Extra Topics 2

1. Nonlinear Regression
2. Logistic Regression
3. Two-Sample Permutation Test
4. Survival Analysis (Kaplan-Meier Curves)
5. Multiple Linear Regression (much more in STAT512!)
6. Principal Components Analysis (PCA)
7. Closing Comments

1. Nonlinear Regression

Velocity Example: (Bates and Watts) “Velocity” of an enzymatic reaction as a function of the substrate concentration. Response variable Y = Velocity of reaction (counts/min). Predictor variable is X = substrate concentration (ppm).

Options: (1) polynomial (quadratic or cubic) regression
(2) nonlinear regression
(3) transform X, Y, or both

The Michaelis-Menten model for enzyme kinetics:

$$y_i = \frac{\beta_0 x_i}{\beta_1 + x_i} + \varepsilon_i$$

The "least squares" method: estimate β_0 and β_1 by values that minimize:

$$SS\,\mathrm{Re}\,sid = \sum_{i=1}^{n} (obs. - pred.)^2 = \sum_{i=1}^{n} \left(y_i - \frac{\beta_0 x_i}{\beta_1 + x_i} \right)^2$$

Nonlinear Regression in R

We will use the nls() function to fit non-linear regression in R.

We have 2 options:

1. Provide the formula and starting values for the parameters.
2. Use a "self-starting" nonlinear model (in this case "SSmicmen"). Available for some popular nonlinear models.

NOTES:

- The nls() procedure is iterative (requires many steps to search).
- The formula is motivated by the subject area! In biochemistry, Michaelis-Menten kinetics is one of the simplest and best-known models of enzyme kinetics.
- The starting values are based on the particular formula.

Finding Starting Values for Velocity Example

$$y_i = \frac{\beta_0 x_i}{\beta_1 + x_i} + \varepsilon_i$$

1. Let x→∞ and see that β_0 is the limit ≈ 200.

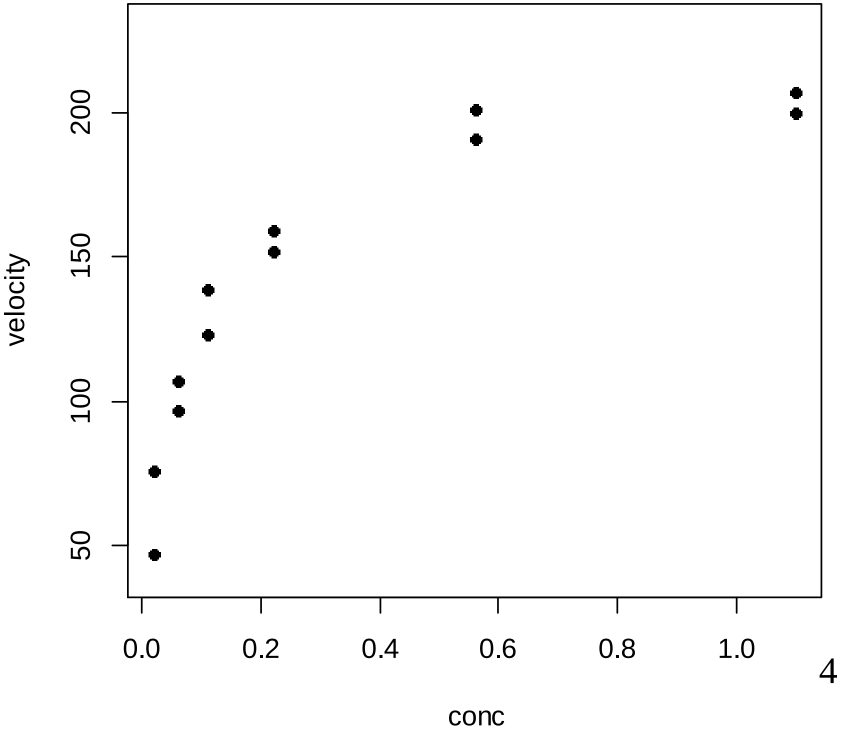
2. Pick an approximate point (x=0.06, y=100).
   Plug in to get:

$$100 = \frac{\beta_0 * 0.06}{\beta_1 + 0.06}$$

$$\Rightarrow 100 = \frac{200 * 0.06}{\beta_1 + 0.06}$$

$$\Rightarrow 100 * \beta_1 + 6 = 12$$

$$\Rightarrow \beta_1 = 6/100 = 0.06$$



Comments on Velocity Example:

1. A very good fit after 5 iterations:
R^2 = 1-(SSResid/SSTotal) =1-(1195.4/30858.9)= 0.961
Note that this is a "pseudo" R^2. It can take values outside the range of (0,1). There are also other definitions of pseudo R^2 in nonlinear regression (e.g. correlation between observed and predicted values.)

