FISH 604
Module 5: Linear models

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

Objectives and Outcomes

- Objectives
  - Review Linear Models (simple/multiple linear regression, ANOVA, ANCOVA)
- Outcomes
  - You should be able to quickly:
    - fit linear models in R
    - extract components from the model for further manipulation & calculations of fitted values etc.
    - perform residual diagnostics
    - plot model results
    - compare different models

Regression analysis

- Introduction
- Classification of regression models
- (General) Linear Models
- Dummy variables (ANOVA)
- Contrasts (ANOVA)
- Interactions
- ANCOVA example / Nested effects
- Model diagnostics

Regression analysis

- Model functional relationship between a response variable (=dependent variable) and explanatory variables (=independent variables)

$$\text{Response} = f\,(\text{explanatory variables}) + \text{error}$$

$$y = f\,(x) + \varepsilon$$

$$f\,(x) = \mathrm{E}(\mathrm{Y} \mid \mathrm{X} = \mathrm{x})$$

We are estimating the expected value of the response at a given level of the independent variables!

Steps in regression analysis

FISH 604

- Problem statement
- Selection of potentially relevant variables
  - Data collection
  - Data reduction (as necessary)
- Model specification
  - Functional relationship(s)
  - Error structure
- Choice of fitting method
- Model fitting
- Model diagnostics
- Model selection
- Conclusions / Inference / Prediction based on "best" model or models

Repeat for alternative models

Regression models

- (General) Linear Models (LM)
  - Simple / multiple linear regression
  - Analysis of variance (ANOVA)
  - Analysis of covariance (ANCOVA)
- Generalized Linear Models (GLM)
  - Binomial models (e.g. logistic regression)
  - Poisson models, etc.
- Generalized Additive Models (GAM)
  - Non-parametric smoothers
- Non-linear models (NLM)
- Linear / Non-Linear Mixed Effects Models

FISH 604

Linear models (LM)

Parameter vector

"Design" matrix

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

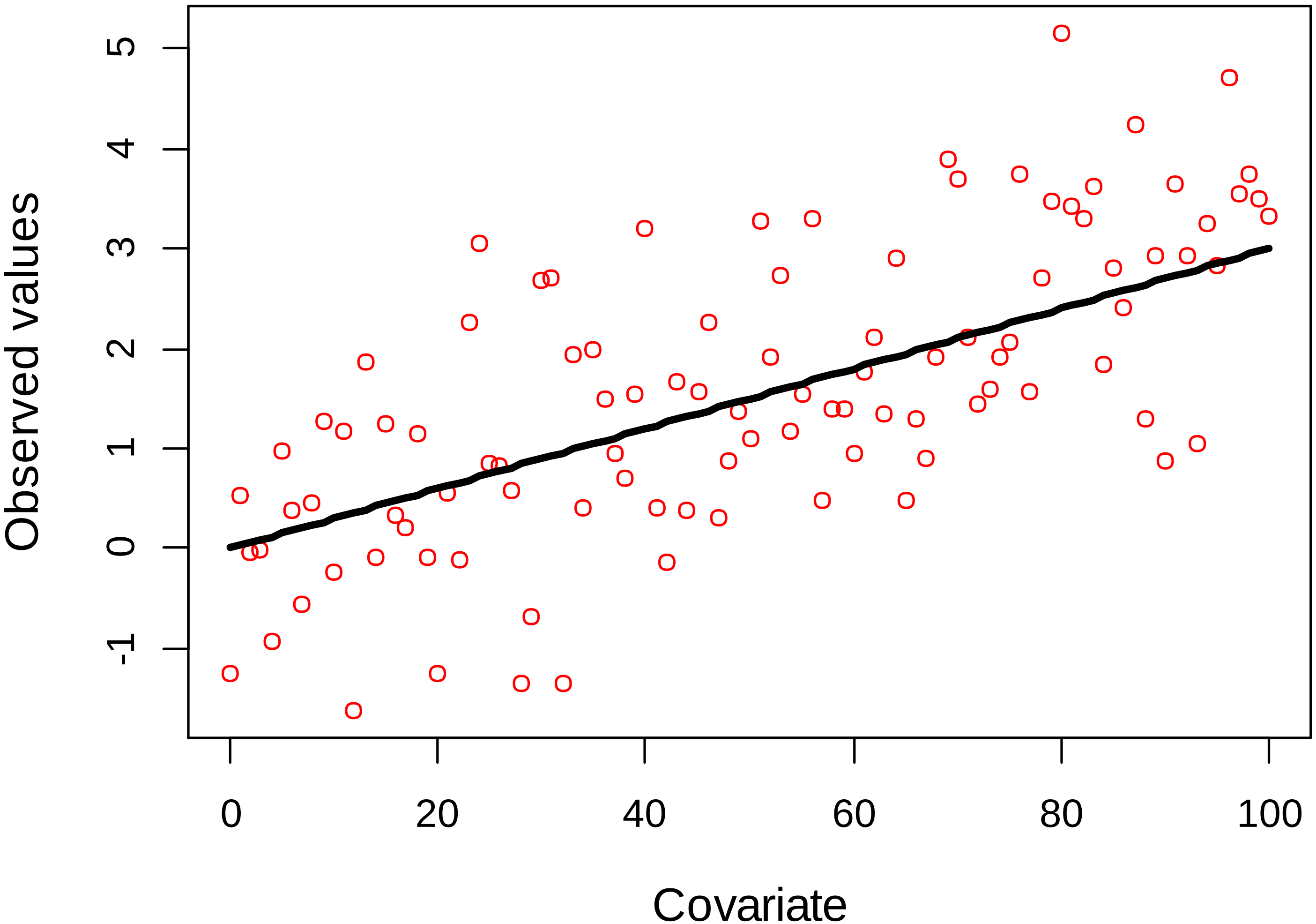
$$\mathbf{y} = \boldsymbol{\beta}\,\mathbf{X} + \boldsymbol{\varepsilon}$$

