Unit 5.
Regression and Correlation

" 'Don't let us quarrel,' the White Queen said in an anxious tone. 'What is the cause of lighting?' 'The cause of lightning, 'Alice said very decidedly, for she felt quite certain about this, 'is the thunder-oh no!', she hastily corrected herself. 'I meant the other way.' 'It's too late to correct it,' said the Red Queen: 'when you've once said a thing, that fixes it, and you must take the consequences.' "
- Carroll

Menopause heralds a complex interplay of hormonal and physiologic changes. Some are temporary discomforts (e.g., hot flashes, sleep disturbances, depression). Others are long-term changes that increase the risk of significant chronic health conditions, bone loss and osteoporosis in particular. Recent observations of an association between ***depressive symptoms*** and ***low bone mineral density (BMD)*** raise the intriguing possibility that alleviation of depression might confer a risk benefit with respect to bone mineral density loss and osteoporosis. However, the finding of an association in a simple (one predictor) linear regression model analysis has multiple possible explanations, only one of which is causal. Others include, but are not limited to: (1) the apparent association is an artifact of the confounding effects of exercise, body fat, education, smoking, etc; (2) there is no relationship and we have observed a chance event of low probability (it can happen!); (3) the pathway is the other way around (low BMD causes depressive symptoms), albeit highly unlikely; and/or (4) the finding is spurious due to study design flaws (selection bias, misclassification, etc).

In settings where multiple, related predictors are associated with the outcome of interest, multiple predictor linear regression analysis allows us to study the joint relationships among the multiple predictors (depressive symptoms, exercise, body fat, etc) and a single continuous outcome (BMD).

In this example, we might be especially interested in using multiple predictor linear regression to isolate the effect of depressive symptoms on BMD, holding all other predictors constant (***adjustment***). Or, we might want to investigate the possibility of synergism or ***interaction***.

Nature ——————— **Population/** ———————**Observation/** ——————— **Relationships/** ——————— **Analysis/**
                    Sample   **Data**   **Modeling**   **Synthesis**

Table of Contents

Datasets used (download from course website)
janka.Rdata
p53paper.Rdata
framingham_1000.Rdata

Packages used (one time installation required)
ggplot2   Hmisc
stargazer   car
gridExtra   lmtest
GGally   summarytools

Tip! **Don't forget that R is case sensitive …**

Nature —————— **Population/** —————— **Observation/** —————— **Relationships/** —————— **Analysis/**
Sample **Data** **Modeling** **Synthesis**

1. Learning Objectives

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

- Explain the concepts of <u>association, causation, confounding, mediation,</u> and <u>effect modification;</u>

- Construct and interpret a <u>scatter plot</u> with respect to: evidence of association, assessment of linearity, and the presence of outlying values;

- <u>State</u> the multiple predictor linear regression model and the <u>assumptions</u> necessary for its use;

- Perform and interpret the <u>Shapiro-Wilk</u> and <u>Kolmogorov-Smirnov</u> tests of <u>normality;</u>

- Explain the <u>relevance</u> of the normal probability distribution;

- Explain and interpret the <u>coefficients (and standard error)</u> and <u>analysis of variance</u> tables outputs of a single or multiple predictor regression model estimation;.

- Explain and compare <u>crude</u> versus <u>adjusted estimates (betas) of association;</u>

- Explain and interpret regression model estimates of <u>effect modification (interaction);</u>

- Explain and interpret <u>overall</u> and <u>adjusted R-squared</u> measures of association;

- Explain and interpret <u>overall</u> and <u>partial F-tests;</u>

- Draft an <u>analysis plan</u> for a multiple predictor regression model analysis; and

- Explain and interpret selected regression model diagnostics: <u>residuals, leverage,</u> and <u>Cook's distance</u>.

1. Review

Simple linear regression and correlation were introduced in <u>BIOSTATS 540, Unit 12.</u>

a. Settings Where Regression Might Be Considered

Example #1
Is the density of wood a predictor of hardness of timber?
<u>Source:</u>
Williams, E.J. (1959) Regression Analysis, New York: John Wiley & Sons

Wood density and timber hardness are two different things, with timber hardness being important in many of the products of wood processing. Wood density is pounds of weight per cubic foot of volume, while timber hardness is measure of force. One measure of the latter is the Janka Scale; it defines hardness as the number of pounds required to push a ball bearing into a timber sample using a machine press.   So, as you might imagine, it might be of interest to estimate the parameters that define the relationship between the two so as to obtain a **prediction equation**. Thus, in this example, the predictor (explanatory variable) is wood density and the outcome (response variable) is the Janka Scale hardness score:

> Y = hardness
> X = density

Example #2
Does the expression of p53 change with parity and age?
<u>Source:</u>
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, <u>after adjustment for</u> 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

Nature ——————— Population/ ——————Observation/ ——————— Relationships/ ——————— Analysis/
Sample   Data   Modeling   Synthesis

- 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_1, X_2, X_3, X_4, X_5, X_6):

  Y = p53
  X1 = Age
  X2 = Parity
  X3 = Age at first mensis
  X4 = Family history of breast cancer
  X5 = Menopausal status
  X6 = 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)

- A **simple linear regression** might be performed to characterize the relationship between FEV and height:

  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 roles of other influences on lung function, such as age, height, 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_1 = Exposure to air pollution
  X_2 = Height, inches
  X_3 = Smoking (1=yes, 0=no)

Nature ————— **Population/** —————**Observation/** ————— **Relationships/** ————— **Analysis/**
                  Sample   **Data**   **Modeling**   **Synthesis**

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 are 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 McCullogh (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
  $X_1 =$ Exercise (1=yes if 3x/week or more, 0 = no)
  $X_2 =$ Age, years
  $X_3 =$ Body Mass Index (BMI)
  $X_4 =$ Alcohol Use (1=yes, 0=no)

Nature ——————— **Population/** ——————— **Observation/** ——————— **Relationships/** ——————— **Analysis/**
                     Sample   **Data**   **Modeling**   **Synthesis**

b. 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 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.**

Nature ——————— Population/ ——————— Observation/ ——————— Relationships/ ——————— Analysis/
Sample   Data   Modeling   Synthesis

c. A General Approach for Model Development

There are <u>**no**</u> rules <u>***nor a single best strategy.***</u> Different study designs and research questions call for different approaches. Tip – Before you begin your model development, make alist of your study design, research aims, outcom variable, primary predictor(s), and covariates. As a general suggestion, the following approach has the advantages of providinga reasonably thorough exploration of the data and relatively small risk of missing something important.

Preliminary – **Be sure you have: (1) checked, cleaned and described your data, (2) screened the data for multivariable associations, and (3) thoroughly explored the bivariate 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.
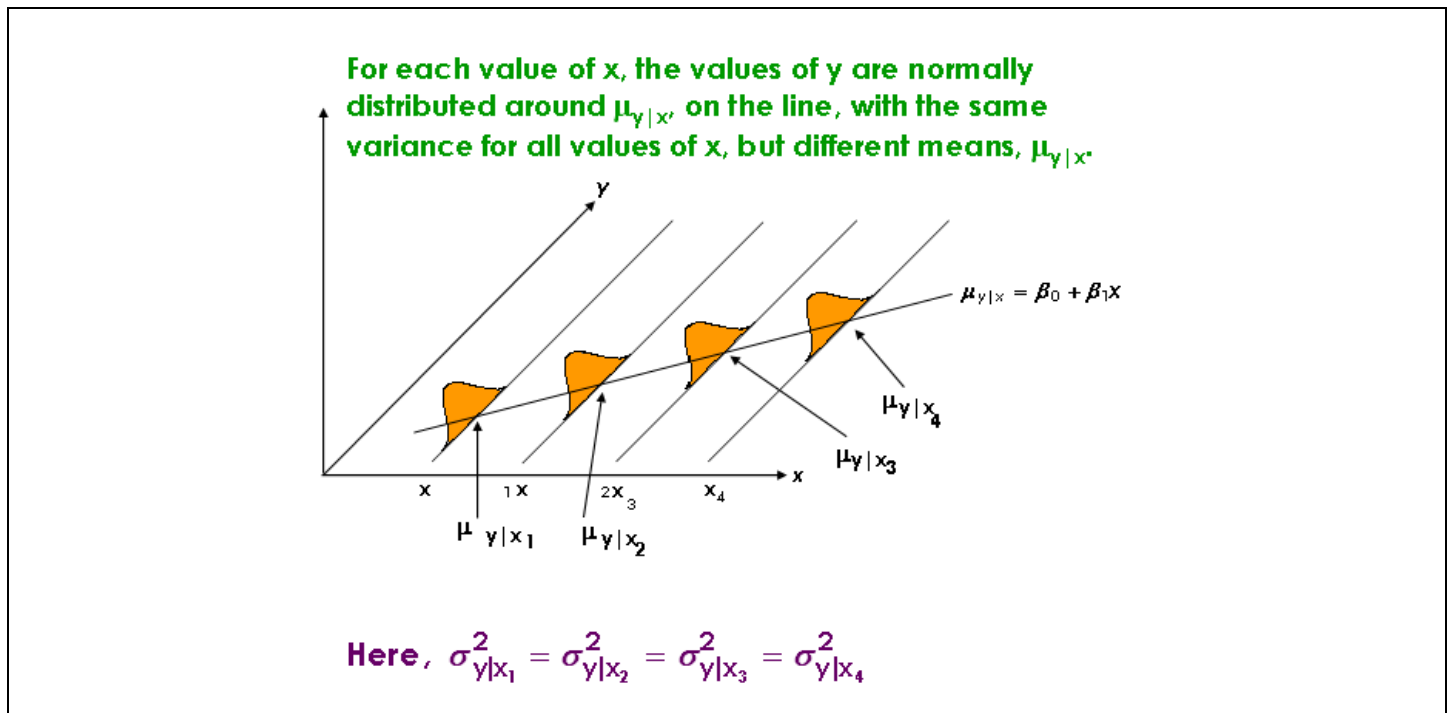
Nature ———— **Population/** ————**Observation/** ———— **Relationships/** ———— **Analysis/**
                     Sample   **Data**   **Modeling**   **Synthesis**

d. Review - Normal Theory Regression

Normal theory regression analysis can be used used to model/investigate possibly complex relationships when:

- The outcome is a **single continuous variable (Y)** that is assumed to be **distributed normal**; and

- The outcome is potentially related to possibly **several predictor variables ($X_1, X_2, …, X_p$)** 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 μ (the average value) of **one continuous, and normally distributed, outcome** random variable Y (e.g. **Y= FEV** for forced expiratory volume) varies linearly with changes in **one continuous predictor** variable X (e.g. **X=Height**). It says that the expected values of the outcome Y, as X changes, lie on a straight line ("regression line").

For each value of x, the values of y are normally distributed around $\mu_{y|x}$, on the line, with the same variance for all values of x, but different means, $\mu_{y|x}$.

$$\mu_{y|x} = \beta_0 + \beta_1 x$$

$$\mu_{y|x_4}$$
$$\mu_{y|x_3}$$
$$\mu_{y|x_1} \quad \mu_{y|x_2}$$

Here, $\sigma^2_{y|x_1} = \sigma^2_{y|x_2} = \sigma^2_{y|x_3} = \sigma^2_{y|x_4}$

Nature ——————— Population/ ———————Observation/ ——————— Relationships/ ——————— Analysis/
                Sample   Data   Modeling   Synthesis

Assumptions of Simple Linear Regression

1.   The outcomes $Y_1, Y_2, \cdots, Y_n$ 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

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

Model
These assumptions say that we are modeling the observed outcome for the ith subject as the sum of two pieces:
1) a model piece; plus 2) an error piece.

Observed for "i"th subject = [Systematic or Predicted for "i"] + [Error or Departure from Predicted for "i"]

that is:

$$Y_i \ = \ [\ \beta_0 + \beta_1 x_i\ ] + \varepsilon_i$$

1.   The errors $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ are **independent**.

2.   Each error ε_i is distributed is **normal** with

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

"Least Squares", "Close" and Least Squares Estimation

Theoretically, it is possible to draw many lines through an X-Y scatter of points. Which to choose? "Least
squares" estimation is one approach to choosing a line that is "closest" to the data.

