For the data shown here:
\[x = 45, 58, 71, 71, 85, 98, 108\]
\[y = 3.20, 3.40, 3.47, 3.55, 3.60, 3.70, 3.80\]

a) Compute the eq. of the OLS fit

b) Compute the total variation, SST.

c) Decompose it into explained and unexplained.

d) Compute R2 and interpret it (in English),

e) Compute The std. dev of errors, and interpret it (in English).

All by hand. You may use R to compute sums, means, std. deviations, but not a function that does regression or analysis of variance.

\[x = c(45, 58, 71, 71, 85, 98, 108)\]
\[y = c(3.20, 3.40, 3.47, 3.55, 3.60, 3.70, 3.80)\]

# a)  
\[
\beta = \frac{\text{mean}(x*y) - \text{mean}(x)*\text{mean}(y)}{\text{mean}(x^2) - (\text{mean}(x))^2} \\
\alpha = \text{mean}(y) - \beta*\text{mean}(x)
\]

\beta
\alpha
# equation of line is \( y = 2.856 + .0088 \times \)

# b)  
\[
\text{SST} = \sum (y - \text{mean}(y))^2 \\
\text{SST} \quad # 0.2364857
\]

# c)  
\[
\hat{y} = \alpha + \beta \times x \\
\text{SSexplained} = \sum (\hat{y} - \text{mean}(y))^2 \\
\text{SSunexplained} = \sum (y - \hat{y})^2
\]

\text{SSexplained} \quad # 0.2273712
\text{SSunexplained} \quad # 0.009114473

# Note that SSexplained + SSunexplained = SST

#d)  
\[
R^2 = 1 - \frac{\text{SSunexplained}}{\text{SST}} \\
R^2 \quad # 0.96
\]

# 96% of the variation in y is explained through its linear relation with x.

# e)  
\[
\text{se} = \sqrt{\frac{\text{SSunexplained}}{\text{length}(x) - 2}} \\
\text{se} \quad # 0.042
\]

# The typical error (or deviation) about the line is 0.042 .

# Note, all of the above can be obtained by
\[
\text{lm.1} = \text{lm}(y \sim x) \\
\text{summary(lm.1)} \\
\text{anova(lm.1)}
\]
Consider the following decomposition:

\[
\sum_i (y_i - \bar{y})^2 = \sum_i [(\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)]^2
\]

\[
= \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2 + 2 \sum_i (\hat{y}_i - \bar{y})(y_i - \hat{y}_i)
\]

In past hws I have asked students to prove that the last term is zero if \( \hat{y}_i = \hat{\alpha} + \hat{\beta} x_i \), with \( \hat{\alpha}, \hat{\beta} \) being the OLS estimates (ie. \( \hat{\alpha}, \hat{\beta} \) given in lecture, book). Unfortunately, it's a long calculation; so this time we'll try to show that it's zero using simulation in R. Write code to

a) generate a sample of size 100 from the uniform dist. between -1 and +1. Call it \( x \).

b) generate \( y \) such that \( y = 2 + 3 x + \epsilon \) with \( \epsilon \) having a normal distrib. with \( \mu = 0, \sigma = 0.5 \).

c) Do regression on \( x, y \), and call the predictions \( \hat{y} \).

d) compute \( \sum_i (\hat{y}_i - \bar{y})(y_i - \hat{y}_i) \). It should be (very) zero.

```r
# a)
> n = 100
> x = runif(n,-1,1)
# b)
> y = 2 + 3*x + rnorm(n,0,0.5)
# c)
> lm.1 = lm(y~x)
> yhat = predict(lm.1)
# d)
> sum((yhat-mean(y))*(y - yhat))
```
SS_{exp.} can be computed from its defining relation \( \sum_i (\hat{y}_i - \bar{y})^2 \)
Or from \((SS\mathcal{T} - SSE)\), or from \(\hat{\beta}\) and \(S_{xx}\), as follows.

