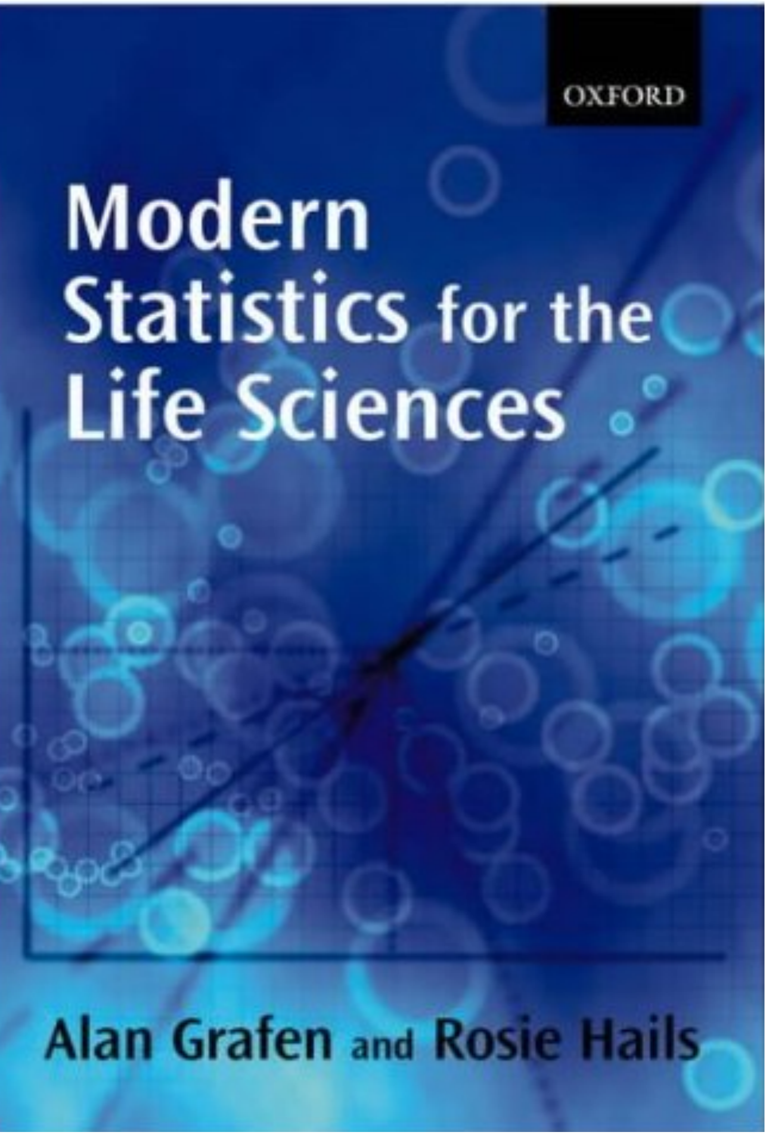
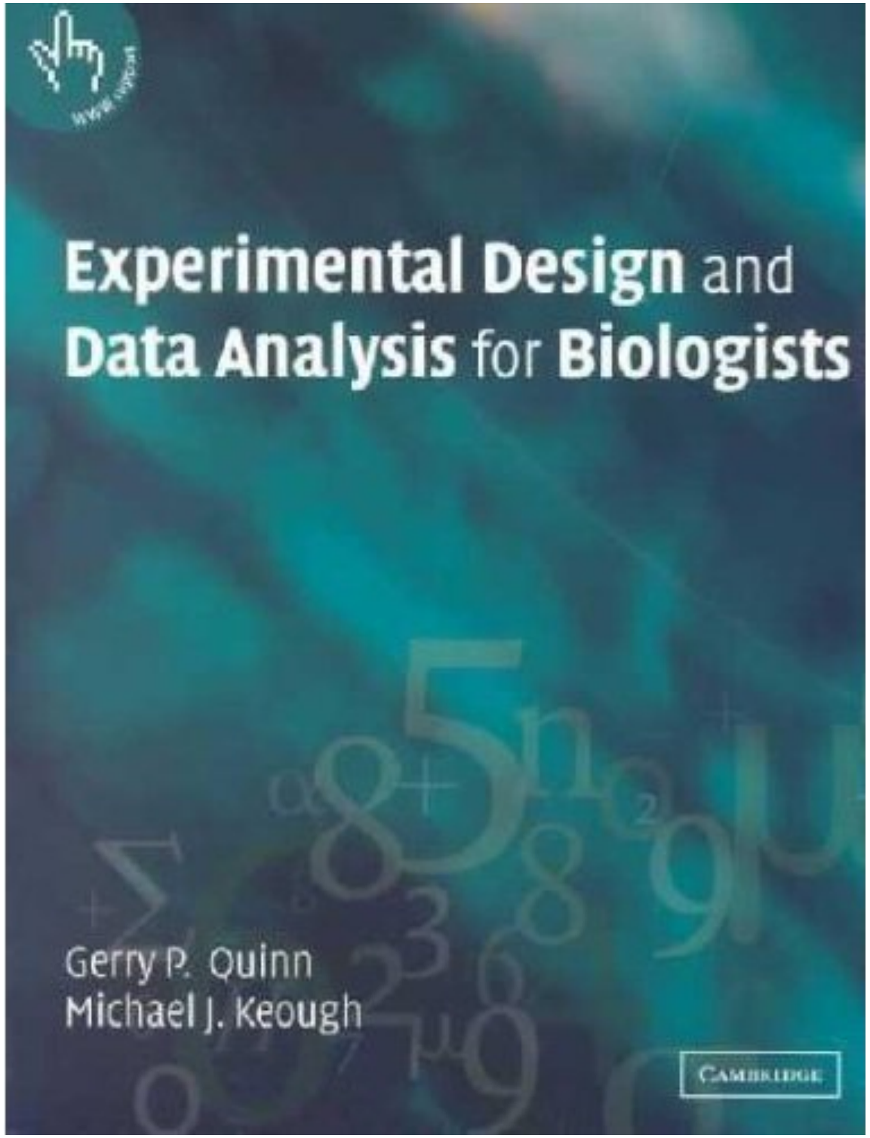
Introduction to General Linear Models
BIO782P 2017

