RStatsbook

Andreas Busjahn

2023-03-07

Table of contents

Preface

This is a Quarto book.

To learn more about Quarto books visit https://quarto.org/docs/books.

```
1 + 1
```

```
[1] 2
```

1 Introduction

This is a book created from markdown and executable code.

See Knuth (1984) for additional discussion of literate programming.

```
1 + 1
```

```
[1] 2
```

2 Intro to lm

In this script, linear models (including linear regression and ANOVA) will be introduced. Output is not optimized for word, but rather for interactive use.

2.1 Setup

All packages necessary will be invoked by p_load. Packages with only a single function call or potential for name conflicts can be unload, this way we still checked for their existence and installed them if need be.

```
pacman::p_load(conflicted,wrappedtools,car,nlme,broom,
               multcomp,tidyverse,foreign,DescTools, ez,
               ggbeeswarm,
               lme4, nlme,merTools,
               easystats, patchwork,here)#conflicted,
# rayshader,av)
# pacman::p_unload(DescTools, foreign)
# conflict_scout()
conflicts_prefer(dplyr::select,
                 dplyr::filter,
                 modelbased::standardize)
```

```
[conflicted] Will prefer dplyr::select over any other package.
[conflicted] Will prefer dplyr::filter over any other package.
[conflicted] Will prefer modelbased::standardize over any other package.
```

```
base_dir <- here::here()
```

2.2 Import / Preparation

Data are read from an SPSS file. Numeric column Passage is mutated into a factor as Passage_F, this is necessary for group comparisons in ANOVA. The call to here() expands the path to a file from the project directory to the full system path.

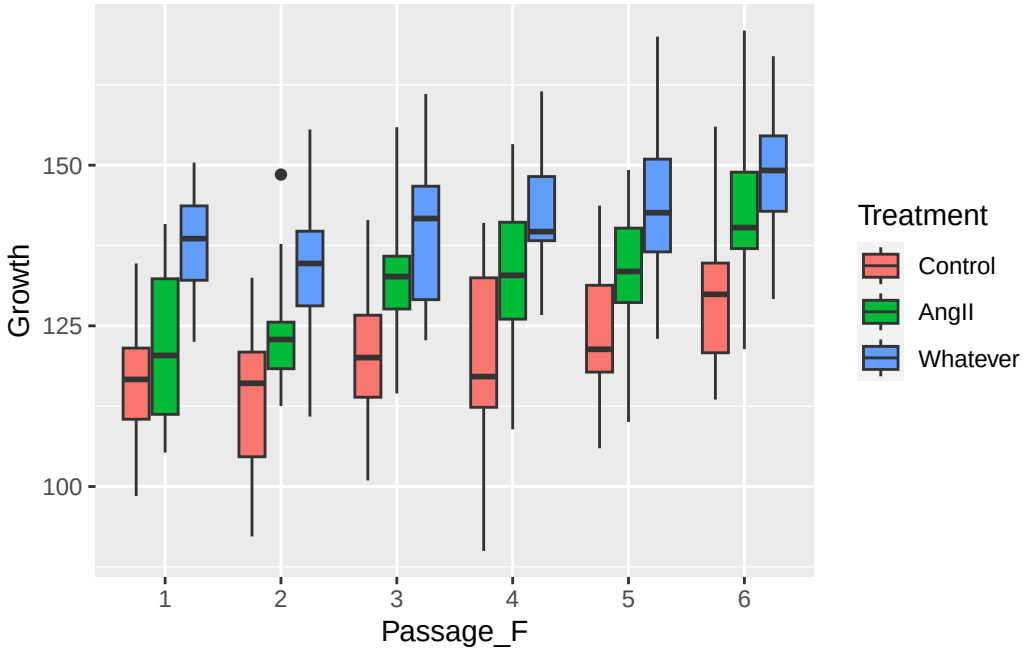
```
rawdata<-foreign::read.spss(file=here('Data/Zellbeads.sav'),
                            use.value.labels=T,to.data.frame=T) %>%
  as_tibble() %>%
  dplyr::select(-ZahlZellen) |>
  rename(Growth=Wachstum,Treatment=Bedingung) |>
  mutate(Passage_F=factor(Passage),
         Treatment=fct_recode(Treatment,
                              Control="Kontrolle"))
```

```
re-encoding from CP1252
```

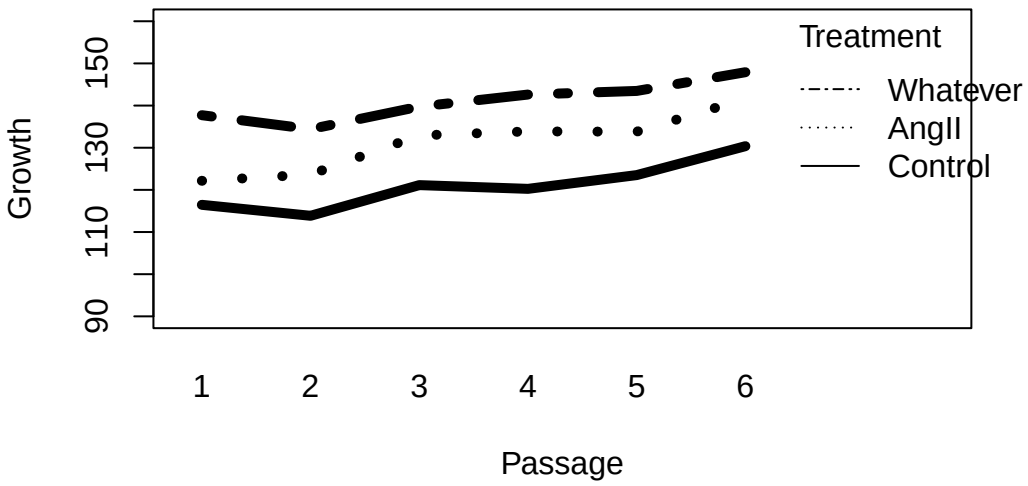
2.3 Graphical exploration

First impression of the data will be attempted by grouped boxplot, followed by interaction plots, both as basic and ggplot with variations.

```
ggplot(rawdata,aes(Passage_F,Growth, fill=Treatment))+
  geom_boxplot(coef=3)
```

