STATS 2107
Statistical Modelling and Inference II
Solutions
Workshop 11:
From ANOVA to ANCOVA

Matt Ryan

Semester 2 2022

Contents

The data **2**
- Where to get it . . . 2
- What do we have here? . . . 2
- score . . . 2
- treatment . . . 3
- exercise . . . 3
- age . . . 3

One-way ANOVA **4**
- What to consider . . . 4
- Is this supported by EDA . . . 4
- Fit it in R . . . 5
- Do the ANOVA . . . 5

Your turn **6**
- What to do . . . 6

Two-way ANOVA **7**
- What to consider . . . 7
- Do we have the data for an interaction? . . . 8
- Is this model supported by EDA - a treatment effect . . . 8
- Is this model supported by EDA - an interaction . . . 8
- Fit it in R . . . 9

Your turn **10**
- What to do . . . 10

Two-way ANCOVA **10**
- What to consider . . . 10
- Is this supported by EDA . . . 11
- Fit it in R . . . 11
- Can we simplify? . . . 12
- Yes we can! . . . 12

Your turn **13**
What to do . . . 13

The data

Where to get it

```
install.packages("datarium")
data("stress", package = "datarium")
stress <- as_tibble(stress) %>%
  mutate(treatment = fct_rev(treatment))
```

```
data("stress", package = "datarium")
stress <- as_tibble(stress) %>%
  mutate(treatment = fct_rev(treatment))
```

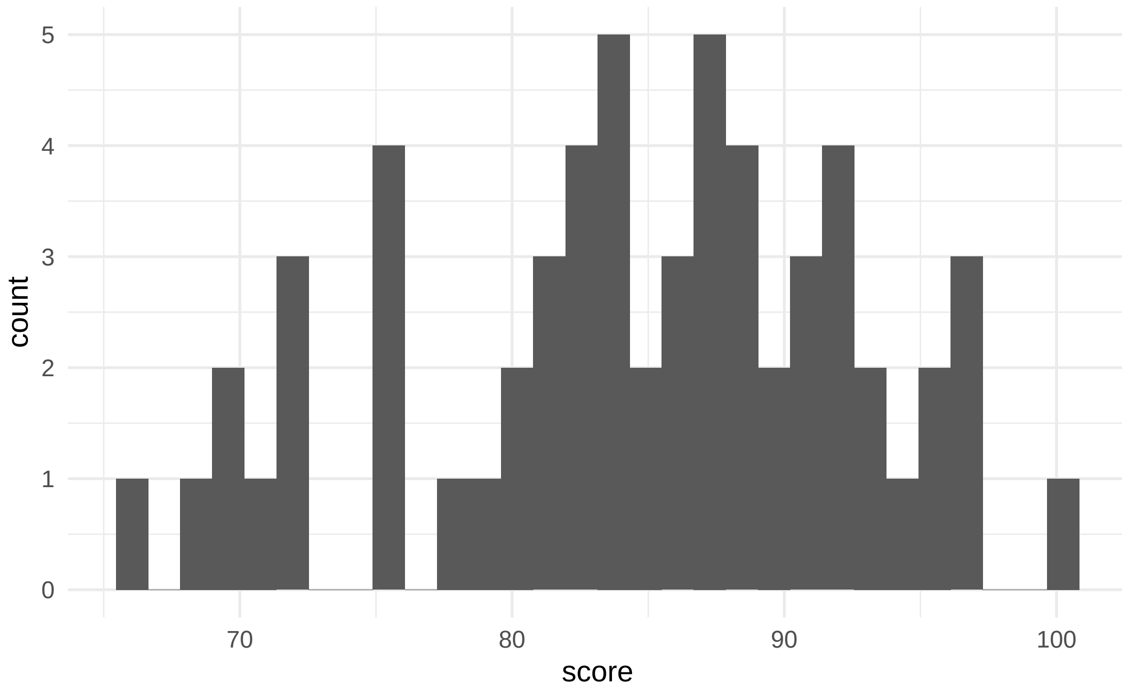
What do we have here?

```
tribble(
  ~Variable, ~Description, ~Type,
  "id", "A unique identifier", "ID variable",
  "score", "Stress score out of 100", "Continuous numeric (response variable)",
  "treatment", "Are they in the treatement group?", "Categorical nominal",
  "exercise", "What level of exercise do they do?",  "Categorical nominal",
  "age", "Age of participant", "Continuous numeric"
) %>%
  knitr::kable()
```