Recap

Textbooks


OXFORD
Modern Statistics for the Life Sciences
Alan Grafen and Rosie Hails


Experimental Design and Data Analysis for Biologists
Gerry P. Quinn
Michael J. Keough
CAMBRIDGE

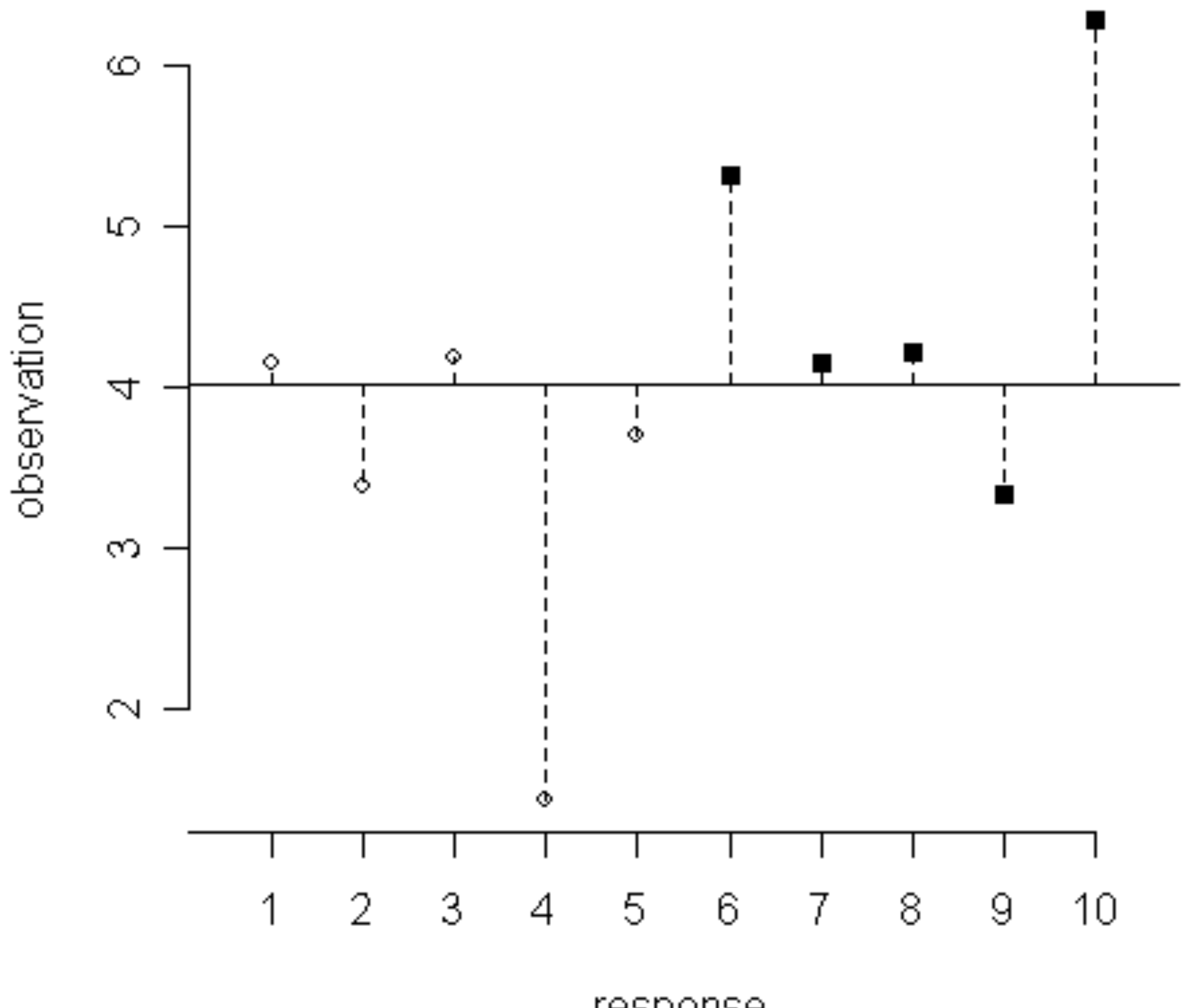
One-way ANOVA

Divide the variation in the response variable into two parts:

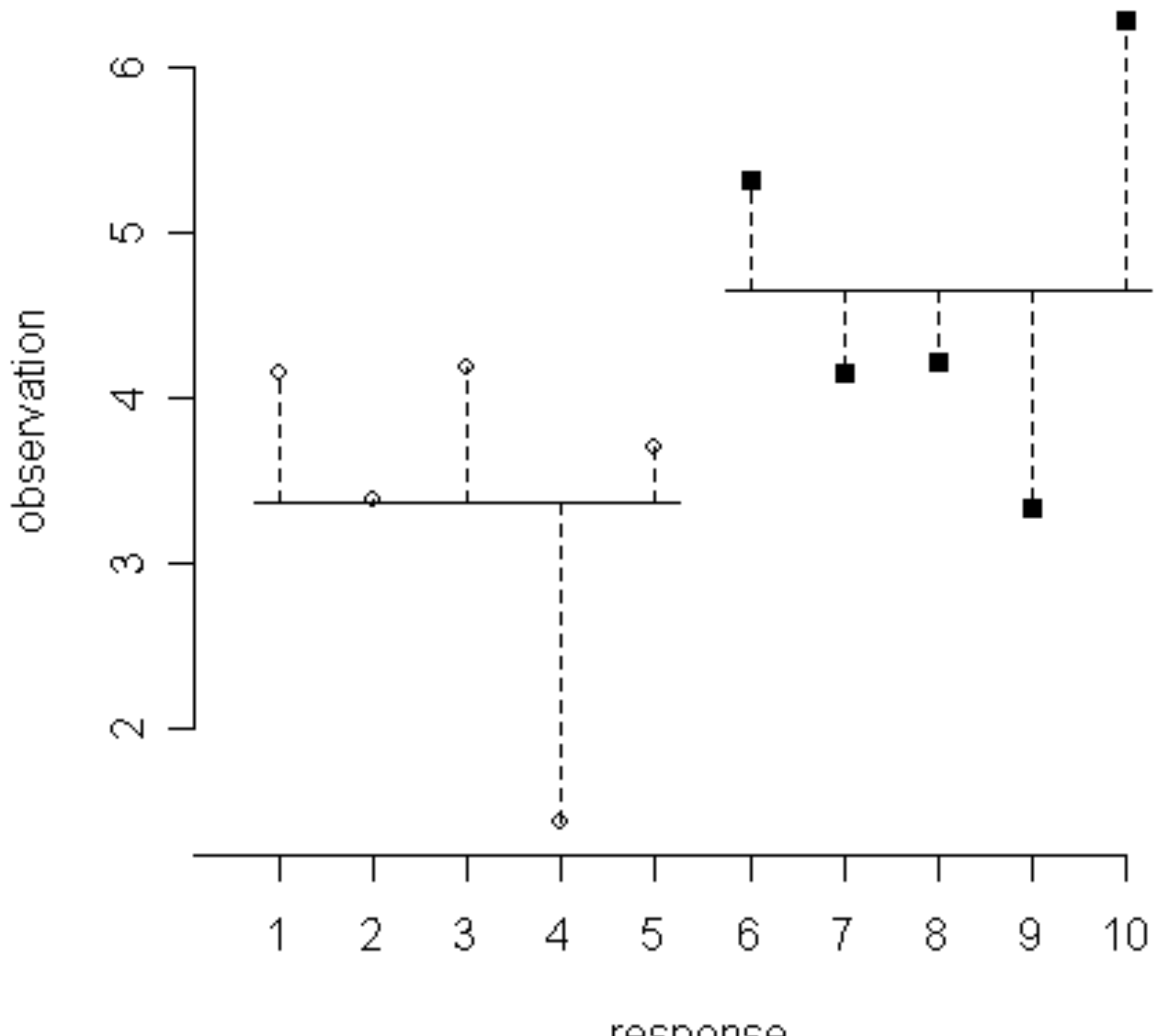
Variation between groups

Variation within groups

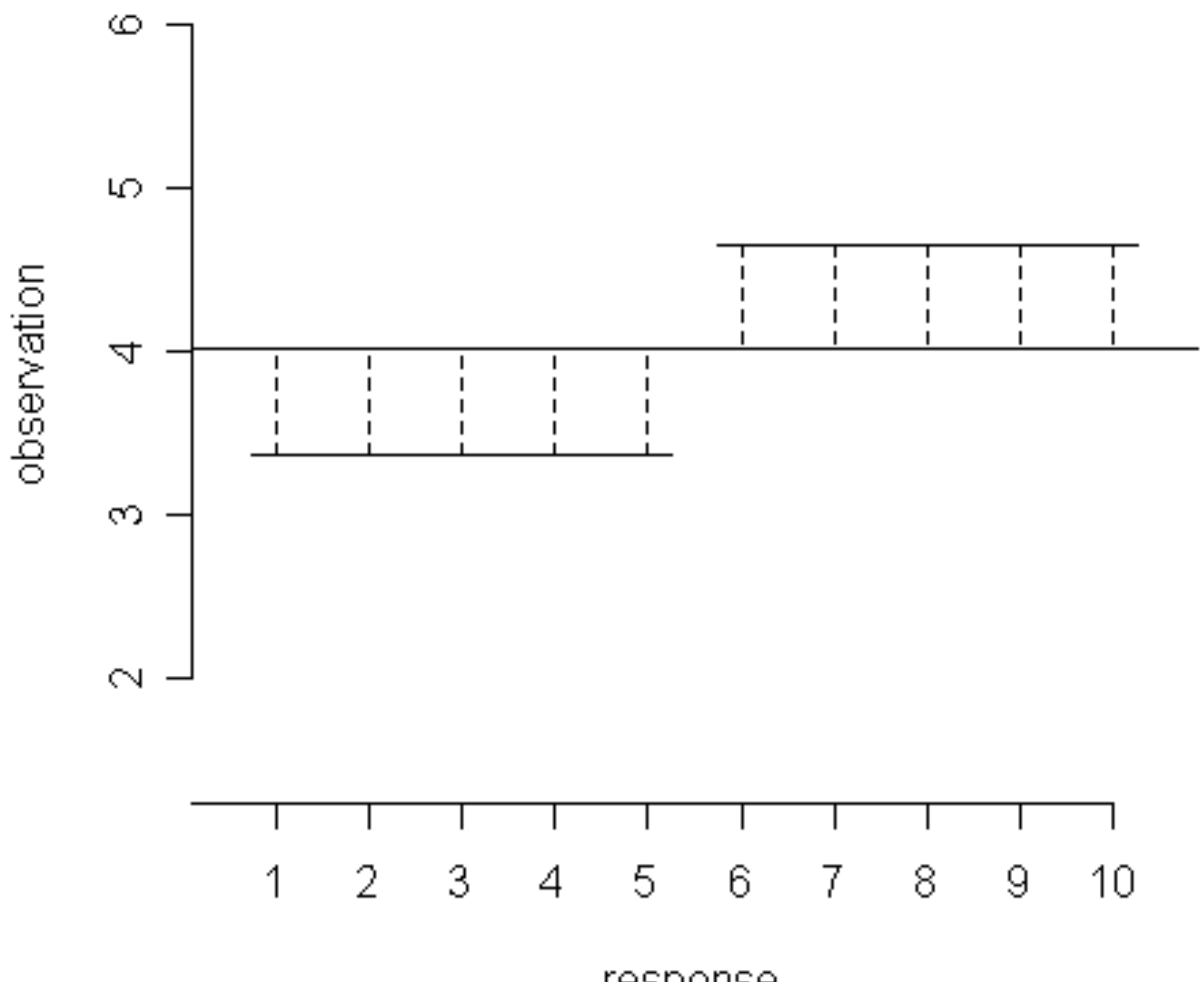
Total SS



Error SS



Treatment SS



One-way ANOVA

```
          Df Sum Sq Mean Sq F value Pr(>
atment     1  1.721   1.721   1.694  0.2
iduals     7  7.109   1.016
```

