STA 303/1002-Methods of Data Analysis II

Sections L0101& L0201, Winter 2018

Shivon Sue-Chee

UNIVERSITY OF TORONTO

March 6, 2018

Case Study V Example: Mating Success of Elephants

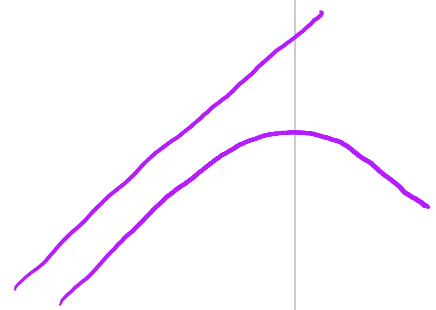
- Data: $n = 41$ male elephants, followed for 8 years

AGE	MATINGS
27	0
28	1
28	1
28	1
28	3
...	
47	7
48	2
52	9

x
  - Predictor: Age- age at beginning (27-52yrs)

y
  - Outcome: Matings- # of successful matings (from 0)

- Question: What is the relationship between mating success and age? Do males have diminished success after reaching some optimal age?



Case Study V Model: Why Poisson?

- Why not linear regression?
  - Outcome is counts and small numbers
  - Won't have a normal distribution conditional on age
- Why not logistic regression?
  - Not a binary outcome
  - Not a binomial outcome since not a fixed number of trials

π_i

m_i

- Poisson distribution- useful for counts of rare events

Case Study V Model: Why Poisson?

- Other examples:
  1. Relationship between family's number of trips to grocery store during a particular week and the family's income, number of children and distance from store
  2. Relation between the number of hospitalizations of a member of a health organization during the past year and the member's age, income and previous health status
  3. Is the count of Del Norte salamanders in northwest California related to canopy cover and forest age?

Case Study V: Poisson model

If $Y \sim \text{Poisson}(\mu)$, then

- Probability mass function:

$$P(Y = y) = \frac{\mu^y e^{-\mu}}{y!}, \; y = 0, 1, \ldots$$

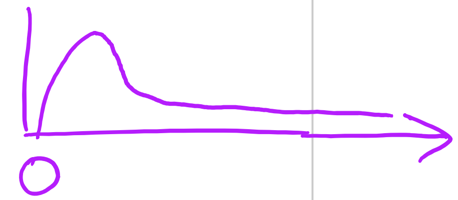
- Expectation and Variance:

$$E(Y) = \mu = Var(Y)$$

$y_i \sim P(\mu_i)$

$Var(y_i) = m_i \pi_i (1 - \pi_i)$.

- Distribution tends to be right skewed, especially when the mean is small



- When mean is large, Poisson can be approximated by a Normal.
- Poisson regression model is an example of a **Generalized Linear Model**.

Poisson model: A generalized linear model

$E(\log(y)) = X\beta$

- Model $E(Y)$ as linear in the parameters,

$$g(E(Y)) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p = \mathbf{X}\beta$$

where g is the link function

$\log\left(\frac{\mu}{1-\mu}\right)$

- Poisson link function: $g(\mu) = \log(\mu)$
- Also called "log-linear" model
- Interpretation of β's: Increase x_j by one unit, holding other predictors constant, μ_j changes by a factor of $\exp(\beta_j)$

transform y
" X
" y & X
— Poisson model.

Poisson model: Estimation of model parameters

- Estimation method: Maximum likelihood estimation by IRLS algorithm
- Inference: Wald procedures and likelihood ratio tests (as in logistic regression)

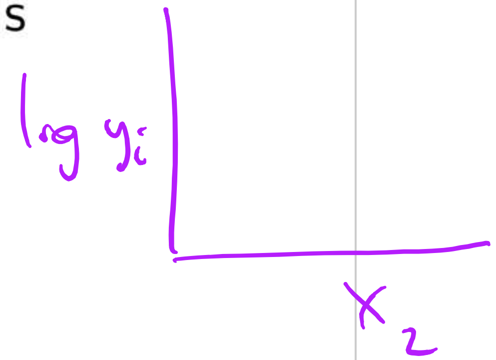
Poisson model: Checking Model Adequacy

(Similar procedures as in binomial logistic regression):

- Linear in β's: Plot $\log(y_i)$ versus x's to see if linear relationship seems appropriate. Jitter if many $y_i = 0$.
  - use $\log(y_i + k)$ if $y_i = 0$, k is a small positive value

