FISH 604
Module 3: Exploratory data analysis

Instructor: Franz Mueter
Lena Point, Rm 315
796-5448
fmueter@alaska.edu

MSL FISH 604

Review/preview

- Visualizing data
- Assessing distributions
- Outliers
- Standardization
- Transformations
- Correlations

Today

MSL
FISH 604

Objectives & outcomes

Objectives

- Review standard methods for **Exploratory Data Analysis** to conduct prior to statistical modeling

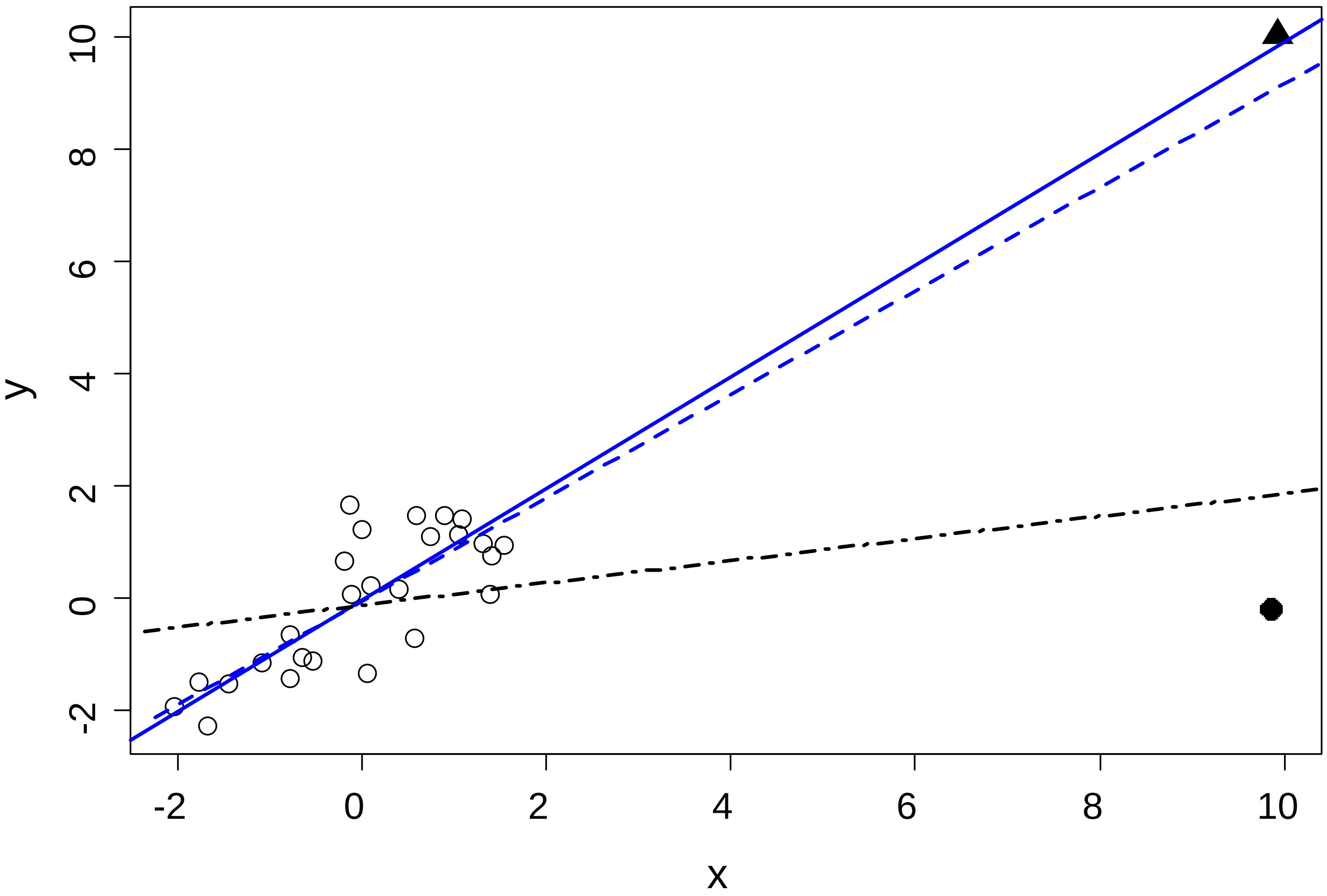
Outcomes

- Know how to detect outliers and what to do in the presence of outliers
- Be able to identify when and how to apply appropriate <u>data standardizations</u>
- Be able to identify when and how to apply appropriate <u>data transformations</u>
- Be proficient in quickly exploring correlations among a set of variables (graphically & statistically)

Outlier detection

- Why?
  - Tests and model fitting algorithms (least-squares, likelihood) are often sensitive to outliers
- How?
  - Graphical, distance from mean
- What to do?
  - Ignore
  - Eliminate (Compare results)
  - Use robust measures / estimation

Outliers: Influence & leverage



high leverage (low influence)

high influence (high leverage)

