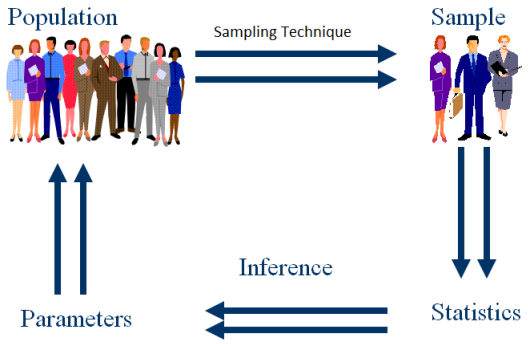
Chapter 1

ST 512 - Review

Readings: Chapters 1-8 as needed

Big ideas in stats:

- Population - all the values, items, or individuals of interest
- Parameter - a (usually) unknown summary value about the population
- Sample - a subset of the population we observe data on
- Statistic - a summary value calculated from the sample observations



Examples of paramters - (true) mean μ, (true) variance σ^2.
Examples of statistics - sample mean $\bar{y}$, sample variance $s^2 = \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}$
Inference - Making mathematically sound claims about the poulation using sample data.

Scales (Types) of Data:

- Qualitative or Categorical - A variable that is described by attributes or labels
  Subscales:
  Nominal - categories have no ordering (Male, Female)
  Ordinal - can order categories (Lickert scale data)

- Quantitative - A variable that is described by numerical measurements where arithmetic can be performed
  Subscales:
  Discrete - finite or countable finite number of values (# of flowers on a plant, 0, 1, 2, ...)
  Continuous - any value in an interval is possibel (Temperature, $(-459.67 \deg F, \infty)$

Random Variables and Things of Interest:

- Random Variable (RV) - Function that takes in outcomes from an experiment and outputs real numbers, or a numeric outcome to a random process

  Things of interest

  - Distribution - pattern and frequency of observable values
    For continuous RVs, visualized with a smooth curve.
  - Mean/Median - measures of center of the distribution

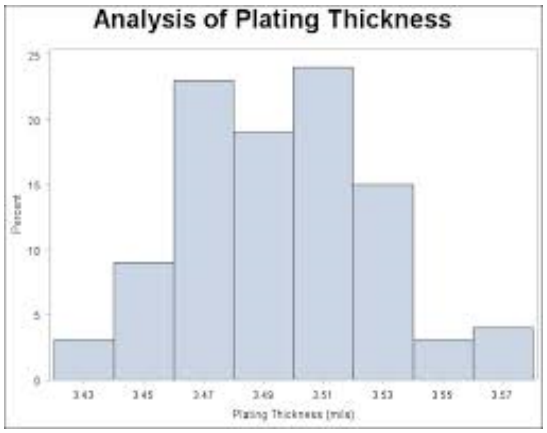
    Focus on mean: true mean μ, RV sample mean $\bar{Y}$, observed sample mean $\bar{y}$
  - Standard Deviation, Variance, IQR, Range - measures of spread for the distribution

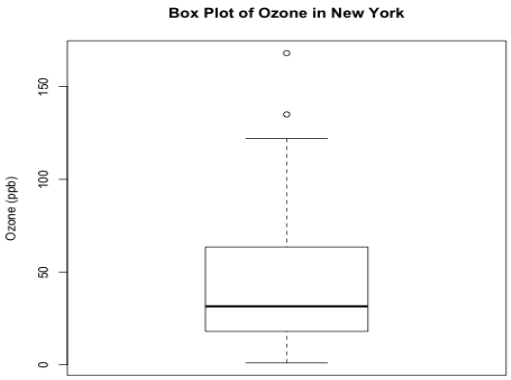
    Focus on SD and Variance: true variance σ^2, true SD σ, observed sample variance s^2, observed SD s

Graphical Descriptions of RV's:

- Histogram - Graphs the frequencies or relative frequencies of realizations of a RV

- Boxplot - Uses the Five Number Summary to display the realizations of a RV
  Five number summary: $min,\ Q_1,\ M,\ Q_3,\ max$





Statistics are also RVs. The distribution of a statistic is called a sampling distribution

Central Limit Theorem (CLT):

If a RV Y has a (true) mean μ and (true) variance σ^2, and a random sample is of size $n \geq 30$ is taken then

$$\bar{Y} \sim N(\mu, \sigma^2/n)$$

Note: If $Y \sim N(\mu, \sigma^2)$ then $\bar{Y} \sim N(\mu, \sigma^2/n)$ for any n.

2 main ways to make inference about a (true) mean, μ:

1. When the true SD, σ, is known we looked at the sampling distribution of the statistic

   $Z = \frac{\bar{Y}-\mu}{\sigma/n} \sim N(0,1)$ valid if $\bar{Y}$ has a normal distribution

   Allows us to form a CI: And a test statistic: Testing $H_0 : \mu = \mu_0$

   $\bar{y} \pm z_{\alpha/2}\sigma/\sqrt{n}$ $z_{obs} = \frac{\bar{y}-\mu_0}{\sigma/\sqrt{n}}$

2. When the true SD, σ, is unknown we looked at the sampling distribution of the statistic

   $T = \frac{\bar{Y}-\mu}{s/n} \sim t_{n-1}$ valid if $\bar{Y}$ has a normal distribution, allow for $n \geq 15$ or so in CLT

   Allows us to form a CI: And a test statistic: Testing $H_0 : \mu = \mu_0$

   $\bar{y} \pm t_{(n-1,\alpha/2)}s/\sqrt{n}$ $t_{obs} = \frac{\bar{y}-\mu_0}{s/\sqrt{n}}$

Inference about two (true) means, μ_1 and μ_2:

- From paired samples, $x_1, x_2, ..., x_n$ and $y_1, y_2, ..., y_n$ where difference is normally distributed

  CI: $(\bar{x} - \bar{y}) \pm t_{(n-1,\alpha/2)} s_{diff}/\sqrt{n}$

  HT: $H_0 : \mu_1 = \mu_2$, i.e. $\mu_1 - \mu_2 = 0 \quad t_{obs} = \frac{(\bar{x}-\bar{y})-0}{s_{diff}/\sqrt{n}}$

- Two separate samples from normal populations, $x_1, x_2, ..., x_n$ and $y_1, y_2, ..., y_n$

  CI: $(\bar{x} - \bar{y}) \pm t_{(\nu,\alpha/2)} \sqrt{s_X^2/n + s_Y^2/m}$ where ν is an estimate of df

  HT: $H_0 : \mu_1 = \mu_2$, i.e. $\mu_1 - \mu_2 = 0 \quad t_{obs} = \frac{(\bar{x}-\bar{y})-0}{\sqrt{s_X^2/n + s_Y^2/m}}$

Extension to inference about t (true) means, $\mu_1, \mu_2, ..., \mu_t$:

Balanced One-way ANOVA table (same number of replicates per group)

Source	DF	SS	MS	F-stat	P-value
Treatment	$t-1$	$n\sum_{i=1}^{t}(\bar{Y}_{i+} - \bar{Y}_{++})^2$	$\frac{SS(Trt)}{t-1}$	$\frac{MS(Trt)}{MS(E)}$	Use $F(t-1, t(n-1))$
Error	$t(n-1)$	$\sum_{i=1}^{t}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_{i+})^2$	$\frac{SS(E)}{t(n-1)}$		
Total	$nt-1$	$\sum_{i=1}^{t}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_{++})^2$			

Analysis used for a completely randomized design.

P-value tests $H_0 : \mu_1 = \mu_2 = ... = \mu_t$ vs H_A : at least 1 mean differs

One Way ANOVA model:

$$Y_{ij} = \mu + \alpha_i + E_{ij}$$

where $i = 1, 2, ...t$ and $j = 1, 2, ..., n$ (total sample size $= nt = N$)

μ = overall mean
α_i = effect from group i
E_{ij} = random error assumed to be $iid\ N(0, \sigma^2)$

For two quantitative variables measured on the same units, the linear relationship can be investigated:

Simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + E_i$ or use correlation.

For a hypothesis test, the p-value means

probability of obsevign a test statistic as extreme or more extreme than the one observed, assuming the null hypothesis is true.

For a given a null hypothesis, statistical significance implies

the observed value was unlikely to have occurred by random chance alone (assuming the null hypothesis is true).

For an observed confidence interval (cL, cU) we can say

We are ___% confident the true parameter value is contained in the interval. (***Do not say probability or chance!)

The idea of Confidence means

The procedure used to create the interval has a __% probability of producing an interval that contains the parameter.

i.e. If the experiment were done repeatedly and an interval made for each sample, __% of the intervals would contain the parameter value.

Chapter 2

ST 512 - Experiments

Readings: 7.2 and 7.3, pg 244-255

Example: An experiment was run to determine the effects of adding phosphorous ($0, 147, 294, 441$ kg/m^2) and nitrogen ($0, 45, 90, 135$ kg/m^2) to the soil of a certain type of grass (a Miscanthus species). The growth of the plant was of interest and at the end of the growing period the plant was dried and the weight recorded with the final measurement being recorded in megagram per hectare (0.1 kg/m^2). Four plots of grass were used in total. Within each plot, each combination of phosphorous and nitrogen was observed. A partial data table is given here:

Plot	P	N	Dry yield
1	0	135	1.95
1	0	45	3.51
1	0	90	2.87
1	0	0	2.88
1	294	45	2.37
1	294	0	3.5
1	294	135	3.55
1	294	90	4.4
...	...	...	...

Let's identify (if possible) the response, explanatory variable(s), factor(s), level(s), confounding factor(s), treatment(s), number of replicates, and experimental units.

Sources of Variation in the responses of an experiment:

1. **Treatment effect** - we hope there is an effect due to the variables we control
2. **Identified confounding variables** - We record some variables that are not of interest, but we think may have an effect on the response.
3. **Unidentified sources (Experimental Error or error variation)** -
   (a) Inherent variability in experimental units - Experimental units are different!
   Ex: No two people, paper towels, concrete blocks, or even lab rats are exactly the same.
   Consequence: Experimental units respond differently to the same treatment
   (b) Measurement error - Multiple measurements of a same experimental unit typically contain error.
   If the same experimental unit is measured more than once, will the value be the same?
   Ex: Blood Pressure, Quality Ratings of food, Break a water sample in two, measure each for bacteria
   (c) Variations in applying/creating treatments
   The treatment is not clearly defined, leaving room for interpretation.
   Ex: Two researchers mix concrete, will it come out exactly the same? Ovens don‘t heat exactly the same, etc.
   (d) Effects from any other extraneous (or lurking) variables - Extraneous variables are those variables that are not part of the treatment, but may influence the response.

Let‘s identify these in the previous example.

