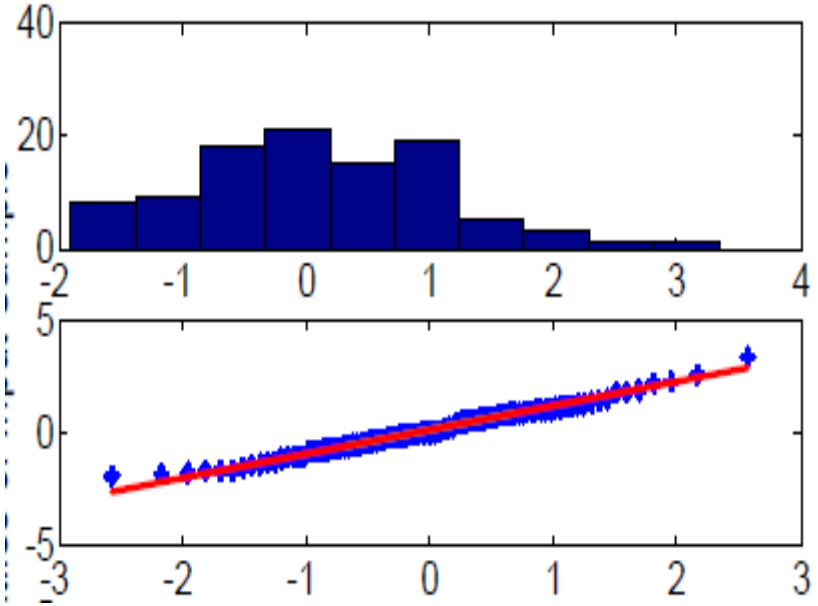
R Exercise 1: Checking linear model assumptions; outliers

SLIDES BY MAX LI, BASED ON BRIAN MCGILL'S MATERIALS AT STATS.BRIANMCGILL.ORG

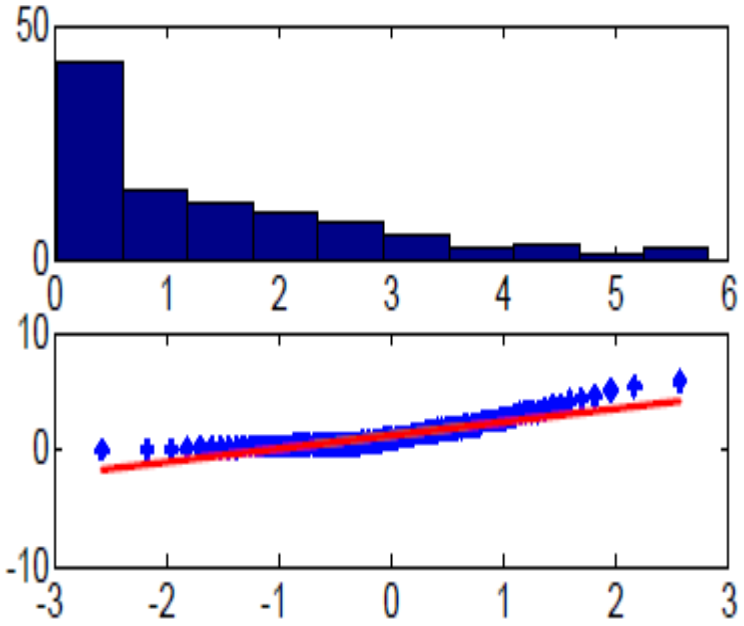
Checking the GLM assumptions (normality)

Histogram and Quantile-Quantile plot of residuals $\varepsilon_{i(j)}$: check the normality of errors

Normal



Non-normal (needs log transformation)



If non-normal

GLM are quite robust even if the errors are not normal.

As long as the distribution of errors has a peak with reasonable symmetry.

If the distribution is heavily skewed

- Transform the data (but see Zuur et al. 2010)
  - Log(Y) – If Y is the product (instead of sum) of independent factors
  - Sqrt(Y) – If expect y to be a Poisson count
  - 1/Y – common for rates (e.g. # offspring/female)
  - Arcsin(sqrt(Y)) – for proportion (0-1)
  - Power (Box-Cox) Transformation: more general, includes the above first 3. Available in MASS package in R.
- Report backtransform parameters (e.g. effect size)
- Non-parametric tests (but see Stewart-Oaten 1995)
- Generalized linear model

If heteroscedastic

ANOVA is robust if design is nearly balanced

Regression moderately robust, but the estimates are biased by one side of the regression

Good transformation may fix the problem

Using other methods that do not require homoscedasticity (e.g. generalized least square)

Outliers in GLM

The normal distribution assumes the tails are small, the chance of observing extreme values (~3+ standard deviations) is extremely low, extreme outliers “shouldn’t” occur given the often small sample size

Can heavily skew estimates

Detecting outliers

- Obvious in residual plots
- Outside the whiskers in Box plots
- Also calculate leverage of influence (or Cook’s distance)

How to handle outliers

Revisit the data point

- Most often a data entry or other human error

Revisit notes about that site/experiment/data point

Remove if:

- Has high leverage and care about estimating parameters
- Obvious experimental issue

If you remove – grossly unethical to fail to report this. OK to report and explain why

Checking GLM assumptions and outliers in R

```
birdsdiet <- read.csv('http://130.111.193.18/stats/birdsdiet.csv')
lm1 <- lm(MaxAbund~Mass,data=birdsdiet)
```

plot (lm1)

1st graph: How do the residuals change with the fitted value? Is the model linear? Errors independent? Homoscedastic?

2nd graph: Are the residuals normally distributed?

3rd graph: Similar to first graph except it uses sqrt(|ε_i|). Are the errors homoscedastic?

4th graph: Is there any data point that has strong influence on the model? (Those close or beyond the Cook's distance contour)