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$

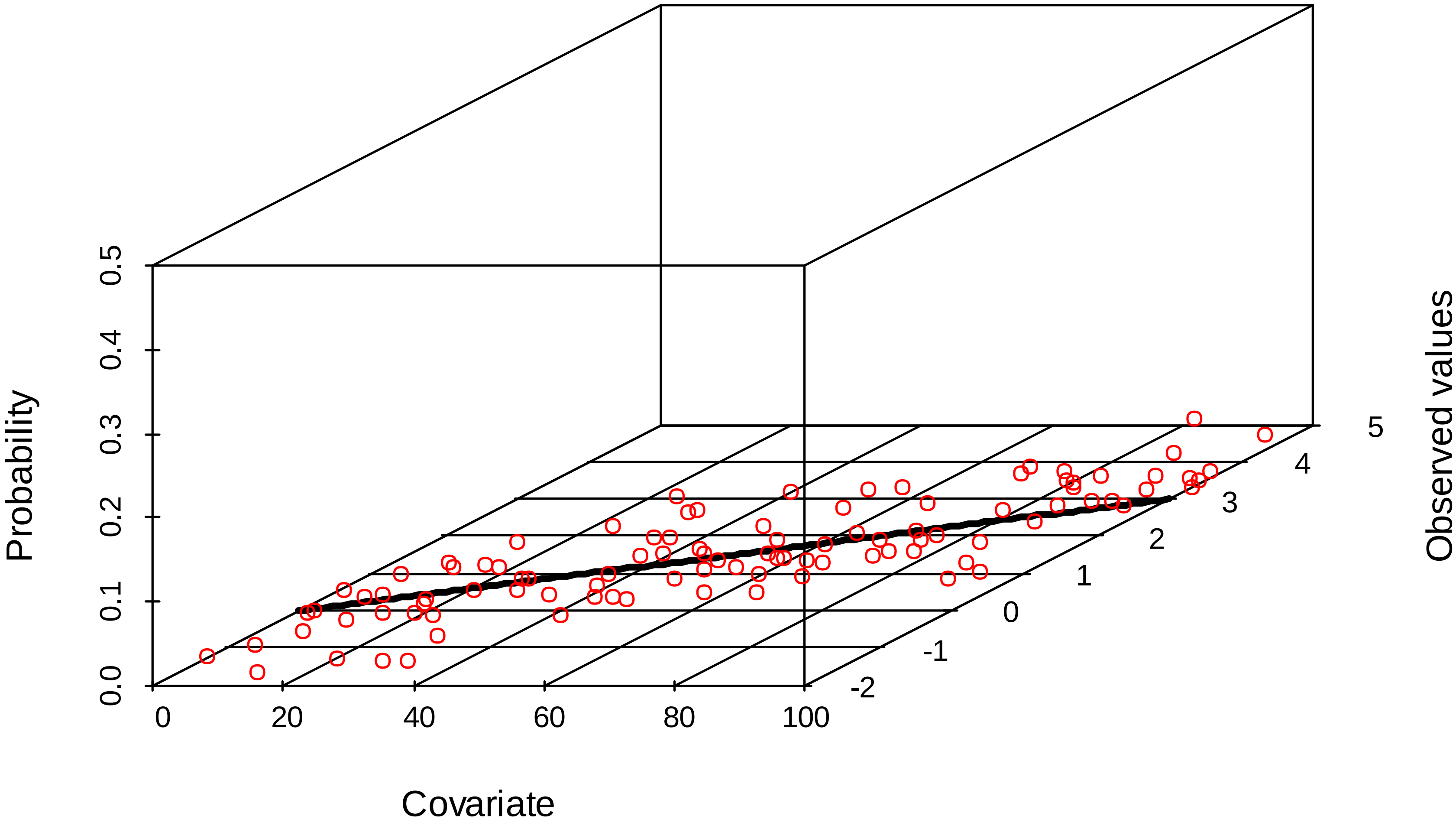
Model is <u>linear in the parameters!</u>
Independent variables may take on any form (x^2, $\log(x)$, a^x, *etc.*)

