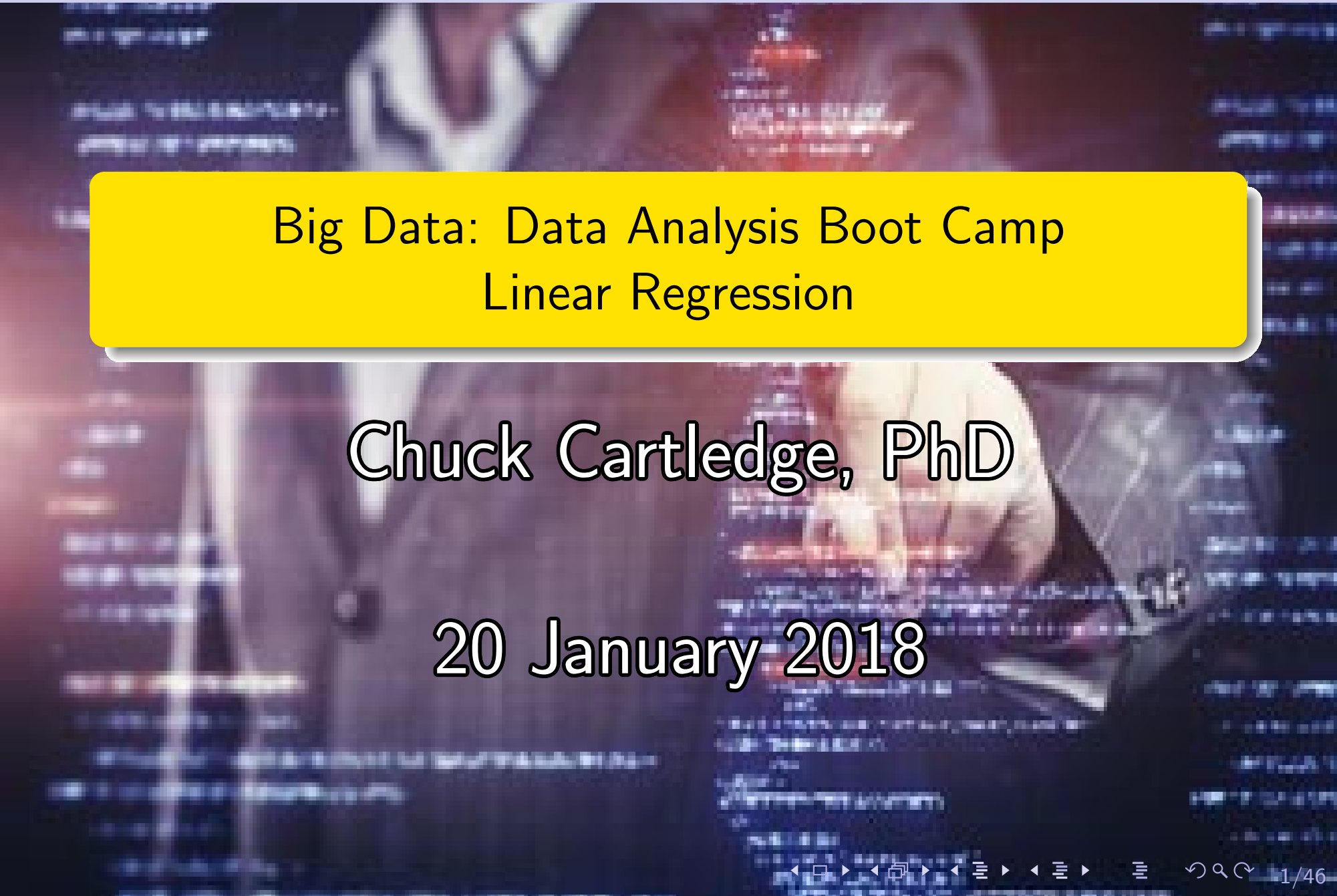


Big Data: Data Analysis Boot Camp
Linear Regression

Chuck Cartledge, PhD

20 January 2018

Table of contents (1 of 1)

What are we going to cover?

We're going to talk about:

- Ideas and assumptions that are built into linear regression (LR).
- How different techniques mitigate the effects of outliers in LR.
- How resampling a dataset can help dampen the effects of "noise" in a dataset.

Attribute types (aka values)[11][Table 2.2]

Attribute Type		Description	Operations
Categorical (Qualitative)	Nominal	The values are just different names. Nominal values provide only enough information to distinguish one object from another. $(=, \neq)$	mode, entropy, contingency, correlation χ^2 test
	Ordinal	The values provide enough information to order objects. $(<, >)$	median, percentiles, rank correlation, run tests, sign tests
Numeric (Quantitative)	Interval	The differences between values are meaningful. $(+, -)$	mean, standard deviation, Pearson's correlation, *t* and *f* tests
	Ratio	Differences and ratios are meaningful. $(*, /)$	geometric mean, harmonic mean, percent, variation

Attribute transformations (aka values)[11][Table 2.3]

Attribute Type		Transformation	Comment
Categorical (Qualitative)	Nominal	Any one-to-one mapping.	If all employee ID numbers are reassigned, it will not make a difference.
	Ordinal	An order-preserving change of values: $newValue = f(oldValue)$, where f is a monotonic function.	An attribute encompassing the notion of good better best can be represented equally well by the values $\{1, 2, 3\}$ or by $\{0.5, 1, 10\}$.
Numeric (Quantitative)	Interval	$newValue = a * oldValue + b$, a and b are constants.	The Fahrenheit and Celsius temperature scales differ in the location of their zero value and the size of a degree.
	Ratio	$newValue = a * oldValue$	Length can be measured in meters or feet.

Identifying attributes[2]

- Which variable is the response variable;
- Which variables are the explanatory variables;
- Are the explanatory variables continuous, categorical, or a mixture of both;
- What is the nature of the response variable – is it a continuous measurement, a count, a proportion, a category, or a time-at-death?

Some modeling approaches only make sense with certain types of attributes.

A little vocabulary

Sometimes there are many words for the same thing[4].

- The thing (variable) that is controlled:
  - Criterion
  - Dependent
  - Endogenous
  - Measured
  - Output
  - Regressand
  - Response
- The thing(s) (variables) that control:
  - Covariates
  - Exogenous
  - Explanatory
  - Features
  - Independent
  - Input
  - Predictor
  - Regressors
  - Variables

The type of variable determines type of model to use. (1 of 2)[2]

The explanatory variables:

Explanatory variable(s)	**Model**
All continuous	Regression
All categorical	Analysis of variance (Anova)
Both continuous and categorical	Analysis of covariance (Ancova)

The type of variable determines type of model to use. (2 of 2)[2]

The response variable:

Response variable	**Model**
Continuous	Normal Regression, Anova, Ancova
Proportion	Logistic regression
Count	Log linear models
Binary	Binary logistic analysis
Time-at-death	Survival analysis

How do all these models fit together?

"...the generalized linear model (GLM) is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution."

W. Staff [7]

Part of the regression zoo

Generalized linear models =
- General linear models =
  - t-test
  - ANCOVA
  - ANOVA
  - F-test
  - Linear regression
  - MANCOVA
  - MANOVA
- Hierarchical linear models
- Logistic regression
- Loglinear modeling
- Multinomial logistic regression
- Multinomial probit regression
- Ordered probit
- Poisson regression
- Probit regression

- ANCOVA – Analysis of covariance
- ANOVA – Analysis of variance
- MANCOVA – Multivariate analysis of covariance
- MANOVA – Multivariate analysis of variance

Math behind linear regression[9]

$$y_i = \beta_0 1 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where $\top$ denotes the transpose, so that $x_i^\top \boldsymbol{\beta}$ is the inner product between vectors x_i and β.

Often these n equations are stacked together and written in vector form as $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$,

$$X = \begin{pmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_n^\top \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

What is this ε term?