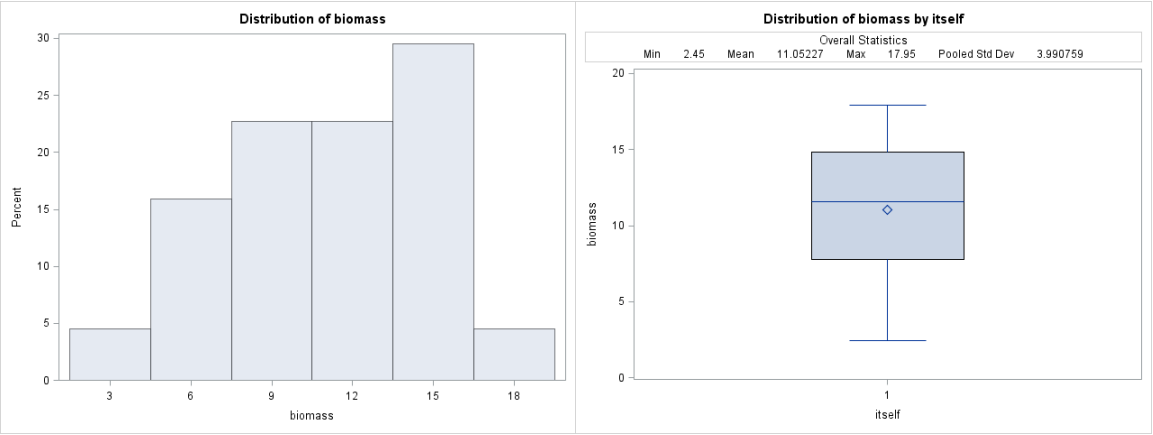
Chapter 3

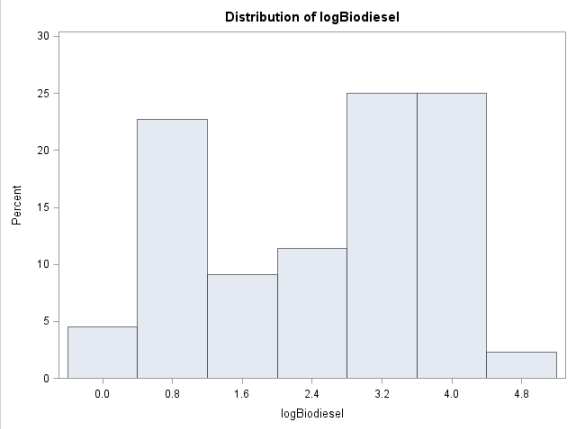
ST 512 - Correlation

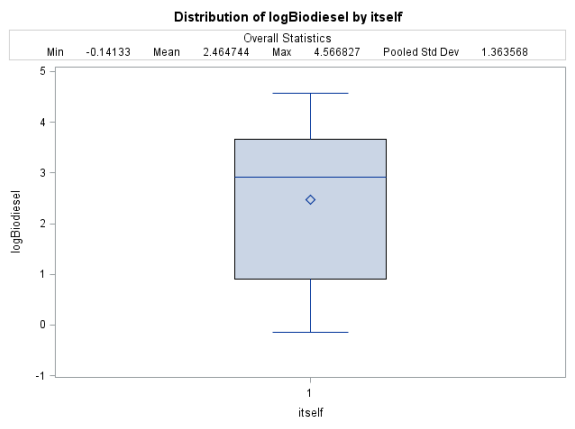
Readings for Correlation and SLR: 10.1-10.5 pg 378-420 and 10.7-10.8 pg 425-444 and 8.7 pg 305-311

Motivating example: One type of fuel is biodiesel, which comes from plants. An experiment was done to determine how much biodiesel could be generated from a certain type of plant grown in different medias. The final biomass was also recorded on 44 the plants from the experiment. Let's consider these two variables, the log of biodiesel and biomass.

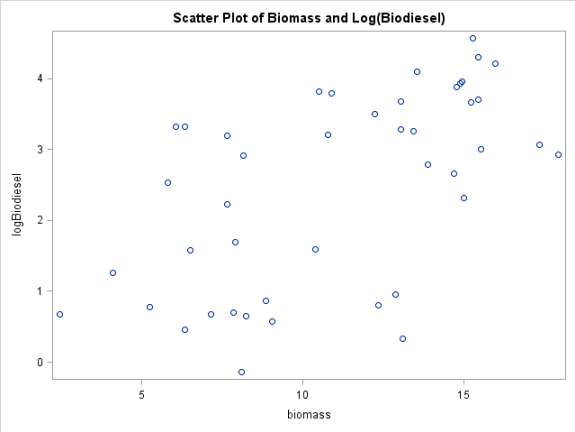
We can look at the distribution of each individually using our univariate methods (histogram, boxplot, mean/sd, etc.)

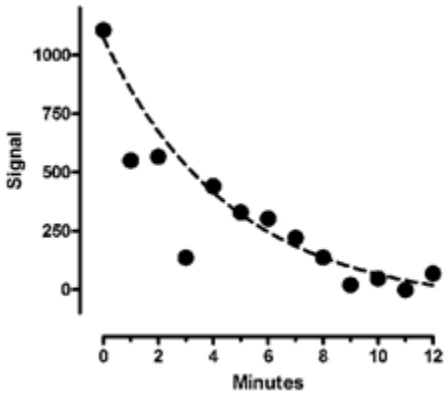






How can we visually inspect the association between the two? A **Scatter plot** gives a visual approximation of the "joint distribution" between two variables.





Properties of r_{XY}

- r_{XY} is an observed measure of the linear assn. between X and Y in a dataset.
- correlation coefficient is unitless and always between -1 and 1:

$$-1 \leq r_{XY} \leq 1$$

- The closer r_{XY} is to 1, the stronger the positive linear association
- The closer r_{XY} is to -1, the stronger the negative linear association
- The bigger $|r_{XY}|$, the stronger the linear association
- If $|r_{XY}| = 1$, then X and Y are said to be perfectly correlated (relationship is deterministic)

For the log(Biodiesel) (call this Y) and Biomass (call this X) example we can compute the sample correlation coefficient using summary statistics:

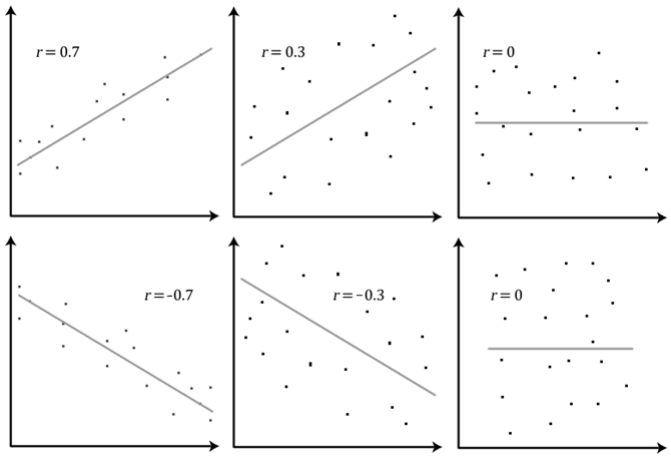
$$\bar{x} = 11.0523, \quad s_X = 3.9908, \quad \bar{y} = 2.4647, \quad s_Y = 1.3636$$

$$s_{XY} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1} = 3.1485$$

Applying the formula for r_{XY}, we get

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{3.1485}{\sqrt{3.9908 \times 1.3636}} = 0.5786$$

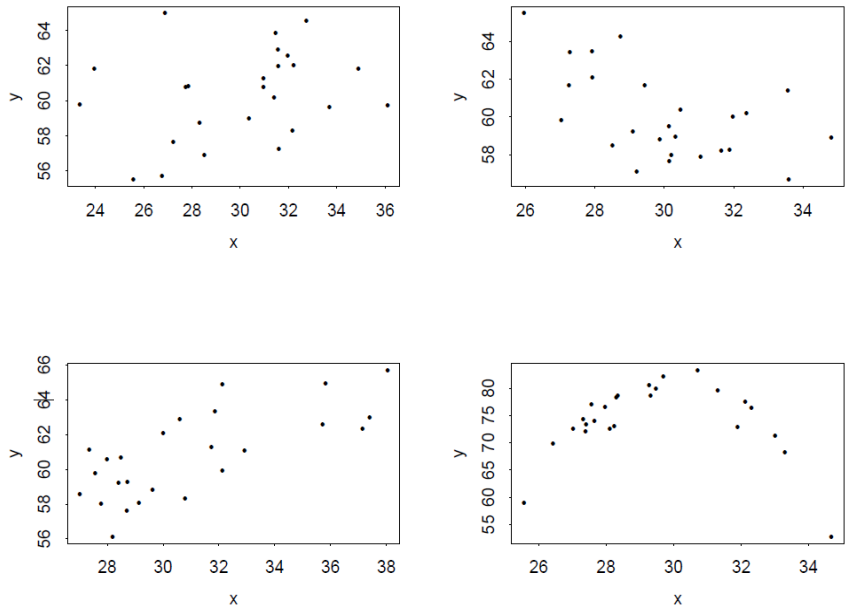
Some example scatter plots



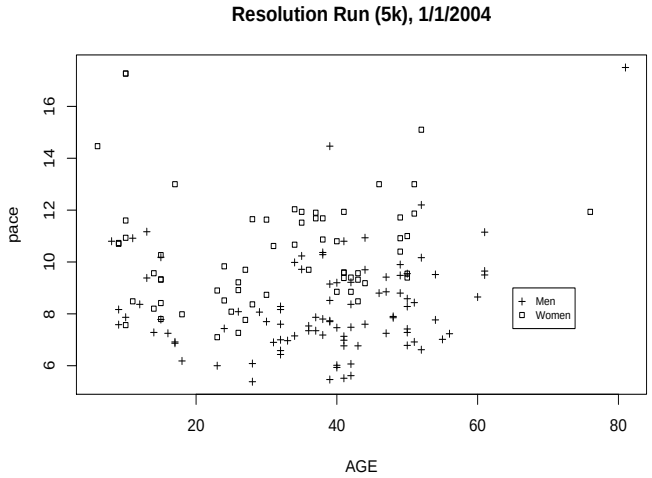
An exercise/activity:

Label the four plots below with the four sample correlation coefficients:

- $r = 0.3$ $\qquad$ $r = 0.7$
- $r = 0.1$ $\qquad$ $r = -0.6$



Would it be appropriate to use correlation to summarize the relationship between age and pace in the following scatter plot? Why or why not?



To perform a Hypothesis Test about ρ:
We often want to test the following hypotheses,

$$H_0 : \rho = 0 \qquad H_A : \rho \neq 0$$

Assuming H_0 is true, the test statistic is

$$z_{obs} = \left(\frac{1}{2}\sqrt{n-3}\right) \log \frac{1+r}{1-r}$$

and the reference distribution is the standard normal distribution, i.e. reject if $z_{obs} > z_{\alpha/2}$ or if $z_{obs} < z_{1-\alpha/2}$ where z_α satisfies $\alpha = \Pr(Z > z_\alpha)$ with $Z \sim N(0,1)$.

The p-value if found by finding $2P(Z > |z_{obs}|)$. Why do we multiply by 2?

To find a Confidence Interval for ρ:
An approximate $100(1-\alpha)\%$ confidence interval for ρ can be obtained by inverting the *Fisher transformation*:

$$\left(\frac{\frac{1+r}{1-r}e^{-2z_{\alpha/2}/\sqrt{n-3}} - 1}{\frac{1+r}{1-r}e^{-2z_{\alpha/2}/\sqrt{n-3}} + 1}, \frac{\frac{1+r}{1-r}e^{2z_{\alpha/2}/\sqrt{n-3}} - 1}{\frac{1+r}{1-r}e^{2z_{\alpha/2}/\sqrt{n-3}} + 1}\right).$$

For the log(Biodiesel) and Biomass example our hypothesis test is:

$$H_0 : \rho = 0 \qquad H_A : \rho \neq 0$$

$$\text{giving a test statistic of } z_{obs} = \frac{1}{2}\sqrt{44-3}\ log\left(\frac{1+0.5786}{1-0.5786}\right) = 4.228$$

Using an $\alpha = 0.05$ our rejection region is any z_{obs} outside of ± 1.96.

Our p-value $= 2P(Z > 4.228) = 2(0.00001) = 0.00002 < \alpha = 0.05$ so we reject our null hypothesis in favor of the alternative.

What is the interpretation of the p-value=0.00002?
The probability of getting a sample correlation (r) further (in magnitude) from 0 than 0.5786 assuming the true correlation (ρ) is 0 is 0.00002.

The corresponding 95% confidence interval is

$$\left(\frac{\frac{1+0.5786}{1-0.5786}e^{-2*1.96/\sqrt{44-3}}-1}{\frac{1+0.5786}{1-0.5786}e^{-2*1.96/\sqrt{44-3}}+1}, \frac{\frac{1+0.5786}{1-0.5786}e^{2*1.96/\sqrt{44-3}}-1}{\frac{1+0.5786}{1-0.5786}e^{2*1.96/\sqrt{44-3}}+1}\right) = (0.3401, 0.7471)$$

We can say that we are 95% confident that the true correlation (ρ) is between 0.3401 and 0.7471.

When we say confident, we mean that if we did this experiment repeatedly and made an interval for each experiment, the true correlation would fall in 95% of the intervals created.

