The procedure for estimating the regression coefficients in polynomial regression is the same as before, i.e. by minimizing MSE w.r.t. $\alpha, \beta_1, \beta_2, \ldots$. Each derivative leads to a linear equation, and the system of equations can be uniquely solved to give $\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \ldots$.

For this homework, consider a quadratic regression, and derive the linear equations that must be satisfied by $\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2$. Write these equations in terms of the following means: $\bar{x}, \bar{x}^2, \bar{x}^3, \bar{x}^4, \bar{xy}, \bar{x^2y}, \bar{y}$.

Do not solve the system of equations.

\[
\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta}_1 x_i - \hat{\beta}_2 x_i^2)^2 ,
\]

\[
\frac{\partial}{\partial \alpha} \text{MSE} \bigg|_{\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2} = 0 , \quad \frac{\partial}{\partial \beta_1} = 0 , \quad \frac{\partial}{\partial \beta_2} = 0
\]

\[
\frac{\partial}{\partial \alpha} = 0 \Rightarrow \frac{2}{n} \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta}_1 x_i - \hat{\beta}_2 x_i^2) = 2(\bar{y} - \hat{\alpha} - \hat{\beta}_1 \bar{x} - \hat{\beta}_2 \bar{x}^2) = 0
\]

\[
\frac{\partial}{\partial \beta_1} = 0 \Rightarrow \frac{2}{n} \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta}_1 x_i - \hat{\beta}_2 x_i^2) (-x_i) = -2(\bar{xy} - \hat{\alpha} \bar{x} - \hat{\beta}_1 \bar{x}^2 - \hat{\beta}_2 \bar{x}^3) = 0
\]

\[
\frac{\partial}{\partial \beta_2} = 0 \Rightarrow \frac{2}{n} \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta}_1 x_i - \hat{\beta}_2 x_i^2) (-x_i^2) = -2(\bar{x^2y} - \hat{\alpha} \bar{x^2} - \hat{\beta}_1 \bar{x^3} - \hat{\beta}_2 \bar{x^4}) = 0
\]
a) $R^2 = 0.78$
I.e., 78% of the variability in $y$ can be explained (or attributed to)
the linear relation $y = a + b_1 x_1 + b_2 x_2 + b_3 x_3$.

b) $y = 185.49 - 45.97 \times 2.6 - 0.3 \times 250 + 0.089 \times 2.6 \times 250 = 48.8$

c) Were it not for the interaction term, the answer would be "yes."
However, because $b_3$ is not zero, the interaction term exists, and so, the
coefficients cannot be interpreted in this way.

d) The question is asking for $s_e$. Most standard computer printouts
would report $s_e$ on the same line (or near) $R^2$. This printout does NOT!
You just have to get used to these things. Now,

\[
s_e = \sqrt{\frac{SSE}{n-(k+1)}} = \sqrt{\frac{4290.5}{20 - (3 + 1)}} = 16.4
\]

e) $x_1 = c(0.6, 0.6, 0.6, 0.6, 0.6, 0.6, 0.6, 0.6, 0.6, 0.6)$
$x_2 = c(200, 250, 400, 500, 600, 200, 250, 400, 500, 600)$
$y = c(90.6, 82.7, 58.7, 43.2, 25.0, 127.1, 112.3, 19.6, 17.8, 9.1, 53.1, 52.0, 43.4, 42.4, 31.6, 40.9, 37.9, 27.5, 27.3, 19.0)$
$x_3 = x_1 \times x_2$

Even though the data are provided through $x_1$ and $x_2$, the regression
equation is $y = a + b_1 x_1 + b_2 x_2 + b_3 x_3$. So, one might treat $x_3$ as
another predictor. In that case, the scatterplots and the correlations
that should be check are the following:

```r
par(mfrow=c(2,2))
plot(x1,x2)
plot(x1,x3)
plot(x2,x3)
cor(x1,x2)  # 0
cor(x1,x3)  # 0.78241
cor(x2,x3)  # 0.5456007
```

# $x_1$ and $x_2$ are not at all collinear.
# But $x_1$ and $x_3$, and $x_2$ and $x_3$, do show some collinearity. And their
# correlation coefficients suggest that too. The implication of this
# collinearity is that the regression coefficients are going to have
# large variability, and so their interpretation (as rate of ....) is
# going to be unreliable.

