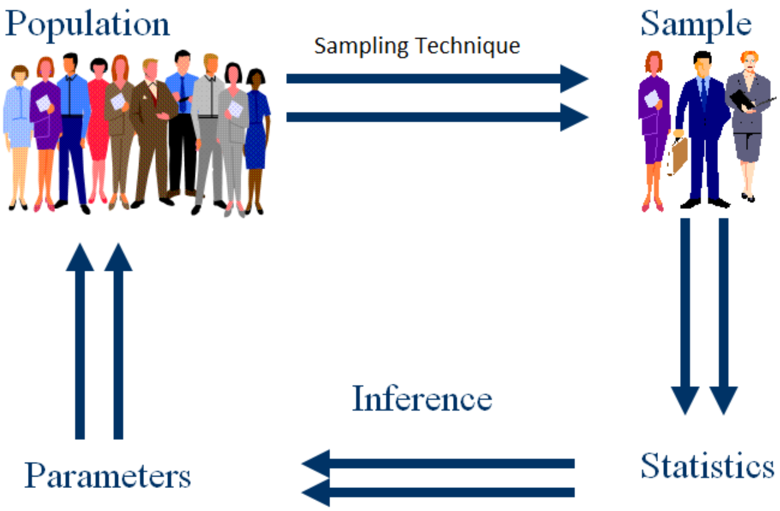
Chapter 1

ST 512 - Review

Readings: Chapters 1-8 as needed

Big ideas in stats:

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



Scales (Types) of Data:

- ______________________ - A variable that is described by attributes or labels
  Subscales:

- ______________________ - A variable that is described by numerical measurements where arithmetic can be performed
  Subscales:

Random Variables and Things of Interest:

- ______________________ - Function that takes in outcomes from an experiment and outputs real numbers, or a numeric outcome to a random process

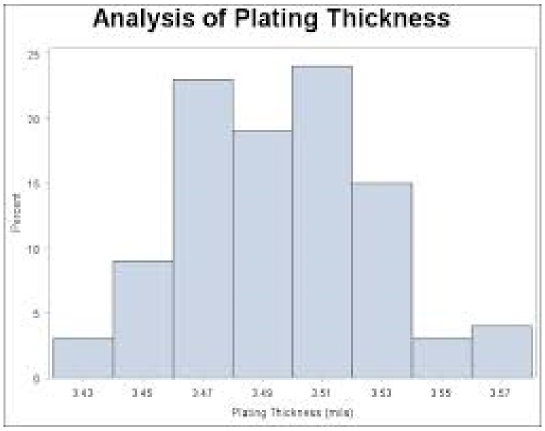
  Things of interest

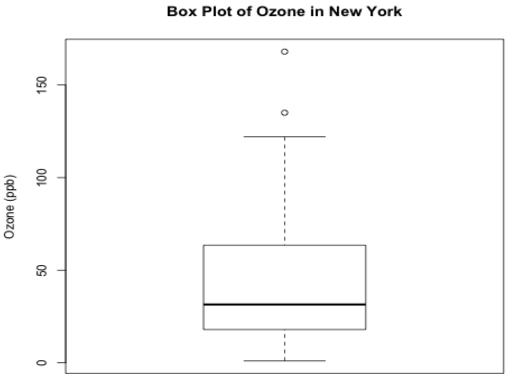
  - ______________________ - pattern and frequency of observable values
  - ______________________ - measures of center of the distribution
  - ______________________ - measures of spread for the distribution

Graphical Descriptions of RV‘s:

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

- ______________________ - Uses the Five Number Summary to display the realizations of a RV





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

Central Limit Theorem (CLT):

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

   Allows us to form a CI:

   And a test statistic:

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

   Allows us to form a CI:

   And a test statistic:

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

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

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	$nt-1$	$\sum_{i=1}^{t}\sum_{j=1}^{n}(Y_{ij}-\bar{Y}_{++})^2$			

Mention used for a completely randomized design

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

For a hypothesis test, the p-value means

For a given a null hypothesis, statistical significance implies

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

The idea of Confidence means

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.