$k=0.5$, $k=0.1$ often $0<k<1$

- Outliers: Look at residuals- Deviance and Pearson residuals
- Correct form: Use Wald and LRT tests
- Adequate fit: Use Deviance GOF test



Common problem: Variance is larger than mean
SOLUTION: Add an extra (dispersion) parameter to the model

Poisson Regression Model Log Likelihood

$y_i \sim P(\mu_i)$

$P(Y_i = y_i) = \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}$

$y_i = 0, 1, 2, \ldots$

- Show that the Log-likelihood is

$$\log \mathcal{L} = \sum_{i=1}^{n} \left\{ y_i \log(\mu_i) - \mu_i + \text{constant} \right\}$$

Likelihood (Joint probability) $L = \prod_{i=1}^{n} P(Y_i = y_i)$ if y_i are indep.

$= \prod_{i=1}^{n} \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}$

Log-likelihood, $\log L = \sum_{i=1}^{n} \left(-\mu_i + y_i \log \mu_i - \log y_i! \right)$

- What is the constant term?

$\rightarrow \sum_{i=1}^{n} -\log y_i!$

Poisson model for Case V

- Count, Y_i for an elephant of age_i follow Poisson(μ_i)
- Assume that all responses, Y_i pertain to the same unit of time or space
- Model $E(Y_i) = \mu_i$ as a linear function,

$$g(\mu) = \underline{\log(\mu_i)} = \beta_0 + \beta_1 age_{i1}, \quad i = 1, \dots, 41$$

(1)

  - where μ_i- mean # of matings for an elephant of $Age = age_i$
  - Then

non-linear reg.

$$\boxed{\mu_i = \exp\{\beta_0 + \beta_1 age_{i1}\}}$$

Linear Reg:

$\mu_i = X\beta$.

- Interpretation of β's: Increasing Age by one unit, changes μ by a factor of $\exp(\beta_1)$

Likelihoods

- The likelihood function is

$$\mathcal{L} = \prod_{i=1}^{41} \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}$$

$n = 41.$

- The log-likelihood function is

$$\sum_{i=1}^{41} \left\{ y_i \log(\mu_i) - \mu_i - \log(y_i!) \right\}$$

- Hence, $\underline{-2\log\mathcal{L}} = 2\sum_{i=1}^{41}\{\mu_i - y_i\log(\mu_i) + \log(y_i!)\}$

In R: Log-linear models

- R syntax:

$$\texttt{glm(formula, family = poisson, data)}$$

  - Can be used for any generalized linear model
  - For Poisson, use $\texttt{family = poisson}$ in glm

- Plot of $\log y$ versus x:

- Wald procedures:

  - Chi-square test statistic: $\left(\frac{\widehat{\beta}_j}{SE(\widehat{\beta}_j)}\right)^2$
  - 95% CI for β_j:

$$\widehat{\beta}_j \pm 1.96 * SE(\widehat{\beta}_j)$$

- Plot of residuals:

Case Study V: Deviance GOF test

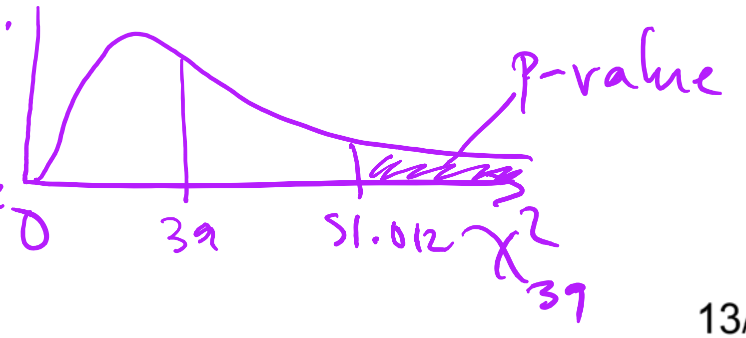
Q: Determine whether the fitted model fits as well as the saturated model.

- Hypotheses:
  - Ho: Fitted model fits as well as saturated model
  - Ha: Saturated model fits better.
    (uses indicator variables for each value of *Age*)
- Test Statistic: Deviance (In R Residual deviance) $= -2\log\left(\frac{L_M}{L_S}\right) = \boxed{51.012}$
- Distribution of TS under H_0: $\chi^2_{n-(p+1)}$, $p=1$, $n=41$ $\Rightarrow$ $\boxed{df=39}$
- p-value: ≈ 0.094

Weak evidence in support of H_0.