Normality assumption

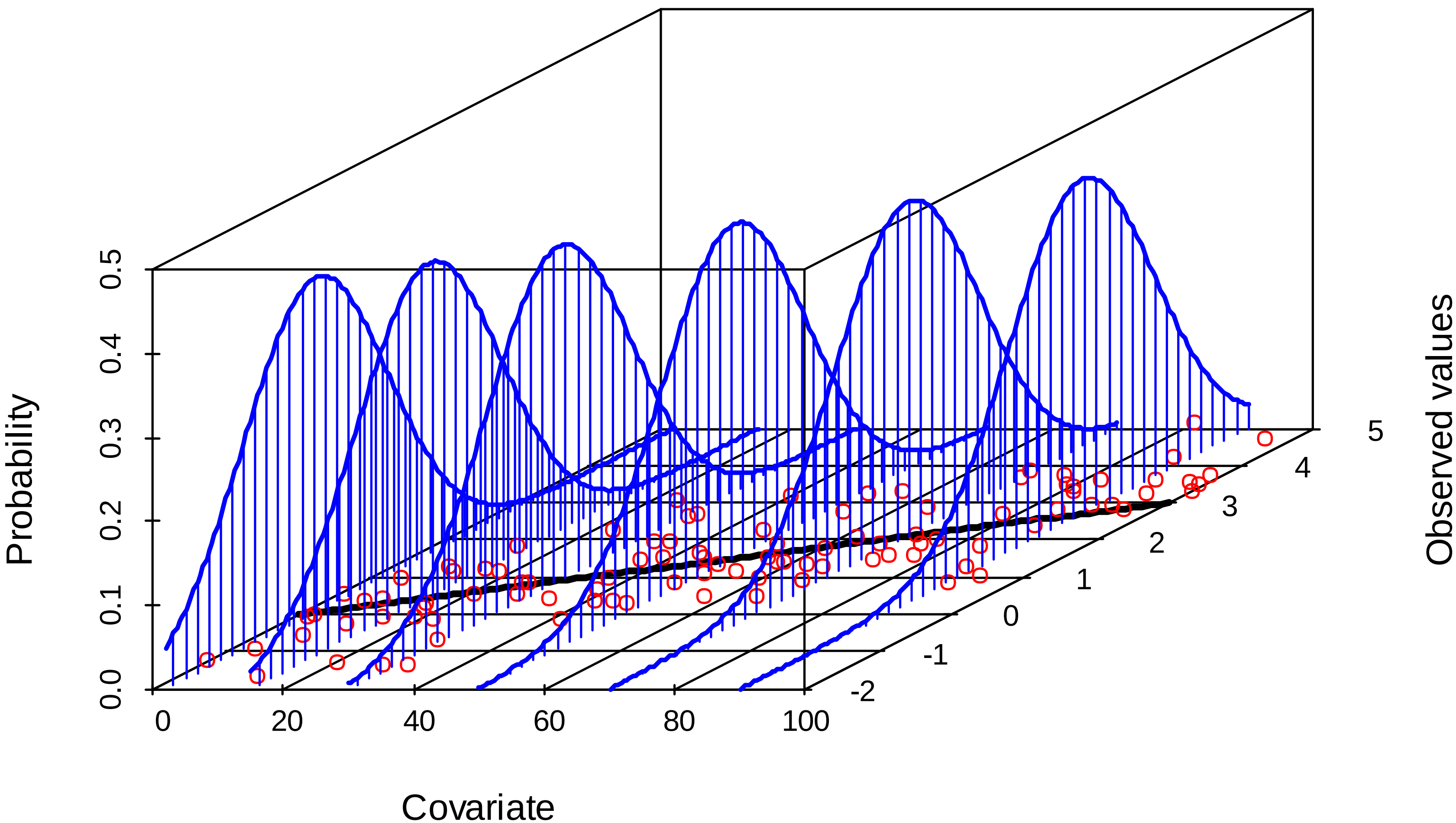
Example: Simple linear regression



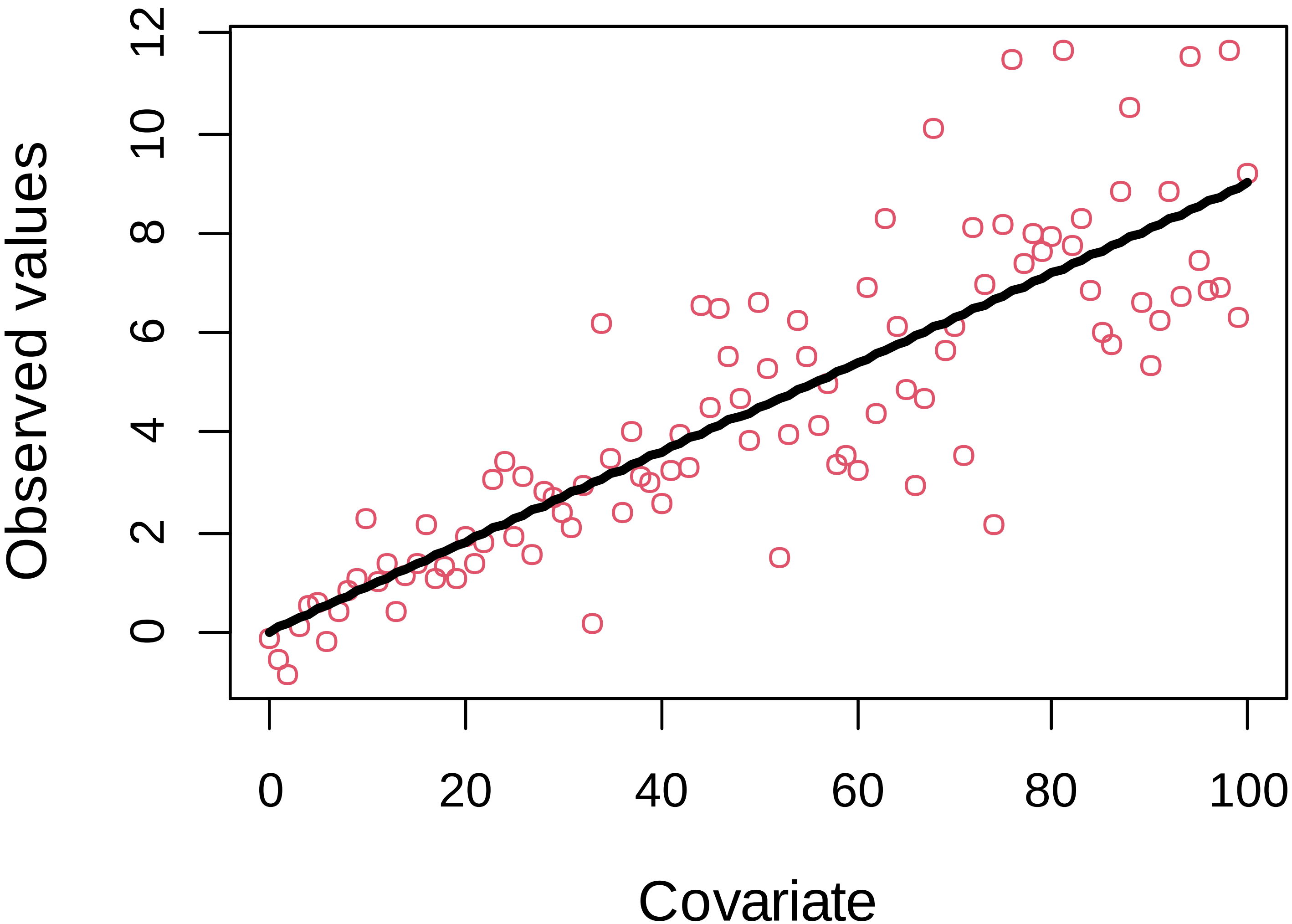
Normality assumption



Normality assumption

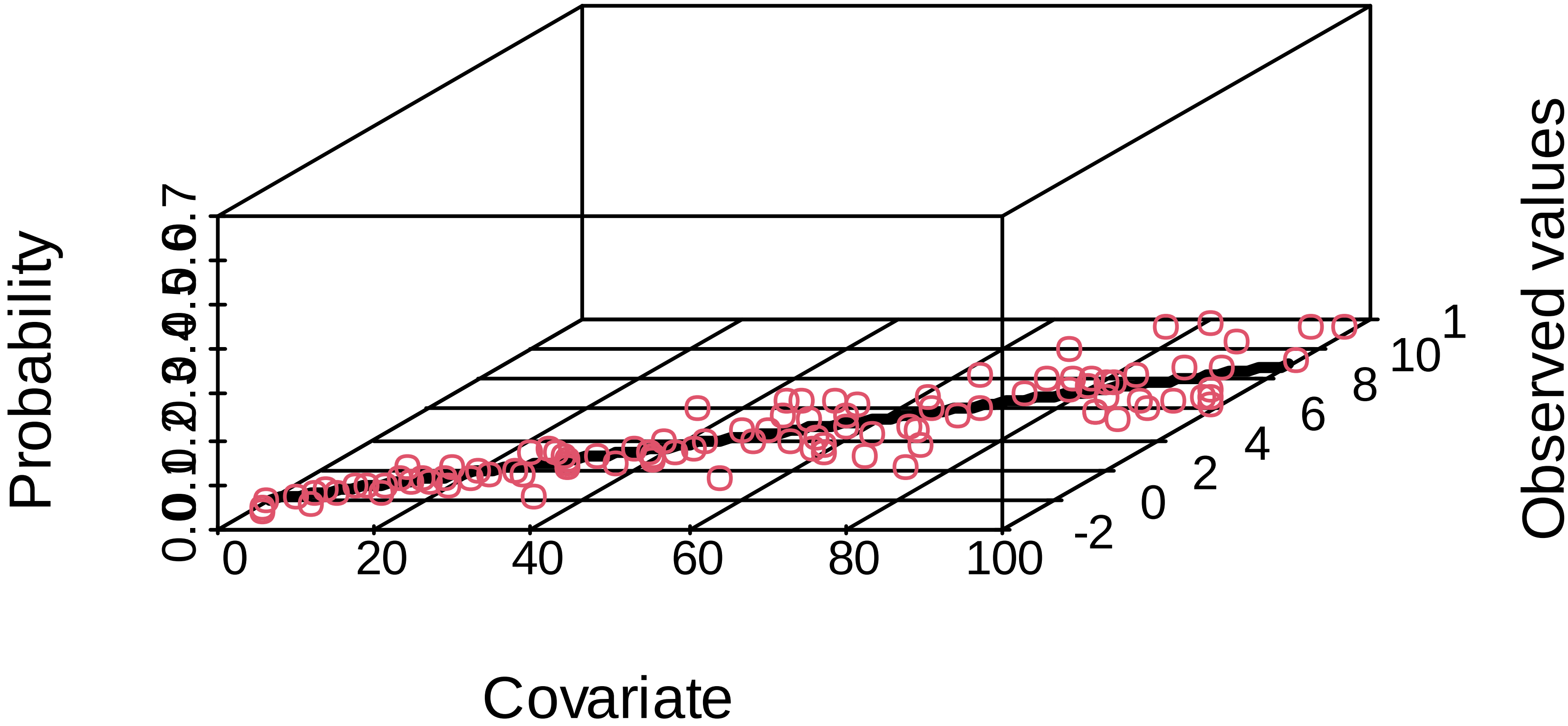


Normality assumption
(unequal variances)

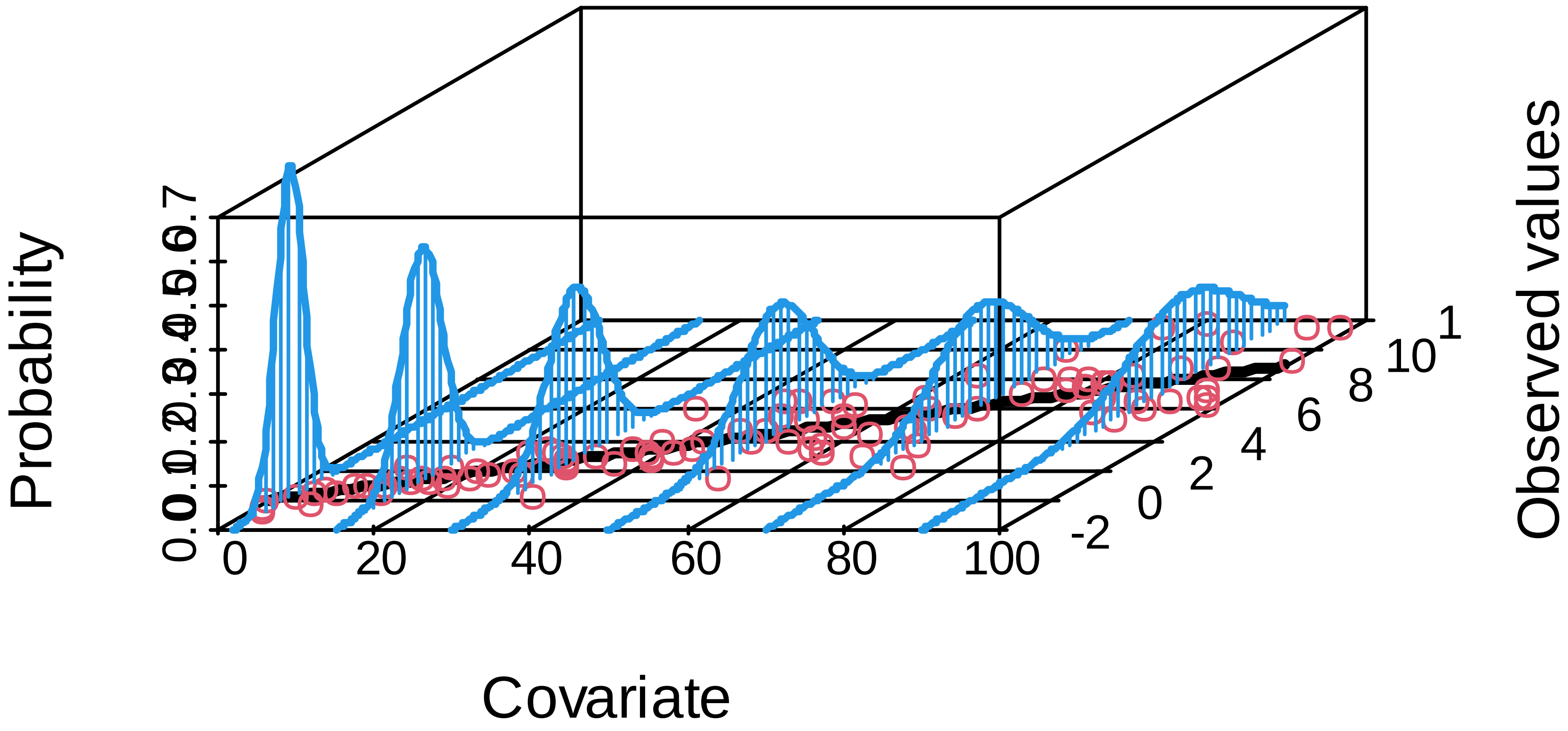


FISH 604

Normality assumption (unequal variances)



Normality assumption (unequal variances)



Linear models (LM)

- Continuous independent variables

  ⟹ Multiple linear regression

- Categorical independent variables

  ⟹ Analysis of variance (ANOVA)

- Continuous + categorical variables

  ⟹ Analysis of covariance (ANCOVA)

Multiple Linear Regression

- <u>Linear regression models</u>

$$y_i = \alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_p x_{p,i} + \varepsilon$$

Interaction: $y_i = \alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 (x_{1,i} \cdot x_{2,i}) + \varepsilon$

- Any linear model (multiple regression, ANOVA, ANCOVA) can be formulated as above using 'dummy' variables (contrasts).