How can we get SAS to do this for us?

```
proc corr data=bioexp FISHER(biasadj=NO);
var butterfat temp;
run;
```

Output From Proc Corr for Biomass and Log(Biodiesel) Example **1**

The CORR Procedure

2 Variables:	biomass	logBiodiesel

Covariance Matrix, DF = 43		
	biomass	logBiodiesel
biomass	15.92615751	3.14851427
logBiodiesel	3.14851427	1.85931767

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
biomass	44	11.05227	3.99076	486.30000	2.45000	17.95000
logBiodiesel	44	2.46474	1.36357	108.44873	-0.14133	4.56683

Pearson Correlation Coefficients, N = 44 Prob > |r| under H0: Rho=0		
	biomass	logBiodiesel
biomass	1.00000	0.57859 <.0001
logBiodiesel	0.57859 <.0001	1.00000

Pearson Correlation Statistics (Fisher's z Transformation)							
Variable	With Variable	N	Sample Correlation	Fisher's z	95% Confidence Limits		p Value for H0:Rho=0
biomass	logBiodiesel	44	0.57859	0.66035	0.340140	0.747136	<.0001

Note: Significant correlation does NOT imply causation

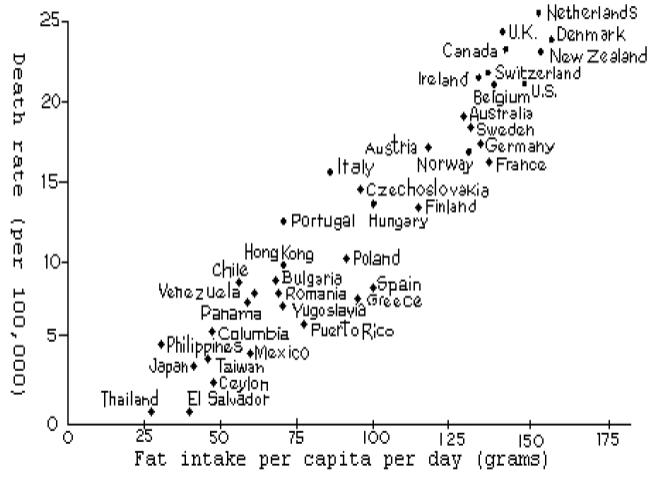
Famous examples of *spurious correlations*:

- A study finds a high positive correlation between coffee drinking and coronary heart disease. Newspaper reports say the fragrant essence of the roasted beans of *Coffea arabica* are a menace to public health.
- In a city, if you were to observe the amount of damage and the number of fire engines for enough recent fires, you would likely see a positive and significant correlation among these variables. Obviously, it would be erroneous to conclude that fire engines cause damage.
- *Lurking variable* - a third variable that is responsible for a correlation between two others. (A.k.a. confounding factor.)
  An example would be to assess the association between say the reading skills of children and other measurements taken on them, such as shoesize. There may be a statistically significant association between shoe size and reading skills, but that doesn't imply that one causes the other. Rather, both are positively associated with a third variable, *age*.
- Among 50 countries examined in a dietary study, high positive correlation among fat intake and cancer (see figure, next page). This example is taken from from *Statistics* by Freedman, Pisani and Purves.

> In countries where people eat lots of fat like the United States rates of breast cancer and colon cancer are high. This correlation is often used to argue that fat in the diet causes cancer. How good is the evidence?
>
> Discussion. If fat in the diet causes cancer, then the points in the diagram should slope up, other things being equal. So the diagram is some evidence for the theory. But the evidence is quite weak, because other things aren't equal. For example, the countries with lots of fat in the diet also have lots of sugar. A plot of colon cancer rates against sugar consumption would look just like figure 8, and nobody thinks that sugar causes colon cancer. As it turns out, fat and sugar are relatively expensive. In rich countries, people can afford to eat fat and sugar rather than starchier grain products. Some aspects of the diet in these countries, or other factors in the life-style, probably do cause certain kinds of cancer and protect against other kinds. So far, epidemiologists can identify only a few of these factors with any real confidence. Fat is not among them.

(p. 152, *Statistics* by Friedman, Pisani, Purves and Adhikari)

Figure 8. Cancer rates plotted against fat in the diet for a sample of countries



Source: K. Carroll. "Experimental evidence of dietary factors and hormone-dependent cancers Cancer Research vol. 35 (1975) p.3379. Copyright by Cancer Research. Reproduced by permission

Chapter 4

ST 512 - Simple Linear Regression

Readings for Correlation and SLR: 10.1-10.5 pg 378-420 and 10.7-10.8 pg 425-444 and 8.7 pg 305-311

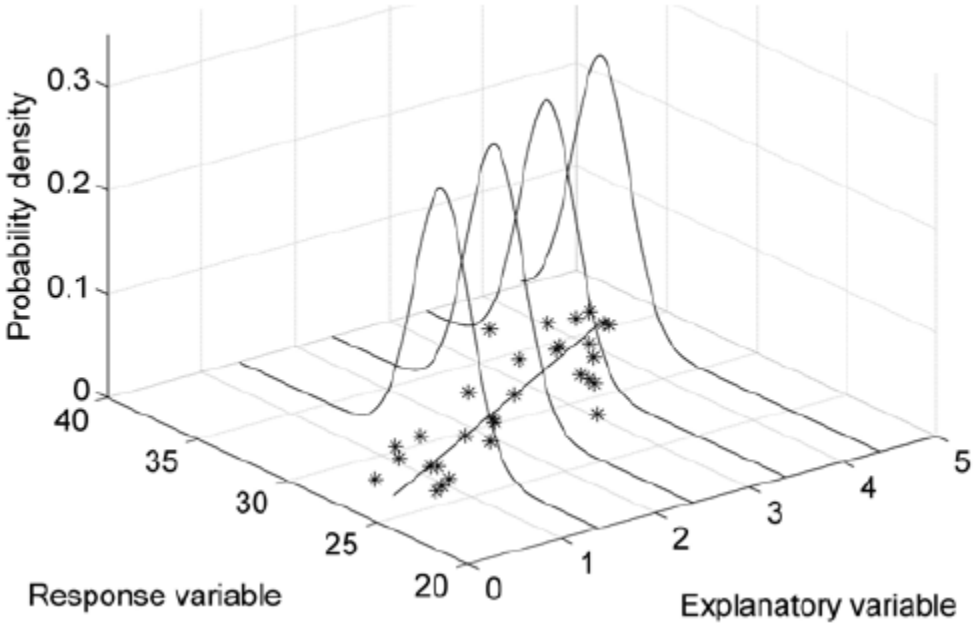
Fit a linear regression model - A probabilistic model for Y conditional on $X = x$:

$$Y_i = \beta_0 + \beta_1 x_i + E_i$$

Defintions:

- Y_i - response (also called dependent variable)
- x_i - explanatory variable (also called independent variable or predictor variable)
- E_i - random error for observation i
- $\beta_0 = E(Y|X = 0)$ - True population intercept (average value of response when $X = 0$
- β_1 - True population slope (average change in Y per unit increase in x)
- σ^2 - Error variance (variance due to experimental error)

Note: We make the assumption that $E_1, \ldots, E_n$ are independent and identically distributed normal random variables with mean 0 and variance σ^2. We write $E_i \overset{iid}{\sim} N(0, \sigma^2)$. This variance is assumed the same for all x, called assumption of **homoskedasticity**.



1. $E(Y|X = x) = \beta_0 + \beta_1 x = \mu(x)$ (The line describes the mean Y for a given X.)
2. $\text{Var}(Y|X = x) = \sigma^2$

For the log(Biodiesel) and Biomass example let's find our fitted line. Recall the summary stats on page 10.

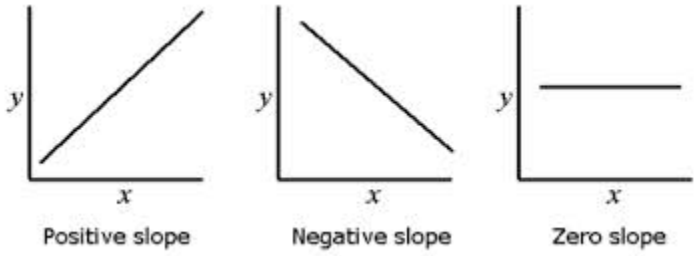
$$\hat{\beta}_1 = s_{XY}/s_X^2 = 3.1485/3.9908^2 = 0.1977$$
$$\hat{\beta}_0 = 2.4647 - 11.0523 * 0.1977 = 0.2797$$
$$\hat{y} = 0.2808 + 0.1977x$$

This line can now be used to make predictions for new X values by simply plugging in the x!

Again, we have now have point estimates for our true parameters. How can we make inference (claims about the true values)? Do we have a *significant linear relationship*?
Under the normal distribution assumption on the errors, the RV's $\hat{\beta}_0$ and $\hat{\beta}_1$ follow normal distributions. Thus, we can use this as a basis for inference.

What value of the slope do we test?

- If a linear relationship, Y will tend to change with X (i.e. $\beta_1 \neq 0$)
- If no linear relationship, Y won't tend to change with X (i.e. $\beta_1 = 0$).



Any hypothetical slope, like $\boxed{H_0 : \beta_1 = \text{slope}_0}$ may be tested using the T-statistic below with $df = n - 2$:

$$T = \frac{\hat{\beta}_1 - \text{slope}_0}{\widehat{SE}(\hat{\beta}_1)}$$

and any hypothetical intercept, like $\boxed{H_0 : \beta_0 = \text{intercept}_0}$ may be tested using the T-statistic below with $df = n - 2$:

$$T = \frac{\hat{\beta}_0 - \text{intercept}_0}{\widehat{SE}(\hat{\beta}_0)}$$

Confidence intervals for β_0, β_1

$100(1-\alpha)\%$ confidence intervals for β_0 and β_1 are given by

$$\hat{\beta}_0 \pm t(n-2, \alpha/2)\sqrt{MS[E]\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}.$$

$$\hat{\beta}_1 \pm t(n-2, \alpha/2)\sqrt{\frac{MS[E]}{S_{xx}}}.$$

Often we will only care about the test and CI for the slope. The hypothesis test is equivalent to checking if 0 is in the confidence interval. It will depend on the context of the question if testing β_0=0 makes sense.

Confidence interval for $\mu(x_0) = E(Y|X = x_0)$
The point estimate for $\mu(x_0)$ is simply $\hat{\beta}_0 + \hat{\beta}_1 x_0$. We need to know about the variability of this estimate and we can again use the t-distribution for inference.

$$\text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0 | X = x_0) \quad =$$

This yields a confidence interval of the form

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t(n-2, \alpha/2)\sqrt{MS[E]\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

Note: We are attempting to capture the true *mean* at x_0 in this interval.

Prediction interval for a new observation x_0
The point estimate for at x_0 is still $\hat{Y}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$. However, the variability will change.

$$\text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0 + E_{new} | X = x_0) \quad =$$

Thus we can form a PI using

$$\hat{Y}(x_0) \pm t(n-2, \alpha/2)\sqrt{MS[E]\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}.$$

Note: In this interval we are attempting to capture the next Y value that takes on x_0. As this is a much more difficult task, PI's are wider than CI's.

The ANOVA table from simple linear regression

The full ANOVA table for SLR is given below:

Source	Sum of squares	df	Mean Square	F-Ratio
Regression	$SS(R)$	1	$MS(R)$	$MS(R)/MS(E)$
Error	$SS(E)$	$n-2$	$MS(E)$	
Total	$SS(Tot)$	$n-1$		

The mean squares represent standardized measures of variation due to the different sources and are given by $SS(source)/dfsource$. Ratios of mean squares often follow an F-distribution and are appropriate for testing different hypotheses of interest.

In this case, to test

$$H_0 : \beta_1 = 0 \qquad \text{vs} \qquad H_1 : \beta_1 \neq 0$$

$F = MS(R)/MS(E) \sim F(1, n-2)$.

That is, the F statistic follows an F-distribution with 1 numerator df and $n-2$ denominator df. In SLR, this F test is equivalent to the T test we already looked at. The relationship is that $T^2 = F$.

Note: The mean square for error, $MS[E]$, is an unbiased estimator for σ^2. It is an estimate of the variability due left over once we account for our explanatory variable.

