

## Unit 9 Stata for Normal Theory Regression version 16

*“Assume that a statistical model such as a linear model  
is a good first start only”*

*- Gerald van Belle*

**Normal theory regression analysis** explores the relationship of one outcome that is continuous (e.g.  $Y$  = birth weight) with one or more predictors that can be continuous or discrete (e.g.  $X_1$  = months gestation,  $X_2$  = yes/no indicator of mother’s smoking status,  $X_3$  = mother’s systolic blood pressure, and so on).

In **simple linear regression**, the number of predictors is **one** and **continuous** (eg  $X$ =mother’s systolic blood pressure).

In **multiple linear regression**, the number of predictors is **two or more** and can be both **continuous and discrete**

The **goal** is to explain the variation in the outcomes (the  $Y$  variable) with a “good” model that is a function of the predictors (the  $X$  variables) that is as “small” as possible. The challenge is in how to achieve both “good” (close fit) and “small” (parsimony) simultaneously.

Ultimately, we don’t know if our model is correct and most likely it is not. Nevertheless, a model that is “good” and “small” has a variety of **uses**:

### **Hypothesis Tests and Confidence Intervals**

We can ask such questions as: “Is the experimental treatment is associated with a statistically significant benefit?”

### **Prediction**

We can use the estimating equation to make confidence interval predictions such as: the survival time following surgery of a future patient undergoing coronary bypass surgery.

### **Insights into Nature**

Sometimes, the fitted model derives from a physical-equation. An example is Michaelis-Menton kinetics. A Michaelis-Menton model is fit to the data for the purpose of estimating the actual rate of a particular chemical reaction.

## Table of Contents

Topic	Page
Learning Objectives .....	3
1. Introduction .....	4
1.1 Settings Where Regression Might be Considered .....	4
1.2 Review - What is Statistical Modeling .....	7
1.3 A General Approach for Model Development .....	8
1.4 Review - Normal Theory Regression .....	9
2. Example of Stata to Perform Normal Theory Regression .....	12
3. Exploratory Data Analysis, Indicator Variables, and Interactions .....	29
3.1 Exploratory Data Analysis .....	29
3.2 How to Create Indicator Variables .....	33
3.3 How to Create Interactions .....	35
3.4 How to Create Quartiles (or other groupings).....	35
4. Simple Linear Regression (Bivariate Analyses) .....	36
5. Multiple Linear Regression and Choose “Tentative” Final Model .....	37
5.1 Estimation .....	37
5.2 Hierarchical Model Comparisons .....	38
6. Regression Diagnostics: Model Assumptions and Model Adequacy .....	40
6.1 Linearity :.....	41
6.2 Normality of Residuals.....	41
6.3 Multicollinearity .....	42
6.4 Model Misspecification .....	42
6.5 Constant Variance .....	43
6.6 Outlying, High Leverage and Influential Points .....	44
7. Post Regression: Prediction and Reporting .....	46
7.1 Predictions .....	46
7.2 Show Models Side-by-Side .....	47
7.3 Plot Predicted Values .....	48

## Learning Objectives

When you have finished this unit, you should be able to:

- **Define** the simple and multiple linear regression models;
- State and explain the **assumptions** for normal theory linear regression analysis;
- Use Stata to **explore a data set** (numerical descriptions, scatterplots, etc) prior to model estimation;
- Use Stata to **create design variables** for use in the modeling of categorical explanatory variables;
- Use Stata to **fit (estimate)** a normal theory regression model;
- **Interpret a fitted model**, including the regression coefficients, standard errors,  $R^2$ , sums of squares, analysis of variance, t-tests, and F-tests;
- Explain **confounding** and **effect modification**;
- Use Stata to **assess confounding and modification** in a normal theory regression;
- Use Stata to perform **hypothesis tests** and obtain **confidence intervals**;
- Use Stata to produce **post-estimation graphical summaries** of model fit;
- Use Stata to perform **regression diagnostics** to assess model adequacy for a normal theory regression; and
- **Write a 1-2 paragraph interpretation** of a normal theory regression analysis.

# 1. Introduction

## 1.1 Settings Where Regression Might Be Considered

### Example #1

#### Are Emergency Calls to the New York Auto Club Related to the Weather?

Source:

Chatterjee, S; Handcock MS and Simonoff JS *A Casebook for a First Course in Statistics and Data Analysis*. New York, John Wiley, 1995, pp 145-152.

Are calls to the New York Auto Club related to the weather, with more calls occurring during bad weather? To explore this possibility, the NY Auto Club obtained observations on numbers of calls to the New York Auto Club (Y=calls) together with several kinds of information about the weather on the day of the call. Among the analyses they performed was a **simple linear regression** with outcome (dependent) variable Y and predictor (explanatory) variable X, both continuous, defined:

Y = calls (number of calls)  
X = low (the lowest temperature of the day).

Dear reader: Strictly speaking, the variable Y=calls is discrete, not continuous. In this example, however, the sample size was large and the distribution of calls was approximated well with the assumption of normality. So, the normal theory linear regression went forward!

### Example #2

#### Does the expression of p53 change with parity and age?

Source:

Matthews et al. *Parity Induced Protection Against Breast Cancer* 2007.

P53 is a human gene that is a tumor suppressor gene. Malfunctions of this gene have been implicated in the development and progression of many cancers, including breast cancer. Matthews et al were interested in exploring the relationship of Y=p53 expression to parity and age at first pregnancy, after adjustment for other, established, risk factors for breast cancer, including: age at first mensis, family history of breast cancer, menopausal status, and history of oral contraceptive use.

- Among the initial analyses, a **simple linear regression** might be performed to obtain a thorough understanding of the relationship of p53 expression and age. Both the outcome (Y) and the predictor (X) are continuous.

Y = p53 expression  
X = Age

- A **multiple linear regression** might then be performed to see if age and parity retain their predictive significance, after controlling for the other, known, risk factors for breast cancer. Thus, the analysis would consider one outcome variable (Y) and 6 predictor variables (X<sub>1</sub>, X<sub>2</sub>, X<sub>3</sub>, X<sub>4</sub>, X<sub>5</sub>, X<sub>6</sub>):

Y = p53  
 X<sub>1</sub> = Age  
 X<sub>2</sub> = Parity  
 X<sub>3</sub> = Age at first menses  
 X<sub>4</sub> = Family history of breast cancer  
 X<sub>5</sub> = Menopausal status  
 X<sub>6</sub> = History of oral contraceptive use

### Example #3

#### Does Air Pollution Reduce Lung Function?

##### Source:

Detels et al (1979) *The UCLA population studies of chronic obstructive respiratory disease. I. Methodology and comparison of lung function in areas of high and low pollution. Am. J. Epidemiol. 109: 33-58.*

Detels et al (1979) investigated the relationship of lung function to exposure to air pollution among residents of Los Angeles in the 1970's. Baseline and follow-up measurements of exposure and lung function were obtained. Also obtained were measurements of selected other variables that the investigators suspected might confound or modify the effects of pollution on lung function: age, sex, height, weight, etc. Afifi, Clark and May (2004) consider portions of this data in their 2004 text, Computer-Aided Multivariate Analysis, Fourth Edition (Chapman & Hall)

- It is already known that a person's FEV is related to their height. Thus, an analysis of the effects of air pollution might begin with a **simple linear regression** analysis of the relationship between FEV and height before moving on to an examination of the effects of exposure to air pollution:

Y = FEV, liters  
 X = Height, inches

- A **multiple linear regression** might then be performed to determine the nature and strength of exposure to pollution for the prediction of lung function, taking into account the role of height and other influences on lung function, such as age, smoking, etc. For example, the relationship of lung function to exposure to air pollution might be different for smokers and non-smokers; this would be an example of effect modification (interaction). It might also be the case that the relationship of lung function to exposure to air pollution is confounded by height. Here, we would have something like:

Y = FEV, liters  
 X<sub>1</sub> = Exposure to air pollution  
 X<sub>2</sub> = Height, inches  
 X<sub>3</sub> = Smoking (1=yes, 0=no)

## Example #4

### Exercise and Glucose for the Prevention of Diabetes

#### Source:

Hulley et al (1998) *Randomized trial of estrogen plus progestin for secondary prevention of heart disease in postmenopausal women. The Heart and Estrogen/progestin Study. JAMA 280(7): 605-13.*

In the HERS study, Hulley et al. (1998) sought to determine if exercise, a modifiable behavior, might lower the risk of diabetes in non-diabetic women who were at risk of developing the disease. The question is a complex one because there are many risk factors for diabetes. Moreover, the type of woman who chooses to exercise may be related in other ways to risk of diabetes, apart from the fact of her exercise habit. For example, women who exercise regularly are typically younger and have lower body mass index (BMI); these characteristics also confer a risk benefit with respect to diabetes. Finally, the benefit of exercise may be mediated through a reduction of body mass index. Vittinghoff, Glidden, Shiboski and McCulloch (2005) consider portions of this data in their 2005 text, *Regression Methods in Biostatistics: Linear, Logistic, Survival and Repeated Measures Models* (Springer).

- A **multiple linear regression** was performed to assess the benefit of exercising at least three times/week, compared to no exercise, on blood glucose, after controlling for other factors associated with blood glucose levels. Thus, here we would have something like:

Y = Glucose, mg/dL

X<sub>1</sub> = Exercise (1=yes if 3x/week or more, 0 = no)

X<sub>2</sub> = Age, years

X<sub>3</sub> = Body Mass Index (BMI)

X<sub>4</sub> = Alcohol Use (1=yes, 0=no)

## 1.2 Review - What is Statistical Modeling

George E.P. Box, a very famous statistician, once said, “*All models are wrong, but some are useful.*” Incorrectness notwithstanding, we do statistical modeling for a very good reason: we seek an understanding of the natures and strengths of the relationships (if any) that might exist in a set of observations that co-vary.

For any set of observations, theoretically, lots of models are possible. So, how to choose? The **goal** of statistical modeling is to obtain a model that is simultaneously **minimally adequate** and a **good fit**. **The model should also make sense.**

### Minimally adequate

- Each predictor is “important” in its own right
- Each extra predictor is retained in the model only if it yields a significant improvement (in fit and in variation explained).
- The model should not contain any redundant parameters.

### Good Fit

- The amount of variability in the outcomes (the Y variable) explained is a lot
- The outcomes that are predicted by the model are close to what was actually observed.

### The model should also make sense

- A preferred model is one based on “subject matter” considerations
- The preferred predictors are the ones that are simply and conveniently measured.

It is not possible to choose a model that is simultaneously minimally adequate and a perfect fit.  
Model estimation and selection must achieve an appropriate balance.

### 1.3 A General Approach for Model Development

There are **no** rules **nor a single best strategy**. Different study designs and research questions call for different approaches for model development. **Tip** – Before you begin model development, make a list of your study design, research aims, outcome variable, primary predictor variables, and covariates.

As a general suggestion, the following approach has the advantages of providing a reasonably thorough **exploration of the data and relatively little risk of missing something important**

**Preliminary** – Be sure you have: (1) checked, cleaned and described your data, (2) screened the data for possible associations, and (3) thoroughly explored the bivariate (also called “single predictor”, “unadjusted”, “crude”) relationships.

#### **Step 1 – Fit the “maximal” model.**

The maximal model is the large model that contains all the explanatory variables of interest as predictors. This model also contains all the covariates that might be of interest. It also contains all the interactions that might be of interest. Note the amount of variation explained.

#### **Step 2 – Begin simplifying the model.**

Inspect each of the terms in the “maximal” model with the goal of removing the predictor that is the least significant. Drop from the model the predictors that are the least significant, beginning with the higher order interactions (**Tip** -interactions are complicated and we are aiming for a simple model). Fit the reduced model. Compare the amount of variation explained by the reduced model with the amount of variation explained by the “maximal” model.

If the deletion of a predictor has little effect on the variation explained  
Then leave that predictor out of the model.  
And inspect each of the terms in the model again.

If the deletion of a predictor has a significant effect on the variation explained  
Then put that predictor back into the model.

#### **Step 3 – Keep simplifying the model.**

Repeat step 2, over and over, until the model remaining contains nothing but significant predictor variables.

