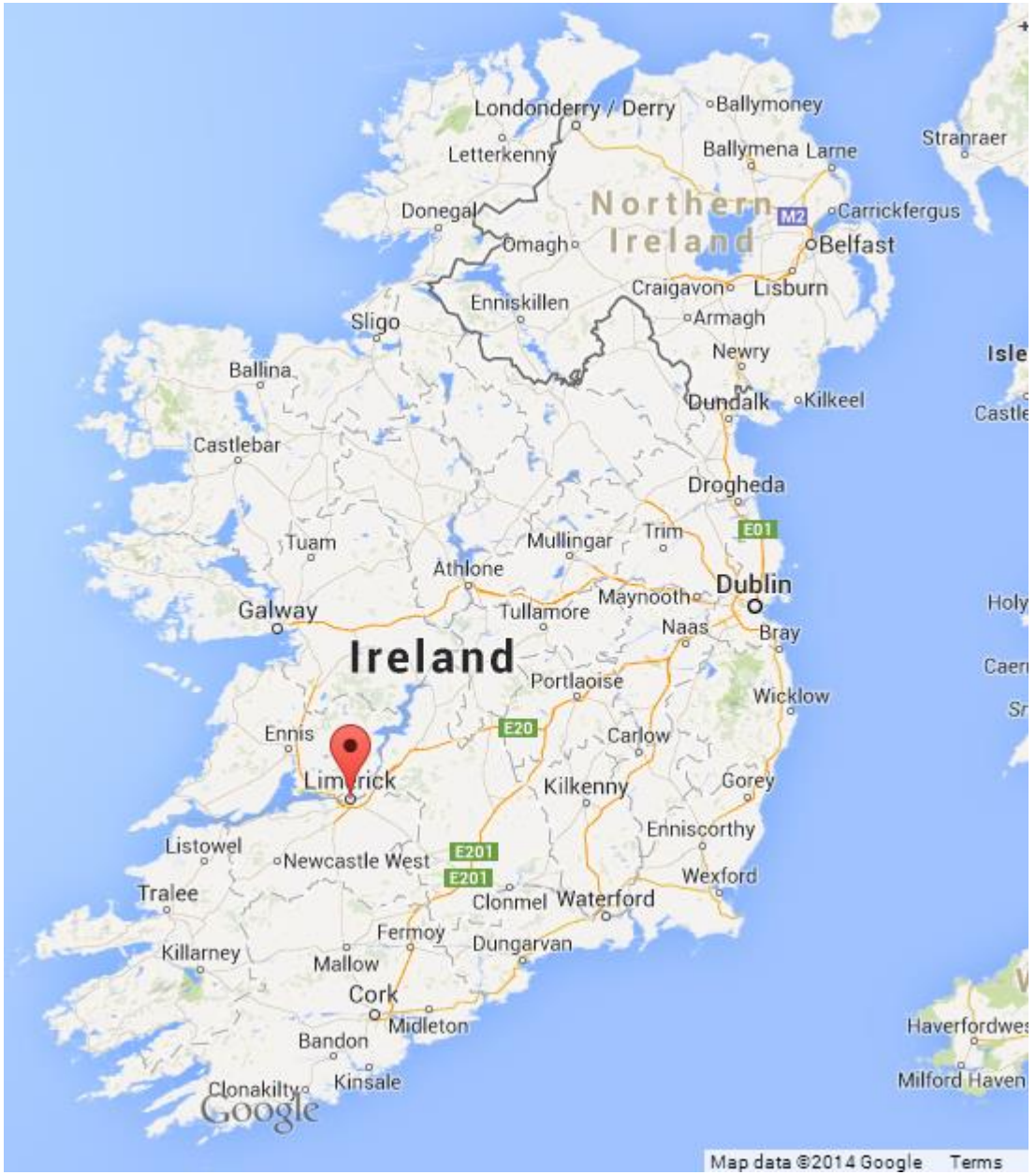
OSUN R Users Community

Kevin O'Brien

Kevin O'Brien

- Forestry Data Scientist based in the West of Ireland
  - Also in London, UK a lot
- Why R? - Community Team lead & Webinars co-ordinator
- Python Ireland - Director
- JuliaCon 2022 - Social Media Chair
- was previously (what is now known) as a "Research Software Engineer" in a University.


Londonderry / Derry
Ballymoney
Letterkenny
Ballymena
Larne
Stranraer
Donegal
Northern Ireland
M2
Carrickfergus
Omagh
Belfast
Craigavon
Lisburn
Enniskillen
Armagh
Sligo
Newry
Ballina
Dundalk
Kilkeel
Castlebar
Drogheda
E01
Tuam
Mullingar
Trim
Athlone
Dublin
Maynooth
Galway
Tullamore
Naas
Bray
Ireland
Portlaoise
Wicklow
Ennis
E20
Carlow
Limerick
Kilkenny
Gorey
Enniscorthy
Listowel
Newcastle West
E201
Wexford
Tralee
Clonmel
Waterford
Fermoy
Dungarvan
Killarney
Mallow
Cork
Midleton
Bandon
Kinsale
Clonakilty
Google
Map data ©2014 Google
Terms

Forestry

- R is very useful in Forestry
    - Diameter at Breast Height / Height
    - Growth Curves and Yield Models
    - Statistical Analysis and Data Visualization
- Other R packages have been ***VERY*** useful

(Source: Timbeter.Com)

Intended Audience

- Career Young Statisticians and Data Scientists.
- Solid Foundation in R and/or Python.
- Familiarity with Machine Learning / Deep Learning
- Github

What's next?

- Transition to Research Software Engineer?
- Apply skills base to real world problems

Scope

Machine Learning - types of analysis

Many Types of ML Problems

- Clustering
- Classification
- Regression

Focus on Regression

- Predicting a numeric value based on predictors
- Can easily generalize most of the content to all Linear Models (ANOVA)

Scope

- Mainly focusing on item-wise diagnostics

Feature Engineering

Not Covering This - but it is important

- Normalisation
- Scaling
- One-hot Encoding / Model Matrices
- Log Transformation
- Box-Cox Transformation

- {caret} package

{yardstick}

(**Digression**)

Not Really Covering Machine Learning or Binary Classifers

- Accuracy, Precision, Recall and F-Score
- NPV, PPV
- Lift and Gain Curves

Includes a lot of functionality for standard regression model metrics

My career

Former University Lecturer of Mathematics and Statistics.

Motivation

- Career young data scientists, mathematicians and statisticians.
- Job interviews.
- Career advice.
- Professional development.