# Now, on a slightly technical side: We already know that the
# coefficients are not going to be interpretable *because* of the
# interaction term. But, if the model did not have an interaction
# term (i.e., $x_3$), then you would look only at the collinearity between
# $x_1$ and $x_2$. For this data set, given that $x_1$ and $x_2$ are not correlated,
# we would be able to interpret the coefficients.
From the statement of the problem, I copied/pasted the data into an ascii text file called hw_3_37_dat.txt. I did include a line of header. Then,

dat = read.table("hw_3_37_dat.txt", header=T)
x1 = dat[,2]
x2 = dat[,3]
y = dat[,4]

# a)
lm.1 = lm(y ~ x1 + x2 + I(x1^2) + I(x2^2) + I(x1*x2))

# (Intercept)           x1           x2      I(x1^2 )      I(x2^2)   I(x1 * x2)
# -140.22976    -16.47521     12.82710      0.09555      -0.24339      0.49864

# b)
The regression coefficients cannot be interpreted (as the rate of ...) because of the interaction term. If the interaction term were not present, we would still have to check for collinearity before we could attempt to interpret the coefficients.

# c)
summary(lm.1)

# Multiple R-squared: 0.7561
# This means that about 76% of the variability in Strength can be attributed to (or explained by) Depth and Water Content, through the expression given above in lm().

# d) summary(lm.1) also contains se: 7.023.
# This means that the typical error (or deviation) of the data about the fit is about 7 kPa.

# e) Two ways of getting residual plots, that you learned in lab.
# first way:
plot(lm.1) # hit return after this. The first fig is the residual plot.

# second way:
png("3_37_residual.png")
y_hat = predict(lm.1)
plot(y_hat, lm.1$residuals)
dev.off()

# The residual plot is a random scatter of dots about the y=0 line, and so, suggests that the fit is good.

# f)
lm.2 = lm(y ~ x1 + x2 )

# (Intercept)           x1           x2
# 14.8893       0.6607      -0.0284

# g)
summary(lm.2)

# Multiple R-squared: 0.447,
# About 45% of the variance in Strength can be attributed to (or explained by) Depth and and Water Content through the linear expression in lm().

# h) Given that the more complex model has a much larger R2 (76% compared to 45%) it's reasonable to conclude that at least one of the higher-order terms provides useful information about Strength.

# i) According to summary(lm.2), the value of se is 9.019. In other words, the more complex model has a higher R2 and smaller typical error (7.02 compared to 9.02).
Data for 3.37 above:

<table>
<thead>
<tr>
<th>Obs</th>
<th>Depth</th>
<th>Content</th>
<th>Strength</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8.9</td>
<td>31.5</td>
<td>14.7</td>
</tr>
<tr>
<td>2</td>
<td>36.6</td>
<td>27.0</td>
<td>48.0</td>
</tr>
<tr>
<td>3</td>
<td>36.8</td>
<td>25.9</td>
<td>25.6</td>
</tr>
<tr>
<td>4</td>
<td>6.1</td>
<td>39.1</td>
<td>10.0</td>
</tr>
<tr>
<td>5</td>
<td>6.9</td>
<td>39.2</td>
<td>16.0</td>
</tr>
<tr>
<td>6</td>
<td>6.9</td>
<td>38.3</td>
<td>16.8</td>
</tr>
<tr>
<td>7</td>
<td>7.3</td>
<td>33.9</td>
<td>20.7</td>
</tr>
<tr>
<td>8</td>
<td>8.4</td>
<td>33.8</td>
<td>38.8</td>
</tr>
<tr>
<td>9</td>
<td>6.5</td>
<td>27.9</td>
<td>16.9</td>
</tr>
<tr>
<td>10</td>
<td>8.0</td>
<td>33.1</td>
<td>27.0</td>
</tr>
<tr>
<td>11</td>
<td>4.5</td>
<td>26.3</td>
<td>16.0</td>
</tr>
<tr>
<td>12</td>
<td>9.9</td>
<td>37.0</td>
<td>24.9</td>
</tr>
<tr>
<td>13</td>
<td>2.9</td>
<td>34.6</td>
<td>7.3</td>
</tr>
<tr>
<td>14</td>
<td>2.0</td>
<td>36.4</td>
<td>12.8</td>
</tr>
</tbody>
</table>
For each of the data sets a) hw_3_dat1.txt and b) hw_3_dat2.txt, find the "best" fit, and report R-squared and the standard deviation of the errors. Do not use some ad hoc criterion to determine what is the "best" fit. Instead, use your knowledge of regression to find the best fit, and explain in words why you think you have the best fit.

