# Unit 9 – STATA for Normal Theory Regression
## Homework

**Due: Friday December 4, 2020**

**Before you Begin**
**Download from the course website**
**companies.dta**
**hersdata_small.dta**

### Description of companies.dta

This data set contains 30 observations. It is from a study reported in the January 1981 issue of *Forbes* magazine of the characteristics of the 30 largest chemical companies. The two variables that we will use in a simple linear regression analysis are the following:
**eps5** and **salesgr5**

### Codebook

| Variable Name | Variable Coding | Description |
|---|---|---|
| **salesgr5** | Continuous | **Independent (X)**: Per cent annual compound growth rate of sales, computed from the most recent five years compared with the previous five years. |
| **eps5** | Continuous | **Dependent (Y)**: Per cent annual compound growth in earnings per share, computed from the most recent five years compared with the preceding five years. |

---

**Simple Linear Regression Using companies.dta**

___1. Label the variable eps5 as y: earnings per share
Label the variable salesgr5 as x: growth in sales

___2. Produce an X-Y scatter plot of your data with any aesthetics you like!
**In 1-2 sentences, interpret**.

___3. Using any graphical approach of your choosing, assess the normality of Y = eps5
**In 1-2 sentences, interpret**.

___4. Using any numerical test of your choosing, test the normality of Y = eps5
**In 1-2 sentences, interpret**.

___ 5. Estimate the straight-line regression.
**In 1-2 sentences, provide annotations of the results**.

___6 (a) How much of the variability in Y = eps5 is explained by the fitted model?
(b) Perform the overall F-test of the significance of the fitted line.
**In 1-2 sentences, interpret**.

__7. (a) Test the null hypothesis that the slope of the regression of Y on X is zero
(b) Produce a 95% confidence interval estimate of the slope of the regression of Y on X.
**In 1-2 sentences, interpret**.

**Multiple Linear Regression Using hersdata_small.dta**

**Description of hersdata_small.dta**

These data are a simple random sample of 1000 observations from the HERS study called *hersdata_small.dta*. The HERS study was a randomized clinical trial of hormone therapy (estrogen plus progestin) for the reduction of cardiovascular disease risk in post-menopausal women with established coronary disease. Study participants were n=2,763 women who were: (1) post-menopausal (2) with coronary disease; and (3) with an intact uterus. The data set for this homework is a simple random sample of n=1000.

**The following variables are considered (*Tip* – Remember that Stata is case sensitive)**

```
. summarize age BMI drinkany exercise glucose HT LDL physact statins
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| age | 1000 | 66.701 | 6.575168 | 45 | 79 |
| BMI | 999 | 28.31502 | 5.45297 | 15.21 | 54.13 |
| drinkany | 999 | .4134134 | .4926924 | 0 | 1 |
| exercise | 1000 | .393 | .4886612 | 0 | 1 |
| glucose | 1000 | 111.214 | 35.46136 | 29 | 298 |
| HT | 1000 | .508 | .5001862 | 0 | 1 |
| LDL | 997 | 144.6558 | 36.84347 | 44.4 | 393.4 |
| physact | 1000 | 3.219 | 1.091893 | 1 | 5 |
| statins | 1000 | .365 | .4816711 | 0 | 1 |

## Exercises # 8-13
## Multiple Linear Regression Analysis of Y=Glucose, among non-diabetics only

___8. Fit a single predictor model of Y=**glucose** to X= **exercise** among non-diabetics ONLY. In 1-2 sentences, report and interpret the output.

___9. Next fit a multiple predictor model of Y= **glucose** among non-diabetics ONLY. Fit the following predictors: **exercise, age, drinkany,** and **BMI.** In 1-2 sentences, interpret the output.

___ 10. Perform a partial F-test for the significance of **exercise** controlling for **age, drinkany,** and **BMI.** In 1-2 sentences, interpret.

___11. Create four 0/1 design variables to represent the 5 possible outcomes of **physact.**
Using a command of your choosing, produce a check on the creation of your design variables.

___12. Fit a multiple predictor model of Y= **glucose** among non-diabetics ONLY. Consider as the predictor ONLY the design variables for **physact.** In 1-2 sentences, interpret the output.

___ 13. Repeat exercise #12. However, this time, use the following stata command

**xi: regress glucose i.physact if diabetes == 0**

## Exercises # 14-19
## Multiple Linear Regression Analysis of Y=LDL, entire sample.

___14. Perform a two-way contingency table analysis of **HT** and **statins.** Interpret.

___15. Fit a single predictor model of Y=**LDL** to X= **HT.** Interpret.

___16 Fit a single predictor model of Y=**LDL** to X= **statins.** Interpret.

___17. Create a new variable called **HTstatins** that is an interaction of **HT** and **statins.**

___18. Fit a three predictor model of Y=**LDL** using the predictors: **HT,statins** and **HTstatins.**
In 1-2 sentences, report and interpret the output.

___ 19. Perform a partial F-test for the significance of **HTstatins** controlling for **HT** and **statins.**
Interpret.