Modeling tools and techniques using R

Paul Esker
Pennsylvania State University
pde6@psu.edu or paul.esker@gmail.com

Felipe Dalla Lana
Ohio State University
dallalanadasilva.2@osu.edu or felipedallalana@gmail.com

Workshop format

- 8 am to 12 pm
- Mix of some lecture material with hands-on learning in R and Rstudio
  - ➢R: www.r-project.org
  - ➢Rstudio: www.rstudio.com
- Focus is on useful tools for modeling, including examining model assumptions and predictions
- Examples draw on collaboration and consulting experiences (PDE) as such not all focus on Plant Path examples
- Assumption: some exposure to R and comfort with at least running code and working with different packages

Materials

- Webpage: https://rtools.netlify.com/
- Github: https://github.com/PSUPlantEpidemiology/APS2019.git
- Material that is available:
  - R scripts
  - R markdown documents
  - Output in word and pdf format for note-taking

Background to notes

- These slides draw on previous teaching experiences including:
  - ➢Majority of examples (except dose-response additional example) focus on a regression framework assuming continuous-type variables
    - All explanatory variables continuous = regression
    - All explanatory variables categorical = ANOVA type methods
    - Combination of continuous and categorical = ANCOVA type methods
  - ➢The combination of the Rmd output and notes should provide a more complete overview (I hope…)
  - ➢Two statistics courses geared to graduate students at the University of Costa Rica
  - ➢Week-long workshops on statistical modeling in epidemiology taught in Toluca, Mexico
  - ➢Consulting across the following programs: plant pathology, agronomy, entomology, soils science, horticulture, biology, molecular biology, plant physiology, engineering, chemistry, and the social sciences

Available material

Background material:
1. Introduction (Rmd) to R (from McRoberts and Esker)
2. Correlations

Primary material:
1. R scripts for linear, multiple, and regression modeling tools.
2. Rmd files for same models
3. PDF outputs from the models.

Additional material:
1. Rmd files and outputs for examples: quadratic, nonlinear, nonparametric, and generalized linear models

More about dditional examples and learning goals

- *Polynomial regression*: examine issues in collinearity in more detail and how to define centered variables
- *Mosquito Dose-Response Final*: generalized linear model (mixed) example that compares different model types and assumptions
- *Nonlinear regression*: introduction to defining nonlinear models and initial parameters
- *Nonparametric regression*: smoothing methods to look at nonlinear responses and the tradeoff with model complexity

Modeling goals

The first step (of seven – note that we will touch on these others with our examples)
Decide on the type of model that is needed in order to achieve the goals of the study. In general, there are five reasons one might want to build a regression model.

They are:
- For **predictive** reasons — that is, the model will be used to predict the response variable from a chosen set of predictors.
- For **theoretical** reasons — that is, the researcher wants to estimate a model based on a known theoretical relationship between the response and predictors.
- For **control** purposes — that is, the model will be used to control a response variable by manipulating the values of the predictor variables.
- For **inferential** reasons — that is, the model will be used to explore the strength of the relationships between the response and the predictors.
- For **data summary** reasons — that is, the model will be used merely as a way to summarize a large set of data by a single equation.

- https://newonlinecourses.science.psu.edu/stat501/node/332/

Modeling thoughts

General principles

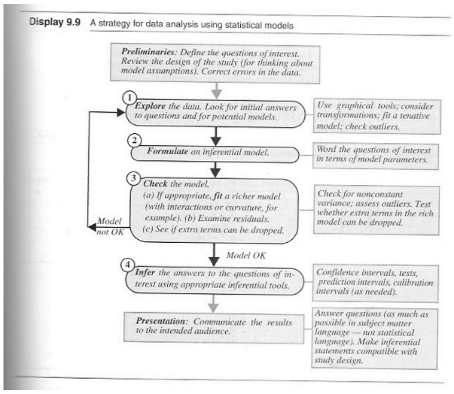
Our general principles for building regression models for prediction are as follows:

1. Include all input variables that, for substantive reasons, might be expected to be important in predicting the outcome.
2. It is not always necessary to include these inputs as separate predictors—for example, sometimes several inputs can be averaged or summed to create a "total score" that can be used as a single predictor in the model.
3. For inputs that have large effects, consider including their interactions as well.
4. We suggest the following strategy for decisions regarding whether to exclude a variable from a prediction model based on expected sign and statistical significance (typically measured at the 5% level; that is, a coefficient is "statistically significant" if its estimate is more than 2 standard errors from zero):
   (a) If a predictor is not statistically significant and has the expected sign, it is generally fine to keep it in. It may not help predictions dramatically but is also probably not hurting them.
   (b) If a predictor is not statistically significant and does not have the expected sign (for example, incumbency having a negative effect on vote share), consider removing it from the model (that is, setting its coefficient to zero).
   (c) If a predictor *is* statistically significant and does not have the expected sign, then think hard if it makes sense. (For example, perhaps this is a country such as India in which incumbents are generally unpopular; see Linden, 2006.) Try to gather data on potential lurking variables and include them in the analysis.
   (d) If a predictor is statistically significant and has the expected sign, then by all means keep it in the model.

These strategies do not completely solve our problems but they help keep us from making mistakes such as discarding important information. They are predicated on having thought hard about these relationships *before* fitting the model. It's always easier to justify a coefficient's sign after the fact than to think hard ahead of time about what we expect. On the other hand, an explanation that is determined after running the model can still be valid. We should be able to adjust our theories in light of new information.

Source. Gelman y Hall. 2007. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge.

Modeling strategy



Source. Ramsey y Schafer. 2002. The Statistical Sleuth – A Course in Methods of Data Analysis, Second Edition. Duxbury.

Starting from a regression framework

- We will use the following structure to explore tools in R:
  - Linear model: understand assumptions and tools for prediction
  - Expanding the model with multiple explanatory variables
  - Comparing methods for comparing larger sets of models (i.e., all combinations, etc.)
  - If time, explore some additional tools for additional model types:
    - Quadratic to understand concepts in collinearity
    - Nonlinear to visualize the fitting algorithms
    - Generalized linear models to look at model assumptions for things like dose-response curves

Exponential family of distributions

TABLE 2.1. Examples of probability distributions that belong to the exponential family. All distributions, except for the log-normal distribution, have been parameterized such that $\mu = E(Y)$ is the mean of the random variable Y. For the log-normal distribution, the distribution of $Z = \log(Y)$ is normally distributed with mean $\mu_z = E[\log(Y)]$ and $\phi = \text{var}[\log(Y)]$.