#### **Beware of some important caveats**

- Sometimes, you will want to keep a predictor in the model regardless of its statistical significance (an example is randomization assignment in a clinical trial)
- The order in which you delete terms from the model matters
- You still need to be flexible to considerations of biology and what makes sense.



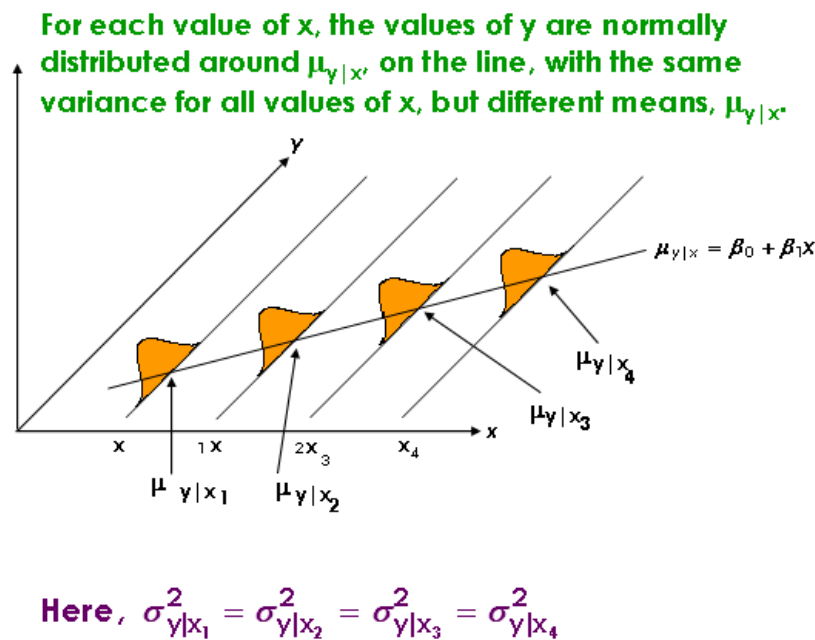
## 1.4 Review - Normal Theory Regression

Normal theory regression analysis is used to investigate possibly complex relationships when:

- The outcome is a **single continuous variable (Y)** that can reasonably be assumed to be **distributed normal**; and
- The outcome is potentially related to possibly **several predictor variables ( $X_1, X_2, \dots, X_p$ )** which can be **continuous or discrete**; and
- Some of the predictor variables might **confound** the prediction role of other explanatory variables; and
- Some of the predictor-outcome relationships may be different (are modified by) depending on the level of one or more different predictor variables (**interaction**)

### Simple Linear Regression:

A simple linear regression model is one for which the mean  $\mu$  (the average value) of **one continuous, and normally distributed, outcome** random variable  $Y$  (e.g.  $Y = \text{FEV}$ ) varies linearly with changes in **one continuous predictor** variable  $X$  (e.g.  $X = \text{Height}$ ). It says that the expected values of the outcome  $Y$ , as  $X$  changes, lie on a straight line (“regression line”).



## Assumptions

1. The outcomes  $Y_1, Y_2, \dots, Y_n$  are **independent**.
2. The values of the predictor variable  $X$  are fixed and measured without error.
3. At each value of the predictor variable  $X=x$ , the distribution of the outcome  $Y$  is **normal** with

$$\begin{aligned}\text{mean} &= \mu_{Y|X=x} = \beta_0 + \beta_1 x \\ \text{variance} &= \sigma_{Y|x}^2.\end{aligned}$$

## Model

These assumptions mean that we are considering the following **model**. For individual “i”,

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \text{ where}$$

1. The errors  $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$  are **independent**.
2. Each error  $\varepsilon_i$  is distributed is **normal** with

$$\begin{aligned}\text{mean} &= 0 \\ \text{variance} &= \sigma_{Y|x}^2.\end{aligned}$$

## Multiple Linear Regression:

In multiple linear regression, there is still just **one outcome variable, continuous**. The term “multiple” refers to there being **more than one predictor variable**.

### Definition

A multiple linear regression model is a particular model in which the mean  $\mu$  of **one continuous** outcome random variable  $Y$  (e.g.  $Y = \text{FEV}$ ) varies linearly with changes in two or more predictor variables  $X_1, X_2$ , etc. (e.g.  $X_1 = \text{Height}$ ,  $X_2 = \text{Smoking}$  ( $1 = \text{yes}$ ,  $0 = \text{no}$ )). The predictor variables can be continuous, discrete, or both. A multiple linear regression model says that the expected values ( $\mu$ ) of the outcome  $Y$ , as  $X_1, X_2$ , etc change, lie on a plane (“regression plane”).

### Assumptions

The assumptions required are an extension of those for simple linear regression.

1. The outcomes  $Y_1, Y_2, \dots, Y_n$  are **independent**.
2. The values of the predictor variables  $X_1 \dots X_p$  are fixed and measured without error.
3. For each fixed profile of values,  $x_1, x_2, \dots, x_p$ , of the  $p$  predictor variables  $X_1 \dots X_p$  (written using vector notation  $\underline{X} = \underline{x}$ ), the distribution of values of  $Y$  is **normal** with

$$\text{mean} = \mu_{Y|\underline{X}=\underline{x}} = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

$$\text{variance} = \sigma_{Y|\underline{X}=\underline{x}}^2.$$

### Model

Our model is now:

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \varepsilon$$

- $p = \#$  predictors, **apart** from the intercept
- Each  $X_1 \dots X_p$  can be either discrete or continuous.

## 2. Example of Stata to Perform Normal Theory Regression

### **How to follow along:**

**Download from the course website.**

**framingham\_1000.dta**

### Source:

Levy (1999) *National Heart Lung and Blood Institute. Center for Bio-Medical Communication. Framingham Heart Study*

### Description:

Cardiovascular disease (CVD) is the leading cause of death and serious illness in the United States. In 1948, the Framingham Heart Study - under the direction of the National Heart Institute (now known as the National Heart, Lung, and Blood Institute or NHLBI) was initiated. The objective of the Framingham Heart Study was to identify the common factors or characteristics that contribute to CVD by following its development over a long period of time in a large group of participants who had not yet developed overt symptoms of CVD or suffered a heart attack or stroke.

Here we will use a subset of the data comprised of information on 9 variables in a subset of n=1000.

**Note – some of the variables shown here will be created in the pages that follow.**

Variable	Label	Codings
sbp	Systolic Blood Pressure (mm Hg)	
ln_sbp	Natural logarithm of sbp	ln_sbp=ln(sbp)
age	Age, years	
bmi	Body Mass index (kg/m <sup>2</sup> )	
ln_bmi	Natural logarithm of bmi	ln_bmi=ln(bmi)
sex	Gender	1=male 2=female
female	Female Indicator	0 = male 1 = female
scl	Serum Cholesterol (mg/100 ml)	
ln_scl	Natural logarithm of scl	ln_scl=ln(scl)

### Multiple Regression Variables:

Outcome Y = ln\_sbp

Predictor Variables: ln\_bmi, ln\_scl, age, sex

### Research Question:

From among these 4 “candidate” predictors, what are the important “risk” factors and what is the nature of their association with Y=ln\_sbp?

The basic steps in this illustration are the following and correspond to the general approach to model development introduced on page 8.

### Plan of Illustration

#### **Step 1 – Exploratory Data Analysis, Indicator Variables, and Interactions.**

Examine descriptive statistics, assess normality of the dependent variable, consider a “normalizing” transformation if needed, create indicator variables, create interaction variables

#### **Step 2 – Examine Bivariate Relationships.**

Look at the relationship of the dependent variables (Y) with each of the candidate predictor variables (X). Look at these relationships graphically and test correlations. Consider transformations of the predictor variables if needed.

#### **Step 3 – Fit Models and Choose “Tentative” Final Model.**

Fit an initial model. Fit alternative models. Compare competing models with partial F-tests and side-by-side comparisons of estimated regression coefficients, percent variance explained (R-squared), and mean squared error. Choose a “tentative” final model.

#### **Step 4 – Regression Diagnostics.**

Fit again the “tentative” final model; this is a necessary preliminary to doing most regression diagnostics. Check model assumptions. Check model adequacy.

#### **Step 5 – Repeat steps #3 and #4 as needed.**

#### **Step 6 – Report Regression Results.**

Produce appropriate tabulations of regression results. Produce graphical summaries of the “final” model. Interpret.

### Step 1 – Exploratory Data Analysis, Indicator Variables, and Interactions.

Examine descriptive statistics, assess normality of the dependent variable, consider a “normalizing” transformation if needed, create indicator variables, create interaction variables.

```
. * ----- Preliminary: Check variables with respect to definition, # obs, missing, range, etc.
. codebook sex sbp scl age bmi id, compact
```

Variable	Obs	Unique	Mean	Min	Max	Label
sex	1000	2	1.557	1	2	Sex
sbp	1000	87	132.35	80	270	Systolic Blood Pressure
scl	996	182	227.8464	115	493	Serum Cholesterol
age	1000	36	45.922	30	66	Age in Years
bmi	998	186	25.56623	16.4	43.4	Body Mass Index
id	1000	1000	2410.031	1	4697	

Interpretation: 1) sex is coded 1 or 2; 2) we are missing 4 observations of scl and 2 observations of bmi.

```
. * ----- 1) Create a data set comprised of complete observations ONLY. Create new vars. Save.
. drop if scl>=.|bmi>=.
(6 observations deleted)
```

```
. codebook sex sbp scl age bmi id, compact
```

Variable	Obs	Unique	Mean	Min	Max	Label
sex	994	2	1.557344	1	2	Sex
sbp	994	87	132.3702	80	270	Systolic Blood Pressure
scl	994	182	227.8773	115	493	Serum Cholesterol
age	994	36	45.92153	30	66	Age in Years
bmi	994	186	25.57706	16.4	43.4	Body Mass Index
id	994	994	2409.462	1	4697	Subject id

```
. generate ln_scl=log(scl)
. generate ln_sbp=ln(sbp)
. generate ln_bmi=ln(bmi)
. generate female=(sex==2)
```

```
. label variable ln_sbp "Natural logarithm (sbp)"
. label variable ln_bmi "Natural logarithm (bmi)"
. label variable ln_scl "Natural logarithm (scl)"
. label variable female "Female (0/1)"
```

```
. * ----- Save complete data as framingham_complete.dta
. save "/Users/cbigelow/Desktop/framingham_complete.dta"
file /Users/cbigelow/Desktop/framingham_complete.dta saved
```

. \* ----- 2) Numerical descriptives to examine data for shape, range, outliers and completeness.

```
. tabstat sbp ln_sbp age bmi ln_bmi scl, statistics(n mean sd min q max) columns(statistics)
format(%8.2f)
```

variable	N	mean	sd	min	p25	p50	p75	max
sbp	994.00	132.37	22.99	80.00	116.00	128.00	144.00	270.00
ln_sbp	994.00	4.87	0.16	4.38	4.75	4.85	4.97	5.60
age	994.00	45.92	8.53	30.00	39.00	45.00	53.00	66.00
bmi	994.00	25.58	3.85	16.40	23.00	25.10	27.80	43.40
ln_bmi	994.00	3.23	0.15	2.80	3.14	3.22	3.33	3.77
scl	994.00	227.88	45.10	115.00	197.00	225.00	255.00	493.00

. fre sex

sex -- Sex

		Freq.	Percent	Valid	Cum.
Valid	1 Men	440	44.27	44.27	44.27
	2 Women	554	55.73	55.73	100.00
	Total	994	100.00	100.00	

Dear Reader: The following assessment of normality is for illustration. In actuality, we already know that we will be using  $Y=\ln(\text{sbp})$  as our dependent variable.

```
. * ----- 3) Assess normality of "candidate" dependent variable Y=ln_sbp
. * sfrancia test of normality (Null: distribution is normal)
. sfrancia sbp
```

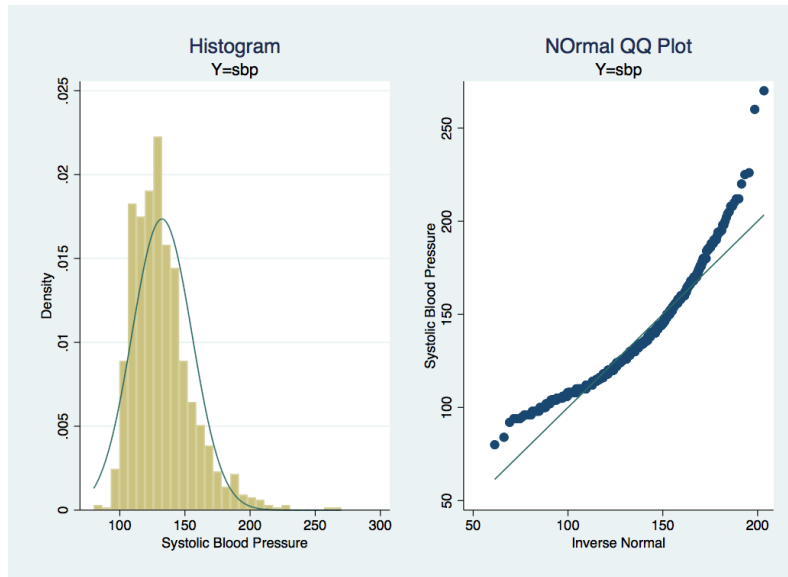
Shapiro-Francia W' test for normal data

Variable	Obs	W'	V'	z	Prob>z
sbp	994	0.92193	52.000	9.055	0.00001

Interpretation: The null hypothesis of normality is rejected ( $p = .00001$ ) → consider a transformation.