The ε term is a "random noise" that is part of each response value. There are some inherent assumptions about ε[1]:

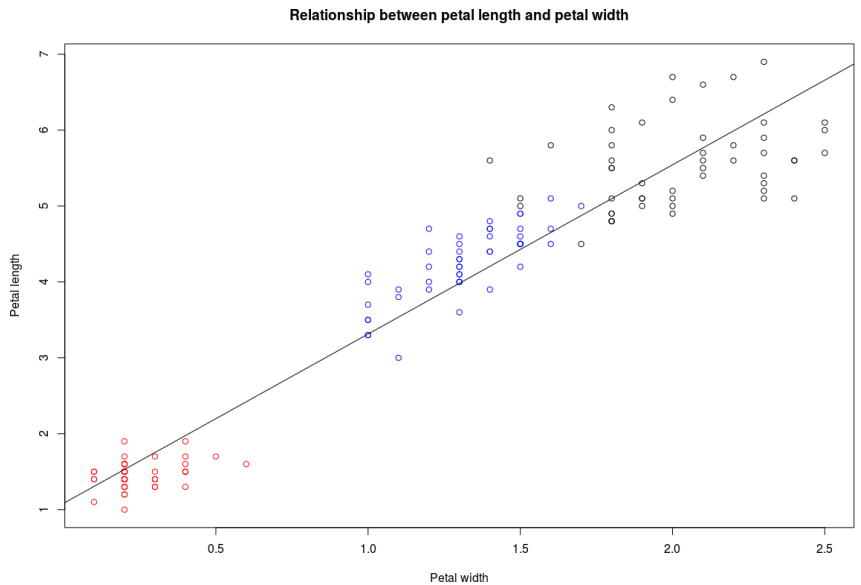
1. Linearity - the dependent variable y is a linear function of x plus random distribution ε
2. Mean independence - the mean of ε is always 0
3. Homoscedastcity (variance independence) - ε does not depend on x
4. Uncorrelated disturbances - each ε is independent of any other ε
5. Normal disturbance - ε has a normal distribution

If these assumptions are not met, then linear regression should not be used[1].

[1] See also[5].

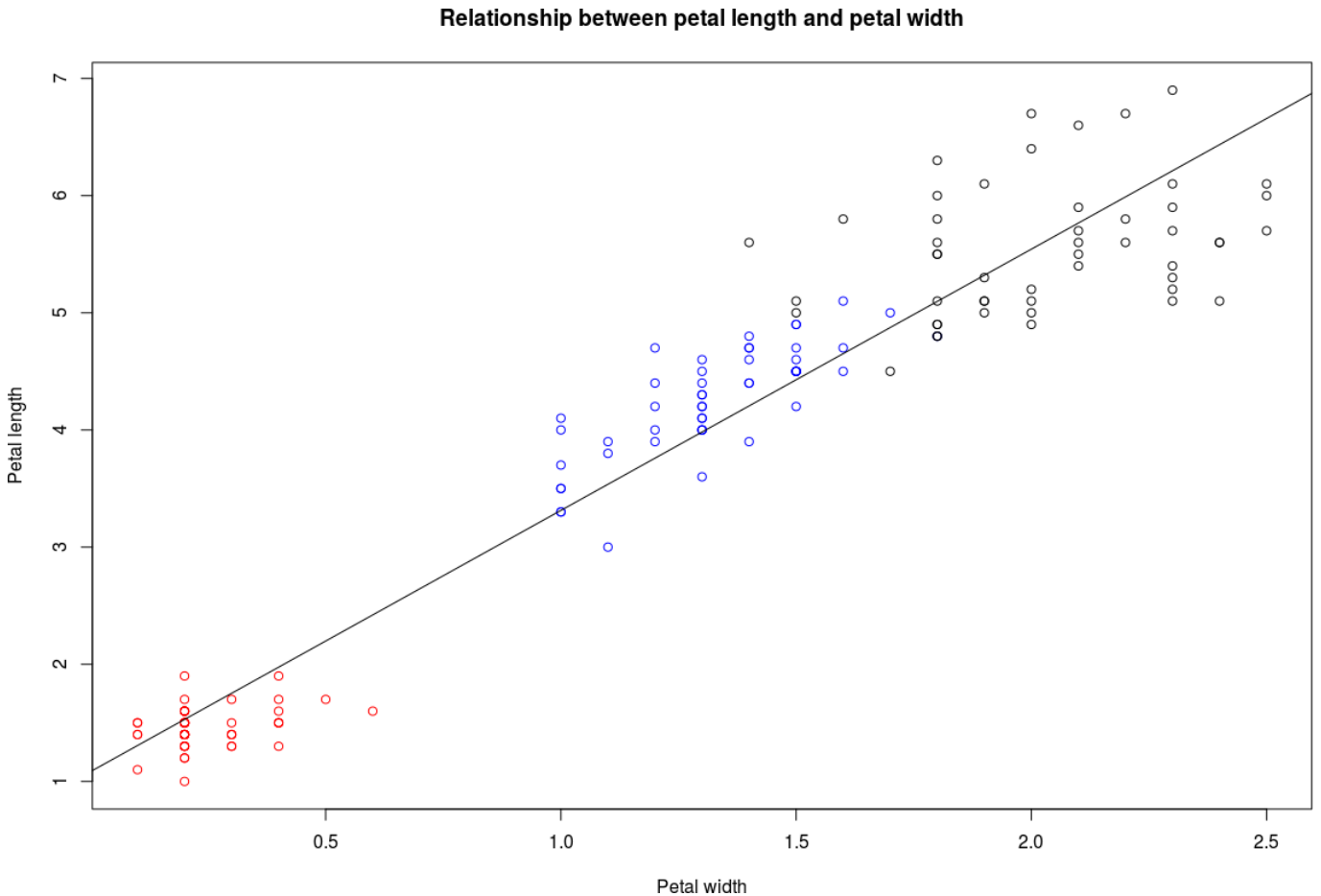
Looking at petal length and width

```
myIris <- iris

model <- myIris$Petal.Length ~myIris$Petal.Width

myIris$SpCol <- ifelse(myIris$Species ==
''setosa'', ''red'', ifelse(myIris$Species ==
''versicolor'', ''blue'', ''black''))

plot(model,
main = ''Relationship between petal length and
petal width'',
xlab = ''Petal width'', ylab = ''Petal length'',
col=myIris$SpCol)

iris.lm = lm(model)

abline(iris.lm)
```



Slope ≈ 2.23, intercept ≈ 1.08

Same image.