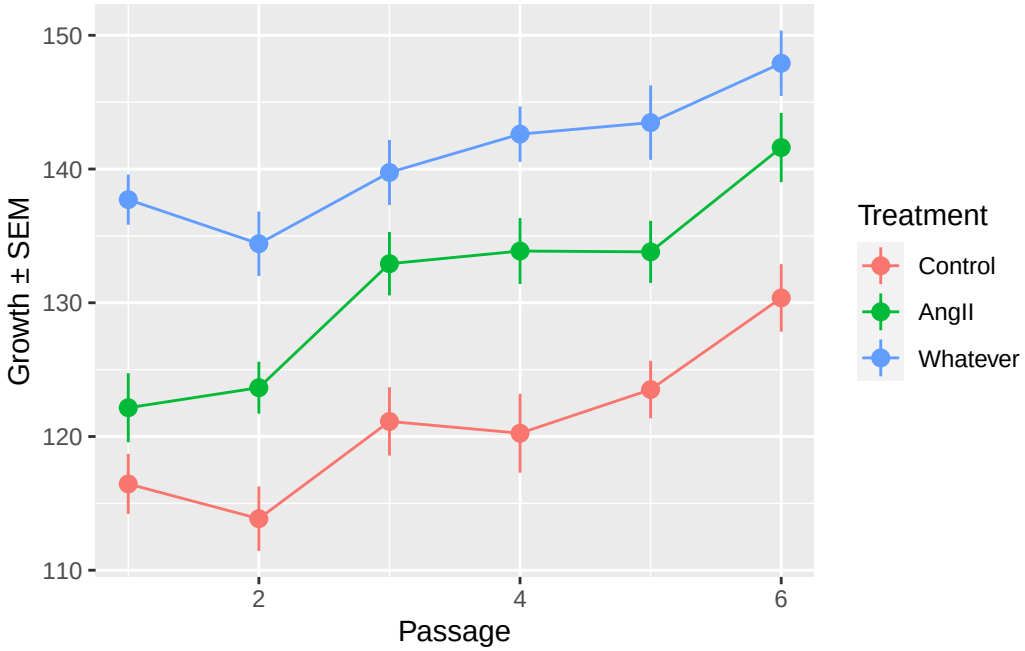



```
with(rawdata, interaction.plot(
  x.factor=Passage, trace.factor=Treatment, response=Growth,
  ylim = c(90, 160), lty = c(1,3,12), lwd = 5,
  ylab = "Growth", xlab = "Passage",
  trace.label = "Treatment"))
```



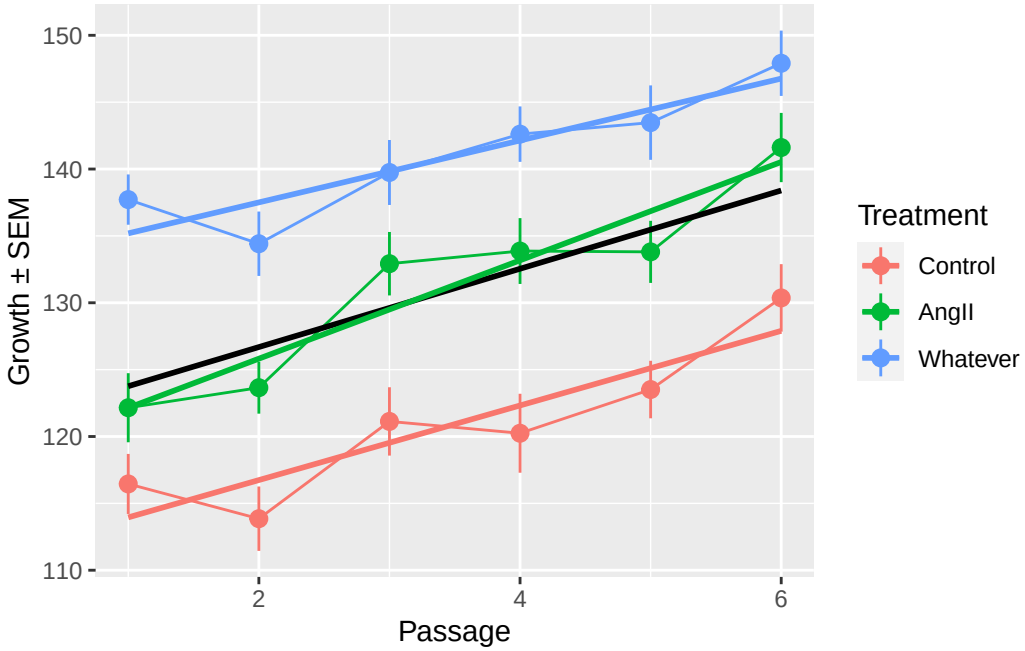

```
# p1<-ggplot(rawdata,aes(x=Passage,y=Growth))+
#    stat_summary(geom='line',fun='mean',aes(color=Treatment))+
#    stat_summary(geom='line',fun='mean')
p1<-ggplot(rawdata,aes(x=Passage,y=Growth))+
  stat_summary(geom='line',fun='mean',aes(color=Treatment))+
  stat_summary(aes(color=Treatment))+
  ylab('Growth \u00b1 SEM')
p1
```

```
No summary function supplied, defaulting to `mean_se()`
```