We choose values for $\hat{\beta}_0$ and $\hat{\beta}_1$ that, upon insertion, minimizes the total

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2 = \sum_{i=1}^{n} \left(Y_i - \left[\hat{\beta}_0 + \hat{\beta}_1 X_i \right] \right)^2$$

Nature ───────── Population/ ─────────Observation/ ───────── Relationships/ ───────── Analysis/
                 Sample   Data   Modeling   Synthesis

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2 = \sum_{i=1}^{n} \left(Y_i - \left[\hat{\beta}_0 + \hat{\beta}_1 X_i \right] \right)^2 \textbf{ has a variety of names:}$$

- ♦ residual sum of squares, SSE or SSQ(residual)
- ♦ sum of squares about the regression line
- ♦ sum of squares due error (SSE)

Least Squares Estimation Solutions

Note – the estimates are denoted either using Greek letters with a caret or with Roman letters

Estimate of Slope $\hat{\beta}_1$ **or** b_1	$$\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} \left(X_i - \overline{X} \right)\left(Y_i - \overline{Y} \right)}{\sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2}$$
Intercept $\hat{\beta}_0$ **or** b_0	$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

Analysis of Variance

Partitioning the Total Variance and all things sum of squares and mean squares

Source	df	Sum of Squares A measure of variability	Mean Square = Sum of Squares / df A measure of average/typical/mean variability
Regression **due model**	1	$SSR = \sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y} \right)^2$	MSR = SSR/1
Residual **due error**	(n-2)	$SSE = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2$	MSE = SSE/(n-2) = $\hat{\sigma}^2_{Y|X}$ Note: also called "mean squared error"
Total, corrected	(n-1)	$SST = \sum_{i=1}^{n} \left(Y_i - \overline{Y} \right)^2$	

Nature ————— Population/ ————— Observation/ ————— Relationships/ ————— Analysis/
Sample   Data   Modeling   Synthesis

2.
Illustration: Fit a Simple Linear Regression Model

Preliminary - Set working directory (change <mark>mine in yellow</mark> to yours)

```
setwd("/Users/cbigelow/Desktop/")
```

Input R dataset `janka.Rdata`. Inspect.
```
library(tidyverse)
load(file="janka.Rdata")                          # Assumes the data are in the working directory
janka$hardness <- as.numeric(janka$hardness)
glimpse(janka)

## Observations: 36
## Variables: 2
## $ density  <dbl> 24.7, 24.8, 27.3, 28.4, 28.4, 29.0, 30.3, 32.7, 35.6, 3…
## $ hardness <dbl> 484, 427, 413, 517, 549, 648, 587, 704, 979, 914, 1070,…

janka

##      density hardness
## 1      24.7      484
## 2      24.8      427
## 3      27.3      413
## 4      28.4      517

…. Rows omitted …

## 34     68.8     2890
## 35     69.1     2740
## 36     69.1     3140
```
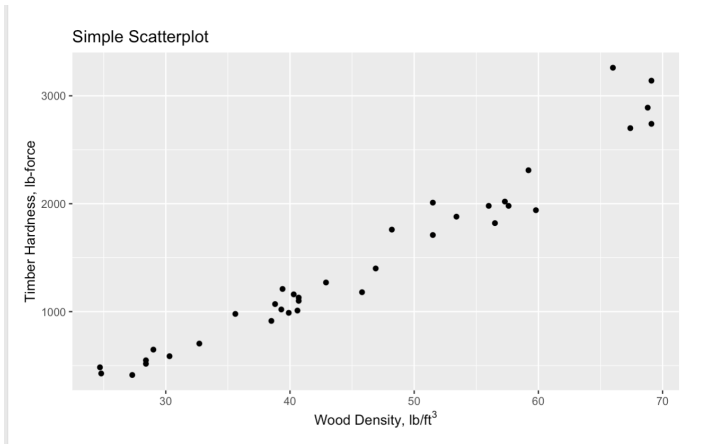
Descriptives using command `stargazer()` in package `stargazer`

```
library(stargazer)
stargazer::stargazer(data=janka,type="text",median=TRUE)
##
## ========================================================================
## Statistic  N     Mean      St. Dev.  Min   Pctl(25) Median Pctl(75)  Max
## ------------------------------------------------------------------------
## density    36   45.733    13.580    24.700  37.775  41.800  56.700  69.100
## hardness   36 1,469.472  801.517    413     962.8   1,195   1,980   3,260
## ------------------------------------------------------------------------
```

Scatterplot using command `ggplot()` and option `geom_point( )` in package `ggplot2`

```
library(ggplot2)
library(ggplot2)
p <- ggplot(data=janka) +                          # specify dataset
    aes(x=density,y=hardness) +                    # define x- and y-axis
    geom_point()  +                                # produce XY scatterplot
    xlab(expression("Wood Density, lb/ft"^{3}))  + # label the x-axis
    ylab("Timber Hardness, lb-force") +            # label the y-axis
    ggtitle("Simple Scatterplot")
p
```

Nature ——————— Population/ —————————Observation/ ——————— Relationships/ —————— Analysis/
                      Sample   Data   Modeling   Synthesis

Simple Scatterplot

Fit Simple Linear Regression. Obtain Coefficients Table. Obtain Analysis of Variance Table

```
# KEY:
# lm( ) fits the model.  Example: MODELNAME <- lm(data=DATAFRAMENAME, YVARIABLE ~ XPREDICTOR)
# summary( ) provides coefficients table and some other info.  Example:  summary(MODELNAME)
# anova( ) produces anova table.               Example:  anova(MODELNAME)

model1 <- lm(data=janka, hardness~density)
summary(model1)
## lm(formula = hardness ~ density, data = janka)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -338.40  -96.98  -15.71   92.71  625.06
##
## Coefficients:
##              Estimate Std. Error t value          Pr(>|t|)
## (Intercept) -1160.500    108.580  -10.69      0.00000000000207 ***
## density        57.507      2.279   25.24 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 183.1 on 34 degrees of freedom
## Multiple R-squared:  0.9493, Adjusted R-squared:  0.9478
## F-statistic:   637 on 1 and 34 DF,  p-value: < 0.00000000000000022
```