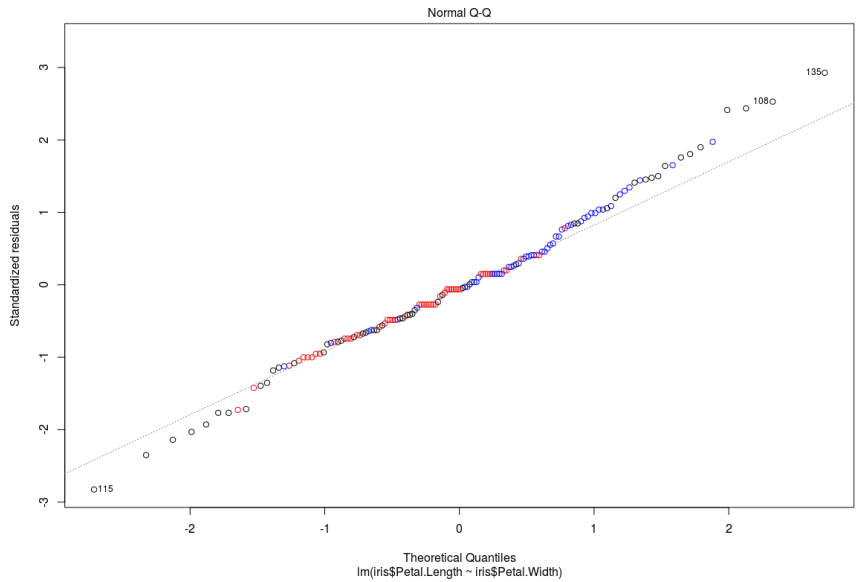
Slope ≈ 2.23, intercept ≈ 1.08

Iris linear regression Q-Q plot

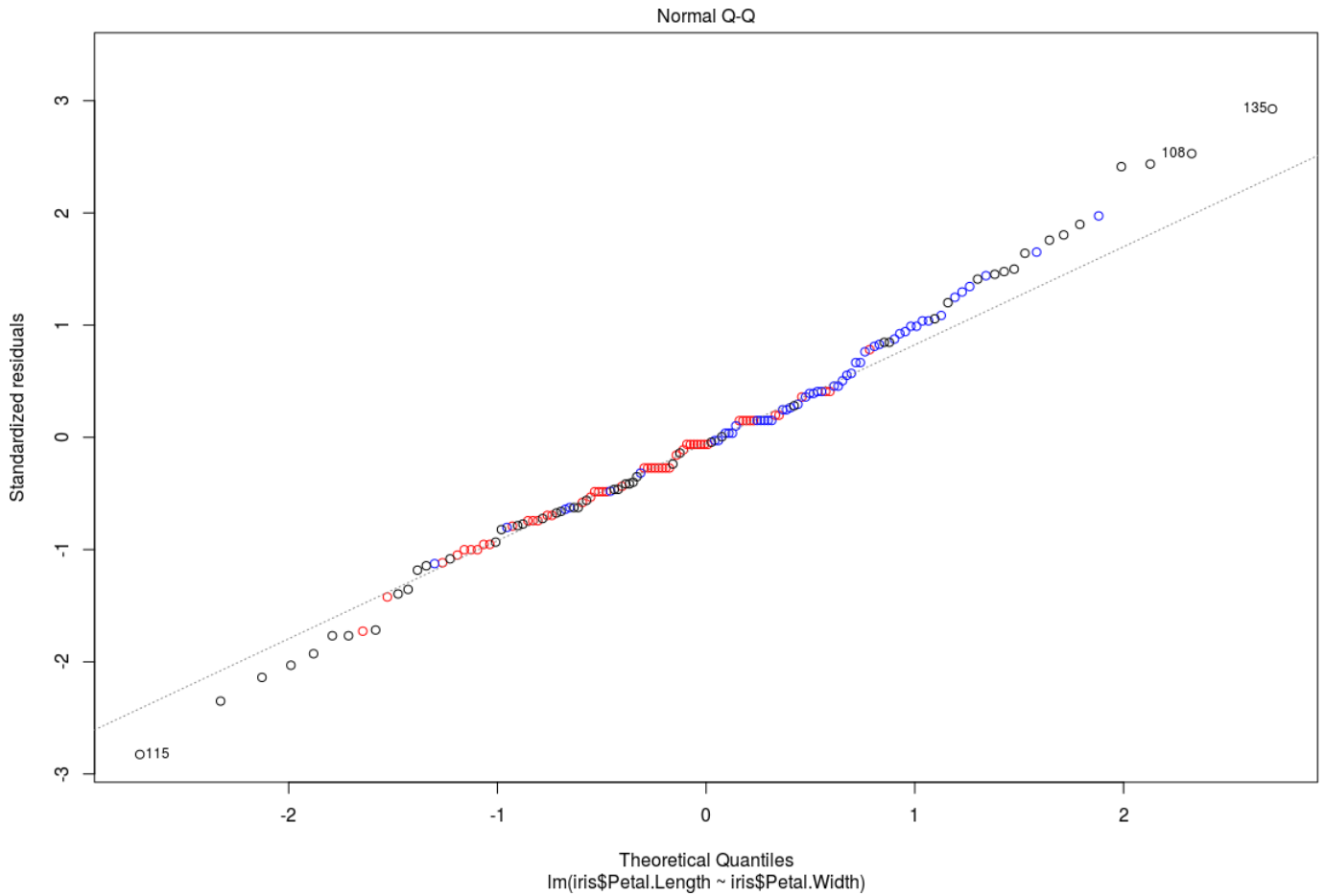
Continuing from the previous slide:

```
plot(iris.lm,
col=myIris$SpCol)
```

The third plot of interest.
A Quantile-Quantile (Q-Q) plot compares the quantiles of a theoretical (modeled) distribution and empirical data.[10]



Same image.


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(iris$Petal.Length ~ iris$Petal.Width)
135
108
115
3
2
1
0
-1
-2
-3
-2
-1
0
1
2

Details about the data and how good it is.

Execute this command:

summary(iris.lm)

	Estimate	Std. Error	t value	Pr($>\|t\|$)	
(Intercept)	1.08356	0.07297	14.85	<2e-16	***
myIris$Petal.Width	2.22994	0.05140	43.39	<2e-16	***

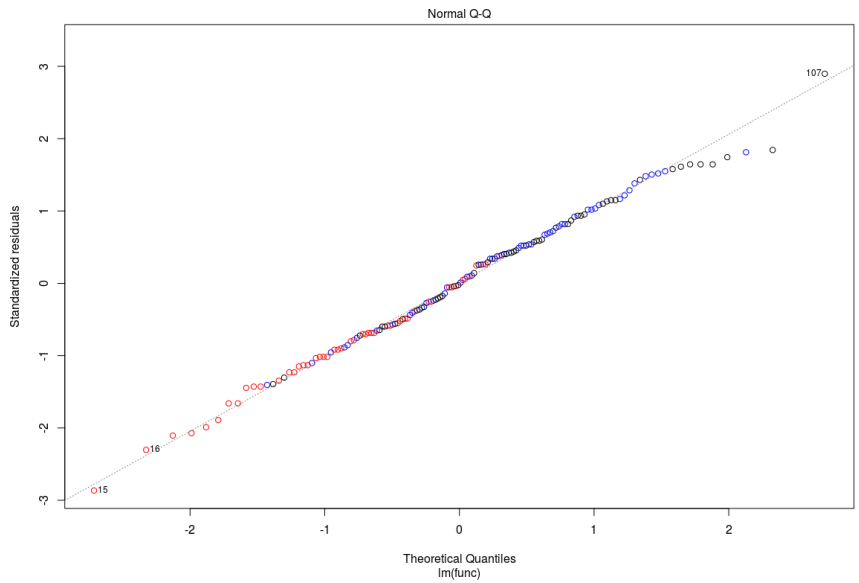
Things of note:

- Intercept and Petal.Width estimates are in close agreement with plot
- Standard errors for intercept and slope are small (implying that errors are "normal")
- High t values imply high likelihood of "normal distribution"

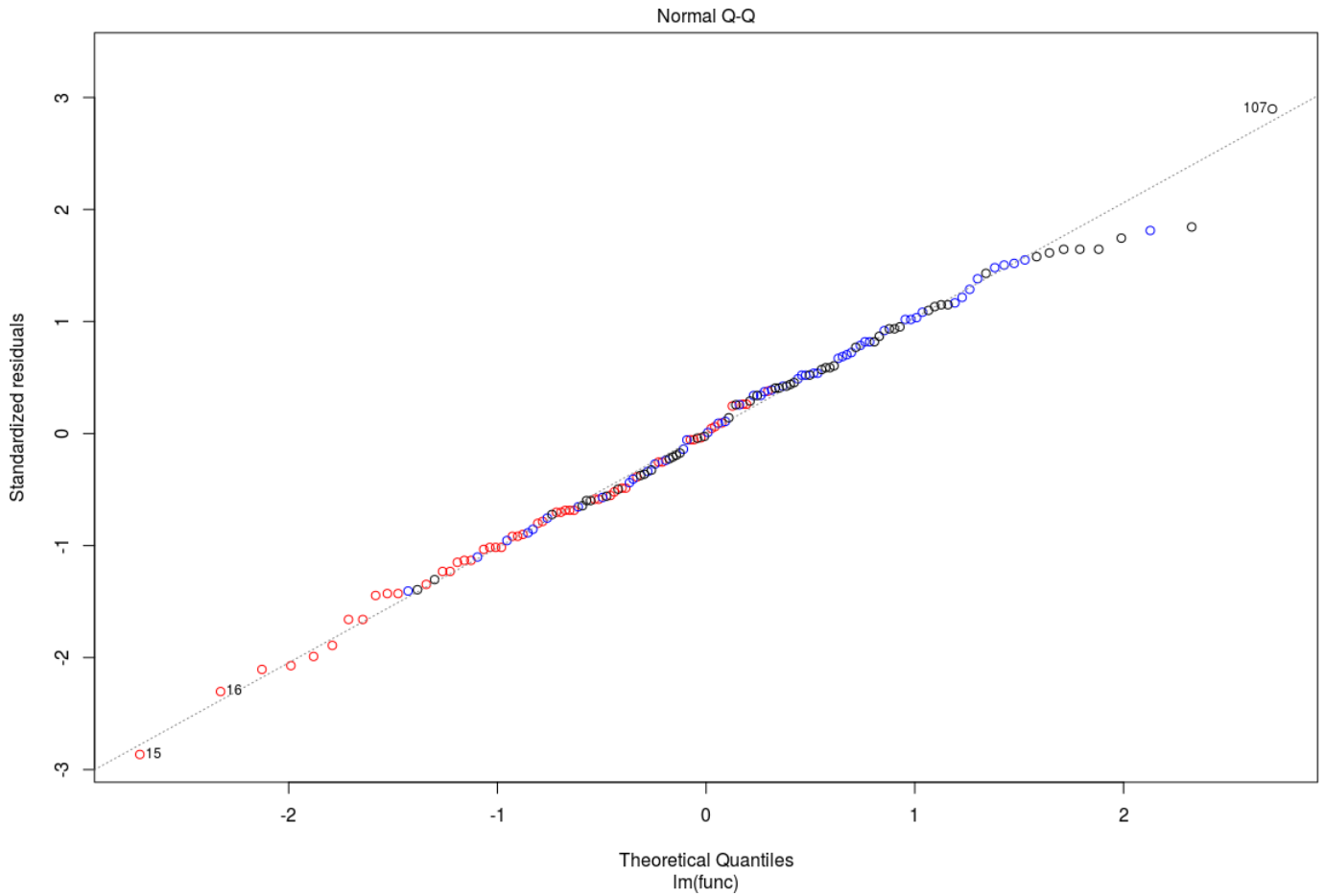
Change the model and test another relationship