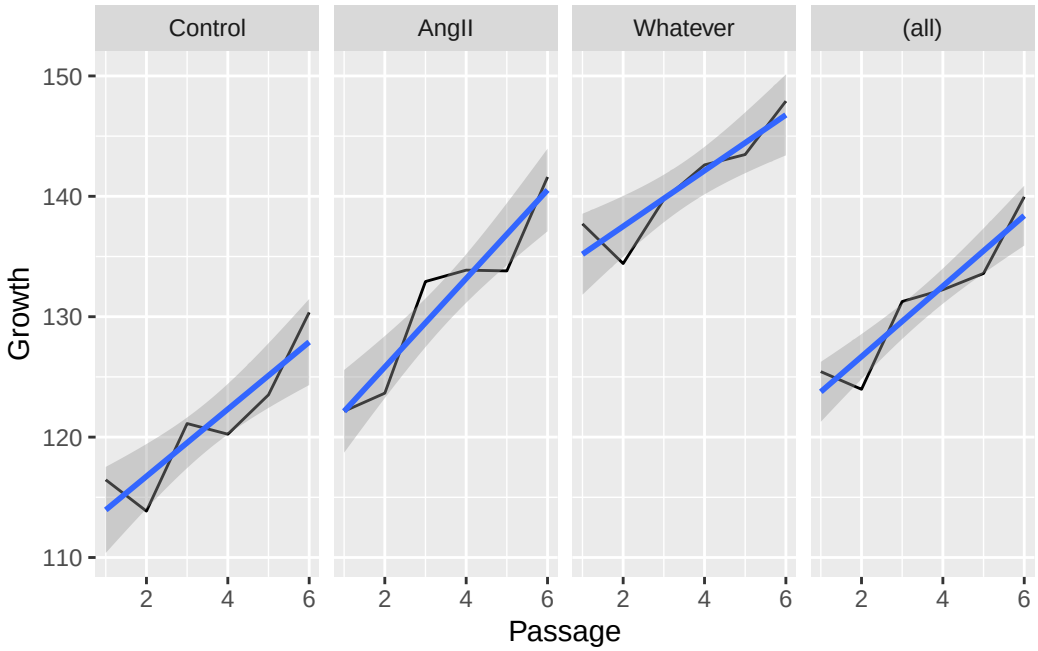

```
p1+geom_smooth(method='lm',color='black',se=F)+
  geom_smooth(method='lm',aes(color=Treatment),se=F)
```

```
No summary function supplied, defaulting to `mean_se()`
`geom_smooth()` using formula = 'y ~ x'
`geom_smooth()` using formula = 'y ~ x'
```




```
ggplot(rawdata,aes(x=Passage,y=Growth))+
  stat_summary(geom='line',fun='mean')+
  geom_smooth(method='lm')+
  facet_grid(cols = vars(Treatment), margins=T)
```

```
`geom_smooth()` using formula = 'y ~ x'
```



2.4 Linear Models

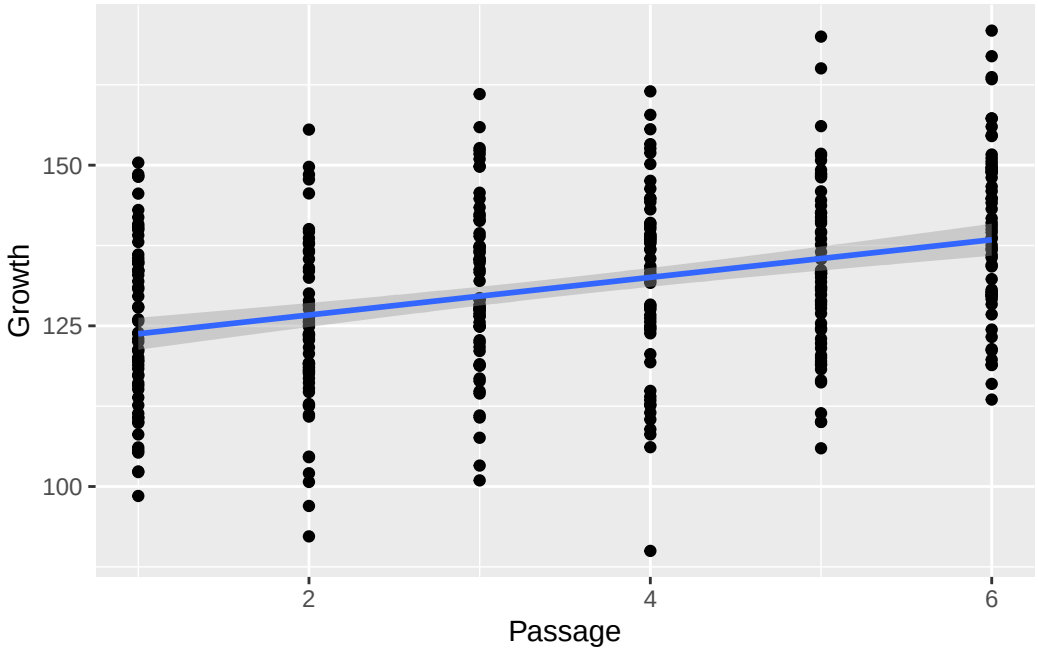
2.4.1 Linear regression

We will analyse the relation between indepemdent variable (IV) Passage and dependent variable (DV) Growth.

2.4.1.1 Graphical exploration

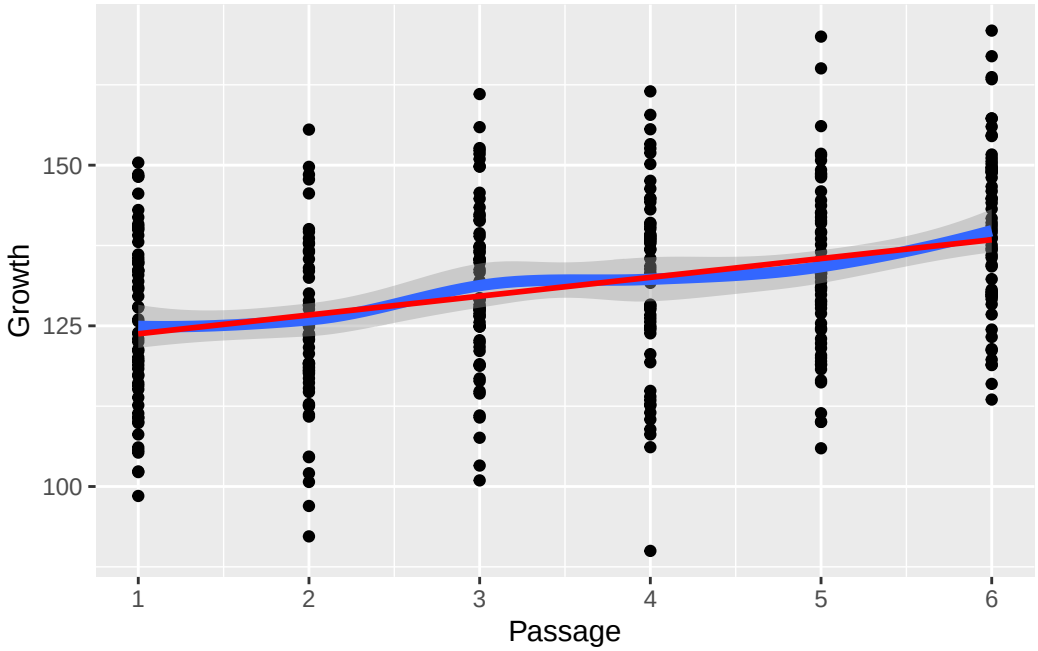
```
ggplot(rawdata,aes(Passage,Growth))+
  geom_point()+
  geom_smooth(method=lm)
```

```
`geom_smooth()` using formula = 'y ~ x'
```




```
ggplot(rawdata,aes(Passage,Growth))+
  geom_point()+
  scale_x_continuous(breaks=seq(0,10,1))+
  geom_smooth(linewidth=2)+
  geom_smooth(method=lm,se=F,color='red')
```

```
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
`geom_smooth()` using formula = 'y ~ x'
```



