Introduction to R for Beginners

Lyda Hill department of Bioinformatics 2022 Nanocourse Series

Date & Time: May 9-10: 9AM-5PM (NG3.202)

Course Instructors: Christopher Chaney, Amit Amritkar, & Micah Thornton

UT Southwestern
Medical Center

Day 2: Section III (Transforming Data, Modeling, Regularization, Variable Selection and Tabulating/Displaying Results)

UT Southwestern
Medical Center

What you will learn in this section

- The generalized Linear Model (R glm)
  - What is a GLM?
  - Which options are available in R?
  - What are the assumptions for using GLMs?
  - How can I check these assumptions in R?
    - Graphical qq plots & histograms (refresher)
  - How can I estimate the parameters of these models for a dataset in R?
- What to do if my data is not normal or linearly related (Transforming data)?
  - Standard data transformations in R
- How can I determine which variables should be included in the model (Regularization)?
  - What are variable selection approaches?
  - Forward/Backwards/Stepwise based on different criteria in R
  - Elastic-Net, Ridge, LASSO, and penalized-regression in R
- How can I store and sort through the results of running regression models?
  - How to neatly display data in labeled and formatted dataframes.
  - How to produce html tables using the htmlTable() package.
- *Tentative Bonus - Survival Modeling:*
  - *How can I perform Cox Proportional Hazards Regression and check the proportional hazards assumptions in R?*

UT Southwestern Medical Center

What you will *NOT* learn in this section but have available in R!

- What are the differences between fixed effects and random effects
  - Can I produce Random Effects models using the R Languages **lme4 (Linear Mixed-Effects models) package**?
  - Can I put random effects and mixed effects in the same model for a ***mixed* effects model?**
  - How can I tell if an effect should be modeled as **fixed vs. random**.
- What are principal components and how can they be used in regression?
  - The **principal components regression** approach.
  - Assessment of principal components.
- Alternative kinds of regression available in R:
  - Isotonic Regression (Monotonic Regression)
  - AR-type models (Auto-Regressive models for time series data)
  - Projection Pursuit Regression (Generalization of Additive model)
  - Nonlinear regression and modeling (nlm() in R)
  - Partial Least Squares Regression (or Discriminant Analysis)
  - Machine Learning (Perceptron Modeling, Random Forests, Naïve Bayes, etc…)
- Bayesian + Hierarchical (Random Effects) Modeling approaches
  - How to effectively select prior distribution models
    - What are uninformative priors?
    - What are conjugate priors?
    - What is Jeffries Prior?
  - How to determine the likelihood model, and posterior distribution/ posterior probable intervals etc…
- Time Series/Longitudinal Data Modeling Approaches
- Causal Inference + Exact Testing.
- Much much more…

UT Southwestern
Medical Center

Part I:
The Multivariate Linear Model
Specification, Estimation, & Validation in R

UT Southwestern
Medical Center

The Multivariate Linear Model

- Standard multivariate regression modeling procedure which seeks to relate two or more random variables through a linear combination of estimable effects.

$$Y_i = \beta_0 + \beta_1 X_{1_i} + \beta_2 X_{2_i} + \cdots + \beta_N X_{N_i} + \varepsilon_i$$

- In the above, Y_i indicates the i^{th} observation of a dependent variable, and the series of $X_{1_i}, X_{2_i}, \ldots, X_{N_i}$ indicate the i^{th} observations of independent variables, in the modeling sense. β_0 is an estimable intercept, and the $\beta_1, \beta_2, \ldots, \beta_N$ are estimable effects (or coefficients).

- Generally the intercept and coefficients are determined by attempting to minimize the residuals for a specific set of observations, which reduces to an optimization problem, luckily `R` has in-built methods for estimating these parameters in a variety of different situations.

- Due to the manner in which the parameters of the linear model $(\alpha, \beta_0, \beta_1, \ldots, \beta_N)$ are estimated, there are a few assumptions that must hold, but these will be discussed in more detail in a later slide.

The Multivariate Linear Model in R Example (Iris Data)

- There are many built in datasets in R, for this example I will use the `iris` dataset.

```
R> summary(iris)
```

```
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

- From the summary, we can see that there are three numerical variables and one categorical, for this first part we will only use the numerical variables, and so we can now subset these into a separate dataset.

```
R> irisn=iris[,1:4]
R> summary(irisn)
```

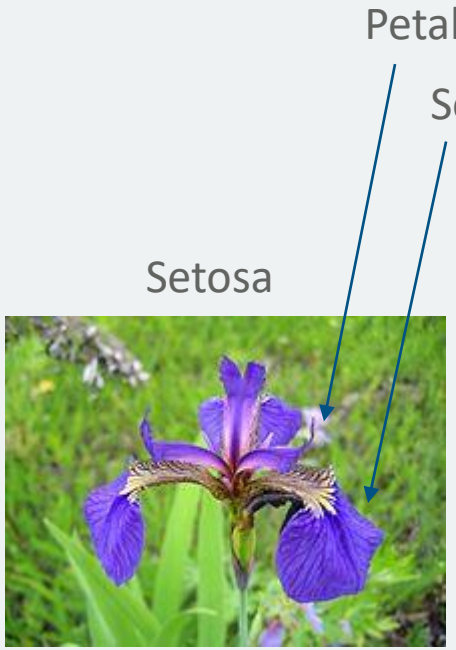
```
> irisn <- iris[,1:4]
> summary(irisn)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
 Median :5.800   Median :3.000   Median :4.350   Median :1.300
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

UT Southwestern Medical Center

Iris Dataset Guide

N=150 samples, 50 of each type

$1\ \text{cm} \approx 0.39370\ \text{in.}$



Images Wikipedia.

Dataset captured by Edgar Anderson (English Botanist) and Popularized by Fisher.

UT Southwestern Medical Center

The Multivariate Linear Model in R Example (Iris Data)

- To get even more information about a dataset using the help(·) function with · replaced by the dataset name will bring up the help entry which usually will contain information about the dataset.

```
R> help(iris)
```

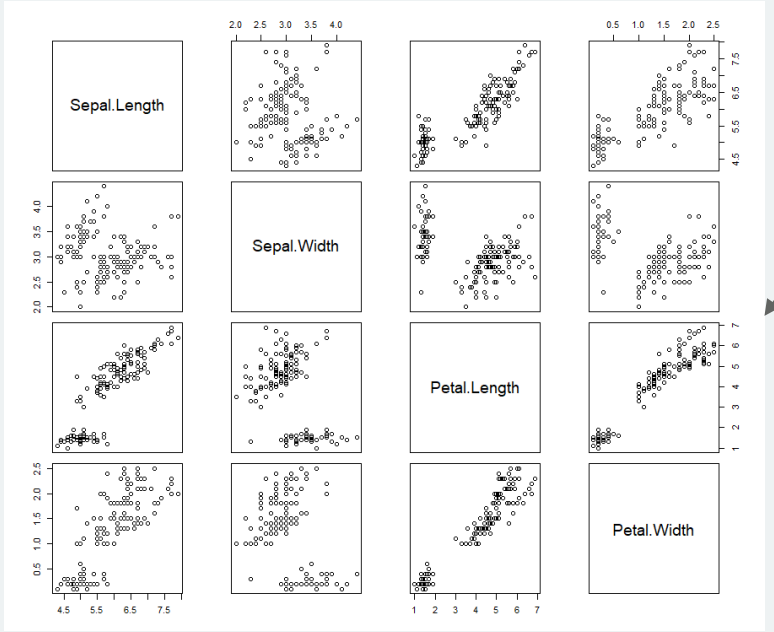
iris {datasets} R Documentation

Edgar Anderson's Iris Data

Description

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

- When considering a linear model, it is often very helpful to produce scatterplots of the various pairings of variables to determine the dependence structure, which we can do with the built-in pairs() function in base R.




```
R> pairs(irisn)
```

- We can see an apparent linear trend between sepal.length and the other variables, indicating a linear model may be appropriate.

UT Southwestern Medical Center

The Multivariate Linear Model in R Example (Iris Data)

- Its always a good idea to write down the model we will be estimating the effects of ahead of time, and label everything to make sure that we have a good understanding of the results of the estimation procedure in R (note that this is not practical in some cases with many variables).

$$Y_i = \beta_0 + \beta_1 X_{1_i} + \beta_2 X_{2_i} + \beta_3 X_{3_i} + \varepsilon_i$$

Y_i : sepal length (cm) of the i^{th} observed iris.

X_{1_i} : sepal width (cm) of the i^{th} observed iris.

X_{2_i} : petal length (cm) of the i^{th} observed iris.

X_{3_i} : petal width (cm) of the i^{th} observed iris.

ε_i : The random error associated with the i^{th} observation.

- Writing the full model ahead of time also helps when interpreting the results, and in potentially finding alternative modeling approaches for the data. This linear model can be specified in R using an R object known as a formula.

```
R> formula(Sepal.Length~Sepal.Width+Petal.Length+Petal.Width) -> irisn.form1;
```

- You can specify formulae within the glm command itself, but you may wish to construct a procedure which automatically builds formulae from data with columns that you do not know ahead of time, which is where these separate formula objects come in handy.

UT Southwestern Medical Center

The Multivariate Linear Model in R Example (Iris Data)

- To finally estimate the parameters of the model, and store the resulting calculations in an R object for easy retrieval later use the glm() function in R.

```
R> glm(irisn.form1,data=irisn) -> irisn.mod1;
```

- To get a text-based summary of the results of the model estimation procedure (stored in irisn.mod1) do either:

```
R> summary(irisn.mod1);
```

```
Call:
glm(formula = irisn.form1, data = irisn)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.82816  -0.21989   0.01875   0.19709   0.84570

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.85600    0.25078   7.401 9.85e-12 ***
Sepal.Width   0.65084    0.06665   9.765  < 2e-16 ***
Petal.Length  0.70913    0.05672  12.502  < 2e-16 ***
Petal.Width  -0.55648    0.12755  -4.363 2.41e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.09894113)

    Null deviance: 102.168  on 149  degrees of freedom
Residual deviance:  14.445  on 146  degrees of freedom
AIC: 84.643

Number of Fisher Scoring iterations: 2
```

```
R> irisn.mod1;
```

```
Call:  glm(formula = irisn.form1, data = irisn)

Coefficients:
 (Intercept)   Sepal.Width  Petal.Length   Petal.Width
      1.8560        0.6508        0.7091       -0.5565

Degrees of Freedom: 149 Total (i.e. Null);  146 Residual
Null Deviance:      102.2
Residual Deviance: 14.45        AIC: 84.64
```

UT Southwestern Medical Center

Interpreting the output of glm

```
R> summary(irisn.mod1);
```

```
Call:
glm(formula = irisn.form1, data = irisn)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.82816  -0.21989   0.01875   0.19709   0.84570

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.85600    0.25078   7.401 9.85e-12 ***
Sepal.Width   0.65084    0.06665   9.765  < 2e-16 ***
Petal.Length  0.70913    0.05672  12.502  < 2e-16 ***
Petal.Width  -0.55648    0.12755  -4.363 2.41e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.09894113)

    Null deviance: 102.168  on 149  degrees of freedom
Residual deviance:  14.445  on 146  degrees of freedom
AIC: 84.643

Number of Fisher Scoring iterations: 2
```

Deviance is a measure of how much the likelihood for the estimated model, $\ell(\hat{\beta})$ differs with respect to the perfect (or 'saturated' model, ℓ_s)

$$D \equiv -2 \cdot \left(\ell\left(\hat{\beta}\right) - \ell_s\right)\phi \text{ [Residual Deviance]}$$

The Null deviance measures how much the likelihood function of only the estimated intercept (average) differs from the perfect (saturated model)

$$D_0 \equiv -2 \cdot \left(\ell\left(\widehat{\beta_0}\right) - \ell_s\right)\phi \text{ [Null Deviance]}$$

Residual Deviance can be subdivided into a separate deviance contribution for each of the points observed (ie likelihood of that particular observation under the estimated and saturated models).

Akaike Information Criterion: $2 \cdot k - 2\ln(\hat{L})$, where k is the number of parameters estimated and $\hat{L}$ is the maximum likelihood.