- Conclusion:

Fitted model is adequate

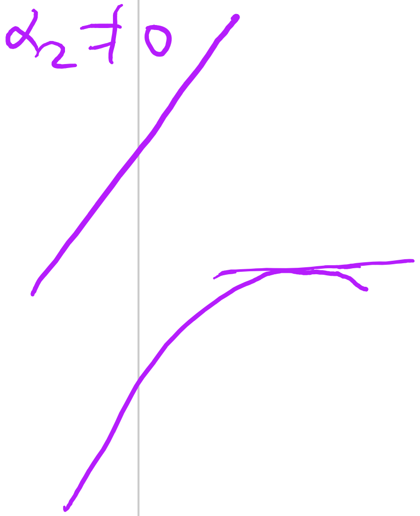


Case Study V: Wald or LRT

$H_0: \alpha_2 = 0$

$H_a: \alpha_2 \neq 0$

Q: Determine whether the mean number of successful matings tends to peak at some age and then decrease or whether the mean continues to increase with age.



- Hypotheses:
  - Ho: Reduced : $\boxed{\log \mu_i = \beta_0 + \beta_1 Age_i}$
  - Ha: Full : $\log \mu_i = \alpha_0 + \alpha_1 Age_i + \alpha_2 Age_i^2$
- Test Statistic:
- Distribution of TS under H_0:
- p-value:
- Conclusion:

Wald	LRT
$0.182 = (-0.427)^2$	$G^2 = 51.012 - 50.826 = 0.186$
χ^2_1	χ^2_1
0.669	0.666
Reduced	Reduced

Case Study V: Other model assessments

- Dispersion parameter, $\psi = 1$?:
  - Assess the situation
  - Plot Variances against Averages
  - Perform Deviance GOF test on a rich model
  - Check for outliers using Pearson or Deviance residuals

- AIC$= -2\text{Log}\mathcal{L} + 2(p+1) = -2(-76.2289) + 2(2)$
- BIC$= -2\text{Log}\mathcal{L} + (p+1)\log(N) = -2(-76.2289) + 2\log(41)$

Case Study V: Summary of findings

- Fitted model:

$$\boxed{\widehat{\log(\mu)} = -1.5820 + 0.0687 * Age}$$

- Wald test conclusion: Strong evidence that the mean # of successful matings depends on *Age* ($p < 0.0001$)

- Interpretation: For every 1-year increase in *Age*, the mean number of successful matings increased by a factor of $\exp(0.0687) = 1.071$ ($\sim 7\%$ increase).

STA303/1002 - Class 15 R Markdown

March 6, 2018

Case Study V: The Data

Get the data (from R library):

```
#load Sleuth3 R data library; see case2101
library(Sleuth3); elmasu = case2201
str(elmasu)
```

