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 investigate 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*.

Table of Contents

Datasets used (download from course website)	**Packages used** (one time installation)
janka.Rdata	ggplot2 Hmisc
p53paper.Rdata	stargazer car
framingham_1000.Rdata	gridExtra lmtest
	GGally summarytools
	Tip! **Don't forget that R is case sensitive …**

1. Learning Objectives

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

- Explain the concepts of association, causation, confounding, mediation, and effect modification;
- Construct and interpret a scatter plot with respect to: evidence of association, assessment of linearity, and the presence of outlying values;
- State the multiple predictor linear regression model and the assumptions necessary for its use;
- Perform and interpret the Shapiro-Wilk and Kolmogorov-Smirnov tests of normality;
- Explain the relevance of the normal probability distribution;
- Explain and interpret the coefficients (and standard error) and analysis of variance tables outputs of a single or multiple predictor regression model estimation;.
- Explain and compare crude *versus* adjusted estimates (betas) of association;
- Explain and interpret regression model estimates of effect modification (interaction);
- Explain and interpret overall and adjusted R-squared measures of association;
- Explain and interpret overall and partial F-tests;
- Draft an analysis plan for a multiple predictor regression model analysis; and
- Explain and interpret selected regression model diagnostics: residuals, leverage, and Cook's distance.

1. Review

Before you begin - review your prior introduction to regression and correlation, in particular simple linear regression. In your previous biostatistics or statistics course, you were likely introduced to simple linear regression. There are many excellent introductions to simple linear regression. And you may already have a favorite! Alternatively, you might consider the Fall 2022 the UMass/Amherst BIOSTATS 540 (Introductory Biostatistics) course introduction here: BIOSTATS 540, Unit 12 - Simple Linear Regression.

a. Settings Where Regression Might Be Considered

Example #1
Is the density of wood a predictor of hardness of timber?

Source:
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 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?

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* selected 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_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 other variables that 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)

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)

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 of models notwithstanding, we do statistical modeling for very good reasons. Among them is 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 *(more on this later)*.

Good Fit

- Variance explained. The variability in the outcomes (the Y variable) explained is a lot
- Prediction. The outcomes predicted by the model are close to the observed outcomes.

The model should also make sense

- Biological sense. A preferred model is one based on "subject matter" considerations
- Useful. The preferred predictors are simple, measurable and convenient.

Sigh.

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.

c. A General Strategy for Model Selection

There are ***no*** rules ***nor a single best strategy.*** Different study designs and research questions call for different strategies for building a regression model. Essential. Before you begin your model development, make a list of your study design, research aims, outcome variable (Y), primary predictors(X_1, X_2, ..., X_p), and covariates.

As a general strategy (and by no means the only one!), the following approach has the advantages of providing a reasonably thorough exploration of the data and a 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 the variability in the outcome that is explained. Note - we really hope that our final model is not as complicated as this!

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 essential considerations (this is not "backward elimination")

- Prioritorize considerations of biology and what makes sense. In particular,
- 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!

d. Review - Normal Theory Regression

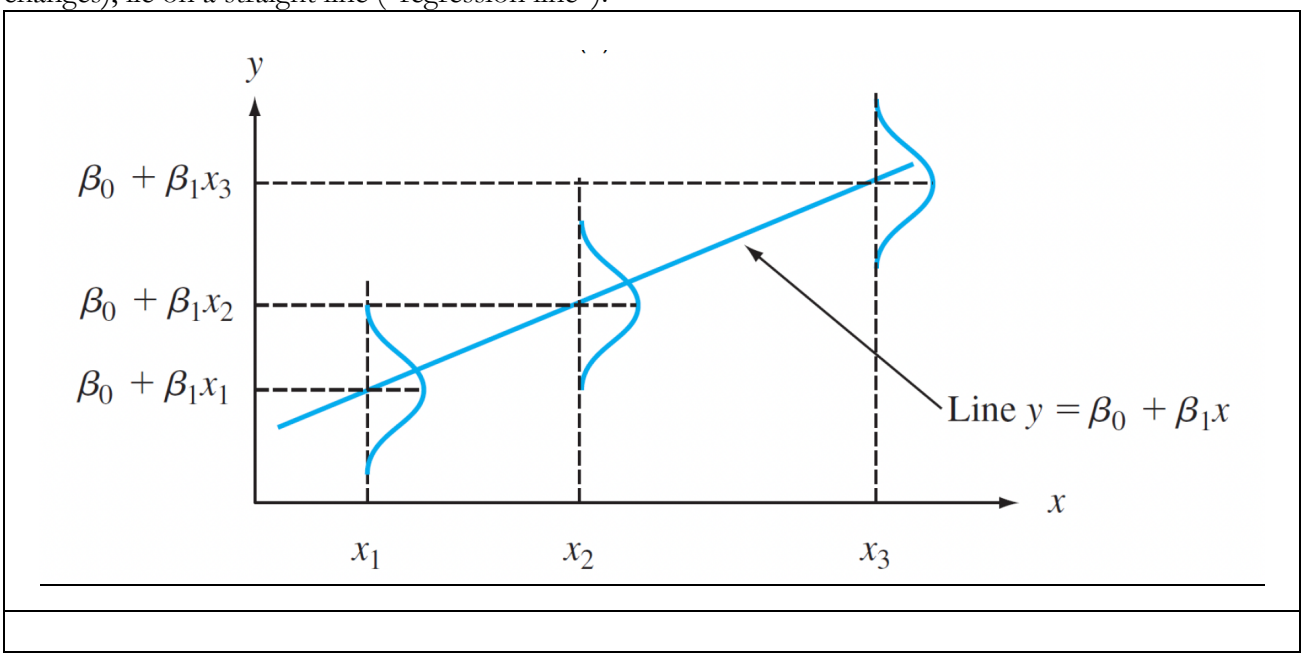
When can I perform Normal theory regresstion? Normal theory regression analysis can be used used to model/investigate possibly complex relationships when:

- The outcome is a **single continuous variable (Y)** that is assumed to be **distributed normal**; *and*
- The outcome is potentially related to possibly **several predictors (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:

We're modeling the means of several (assumed Normally distributed) subpopulations, each defined by a particular X=x. 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**).

Here is a picture. It says that the subpopulation means $\mu_{Y|X=x}$(the expected values of the outcome Y, as X=x changes), lie on a straight line ("regression line").



- At each value of X (e.g., x_1, x_2, and x_3) the corresponding outcomes Y are modeled as random draws from a Normal distribution with mean = $\mu_{X=x} = \beta_0 + \beta_1 \cdot x$
- The means of these Normal distributions lie on the simple linear regression line; and
- The variances of these Normal distributions, $\sigma^2_{Y|x}$ are assumed to be the same (notice similar "bells"!)

Source: *Devore, J.L. (2011) Probability and Statistics: For Engineering and the Sciences. 8th Edition, Cengage Learning, Boston.*

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 for the subpopulation with X=x is modeled as distributed **Normal** with

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

Model

A linear model says "Observed = Model + Error." 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.

Linear Model

Observed = [modeled] + [error]

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

Y_i	$[\beta_0 + \beta_1 x_i]$	ε_i
Observed Ourcome Y for ith person	**Modeled mean $\mu_{Y\|X=x_i}$ for subpopulation with X = x_i**	**Error in observation of mean**

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

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

How to estimate β_0, β_1: "Least Squares", "Close" and Least Squares Estimation

It's possible to draw lots of lines through an X-Y scatter of points! So, which one should we choose? "Least squares" estimation is one approach to choosing a line that is "closest" to the data. Least squares estimation says choose the values for $\hat{\beta}_0$ and $\hat{\beta}_1$ that, upon insertion, minimizes the sum total of the squared vertical differences, d_i^2

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

The sum total, $\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n}\left(Y_i - \left[\hat{\beta}_0 + \hat{\beta}_1 X_i\right]\right)^2$ **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	$\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}$
Intercept $\hat{\beta}_0$ **or** b_0	$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

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**	$SSR = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2$	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)**	$SST = \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2$	

2.
R Illustration: Fit a Simple Linear Regression Model

Preliminary - Set working directory (user edits)