Students

- Maths and Statistics Students (Teaching R)
- Health sciences, Life Sciences, Equine science, Sports Science, Food science, Biochemistry

(Emphasizing the second group more - as those subjects area have very interesting real-world applications.)

Topics

- Statistics 101
- Exploratory Data Analysis
- Linear models.
  - Robust, Polynomial Regression
  - Model Metrics
- Experimental design.
- Linear Mixed Effects models.
- Non-Parametric Statistics.
- Statistical Process Control.

Statistical Knowledge

- The talk is aimed at students and early career data professionals who have already encountered conventional regression analyses, and are familiar with the model fitting process in R (i.e. the *lm()* function).
- The talk will introduce a mixture of graphical procedures, statistical measures and hypothesis tests, which the attendees are invited to learn more about beyond the talk.
- The talk will feature the {CAR} R package [1], but all of the other functionality is available in Base R or Tidyverse.

[1] Fox, John, et al. "*The car package.*" R Foundation for Statistical Computing (2007).

Key Motivations

- All statistical models and tests have underlying mathematical assumptions on the types of conditions upon we can generate reliable results (**Hoekstra et al., 2012**).
- What this means is that before we go off and generate predictions, p-values and correlation coefficients, we need to understand whether our data fits certain assumption criteria in order to avoid Type I or II errors given the technique at hand.

Domain Knowledge

(Digression)

https://www.linkedin.com/feed/update/urn:li:activity:6765853465861267456/

- Agriculture and Food
- Health & Life Sciences
- Natural Sciences & GIS

My own career: Audiology, Equine Science, Water Quality, Milk Production, Sport Science, Forestry

Life Sciences

- Models must scrutinized thoroughly - particularly the effect of "Influential cases" and "outliers".
- Random Forest Models are insufficent in this regards - can deduce important variables, but not important cases

Introduction to Model Validation Procedures with R

- Model validation is a vital part of the statistical modelling process, but is often overlooked in statistical courses.
- This process allows the analyst to properly validate the assumptions underlying the model, once applied to the data.
- In this presentation, we will look at residual analysis and influence measures for linear models, with some associated topics.

R Packages


modelbased
report
see
6
3
0
correlation
effectsize
easystats
bayestestR
insight
performance
8
9
7
parameters
datawizard

Pedagogical Effects on Statistics

(More the general science students, rather than the Maths Science students)

- Designing a Statistics course is an exercises in compromise, particularly Statistics 101 courses. For the curriculum design in universities, some choices are made that reflect scheduling needs.
- 12-14 Week semesters with 2 or 3 lectures per week. Many students will have only 1, but maybe 2 Statistics Modules in their 4 year degree programme.
- Many would benefit from far more Statistics and Statistical Computing content on their curriculum.
- Lecturers must create a pen-and-paper exam paper at the end of the semester, with a transparent rubric. Conventions must be followed.
- Example: You ***MUST*** cover the Normal Distribution - but you are caught for time to explain all the reasons ***WHY*** it is so important to know.
- DANGER! Pedagogical design can create misconceptions about the relative importance of various statistical topics.

Job Advertisement (Financial Services)

The Data Scientist has a deep statistical knowledge and strong quantitative skills who will work as part of the Model Validation team to independently evaluate the efficacy of the design and implementation of [COMPANY]'s high performance modeling solutions across our Big Data analytics landscape.

Quantitative models support some of the most important processes and decisions at [COMPANY], and the Model Validation team is responsible for their effective challenge.

The Model Validation team conducts testing and provides a ***critical review of conceptual soundness and model performance***, producing technical reports describing the results of validation, and interface with internal stakeholders and regulators to communicate findings in model risk.

The Model Validation team assesses all models across [COMPANY] and this role offers the unique opportunity to acquire a wide variety of experience of different models supporting a wide range of relevant processes including credit risk, fraud etc.

Job Advertisement (Financial Services)

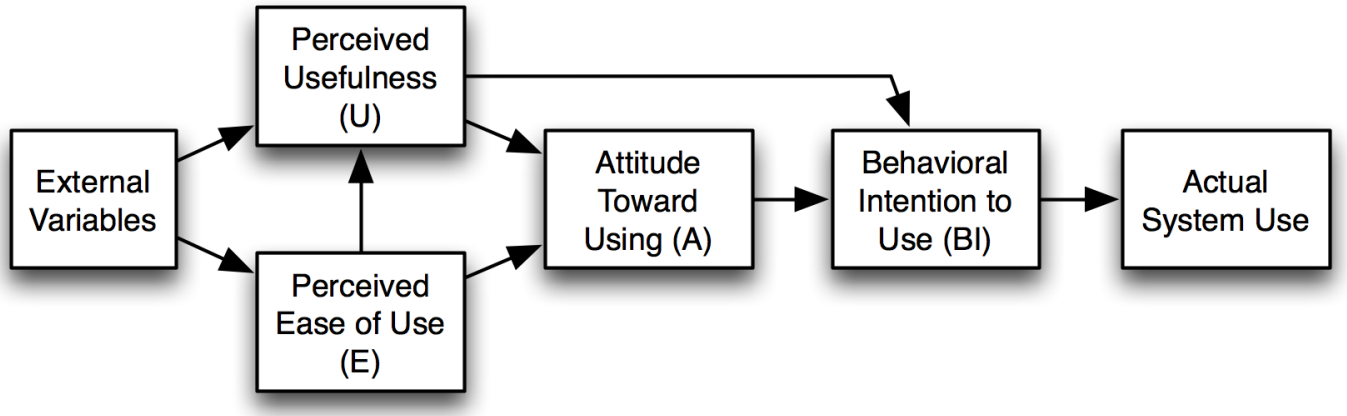
(Continued)

You will be responsible for sourcing and generating validation datasets, analysis and critical review of related data, assessment of model performance and ultimately delivering validation reports on the status of the underlying models. You will work with a wide, diverse range of data science/model development teams.

You will have the opportunity to work with one of the world's largest financial datasets and the most advanced analytics and machine learning approaches that deliver near real-time predictions and recommendations.

Technology Acceptance Model

Why do people use some tools and not others?



(Source: Wikipedia)

Technology Acceptance Model

Perceived usefulness (PU) – This was defined by Fred Davis as "the degree to which a person believes that using a particular system would enhance their job performance". It means whether or not someone perceives that technology to be useful for what they want to do.

Perceived ease-of-use (PEOU) – Davis defined this as "the degree to which a person believes that using a particular system would be free from effort" (Davis 1989). If the technology is easy to use, then the barriers conquered. If it's not easy to use and the interface is complicated, no one has a positive attitude towards it.