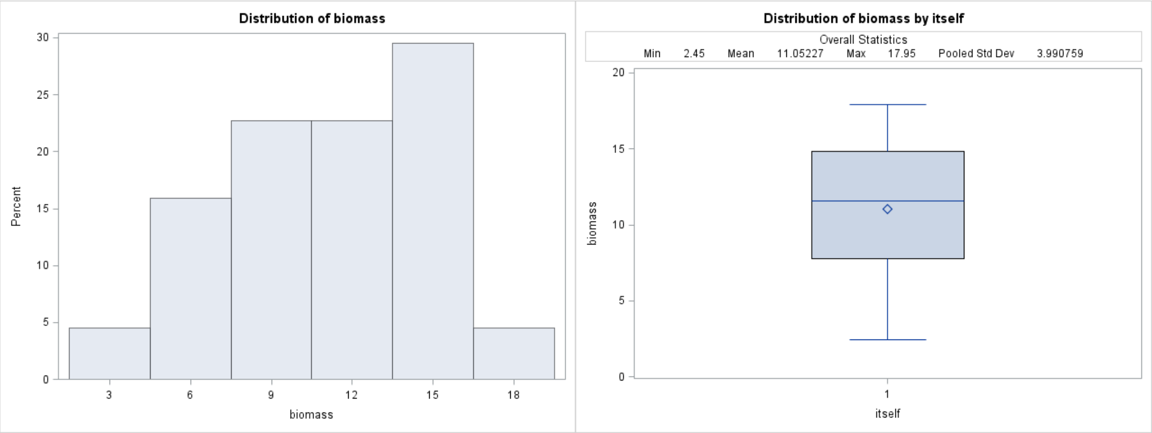
Chapter 3

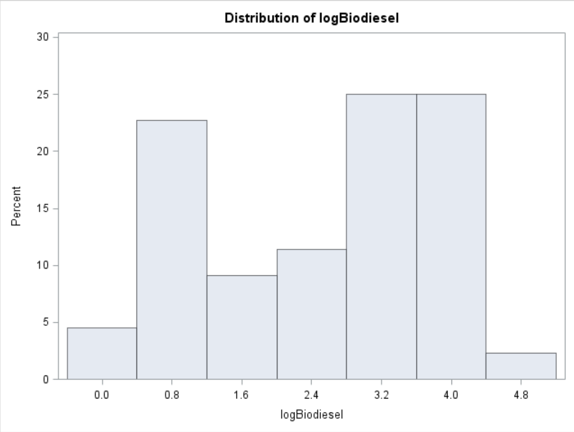
ST 512 - Correlation

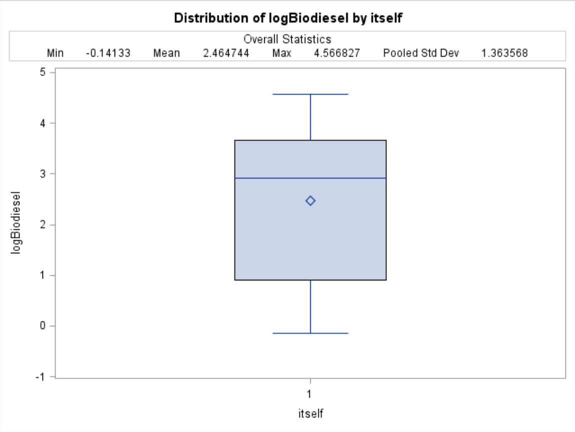
Readings for Correlation and SLR: 10.1-10.5 pg 378-420 and 10.7-10.8 pg 425-444 and 8.7 pg 305-311

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.

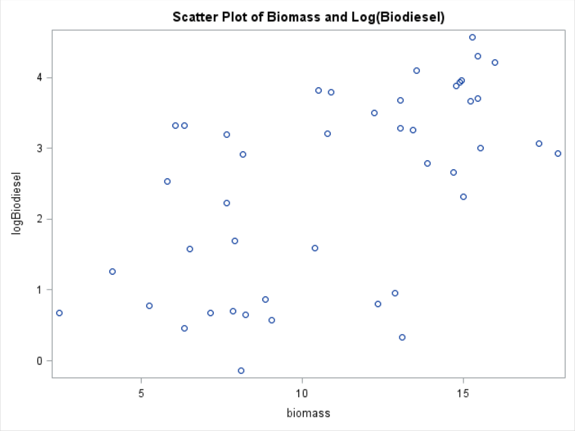
We can look at the distribution of each individually using our univariate methods (histogram, boxplot, mean/sd, etc.)