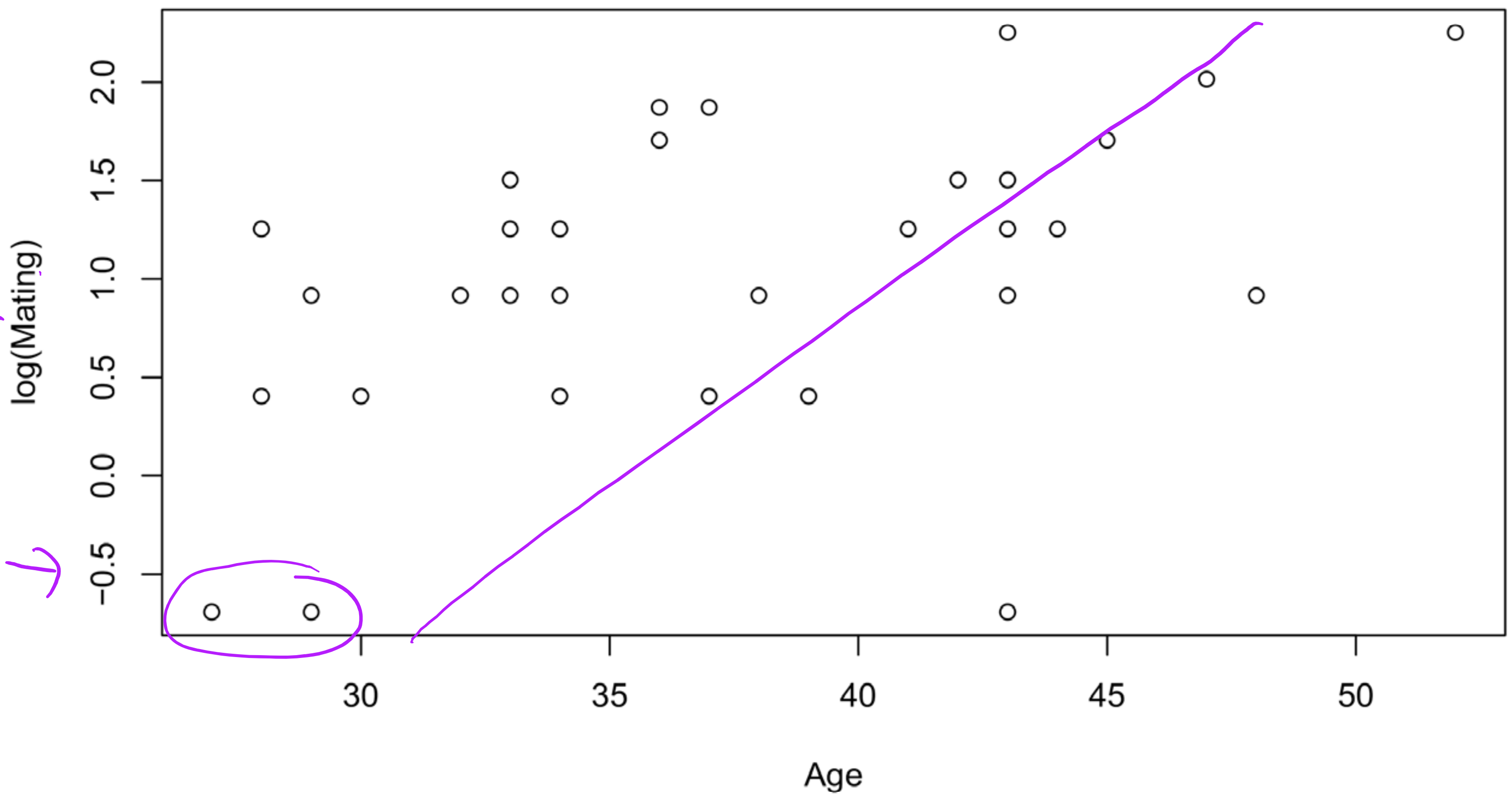
```
## 'data.frame':    41 obs. of  2 variables:
##  $ Age     : int  27 28 28 28 28 29 29 29 29 29 ...
##  $ Matings: int  0 1 1 1 3 0 0 0 2 2 ...
```

```
attach(elmasu)
```

27
28
29

Case Study V: Data Visualization

jittered $y_i = 0$



+ve linear trend.

Case Study V: Log Linear Model

```
fitllm<-glm(Matings~Age, family=poisson, data=elmasu)
summary(fitllm)
```

```
##
## Call:
## glm(formula = Matings ~ Age, family = poisson, data = elmasu)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.80798  -0.86137  -0.08629   0.60087   2.17777
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.58201    0.54462  -2.905  0.00368 **
## Age          0.06869    0.01375   4.997 5.81e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 75.372  on 40  degrees of freedom
## Residual deviance: 51.012  on 39  degrees of freedom
## AIC: 156.46
##
## Number of Fisher Scoring iterations: 5
```

$\hat{\beta}_0$

$\hat{\beta}_1$

0.0069

IRLS

Case Study V: Richer Log Linear Model

```
fitllm2<-glm(Matings~Age+I(Age^2), family=poisson, data=elmasu)
summary(fitllm2)
```

```
##
## Call:
## glm(formula = Matings ~ Age + I(Age^2), family = poisson, data = elmasu)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8470  -0.8848  -0.1122   0.6580   2.1134
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.8574060  3.0356383  -0.941    0.347
## Age          0.1359544  0.1580095   0.860    0.390
## I(Age^2)    -0.0008595  0.0020124  -0.427    0.669
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 75.372  on 40  degrees of freedom
## Residual deviance: 50.826  on 38  degrees of freedom
## AIC: 158.27
##
## Number of Fisher Scoring iterations: 5
```