# a)
```r
dat = read.table("hw_3_dat1.txt",header=T)
x1 = dat[,1] ; x2 = dat[,2] ; y = dat[,3]
plot(dat)
```

# There is no collinearity, but there is clearly a signature of interaction (see lecture 13).
```r
lm.1 = lm(y~x1+x2 + I(x1*x2))
summary(lm.1)  # R2 = 0.94, s_e = 2.0
```

# Note that without the interaction term, the fit is much worse:
```
lm.1 = lm(y~x1+x2)
summary(lm.1)  # R2 = 0.23, s_e = 6.9
```

# b)
```r
dat = read.table("hw_3_dat2.txt",header=T)
plot(dat)
```

# There is significant collinearity, and so one of the two predictors should be ignored.
```r
plot(x1,y)
```

# There is still a quadratic behavior, though:
```r
x1 = dat[,1] ; x2 = dat[,2] ; y = dat[,3]
```
```r
lm.2 = lm(y~x1 + I(x1^2))
summary(lm.2)  # R2 = 0.87, s_e = 3.8
```

# Note that without the quadratic term, the fit is much worse:
```
lm.0 = lm(y~x1)
summary(lm.0)  # R2 = 0.25, s_e = 9.0
```
```
lm.00 = lm(y~ x1 + x2)
summary(lm.00) # R2 = 0.27, s_e = 8.9
```
Generate data on $x_1$, $x_2$, and $y$, such that

1) $n$ (= sample size) = 100,
2) $x_1$ and $x_2$ are uncorrelated, and from a uniform distribution between 0 and 1,

a) Let $y$ be given by $y = 2 + 3x_1 + 4x_2 + \text{error}$, where error is from a normal distribution with mean = 0 and sigma = 0.5. Fit the model \( y = \alpha + \beta_1 x_1 + \beta_2 x_2 \) to the above data, and report $R^2$ and $s_e$.

b) Let $y$ be given by $y = 2 + 3x_1 + 4x_2 + 50(x_1 x_2) + \text{error}$, where error is from a normal distribution with mean = 0 and sigma = 0.5. Fit the model \( y = \alpha + \beta_1 x_1 + \beta_2 x_2 \) to the above data, and report $R^2$ and $s_e$.

c) Fit the model \( y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 x_2) \) to the data from part b, and report $R^2$ and $s_e$.

d) Install the R package called "rgl" on your computer, by typing \texttt{install.packages("rgl",dep=T)}, and following the instructions. If you have trouble with this, ask the TAs or I during office hours.

Then, at the R prompt, type

\begin{verbatim}
library(rgl)
plot3d(x1,x2,y)
\end{verbatim}

The panel you will see is interactive. By holding down the left-button, and moving the mouse around, you will be able to "turn" the figure around in different ways. Have some fun with it, THEN based on what you see, provide an explanation for why the quality (in terms of $R^2$ and/or $s_e$) of the fit in part c is better than that in part b.

# Soln

\begin{verbatim}
set.seed(123)
n = 100x1 = runif(n,0,1)x2 = runif(n,0,1)error = rnorm(n,0,0.5)

# a)
y = 2 + 3*x1 + 4*x2 + error
summary(lm(y ~ x1 + x2)) # R2 = 0.8792, se = 0.4882

# b)
y = 2 + 3*x1 + 4*x2 + 50*x1*x2 + error
summary(lm(y ~ x1 + x2)) # R2 = 0.9118, se = 3.549

# c)
summury(lm(y ~ x1 + x2 + I(x1*x2))) # R2 = 0.9983, se = 0.4897

# d)
# install.packages("rgl",dep=T) # follow the instructions.
library(rgl)
plot3d(x1,x2,y)
\end{verbatim}

Clearly, the data don't look planar. They are "warped" in the shape of a saddle surface. For this reason, a model with an interaction term will provide a better fit to the data.
Consider fitting a model \( y_i = \beta x_{1i} x_{2i} + \varepsilon_i \), \( i = 1, \ldots, n \), to data on \( x_1, x_2, y \). Show that the OLS estimate of \( \beta \) is

\[
\hat{\beta} = \frac{x_1 x_2 y}{(x_1)^2 (x_2)^2}
\]