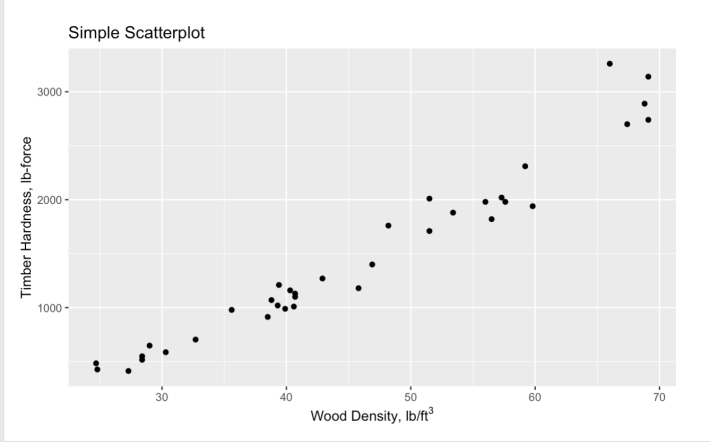
```
setwd("/Users/cbigelow/Desktop/")                # setwd( )to set working directory = folder to read from and write to

Input R dataset janka.Rdata.  Inspect.
library(tidyverse)                               # glimpse() in package {tidyverse}
load(file="janka.Rdata")                         # Assumes the data are in the working directory
janka$hardness <- as.numeric(janka$hardness)
glimpse(janka)                                   # glimpse( ) to view dataset structure. Could also do str( ) in {base}

## 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                                            # Recommended:  Look at the data!

##    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)
ggplot(data=janka) +                             # required layer:  data = to specify dataset
    aes(x=density,y=hardness) +                  # required layer:  aes( ) to define x- and y-axis
    geom_point()  +                              # required layer:  geom_point( ) to produce XY scatterplot
    xlab(expression("Wood Density, lb/ft"^{3}))  + # optional:  label the x-axis
    ylab("Timber Hardness, lb-force") +          # optional:  label the y-axis
    ggtitle("Simple Scatterplot")                # optional:  provide a title
```



looks linear with no influential observations

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 ***    Intercept = β̂0 = b0 = −1160.500
## density        57.507      2.279   25.24 < 0.0000000000000002 ***       Slope = β̂1 = b1 = 57.507
## ---
## 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       R2 = % Variance explained = .9493 or ~ 95%
## F-statistic:   637 on 1 and 34 DF,  p-value: < 0.00000000000000022

The fitted line is thus:    Predicted hardness = hardness^ = −1160.500 + 57.507 · 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 ***      SSQ(model) = SSR = 21,345,674
## Residuals 34  1139366    33511                                       SSQ(residual) = SSE = 1,139,366
## ---                                                                   Overall F-test of null: slope=0
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1       F = 636.98 with df = 1, 34
                                                                         p-value <<< .0001 REJECT null
                                                                         Conclude fitted line is significant
```

3. Multivariable Linear Regression

a. Introduction

In multiple linear regression, the number of explanatory (predictor) variables is more than 1.

- There is just one outcome, Y; and
- The predictors can be several: $X_1, X_2, ..., X_p$; and
- $X_1, X_2, ..., X_p$ can be a mixture of continuous and discrete (hooray); however,
- Care needs to be taken in the modeling of discrete predictors. Stay tuned.

Definition
Normal theory regression models **means** of Normal distributions.

In simple linear regression, we modeled the outcomes Y as random draws from several Normal distributions for which the means lie on the line: $\mu_{Y|X=x} = \beta_0 + \beta_1 \cdot x$.

In multiple linear regresssion, we model the outcomes Y as random draws from several Normal distributions for which the means lie on linear regression plane: $\mu_{Y \mid X_1 X_2 \ldots X_p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$

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 number of pregnancies and age at first pregnancy.

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

This is a multivariable linear model with number of predictors p = 3:

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

The General Multivariable Linear Model $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \text{error}$

- 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 the subscripts 1, 2, …, p are indexing the predictors
- For the i^{th} 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 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 vector value of the predictor variable $\underline{X}=\underline{x}$, the distribution of values of Y is modeled as distributed **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, $x_1, x_2, \ldots, x_p$, of the p predictor variables $X_1 \cdots X_p$ (written using vector notation $\underline{X}=\underline{x}$), the distribution of values of Y modeled as distributed **normal** with

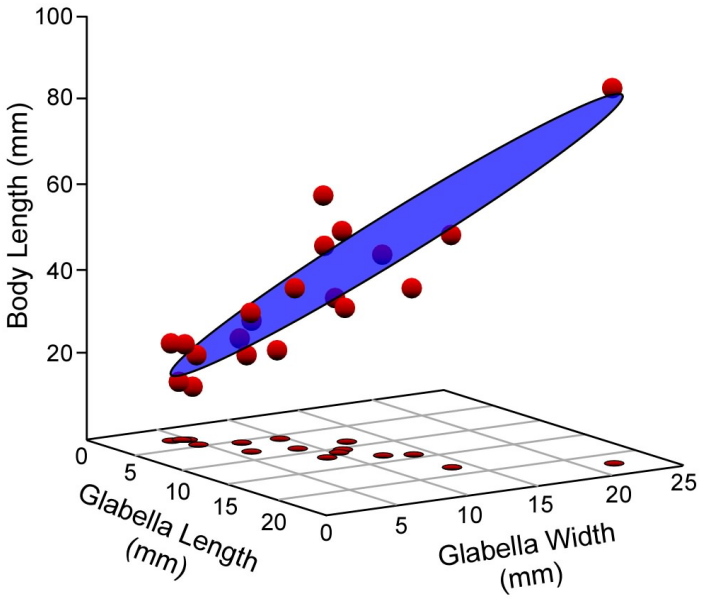
$$\text{mean} = \mu_{Y|\underline{X}=\underline{x}} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

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

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 (sort of anyway) to show a schematic of the fitted plane that results from least squares estimation.

Consider 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 analogous to 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 do we need indicator variables?
Indicator variables are very useful when we want to include discrete variables as predictors in our model.

Example - Suppose you want to model some outcome (Y = duration of stay in ICU, in days) in relationship to a nominal predictor, type of surgery X. X might be lazily stored in the data using "1", "2", and "3" as placeholders for the names of the type of surgery; e.g., 1=medical therapy, 2=angioplasty, and 3=coronary bypass surgery).

The following would be incorrect. Not really appreciating that "1", "2", and "3" are your lazy placeholders/names and not actually bona fide numbers, you might just forge on and fit a simple linear model.

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

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

β_1 = Δ 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.

Indicator Variables are Variables that are coded 0 or 1.
Indicator variables are commonly used as predictors in multivariable regression models. We let

1 = value of indicator when "trait" is present
0 = 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"
  (estimated change in Y associated with unit change in trait: from "0=absent" to "1=present")

Design variables. One separator distinguishes 2 groups, 2 separators distinguish 3 groups, and so on. Thus, **(k-1) separators are needed to distinguish k groups**. This is the idea of design variables

Consider sex at birth, which is nominal discrete. For illustration purposes only (because, in reality, more than 2 outcomes are possible), assume there are just 2 nominal levels: male sex at birth and female sex at birth. To distinguish these 2 groups, just one separator is required. This will be one indicator/dummy variable. By convention, we assign the value=0 to the level that we think of as the referent. The value=1 is assigned to the comparison group. Thus, and because I like self explanatory variable names, I might create the following:

```
female  =   missing if "sex at birth" = missing
                0  if "sex at birth" = "male"; and
                1 if "sex at birth" = female.
```

IMPORTANT!!!
Always handle missing values explicity.

Consider pain level which is ordinal discrete and having 3 levels, "low", "medium", and "high". To distinguish 3 levels, we need 2 0/1 indicator variables. Again the convention is to assign the value 0 to the level we think of as the referent. Thus, I might create the following:

```
pain_medium  =   missing if "pain level" = missing
                          0  if "pain level" = "low" OR "high"; and
                          1 if "pain_level" = "medium"


pain_high  =   missing if "pain level" = missing
                          0  if "pain level" = "low" OR "medium"; and
                          1 if "pain_level" = "high"
```

The resulting set of 2 0/1 indicators (`pain_medium` and `pain_high`) which collectively distinguish all 3 levels of pain are called **design variables**.

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 if treatment is medical therapy
2 if treatment is angioplasty
3 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.