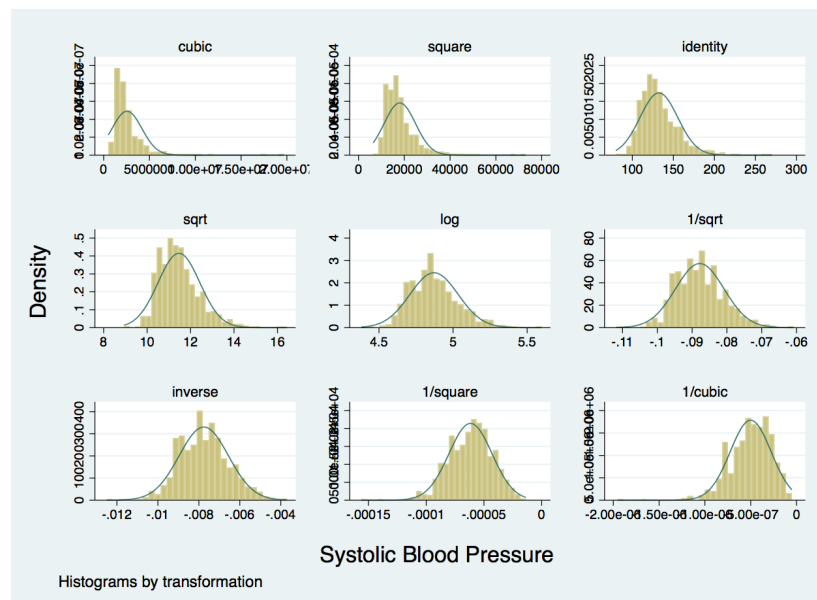
```
. * histogram with overlay normal and quantile-normal plot
. * LOOK FOR: points in quantile-normal plot should fall on the line
. histogram sbp, normal title("Histogram") subtitle("Y=sbp") name(histogram, replace)
(bin=29, start=80, width=6.5517241)

. qnorm sbp, title("Normal QQ Plot") subtitle("Y=sbp") name(qqplot, replace)
. graph combine histogram qqplot
```



Interpretation: The distribution of  $Y=sbp$  departs from normality → confirming that we should consider a transformation.

```
. * command gladder to explore appropriate transformations of Y=sbp
. * NOTE – You may need to issue the command findit gladder and download the routine sed2
. gladder sbp
```





Interpretation: 3 transformations look promising: log, 1/sqrt, and inverse. For this illustration we already know we will use the natural log transformation.

```
. * ----- 4) Create Interactions

. * Interaction age * female sex
. generate age_female=age*female
(0 missing values generated)

. * Interaction ln(scl) * female sex
. generate lnscl_female=ln_scl*female
(4 missing values generated)

. * Interaction ln(bmi) * female sex
. generate lnBMI_female=ln_bmi*female
(2 missing values generated)

. label variable age_female "Age x Female Interaction"
. label variable lnscl_female "ln(scl) x Female Interaction"
. label variable lnBMI_female "ln(bmi) x Female Interaction"
```

## Step 2 – Examine Bivariate Relationships.

Look at the relationship of the dependent variables (Y) with each of the candidate predictor variables (X). Look at these relationships graphically and test correlations. Consider transformations of the predictor variables if needed.

p-value for Null: zero correlation < .0001 → Reject null.

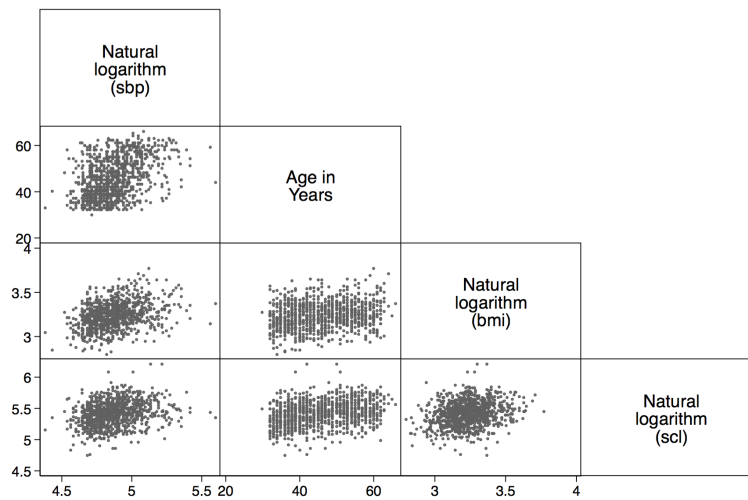
```
. * ----- 1) Command pwcorr to obtain pairwise correlations of Y with each X
. * Command pwcorr YVARIABLE X1 X2 etc
. pwcorr ln_sbp age ln_bmi ln_scl sex, obs sig
```

	ln_sbp	age	ln_bmi	ln_scl	sex
ln_sbp	1.0000				
	994				
age	0.4103	1.0000			
	0.0000	994			
	994	994			
ln_bmi	0.3508	0.1988	1.0000		
	0.0000	0.0000	994		
	994	994	994		
ln_scl	0.2524	0.3055	0.2358	1.0000	
	0.0000	0.0000	0.0000	994	
	994	994	994	994	
sex	0.0119	0.0250	-0.0689	0.0095	1.0000
	0.7077	0.4303	0.0298	0.7642	994
	994	994	994	994	994

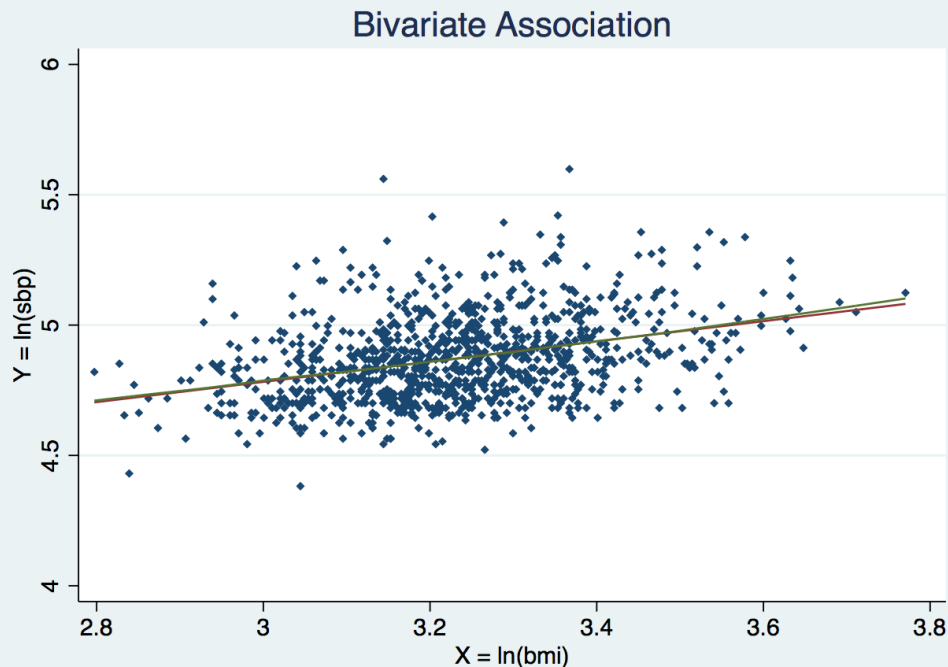
correlation(ln\_sbp, age) = .4103 (Thus, R-squared = .4103<sup>2</sup> = .1683)

p-value for Null: zero correlation < .0001 → Reject null.

```
. * ----- 2) Command graph matrix to obtain pairwise scatterplots of Y with each X
. * graph matrix yvar xvar1 xvar2
. graph matrix ln_sbp age ln_bmi ln_scl, half msize(vsmall)
```



```
. * ----- 3) Command graph twoway to obtain pairwise scatterplots of Y with ONE X
. * Tip! Consider doing an overlay of 3 plots: 1) scatter, 2) least squares line, and 3) lowess
. * graph twoway (scatter yvar xvar) (lfit yvar xvar) (lowess yvar xvar)
. graph twoway (scatter ln_sbp ln_bmi, symbol(d) msize(vsmall)) (lfit ln_sbp ln_bmi) (lowess ln_sbp
ln_bmi), title("Bivariate Association") ylabel4(.5)6 ylabel("Y = ln_sbp") xtitle("X = ln(bmi)")
legend(off)
```



Note:  $Y = \ln\_sbp$  looks to be linearly related to  $X = \ln\_bmi$ . The lowess fit does not depart appreciably from the least squares linear fit.

### Step 3 – Fit Models and Choose “Tentative” Final Model.

Fit an initial model. Fit alternative models. Compare competing models with partial F-tests and side-by-side comparisons of estimated regression coefficients, percent variance explained (R-squared), and mean squared error. Choose a “tentative” final model.

```
. * ----- 1) Fit of initial "maximal" model and tests of interactions.
.* regress yvar xvar1 xvar2
```

Y=ln\_sbp    coef. =  $\hat{\beta}$     \_cons = intercept

```
. regress ln_sbp ln_bmi ln_scl age female lnbmi_female lnscl_female age_female
```

Source	SS	df	MS	Number of obs	=	994
Model	7.01711933	7	1.00244562	F(7, 986)	=	51.21
Residual	19.3006621	986	.019574709	Prob > F	=	0.0000
Total	26.3177825	993	.026503306	R-squared	=	0.2666
				Adj R-squared	=	0.2614
				Root MSE	=	.13991

ln_sbp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
ln_bmi	.303811	.0549107	5.53	0.000	.1960557 .4115663
ln_scl	.0591585	.0368291	1.61	0.109	-.013114 .131431
age	.003694	.0008046	4.59	0.000	.002115 .0052729
female	-.0109333	.3043505	-0.04	0.971	-.6081825 .5863159
lnbmi_female	-.0507228	.0674812	-0.75	0.452	-.1831461 .0817005
lnscl_female	-.0091802	.0498751	-0.18	0.854	-.1070538 .0886934
age_female	.0050381	.0011343	4.44	0.000	.0028121 .0072641
_cons	3.396028	.233872	14.52	0.000	2.937084 3.854972

The fitted line is thus the following.

$$\ln\_sbp = 3.4 + 0.30*\ln\_bmi + 0.06*\ln\_scl + 0.003*age - 0.01*female - 0.05*\lnbmi\_female - 0.009*lnscl\_female + 0.005*age\_female$$

```
. * ----- command testparm to all the interaction terms (3 df Partial F) (NULL: zero)
. testparm lnbmi_female lnscl_female age_female
```

```
( 1)  lnbmi_female = 0
( 2)  lnscl_female = 0
( 3)  age_female = 0
```

```
F( 3, 986) = 6.89
Prob > F = 0.0001
```