Case V Fit Statistics

Red

```
AIC(fitllm)
```

```
## [1] 156.4578
```

```
BIC(fitllm)
```

BIC > AIC

```
## [1] 159.8849
```

```
AIC(fitllm2)
```

Full

```
## [1] 158.2723
```

```
BIC(fitllm2)
```

```
## [1] 163.4131
```

Case V Residuals

```
yhats<-predict.glm(fitllm, type="response") # estimated means
rres<-residuals(fitllm, type=c("response"))
pres<-residuals(fitllm, type=c("pearson"))
dres<-residuals(fitllm, type=c("deviance"))
options(digits=4)
rbind(Matings, yhats, rres)
```

y_i

$\hat{\mu}_i = \hat{y}_i$

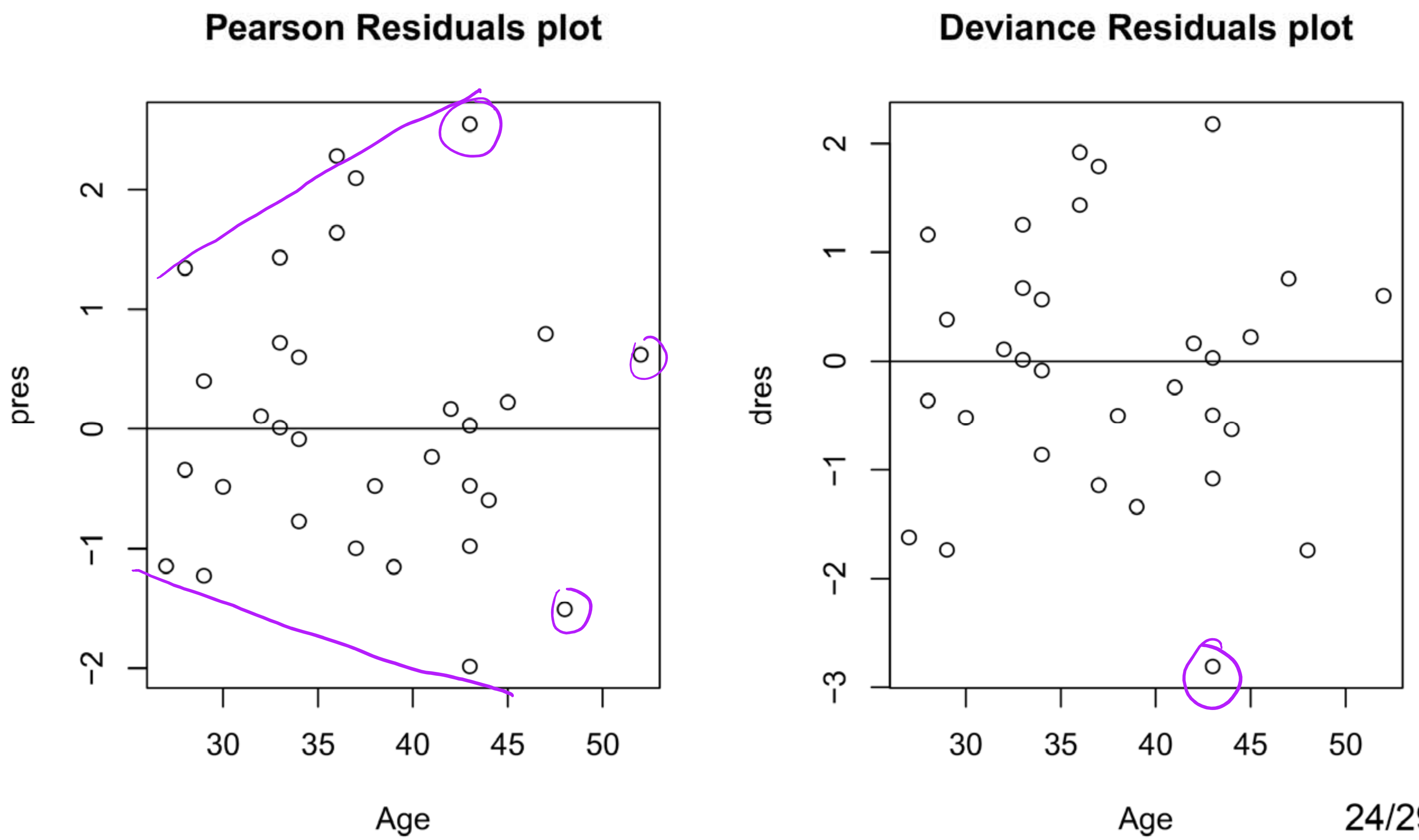
$(y_i - \hat{y}_i)$

```
##              1       2       3       4     5      6      7      8      9
## Matings  0.000  1.0000  1.0000  1.0000 3.000  0.000  0.000  0.000 2.0000
## yhats    1.314  1.4069  1.4069  1.4069 1.407  1.507  1.507  1.507 1.5069
## rres    -1.314 -0.4069 -0.4069 -0.4069 1.593 -1.507 -1.507 -1.507 0.4931
##             10     11      12     13    14    15    16    17      18
## Matings 2.0000 2.0000  1.0000 2.0000 4.000 3.000 3.000 3.000 2.00000
## yhats   1.5069 1.5069  1.6141 1.8518 1.983 1.983 1.983 1.983 1.98348
## rres    0.4931 0.4931 -0.6141 0.1482 2.017 1.017 1.017 1.017 0.01652
##             19     20      21     22    23    24     25     26    27
## Matings  1.000  1.000  2.0000 3.0000 5.000 6.000  1.000  1.000 6.000
## yhats    2.125  2.125  2.1245 2.1245 2.437 2.437  2.611  2.611 2.611
## rres    -1.125 -1.125 -0.1245 0.8755 2.563 3.563 -1.611 -1.611 3.389
##              28     29      30     31     32     33      34      35    36
## Matings  2.0000  1.000  3.0000 4.0000  0.000  2.000  3.0000 4.00000 9.000
## yhats    2.7964  2.995  3.4363 3.6807  3.942  3.942  3.9424 3.94237 3.942
## rres    -0.7964 -1.995 -0.4363 0.3193 -3.942 -1.942 -0.9424 0.05763 5.058
##             37    38    39     40    41
## Matings  3.000 5.000 7.000  2.000 9.000
## yhats    4.223 4.523 5.189  5.558 7.316
```