Execute these commands:

```
func <-
myIris$Petal.Length
~myIris$Sepal.Length


iris.lm <- lm(func)
plot(iris.lm,

col=myIris$SpCol)


summary(iris.lm)
```



Same image.

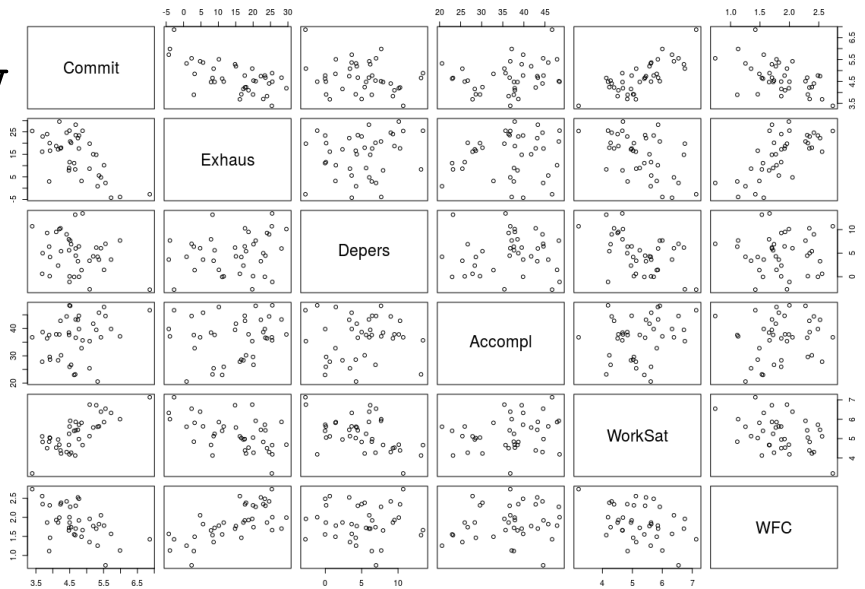

Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(func)
107
16
15
3
2
1
0
-1
-2
-3

Multiple regression using artificial data about nurse's commitment

Load the file `chapter-09-multiregression-rev` into the editor.

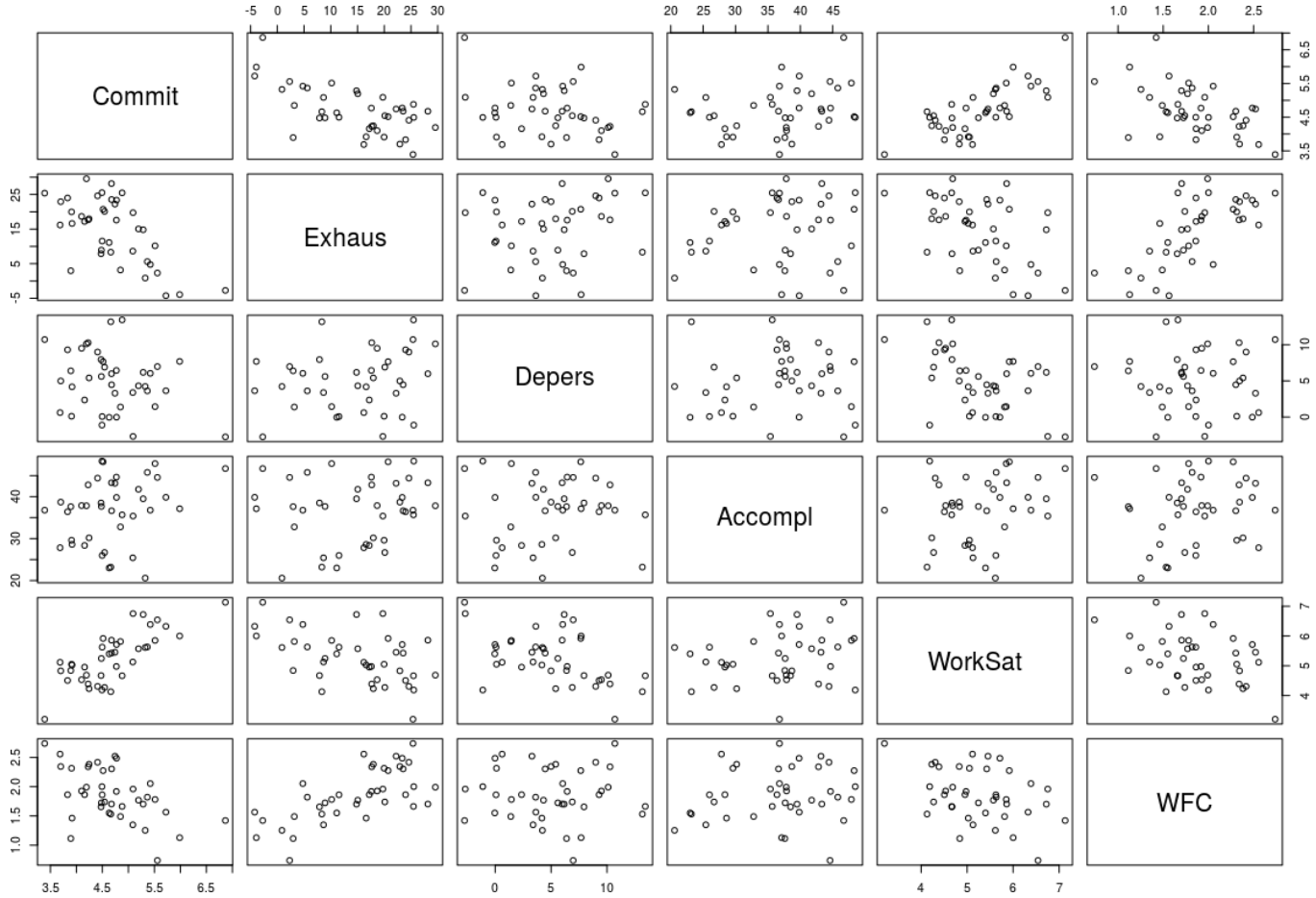
Execute: `main()`

The `shapiro.test` can return false values, so should be used with caution.