How to get tests in SAS?
For our Biodiesel and Biomass example we can get much of our output from SAS using the following commands:

```
proc reg data=bioexp ;
model logbiodiesel=biomass/clb;
run;
```

Output From Proc Reg for Biomass and Log(Biodiesel) Example 1

The REG Procedure
Model: MODEL1
Dependent Variable: logBiodiesel

Number of Observations Read	44
Number of Observations Used	44

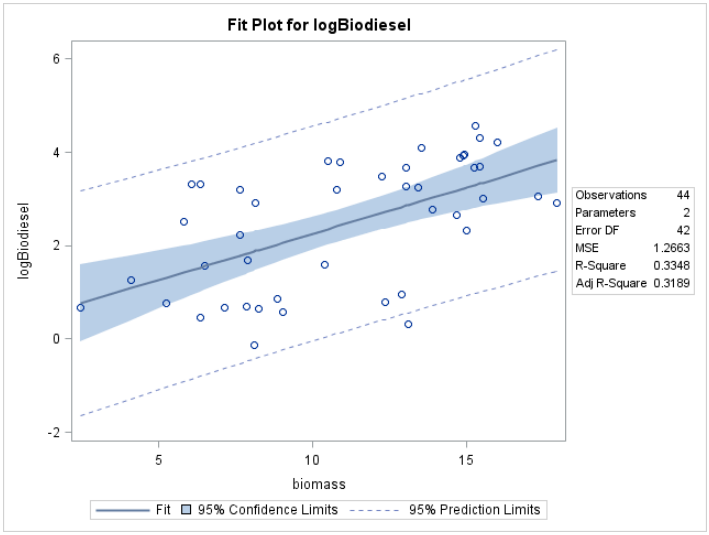
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	26.76509	26.76509	21.14	<.0001
Error	42	53.18557	1.26632		
Corrected Total	43	79.95066			

Root MSE	1.12531	R-Square	0.3348
Dependent Mean	2.46474	Adj R-Sq	0.3189
Coeff Var	45.65627		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	0.27977	0.50463	0.55	0.5822	-0.73862	1.29816
biomass	1	0.19769	0.04300	4.60	<.0001	0.11091	0.28447

Using $\alpha = 0.05$, (1) let's find the CI for the slope by hand, (2) form a CI for the mean log of biodiesel when biomass is 12, and (3) form a PI for a future log biodiesel measurement for a biomass of 12.

SAS will also produce a very nice plot that includes *pointwise* confidence and prediction bands at all points. Notice that the bands get wider the further x_0 is from $\bar{x}$. Why?

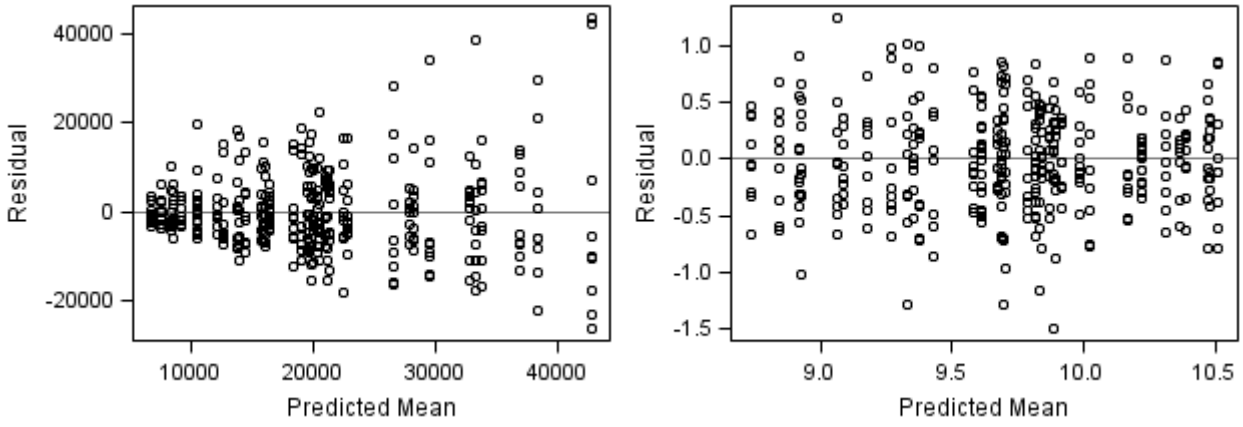


Checking assumptions

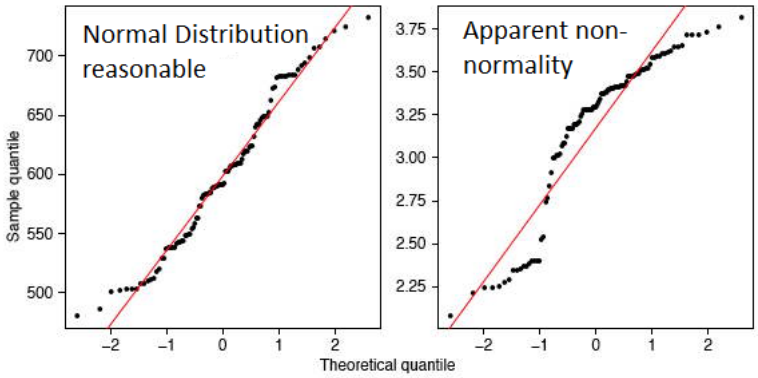
Firstly, we should always inspect a scatter plot to determine if the linear relationship we are assuming in our model is appropriate.

Secondly we can check our assumption of $iidN(0, \sigma^2)$ errors.

- Independence - There is not a check for independence of errors, we simply need to consider whether or not our EUs can be considered independent.
- Constant variance - A residuals vs fitted (predicted) values plot or a residual vs independent variable plot are tools for detecting heteroskedasticity (non-constant variance).



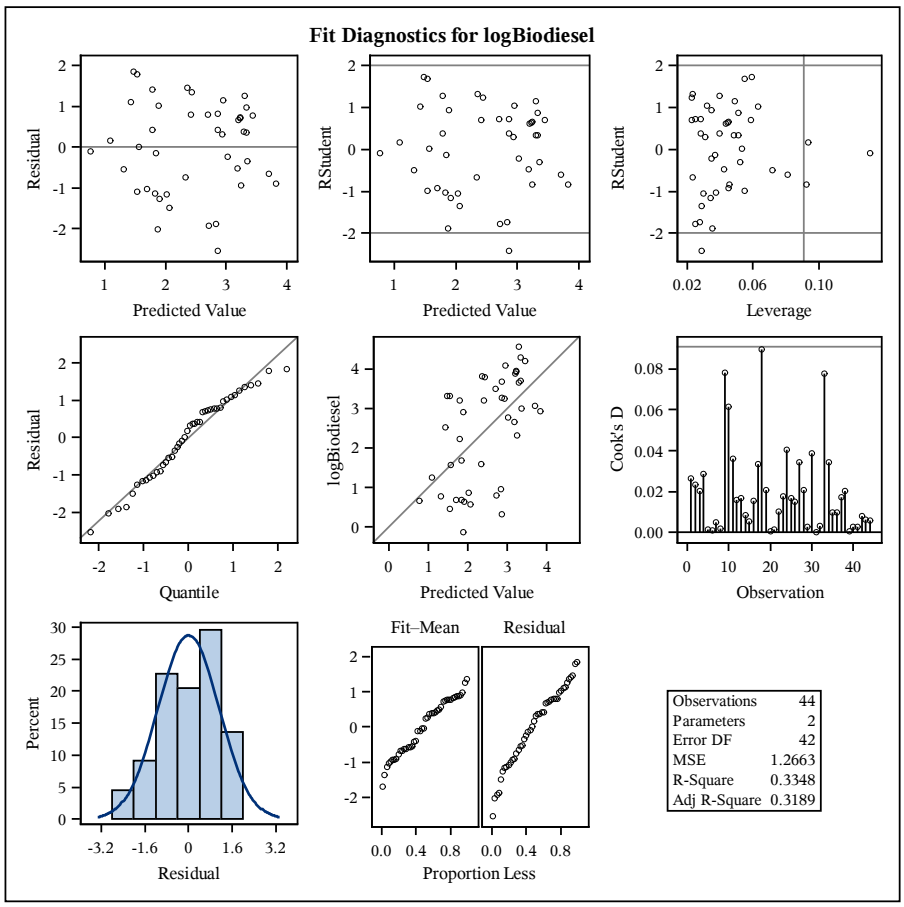
- Normality of errors - A quantile-quantile plot (or qq-plot for short) can be inspected to see if normality is reasonable.



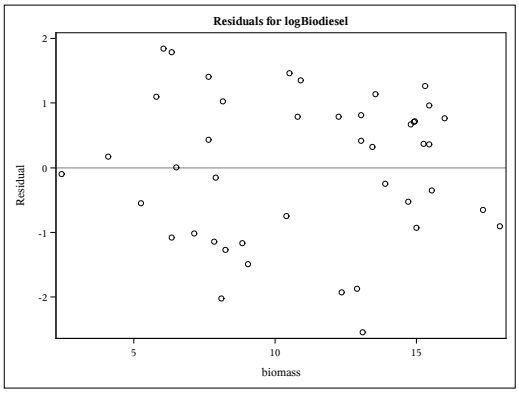
We can inspect the diagnostic plots that SAS produces when the reg procedure is used:

Output From Proc Reg for Biomass and Log(Biodiesel) Example **2**

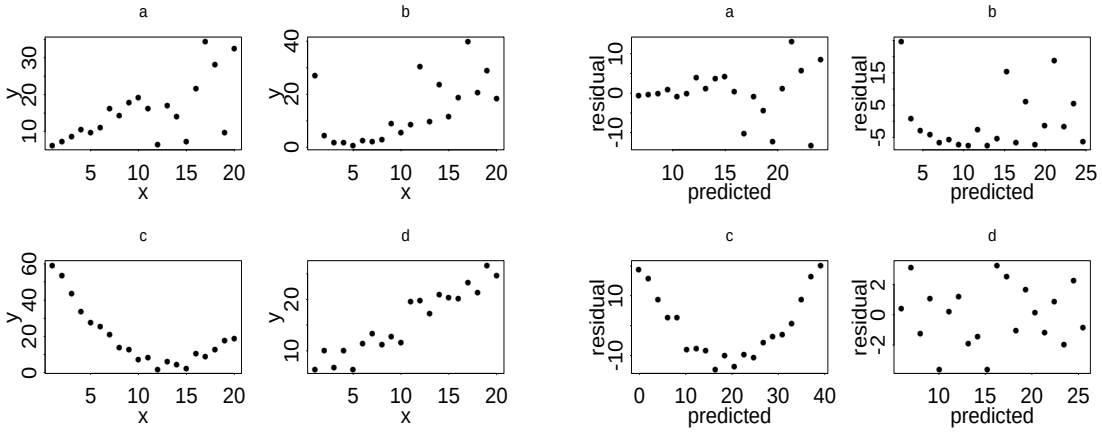
The REG Procedure
Model: MODEL1
Dependent Variable: logBiodiesel

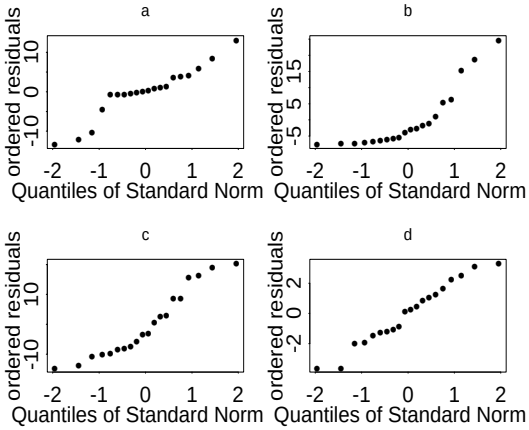


The REG Procedure
Model: MODEL1
Dependent Variable: logBiodiesel



An exercise: Match up letters a,b,c,d with the model violation - Heteroscedasticity, Nonlinearity, Nonnormality, Model fits