Exploratory Data Analysis

BE THOROUGH

- Domain Knowledge
- Summary Statistics
- Data Visualization
- Outlier Detection
- Missing Data Analysis

Remark:

Cluster Analysis (e.g. K-means) can be a very useful EDA procedure.

Exploratory Data Analysis

Data Inspection

- {inspectdf}
- {janitor}

Data Visualization

- {WVPlots}
- {ggally}

tidyverse

- {broom} and {modelr} (succeeded by {tidymodels})

- `dplyr::tally()`
- `dplyr::distinct()`

Useful Packages (a selection)

- {A3}
- {arsenal}
- {analytics}
- {gdata}
- {descriptr}
- {furniture}
- {rockchalk}
- {yardstick}

{inspectdf}

inspectdf: Inspection, Comparison and Visualisation of Data Frames

inspectdf is collection of utilities for columnwise summary, comparison and visualisation of data frames. Functions are provided to summarise missingness, categorical levels, numeric distribution, correlation, column types and memory usage. The package has three aims: to speed up repetitive checking and exploratory tasks for data frames

{inspectdf}

Key functions

- `inspect_types()`: summary of column types
- `inspect_mem()`: summary of memory usage of columns
- `inspect_na()`: columnwise prevalence of missing values
- `inspect_cor()`: correlation coefficients of numeric columns
- `inspect_imb()`: feature imbalance of categorical columns
- `inspect_num()`: summaries of numeric columns
- `inspect_cat()`: summaries of categorical columns

{inspectdf}

```
library(inspectdf)

# Load dplyr for starwars data & pipe
library(dplyr)
# Single dataframe summary
inspect_cat(starwars)
```

```
## # A tibble: 8 x 5
##   col_name     cnt common    common_pcnt levels
##   <chr>      <int> <chr>           <dbl> <named list>
## 1 eye_color     15 brown           24.1  <tibble [15 x 3]>
## 2 gender         3 masculine      75.9  <tibble [3 x 3]>
## 3 hair_color    13 none            42.5  <tibble [13 x 3]>
## 4 homeworld     49 Naboo           12.6  <tibble [49 x 3]>
## 5 name          87 Ackbar           1.15 <tibble [87 x 3]>
## 6 sex            5 male            69.0  <tibble [5 x 3]>
## 7 skin_color    31 fair            19.5  <tibble [31 x 3]>
## 8 species       38 Human           40.2  <tibble [38 x 3]>
```

{inspectdf}

```
# Paired dataframe comparison
inspect_cat(starwars, starwars[1:20, ])
```

```
## # A tibble: 8 x 5
##   col_name      jsd     pval lvls_1             lvls_2
##   <chr>       <dbl>    <dbl> <named list>      <named list>
## 1 eye_color  0.0936 7.08e- 1 <tibble [15 x 3]> <tibble [8 x 3]>
## 2 gender     0.0387 3.38e- 1 <tibble [3 x 3]>  <tibble [2 x 3]>
## 3 hair_color 0.261  9.04e- 4 <tibble [13 x 3]> <tibble [10 x 3]>
## 4 homeworld  0.394  2.21e- 2 <tibble [49 x 3]> <tibble [11 x 3]>
## 5 name       0.573  9.35e-11 <tibble [87 x 3]> <tibble [20 x 3]>
## 6 sex        0.0526 5.19e- 1 <tibble [5 x 3]>  <tibble [4 x 3]>
## 7 skin_color 0.288  1.58e- 1 <tibble [31 x 3]> <tibble [10 x 3]>
## 8 species    0.300  7.86e- 2 <tibble [38 x 3]> <tibble [6 x 3]>
```

{inspectdf}

```
# Grouped dataframe summary
starwars %>% group_by(species) %>% inspect_cat()
```

```
## # A tibble: 266 x 6
## # Groups:   species [38]
##    species  col_name     cnt common        common_pcnt levels
##    <chr>    <chr>      <int> <chr>               <dbl> <named list>
##  1 Aleena   eye_color      1 unknown               100 <tibble [1 x 3]>
##  2 Aleena   gender         1 masculine             100 <tibble [1 x 3]>
##  3 Aleena   hair_color     1 none                  100 <tibble [1 x 3]>
##  4 Aleena   homeworld      1 Aleen Minor           100 <tibble [1 x 3]>
##  5 Aleena   name           1 Ratts Tyerell         100 <tibble [1 x 3]>
##  6 Aleena   sex            1 male                  100 <tibble [1 x 3]>
##  7 Aleena   skin_color     1 grey, blue            100 <tibble [1 x 3]>
##  8 Besalisk eye_color      1 yellow                100 <tibble [1 x 3]>
##  9 Besalisk gender         1 masculine             100 <tibble [1 x 3]>
## 10 Besalisk hair_color     1 none                  100 <tibble [1 x 3]>
## # ... with 256 more rows
```