Intercept $= \hat{\beta}_0 = b_0 = -1160.500$

Slope $= \hat{\beta}_1 = b_1 = 57.507$

R^2 = % Variance explained = .9493 or ~ 95%

The fitted line is thus: Predicted hardness $=\hat{hardness} = -1160.500 + 57.507 \cdot density$

```
anova(model1)

## Analysis of Variance Table
##
## Response: hardness
##           Df  Sum Sq  Mean Sq F value            Pr(>F)
## density    1 21345674 21345674  636.98 < 0.00000000000000022 ***
## Residuals 34  1139366    33511
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

SSQ(model) = SSR = 21,345,674
SSQ(residual) = SSE = 1,139,366
Overall F-test of null: slope=0
F = 636.98 with df = 1, 34
p-value <<< .0001 REJECT null
Conclude fitted line is significant

Nature ———————— **Population/ Sample** ———————— **Observation/ Data** ———————— **Relationships/ Modeling** ———————— **Analysis/ Synthesis**

3. Multivariable Linear Regression

a. Introduction

In multiple linear regression, the number of explanatory (predictor) variables is > 1. There is still just one outcome (response) variable Y, continuous and assumed distributed <u>normal</u>. The <u>multiple</u> predictors in a linear regression model can be any mix of <u>continuous</u> or <u>discrete</u>.

Definition
A multiple linear regression model is a particular model of how the mean μ (the average value) of one <u>***continuous***</u> outcome random variable Y (e.g. ***Y= length of hospital stay***) varies, depending on the value of two or more (these can be a mixture of continuous and discrete) predictor variables X (e.g. ***X_1=age, X_2=0/1 history of vertebral factures, etc..)*** Specifically, it says that the average values of the outcome variable, as the profiles of predictors X_1 , X_2 , … etc change, lie on a "plane" ("regression plane").

Example
P53 is a tumor suppressor gene that has been extensively studied in breast cancer research. Suppose we are interested in understanding the correlates of p53 expression, especially those that are known breast cancer risk variables. We might hypothesize that p53 expression is related to, among other things, number of pregnancies and age at first pregnancy.

$$Y = \text{p53 expression level}$$
$$X_1 = \text{number of pregnancies (coded 0, 1, 2, etc)}$$
$$X_2 = \text{age at first pregnancy} \leq 24 \text{ years (1=yes, 0=no)}$$
$$X_3 = \text{age at first pregnancy} > 24 \text{ years (1=yes, 0=no)}$$

A multivariable linear model that relates Y to X_1 X_2, and X_3 is the following

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \text{error}$$

Nature ———————— **Population/** ————————**Observation/** ———————— **Relationships/** ———————— **Analysis/**
                      Sample   **Data**   **Modeling**   **Synthesis**

The General Multivariable Linear Model
Similarly, it is possible to consider a multivariable model that includes p predictors:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \text{error}$$

- p = # predictors, _**apart**_ from the intercept

- Each $X_1 \cdots X_p$ can be either discrete or continuous.

- Data are comprised of n data points of the form $(Y_i, X_{1i}, \cdots, X_{pi})$
  Note: The subscript "i" is indexing the individual, while 1, 2, …, p is indexing the predictor

- For the ith individual, we have a vector of predictor variable values
  that is represented $X_i' = \left[X_{1i}, X_{2i}, ..., X_{pi} \right]$

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

1. The sample size = n separate observations $Y_1, Y_2, \cdots, Y_n$ are **independent**.

2. The values of the predictor variables $X_1 \cdots X_p$ are **fixed** and measured without error.

3. For each <u>vector</u> value of the predictor variable $\underline{X}=\underline{x}$, the distribution of values of Y follows a **normal** distribution with mean equal to $\mu_{Y|\underline{X}=\underline{x}}$ and common variance equal to $\sigma_{Y|\underline{x}}^2$.

4. For each profile of values, $x_1, x_2, ..., x_p$, of the p predictor variables $X_1 \cdots X_p$ (written using vector notation $\underline{X}=\underline{x}$), the distribution of values of Y is **normal** with

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

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

Model Fitting (Estimation)

When there are multiple predictors, the least squares fit is multi-dimensional. In the setting of just 2 predictors, it's possible to show a schematic of the fitted plane that results from least squares estimation.

In the picture below, the outcome (dependent variable) is Y=body length and there are two predictors: X_1=glabella length and X_2=glabella width. The purple ellipse is the least squares fit and is a **2-dimensional plane** in 3-dimensional space; it is the analogue of the straight line fit that was explained in simple linear regression.
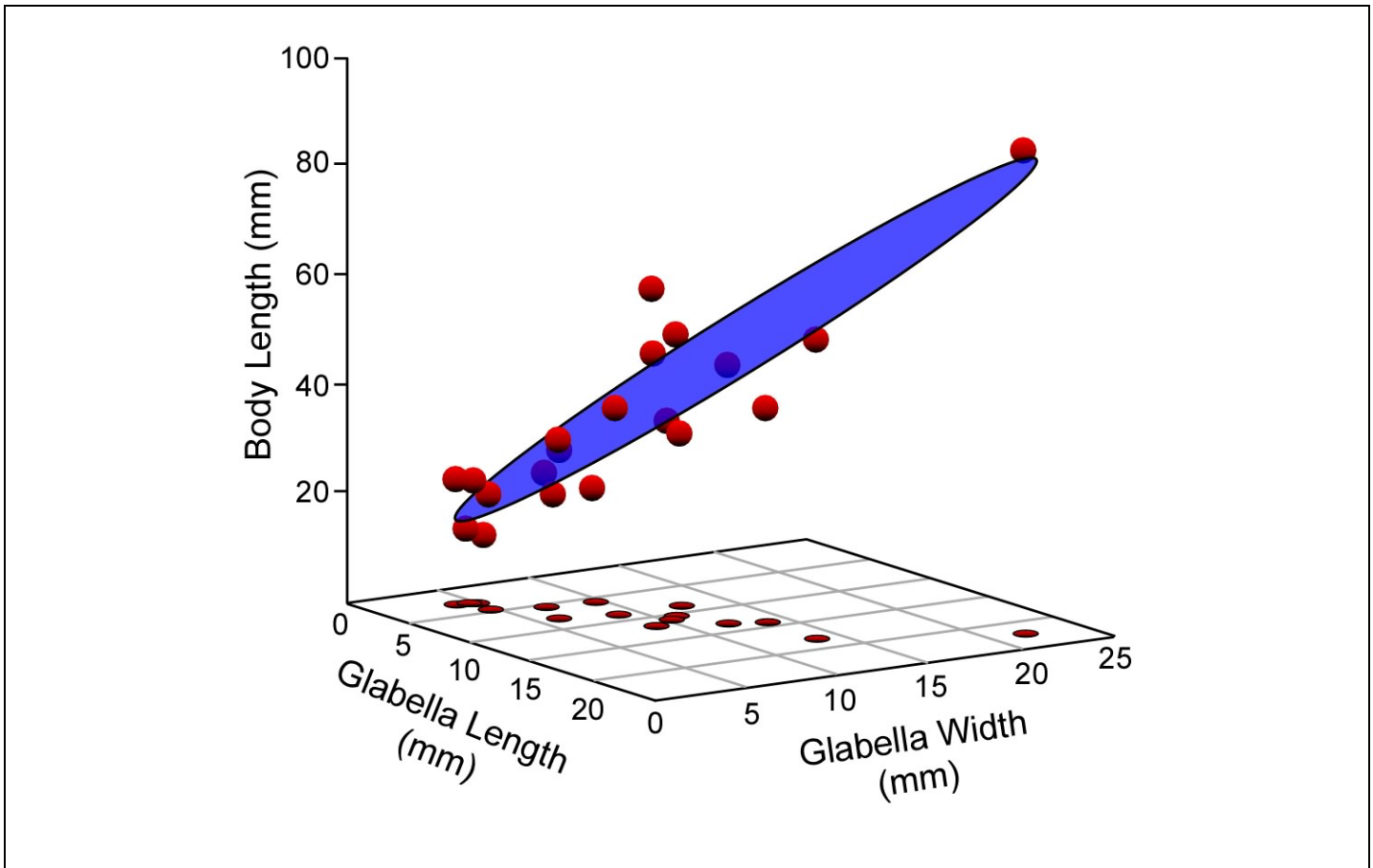
Source: www.palass.org

b. Indicator Variables (also called "dummy variables") and Design Variables

Why Indicator Variables?

Example - Suppose you wanted to model some outcome (Y = duration of stay in ICU, in days) in relationship to type of surgery X with X coded as follows: 1=medical therapy, 2=angioplasty, and 3=coronary bypass surgery).

You just forge on and fit a simple linear model. **Spoiler alert – the following would be incorrect**):

$$days_i = [\beta_0 + \beta_1 * (\text{type of surgery})_i] + \varepsilon_i$$

The notion of slope representing the change in Y=days per 1 unit increase in X=type of surgery doesn't work!

$\beta_1 = \Delta$ Y per 1 unit increase in X, by definition

$= $ Predicted change in duration of stay in ICU per 1 unit increase in TYPE OF SURGERY???

$= $ "makes no sense"

So, what to do?   Answer: 1) we will NOT put X=type of surgery into the model; and 2) instead, we will substitute a set of what are called indicator variables, as described below.

Introduction to Indicator Variables.
Indicator variables are commonly used as predictors in multivariable regression models. We let

$$1 \quad = \text{value of indicator when "trait" is present}$$
$$0 \quad = \text{value of indicator when "trait" is not present}$$

♦   The estimated regression coefficient β associated with an indicator
    variable has a straightforward interpretation, namely:

♦   β = predicted change in outcome Y that accompanies presence of "trait"

Examples of Simple Indicator Variables

SEXF = 1 if individual is female
           0 otherwise

TREAT = 1 if individual received experimental treatment
            0 otherwise

Nature ——————— Population/ ——————Observation/ ——————— Relationships/ ——————— Analysis/
                    Sample   Data   Modeling   Synthesis

When the Nominal Predictor has MORE THAN 2 Possibilities, the set of 0/1 indicator variables are called Design Variables.

If a nominal variable has k possible values, (k-1) indicator variables are needed to distinguish the entire range of possibilities.

Returning to our Example (Y=duration of stay in ICU, X = type of surgery)
Our original predictor variable X is nominal with 3 possible values:

$$X = 1 \text{ if treatment is medical therapy}$$
$$2 \text{ if treatment is angioplasty}$$
$$3 \text{ if treatment is bypass surgery}$$

So, we've agreed that we ***cannot*** put X = type of surgery into a regression model "as is" because the resulting estimated slope makes no sense. Instead, we will put into the model TWO 0/1 indicator variables. One is a "flag" of angioplasty. The other is a "flag" of bypass surgery. Folks receiving medical therapy are then those for whom NEITHER flag occurs.   For example, we might include the following set:

$$TR\_ANG = 1 \text{ if treatment is angioplasty (X=2)}$$
$$0 \text{ otherwise}$$

$$TR\_SUR = 1 \text{ if treatment is bypass surgery (X=3)}$$
$$0 \text{ otherwise}$$

A set of design variables comprised of (3-1) = 2 indicator variables summarize three possible values of treatment. The reference category is medical therapy.

Value of Original Predictor X	Value of 0/1 Indicator TR_ANG	Value of 0/1 Indicator TR_SUR
X=1 ("medical"), the **"referent"**	0	0
X=2 ("angioplasty")	1	0
X=3 ("surgery")	0	1

Guidelines for the Definition of Indicator and Design Variables

1) Consider the choice of the referent group.
Often this choice will be straightforward. It might be one of the following categories of values of the nominal variable:

- The unexposed
- The placebo
- The standard
- The most frequent

Nature ──────── Population/ ────────Observation/ ──────── Relationships/ ──────── Analysis/
Sample   Data   Modeling   Synthesis

2) <u>K levels of the nominal predictor → (K-1) indicator variables</u>
When the number of levels of the nominal predictor variable = k, define (k-1) indicator variables that will identify persons in each of the separate groups, apart from the reference group.

3) <u>In general (this is not hard and fast), treat the (k-1) design variables as a set.</u>

- Enter the set together
- Remove the set together
- In general, retain all (k-1) of the indicator variables, even when only a subset are significant.

c. Interaction Variables

Sometimes the nature of an X-Y relationship is ***different (meaning the slope is different)***, depending on the level of some third variable which, for now, we'll call Z. This is **interaction.** To capture how an X-Y relationship is "different" (or "modified by"), depending on the level of Z, we can define an **interaction variable** and then incorporate it as an additional predictor in the model.

Interaction of predictor X with third variable $Z = XZ = X*Z$

Example: Y = length of stay
X = age (years)
Z = 0/1 indicator of history of vertebral fracture (Z=0 for NON fractures and Z=1 for fractures)
XZ = [X] * [Z] = interaction of X and Z

Our full model is thus the following:

$$Y = \beta_0 + \beta_1 Z + \beta_2 X + \beta_3 XZ$$

Key to the betas:

β_0 = intercept for referent (the referent group are patients with $Z = 0$, the non-vertebral fracture folks)
β_1 = CHANGE in INTERCEPT (associated with Z=1, that is - associated with vertebral fracture)
β_2 = slope of change in Y per unit X for referent group
β_3 = CHANGE in SLOPE associated with Z=1 (that is - associated with vertebral fracture)

What does the model become when Z=0?
This yields the model for the non-vertebral fractures patients. Insertion of Z=0 yields

$$Y = \beta_0 + \beta_2 X$$
$$\text{Intercept} = \beta_0$$
$$\text{Slope} = \beta_2$$

What does the model become when Z=1.
This yields the model for the vertebral fractures patients. Insertion of Z=1 yields

$$Y = [\beta_0 + \beta_1] + [\beta_2 + \beta_3]X$$
$$\text{Intercept} = [\beta_0 + \beta_1]$$
$$\text{Slope} = [\beta_2 + \beta_3]$$

Nature ——————— Population/ ——————— Observation/ ——————— Relationships/ ——————— Analysis/
Sample   Data   Modeling   Synthesis

d. *Look!* Schematic of Confounding and Effect Modification

The tools of indicator variables and interaction variables are helpful (but not without important caveats) in exploring the data for evidence of confounding and effect modification.
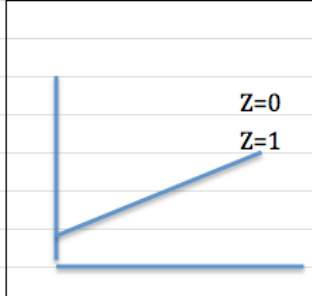
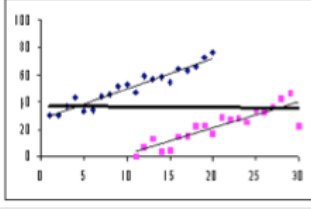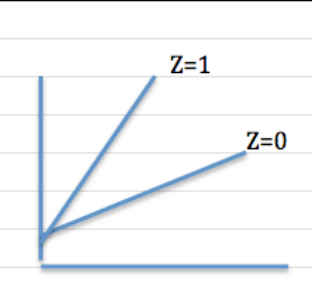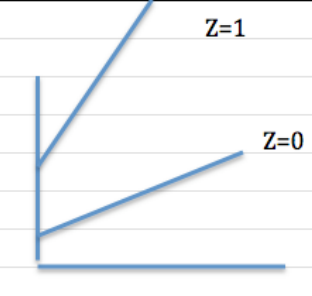Consider a similar regression setting:
Y = length of hospital stay
X = duration of surgery, continuous
Z = a nominal predictor coded 0 for "no comorbidities" and coded 1 for "one or more comorbidities".

Associated with Z=1 (the patient has comorbidities), relative to Z=0 (the referent patient with no comorbidities), the X-Y relationship might have a different intercept, or a different slope, or a different intercept and a different slope.

Take a look.

Coincident	Confounding (admittedly extreme!)	Effect Modification	Effect Modification
$Y = \beta_0 + \beta_1 X$	$Y = \beta_0 + \beta_1 X + \beta_2 Z$	$Y = \beta_0 + \beta_1 X + \beta_2 XZ$ where $XZ = X*Z$	$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ$ where $XZ = X*Z$
Comorbidities=0: $Y = \beta_0 + \beta_1 X$	**Comorbidities=0:** $Y = \beta_0 + \beta_1 X$	**Comorbidities=0:** $Y = \beta_0 + \beta_1 X$	**Comorbidities=0:** $Y = \beta_0 + \beta_1 X$
Comorbidities=1: $Y = \beta_0 + \beta_1 X$	**Comorbidities=1:** $Y = (\beta_0 + \beta_2) + \beta_1 X$	**Comorbidities=1:** $Y = \beta_0 + (\beta_1 + \beta_2)X$	**Comorbidities=1:** $Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)X$
	$\beta_2 =$ **change** in intercept associated with presence of any comorbidities	$\beta_2 =$ **change** in slope of Y on X associated with presence of any comorbidities	$\beta_2 =$ **change** in intercept $\beta_3 =$ **change** in slope of Y on X

Nature ————— Population/ ————— Observation/ ————— Relationships/ ————— Analysis/
          Sample   Data   Modeling   Synthesis

e. The Analysis of Variance Table

The ideas of the analysis of variance table introduced in BIOSTATS 540 (*Unit 12, Simple Linear Regression and Correlation*) apply here, as well.

1. **SST:** "Total" or "total, corrected"

     ◆ $SST = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$ is the variability of Y about $\overline{Y}$

     ◆   Degrees of freedom = df = (n-1).

2. **SSR** "Regression" or "due model"

     ◆ $SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$ is the variability of $\hat{Y}$ about $\overline{Y}$

     ◆   Degrees of freedom = df = p = # predictors **_apart_** from intercept

3. **SSE:** "Residual" or "due error" refers to the

     ◆ $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ is the variability of Y about $\hat{Y}$

     ◆   Degrees of freedom = df = (n-1) – (p)

Source	df	Sum of Squares	Mean Square
Model	p	$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$	MSR = SSR/p
Residual	(n-1) - p	$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	MSE = SSE)/(n-1-p)
Total, corrected	(n-1)	$SST = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$	

Overall F Test

The overall F test also applies, yielding an overall F-test to assess the significance of the variance explained by the model. Note that the degrees of freedom is different here; this is because there are now "p" predictors instead of 1 predictor.

$$H_O: \quad \beta_1 = \beta_2 = \; ... \; = \beta_p = 0$$

$$H_A: \quad \text{At least one } \beta_i \; \neq \; 0$$

$$F_{OVERALL} \quad = \quad \frac{\text{mean square due model}}{\text{mean square due residual}} = \frac{MSR}{MSE} = \frac{SSR/(p)}{SSE/(n-1-p)} \quad \text{with df} = p, (n\text{-}1\text{-}p)$$

Rejection of the null occurs for large values of $F_{OVERALL}$ with accompanying small p-value. With rejection of the null, we conclude at least one predictor (we don't know which ones) has a slope that is statistically significantly different from zero.

Example - Consider a multiple linear regression analysis of the relationship of Y=p53 expression to age at first pregnancy (**pregnum**), 1st pregnancy at age $\leq$ 24 (**early**), and 1st pregnancy at age > 24 (**late**). The variables **early** and **late** are each 0/1. The referent group is nulliparous.

R illustration
The following assumes that you have downloaded **p53paper.Rdata** from the course website

```
load(file="p53paper.Rdata")
fit <- lm(p53 ~ pregnum + early + late, data=p53paper)
summary(fit)

##
## Call:
## lm(formula = p53 ~ pregnum + early + late, data = p53paper)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.86030 -0.57031  0.01611  0.51611  2.62100
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.57031    0.24088  10.671 9.36e-16 ***
## pregnum      0.37641    0.20087   1.874   0.0656 .
## early        0.16076    0.55559   0.289   0.7733
## late        -0.06772    0.50174  -0.135   0.8931
## ---
##
## The fitted line is:  p̂53 = 2.57 + 0.38·pregnum + 0.16·early − 0.07·late
##
## Residual standard error: 0.9635 on 63 degrees of freedom
##    (1 observation deleted due to missingness)
## Multiple R-squared:  0.203,  Adjusted R-squared:  0.165
## F-statistic: 5.349 on 3 and 63 DF,  p-value: 0.002402    The overall F-test of the null hypothesis of
zero slopes on every predictor is rejected.  Conclude at least one slope is statistically significantly
different from zero.
```

Nature ——————— Population/ ———————Observation/ ——————— Relationships/ ——————— Analysis/
                        Sample   Data   Modeling   Synthesis

```
anova(fit)

## Analysis of Variance Table
##
## Response: p53
##           Df Sum Sq Mean Sq F value    Pr(>F)
## pregnum    1 14.330 14.3301 15.4359 0.0002146 ***
## early      1  0.550  0.5497  0.5921 0.4444682
## late       1  0.017  0.0169  0.0182 0.8930686
## Residuals 63 58.487  0.9284
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

H0: $\beta_{PREGNUM} = 0$ and $\beta_{EARLY} = 0$ and $\beta_{LATE} = 0$
HA: At least one slope $\neq 0$

$$F_{3,63} = \frac{\text{mean square due model}}{\text{mean square due residual}} = \frac{MSR}{MSE} = \frac{SSR/(p)}{SSE/(n-1-p)}$$

$$F_{3,63} = \frac{\text{msq(Model)}}{\text{msq(Residual)}} = \frac{(14.330 + 0.550 + 0.017)/3}{(58.487)/63} = \frac{4.96557054}{0.9284} = 5.349$$

The overall F-test of the null hypothesis of zero slopes on every predictor is rejected (p-value = .002;
see previous page). Conclude at least one slope is statistically significantly different from zero.
different from zero. Important: all we can say at this point, however, is that the model that was fit
explains statistically significantly more of the variability in Y = p53 than is explained by "no model" at
all (the intercept only model).