Chapter 5

ST 512 - Multiple Linear Regression

Readings: 11.1-11.6 and 11.9-11.11 pg 463 - 515 and 529 - 539

Motivating Example
(Taken from Probabilitiy and Statistics, Devore) Soil and sediment adsorption, the extent to which chemicals collect in a condensed form on the surface, is an important characteristic influencing the effectiveness of pesticides and various agricultural chemicals. A study was done on 13 soil samples that measured Y = phosphate adsorption index, X_1 = amount of extractable aluminum, and X_2 = amount of extractable iron. The data are given below:

Adsorption	Aluminum	Iron
4	13	61
18	21	175
14	24	111
18	23	124
26	64	130
26	38	173
21	33	169
30	61	169
28	39	160
36	71	244
65	112	257
62	88	333
40	54	199

MLR model for two quantitative explantory variables:
For observation i we can use the model

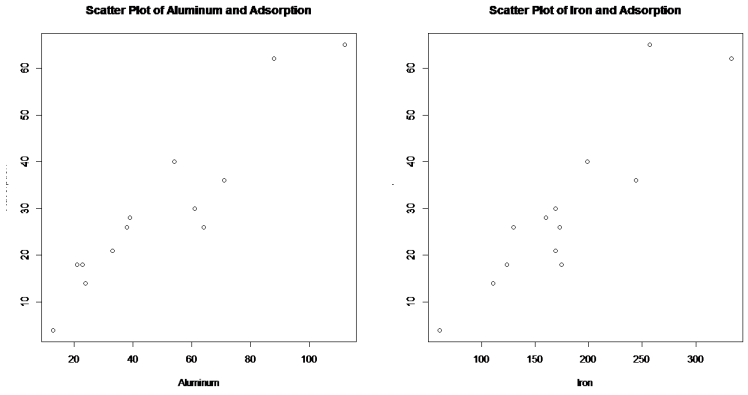
$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + E_i$$

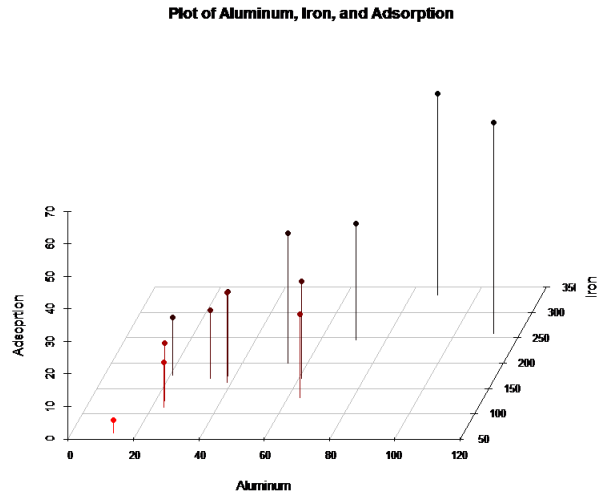
For clarity we can write the model for each subject

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + E_1 \\ Y_2 &= \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + E_2 \\ \vdots &= \vdots \\ Y_{13} &= \beta_0 + \beta_1 X_{13,1} + \beta_2 X_{13,2} + E_{13} \end{aligned}$$

Generally our model is

$$Adsorption = \beta_0 + \beta_1 Aluminum + \beta_2 Iron + Experimental\ Error$$

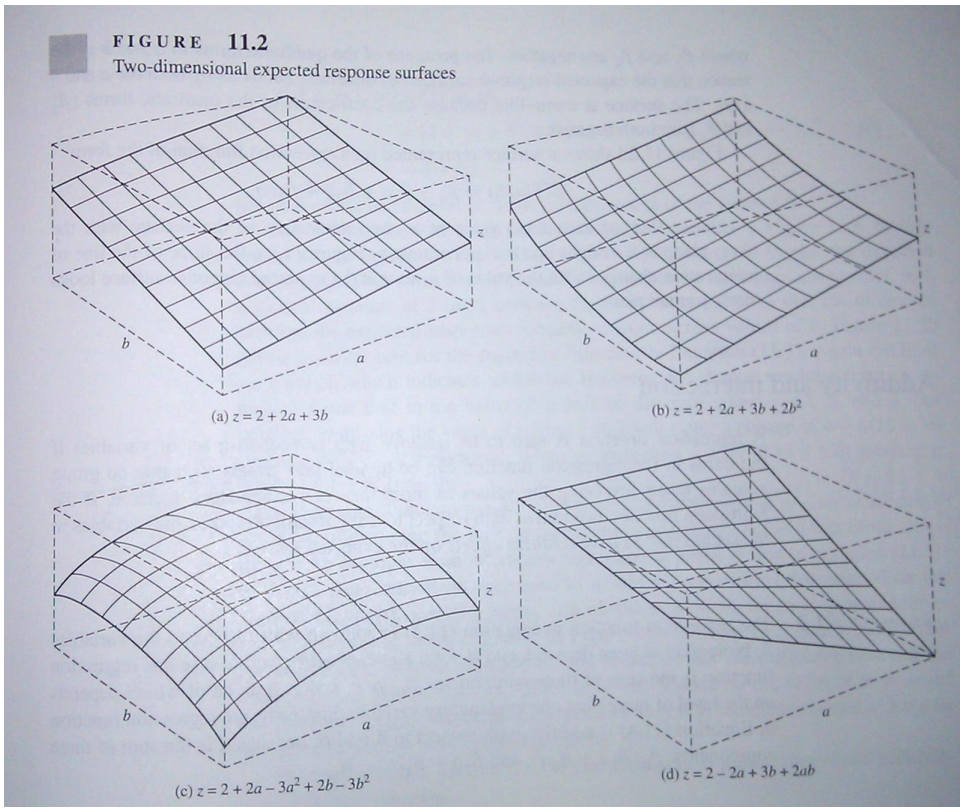




In this case, we don't want to find the best fitting line, but rather the best fitting *plane* (the one the minimizes the squared distances between the plane and the data points). Our hypothesis of interest is that at least one of our variables is useful (i.e. at least one partial slope is truly non-zero). We can then test

$$H_0 : \beta_1 = \beta_2 = 0 \quad vs \quad H_A : \text{ at least one is non-zero}$$

When we fit an MLR model with p different predictors we are really attempting to find the best 'response surface' of degree p in a $p + 1$ dimensional space. For instance, with one predictor, we are fitting the best line in a 2-d space. The plots below give a number of surfaces that can be fit using two predictors when quadratic or interaction terms are included.



A link to visualizing different surfaces:

`http://www.ats.ucla.edu/stat/sas/teach/reg_int/reg_int_cont.htm`

Very brief matrix review:
Note: Capital boldface letters are usually used for matrices and boldface lower case letters are usually used for vectors (matrices where the number of rows or the number of columns is 1).

Matrices - rectangular arrays of numbers that have a great many uses. Some matrices, (with *dimension* in parentheses):

$$\begin{aligned}
\mathbf{A} &= \begin{pmatrix} 7 & 5 \\ 5 & 2 \\ 3 & 2 \end{pmatrix} \quad (3 \times 2) \\
\mathbf{B} &= \begin{pmatrix} 4 & 2 & 1 \\ 3 & 1 & 1 \end{pmatrix} \quad (2 \times 3) \\
\mathbf{C} &= \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \quad (2 \times 2) \\
\mathbf{I}_2 &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad (2 \times 2)
\end{aligned}$$

Matrix operations

1. Transposition - swap rows for columns, columns for rows:

$$t(\mathbf{A}) = \mathbf{A}' = \begin{pmatrix} 7 & 5 & 3 \\ 5 & 2 & 2 \end{pmatrix} \quad \text{“transpose of } \mathbf{A}\text{”}$$

2. Addition is elementwise, matrices must have same *dimension*

$$\mathbf{C} + \mathbf{I}_2 = \begin{pmatrix} 1+1 & 1+0 \\ -1+0 & 1+1 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ -1 & 2 \end{pmatrix} \quad (2 \times 2)$$

   Subtraction, same deal

$$\mathbf{C} - \mathbf{I}_2 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \quad (2 \times 2)$$

3. Multiplication requires *conformability*. Element in i^{th} row, j^{th} column of $\mathbf{AB}$ is dot-product of i^{th} row of $\mathbf{A}$, j^{th} column of $\mathbf{B}$:

$$\begin{aligned}
\mathbf{AB} &= \begin{pmatrix} 7 & 5 \\ 5 & 2 \\ 3 & 2 \end{pmatrix} \begin{pmatrix} 4 & 2 & 1 \\ 3 & 1 & 1 \end{pmatrix} \\
\mathbf{AB} &= \begin{pmatrix} 7\cdot 4 + 5\cdot 3, & 7\cdot 2 + 5\cdot 1, & 7\cdot 1 + 5\cdot 1 \\ 5\cdot 4 + 2\cdot 3, & 5\cdot 2 + 2\cdot 1, & 5\cdot 1 + 2\cdot 1 \\ 3\cdot 4 + 2\cdot 3, & 3\cdot 2 + 2\cdot 1, & 3\cdot 1 + 2\cdot 1 \end{pmatrix} \\
&= \begin{pmatrix} 43 & 19 & 12 \\ 26 & 12 & 7 \\ 18 & 8 & 5 \end{pmatrix}
\end{aligned}$$

(The product **DE** is not necessarily equal to **ED**). The matrices **D** and **E** are conformable for the product **DE** if **D** has the same number of columns as **E** has rows. Note that in the product of **AB** of the matrices given above, $\mathbf{A}(3 \times 2)$ and $\mathbf{B}(2 \times 3)$ are conformable.

I is reserved for the *identity* matrix, which is *square*, *symmetric*, *diagonal* with $1's$ along the diagonal and $0's$ elsewhere:

$$\mathbf{I}_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Multiplication of any (conformable) matrix **M** by **I** gives **M**: $\mathbf{AI}_3 = \mathbf{A} = \mathbf{I}_2\mathbf{A}$

4. Inversion. The *inverse* $\mathbf{M}^{-1}$ of a *square* $(r \times r)$ matrix **M**, if it exists, satisfies $\mathbf{MM}^{-1} = \mathbf{I}_r$ (similar to the reciprocal of real number). A square matrix with an inverse is called *non-singular*.
   Inversion can be computationally challenging, but not for (2×2) case:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

   Find $\mathbf{C}^{-1}$.

$$\mathbf{C}^{-1} = \frac{1}{2} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$$

   The *rank* of a matrix is equal to the number of *linearly independent* rows or columns of the matrix. Vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are linearly independent if $\sum_i a_i \mathbf{x}_i = 0$ implies $a_1 = a_2 = \cdots = a_n = 0$.

   Matrix uses: model statements, systems of linear equations, covariance matrices of random vectors,....

   Consider two lines $y_1 = 5 - x, y_2 = 3 + x$. Do these lines intersect? Where? Write this system of two equations in two unknowns using a matrix:

$$\mathbf{C} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 5 \\ 3 \end{pmatrix}$$

   (left-multiply both sides by $\mathbf{C}^{-1}$)

$$\begin{pmatrix} x \\ y \end{pmatrix} = \mathbf{C}^{-1} \begin{pmatrix} 5 \\ 3 \end{pmatrix} = \begin{pmatrix} 1 \\ 4 \end{pmatrix}$$

   The lines intersect at the solution, $(x = 1, y = 4)$.

Matrices are cool and very useful!

If we have a random vector (just like a random variable but in vector form, i.e. components yield numeric answers that are random), call it $\mathbf{Y}$, and a constant vector, call it $\mathbf{a}$, then

$$E(\mathbf{a}'\mathbf{Y}) = \mathbf{a}'E(\mathbf{Y})$$

$$\text{Var}(\mathbf{a}'\mathbf{Y}) = \mathbf{a}'\text{Var}(\mathbf{Y})\mathbf{a}$$

Understanding matrices if very important as this is how we will look at our models for much of the rest of the class. Also, SAS and other statistical programs use matrices in their calculations and in their output.