MSL FISH 604

Outliers: Leverage & Influence

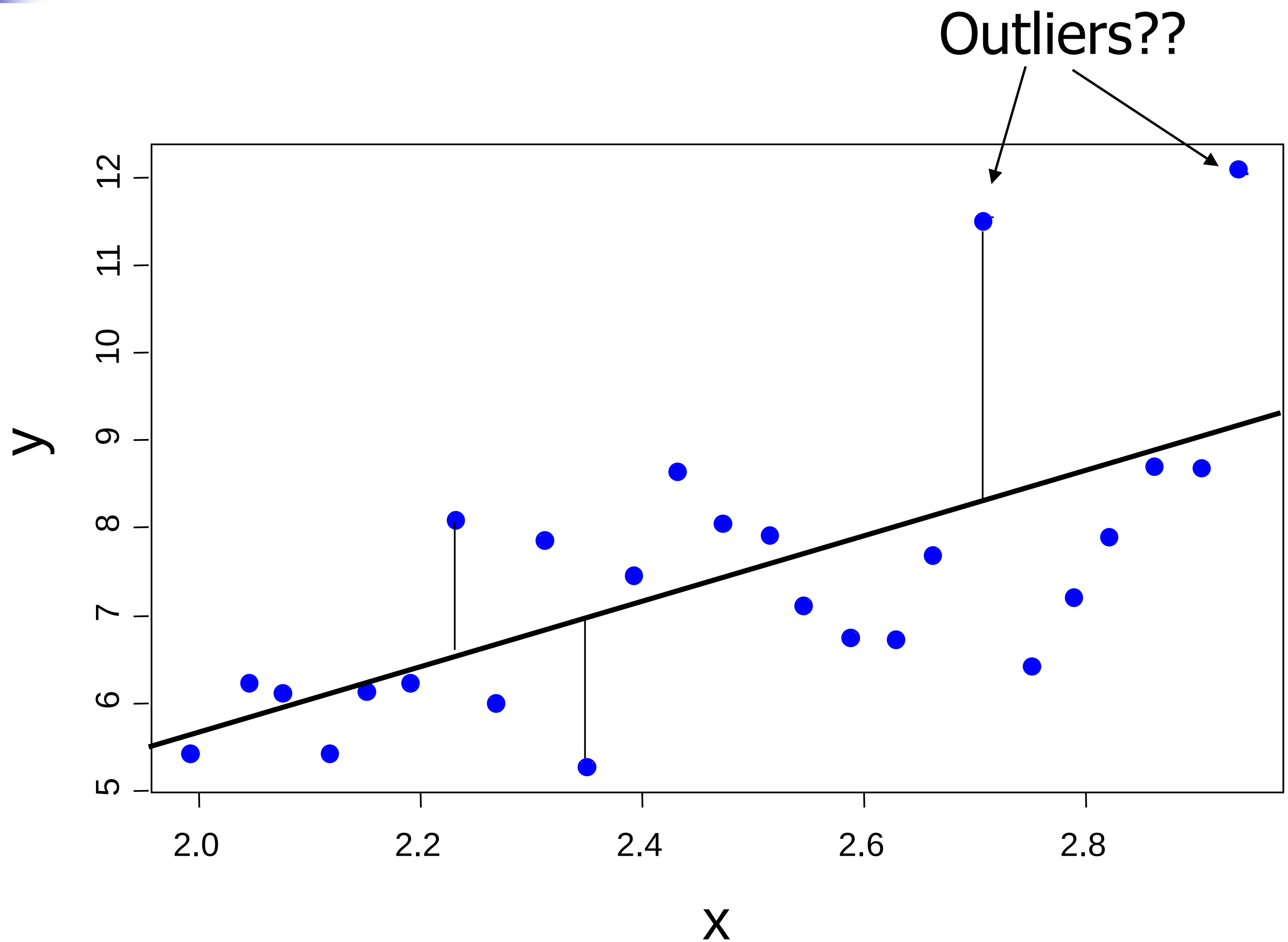
Leverage

- Defined as diagonals of "hat matrix": H_{ii}
- Large values often due to extreme values in X
- Rule of thumb: If leverage > 2p/n, look at the data point more closely
- Point with high leverage may or may not also be influential!

Influence

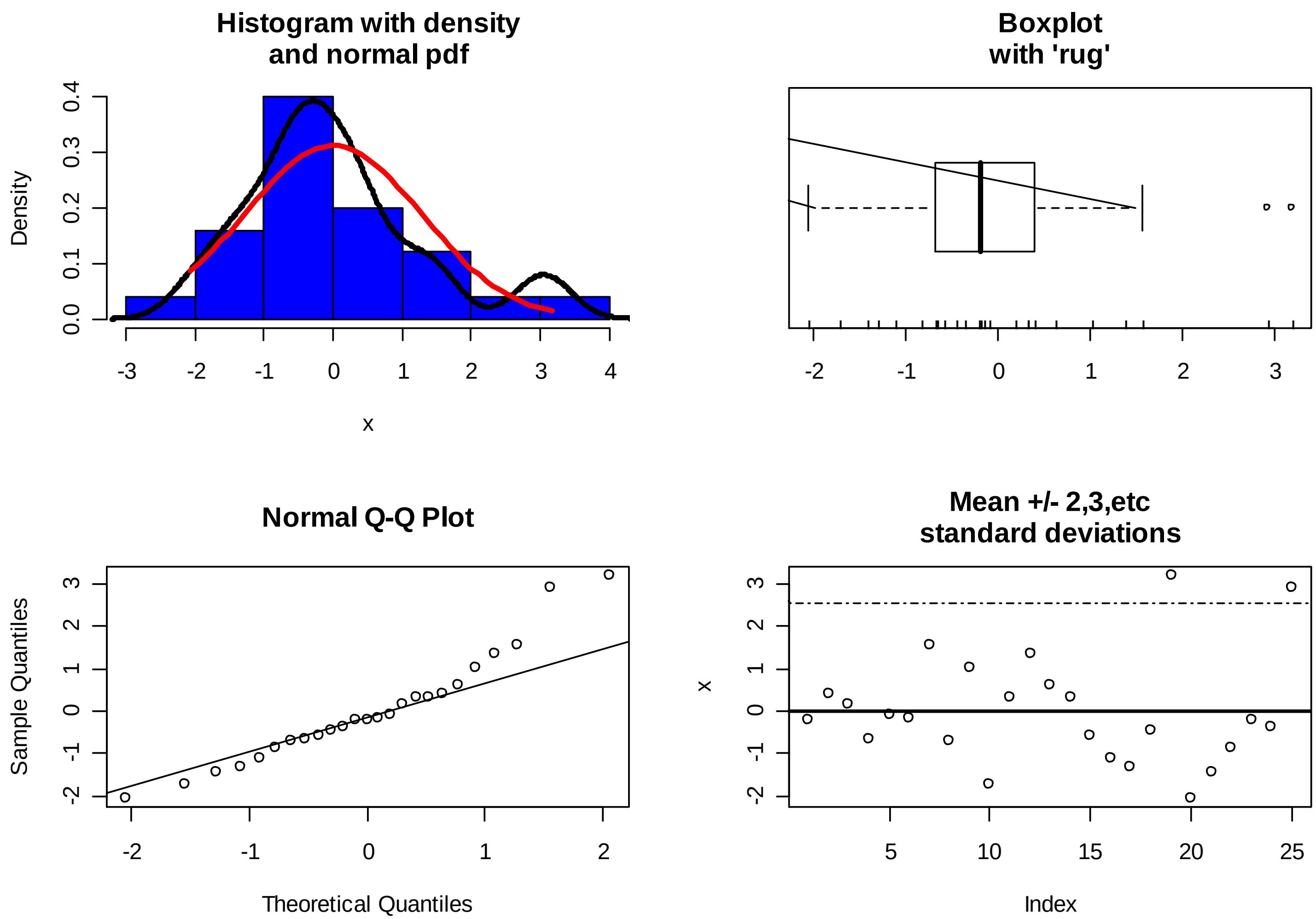
- Removal of influential point results in large change in fit!
- Usually measured by change in predicted values or regression coefficient(s) when i^{th} observation is removed,
- Cook's distance is a common measure of influence