Distribution	$f(y \mid \mu)$	$\theta = \eta(\mu)$	Variance	ϕ
Normal (μ, ϕ) $-\infty < y < \infty$	$\frac{1}{\sqrt{2\pi\phi}}\exp\left[\frac{-(y-\mu)^2}{2\phi}\right]$	μ	ϕ	$\phi > 0$
Inverse normal (μ, ϕ) $-\infty < y < \infty$	$\left(\frac{1}{2\pi\phi y^3}\right)^{1/2}\exp\left[\frac{-(y-\mu)^2}{2y\phi\mu^2}\right]$	$1/\mu^2$	$\phi\mu^3$	$\phi > 0$
Log-normal (μ, ϕ) $-\infty < \log(y) < \infty$	$f[\log(y) \mid \mu] = \frac{1}{\sqrt{2\pi\phi}}\exp\left\{\frac{-[\log(y)-\mu]^2}{2\phi}\right\}$	μ	ϕ	$\phi > 0$
Gamma (μ, ϕ)† $y \geq 0$	$\frac{y^{\phi-1}}{\Gamma(\phi)}\left(\frac{\phi}{\mu}\right)^{\phi}\exp\left(\frac{-\phi y}{\mu}\right)$	$1/\mu$	$\phi\mu^2$	$\phi > 0$
Exponential (μ) $y \geq 0$	$\frac{1}{\mu}\exp\left(\frac{-y}{\mu}\right)$	$1/\mu$	μ^2	$\phi \equiv 1$
Beta (μ, ϕ)† $0 \leq y \leq 1$	$\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma[(1-\mu)\phi]}y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}$	$\log\left(\frac{\mu}{1-\mu}\right)$	$\frac{\mu(1-\mu)}{(1+\phi)}$	$\phi > 0$
Binomial (n, π) $y = 0, \ldots, n$ where $\pi = \mu/n$	$\binom{n}{y}\left(\frac{\mu}{n}\right)^y\left(1-\frac{\mu}{n}\right)^{n-y}$	$\log\left(\frac{\mu}{n-\mu}\right)$	$\mu\left(1-\frac{\mu}{n}\right)$	$\phi \equiv 1$
Geometric (μ, ϕ) $y = 0, 1, 2, \ldots$	$\left(\frac{\mu}{1+\mu}\right)^y\left(\frac{1}{1+\mu}\right)$	$\log(\mu)$	$\mu + \mu^2$	$\phi \equiv 1$
Poisson (μ)‡ $y = 0, 1, 2, \ldots$	$\frac{\mu^y e^{-\mu}}{y!}$	$\log(\mu)$	μ	$\phi \equiv 1$

† The gamma function $\Gamma(x)$ equals $(x-1)!$ when x is an integer but otherwise equals $\int_0^\infty t^{x-1}e^{-t}dt$.
‡ In the case of an over-dispersed Poisson distribution, the variance of Y is $\phi\mu$ where $\phi > 0$ and often $\phi > 1$.

Some basic syntax

Syntax in R form	Interpretation
Y ~ A	Linear regression that includes both the intercept and slope
Y ~ -1 + A	Linear regression that does not include the intercept (= regression forced through intercept)
Y ~ A + I(A^2)	Polynomial model [**I()** = identity function]
Y ~ A + B	First order model for factors A and B, without interaction
Y ~ A:B	First order model that only includes the interaction term
Y ~ A*B	Full first order model Y ~ A + B + A:B
Y ~ (A + B + C) ^2	Model that includes all first order effects plus the interactions through order "X" (= second order in this example) : Y ~ A + B + C + AB + AC + BC

Background

- Our interest includes:
  1. Understanding the inherent relationships between different variables.
  2. Developing methods for predictions based on estimating a dependent variable (risk model, forecast model, …)
- Given that, we are interested in exploring these relationships based on quantitative variables
  - It should be obvious though that we can also incorporate qualitative factors and create conditional models (i.e., dependent on the factor of interest, *dummy variables*)

Starting point…

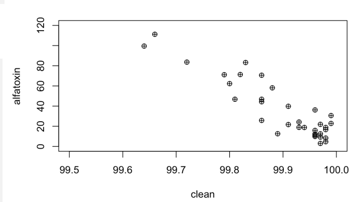
- We need to define the independent-dependent relationship
  - Dependent variable = response
  - Independent variable(s) = regressor(s), predictor(s), …
- *Linearity assumption* = the rate of change (slope) does not change at different levels of *X*

R example 1.

- File name (R script): “Linear regression.R”
  - Data come from the analysis of alfatoxin in peanut
    - ○ Percentage clean grain
    - ○ Concentration of alfatoxin
- Objectives:
  - Quantify the relationship between the percentage clean grain and alfatoxin concentration
  - Determine if this model can be used for future predictions

```
mean(alfatoxin)
## [1] 36.60294
sd(alfatoxin)
## [1] 29.3194
sd(alfatoxin)/mean(alfatoxin)*100
## [1] 80.1012
```

```
rcorr(clean, alfatoxin)

##        x     y
## x  1.00 -0.91
## y -0.91  1.00
##
## n= 34
##
##
## P
##    x  y
## x     0
## y 0
```



Model structure

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

Response

Intercept = intersection with vertical axis

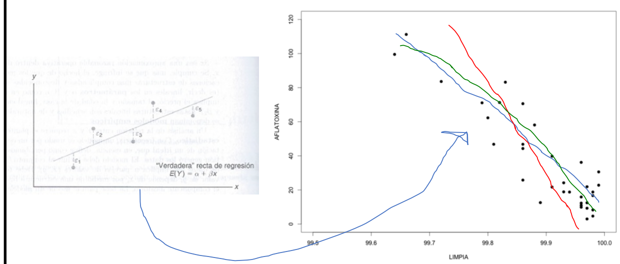
Slope = rate of change

Random variable = error or residual variance, which includes a combination of known and unknown factors

Independent variable as a function of the dependent variable:

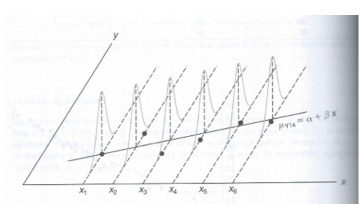
- Dependent variable = *random variable*, since the residual variable is random
- Independent variable = not random, but it is measured with some minimal error

What are we trying to do?



Model assumptions

- Errores are distributed normally
- Mean error is 0
- Variance is the same for all errors



Modeling fitting method: least squares

- Most common method for the majority of statistical packages
- We can use likelihood based methods (generalized linear models) also
  - Dose-response example illustrates that philosophy
  - Some of our automated methods also are based on this approach
- Objective: Minimize the residual sum of squares
  - Reduce the amount of error between the observed value and the model adjusted value (i.e., predicted or estimated value)

$$e_i = Y_i - \widehat{Y}_i$$

$$\sum_i (Y_i - \widehat{Y}_i) = \sum_i \left[Y_i - (\widehat{\beta}_0 - \widehat{\beta}_1 \overline{X})\right]^2$$

Estimating the coefficients based on least squares

- Slope:

This is pure mathematics…

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

- Summation of the deviations of X multiplied by the deviations in Y
- Summation of the deviations in X values

- Intercept:

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X} = \frac{\sum_{i=1}^{n} Y_i - \widehat{\beta}_1 \sum_{i=1}^{n} X_i}{n}$$

Regression results

```
> linreg <- with(peanut, lm(alfatoxin~clean)) #Format, Y <- X
> anova(linreg) #ANOVA table to see how the model fit looks
Analysis of Variance Table

Response: alfatoxin
          Df  Sum Sq Mean Sq F value    Pr(>F)
clean      1 23334.5 23334.5  148.36 1.479e-13 ***
Residuals 32  5033.2   157.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(linreg) #Another way to see results of the model, with a few more details.

Call:
lm(formula = alfatoxin ~ clean)

Residuals:
    Min      1Q  Median      3Q     Max
-25.843  -7.997  -2.771   6.835  27.695

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 28443.18    2332.21   12.20 1.43e-13 ***
clean        -284.36      23.35  -12.18 1.48e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.54 on 32 degrees of freedom
Multiple R-squared:  0.8226,  Adjusted R-squared:  0.817
F-statistic: 148.4 on 1 and 32 DF,  p-value: 1.479e-13
```