Matrix formulation of MLR

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}$$

All of the response RVs are placed into the **response vector**:

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

For observation i we can group all of the explanatory variables into a vector

$$\mathbf{x}_i = (1, x_{i1}, x_{i2}, x_{i3}, \ldots, x_{ip}).$$

The 1 in the first spot of the vector is for the intercept. If we 'stack' these row vectors on top of each other we can make a matrix called the **design matrix**:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}$$

We also form a column vector corresponding to the regression parameters, called the **'beta vector'**:

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

and a column vector for the error terms, called the **error vector**:

$$\mathbf{E} = \begin{pmatrix} E_0 \\ E_1 \\ \vdots \\ E_n \end{pmatrix}$$

Now we can see that our MLR model (a system of n equations with $p+1$ unknowns)

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} + E_1 \\ Y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} + E_2 \\ \vdots &= \vdots \\ Y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} + E_n \end{aligned}$$

can be easily rewritten as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}$$

Our assumptions on the errors can now be specified as $\mathbf{E} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$ (multivariate normal distribution). $\sigma^2\mathbf{I}_n$ is called the variance-covariance matrix:

$$Var(\mathbf{E}) = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix}$$

The diagonals of the matrix give the variances for the E_i's $(Var(E_1), Var(E_2), \dots, Var(E_n))$ and the off-diagonals (say row i column j) give the covariances between E_i's and the E_j's $(Cov(E_i, E_j))$. As the off-diagonals are all 0, our errors are uncorrelated (which for the multivariate normal distribution implies independence).

Let's look at some of these quantities for our adsorption example. We have $n = 13$ and $p = 2$.

$$\mathbf{y} = \begin{pmatrix} 4 \\ 18 \\ 14 \\ 18 \\ 26 \\ 26 \\ 21 \\ 30 \\ 28 \\ 36 \\ 65 \\ 62 \\ 40 \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & 13 & 61 \\ 1 & 21 & 175 \\ 1 & 24 & 111 \\ 1 & 23 & 124 \\ 1 & 64 & 130 \\ 1 & 38 & 173 \\ 1 & 33 & 169 \\ 1 & 61 & 169 \\ 1 & 39 & 160 \\ 1 & 71 & 244 \\ 1 & 112 & 257 \\ 1 & 88 & 333 \\ 1 & 54 & 199 \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

The predicted values can be written as

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

and residuals as

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

- $\hat{\mathbf{y}}$ is called the vector of fitted or predicted values
- $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}$ is called the hat matrix as it 'places' the hat on $\mathbf{y}$
- $\mathbf{e}$ is the vector of residuals

We will still use least squares to select the parameters, which can be written as the minimum of:

$$SS(E) = \sum_{i=1}^{n}(obs_i - pred_i)^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2 = \mathbf{e}'\mathbf{e}$$

For the adsorption example, many of these matrices are given below:

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 13 & 641 & 2305 \\ 641 & 41831 & 133162 \\ 2305 & 133162 & 467669 \end{pmatrix} \quad (\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 0.633138 & 0.002477 & -0.003826 \\ 0.002477 & 0.000265 & -0.000088 \\ -0.003826 & -0.000088 & 0.000046 \end{pmatrix}$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{pmatrix} -7.3507 \\ 0.3490 \\ 0.1127 \end{pmatrix} \quad \hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \begin{pmatrix} 4.0610 \\ 19.7008 \\ 13.5350 \\ 14.6511 \\ 29.6363 \\ 25.4084 \\ 23.2126 \\ 32.9846 \\ 24.2923 \\ 44.9271 \\ 60.7012 \\ 60.8904 \\ 33.9226 \end{pmatrix}$$

$$SS(E) = \mathbf{e}'\mathbf{e} = 191.7897 \qquad \hat{\sigma}^2 = MS(E) = SS(E)/(n-p-1) = 191.7897/10 = 19.17897$$

$$\hat{\boldsymbol{\Sigma}} = MS(E)(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 12.14294 & 0.04750 & -0.07337 \\ 0.04750 & 0.00508 & -0.00168 \\ -0.07337 & -0.00168 & 0.00088 \end{pmatrix}$$

The parameter estimates and the variance-covariance matrix are very useful for making inference about our intercept and partial slope parameters (done very similary to SLR). Let's use the above to find the following

1. What is the estimate for β_2? What is the interpretation?
2. What is the standard error of $\hat{\beta}_2$?
3. Conduct a test to determine if $\beta_2 = 0$ plausible (technically, after accounting for the linear association between extractable aluminum and adsorption index). Hint: $t(0.025, 10) = 2.228$
4. Estimate the mean adsorption index among the population of ALL soil with extractable aluminum = 100 and extractable iron = 150. Report a standard error for this estimate and a 95% confidence interval and a 95% prediction interval.

Recall that the overall hypotheses we want to test are

$$H_0 : \beta_1 = \beta_2 = 0 \quad vs \quad H_A : \text{ at least one is non-zero}$$

This is the test done in the ANOVA table given in the output from a MLR model. This is called the **global F-test** as it tests whether at least one of the terms in the model is important for predicting the response.

The ANOVA table for MLR follows the same ideas as in SLR. We are taking the total amount of variation in the response (SS(Tot)) and partitioning it into a part due to the model (SS(R)) and a part due to experimental error (SS(E)). In fact, the formulas for the sums of squares remain the same, only the degrees of freedom and the F-distribution used for finding the p-value change.

The full ANOVA table for MLR is given below:

Source	Sum of squares	df	Mean Square	F-Ratio
Regression	$SS(R)$	p	$MS(R)$	$MS(R)/MS(E)$
Error	$SS(E)$	$n-p-1$	$MS(E)$	
Total	$SS(Tot)$	$n-1$		

How to do MLR in SAS?

The following code will produce output appropriate for analysis:

```
proc reg data=adexp ;
model adsorp=aluminum iron/clb;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: adsorp

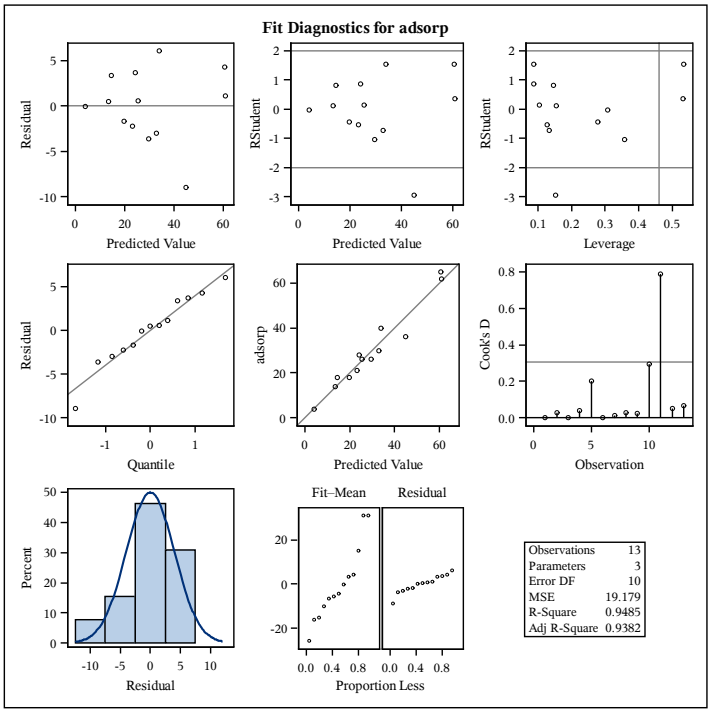
Number of Observations Read	14
Number of Observations Used	13
Number of Observations with Missing Values	1

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	3529.90308	1764.95154	92.03	<.0001
Error	10	191.78922	19.17892		
Corrected Total	12	3721.69231			

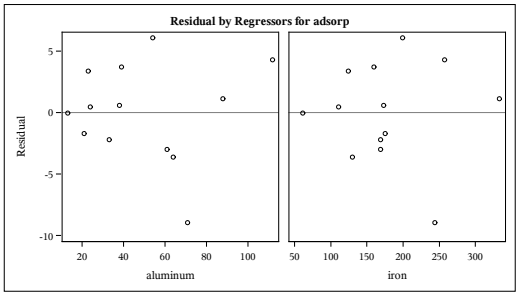
Root MSE	4.37937	**R-Square**	0.9485
Dependent Mean	29.84615	**Adj R-Sq**	0.9382
Coeff Var	14.67316		

Parameter Estimates							
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**95% Confidence Limits**	
Intercept	1	-7.35066	3.48467	-2.11	0.0611	-15.11498	0.41366
aluminum	1	0.34900	0.07131	4.89	0.0006	0.19012	0.50788
iron	1	0.11273	0.02969	3.80	0.0035	0.04658	0.17889

The REG Procedure
Model: MODEL1
Dependent Variable: adsorp



The REG Procedure
Model: MODEL1
Dependent Variable: adsorp



A non-additive model example:
A random sample of students taking the same exam:

IQ	Study TIME	GRADE
105	10	75
110	12	79
120	6	68
116	13	85
122	16	91
130	8	79
114	20	98
102	15	76

Consider regressing GRADE on IQ (X_1), TIME(X_2), and TI ($X_1 * X_2$), where TI = TIME*IQ. That is, we fit the model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + E$$

```
proc reg;
model Grade = IQ Time TI;
run;
```

```
                                The SAS System                                1
                              The REG Procedure

                           Analysis of Variance

                                    Sum of           Mean
Source                      DF     Squares         Square    F Value    Pr > F

Model                        3      610.81033      203.60344      26.22    0.0043
Error                        4       31.06467        7.76617
Corrected Total              7      641.87500