2. No obvious lack of fit in graph. Could also plot residuals versus predicted values.

3. Hypothesis tests and CI's require the assumption of normal, independent errors, with homogeneous variance. These assumptions can be assessed using standard diagnostic plots.

4. Asymptotic (accurate for large samples) CI's look fairly good, but are probably not too accurate for n=12.

A **transformably linear** model is a model that can be made into a linear model by transformation.

Velocity Example:

$$\text{Transform the equation: } y_i = \frac{\beta_0 x_i}{\beta_1 + x_i} + \varepsilon_i$$

$$\text{by inverting both sides: } \frac{1}{y_i} = \frac{\beta_1 + x_i}{\beta_0 x_i} + f_i$$

$$\frac{1}{y_i} = \frac{\beta_1}{\beta_0}\frac{1}{x_i} + \frac{1}{\beta_0} + f_i$$

A linear regression of $\frac{1}{y_i}$ on $\frac{1}{x_i}$, where $\frac{\beta_1}{\beta_0}$ is the slope and $\frac{1}{\beta_0}$ is the intercept.

Estimate the slope and intercept on the transformed scale. Then solve to get estimates of the parameters on the original scale.

Note that the two approaches yield different estimates. This will generally be the case.

Nonlinear Regression:
$\widehat{\beta_0} = 212.7$, $\widehat{\beta_1} = 0.064$

Transformation Approach:
$\widehat{\beta_0} = 195.8$, $\widehat{\beta_1} = 0.048$

Can consider whether assumptions of equal variance and normality are better satisfied for one approach over another.

2. Logistic Regression

In many research studies, the response variable may be represented as one of two possible values (0 or 1, yes or no, dead or alive). In other words, the response variable is binary.

When the response variable is binary, we are interested in probability of an “event” occurring = p = P(Y=1). We want to relate p to a linear combination of predictor variables. However, p varies between 0 and 1.

The model often used to study the association between a binary response and a set of predictor variables is **Logistic Regression**.

Simple Logistic Regression Model:

Let p(x) be the probability that y equals 1 when the predictor variable equals x.

$$\ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x$$

$$\left(\frac{p(x)}{1-p(x)}\right) = e^{(\beta_0+\beta_1 x)}$$

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

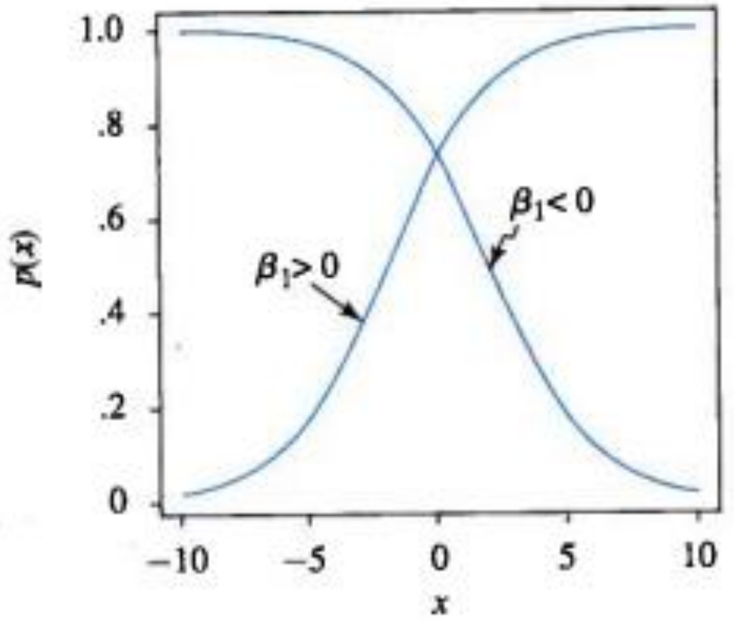


FIGURE 12.5

Logistic regression functions

Parameter Interpretation in Logistic Regression:

$$Odds = \frac{p(x)}{1-p(x)} = e^{(\beta_0+\beta_1 x)}$$

Intercept (β_0):

When x=0: $Odds = e^{(\beta_0)}$

So, when x=0 the odds of the event is a function of just β_0.