Regression with outliers?

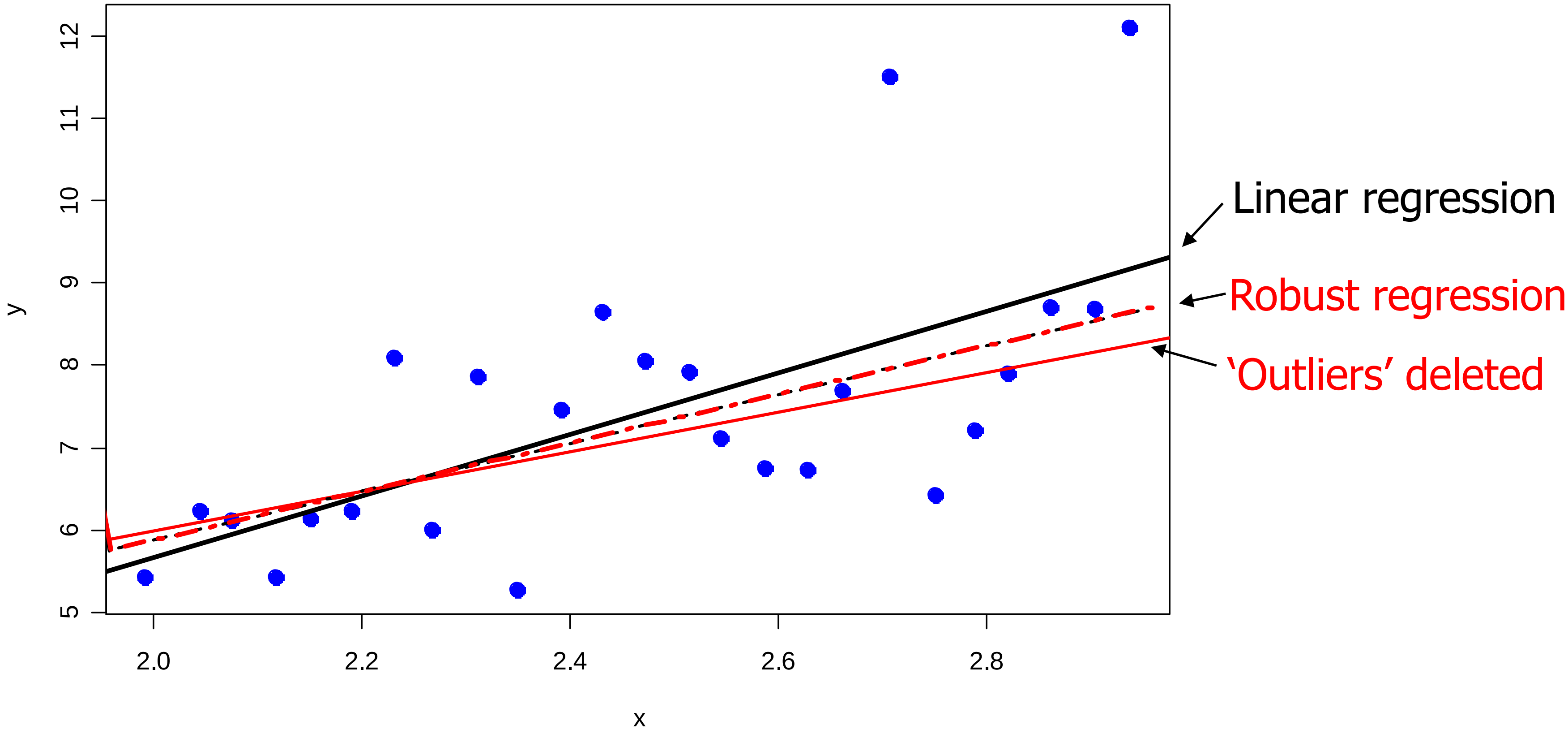


Outlier detection
(linear models or 'normalized' residuals)

Examine distribution of <u>residuals</u> from a model fit



Robust regression



<u>R code</u>:
```
plot(x,y)
fit <- lm(y~x); abline(fit)
fit.rob <- rlm(y~x); abline(fit.rob)
```

(`rlm` in MASS package)

Dealing with outliers:
Re-run analysis with outliers removed

Do the conclusions change when the case is deleted?

NO

Proceed with case included. Examine case to see if anything can be learned from it.

(acknowledge outlier)

YES

Is there a reason to believe that outlier comes from a different population (or is an error)?

