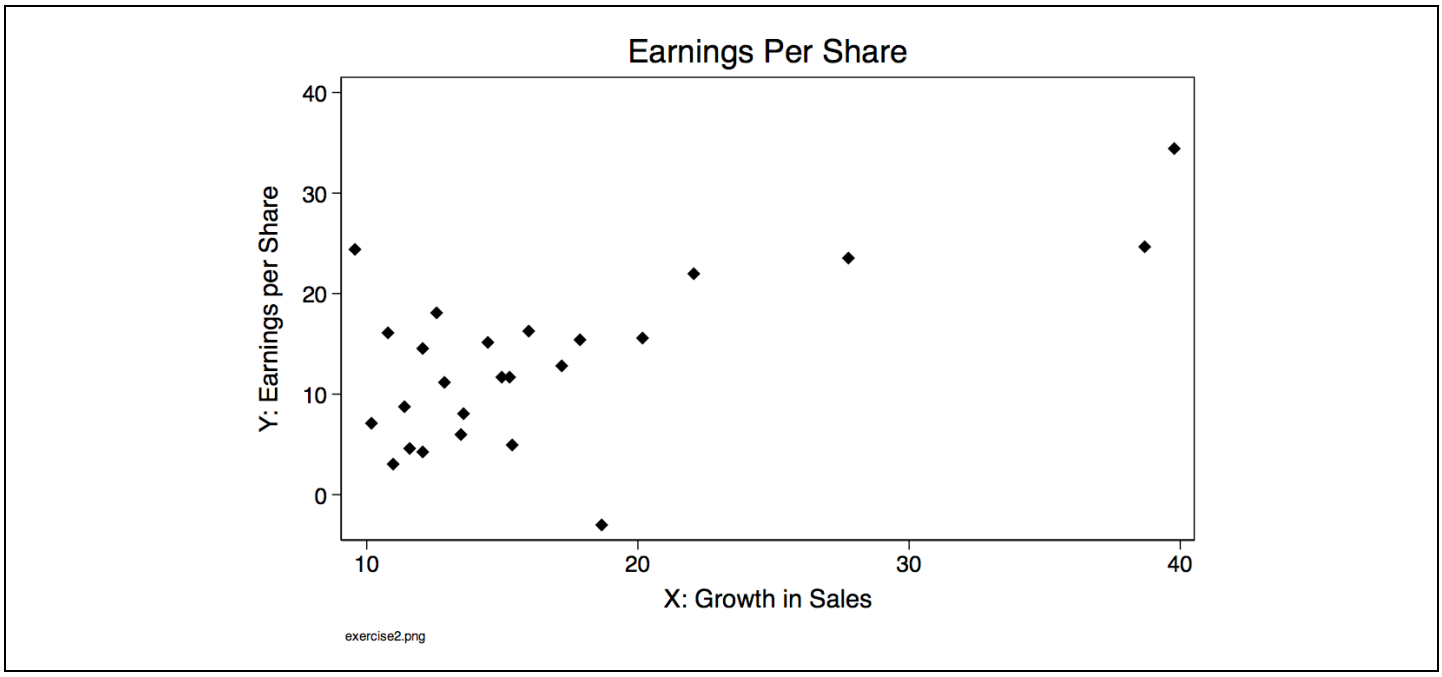
Unit 9 – STATA for Normal Theory Regression
Homework/Practice #11

