogonal polynomial basis with respect to $e^{-x^2}$ in $\mathbb{R}$, i.e.

$$\int_{-\infty}^{\infty} H_i(x)H_j(x)e^{-x^2} \, dx = \delta_{ij} \sqrt{\pi}.$$  

The first few Hermite polynomials are $H_0(x) = 1$, $H_1(x) = 2x$, $H_2(x) = 4x^2 - 2$.

3. For inference on $\beta_j$, we estimate $\hat{\sigma}^2 = \frac{SSE}{n-(m+1)}$. Also $R^2$ is now called the coefficient of multiple determination and it measures the goodness of the polynomial model fit. If $SSE_m$ is the SSE of the $m$-th degree model, we can show that $SSE_{m+1} \leq SSE_m$ and $R^2_{m+1} \geq R^2_m$. If $\hat{S}_{\beta_j}$ is the estimator for $\sigma^{\beta_j}$,

$$T = \frac{\hat{\beta}_j - \beta_j}{\hat{S}_{\beta_j}} \sim t_{n-(m+1)}.$$  

Determining the degree of a polynomial model is done by performing a test on $H_0 : \beta_m = 0$ vs. $H_1 : \beta_m \neq 0$.

```r
> summary(polyfit)
```
Call: \( \text{lm(formula } = \text{ yield } \sim \text{ days } + \text{ I(days}^2) + \text{ I(days}^3) \text{)} \)

Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | -203.60852 | 2285.13020 | -0.089 | 0.930 |
| days | 199.07674 | 242.92513 | 0.819 | 0.428 |
| I(days\(^2\)) | -1.32071 | 8.16843 | -0.162 | 0.874 |
| I(days\(^3\)) | -0.03457 | 0.08751 | -0.395 | 0.700 |

Residual standard error: 210.8 on 12 degrees of freedom

Multiple R-Squared: 0.7968, Adjusted R-squared: 0.7468

F-statistic: 15.68 on 3 and 12 DF, p-value: 0.0001876

4. **Logistic regression.** Consider model \( Y_j = \beta_0 + \beta_1 x_j + \epsilon_j \) with \( \mathbb{E}\epsilon_j = 0, \forall \epsilon_j = \sigma^2 \). If the response variable \( Y_j \) is distributed as \( \text{Bernoulli}(p) \) with the probability of failure \( p \), the above linear model is no longer appropriate since

\[
\mathbb{E}Y_j = p(x_j) = \beta_0 + \beta_1 x_j
\]

but \( \beta_0 + \beta_1 x_j \) may not be in the range \([0, 1]\). To address this problem, we introduce a nonlinear model on the probability of failure \( p \):

\[
p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}.
\]

This is called the logit function. Note that the odd ratio is given by

\[
\frac{p(x)}{1 - p(x)} = e^{\beta_0 + \beta_1 x}.
\]

We estimate the parameters via MLE. For \( Y_j \sim \text{Bernoulli}(p(x_j)) \),

\[
L(\beta_0, \beta_1) = \prod_{j=1}^{n} P(Y_j = y_j) = \prod_{j=1}^{n} \left( \frac{e^{\beta_0 + \beta_1 x_j}}{1 + e^{\beta_0 + \beta_1 x_j}} \right)^{y_j} \prod_{j=1}^{n} \left( \frac{1}{1 + e^{\beta_0 + \beta_1 x_j}} \right)^{1-y_j}.
\]

```r
> xmp13.06
  Temperature Failure
  1 53 Y
  2 56 Y
  3 57 Y
  ...
oring<-glm(Failure~Temperature,family=binomial(link=logit))
beta <- coef(oring)
x<-40:90
plot(x,p,type="l")
```

Coefficients:

| Estimate | Std. Error | z value | Pr(>|z|) |
|----------|------------|---------|----------|
| (Intercept) | 11.74641 | 6.02142 | 1.951 | 0.0511 . |
| Temperature | -0.18843 | 0.08909 | -2.115 | 0.0344 * |

**Assigned problems.** Exercise 13.32., Exercise 13.25 use R package to generate a similar computer output. HW 6 due April 28th. Solve the assigned problems in Lecture 19-22.