YES

Omit case and proceed!

NO

Does the case have unusually "distant" explanatory variables?

YES

Omit the case and proceed! Report conclusions for the reduced range of explanatory variables

NO

Need more data or info to resolve! Report conclusions for analysis w/ and w/out the influential case!

Modified from Ramsey & Schafer (2002)

Standardizations

- Why?
  - Plot / compare variables on common scale
  - Variables in multivariate analyses
    - To make sure each variable has the same influence
  - Multiple regression
    - Stabilize numerical estimation methods
    - **Make coefficients comparable**
- How?
  - **Standardize to $N(0,1)$**
  - Standardize to common maximum
  - Standardize relative to mean

Standardization / scaling

MSL FISH 604

- Normalize to $N(0,1)$: (any continuous variable)

$$x' = (x_i - \bar{x}) / sd(x)$$

→ R function `scale()`

- Relative to mean (x > 0):

$$x' = (x_i - \bar{x}) / \bar{x} \quad \Rightarrow \quad \bar{x}' = 0$$

$$x' = x_i / \bar{x} \quad \Rightarrow \quad \bar{x}' = 1$$

- Same maximum (x > 0):

$$x' = x_i / \max(x) \quad \Rightarrow \quad 0 < x' < 1$$

Transformations

- Why?
  - To meet regression assumptions:
    - Normalize residuals
    - Eliminate / reduce heteroscedasticity
    - Ensure additivity of effects
  - Reduce leverage of extreme values
  - Linearize relationships (e.g. log-log plots)
- How?
  - Logarithmic (natural log) transformation
  - Square-root transformation
  - Arc-sine transformation
  - Power transformations
  - Box-Cox transformations

MSL FISH 604

Log-transformation

Add small constant if x may include 0

- $x' = \log(x + 1)$ OR $x' = \log(x)$ (typically base 10 or **natural logarithm**)
- Log-transform dependent variable in ANOVA when
  - Multiplicative effects are present (Non-additive)
- Log-transform dependent variable in regression when:
  - standard deviations are proportional to mean, i.e. CV is constant (heteroscedasticity)
- Try log-transforming both dependent (y) and independent (x) variable when
  - Errors are multiplicative: $y = f(x) * e^{\varepsilon}$

$$\rightarrow \log(y) = \log\{f(x)\} + \varepsilon$$

where $\varepsilon \sim N(0,\sigma^2)$

MSL
FISH 604

Log-transform for additivity

Example: two-way ANOVA

1. Additive effects

	Factor A		
Factor B	Level 1	Level 2	Level 3
Level 1	10	20	25
Level 2	20	30	35

+10 +5 +10

2. Multiplicative effects

	Factor A		
Factor B	Level 1	Level 2	Level 3
Level 1	10	30	60
Level 2	20	60	120

*3 *2

violates ANOVA assumption

2. Log-transformed (log10)
-> additivity restored!

	Factor A		
Factor B	Level 1	Level 2	Level 3
Level 1	1.00	1.48	1.78
Level 2	1.30	1.78	2.08

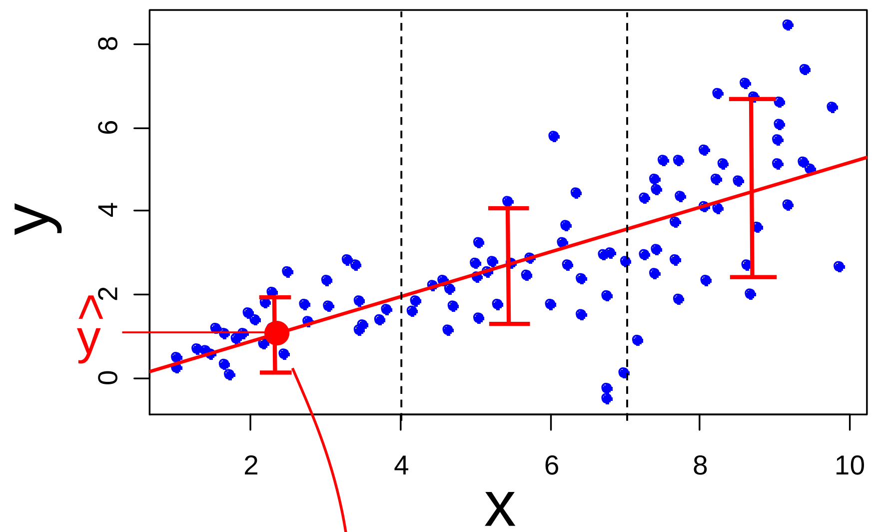
+.48 +.3

meets ANOVA assumption

Log-transform for homoscedasticity

Example: simple linear regression

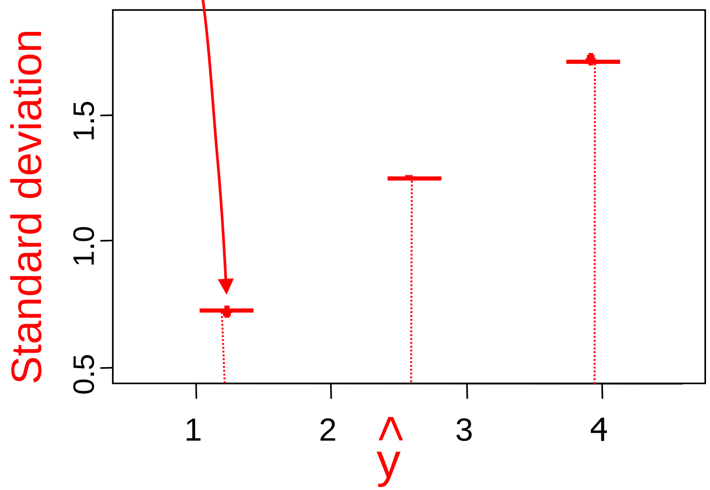
Standard deviation of y increases in proportion to mean of y