Nature ——————— **Population/** ——————— **Observation/** ——————— **Relationships/** ——————— **Analysis/**
                      Sample   **Data**   **Modeling**   **Synthesis**

f. The Partial F Test
The partial F test is among our key tools in model development in multiple linear regression.

<u>What if we want to compare and choose between two models?</u>
The partial F test is a statistical technique for comparing two models that are "hierarchical." It permits the assessment of associations while controlling for confounding.

<u>What are hierarchical models?</u>

- Hierarchical models are <u>two models of a particular type</u>.   One model is called "smaller" or "reduced" or "reference". The other model is called "larger" or "comparison."

- <u>"Hierarchical"</u> means that all of the predictors in the smaller (reduced, reference) are contained in the larger (comparison) model.

- In the **Y = p53** example, we might be interested in comparing the following two hierarchical models:

     Predictors in smaller model = { **pregnum** }
     Predictors in larger model = { **pregnum** + **early** + **late**}

- <u>"Hierarchical"</u> is satisfied because all of the predictors (here there is just one - **pregnum**) that are contained in the smaller model are contained in the larger model.

- The important point to note is this. The comparison of these two models is an analysis of the nature and significance of the <u>extra predictors</u>, (here - **early** and **late**) for the prediction of Y=p53, ***<u>adjusting for (controlling for)</u>*** all of the variables in the smaller model (**pregnum**).

Thus, the comparison of the hierarchical models is addressing the following question:

     What is the significance of **early** and **late** for the prediction of **Y = p53**, *<u>after controlling for</u>* the effects of **pregnum**?

Statistical Definition of the Partial F Test

Research Question: Does inclusion of the _"extra"_ predictors explain significantly more of the variability in outcome compared to the variability that is explained by the predictors that are already in the model?

<div style="border:1px solid;">

Partial F Test

H$_O$: Addition of $X_{p+1} \cdots X_{p+k}$ is of no statistical significance for the prediction of Y
after controlling for the predictors $X_1 \cdots X_p$ meaning that:

$$\beta_{p+1} = \beta_{p+2} = \ldots = \beta_{p+k} = 0 \text{ after adjustment for } X_1 \ldots X_p$$

H$_A$: Not

$$F_{PARTIAL} \quad = \quad \frac{\{ \text{Extra regression sum of squares} \} / \{ \text{Extra regression df} \}}{\{ \text{Residual sum of squares } \underline{larger} \text{ model}\} / \{ \text{Residual df } \underline{larger} \text{ model} \}}$$

$$= \quad \frac{\left[SSR(X_1 \ldots X_p, X_{p+1}, \ldots X_{p+k}) - SSR(X_1 \ldots X_p)\right] \Big/ [(p+k)-p]}{\left[SSE(X_1 \ldots X_p, X_{p+1}, \ldots X_{p+k})\right] \Big/ [(n-1)-(p+k)]}$$

Numerator df = $(p+k) - (p) = k$
Denominator df = $(n-1) - (p+k)$

</div>

H$_O$ true: The extra predictors are not significant in adjusted analysis	F value = small p-value = large
H$_O$ false: The extra predictors are significant in adjusted analysis	F value = large p-value = small

R illustration Example – continued.

```
reduced <- lm(data=p53paper, p53 ~ pregnum)              HO:  Controlling for pregnum,
full <- lm(data=p53paper, p53 ~ pregnum + early + late)       βEARLY = 0 and βLATE = 0
anova(reduced, full)                                    HA:  At least one is of "ADDED" significance,
                                                             after adjustment (controlling for) pregnum


Analysis of Variance Table
## Model 1: p53 ~ pregnum
## Model 2: p53 ~ pregnum + early + late
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     65 59.054
## 2     63 58.487  2   0.56663 0.3052 0.7381        telling us  →     Fpartial = .3052      p-value = .7361
```

The null hypothesis is NOT rejected (p-value = .74). Conclude that early and late are not predictive after adjustment for pregnum. Specifically, their addition to the model does not explain statistically significantly more of the variability in Y=p53 beyond that explained by pregnum.

Nature ———————— Population/ ————————— Observation/ ——————————— Relationships/ ——————————— Analysis/
Sample   Data   Modeling   Synthesis

g. Multiple Partial Correlation

Good to know. The partial F test is a hypothesis test. In contrast, a partial correlation is a measure (expressed as a percent) of what is explained. The two are not the same thing!

A partial correlation addresses the question: *"To what extent is Y correlated with X (or multiple X), after accounting for some control variable Z (or multiple control variables Z)?*

One way to think about a partial correlation is to think along the following lines

- **Preliminary 1**: Regress the predictor X on the control variable Z
  - Obtain the residuals
  - These residuals represent the information in the predictor X that is independent of Z

- **Preliminary 2**: Now regress the outcome Y on the control variable Z
  - Obtain the residuals
  - These residuals represent the information in Y that is independent of Z

- The correlation between these two sets of residuals, correlation(residuals of Y on Z, residuals of X on Z) give you a Z-controlled assessment of the relationship between X and Y, that is, independent of Z.

Two ways to think about it.

Method 1 – as a correlation
R $_{XY|Z}$ = Multiple Partial correlation (X,Y | controlling for Z)
= Correlation (residuals of X regressed on Z, residuals of Y regressed on Z)

Method 2 – as a squared correlation
R^2 $_{XY|Z}$ = Multiple *Squared* Partial correlation (X,Y | controlling for Z)

$$= \frac{SSR(\text{due Model with Z and X}) - SSR(\text{due Model with Z alone})}{SSE(\text{due residual in Z only model})}$$

Putting this all together, and keeping track of the distinctions …

Partial F Test	**Partial Multiple Correlation Squared**
Goal: Hypothesis test of significance of extra variables, after adjustment for the control variables.	Goal: Estimation of percent of variability in outcome Y that is explained by the extra variables, independent of the control variables.

Control variables: X_1 …. X_p
   Extra variables: X_{p+1} … X_{p+k}
	Control variables: X_1 …. X_p
	   Extra variables: X_{p+1} … X_{p+k}

$F_{PARTIAL}$ hypothesis test compares mean squares to mean squares	$R^2{}_{partial}$ multiple partial correlation squared compares sum of squares to sum of squares
The denominator has the FULL model	The denominator has the REDUCED model
$$= \dfrac{\left[SSR(X_1 \ldots X_p, X_{p+1}, \ldots X_{p+k}) - SSR(X_1 \ldots X_p) \right] \big/ [(p+k) - p]}{\left[SSE(X_1 \ldots X_p, X_{p+1}, \ldots X_{p+k}) \right] \big/ [(n-1) - (p+k)]}$$	$$= \dfrac{SSR(\text{due Model with all}) \ - \ SSR(\text{due Model control only})}{SSE(\text{due residual in Z only model})}$$

4. Multivariable Model Development

a. Introduction

Recall from page 7 …. The **goal** of statistical modeling is to obtain a model that is simultaneously **minimally adequate** and a **good fit**.   **And the model should make sense.**

We also learned some general guidelines for how to proceed (Take care – there is no single right answer here).

> **Preliminary –**
> Be sure you have: (1) checked, cleaned and described your data, (2) screened the data for multivariate associations, and (3) thoroughly explored the bivariate relationships.

> **Step 1 –**
> **Fit the "maximal" model.**.

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

> **Step 3 –**
> **Keep simplifying the model.**
> Repeat step 2, over and over, until the model remaining contains nothing but significant predictor variables.

> **Then there is a Step 4 -**
> **Perform regression diagnostics**
> We'll get to this later, *Section 5. Goodness-of-Fit and Regression Diagnostics*

b. Example

Framingham Study

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 use a subset of the data in a subset of n=1000.

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/m2)	
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?

Nature ——————— Population/ ——————— Observation/ ——————— Relationships/ ——————— Analysis/
                    Sample   Data   Modeling   Synthesis

Input Data. Check. Produce descriptives:

```
# KEY:
# rm(list=ls()) clears the workspace

rm(list=ls())
setwd("/Users/cbigelow/Desktop/")

load(file="framingham_1000.Rdata")

# Dear class – I don't want to have to keep writing out Framingham_1000
# So I'm creating a new dataframe with a shorter name
framingham <- framingham_1000

summary(framingham)

##      sex           sbp             scl              age
##   Men  :443   Min.   : 80.0   Min.   :115.0   Min.   :30.00
##   Women:557   1st Qu.:116.0   1st Qu.:197.0   1st Qu.:38.75
##               Median :128.0   Median :225.0   Median :45.00
##               Mean   :132.3   Mean   :227.8   Mean   :45.92
##               3rd Qu.:144.0   3rd Qu.:255.0   3rd Qu.:53.00
##               Max.   :270.0   Max.   :493.0   Max.   :66.00
##                               NA's   :4                        There are 4 missing values of scl
##      bmi             id            ln_bmi           ln_sbp
##   Min.   :16.40   Min.   :   1   Min.   :2.797   Min.   :4.382
##   1st Qu.:23.00   1st Qu.:1246   1st Qu.:3.135   1st Qu.:4.754
##   Median :25.10   Median :2488   Median :3.223   Median :4.852
##   Mean   :25.57   Mean   :2410   Mean   :3.230   Mean   :4.872
##   3rd Qu.:27.80   3rd Qu.:3605   3rd Qu.:3.325   3rd Qu.:4.970
##   Max.   :43.40   Max.   :4697   Max.   :3.770   Max.   :5.598
##   NA's   :2                      NA's   :2                       There are 2 missing values of bmi, ln_bmi
##      ln_scl
##   Min.   :4.745
##   1st Qu.:5.283
##   Median :5.416
##   Mean   :5.410
##   3rd Qu.:5.541
##   Max.   :6.201
##   NA's   :4                                                      There are 4 missing values of ln_scl

library(stargazer)
stargazer::stargazer(framingham, type="text", median=TRUE)

##
## =========================================================
## Statistic   N     Mean    St. Dev.   Min   Median   Max
## ---------------------------------------------------------
## sbp        1,000  132.350  23.043     80    128     270
## scl          996  227.846  45.087    115    225     493
## age        1,000  45.922    8.545     30     45      66
## bmi          998  25.566    3.848   16.400 25.100  43.400
## id         1,000 2,410.031 1,363.439   1   2,487.5 4,697
## ln_bmi       998   3.230    0.147   2.797   3.223   3.770
## ln_sbp     1,000   4.872    0.163   4.382   4.852   5.598
## ln_scl       996   5.410    0.195   4.745   5.416   6.201     Nicer layout, slightly different info
## ---------------------------------------------------------

Examination of the ranges of systolic bp, age, bmi look to be all plausible; no suggestion of significant
errors in the data itself.
```

Nature —————— **Population/** —————— **Observation/** —————— **Relationships/** —————— **Analysis/**
                        Sample   **Data**   **Modeling**   **Synthesis**

```
library(summarytools)
summarytools::freq(framingham$sex)

## Frequencies
## framingham$sex
## Type: Factor (unordered)
##
##              Freq    % Valid   % Valid Cum.    % Total   % Total Cum.
## ----------- ------ --------- -------------- --------- --------------
##        Men    443     44.30          44.30     44.30          44.30
##      Women    557     55.70         100.00     55.70         100.00
##       <NA>      0                               0.00         100.00
##      Total   1000    100.00         100.00    100.00         100.00

library(summarytools)
summarytools::descr(framingham$sbp, stats = c("n.valid","mean", "sd", "min","q1", "med", "q3", "max","CV"),
transpose = TRUE)

## Descriptive Statistics
## framingham$sbp
## N: 1000
##
##            N.Valid      Mean   Std.Dev      Min      Q1    Median       Q3      Max      CV
## --------- --------- --------- --------- ------- -------- -------- -------- -------- ------
##      sbp   1000.00    132.35     23.04    80.00   116.00   128.00   144.00   270.00    0.17
```