Analysis of variance (ANOVA)

- ANOVA models

categorical variable (factor)

r 'levels'

One-way: $y_{ik} = \alpha + \mu_i + \varepsilon_{ik} \qquad (i = 1,\ldots,r; \;\; k = 1,\ldots,n)$

Two-way: $y_{ijk} = \alpha + \mu_i + \nu_j + \varepsilon_{ijk} \qquad (j = 1,\ldots,c)$

Interaction: $y_{ijk} = \alpha + \mu_i + \nu_j + \gamma_{ij} + \varepsilon_{ijk}$

Interaction term

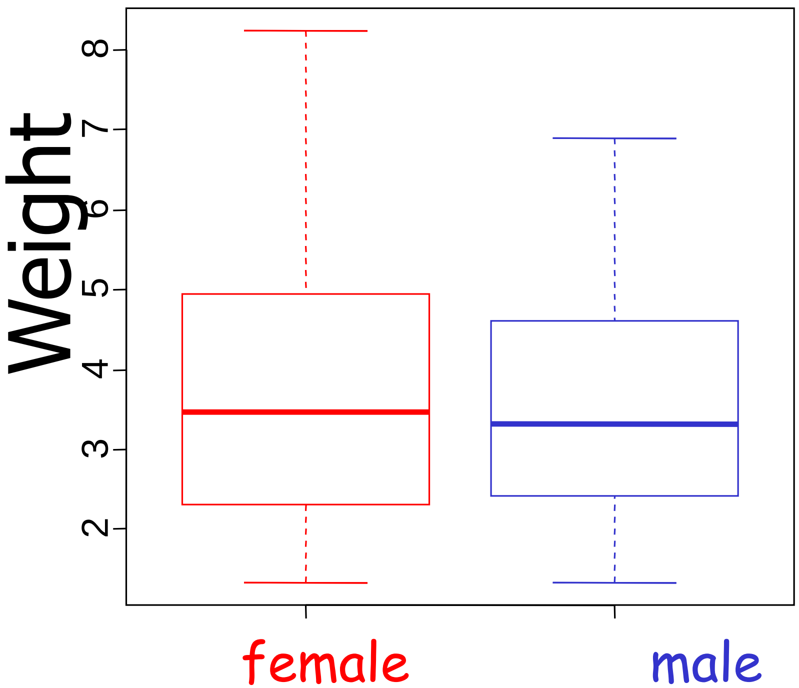
ANCOVA

- Combines continuous & categorical covariates
- Test for effect of a categorical variable on a response variable, while accounting / controlling for effect of one or more other variables, called "covariates" that also affect the response.
- In "true" ANCOVA:
  - Regression on the covariates is used to predict the response
  - ANOVA is done on the residuals to see if factors are still significantly related to the response after any variation due to the covariates has been removed
- In regression setting, effects of categorical and continuous variables are estimated simultaneously

ANCOVA example

Weight of fish by sex

Is there a difference in weight between males and females?



No covariate (one-way ANOVA):

Model: $logW = \alpha_s$

or: $logW = \alpha_F + \alpha_M \cdot D_M$

R code & model summary:

```
fit1 <- lm(log(Weight) ~ Sex)
                Estimate Std.Err. t value P-value
Intercept αF     1.216   0.0508     23.93  <2e-16
SexM    αM      -0.056   0.07185   -0.785   0.433
```

not signif!

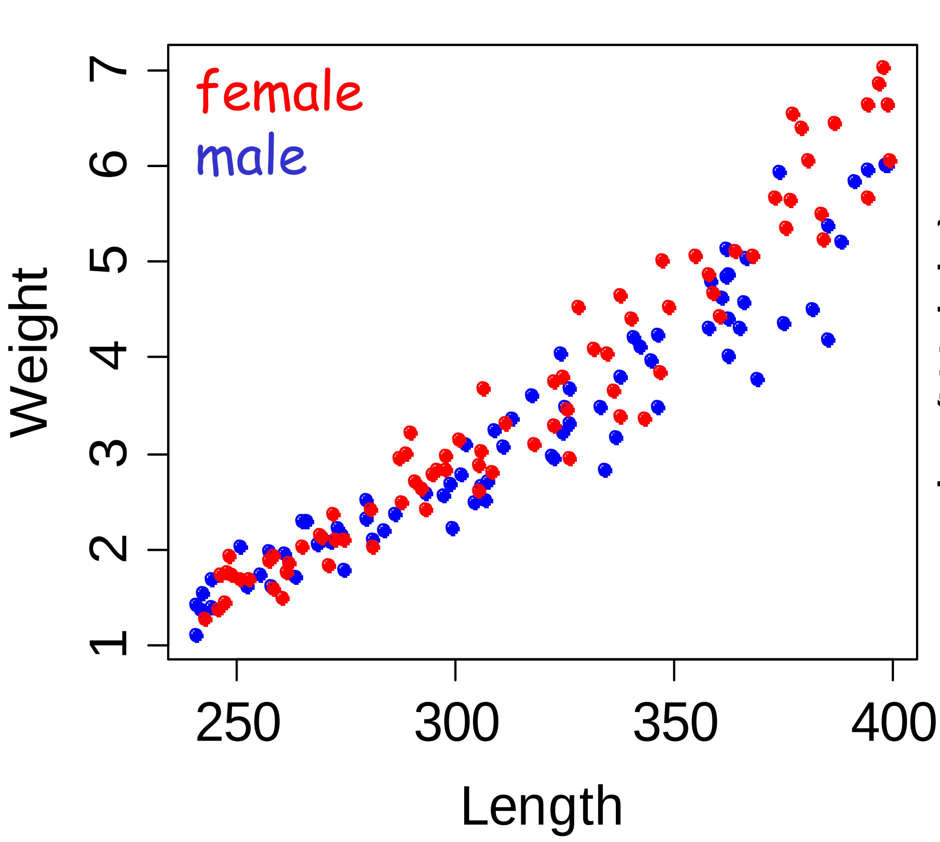
Dummy variable: D_M=0 for females
D_M=1 for males