Heteroscedasticity violates regression assumption! (affects CIs & p-values)

Plot standard deviation of residuals against predicted y

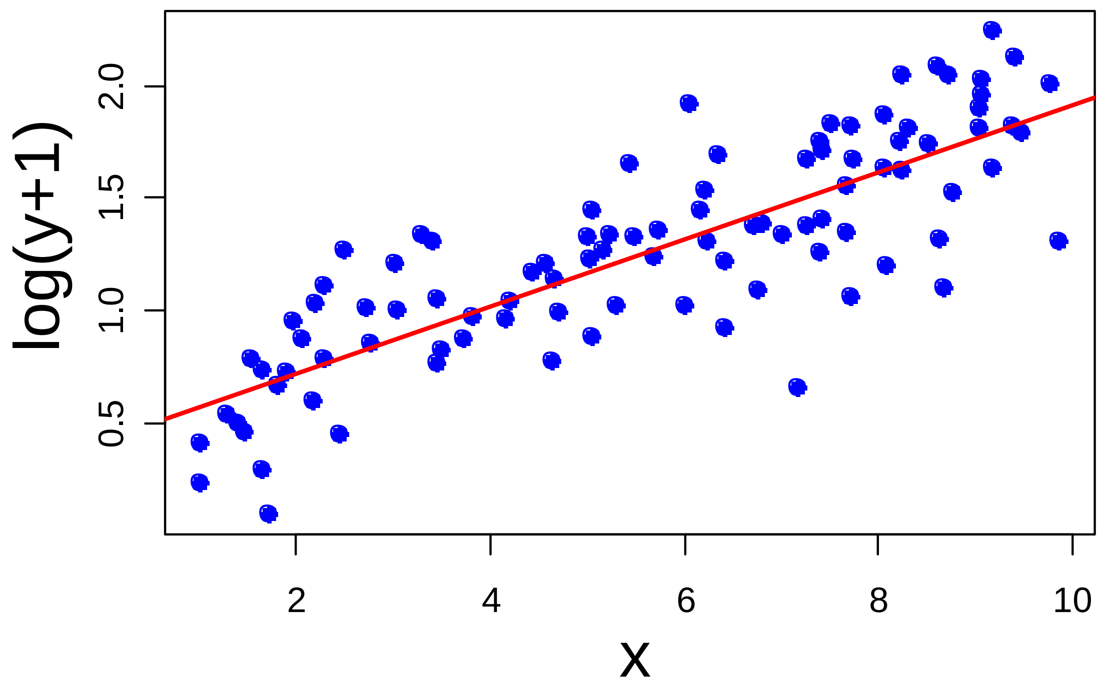
→ Linear increase



Visualize distribution of responses in vertical 'slices' to determine relationship between SD of response and predicted values