Relationship between the two variables exists

Regression results

```
> linreg <- with(peanut, lm(alfatoxin~clean)) #Format, Y <- X
> anova(linreg) #ANOVA table to see how the model fit looks
Analysis of Variance Table

Response: alfatoxin
          Df  Sum Sq Mean Sq F value    Pr(>F)
clean      1 23334.5 23334.5  148.36 1.479e-13 ***
Residuals 32  5033.2   157.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(linreg) #Another way to see results of the model, with a few more details.

Call:
lm(formula = alfatoxin ~ clean)

Residuals:
    Min      1Q  Median      3Q     Max
-25.843  -7.997  -2.771   6.835  27.695

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 28443.18    2332.21   12.20 1.43e-13 ***
clean        -284.36      23.35  -12.18 1.48e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.54 on 32 degrees of freedom
Multiple R-squared:  0.8226,  Adjusted R-squared:  0.817
F-statistic: 148.4 on 1 and 32 DF,  p-value: 1.479e-13
```

$$\widehat{\sigma} = \sqrt{\frac{SC_{error}}{g.l.}}$$

Important for looking at the overall distributions of the intercept and slope

Regression results

```
> linreg <- with(peanut, lm(alfatoxin~clean)) #Format, Y <- X
> anova(linreg) #ANOVA table to see how the model fit looks
Analysis of Variance Table

Response: alfatoxin
          Df  Sum Sq Mean Sq F value    Pr(>F)
clean      1 23334.5 23334.5  148.36 1.479e-13 ***
Residuals 32  5033.2   157.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(linreg) #Another way to see results of the model, with a few more details.

Call:
lm(formula = alfatoxin ~ clean)

Residuals:
    Min      1Q  Median      3Q     Max
-25.843  -7.997  -2.771   6.835  27.695

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 28443.18    2332.21   12.20 1.43e-13 ***
clean        -284.36      23.35  -12.18 1.48e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.54 on 32 degrees of freedom
Multiple R-squared:  0.8226,  Adjusted R-squared:  0.817
F-statistic: 148.4 on 1 and 32 DF,  p-value: 1.479e-13
```

You can change the value for β_0 to test other hypotheses (see additional notes)

$$H_0: \beta_0 = 0; H_1: \beta_0 \neq 0$$

$$T = \frac{\widehat{\beta}_0 - \beta_0}{EE(\widehat{\beta}_0)} = \frac{\widehat{\beta}_0 - \beta_0}{\widehat{\sigma}\sqrt{\frac{1}{(n-1)s_X^2}}}, g.l. = n-2$$

Interpretation?

Regression results

```
> linreg <- with(peanut, lm(alfatoxin~clean)) #Format, Y <- X
> anova(linreg) #ANOVA table to see how the model fit looks
Analysis of Variance Table

Response: alfatoxin
          Df  Sum Sq Mean Sq F value    Pr(>F)
clean      1 23334.5 23334.5  148.36 1.479e-13 ***
Residuals 32  5033.2   157.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(linreg) #Another way to see results of the model, with a few more details.

Call:
lm(formula = alfatoxin ~ clean)

Residuals:
    Min      1Q  Median      3Q     Max
-25.843  -7.997  -2.771   6.835  27.695

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 28443.18    2332.21   12.20 1.43e-13 ***
clean        -284.36      23.35  -12.18 1.48e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.54 on 32 degrees of freedom
Multiple R-squared:  0.8226,  Adjusted R-squared:  0.817
F-statistic: 148.4 on 1 and 32 DF,  p-value: 1.479e-13
```

The β_1 can change to test additional hypotheses (see addtional materials)

$H_0: \beta_1 = 0; H_1: \beta_1 \neq 0$

$$T = \frac{\widehat{\beta}_1 - \beta_1}{EE(\widehat{\beta}_1)} = \frac{\widehat{\beta}_1 - \beta_1}{\widehat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s_X^2}}}, g.l. = n-2$$

Interpretation?

Regression results

```
> linreg <- with(peanut, lm(alfatoxin~clean)) #Format, Y <- X
> anova(linreg) #ANOVA table to see how the model fit looks
Analysis of Variance Table

Response: alfatoxin
          Df  Sum Sq Mean Sq F value    Pr(>F)
clean      1 23334.5 23334.5  148.36 1.479e-13 ***
Residuals 32  5033.2   157.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(linreg) #Another way to see results of the model, with a few more details.

Call:
lm(formula = alfatoxin ~ clean)

Residuals:
    Min      1Q  Median      3Q     Max
-25.843  -7.997  -2.771   6.835  27.695

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 28443.18    2332.21   12.20 1.43e-13 ***
clean        -284.36      23.35  -12.18 1.48e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.54 on 32 degrees of freedom
Multiple R-squared:  0.8226,  Adjusted R-squared:  0.817
F-statistic: 148.4 on 1 and 32 DF,  p-value: 1.479e-13
```

Coefficient of determination = measure of the proportion of variability explained by adjusted model

$$R^2 = 1 - \frac{SC_{error}}{SC_{total}} = \frac{SC_{regresión}}{SC_{total}}$$

Regression results

```
> linreg <- with(peanut, lm(alfatoxin~clean)) #Format, Y <- X
> anova(linreg) #ANOVA table to see how the model fit looks
Analysis of Variance Table

Response: alfatoxin
          Df  Sum Sq Mean Sq F value    Pr(>F)
clean      1 23334.5 23334.5  148.36 1.479e-13 ***
Residuals 32  5033.2   157.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(linreg) #Another way to see results of the model, with a few more details.

Call:
lm(formula = alfatoxin ~ clean)

Residuals:
    Min      1Q  Median      3Q     Max
-25.843  -7.997  -2.771   6.835  27.695

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 28443.18    2332.21   12.20 1.43e-13 ***
clean        -284.36      23.35  -12.18 1.48e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.54 on 32 degrees of freedom
Multiple R-squared:  0.8226,  Adjusted R-squared:  0.817
F-statistic: 148.4 on 1 and 32 DF,  p-value: 1.479e-13
```

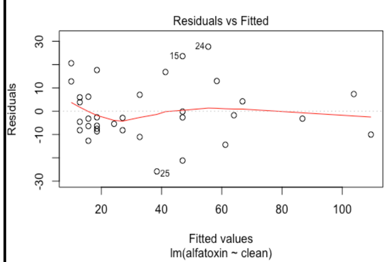
$R^2_{adjusted}$= takes into account the number of factors in the to reduce the effect of just seeing an improved R^2 with more variables)

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

Interpretation for R^2

- Realibility for R^2 is a function of:
  - Database size
  - Type of application
- Final interpretation will vary depending on the system under study:
  - 0.95 (biology) = good model fit
  - 0.95 (chemistry) = poor model fit

Model assumptions based on plot(): residuals versus fitted values

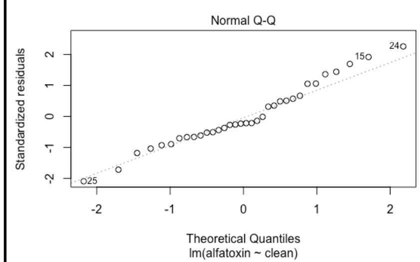


plot(linreg, which=1)

For those familiar with ANOVA, this has a similar interpretation

Note: These are the raw differences between observed and predicted so interpret the results with caution

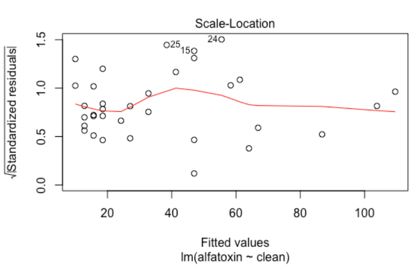
Quantile-quantile (Q-Q plot)



plot(linreg, which=2)

Are the residuals distributed normally?

Scale-location plot

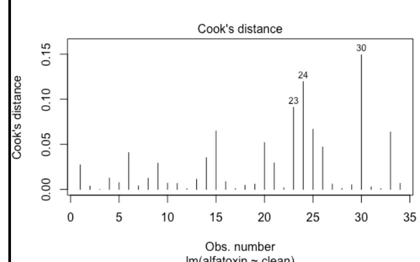


plot(linreg, which=3)

Are the residuals distributed randomly?