$n = 41$

Case V Residuals Plot

```
par(mfrow=c(1,2))
plot(Age, pres, main="Pearson Residuals plot")
abline(h=0)
plot(Age, dres, main="Deviance Residuals plot")
abline(h=0)
```



No obvious outliers.

Case V Estimating ψ

```
(psihat=sum(residuals(fitllm, type="pearson")^2/fitllm$df.residual))
```

```
## [1] 1.157
```

Assume $\psi = 1$

```
summary(fitllm, dispersion=psihat)
```

```
##
## Call:
## glm(formula = Matings ~ Age, family = poisson, data = elmasu)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8080  -0.8614  -0.0863   0.6009   2.1778
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.5820     0.5859   -2.70   0.0069 **
## Age           0.0687     0.0148    4.65  3.4e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1.157)
##
##     Null deviance: 75.372  on 40  degrees of freedom
## Residual deviance: 51.012  on 39  degrees of freedom
## AIC: 156.5
```

$Var(y_i) = \psi \mu_i$

$\hat{\psi}$

$se(\hat{\beta}_{\hat{\psi}}) > se(\hat{\beta}_{\psi=1})$

$se(\hat{\beta}_{\hat{\psi}}) = \sqrt{\hat{\psi}}\, se(\hat{\beta}_{\psi=1})$

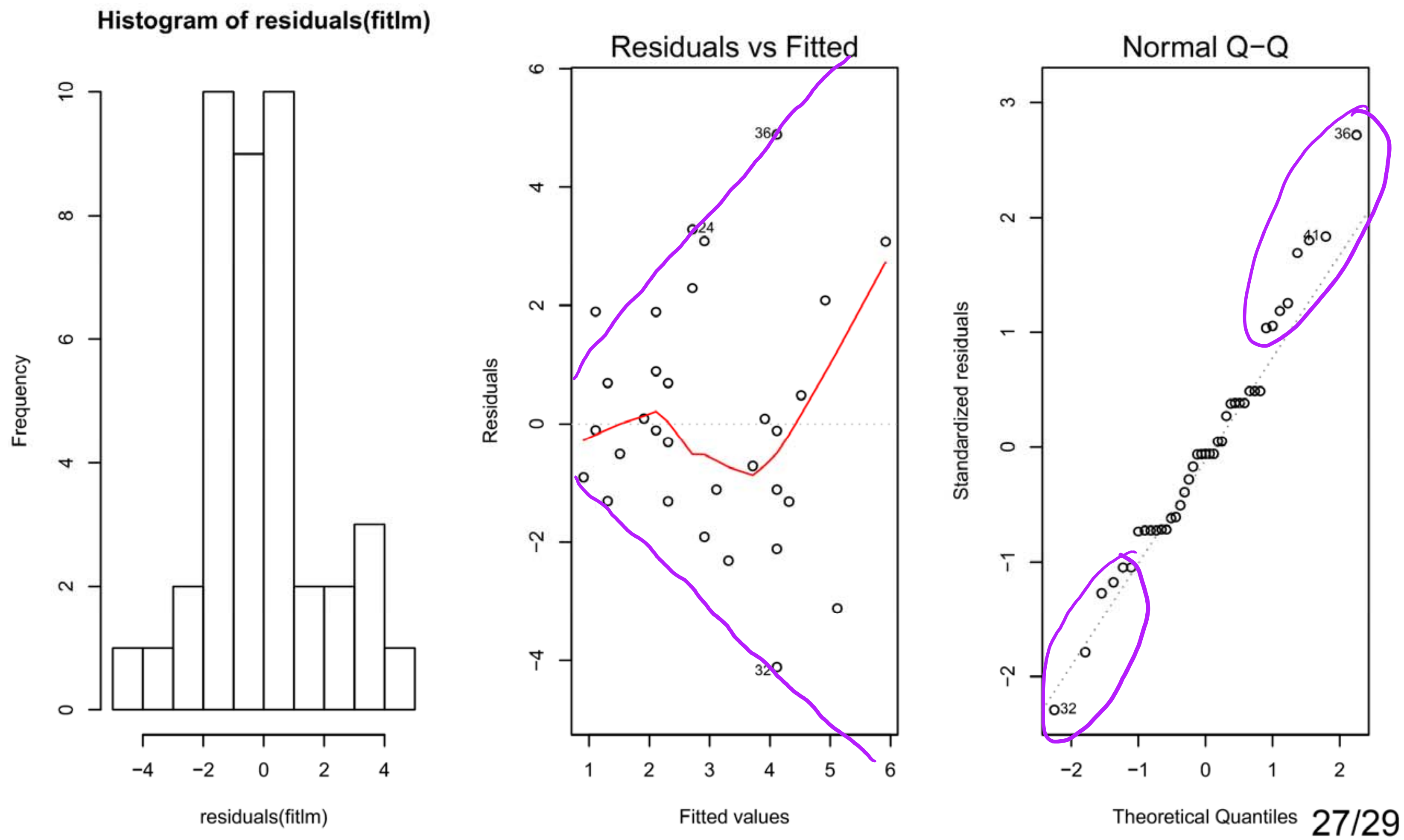
Case V: Simple Linear model

```r
fitlm= lm(Matings~Age)
summary(fitlm)
```

```