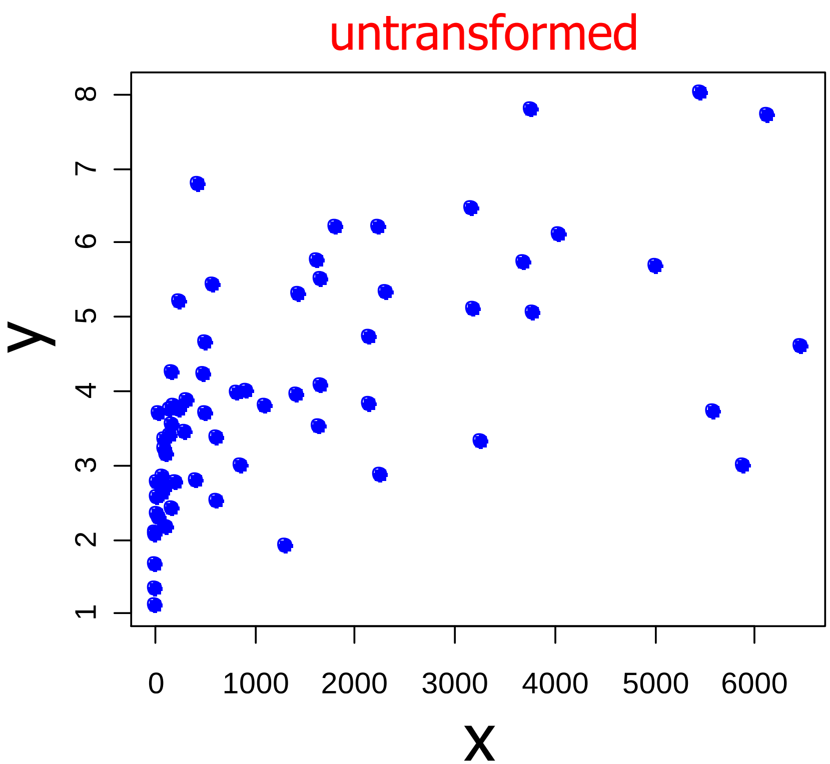
Log-transform for homoscedasticity

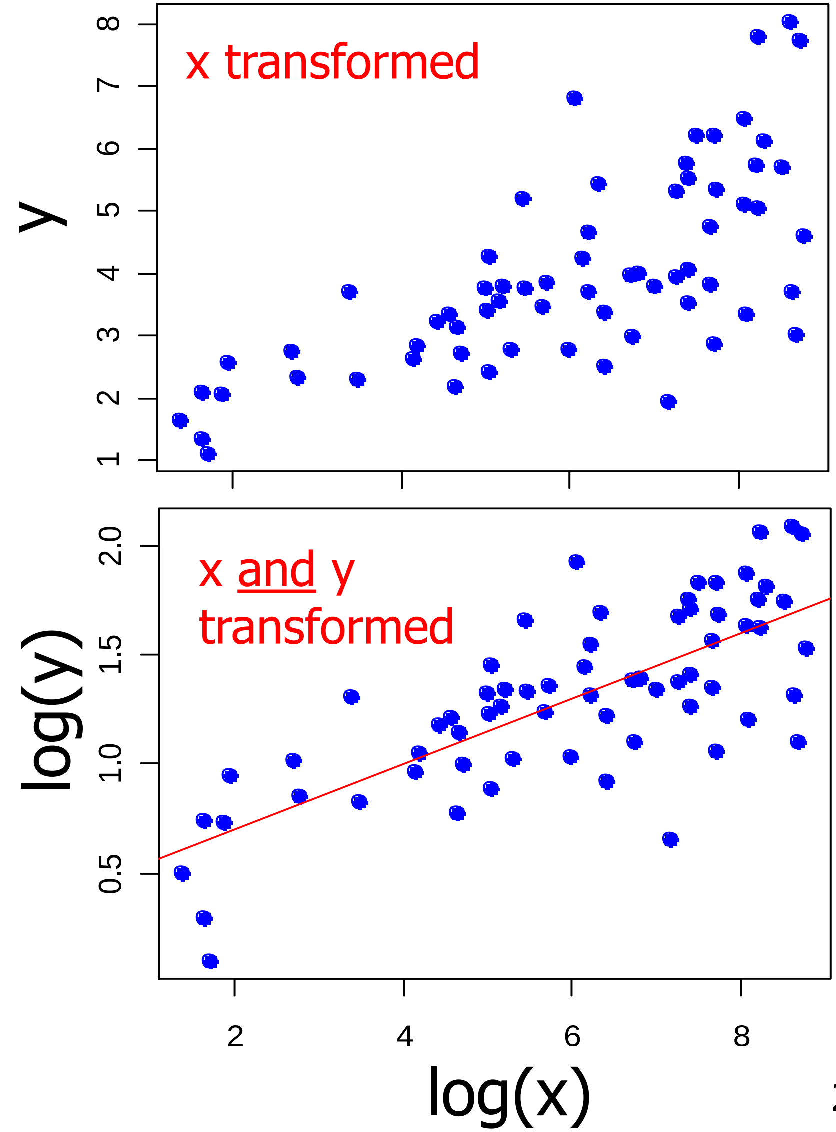
Example: simple linear regression with log-transformed y-variable



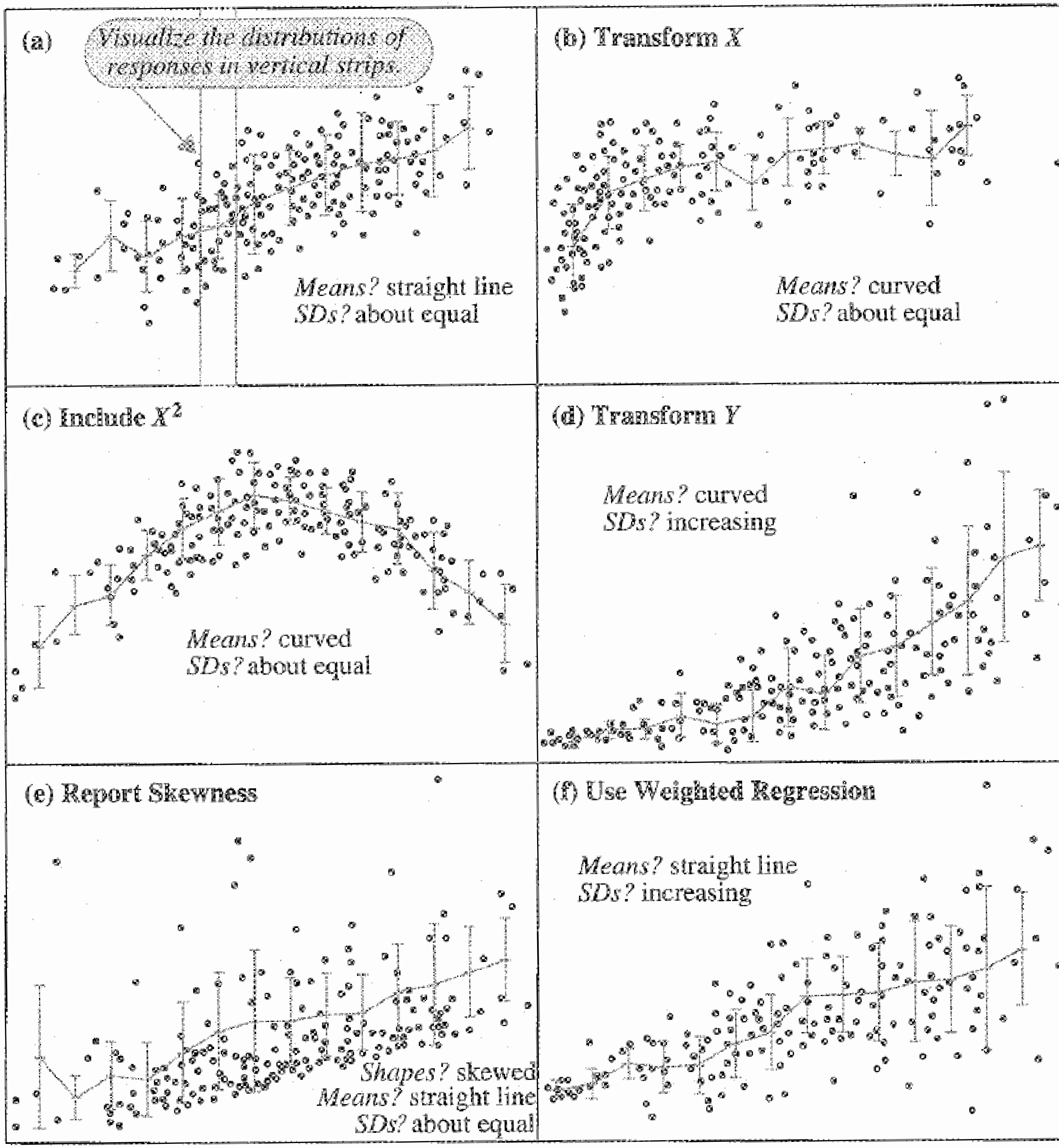
Transformed variable has <u>constant</u> standard deviation (Homoscedasticity)