One can also use in the MASS library options to obtain studentized residuals

Cook Distance



plot(linreg, which=4)

Measures the effect of each observation when eliminated from the model (relative importance or leverage)

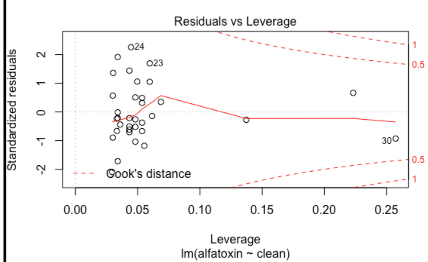
Different methods to define the critical value:

$D_i > 1$

$D_i > 4/n$ (= 4/34 = 0,12)

$D_i > 4/(n-p-1)$ (= 4/(34-2-1) = 0,129)

Leverage plot



plot(linreg, which=5)

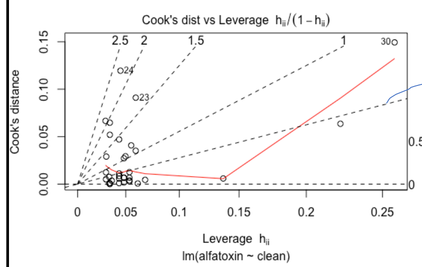
Are there values that have greater influence and would impact the model fit?

Values: 0 – 1

- 0 = no affect on the model;
- 1 = that the regression line should follow this point perfectly
- If there are observations with leverage = outlier observation since it will have a large effect on the model fit

Cook's Distance and Leverage

(linreg, which=6)



Information about the residuals

Estimation and prediction

- Estimation = we are interested in study the response variable for specific values of X that are within the range of observed values
- Example: What is the mean concentration of alfatoxin when the percentage clean seed is 99.68%?
- Based on confidence intervals for probable values
  - ➢$X=X_0$
  - ➢Y is distributed as follows:
    - Mean = $\beta_0+\beta_1X$
    - Standard deviation = σ

Estimation and prediction

- Prediction = the objective is to predict a new value(s) assuming a future occurrence (new lots, new forecast year, etc.)
  - ➢Example: What is the mean concentration of alfatoxin when the percentage clean seed is 99.68% is we obtained an unknown sample from a different location?
  - ➢In this case, the prediction takes into account two sources of uncertainty:
    - About the general location of population mean
    - About the location of the new value in the future as related to the mean value

$$Pred[Y \mid X_0] = \widehat{\mu}\{Y \mid X_0\} = \widehat{\beta}_0 + \widehat{\beta}_1 X_0$$

Confidence intervals

$$EE[\widehat{\mu}\{Y \mid X_0\}] = \sigma^2 \sqrt{\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{(n-1)s_X^2}}, g.l. = n-2$$

Prediction intervals

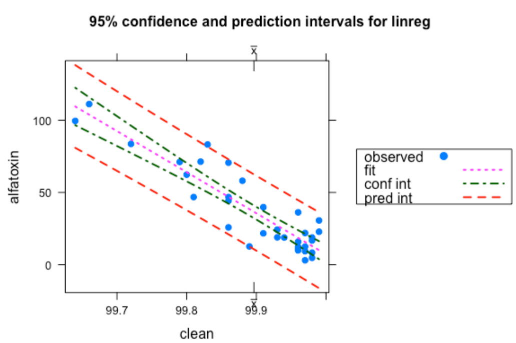
$$EE[Pred\{Y \mid X_0\}] = \sqrt{\widehat{\sigma}^2 + EE[\widehat{\mu}\{Y \mid X_0\}]^2}$$

Peanut example

```
observation <- data.frame(clean=99.68)

predict(object=linreg, newdata=observation, interval="confidence")
##        fit      lwr      upr
## 1 98.15855 86.97085 109.3462


predict(object=linreg, newdata=observation, interval="predict")
##        fit      lwr     upr
## 1 98.15855 70.27011 126.047
```

Additional R tools for examining model fit and assumptions (package = HH)

```
ci.plot(linreg)
```



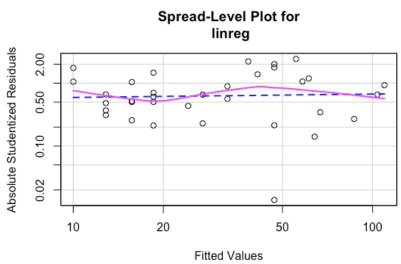
Additional R tools for examining model fit and assumptions (package = HH)

```
# Method to look for outliers using a Bonferroni adjustment
outlierTest(linreg)

## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##    rstudent unadjusted p-value Bonferroni p
## 24 2.425727           0.021292      0.72394
```

```
# Test of homoscedasticity
ncvTest(linreg)

## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.183475, Df = 1, p = 0.6684
```

Additional R tools for examining model fit and assumptions (package = HH)

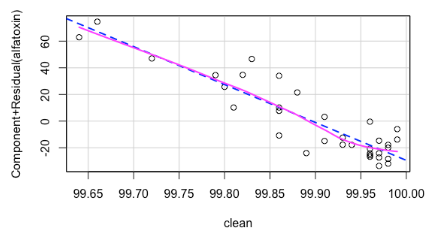
```
# Method to verify if there is dependency in the model, which means
that a transformation may be appropriate to model the relationship
spreadLevelPlot(linreg)
```




```
## Suggested power transformation:  0.9466765
```

Additional R tools for examining model fit and assumptions (package = HH)

```
# Method to verify if there is evidence that the relationship
is not linear
crPlots(linreg)
```



Transformations

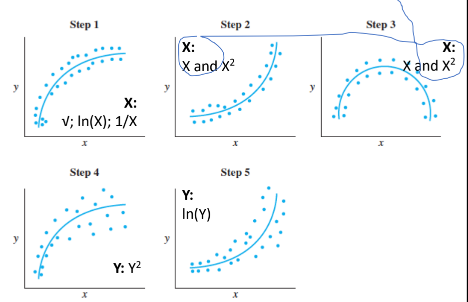
- Can involve transforming:
  - Response variable
  - Predictor variable
  - Both variables

For further information, see:
11.1 = Lyman Ott and Longnecker.
https://onlinecourses.science.psu.edu/stat501/node/48

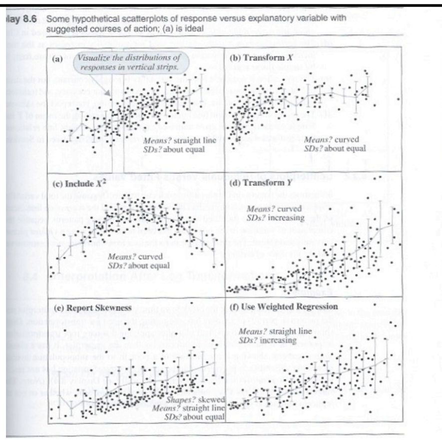
Examples

Links with multiple regression concepts

FIGURE 11.6
Plots corresponding to steps in Definition 11.2

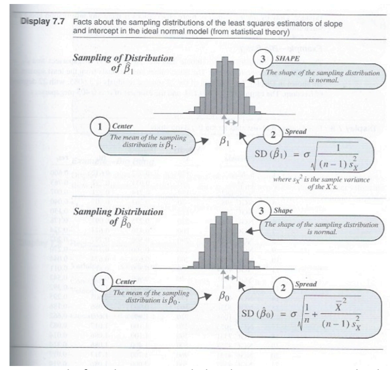


Source: Lyman Ott and Longnecker.