Slope (β_1):

A 1 unit increase in x gives:

$$Odds = e^{\beta_0+\beta_1(x+1)} = e^{(\beta_0+\beta_1 x)}e^{\beta_1}$$

So, a one unit increase in x multiplies the odds by e^{β_1}.

Beetle Kill Example: In a pesticide study, approximately sixty beetles were tested at 8 doses of a pesticide. A particular beetle is found to be either dead (“event”) or alive.

X = log (dose)
N = number tested at each dose
Y = number that died at each dose (out of N)
p = true probability that an individual beetle will die.

Then Y is binomial (N, p) at each dose and X is a continuous predictor variable.

This is an example of designed experiment with **grouped** data.

In R, we will use

```
glm(Y ~ X, family=binomial(link=“logit”))
```

to fit the logistic regression model.

Beetle Kill Example:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -59.282      4.995  -11.87   <2e-16 ***
lgdose        33.519      2.814   11.91   <2e-16 ***
```

Estimated “slope” is 33.519.
Interpretation: A one unit increase in lgdose (x) multiplies the odds of death by $e^{33.519} = 3.6072\text{x}10^{14}$!
Note that the range of lgdose is only 1.69 – 1.88 for this study!

A confidence interval for the odds ratio is found by exponentiating the confidence interval for the slope.
95% CI for the slope: (28.27950, 39.34206)
95% CI for the odds ratio: ($1.9126\text{x}10^{12}$, $1.2190\text{x}10^{17}$)

Notes about logistic regression:

1. Note that the responses may or may not be grouped. The glm() function can handle both formats, but be careful about formatting!

2. Parameter estimation is based on the "maximum likelihood" (ML) method, in which the parameters are estimated by the values of β_0 and β_1 that maximize the probability of observing the data that you actually did observe. The probability is calculated based on the distribution of Y (which is binomial).

3. Finding the ML estimates is an iterative procedure, but is better behaved than nonlinear regression, and doesn't require that you provide starting values.

4. Because we have "grouped" data, the estimated response can be plotted by putting the estimates into the equation.

5. Approximate tests of parameters are given in the output.
6. Approximate CIs can be found using the `confint()` function.
7. The `dose.p()` function from the `MASS` package, can be used to find the log dose that is required to achieve a given percentage mortality and construct confidence intervals. Example: LD_{50} is the log dose required to achieve 50% mortality.
8. The `glm()` function accepts categorical and continuous predictor variables.
9. The `step()` function can be used to perform stepwise logistic regression based on AIC criteria. We will discuss AIC and stepwise selection procedures in STAT512.

3. Two-Sample Permutation Test

The idea of the permutation tests for statistical inference dates back to Fisher in 1935.

Example (Higgins 2004): Suppose a company is trying to decide whether to augment its traditional instruction for new employees with computer assisted instruction. Seven new employees are selected for a trial. Four are randomly assigned to the new method and the other three are given the traditional instruction.

New: 37, 49, 55, 57
Traditional: 23, 31, 46

The mean for the New method is 49.5 and the mean for the Traditional method is 33.3. The difference is 16.2.

Suppose we are interested in testing
H_0: μ_{New}-$\mu_{Trad} \leq 0$ versus H_A: μ_{New}-$\mu_{Trad} > 0$

Using the **two-sample t-test**, the p-value is found to be 0.046. So we Reject H_0 and conclude that the New method has a higher mean response.

The two-sample t-test requires several assumptions (1) independent observations (2) normal distribution (3) equal variances for the two groups.

We have already discussed using the **two-sample Wilcoxon test** as a nonparametric alternative to the two-sample t-test. The Wilcoxon p-value is found to be 0.057.