```
tr_ang  =   missing if "treatment" = missing  (unlikely but code this anyway!)
            0  if "treatment" = "medical therapy OR "bypass surgery"; and
            1  if "treatment" = "angioplasty"

tr_sur  =   missing if "treatment" = missing
            0  if "treatment" = "medical therapy OR "angioplasty"; and
            1  if "treatment" = "surgery"
```

Check that all k=3 levels are reprsented by a set of (k-1) = 2 design variables comprised. The reference category is medical therapy.

Value of original X = Type of Surgery	Value of 0/1 Indicator tr_ang	Value of 0/1 Indicator tr_sur
X="1" for "medical", the "referent"	0	0
X="2" for "angioplasty"	1	0
X="3" for "surgery"	0	1

Guidelines for the Definition of Indicator and Design Variables

1) K levels of the nominal predictors requires (K-1) separators to distinguish.

   Create (K-1) 0/1 indicator variables.

2) Take care in choosing 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

3) In general (this is not hard and fast), treat the (k-1) design variables as a set. This means that you.

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

c. Interaction Variables

Previously, we've talked about "effect modification" (in the lab sciences, this might be called "synergism"). It refers to the phenomenon that 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. In regression, we call this **interaction.**

How to create a predictor that will model the interaction of a continuous predictor X and a 0/1 predictor Z. The solution is straightfoward. Use the product of X and Z. Here, I've named this new variable XZ.

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)

Your turn. What is the model of Y for non-vertebral fractures patients (Z=0)?
For the non-vertebral fractures patients, setting Z=0 into the expression for the full model yields

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

Your turn. What is the model of Y for vertebral fractures patients (Z=1)?
For the vertebral fractures patients, setting Z=1 into the expression for the full model yields

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

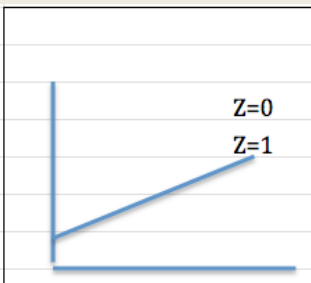
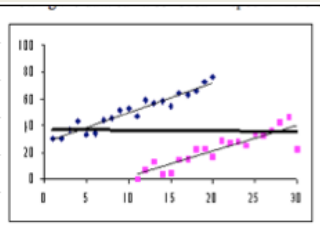
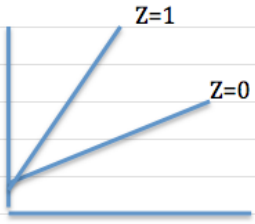
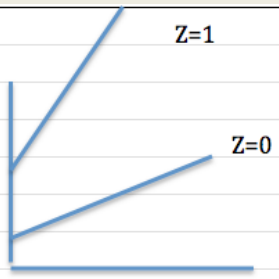
d. *Look!* Schematic of Confounding and Effect Modification

The use of indicator variables and interaction variables are helpful (but not without important caveats) in assessing 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=1**: $Y = \beta_0 + \beta_1 X$	**Comorbidities=0:** $Y = \beta_0 + \beta_1 X$ **Comorbidities=1**: $Y = (\beta_0 + \beta_2) + \beta_1 X$	**Comorbidities=0:** $Y = \beta_0 + \beta_1 X$ **Comorbidities=1**: $Y = \beta_0 + (\beta_1 + \beta_2) X$	**Comorbidities=0:** $Y = \beta_0 + \beta_1 X$ **Comorbidities=1**: $Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X$
	β_2 = **change** in intercept associated with presence of any comorbidities	β_2 = **change** in slope of Y on X associated with presence of any comorbidities	β_2 = **change** in intercept β_3 = **change** in slope of Y on X

e. The Analysis of Variance Table

Big idea. The total variability in the outcome (whole pie) is partitioned into two component sources (1st = slice taken and 2nd = remainder of the pie):

1. **SST (whole pie):** "Total" or "total, corrected"
   - SST = $\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2$ is the variability of Y about $\overline{Y}$
   - Degrees of freedom = df = (n-1).

2. **SSR (pie slice taken):** "Regression" or "due model"
   - SSR = $\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2$ is the variability of $\hat{Y}$ about $\overline{Y}$
   - Degrees of freedom = df = p = # predictors ***apart*** from intercept

3. **SSE (remainder of pie):** "Residual" or "due error" refers to the
   - SSE = $\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$ is the variability of Y about $\hat{Y}$
   - Degrees of freedom = df = (n-1) – (p)

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

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.

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

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

Eureka!!! When the null is true, the best model is "intercept only"

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

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 (Sigh - we don't know which ones) has a slope that is statistically significantly different from zero.

Example - Earlier age at first full term pregnancy is thought to confer a reduction in risk for breast cancer later in life. The mechanism for this is not clear, but it is thought be related to expression of the hormone p53. Consider a multiple linear regression analysis of the relationship of outcome Y=p53 expression to number of pregnancies and two 0/1 indicator variables 1st pregnancy at age $\leq$ 24 (early), and 1st pregnancy at age > 24 (late). Women who are nulliparous (never delivered a live fetus) comprise the referent group.

Example - R

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:  p53^ = 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.  Upon inspection of the estimates, their standard errors, their t-values, what do you think?
```

```
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
```

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

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

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

This matches "F-Statistic" p 23

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).

f. The Partial F Test

When to use. The partial F test is used when we want to know if our current model should be enhanced to include additional predictors. The two models being compared are called **"hierarchical"** because they both contain all the predictors in the smaller model, with the larger model containing the extra predictors of interest. One convention in the naming of hierarchical models is to call the larger model "full" and the smaller model "reduced".

How does the Partial F Test work. The Partial F Test tests the null hypothesis that says "controlling for the variables in the reduced model, the 'extra' predictors explain no additional variability in outcome (all their regression coefficients are 0)". Thus, the partial F test is used to assess if the "extra" predictors are statistically significant, "above and beyond" the control variables.

Some more details of hierarchical models.

- **"Hierarchical"** means we are comparing just 2 models, and one model is an enhancement of the other.
- **An analysis of hierarchical models** requires that all of the predictors in the smaller (reduced, reference) are contained in the larger (comparison) model.
- **The smaller and larger models have various names**: The smaller model might be termed "reduced", "reference", "smaller". The larger (enhanced) model might be termed "full", "comparison", "larger"

Example. 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** }

"Hierarchical" 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.

In a partial F test, we are assessing the nature and significance of the extra predictors, (**early** and **late**) for the prediction of Y=p53, ***adjusting for (controlling for)*** all of the variables in the smaller model (**pregnum**).

Tip. Identify the question that is being asked in the comparison of the hierarchical models you are performing In this example, the Partial F test is addressing the following question:

What is the statistical significance of **early** and **late** for the prediction of **Y = p53**, *after controlling for* the association of Y=p53 with the control variable **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?

Partial F Test

Ho: 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:

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

HA: Not

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

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

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

Ho true: The extra predictors are not significant in adjusted analysis	F statistic = small (close to 1) p-value = large
Ho false: The extra predictors are significant in adjusted analysis	F statistic = large (bigger than 1) p-value = small

Example - R, continued.

```
reduced <- lm(data=p53paper, p53 ~ pregnum)
full <- lm(data=p53paper, p53 ~ pregnum + early + late)
anova(reduced, full)
```

HO: Controlling for pregnum, the additional predictors have $\beta_{EARLY} = 0$ and $\beta_{LATE} = 0$
HA: At least one extra predictor 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 → $F_{partial} = .3052$ **p-value = .7361**

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

g. Multiple Partial Correlation

Beware. Partial F test ≠ partial correlation

- The partial F test is a hypothesis test; whereas.
- A partial correlation is a statistic, measuring the what is explained (and expressed as a percent if squared)

Partial correlation. ***"To what extent is Y correlated with X (or multiple X), after accounting for some control variable Z (or multiple control variables Z)?***

In a partial correlation, we are removing the influence of the control variable (Z). A partial correlation is the correlation of (residuals of Y on Z) with the (residuals of X on Z). To appreciate what this means, consider:

- **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 partial correlation of Y on X controlling for Z as the correlation between these two sets of residuals: (residuals of Y on Z) and (residuals of X on Z) give you a Z-controlled assessment of the relationship between X and Y, that is, *independent of Z*.

Partial Correlation

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)

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

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

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

$F_{partial}$ =Partial F Test	**$R^2_{partial}$ = Partial Multiple Correlation Squared**
Goal is a hypothesis test - Hypothesis test of significance of extra variables, after adjustment for the control variables.	Goal is estimation - Estimation of percent of variability in outcome Y that is explained by the extra variables, independent of the control variables.
Control variables: $X_1 \ldots X_p$ Extra variables: $X_{p+1} \ldots X_{p+k}$	Control variables: $X_1 \ldots X_p$ Extra variables: $X_{p+1} \ldots X_{p+k}$
$F_{PARTIAL}$ 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
$= \dfrac{[SSR(X_1 \ldots X_p, X_{p+1}, \ldots X_{p+k}) - SSR(X_1 \ldots X_p)]/[(p+k)-p]}{[SSE(X_1 \ldots X_p, X_{p+1}, \ldots X_{p+k})]/[(n-1)-(p+k)]}$	$= \dfrac{SSR(\text{due Model with all}) - SSR(\text{due Model control only})}{SSE(\text{due residual in Z only model})}$

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.

Recall. Some general guidelines **(Again and very important. There is no single right answer)**

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

Data set used
`framingham_1000.Rdata`

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, 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?

R

Input Data. Check. Produce descriptives:

```
# User edits

rm(list=ls())                                                    # Clear the environment (workspace)
setwd("/cloud/project/")                                         # Tell R where to "read from" and "write to"


load(file="framingham_1000.Rdata")

framingham <- framingham_1000                                    # I'm lazy.  So, I'm creating a shorter name

summary(framingham)                                              # Inspect distributions of all study variables

##     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)       # Another way to inspect distributions (more succinct).

##
## ===========================================================
## 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.