"Holding Petal Length, and Width Constant, we would expect that on average, for every one centimeter increase of sepal width, the sepal length will be 0.65084 centimeters longer"

Values computed for testing whether effects are statistically significantly different from zero.

$$Y_i = \boxed{1.856} + \boxed{0.651} \cdot X_{1_i} + \boxed{0.709} \cdot X_{2_i} - \boxed{0.556} \cdot X_{3_i} + \varepsilon_i$$

For Gaussian Models, Dispersion = MSE (estimated from data)

$Y_i \equiv$ Sepal Length (cm) of i^{th} observation.
$X_{1_i} \equiv$ Sepal Width (cm) of i^{th} observation.
$X_{2_i} \equiv$ Petal Length (cm) of i^{th} observation.
$X_{3_i} \equiv$ Petal width (cm) of i^{th} observation.
$\varepsilon_i \equiv$ Random error associated with i^{th} observation.
$\varepsilon_i \sim N(0, \sigma_i)$

Iterations of Fisher Scoring needed to solve for MLEs (Type of Newton's Method)

UT Southwestern Medical Center

Interpreting the output of glm

```
R> summary(irisn.mod1);
```

```
Call:
glm(formula = irisn.form1, data = irisn)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.82816  -0.21989   0.01875   0.19709   0.84570

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.85600    0.25078   7.401 9.85e-12 ***
Sepal.Width   0.65084    0.06665   9.765  < 2e-16 ***
Petal.Length  0.70913    0.05672  12.502  < 2e-16 ***
Petal.Width  -0.55648    0.12755  -4.363 2.41e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.09894113)

    Null deviance: 102.168  on 149  degrees of freedom
Residual deviance:  14.445  on 146  degrees of freedom
AIC: 84.643

Number of Fisher Scoring iterations: 2
```

Questions:

- **What/How can the intercept be interpreted in this model?**
- **Does it make sense to have an intercept included in the model?**

$$Y_i = 1.856 + 0.651 \cdot X_{1_i} + 0.709 \cdot X_{2_i} - 0.556 \cdot X_{3_i} + \varepsilon_i$$

$Y_i \equiv$ Sepal Length (cm) of i^{th} observation.
$X_{1_i} \equiv$ Sepal Width (cm) of i^{th} observation.
$X_{2_i} \equiv$ Petal Length (cm) of i^{th} observation.
$X_{3_i} \equiv$ Petal width (cm) of i^{th} observation.
$\varepsilon_i \equiv$ Random error associated with i^{th} observation.
$\varepsilon_i \sim N(0, \sigma_i)$

UT Southwestern Medical Center

Interpreting the output of glm

```
R> summary(irisn.mod1);
```

```
Call:
glm(formula = irisn.form1, data = irisn)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.82816  -0.21989   0.01875   0.19709   0.84570

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.85600    0.25078   7.401 9.85e-12 ***
Sepal.Width   0.65084    0.06665   9.765  < 2e-16 ***
Petal.Length  0.70913    0.05672  12.502  < 2e-16 ***
Petal.Width  -0.55648    0.12755  -4.363 2.41e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.09894113)

    Null deviance: 102.168  on 149  degrees of freedom
Residual deviance:  14.445  on 146  degrees of freedom
AIC: 84.643

Number of Fisher Scoring iterations: 2
```

$$Y_i = 1.856 + 0.651 \cdot X_{1_i} + 0.709 \cdot X_{2_i} - 0.556 \cdot X_{3_i} + \varepsilon_i$$

$Y_i \equiv$ Sepal Length (cm) of i^{th} observation.
$X_{1_i} \equiv$ Sepal Width (cm) of i^{th} observation.
$X_{2_i} \equiv$ Petal Length (cm) of i^{th} observation.
$X_{3_i} \equiv$ Petal width (cm) of i^{th} observation.
$\varepsilon_i \equiv$ Random error associated with i^{th} observation.
$\varepsilon_i \sim N(0, \sigma_i)$

Questions:

- **What/How can the intercept be interpreted in this model?**
- **Does it make sense to have an intercept included in the model?**

When the sepal width is zero, it would not be possible to measure the sepal length, hence it makes no sense to include an intercept in the model! Let us specify a new model that does not include an intercept.

UT Southwestern Medical Center

The Multivariate Zero-Intercept Linear Model in R Example (Iris Data)

- The zero-intercept model may be written by slightly modifying the previous fixed effects model.

$$Y_i = \beta_1 X_{1_i} + \beta_2 X_{2_i} + \beta_3 X_{3_i} + \varepsilon_i$$

Y_i : sepal length (cm) of the i^{th} observed iris.
X_{1_i} : sepal width (cm) of the i^{th} observed iris.
X_{2_i} : petal length (cm) of the i^{th} observed iris.
X_{3_i} : petal width (cm) of the i^{th} observed iris.
ε_i : The random error associated with the i^{th} observation.

- We can specify the **zero-intercept fixed effects model** formula by utilizing the **0+** syntax in an R-formulae.

```
R> formula(Sepal.Length~0+Sepal.Width+Petal.Length+Petal.Width) -> irisn.form2;
```

UT Southwestern Medical Center

Revising the GLM Formula, and glm options (1)

Recall from Intercept Model:

```
    Null deviance: 102.168  on 149  degrees of freedom
Residual deviance:  14.445  on 146  degrees of freedom
AIC: 84.643
```

```
Call:
glm(formula = irisn.form2, data = irisn)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7772  -0.2372   0.0517   0.2637   0.9897

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
Sepal.Width   1.12106    0.02352  47.658  < 2e-16 ***
Petal.Length  0.92353    0.05699  16.205  < 2e-16 ***
Petal.Width  -0.89568    0.13911  -6.439 1.61e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.1351351)

    Null deviance: 5223.850  on 150  degrees of freedom
Residual deviance:   19.865  on 147  degrees of freedom
AIC: 130.43

Number of Fisher Scoring iterations: 2
```

- Even though the Models Residual deviance and AIC increased relative to the intercept model, this model makes more physical sense.
- This model is still not perfectly indicative of internal relationships among the variables.

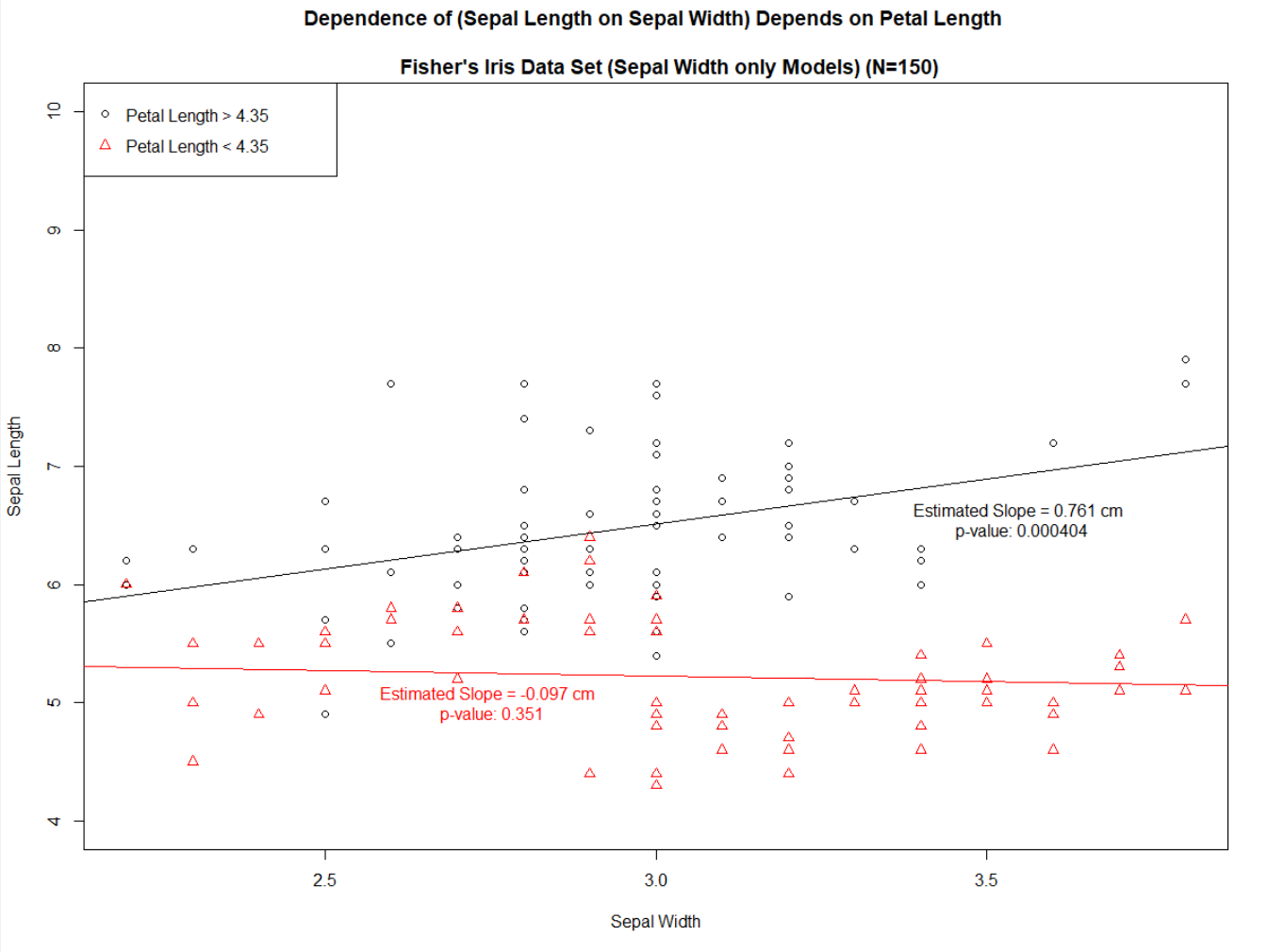
Question:

- **How else is the model inaccurate?**

UT Southwestern Medical Center

The Multivariate Joint Effects Linear Model in R Example (Iris Data)

- If there is reason to believe that the dependent variables might jointly affect the outcome, we may wish to include specific variables in the model



- Split the 150 iris dataset into two equal sized sets of 75, based on whether petal length was higher than the median (4.35 cm) for the set.
  - (Petal Length > 4.35 cm): On average Sepal length will be 0.761 cm longer for each centimeter of length of sepal width.
  - (Petal Length < 4.35 cm): On average the Sepal Length is **Not** statistically significantly associated with the Sepal Width.

- This relationship may be captured by examining **joint-effects** in the general linear model.

UT Southwestern Medical Center

The Multivariate Joint Effects Linear Model in R Example (Iris Data)

- The Full joint effects model may be written by slightly modifying the previous fixed effects model. $(\cdot)$ added for readability.

$$Y_i = \beta_1 X_{1_i} + \beta_2 X_{2_i} + \beta_3 X_{3_i} + \beta_{12}(X_{1_i} X_{2_i}) + \beta_{13}(X_{1_i} X_{3_i}) + \beta_{23}(X_{2_i} X_{3_i}) + \beta_{123}(X_{1_i} X_{2_i} X_{3_i}) + \varepsilon_i$$

Y_i : sepal length (cm) of the i^{th} observed iris.
X_{1_i} : sepal width (cm) of the i^{th} observed iris.
X_{2_i} : petal length (cm) of the i^{th} observed iris.
X_{3_i} : petal width (cm) of the i^{th} observed iris.
ε_i : The random error associated with the i^{th} observation.

- Note that here four additional parameters must be estimated, but due to the fact that we have 150 observations (and hence degrees of freedom), this is a reasonable number of parameters to estimated.

- We can specify the **full joint zero-intercept fixed effects model** formula by utilizing the **0+** syntax in an R-formulae. In a few ways.

The Multivariate Joint Effects Linear Model in R Example (Iris Data)

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_{12}(X_{1i}X_{2i}) + \beta_{13}(X_{1i}X_{3i}) + \beta_{23}(X_{2i}X_{3i}) + \beta_{123}(X_{1i}X_{2i}X_{3i}) + \varepsilon_i$$

- All effects can be added with one term using the `*' operator (which adds the joint effects denoted by the combinations of the operands specified, as well as all lower ordered effects).

```
R> formula(Sepal.Length~0+Sepal.Width*Petal.Length*Petal.Width) -> irisn.form3.a;
```

- The above creates a formula which will estimate effects for the joint effect indicated (three-way joint effect between Sepal Width, and Petal Length, and Width) and all two-joint effects (sepal width and petal length, sepal width and petal width, and petal length and petal width) and the singleton effects for each of sepal width, petal length, and petal width.
- Remove unwanted but automatically included effects with `-':

```
R> formula(Sepal.Length~0+Sepal.Width*Petal.Length - Petal.Length) -> irisn.form4;
```

- Joint effects can be added one at a time using the `:' operator.

```
R> formula(Sepal.Length~0+Sepal.Width+Petal.Length+Petal.Width+
                                        Sepal.Width:Petal.Length +
                                        Sepal.Width:Petal.Width +
                                        Petal.Length:Petal.Width +
                                        Sepal.Width:Petal.Length:Petal.Width) -> irisn.form3.b;
```

UT Southwestern Medical Center

The Multivariate Joint Effects Linear Model in R Example (Iris Data)

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_{12}(X_{1i}X_{2i}) + \beta_{13}(X_{1i}X_{3i}) + \beta_{23}(X_{2i}X_{3i}) + \beta_{123}(X_{1i}X_{2i}X_{3i}) + \varepsilon_i$$

```
Call:
glm(formula = irisn.form3.b, data = irisn)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.82784  -0.20332  -0.01709   0.22545   0.75678

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
Sepal.Width                       1.27703    0.03805  33.562  < 2e-16 ***
Petal.Length                      1.77850    0.33157   5.364  3.2e-07 ***
Petal.Width                      -1.53678    1.46791  -1.047  0.29690
Sepal.Width:Petal.Length         -0.36694    0.11357  -3.231  0.00153 **
Sepal.Width:Petal.Width           0.24719    0.48389   0.511  0.61025
Petal.Length:Petal.Width         -0.14219    0.15576  -0.913  0.36284
Sepal.Width:Petal.Length:Petal.Width  0.06658    0.05359   1.242  0.21616
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.09930425)

    Null deviance: 5223.850  on 150  degrees of freedom
Residual deviance:   14.201  on 143  degrees of freedom
AIC: 88.078

Number of Fisher Scoring iterations: 2
```

"For each one cm increase in Petal Length, the expected change in Sepal Length per one cm increase in Septal Width will decrease (on average, or 'an expected') from 1.27703 cm by 0.367 cm"

-or-

"For each one cm increase in Sepal Width, the expected change in Sepal Length per one cm increase in Petal Length will decrease (on average from, or 'an expected') from 1.77850 cm by 0.367 cm"

Note Number of Potential interpretations of Joint effect = number variables involved.