Variable	Description	Type
id	A unique identifier	ID variable
score	Stress score out of 100	Continuous numeric (response variable)
treatment	Are they in the treatement group?	Categorical nominal
exercise	What level of exercise do they do?	Categorical nominal
age	Age of participant	Continuous numeric

score

```
stress %>%
  ggplot(aes(x = score)) +
  geom_histogram(bins = 30) +
  theme_minimal()
```



treatment

```
stress %>%
  count(treatment) %>%
  knitr::kable()
```

treatment	n
no	30
yes	30

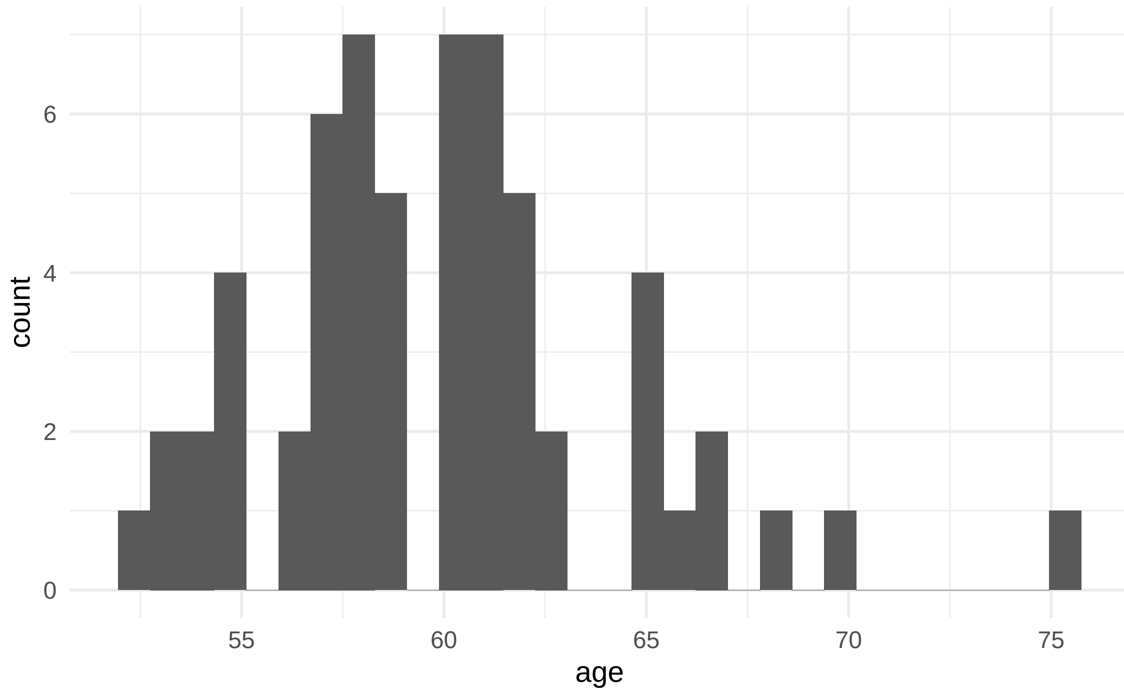
exercise

```
stress %>%
  count(exercise) %>%
  knitr::kable()
```

exercise	n
low	20
moderate	20
high	20

age

```
stress %>%
  ggplot(aes(x = age)) +
  geom_histogram(bins = 30) +
  theme_minimal()
```



