The (normal) simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, 2, \ldots, n,$$

where ε_i's are independent Normal$(0, \sigma^2)$ (iid Normal$(0, \sigma^2)$).

β_0, β_1, and σ^2 are unknown model parameters.

1. The owner of *Momma Leona's Pizza* restaurant chain believes that if a restaurant is located near a college campus, then there is a linear relationship between sales and the size of the student population. Suppose data were collected from a sample of 10 *Momma Leona's Pizza* restaurants located near college campuses. For the i th restaurant in the sample, x_i is the size of the student population (in thousands) and y_i is the quarterly sales (in thousands of dollars). The values of x_i and y_i for the 10 restaurants in the sample are summarized in the following table:

Restaurant	Student Population (1000s)	Quarterly Sales ($1000s)
i	x_i	y_i
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

Recall:

$\bar{x} = 14$

$\bar{y} = 130$

$SXX = 568$

$SXY = 2{,}840$

$SYY = 15{,}730$

$\hat{\beta}_1 = 5$

$\hat{\beta}_0 = 60$

$\hat{y} = 60 + 5 \cdot x$

DATA = PREDICTION OF MODEL + RESIDUAL

$$y = \hat{y} + e$$

$$e = y - \hat{y} = y - (\hat{\beta}_0 + \hat{\beta}_1 \cdot x)$$

i	x	y	$\hat{y}$	$e = y - \hat{y}$	e^2
1	2	58	70	-12	144
2	6	105	90	15	225
3	8	88	100	-12	144
4	8	118	100	18	324
5	12	117	120	-3	9
6	16	137	140	-3	9
7	20	157	160	-3	9
8	20	169	160	9	81
9	22	149	170	-21	441
10	26	202	190	12	144
			Total	**0**	**1530**

$RSS = \sum e_i^2 = \sum \left(y_i - \hat{y}_i \right)^2 = 1530$ (another common notation is SSE)

$$\underbrace{\sum \left(y_i - \bar{y} \right)^2}_{\text{Total variation}} = \underbrace{\sum \left(y_i - \hat{y}_i \right)^2}_{\text{Unexplained variation}} + \underbrace{\sum \left(\hat{y}_i - \bar{y} \right)^2}_{\text{Explained variation}}$$

coefficient of determination:

$$R^2 = \frac{\Sigma \left(\hat{y}_i - \bar{y} \right)^2}{\Sigma \left(y_i - \bar{y} \right)^2} = 1 - \frac{\Sigma \left(y_i - \hat{y}_i \right)^2}{\Sigma \left(y_i - \bar{y} \right)^2} = 1 - \frac{\Sigma e_i^2}{\Sigma \left(y_i - \bar{y} \right)^2}$$

$$0 \le R^2 \le 1$$

Coefficient of determination is interpreted as the proportion of observed y variation that can be explained by the simple linear regression model (attributed to an approximate linear relationship between y and x).

$R^2 = 1 - \frac{1530}{15730} \approx \mathbf{0.9027}$. **90.27%**

To estimate σ^2:

Maximum Likelihood Estimator:

$$\hat{\sigma}^2 = \frac{1}{n} \sum \left(y_i - \hat{y}_i \right)^2 .$$

Simple linear regression sample variance:

$$s_e^2 = \frac{1}{n-2} \sum \left(y_i - \hat{y}_i \right)^2 .$$

$$s_e^2 = \frac{1}{n-2} \sum \left(y_i - \hat{y}_i \right)^2 = \frac{1}{n-2} \sum e_i^2 = \frac{1}{8} \cdot 1530 = \mathbf{191.25}.$$