2.4.1.2 Modelling

This takes 2 steps, building the model and computing p-values.

```
# model
(regressionOut<-lm(Growth~Passage,data=rawdata))
```

```
Call:
lm(formula = Growth ~ Passage, data = rawdata)

Coefficients:
(Intercept)      Passage
    120.834        2.927
```

```
# model and p.value for slope, not recommended
tidy(regressionOut)
```

```
# A tibble: 2 x 5
```

```
  term        estimate std.error statistic   p.value
  <chr>          <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)   121.       1.63      74.2  1.77e-219
2 Passage         2.93     0.418      7.00 1.26e- 11
```

```
# computation of SSQs and p-values, use this!
(anova_out<-anova(regressionOut))
```

```
Analysis of Variance Table

Response: Growth
           Df Sum Sq Mean Sq F value    Pr(>F)
Passage     1   8996  8996.2  49.022 1.257e-11 ***
Residuals 358  65698   183.5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova_out$`Pr(>F)` #|> na.omit()
```

```
[1] 1.257266e-11           NA
```

```
tidy(anova_out)
```

```
# A tibble: 2 x 6
  term         df  sumsq meansq statistic   p.value
  <chr>     <int>  <dbl>  <dbl>     <dbl>     <dbl>
1 Passage       1  8996.  8996.      49.0  1.26e-11
2 Residuals   358 65698.   184.      NA   NA
```

```
# summary(regressionOut)
# str(regressionOut)
```

2.4.1.3 Adjusting

To take out the variance due to Passage effects, we can use the residuals and shift them to the original mean:

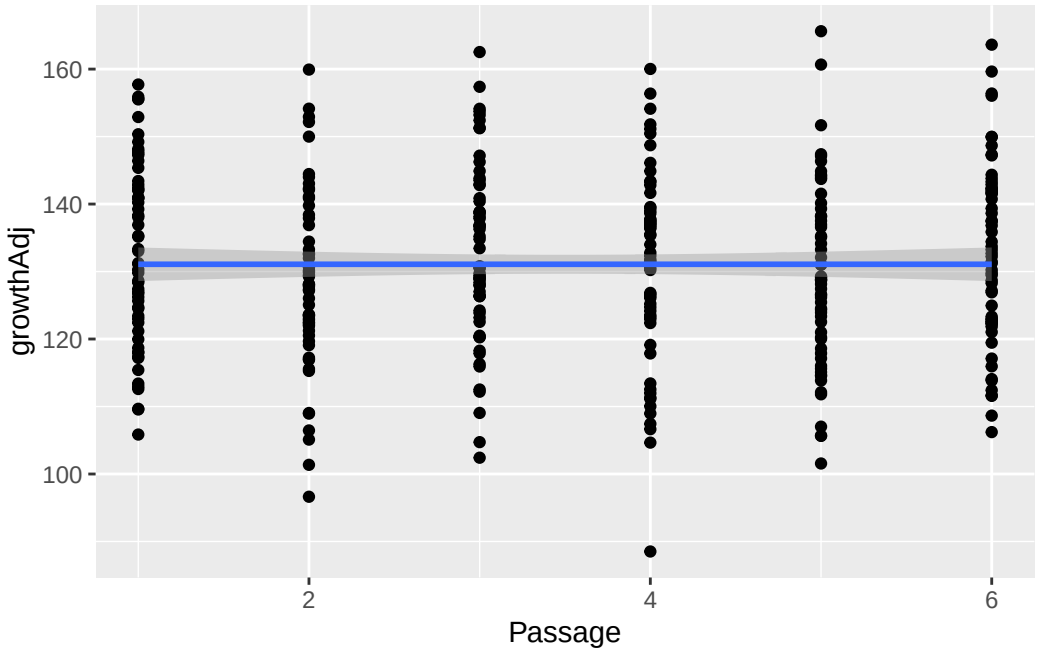
```
  rawdata <-
    mutate(rawdata,
           growthAdj = regressionOut$residuals+mean(Growth))

  summarise(rawdata,
            across(c('Growth','growthAdj'),
                   ~meansd(.x,roundDig =4)))
```

```
# A tibble: 1 x 2
  Growth       growthAdj
  <chr>        <chr>
1 131.1 ± 14.4 131.1 ± 13.5
```

```
  ggplot(rawdata,aes(Passage,growthAdj))+
    geom_point()+
    geom_smooth(method = 'lm')
```

```
`geom_smooth()` using formula = 'y ~ x'
```



```
lm(growthAdj~Passage,data=rawdata) |> tidy()
```