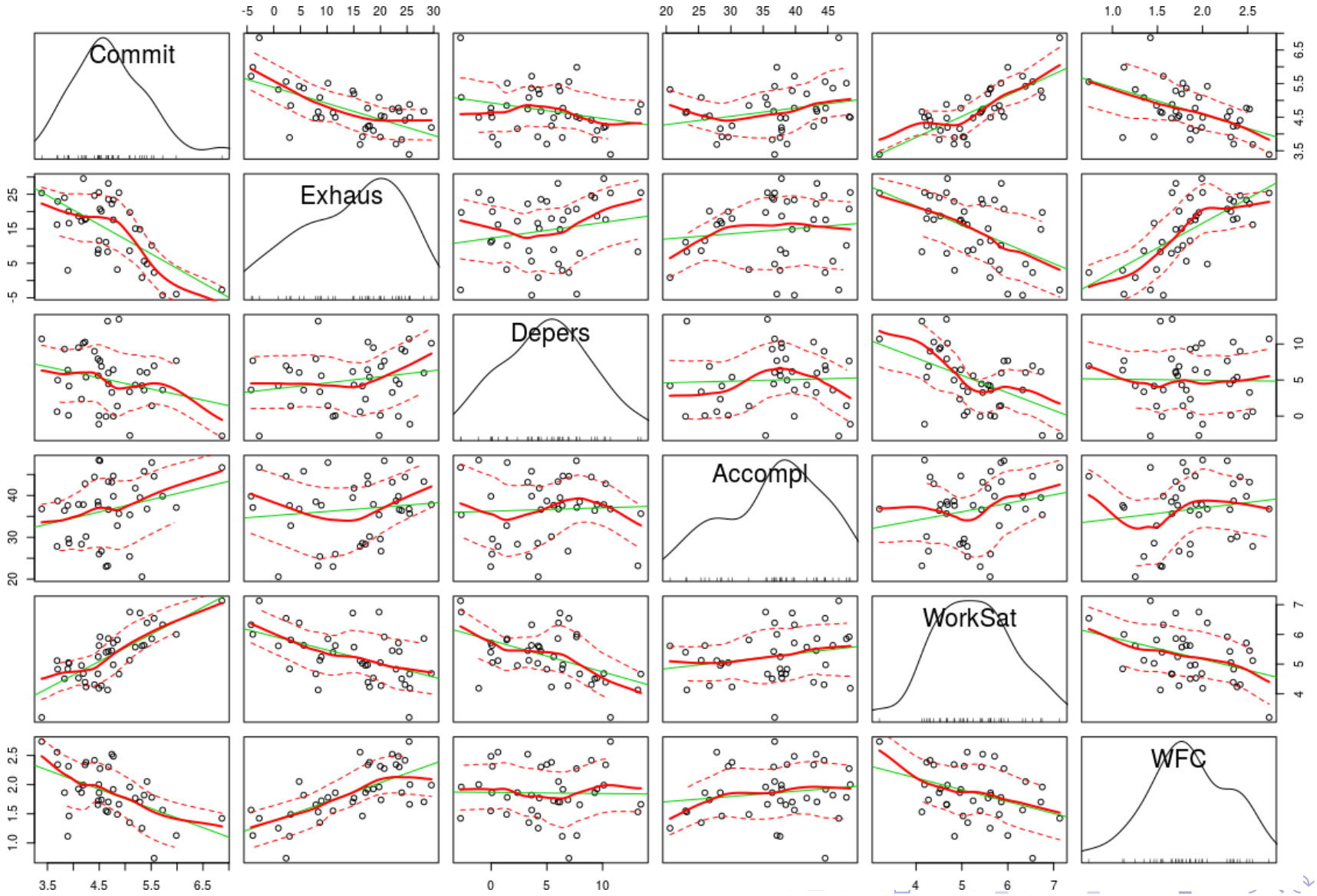
Attached file.

Same image.

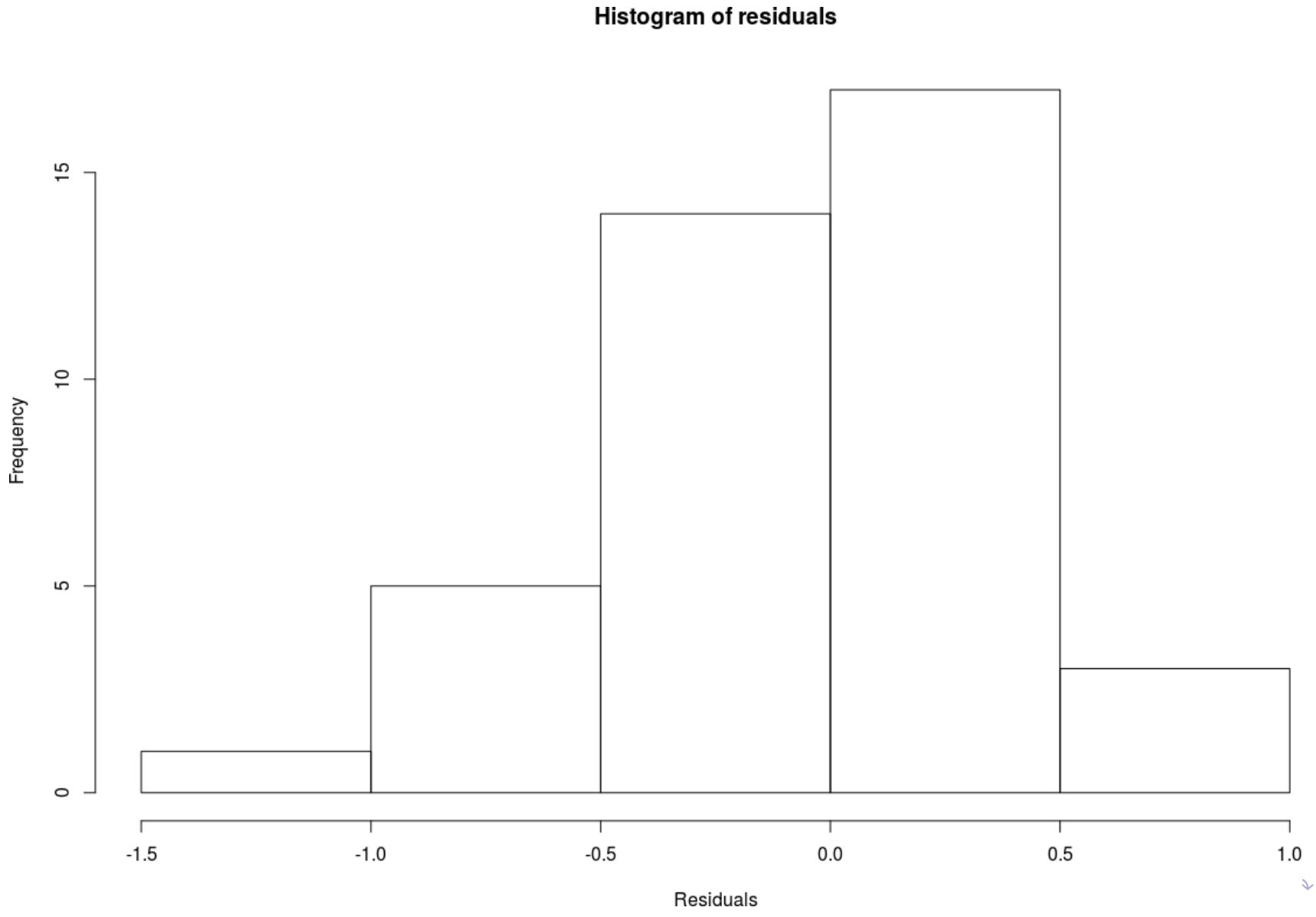


Attached file.