```
library(summarytools)                                                    # freq() is in package {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)                                                    # descr() is in package {summarytools}
summarytools::descr(framingham$sbp, stats = c("n.valid","mean", "sd", "min","q1", "med", "q3", "max","CV"),
transpose = TRUE)                                                        # option stats=c( )to choose statistics to show

## 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
```

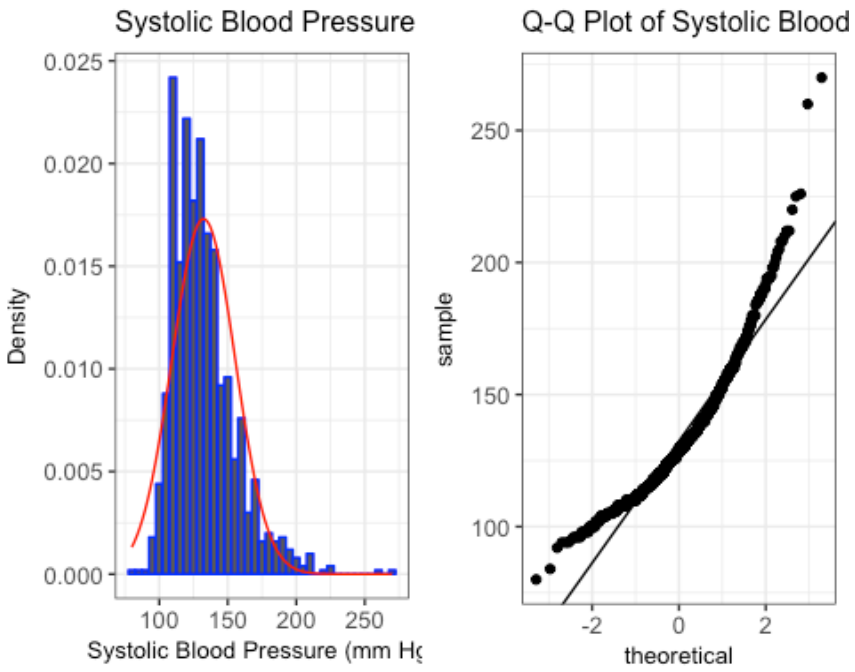
Assess Normality of Candidate Dependent Variable = sbp. Shapiro-Wilk Test (Null: normality)
Histogram w Overlay Normal and QQ Plot

```
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). Boo hoo.

```
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)         # grid.arrange( ) in package {gridExtra} to lay out panels in figure
```