                           Parameter Estimates

                     Parameter     Standard
Variable     DF       Estimate        Error  t Value  Pr > |t|
Intercept     1      72.20608     54.07278     1.34    0.2527
IQ            1      -0.13117      0.45530    -0.29    0.7876
Time          1      -4.11107      4.52430    -0.91    0.4149
TI            1       0.05307      0.03858     1.38    0.2410
```

Discussion of the interaction model:
We call the product TI = Time*IQ an "interaction" term. That is, our explanatory variables do not have an independent effect on the response.

$$\widehat{MeanGrade} = 72.21 - 0.13 * IQ - 4.11 * Time + 0.0531 * TI$$

Now if IQ = 100 we get

$$\widehat{MeanGrade} = (72.21 - 13.1) + (-4.11 + 5.31) * Time$$

and if IQ 120 we get

$$\widehat{MeanGrade} = (72.21 - 15.7) + (-4.11 + 6.37) * Time.$$

Thus we expect an extra hour of study to increase the grade by 1.20 points for someone with IQ = 100 and by 2.26 points for someone with IQ = 120 if we use this interaction model.

Generally, we can interpret the (true) β parameters in the model as:

- β_0 - Average value of Grade when IQ and Study Time are 0
- β_1 - Average change in Grade for a unit increase in IQ when Study Time is 0
- β_2 - Average change in Grade for a unit increase in Study Time when IQ is 0
- β_3 - Average change in the slope for IQ (or Study Time) for a given value of Study Time (or IQ).

The interpretation of the interaction 'slope' can be seen by looking at the following:

$$\mu(x_1+1, x_2) - \mu(x_1, x_2) = \beta_0 + \beta_1(x_1+1) + \beta_2 x_2 + \beta_3(x_1+1)(x_2) - \beta_0 - \beta_1 x_1 - \beta_2 x_2 - \beta_3 x_1(x_2)$$

$$= \beta_1 + \beta_3 x_2$$

So β_3 is the amount the slope for x_1 changes per unit change in x_1 while x_2 is held constant.

Note: The global p-value is significant, but none of our individual terms are. This gives evidence that our model is over-fit. we may want to go back to the simpler "main effects" model.

Model Selection:

x_1, x_2, x_3 denote p independent variables. Consider several models:

1. $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_1 x_1$
2. $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_2 x_2$
3. $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_3 x_3$
4. $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
5. $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_3 x_3$
6. $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
7. $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_2 x_2 + \beta_3 x_3$

A is nested in B means model A can be obtained by restricting (e.g. setting to 0) parameter values in model B.
True or false:

- Model 1 nested in Model 4
- Model 1 nested in Model 5
- Model 2 nested in Model 4
- Model 4 nested in Model 1
- Model 3 nested in Model 4
- Model 5 nested in Model 4
- Model 3 nested in Model 7
- Model 1 nested in Model 7

A nested in $B \longrightarrow A$ called *reduced model*, B called *full model*.

p - number of regression parameters in full model
q - number of regression parameters in reduced model
$p - q$ - number of regression parameters being tested.

Let's get a handle on this notation. Give the extra regression SS terms for comparing some of the nested models on preceding page:

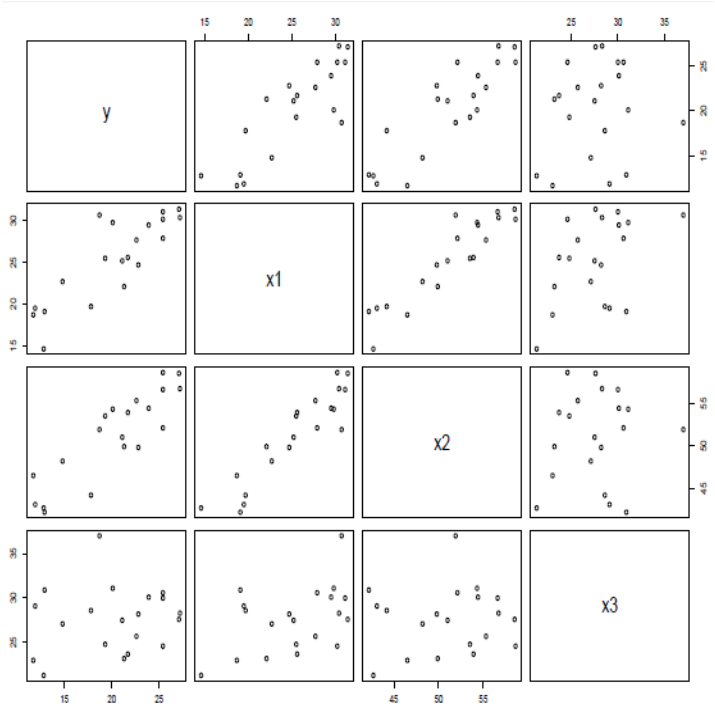
- Model 1 in model 4: $R(\beta_2, \beta_3|\beta_1)$
- Model 2 in model 4:
- Model 3 in model 4:
- Model 1 in model 5: $R(\beta_3|\beta_1)$
- Model 5 in model 4:

An example: How to measure body fat?

For each of $n = 20$ healthy individuals, the following measurements were made: bodyfat percentage y_i, triceps skinfold thickness, x_1, thigh circumference x_2, midarm circumference x_3.

x1	x2	x3	y
19.5	43.1	29.1	11.9
24.7	49.8	28.2	22.8
30.7	51.9	37.0	18.7
29.8	54.3	31.1	20.1
19.1	42.2	30.9	12.9
25.6	53.9	23.7	21.7
31.4	58.5	27.6	27.1
27.9	52.1	30.6	25.4
22.1	49.9	23.2	21.3
25.5	53.5	24.8	19.3
31.1	56.6	30.0	25.4
30.4	56.7	28.3	27.2
18.7	46.5	23.0	11.7
19.7	44.2	28.6	17.8
14.6	42.7	21.3	12.8
29.5	54.4	30.1	23.9
27.7	55.3	25.7	22.6
30.2	58.6	24.6	25.4
22.7	48.2	27.1	14.8
25.2	51.0	27.5	21.1

```
ods graphics on;
proc corr plots=matrix;
var y x1 x2 x3;
run;
```



Pearson Correlation Coefficients, N = 20
Prob > |r| under H0: Rho=0

	y	x1	x2	x3
y	1.00000	0.84327	0.87809	0.14244
		<.0001	<.0001	0.5491
x1	0.84327	1.00000	0.92384	0.45778
	<.0001		<.0001	0.0424
x2	0.87809	0.92384	1.00000	0.08467
	<.0001	<.0001		0.7227
x3	0.14244	0.45778	0.08467	1.00000
	0.5491	0.0424	0.7227	

Looking at the scatter plots and the correlation output, marginal associations between y and x_1 and between y and x_2 are highly significant, providing evidence of a strong $r \approx 0.85$ linear association between average bodyfat and triceps skinfold and between average bodyfat and thigh circumference.

Notice the scatter plot between x_1 and x_2, there is a strong linear relationship. This means that triceps skinfold and thigh circumference are giving some of the same information. This can lead to issues when fitting a model.

Multicollinearity: linear associations among the independent variables; causes problems such as inflated sampling variances for $\hat{\boldsymbol{\beta}}$.

```
proc reg data=bodyfat;
   model y=x1/covb;
   model y=x2/covb;
   model y=x3/covb;
   model y=x1 x2/covb;
   model y=x1 x2 x3/covb;
run;
```

Yields the following output:

Output From Proc Reg for Bodyfat Example

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read	21
Number of Observations Used	20
Number of Observations with Missing Values	1

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	352.26980	352.26980	44.30	<.0001
Error	18	143.11970	7.95109		
Corrected Total	19	495.38950			

Root MSE	2.81977	R-Square	0.7111
Dependent Mean	20.19500	Adj R-Sq	0.6950
Coeff Var	13.96271		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-1.49610	3.31923	-0.45	0.6576
x1	1	0.85719	0.12878	6.66	<.0001

Covariance of Estimates		
Variable	Intercept	x1
Intercept	11.01731839	-0.419670565
x1	-0.419670565	0.0165844918

Output From Proc Reg for Bodyfat Example

The REG Procedure
Model: MODEL2
Dependent Variable: y

Number of Observations Read	21
Number of Observations Used	20
Number of Observations with Missing Values	1

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	381.96582	381.96582	60.62	<.0001
Error	18	113.42368	6.30132		
Corrected Total	19	495.38950			

Root MSE	2.51024	R-Square	0.7710
Dependent Mean	20.19500	Adj R-Sq	0.7583
Coeff Var	12.43002		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-23.63449	5.65741	-4.18	0.0006
x2	1	0.85655	0.11002	7.79	<.0001

Covariance of Estimates		
Variable	Intercept	x2
Intercept	32.006329324	-0.619332881
x2	-0.619332881	0.0121034372

Output From Proc Reg for Bodyfat Example

The REG Procedure
Model: MODEL3
Dependent Variable: y

Number of Observations Read	21
Number of Observations Used	20
Number of Observations with Missing Values	1

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	10.05160	10.05160	0.37	0.5491
Error	18	485.33790	26.96322		
Corrected Total	19	495.38950			

Root MSE	5.19261	R-Square	0.0203
Dependent Mean	20.19500	Adj R-Sq	-0.0341
Coeff Var	25.71236		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	14.68678	9.09593	1.61	0.1238
x3	1	0.19943	0.32663	0.61	0.5491

Covariance of Estimates		
Variable	Intercept	x3
Intercept	82.735867956	-2.946694682
x3	-2.946694682	0.1066869907

Output From Proc Reg for Bodyfat Example

The REG Procedure
Model: MODEL4
Dependent Variable: y

Number of Observations Read	21
Number of Observations Used	20
Number of Observations with Missing Values	1

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	385.43871	192.71935	29.80	<.0001
Error	17	109.95079	6.46769		
Corrected Total	19	495.38950			

Root MSE	2.54317	R-Square	0.7781
Dependent Mean	20.19500	Adj R-Sq	0.7519
Coeff Var	12.59305		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-19.17425	8.36064	-2.29	0.0348
x1	1	0.22235	0.30344	0.73	0.4737
x2	1	0.65942	0.29119	2.26	0.0369

Covariance of Estimates			
Variable	Intercept	x1	x2
Intercept	69.900312587	1.8469661215	-2.273097628
x1	1.8469661215	0.0920751757	-0.081628463
x2	-2.273097628	-0.081628463	0.0847900309

The REG Procedure
Model: MODEL5
Dependent Variable: y

Number of Observations Read	21
Number of Observations Used	20
Number of Observations with Missing Values	1

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	396.98461	132.32820	21.52	<.0001
Error	16	98.40489	6.15031		
Corrected Total	19	495.38950			

Root MSE	2.47998	R-Square	0.8014
Dependent Mean	20.19500	Adj R-Sq	0.7641
Coeff Var	12.28017		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	117.08469	99.78240	1.17	0.2578
x1	1	4.33409	3.01551	1.44	0.1699
x2	1	-2.85685	2.58202	-1.11	0.2849
x3	1	-2.18606	1.59550	-1.37	0.1896

Covariance of Estimates

Variable	Intercept	x1	x2	x3
Intercept	9956.5279384	300.1979628	-257.3823153	-158.6704127
x1	300.1979628	9.0933087788	-7.779145105	-4.7880263
x2	-257.3823153	-7.779145105	6.6668028532	4.0946155019
x3	-158.6704127	-4.7880263	4.0946155019	2.545617053

Question: Why is the global p-value in the last model significant, i.e. at least one predictor is useful, but the individual tests are all nonsignificant?

In the bodyfat data, consider comparing the simple model that Y depends only on x_1 (triceps) versus the full model that it depends on all three.

$$\begin{aligned} \text{Model } A: \mu(x_1, x_2, x_3) &= \beta_0 + \beta_1 x_1 \\ \text{Model } B: \mu(x_1, x_2, x_3) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \end{aligned}$$

or the null hypothesis

$$H_0: \beta_2 = \beta_3 = 0 \quad \text{vs} \quad H_1: \beta_2, \beta_3 \text{ not both } 0$$

after accounting for x_1. Our F statistic can be used

$$F = \frac{(396.9 - 352.3)/2}{6.15} = \frac{22.3}{6.15} = 3.64$$

How many *df* for numerator and denominator?

The 95^{th} percentile is $F(0.05, \quad , \quad) = 3.63$.

Our conclusion about the hypotheses?

That is, after accounting for the linear dependence between triceps and bodyfat, there is still some linear association between mean bodyfat and at least one of x_2, x_3 (thigh,midarm).

To get the nested model F-ratio in SAS:

```
proc reg data=bodyfat;
     model y=x1 x2 x3;
     test x2=0,x3=0;
run;
```

Full mode vs only Triceps 4

The REG Procedure
Model: MODEL1

Test 1 Results for Dependent Variable y

Source	DF	Mean Square	F Value	Pr > F
Numerator	2	22.35741	3.64	0.0500
Denominator	16	6.15031		

However, we saw in the previous output that a model with all three variables is no good. This is due to the multicollinearity. We will now very briefly look at a few automated model selection techniques.

Using proc reg to perform variable selection:
We'll discuss three hypothesis testing methods for selecting variables (there are many other ways to accomplish this we won't discuss).

1. **Forward Selection** - Start with nothing and work forward.
   (a) Begin with a model with only β_0
   (b) Calculate $R(\beta_i|\beta_0)$ for all possible predictors and find p-values for each
   (c) Take most significant p-value less than a cutoff (say 0.3), add predictor into model.
   (d) Say β_j was added in the last step, repeat above process with added predictor. That is, calculate $R(\beta_i|\beta_0, \beta_j)$ for all other predictors, etc.
   (e) Stop when no predictors are below the cutoff or if the full model is selected.
2. **Backward Selection** - Start with everything and work backward.
   (a) Start with full model.
   (b) Locate variable with largest p-value greater than a cutoff (say 0.1), remove that variable.
   (c) Repeat until all p-values are less than the cut off or the null model (intercept only model) is chosen.
3. **Subset Selection** - Compute all possible models, pick best.
   (a) Compare each of the models using a criterion.
   (b) Choose model that minimizes that criterion. Possible criteria include:
      - *Adjusted* $R^2 = 1 - \frac{n-1}{n-p-1}(1 - R^2)$ (takes into account the addition of more predictors)
      - Mallow's C_P, AIC, AICc, or BIC (all take into account the model complexity, not just how well the model fits the data)

How to do these model selection methods in SAS?