One-way ANOVA

What to consider

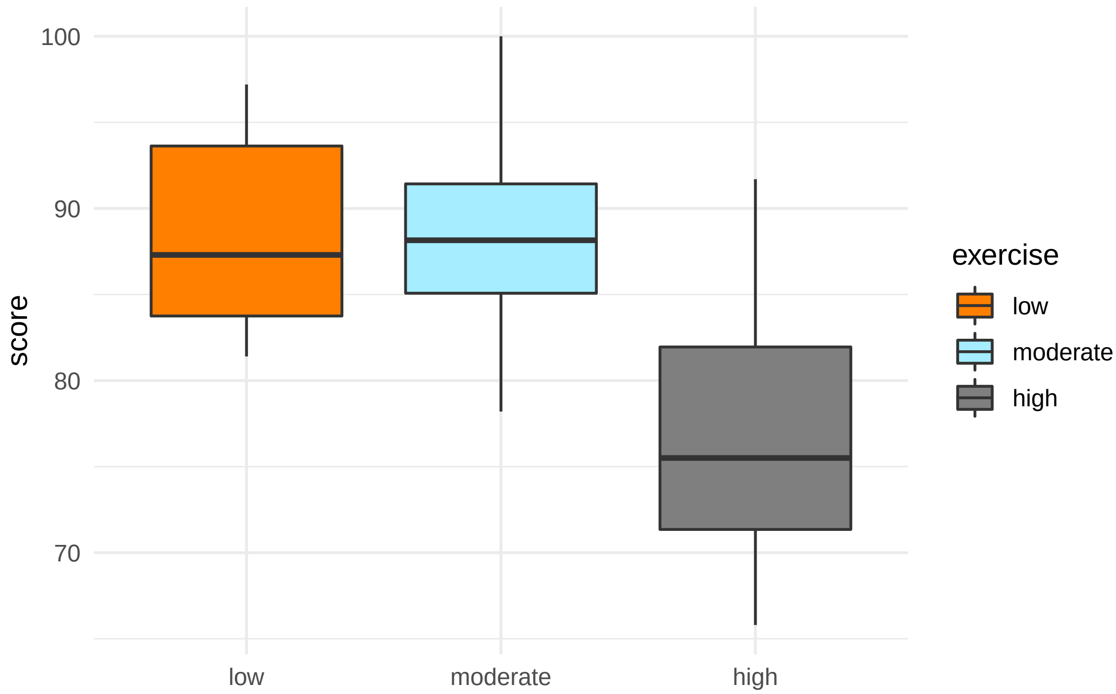
Let's suppose that there is a relationship between stress levels and exercise:

$$\text{score}_{ik} = \mu + \alpha_i + \varepsilon_{ik}$$

where α_i is the effect of each exercise group

Is this supported by EDA

```
stress %>%
  ggplot(aes(x = exercise, y = score, fill = exercise)) +
  geom_boxplot() +
  labs(x = NULL) +
  scale_fill_manual(values = cols) + # This wont work for you
  theme_minimal()
```



Fit it in R

We fit using the `lm` command:

```
stress_anova <- lm(score ~ exercise, data = stress)
summary(stress_anova)
```

```
##
## Call:
## lm(formula = score ~ exercise, data = stress)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.090  -4.674  -1.107   4.628  14.810
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        88.725      1.361  65.186  < 2e-16 ***
## exercisemoderate   -0.610      1.925  -0.317    0.752
## exercisehigh      -11.835      1.925  -6.148 8.21e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.087 on 57 degrees of freedom
## Multiple R-squared:  0.4568, Adjusted R-squared:  0.4378
## F-statistic: 23.97 on 2 and 57 DF,  p-value: 2.791e-08
```