Interpretation: The null hypothesis that all 3 interactions are zero is rejected ( $p=.0001$ ). Examination of the coefficients table (see  $P > |t|$ ) suggests that this significance is associated with just one interaction, *age\_female*. Perhaps the other 2 interactions could be dropped.

```
. * ----- Command testparm xvar1 xvar2 to test 2 interactions (2 df Partial F) (NULL: zero)
. testparm ln_bmi_female ln_scl_female
```

```
( 1) ln_bmi_female = 0
( 2) ln_scl_female = 0

F( 2, 986) = 0.34
Prob > F = 0.7144
```

Interpretation: Nice! The null hypothesis that the 2 interactions are zero is NOT rejected → So, tentatively, we think it's okay to drop ln\_bmi\_female and ln\_scl\_female

```
. * ---- 2) Fit of reduced multiple predictor model (this is the candidate/tentative final model)
```

```
. regress ln_sbp ln_bmi ln_scl age female age_female
```

Source	SS	df	MS	Number of obs	=	994
Model	7.00394663	5	1.40078933	F(5, 988)	=	71.66
Residual	19.3138358	988	.019548417	Prob > F	=	0.0000
Total	26.3177825	993	.026503306	R-squared	=	0.2661
				Adj R-squared	=	0.2624
				Root MSE	=	.13982

ln_sbp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
ln_bmi	.2707647	.0318537	8.50	0.000	.208256 .3332734
ln_scl	.0559982	.024711	2.27	0.024	.0075061 .1044902
age	.0036879	.0008017	4.60	0.000	.0021147 .0052612
female	-.2169167	.0508166	-4.27	0.000	-.3166377 -.1171957
age_female	.0048696	.0010882	4.47	0.000	.0027341 .0070051
_cons	3.520535	.1586124	22.20	0.000	3.209279 3.831791

Interpretation: The overall F test (F=71.66) is highly statistically significant. The percent variance explained by this fitted model is 26.6%. Each predictor, controlling for all the other predictors in the model, has a slope that is statistically significantly different from the null value of zero.

```
.
. * ---- 3) Compare Some Competing Models. Use commands eststo and esttab to make a nice table.
. * NOTE - You may need to issue the command findit eststo and download
. * TIP! - Here, I'm using the prefix "quietly:" to suppress all the output. I don't need
. * to see it all again and, besides, I'm showing you how to produce a nifty table.
```

```
. *-- model 1 - Initial "maximal" model
. quietly: regress ln_sbp ln_bmi ln_scl age female age_female ln_bmi_female ln_scl_female
. eststo model1
```

```
. *-- model 2 - Candidate final multiple predictor model
. quietly: regress ln_sbp ln_bmi ln_scl age female age_female
. eststo model2
```

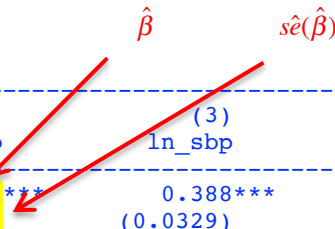
```
. * --- model 3 - Single Predictor model, X=ln(bmi)
. quietly: regress ln_sbp ln_bmi
. eststo model3
```

```
. * ---- model 4 - Single Predictor model, X=ln(scl)
. quietly: regress ln_sbp ln_scl
. eststo model4
```

Design ..... Data Collection ..... Data Management ..... Data Summarization ..... Statistical Analysis ..... Reporting

```
.
. * ---- model 5 - Two Predictor model + Interaction: age, female, and [age x female]
. quietly: regress ln_sbp age female age_female
. eststo model5

. * -- Show comparison of models #1 - #5
. esttab, r2 se scalar(rmse)
```



	(1) ln_sbp	(2) ln_sbp	(3) ln_sbp	(4) ln_sbp	(5) ln_sbp
ln_bmi	0.304*** (0.0549)	0.271*** (0.0319)	0.388*** (0.0329)		
ln_scl	0.0592 (0.0368)	0.0560* (0.0247)		0.211*** (0.0257)	
age	0.00369*** (0.000805)	0.00369*** (0.000802)			0.00370*** (0.000833)
female	-0.0109 (0.304)	-0.217*** (0.0508)			-0.327*** (0.0511)
age_female	0.00504*** (0.00113)	0.00487*** (0.00109)			0.00715*** (0.00110)
lnbmi_female	-0.0507 (0.0675)				
lnscl_female	-0.00918 (0.0499)				
_cons	3.396*** (0.234)	3.521*** (0.159)	3.618*** (0.106)	3.730*** (0.139)	4.701*** (0.0387)
N	994	994	994	994	994
R-sq	0.267	0.266	0.123	0.064	0.203
rmse	0.140	0.140	0.153	0.158	0.146

Standard errors in parentheses  
 \* p<0.05, \*\* p<0.01, \*\*\* p<0.001

Interpretation: Model #2 is our “tentative” final model.

#### Step 4 – Regression Diagnostics.

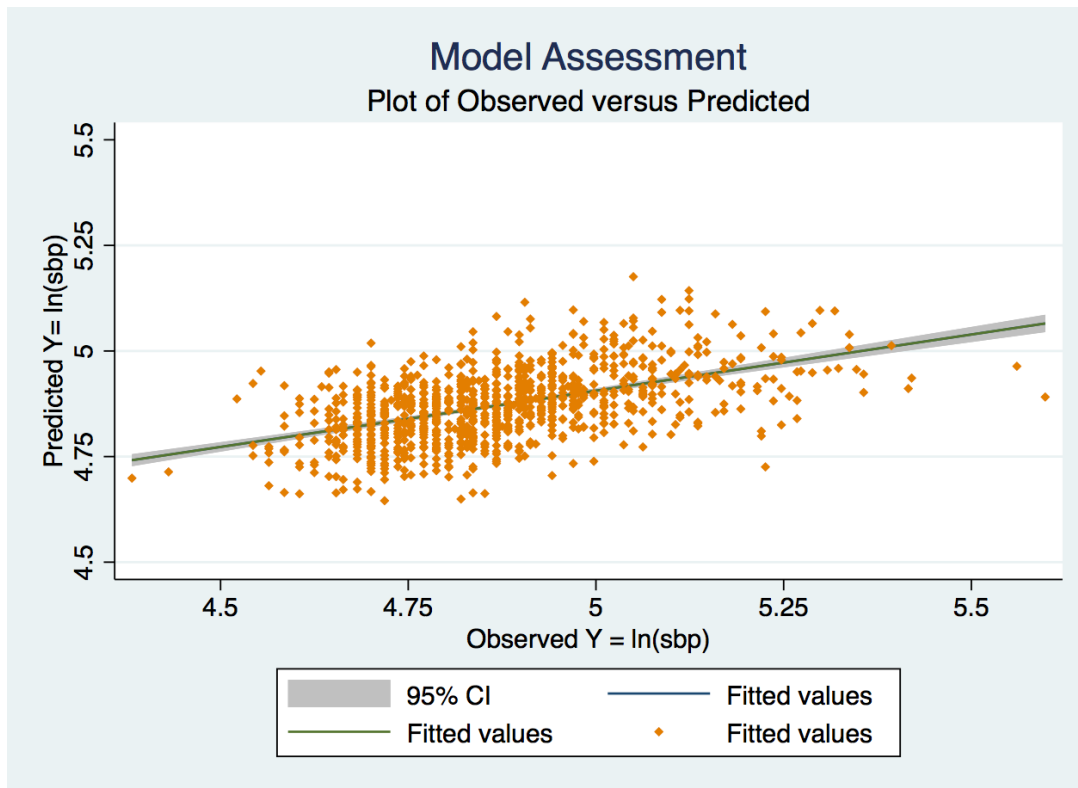
Tip – You must fit your model before any diagnostics on it. The diagnostics you run are called “post-estimation” commands (that makes sense, yes?). Fit again the “tentative” final model; this is a necessary preliminary to doing most regression diagnostics. Check model assumptions. Check model adequacy.

```
. * ----- Preliminary: Must fit the model before doing regression diagnostics (Okay to do quietly)
. quietly: regress ln_sbp ln_bmi ln_scl age female age_female

. * ----- 1) Linearity: Plot of Observed v Predicted
. * LOOK FOR: Points along a straight line (this suggests all is well)

. * Command predict to create a new variable=ypredicted that contains the predicted Y values
. predict ypredicted, xb

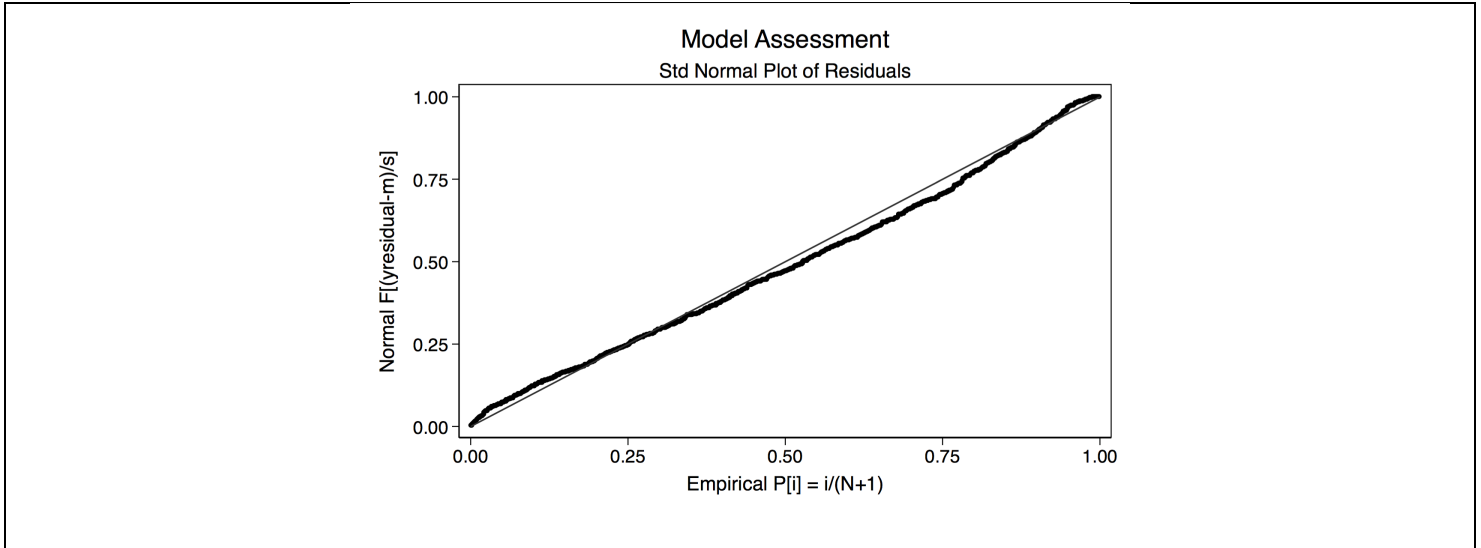
. graph twoway (scatter ypredicted ln_sbp, symbol(d) msize(vsmall)) (lfit ypredicted ln_sbp) (lfitci
ypredicted ln_sbp), title("Model Assessment") subtitle("Plot of Observed versus Predicted")
xtitle("Observed Y = ln(sbp)") ytitle("Predicted Y=ln(sbp)") xlabel(4.5(.25)5.5) ylabel(4.5(.25)5.5)
```



