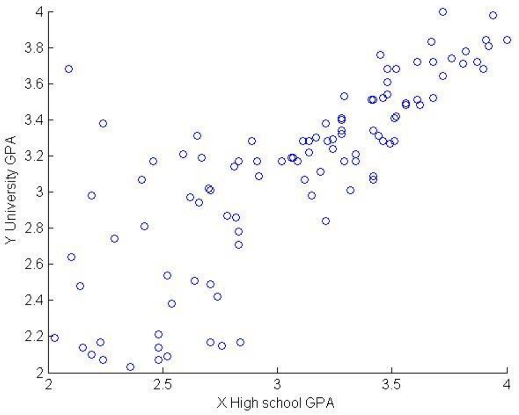
## Solutions

1 Copy-paste the data on pages 100-101 below into your Matlab editor, and define the variables

```
x = data(:,1); % High school GPA
y = data(:,4); % University GPA
```

a. Scatterplot

```
scatter(x,y)
```



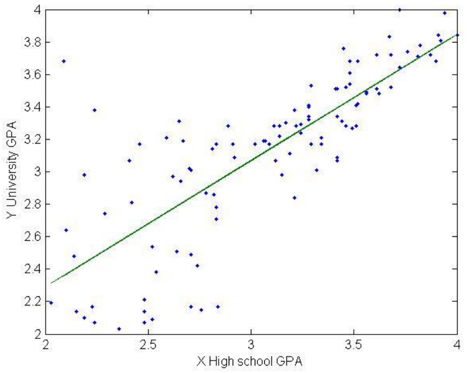
The scatter plot shows that high school GPAs are positively correlated with university GPAs.

b. Sample correlation coefficient

```
>> corr2(x,y)
ans =
    0.7915
```

c. Estimates of slope and intercept:

```
beta = sum((x-mean(x)).*(y-mean(y)))/sum((x-mean(x)).^2)
alpha = mean(y) - beta*mean(x)
figure
plot(x,y,'.',...
     x,alpha+beta*x)
xlabel('X High school GPA')
ylabel('Y University GPA')
```



d. What is the slope?

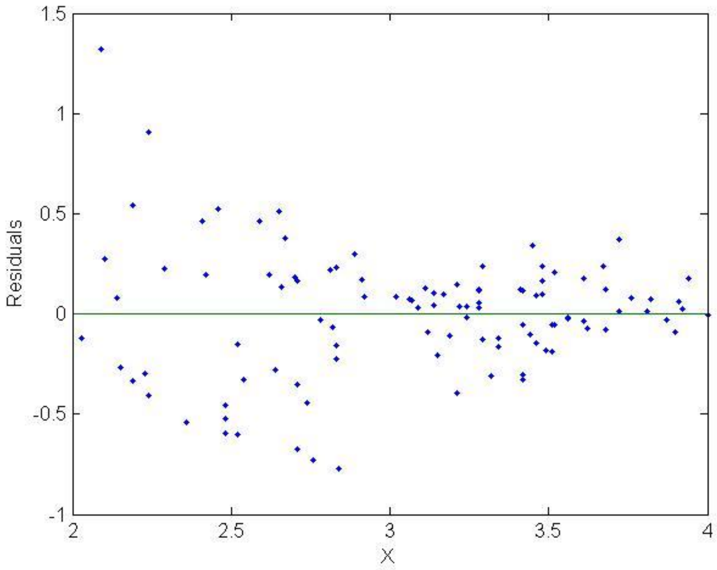
```
>> beta
beta =
    0.7799
```

e. What is the y-intercept?

```
>> alpha
alpha =
    0.7287
```

f. Residual plot

```
res=y-alpha-beta*x
figure
plot(x,res,'.',...
    [2 4],[0 0])
xlabel('X')
ylabel('Residuals')
```



The residual plot shows that there might be a relationship between x and the residuals, suggesting that there might be a better alternative to the linear model. The residuals are randomly distributed around zero, but the variance depends on x (low variance for high x, high variance for low x).

g. If someone had a 2.2 GPA in high school, what is the best estimate of his or her college GPA?

```
x0 = 2.2
y0 = alpha + beta*x0
y0 =
    2.4445
```

h. If someone had a 4.0 GPA in high school, what is the best estimate of his or her college GPA?

```
x0 = 4.0
y0 = alpha + beta*x0
y0 =
    3.8483
```

i. Test the null hypothesis that the slope is zero, $H_0\colon \beta = 0$.

```
N = length(x)
sr2 = 1/(n-2)*sum((y-(alpha+beta*x)).^2)
sr = sqrt(sr2)
t = (beta-0)/(sr*sqrt(1/sum((x-mean(x)).^2)))
pval=2*(1-tcdf(abs(t),n-2))

n =
   105
sr2 =
    0.0977
sr =
    0.3126
t =
    13.1422
pval =
     0
```

Since p<0.05, we reject the null hypothesis that the slope is zero.

j. Compute the 95% confidence interval for the intercept ($\alpha$).

```
t0 = tinv(0.975,n-2)
alpha_minus = alpha-t0*sr*sqrt(1/n+mean(x)^2/sum((x-mean(x)).^2))
alpha_plus = alpha+t0*sr*sqrt(1/n+mean(x)^2/sum((x-mean(x)).^2))

t0 =
    1.9833
alpha_minus =
    0.3616
alpha_plus =
    1.0958
```

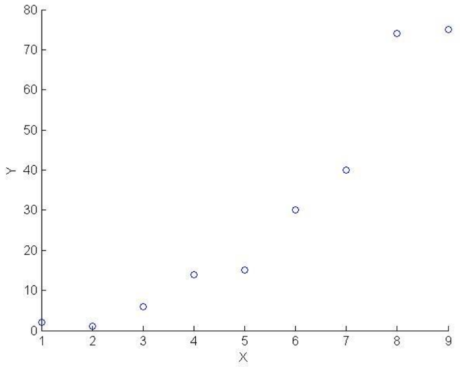
The endpoints of the 95% confidence interval are $\alpha_- = 0.3616$ and $\alpha_+ =$ 1.0958.