Log-transform both *y* and *x*





→ In some cases, this can result in both linearity & homoscedasticity (and simplify analyses)



If possible, aim for equal variances and a 'linear' relationship

Other transformations

MSL FISH 604

- Square-root transformation
  - When variances are proportional to the mean (e.g. in biology: samples from a Poisson distribution)

$$y' = \sqrt{y + 3/8}$$

- Arc-sine
  - To normalize proportions, as long as they are not too close to zero (~ 0.2 – 0.8)

$$p' = \arcsin\sqrt{p}$$

Other transformations

- Power transformation
  - When standard deviations decrease with the mean and/or if distribution is left-skewed:

(rare!)

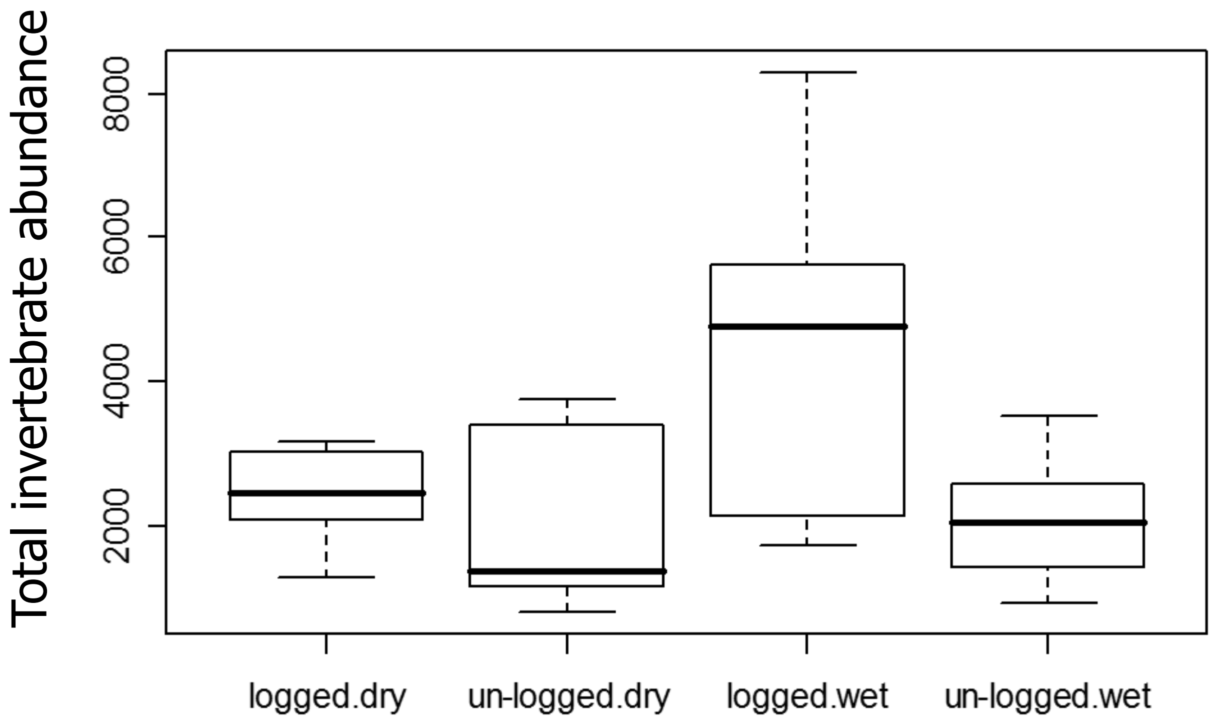
$$y' = y^2$$

- Box-cox transformation
  - General approach to determine optimal transformation for normalizing data

$$y' = \begin{cases} y^{\lambda} - 1/\lambda & \lambda \neq 0 \\ \log(y) & \lambda = 0 \end{cases}$$

Box-Cox transformation

- Which transformation, if any?



R-code: `boxplot(totalN ~ treatment * ecoregion, data=logging)`

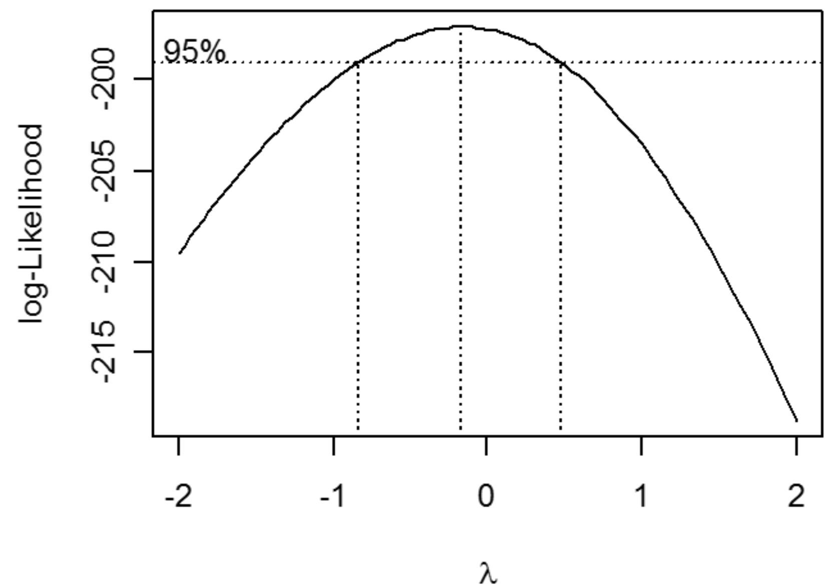
Box-Cox transformation

- Fit desired model:

```
fit <- aov(totalN ~ treatment * ecoregion,
     data=logging)
```