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

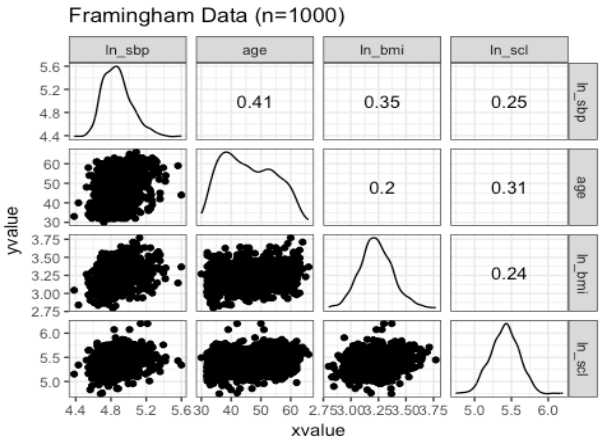
Create "regression-friendly" indicator variables and interactions. Check.

```
library(summarytools)
library(Hmisc)


# Create 0/1 indicator/dummy variable using logical operator:
# If sex="Women" is TRUE, code new variable female=1.  Otherwise, code new variable female=0
# option na.rm=TRUE ensures that missing values will not be considered and instead will be retained as missing.


framingham$female <- as.numeric(framingham$sex == "Women", na.rm=TRUE)
summarytools::ctable(framingham$sex,framingham$female,prop = 'n', totals = FALSE)              # xtab check

## 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!
```

```
Hmisc::label(framingham$female) <- "female01"                          # label( ) is in package {Hmisc} to label variables

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



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)

# na.omit( ) to omit observations with anything missing; the resulting object named complete contains complete data only
# cols=c("var1", "var2", etc) to specify variables to keep
complete <- na.omit(framingham, cols=c("ln_sbp", "ln_bmi", "age", "female", "lnbmiXfemale", "lnsclXfemale","ageXfemale"))

# Fit each model of interest to the SAME dataset comprised of complete data only
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)
```

```
# stargazer( ) in package {stargazer} for nice display of 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  Models 1 & 2 have nearly identical R^2 = % variance
Adjusted R2   0.261     0.262    0.122    0.063     0.200  explained (26.7%, 26.6%).  This suggests the extra predictors
                                                           in model 1 are not needed. -> Model 2 is preferred (simpler!)
============================================================
Note:                           *p<0.1; **p<0.05; ***p<0.01
```

```
# anova(reduced,full) to obtain Partial F Test
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*).

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.

Important. Don't forget to consider what makes sense biologically!

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.

Again, important. Don't forget to consider what makes sense biologically!

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.

In addition to identifying the variables, for each variable, take care to note its hypothesized role **(important!)**.

- 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. There is a lot to look for, including:

- Completeness
- Occurrence of logical errors
- Within form consistency
- Between form consistency
- Range

#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. Assess 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. Fit Single Predictor Regression Models

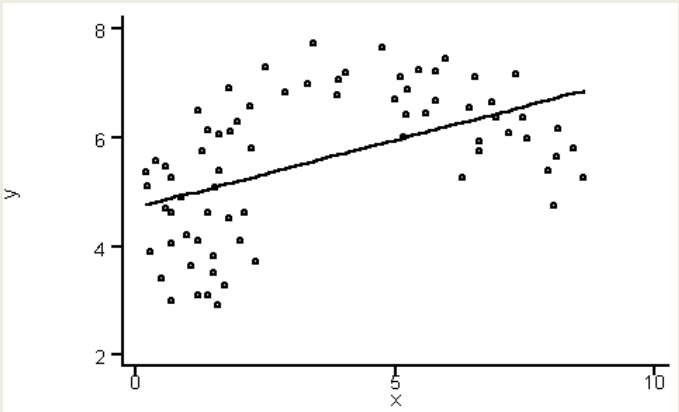
The fit of these models identifies the nature and magnitude of crude associations. It also permits assessment of the appropriateness of the assumed functional form of the predictor-outcome relationship.

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

5. Regression Diagnostics

a. Rationale and Terminology

Rationale. Neither prediction nor estimation have meaning when the estimated model is a poor fit to the data:



There are lots to see in this picture.

- A better fitting relationship between X and Y is quadratic;
- We notice different sizes of discrepancies; in particular:
- While some observed Y are close to the fitted line $\hat{Y}$ (e.g. near X=1 or X=8),
- Other observed Y are very far from the fitted line $\hat{Y}$ (e.g. near X=5)

A 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 *(more on this later)* was fit;
- Extreme values (outliers) exhibit uniquely large discrepancies between observed and fitted value;.
- One or more important explanatory variables have been omitted; and/or
- One or more model assumptions have been violated.

Consequences of a poor fit include:

- You might conclude the wrong biology.
- With a poor fit, comparison of group differences are not "fair" because they are unduly influenced by a minority.
- With a poor fit, comparison of group means will be based on incorrect estimates of standard error.
- With a poor fit, predictions will be wrong in at least some instances, as when the poor fit does not apply to the case of interest.

Rationale. Neither prediction nor estimation has meaning when model assumptions are violated.

Sigh. Before detailing the model assumptions, some terminology is needed. Sorry about that.

Terminology. Available techniques of regression diagnostics and goodness-of-fit assessment are of two types, systematic and case analysis

1. **Systematic.** These techniques assess the reasonableness of the model itself.

   Have we fit the correct functional form model (linear v polynomial or other, etc.)?

   Have any important predictors been omitted?

   Are some of the predictors in the model unnecessary?

   Can the errors be (reasonably assumed independent, of constant variance, and distributed Normal with mean=0?

2. **Case Analysis.** These techniques investigate the influence of individual observations

   Are there a small number of individuals whose inclusion in the analysis exert TOO MUCH influence on the choice of the final model?

Systematic versus case analysis

Consider again the assumed normal theory multiple predictor regression model.

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

Observed = systematic + error

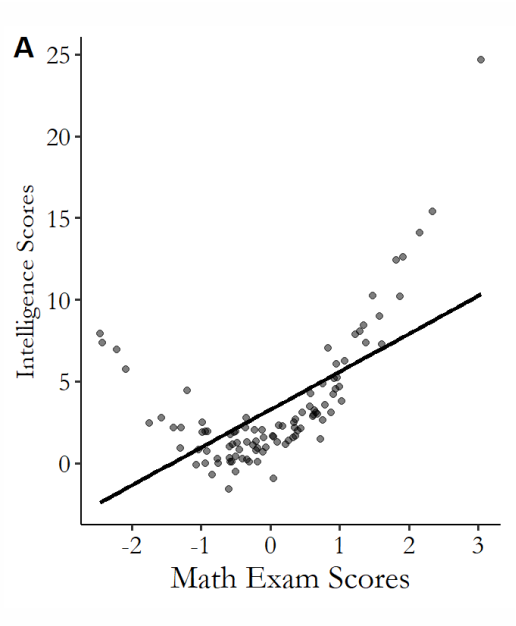
This is the mean of Y at $X_1, X_2, \ldots, X_p$ = E[Y at $\underline{X}$] =

Terminology - systematic

Link: **Have we fit the right functional form (be it a line or a quadratic, etc.)?**	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 $\mu_{Y\|\underline{X}} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$ This is called the **natural** or **identity** link. **Example**: When μ is a proportion, we might model $\ln [\mu_{Y\|\underline{X}} / (1-\mu_{Y\|\underline{X}})] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$. This is called the **logit** link.
Normality: **Is the assumption of normality of errors reasonably satisfied?**	In the linear model regression analysis, we assume that the errors are independent and follow a Normal$(0, \sigma^2_{Y\|\underline{X}})$ distribution. Recall: The unobservable true errors ε are estimated by the residuals e.
Heteroscedasticity: **Is the assumption of a constant variance of errors reasonably satisfied?**	We also assume that the errors have constant variance. If this assumption is not true, we say there is heteroscedasticity of errors, or non-homogeneity (non constancy) of errors.

A Feel for Systematic

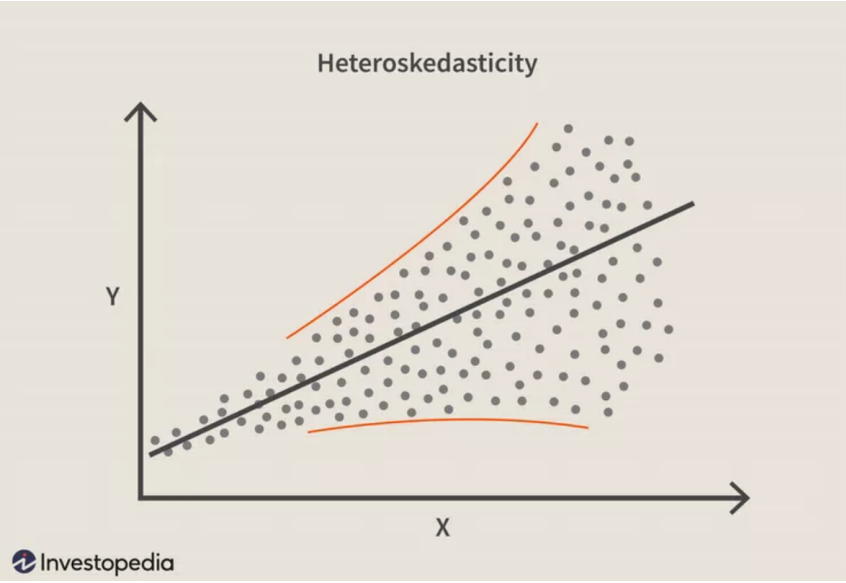
The correct functional form may not be linear



(source: https://theeffectbook.net/ch-StatisticalAdjustment.html)

A Feel for Systematic

The errors do not have constant variance



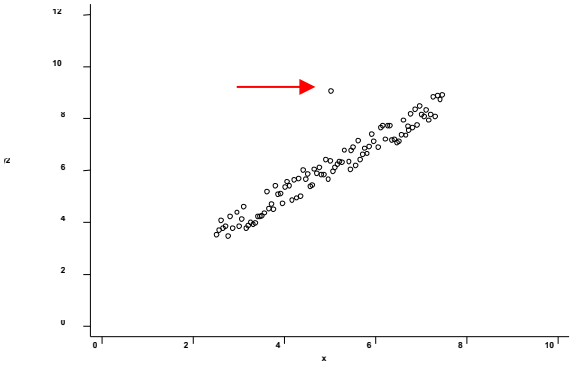
(source: https://www.investopedia.com/terms/h/heteroskedasticity.asp)

Terminology - 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 true error ε.
Outlier: **Unusualness in Y direction**	An **outlier** is a residual that is unusually 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: **Unusualness in X direction**	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:	An observation with **influence** changes the fitted line, depending on whether or not they are included in the estimation of the model fit. **Example:** One measure (there are others) of influence is **Cook's Distance**

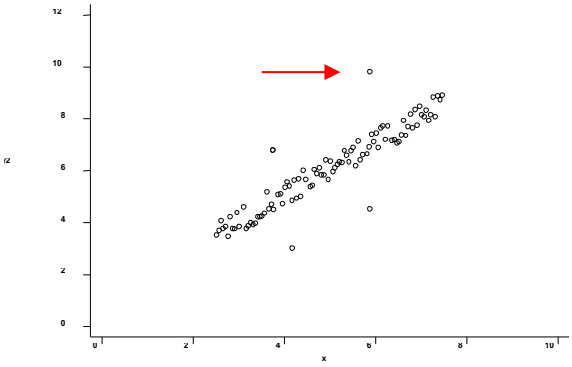
A Feel for Case Analysis

Large residuals may or may not be influential



Large residual (Y direction)
Low leverage (X direction)

The large residual effects a large influence.

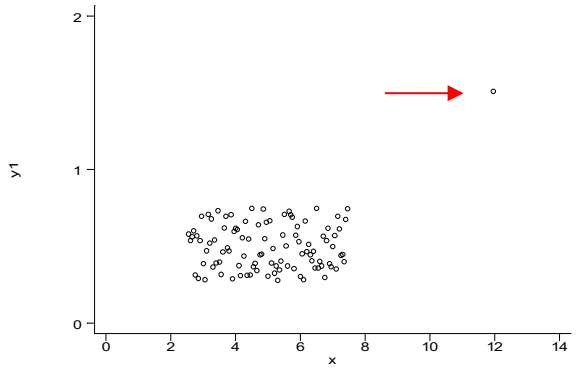


Large residual (Y direction)
Low leverage (X direction)

Despite its size, the large residual effects only small influence.

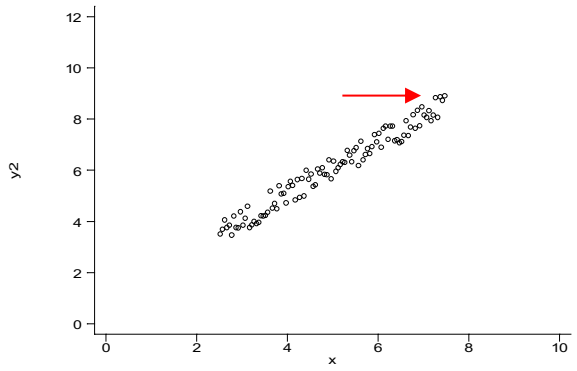
A Feel for Case Analysis

High leverage may or may not be influential



High leverage (Y direction)
Small residual (X direction)

The high leverage effects a large influence.



High leverage (X direction)
Small residual (Y direction)

Despite its size, the large leverage effects only small influence.

Case analysis is needed to discover all of:

- high leverage;
- large residuals; and
- large influence

b. Assumptions of Normal Theory Multiple Linear Regression

It's useful to consider the implications of each assumption. Depending, this helps us in model building. Plus, it helps us in developing diagnostic assessments.

__1. Systematic. *The functional form of the true model is linear in the predictors and error is additive.*

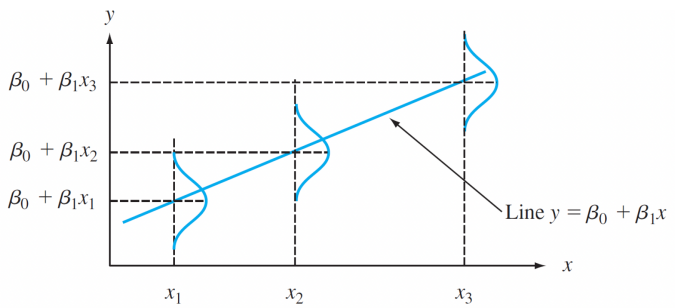
$$Y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \ldots + \beta_p \cdot x_p + \varepsilon$$

- **Nothing fancier is needed.** This assumption says that linearity in X_1, X_2, ..., X_p is the true functional form (even allowing for predictors such as $X_2 = X_1 * X_1$); **in reality, this is unlikely to be true.**
- **Ideally, we want our final model to be NEITHER overfit NOR underfit.** An overfit model is overly complex (making it hard to interpret), and too close to the observed data (making it not very generalizable). An underfit model too simple (we might miss discovering important biology).
- **This assumption is also saying** that there are no missing/omitted variables
- **This assumption is also saying** that ther are no extraneous/unnecessary variables
- **Implication.** Variability in the predictors that belong in the model (but are not included) will be incorporated into (become part of) the error term.
- **It is also assumed** that the excluded predictors are independent of the predictors included in the model.

__2. Systematic. *The errors are independent and distributed Normal (mean=0, variance=constant)*

- **Independence.**
  The errors ε are independent (Note - this is the same as saying the observations are independent)
- **Normality.**
  The errors ε are distributed Normal $(0, \sigma^2_{Y \mid X1,\ X2,\ \ldots,\ Xp})$
- **Zero expected value.**
  The probability distribution of the error terms has expected value $E[\varepsilon] = 0$
- **Homogeneity of Error Variance.**
  The collection of all sub-populations of Y at each set of profiles of values of the predictors $X_1=x_1$, $X_2=x_2$, ..., $X_p = x_p$, are called **conditional distributions of Y**. These are assumed to have constant variance which is denoted $\sigma^2_{Y \mid X1,\ X2,\ \ldots,\ Xp} = constant$

Consider again the picture from page 10 (duplicated here). We can say a bit more.



Source: Devore, J.L. (2011) Probability and Statistics: For Engineering and the Sciences. 8th Edition, Cengage Learning, Boston

Conditional distributions of Y. At each value of X (e.g., x_1, x_2, and x_3) the corresponding outcomes Y are modeled as random draws from a Normal distribution with mean = $\mu_{Y \mid X=x} = \beta_0 + \beta_1 \cdot x$

Normal theory regression models means. The means of these Normal distributions lie on the linear regression line (or plane); and

Assumption of Constant Variance. A consequence of the variances of these conditional Normal distributions, $\sigma^2_{Y \mid x}$ being assumed to be constant means the following:

We can also (reasonably, that is) assume that the observable errors $e = [y - \hat{y}]$ are distributed Normal $(0, \sigma^2_{Y \mid x})$ with the same constant variance.

__3. **Systematic.** ***The predictors (also called covariates) are linearly independent and fixed.***

- **The predictors are linearly independent.**
  When any of the predictors are NOT linearly independent of each other, this is **multicollinearity**. In the presence of multicollinearity, the fitted model is very unstable (if it can be estimated at all), resulting in large estimated standard errors of the regression coefficients (betas).

- **The values of the predictors are treated as fixed.**
  This assumption is saying that the predictor variables are measured without error. Note - in reality, there is likely to be some measurement error.

__4. **Case analysis.** ***No influence. There are no observations whose inclusion changes substantially the fitted regression model.***

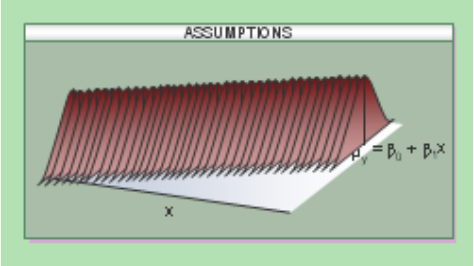
- **This assumption is saying** that the estimated regression coefficients (betas) are not affected appreciably by whether or not the observation is included in model fitting.

c. At a Glance: Regression Diagnostics (partial listing!)

Normal Theory Multiple Predictor Linear Regression

Assumption	Why is this important	Regression Diagnostics
Systematic **The errors ε are distributed Normal**	The validity of hypothesis tests and confidence intervals depend on the assumption of normality of errors.	(1) Shapiro-Wilk test of normality Null, H_O: error ~ Normal; or Kolmogorov-Smirnov test of normality Null, H_O: error ~ Normal (2) Visualization: histogram of residuals, e
Systematic **The errors ε have constant variance**	The validity of hypothesis tests and confidence intervals and, in particular the estimated standard errors, depend on the assumption of constant variance of errors.	(1) Cook-Weisberg test of constant variance. Null, H_O: constant variance (2) Visualization: Scatterplot of X=residual versus Y = predicted
Systematic **The true model functional form is linear in the predictors**	The final model should be as simple as possible, but not underfit. Thus, the question becomes: Is the choice of functional form relating the predictors to outcome a "good" one?	(1) Method of Fractional polynomials for the assessment of functional form; and (2) Visualization: Lowess smoothing
Systematic **The model is correctly specified**	Questions. Is the model correctly specified? Have we failed to include any important explanatory (predictor) variables? We might not know what the correct specification is nor what the omitted variables are. But it would be nice get an alert!	(1) Ramsey Test of Model Misspecification Null, H_O: current model is adequate
Systematic **The predictors are linearly independent (No multicollinearity)**	This also relates, in one respect, to the assumption that there are no unnecessary predictors in the model.	(1) Variance Inflation Factor (VIF) (2) Visualization: Added Variables Plot
Case analysis **There are no "inappropriate" influential observations**	Take care!! Sometimes, as when investigating rare outcomes that are important to detect, it just might be that the supposedly "influential" observation that is the discovery you are after! Soooooo .... Tip. Be thoughtful in your investigation of "influence"	(1) Studentized residuals - are there outliers with respect to the Y outcomes value? (2) Leverage - are there outliers with respect to the X predictor outcomes? (3) Cook's distance - are there observations that are exerting too much influence on the fitted model (betas)

d. Assessment of Normality of Errors



Subpopulations of Y are defined by each profile of predictors values, $X_1=x_1$, $X_2=x_2$, ... $X_p=x_p$. The distributions of Y in these subpopulations are called **conditional distributions of $Y_{\underline{X}}$** .

The conditional probability distributions of $Y_{\underline{X}}$ are assumed to be Normal distributions with

mean = $\mu_{Y|\underline{x}} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$; and

constant variance = $\sigma^2_{Y|x}$

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.

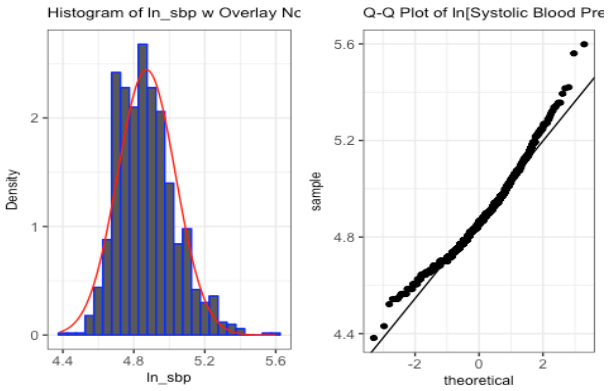
Graphical assessments of normality

Method	What to watch out for:
Scatterplot of residuals, e, with overlay normal.	Look for: (1) nice bell shape, (2) center at zero; and (30 no gross outliers
Histogram of outcome variable Y and/or Histogram of residuals with overlay normal	Look for normal shape of the histogram.
Histogram of residuals (or studentized or jackknife residuals) with overlay normal	Look for normal shape of the histogram.
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.

Example - R

Histogram w Overlay Normal and QQ plot