Interpretation: Looks reasonable

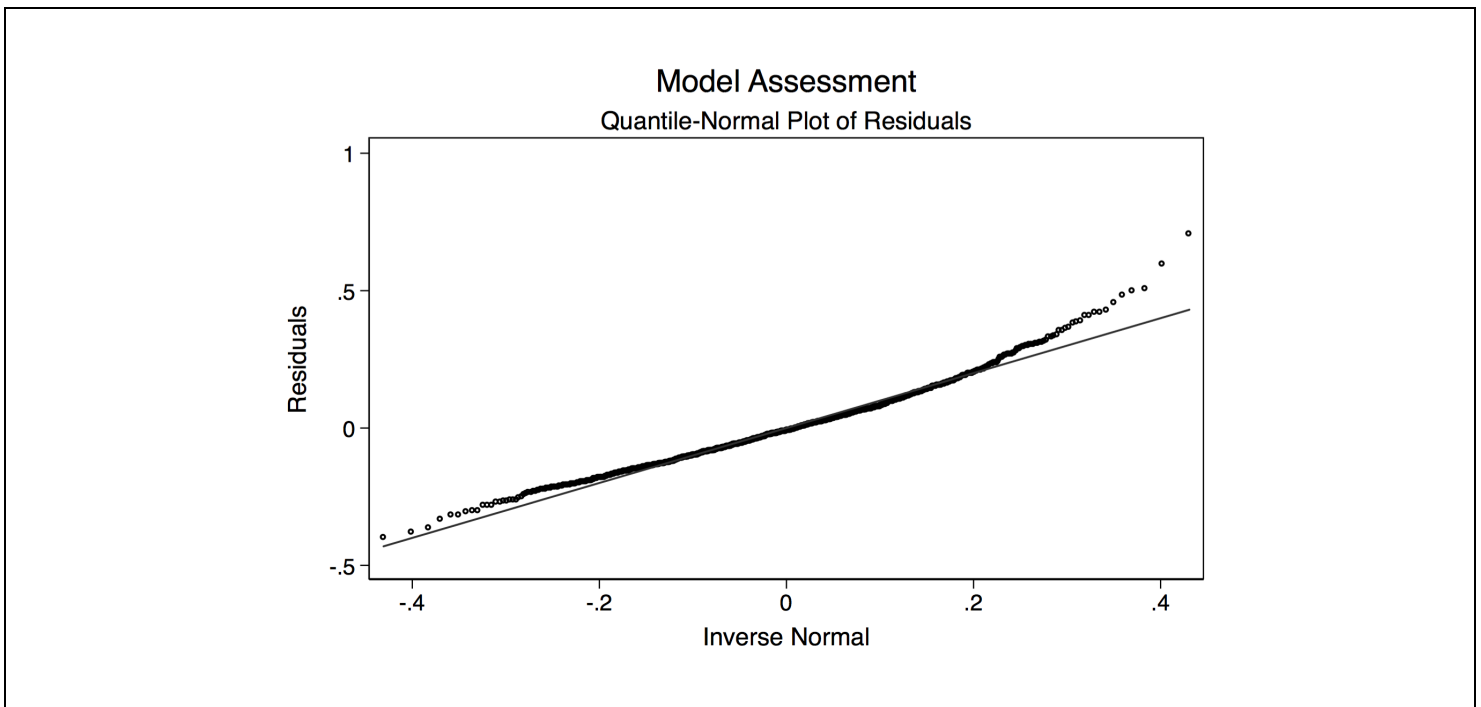
```
. * ---- 2) Normality of residuals: Graphical Assessment
. * LOOK FOR: Points lying on the line (this suggests all is well)
. * Command predict with option resid to create yresidual that contains the residuals
. predict yresidual, resid
(6 missing values generated)

. pnorm yresidual, msize(vsmall) title("Model Assessment") subtitle("Std Normal Plot of Residuals")
```



Interpretation: Again. Looks reasonable

```
. qnorm yresidual, msize(vsmall) title("Model Assessment") subtitle("Quantile-Normal Plot of Residuals")
```



Interpretation: While there is some slight departure of points from the ideal line, both plots are okay for now.

```
. * ---- 3) Normality of residuals: Hypothesis Test (Null: distribution is normal)
. * LOOK FOR: large p-value, not significant (this suggests it is okay to assume normality)

. sfrancia yresidual
```

Shapiro-Francia W' test for normal data

Variable	Obs	W'	V'	z	Prob>z
-----+-----					
yresidual	994	0.97683	15.434	6.271	0.00001

Interpretation: The null hypothesis of normality of the residuals is rejected (not what we want, but okay for now).

```
. * ---- 4) Assessment of Multicollinearity: variance inflation factor (VIF)
. * LOOK FOR: VIF <10 OR 1/VIF > 0.10 (this suggests all is well)
. vif
```

Variable	VIF	1/VIF
-----+-----		
age_female	34.12	0.029311
female	32.39	0.030869
age	2.38	0.420495
ln_scl	1.18	0.850679
ln_bmi	1.12	0.896450
-----+-----		
Mean VIF	14.24	

Interpretation: We have two VIF > 10 (and two 1/VIF < .10) → We may have a multicollinearity problem with age\_female and female

```
. * ---- 5) 2 Tests of Model Misspecification
. * ---- 5a) LINK test (Null: No misspecification. _htsq is NOT significant)
. * LOOK FOR: large p-value, not significant (this suggests all is well)

. linktest
```

Source	SS	df	MS	Number of obs	=	994
-----+-----				F(2, 991)	=	180.45
Model	7.02566005	2	3.51283003	Prob > F	=	0.0000
Residual	19.2921224	991	.019467328	R-squared	=	0.2670
-----+-----				Adj R-squared	=	0.2655
Total	26.3177825	993	.026503306	Root MSE	=	.13953

ln_sbp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
-----+-----					
_hat	-3.398892	4.165516	-0.82	0.415	-11.57314 4.775353
_htsq	.4506253	.4266839	1.06	0.291	-.3866825 1.287933
_cons	10.73201	10.16508	1.06	0.291	-9.215532 30.67956
-----+-----					

Interpretation: Nice! The null hypothesis is NOT rejected → This test does not suggest a model misspecification problem.



```
. * ---- 5b) Ommitted variables (NULL: no variables omitted. All is well)
. * LOOK FOR: large p-value, not significant (this suggests all is well)
. ovtest
```

```
Ramsey RESET test using powers of the fitted values of ln_sbp
Ho: model has no omitted variables
F(3, 985) = 1.81
Prob > F = 0.1442
```

Interpretation: Also nice! The null hypothesis is NOT rejected → This test does not suggest we've omitted any important predictors

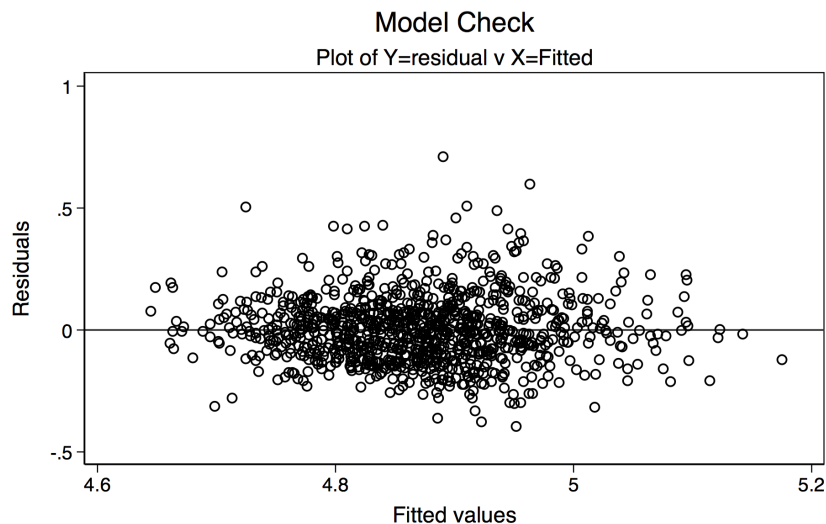
```
. * ----- 6) Hypothesis Test of Constant Variance of the Residuals
. * Breusch-Pagan/Cook Weisberg Test (NULL: constant variance)
. * LOOK FOR: large p-value, not significant (this suggests all is well)
. hettest
```

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
Ho: Constant variance
Variables: fitted values of ln_sbp

chi2(1) = 14.56
Prob > chi2 = 0.0001
```

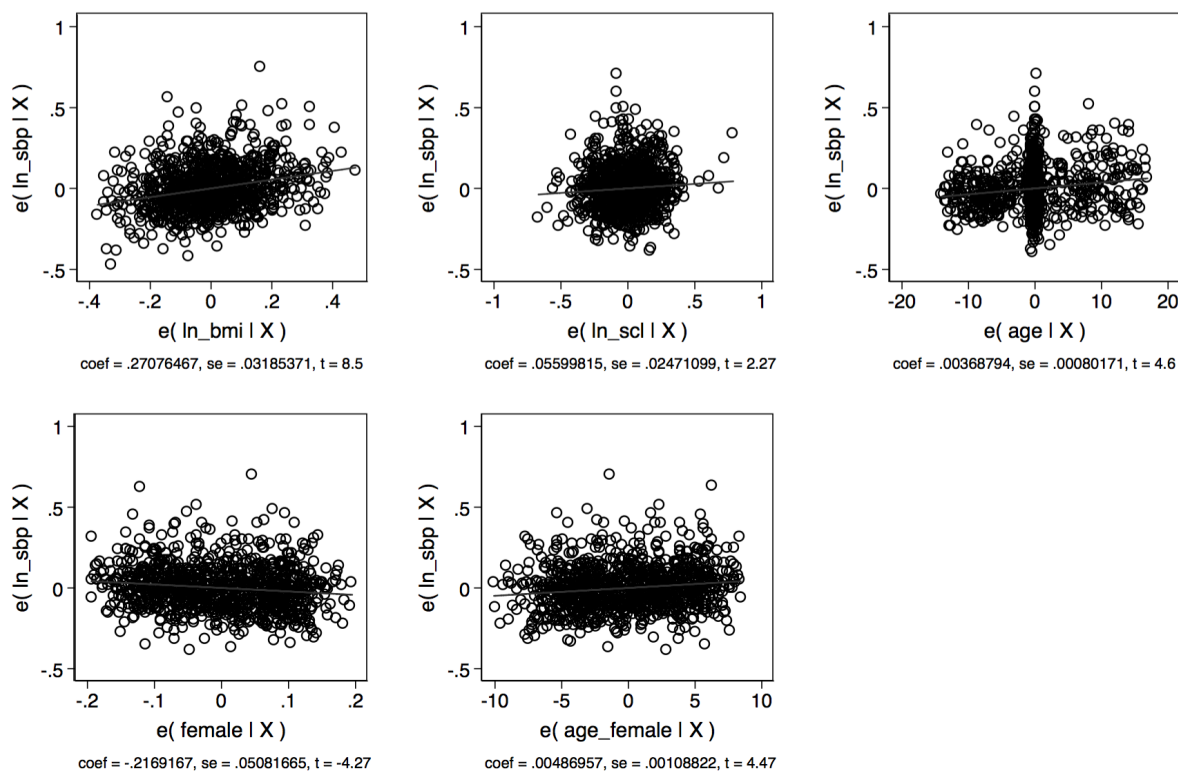
Interpretation: The null hypothesis of constant variance is highly statistically significant, suggesting rejection of the null hypothesis of constant variance of the residuals. So, next, we will look at things graphically