Assess Normality of Candidate Dependent Variable = sbp. Shapiro-Wilk Test (Null: normality)
Histogram w Overlay Normal and QQ Plot

```
options(scipen=1000)
shapiro.test(framingham$sbp)

##
##  Shapiro-Wilk normality test
##
## data:  framingham$sbp
## W = 0.92121, p-value < 0.00000000000000022
```

Interpretation: The null hypothesis of normality of the distribution of sbp is rejected (p << .00001)
```
library(ggplot2)
library(gridExtra)
# p1 is panel 1 = histogram w overlay normal

p1 <- ggplot(data=framingham, aes(x=sbp)) +
      geom_histogram(binwidth=5, colour="blue",
                       aes(y=..density..)) +
      stat_function(fun=dnorm,
                      color="red",
                      args=list(mean=mean(framingham$sbp),
                                 sd=sd(framingham$sbp))) +
     ggtitle("Systolic Blood Pressure (sbp)") +
     xlab("Systolic Blood Pressure (mm Hg)") +
     ylab("Density") +
     theme_bw() +
     theme(axis.text = element_text(size = 10),
        axis.title = element_text(size = 10),
        plot.title = element_text(size = 12))
```
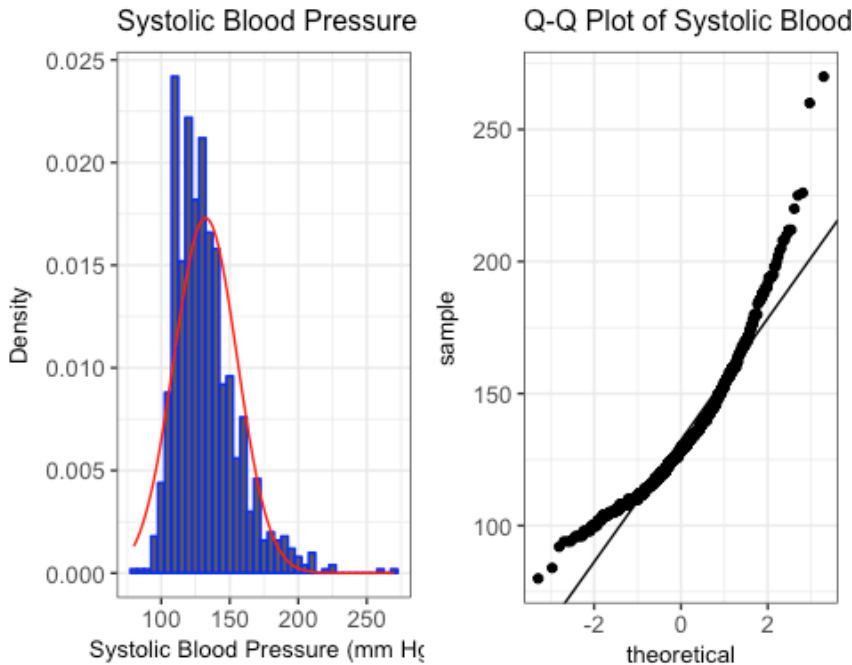
```
# p2 is panel 2 = quantile-quantile plot
p2 <- ggplot(data=framingham, aes(sample=sbp)) +
      stat_qq() +
      geom_abline(intercept=mean(framingham$sbp), slope = sd(framingham$sbp)) +
      ggtitle("Q-Q Plot of Systolic Blood Pressure (sbp)") +
      theme_bw() +
      theme(axis.text = element_text(size = 10),
        axis.title = element_text(size = 10),
        plot.title = element_text(size = 12))

gridExtra::grid.arrange(p1, p2, ncol=2)
```


Interpretation: This confirms what the Shapiro Wilk test suggests. The null hypothesis of normality of the distribution of sbp is not supported.

Create "regression-friendly" indicator variables and interactions. Check.

```
library(summarytools)
library(Hmisc)
# KEY to true/false operator to define indicator:
# If original variable sex="Women" then code new variable female=1.  Otherwise, code new variable female=0
# na.rm=TRUE ensures that missing values will not be considered.
framingham$female <- as.numeric(framingham$sex == "Women", na.rm=TRUE)
summarytools::ctable(framingham$sex,framingham$female,prop = 'n', totals = FALSE)

## Cross-Tabulation
## Variables: sex * female
## Data Frame: framingham
##
## ------- -------- ----- -----
##          female    0    1             female is the new indicator variable created and is coded 0/1
##     sex                            sex is the original variable used to create female
##     Men            443    0
##   Women             0   557          It worked!
## ------- -------- ----- -----
```

| Nature | ——————— | Population/ Sample | ——————— | Observation/ Data | ——————— | Relationships/ Modeling | ——————— | Analysis/ Synthesis |

```
Hmisc::label(framingham$female) <- "female01"

framingham$ageXfemale <- framingham$age*framingham$female
Hmisc::label(framingham$ageXfemale) <- "AGE x FEMALE interaction"


framingham$lnsclXfemale <- framingham$ln_scl*framingham$female
Hmisc::label(framingham$lnsclXfemale) <- "ln(scl) x FEMALE interaction"


framingham$lnbmiXfemale <- framingham$ln_bmi*framingham$female
Hmisc::label(framingham$lnbmiXfemale) <- "ln(bmi) x FEMALE interaction"
```
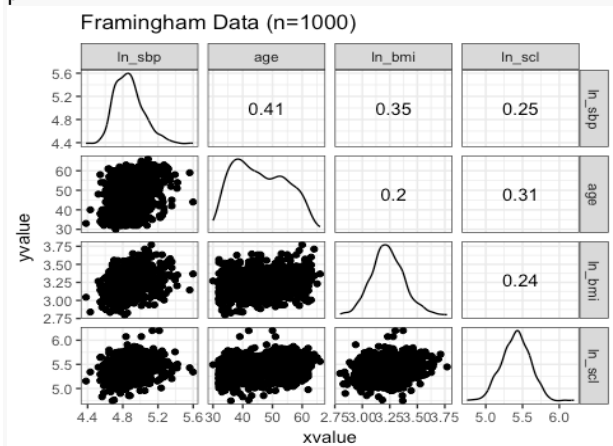
Examine Pairwise Relationships: 1) Y with X's; and 2) X's with X's

```
library(GGally)
p <- GGally::ggscatmat(data=framingham, columns=c("ln_sbp","age","ln_bmi","ln_scl"))
p <- p + ggtitle("Framingham Data (n=1000)")
p <- p + theme_bw()
p
```


Create a dataset that has no missing values on any variables of interest. Name this dataset complete.
Then fit the following five (5) models named as follows
m_maximal: Contains all predictors
m_2: Drops 2 interactions - lnbmiXfemale and lnsclXfemale
m_3: One predictor model w predictor = ln_bmi
m_4: One predictor model w predictor = ln_scl
m_5: Three predictor model w predictors = age, female, and ageXfemale

```
library(stargazer)
# KEY to creating a dataset that has NO missing values on any variables of interest
# NEWDATATFRAME <- na.omit(data=SOURCEDATA, cols=c("var","var", …"var"))
complete <- na.omit(framingham, cols=c("ln_sbp", "ln_bmi", "age", "female", "lnbmiXfemale", "lnsclXfemale","ageXfemale"))

# Now we can fit the 5 models to the EXACTLY the same observations
m_maximal <- lm(data=complete, ln_sbp ~ ln_bmi + ln_scl + age + female + lnbmiXfemale + lnsclXfemale + ageXfemale)
m_2 <- lm(data=complete, ln_sbp ~ ln_bmi + ln_scl + age + female + ageXfemale)
m_3 <- lm(data=complete, ln_sbp ~ ln_bmi)
m_4 <- lm(data=complete, ln_sbp ~ ln_scl)
m_5 <- lm(data=complete, ln_sbp ~ age + female + ageXfemale)
```

Nature ——————— Population/ ——————— Observation/ ——————— Relationships/ ——————— Analysis/
                         Sample   Data   Modeling   Synthesis

```
# Use command stargazer() in package stargazer to display models side by side
stargazer::stargazer(m_maximal,m_2,m_3,m_4,m_5,type="text",font.size="small", align=TRUE, omit.stat=c("f", "ser"))
```

```
===========================================================
                          Dependent variable:
              ---------------------------------------------
                                 ln_sbp
                 (1)       (2)       (3)       (4)       (5)
-----------------------------------------------------------
ln_bmi        0.304*** 0.271*** 0.388***
              (0.055)   (0.032)   (0.033)

ln_scl         0.059    0.056**            0.211***
              (0.037)   (0.025)             (0.026)

age           0.004*** 0.004***                      0.004***
              (0.001)   (0.001)                       (0.001)

female        -0.011   -0.217***                     -0.327***
              (0.304)   (0.051)                       (0.051)

lnbmiXfemale  -0.051
              (0.067)

lnsclXfemale  -0.009
              (0.050)

ageXfemale    0.005*** 0.005***                      0.007***
              (0.001)   (0.001)                       (0.001)

Constant      3.396*** 3.521*** 3.618*** 3.730*** 4.701***
              (0.234)   (0.159)   (0.106)   (0.139)   (0.039)

-----------------------------------------------------------
Observations    994       994       994       994       994
R2             0.267     0.266     0.123     0.064     0.203
Adjusted R2    0.261     0.262     0.122     0.063     0.200
===========================================================
Note:                       *p<0.1; **p<0.05; ***p<0.01
```

Models 1 & 2 have nearly identical R^2 = % variance explained. This suggests the extra predictors are not needed. -> Model 2 is preferred (simpler!)

```
# Use command anova(REDUCEDMODEL, FULLMODEL) to obtain partial F Tests
paste("Partial F-test, 2df:  Null:  lnbmiXfemale=0 lnsclXfemale=0")
anova(m_2, m_maximal)
```

```
[1] "Partial F-test, 2df:  Null:  lnbmiXfemale=0 lnsclXfemale=0"
Analysis of Variance Table

Model 1: ln_sbp ~ ln_bmi + ln_scl + age + female + ageXfemale
Model 2: ln_sbp ~ ln_bmi + ln_scl + age + female + lnbmiXfemale + lnsclXfemale + ageXfemale

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    988 19.314
2    986 19.301  2  0.013173    0.3365 0.7144
```

Interpretation – This confirms that it is okay to DROP lnbmi_female and lnscl_female (Partial F = 0.34, p-value = .71) nsSo, model 2 is our "tentative" final model

Further work, regression diagnostics, are needed next (See, section 5. *Goodness-of-Fit and Regression Diagnostics*).

Nature ───────── Population/ ───────Observation/ ───────── Relationships/ ───────── Analysis/
                 Sample   Data   Modeling   Synthesis

c. Suggested Criteria for Confounding and Interaction

A Suggested Statistical Criterion for Determination of Confounding

A variable Z might be judged to be a confounder of an X-Y relationship if
BOTH of the following are satisfied:

1) Its inclusion in a model that already contains X as a predictor has
   adjusted significance level < .10 or < .05; _and_

2) Its inclusion in the model changes the estimated regression coefficient for
   X by 15-20% or more, relative to the model that contains only X as a
   predictor.

A Suggested Statistical Criterion for Assessment of Interaction

**A "candidate" interaction variable might be judged to be worth retaining in
the model if BOTH of the following are satisfied:**

1) The partial F test for its inclusion has significance level < .05; _and_

2) Its inclusion in the model alters the estimated regression coefficient for
   the main effects by 15-20% or more.

d. Additional Tips for Multivariable Analysis of Large Data Sets

#1. State the Research Questions.
Aim for a focus that is explicit, complete, and focused, including:

- Statement of population

- Definition of outcome

- Specification of hypotheses (predictor-outcome relationships)

- Identification of (including nature of) hypothesized covariate relationships

#2. Define the Analysis Variables.
For each research question, note for each analysis variable, its hypothesized role.

- Outcome

- Predictor

- Confounder

- Effect Modifier

- Intermediary (also called intervening)

#3. Prepare a "Clean" Data Set Ready for Analysis (Data Management)
For each variable, check its distribution, especially:

- Completeness

- Occurrence of logical errors

- Within form consistency

- Between form consistency

- Range

Nature ————— Population/ ————— Observation/ ————— Relationships/ ————— Analysis/
              Sample   Data   Modeling   Synthesis

#4. Describe the Analysis Sample
This description serves three purposes:

1) Identifies the population actually represented by the sample

2) Defines the range(s) of relationships that can be explored

3) Identifies, tentatively, the function form of the relationships

Methods include:

♦ Frequency distributions for discrete variables

♦ Mean, standard deviation, percentiles for continuous variables

♦ Bar charts

♦ Box and whisker plots

♦ Scatter plots

#5. Assessment of Confounding
The identification of confounders is needed for the correct interpretation of the predictor-outcome relationships. Confounders need to be controlled in analyses of predictor-outcome relationships.