WVPlots: Common Plots for Analysis

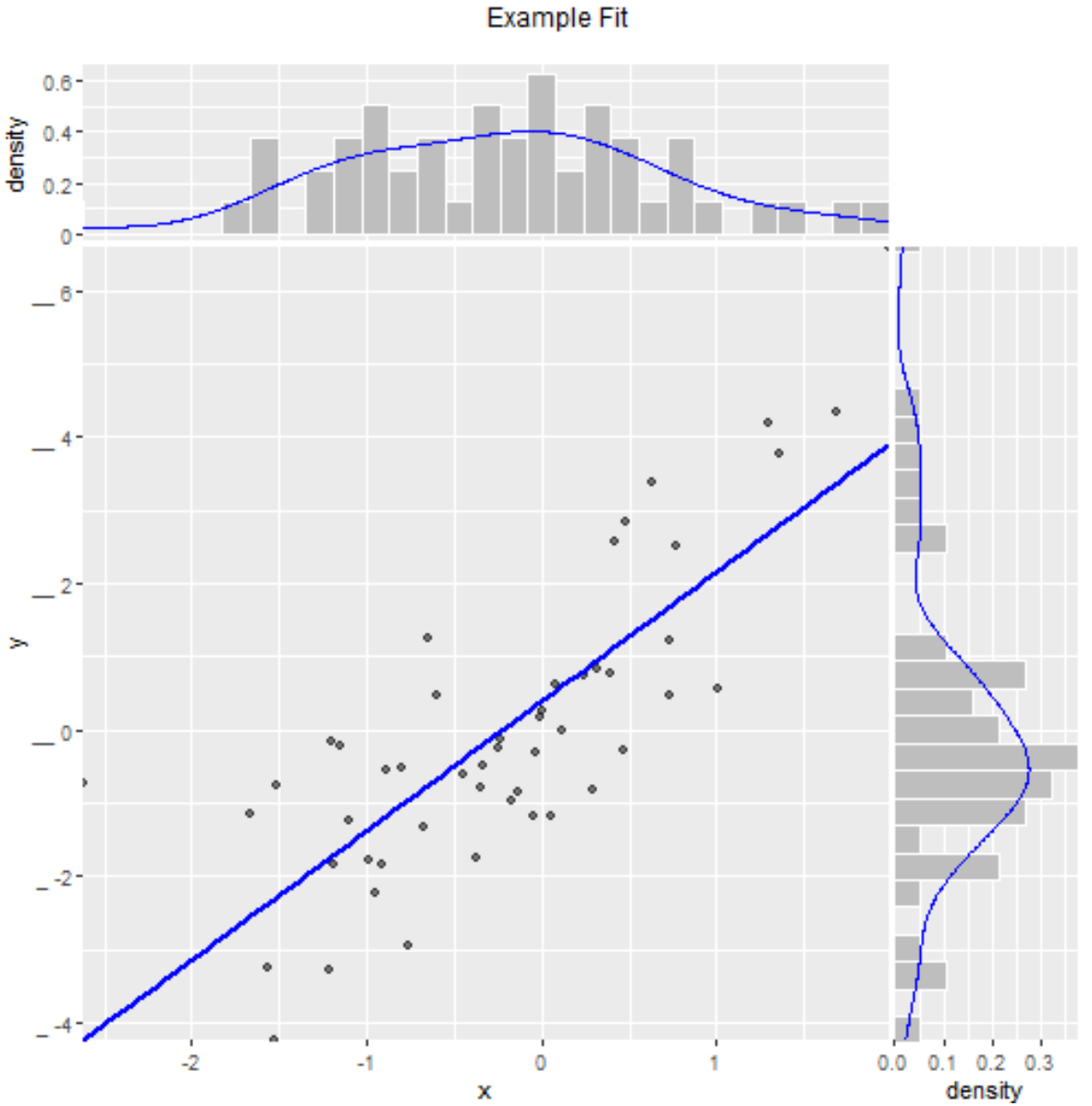
Select data analysis plots, under a standardized calling interface implemented on top of 'ggplot2' and 'plotly'. Plots of interest include: 'ROC', gain curve, scatter plot with marginal distributions, conditioned scatter plot with marginal densities, box and stem with matching theoretical distribution, and density with matching theoretical distribution.

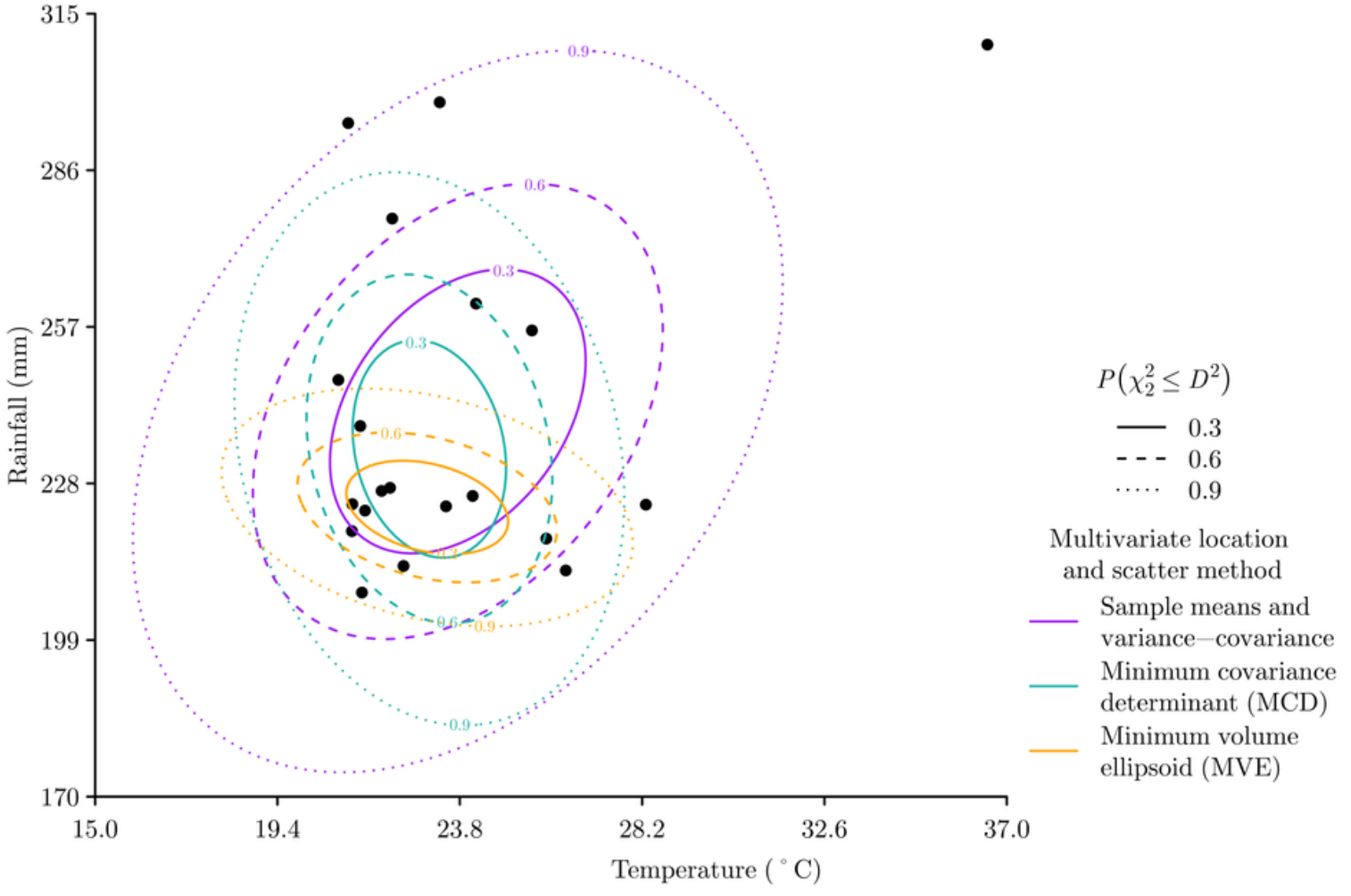
```
set.seed(34903490)
x = rnorm(50)
y = 0.5*x^2 + 2*x + rnorm(length(x))
frm = data.frame(
  x = x,
  y = y,
  yC = y>=as.numeric(quantile(y,probs=0.8)),
  stringsAsFactors = FALSE)

frm$absY <- abs(frm$y)
frm$posY = frm$y > 0
```