2. Given the following data

```
x = [ 1 2 3 4 5 6 7 8 9 ]
y = [ 2 1 6 14 15 30 40 74 75 ]
```

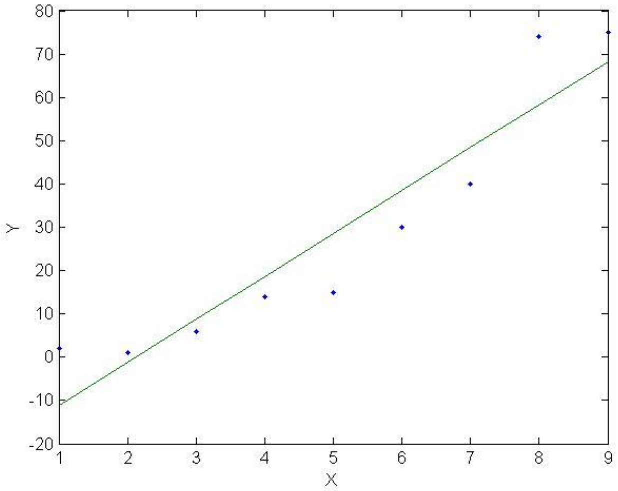
a. Scatter plot – looks like a non-linear relationship between x and y…

```
scatter(x,y)
```



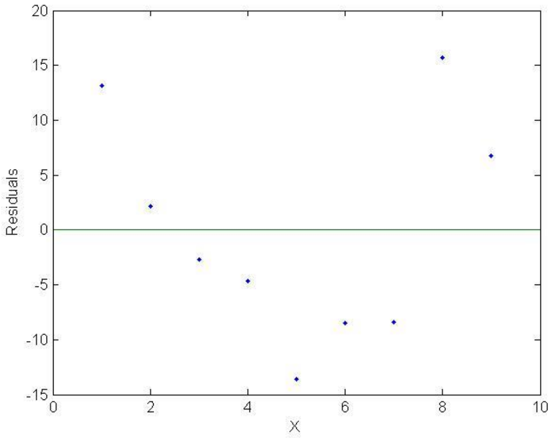
b. Find the regression line for predicting y from x.

```
beta = sum((x-mean(x)).*(y-mean(y)))/sum((x-mean(x)).^2)
alpha = mean(y) - beta*mean(x)
figure
plot(x,y,'.',...
     x,alpha+beta*x)
xlabel('X')
ylabel('Y')
```



c. Make a residual plot. Does the assumption of linearity seem to hold?

```
res=y-alpha-beta*x
plot(x,res,'.',...
    [0 10],[0 0])
xlabel('X')
ylabel('Residuals')
```



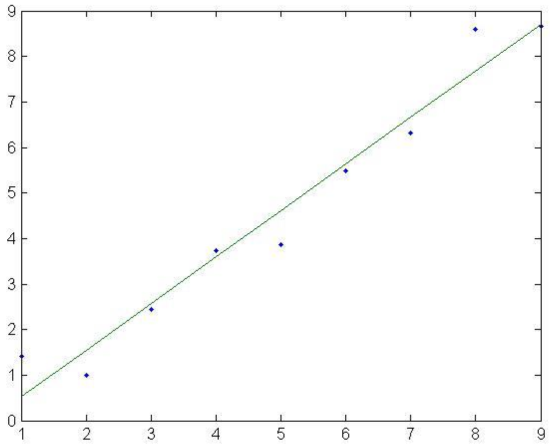
Notice the U-shape of the residual plot. This shape indicates that there is a non-linear relationship between x and y; the residuals are not randomly distributed around zero, they depend on x.

d. Transform y by taking the square root:

```
y = sqrt([ 2 1 6 14 15 30 40 74 75 ])
```

e. Find the regression line for predicting sqrt(y) from x.

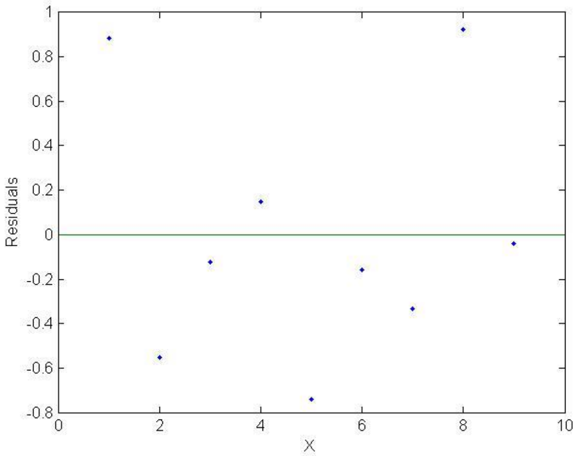
```
x = [ 1 2 3 4 5 6 7 8 9 ]
y = [ 2 1 6 14 15 30 40 74 75 ]
y = sqrt(y);
beta = sum((x-mean(x)).*(y-mean(y)))/sum((x-mean(x)).^2)
alpha = mean(y) - beta*mean(x)
plot(x,y,'.',...
     x,alpha+beta*x)
```



f. Make a residual plot. Does the assumption of linearity seem to hold?

```
res=y-alpha-beta*x
plot(x,res,'.',...
    [0 10],[0 0])
```

```
xlabel('X'),ylabel('Residuals')
```



The residuals are now randomly distributed around zero and do not depend on x. The variance of the estimates also appears to be independent of x. Hence, the assumption of linearity seems to hold.