- Determine "best" transformation:

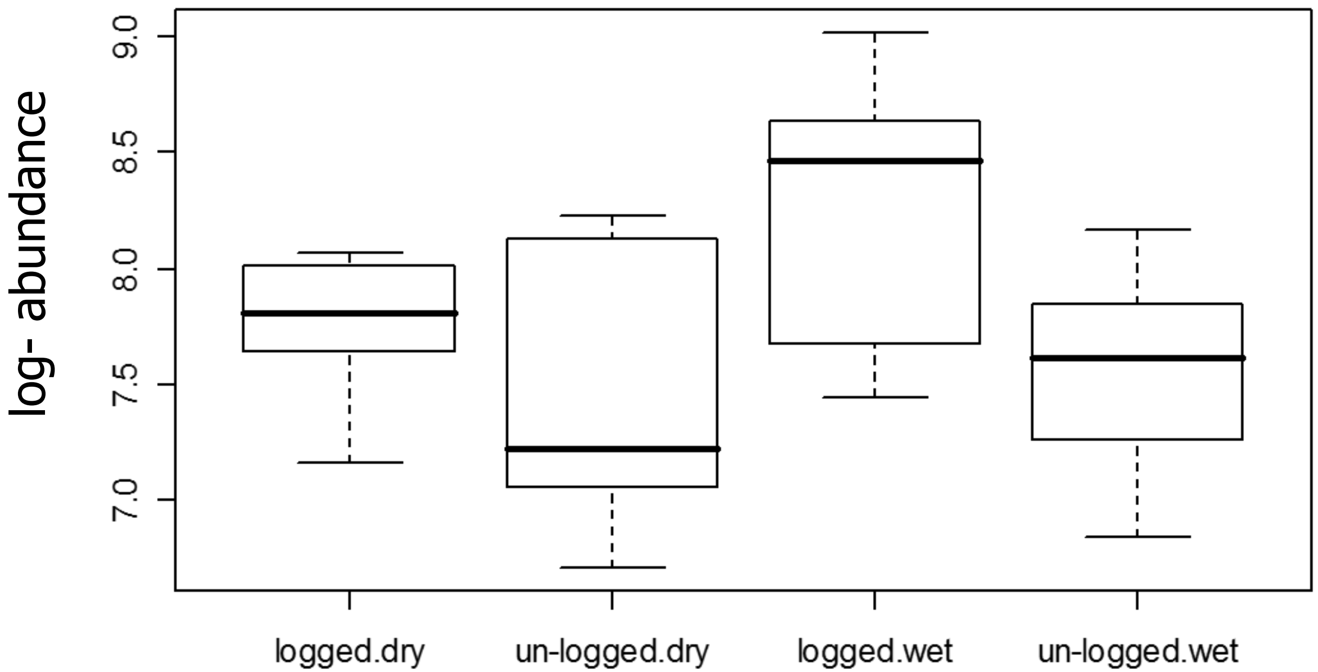
```
boxcox(fit)
```



- log-transform!?

Box-Cox transformation

- Approximately normal distribution <u>and</u> equal variances:



R-code: `boxplot(log(totalN) ~ treatment * ecoregion, data=logging)`

Correlations

- Correlation matrix
- Pearson's product moment correlation
- Robust correlations (rank based)
  - Spearman's rho
  - Kendall's tau
- Testing significance of correlations

<u>Some useful R functions</u>

```
cor(); cor.test(x,y,method="pearson")
cor.test(x,y,method="spearman")
rcorr() # library(Hmisc)]
corrplot() # library(corrplot)
ggcorrplot()  # library(ggcorrplot)
```

FISH 604 MSL

Correlation matrix

p < 0.05 (light blue)
p < 0.01 (dark blue)

	airT.Kodiak	airT.Sitka	GAK1.SST	GAK1.SSS	GAK1.BS	FW.spring	FW.annual	DW.win	DW.sum
airT.Kodiak	1	0.68	0.69	-0.58	0.08	0.50	0.27	0.04	-0.17
airT.Sitka	0.68	1	0.65	-0.83	0.19	0.69	0.60	-0.30	-0.10
GAK1.SST	0.69	0.65	1	-0.55	-0.20	0.70	0.36	0.01	-0.20
GAK1.SSS	-0.58	-0.83	-0.55	1	-0.35	-0.68	-0.38	0.44	0.16
GAK1.BS	0.08	0.19	-0.20	-0.35	1	-0.08	0.01	-0.39	0.31
FW.spring	0.50	0.69	0.70	-0.68	-0.08	1	0.40	-0.04	-0.13
FW.annual	0.27	0.60	0.36	-0.38	0.01	0.40	1	-0.26	-0.30
DW.win	0.04	-0.30	0.01	0.44	-0.39	-0.04	-0.26	1	0.04
DW.sum	-0.17	-0.10	-0.20	0.16	0.31	-0.13	-0.30	0.04	1

Reading assignment

- Zuur et al. (2007). **Analyzing Ecological Data**. Springer. *Chapter 4: Exploration* (see pdf posted on Canvas)

MSL FISH 604

Further reading

Graphical analysis

- Cleveland, W.S., 1993. Visualizing Data. AT&T Bell Laboratories, Murray Hill, NJ.
- Wilkinson, L. 1999. The Grammar of Graphics. Springer, New York
- Yau, Nathan 2013. Data Points: Visualization that means something. Wiley.

Multivariate data

- Cook, D., Swayne, D.F., and Buja, A. 2007. Interactive and Dynamic Graphics for Data Analysis: With R and GGobi. Springer, New York.

General exploratory analyses

- Zar, J.H., 1984. Biostatistical Analysis. Prentice-Hall, Englewood Cliffs, NJ.
- Barnett V, 2004. Environmental Statistics: methods and applications, John Wiley & Sons, Chichester, England.