Explain what has happened at every step:

\[
SS_{exp.} = \sum \left( \hat{y}_i - \bar{y} \right)^2 \]
\[
= \sum \left( \hat{\alpha} + \hat{\beta} x_i - \bar{y} \right)^2 \]
\[
= \sum \left( \bar{y} - \hat{\beta} \bar{x} + \hat{\beta} x_i - \bar{y} \right)^2 \]
\[
= \sum (\hat{\beta})^2 (x_i - \bar{x})^2 \]
\[
= (\hat{\beta})^2 \sum (x_i - \bar{x})^2 \]
\[
= (\hat{\beta})^2 S_{xx} \]

? \( \hat{\beta} \) is the slope of the line.

\( \hat{\beta} \) has no "i"; factors out.

\( \bar{y} \)'s cancel, \((\hat{\beta})^2\) factors.
a) Read the data file transform.dat.txt from the course website into R, and
b) Make a scatterplot of y vs. x.
c) Transform x and/or y to linearize the relationship.
d) Perform regression on the transformed data, i.e., do \( \text{lm} \),
e) Overlay the corresponding line on the scatterplot
f) What percentage of the variability in the transformed y is explained by the transformed x, and what's the typical error in the prediction of the transformed y.

\begin{verbatim}
# a)
dat = read.table("transform_dat.txt", header=T)
attach(dat)
# b)
plot(x,y) # nonlinear, but monotonic.
# c)
plot(sqrt(x), sqrt(y)) # linear
# d)
lm.1 = lm(sqrt(y) ~ sqrt(x))
# e)
abline(lm.1, col=2)
# f)
summary(lm.1) # R^2 = 0.9922, 99% of the variability in sqrt(y) is explained by sqrt(x)
# se = 0.01911 = typical error in sqrt(y)

# FYI: If you had not performed the transformation, then R^2 would have been smaller, and s_e would have been larger. Look

summary(lm(y ~ x)) # R^2 = 95.7%, s_e = 0.2272

But, you should keep in mind that the 0.2271 is the typical error in y, while the 0.019 we got previously is the typical error in sqrt(y). So, they are hard to compare.
\end{verbatim}
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, 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: $\bar{x}, \bar{x}^2, \bar{x}^3, \bar{x}^4, \bar{x}\bar{y}, \bar{x}^2\bar{y}, \bar{y}$.

Do not solve the system of equations.

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

\[ \frac{2}{\hat{\alpha}} \text{MSE} \bigg|_{\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2} = 0, \quad \frac{2}{\hat{\beta}_1} = 0, \quad \frac{2}{\hat{\beta}_2} = 0, \]

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

\[ \frac{2}{\hat{\beta}_1} = 0 \Rightarrow \frac{2}{n} \sum_i (y_i - \hat{\alpha} - \hat{\beta}_1 x_i - \hat{\beta}_2 x_i^2)(-x_i) = -2(\bar{xy} - \bar{x} \bar{y} - \hat{\beta}_1 \bar{x}^2 - \hat{\beta}_2 \bar{x}^3) = 0 \]

\[ \frac{2}{\hat{\beta}_2} = 0 \Rightarrow \frac{2}{n} \sum_i (y_i - \hat{\alpha} - \hat{\beta}_1 x_i - \hat{\beta}_2 x_i^2)(-x_i^2) = -2(\bar{x}^2y - 2\bar{x}^2\bar{y} - \hat{\beta}_1 \bar{x}^4 - \hat{\beta}_2 \bar{x}^5) = 0 \]
Consider fitting a model \( y_i = \beta x_{i1} x_{i2} + \epsilon_i \), \( i = 1, \ldots, n \), 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

\[
\text{SSE} = \sum_{i=1}^{n} (y_i - \beta x_{i1} x_{i2})^2
\]

\[
\frac{\partial \text{SSE}}{\partial \beta} = 2 \sum_{i=1}^{n} (y_i - \hat{\beta} x_{i1} x_{i2}) x_{i1} x_{i2} = 0
\]