The random assignment of the experimental subjects to the two methods provides a basis for drawing statistical inferences about the effect of the new method without the assumptions associated with the two-sample t-test.

If there is no difference between the two methods, then all data sets obtained by randomly assigning four of these scores to the New method and the other three to the Traditional method would have an equal chance of being observed. There are 35 such sets.

We can calculate a difference between means for each of these "permuted" data sets. Then compare our observed difference to this distribution. The difference between means is our test statistic.

Steps in the Two-Sample Permutation Test

1. Randomly assign experimental units to one of two treatments with *m* units in trt1 and *n* units in trt2. Obtain data and compute the difference between means, D_{obs}.
2. Permute the *m*+*n* observations between the two treatments maintaining the sample sizes. The number of possibilities is $\binom{m+n}{m} = \frac{(m+n)!}{m!n!}$
3. For each permutation of the data, compute the difference, D, between the mean for trt1 vs mean for trt2.
4. For a one-sided ("greater than") alternative, compute the p-value as the proportion of D's greater than or equal to D_{obs}: $P_{upper\ tail} = \frac{number\ of\ D's \geq D_{obs}}{\binom{m+n}{n}}$

5. Compare the p-value to your stated alpha.

Notes on the Two-Sample Permutation Test

1. The exact permutation test can be carried out using the `oneway_test()` function from the `coin` package. See "**Two Sample Permutation Test**" example.
2. We find that the permutation test is equivalent to the exact two-sample Wilcoxon test. Exact rank tests are just one type of permutation test.
3. In fact, some other test statistics (for example difference between medians) can be substituted and we will find the same result.
4. Permutation tests are sometimes called randomization tests. There is some ambiguity in these terms.
5. When sample sizes are large, you can use a random subset of permutations. The `coin` package can do this. Some people refer to this as a Monte Carlo test.

4. Survival Analysis

Survival analysis is a set of statistical methods for examining not only event *occurrence* but also the *timing* of events.

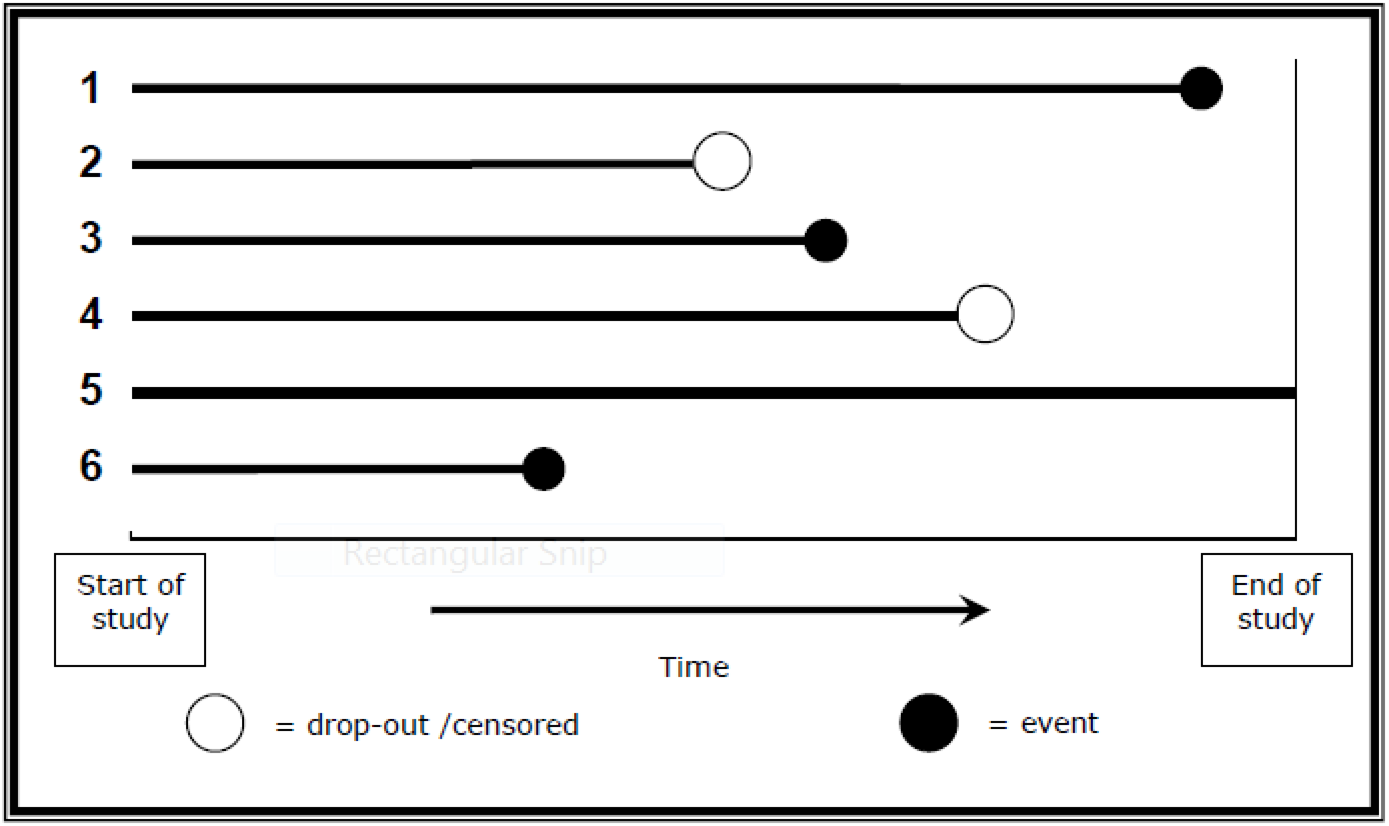
Survival analysis is often (but not exclusively) used for studying death. In this case the "**event**" of interest is death.

Censored data means that an event did not occur during the study for a particular subject. This can happen for many reasons. For example a participant may decide they no longer want to continue in the study. Or no event may have occurred by the time follow-up ended.

One of the benefits of survival analysis is that information can be used about subjects up till the time they are censored.

Censoring Example

(from Surviving Survival Analysis – An Applied Introduction by Christianna S. Williams)



Formatting Data for Survival Analysis

For every subject we need:

1. "Survival" Time: Time from the start of a study to when one of three things occurs:

- He/she has the event of interest
- He/she has an event that make him/her no longer at risk for the event (for example dropping out of the study).
- The study ends.

Survival is typically measured in days or months, but essentially needs to be continuous!

2. Censoring Indicator: An indicator variable that allows us to distinguish whether a given individual's survival time represents time to the event of interest (#1) or time until some other competing event or end of study (#2 or 3).

Be careful: Censor=0 often represents censored data!

VA Lung Cancer Example: Randomized trial of two treatment regimens for lung cancer. Other variables are included in the data set, but we will focus just on trt (1 or 2).

We will use the survival package to run the analysis.