```
library(ggplot2)
library(gridExtra)
# Left Panel
# ggplot(data= DATAFRAME, aes(x=VARIABLENAME)) + geom_hisotgram() + stat_function( ) + options
p1 <- ggplot(data=framingham, aes(x=ln_sbp)) +
     geom_histogram(binwidth=.05, colour="blue",                    # TIP - You may want to tweak binwidth =
                            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)                              # ncol=2 arranges panels in 2 columns, 1 row
```



The histogram looks reasonably bell shaped but we see some departure from the ideal straight line in the QQ plot. It is not too terrible, however, so we will proceed.

Hypothesis test assessments of normality

Null Hypothesis H_O: Outcomes **Y** are distributed Normal
Alternative Hypothesis H_A: Not, two sided

Test Statistic	What to watch out for:
Shapiro Wilk (W) W is a measure of the correlation between the values in the sample and their associated normal scores	**Null Hypothesis H_O true** W = 1 p-value = large **Alternative Hypothesis H_A true** W < 1 p-value = small
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	**Null Hypothesis H_O true** D ~ 0 K-S ~ 0 p-value = large **Alternative Hypothesis H_A true** D > 0 K-S > 0 p-value = small

Good to Know

- 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)

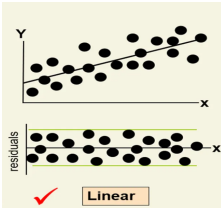
Example - R

Shapiro Wilk Test of normality

```
shapiro.test(saved.residuals)
shapiro.test(ready$fit.resid)