Methods include:

♦ Cross-tabulations and single predictor regression models to determine whether suspected confounders are predictive of outcome and are related to the predictor of interest.

♦ This step should include a determination that there is a confounder-exposure relationship among controls.

#6. Single Predictor Regression Model Analyses
The fit of these models identifies the nature and magnitude of crude associations. It also permits assessment of the appropriateness of the assumed functional form of the predictor-outcome relationship.

♦ Cross-tabulations
♦ Graphical displays (Scatter plots)
♦ Estimation of single predictor models

Nature ——————— **Population/** ——————— **Observation/** ——————— **Relationships/** ——————— **Analysis/**
Sample   **Data**   **Modeling**   **Synthesis**

5. Goodness-of-Fit and Regression Diagnostics

a. Introduction and Terminology

Neither prediction nor estimation has meaning when the estimated model is a poor fit to the data:
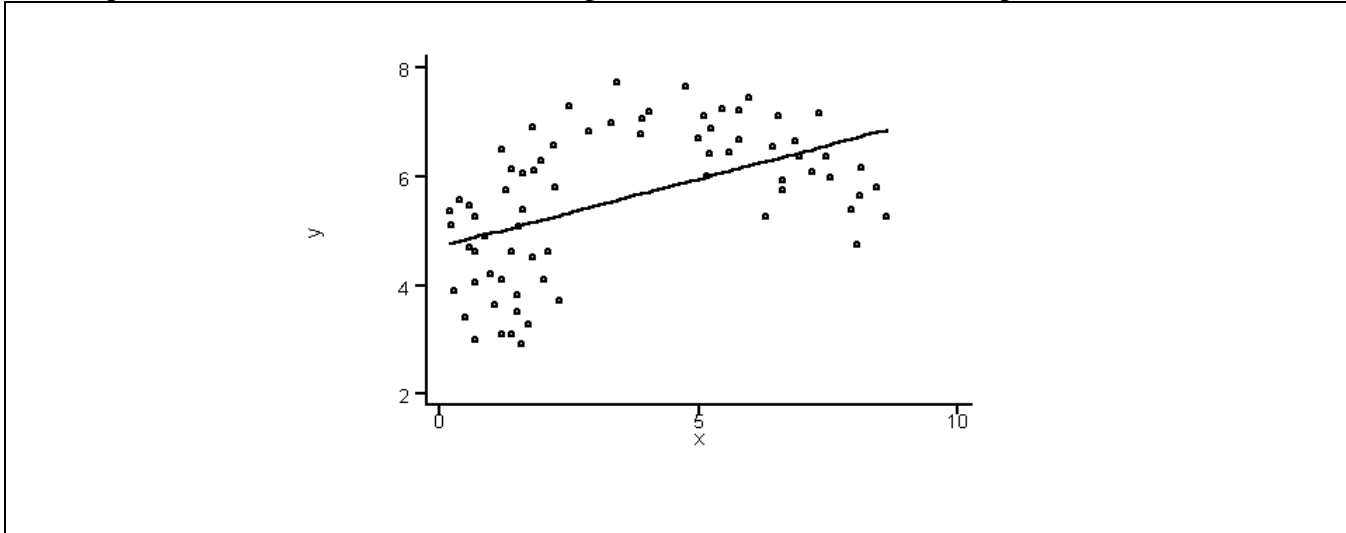
Our eye "tells" us:

♦   A better fitting relationship between X and Y is quadratic
♦   We notice different sizes of discrepancies
♦   Some observed Y are close to the fitted $\hat{Y}$ (e.g. near X=1 or X=8)
♦   Other observed Y are very far from the fitted $\hat{Y}$ (e.g. near X=5)

Poor fits of the data to a fitted line can occur for several reasons and can occur even when the fitted line explains a large proportion (R^2) of the total variability in response:

♦   The wrong functional form (link function) was fit.

♦   Extreme values (outliers) exhibit uniquely large discrepancies between observed and fitted values.

♦   One or more important explanatory variables have been omitted.

♦   One or more model assumptions have been violated.

Consequences of a poor fit include:

- ♦ We learn the wrong biology.

- ♦ Comparison of group differences aren't "fair" because they are unduly influenced by a minority.

- ♦ Comparison of group means aren't "fair" because we used the wrong standard error.

- ♦ Predictions are wrong because the fitted model does not apply to the case of interest.

Available techniques of goodness-of-fit assessment are of two types:

1. **Systematic** - those that explore the appropriateness of the model itself

    Have we fit the correct model?
    Should we fit another model?

2. **Case Analysis** – those that investigate the influence of individual data points

    Are there a small number of individuals whose inclusion in the analysis influences excessively the choice of the fitted model?

Goodness-of-Fit Assessment
Some Terminology

The Multiple Linear Regression Model, again:

$$Y_{\underline{X}} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \text{error}$$

Observed   =   **systematic**   +   **error**
This is the mean of Y

at X_1, X_2, …, X_p
= E[Y at $\underline{X}$] =

Systematic Component

Link:	The functional form (and the assumed underlying distribution of the errors) is sometimes called the link.
	Example: When μ is the mean of a normal distribution, we model $\mu_{Y\vert\underline{X}} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$ This is called the **natural** or **linear** link.
	Example: When μ is a proportion, we might model $\ln[\mu_{Y\vert\underline{X}}/(1-\mu_{Y\vert\underline{X}})] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$. This is called the **logit** link.
Normality:	In the linear model regression analysis, we assume that the errors E follow a Normal(0, $\sigma^2{}_{Y\vert\underline{X}}$) distribution. Recall: The errors ε are estimated by the residuals e.
Heteroscedasticity:	If the assumption of constant variance of the errors E is not true, we say there is heteroscedasticity of errors, or non-homogeneity of errors.

Goodness-of-Fit Assessment
**Some Terminology - continued**

Case Analysis

Residual:	The residual is the difference between the observed outcome Y and the fitted outcome $\hat{Y}$. $$e = \left[Y - \hat{Y} \right]$$ It estimates the unobservable error ε.
Outlier:	An outlier is a residual that is <u>unusually</u> large. *Note:* As before, we will rescale the sizes of the residuals via standardization so that we can interpret their magnitudes on the scale of SE units.
Leverage:	The leverage is a measure of the unusualness of the value of the predictor X. Leverage = distance (observed X, center of X in sample) Predictor values with high leverages have, potentially, a large influence on the choice of the fitted model.
Influence:	Measures of influence gauge the change in the fitted model with the omission of the data point. **Example: Cook's Distance**

A Feel for Residual, Leverage, Influence
Large residuals may or may not be influential
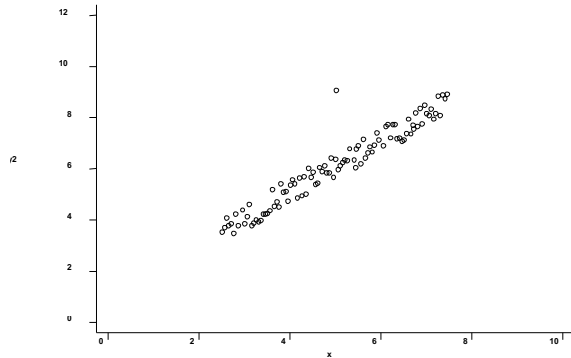
Large residual
Low leverage

The large residual effects a
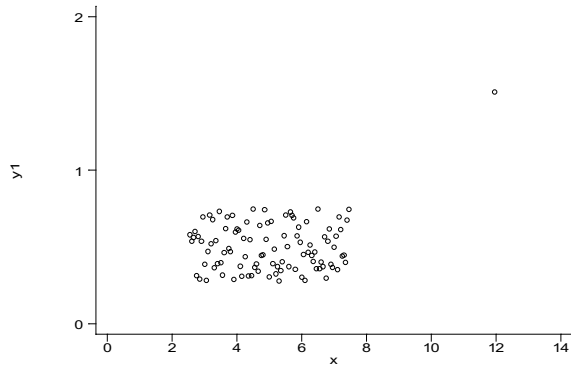<u>large</u> influence.
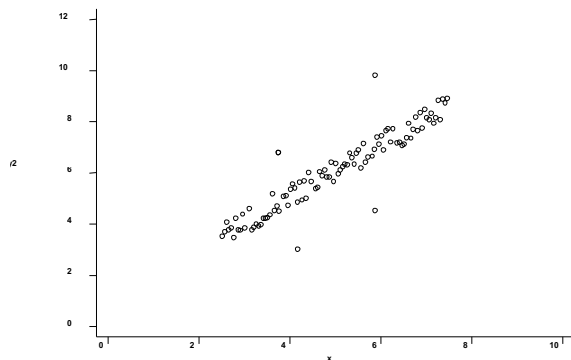
Large residual
Low leverage

Despite its size, the large residual
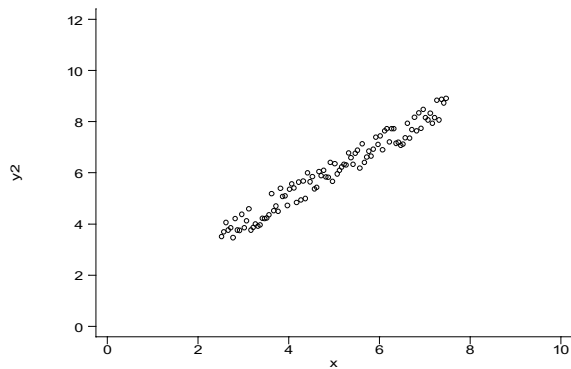effects only <u>small</u> influence.

A Feel for Residual, Leverage, Influence

High leverage may or may not be influential

High leverage
Small residual

The high leverage effects a <u>large</u> influence.

High leverage
Small residual

Despite its size, the large leverage effects only <u>small</u> influence.

Thus, case analysis is needed to discover all of:

♦   high leverage
♦   large residuals
♦   large influence

Overview of Techniques of Goodness-of-Fit Assessment
Linear Model

	Question Addressed	**Procedure**
Systematic Component	Error Distribution: Is it reasonable to assume a normal distribution of errors with a constant variance? H_O: $E \sim$ Normal $(0, \sigma^2)$	Shapiro-Wilk test of normality Cook-Weisberg test of heteroscedasticity
	Functional Form: Is the choice of functional form relating the predictors to outcome a "good" one?	Method of fractional polynomials.
	Systematic Violation: Have we failed to include any important explanatory (predictor) variables?	Ramsey Test for omitted variables.
Case Analysis	Are there outliers with respect to the <u>outcome</u> values?	Studentized residuals
	Are there outliers with respect to the <u>predictor</u> variable values?	Leverage
	Are there individual observations with unduly large influence on the fitted model?	Cook's distance (influence)

Nature ————— Population/ ————— Observation/ ————— Relationships/ ————— Analysis/
Sample   Data   Modeling   Synthesis

## b. Assessment of Normality

Recall what we are assuming with respect to normality:

- **Simple Linear Regression**:
  At each level "x" of the predictor variable X, the outcomes Y_X are distributed normal with mean = $\mu_{Y|x} = \beta_0 + \beta_1 x$ and constant variance $\sigma^2_{Y|x}$

- **Multiple Linear Regression:**
  At each vector level "$\underline{x} = [x_1, x_2, \ldots, x_p]$" of the predictor vector $\underline{X}$, the outcomes Y_X are distributed normal with mean = $\mu_{Y|\underline{x}} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$ and constant variance $\sigma^2_{Y|x}$

This is what it looks like (courtesy of a picture on the web!)
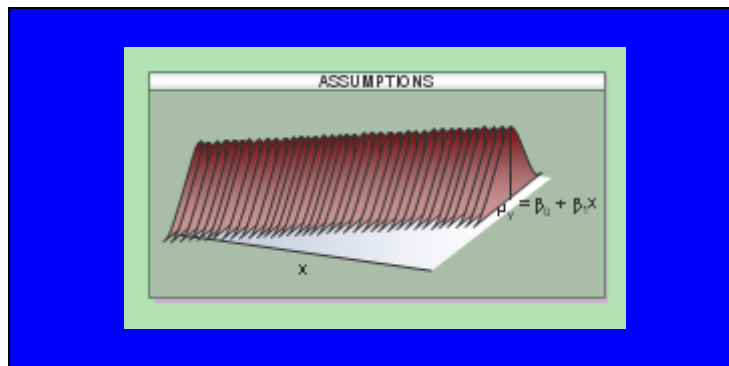
Violations of Normality are sometimes, but not always, a serious problem

- **When not to worry**: Estimation and hypothesis tests of regression parameters are fairly robust to modest violations of normality

- **When to worry**: Predictions are sensitive to violations of normality

- *Beware*: Sometimes the cure for violations of normality is worse than the problem.