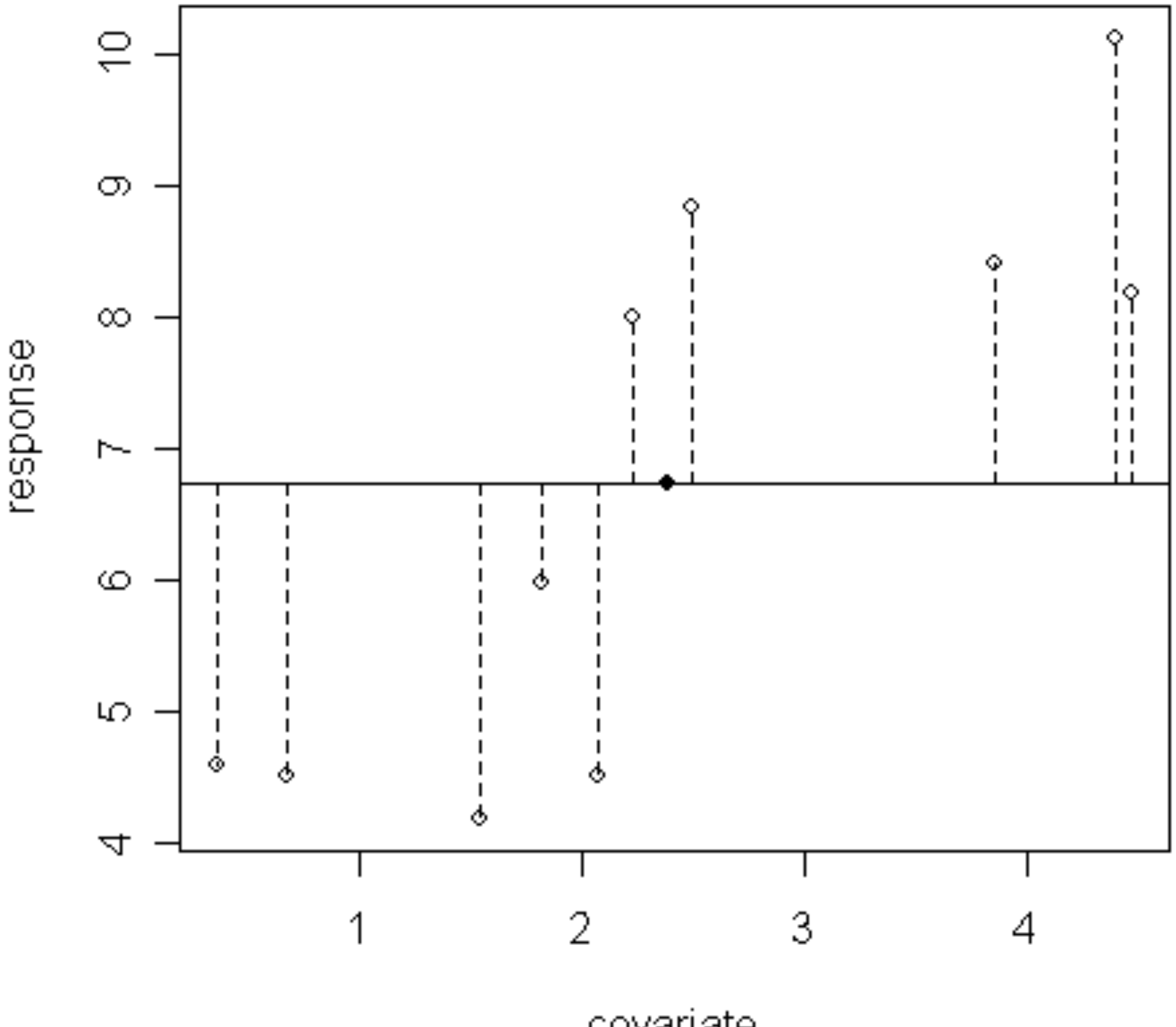
Regression

We divide the variation in the response variable into two parts:

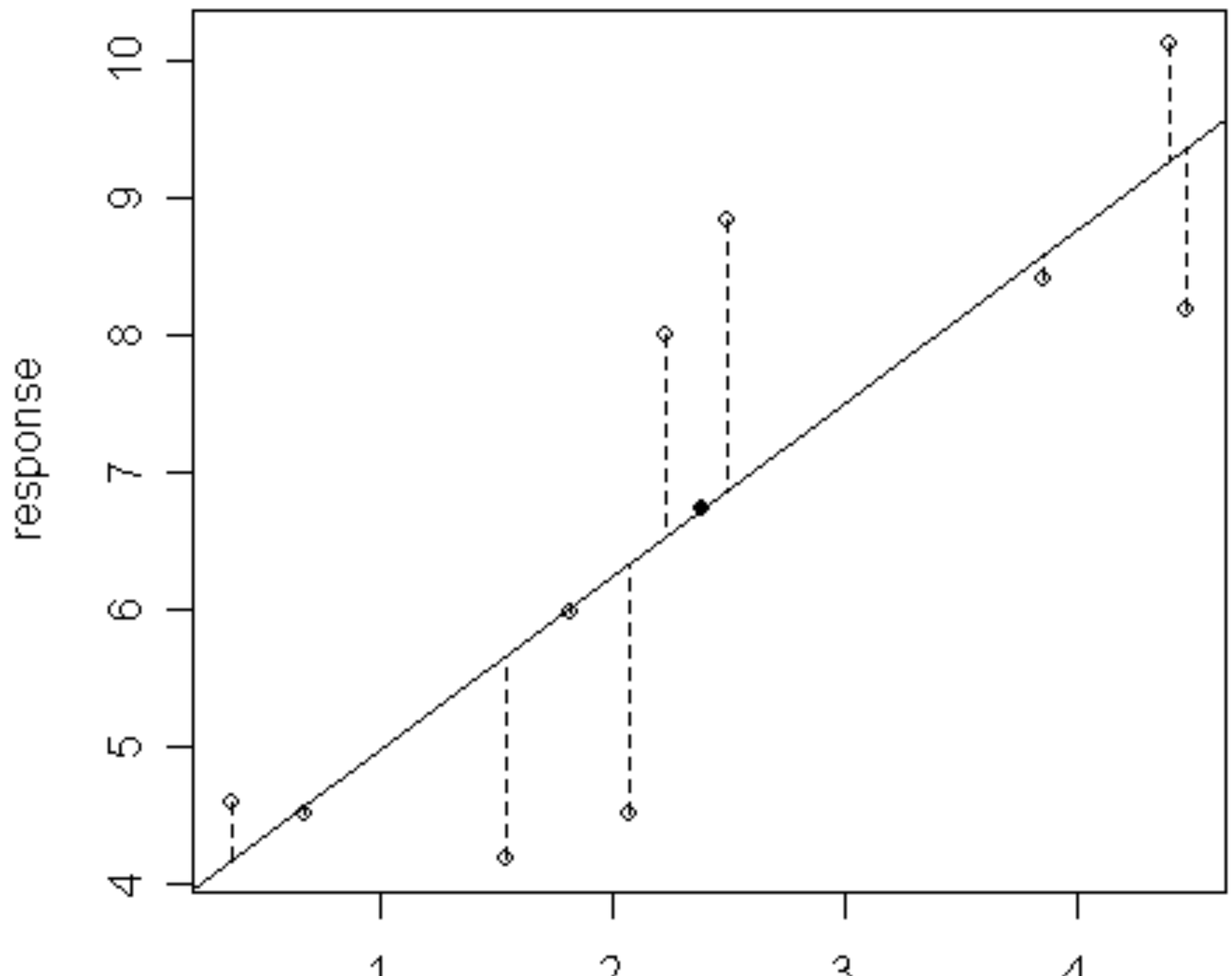
Variation of the fitted line from the mean