Scatterplot

Scatterplot with smoothing line through points.

```
WVPlots::ScatterHist(frm, "x", "y", title="Example Fit")
```


Example Fit
density
0.6
0.4
0.2
0
y
6
4
2
0
-2
-4
-2
-1
0
1
x
0.0
0.1
0.2
0.3
density


315
286
257
228
199
170
Rainfall (mm)
15.0
19.4
23.8
28.2
32.6
37.0
Temperature (°C)
0.9
0.6
0.3
0.3
0.6
0.6
0.9
0.9
$P(\chi_2^2 \leq D^2)$
0.3
0.6
0.9
Multivariate location and scatter method
Sample means and variance–covariance
Minimum covariance determinant (MCD)
Minimum volume ellipsoid (MVE)

Mahalanobis Distance

```
library(faraway)

data(cheddar)
```

cheddar: Taste of Cheddar cheese

In **{faraway}**: Functions and Datasets for Books by Julian Faraway

Description In a study of cheddar cheese from the LaTrobe Valley of Victoria, Australia, samples of cheese were analyzed for their chemical composition and were subjected to taste tests. Overall taste scores were obtained by combining the scores from several tasters.

- **taste** - a subjective taste score
- **Acetic** - concentration of acetic acid (log scale)
- **H2S** - concentration of hydrogen sulfice (log scale)
- **Lactic** - concentration of lactic acid

Step 1: Create the dataset.

```
head(cheddar)
```

```
##   taste Acetic   H2S Lactic
## 1  12.3  4.543 3.135   0.86
## 2  20.9  5.159 5.043   1.53
## 3  39.0  5.366 5.438   1.57
## 4  47.9  5.759 7.496   1.81
## 5   5.6  4.663 3.807   0.99
## 6  25.9  5.697 7.601   1.09
```

Step 2: Calculate the Mahalanobis distance for each observation.

Next, we'll use the built-in `mahalanobis()` function in R to calculate the Mahalanobis distance for each observation, which uses the following syntax:

```
mahalanobis(x, center, cov)
```

where:

- `x`: matrix of data
- `center`: mean vector of the distribution
- `cov`: covariance matrix of the distribution

Implementation

The following code shows how to implement this function for our dataset:

```
df <- cheddar[,2:4]

#calculate Mahalanobis distance for each observation

mahalanobis(df, colMeans(df), cov(df)) %>%
  head() %>%
  t()
```

```
##          [,1]     [,2]      [,3]     [,4]     [,5]     [,6]
## [1,] 4.115811 1.235341 0.7716917 1.593862 2.768398 5.713978
```

Step 3: Calculate the p-value for each Mahalanobis distance.

We can see that some of the Mahalanobis distances are much larger than others. To determine if any of the distances are statistically significant, we need to calculate their p-values.

The p-value for each distance is calculated as the p-value that corresponds to the Chi-Square statistic of the Mahalanobis distance with k-1 degrees of freedom, where k = number of variables. So, in this case we'll use a degrees of freedom of 3-1 = 2.

Step 3: Calculate the p-value for each Mahalanobis distance.

```
#create new column in data frame to hold Mahalanobis distances
df$mahal <- mahalanobis(df, colMeans(df), cov(df))

#create new column in data frame to hold p-value for each Mahalanobis
df$p <- pchisq(df$mahal, df=2, lower.tail=FALSE)
```

Mahalanobis distance.

Step 3: Calculate the p-value for each Mahalanobis distance.

```
#view data frame
df %>%
  head() %>%
  kable(format="markdown")
```

Acetic	H2S	Lactic	mahal	p
4.543	3.135	0.86	4.1158108	0.1277212
5.159	5.043	1.53	1.2353409	0.5391991
5.366	5.438	1.57	0.7716917	0.6798753
5.759	7.496	1.81	1.5938621	0.4507101
4.663	3.807	0.99	2.7683980	0.2505244
5.697	7.601	1.09	5.7139779	0.0574415

Intrepretating the output

- Typically a p-value that is less than some threshold (e.g. 0.001) is considered to be an outlier.
- In this case, all the p values are greater than 0.001
- Depending on the context of the problem, you may *omit* any outlier observation from the dataset, as they could affect the results of the analysis. (Domain knowledge is vital).

Linear Regression

Simple Linear Regression

- In simple linear regression, we predict values on one variable from the values of a second variable.
- The variable we are predicting is called the ***dependent variable*** (or response variable) and is referred to as Y.
- The variable we are basing our predictions on is called the ***independent variable*** (or predictor variable) and is referred to as X.
- Remark: When there is only one predictor variable, the prediction method is called simple regression. Linear regression can have more than one predictor variable, i.e. Multiple Linear Regression.

Simple Linear Regression

- Suppose we construct our model using n observed values of the response variable: $\{y_1, y_2, \ldots y_i \ldots y_n\}$.
- For the original data set, there is a predicted value of each case of Y that corresponds to an observed value of Y.
- The difference between an observed value of the dependent variable (y_i) and the corresponding predicted value ($\hat{y}$) is called the residual (e_i). Each data point from the data set has one residual.

Residuals

Simply put, the values of the residuals are derived as follows:

$$\text{Residual} = \text{Observed value - Predicted value}$$

$$e_i = y_i - \hat{y}_i$$

- Important theoretical assumption underlying the OLS model: the sum of the residuals should equal to zero.

$$\sum e_i = 0$$

- An extension of this is that the expected value of the residuals is 0: $\mathrm{E}(e) = 0$.

- Another Important Theoretical Assumption - The residuals are normally distributed. (more on that later)

Residual Plots

- A residual plot is a graph that shows the residuals on the vertical axis and the independent variable on the horizontal axis.
- If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a non-linear model is more appropriate.

Summary of Important Terms

Some important terms in model diagnostics, essentially a plan for this talk.

- ***Residual:*** The difference between the predicted value (based on the regression equation) and the actual, observed value.
- ***Outlier:*** In linear regression, an outlier is an observation with large residual. In other words, it is an observation whose dependent-variable value is unusual given its value on the predictor variables. An outlier may indicate a sample peculiarity or may indicate a data entry error or other problem.