Some graphical assessments of normality and what to watch out for:

Method	What to watch out for:
1. Histogram of outcome variable Y and/or Histogram of residuals	Look for normal shape of the histogram.
2. Histogram of residuals (or studentized or jackknife residuals)	Look for normal shape of the histogram.
3. Quantile quantile plot of the quantiles of the residuals versus the quantiles of the assumed normal distribution of the residuals.	Normally distributed residuals will appear, approximately, linear.

Two Panel Graph: 1) Histogram w Overlay Normal + 2) QQ plot

```
# Tip- Might want to tweak binwidth
library(ggplot2)
library(gridExtra)
# Left Panel
# ggplot(data= DATAFRAME, aes(x=VARIABLENAME)) + stuff below
p1 <- ggplot(data=framingham, aes(x=ln_sbp)) +
    geom_histogram(binwidth=.05, colour="blue",
                       aes(y=..density..)) +
    stat_function(fun=dnorm,
                      color="red",
                      args=list(mean=mean(framingham$ln_sbp),
                                    sd=sd(framingham$ln_sbp))) +
    ggtitle("Histogram of ln_sbp w Overlay Normal") +
    xlab("ln_sbp") +
    ylab("Density") +
    theme_bw() +
    theme(axis.text = element_text(size = 9),
       axis.title = element_text(size = 9),
       plot.title = element_text(size = 10))


# Right Panel
p2 <- ggplot(data=framingham, aes(sample=ln_sbp)) +
    stat_qq()   +
    geom_abline(intercept=mean(framingham$ln_sbp), slope = sd(framingham$ln_sbp)) +
    ggtitle("Q-Q Plot of ln[Systolic Blood Pressure (ln_sbp)]") +
    theme_bw() +
    theme(axis.text = element_text(size = 9),
       axis.title = element_text(size = 9),
       plot.title = element_text(size = 10))


gridExtra::grid.arrange(p1, p2, ncol=2)
```
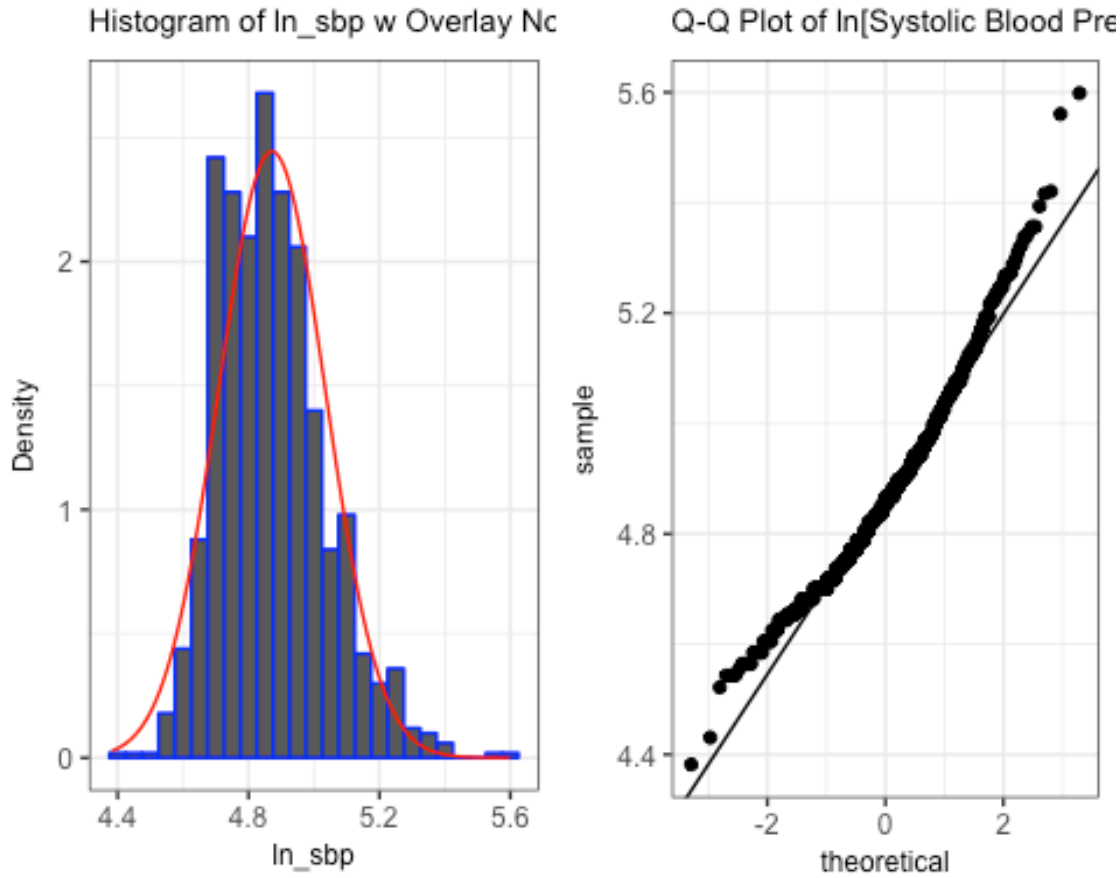
Nature ——————— Population/ ——————— Observation/ ——————— Relationships/ ——————— Analysis/
               Sample   Data   Modeling   Synthesis

Skewness and Kurtosis Statistics for Assessing Normality:

	What to watch out for:
Skewness - symmetry of the curve Standardization of the 3rd sample moment about the mean $m_2 = E\left[(Y\text{-}\mu)^2\right]$ $m_3 = E\left[(Y\text{-}\mu)^3\right]$ What is actually examined is $a_3 = \dfrac{m_3}{(m_2)^{3/2}}$ because it is unitless $a_3 = 0$ indicates symmetry $a_3 < 0$ indicates lefthand skew (tail to left) $a_3 > 0$ indicates right hand skew (tail to right)	When *yvariable* is distributed normal: 　Skewness = 0 　**Look for** skewness between -2 and +2, roughly.

Nature ——————— **Population/** ——————— **Observation/** ——————— **Relationships/** ——————— **Analysis/**
　　　　　　　　　　　　　　Sample　　　　　　　　　　**Data**　　　　　　　　　　**Modeling**　　　　　　　　　　**Synthesis**

Kurtosis – flatness versus peakedness of the curve
Standardization of the 4th sample moment about the mean

$$m_2 = E\left[(Y-\mu)^2 \right]$$

$$m_4 = E\left[(Y-\mu)^4 \right]$$

Pearson kurtosis is $a_4 = \dfrac{m_4}{\left(m_2\right)^2}$

$a_4 = 3$ when distribution is normal
$a_4 < 3$ is "leptokurtic" is too little in the tails
$a_4 > 3$ is "platykurtic" is too much in the tails

When **yvariable** is distributed normal:

Kurtosis = 3

Hypothesis Tests of Normality and what to watch out for:

Test Statistic	What to watch out for:

1. Shapiro Wilk (W)
 	
 W is a measure of the correlation between the values in the sample and their associated normal scores (for review of Normal Scores, see BE540 Topic 5 – Normal Distribution)
 	
 W = 1 under normality
	Null Hypothesis H_O: **yvariable** is distributed normal:
	Alternative Hypothesis H_A: Not.
	
	Evidence of violation of normality is reflected in
	    W < 1
	    small p-value

2. Kolmogorov-Smirnov (D). See also Lilliefors (K-S)
 	
 This is a goodness of fit test that compares the distribution of the residuals to that of a reference normal distribution using a chi square test.
 	
 Lilliefors utilizes a correction
	Evidence of violation of normality is reflected in
	
	   D > 0
	
	   K-S > 0
	
	    small p-value

Guidelines

In practice, the assessment of normality is made after assessment of other model assumption violations.
The linear model is often more robust to violations of the assumption of normality.
The cure, is often worse than the problem. (e.g. – transformation of the outcome variable)

Consider doing a <u>scatterplot of the residuals</u>. Look for

- ♦ Bell shaped pattern
- ♦ Center at zero
- ♦ No gross outliers

c. Cook-Weisberg Test of Heteroscedasticity

Recall what we are assuming with respect to homogeneity of variance:

- **In Simple Linear Regression**:
  At each level "x" of the predictor variable X, the outcomes Y are
  distributed normal with mean = $\mu_{Y|x}$ = $\beta_0 + \beta_1 x$ and constant variance $\sigma^2_{Y|x}$

Evidence of a violation of homogeneity (this is heteroscecasticity) is seen when

- There is increasing or decreasing variation in the residuals with fitted $\hat{Y}$

- There is increasing or decreasing variation in the residuals with predictor X

Some graphical assessments of homogeneity of variance and what to watch out for:

Method	What to watch out for:
1. Plot Residuals or standardized residuals or studentized residuals on the vertical – versus - Predicted outcomes $\hat{Y}$ on the horizontal	Look for even band at zero
2. Plot Residuals or standardized residuals or studentized residuals on the vertical – versus - Predictor values X	Look for even band at zero

Hypothesis Test of homogeneity of variance is Cook-Weisberg

Cook-Weisberg Test	What to watch out for:
This test is based on a model of the variance as a function of the fitted values (or the predictor X). Specifically, it is a chi square test of whether the squared standardized residuals are linearly related to the fitted values (or the predictor X).	Evidence of violation of homogeneity of variance is reflected in Large test statistic > 0 small p-value

d. The Method of Fractional Polynomials

This method is beyond the scope of this course. However, it's helpful to understand the ideas.

Goal: The goal is to select a "good" functional form that relates Y to X from a collection of candidate models. Candidates are lower polynomials and members of the Box-Tidwell family.

Fractional Polynomials: Instead of $Y = \beta_0 + \beta_1 X$, we consider the following:

Instead of fitting a
simple linear relationship of the form
$$\beta_1 X$$

We consider fitting a
fractional polynomial relationship of the form
$$\beta_1 X^{P_1} + \beta_2 X^{P_2} + \beta_3 X^{P_3} + ... + \beta_m X^{P_m}$$

where

m = number of powers ("degree")
p₁,p₂, p₃, ...pₘ are choices from a special set of 8 candidate powers = **{ -2, -1, -0.5, 0, 0.5, 1, 2, 3 }**

And where, when powers repeat
E.g. - when **p₂ = p₁** we consider $\beta_1 X^{P_1} + \beta_2 X^{P_1} \ln(X)$.

Example: Suppose **m=1** with **p₁ = 1** . This yields
$$Y = \beta_0 + \beta_1 X$$

Example: Next, suppose **m=2** with **p₁ = 0.5** and **p₂ = 0.5.** Because **p₂ = p₁** this yields
$$Y = \beta_0 + \beta_1 \sqrt{X} + \beta_2 \sqrt{X} \ln(X)$$

Nature ─────── Population/ ───────Observation/ ─────── Relationships/ ─────── Analysis/
                        Sample   Data   Modeling   Synthesis

The Method of Fractional Polynomials - Continued

Guidelines

Competing models are assessed using a chi square statistic that compares the likelihoods of the data under each of the two models using what is called a "deviance" statistic.

Don't worry: We will learn about the "deviance" statistic in BIOSTATS 640 Unit 5 in the context of the logistic regression model.

Search begins with examination of all models for which m=1. We choose the one model in this class that has the smallest deviance.

♦   We compare the best m=1 model to the specific model for which
m=1 and **p_1= 1** because the latter is the simple linear model.

♦   Thus, we are asking whether it is really necessary to abandon the simple linear model.

Next, we compare the best m=1 model to the best m=2 model. And so on …

♦   In general, we must choose between two costs:

1) A smaller model has a lower goodness-of-fit but more generalizability

2) A larger model has a higher goodness-of-fit but less generalizability

♦ Our goal is to choose the smallest model for which the goodness-of-fit is acceptable.

e. Ramsey Test for Omitted Variables

A fitted model that fails to include an important explanatory variable is problematic.

♦ Our understanding of the outcomes is incomplete.
♦ Estimated associations may be biased due to confounding.
♦ Model assumptions may be violated.

Method of the Ramsey Test

♦ H$_O$: Predicted values from the fitted model are unrelated to powers
    of the fitted model, after adjustment for the predictor variables in
    the model.

$$\text{corr}\left(\hat{Y}, \hat{Y}^p\right) = 0$$

♦ For example, we might fit the model $\hat{Y} = \beta_0 + \beta_1 \hat{Y} + \beta_2 \hat{Y}^2 + \beta_3 X + \text{error}$
    and test the significance of $\hat{\beta}_1$ and $\hat{\beta}_2$.

♦ The test statistic is an F statistic.

Guidelines

Evidence of a failure to include one or more explanatory variables is reflected in a large F statistic value.