```
. * ----- 6) Graphical Assessment of Constant Variance of the Residuals
. * Plot of Y=residual versus X=fitted
. * LOOK FOR: Even band, centered at zero (this suggests all is well).
. * Command rvfplot, yline(0)
. rvfplot, yline(0) title(Model Check) subtitle(Plot of Y=residual v X=Fitted)
```



Interpretation: Assessed graphically, things don't look so bad. We'll forge on.

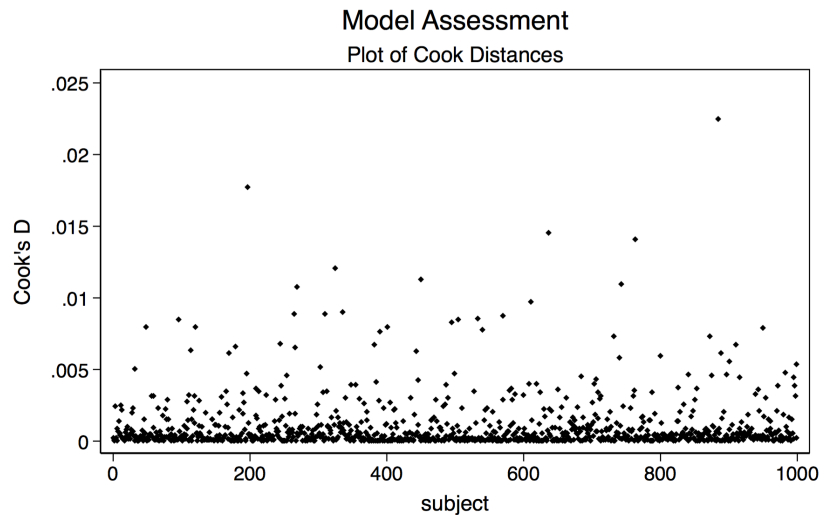
```
. * ----- 7) Checks for Outliers, High Leverage and Influential Points
. * ---- 7a) Added Variable Plots to Look for Unusual/Influential Points
. * Command is avplots.
. * LOOK FOR: Points that are unusual/influential (suggests a problem)
. avplots
```



Interpretation: None of these reveal anything alarmingly unusual/influential

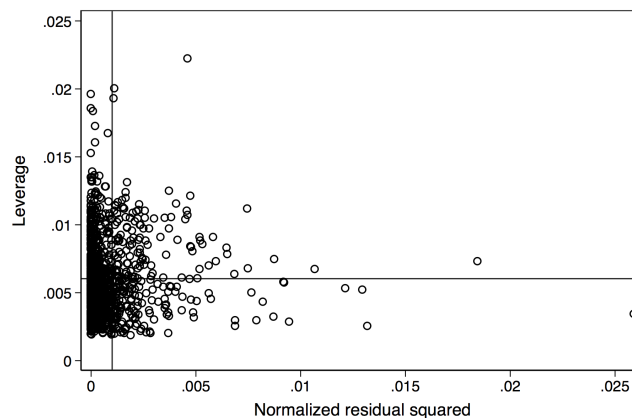
```
. * ---- 7b) Cooks Distances Plot Y=Cook's distance with X=study id
. * LOOK FOR: all to be less than 4/N (this suggests all is well)
. * Command predict with option cooks to create cook that contains the Cook's distances
. predict cook, cooks
(6 missing values generated)

. * Command generate subject=_n to create subject id for nice plotting on x-axis
. generate subject=_n
. graph twoway (scatter cook subject, symbol(d) msize(vsmall)), title("Model Assessment")
  subtitle("Plot of Cook Distances")
```



Interpretation:  $4/N = 4/1000 = .004$ . → We do see some Cook's distances that are larger than  $4/N$ .

```
. * ---- 7c) Plot of Y=leverage versus X=residual squared
. * Command is lvr2plot
. * LOOK FOR: points that are outlying on both (this suggests a problem)
. lvr2plot
```



Interpretation: This confirms what we saw in the plot of Cook's distances

### Step 6 – Report Regression Results.

Produce appropriate tabulations of regression results. Produce graphical summaries of the “final” model. Interpret.

```
. * ----- Again: Must fit the model before doing these reporting commands
. quietly: regress ln_sbp ln_bmi ln_scl age female age_female

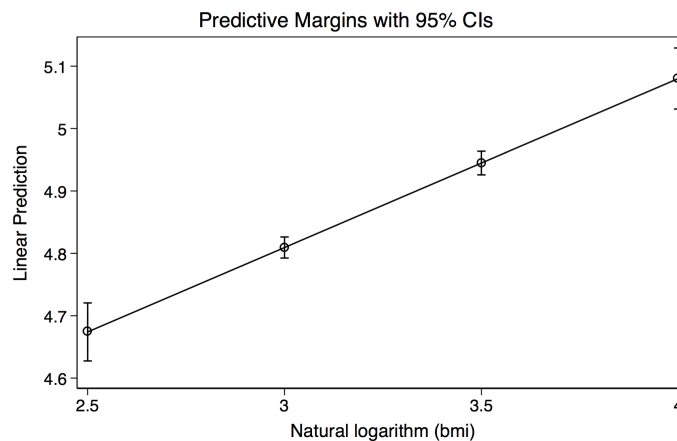
. * ---- 1) Plot predicted Y=ln(sbp) with increasing X = ln(bmi). Option vsquish suppresses blanks.
. margins, at(ln_bmi=(2.6(.2)3.8)) vsquish
```

```
Predictive margins                                Number of obs      =          994
Model VCE      : OLS
```

```
Expression   : Linear prediction, predict()
1._at        : ln_bmi          =          2.6
2._at        : ln_bmi          =          2.8
3._at        : ln_bmi          =           3
4._at        : ln_bmi          =          3.2
5._at        : ln_bmi          =          3.4
6._at        : ln_bmi          =          3.6
7._at        : ln_bmi          =          3.8
```

		Delta-method				
	Margin	Std. Err.	t	P> t	[95% Conf. Interval]	
_at						
1	4.701072	.0205758	228.48	0.000	4.660695	4.741449
2	4.755225	.0144203	329.76	0.000	4.726927	4.783523
3	4.809378	.0085848	560.22	0.000	4.792531	4.826224
4	4.863531	.0045417	1070.87	0.000	4.854618	4.872443
5	4.917684	.0069805	704.49	0.000	4.903985	4.931382
6	4.971836	.0125698	395.54	0.000	4.94717	4.996503
7	5.025989	.0186667	269.25	0.000	4.989359	5.06262