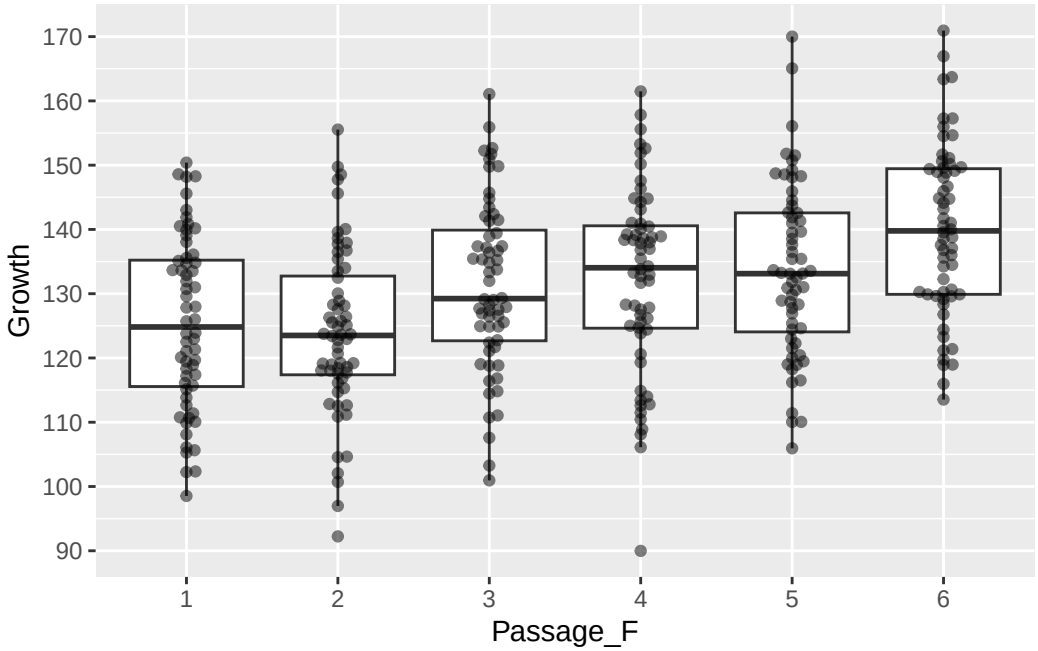
```
# A tibble: 2 x 5
  term        estimate std.error statistic   p.value
  <chr>          <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept) 1.31e+ 2     1.63   8.05e+ 1 2.04e-231
2 Passage     6.80e-15     0.418  1.63e-14 1.00e+  0
```

2.4.2 ANOVA

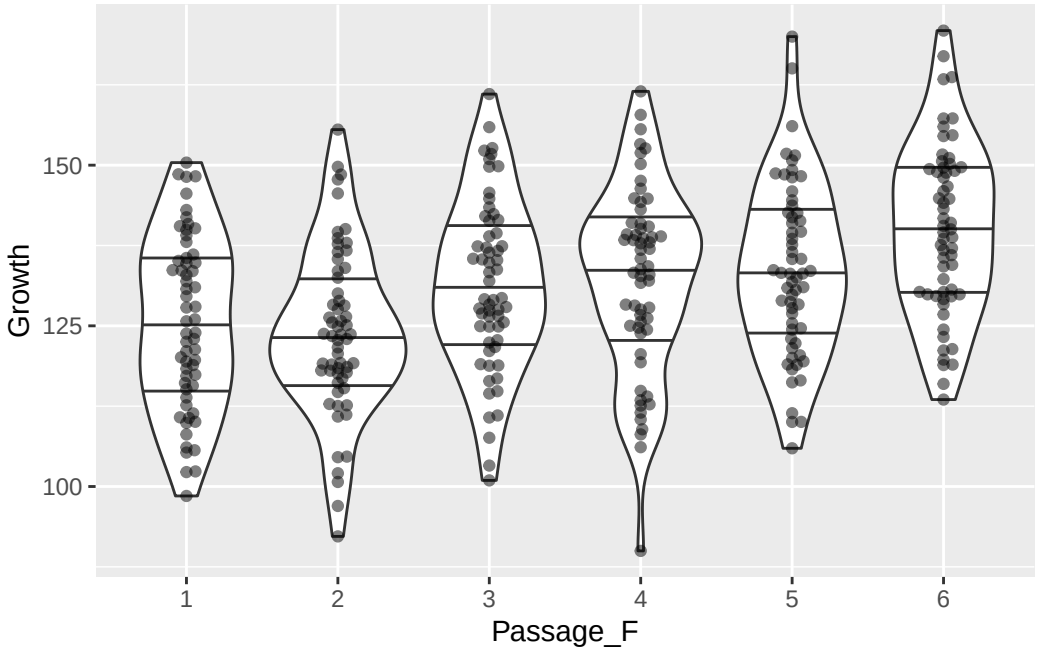
In the linear regression, we had Passage as a continuous IV, estimating a global 'universal' effect supposed to be constant. Now we look at Passege_F and model a discrete IV, allowing for specific effects, and thereby comparing means between groups.

2.4.2.1 Graphical exploration

```
ggplot(rawdata,aes(x = Passage_F, y = Growth))+
  geom_boxplot(outlier.alpha = 0)+
  geom_beeswarm(alpha=.5)+
  scale_y_continuous(breaks=seq(0,1000,10))
```




```
ggplot(rawdata,aes(x = Passage_F, y = Growth))+
  geom_violin(draw_quantiles = c(.25,.5,.75))+
  geom_beeswarm(alpha=.5)
```



2.4.2.2 Modelling

```
(AnovaOut<-lm(Growth~Passage_F,data=rawdata))
```

```
Call:
lm(formula = Growth ~ Passage_F, data = rawdata)

Coefficients:
(Intercept)   Passage_F2   Passage_F3   Passage_F4   Passage_F5   Passage_F6
    125.440       -1.467        5.824        6.801        8.156       14.520
```

```
tidy(AnovaOut)
```

```
# A tibble: 6 x 5
  term        estimate std.error statistic   p.value
  <chr>          <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)   125.        1.74    71.9   2.20e-213
2 Passage_F2     -1.47      2.47    -0.595 5.52e-  1
```

```
3 Passage_F3      5.82      2.47     2.36  1.87e-  2
4 Passage_F4      6.80      2.47     2.76  6.11e-  3
5 Passage_F5      8.16      2.47     3.31  1.04e-  3
6 Passage_F6     14.5       2.47     5.89  9.03e-  9
```

```
# summary(AnovaOut)
(t <- anova(AnovaOut))
```

```
Analysis of Variance Table

Response: Growth
           Df Sum Sq Mean Sq F value    Pr(>F)
Passage_F   5  10134 2026.71  11.113 5.852e-10 ***
Residuals 354  64561  182.38
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
t$`Pr(>F)`
```

```
[1] 5.851856e-10           NA
```

```
tidy(t)
```

```
# A tibble: 2 x 6
  term         df  sumsq meansq statistic   p.value
  <chr>     <int>  <dbl>  <dbl>     <dbl>     <dbl>
1 Passage_F     5 10134.  2027.      11.1  5.85e-10
2 Residuals   354 64561.   182.      NA   NA
```