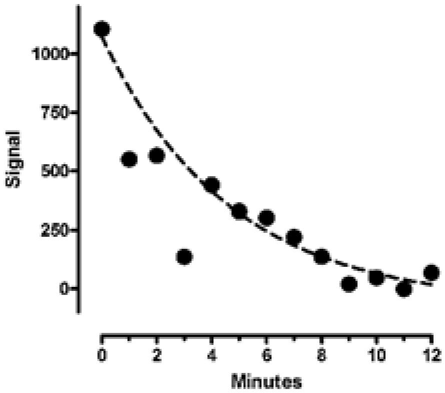






How can we visually inspect the association between the two? A **Scatter plot** gives a visual approximation of the "joint distribution" between two variables.





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:

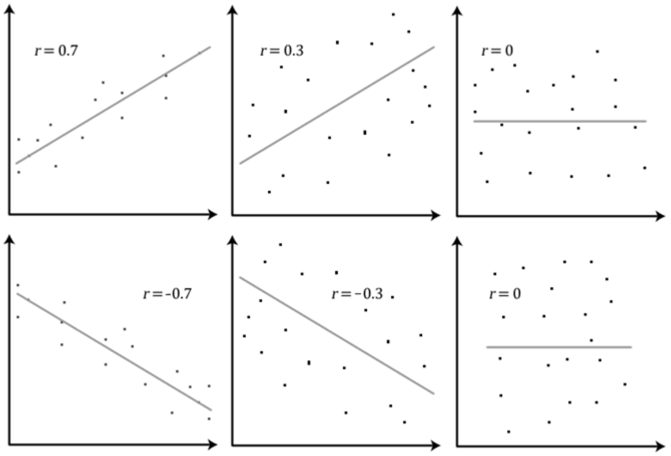
$$\bar{x} = 11.0523, \quad s_X = 3.9908, \quad \bar{y} = 2.4647, \quad s_Y = 1.3636$$

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

Applying the formula for r_{XY}, we get

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

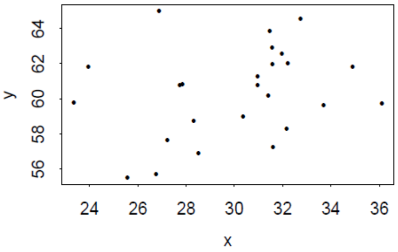
Some example scatter plots

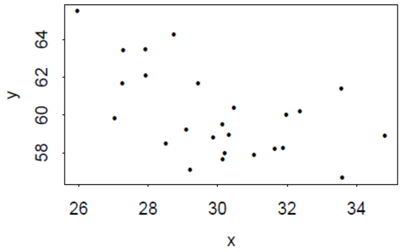


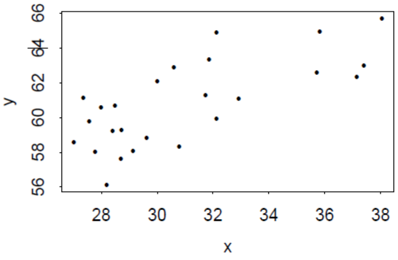
An exercise/activity:

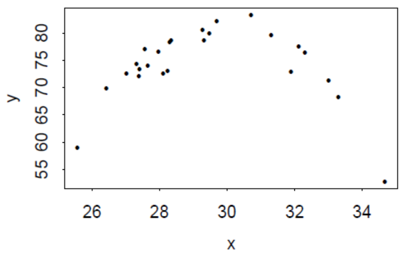
Label the four plots below with the four sample correlation coefficients:

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

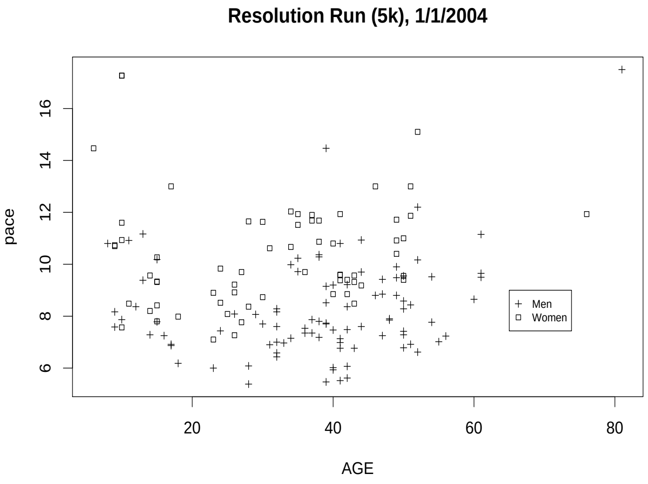








Would it be appropriate to use correlation to summarize the relationship between age and pace in the following scatter plot? Why or why not?