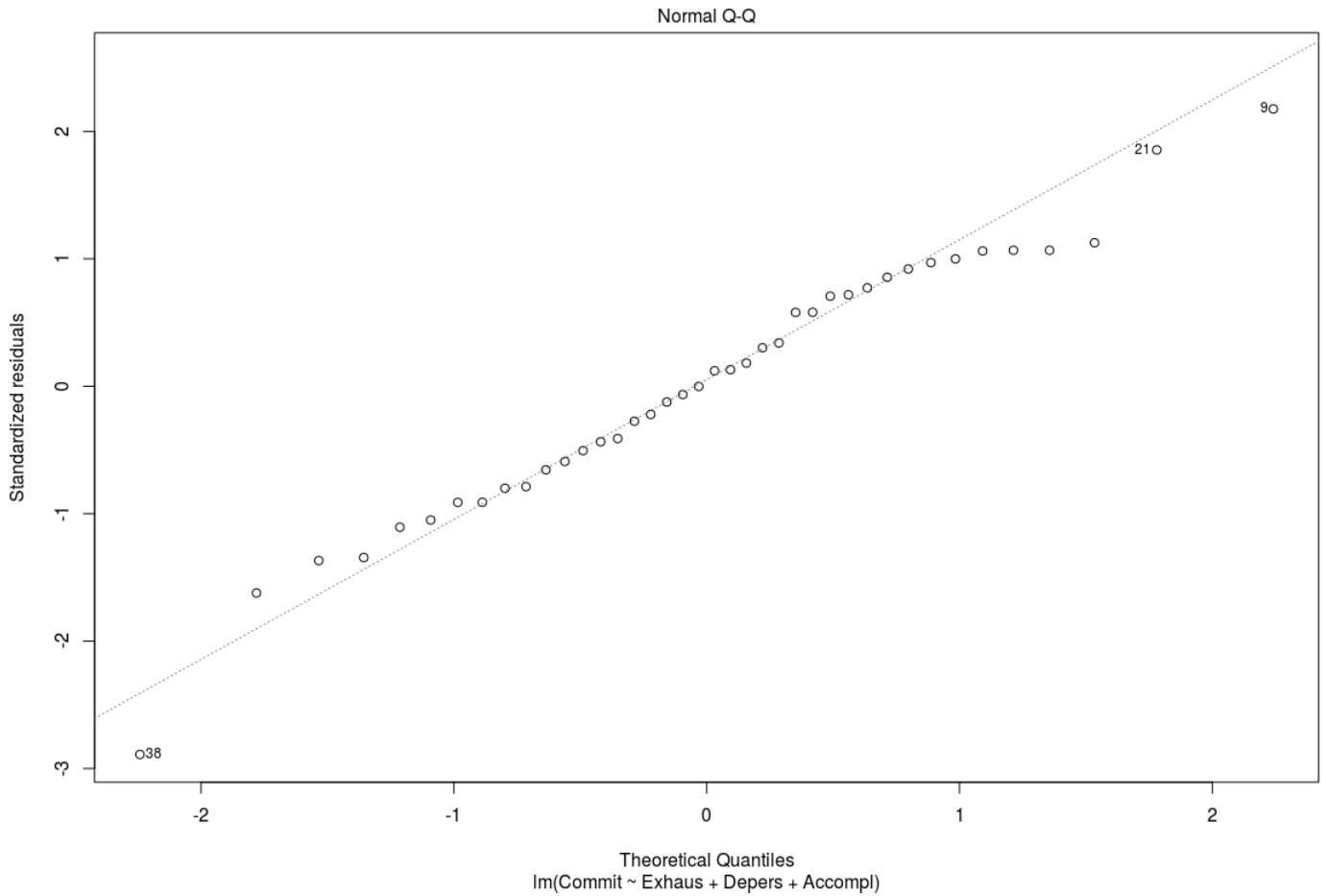
Another presentation



Are the residuals distributed "normally"?



What does our friend QQ plot show?



Mostly normal distribution.

Are there mediating variables?

Executing the code returned these values:

	Sobel	Aroian	Goodman
z.value	-2.972270400	-2.936471185	-3.009411683
p.value	0.002956062	0.003319697	0.002617542

The **Sobel**, **Aroian**, and **Goodman** are specialized t tests to determine whether the reduction in the effect of the independent variable, after including the mediator in the model, is a significant reduction and therefore whether the mediation effect is statistically significant.

Iterated Weighted Least Squares (IWLS)

One of the basic assumptions about linear regression is **homoscedastcity**. If that is TRUE, then dependent values errors are independent of the independent values. Outliers will "draw" the solution to them.
What if we suspect that homoscedastcity is violated?
Heteroscedasticity exists. Therefore dependent values errors are some how dependent on the independent values. Outliers should be ignored.
But how?
Weighted least squares is an iterative technique that discounts (reduces the "weight of") dependent values that are outliers.

Math behind IWLS:[8]

From 50,000 feet, we are:

- Looking to find a function $f(x)$ that "best fits" all the data points y,
- By not paying as much attention to outliers as we do to points close to $f(x)$.

We do this by assigning weights to each data point, and then adjusting the weights until we are satisfied.

$$\text{objective function} \equiv \arg\min_{\beta} \sum_{i=1}^{n} \left| y_i - f_i(\beta) \right|^p$$

$$\beta^{(t+1)} = \arg\min_{\beta} \sum_{i=1}^{n} w_i(\beta^{(t)}) \left| y_i - f_i(\beta) \right|^2$$

The R function MASS:rlm hides all this messiness and returns answers.

The results:

Call: lm(formula = model, data = nurses)			
Coefficients:	Estimate	Std. Error	t value
(Intercept)	4.331261	0.398985	10.856
Exhaus	-0.048725	0.008625	-5.649
Depers	-0.027053	0.019795	-1.367
Accompl	0.032923	0.010392	3.168
Call: rlm(formula = model, data = nurses)			
Coefficients:	Value	Std. Error	t value
(Intercept)	4.3602	0.3849	11.3271
Exhaus	-0.0518	0.0083	-6.2306
Depers	-0.0279	0.0191	-1.4602
Accompl	0.0338	0.0100	3.3676

`rlm()` improved the "t value" and is a better solution.

Some basic information

Bootstrapping repeatedly uses parts of given data to help qualify the variability of parameters.
Basically:

- Sample N values from the given data (the population)
- Compute regression on the sample
- Repeat above steps K times

A "better" estimate of the parameter will be the mean of computed regression parameters.

How does this affect our "nurses" data?

	95% C.I. lower bound	95% C.I. upper bound
Intercept	4.297	4.325
Exhaus	-0.048	-0.048
Depers	-0.029	-0.027
Accomp	0.033	0.033
R2	0.558	0.570
F	18.179	19.139

We can now say what the values are with 95% confidence.

Some simple exercises to get familiar with linear regression

Extract Anscombe's third dataset using something like:[2]

```
temp <- data.frame( anscombe[, "x3"], anscombe[, "y3"])
```

1. Plot the raw data
2. Plot a simple linear model of the data (think `abline`)
3. Plot an IWLS model of the data (think `abline`)
4. Compare the residuals of both models

[2] See presentation "005-what-is-da" about Anscombe's data.

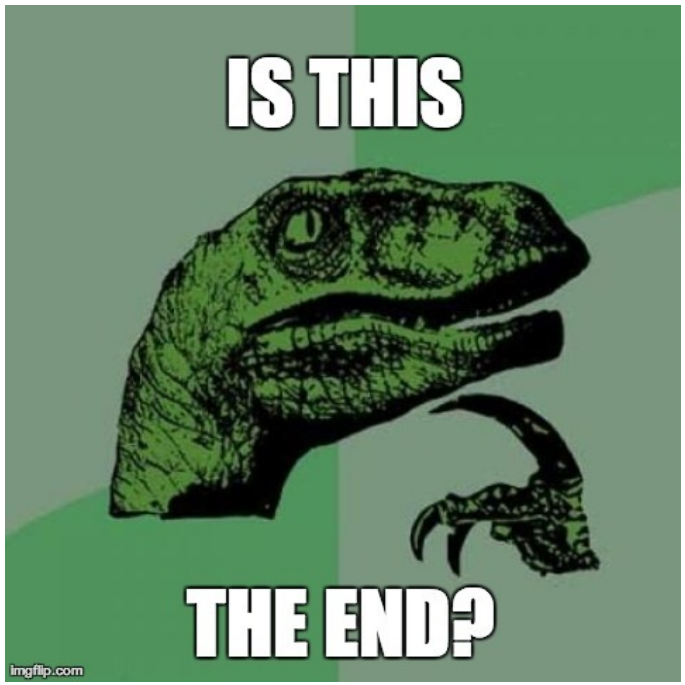
Q & A time.

Q: How many bureaucrats does it take to screw in a light bulb?
A: Two. One to assure everyone that everything possible is being done while the other screws the bulb into the water faucet.

What have we covered?

- Covered a lot ideas and assumptions about linear regression (LR)
- Looked at iris data to start to understand LR
- Looked at how mediating parameters can affect solutions
- Looked at how weighted least squares can affect solutions
- Looked at how bootstrapping can improve solutions

Next: LPAR Chapter 10, classification



References (1 of 3)

[1] Paul D. Allison, Multiple Regression: A Primer, Pine Forge Press, 1999.