2.4.2.3 Post-hoc analyses

The p-value from our model only tests the global Null hypothesis of no differences between any group (all means are the same / all groups come from the same population). Post-hoc tests are used to figure out which groups are different. Those tests need to take multiple testing into account. Try to limit selection of tests!

```r
# possible in a loop, but nominal p
t.test(rawdata$Growth[which(rawdata$Passage==1)],
       rawdata$Growth[which(rawdata$Passage==2)],
       var.equal = T)
```

```
    Two Sample t-test

data:  rawdata$Growth[which(rawdata$Passage == 1)] and rawdata$Growth[which(rawdata$Passage =
t = 0.60679, df = 118, p-value = 0.5452
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.321297  6.255936
sample estimates:
mean of x mean of y
 125.4396  123.9723
```

```r
# all pairwise group combinations
pt_out<-pairwise.t.test(x=rawdata$Growth,
                        g=rawdata$Passage_F,
                        p.adjust.method='none')
pt_out
```

```
    Pairwise comparisons using t tests with pooled SD

data:  rawdata$Growth and rawdata$Passage_F

  1       2       3       4       5
2 0.55215 -       -       -       -
3 0.01871 0.00331 -       -       -
4 0.00611 0.00088 0.69214 -       -
5 0.00104 0.00011 0.34487 0.58296 -
6 9e-09   3e-10   0.00048 0.00189 0.01025

P value adjustment method: none
```

```r
pairwise.t.test(x=rawdata$Growth,g=rawdata$Passage,
                p.adjust.method='fdr')
```

```
    Pairwise comparisons using t tests with pooled SD

data:  rawdata$Growth and rawdata$Passage

  1       2       3       4       5
2 0.62460 -       -       -       -
3 0.02552 0.00621 -       -       -
4 0.01018 0.00259 0.69214 -       -
5 0.00259 0.00057 0.43109 0.62460 -
6 6.8e-08 4.5e-09 0.00178 0.00405 0.01538

P value adjustment method: fdr
```

```
pairwise.t.test(x=rawdata$Growth,g=rawdata$Passage,
                p.adjust.method='bonferroni')
```

```
    Pairwise comparisons using t tests with pooled SD

data:  rawdata$Growth and rawdata$Passage

  1       2       3      4      5
2 1.0000  -       -      -      -
3 0.2807  0.0497  -      -      -
4 0.0917  0.0133  1.0000 -      -
5 0.0155  0.0017  1.0000 1.0000 -
6 1.4e-07 4.5e-09 0.0071 0.0283 0.1538

P value adjustment method: bonferroni
```

```
# comparison against reference group 1
pt_out$p.value[,1]
```

```
           2            3            4            5            6
5.521460e-01 1.871115e-02 6.110172e-03 1.036173e-03 9.031123e-09
```

```r
# comparison against reference group 6
pt_out$p.value[5,]
```

```
           1            2            3            4            5
9.031123e-09 3.001066e-10 4.757018e-04 1.889098e-03 1.025037e-02
```

```r
# comparison for selection
c(pt_out$p.value[1,1],pt_out$p.value[3,2],
  pt_out$p.value[5,1])
```

```
[1] 5.521460e-01 8.842382e-04 9.031123e-09
```

```r
# comparison against next level
diag(pt_out$p.value)
```

```
[1] 0.55214600 0.00331248 0.69214393 0.58295615 0.01025037
```

```r
# adjusting for multiple testing for selected comparisons
p.adjust(diag(pt_out$p.value),method='fdr')
```

```
[1] 0.69214393 0.01656240 0.69214393 0.69214393 0.02562592
```

```r
formatP(p.adjust(pt_out$p.value[,1],method='fdr'))
```

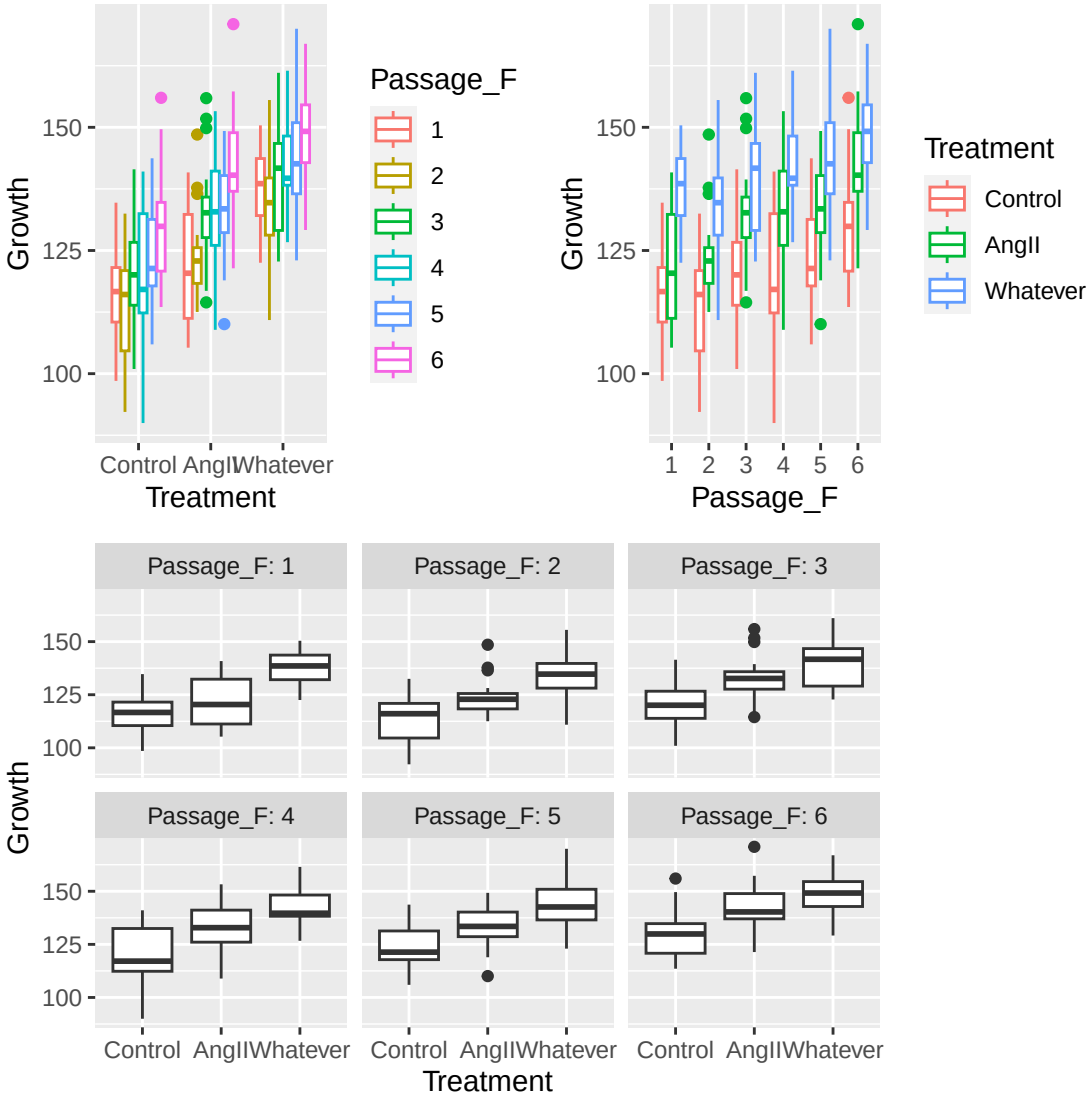
```
[1] "0.552" "0.023" "0.010" "0.003" "0.001"
```