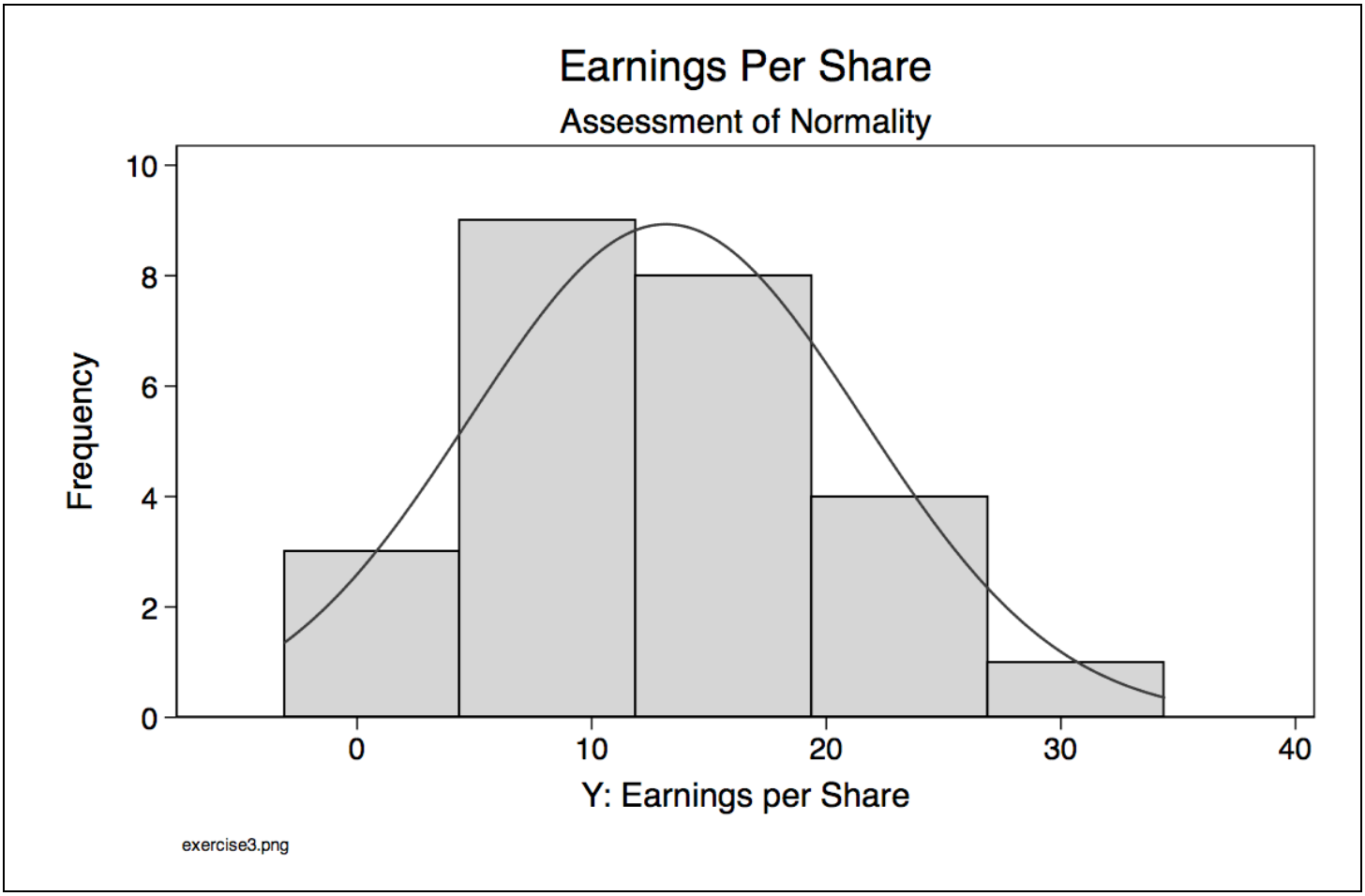
SOLUTIONS

```
. set more off

. *
. ***** Input data
. use "http://people.umass.edu/biep691f/data/companies.dta", clear


. *
. ***** 1).  Label variables
. label variable eps5 "Y: Earnings per Share"
. label variable salesgr5 "X: Growth in Sales"


. *
. ***** 2)  Scatterplot
. graph twoway (scatter eps5 salesgr5, symbol(d)), title("Earnings Per Share")
caption("exercise2.png", size(vsmall))
```