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

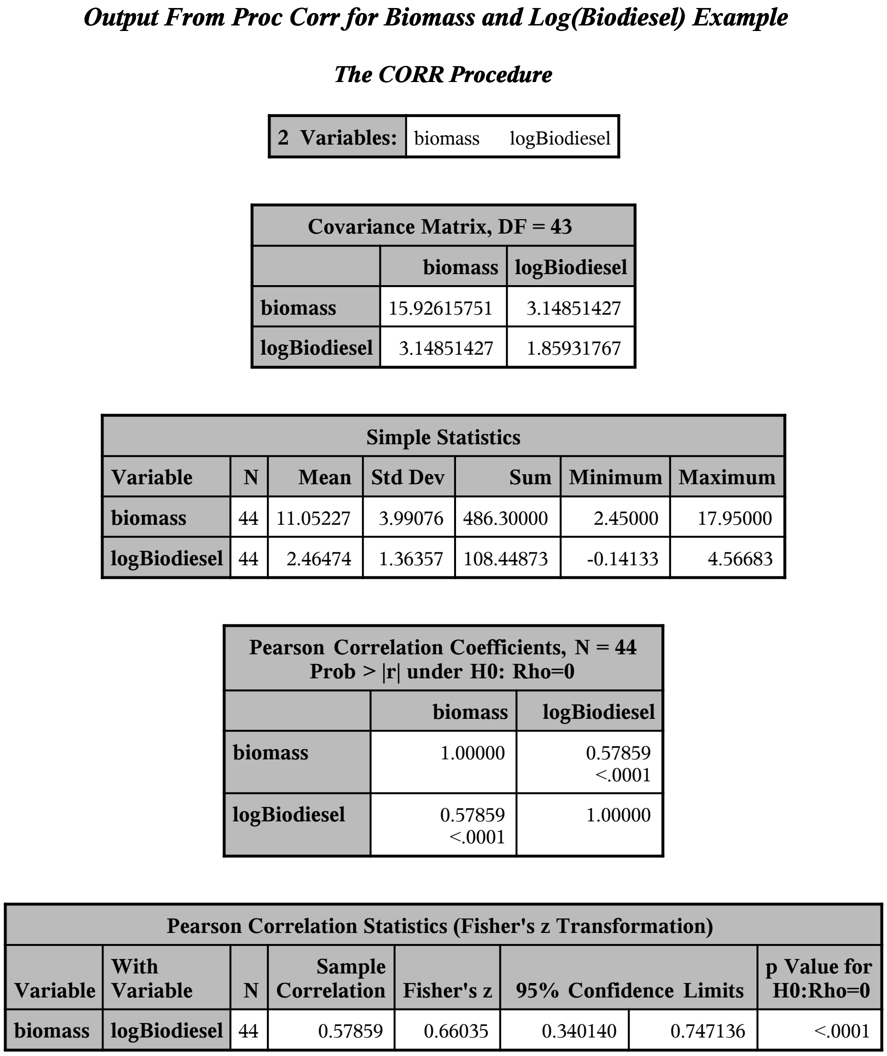
$$\left(\frac{\frac{1+0.5786}{1-0.5786}e^{-2*1.96/\sqrt{44-3}}-1}{\frac{1+0.5786}{1-0.5786}e^{-2*1.96/\sqrt{44-3}}+1},\frac{\frac{1+0.5786}{1-0.5786}e^{2*1.96/\sqrt{44-3}}-1}{\frac{1+0.5786}{1-0.5786}e^{2*1.96/\sqrt{44-3}}+1}\right)=(0.3401,0.7471)$$