Variation around the fitted line

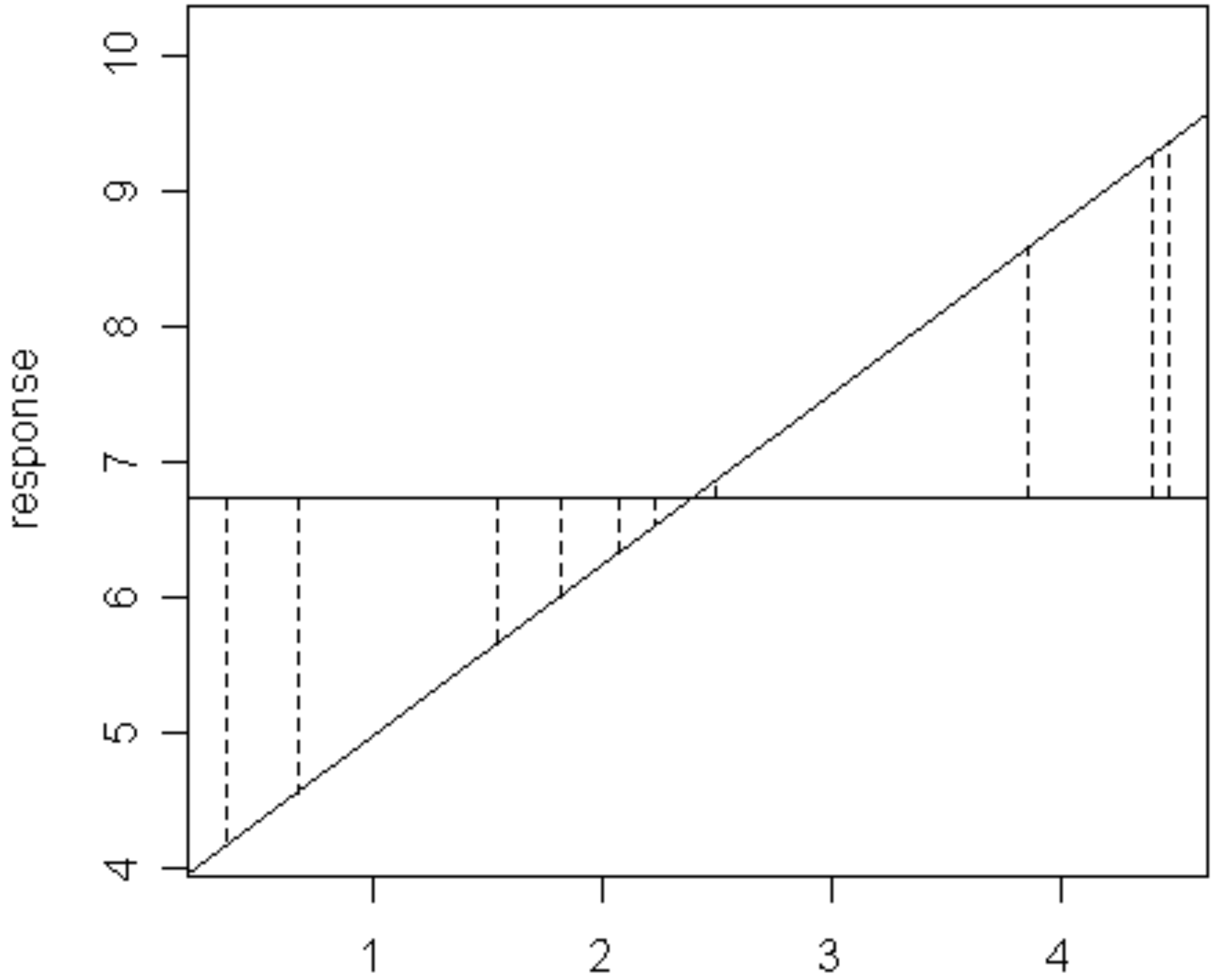
Total SS



Error SS



Effect SS



```
> anova(lm(dum2~dum1))
Analysis of Variance Table

Response: dum2
          Df Sum Sq Mean Sq F value   Pr(>F)
dum1       1 30.164  30.164  17.316 0.00316 **
Residuals  8 13.936   1.742
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’
0.05 ‘.’ 0.1 ‘ ’ 1
```

The General Linear Mode

It should now be clear that both ANOVA and regression have a very similar structure.