```
. marginsplot, recast(line) recastci(rarea)
```



Variables that uniquely identify margins: ln\_bmi

### 3. Exploratory Data Analysis, Indicator Variables and Interactions

Examine the data to assess:

1. The range and pattern of variability in the outcome variable, Y
2. The range and pattern of variability in the predictor variable X
3. The nature and strength of the presumed linear relationship, Y on X
4. The occurrence of unusual data points requiring further examination; these could be either important data points that are influential or errors.

#### 3.1 Exploratory Data Analysis

##### Familiarize Yourself with the Dataset

Stata Syntax	Notes
<b>describe</b>	
<b>codebook</b>	
<b>codebook, compact</b>	
<b>notes</b>	Stata returns any notes that the dataset creator attached to this dataset.
<b>label list</b>	Stata shows you the labels attached to discrete variable values.

##### One Variable Descriptions – Continuous Variables

Stata Syntax	Notes
<b>summarize</b> <i>var1 var2</i>	
<b>summarize</b> <i>var1 var2</i> , detail	
<b>codebook</b> <i>var1 var2</i>	
<b>codebook</b> <i>var1 var2</i> , compact	
<b>tabstat</b> <i>var1 var2</i> , statistics(n mean sd min max)	

## Glossary of Choices of Options for tabstat

<b>tabstat variable, statistics( )</b>	
statname	Definition
<u>m</u> ean	mean
<u>c</u> ount	count of nonmissing observations
<u>n</u>	same as count
<u>s</u> um	sum
<u>m</u> ax	maximum
<u>m</u> in	minimum
<u>r</u> ange	range = max - min
<u>s</u> d	standard deviation
<u>v</u> ariance	variance
<u>c</u> v	coefficient of variation (sd/mean)
<u>s</u> em <sup>e</sup> an	standard error of mean (sd/sqrt(n))
<u>s</u> ke <sup>w</sup> ness	skewness
<u>k</u> urtosis	kurtosis
<u>p</u> 1	1st percentile
<u>p</u> 5	5th percentile
<u>p</u> 10	10th percentile
<u>p</u> 25	25th percentile
<u>m</u> edian	median (same as p50)
<u>p</u> 50	50th percentile (same as median)
<u>p</u> 75	75th percentile
<u>p</u> 90	90th percentile
<u>p</u> 95	95th percentile
<u>p</u> 99	99th percentile
<u>i</u> qr	interquartile range = p75 - p25
<u>q</u>	equivalent to specifying p25 p50 p75

## One Variable Descriptions – Discrete Variables

Stata Syntax	Notes
<b>tabulate var1</b> <b>tabulate var1, missing</b>	For single variable frequency tables, the command tabulate allows ONE discrete variable only. If you issue the command tabulate var1 var2, Stata will return a cross-tabulation. This may not be what you want
<b>ssc install fre</b> <b>fre var1 var2</b>	Issue the command ssc install fre ONLY ONCE; this will download and install the command fre
<b>tab1 var1 var2</b> <b>tab1 var1 var2, missing</b> <b>tab1 var1 var2, plot</b>	tab1 with more than one variable will produce separate one way frequency tables.
<b>ssc install groups</b> <b>groups var1 var2</b>	Issue the command ssc install groups ONLY ONCE; this will download and install the command groups

## Two Variable Descriptions – One Continuous, One Discrete (Grouping Variable)

Stata Syntax	Notes
<b>summarize</b> <i>var1</i> if <i>groupvar</i> == <i>expression</i>	Obtain one variable description, for single group defined by <i>groupvar</i> == <i>expression</i> .
<b>bysort</b> <i>groupvar</i> : <b>summarize</b> <i>var1</i>	Obtain one variable description for all groups defined by <i>groupvar</i> .
<b>sort</b> <i>groupvar</i> <b>tabstat</b> <i>var1</i> , by( <i>groupvar</i> ) <b>statistics</b> (n mean sd min max)	
<b>sort</b> <i>groupvar</i> <b>table</b> <i>groupvar</i> , <b>contents</b> (n <i>var1</i> mean <i>var1</i> sd <i>var1</i> )	

## Glossary of Choices of Options for **table**

<b>table</b> <i>xvariable</i> , <b>contents</b> ( )	
<b>freq</b>	frequency
<b>mean</b> <i>varname</i>	mean of <i>varname</i>
<b>sd</b> <i>varname</i>	standard deviation
<b>semean</b> <i>varname</i>	standard error of the mean (sd/sqrt(n))
<b>sebinomial</b> <i>varname</i>	standard error of the mean, binomial distribution (sqrt(p(1-p)/n))
<b>sepoisson</b> <i>varname</i>	standard error of the mean, Poisson distribution (sqrt(mean))
<b>sum</b> <i>varname</i>	sum
<b>rawsum</b> <i>varname</i>	sums ignoring optionally specified weight
<b>count</b> <i>varname</i>	count of nonmissing observations
<b>n</b> <i>varname</i>	same as count
<b>max</b> <i>varname</i>	maximum
<b>min</b> <i>varname</i>	minimum
<b>median</b> <i>varname</i>	median
<b>p1</b> <i>varname</i>	1st percentile
<b>p2</b> <i>varname</i>	2nd percentile
...	3rd-49th percentile
<b>p50</b> <i>varname</i>	50th percentile (median)
...	51st-97th percentile
<b>p98</b> <i>varname</i>	98th percentile
<b>p99</b> <i>varname</i>	99th percentile
<b>iqr</b> <i>varname</i>	interquartile range

### Graphical Assessments

Stata Syntax	Notes
<p><b><u>Y with one predictor</u></b></p> <p><b>graph twoway</b> (scatter <i>yvar xvar</i>)  <b>graph twoway</b> (scatter <i>yvar xvar</i>) (lfit <i>yvar xvar</i>)  <b>(lowess</b> <i>yvar xvar</i>)</p> <p><b><u>Y with several predictors (handy and compact)</u></b></p> <p><b>graph matrix</b> <i>yvar xvar1 xvar2</i>  <b>graph matrix</b> <i>yvar xvar1 xvar2</i>, half  <b>graph matrix</b> <i>yvar xvar1 xvar2</i>, half  <b>maxis</b>(ylabel(none) xlabel(none))</p>	<p>lfit produces least squares linear fit  lowess produces lowess smoothing fit.</p> <p><b>Tip</b> - graph matrix produces pairwise scatter of the predictor variables</p>

### Assess Normality of Y

Stata Syntax	Notes
<p><b>histogram</b> <i>yvar</i>  <b>histogram</b> <i>yvar</i>, normal</p> <p><b>qnorm</b> <i>yvar</i></p> <p><b><u>Hypothesis Tests of Normality (Null: Normality)</u></b>  <b>swilk</b> <i>yvar</i>  <b>sfrancia</b> <i>yvar</i></p>	<p>Look for: bell shape distribution (all is well)</p> <p>Look for : points falling on a line (all is well)</p> <p>Look for: non-significant (large) p-value (all is well)</p>

### Search for Normalizing Transformation of Y

Stata Syntax	Notes
<p><b>ladder</b> <i>yvar</i></p> <p><b>gladder</b> <i>yvar</i></p>	<p>Produces <u>table</u> of transformations of Y that might be normalizing. Choose one(s) with the smallest chi square value</p> <p>Produces <u>histogram plots</u> of transformations of Y that might be normalizing.</p>



## 3.2 How to Create Indicator Variables

**NEVER!!!**

Use a NOMINAL Predictor in a regress Command

The estimated slope will be meaningless.

**Example:**

**Party** (1=Republican, 2=Democratic, 3 = Libertarian, 4 = Green) is a nominal variable. Because the numbers “1”, “2”, “3” and “4” are just labels, a unit change in race has no meaning. Therefore, an estimated slope for party also has no meaning.

## Review of How to Model Discrete Predictors

- (1) A discrete predictor might be nominal (eg. – race) or ordinal (eg – age, grouped)
- (2) Note the number of levels (eg – party has 4 levels)
- (3) Choose one level to be the referent (eg – the group “1=Republican”, if this is the most numerous)
- (3) K levels require (K-1) design variables (eg – For race, we need  $[4-1]=3$  design variables)
- (4) Use ONLY design variables as predictors.

## How to Create 0/1 Indicator Variables

**Example – Create 0/1 Indicator variables for the 4 levels of party and use these in a regression. Recall:**

party = 1 if Republican  
           2 if Democratic  
           3 if Libertarian  
           4 if Green

### 0/1 Indicators and Regression on 0/1 Indicators

Stata Syntax	Notes
<p><b><u>Brute force (Method 1)</u></b></p> <p>generate repub=(party==1) &amp; !missing(party)  generate democrat=(party==2) &amp; !missing(party)  generate libert=(party==3) &amp; !missing(party)  generate green=(party==4) &amp; !missing(party)</p> <p>regress yvar democrat libert green  regress yvar repub democrat libert</p>	<p>Reminder: Stata requires double equal signs in logical operators.</p>
<p><b><u>Brute force (Method 2)</u></b></p> <p>quietly: tab1 party, generate(party)  groups party*  rename party1 repub  rename party2 democrat  rename party3 libert  rename party4 green</p> <p>regress yvar democrat libert green  regress yvar repub democrat libert</p>	<p>Referent is party = 1 (Republican)  Referent is party = 4 (Green)</p>
<p><b><u>Using the xi: prefix and the i.predictor</u></b></p> <p>xi: regress yvar i.party</p> <p>xi: regress yvar ib4.party</p>	<p>Referent is party = 1 (Republican)  Referent is party = 4 (Green)</p> <p>Default referent is 1<sup>st</sup> group, party = 1 (Republican)</p> <p><b>Tip – You can choose your own referent</b>  Use ib4 to instead specify 4<sup>th</sup> group as baseline/referent, party = 4 (Green)</p>

### 3.3 How to Create Interactions

#### Interactions and Regression on Interactions

Stata Syntax	Notes
<p><b>Brute force</b>  <code>generate dummyx=dummy*xvar</code></p> <p><code>regress yvar dummy xvar dummyx</code></p>	<p>Here, dummy is a 0/1 indicator.  xvar is the other predictor of interest</p> <p>If you include an interaction as a predictor, your model must also contain the main effects.</p>
<p><b>Using the xi: prefix and the i.predictor</b>  <u>ONLY intercepts vary</u>  <code>xi: regress yvar i.groupvar xvar</code></p>	<p><u>ONLY intercepts vary</u>  This will yield a separate intercept for each group defined by <i>groupvar</i> and a common slope of <i>yvar</i> on <i>xvar</i>.</p>
<p><b>Using the xi: prefix and the i.predictor</b>  <u>Intercepts AND Slopes vary</u>  <code>xi: regress yvar i.groupvar xvar i.groupvar#c. xvar</code></p>	<p><u>Intercepts AND Slopes intercepts vary</u>  This will yield a separate intercept for each group defined by <i>groupvar</i> and a separate slope of <i>yvar</i> on <i>xvar</i>.</p>

### 3.4 How to Created Quartiles (or other groupings)

Dear reader: There are lots of ways to do this, some quite slick. I prefer a “brute force” approach so that I’m sure of what I’ve got.

Stata Syntax	Example																
<p><b><u>Brute force</u></b></p> <p>Step 1: Obtain values of quartiles (or centiles) <b>centile</b> <i>variable</i>, <b>centile(0 25 50 75 100)</b></p> <p>Step 2: Create grouped variable as copy of original <b>generate</b> <i>newvariable</i> = <i>variable</i></p> <p>Step 3: Recode new variable according to quartile (or centile) boundaries <b>recode</b> <i>newvariable</i> (<b>##=1</b>) (<b>##=2</b>) (<b>##=3</b>) (<b>##=4</b>)</p> <p>Step 4: (Just to be sure) Set new variable = missing whenever original variable=missing <b>replace</b> <i>newvariable</i>=. <b>if</b> <i>variable</i>==.</p> <p>Step 5: Check that all is well <b>table</b> <i>newvariable</i>, <b>contents(min</b> <i>variable</i> <b>max</b> <i>variable</i>)</p>	<pre>. centile price, centile(0 25 50 75 100)</pre> <table><thead><tr><th>Variable</th><th>Obs</th><th>Percentile</th><th>Centile</th></tr></thead><tbody><tr><td rowspan="5">price</td><td rowspan="5">74</td><td>0</td><td>3291</td></tr><tr><td>25</td><td>4193</td></tr><tr><td>50</td><td>5006.5</td></tr><tr><td>75</td><td>6378</td></tr><tr><td>100</td><td>15906</td></tr></tbody></table> <pre>. generate quartile_price=price if !missing(price)</pre> <pre>. recode quartile_price (3291/4193=1) (4193.01/5006.5=2) (5006.6/6378=3) (6378.1/15906=4)</pre> <pre>. replace quartile_price=. if price==.</pre> <pre>. table quartile_price, contents(min price max price)</pre>	Variable	Obs	Percentile	Centile	price	74	0	3291	25	4193	50	5006.5	75	6378	100	15906
Variable	Obs	Percentile	Centile														
price	74	0	3291														
		25	4193														
		50	5006.5														
		75	6378														
		100	15906														

## 4. Simple Linear Regression (Bivariate Analyses)

### Simple Linear Regression

Stata Syntax	Notes
<p><b><u>Graph Simple Linear Regression</u></b>  graph twoway (scatter <i>yvar xvar</i>)  graph twoway (scatter <i>yvar xvar</i>) (lfit <i>yvar xvar</i>)  graph twoway (scatter <i>yvar xvar</i>) (lfitci <i>yvar xvar</i>)  (lfit <i>yvar xvar</i>)</p> <p><b><u>Fit Simple Linear Regression</u></b>  regress <i>yvar xvar</i></p> <p><b><u>The Residuals Should be Normally Distributed</u></b>  quietly: regress <i>yvar xvar</i>  predict <i>newvar1</i>, residuals  swilk <i>newvar1</i>  sfrancia <i>newvar1</i>  histogram <i>newvar1</i>, normal  qnorm <i>newvar1</i>  pnorm <i>newvar1</i></p> <p><b><u>The Variance of the Residuals Should be Constant</u></b>  predict <i>newvar1</i>, residuals  predict <i>newvar2</i>, xb  rvplot, yline(0)  graph twoway (scatter <i>newvar1 studyid</i>, yline(0))  graph twoway (scatter <i>newvar1 newvar2</i>, yline(0))</p> <p><b><u>The Cook Distances Should be Small (&lt; 1) w NO Spikes</u></b>  predict <i>newvar3</i>, cooks  graph twoway (scatter <i>newvar3 studyid</i>)</p>	<p>Prefix “quietly:” tells stata to suppress output.  The regress command that follows must be issued before the command “predict” will work.</p>



## 5.2 Hierarchical Model Comparisons

**Optional** (because `testparm` does the same)

**You can use the command `ftest` (but you may need to install it first)**

The `ftest` command performs a partial F-test

Step 1: In Stata issue the command  
**`findit ftest`**

Step 2: From the `findit` screen  
Scroll down to locate the package at [fmwww.bc.edu](http://fmwww.bc.edu).

Step 3: Follow the instructions to download

### Review of Hierarchical Models

Two models, conveniently referred to as “reduced” and “full”, are hierarchical if the all of the predictors in the “reduced” model are contained in the “full” model. Their comparison then addresses the question: are the additional variables in the “full” model significant after adjustment for all the variables in the “reduced” model?

The comparison of hierarchical models is an essential tool in regression model development.

**Hierararchical model comparison requires that the fitted models are to the SAME observations** – This glitch arises if a smaller set of observations is used to fit a model with lots of predictors (because of missing values). **Tip** – Fit your full model first, create an indicator of data completeness as a post-regression command using the internal Stata variable `e(sample)`, and then use this indicator in the fitting of the smaller model. This is illustrated on the next page.

### Comparison of Hierarchical Models

Stata Syntax	Notes