UT Southwestern Medical Center

The Multivariate Joint Effects Linear Model in R Example (Iris Data)

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_{12}(X_{1i}X_{2i}) + \beta_{13}(X_{1i}X_{3i}) + \beta_{23}(X_{2i}X_{3i}) + \beta_{123}(X_{1i}X_{2i}X_{3i}) + \varepsilon_i$$

- We Also can 'set fixed values' for the coefficients ahead of time (to remove their estimation from the model by effectively reducing the degrees of freedom. We can do this using the offset() function in R as follows:

```
R> formula(Sepal.Length~0+offset(1.3*Sepal.Width)+Petal.Length+Petal.Width+
                                                  Sepal.Width:Petal.Length +
                                                  Sepal.Width:Petal.Width +
                                                  Petal.Length:Petal.Width +
                                                  Sepal.Width:Petal.Length:Petal.Width) -> irisn.form5;
```

- This will fix the estimate of Sepal Width's effect in the model at 1.3 (recall from earlier that 1.27703 was the mle).

```
Call:
glm(formula = irisn.form5, data = iris)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.82085  -0.20750  -0.01842   0.21217   0.74237

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
Petal.Length                     1.84484    0.31214   5.910 2.36e-08 ***
Petal.Width                     -1.58898    1.46212  -1.087 0.278955
Petal.Length:Sepal.Width        -0.39898    0.10018  -3.982 0.000108 ***
Petal.Width:Sepal.Width          0.25400    0.48268   0.526 0.599536
Petal.Length:Petal.Width        -0.16170    0.15203  -1.064 0.289279
Petal.Length:Petal.Width:Sepal.Width  0.07784    0.05013   1.553 0.122685
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.09886593)

    Null deviance: 690.308  on 150  degrees of freedom
Residual deviance:  14.237  on 144  degrees of freedom
AIC: 86.46

Number of Fisher Scoring iterations: 2
```

- Since nothing was estimated (no parameter anyway) nothing is reported for sepal.Width in this case.
- In this way, this model is constructed using one sliver of the likelihood from the full model, the sliver where Sepal Width = 1.3.
- If we had fixed Sepal Width at 1.27703 (the MLE) the model would be based on that slivers Likelihood, which has a specific name, the "profile likelihood".
- So when we see "waiting for profiling to be done", with confint, it is fitting many models near the MLE.

UT Southwestern Medical Center

Including categorical predictors in model (Iris Data)

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3_i} + \beta_{12}(X_{1_i} X_{2_i}) + \beta_{13}(X_{1_i} X_{3_i}) + \beta_{23}(X_{2_i} X_{3_i}) + \beta_{123}(X_{1_i} X_{2_i} X_{3_i}) + \varepsilon_i$$

Suppose that the species of the iris (of three known species) is also identified for the data, and we would like to include this information in our model for predicting the Sepal Length.

The `species' can be thought of as potentially having an effect on every one of the model terms involved in the original model above, to the point where we might consider three separate models (the above estimated within each species)

We can easily do this, and store the results separately for comparison.

```
R> glm(irisn.form3.a,data=iris[iris$Species=="setosa",])-> irisn.setosa.full;
R> glm(irisn.form3.a,data=iris[iris$Species=="versicolor",])-> irisn.versicolor.full;
R> glm(irisn.form3.a,data=iris[iris$Species=="virginica",])-> irisn.virginica.full;
```

UT Southwestern Medical Center

Including categorical predictors in model (Iris Data)

For joint effects, it appears that only sepal.width and petal.length are statistically significantly associated with Sepal.length only for the Setosa data

	Setosa Data		Versicolor Data		Virginica Data	
	Coefficient	**P-Value**	**Coefficient**	**P-Value**	**Coefficient**	**P-Value**
Singleton Effects						
Sepal.Width	1.09421764	0.00006382	1.04536391	0.51501824	-0.43214247	0.79442199
Petal.Length	1.67576250	0.01203815	0.10393301	0.93175056	0.34316043	0.59183366
Petal.Width	4.37981548	0.72735137	6.65459693	0.35813542	-0.38682456	0.86235329
Bivariate Effects						
Sepal.Width:Petal.Length	-0.25787355	0.30352281	0.29561687	0.58652482	0.38794463	0.21051051
Sepal.Width:Petal.Width	-0.16976886	0.96001151	-2.63673077	0.26297065	0.58467794	0.54566111
Petal.Length:Petal.Width	-1.93346198	0.81956704	-0.77341468	0.52143425	0.39196892	0.33278623
Trivariate Effects						
Sepal.Width:Petal.Length:Petal.Width	-0.09867646	0.96548911	0.27777579	0.40545117	-0.22532492	0.13821805

When estimating the model effects in each of the different data sets separately we only find a few significant effects.

When adding categorical variables such as species to the model there is essentially one primary consideration, whether or not the categories are ordinal, in this case, they are not.

```
R> iris.form.1 <- formula(Sepal.Length~Sepal.Width*Petal.Length*Petal.Width + as.factor(Species));
R> iris.mod.1 <- glm(iris.form.1,data=iris);
```

UT Southwestern Medical Center

Including categorical predictors in model (Iris Data)

```
Call:
glm(formula = iris.form.1, data = iris)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7758  -0.2440   0.0075   0.2023   0.7204

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                        1.53971    0.93369   1.649  0.10138
Sepal.Width                        0.79080    0.27184   2.909  0.00422 **
Petal.Length                       1.01238    0.53125   1.906  0.05874 .
Petal.Width                       -1.21721    1.76807  -0.688  0.49231
Speciesversicolor                 -0.28039    0.45134  -0.621  0.53546
Speciesvirginica                  -0.63456    0.51171  -1.240  0.21702
Sepal.Width:Petal.Length          -0.11740    0.16310  -0.720  0.47286
Sepal.Width:Petal.Width            0.14887    0.52907   0.281  0.77883
Petal.Length:Petal.Width           0.14119    0.30068   0.470  0.63938
Sepal.Width:Petal.Length:Petal.Width -0.01106    0.08753  -0.126  0.89964
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.09327036)

    Null deviance: 102.168  on 149  degrees of freedom
Residual deviance:  13.058  on 140  degrees of freedom
AIC: 81.495

Number of Fisher Scoring iterations: 2
```

- Notice how two entries were added automatically (one for each category other than the reference of the categorical variable).
- By default the category level name is concatenated with the variable name and included in the list.
- 'Species' is handled correctly because it is already a factor variable, character variables are treated as categorical by default. [use as.factor() to force this]
- Notice also how setosa is omitted from the list, this is because that is the reference level (for which the intercept makes sense again).

Questions:

(1) Why does including this categorical variable in the model cause the intercept estimation to make logical sense again?

(2) What is the interpretation of a coefficient estimated for a categorical variable?

UT Southwestern Medical Center

Interpreting categorical variable effects (Iris Data)

```
Call:
glm(formula = iris.form.1, data = iris)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7758  -0.2440   0.0075   0.2023   0.7204

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                        1.53971    0.93369   1.649  0.10138
Sepal.Width                        0.79080    0.27184   2.909  0.00422 **
Petal.Length                       1.01238    0.53125   1.906  0.05874 .
Petal.Width                       -1.21721    1.76807  -0.688  0.49231
Speciesversicolor                 -0.28039    0.45134  -0.621  0.53546
Speciesvirginica                  -0.63456    0.51171  -1.240  0.21702
Sepal.Width:Petal.Length          -0.11740    0.16310  -0.720  0.47286
Sepal.Width:Petal.Width            0.14887    0.52907   0.281  0.77883
Petal.Length:Petal.Width           0.14119    0.30068   0.470  0.63938
Sepal.Width:Petal.Length:Petal.Width -0.01106    0.08753  -0.126  0.89964
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.09327036)

    Null deviance: 102.168  on 149  degrees of freedom
Residual deviance:  13.058  on 140  degrees of freedom
AIC: 81.495

Number of Fisher Scoring iterations: 2
```

- Neither of these effects is significant, but nonetheless the effects were estimated and can be interpreted as saying that “on average when compared with those of the setosa species, irises of the versicolor species had sepal lengths that were 0.2804 cm shorter, and those of virginica had 0.63456 cm shorter lengths”.
- Note that the interpretation depends on the reference level.

UT Southwestern Medical Center

Changing the reference for a categorical effect in R (Iris Data)

```
R> iris$Species <- relevel(as.factor(iris$Species),ref='versicolor')
R> iris.mod.1.ref2 <- glm(iris.form.1,data=iris);
```

```
Call:
glm(formula = iris.form.1, data = iris)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7758  -0.2440   0.0075   0.2023   0.7204

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                        1.25933    1.19023   1.058  0.29185
Sepal.Width                        0.79080    0.27184   2.909  0.00422 **
Petal.Length                       1.01238    0.53125   1.906  0.05874 .
Petal.Width                       -1.21721    1.76807  -0.688  0.49231
Speciessetosa                      0.28039    0.45134   0.621  0.53546
Speciesvirginica                  -0.35418    0.12585  -2.814  0.00559 **
Sepal.Width:Petal.Length          -0.11740    0.16310  -0.720  0.47286
Sepal.Width:Petal.Width            0.14887    0.52907   0.281  0.77883
Petal.Length:Petal.Width           0.14119    0.30068   0.470  0.63938
Sepal.Width:Petal.Length:Petal.Width -0.01106    0.08753  -0.126  0.89964
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.09327036)

    Null deviance: 102.168  on 149  degrees of freedom
Residual deviance:  13.058  on 140  degrees of freedom
AIC: 81.495

Number of Fisher Scoring iterations: 2
```

- Observe that the statistical significance of the difference in sepal length between versicolor and virginica was obfuscated in the previous model where the reference was setosa.
- Choice of reference is very important in model interpretation!

UT Southwestern Medical Center

Adding Ordinal Effects in an R Model (Iris Data)

- Just as we used the 'as.factor()' function in the formula for the categorical variables which were unordered, we may use the 'ordered()' function for those which are ordered.
- First let us create an ordinal variable in the iris dataset based on the value of petal-width. Let us create three equally sized groups with small, medium and large petal sizes. In R we can easily do this using the 'cut()' function.

```
R> iris$PW.O <- cut(iris$Petal.Width,3)
```

- We can Now regress Petal Length on this ordinal variable, and investigate the results.

```
R> iris.ordinal.form <- formula(Petal.Length ~ ordered(PW.O),
                                levels = c("(0.0976,0.9]",
                                           "(0.9,1.7]",
                                           "(1.7,2.5]"));

R> iris.ordinal.mod <- glm(iris.ordinal.form);
R> summary(iris.ordinal.mod);
```

".L" is Linear ".Q" quadratic, ".C" cubic, "^4" and so on, fourth and higher ordered trends

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.79098    0.03755  100.97   <2e-16 ***
ordered(PW.O).L   2.90756    0.06629   43.86   <2e-16 ***
ordered(PW.O).Q  -0.66878    0.06375  -10.49   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation of Estimates is hard in this case, A significant result really only indicates that there is statistical evidence of a Linear/Quadratic/Cubic/etc… trend by level.

For more information on the ordinal contrast beta estimates.

https://blogs.uoregon.edu/rclub/2015/11/03/anova-contrasts-in-r/

https://stackoverflow.com/questions/57297771/interpretation-of-l-q-c-4-for-logistic-regression