```
proc reg data=bodyfat plots=none;
    model y=x1 x2 x3/selection=cp ;
    model y=x1 x2 x3/selection=forward SLentry=0.3;
    model y=x1 x2 x3/selection=backward SLstay=0.1;
   model y=x1 x2 x3/selection=adjrsq;
run;
```

Variable Selection Methods on Bodyfat Example

The REG Procedure
Model: MODEL1
Dependent Variable: y

C(p) Selection Method

Number of Observations Read	21
Number of Observations Used	20
Number of Observations with Missing Values	1

Number in Model	C(p)	R-Square	Variables in Model
1	2.4420	0.7710	x2
2	3.2242	0.7862	x1 x3
2	3.8773	0.7781	x1 x2
3	4.0000	0.8014	x1 x2 x3
2	4.0657	0.7757	x2 x3
1	7.2703	0.7111	x1
1	62.9128	0.0203	x3

Variable Selection Methods on Bodyfat Example

The REG Procedure
Model: MODEL2
Dependent Variable: y

Number of Observations Read	21
Number of Observations Used	20
Number of Observations with Missing Values	1

Forward Selection: Step 1

Variable x2 Entered: R-Square = 0.7710 and C(p) = 2.4420

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	381.96582	381.96582	60.62	<.0001
Error	18	113.42368	6.30132		
Corrected Total	19	495.38950			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-23.63449	5.65741	109.97344	17.45	0.0006
x2	0.85655	0.11002	381.96582	60.62	<.0001

Bounds on condition number: 1, 1

No other variable met the 0.3000 significance level for entry into the model.

Summary of Forward Selection							
Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x2	1	0.7710	0.7710	2.4420	60.62	<.0001

Variable Selection Methods on Bodyfat Example

The REG Procedure
Model: MODEL3
Dependent Variable: y

Number of Observations Read	21
Number of Observations Used	20
Number of Observations with Missing Values	1

Backward Elimination: Step 0

All Variables Entered: R-Square = 0.8014 and C(p) = 4.0000

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	396.98461	132.32820	21.52	<.0001
Error	16	98.40489	6.15031		
Corrected Total	19	495.38950			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	117.08469	99.78240	8.46816	1.38	0.2578
x1	4.33409	3.01551	12.70489	2.07	0.1699
x2	-2.85685	2.58202	7.52928	1.22	0.2849
x3	-2.18606	1.59550	11.54590	1.88	0.1896

Bounds on condition number: 708.84, 4133.4

Backward Elimination: Step 1

Variable x2 Removed: R-Square = 0.7862 and C(p) = 3.2242

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	389.45533	194.72767	31.25	<.0001
Error	17	105.93417	6.23142		
Corrected Total	19	495.38950			

Variable Selection Methods on Bodyfat Example

The REG Procedure
Model: MODEL3
Dependent Variable: y

Backward Elimination: Step 1

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	6.79163	4.48829	14.26834	2.29	0.1486
x1	1.00058	0.12823	379.40373	60.89	<.0001
x3	-0.43144	0.17662	37.18554	5.97	0.0258

Bounds on condition number: 1.2651, 5.0605

All variables left in the model are significant at the 0.1000 level.

Summary of Backward Elimination							
Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x2	2	0.0152	0.7862	3.2242	1.22	0.2849

The REG Procedure
Model: MODEL4
Dependent Variable: y

Adjusted R-Square Selection Method

Number of Observations Read	21
Number of Observations Used	20
Number of Observations with Missing Values	1

Number in Model	Adjusted R-Square	R-Square	Variables in Model
3	0.7641	0.8014	x1 x2 x3
2	0.7610	0.7862	x1 x3
1	0.7583	0.7710	x2
2	0.7519	0.7781	x1 x2
2	0.7493	0.7757	x2 x3
1	0.6950	0.7111	x1
1	-.0341	0.0203	x3

Types of Sums of Squares

Given that we have 4 predictors, $X_1 - X_4$ we really can have a number of tests based on nested models for $\beta_4 = 0$ (or for any other β for that matter). Let's write them down in terms of extra regression sums of squares:
$R(\beta_4|\beta_0)$ (SLR test)
$R(\beta_4|\beta_0, \beta_1)$ (test after accounting for X_1)
$R(\beta_4|\beta_0, \beta_2)$ (test after accounting for X_2)
$R(\beta_4|\beta_0, \beta_3)$ (test after accounting for X_3)
$R(\beta_4|\beta_0, \beta_1, \beta_2)$ (test after accounting for X_1 and X_2)
$R(\beta_4|\beta_0, \beta_1, \beta_3)$ (test after accounting for X_1 and X_3)
$R(\beta_4|\beta_0, \beta_2, \beta_3)$ (test after accounting for X_2 and X_3)
$R(\beta_4|\beta_0, \beta_1, \beta_2, \beta_3)$ (test after accounting for X_1, X_2, and X_3)

Some of these tests can be easily found using different types of sums of squares.

The tests given for the parameter estimates are all type III tests and this is the test usually done to determine if a slope term has significance. However, type I tests are very useful for model building. For example, if we wanted to look at building a model for the bodyfat example and we thought the order of importance for the variables was X_1 (triceps), X_3 (midarm), and X_2 (thigh), we could get sequential tests for these models using type I sums of squares.

In SAS proc reg use the following code:

```
proc reg data=bodyfat;
   model y=x1 x3 x2/ss1; *Note the order of variables is important for Type I;
run;
```

Sequential tests for bodyfat example **1**

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read	21
Number of Observations Used	20
Number of Observations with Missing Values	1

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	396.98461	132.32820	21.52	<.0001
Error	16	98.40489	6.15031		
Corrected Total	19	495.38950			

Root MSE	2.47998	**R-Square**	0.8014
Dependent Mean	20.19500	**Adj R-Sq**	0.7641
Coeff Var	12.28017		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS
Intercept	1	117.08469	99.78240	1.17	0.2578	8156.76050
x1	1	4.33409	3.01551	1.44	0.1699	352.26980
x3	1	-2.18606	1.59550	-1.37	0.1896	37.18554
x2	1	-2.85685	2.58202	-1.11	0.2849	7.52928

Let‘s label the Type I SS in terms of extra regression sums of squares (R notation).

Note: we will soon use proc glm for our model analysis and this gives even better output for type I sums of squares. (The tests given for type I sums of squares use the *full model* MS(E) rather than the full model MS(E) up to that point. This test still works because MS(E) from each model is an unbiased estimate of σ^2. The tests using the different MS(E) terms could give different results, but will usually agree.

```
proc glm data=bodyfat;
   model y=x1 x3 x2;
run;
```

Sequential tests for bodyfat example using GLM 2

The GLM Procedure

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	396.9846118	132.3282039	21.52	<.0001
Error	16	98.4048882	6.1503055		
Corrected Total	19	495.3895000			

R-Square	Coeff Var	Root MSE	y Mean
0.801359	12.28017	2.479981	20.19500

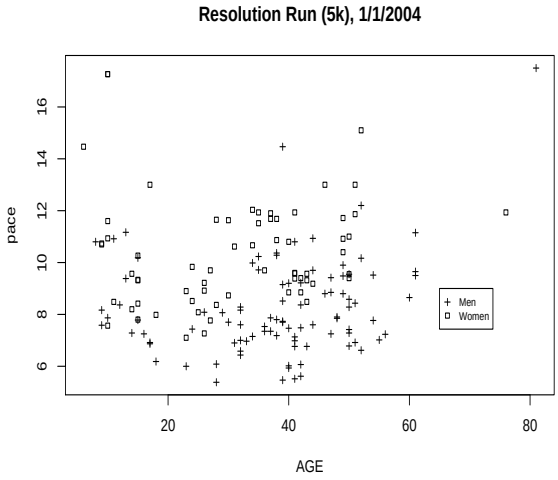
Source	DF	Type I SS	Mean Square	F Value	Pr > F
x1	1	352.2697968	352.2697968	57.28	<.0001
x3	1	37.1855371	37.1855371	6.05	0.0257
x2	1	7.5292779	7.5292779	1.22	0.2849

Source	DF	Type III SS	Mean Square	F Value	Pr > F
x1	1	12.70489278	12.70489278	2.07	0.1699
x3	1	11.54590217	11.54590217	1.88	0.1896
x2	1	7.52927788	7.52927788	1.22	0.2849

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	117.0846948	99.78240295	1.17	0.2578
x1	4.3340920	3.01551136	1.44	0.1699
x3	-2.1860603	1.59549900	-1.37	0.1896
x2	-2.8568479	2.58201527	-1.11	0.2849

A linear regression example with a quadratic explanatory variable:
Data was collected on 5 kilometer run times. The variables collected were age, sex, and pace.

	Obs	age	sex	race	pace
	1	28	M	16.6833	5.38333
	2	39	M	16.9500	5.46667
	3	41	M	17.1333	5.51667
	4	42	M	17.4000	5.61667
(abbreviated)	...	...	...	...	...
	157	52	F	46.8833	15.1000
	158	10	F	53.6000	17.2667
	159	10	F	53.6167	17.2667
	160	81	M	54.3167	17.5000



Quadratic model for pace (Y) as a function of age (x):

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + E_i \quad \text{for } i = 1, \ldots, 160$$

where $E_i \overset{iid}{\sim} N(0, \sigma^2)$.

Question: What does σ^2 represent in the model?

Question: What do the parameters mean, i.e. what is their interpretation?

We may want to compare this model with a SLR model

$$Y_i = \beta_0 + \beta_1 x_i + E_i \text{ for } i = 1, \ldots, 160$$

Question: How can we compare the two models?

```
/* age2 defined in data step as age*age */
PROC REG;
   MODEL pace=age;
   MODEL pace=age age2/ss1 covb;
RUN;
```

Model: MODEL1
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1.09650	1.09650	0.22	0.6396
Error	158	786.99821	4.98100		
Corrected Total	159	788.09472			

Root MSE	2.23182	R-Square	0.0014
Dependent Mean	9.12063	Adj R-Sq	-0.0049

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	8.92271	0.45724	19.51	<.0001
age	1	0.00564	0.01203	0.47	0.6396

Model: MODEL2
Analysis of Variance

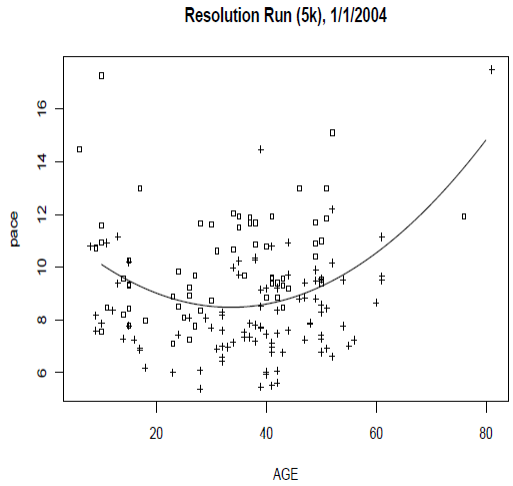
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	113.64500	56.82250	13.23	<.0001
Error	157	674.44972	4.29586		
Corrected Total	159	788.09472			

Root MSE	2.07265	R-Square	0.1442
Dependent Mean	9.12063	Adj R-Sq	0.1333

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS
Intercept	1	11.78503	0.70216	16.78	<.0001	13310
age	1	-0.19699	0.04113	-4.79	<.0001	1.09650
age2	1	0.00294	0.00057380	5.12	<.0001	112.54850

Covariance of Estimates

Variable	Intercept	age	age2
Intercept	0.4930258	-0.0265145	0.0003209
age	-0.0265145	0.0016921	-0.0000227
age2	0.0003209	-0.0000227	0.0000003



Fitted models are:

$$\text{Model 1: } \hat{\mu}(x) = 8.923 + 0.0056age$$

$$\text{Model 2: } \hat{\mu}(age) = 11.785 - 0.197age + 0.00294age^2$$

$$
\begin{aligned}
F &= \frac{R(\beta_2|\beta_0,\beta_1)}{MS(E)_{full}} \\
&= \frac{(SS(R)_{full} - SS(R)_{red})/1}{MS(E)_{full}} \\
&= \frac{(113.6 - 1.1)/1}{4.3} \\
&= \frac{(SS(E)_{red} - SS(E)_{full})/1}{MS(E)_{full}} \\
&= \frac{(787.0 - 674.4)/1}{4.3} = 26.2 \\
&= \left(\frac{\hat{\beta}_2}{SE}\right)^2 = (5.12)^2
\end{aligned}
$$

with $F(0.05, 1, 157) = 3.90$. Since $26.2 >> 3.9$, we reject that the linear model is appropriate when compared to the quadratic model. This is the same test as the t-test for age2!