[2] Steven Buechler, Statistical Models in R, `https://www3.nd.edu/~steve/Rcourse/Lecture7v1.pdf`, 2007.

[3] Award design Staff, Standard Score, `https://competition.adesignaward.com/normalizingscore.html`, 2017.

[4] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, An Introduction to Statistical Learning, vol. 6, Springer, 2013.

References (2 of 3)

[5] Sunil Ray,
7 Types of Regression Techniques you should know!,
https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/, 2015.

[6] Wikipedia Staff, F-test,
https://en.wikipedia.org/wiki/F-test, 2017.

[7] ———, Generalized linear model, https://en.wikipedia.org/wiki/Generalized_linear_model, 2017.

[8] ———, Iteratively reweighted least squares,
https://en.wikipedia.org/wiki/Iteratively_reweighted_least_squares, 2017.

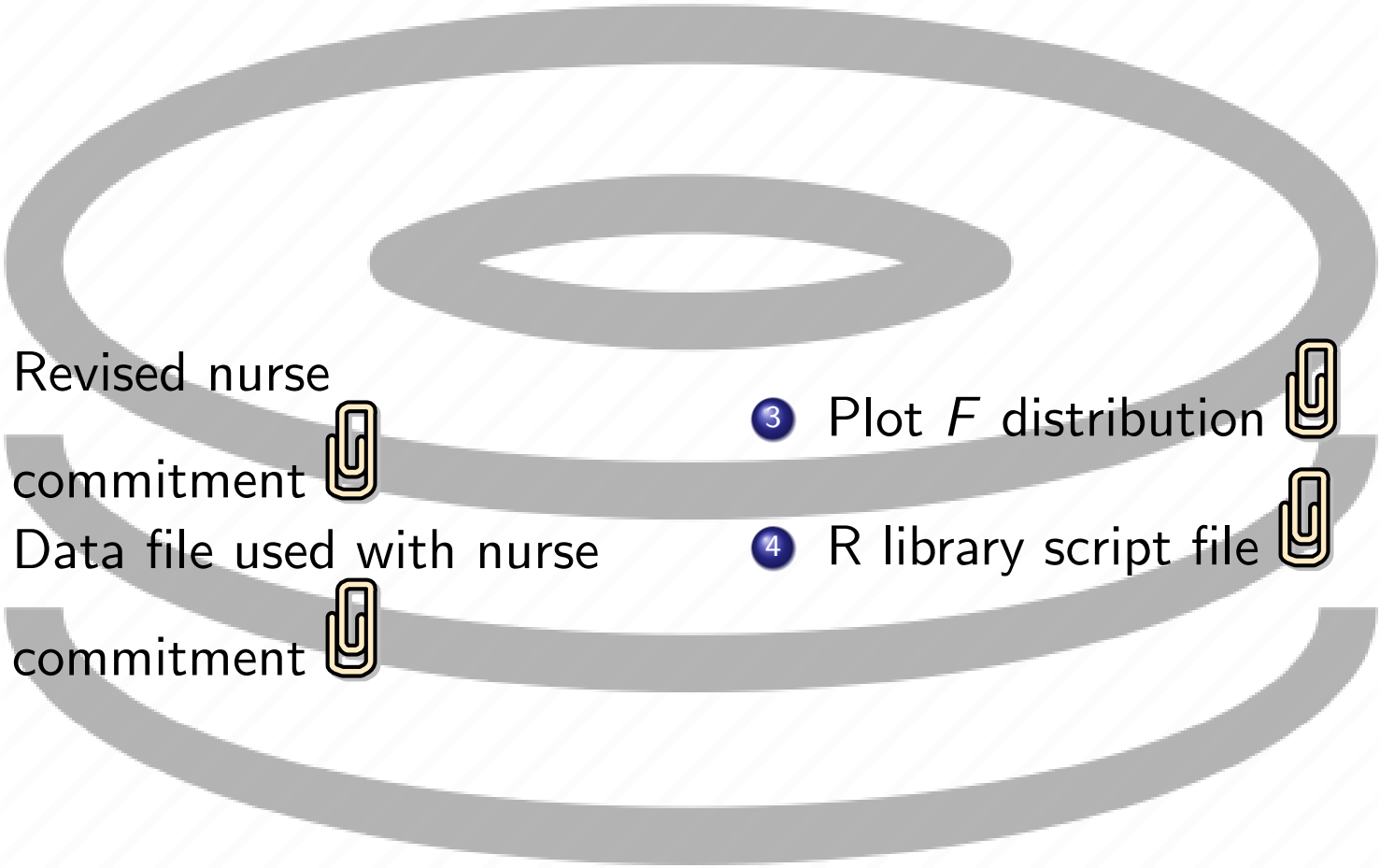
References (3 of 3)

[9] _____, Linear regression, https://en.wikipedia.org/wiki/Linear_regression, 2017.

[10] JB Statistics, Normal Quantile-Quantile Plots, https://www.youtube.com/watch?v=X9_ISJ0YpGw, 2013.

[11] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, Introduction to Data Mining, Pearson Education India, 2006.

Files of interest



1. Revised nurse commitment
2. Data file used with nurse commitment
3. Plot F distribution
4. R library script file

A few useful equations (1 of 2)

$$
\begin{aligned}
\sigma^2 \text{ (variance)} &= \sum_{i=1}^{n}(x_i - \bar{x})^2 \\
\sigma \text{ (standard deviation)} &= \sqrt{\sigma^2} \\
\beta \text{ (slope)} &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
\alpha \text{ (intercept)} &= \bar{y} - \beta * \bar{x} \\
R^2 \text{ (R-squared)} &\equiv 1 - \frac{RSE}{RSS} \\
\bar{R}^2 \text{ (adjusted R-squared)} &\equiv R^2 - (1 - R^2)\frac{p}{n - p - 1} \\
RSE \text{ (residual standard error)} &= \sqrt{\frac{RSS}{n - 2}}
\end{aligned}
$$

A few useful equations (2 of 2)

$$RSS \text{ (residual sum of squares)} = \sum_{i=1}^{n}(\varepsilon_i)^2 = \sum_{i=1}^{n}(y_i - (\alpha + \beta x_i))^2$$

$$SE(\beta)^2 \text{ (standard error slope)} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})}$$

$$SE(\alpha)^2 \text{ (standard error slope)} = \sigma^2 * (\frac{1}{n} + \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})})$$

$$t \text{ (t-statistic)} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$TSS \text{ (total sum of squares)} = \sum_{i=1}^{n}(y_i - \bar{y})$$

Confidence intervals

$$\text{Lower bound} \equiv \bar{x} - z * \frac{\sigma}{\sqrt{n}}$$

$$\text{Upper bound} \equiv \bar{x} + z * \frac{\sigma}{\sqrt{n}}$$

Where:

- σ is the standard deviation of the population
- n is the size population
- $\bar{x}$ is the mean of the population
- z is the threshold value $z - \textit{distribution}$

F-distribution

$$f(x) = \frac{\Gamma(\frac{n1+n2}{2})(\frac{n1}{n2})^{\frac{n1}{2}} x^{\frac{n1}{2}-1}}{\Gamma(\frac{n1}{2})\Gamma(\frac{n2}{2})(1+\frac{n1}{n2}x)^{\frac{n1+n2}{2}}}$$

The parameters for the degrees of freedom are not interchangeable. That is, the F-distribution with 3 and 5 degrees of freedom is different than the F-distribution with 5 and 3 degrees of freedom.

F statistic

$$F = \frac{\text{explained variance}}{\text{unexplained variance}}$$

$$\text{explained variance} = \sum_{i=1}^{K} n_i(\bar{Y}_{i\cdot} - \bar{Y})^2/(K-1)$$

$$\text{unexplained variance} = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot})^2/(N-K)$$

Where:

- K is the number of groups
- N is the overall sample size

"The statistic will be large if the between-group variability is large relative to the within-group variability, which is unlikely to happen if the population means of the groups all have the same value."