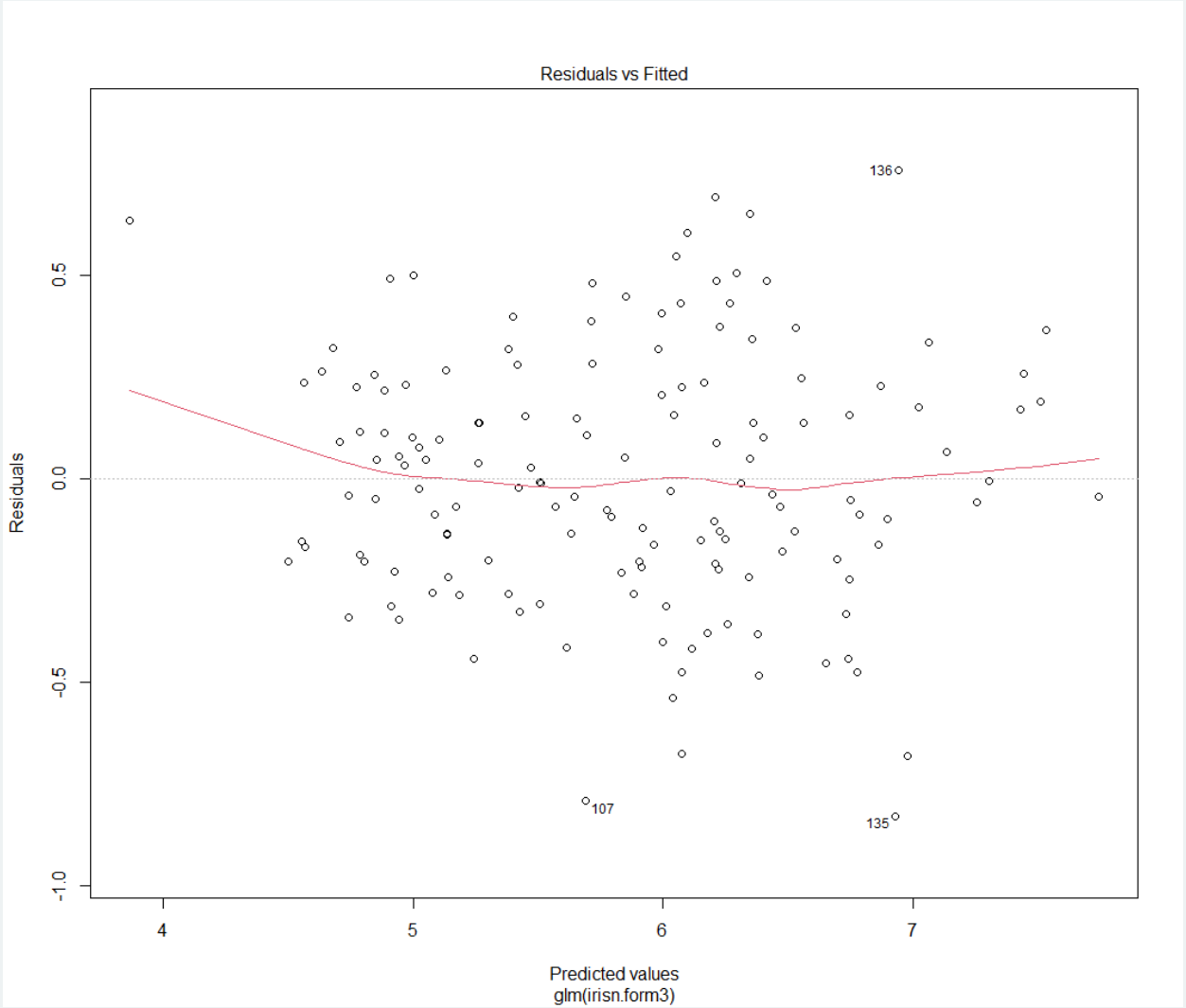
https://github.com/RInterested/SIMULATIONS_and_PROOFS/blob/master/Contrasts%20Polynomial

https://stats.stackexchange.com/questions/105115/polynomial-contrasts-for-regression

UT Southwestern Medical Center

Plots and Interpretations associated with the glm() object (1)

```
R> iris.form.1 <- formula(Sepal.Length~Sepal.Width*Petal.Length*Petal.Width+Species);
R> iris.mod.1 <- glm(iris.form.1,data=iris);
R> plot(iris.mod.1);
```

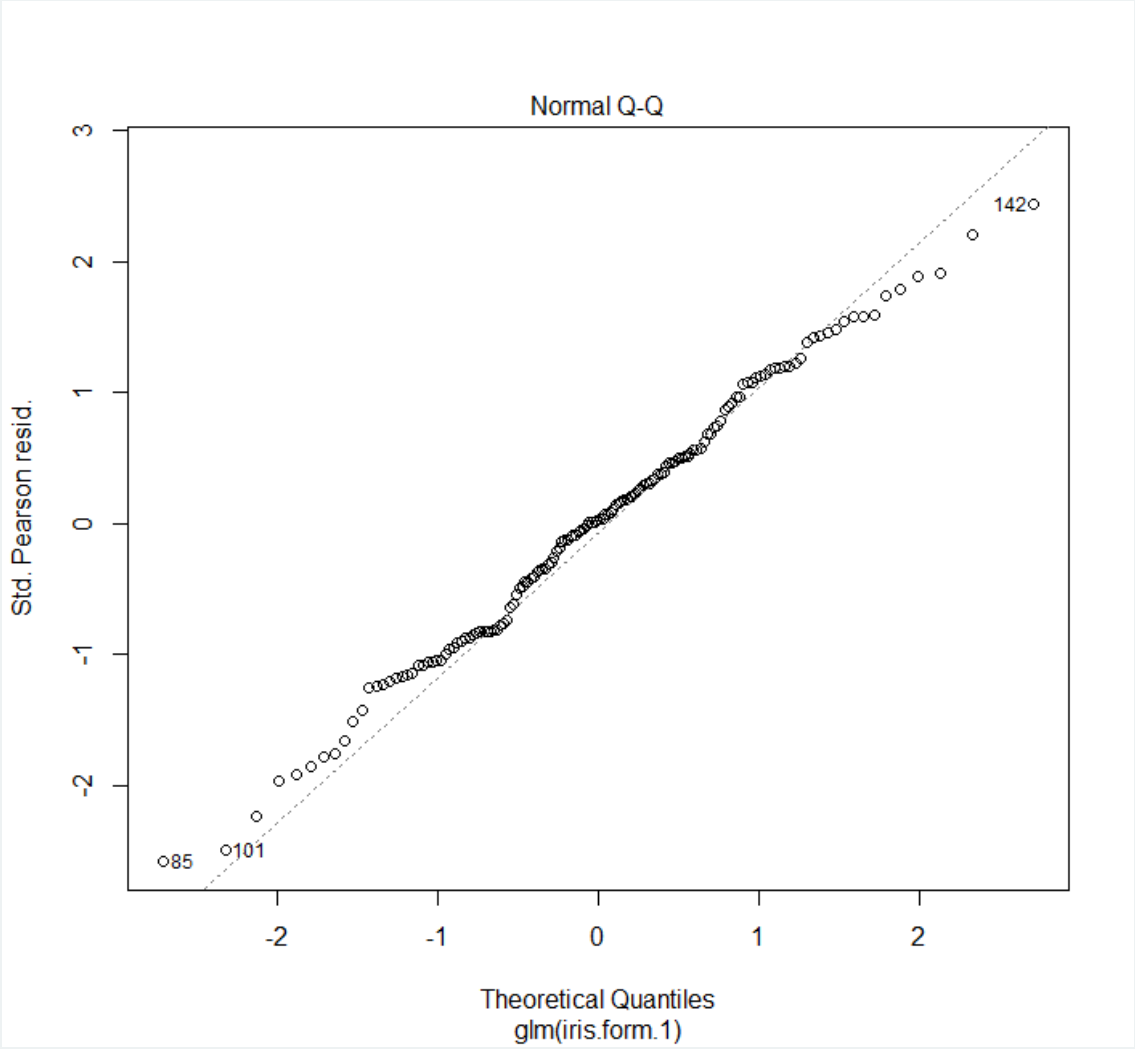


- Using plot() on a glm object will give useful information and diagnostics for validating and assessing the validity of the assumptions for using a generalized linear model.
- The first plot gives the actual predicted value from the model for each datapoint on the x-axis, and the associated residual on the y-axis.
- This gives the user the ability to determine whether the residuals tended to depend in any obvious or meaningful way on the actual value that is predicted.
- All points which produce a particularly high residual value will be labeled on this plot.

UT Southwestern Medical Center

Plots and Interpretations associated with the glm() object (2)

```
R> iris.form.1 <- formula(Sepal.Length~Sepal.Width*Petal.Length*Petal.Width+Species);
R> iris.mod.1 <- glm(iris.form.1,data=iris);
R> plot(iris.mod.1);
```

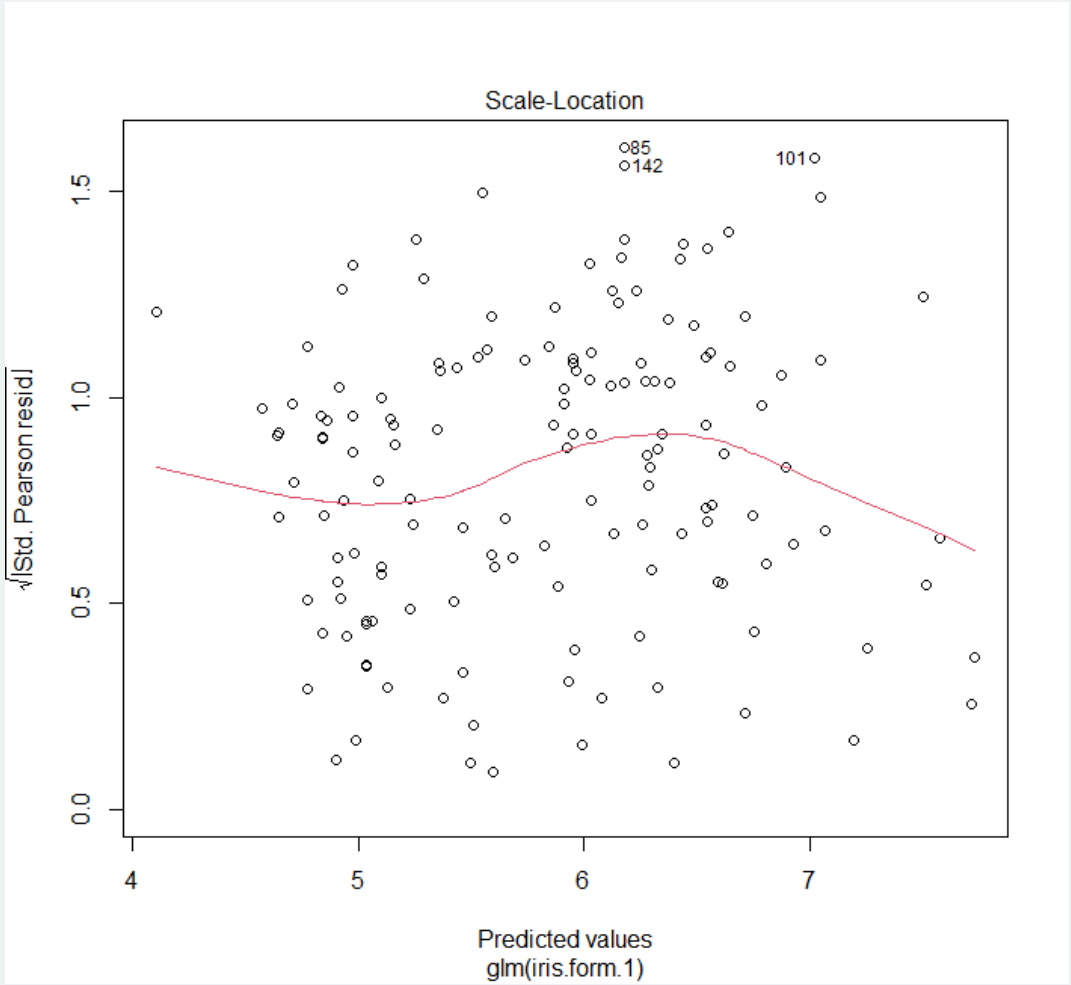


- In the next plot, the standardized Pearson Residuals from the model (Observed-Expected/Standard Error) are plotted against what would be expected from the standard normal distribution for specific quantiles.
- In this plot we can determine whether the models residuals tend to follow the normal distribution (a key assumption for the multivariate Linear Model – which assumes the Gaussian distribution for the errors) or whether and by how much they diverge.
- Again, the extreme observations of the data here are labeled by their ID.
- The diagonal line indicates perfect correspondence to the normal distribution.

UT Southwestern
Medical Center

Plots and Interpretations associated with the glm() object (2)

```
R> iris.form.1 <- formula(Sepal.Length~Sepal.Width*Petal.Length*Petal.Width+Species);
R> iris.mod.1 <- glm(iris.form.1,data=iris);
R> plot(iris.mod.1);
```

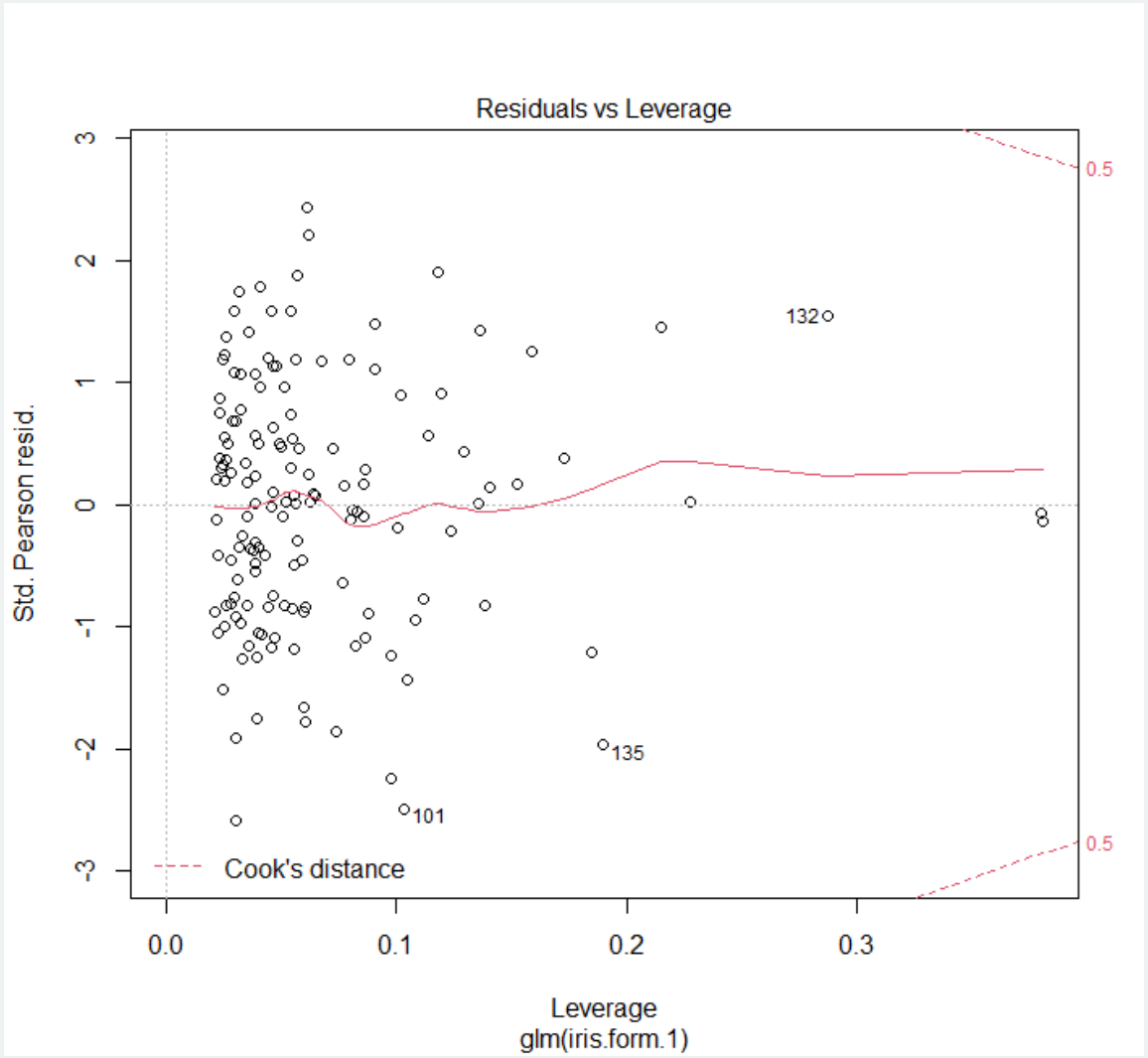


- Next the user gets the opportunity to assess the heteroscedasticity of the errors again (the dependence of the errors – in this case represented as the square root of the standard pearson residual, a quantity expected to be roughly normally distributed – on the actual predicted value).
- The distribution should tend to show more observations nearer the moving average (indicated by the read line) and fewer observations as the residual moves away from the mean.
- This plot also allows the assessment of the distribution of the errors across the predicted values (the scale and location across predicted values is visualized).