Source: Ramsey y Schafer. The Statistical Sleuth – A Course in Methods of Data Analysis

Appendix 1



(Source: Ramsey y Schafer. The Statistical Sleuth – A Course in Methods of Data Analysis)

Appendix 2 – Cook's Distance

$$D_i = \sum_{j=1}^{n} \frac{(\widehat{Y}_{j(i)} - \widehat{Y}_j)^2}{p\widehat{\sigma}^2}$$

Adjusted value with observation i

Number of predictors

Estimated variance based on the complete set of observations

Adjusted value for complete observations

Appendix 2 – Cook's Distance (Version 2, better for computational use)

$$D_i = \frac{1}{p}(studres_i)^2\left(\frac{h_i}{1-h_i}\right)$$

Number of predictors

Student residual

Hat matrix value

http://en.wikipedia.org/wiki/Hat_matrix

Multiple regression modeling as a next step

- Components:
  - Response variable (dependent variable)
  - >1 independent varible (multiple factors)
- Model types include:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2$$
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1^3$$

Assumptions

- Model is properly defined
- Assumptions about the errors:

$$\varepsilon_i \sim Normal$$
$$Var(\varepsilon_i) = \sigma_\varepsilon^2$$
$$\varepsilon_i = \text{independent}$$

Partial slopes

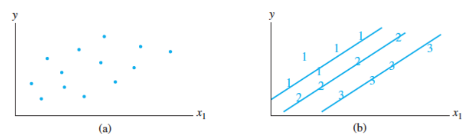
- Parameters for the independent variables:
  - Are called partial slopes because these values represetn the change in Y as a unit change in X_i (factor i) but by maintaining constant the other factors

Additive effects

- When the effects between the X factors are independents

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

FIGURE 12.2
(a) Scatterplot of y versus x_1; (b) scatterplot of y versus x_1, indicating additivity of effects for x_1 and x_2

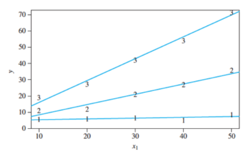


Interactions

- When there are changes in Y with different levels of X_1, but the magnitud of this change depends on the level of X_2

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 X_2) + \varepsilon$$

FIGURE 12.3
Scatterplot of y versus x_1 at three levels of x_2



Dummy variables

- Qualititative factors: We can use a 0-1 representtion to define the variable
  - Example: for two factors, A and B,
    - X_1 = 1 if treatment A
    - X_1 = 0 if treatment B
  - Which results in a model form:

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon = \begin{cases} Y = \beta_0 + \varepsilon, \text{ if treatment = B} \\ Y = \beta_0 + \beta_1 X_1 + \varepsilon, \text{ if treatment = A} \end{cases}$$

General linear model form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$$

- For the independent variables, we can have:
  - Quantitative factors
  - Qualitative factors
  - Quadratic forms for the factors
  - Interactions among factors (multiplicative form)

Estimation

- Is based on the use of normal equations
  - Which are simultaneously solved

TABLE 12.5
Normal equations for a multiple regression model

	y_i	$\hat{\beta}_0$	$x_{i1}\hat{\beta}_1$	$\cdots$	$x_{ik}\hat{\beta}_k$
1	$\Sigma y_i =$	$n\hat{\beta}_0$	$+ \; \Sigma x_{i1}\hat{\beta}_1$	$+ \cdots +$	$\Sigma x_{ik}\hat{\beta}_k$
x_{i1}	$\Sigma x_{i1} y_i =$	$\Sigma x_{i1}\hat{\beta}_0$	$+ \; \Sigma x_{i1}^2\hat{\beta}_1$	$+ \cdots +$	$\Sigma x_{i1} x_{ik}\hat{\beta}_k$
$\vdots$	$\vdots$				
x_{ik}	$\Sigma x_{ik} y_i =$	$\Sigma x_{ik}\hat{\beta}_0$	$+ \; \Sigma x_{ik} x_{i1}\hat{\beta}_1$	$+ \cdots +$	$\Sigma x_{ik}^2\hat{\beta}_k$

Tests

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$
$$H_1 : \text{at least one } \beta \neq 0$$

$$F = \frac{SS_{regression} / k}{SS_{error} / \left[n - \left(k+1\right)\right]} = \frac{MS_{regression}}{MS_{error}}$$

Reject H_0 means that there is some degree of predictive value, meaning some of the factors are important (statistically)

Let's move into R to see this…

- Data source: aphid counts in different lots
- Additional measures include:
  - Average daily temperature (C)
  - Average daily relative humitidy (%)
- Can we define a "best" model that describes aphid numbers as a combination of these two factors?

```
> summary(aphids_data)
      lot             aphids         temperature        humidity    
 Min.   : 1.00   Min.   :  6.00   Min.   :16.30   Min.   : 6.00  
 1st Qu.: 9.25   1st Qu.: 27.75   1st Qu.:26.00   1st Qu.:21.88  
 Median :17.50   Median : 62.00   Median :28.30   Median :32.50  
 Mean   :17.50   Mean   : 61.91   Mean   :28.09   Mean   :35.19  
 3rd Qu.:25.75   3rd Qu.: 92.00   3rd Qu.:31.95   3rd Qu.:46.38  
 Max.   :34.00   Max.   :118.00   Max.   :34.50   Max.   :79.50  
```

Let's move into R to see this…below is the additive model

```
> model3<-with(aphids_data, lm(aphids~temperature+humidity))
> anova(model3)
Analysis of Variance Table

Response: aphids
            Df  Sum Sq Mean Sq F value    Pr(>F)    
temperature  1 15194.8 15194.8 28.7765 7.554e-06 ***
humidity     1  4813.1  4813.1  9.1151  0.005038 ** 
Residuals   31 16368.9   528.0                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$SS_{regression}$ = 15194,8 + 4813,1 = 20007,9
$df_{regression}$ = 2

F = (20007,9/2) / 528 = 18,946
Prob(F) = 8,42 x 10^{-6}

Evidence that there is a relationship between aphid numbers and temperature and relative humidity

Model standard deviation

Residuals

$$Y_i - \widehat{Y}_i = Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_{i1} + \widehat{\beta}_2 X_{i2} + \ldots + \widehat{\beta}_k X_{ik})$$

Model standard deviation

$$s_\varepsilon = \sqrt{MS_{error}} = \sqrt{\frac{SS_{error}}{n - \left(k+1\right)}}$$

Also known as: residual standard error, standard error of the estimate, root mean square error

Coefficient of determination

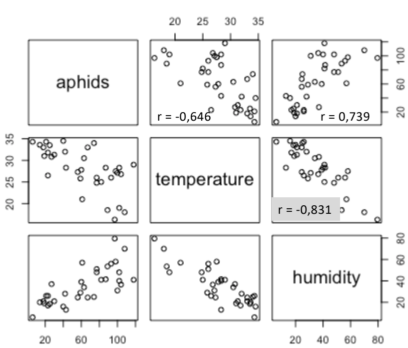
$$R^2_{Y \cdot X_1 \cdots X_k} = \frac{SS_{total} - SS_{error}}{SS_{total}}$$

- Interpretation is the same as a simple linear model
- By default, R^2 will increase with the additional of further factors
- If the factors are not correlated, this represents the summation of each individual R^2
- When the variables are correlated = collinearity

Collinearity

- When there exists some correlation between independent variables
  - One factor may be explained well by another factors
  - May not impact model if the correlations are small
  - When correlation is high, could impact model fit (overfitted)

Back to the example



One way to look at this: variance inflation factor (VIF)

$$VIF = \frac{1}{(1 - R_X^2)}$$

= proption of the variance in X explained by the liner relation with the other variables in the model

Interpretation

VIF = 1, no evidence of collinearity
1 < VIF < 5, moderate evidence of collinearity
VIF > 5, strong evidence of collinearity

```
> vif(lm(aphids~temperature+humidity, data=aphids_data))
temperature    humidity
   3.238084    3.238084
```

Model comparision: nested models

- F-test: complete model versus reduced model
  - Question, is there still predictive value?

$$F = \frac{\left[SS_{full} - SS_{reduced}\right] / (k-g)}{SS_{full} / \left[n-(k+1)\right]}$$

With,

$$df_1 = k - g$$

$$df_2 = \left[n-(k+1)\right]$$

Let's compare the additive model with the model that includes an interaction term between temp and RH