<p><b>Quick Look at Significance Of Some Predictors in a Fitted Model (Easy)</b></p> <p><u>Partial F-test of added variables xvar3 xvar4</u>  <b>regress</b> <i>yvar xvar1 xvar2 xvar3 xvar4</i>  <b>testparm</b> <i>xvar3 xvar4</i></p>	<p>testparm produces a partial F-test of inclusion of xvar3 and xvar4 controlling for xvar1 and xvar2 already in the model (NULL: zero)</p> <p>Look for: small p-value → after adjustment /controlling for xvar1 and xvar2, the additional inclusion of xvar3 and xvar2 is statistically significant and should be in model.</p>
<p><b>Hierarchical Comparison of “Full” v “Reduced” Models</b></p> <p><u>STEP 1: Fit full model. Store.</u>  <b>regress</b> <i>yvar xvar1 xvar2 xvar3 xvar4</i>  <b>estimates store</b> <i>full</i></p> <p><u>STEP 2: Generate indicator of data completeness</u>  <b>generate</b> <i>complete=e(sample)</i></p> <p><u>STEP 3: Fit reduced model on SAME observations as for the full model. Store.</u>  <b>regress</b> <i>yvar xvar1 xvar2 if complete==1</i>  <b>estimates store</b> <i>reduced</i></p> <p><u>Partial F-test Comparing Full v Reduced Model</u>  <b>ftest</b> <i>full reduced</i></p> <p><u>Likelihood Ratio Test Comparing Full v Reduced Model</u>  <b>lrtest</b> <i>full reduced</i></p>	<p>I chose to name my full model <i>full</i></p> <p>I chose to name my indicator variable of complete on all predictors <i>complete</i></p> <p><i>Same as testparm.</i> Look for: small p-value → after adjustment /controlling for predictors in the reduced model, the extra predictors in the full model are statistically significant and should be in model.</p> <p>Look for: small p-value → after adjustment /controlling for predictors in the reduced model, the extra predictors in the full are statistically significant and should be in model.</p>

## 6. Regression Diagnostics: Model Assumptions and Model Adequacy

Here are Several Useful Variables that STATA Creates for You AFTER Fitting a Model

Stata Syntax	Notes
<u>Predicted Values</u> <b>predict</b> <i>var1</i> , xb <b>predict</b> <i>var1</i> if e(sample)==1, xb	Saves predicted Y If e(sample)==1 tells Stata to use ONLY the observations that were included in model estimation. Note – This is not necessary if you have already restricted your modeling to complete data only, as we did here.
<u>Standard Error of Predicted Mean of Y</u> <b>predict</b> <i>var2</i> , stdp	
<u>Standard Error of Predicted Individual Y</u> <b>predict</b> <i>var3</i> , stdf	
<u>Residuals</u> <b>predict</b> <i>var4</i> , residuals <b>predict</b> <i>var4</i> if e(sample)==1, residuals	Saves residuals = (observed Y) - (fitted Y)
<u>Standard Errors of Residuals</u> <b>predict</b> <i>var5</i> , stdr	
<u>Standardized Residuals</u> <b>predict</b> <i>var6</i> , rstandard <b>predict</b> <i>var6</i> if e(sample)==1, rstandard	Saves standardized residuals = (residual)/SE(residual)
<u>Jackknife (Studentized) Residuals</u> <b>predict</b> <i>var7</i> , rstudent	Saves studentized (jackknife) residuals; the SE is slightly different
<u>Leverage</u> <b>Predicts</b> <i>var8</i> , leverage	Saves leverage
<u>Cook Distances</u> <b>predict</b> <i>var7</i> , cooksd	Saves cook distances
<b>generate</b> <i>newid</i> =_n	<b>Tip</b> – Use this command ONLY IF your data does not contain a studyid variable. We use this in a plot of cook distances versus study id (see page 27) which we named <i>subject</i>



## 6.1 Linearity

### Linearity

Stata Syntax	Notes
<b>For Simple Linear Regression</b>  <b>graph twoway (scatter yvar xvar) (lfit yvar xvar)</b>  <b>graph twoway (scatter yvar xvar) (lfitci yvar xvar) (lfit yvar xvar)</b>  <b>graph twoway (scatter yvar xvar) (lowess yvar xvar) (lfit yvar xvar)</b>	Look for: linearity  Produces line and 95% confidence band  Produces both line and lowess fit. Departure of lowess fit from the fitted line suggests a problem.

## 6.2 Normality of Residuals

### Normality of Residuals

Stata Syntax	Notes
<u><b>Histogram of Residuals</b></u> <b>predict yresid, residuals</b> <b>histogram yresid, normal</b>	I chose the name yresid for the residuals.
<u><b>Standardized Normal Probability Plot of Residuals</b></u> <b>predict yresid, residuals</b> <b>pnorm yresid</b>	Assesses normality of residuals Look for: points lie on line (all is well)
<u><b>Test of Normality (NULL: Normal)</b></u> <b>swilk yresid</b> <b>sfrancia yresid</b>	

### 6.3 Multicollinearity

Multicollinearity occurs when the predictor variables themselves are linearly interrelated. This is a problem because it makes it difficult to extract the separate effect of each predictor; the betas are unstable.

Multicollinearity also has the effect of inflating the variances of the estimated betas. For example, if `xvar1` and `xvar2` are themselves highly linearly interrelated, then the variance of the beta for `xvar1` will be inflated!

We use the variance inflation factor (VIF) to assess the data for evidence of multicollinearity.

#### Multicollinearity

Stata Syntax	Notes
<u>Pairwise Scatterplots of Predictor Variables</u> <code>graph matrix xvar1 xvar2 xvar3 xvar4, half</code>  <u>Variance Inflation Factor (VIF) Values</u> <code>vif</code>	Look for: VIF values > 10 suggest a problem 1/VIF values < 0.10 suggest a problem

### 6.4 Model Misspecification

#### Model Misspecification

Stata Syntax	Notes
<u>Test of Model Misspecification (NULL: none)</u> <code>linktest</code>  <u>Test for Omitted Variables (NULL: none forgotten)</u> <code>ovtest</code>	Look for: Predictor <code>_hatsq</code> should be NOT significant (all is well)  Look for: NON significance (all is well)

## 6.5 Constant Variance

### Constant Variance

Stata Syntax	Notes
<u>Plot Residuals versus Predicted Y (Method I)</u> <b>rvfplot, yline(0)</b>	Look for residuals randomly distributed in an even band centered at 0.
<u>Plot of Residuals versus Predicted Y (Method II)</u> <b>predict yhat, xb</b> <b>predict yresid, residuals</b> <b>graph twoway (scatter yresid yhat)</b>	I chose the name <b>yhat</b> for the predicted y I chose the name <b>yresid</b> for the residuals Assesses constant variance Look for: even band centered at zero (all is well)
<u>Plot Y=residuals versus X=Predictor variable</u> <b>predict yresid, residuals</b> <b>graph twoway (scatter yresid xvar1)</b>	I chose the name <b>yresid</b> for the residuals Assesses constant variance Look for: even band centered at zero (all is well)
<u>Hypothesis Test (Null: Constant variance)</u> <b>hettest</b>	Reject constant variance for small p-values

## 6.6 Outlying, High Leverage, and Influential Points

### Preliminary (if you don't already have it): Download the command hilo

The hilo command lets you list the highest and lowest values of a variable together with whatever companion data you might want. Handy for regression diagnostics!

Step 1: In Stata issue the command  
findit hilo

Step 2: From the findit screen  
Scroll down to locate the package at [www.ats.ucla](http://www.ats.ucla).

Step 3: Follow the instructions to download

### Outliers are Observations with Large Residuals

Stata Syntax	Notes
<b>predict</b> <i>yresid</i> , residuals	I chose the name <i>yresid</i> for the residuals
<b>stem</b> <i>yresid</i>	Produces a stem and leaf, good for detecting outliers
<b>hilo</b> <i>yresid studyidvar</i> , high show(#)	Lists the # observations that have the highest values of <i>yresid</i> , together with the <i>studyid</i>

### Leverage are Observations with Extreme Values on the Predictor Variables

Stata Syntax	Notes
<b>predict</b> <i>xleverage</i> , leverage	I chose the name <i>xleverage</i> for the leverages
<b>stem</b> <i>xleverage</i>	Produces a stem and leaf, good for detecting high leverage observations. Extreme is leverage > (2p + 2) / n where p=# predictors.
<b>hilo</b> <i>xleverage studyidvar</i> , high show(#)	Lists the # observations that have the highest leverages, together with the <i>studyid</i>

### Influence are Observations that Influence the Estimated Betas

Stata Syntax	Notes
<p><u>Cook's Distances</u></p> <p><b>predict</b> <i>cookvar</i>, <b>cooks</b></p> <p><b>graph twoway</b> (<b>scatter</b> <i>cooksvar idvar</i>)</p>	<p>I chose the name <i>cookvar</i> for the cook's distances</p> <p>Plot of Y=Cook distances versus X=Study id (<i>idvar</i>) Look for: nothing extreme (all is well). Extreme is cook distance <math>&gt; 4/n</math></p>
<p><u>dfbeta</u></p> <p><b>dfbeta</b></p> <p><b>graph twoway</b> (<b>scatter</b> <i>DFvar1 studyid</i>)</p>	<p>Command dfbeta produces several variables, one for each predictor: <i>var1</i>, <i>var2</i>, etc. The names of these will be <b>DFvar1</b>, <b>DFvar2</b>, etc. Thus, you can assess the influence of an observation on the beta for each predictor separately. Extreme is dfbeta <math>&gt; 2 / \sqrt{n}</math></p> <p>Plot of <b>DFvar1</b> for the predictor <i>var1</i> versus X=Study id Extreme is <b>DFvar1</b> <math>&gt; 2 / \sqrt{n}</math></p>
<p><b>lvr2plot</b></p> <p><b>lvr2plot</b>, <b>mlabel</b>(<i>studyidvar</i>)</p>	<p>Plot of Y=leverage versus X=squared residual Look for: Observations that are high on both (suggests a problem).</p> <p>Use option mlabel to identify the observations that are problematic.</p>

## 7. Post Regression: Prediction and Reporting

### 7.1 Predictions

#### Predictions

Stata Syntax	Notes
<p><b>Prediction of Mean of Y</b>  <b>Point Estimates</b>  <code>predict newvar1, xb</code>  <code>predict newvar1 if e(sample)==1, xb</code></p> <p><b>Standard Error of Predicted Means</b>  <code>predict newvar2, stdp</code></p> <p><b>Predicted Mean of Y at new value of x</b>  <code>margins, at(xvar=newvalue) atmeans vsquish</code></p> <p><b>Predicted Mean of Y at more than one new value</b>  <code>margins, at(xvar=(value1 value2 etc)) atmeans vsquish</code></p> <p><b>Predicted Mean of Y for values in a cross-tabulation of 2 categorical predictors and all other predictors at their means</b>  <code>margins catvar1 catvar2, atmeans</code></p>	<p>Save predicted Y to newvar            If e(sample)==1 tells Stata to use only the observations that were included in the regression.</p> <p>Save standard errors of predicted means</p> <p>Example: Predicted mean of Y when <code>x=newvalue</code> and all other predictors are at the value of their mean  <b>NOTE:</b> <code>vsquish</code> is just an aesthetic thing; this option eliminates blank lines in tables.</p> <p>Example: Predicted mean of Y when <code>x=value1, value2, etc</code> and all other predictors are at the value of their mean</p> <p>Example:  <code>margins gender party, atmeans</code></p>
<p><b>Prediction of Individual Y</b>  <b>Point Estimates</b>  <code>predict newvar1, xb</code>  <code>predict newvar1 if e(sample)==1, xb</code></p> <p><b>Standard Error of Predicted Individual Values</b>  <code>predict newvar3, stdf</code></p>	<p>Save predicted Y to newvar            If e(sample)==1 tells Stata to use only the observations that were included in the regression.</p> <p>Save standard errors of predicted individual values</p>

## 7.2 Show Models Side-by-Side

**I Highly Recommend!** – Consider showing side-by-side the various models that you fit and assessed. Tip – Take care that each model is fit to the same observations. To do this, fit the model with the most predictors first and, from this model, create an indicator variable denoting complete data (See again page 40)

### Example

Model 1 Predictors (smallest): dose age

Model 2 Predictors (intermediate): dose age female

Model 3 Predictors (largest): dose age female doseage

We think “female” might be a confounder

We considered an interaction of dose with age (doseage=dose\*age)

### Show Models Side-by-Side

Stata Syntax	Notes
<p><b>Step 1: Obtain estimation sample for use in all 3 models</b></p> <p>quietly: regress <i>yvar dose age female doseage</i> generate <i>complete</i>=e(sample)</p> <p><b>Step 2: Fit models and store</b></p> <p>quietly: regress <i>yvar dose age</i> if <i>complete==1</i> estimates store <i>model1</i></p> <p>quietly: regress <i>yvar dose age female</i> if <i>complete==1</i> estimates store <i>model2</i></p> <p>quietly: regress <i>yvar dose age female doseage</i> if <i>complete==1</i> estimates store <i>model3</i></p> <p><b>Step 3: Show models side-by-side</b></p> <p>esttab <i>model1 model2 model3</i>, r2 ar2 se scalar(rmse)</p>	<p>Your choice whether to do this “quietly” (suppress output).</p>

### 7.3 Plot Predicted Values

**Also Highly Recommended!** – Consider producing plots of predicted values from your “final” model.

Two commands are needed here, **margins** and **marginsplot**.

Command **margins** produces predicted Y (fitted values) at the values of the predictors that you specify. The output is a bit hard to read; consider doing this quietly.

Command **marginsplot** produces a plot of the predicted Y (fitted values) at the values of the predictors that you specify. Typically, the other covariates are set to their mean values using the option **atmeans**. Your choice.

#### Example

Suppose we have fit a model (any old *yvar*) to the predictors: *female01* and *age*

#### Obtain and Plot Adjusted Predicted Values

Stata Syntax	Notes
<p><b>Predicted Y at Specified X (age), with 95% CI</b></p> <p><b>margins, at(<i>age</i>==(20(5)75)) atmeans</b>  <b>quietly: margins, at(<i>age</i>==(20(5)75)) atmeans</b>  <b>marginsplot</b></p> <p><b>Predicted Y at Specified X (age), no CONFIDENCE INTERVAL</b></p> <p><b>margins, at(<i>age</i>==(20(5)75)) atmeans</b>  <b>quietly: margins, at(<i>age</i>==(20(5)75)) atmeans</b>  <b>marginsplot, noci</b></p>	<p>Produces predicted Y at ages 20, 25, ..., 70, 75 together with associated 95% confidence limits. All the other covariate values are set to their mean values</p>
<p><b>Predicted Y at Specified X<sub>1</sub> (age), separately for Groups Defined by X<sub>2</sub> (<i>female01</i>), no 95% CI</b></p> <p><b>margins <i>female01</i>, at(<i>age</i>==(20(5)75)) atmeans</b>  <b>quietly: margins <i>female01</i>, at(<i>age</i>==(20(5)75)) atmeans</b>  <b>marginsplot, legend(row(1))</b>  <b>marginsplot, noci legend(row(1))</b></p>	<p>Produces predicted Y at ages 20, 25, ..., 70, 75 separately for <i>female01</i>=0( denotes males) and <i>female01</i>=1(females). All other covariate values are set to their mean values.</p> <p><b>legend(row(1))</b> produces legend at the bottom, rather than on the side.</p>