$$s_e = \sqrt{191.25} \approx \mathbf{13.83}.$$

OR

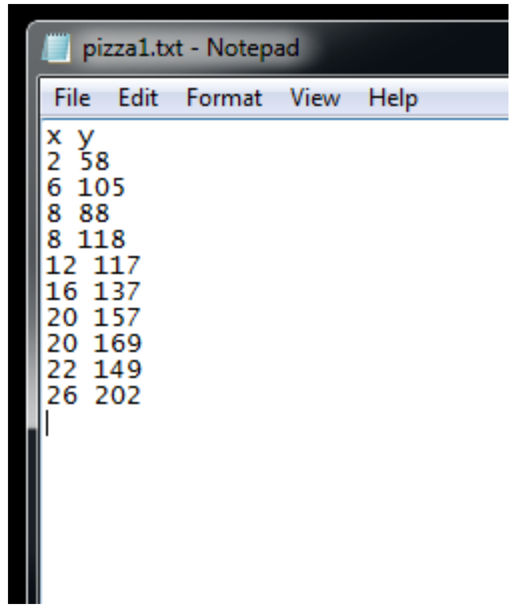
$$\hat{\sigma}^2 = \frac{1}{n} \sum \left(y_i - \hat{y}_i \right)^2 = \frac{1}{n} \sum e_i^2 = \frac{1}{10} \cdot 1530 = \mathbf{153}.$$

$$\hat{\sigma} = \sqrt{153} \approx \mathbf{12.37}.$$

STAT 420

```
## Simple Linear Regression
## Data
x = c(2,6,8,8,12,16,20,20,22,26)
y = c(58,105,88,118,117,137,157,169,149,202)
N = length(x)

## Estimate the regression line using pre-defined R functions
fit = lm(y ~ x)
fit
summary(fit)
names(fit)
names(summary(fit))

plot(x,y)
abline(fit$coefficients)

fit$fitted.values
fit$residuals
summary(fit)$sigma

sum(fit$residuals^2)
sum((y-fit$fitted.values)^2)

s2 = sum(fit$residuals^2)/(N-2); s2
s = sqrt(s2); s

## SXX, SXY, SYY
SXX = sum((x-mean(x))^2); SXX
SXY = sum((x-mean(x))*(y-mean(y))); SXY
SYY = sum((y-mean(y))^2); SYY
beta1hat = SXY/SXX; beta1hat
beta0hat = mean(y)-beta1hat*mean(x); beta0hat

## Matrix approach
Xmat = cbind(rep(1,N), x); Xmat
XX = t(Xmat) %*% Xmat; XX
XXinv = solve(XX); XXinv
XY = t(Xmat) %*% y; XY
betahat = XXinv %*% XY; betahat

## Predicting Y values
predict(fit,data.frame(x=10))
predict(fit,data.frame(x=38))


## Estimate the regression line using pre-defined R functions
fit1 = glm(y ~ x)
fit1
summary(fit1)
names(fit1)
```

```
## True relationship
beta0 = 10
beta1 = 5
truevar = 10

## x's fixed
N = 20
x = seq(25, 30, length=N)
trueline = beta0 + beta1*x

## Y data
yobs = trueline + rnorm(N, 0, sqrt(truevar))

plot(x, yobs)
lines(x, trueline, col=2)

regout = lm(yobs~x)
estline = regout$coeff[1] + regout$coeff[2]*x
lines(x, estline, col=1)


##### Sampling - plot of many estimates

plot(x, trueline, type="l", ylim=c(125, 170), col=2)
S = 100
for(s in 1:S){
yobs = trueline + rnorm(N, 0, sqrt(truevar))
regout = lm(yobs~x)
estline = regout$coeff[1] + regout$coeff[2]*x
lines(x, estline, col=1)
}
lines(x, trueline, type="l",col=2)


##### Sampling distribution of the estimators

simsize=1000
beta0est = c(1:simsize)
beta1est = c(1:simsize)
varest = c(1:simsize)

for (s in 1:simsize){
yobs = trueline + rnorm(N, 0, sqrt(truevar))
regout = lm(yobs~x)
beta0est[s] = regout$coeff[1]
beta1est[s] = regout$coeff[2]
varest[s] = sum((regout$resid)^2)/(N-2)
}

## Histogram of beta0est
hist(beta0est, nclass=10)

## Histogram of beta1est
hist(beta1est, nclass=10)

## Histogram of varest
hist(varest, nclass=10)
```