2.4.3 LM with continuous AND categorical IV

Traditionally you may thing of *regression* **OR** *ANOVA*, but they are no different and can be combined. This is called a general linear model. Multivariable models may contain interactions between independent variables V IV1*IV2.

2.4.3.1 Graphical exploration

```r
p0 <- ggplot(rawdata,aes(Treatment,Growth))+
  geom_boxplot()
p1 <- ggplot(rawdata,aes(Treatment,Growth, color=Passage_F))+
  geom_boxplot()
p2 <- ggplot(rawdata,aes(color=Treatment,Growth, x=Passage_F))+
  geom_boxplot()
p3 <- ggplot(rawdata,aes(Treatment,Growth))+
  geom_boxplot()+
  facet_wrap(facets = vars(Passage_F), labeller='label_both')
# from patchwork
(p1+p2)/p3
```



2.4.3.2 Modelling

Models with (*) and without (+) interaction are build and tested.

```
lmOut_interaction<-lm(Growth~Passage*Treatment,data=rawdata)
Anova(lmOut_interaction,type = 3)
```

```
Anova Table (Type III tests)
```

```
Response: Growth
                  Sum Sq  Df   F value    Pr(>F)
(Intercept)       285160   1 2448.5613 < 2.2e-16 ***
Passage             2723   1   23.3855 1.981e-06 ***
Treatment           5635   2   24.1924 1.419e-10 ***
Passage:Treatment    335   2    1.4376    0.2389
Residuals          41227 354
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#
lmOut_additive<-lm(Growth~Passage+Treatment,data=rawdata)
Anova(lmOut_additive,type=2)
```

```
Anova Table (Type II tests)

Response: Growth
          Sum Sq  Df F value    Pr(>F)
Passage     8996   1  77.058 < 2.2e-16 ***
Treatment  24137   2 103.372 < 2.2e-16 ***
Residuals  41562 356
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# for comparison, here is the univariable model
lmOut_uni<-lm(Growth~Treatment,data=rawdata)
aOut<-Anova(lmOut_uni,type=3)
a_uni <- anova(lmOut_uni)
a_uni$`Pr(>F)`
```

```
[1] 5.549803e-31           NA
```

2.4.3.3 Post-hoc analyses

For multivariable models, pairwise.t.test() is not appropriate, Dunnet or Tukey tests (depending on hypothesis) are typical solutions.

```
summary(glht(model=lmOut_additive,linfct=mcp(Treatment='Dunnett')))
```

```
     Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Dunnett Contrasts


Fit: lm(formula = Growth ~ Passage + Treatment, data = rawdata)

Linear Hypotheses:
                        Estimate Std. Error t value Pr(>|t|)
AngII - Control == 0      10.409      1.395   7.462   <1e-10 ***
Whatever - Control == 0   20.052      1.395  14.375   <1e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

```
summary(glht(model=lmOut_additive,linfct=mcp(Treatment='Tukey')))
```

```
     Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = Growth ~ Passage + Treatment, data = rawdata)

Linear Hypotheses:
                        Estimate Std. Error t value Pr(>|t|)
AngII - Control == 0      10.409      1.395   7.462   <1e-10 ***
Whatever - Control == 0   20.052      1.395  14.375   <1e-10 ***
Whatever - AngII == 0      9.643      1.395   6.913   <1e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

```
DescTools::DunnettTest(Growth~Passage_F,data=rawdata)
```

```
  Dunnett's test for comparing several treatments with a control :
    95% family-wise confidence level

$`1`
         diff     lwr.ci    upr.ci    pval
2-1 -1.467320 -7.6899251  4.755285  0.9648
3-1  5.824059 -0.3985468 12.046664  0.0750 .
4-1  6.801105  0.5784996 13.023710  0.0263 *
5-1  8.156143  1.9335375 14.378748  0.0047 **
6-1 14.520106  8.2975011 20.742712 4.8e-08 ***

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
pairwise.t.test(rawdata$Growth,rawdata$Treatment,p.adjust.method = 'n')
```

```
    Pairwise comparisons using t tests with pooled SD

data:  rawdata$Growth and rawdata$Treatment

         Control AngII
AngII    5.1e-11 -
Whatever < 2e-16 1.0e-09

P value adjustment method: none
```

```
mean(rawdata$Growth[which(rawdata$Passage==1 &
                          rawdata$Treatment=='Control')])
```

```
[1] 116.4531
```

```
aOut$'Pr(>F)'
```

```
[1] 2.909117e-279  5.549803e-31            NA
```

```
aOut$`Sum Sq`
```

```
[1] 1754742.53   24136.66   50557.95
```

```
summary(lmOut_additive)
```

```
Call:
lm(formula = Growth ~ Passage + Treatment, data = rawdata)

Residuals:
    Min      1Q  Median      3Q     Max
-32.407  -7.793  -0.281   7.255  32.283

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       110.6802     1.5280  72.432  < 2e-16 ***
Passage             2.9271     0.3334   8.778  < 2e-16 ***
TreatmentAngII     10.4089     1.3949   7.462 6.59e-13 ***
TreatmentWhatever  20.0520     1.3949  14.375  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.8 on 356 degrees of freedom
Multiple R-squared:  0.4436,    Adjusted R-squared:  0.4389
F-statistic:  94.6 on 3 and 356 DF,  p-value: < 2.2e-16
```

```
(result<-tibble(predictor=rownames(aOut),
                    p=formatP(aOut$'Pr(>F)',ndigits=5)))
```

```
# A tibble: 3 x 2
  predictor   p
  <chr>       <chr>
1 (Intercept) "0.00001"
2 Treatment   "0.00001"
3 Residuals   "    NA"
```