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

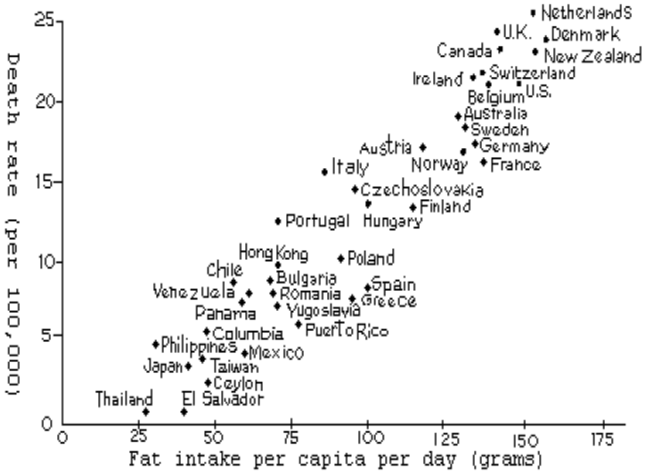
Famous examples of *spurious correlations*:

- 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. "Experimentalevidence 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

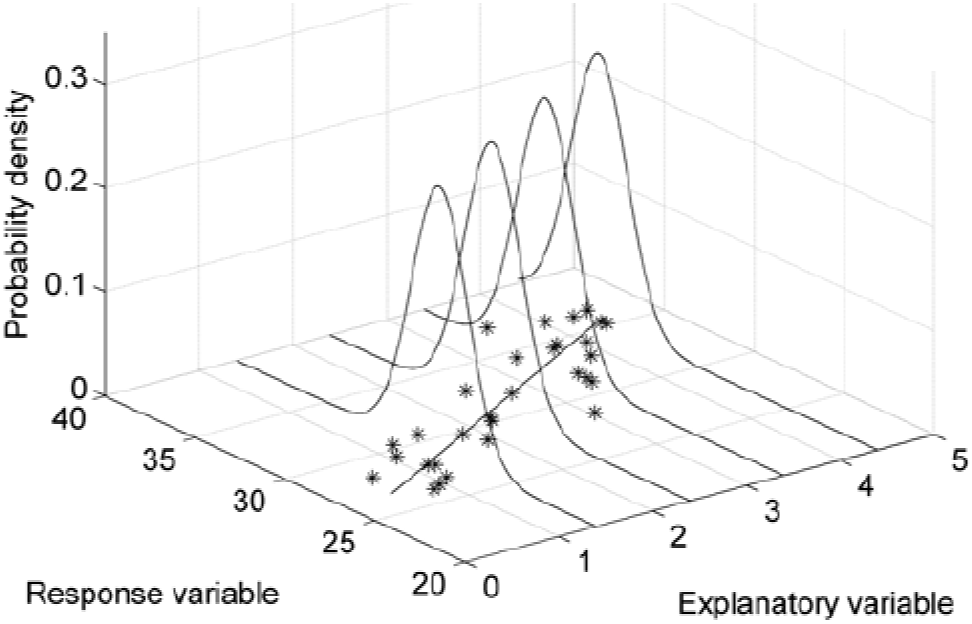
Fit a linear regression model - A probabilistic model for Y conditional on $X = x$:

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

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.

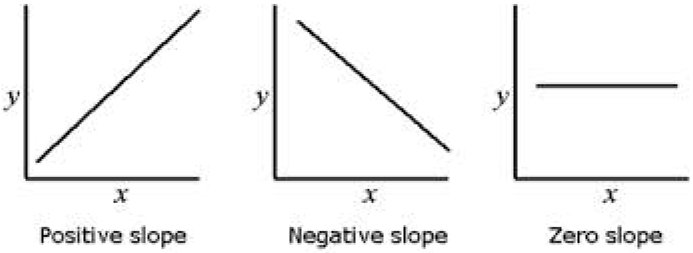
$$\hat{\beta}_1 = s_{XY}/s_X^2 = 3.1485/3.9908^2 = 0.1977$$
$$\hat{\beta}_0 = 2.4647 - 11.0523 * 0.1977 = 0.2797$$
$$\hat{y} = 0.2808 + 0.1977x$$

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 $SS(source)/df\,source$. Ratios of mean squares often follow an F-distribution and are appropriate for testing different hypotheses of interest.