As a suggestion, do also a scatterplot of the squared standardized residuals versus the leverage values.
Omission of an important explanatory variables is suggested by

♦ Extreme values
♦ Any systematic pattern

f. Residuals, Leverage, and Cook's Distance

Residuals - There are multiple measures of "residual".

Ordinary residual	Standardized residual
$e = \left(Y - \hat{Y}\right)$	$e^* = \dfrac{e}{\sqrt{ms(residual)}} = \dfrac{e}{\sqrt{\hat{\sigma}^2_{Y\mid x}}}$
Studentized residual	Jacknife residual, also called Standardized deleted residual
$e^* = \dfrac{e}{\sqrt{ms(residual)}\sqrt{1-h}} = \dfrac{e}{\sqrt{\hat{\sigma}^2_{Y\mid x}}\sqrt{1-h}}$	$e^* = \dfrac{e}{\sqrt{ms(residual)_{-i}}\sqrt{1-h}} = \dfrac{e}{\sqrt{\hat{\sigma}^2_{Y\mid \dot{x}}}\sqrt{1-h}}$

Which one or ones should we use?

- Standardized residuals can be appreciated as we do z-scores.

- Studentized residuals are distributed Student's t (df=n-p-1) when regression assumptions hold.

- Jacknife residuals are distributed Student's t (df=n-p-2) when regression assumptions hold. These also have the advantage of correcting the magnitude of the $\sqrt{MS(residual)}$ when it is otherwise too big because of the effects of influential points.

Leverage, h:

Leverage is the distance of a predictor value X=x from the center of the values of the predictor value $X = \bar{x}$. This distance is denoted h_i.

For simple linear regression, $h_i = \dfrac{1}{n} + \dfrac{\left(x_i - \bar{x}\right)^2}{\sum\limits_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$

For simple linear regression, a "large" leverage value is $h_i \geq \dfrac{4}{n}$

Cook's Distance, d

Recall from our pictures that neither a large residual alone nor a high leverage determine the influence of an individual data point.

Cook's distance provides a measure of the influence of an individual data point on the fitted model and is a function of the values of both the residual and leverage:

Cook's Distance = Change in estimated regression coefficient
value, expressed in standard error units.

1) For simple linear regression $d = \dfrac{e^2 h}{2s^2 (1-h)^2}$

2) For multivariable linear regression models $d_i = \dfrac{\left(\hat{\beta}_{-i} - \hat{\beta}\right)' \left(X'X\right)\left(\hat{\beta}_{-i} - \hat{\beta}\right)}{p's_{Y|x}^2}$

where

i indexes the individual for which measure of influence is sought

$\hat{\beta}$ = vector of estimated regression coefficients using the entire sample

$\hat{\beta}_{-i}$ = vector of estimated regression coefficients with omission of the
    i^{th} data point

X = matrix of values of the predictor variables

p' = rank (X) = number of predictors + 1

Guidelines

♦ For the linear regression model, a "noteworthy" influential data point is one for which $d \geq 1$.

For a multivariable regression model, a "noteworthy" influential data
point is one for which $d \geq 2(p+1)/n$ where p=# predictors.

Nature ——————— **Population/** ——————— **Observation/** ——————— **Relationships/** ——————— **Analysis/**
                    Sample   **Data**   **Modeling**   **Synthesis**

g. Example

Framingham Study – model #2
Plot of Observed v Predicted. Look for: Points along a straight line ("all is well")

```
library(ggplot2)
library(Hmisc)
m_best <- lm(data=complete, ln_sbp ~ ln_bmi + ln_scl + age + female + ageXfemale)    # Fit model
complete$yhat <- predict(m_best)                                                     # Add predicted values to dataset
Hmisc::label(complete$yhat) <- "Predicted ln(sbp)"
p <- ggplot(data=complete, aes(x=ln_sbp,y=yhat)) +
    geom_smooth(method=lm, se=FALSE) +                                               # TIP - plot line first
    geom_point() +                                                                   # Then plot your points on top
    xlab("Observed ln_sbp") +
    ylab("Predicted ln_sbp") +
    ggtitle("Framingham Regression of ln_sbp (Model #2)") +
    theme_bw()
p
```


Framingham Regression of ln_sbp (Model #2)
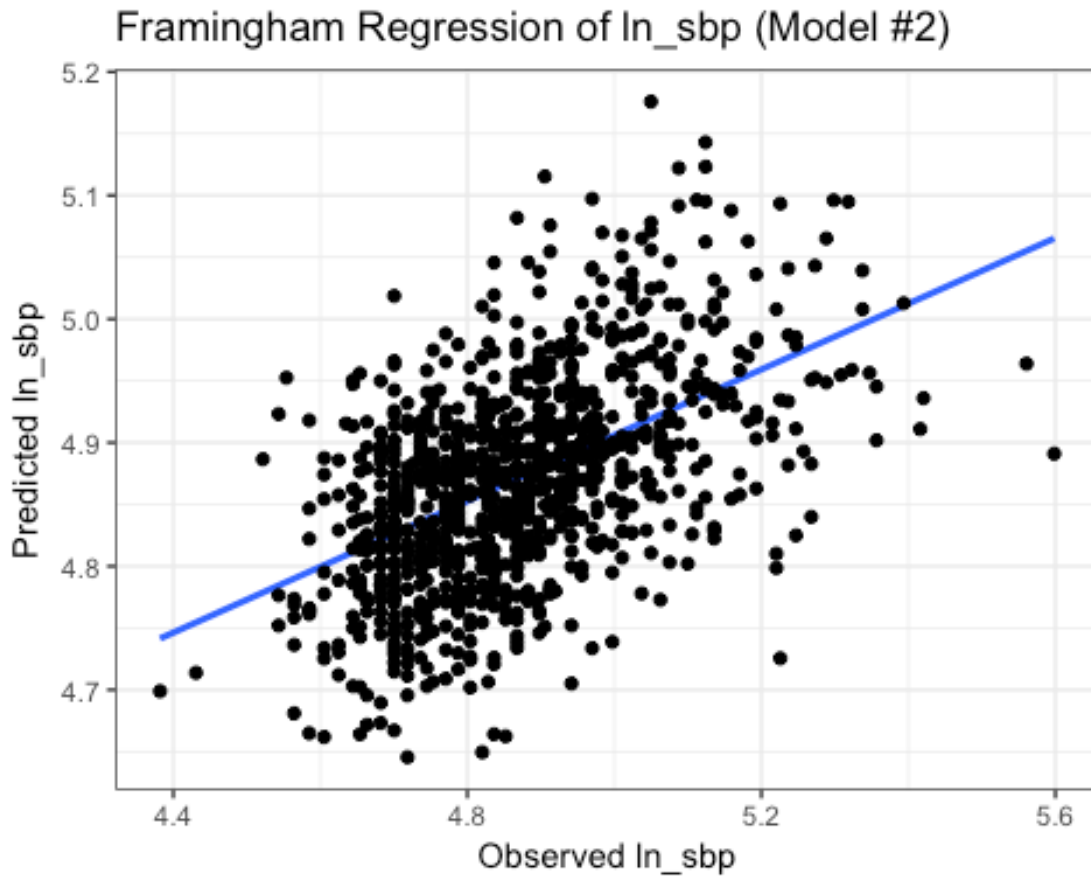
Interpretation – Not bad! Ideally, the scatter lies on the line defined by 45 degrees. We expect some widening of the confidence intervals at the ends of the range but not too much. What we see here is reasonable.

Nature —————— **Population/ Sample** —————— **Observation/ Data** —————— **Relationships/ Modeling** —————— **Analysis/ Synthesis**

Normality of Residuals – QQ Plot and Shapiro Wilk Test. Null: Normality ("all is well")

```
library(ggplot2)
complete$residuals <- residuals(m_best)                          # Add the residuals to the dataset
p <- ggplot(data=complete, aes(sample=residuals)) +
    stat_qq() +
    geom_abline(intercept=mean(complete$residuals),
                    slope = sd(complete$residuals)) +
    ggtitle("Q-Q Plot of Residuals (model #2)") +
    theme_bw() +
    theme(axis.text = element_text(size = 9),
        axis.title = element_text(size = 9),
        plot.title = element_text(size = 10))
p
```


Q-Q Plot of Residuals (model #2)
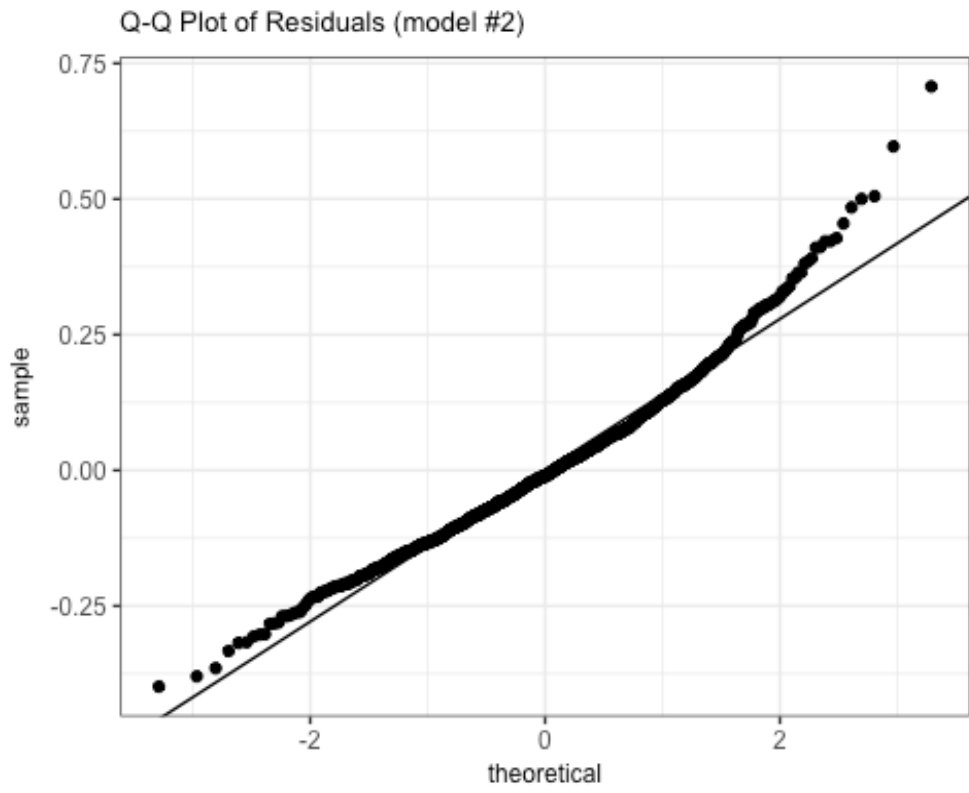
```
options(scipen=1000)
shapiro.test(complete$residuals)

##
##   Shapiro-Wilk normality test
##
## data:  complete$residuals
## W = 0.9775, p-value = 0.000000000028
```

Interpretation – Here too, we hope to see a scatter on the 45 degree line. Not bad!

Nature ———————— Population/ ———————— Observation/ ———————— Relationships/ ———————— Analysis/
                              Sample   Data   Modeling   Synthesis

Ramsay Test of Omitted Variables. Null: No omissions ("all is well")

```
library(lmtest)
lmtest::resettest(m_best,power=2,type="regressor")
##  RESET test
##
## data:  m_best
## RESET = 0.42467, df1 = 5, df2 = 983, p-value = 0.8317
```

Interpretation – Ramsey test is NOT significant (p=.83) suggesting we're okay!

Assessment of Multicollinearity ("all is well" if VIF < 10)

```
library(car)
car::vif(m_best)

##     ln_bmi    ln_scl        age      female  ageXfemale
##   1.115511  1.175531   2.378150  32.394888   34.116761
```

Interpretation – female and ageXfemale appear to be collinear suggesting some concern about the extent to which there is adequacy of range of age in the 2 genders.

Cook's Distances (flag observations for which Cook distance > 4/(n-p-1). Other definitions possible.

```
library(Hmisc)
library(ggplot2)
complete$ID <- as.numeric(row.names(complete))          # Create ID variable and add it to the dataset
Hmisc::label(complete$ID) <- "Observation Number"

complete$cooks <- cooks.distance(m_best)                        # Add cooks distances to the dataset

cutoff <- 4/((nrow(complete)-length(m_best$coefficients)-2)) # Solve for cutoff  as equal to = 4 / (n-p-1).

p1 <- ggplot(data=complete, aes(x=ID, y=cooks)) +
     geom_bar(stat="identity", position="identity") +
     xlab("Observation Number") +
     ylab("Cooks Distance") +
     geom_hline(yintercept=cutoff) +
     geom_text(aes(label=ifelse((cooks>cutoff), ID, "")),vjust=-0.2, hjust=0.5) +
     ggtitle("Cooks Distances > 4 / (n-p-1)") +
     theme_bw()
p1
```


Nature ——————— Population/ ———————Observation/ ——————— Relationships/ ——————— Analysis/
                    Sample   Data   Modeling   Synthesis