UT Southwestern
Medical Center

Plots and Interpretations associated with the glm() object (2)

```
R> iris.form.1 <- formula(Sepal.Length~Sepal.Width*Petal.Length*Petal.Width+Species);
R> iris.mod.1 <- glm(iris.form.1,data=iris);
R> plot(iris.mod.1);
```



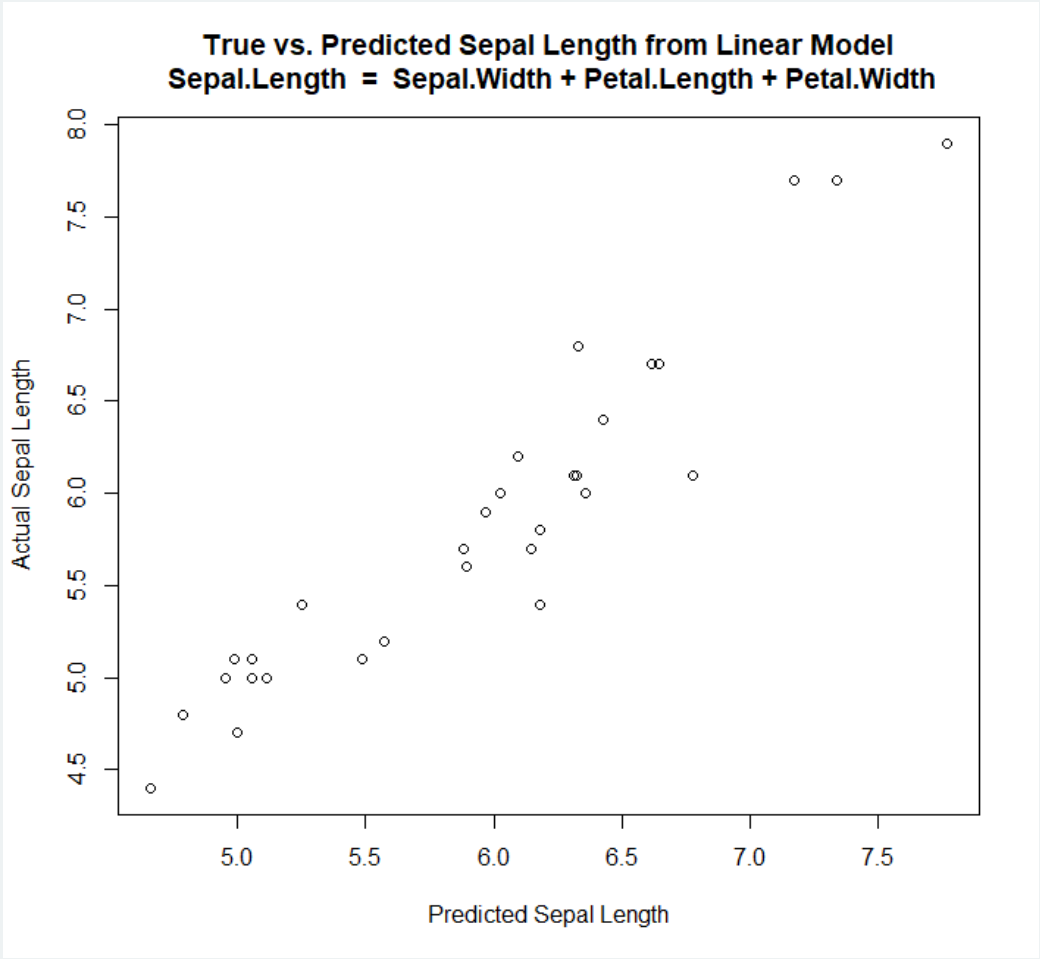
- The final of the four plots produced by using plot() on a glm object shows a plot of the residuals vs. the leverage of an observation.
- The leverage is a measure of how influential an observation is on the estimation of the parameters in the model, such that the higher the leverage the more influence the observation has on the final estimation of the model effects.
- In this plot we can determine whether there are any outlying values which have a large influence on the parameter estimation, while retaining a large error.
- Such observations could stand to be removed when generalizing the model, as these indicate special-case observations not appropriate for use in forming general models.

UT Southwestern Medical Center

Using an estimated Linear Model for Prediction (Interpolation)

- We can use a fitted model in R to predicted outcome variables for new data (or data held out, or included in the original model effect estimation).

```
R> f.p <- formula(Sepal.Length ~ Sepal.Width+Petal.Length+Petal.Width);
R> iris.pmod <- glm(f.p,data=iris);
```



- We can use the predict() function on a glm() object, and some new data in order to make predictions about the dependent variable.
- For example in the R code below, replace 1,2, and 3 with the new data for prediction.

```
R> predict(iris.pmod,data.frame(Sepal.Width=1,
                                Petal.Length=2,
                                 Petal.Width=3)
```

- Left we can see a plot of the Sepal Length predictions on 20 previously unseen observations from a LM trained on 130 observations, with the true values plotted on the Y-axis.

UT Southwestern Medical Center

Functions for Extracting Information from glm objects

```
R> f.p <- formula(Sepal.Length ~ Sepal.Width+Petal.Length+Petal.Width);
R> iris.pmod <- glm(f.p,data=iris);
```

- Coef(), coefficients() – Will return vector with numerical coefficient estimates (including intercept).

```
R> coef(iris.pmod)
```

```
(Intercept) Petal.Length  Sepal.Width  Petal.Width
  1.8559975    0.7091320    0.6508372   -0.5564827
```

- AIC(), BIC(), deviance(), and logLik() – Will return each of these values for the model.

```
R> c(AIC(iris.pmod),BIC(iris.pmod),deviance(iris.pmod),logLik(iris.pmod))
```

```
[1]  84.64272  99.69590
[3]  14.44540 -37.32136
```

- Aov(), anova(). – give information about the analysis of variance (or deviance) in the model.
- fitted.values(). – provides a vector with estimates of the dependent effect for each observation used in estimation.
- Resid(), residuals(). – gives the true value minus the predicted value for all observations used in the estimation.
- Confint(). – produce confidence intervals for estimated effects using profile likelihood
- Case.names(), variable.names() – retrieve the row and column names used in training the model.
- Influence.measures() – Retrieve matrix with measures of how influential each observation was on overall estimation
- Model.frame(), - dataframe originally used to train model
- Model.matrix(),- dataframe with training variables (and all 1's column for intercept – desing matrix)

UT Southwestern Medical Center

Functions for Extracting Information from glm objects (continued)

- Effects(), - get the effects (for all terms in the model, and the resulting orthogonal effects (for residuals))
- Proj(), - Will produce projections (evaluated or fitted.values in glm) for all modeled data.
- Vcov(), - Get the variance covariance matrix for the model terms.
- Weights(), - extract weights (should be all ones in linear models such as this).
- Confint(),- assumes normality + builds confidence intervals correspondingly if vcov and coef functions available for object. https://www.clayford.net/statistics/profile-likelihood-ratio-confidence-intervals/

```
R> confint(iris.pmod)
```

```
                2.5 %     97.5 %
(Intercept)   1.3644834  2.3475116
Petal.Length  0.5979642  0.8202997
Sepal.Width   0.5202107  0.7814637
Petal.Width  -0.8064720 -0.3064933
```

The profiling method amounts to trying many values around the chosen value (so building many models).

"We are 95% confident for irises the true average change in Sepal Length will be an increase of between 0.598 and 0.820 cm per centimeter of Petal Length"

-- Confidence is in the procedure used to arrive at the interval estimates.

```
R> vcov(iris.pmod)
```

```
              (Intercept) Petal.Length  Sepal.Width  Petal.Width
(Intercept)   0.062889160 -0.007264687 -0.015933225  0.011493320
Petal.Length -0.007264687  0.003217078  0.001142174 -0.006934766
Sepal.Width  -0.015933225  0.001142174  0.004441875 -0.001617029
Petal.Width   0.011493320 -0.006934766 -0.001617029  0.016268479
```

UT Southwestern Medical Center

Getting Effect-Specific T-Test Result Data from GLM

- To get the information regarding the specifics of the t-tests for each of the effects separately, we need to use the coefficients of the summary of the glm object.

- For instance, to get the coefficients, and their associated p-values, we and adjust these p-values using the bonferroni correction, finally displaying the three (coefficients, unadjusted, and adjusted p-values) together in a table we can use the following:

```
R> f.p <- formula(Sepal.Length ~ Sepal.Width+Petal.Length+Petal.Width);
R> iris.pmod <- glm(f.p,data=iris);
R> iris.pmod.estim <- coefficients(summary(iris.pmod))[,1];
R> iris.pmod.pvals <- coefficients(summary(iris.pmod))[,4];
R> iris.pmod.pvals.adj <- p.adjust(iris.pmod.pvals,method = "bonferroni");
R> cbind(iris.pmod.estim,iris.pmod.pvals,iris.pmod.pvals.adj);
```

	iris.pmod.estim	iris.pmod.pvals	iris.pmod.pvals.adj
(Intercept)	1.8559975	9.853855e-12	3.941542e-11
Petal.Length	0.7091320	7.656980e-25	3.062792e-24
Sepal.Width	0.6508372	1.199846e-17	4.799383e-17
Petal.Width	-0.5564827	2.412876e-05	9.651503e-05

Current Multiple Adjustment Methods Available in R (4.1.1) [use p.adjust.methods() to see]

"holm", "Hochberg", "Hommel", "Bonferroni", "BH", "BY", "FDR", "none".

UT Southwestern Medical Center

Introduction to R for Beginners Session III Part I Review

We have seen…

- How to setup formulas for describing model relationships in R's formula language
  - **The zero-intercept vs. intercept models**
  - **Adding Joint Effects**
  - **Adding Categorical Effects**
- What a linear model is.
- How to apply glm() to estimate model effects in linear models.
- How to view and interpret the output of the summary of the glm object.

We are going to see…

- How the generalized linear model contains these multivariate linear models, and extends them:
  - Logistic & Count Regression
- What does plotting the output provide
- What are the assumptions for using these models, and how can I assess them in R?
  - Adjusting data through transforms to fit the data.
- How can we select the most useful variables for our models?
  - Elastic-net (ridge/LASSO – shrinkage estimation), Stepwise selection based on AIC

Notes!

- So far everything we have done can be done with `lm' in R, `lm' stands for linear model, and is used for fitting linear models (which include anova, ancova, and regression), it fits using ordinary least squares (that is minimizing the residual difference) this is why instead of deviance residuals, the actual residuals are reported by default when 'lm' is used.
- 'glm' is more general and implements what is known as the generalized linear model, of which the linear regression we have been doing is just one example.

UT Southwestern Medical Center

Creating a Multivariate Linear Model in R with glm(): Exercise 1

- Use R, and what we have learned to help you answer these questions about the swiss dataset in a .txt file:
  1. What year that the dataset “swiss” was collected?
  2. Adjusting for Education, Catholic, and infant.Mortality,
     1. What is the expected average increase/decrease in Fertility corresponding to a 1% increase in the percentage of males working in agriculture (while also adjusting for Examination)?
     2. What is the expected average increase/decrease in Fertility corresponding to a 1% increase in the percentage of draftees into the military who had education beyond highschool, (while also adjusting for Examination, and Agriculture)?
  3. What is the AIC for a linear model relating Fertility to each of and all combinations involving the variables Education, Agriculture, and Infant.Mortality?
  4. What is the AIC for a linear model relating Fertility to each of and all combinations involving the variables Education, Agriculture, and Infant.Mortality, except for the singleton effect of Education, and the binary joint effect of Agriculture and Infant.Mortality?
     1. Can the removal of variables from the model increase the models effectiveness in terms of AIC?