```
> summary(model3)

Call:
lm(formula = aphids ~ temperature + humidity)

Residuals:
    Min      1Q  Median      3Q     Max 
-35.393 -14.006  -3.198  10.335  49.265 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)  35.8255    53.5388   0.669  0.50835   
temperature  -0.6765     1.4360  -0.471  0.64089   
humidity      1.2811     0.4243   3.019  0.00504 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22.98 on 31 degrees of freedom
Multiple R-squared:   0.55,	Adjusted R-squared:  0.521 
F-statistic: 18.95 on 2 and 31 DF,  p-value: 4.212e-06
```

```
> summary(model4)

Call:
lm(formula = aphids ~ temperature + humidity + temperature:humidity)

Residuals:
   Min     1Q Median     3Q    Max 
-41.13 -12.87  -2.02  10.25  41.75 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)  
(Intercept)          150.70989   68.02395   2.216   0.0345 *
temperature           -4.72276    2.11121  -2.237   0.0329 *
humidity              -1.29670    1.11576  -1.162   0.2543  
temperature:humidity   0.09728    0.03940   2.469   0.0195 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.3 on 30 degrees of freedom
Multiple R-squared:  0.626,	Adjusted R-squared:  0.5886 
F-statistic: 16.74 on 3 and 30 DF,  p-value: 1.414e-06
```

anova(model3, model4)

```
> anova(model3, model4) # the interaction improved the model
Analysis of Variance Table

Model 1: aphids ~ temperature + humidity
Model 2: aphids ~ temperature + humidity + temperature:humidity
  Res.Df   RSS Df Sum of Sq     F  Pr(>F)  
1     31 16369                              
2     30 13605  1      2764 6.095 0.01947 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Diagnostics

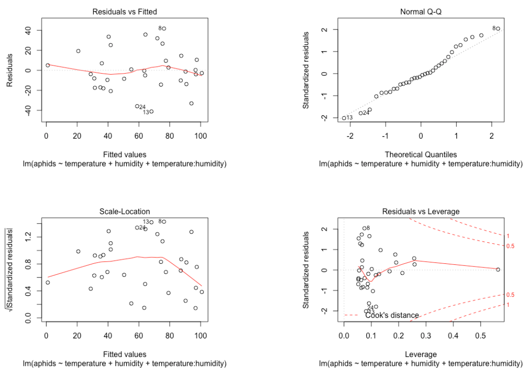
- plot()
- rstudent()
- dbetas()
- dffits()
- covratio()
- cooks.distance()

Plot provides the graphical representation

The other functions provide calculated values (and graphical tools with olsrr) for the respective measures, which can be useful if you would like to search for specific values that are beyond threshold values, etc.

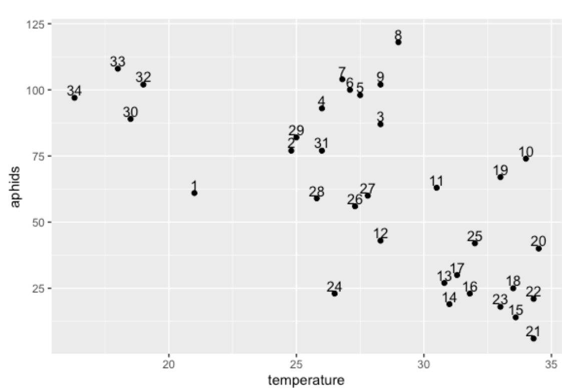
Follows from the same ideas in our peanut example, are there outliers, influential points, etc.?

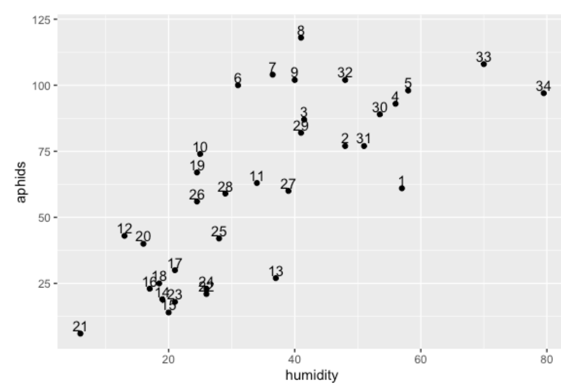
Using model 4

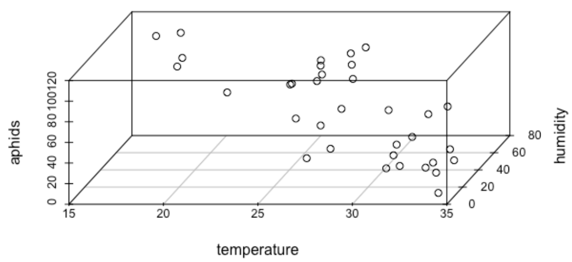


Exploring the values in a bit more detail

```
> aphids_data
   lot aphids temperature humidity
1    1     61        21.0     57.0      21  21      6        34.3      6.0
2    2     77        24.8     48.0      22  22     21        34.3     26.0
3    3     87        28.3     41.5      23  23     18        33.0     21.0
4    4     93        26.0     56.0      24  24     23        26.5     26.0
5    5     98        27.5     58.0      25  25     42        32.0     28.0
6    6    100        27.1     31.0      26  26     56        27.3     24.5
7    7    104        26.8     36.5      27  27     60        27.8     39.0
8    8    118        29.0     41.0      28  28     59        25.8     29.0
9    9    102        28.3     40.0      29  29     82        25.0     41.0
10  10     74        34.0     25.0      30  30     89        18.5     53.5
11  11     63        30.5     34.0      31  31     77        26.0     51.0
12  12     43        28.3     13.0      32  32    102        19.0     48.0
13  13     27        30.8     37.0      33  33    108        18.0     70.0
14  14     19        31.0     19.0      34  34     97        16.3     79.5
15  15     14        33.6     20.0
16  16     23        31.8     17.0
17  17     30        31.3     21.0
18  18     25        33.5     18.5
19  19     67        33.0     24.5
20  20     40        34.5     16.0
```





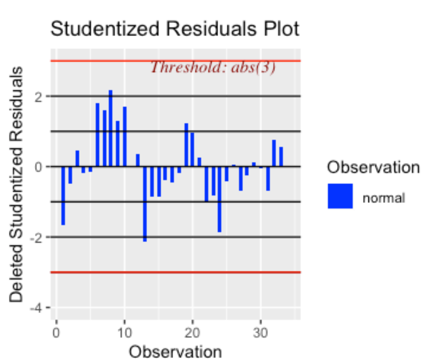


Visual examination of key measures using *olsrr* package

- Interesting package that takes the calculations and offers graphical visualization of the results for each assumption (you can teach an old dog new tricks…)

- https://www.rdocumentation.org/packages/olsrr/versions/0.5.2
  - Recommended for the beginner/intermediate R user and focused on ordinary least square regression models

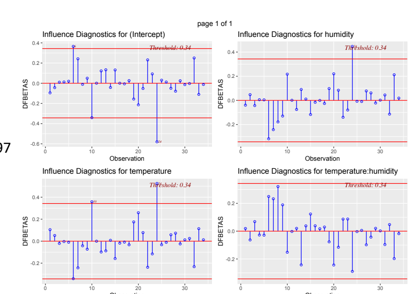
Student residuals



dfbetas

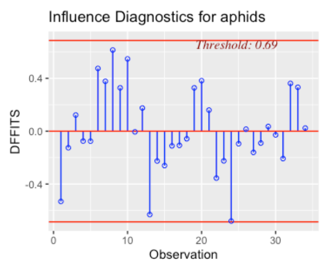
- Measures the effect (change) in the regression coefficient
  - We are looking for observations that may influence parameter estimation