To create a data frame in R, use

```
> pizza1 = read.table(" ... /pizza1.txt", header = T)
```

or

```
> pizza1 = read.table(" ... /pizza1.txt", sep="", header = T)
```

`sep = ""` (the default for `read.table`) – the separator is 'white space'

`header = T` indicates that the first line of the data file contains the names for the variables (as opposed to `header = F`)

```
> pizza1
    x   y
1   2  58
2   6 105
3   8  88
4   8 118
5  12 117
6  16 137
7  20 157
8  20 169
9  22 149
10 26 202
```

You can then access individual variables in the data frame `pizza1` by using `pizza1$x` and `pizza1$y`:

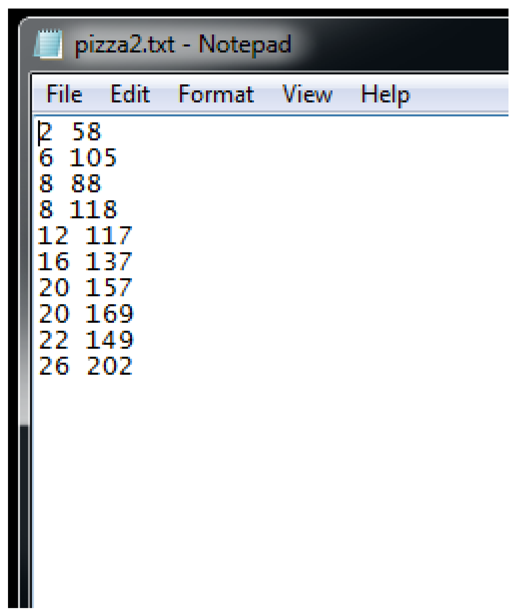
```
> pizza1$x
 [1]  2  6  8  8 12 16 20 20 22 26
> pizza1$y
 [1]  58 105  88 118 117 137 157 169 149 202
```

or

```
> fit = lm(y ~ x, data = pizza1)
```

```
> attach(pizza1)
```

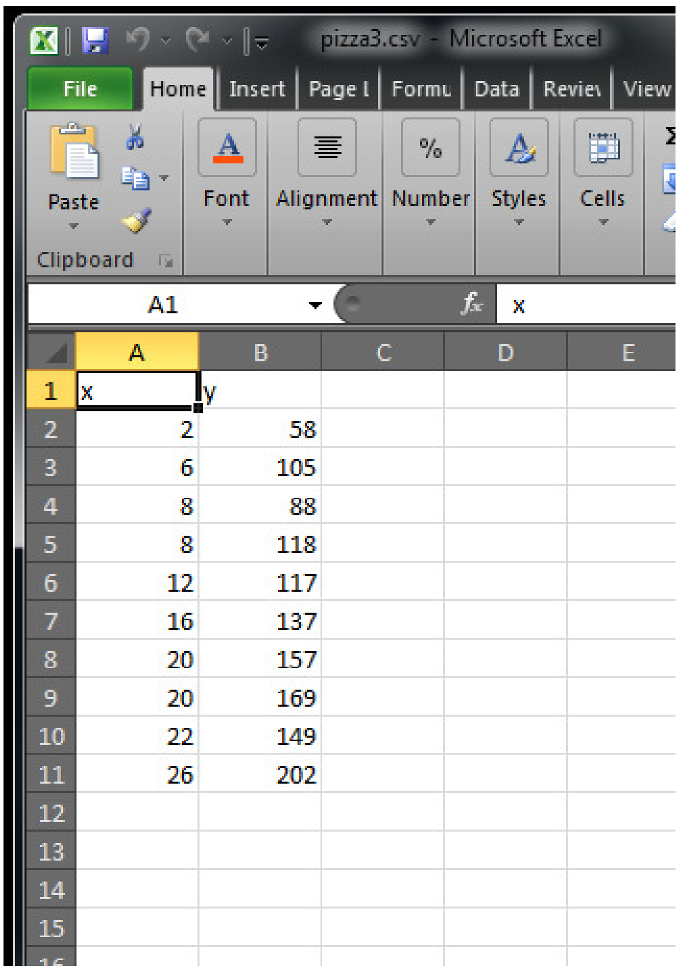
After `attach(pizza1)`, you may refer to the variables in the data set directly, i.e., `x` instead of `pizza1$x`.