In this case, to test

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

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

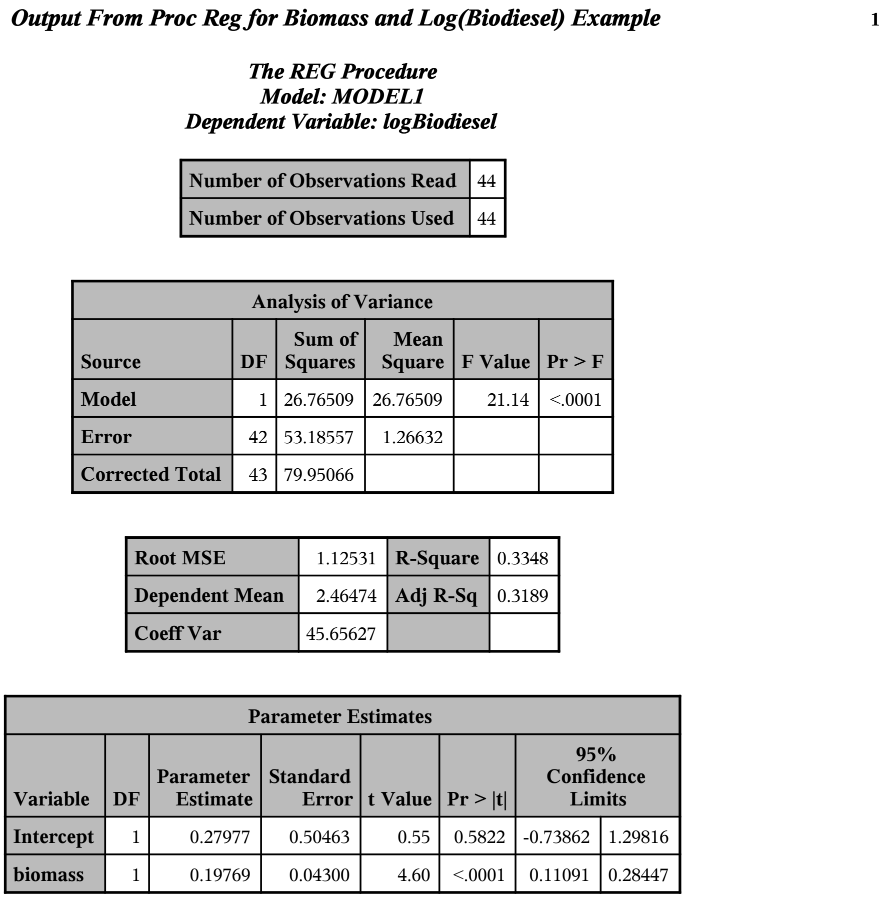
That is, the F statistic follows an F-distribution with 1 numerator df and $n-2$ denominator df. In SLR, this F test is equivalent to the T test we already looked at. The relationship is that $T^2 = F$.

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			

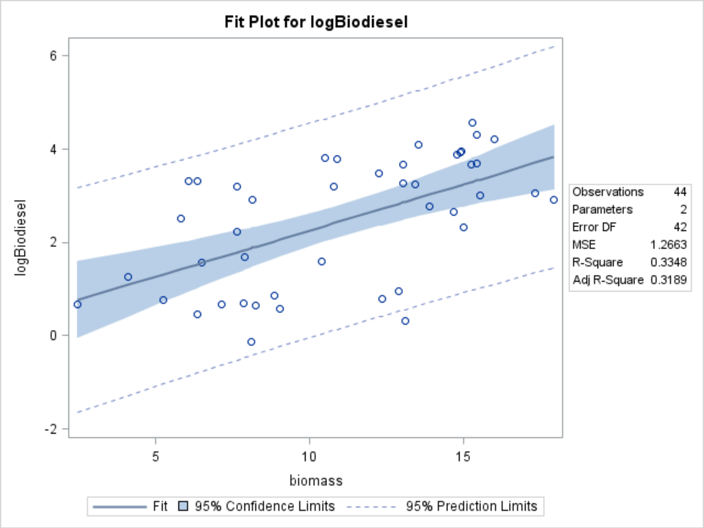
Root MSE	1.12531	**R-Square**	0.3348
Dependent Mean	2.46474	**Adj R-Sq**	0.3189
Coeff Var	45.65627		

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?