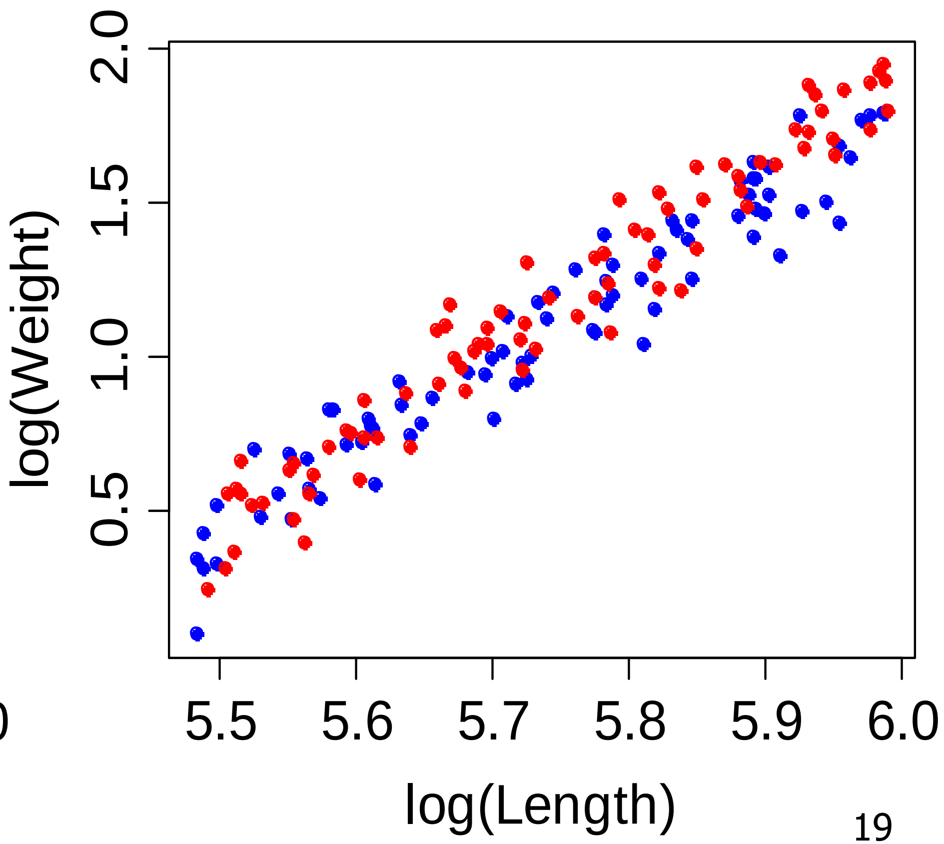
ANCOVA example

Weight at length of fish by sex

Is there a difference in weight between males and females?

Need to account for effect of length!!





ANCOVA example

Weight at length of fish by sex

Is there a difference in weight between males and females?

Need to account for effect of length!!

log(Length) as covariate (ANCOVA):

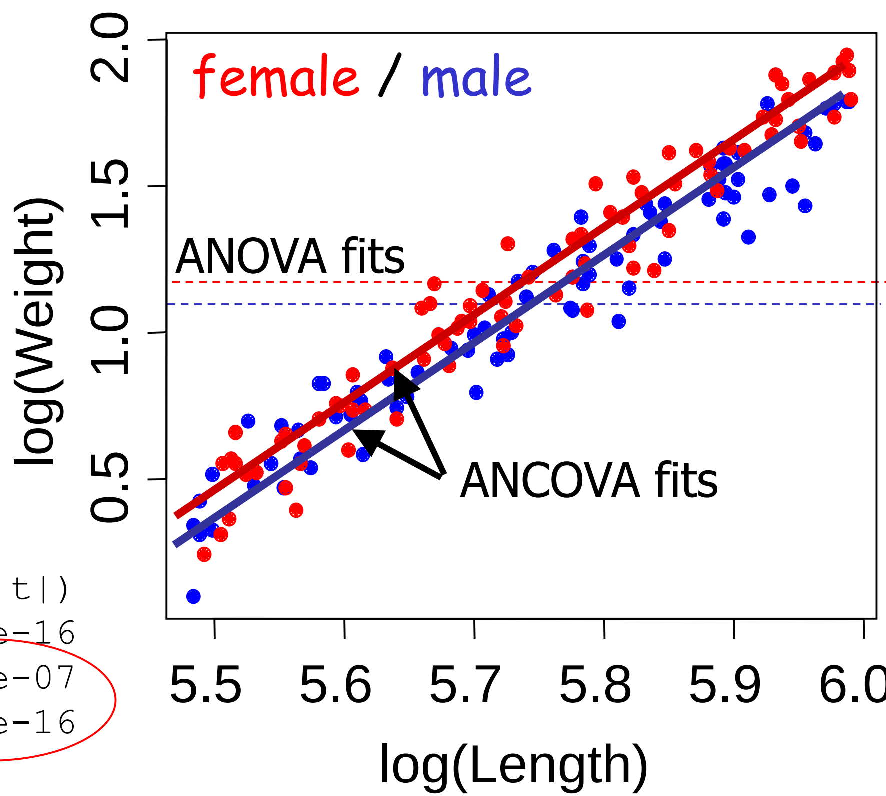
$$logW = \alpha_S + \beta \cdot logL$$

or:

$$logW = \alpha_F + \alpha_M \cdot D_M + \beta \cdot logL$$

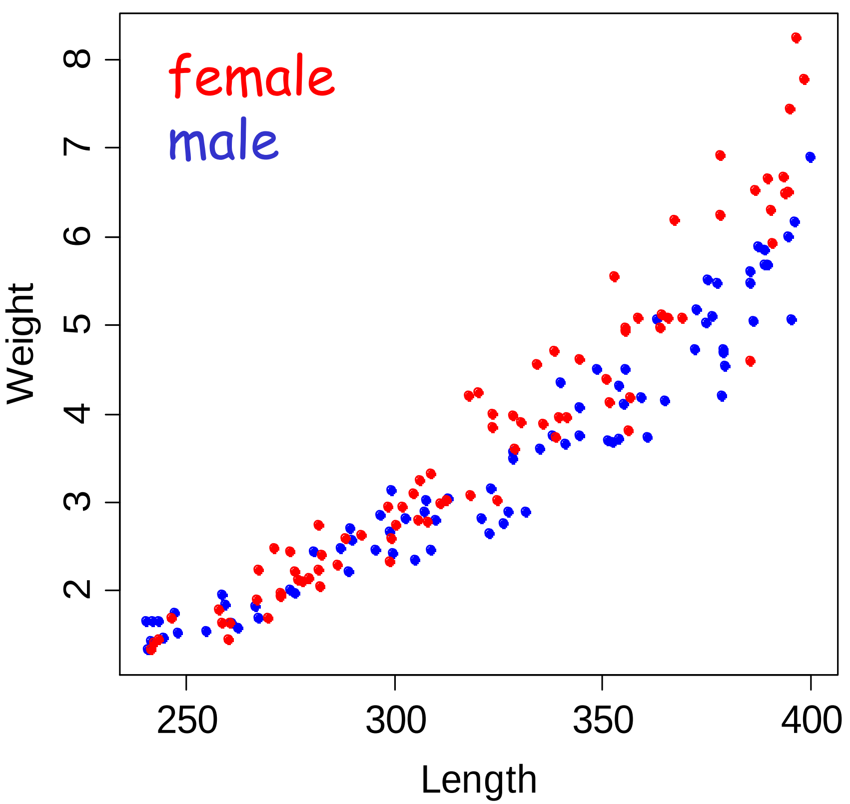
```
fit2 <- lm(log(Weight) ~ Sex + log(Length))
summary(fit2)
                  Estimate Std. Error t value Pr(>|t|)
Intercept αF  -15.93938    0.33231  -47.97   < 2e-16
SexM      αM   -0.09164    0.01700   -5.39  2.54e-07
log(Length) β   2.97707    0.05763   51.66   < 2e-16
```

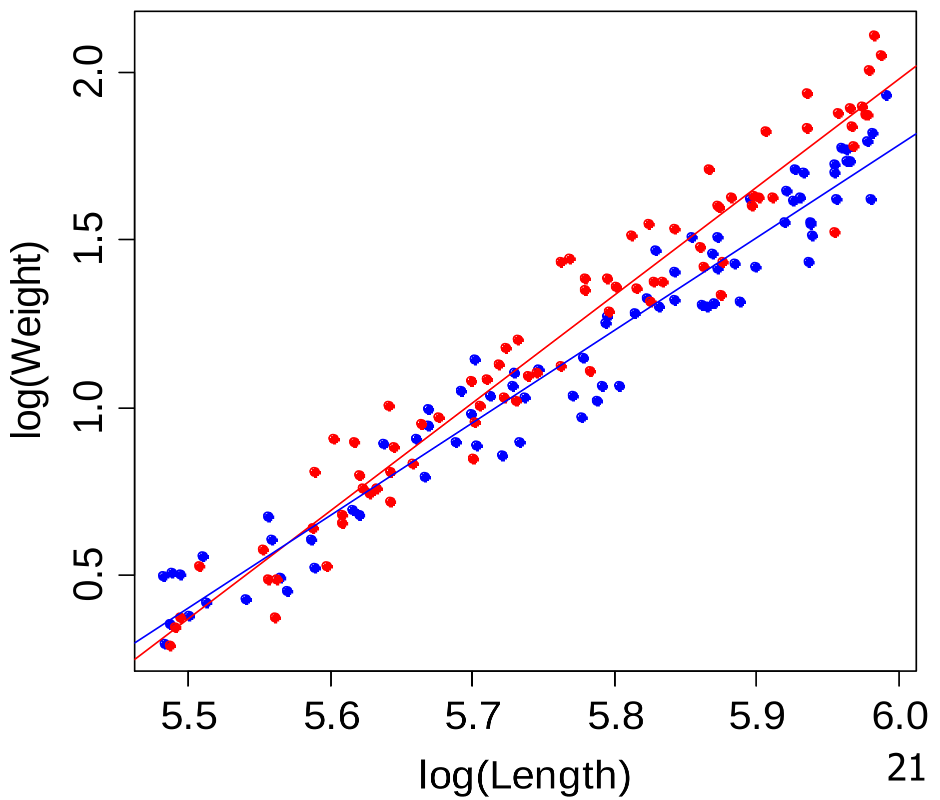
highly significant!



ANCOVA example
(with nested effect or interaction)

- Growth rate of fish by sex
  - Is there a difference in growth (slope and/or intercept)?
  - Nested effect = Interaction between numeric variable (length) and categorical variable (sex)





ANCOVA example

- Separate models by sex (Intercept, slope, and standard error by sex):

```
> Females <- lm(log(Weight) ~ log(Length), subset=Sex=="F")
> Males <- lm(log(Weight) ~ log(Length), subset=Sex=="M")
> summary(Females)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.36107    0.48805  -35.57   <2e-16
log(Length)   3.22378    0.08467   38.08   <2e-16
Residual standard error: 0.1067 on 78 degrees of freedom

> summary(Males)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -14.82832    0.41201  -35.99   <2e-16
log(Length)   2.76877    0.07133   38.82   <2e-16
Residual standard error: 0.09773 on 78 degrees of freedom
```

(Selected output only)

(→ separate residual standard errors

ANCOVA example
(with nested effect or interaction)

- Simultaneous fit, both sexes (Intercept and slope by sex, <u>same standard error</u>):

$$\boxed{logW = \alpha_F + \alpha_M \cdot D_M + \beta_S \cdot logL}$$ Model

denotes nesting

```
> Both <- lm(log(Weight) ~ Sex / log(Length))
> summary(Both)
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          -17.36107    0.46811  -37.09  < 2e-16
SexM                   2.53275    0.63644    3.98 0.000105
SexF:log(Length)       3.22378    0.08121   39.70  < 2e-16
SexM:log(Length)       2.76877    0.07465   37.09  < 2e-16
Residual standard error: 0.1025 on 156 degrees of freedom
```

Intercept for F (α_F)

α_M

β_S

(Selected output only)

combined SE

Difference between M & F intercept

Slopes for F & M

LM

ANCOVA example
(with nested effect or interaction)

FISH 604

- Simultaneous fit, both sexes (Intercept and slope by sex, <u>same standard error</u>).
- Sometimes it can be useful to re-parameterize the model (same model):

Model: $\boxed{logW = \alpha_S + \beta_S \cdot logL}$

```
> both <- lm(log(Weight) ~ Sex / log(Length) - 1)
> summary(both)
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
SexF               -17.36107    0.46811  -37.09   <2e-16
SexM               -14.82832    0.43119  -34.39   <2e-16
SexF:log(Length)     3.22378    0.08121   39.70   <2e-16
SexM:log(Length)     2.76877    0.07465   37.09   <2e-16
Residual standard error: 0.1025 on 156 degrees of freedom
```

<u>removes</u> intercept!