Critical value: 2/sqrt(n)
Model: 2/sqrt(34) = 0,342997

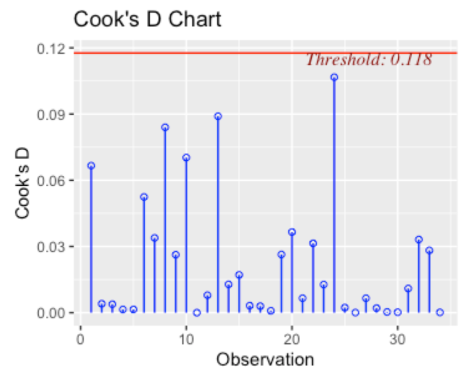


dffits

- Standardized measure (scaled) that represents the change in predicted value for each observation when it is eliminated
- Large value = high influence
- Crticial value:
  - 2
  - 2*sqrt(p/n)
- Ejemplo:
  - 2*sqrt(4/34) = 0.69



Cook's Distance

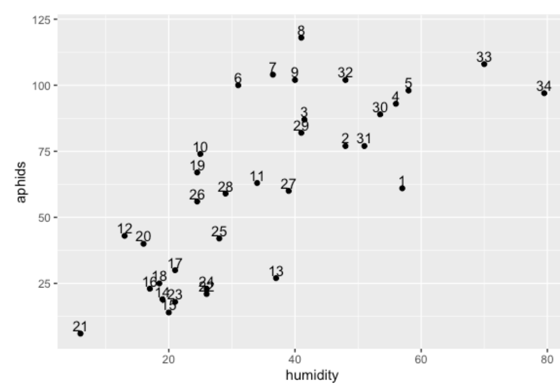


covratio

- Measure of the change in the determinant function of the covariance matrix for each value when eliminated

Valor crítico:
covratio>1+(3*p)/n
covratio<1-(3*p)/n
Ejemplo:
1+(3*4/34) = 1.3529
1-(3*4/34) = 0.6471

```
> covratio(model4)
        1         2         3         4         5         6         7 
0.8701623 1.1826549 1.1927973 1.3069641 1.4520128 0.8020052 0.8681666 
        8         9        10        11        12        13        14 
0.6819798 0.9680977 0.8603410 1.2120133 1.3903653 0.6965068 1.1033099 
       15        16        17        18        19        20        21 
1.1335691 1.2134151 1.1742397 1.2513168 0.9974108 1.1632187 1.5283511 
       22        23        24        25        26        27        28 
1.1085842 1.1210624 0.8245063 1.1827248 1.2741137 1.1331044 1.2849849 
       29        30        31        32        33        34 
1.2223690 1.3735892 1.1795590 1.3049070 1.4748564 2.6250656 
```



Let’s take a pause…next step

- Aphid example: how well does this model look?
- Are there other models we should consider?
- In the next step, we will expand on the model by looking at different tools to identify “best” models and compare those based on tools like AIC, BIC, and Mallow’s Cp

Let’s build this considering what we have seen so far…

- Model a (additive):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

- Model b (full model with quadratic terms):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2 + \varepsilon$$

- Transformation? Model c [count data -> ln()]

$$\ln(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Model a (additive model):

```
> summary(model_a) #R^2 = 0.55

Call:
lm(formula = aphids ~ temperature + humidity)

Residuals:
    Min      1Q  Median      3Q     Max 
-35.393 -14.006  -3.198  10.335  49.265 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)  35.8255    53.5388   0.669  0.50835   
temperature  -0.6765     1.4360  -0.471  0.64089   
humidity      1.2811     0.4243   3.019  0.00504 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.98 on 31 degrees of freedom
Multiple R-squared:   0.55,  Adjusted R-squared:  0.521 
F-statistic: 18.95 on 2 and 31 DF,  p-value: 4.212e-06
```

Model b (full model):

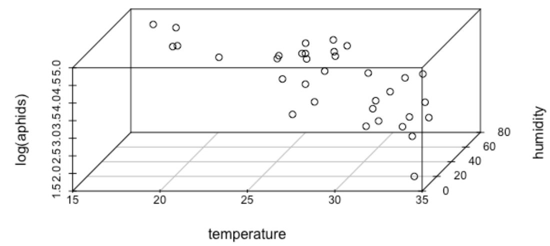
```
> summary(model_b) #R^2 = 0.63

Call:
lm(formula = aphids ~ temperature + humidity + I(temperature^2) + 
    I(humidity^2) + temperature:humidity)

Residuals:
    Min      1Q  Median      3Q     Max 
-41.700 -12.220  -1.462  10.894  41.673 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          143.069144 610.542500   0.234    0.816
temperature           -5.639044  33.900957  -0.166    0.869
humidity              -0.182206   8.875236  -0.021    0.984
I(temperature^2)       0.029174   0.476345   0.061    0.952
I(humidity^2)         -0.008121   0.036214  -0.224    0.824
temperature:humidity   0.078534   0.233701   0.336    0.739

Residual standard error: 21.98 on 28 degrees of freedom
Multiple R-squared:  0.6281,  Adjusted R-squared:  0.5617 
F-statistic:  9.46 on 5 and 28 DF,  p-value: 2.285e-05
```

```
> model_c<-with(aphids_data, lm(log(aphids)~temperature+humidity))
> anova(model_c)
Analysis of Variance Table

Response: log(aphids)
            Df Sum Sq Mean Sq F value    Pr(>F)
temperature  1 6.8912  6.8912 24.9995 2.148e-05 ***
humidity     1 2.1424  2.1424  7.7722  0.008982 **
Residuals   31 8.5453  0.2757
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(model_c)

Call:
lm(formula = log(aphids) ~ temperature + humidity)

Residuals:
     Min       1Q   Median       3Q      Max
-1.24697 -0.45491  0.03205  0.37979  0.78184

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.395151   1.223267   2.775  0.00926 **
temperature -0.015120   0.032810  -0.461  0.64814
humidity     0.027030   0.009695   2.788  0.00898 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.525 on 31 degrees of freedom
Multiple R-squared:  0.5139,  Adjusted R-squared:  0.4825
F-statistic: 16.39 on 2 and 31 DF,  p-value: 1.394e-05
```

Model c (transformation)

Modelo c:

ln(aphids) = temperature + humidity



Comparison of model a and b

```
> anova(model_a, model_b)
Analysis of Variance Table

Model 1: aphids ~ temperature + humidity
Model 2: aphids ~ temperature + humidity + I(temperature^2) +
I(humidity^2) +
    temperature:humidity
  Res.Df   RSS Df Sum of Sq     F Pr(>F)
1     31 16369
2     28 13527  3    2842.1 1.961 0.1427
```

Indicates that model b does not significantly improve the model (over-parameterized?)

Summary (partial)

- Of the three models (a, b, and c)
  - a versus b: while the model improved some of the predictive value with b, there are probably factors that are not needed to best explain the relationship
  - a and c: the transformation did not improve the model
    - One possibility to consider with count data would be to switch to a generalized linear model and use a Poisson distribution

Is there a model that better represents the observations?