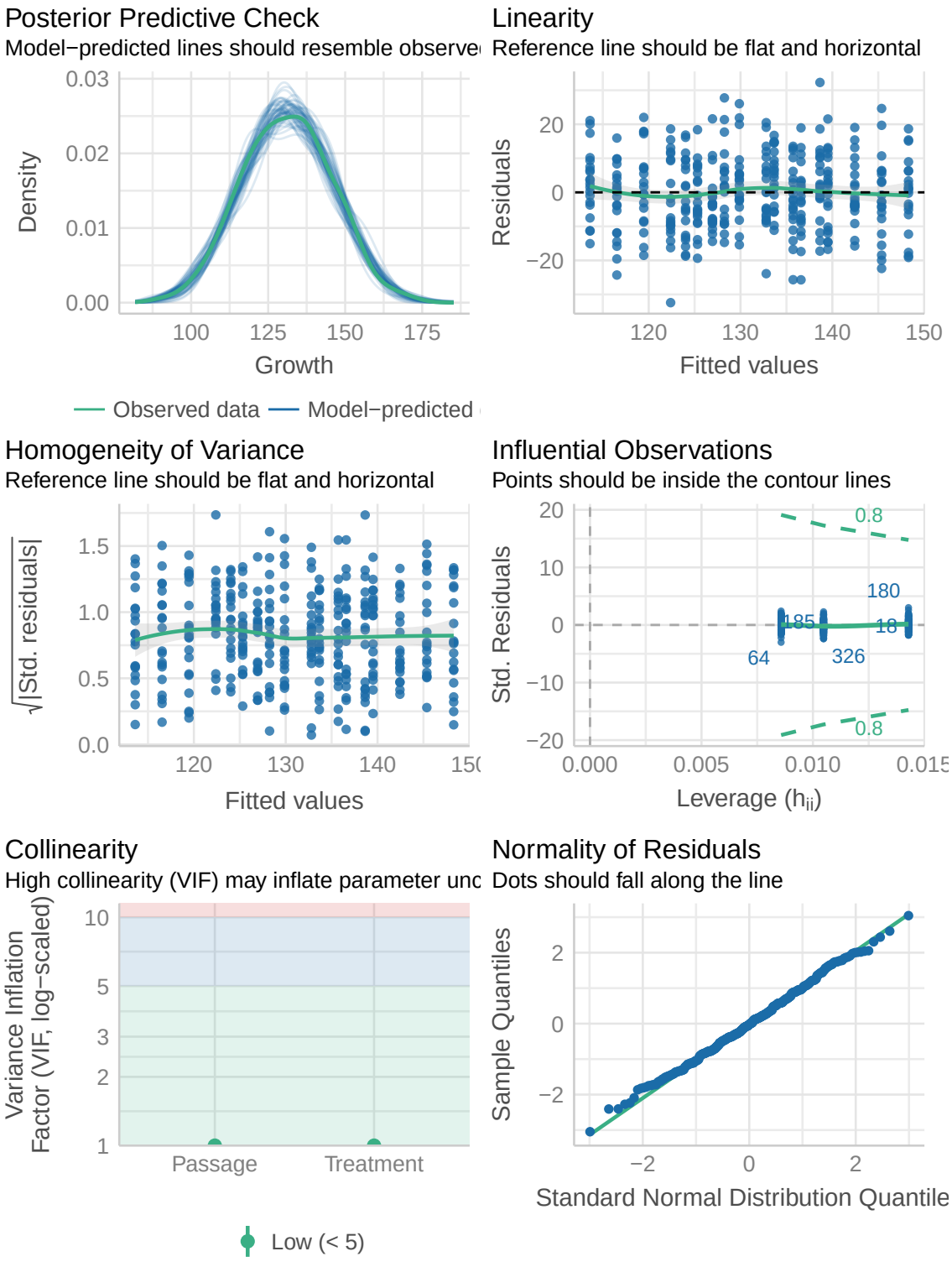
```r
broom::tidy(aOut)
```

```
# A tibble: 3 x 5
  term           sumsq    df statistic    p.value
  <chr>          <dbl> <dbl>     <dbl>      <dbl>
1 (Intercept) 1754743.     1   12391.   2.91e-279
2 Treatment     24137.     2      85.2  5.55e- 31
3 Residuals     50558.   357      NA   NA
```

2.4.4 Model exploration with package performance

```
# x11() #interactive only!

# from package performance
check_model(lmOut_additive)
```


Posterior Predictive Check
Model−predicted lines should resemble observe
Density
0.03
0.02
0.01
0.00
100
125
150
175
Growth
Observed data
Model−predicted
Linearity
Reference line should be flat and horizontal
Residuals
20
0
−20
120
130
140
150
Fitted values
Homogeneity of Variance
Reference line should be flat and horizontal
√|Std. residuals|
1.5
1.0
0.5
0.0
120
130
140
Fitted values
Influential Observations
Points should be inside the contour lines
Std. Residuals
20
10
0
−10
−20
0.8
180
185
18
64
326
0.8
0.000
0.005
0.010
Leverage (hii)
Collinearity
High collinearity (VIF) may inflate parameter unc
Variance Inflation Factor (VIF, log−scaled)
10
5
3
2
1
Passage
Treatment
Low (< 5)
Normality of Residuals
Dots should fall along the line
Sample Quantiles
2
0
−2
−2
0
2
Standard Normal Distribution Quantile

3 Summary

In summary, this book has no content whatsoever.

```
1 + 1
```

```
[1] 2
```

References

Knuth, Donald E. 1984. “Literate Programming.” *Comput. J.* 27 (2): 97–111. https://doi.org/10.1093/comjnl/27.2.97.