(Selected output only)

α_S

β_S

Intercepts for M & F

Slopes for F & M

combined SE

ANCOVA example
(with nested effect or interaction)

- Simultaneous fit, both sexes (Intercept and slope by sex, same standard error).
- Yet another version of the same model, using interaction term:

Model: $$logW = \alpha_F + \alpha_M \cdot D_M + \beta_F \cdot logL + \beta_M \cdot D_M \cdot logL$$

```
> both <- lm(log(Weight) ~ Sex / log(Length) - 1)
> summary(both)
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
SexF               -17.36107    0.46811  -37.09   <2e-16
SexM               -14.82832    0.43119  -34.39   <2e-16
SexF:log(Length)     3.22378    0.08121   39.70   <2e-16
SexM:log(Length)     2.76877    0.07465   37.09   <2e-16
Residual standard error: 0.1025 on 156 degrees of freedom
```

removes intercept!

α_S

β_S

(Selected output only)

Intercepts for M & F

Slopes for F & M

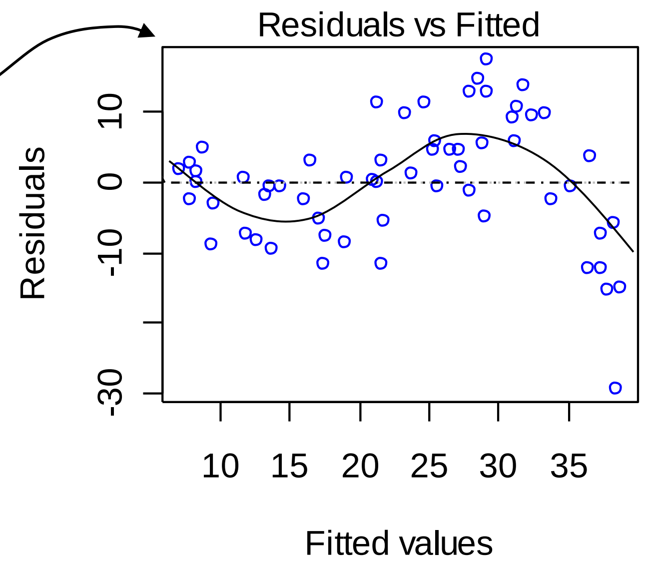
combined SE

FISH 604

Diagnostic problems (Regression)

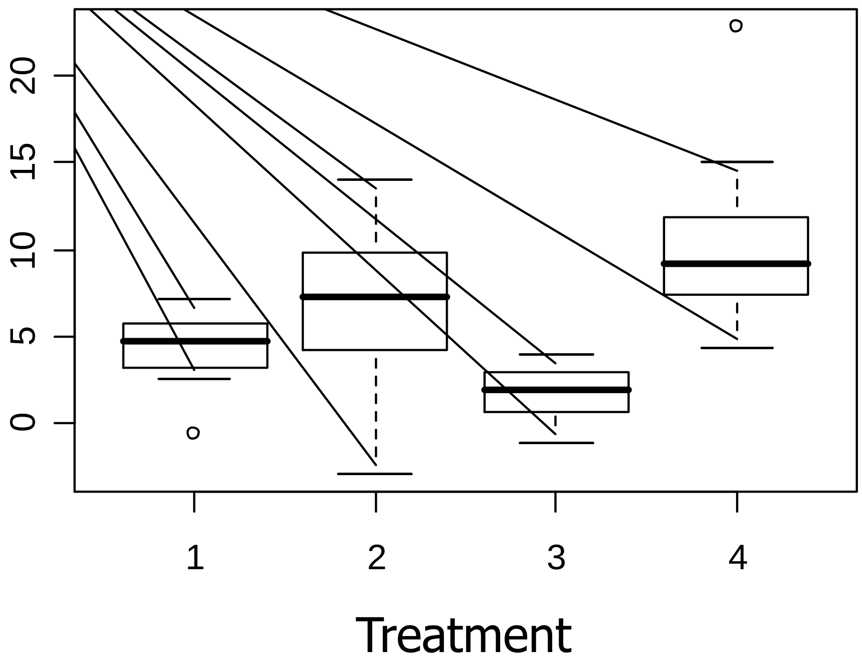
<u>What to do if assumptions are violated?</u>

- Trends / patterns in residuals
  - Typically suggests model mis-specification
  - Try a different / more complex model
  - GAM for non-linear trends (Module 7)
- Heteroscedasticity
  - Transformation (see Module 3)
  - Weighting (see Module 4)
  - Use robust methods (e.g. rank-based)
- Non-normality
  - Transformation (see Module 3)
  - Alternative error structure (GLM, Module 6)
- Outliers, observations with high leverage
  - re-fit model without observations, compare (Module 3)
  - Use robust regression (Module 3)



Diagnostic problems (ANOVA)

- Unequal variances (Heteroscedasticity)



What to do?

- Ignore!
  - ANOVA is robust, i.e. works well even with considerable heteroscedasticity
- Use non-parametric alternative:
  - One-way ANOVA:
    - Kruskal-Wallis test (rank-based)
  - Multi-way ANOVA: Convert observations to ranks, fit ANOVA (may have less power)
  - Randomization tests
- Welch's variance adjusted ANOVA: `oneway.test()` for one-way ANOVA with unequal variances
- Use weights (inverse proportional to variance)

Further reading - Linear models

- Venables & Ripley (The "Yellow Book"), Chapter 6 – simple and advanced R examples
- Jennrich R.I. (1995) An introduction to computational statistics: Regression analysis, Prentice Hall, Englewood Cliffs, NJ (QA278.2.J46) - Good, readable intro to theory
- Neter J., Wasserman W., Kutner M.H. (1990) Applied linear statistical models, Richard D. Irwin, Inc., Burr Ridge, Illinois - The "Bible": Everything you ever need to know
- Zar J.H. (1999) Biostatistical Analysis, fourth edition, Prentice - Hall, Inc., Englewood Cliffs, NJ (QH323.5.Z37) - Readable introduction, lots of practical hints, good index
- Chatterjee, S., Hadi, A.S. Price, B. (2000) Regression analysis by example, third edition. Wiley Series in Probability and Statistics (QA278.2 C5) - Readable introduction to applied regression
- Faraway, J.J. (2004). Linear models with R. Chapman & Hall/CRC - Good basic introduction to applied linear modeling