Do the ANOVA

```
anova(stress_anova)
```

```
## Analysis of Variance Table
```

```
##
## Response: score
##           Df Sum Sq Mean Sq F value    Pr(>F)
## exercise   2 1776.3  888.13   23.97 2.791e-08 ***
## Residuals 57 2112.0   37.05
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Your turn

What to do

1. Based on the ANOVA output, would you reject or retain the null hypothesis that all exercise groups have the same mean pain score.

Solutions:
Out p-value is much less than 0.05, so we would reject this. From the EDA, this seems to be driven by the high exercise group.

---

2. Look at the model summary. What does the intercept term represent?

Solutions:
The intercept term 88.725 represents the mean pain score for the low exercise group.

---

3. Does this data meet the assumptions of ANOVA?

Solutions:
We need:

- Independence
- Constant variance
- Normality of each group

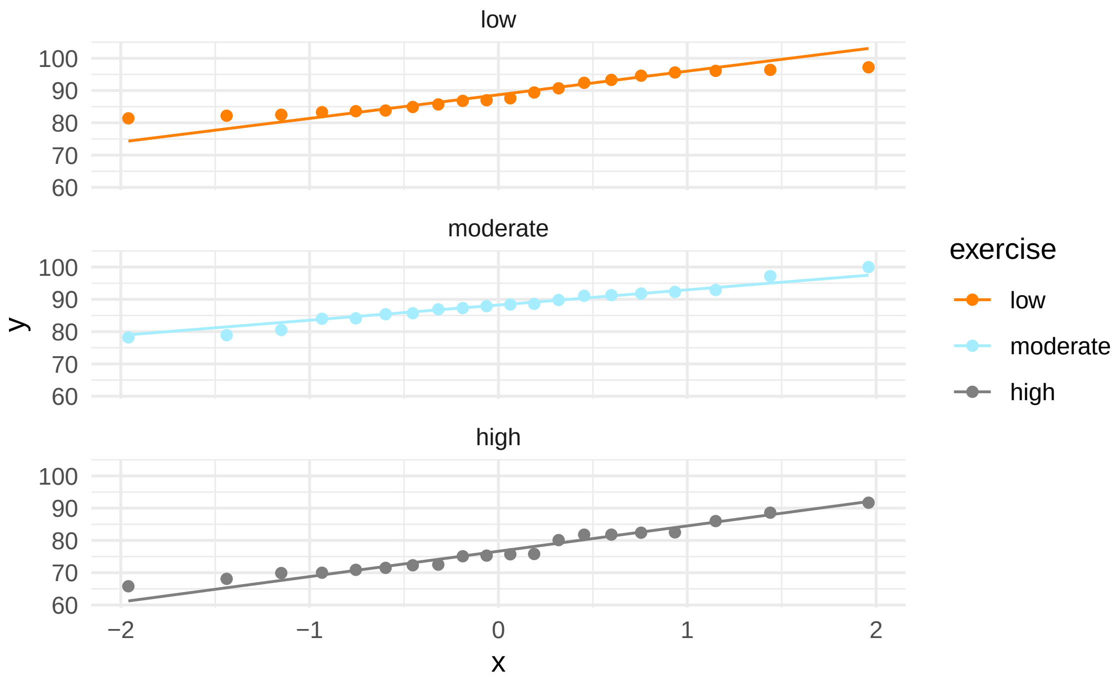
We are not given enough information to test for independence. To test for constant variance, we want our max(sd)/min(sd) $<$ 2, so we get those with:

```
stress %>%
  group_by(exercise) %>%
  summarise(s = sd(score)) %>%
  arrange(s)
```

```
## # A tibble: 3 x 2
##   exercise     s
##   <fct>    <dbl>
## 1 low       5.39
## 2 moderate  5.57
## 3 high      7.15
```

Solutions:
Our ratio is $7.145105/5.391794 < 2$, so constant variance is satisfied. To test normality, we need a normal QQ-plot for each group:

```
stress %>%
  ggplot(aes(sample = score, colour = exercise)) +
  geom_qq() +
  geom_qq_line() +
  scale_colour_manual(values = cols) +
  facet_wrap(~exercise, nrow = 3) +
  theme_minimal()
```



Solutions:
We see some small deviation, but this looks pretty good for the small number of data points.

Hence, our assumptions are likely all reasonable.

Two-way ANOVA

What to consider

Now, we know there is a treatment group, so let's suppose that there is a relationship between stress levels, exercise and treatment:

$$\text{score}_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

where α_i is the effect of each exercise group, β_j is the effect of each treatment group, and γ_{ij} is an interaction between the exercise and treatment.

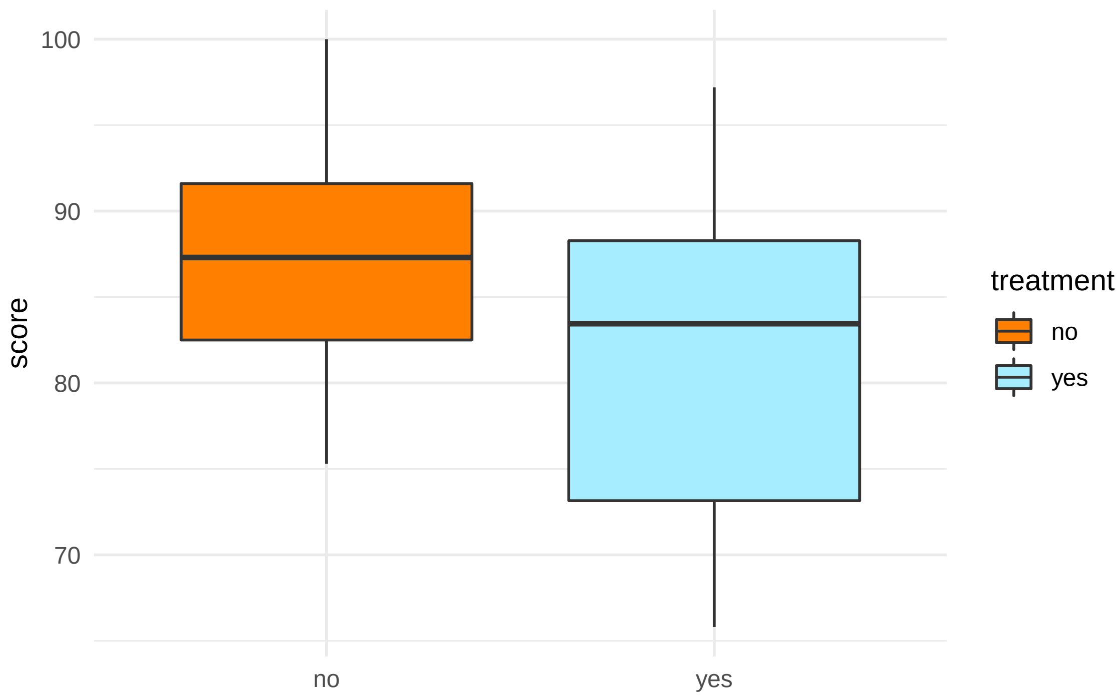
Do we have the data for an interaction?