##
##  Shapiro-Wilk normality test
##
## data:  ready$fit.resid
## W = 0.9873, p-value = 0.000004419
```

Interpretation: This is a nice example of how sample sizes that are very large (here, n=748) can produce statistical significance when, in reality, the data themselves do not suggest a meaningful departure from the null. A great reminder of the importance of looking at the data!

e. Assessment of Constancy of Error Variance



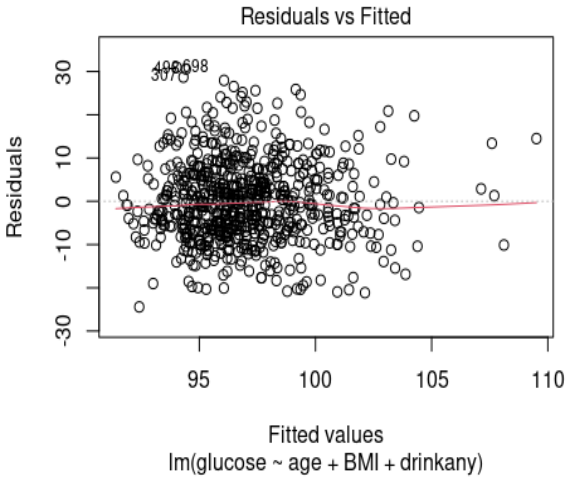
https://sid-sharma1990.medium.com/general-linear-model-3-residual-analysis-892ab34af76

Graphical assessments of constancy of error variance

Method	What to watch out for:
Plot X = Predicted outcomes $\hat{Y}$ on the horizontal Y = Residuals or standardized residuals or studentized residuals on the vertical	Look for even band at zero (all is well!)
Plot X = Predictor values of X on the horizontal Y = Residuals or standardized residuals or studentized residuals on the vertical	Look for even band at zero

Example - R

```
plot(fit, which=1)                    # user names model object fit
plot(fit, which = 1)                  # which=1 plots X=predicted v Y=residual
```



Interpretation: Looks reasonably like an even band centered at zero. So, okay.

Hypothesis Test of homogeneity of variance (Cook-Weisberg)

Null Hypothesis H_O: Errors **e** have constant variance

Alternative Hypothesis H_A: Not, two sided

Cook-Weisberg Test (also known ast Breusch-Pagan 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).	**Null Hypothesis H_O true** Chi square test value = small p-value = large **Alternative Hypothesis H_A true** Chi square test = large p-value = small

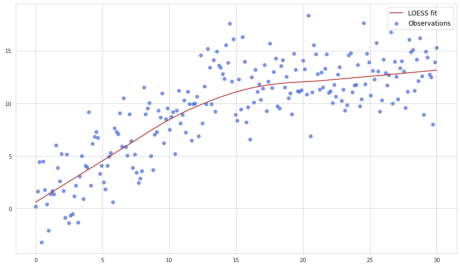
Example - R

Cook-Weisberg Test is function ncvTest( ) in package {car}

```
library(car)
ncvTest(fit)

## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 1.832514, Df = 1, p = 0.17583
```

Interpretation: Do NOT reject the null hypothesis of constant variance (p-value = .18) Assumption of the null (constant error variance) and its application to the data has NOT led to a contradition of the null. All is well.

f. Assessment of Functional Form



source: https://vzahorui.net/regression/loess/

Nature is probably rarely (if ever!) linear. But our hope is that a fitted linear model is a sufficiently reasonable fit (translation "good") so that we can draw inferences about relationships in nature and make predictions. There are many techniques for assessing functional form. Most are beyond the scope of this course.

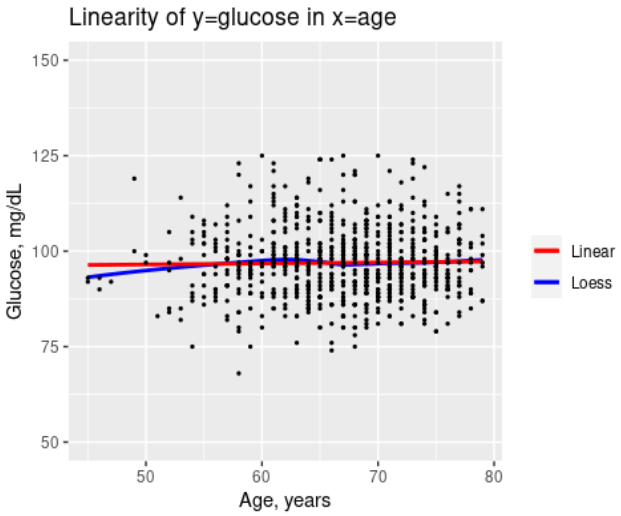
Graphical assessments of functional form

Method	What to watch out for:
loess/lowess smoothing This is a smoothing technique that is useful for graphically assessing departures of the true functional form from linearity. Briefly, at each value of the predictor X, lowess smoothing involves fitting a smooth polynomial regression model to "local" data.	Look for similarity of loess and linear regression plot (all is well!)

Example - R

XY Scatterplot with Overlay Line and Lowess

```
library(ggplot2)
# linearity of Y = glucose in X =age
ggplot(data=ready) +
    aes(y=glucose) +
    aes(x=age) +

    geom_smooth(method="loess", aes(color="Loess"), se=FALSE) +                    # Loess smooth w no CI
    geom_smooth(method="lm", aes(color="Linear"), se=FALSE) +                      # linear fit
    geom_point(size=0.5) +                                                         # X-Y scatter
    scale_colour_manual(name="", values=c("red","blue")) +
    scale_y_continuous(limits = c(50,150), breaks = seq(50,150, by=25))  +         # set y-axis explicitly
    ggtitle("Linearity of y=glucose in x=age") +
    xlab("Age, years") +
    ylab("Glucose, mg/dL")
```



Interpretation: Pretty close! Conclude okay to assume linearity of Y=glucose in X=age.

Method of Fractional Polynomials

The method of fractional polynomials is a technique for selecting a "good" functional form that relates Y to X from a collection of candidate models known as the Box-Tidwell family. Briefly, "under the hood", R (or whatever statistical software you are using) performs maximum likelihood estimation of the data using several polynomial regression models that related the outcome Y to the predictor X. It then reports to you your "choices"!

This method is beyond the scope of this course. However, a brief description follows.

Instead of fitting a
simple linear relationship of the form

$$\beta_1 X$$

We consider fitting a
fractional polynomial relationship of the form

$$\beta_1 X^{p_1} + \beta_2 X^{p_2} + \beta_3 X^{p_3} + \ldots + \beta_m X^{p_m}$$

where
m = number of powers ("degree")
$p_1, p_2, p_3, \ldots p_m$ 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 **$p_2 = p_1$** we consider $\beta_1 X^{p_1} + \beta_2 X^{p_1} \ln(X)$.

Example: Suppose **m=1** with **$p_1 = 1$** . This yields

$$Y = \beta_0 + \beta_1 X$$

Example: Next, suppose **m=2** with **$p_1 = 0.5$** and **$p_2 = 0.5$.** Because **$p_2 = p_1$** this yields

$$Y = \beta_0 + \beta_1 \sqrt{X} + \beta_2 \sqrt{X} \ln(X)$$

Brief Description of Method of Fractional Polynomials

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. (***Stay tuned.*** We will learn about the "deviance" statistic in Unit 7, Logistic Regression.)

The search for a "good" model by the method of fractional polynomials begins with an examination of all the models for which m=1. We choose the one model in this class that is the best according to maximum likelihood estimation.

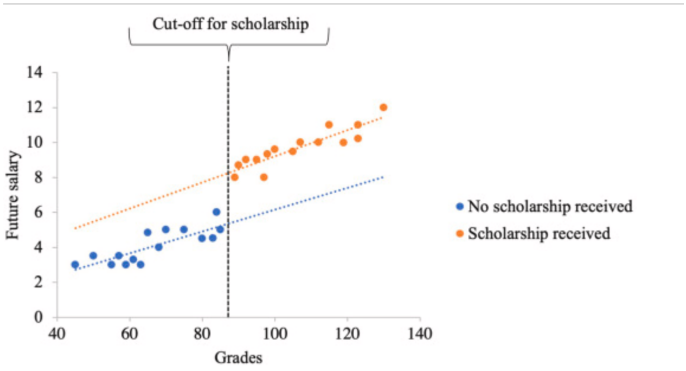
> We compare the best m=1 model to the specific model for which **m=1** and **p_1= 1** because the latter is the simple linear model.
>
> Thus, we are asking whether it is really necessary to abandon the simple linear model.

Next, we compare the best m=1 model to the best m=2 model. And so on … There's always a trade-off:

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 ☹

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

g. Ramsey Test of Model Misspecification



source: https://www.sciencedirect.com/science/article/pii/S2590260121000321

A misspecified model can be a problem

- **Omitted variable bias**. This is uncontrolled confounding and is illustrated in the picture above. The model fit is Y = future Salary and X=Grades. The omitted variable is Z=scholarship;
- **Model misspecification**. We might miss detecting important biology;
- **Violation of assumptions**. We may have violated some model assumptions.

Hypothesis Test of Model Misspecification

Null Hypothesis H_O: The model is correctly specified
Alternative Hypothesis H_A: The model is misspecified (wrong form and/or omitted variables)

Ramsey Test	What to watch out for:
Idea. If the model is reasonably adequate (correctly specified), then the predictors in the current model should suffice. It should NOT be necessary to add to the model any non-linear functions of the predictors that are already in the model. Computation. The computation of the Ramsey Test statistic involves the fit of a new regression model with Outcome = Fitted $\widehat{\mathbf{Y}}$ from current model and Predictors = Fitted $\widehat{\mathbf{Y}}$ + $\widehat{\mathbf{Y}}^2$ + original predictors X. Test Statistic is a Partial F statistic that asseses the extra significance of $\widehat{\mathbf{Y}}^2$ + original predictors X (this makes sense because if the model is adequate, the predictor $\widehat{\mathbf{Y}}$ is all that is needed, nothing else)	**If the model is adequate** F statistic = small p-value = large **Alternative Hypothesis H_A true** F statistic = large p-value = small

Example - R
Ramsey Test of Model Misspecification using resettest( ) in package {lmtest}

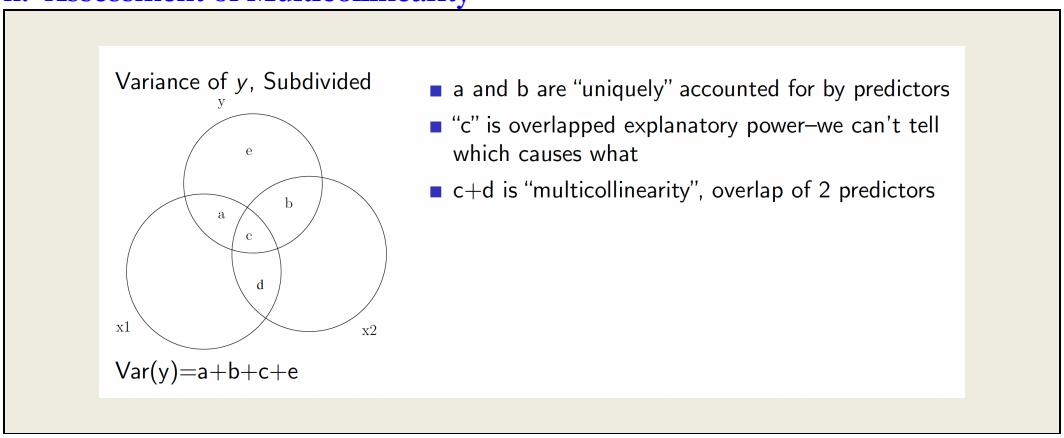
```
library(lmtest)
resettest(fit, power=2, type="regressor")