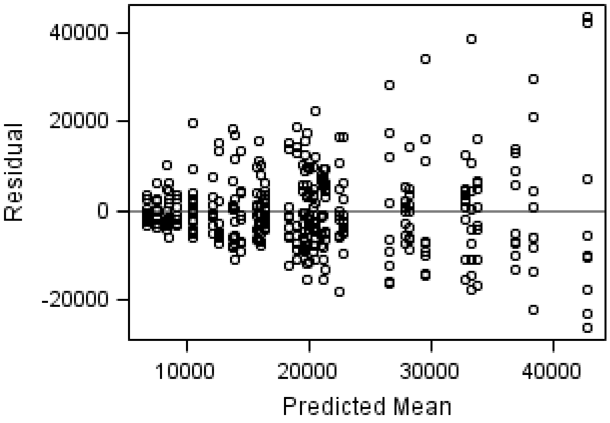


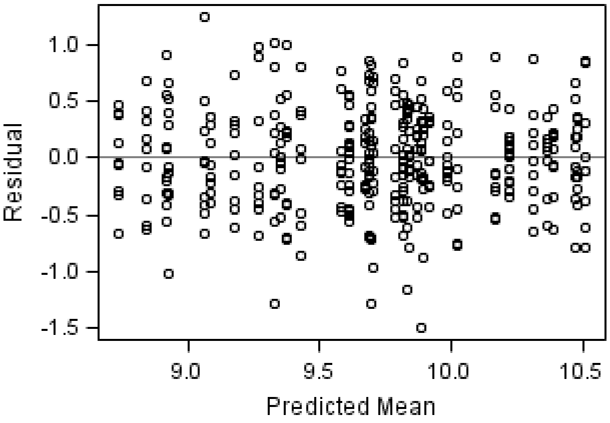
Checking assumptions

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

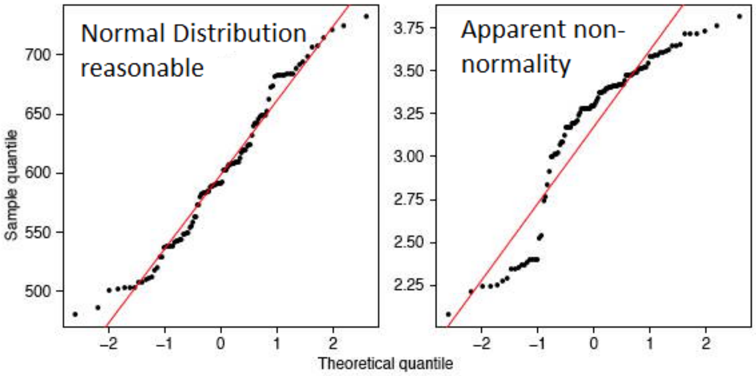
Secondly we can check our assumption of $iidN(0, \sigma^2)$ errors.

- 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).





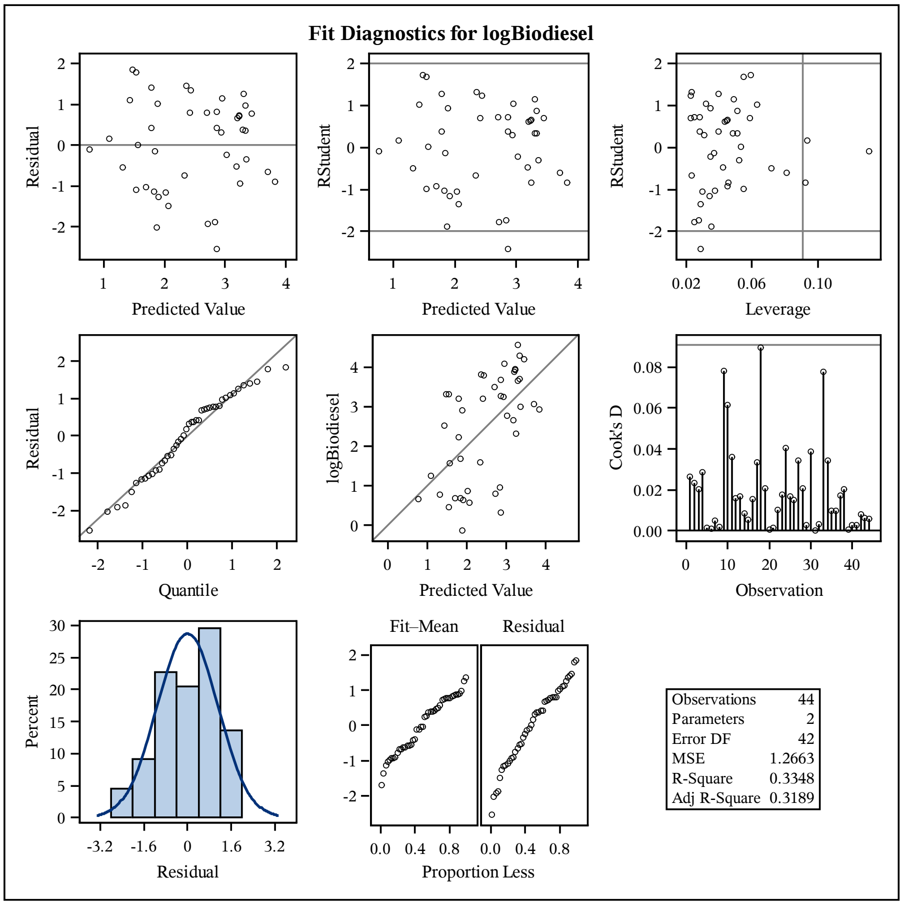
- Normality of errors - A quantile-quantile plot (or qq-plot for short) can be inspected to see if normality is reasonable.