Bonus Question: We will try together.

1. Create a categorical ordinal variable corresponding to Infant.Mortality with three levels: low (less than 18.936), medium (between 18.936 and 20.708), and high (greater than 20.708), and regress Fertility on this variable alone. Does there appear to be either a linear or quadratic trend of association between Infant Mortality and Fertility?
    1. Repeat 5 using 4 categories for Infant Mortality,
    2. 5 Categories for Infant Mortality
    3. Regress Fertility on Infant Mortality and Infant Mortality squared
        1. What do these results indicate about categorizing continuous variables?

UT Southwestern Medical Center

Break

UT Southwestern
Medical Center

Part II:
The Generalized Linear Model and Its Assumptions:
Logistic, Probit, and Count Regressions
&
Predictor Assessment, Transform, and Selection

UT Southwestern
Medical Center

The Generalized Linear Model

- When I talked about the linear model previously, I specifically did not reveal all of the details, but I did mention that all that we have done previously could be done using the 'lm()' function instead of the 'glm()' function.
- The 'g' stands for generalized, of which we have so far been using only a very specific type.

- In the generalized linear model we relate the parameters of the distribution of an outcome variable to some pre specified functions of the explanatory variables estimating effects by minimizing the residuals (or distance to the predicted value). – You may have heard this called Ordinary Least Squares or OLS regression (We pick the model which produces the Least Squared Error between the data and the predictions)
- In the field of statistics we are generally concerned with producing estimates of distributional and model parameters which produce the highest likelihood for a given dataset, it just so happens that in the case of one very particular kind of model the OLS fit is the Maximum Likelihood or MLE fit.

- In a Generalized Linear Model we define three key components.
  1. The distribution of the outcome variable (which is typically a distribution from the exponential family of distributions – this makes the math work out nicely) with parameters θ.
  2. A systematic relationship (function) of the predictors/covariates that is linear in the effects to the linear predictor (a value, usually η).
  3. A link function (generally denoted g) which relates the linear predictor from 2 (η) to the outcome distribution parameter(s) from 1 (θ)

The Generalized Linear Model: Assumptions for R.

- Every observation is independently distributed * usually identically distributed, but do not have to be
  - This was an implied assumption in the previous case.
  - This cannot be too easily determined from the data alone, but can potentially occur as unintentional clusters in the data.

- Far looser assumptions than the original Linear Model,
  - Y_i Does not necessarily have to follow the Normal Distribution, it doesn't even technically *have to* follow an exponential family distribution, but usually does.
  - The residuals aren't required to follow the Normal distribution (but are the distribution of the Likelihood model you choose!)

- Should assess whether the model you choose is feasible for Y_i.

- One way to check if values follow those which would be expected in a general distribution is the quantile-quantile (or qq) plot.

The Generalized Linear Model Components Used Previously

- In the previous Multivariate Linear Modeling we defined three key components.

1. The distribution of the outcome variable as being Normal, with mean μ and variance σ^2.
   - This is specified by default in the 'glm()' function in R, but can be manually input as:

```
R> f.p <- formula(Sepal.Length ~ Sepal.Width+Petal.Length+Petal.Width);
R> iris.pmod <- glm(f.p,data=iris, family=gaussian);
```

2. A very simple direct linear relationship in both the covariates and outcomes (or indicators of levels of covariates/contrasts were applicable). $\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$.
   - This is determined by the formula that was input, note the linear aspect is with regard to the coefficients! We can apply any function to the variables (X_i) which we want. Ex. Sepal.Length ~ log(cos(Sepal.Width/Petal.Length)).
3. Perhaps one of the simplest link functions of all, $\eta = \mu$; that is, $g(\mu) = \mu$.
   - This is the default when family Gaussian is chosen (identity) but is specified using the (link='identity') option in the following:

```
R> f.p <- formula(Sepal.Length ~ Sepal.Width+Petal.Length+Petal.Width);
R> iris.pmod <- glm(f.p,data=iris, family=Gaussian(link='identity'));
```

- *I won't go too much further into the details, but estimation is done using Maximum likelihood estimation (and quasi likelihood in some cases) which hasn't changed per say from the last part, but generalized the OLS procedure (when the appropriate model is selected).*

UT Southwestern Medical Center

The Generalized Linear Model – QQ Plots for assessing Likelihood Fit

- We can determine whether we have chosen a good distributional fit for the model:
  - Plot the Quantiles (ie the 1st percentile, 1.1st percentile – 99.9th percentile, 100th percentile) of one data set against those of datasets generated from known types, can do this manually, or use the quantile function.
  - For example, to compare Petal Length from iris to the normal distribution quantiles we can do this with the following code:

```
R> quantile(rnorm(10000,mean(iris$Petal.Length),sd(iris$Petal.Length)),(0:1000/1000)) -> theoretical.quants;
R> quantile(iris$Petal.Length, (0:1000/1000)) -> experimental.quants;
R> plot(experimental.quants,theoretical.quants);
```

  - Can also just use the build in qqplot function on two samples:

```
R> qqplot(rnorm(10000,mean(iris$Petal.Length),sd(iris$Petal.Length)),iris$Petal.Length);
```

  - We can actually even use the built in qqnorm function for comparing to the normal distribution:

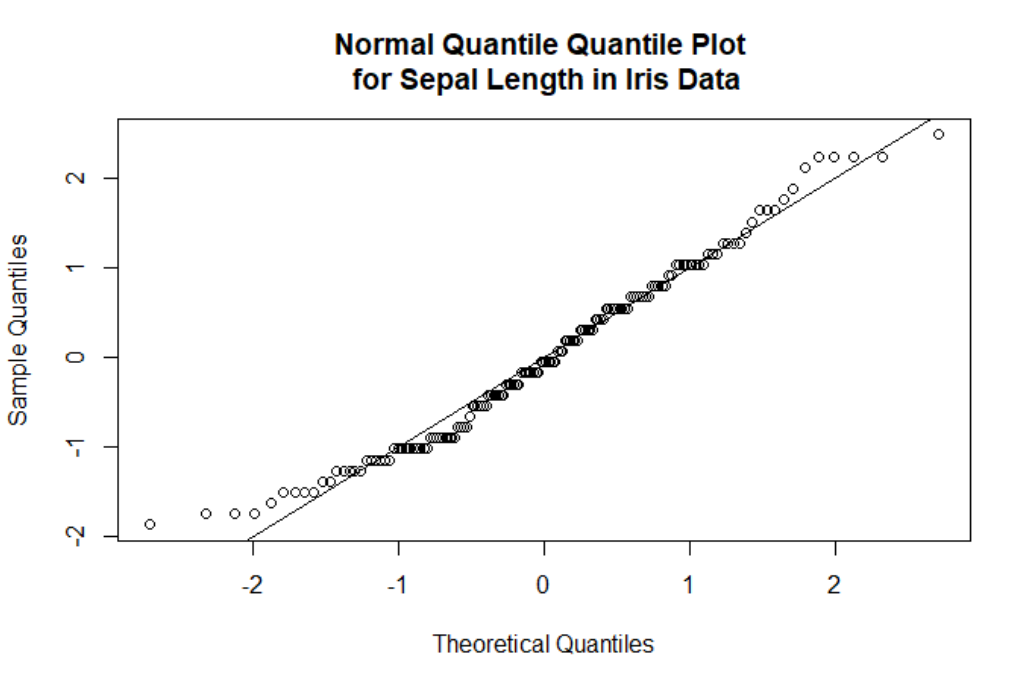
```
R> qqnorm(iris$Petal.Length);
```

UT Southwestern Medical Center

QQ Plots for assessing distributional fit

- Example for normal variate:

```
R> qqnorm(scale(iris$Sepal.Length),main="Normal Quantile Quantile Plot \n for Sepal Length in Iris Data");
R> abline(0,1);
```



- Be careful that axis agree if adding straight comparison line with abline(0,1); - Notice the scale() function.
- The further from the line the points are the further the observed quantiles of the datasets diverge from the theoretical normal quantiles with (qqnorm) or empirical quantiles with (qqplot).

- This plot for instance indicates that the normal distribution is a decent fit for Sepal.Length

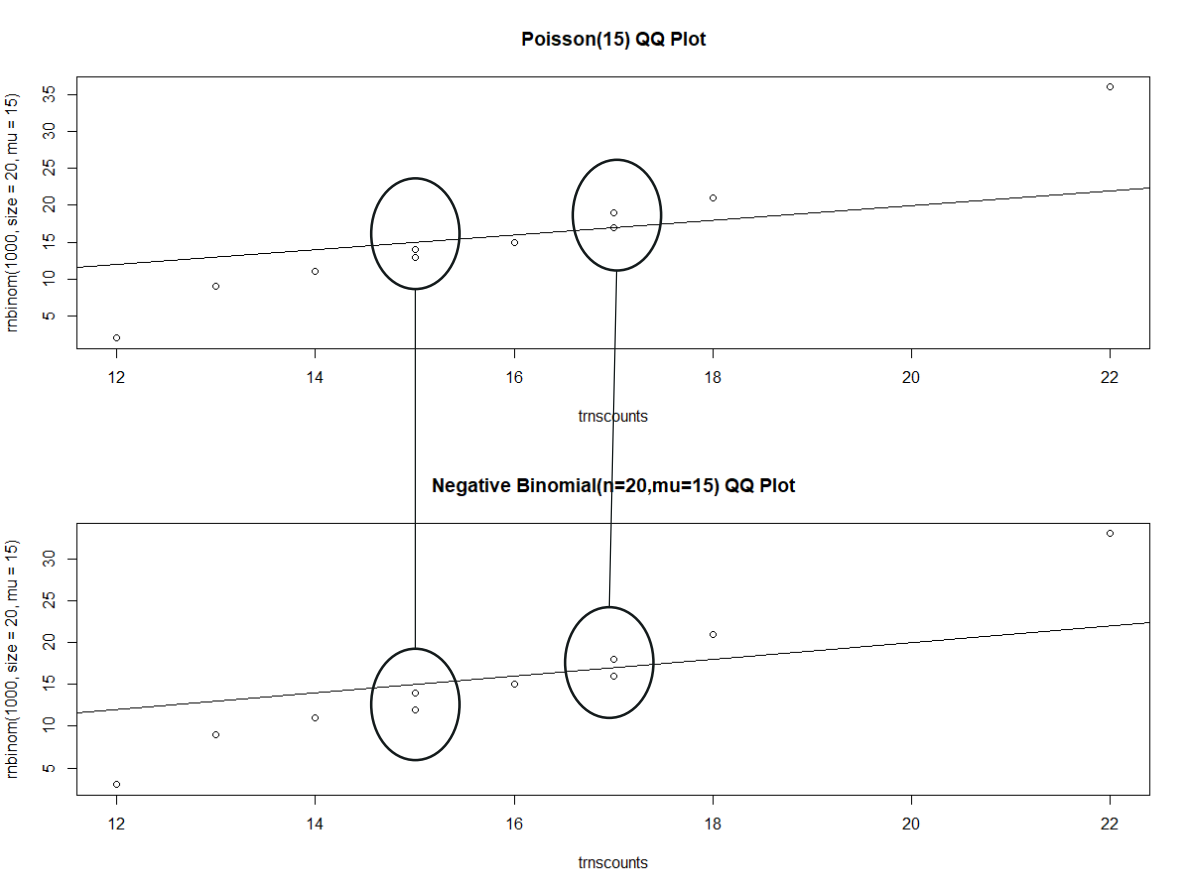
UT Southwestern
Medical Center

QQ Plots for assessing distributional fit (Practice)

- Suppose we want to graphically assess whether negative binomial or Poisson is a better fit for our data.

```
R> trnscounts <- c(18,13,17,14,15,16,15,17,22,12);
R> qqplot(trnscounts,rpois(1000,15), main="Poisson(15) QQ Plot");
R> qqplot(trnscounts,rnbinom(1000,mu=15);
R> par(mfrow=(c(2,1)));
R> qqplot(trnscounts,rnbinom(1000,size=20,mu=15), main="Poisson(15) QQ Plot");
R> abline(0,1);
R> qqplot(trnscounts,rnbinom(1000,size=20,mu=15), main="Negative Binomial(n=20,mu=15) QQ Plot");
R> abline(0,1);
R> par(mfrow=(c(1,1)));
```

QQ Plots for assessing distributional fit (Practice)