\[
\sum_{i=1}^{n} x_{i1} x_{i2} y_i = \hat{\beta} \sum_{i=1}^{n} x_{i1}^2 x_{i2}^2
\]

\[
\sum_{i=1}^{n} x_{i1} x_{i2} y_i = \hat{\beta} \left( \sum_{i=1}^{n} x_{i1}^2 \right) \left( \sum_{i=1}^{n} x_{i2}^2 \right)
\]

\[
\therefore \hat{\beta} = \frac{x_1 x_2 y}{(x_1)^2 (x_2)^2}
\]
The article "The Undrained Strength of Some Thawed Permafrost Soils" (Canadian Geotech. J., 1979: 420-427) contained the accompanying data on y shear strength of sandy soil (kPa), x1 depth (m), and x2 water content (%).

<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>

a) Perform regression to predict y from x1, x2, x3 = x1^2, x4 = x2^2, and x5 = x1*x2; and write down the coefficients of the various terms.

```
# 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.

dat = read.table("hw_3_37.dat.txt", header=T)
x1 = dat[,2]
x2 = dat[,3]
y = dat[,4]

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    -16.4752 1     12.82710      0.09555      -0.24339      0.49864
```

b) Can you interpret the regression coefficients? Explain.

```
# 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) Compute R^2 and explain what it says about goodness-of-fit ("in English").

```
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) Compute s_e, and interpret ("in English").

```
# 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) Produce the residual plot (residuals vs. "predicted" y), and explain what it suggests, if any.

```
# 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 residualplot is a random scatter of dots about the y=0 line, and so, suggests that the fit is good.
```
f) Now perform regression to predict $y$ from $x_1$ and $x_2$ only.

$$\text{lm.2} = \text{lm}(y \sim x_1 + x_2)$$

```
lm.2
# (Intercept)           x1           x2
#     14.8893       0.6607      -0.0284
```

g) Compute $R^2$ and explain what it says about goodness-of-fit.

```
summary(lm.2)
# Multiple R-squared: 0.447; about 45% of the variance in Strength can be attributed to (or explained by) Depth and Water Content through the linear expression in lm() .
```

h) Compare the above two $R^2$ values. Does the comparison suggest that at least one of the higher-order terms in the regression eqn provides useful information about strength?

```
# Given that the more complex model has a much larger $R^2$ (76% compared to 45%) it's reasonable to conclude that at least one of the higher-order terms provides useful information about Strength.
```

i) Compute $s_e$ for the model in part f, and compare it to that in part d. What do you conclude?

```
# i) According to summary(lm.2), the value of $s_e$ is 9.019. In other words, the more complex model has a higher $R^2$ and smaller typical error (7.02 compared to 9.02).
```

---

Data for 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>
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 $y = 2 + 3 \times x_1 + 4 \times x_2 + \text{error}$, 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 $s_e$.

b) Let $y$ be given by $y = 2 + 3 \times x_1 + 4 \times x_2 + 50 \times (x_1 \times x_2) + \text{error}$, 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 $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 \times x_2)$ 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
   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
   library(rgl)
   followed by
   plot3d(x1,x2,y)
   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

set.seed(123)
n = 100
x1 = runif(n,0,1)
x2 = runif(n,0,1)
error = rnorm(n,0,0.5)

# a)
y = 2 + 3*x1 + 4*x2 + \text{error}
summary(lm(y ~ x1 + x2))        # $R^2 = 0.8792$, $se = 0.4882$

# b)
y = 2 + 3*x1 + 4*x2 + 50*(x1*x2) + \text{error}
summary(lm(y ~ x1 + x2))        # $R^2 = 0.9118$, $se = 3.549$

# c)
summary(lm(y ~ x1 + x2 + I(x1*x2))) # $R^2 = 0.9983$, $se = 0.4897$

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

# 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.