W. Staff [6]

ShapiroWilk test[3]

$$
\begin{aligned}
W &= \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \\
(a_1, \ldots, a_n) &= \frac{m^{\mathsf{T}} V^{-1}}{(m^{\mathsf{T}} V^{-1} V^{-1} m)^{1/2}} \\
m &= (m_1, \ldots, m_n)^{\mathsf{T}}
\end{aligned}
$$

Where:

- $m_1, \ldots, m_n$ are the expected values of the independent variables
- V is the covariance matrix

[3] https://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test

Normal distribution with z and t scores

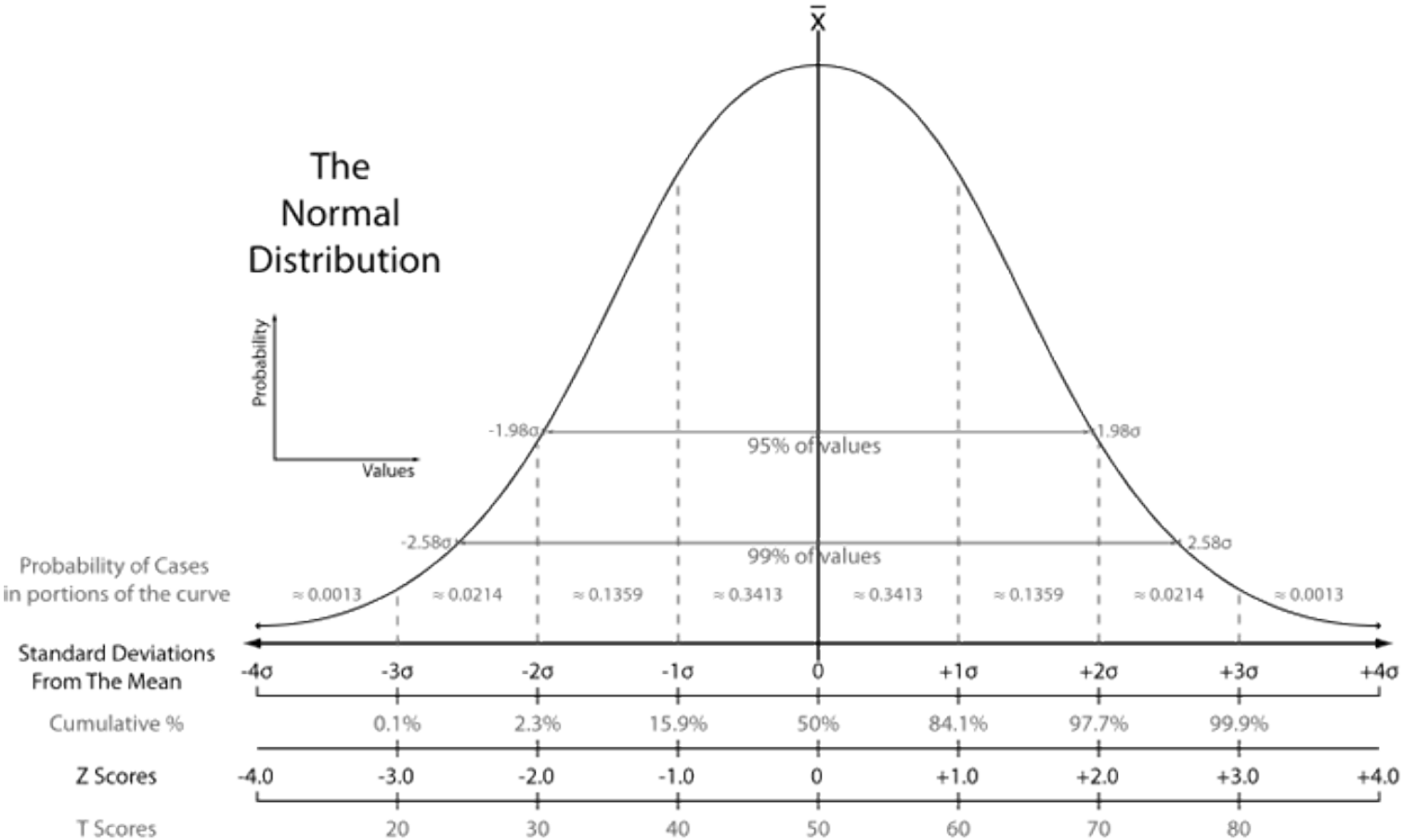
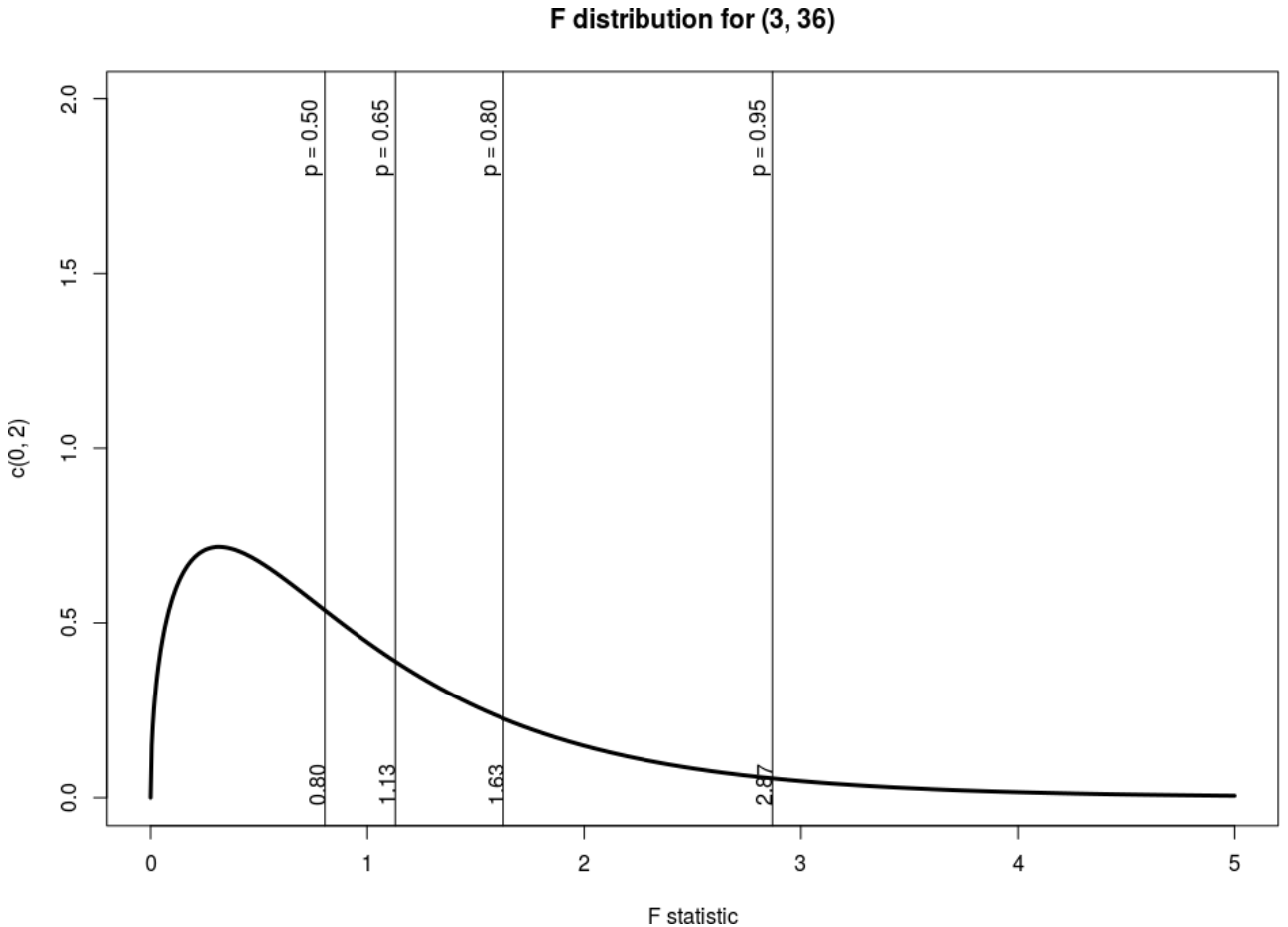


Image from [3].

F distribution with selected degrees of freedom



Attached file.