Summary of Important Terms

- ***Leverage:*** An observation with an extreme value on a predictor variable is a point with high leverage. Leverage is a measure of how far an independent variable deviates from its mean. High leverage points can have a great amount of effect on the estimate of regression coefficients.
- ***Influence:*** An observation is said to be influential if removing the observation substantially changes the estimate of the regression coefficients. Influence can be thought of as the product of leverage and outlierness.
- ***Cook's distance (or Cook's D):*** A measure that combines the information of leverage and residual of the observation.

MultiCollinearity

- An important aspect in model diagnostics is checking for multicollinearity. We are not going to cover this much in this talk - but rather include in a talk about variable selection procedure.

Outliers, leverage and influence

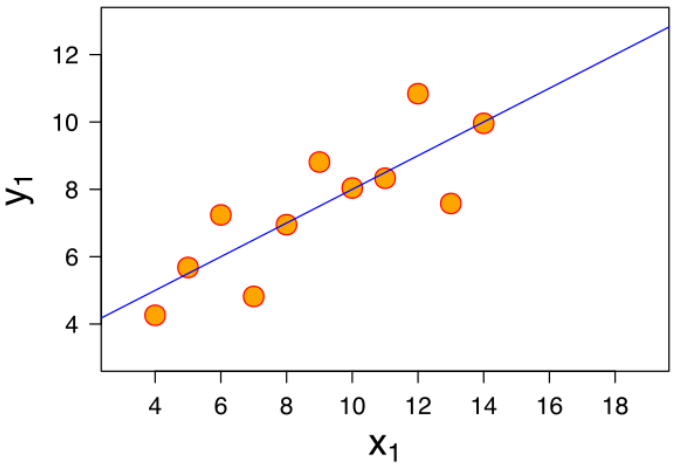
- An outlier may be defined as a data point that differs significantly from other observations.
- A high-leverage point are observations made at extreme values of independent variables.
- Both types of atypical observations will force the regression line to be close to the point.

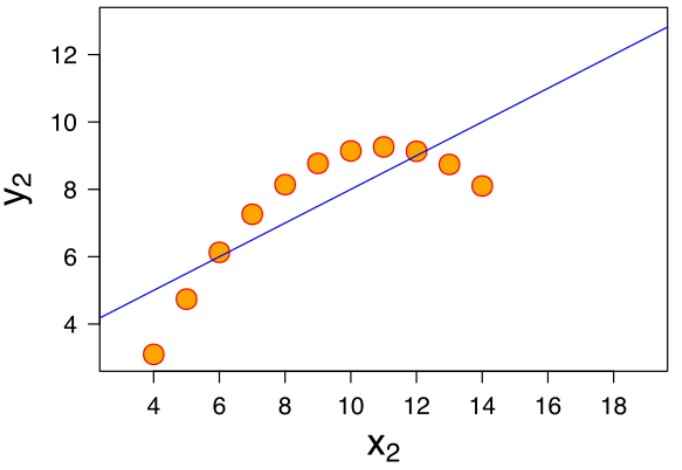
Anscombe's quartet

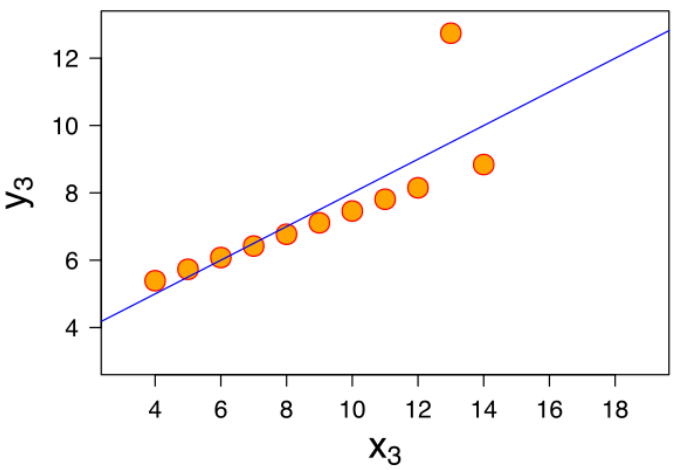
(Next Slide)

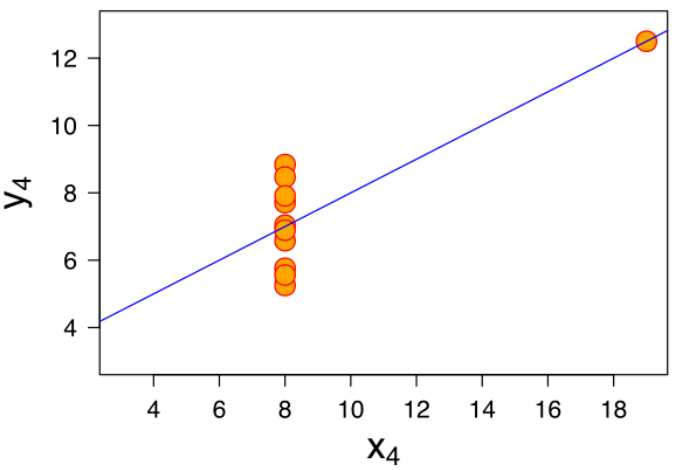
- The bottom right image has a point with high leverage
- The bottom left image has an outlying point.

Anscombe's quartet









Influential observations

- An influential observation is an observation for a statistical calculation whose deletion from the dataset would noticeably change the result of the calculation.
- Equivalently, an influential observation is one whose deletion has a large effect on the parameter estimates in a regression analysis

Metrics

- The DFBETAS are statistics that indicate the effect that deleting each observation has on the estimates for the regression coefficients.
- The DFFITS and Cook's D statistics indicate the effect that deleting each observation has on the predicted values of the model.
- {broom} R package

Model Assumptions

The Distribution of Dependent Variables

- The assumptions of normality and homogeneity of variance for linear models are not about Y, the dependent variable.
- The distributional assumptions for linear regression and ANOVA are for the distribution of $Y|X$ — (Y given X).
- The distribution of Y | X is, by definition, the same as the distribution of the residuals. Hence we can check validity by looking at the residuals.

What are those distributional assumptions of $Y|X$?

1. Independence
2. Normality
3. Constant Variance

Examining the Residual Plots

Recall:

- The mean value of the residuals is zero,
- The variance of residuals are constant across the range of measurements,
- The residuals are normally distributed,
- Residuals are independent.

A residual plot is obtained by plotting the residuals e with respect to the independent variable X or, alternatively with respect to the fitted regression line values $\hat{Y}$. Such a plot can be used to investigate whether the assumptions concerning the residuals appear to be satisfied.