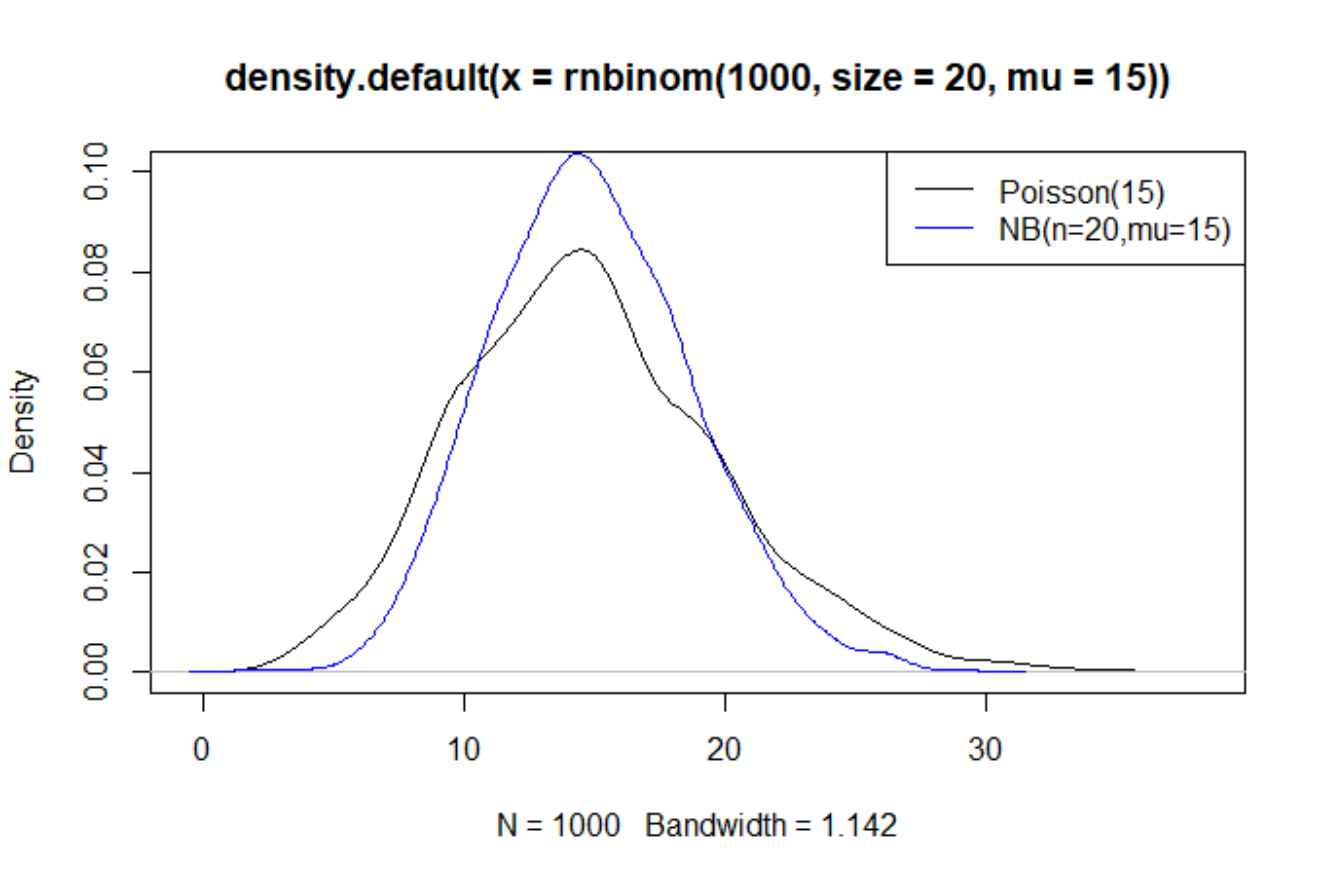


- Model fits nearly identical, but Poisson(15) is slightly closer, and indeed this is what the set was generated from.
- We might use the Poisson distribution for modeling in this case.
- Sometimes when there is a large amount of data we can visualize distribution distances using histograms too.

UT Southwestern Medical Center

Histogram-Like Plots (KDE) for assessing distributional fit (Practice)

```
R> par(mfrow=c(1,1));
R> plot(density(rnbinom(1000,size=20,mu=15)),ylim=c(0,0.1));
R> lines(density(rpois(1000,15)),col='blue');
R> legend('topright',legend=c('Poisson(15)','NB(n=20,mu=15)'),col=c('black','blue'),lty=1);
```



- Here are the densities of these two distributions (1000 values simulated from each) for example.
- We can see for instance that the quantiles in the middle will be more compressed for the negative binomial in this case than the poisson.
- The Poisson is less kurtic in this case.

UT Southwestern Medical Center

Goodness of Fit Tests for statistical distributions in R

- In statistics, there is the concept of numerically assessing how close a sample of values adheres to the probabilities defined in a particular distribution (when choosing one from among multiple models for instance, as we might be doing here).
- There are a few tests for assessing the departure from normality of a sample of data:
  - Shapiro-Wilk test [stats] shapiro.test()
  - Anderson-Darling test [nortest] ad.test()
  - Cramer-von Mises test [twosamples] cvm_stat() – compares two ecdfs – so provide two samples.
    - -> Actually a few of these are more general, but usually used for testing normality specifically
  - KS-test Cramer-von Mises generalizes this.

- **CAVEAT!!! It is really not good practice to base the conduction of the hypothesis test which assumes normality (or other distribution) on the outcome of one of these statistical hypothesis tests at a specific level. Doing so causes the reflected test p-values to become artificially biased.**

Transforming Data

- If we are using a Likelihood model for the data in the Generalized Linear Model (GLM), and we find that the outcome variable is not well fit by a normal distribution (visually, or using a statistical test), and want to transform it we can apply any transform we would like to the variable prior to regression.

- taking the log of the data, or applying some functional transform will sometimes produce normal data.
  - These are simplistic in R, simply define a new variable based on the old and examine it.

- Most of the time we take logs/functional transforms of independent variables in the context of attempting to fit the linearity assumption.
  - There are some models which relax this.

- One particular transform of use is the box cox transform

$$y(\lambda_1, \lambda_2) = \begin{cases} \dfrac{(y+\lambda_2)^{\lambda_1} - 1}{\lambda_1} & \lambda_1 \neq 0 \\ \log(y+\lambda_2) & \lambda_1 = 0 \end{cases}$$

- In practice lambdas can be selected by applying the `boxcox()` function in `R` to a `glm` object.

UT Southwestern Medical Center

Available Model Distributions and Links in R GLM

- binomial – **For Logistic Regression (probability of occurrence)**
  - **'logit' -** $g(x) = log\left(\frac{x}{1-x}\right)$
  - **'probit' -** $g(x) = \Phi^{-1}(\mathbf{x})$
  - **'cauchit' -** $g(x) = \Phi_C^{-1}(x)$
  - **'log' -** $g(x) = log(x)$
  - **'cloglog'-** $g(x) = log(-log(1-x))$
- gaussian – **As in standard MVLR from part 1**
  - **'identity' -** $g(x) = x$
  - **'log' –** $g(x) = log(x)$
  - **'inverse' -** $g(x) = \frac{1}{x}$
- Gamma – **A flexible distribution for fitting a variety of data types**
  - **'identity' -** $g(x) = x$
  - **'log' –** $g(x) = log(x)$
  - **'inverse' -** $g(x) = \frac{1}{x}$

UT Southwestern Medical Center

Available Model Distributions and Links in R GLM (2)

- inverse.gaussian – **Sometimes a good fit for ratio data**
  - **'identity' -** $g(x) = x$
  - **'log' –** $g(x) = log(x)$
  - **'inverse' -** $g(x) = \frac{1}{x}$
  - **'1/mu^2/'** – $g(x) = \frac{1}{x^2}$
- poisson – **Usually used for count regression**
  - **'identity' -** $g(x) = x$
  - **'log' –** $g(x) = log(x)$
  - **'sqrt' -** $g(x) = \sqrt{x}$
- quasi - **gaussian like variance-mean function (can estimate without formal binomial distribution assumptions)**
- quasibinomial – **binomial like variance-mean function (can estimate without formal binomial distribution assumptions)**
- quasipoisson - **poisson like variance-mean function (can estimate without formal binomial distribution assumptions)**

Not Actual Likelihoods.

  - **'identity' -** $g(x) = x$
  - **'log' –** $g(x) = log(x)$
  - **'inverse' -** $g(x) = \frac{1}{x}$
  - **'1/mu^2/'** – $g(x) = \frac{1}{x^2}$
  - **'sqrt' -** $g(x) = \sqrt{x}$
  - **'logit' -** $g(x) = log\left(\frac{x}{1-x}\right)$
  - **'probit' -** $g(x) = \Phi^{-1}(\mathrm{x})$
  - **'cloglog'-** $g(x) = log(-log(1-x))$

UT Southwestern Medical Center

GLM Example: Logistic Regression

- **Using the Iris Data, say we want to fit the following model:**

$$y_i = \begin{cases} 1 & \text{If Iris } i \text{ petal length greater than median} \\ 0 & \text{otherwise} \end{cases}$$

$$Y_i \sim Bern(\pi)$$

$$\pi = P(y_i = 1) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

UT Southwestern
Medical Center

GLM Example: Logistic Regression

- **Using the Iris Data, say we want to fit the following model:**

$$y_i = \begin{cases} 1 & \text{If Iris } i \text{ petal length greater than median} \\ 0 & \text{otherwise} \end{cases}$$

$$Y_i \sim Bern(\pi)$$

$$\pi = P(y_i = 1) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

This quantity is known as the odds, it is the probability of an event, divided by the probability of that event not occurring.

This is problematic (identity link, because we could produce values beyond 1 and below zero, so we have to use the logit link function, hence logistic regression).

$$\log\left(\frac{\pi}{1-\pi}\right) = \log\left(\frac{P(Y_i = 1)}{1 - P(Y_i = 1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

- The logit function will work as a link in this case, because when we take the log we can get values between -∞ and ∞, and dividing by the complement causes the range to become linear.
- We can fit this model in R by first creating an indicator variable for the outcome of interest (when the iris is a setosa)

UT Southwestern Medical Center

GLM Example: Logistic Regression

```
R> iris$LP <- (iris$Petal.Length > median(iris$Petal.Length))
R> logistic.form1 <- formula(LP ~ Petal.Width+Sepal.Length+Sepal.Width);
R> logistic.mod1 <- glm(logistic.form1,data=iris,family=binomial(link='logit'));
R> summary(logistic.mod1);
```

```
Call:
glm(formula = logistic.form1, family = binomial(link = "logit"),
    data = iris)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0239  -0.0001   0.0000   0.0265   2.3912

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    -31.34       9.04   -3.47  0.00053 ***
Petal.Width     14.34       4.55    3.15  0.00161 **
Sepal.Length     2.99       1.21    2.47  0.01368 *
Sepal.Width     -1.98       2.31   -0.85  0.39350
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 207.944  on 149  degrees of freedom
Residual deviance:  31.074  on 146  degrees of freedom
AIC: 39.07

Number of Fisher Scoring iterations: 9
```

Remember we have taken the log odds of the probability, so these values represent the coefficients of the log odds of having a large petal length. Hence at this level we can interpret them as:

"For each increase in centimeters (units) of Petal Width, the expected increase in the log odds of having a Petal Length in the highest 50% increases by 14.34"

most of the time, the odds are interpreted by exponentitating these coefficients.

UT Southwestern Medical Center

GLM Example: Logistic Regression

```
R> exp(coef(logistic.mod1))
```

```
(Intercept)  Petal.Width Sepal.Length  Sepal.Width
   2.44e-14     1.69e+06     2.00e+01     1.39e-01
```

For each centimeter increase in Petal Width the odds of the iris being in the top half of petal length increase by a factor of 1,690,000 – a lot. It Is harder to interpret coefficients on values as directly,

it is usually easier to do in terms of the odds.

Creating general linear Models in R with glm(): Exercise 2

- Use R, and what we have learned to help you answer these questions about the warpbreaks dataset in a .txt file:
  1. Determine whether the probability of having a number of breaks in the top 50 % (greater than the median) is statistically significantly associated with either the type of wool used, or the tension the wool is under.
     1. What is the change in odds (remember to exponentiate) of having a higher number of warp breaks than the median associated with high tension?
  2. Instead of using the binomial Likelihood, and logit link function, use the poisson Likelihood and the identity link function, and use the number of breaks instead of the binary indicator you created in 1.
     1. What is the value in the estimate column for tension [Medium]?

        This value is interpreted as the expected change in count of the dependent variable by a change from the reference level of tension [low] to medium.
  3. Create a qqplot for all of the breaks recorded in the warpbreaks dataset and a poisson distribution with mean equal to that mean of the variable, and place a line on the plot, do the same for the normal distribution with mean and sd the same as the sample from R, which appears to be a closer fit.

     We might be tempted to determine the normal is a closer fit, but this is hazardous the data is not continuous it is count data, but it is the sum of different poissons which makes it look normal.