##
##  RESET test
##
## data:  fit
## RESET = 0.18324, df1 = 3, df2 = 741, p-value = 0.9078
```

Interpretation: Do NOT reject the null hypothesis of adequate model specification (p-value = .91). We have no statistically significant evidence that the model is misspecified (either with respect to its included predictors or with respect to omitting important predictors).

Accompany this test with a visualization. Consider accompanying your Ramsey Test with a scatterplot of the squared standardized residuals versus the leverage values. Omission of any important explanatory variables is suggested by

- Extreme values
- Any systematic pattern

h. Assessment of Multicollinearity



Source: Paul E. Johnson1 2 1Department of Political Science, 2Center for Research Methods and Data Analysis, University of Kansas, 2014

Multicollinearity is a problem

Nature, being what it is, means that there is bound to be some multicollinearity. We hope that it is not too much. The consequences of problematic multicollinearity include:

- **Redundancy.** The multicollinear predictor possesses **too little information** for the modeling of Y;
- **Uninterpretability.** If X_1 and X_2 are collinear, how do you interpret the beta for X_1? You don't actually know what portion of X_1's effect on Y is due to X_1 or X_2.
- **Unstable model.** In the presence of multicollinearity, the fitted model becomes very unstable.
  - The standard errors, SE (beta), will be inflated (and can be quite big)
  - If your primary interest is the predictor X_1, in the presence of multicollinearity of X_1 by X_2, **the beta for X_1 can change dramatically** depending on whether or not X_2 is in the model.

Graphical assessment of multicollinearity

Method	What to watch out for:
matrix scatterplot This is a matrix of all pairwise scatterplots: Y with each predictor and Each predictor with the remaining (p-1) predictors	In assessing multicollinearity, the focus is on the pairwise plots of each predictor with the remaining (p-1) predictors. If there is little to no multicollinearity X-X scatterplots will suggest no linearity and small pairwise correlations. If there is problematic collinearity X-X scatterplots will show a pattern and larger pairwise correlations.

Numerical assessment of multicollinearity: Variance Inflation Factor (VIF) and Tolerance

For each of the p predictors in the model, i = 1, 2, ..., p

- **Fit the (p - 1) predictor regression models**. For the ith predictor
  Outcome = i^{th} predictor X_i
  Predictors = All of the remaining (p-1) predictors.
  Obtain $R^2_{\text{regression of ith predictor on remaining (p – 1) predictors}}$

- **Compute the variance inflation factor (VIF) statistic values**
  The VIF for the i^{th} predictor is defined:

$$VIF_i = \frac{1}{\sqrt{1 - R^2_{regression\ of\ ith\ on\ all\ other\ predictors}}}$$

- **Compute the tolerance statistic values**
  The Tolerance for the i^{th} variable is defined:

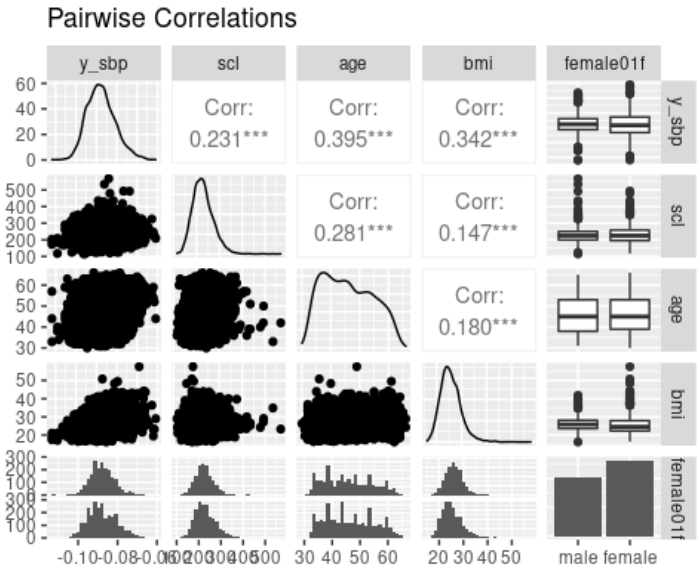
$$Tolerance_i = \frac{1}{VIF_i}$$

Little to no multicollinearity	Suspect multicollinearity
$VIF_i < 10$ or so	$VIF_i > 10$ or so
$Tolerance_i > 0.10$ or so	$Tolerance_i < 0.10$ or so

Example - R

Matrix Plot using ggpairs( ) in package {GGally}

```
library(GGally)
GGally::ggpairs(data=ready,
  columns=c("y_sbp", "scl", "age", "bmi", "female01f")) +
  ggtitle("Pairwise Correlations")
```



Example - R

Variance inflation factor using vif( ) in package {car}

```
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 sex at birth.

i. Case Analysis: Residuals, Leverage, and Cook's Distance

Residuals - There are multiple measures of "residual".

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

Which one or ones should we use?

- Standardized residuals can be (roughly) interpreted as z-scores.
- Studentized residuals can be (roughly) interpreted as t-scores from a Student's t (df=n-p-1) when regression assumptions hold.
- Jacknife residuals can be (roughly) interpreted as t-scores from a 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, $h_i = \dfrac{1}{n} + \dfrac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

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

Cook's Distance, d

Recall. Neither a large residual alone nor a high leverage alone is a guaranteed that an individual data point is influential. To see this, see again the pictures on pp 43-44.

Cook's distance to the rescue. 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 $d_i = \dfrac{\left(\hat{\beta}_{-i} - \hat{\beta}\right)' (X'X) \left(\hat{\beta}_{-i} - \hat{\beta}\right)}{p' s^2_{Y|x}}$ 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

How big should a Cook's Distance be to conclude the data point is influential?

Simple Linear Regression:
Cook's distance $d \geq 1$.

Multiple Linear Regression:
Cook's distance $\geq 2(p+1)/n$ where

n = sample size; and
p = # predictors.

Example - R

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

```
library(Hmisc)
library(ggplot2)
complete$ID <- as.numeric(row.names(complete))                    # create study id using row.names( ) and as.numeric( )
Hmisc::label(complete$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).

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()
```