Asummption of Constant Variance

- Homoscedascity (also known as constant variance) is one of the assumptions required in a regression analysis in order to make valid statistical inferences about population relationships.
- Homoscedasticity requires that the variance of the residuals are constant for all fitted values, indicated by a uniform scatter or dispersion of data points about the trend line (i.e. "The Zero Line").
- From the above plot, we can conclude that the constant variance assumption is valid. We can also see that the mean value of the residuals is close to zero. \textit{(Theoretically it is precisely zero)}.

Residual Plots

Residual Plots

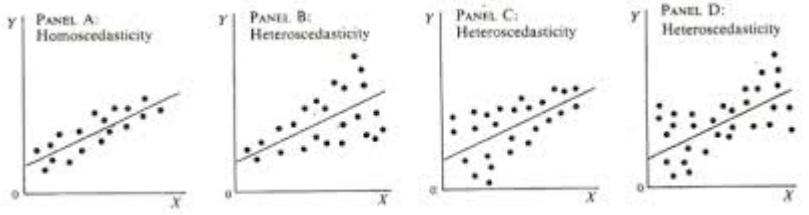
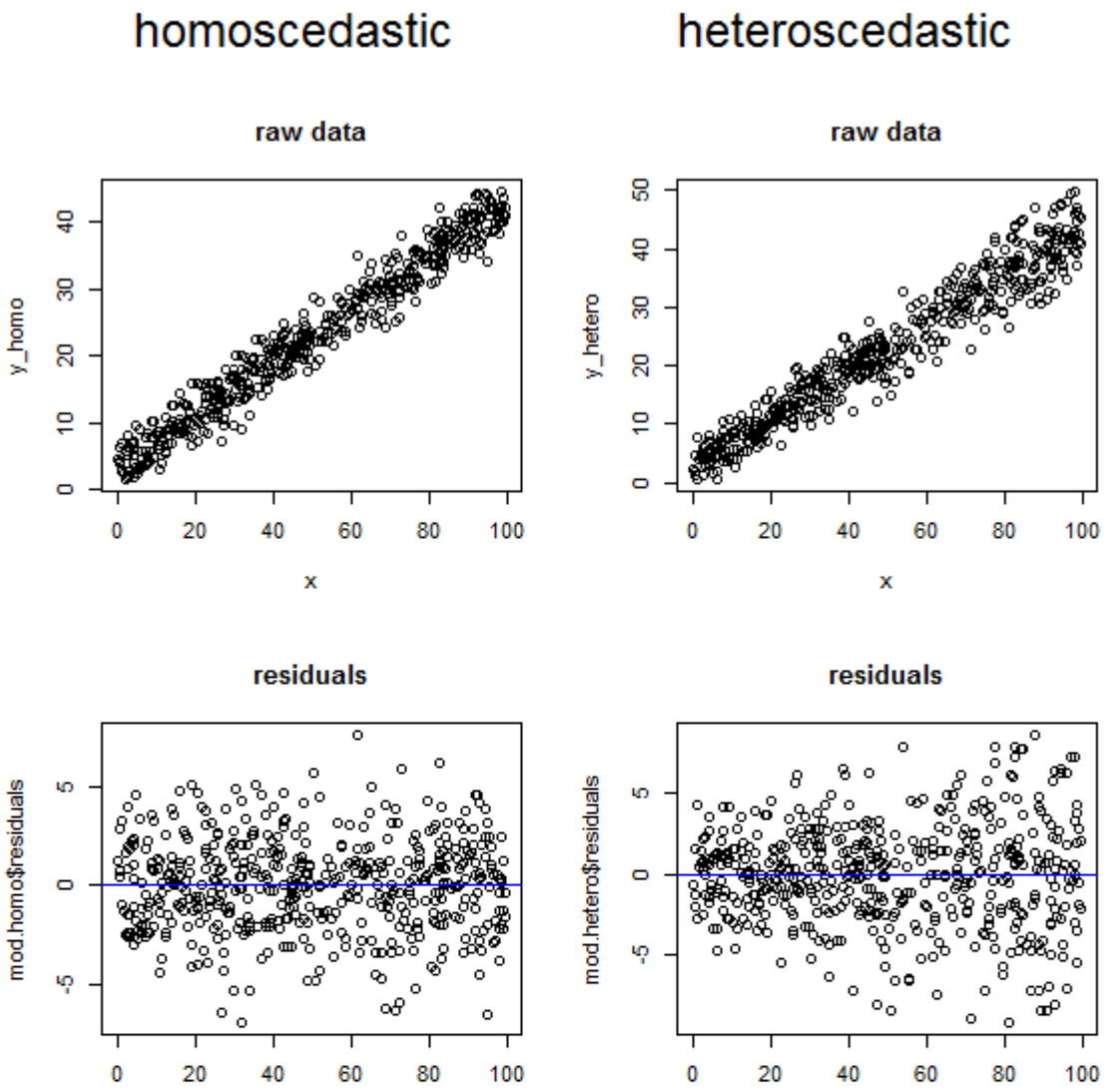


Fig. 9-1

Residual Plots



Linear modelling with R (Cheeses)

Cheddar Cheese taste

- As cheese ages, various chemical processes take place that determine the taste of the final product.
- This dataset contains concentrations of various chemicals in 30 samples of mature cheddar cheese, and a subjective measure of taste for each sample.
- The variables "Acetic" and "H2S" are the natural logarithm of the concentration of acetic asid and hydrogen sulfide respectively.
- The variable "Lactic" has not been transformed.

Reference:

- Moore, David S., and George P. McCabe (1989). Introduction to the Practice of Statistics.

Linear modelling with R (Cheeses)

- Number of cases: 30

Variable Names:

- **Case**: Sample number
- **taste**: Subjective taste test score, obtained by combining the scores of several tasters
- **Acetic**: Natural log of concentration of acetic acid
- **H2S**: Natural log of concentration of hydrogen sulfide
- **Lactic**: Concentration of lactic acid

Linear modelling with R (Cheeses)

```
library(tidyverse)
library(magrittr)
library(faraway)
data(cheddar)


head(cheddar)
```

```
##   taste Acetic   H2S Lactic
## 1  12.3  4.543 3.135   0.86
## 2  20.9  5.159 5.043   1.53
## 3  39.0  5.366 5.438   1.57
## 4  47.9  5.759 7.496   1.81
## 5   5.6  4.663 3.807   0.99
## 6  25.9  5.697 7.601   1.09
```

Linear modelling with R (Cheeses)

```
Fit_1 <- lm(taste ~ Acetic + Lactic, data = cheddar)
Fit_2 <- lm(taste ~ Acetic + H2S, data = cheddar)
Fit_3 <- lm(taste ~ H2S + Lactic, data = cheddar)

Fit_4 <- lm(taste ~ Acetic + H2S + Lactic, data = cheddar)
```

Linear modelling with R (Cheeses)

Aikaike Information Criterion

```
AIC(Fit_1)
```

```
## [1] 237.3884
```

```
AIC(Fit_2)
```

```
## [1] 233.2438
```

```
AIC(Fit_3)
```

```
## [1] 227.7838
```

```
AIC(Fit_4)
```

```
## [1] 229.7775
```

{modelr}

Compute model quality for a given dataset

Three summaries are immediately interpretible on the scale of the response variable:

- `rmse()` is the root-mean-squared-error
- `mae()` is the mean absolute error
- `qae()` is quantiles of absolute error.