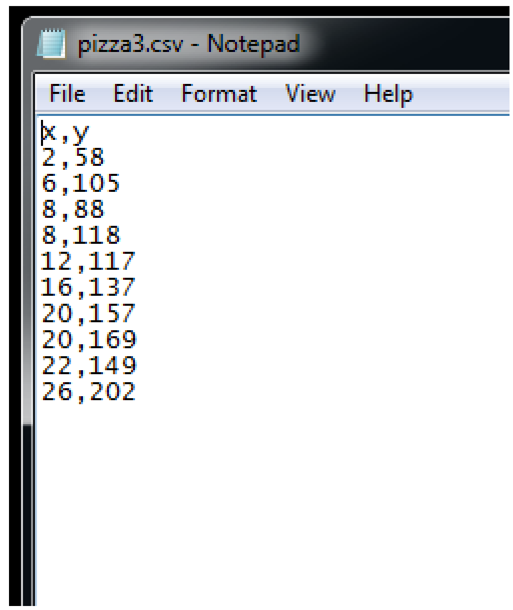

pizza2.txt - Notepad
File Edit Format View Help

```
2 58
6 105
8 88
8 118
12 117
16 137
20 157
20 169
22 149
26 202
```

```
> pizza2 = read.table(" ... /pizza2.txt", header = F)
> pizza2
   V1  V2
1   2  58
2   6 105
3   8  88
4   8 118
5  12 117
6  16 137
7  20 157
8  20 169
9  22 149
10 26 202
```

```
> pizza2$V1
 [1]  2  6  8  8 12 16 20 20 22 26
> pizza2$V2
 [1]  58 105  88 118 117 137 157 169 149 202
```

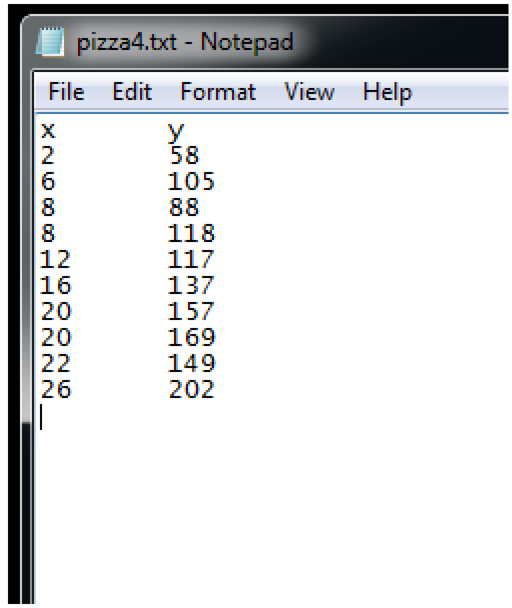




The **c**omma-**s**eparated **v**alues file format (comma-delimited format) is used to store tabular data in which numbers and text are stored in plain textual form. Lines in the text file represent rows of a table, and commas in a line separate what are fields in the row.

```
> pizza3 = read.table(" ... /pizza3.csv", sep = ",", header = T)
```

`sep = ","` indicates that the data in the data file are separated by a comma,



x	y
2	58
6	105
8	88
8	118
12	117
16	137
20	157
20	169
22	149
26	202

```
> pizza4 = read.table(" ... /pizza4.txt", sep = "\t", header = T)
```