Both involve partitioning variance (sums of squares) into those due to the explanatory variable, and those due to error.

Both are special cases of a general linear model, which also includes much more complicated models – much is familiar. We even get an ANOVA table.

The General Linear Mode

ANOVA: for a response *weight_gain* and explanatory factor *diet*

$$Weight\ gain_{i,j} = diet_i + \varepsilon_{i,j}$$

REGRESSION: for a response *height*, with continuous explanatory variable *y*

$$height_i = \alpha + \beta.y + \varepsilon_i$$

In each case, ε is x sampled from a normal distribution with mean 0 and standard deviation *s* (estimated from the data) for every *y*

GLM model formulae

Model formulae are a way of describing how the analysis of variance is to be performed: literally, how the variance is to be divided up. You can use these formulae as input to R. Importantly, many other statistics packages use the same approach.

```
weight_gain ~ diet
height ~ rainfall
leaf.area~height+water
grazing~pH+temperature+oxygen
grazing~pH+temperature+oxygen+pH:oxygen
grazing~pH*temperature*oxygen
```

Why GLM?

The power of the GLM approach is that we can partition the variance in a response variable between any number of continuous and categorical explanatory variables e.g.

height = rainfall + altitude + terrain

rainfall and *altitude* are continuous,
terrain categorical

$$Height_{i,j,k} = rainfall_i + altitude_j + \beta.terrain_{i,j,k} + \varepsilon_{i,j,k}$$