{modelr}

Root Mean Square Error

```
library(modelr)
rmse(Fit_4,cheddar)
```

```
## [1] 9.431174
```

{modelr}

mean absolute error

```
mae(Fit_4,cheddar)
```

```
## [1] 7.586727
```

{modelr}

```
qae(Fit_4,cheddar)
```

```
##         5%        25%        50%        75%        95%
##  1.051164  4.087882  5.238398 10.848030 16.609669
```

{modelr}

Other summaries

- `mape()` mean absolute percentage error.
- `rsae()` is the relative sum of absolute errors.
- `mse()` is the mean-squared-error.
- `rsquare()` is the variance of the predictions divided by the variance of the response.

{modelr}

```
rsquare(Fit_4,cheddar)
```

```
## [1] 0.6517747
```

Diagnostic Plots for Linear Models with R

Plot Diagnostics for an `lm()` Object

There are six plots (selectable by `which=`) are currently available:

- a plot of residuals against fitted values,
- a Normal Q-Q plot,
- a Scale-Location plot of `sqrt( |`residuals`| })` against fitted values,
- a plot of Cook's distances versus row labels,
- a plot of residuals against leverages,
- a plot of Cook's distances against *leverage/(1-leverage)*.

By default, the first three and 5 are provided, if you just type something like .

Diagnostic Plot 1

- The first one displays the residuals vs. the fitted values we use this to evlauate the mean, variance and correlation of residuals.
- If our assumptions of constant variance and uncorrelated residuals are violated we **may** be able to correct this with a variance-stabilizing transformation.
- see `car::ncevTest()`

Diagnostic Plot 1

```
plot(Fit_4,
   which=1,
   pch=16,lwd=1.2)
```

Just increment the "`which=`" argument with any integer between 1 and 6

Diagnostic Plot 1

Diagnostic Plot 2

- The second plot helps us check the normality of the residuals.
- If the residuals are indeed normal, they should fall along the dashed line.
- Remember that the normality assumption for our errors allows us to determine the standard errors of our coefficients and predictions.

Diagnostic Plot 2

Diagnostic Plots 3

- The ***Scale-Location*** plot, also called 'Spread-Location' (or 'S-L' plot), takes the square root of the absolute residuals in order to diminish skewness (**sqrt($\mid E\mid$)**) is much less skewed than $|E|$ for Gaussian zero-mean E).

Diagnostic Plots 3

Diagnostic Plots 4

- This plot details the Cook's Distance for each observation.
- We will revert to this later.

Diagnostic Plots 4

Diagnostic Plots 5

- The ***Residual-Leverage*** plot shows contours of equal Cook's distance, for values of ***cook.levels*** (by default 0.5 and 1) and omits cases with leverage one with a warning.
- If the leverages are constant the plot uses factor level combinations instead of the leverages for the x-axis.
- **(The factor levels are ordered by mean fitted value.)**

Diagnostic Plots 5

Diagnostic Plots 6

- The final plot will display our residuals vs. their leverage.
- The dashed red lines are level curves that denote a particular value of Cook's distance.
- We will pay attention to points lying beyond the distance of 1.
- Notice that when we have data with row labels, the points will be labeled with their names. Otherwise, the row number will be shown.

Diagnostic Plots 6

{broom} R package

- `tidy()` summarizes information about model components such as coefficients of a regression.
- `glance()` reports information about an entire model, such as goodness of fit measures like AIC and BIC.
- `augment()` adds information about individual observations to a dataset, such as fitted values or influence measures.

tidy()

```
library(broom)
tidy(Fit_4)  %>%
  kable( format = "markdown",digits=4)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-28.8768	19.7354	-1.4632	0.1554
Acetic	0.3277	4.4598	0.0735	0.9420
H2S	3.9118	1.2484	3.1334	0.0042
Lactic	19.6705	8.6291	2.2796	0.0311

glance()

```
glance(Fit_4) %>%
  dplyr::select(1:7) %>%
  kable( format = "markdown",digits=3)
```

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik
0.652	0.612	10.131	16.221	0	3	-109.889

```
glance(Fit_4) %>%
  dplyr::select(6:12) %>%
  kable( format = "markdown",digits=4)
```

df	logLik	AIC	BIC	deviance	df.residual	nobs
3	-109.8888	229.7775	236.7835	2668.411	26	30

augment()

```
augment(Fit_4,interval = "confidence") %>%
  kable( format = "markdown",digits=4)
```

taste	Acetic	H2S	Lactic	.fitted	.lower	.upper	.resid	.hat	.sigma	.co
12.3	4.543	3.135	0.86	1.7924	-6.9253	10.5102	10.5076	0.1753	10.0688	0.
20.9	5.159	5.043	1.53	22.6374	16.8992	28.3756	-1.7374	0.0759	10.3250	0.
39.0	5.366	5.438	1.57	25.0372	19.9388	30.1356	13.9628	0.0599	9.9217	0.
47.9	5.759	7.496	1.81	37.9375	31.7498	44.1252	9.9625	0.0883	10.1184	0.
5.6	4.663	3.807	0.99	7.0177	-0.4556	14.4910	-1.4177	0.1288	10.3269	0.
25.9	5.697	7.601	1.09	24.1652	14.1704	34.1600	1.7348	0.2304	10.3238	0.
37.3	5.892	8.726	1.29	32.5640	23.0874	42.0406	4.7360	0.2071	10.2764	0.
21.9	6.078	7.966	1.78	39.2905	33.2790	45.3021	-17.3905	0.0833	9.6716	0.
18.1	4.898	3.850	1.29	13.1641	7.1680	19.1602	4.9359	0.0829	10.2798	0.
21.0	5.242	4.174	1.58	20.2487	13.0309	27.4665	0.7513	0.1201	10.3301	0.
34.9	5.740	6.142	1.68	30.0775	24.7554	35.3996	4.8225	0.0653	10.2831	0.