We can test for a difference between survival times for the two groups using `survdiff`(). This gives the log-rank test. With p-value = 0.928, we cannot conclude there is a difference between the two groups.

We can find summary statistics and a standard Kaplan-Meier curve using `survfit()`.

5. The Multiple Regression Model

The objective of multiple regression is to relate a response variable (Y) simultaneously to multiple predictor variables (X_1, X_2 , X_3 , etc.)

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \ldots + \beta_k X_k + \varepsilon_i$$

k = # predictor variables

The “simple linear model” is a special case (k=1).

The β’s are called “partial regression coefficients”.

Parameter Interpretation: β_1 is the slope in the X_1 direction. Change in mean response corresponding to a one-unit change in X_1, with all other X’s held constant.

Rice Example (Gomez and Gomez): Response variable Y=yield, predictor variables X1= height and X2=tillers. n=8 varieties of rice.

In their respective simple linear regression models, tillers and height are each significant predictors of yield.

In the multiple regression with both tillers and height, neither is significant! This is due to the strong correlation between the two predictors (plus the small sample size).

We have 3 models, which one is “best”?
Note that R^2 cannot be used to determine this because the model with all possible predictors will always have the highest R^2 value!

AIC and AICC for model selection

$$\mathrm{AIC} = n\log\left(\frac{\mathrm{SSResid}}{n}\right) + 2p$$

$$\mathrm{AIC}_{\mathrm{C}} = \mathrm{AIC} + \frac{2p(p+1)}{n-p-1}$$

Notes about AIC and AICC:

1. For a fixed data set, the model with the smallest AIC (or AICC) is the preferred model.
2. p is the # of parameters. For example for multiple regression p = k+1 = # predictor variables +1.
3. AIC: The left term is related to the error variance. The right term is a "penalty" for including parameters. We try to minimize the estimated error variance, subject to a cost (or penalty) for adding variables.

4. AIC: Some people include σ^2 in the count for p, others don’t. For multiple regression, does not matter as long as you are consistent!
5. When the number of parameters (p) is large or the sample size (n) is small, AICC is preferred.
6. AIC is a general method of model selection, used in many types of models, from time series to categorical data, not just regression.
7. Be VERY careful about comparing AIC values across different software packages or even different functions in R (ex: `aic()` versus `extractAIC()`)!
8. AIC, AICC and BIC can also be calculated using the `MuMIn` package.
9. We will talk more about AIC and other model selection methods in STAT512.

6. Principal Components Analysis (PCA)

- Principal components analysis was first described by Karl Pearson in 1901.
- One of the most commonly used multivariate methods.
- The objective is to take *p* variables X_1, X_2,…X_p and find combinations of these to produce indices (principal components) Y_1, Y_2,…Y_p that are uncorrelated and describe the variation in the data.

  ex: PC1 = $Y_1 = a_1X_1 + a_2X_2 + \ldots + a_pX_p$
- The indices are ordered such that $\mathrm{Var}(Y_1)>\mathrm{Var}(Y_2)>\ldots\mathrm{Var}(Y_p)$.
- Classical case: $p<n$

  High Dimension Low Sample Size (HDLSS): $p>n$
- In some cases the first few principal components account for the bulk of the variance of the data.

- The most common use of PCA is for visualization. In other cases, the principal components are used as “new” variables for further analysis.
- In R, two functions can be used for pca analysis `prcomp()` and `princomp()`.

  A benefit of the `prcomp()` function is that it can be used for HDLSS data.
- It is conventional to center and scale the data prior to PCA analysis. After centering and scaling, each variable will have mean=0 and variance=1.
- This can be done within the `prcomp()` function:

  ```
  prcomp(, center=TRUE, scale= TRUE)
  ```

Iris Example: Fisher's (or Anderson's) iris data gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are *Iris setosa*, *versicolor*, and *virginica*.

- There are 4 variables ($p=4$) and 150 observations ($n=150$).
- Some of the variables are highly correlated. For example, correlation between Petal Length and Width is +0.963.
- Note that the species information is not used in the analysis (except for graphing).
- Example: $PC1 = 0.52Z_{SL} - 0.27Z_{SW} + 0.58Z_{PL} + 0.56Z_{PW}$
- The coefficients (or loadings) are labeled “rotation” in the `prcomp()` output. The components (or scores) are labeled “x”.
- I am using Z notation here because the variables have been centered and scaled.

Iris Results:

1. The first 2 principal components together explain 95.8% of the variation in the data.
2. The 4 principal components are uncorrelated (this will always be true).
3. The PCA plot shows good separation between Setosa (S) and the other two species (C and V).

7. Closing Comments

When planning a study:

- Start with specific aims. The specific aims drive the design and analysis!
- 3 R’s of Experimental Design:
    - Randomization: not difficult and just good science.
    - Replication: critical to have independent reps.
    - Reduce Noise: consider technique, selection of subjects, blocking.
- When computing power, keep it simple but realistic.

When entering data:

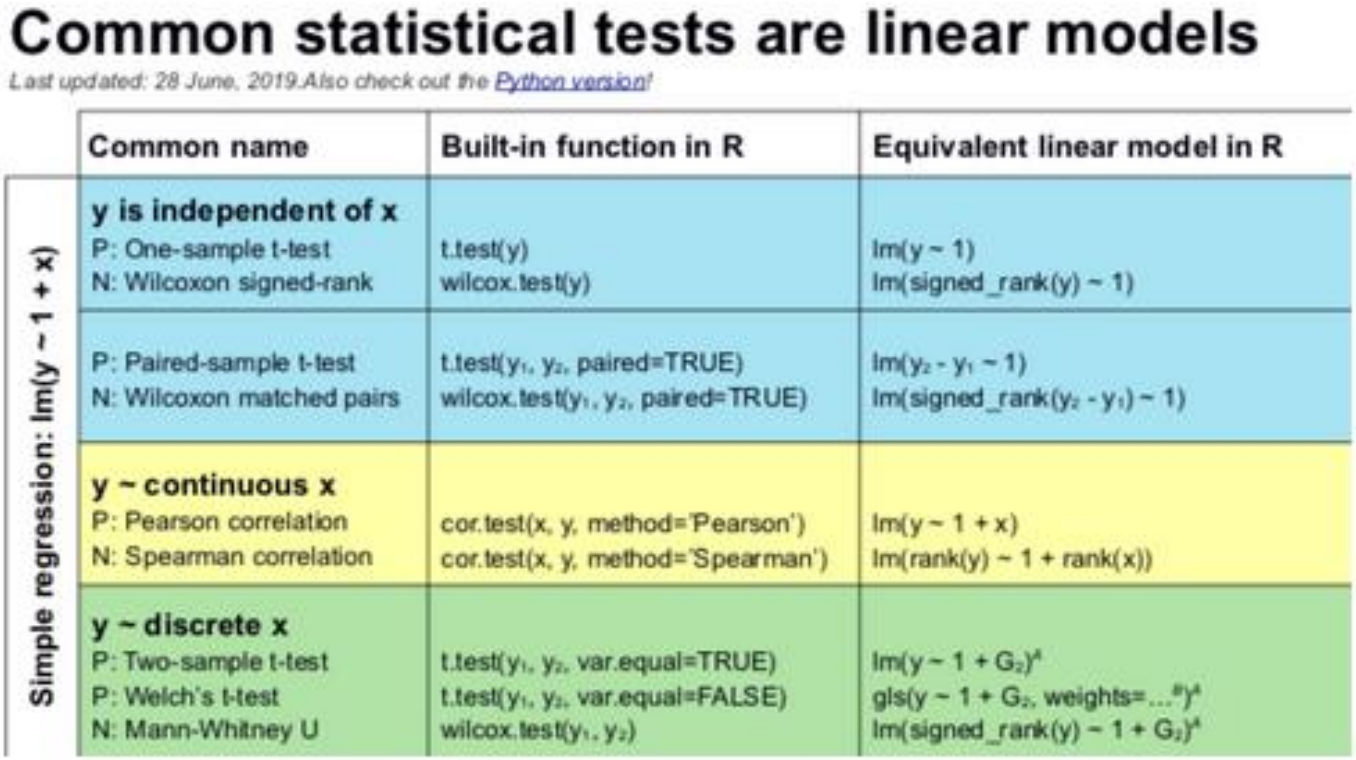
- Not unusual to have multiple versions of the data. But plan ahead to reduce the number of data versions.
- **The original data should never be modified and should be stored in a secure location!**
- Short, unique, informative column names save time with software.
- “Extra” columns not a problem. Would rather have more information than needed, rather than not enough.
- Fewer spreadsheets (with more columns) usually easier to work with.