Tip – I like to put the name of my graph in the caption, for easy remembering later!

```
. *
. ***** 3)  Graphical Assessment of Normality of Y = eps5
. histogram eps5, normal frequency title("Earnings Per Share") subtitle("Assessment of
Normality") caption("exercise3.png", size(vsmall))
(bin=5, start=-3.0999999, width=7.5000003)
```



Interpretation: **The histogram with overlay normal suggests reasonableness of the assumption of normality of Y.**

```
. *
. ***** 4) Hypothesis Test of Normality of Y
. swilk eps5

                    Shapiro-Wilk W test for normal data

    Variable |     Obs       W           V         z       Prob>z
-------------+--------------------------------------------------
        eps5 |      25    0.97412       0.719    -0.674     0.74974
```

```
. sfrancia eps5

                  Shapiro-Francia W' test for normal data

    Variable |     Obs       W'          V'        z       Prob>z
-------------+--------------------------------------------------
        eps5 |      25    0.96853       0.971    -0.053     0.52129
```

Interpretation – These test results are consistent with the histogram and overlay normal. Both the Shapiro-Wilk and Francia tests of the null hypothesis of normality fail to reject (p-values: .75 and .52, respectively). Thus, we may assume that the assumption of normality of Y is reasonably satisfied for these data.

```
. *
. ***** 5) Estimate the straight line regression
. regress eps5 salesgr5

      Source |       SS       df       MS              Number of obs =      25
-------------+------------------------------           F(  1,    23) =   16.18
       Model |  695.264495     1  695.264495           Prob > F      =  0.0005
    Residual |  988.101157    23  42.9609199           R-squared     =  0.4130
-------------+------------------------------           Adj R-squared =  0.3875
       Total |  1683.36565    24  70.1402355           Root MSE      =  6.5545

------------------------------------------------------------------------------
        eps5 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    salesgr5 |   .6792279   .1688408     4.02   0.001     .3299542    1.028502
       _cons |   1.764971   3.124789     0.56   0.578    -4.699148    8.229091
------------------------------------------------------------------------------
```

(b_1 = .6792279, b_0 = 1.764971)

The fitted simple linear regression model is:

$$\widehat{eps5} = 1.76 + 0.679 * salesgr5$$

For each one percentage increase in annual compound growth of sales, it is estimated that there is a 0.679 percentage increase in annual compound growth in earnings per share

```
. *
. ***** 6a) How much of the variability in Y is explained by the model?
```

This is shown at right, as R-squared = 0.4130 which says that 41.3% of the variability in Y is explained by the fitted model

```
. *
. ***** 6b) Overall F-Test
```

This is shown at right, as **Prob > F = 0.0005** which says that the p-value for the overall F test of the null hypothesis that the fitted simple linear regression has slope=0 is 5 chances in 10,000. The null hypothesis of zero slope has led to a very unlikely outcome prompting statistical rejection of the null hypothesis. Conclude that the fitted line explains statistically significantly more of the variability in Y=eps5 than is explained by the fit of no model.
Reminder - The "fit of no model" is really saying that the model fit is instead the average of Y.

```
. *
. ***** 7a) Test of Zero Slope
```

This is shown in the coefficients table, as **P>|t|= 0.001** which says that the p-value for the Student t-test of the null hypothesis of zero slope is 1 chance in 1,000. Again, the null hypothesis has led to a very unlikely outcome, prompting statistical rejection of the null hypothesis. Conclude that the fitted line explains statistically significantly more of the variability in Y=eps5 than is explained by the average of Y (no model).

NOTE - In theory, in simple linear regression, the F-test for the overall regression is equivalent to the Student t-test for zero slope, with [Student t]^2 = [Overall F]. The reason the p-values do not match exactly is the result of rounding.

```
. *
. ***** 7b) Confidence Interval Estimate of the Slope
```

This is shown in the coefficients table, as **(.3299542 , 1.028502)**