UT Southwestern Medical Center

Break

UT Southwestern
Medical Center

Part III (Final):
Using inbuilt R Functions, and `glmnet` for Variable Selection
+
htmlTable() for Displaying Results

UT Southwestern
Medical Center

Stepwise Variable Selection in R (using AIC)

- **R** uses the AIC by default when fitting models, you can tell it to step forward, step backwards, or in both directions.
- R will progress from one model towards another, specified by the user. For example consider again the iris data.
- We use the step() function in R to do forwards, backwards and stepwise selection.
  - In forwards we go from the null model (with only the intercept) towards the model with all terms in it,
  - In Backwards we go from the full model with all terms towards the null model
  - In Both we move in both directions.

```
R> iris.null.formula <- formula(Petal.Length ~ 1);
R> iris.full.formula <- formula(Petal.Length ~ Petal.Width*Sepal.Length*Sepal.Width);

R> iris.null.model <- glm(iris.null.formula,data=iris);
R> iris.full.model <- glm(iris.full.formula,data=iris);

R> step(iris.null.model,direction='forward',scope=iris.full.formula) -> iris.forward.step;
R> step(iris.full.model,direction='backward',scope=iris.full.formula) -> iris.backward.step;
R> step(iris.null.model,direction='both',scope=iris.full.formula) -> iris.both.step;
```

UT Southwestern Medical Center

Stepwise Variable Selection in R (using AIC)

```
R> iris.forward.step$anova
R> iris.backward.step$anova
R> iris.both.step$anova
```

- *Getting the anova object from these shows the anova rows for each model by the step used,*
- *we can see the forward and stepwise selected the same model, but that backwards refused to drop any terms from the full model,*
- *which means these methods arrived at a different optimal variable set.*

```
> iris.forward.step$anova
                        Step Df Deviance Resid. Df Resid. Dev   AIC
1                           NA       NA       149      464.3 599.2
2            + Petal.Width -1  430.481       148       33.8 208.4
3           + Sepal.Length -1    9.943       147       23.9 158.2
4            + Sepal.Width -1    9.049       146       14.9  88.8
5 + Petal.Width:Sepal.Length -1    0.377       145       14.5  87.0
6  + Petal.Width:Sepal.Width -1    0.242       144       14.2  86.4
> iris.backward.step$anova
  Step Df Deviance Resid. Df Resid. Dev  AIC
1      NA       NA       142       13.3 80.8
> iris.both.step$anova
                        Step Df Deviance Resid. Df Resid. Dev   AIC
1                           NA       NA       149      464.3 599.2
2            + Petal.Width -1  430.481       148       33.8 208.4
3           + Sepal.Length -1    9.943       147       23.9 158.2
4            + Sepal.Width -1    9.049       146       14.9  88.8
5 + Petal.Width:Sepal.Length -1    0.377       145       14.5  87.0
6  + Petal.Width:Sepal.Width -1    0.242       144       14.2  86.4
```

UT Southwestern Medical Center

Variable Selection using Elastic Net (generalization of LASSO/Ridge)

- We penalize large coefficients, by adding a constraint to the likelihood,
  - If we penalize the absolute difference with zero this is considered **LASSO** regression
  - If we penalize the squared difference this is **Ridge**
  - If we choose fifty-fifty split between absolute and squared difference from zero (alpha = 0.5 in R) this is **Elastic Net**

```
R> library(glmnet);

R> cbind(iris$Petal.Width,iris$Sepal.Length,iris$Sepal.Width,rnorm(nrow(iris)),rnorm(nrow(iris))) -> dsnmatrix.X;
R> rsp.Y <- iris$Petal.Length;

R> cv.glmnet(dsnmatrix.X,rsp.Y,alpha=1) -> glmnet.iris.cv.lasso;
R> cv.glmnet(dsnmatrix.X,rsp.Y,alpha=0) -> glmnet.iris.cv.ridge;
R> cv.glmnet(dsnmatrix.X,rsp.Y,alpha=0.5) -> glmnet.iris.cv.enet;

R> glmnet(dsnmatrix.X,rsp.Y,lambda=glmnet.iris.cv.lasso$lambda.1se) -> glmnet.iris.cv.lasso.mod;
R> glmnet(dsnmatrix.X,rsp.Y,lambda=glmnet.iris.cv.ridge$lambda.1se) -> glmnet.iris.cv.ridge.mod;
R> glmnet(dsnmatrix.X,rsp.Y,lambda=glmnet.iris.cv.enet$lambda.1se) -> glmnet.iris.cv.enet.mod;
```

Variable Selection using Elastic Net (generalization of LASSO/Ridge)

- We can see the estimated coefficients for each model (and those shrunken to zero) by using the following, which produces and displays a nice html table summary.

```
R> cbind(coef(glmnet.iris.cv.lasso.mod),coef(glmnet.iris.cv.ridge.mod),
      coef(glmnet.iris.cv.enet.mod)) -> coefficients.shrinkage;

R> library(htmlTable);

R> colnames(coefficients.shrinkage) <- c("Lasso", "Ridge", "Elastic Net")
R> rownames(coefficients.shrinkage) <- c("Intercept","Petal Width", "Sepal Length", "Sepal
Width", "Random")

R> format(coefficients.shrinkage,digits=4) -> coefficients.shrinkage;
R> htmlTable(coefficients.shrinkage)
```

	Lasso	Ridge	Elastic Net
Intercept	-0.2271	-0.1220	-0.1938
Petal Width	1.4681	1.5303	1.4877
Sepal Length	0.6430	0.3763	0.5580
Sepal Width	-0.5014	-0.0504	-0.3575
Random	.	.	.

Selected coefficients are very similar, and the random data is shrunken to zero in all cases.

UT Southwestern Medical Center

Producing a basic htmlTable with htmlTable package.

- You can produce tables from any dataframe, for instance to create a custom glm output dataframe, lets select the coefficients and their confidence intervals from an iris model, put them in a dataframe and display them to the user.

```
R> library(htmlTable);

R> iris.table.formula <- formula(Petal.Length~Petal.Width+Sepal.Length+Sepal.Width);
R> iris.table.model <- glm(iris.table.formula,data=iris);

R> cbind(coef(iris.table.model),confint(iris.table.model)) -> table.data;

R> format(table.data,digits=3) -> table.data;

R> colnames(table.data) <- c("est.","L CI", "U CI");
R> htmlTable(table.data);
```

	est.	L CI	U CI
(Intercept)	-0.263	-0.846	0.320
Petal.Width	1.447	1.314	1.579
Sepal.Length	0.729	0.615	0.843
Sepal.Width	-0.646	-0.780	-0.512

https://rpubs.com/tbiggs/RhtmlTable
For More Examples ^^

UT Southwestern Medical Center

Creating Table Results for Swiss Model: Exercise 3

- Use R, and what we have learned to help you answer these questions about the warpbreaks dataset in a .txt file:
    1. Please generate a table for the LASSO and Ridge coefficients for the warpbreaks dataset with breaks regressed as a Poisson count linked linearly by the identity function to a linear combination of the two categorical variables tension and wool.
        1. **Which Model Variable selection technique (LASSO, Ridge, or Elastic Net) drops the wool variable from the model?**

UT Southwestern Medical Center

Additional Resources for Learning Statistics & The R Programming Language.

Statistical Analysis of Genomics Data Course Penn State (STAT 555) [Includes RNA-Seq Data]:

- https://online.stat.psu.edu/stat555/

Penn State University Introduction to R (STAT 484) Textbook, Resources, and Course Notes:

- Textbook: https://online.stat.psu.edu/statprogram/sites/statprogram/files/EssentialR.pdf
- Resources: https://online.stat.psu.edu/statprogram/sites/statprogram/files/EssentialRfiles.zip
- Course Notes: https://online.stat.psu.edu/stat484/

Penn State University Intermediate R (STAT 485) Course Notes:

- Course Notes: https://online.stat.psu.edu/stat485/

Some Basic R Exercises for Practice from geeksforgeeks.org:

- https://www.geeksforgeeks.org/r-programming-exercises-practice-questions-and-solutions/

Some Basic R Exercises for Practicing using the DAAG package (Associated with the textbook "Data Analysis and Graphics Using R"):

- https://maths-people.anu.edu.au/~johnm/courses/r/exercises/pdf/r-exercises.pdf

Some Basic Mathematical Exercises that will strengthen skills in any language including R:

- **https://projecteuler.net/**

Some Good Books for Helping to Learn R:

https://www.oreilly.com/library/view/r-for-data/9781491910382/
https://www.oreilly.com/library/view/learning-r/9781449357160/

UT Southwestern Medical Center

Thank You For Your Attention & Interest In the R Programming Language

Questions/Comments/Responses?

UT Southwestern
Medical Center

Bonus: using the Survival Package for Cox Proportional Hazards Regression

```
R> library(survival)
R> formula(Surv(lung$time,lung$status)~age+as.factor(sex)+meal.cal+wt.loss) -> cox.formula;
R> coxph(cox.formula,data=lung) -> cox.model;
R> summary(cox.model)
```

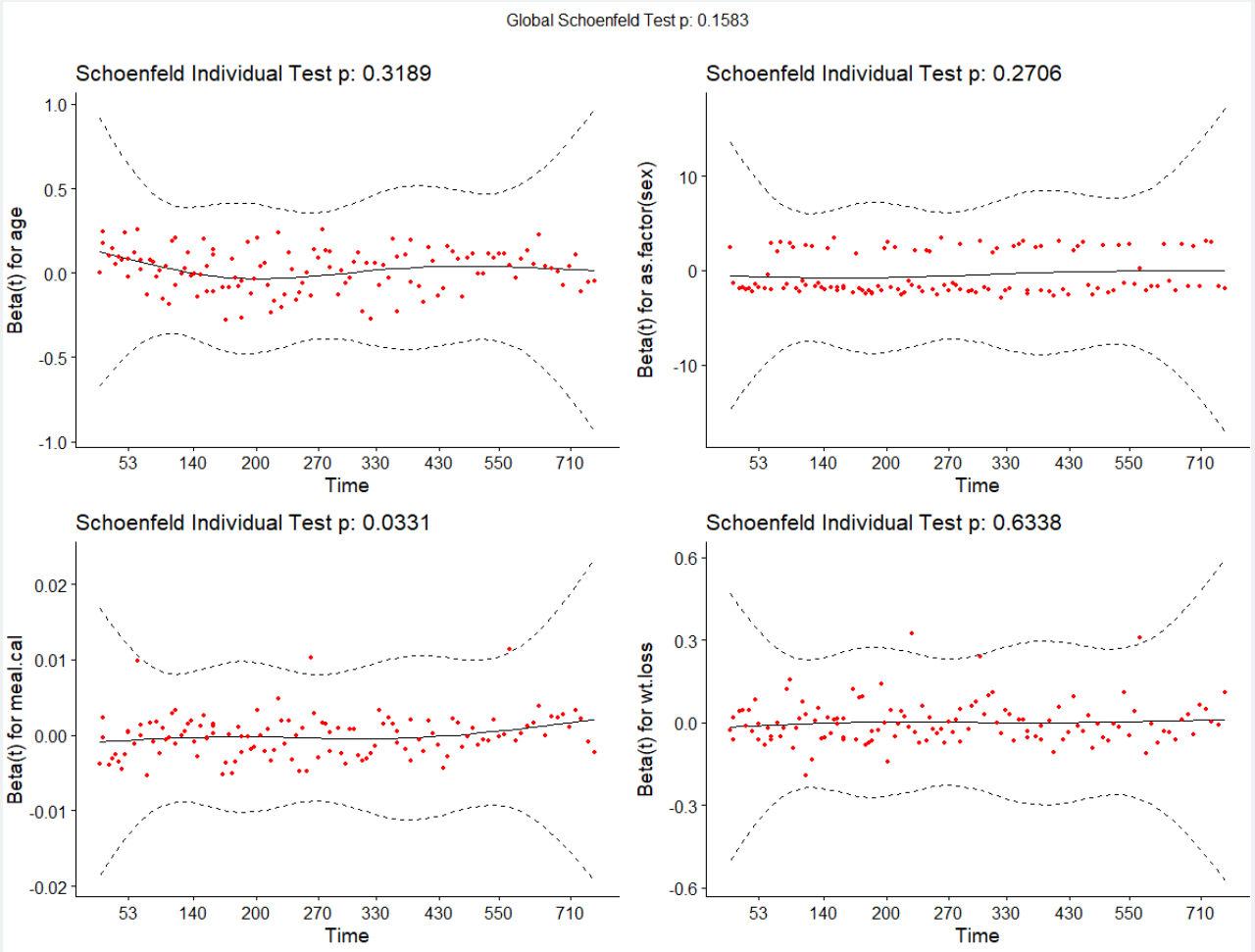
```
Call:
coxph(formula = cox.formula, data = lung)

  n= 171, number of events= 124
   (57 observations deleted due to missingness)

                    coef exp(coef)  se(coef)     z Pr(>|z|)
age              0.017826  1.017986  0.011050  1.61    0.107
as.factor(sex)2 -0.463821  0.628876  0.197542 -2.35    0.019 *
meal.cal        -0.000120  0.999880  0.000247 -0.49    0.627
wt.loss         -0.000543  0.999458  0.006778 -0.08    0.936
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                exp(coef) exp(-coef) lower .95 upper .95
age                 1.018      0.982     0.996     1.040
as.factor(sex)2     0.629      1.590     0.427     0.926
meal.cal            1.000      1.000     0.999     1.000
wt.loss             0.999      1.001     0.986     1.013

Concordance= 0.608  (se = 0.031 )
Likelihood ratio test= 10.1  on 4 df,   p=0.04
Wald test            = 9.63  on 4 df,   p=0.05
Score (logrank) test = 9.78  on 4 df,   p=0.04
```

- The Syntax for Cox Proportional hazards is essentially the same as the other models, however, the outcome variable must first be a survival object (using the Surv function, in the same way we used the as.factor, and as.ordered functions earlier)
- "The hazard (probability of dying in the current instant given survival up to instant) of death from lung cancer is statistically significantly associated with a -0.46381 factor decrease for being female."

UT Southwestern Medical Center

Bonus: Assesing Cox Proportional Hazards Assumptions

```
R> library(survminer)
R> plot(cox.zph(cox.model))
R> ggcoxzph(cox.zph(cox.model))
```



- To determine whether the actual variables do produce a set of values have hazards which are parallel (proportional) across their entire range, we can investigate the Schoenfeld Individual Test results.
- Schoenfeld Individual Test Positive results (pvalue below 0.05) indicate violations of proportional hazards, ie drift from horizontal line in the residuals plots.

UT Southwestern Medical Center