Hint: find the critical value of \( \beta \) that minimizes

\[
SSE = \sum_{i=1}^{n} (y_i - \beta x_{1i} x_{2i})^2
\]

\[
\frac{\partial SSE}{\partial \beta} = 2 \sum_{i=1}^{n} (y_i - \beta x_{1i} x_{2i}) x_{1i} x_{2i} = 0
\]

\[
\sum_{i=1}^{n} x_{1i} x_{2i} y_i = \hat{\beta} \sum_{i=1}^{n} x_{1i}^2 x_{2i}^2
\]

\[
\frac{x_1 x_2 y}{(x_1)^2 (x_2)^2} = \hat{\beta} \frac{x_1 x_2 y}{(x_1)^2 (x_2)^2}
\]

\[
\therefore \hat{\beta} = \frac{x_1 x_2 y}{(x_1)^2 (x_2)^2}
\]
Let \( p \), the sample proportion of girls, be written as \( p = \frac{n_g}{n} \), where \( n \) = sample size, and \( n_g \) = number of girls in \( n \).

Show that \( E[p] = \mu_x \) \( V[p] = \frac{\mu_x(1-\mu_x)}{n} \) where \( \mu_x \) = prop. of girls in the pop.

Do not use sums of 0's and 1's, like the book does. Instead repeat the way we derived \( E[\bar{x}] \) and \( V[\bar{x}] \), above but now keeping in mind that something in this problem is actually binomial. Hint: find out what's binomial, first.

Recall the example from the class:

\[
E[\bar{x}] = E \left[ \frac{1}{n} \sum x_i \right] = \frac{1}{n} \sum E[x_i] = \frac{1}{n} n \mu_x = \mu_x,
\]

\( \mu_x \) because \( x_i \) are normal.

\[
E[p] = E \left[ \frac{n_g}{n} \right] = \frac{1}{n} E[n_g] = \frac{1}{n} n \mu_x = \mu_x.
\]

\( \mu_x \) because \( n_g \) is binomial.

Similarly for the variance:

\[
V[\bar{x}] = V \left[ \frac{1}{n} \sum x_i \right] = \left( \frac{1}{n} \right)^2 \sum V[x_i] = \frac{1}{n^2} n \sigma^2_x = \frac{\sigma^2_x}{n},
\]

\[
V[p] = V \left[ \frac{n_g}{n} \right] = \frac{1}{n^2} V[n_g] = \frac{1}{n^2} n \mu_x(1-\mu_x) = \frac{n \mu_x(1-\mu_x)}{n^2} = \frac{n \mu_x(1-\mu_x)}{n^2}
\]

\( \mu_x(1-\mu_x) \)
A sample of size 36 from a Normal pop. yields \( \bar{x} = 3.5 \).

a) Under the assumption that \( \mu_x = 2.5 \), \( \sigma_x = 2 \), what's the prob of a sample mean larger than the one observed.  

\[
p(\bar{x} > \bar{x}_{\text{obs}}) = p(z > z_{\text{obs}}) = p(z > \frac{3.5 - 2.5}{2/\sqrt{36}}) = p(z > 1.5) = 1 - 0.9332 = 0.0668
\]

If \( \mu_x = 2.5 \), it is unlikely to get a sample mean larger than that observed, and \( \sigma_x = 2 \).

b) Under the assumption that \( \mu_x = 3.5 \), \( \sigma_x = 2 \), what's the prob of a sample mean smaller than the one observed.

\[
p(\bar{x} < \bar{x}_{\text{obs}}) = p(z < z_{\text{obs}}) = p(z < 1.5) = 0.9332
\]

If \( \mu_x = 3.5 \), it is likely to get a sample mean smaller than that observed, and \( \sigma_x = 2 \).

c) Under the assumption that \( \mu_x = 3.5 \), \( \sigma_x = 2 \), what's the prob of a sample mean larger than the one observed.  

\[
p(\bar{x} > \bar{x}_{\text{obs}}) = p(z > z_{\text{obs}}) = p(z > \frac{3.5 - 3.5}{2/\sqrt{36}}) = p(z > -1.5) = 1 - p(z < -1.5) = 1 - 0.0668 = 0.9332
\]

If \( \mu_x = 3.5 \), it is likely to get a sample mean larger than that observed, and \( \sigma_x = 2 \).

d) Under the assumption that \( \mu_x = 3.5 \), \( \sigma_x = 2 \), what's the prob of a sample mean smaller than the one observed.

\[
p(\bar{x} < \bar{x}_{\text{obs}}) = p(z < z_{\text{obs}}) = \text{prob}(z < -1.5) = 0.0668
\]

If \( \mu_x = 3.5 \), it is unlikely to get a sample mean larger than that observed, and \( \sigma_x = 2 \).