##
## Call:
## lm(formula = Matings ~ Age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.116 -1.309 -0.108  0.889  4.884
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -4.5059     1.6190   -2.78   0.0083 **
## Age           0.2005     0.0444    4.51  5.7e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.85 on 39 degrees of freedom
## Multiple R-squared:  0.343,	Adjusted R-squared:  0.326
## F-statistic: 20.4 on 1 and 39 DF,  p-value: 5.75e-05
```

Case V: Simple Linear model assessment

```
par(mfrow=c(1,3))
hist(residuals(fitlm))
plot(fitlm, which=1)
plot(fitlm, which=2)
```



Case V: Log-transformed Linear model

```
fitlml= lm(log(Mating)~Age)
summary(fitlml)
```

```
##
## Call:
## lm(formula = log(Mating) ~ Age)
##
## Residuals:
##    Min      1Q Median     3Q    Max
## -2.036 -0.372  0.139  0.453  0.969
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.3588     0.5988   -2.27  0.02885 *
## Age           0.0628     0.0164    3.82  0.00046 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.684 on 39 degrees of freedom
## Multiple R-squared:  0.273,  Adjusted R-squared:  0.254
## F-statistic: 14.6 on 1 and 39 DF,  p-value: 0.000463
```

Case V: Log-transformed Linear model assessment

```
par(mfrow=c(1,3))
hist(residuals(fitlml))
plot(fitlml, which=1)
plot(fitlml, which=2)
```

Non-normal residuals.