- Given our two models (a and b), is there a better model that reflects the process?
- Our full model includes: main effect terms, interaction term, and quadratic model terms (based on the graphical results)
- We will apply three methods to build the models:
  - Manually = add model parameters at each step and make decisions about the relative fit
    - Here, we will rely extensively on anova(model *X*, model *Y*) to compare the models since there is a natural nesting of one model within another
  - Stepwise method (forward, backward, both)
  - Best subsets (takes the full model and looks at different combinations of factors)

Stepwise methods

- Uses a search algorithm
  - Forward selection: starting from a null model, add variables based on some inclusion critera to keep or remove the variable, and the process continues for the rest of the variables until we arrive at the defined full model
  - Backward selection: similar, but this time we start from the full model and work towards a simpler model
  - Both directions: we apply the search algorithm working simultaneously with a forward and backward mindset
  - Ideally: all methods end the same model (we will see this is not always the case)
  - In R: Variable selection is based on AIC (Akaike Information Criterion)

Best subsets

- Method that looks at different models by considering combinations of the independent variables
  - For example, if we have four possible factors for a model,
    - This approach will look at the best models for only one, for two, for three, and with all factors
    - Comparison methods in R (package = leaps, function = *regsubsets*) :
      - Adjusted coefficient of determination
      - Mallow's Cp
      - Schwartz criterion (Bayesian information criterion)

AIC

- Measure of the relative quality of a statistical model
- Balance the trade-off between the model fit and model complexity

$$AIC = 2p - 2\ln(L)$$

Number of parameters

Maximum value for the likelihood function of the estimated model

"General" = the preferred model will have the mínimum AIC value, what we are doing is penalizing the model for having greater numbers of factors

BIC

- Like AIC this is a method that provides a selection criterion for a finite number of models
  - Based on likelihood functions

$$BIC = -2 * \ln(L) + p\ln(n)$$

Maximum value for the likelihood function of an estimated model

Number of parameters

Number of observations

"General" = the preferred model will have the mínimum value for BIC and the formula penalizes more complex models (i.e., greater number of parameters)

Mallow's C_p

- Equivalent method to AIC

$$C_p = \frac{SC_{error.p}}{CM_{error}} - n + 2p$$

Sum of squares for the model with p factors

Full model

"General" = preferred model with have a mínimum value for C_p

Manually (see R notes)

Model	R2-ajustado
Temperature + Humidity + Temperature^2 + Humidity^2 + Temperature:Humidity	0.5617
Temperature + Humidity + Temperature^2 + Humidity^2	0.5752
Temperature + Humidity + Temperature^2	0.565
Temperature + Humidity + Humidity^2	**0.5832**
Humidity + Humidity^2	**0.5869**
Temperatura + Humedad	0.521

This is a systematic approach to comparing the models. Right now, the best model may be: aphids~temperature + humidity + humidity^2. We could consider a humidity-only model as well?

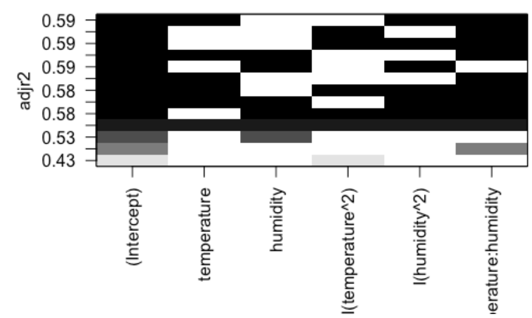
Stepwise algorithms (based on AIC)

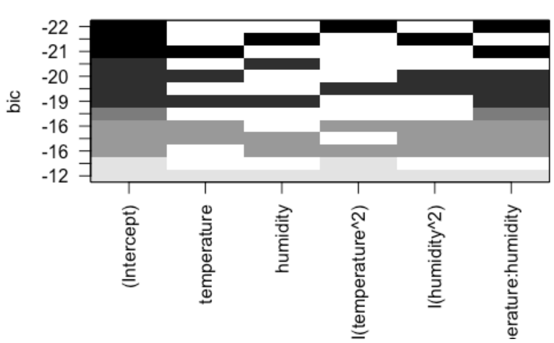
```
model_null <- lm(aphids~1, data=aphids_data)
model_full <- model_b
```

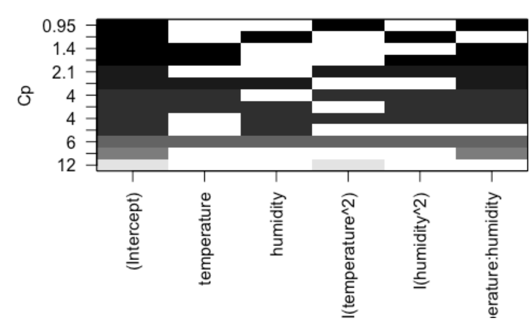
- Forward:
  - **aphids ~ humidity + humidity^2 (AIC = 210.97)**
- Backwards:
  - **aphids ~ humidity + humidity^2 (AIC = 210.97)**
- Both directions
  - **aphids ~ humidity + humidity^2 (AIC = 210.97)**

Best subsets (nbest=3)

- R^2-ajustado (~0,59 para todo):
  - **Temperature + Humidity^2 + Temperature:Humidity**
  - **Temperature^2 + Temperature:Humidity**
  - **Temperature^2 + Humidity^2 + Temperature:Humidity**
  - **Temperature + Humidity + Temperature:Humidity**
  - **Humidity + Humidity^2**
- BIC (-22):
  - **Temperature^2 + Temperature:Humidity**
  - **Humidity + Humidity^2**
- C_p:
  - #1: **Temperature^2 + Temperature:Humidity**
  - #2: **Humidity + Humidity^2**







Modeling thoughts

General principles

Our general principles for building regression models for prediction are as follows:

1. Include all input variables that, for substantive reasons, might be expected to be important in predicting the outcome.
2. It is not always necessary to include these inputs as separate predictors—for example, sometimes several inputs can be averaged or summed to create a "total score" that can be used as a single predictor in the model.
3. For inputs that have large effects, consider including their interactions as well.
4. We suggest the following strategy for decisions regarding whether to exclude a variable from a prediction model based on expected sign and statistical significance (typically measured at the 5% level; that is, a coefficient is "statistically significant" if its estimate is more than 2 standard errors from zero):
   (a) If a predictor is not statistically significant and has the expected sign, it is generally fine to keep it in. It may not help predictions dramatically but is also probably not hurting them.
   (b) If a predictor is not statistically significant and does not have the expected sign (for example, incumbency having a negative effect on vote share), consider removing it from the model (that is, setting its coefficient to zero).
   (c) If a predictor *is* statistically significant and does not have the expected sign, then think hard if it makes sense. (For example, perhaps this is a country such as India in which incumbents are generally unpopular; see Linden, 2006.) Try to gather data on potential lurking variables and include them in the analysis.
   (d) If a predictor is statistically significant and has the expected sign, then by all means keep it in the model.

These strategies do not completely solve our problems but they help keep us from making mistakes such as discarding important information. They are predicated on having thought hard about these relationships *before* fitting the model. It's always easier to justify a coefficient's sign after the fact than to think hard ahead of time about what we expect. On the other hand, an explanation that is determined after running the model can still be valid. We should be able to adjust our theories in light of new information.

Source. Gelman y Hall. 2007. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge.

Which is the best model and why?

- All models have moderate predictive value
- One model consistently found was based on humidity with a quadratic term
- Biologically, does this model make sense?
- We can go back and study the behaviour on the final model using the various tools we saw earlier

Exercise: Fusarium modeling exercise.R

- To put into practice what we saw in the last section
- Data includes:
  - lot = represents an individual plot level observation
  - yield = kilograms per hectare
  - fdk = fusarium damaged kernels (%)
  - incidence = incidence of Fusarium head blight (%)
  - severity = severity of Fusarium head blight in heads (%)
  - moisture = grain moisture (%)
  - don = concentration of vomitoxin (ppm)
- We will read the data into R using "Import Dataset"
- Focus here is on using the automated algorithms

Concluding thoughts

- R has numerous tools to effectively develop, test, and validate models (always has been the software's strength)
- Effective modeling requires decisions based on the overall goal of the model (predictive, theoretical, summary, etc.)
- What are the parameters and what do they mean in the context of the measured variable of interest?
  - Our current work (machine/deep learning) has required substantial time focused on identifying the proper parameters to avoid the "junk in, junk out" result
- What will define a good model is a function of the model goal