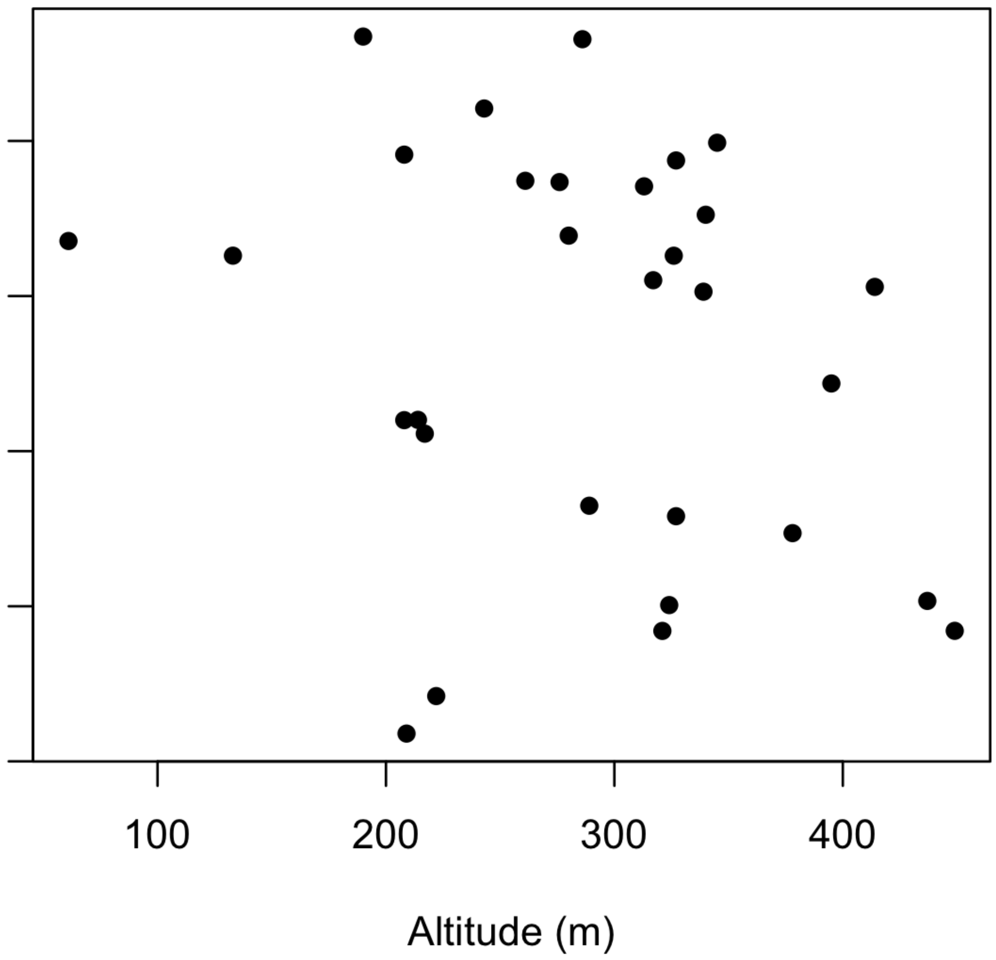
Using GLMs

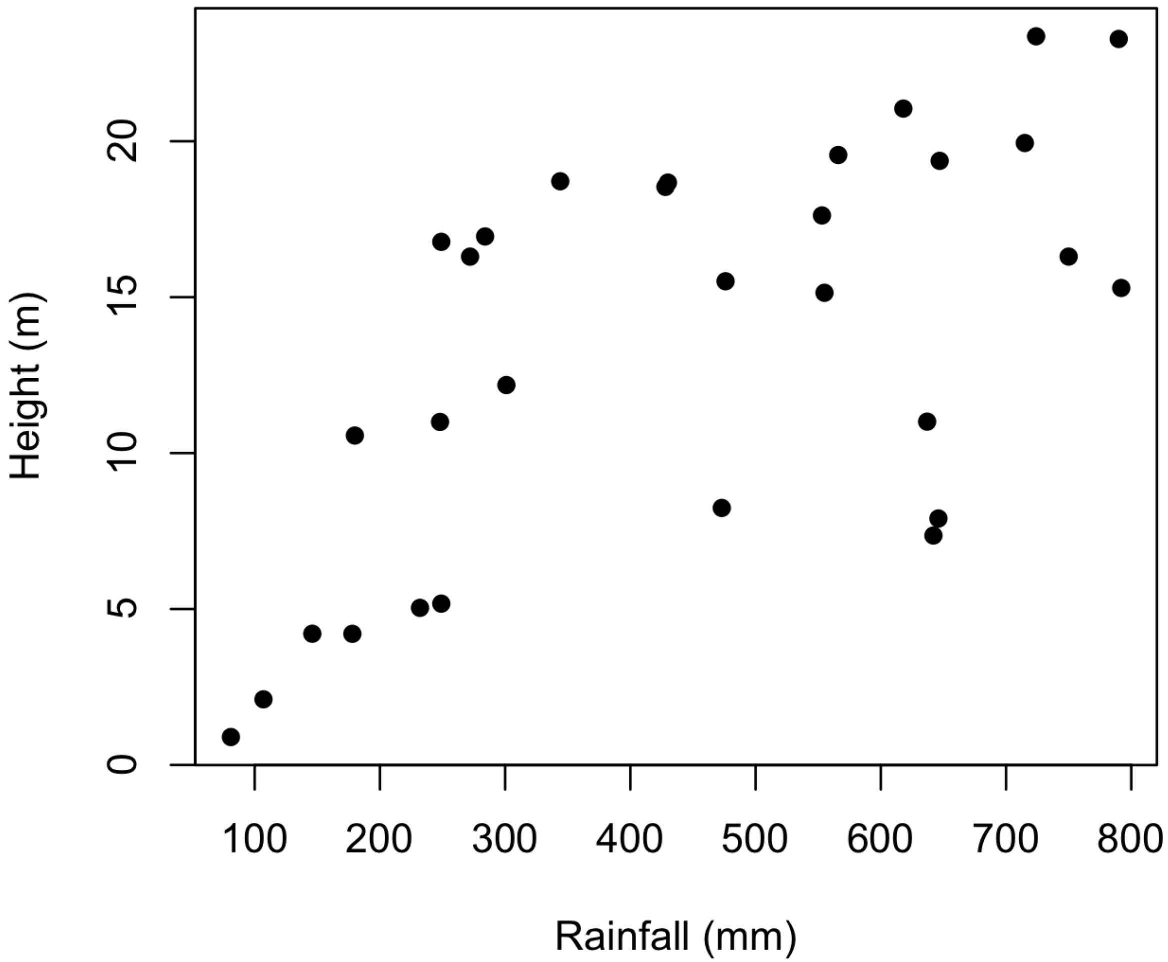
aking the level of one explanatory variable into account will change th
significance of another variable

Total SS is the same in each case,

Error SS decreases because some of the variance that was previousl
culated as error variance is explained by the second explanatory vari

Fit model for two variables





Partitioning SS

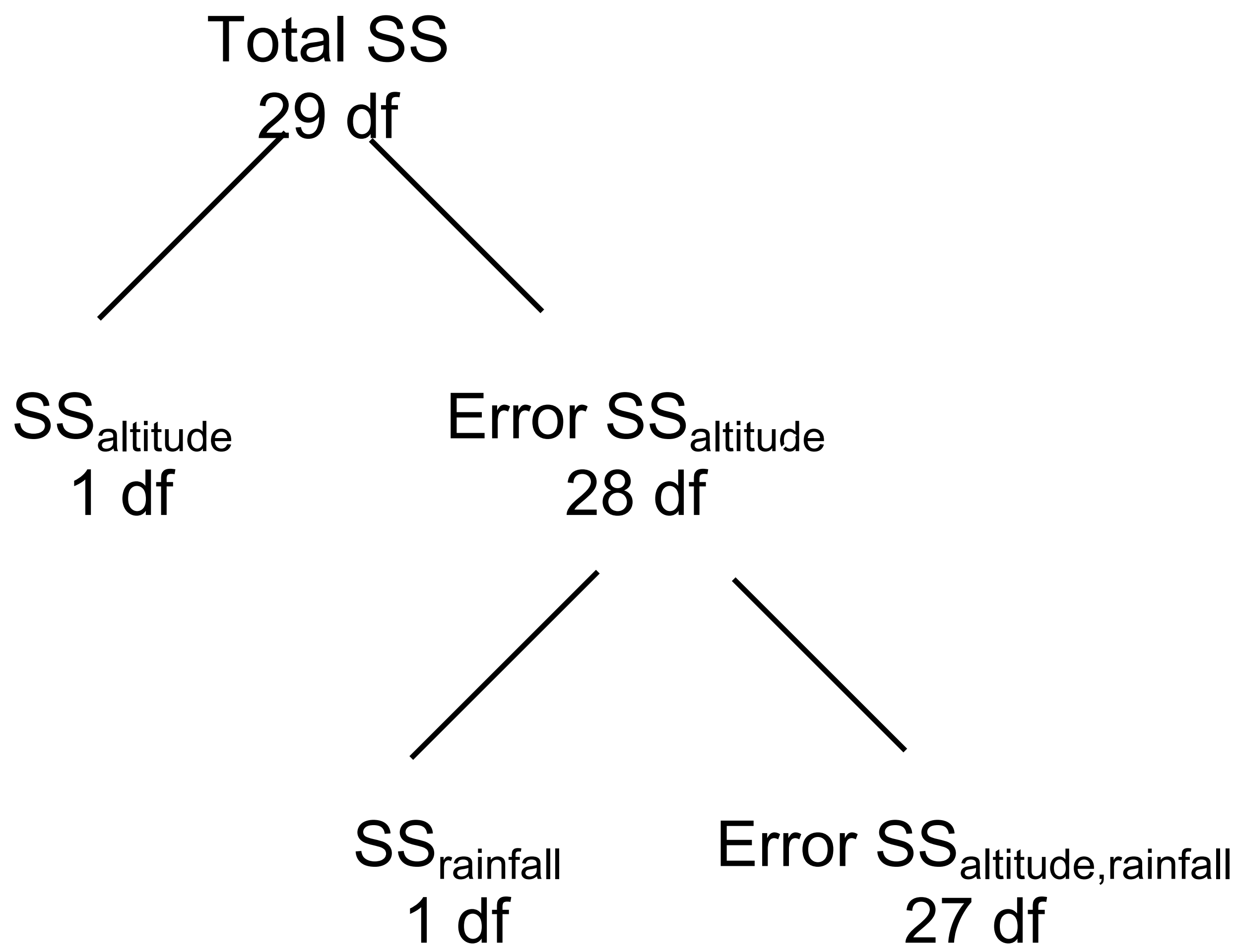
Sequential SS altitude = SS for altitude (first variable fitted)