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**

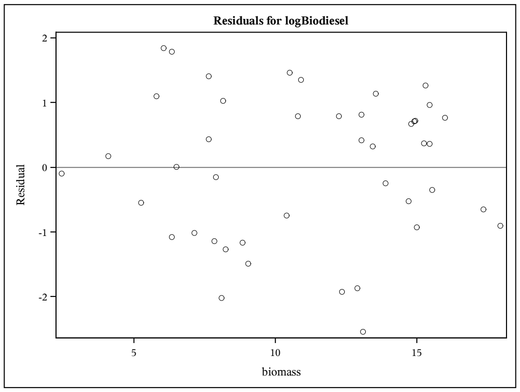
The REG Procedure
Model: MODEL1
Dependent Variable: logBiodiesel



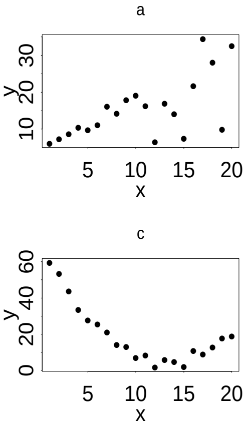
```
Observations        44
Parameters           2
Error DF            42
MSE             1.2663
R-Square        0.3348
Adj R-Square    0.3189
```

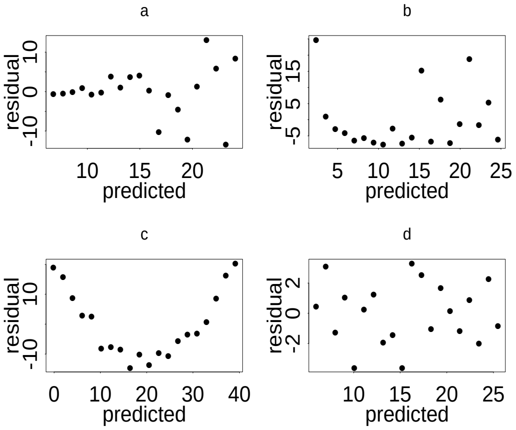
Output From Proc Reg for Biomass and Log(Biodiesel) Example 3

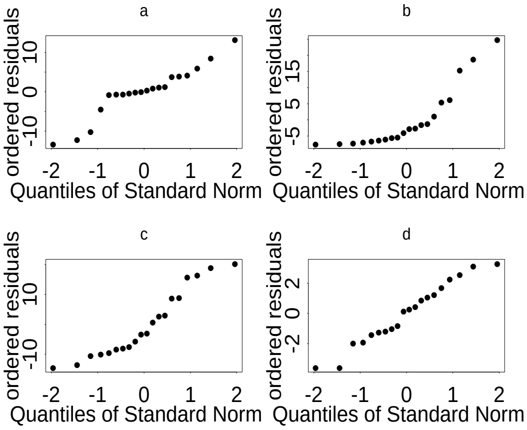
The REG Procedure
Model: MODEL1
Dependent Variable: logBiodiesel



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


a
b
c
d
y
x
5
10
15
20


a
b
c
d
residual
predicted


a
b
c
d
ordered residuals
Quantiles of Standard Norm