Additional Topics

- Residual analysis: Regression and logistic regression
- Weighted least squares
- Mixed models: Regression and logistic regression
- Multinomial models
- Shrinkage and regularization methods
- Dealing with multiple testing

Residual analysis: Linear regression

- Standardized/studentized residuals:

$$r_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{1 - h_i}}$$

Here h_i is the leverage associated to the i-th observation and defined as

$$h_i = H_{i,i}, \quad H = X'(X'X)^{-1}X'$$

With H the “hat” matrix. The leverage is a useful diagnostic tool to determine extreme values and influential observations. Large leverages reduce the variance of $\hat{\epsilon}_i$, forcing the fit to be “close” to y_i. Rule of thumb: Leverages above $2p/n$ indicate potential influential observations/outliers

To obtain the studentized residuals in R:

```
>a=summary(Model)
>a_inf=influence(Model)
>stud=residuals(Model)/(a$sig*sqrt(1-a_inf$hat))
```

This is equivalent to:

```
>stud=rstandard(Model)
```

- Jacknife residuals (externally studentized or cross-validated residuals):

$$t_i = \frac{y_i - \hat{y}_{(i)}}{\hat{\sigma}_{(i)}(1 + x_i'(X_{(i)}'X_{(i)}))^{-1}x_i)^{1/2}}$$

Here the (i) notation refers to estimates obtained from a model that has the same predictors as the original model but excludes the i-th observation. They can also be written as:

$$t_i = r_i \left(\frac{n - p - 1}{n - p - r_i^2} \right)^{1/2}$$

To obtain these residuals in R:

```
>jack=rstudent(Model)
```

Residual analysis: Logistic regression

- Pearson residuals:

$$\frac{y_i - \hat{\theta}_i}{\sqrt{\hat{\theta}_i(1 - \hat{\theta}_i)/n_i}}$$

- Standardized residuals (also called "studentized residuals", "studentized Pearson"…):

$$r_i = \frac{y_i - \hat{\theta}_i}{\sqrt{1 - h_i}}$$

- Deviance residuals and standardized deviance residuals

- Jacknife residuals also available

To obtain the logistic regression residuals in R you can use the function `residuals` and `rstandard`:

```
>residuals(Model,type="pearson")
>residuals(Model,type="deviance")
```

For standardized versions of the Pearson and Deviance residuals you can use the function `rstandard`

Jacknife versions of the residuals are available using the function `rstudent`

Generalized Least Squares

Linear regression models assume $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I)$ or equivalently, $\epsilon_i \sim^{iid} N(0, \sigma^2)$ for all i. This assumption does not always hold.

We can instead assume that $\boldsymbol{\epsilon} \sim N(0, \sigma^2 \Sigma)$ with Σ diagonal (i.e., errors uncorrelated but unequal variances). In this situation we can use generalized least squares which leads to:

$$\hat{\boldsymbol{\beta}}_{GLS} = (\boldsymbol{X}'\Sigma^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\Sigma^{-1}\boldsymbol{y}$$

$$\hat{\sigma}^2_{GLS} = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{GLS})'\Sigma^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{GLS})/(n-p)$$

This can be done in R by specifying the `weights` in the `lm` function:

$$\Sigma = \mathbf{diag}(1/w_1, \ldots, 1/w_n)$$

- Errors proportional to a predictor: $w_i = x_{j,i}^{-1}$, for example:

```
>model=lm(y ~x1 + x2 + x3, weights=1/x1)
```

- When y_i are averages of n_i observations $var(\epsilon_i) = \sigma^2/n_i$, and so $w_i = n_i$

Note that: When using weights the residuals must be modified too so use $\sqrt{w_i}\hat{\epsilon}_t$ for diagnostics

Mixed Models: Fixed and random effects

- **Linear models:** The function `lmer` from the `lme4` R library allows us to fit mixed effects models.

Lets revisit the the exam scores example:

- Fixed effects models:

$$y_{i,j} = \mu + \underbrace{\alpha_i}_{\text{EXAM EFFECT (FIXED)}} + \underbrace{\beta_j}_{\text{STUDENT EFFECT (FIXED)}} + \epsilon_{i,j}, \;\; \epsilon_{i,j} \sim N(0,\sigma^2)$$

```
>Model_Fixed=lm(score ~ exam + student, data=scor.long)
```

```
>Model_Fixed=lm(score ~ exam + student, data=scor.long)
>summary(Model_Fixed)
.
.
.
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   67.927      4.792  14.174  < 2e-16 ***
examvec       11.636      1.580   7.365 1.30e-12 ***
examalg       11.648      1.580   7.372 1.24e-12 ***
examana        7.727      1.580   4.891 1.54e-06 ***
examsta        3.352      1.580   2.122 0.034570 *
student2      -0.400      6.628  -0.060 0.951915
student3      -1.600      6.628  -0.241 0.809401
.
.
.
Residual standard error: 10.48 on 348 degrees of freedom
Multiple R-squared:  0.6389, Adjusted R-squared:  0.5445
F-statistic: 6.766 on 91 and 348 DF,  p-value: < 2.2e-16
```

```
> anova(Model_Fixed)
Analysis of Variance Table

Response: score
           Df Sum Sq Mean Sq F value    Pr(>F)    
exam        4   9315 2328.72  21.201 1.163e-15 ***
student    87  58313  670.26   6.102 < 2.2e-16 ***
Residuals 348  38225  109.84                      
```

- Mixed effects: (Fixed: Exam) + (Random: students)

$$y_{i,j} = \mu + \alpha_i + \beta_j + \epsilon_{i,j}, \ \ \epsilon_{i,j} \sim N(0, \sigma^2)$$

$$\beta_j \sim N(0, \tau^2)$$

```
>Model_Mixed=lmer(score ~ exam + (1 |student), data=scor.long)
>summary(Model_Mixed)
.
.
.
REML criterion at convergence: 3458.3
.
.
.
Random effects:
 Groups   Name        Variance Std.Dev.
 student  (Intercept) 112.1    10.59
 Residual             109.8    10.48
Number of obs: 440, groups:  student, 88

Fixed effects:
            Estimate Std. Error t value
(Intercept)   38.955      1.588  24.530
examvec       11.636      1.580   7.365
examalg       11.648      1.580   7.372
examana        7.727      1.580   4.891
examsta        3.352      1.580   2.122
```

```
>anova(Model_Mixed)
Analysis of Variance Table
     npar Sum Sq Mean Sq F value
exam    4 9314.9  2328.7  21.201

>ranef(Model_Mixed)
$student
    (Intercept)
1   24.22469559
2   23.89024732
3   22.88690254
.
.
.
>coef(Model_Mixed)
$student
   (Intercept)  examvec  examalg  examana  examsta
1     63.17924 11.63636 11.64773 7.727273 3.352273
2     62.84479 11.63636 11.64773 7.727273 3.352273
3     61.84145 11.63636 11.64773 7.727273 3.352273
```

- Note that fixed effects in mixed effects model correspond to fixed effects in the following model:

```
>Model_Fixed_1=lm(score ~ exam, data=scor.long)
>summary(Model_Fixed_1)
.
.
.
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   38.955      1.588  24.530  < 2e-16 ***
examvec       11.636      2.246   5.181 3.38e-07 ***
examalg       11.648      2.246   5.186 3.29e-07 ***
examana        7.727      2.246   3.441 0.000636 ***
examsta        3.352      2.246   1.493 0.136251
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.9 on 435 degrees of freedom
Multiple R-squared:  0.088,  Adjusted R-squared:  0.07961
F-statistic: 10.49 on 4 and 435 DF,  p-value: 4.009e-08
```