Seq SS rainfall = $\text{errorSS}_{\text{altitude}} - \text{errorSS}_{\text{altitude+rainfall}}$

Total SS = $\text{SS}_{\text{altitude +}}\ \text{errorSS}_{\text{altitude}}$

Total SS = $\text{SS}_{\text{altitude}} + \text{SS}_{\text{rainfall}} + \text{errorSS}_{\text{altitude+rainfall}}$

Partitioning SS and df



Partitioning SS and df

```
 mod1<-lm(height~altitude+rainfall)
 anova(mod1)
nalysis of Variance Table

esponse: height
          Df Sum Sq Mean Sq F value    Pr(>F)
ltitude    1  62.60   62.60   2.536 0.1229205
ainfall    1 499.83  499.83  20.249 0.0001167 ***
esiduals  27 666.46   24.68
--
```

Partitioning SS and df

```
 mod2<-lm(height~rainfall+altitude)
 anova(mod2)
nalysis of Variance Table

esponse: height
          Df Sum Sq Mean Sq F value    Pr(>F)    
ainfall    1 489.57  489.57 19.8339 0.0001322 ***
ltitude    1  72.85   72.85  2.9514 0.0972554 .  
esiduals  27 666.46   24.68                      
--
```

Using GLMs

he values for factor SS, F and p will usually change depending on th
er by which the explanatory variable are entered into the model form

This is because they are calculated sequentially - the SS for the first
able is calculated using the raw data but the SS for the second varia
effectively) calculated on the residuals left after the effect of the first
removed

he order in which you enter your terms determines the p-values in ar
ANOVA table

Better to use a deletion test

Using GLMs

etion test: fit model with all explanatory terms, then refit the model w
the term in question removed.

mpare the goodness-of-fit (how well each model explains the data)
each model using a partial F-test.

ves an assessment of how the term in question affects how the mode
cribes the data that is independent of the order that it's entered into
model.

In R can either fit models separately and compare using
anova(model1,model2) or use drop1(model1,test="F").

Using GLMs

```
> drop1(mod1, test="F")
Single term deletions

Model:
height ~ altitude + rainfall
         Df Sum of Sq     RSS     AIC F value    Pr(>F)
<none>                 666.46  99.024
altitude  1     72.85  739.31 100.136  2.9514 0.0972554 .
rainfall  1    499.83 1166.29 113.812 20.2492 0.0001167 ***
---

> drop1(mod2, test="F")
Single term deletions

Model:
height ~ rainfall + altitude
         Df Sum of Sq     RSS     AIC F value    Pr(>F)
<none>                 666.46  99.024
rainfall  1    499.83 1166.29 113.812 20.2492 0.0001167 ***
altitude  1     72.85  739.31 100.136  2.9514 0.0972554
```

Orthogonality

SS and Seq SS will be identical if the information about the response given by
xplanatory variables is independent. In this case, the two explanatory variables a
termed *orthogonal*.

The question to ask yourself is:

Does knowing something about one explanatory variable tell you anything about the level of a second explanatory variable?

This is 'yes':

for two categorical variables if there are unequal numbers of samples for different leve

For two continuous variables if r^2 is not 0 (i.e. Always)

Effect sizes

So far, we've looked at the ANOVA table, which tells us about the significance of the effects we test in a GLM. It doesn't tell us anything about the magnitude of any effects. For this we need to look at the table of coefficients produced by summary()

Effect sizes

```
> summary(mod1)

Call:
lm(formula = height ~ altitude + rainfall)

Residuals:
   Min     1Q Median     3Q    Max 
-8.579 -3.742  1.403  3.303  6.780 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 10.345413   3.622998   2.855 0.008163 ** 
altitude    -0.018103   0.010537  -1.718 0.097255 .  
rainfall     0.018663   0.004148   4.500 0.000117 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.968 on 27 degrees of freedom
Multiple R-squared:  0.4577,	Adjusted R-squared:  0.4175 
F-statistic: 11.39 on 2 and 27 DF,  p-value: 0.0002586
```

Effect sizes

Fitted model:

ght = 10.34 - 0.0181 x altitude + 0.0187 x rainfall + e

ANCOVA type models

Factor: altitude (Low vs High)
Continuous explanatory variable: rainfall

Response variable: height (tree height)

ANCOVA type models

```
> mod1<-lm(height~altitude*rainfall)

> drop1(mod1, test="F")
Single term deletions

Model:
height ~ altitude * rainfall
                  Df Sum of Sq    RSS    AIC F value   Pr(>F)
<none>                         582.81 115.16
altitude:rainfall  1     198.7 781.51 124.89  12.274 0.001247 **
```

ANCOVA type models

```
> summary(mod1)

Call:
lm(formula = height ~ altitude * rainfall)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.0797 -2.4121  0.1078  1.5136 10.3990 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)   
(Intercept)          6.723725   2.176466   3.089  0.00385 **
altitudeLow         -5.554121   2.803289  -1.981  0.05524 . 
rainfall            -0.002779   0.004750  -0.585  0.56219   
altitudeLow:rainfall  0.022706   0.006481   3.503  0.00125 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.024 on 36 degrees of freedom
Multiple R-squared:  0.4055,	Adjusted R-squared:  0.3559
```

ANCOVA type models

Fitted model:

for high altitude: height = 6.72 - 0.00278 x rainfall +e

for low altitude: height = 1.170 + 0.0199 x rainfall + e

Summary

- GLM is a *family* of models
- Linear regression and ANOVA can be thought of as special cases of GLMs
- GLM lets us mix and match any number of factors and variables
- Total variance and d.f. are shared amongst all model terms
- <u>Order of terms in the model affects their power</u>

- Compare model combinations to refine them – we are looking for the <u>minimum adequate model.</u>