- Consistency (within and between sheets) very helpful.
- Data cleaning, formatting can be automated in software (but requires a reasonable starting point).

Exploratory data analysis:

- This step is very important!
- Keep it simple: summary statistics and summary graphics.
- Provides first pass information about specific aims.
- Very helpful for data checking.
- Can help inform the analysis.
- Provides a way to spot check the final results.

A “Unifying” view of the STAT511 topics:

- Jonas Kristoffer Lindeløv provides an elegant summary of how many common statistical tests are linear models.
- https://teachdatascience.com/onemodel/



Common statistical tests are linear models

Last updated: 28 June, 2019.Also check out the Python version!

	Common name	Built-in function in R	Equivalent linear model in R

Simple regression: lm(y ~ 1 + x)	**y is independent of x**
	P: One-sample t-test
	N: Wilcoxon signed-rank
		t.test(y)
		wilcox.test(y)
			lm(y ~ 1)
			lm(signed_rank(y) ~ 1)

	P: Paired-sample t-test
	N: Wilcoxon matched pairs
		t.test(y_1, y_2, paired=TRUE)
		wilcox.test(y_1, y_2, paired=TRUE)
			lm(y_2 - y_1 ~ 1)
			lm(signed_rank(y_2 - y_1) ~ 1)

	y ~ continuous x
	P: Pearson correlation
	N: Spearman correlation
		cor.test(x, y, method='Pearson')
		cor.test(x, y, method='Spearman')
			lm(y ~ 1 + x)
			lm(rank(y) ~ 1 + rank(x))

	y ~ discrete x
	P: Two-sample t-test
	P: Welch's t-test
	N: Mann-Whitney U
		t.test(y_1, y_2, var.equal=TRUE)
		t.test(y_1, y_2, var.equal=FALSE)
		wilcox.test(y_1, y_2)
			lm(y ~ 1 + G_2)^A
			gls(y ~ 1 + G_2, weights=...^{B})^A
			lm(signed_rank(y) ~ 1 + G_2)^A

A Final Thought:

- In consulting, people often ask me if they are doing the “right” analysis.
- Instead, I prefer to focus on finding a “reasonable” approach with the understanding that there may be multiple “reasonable” options.
- My personal philosophy: Use the simplest, reasonable approach that addresses research questions.