```
stress %>%
  count(treatment, exercise) %>%
  pivot_wider(names_from = exercise, values_from = n) %>%
  knitr::kable()
```

treatment	low	moderate	high
no	10	10	10
yes	10	10	10

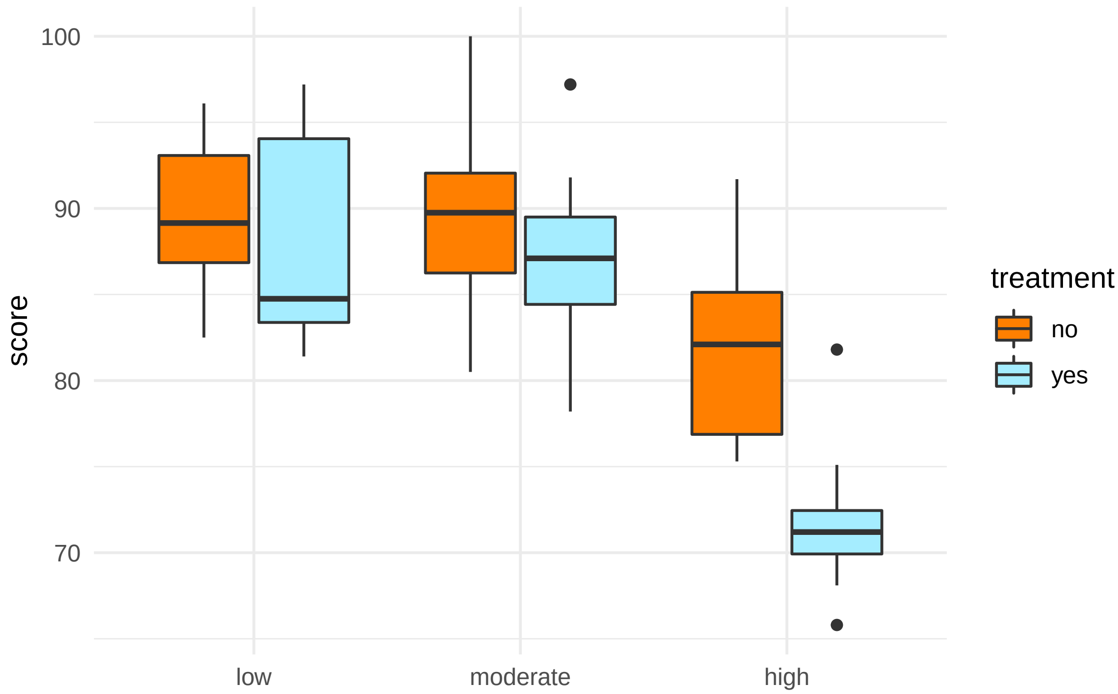
Is this model supported by EDA - a treatment effect

```
stress %>%
  ggplot(aes(x = treatment, y = score, fill = treatment)) +
  geom_boxplot() +
  labs(x = NULL) +
  scale_fill_manual(values = cols) + # This won't work for you
  theme_minimal()
```



Is this model supported by EDA - an interaction

```
stress %>%
  ggplot(aes(x = exercise, y = score, fill = treatment)) +
  geom_boxplot() +
  labs(x = NULL) +
  scale_fill_manual(values = cols) + # This won't work for you
  theme_minimal()
```



Fit it in R

We fit using the `lm` command:

```
stress_two_way_anova <- lm(score ~ exercise * treatment, data = stress)
summary(stress_two_way_anova)
```

```
##
## Call:
## lm(formula = score ~ exercise * treatment, data = stress)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.910 -3.797 -0.240  3.062 10.590
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   89.590      1.690  52.995  < 2e-16 ***
## exercisemoderate              -0.180      2.391  -0.075  0.94026
## exercisehigh                  -7.600      2.391  -3.179  0.00245 **
## treatmentyes                  -1.730      2.391  -0.724  0.47243
## exercisemoderate:treatmentyes  -0.860      3.381  -0.254  0.80019
## exercisehigh:treatmentyes     -8.470      3.381  -2.505  0.01529 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.346 on 54 degrees of freedom
## Multiple R-squared:  0.6031, Adjusted R-squared:  0.5663
## F-statistic: 16.41 on 5 and 54 DF,  p-value: 8.005e-10
```

Your turn

What to do

1. Look at the model summary. What does the intercept term represent?

Solutions:
The intercept term 89.590 represents the mean pain score for people in the low exercise group who are not in the treatment group.

2. Interpret the `exercisehigh` coefficient.

Solutions:
If we are not in the treatment group, we expect people in the high exercise group to have a lower stress score of about 7.6 points on average.

3. Perform the ANOVA. Is the interaction term significant? Interpret this in context.

Solutions:

```
anova(stress_two_way_anova)
```

```
## Analysis of Variance Table
##
## Response: score
##                    Df  Sum Sq Mean Sq F value    Pr(>F)
## exercise            2 1776.27  888.13  31.076 1.045e-09 ***
## treatment           1  351.38  351.38  12.295 0.0009227 ***
## exercise:treatment  2  217.32  108.66   3.802 0.0285218 *
## Residuals          54 1543.30   28.58
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Solutions:
Based on the ANOVA output, we can see the interaction term is significant at the 5% level. This suggest there is a significant relationship between what treatment you are on and which exercise group you are in.

Two-way ANCOVA

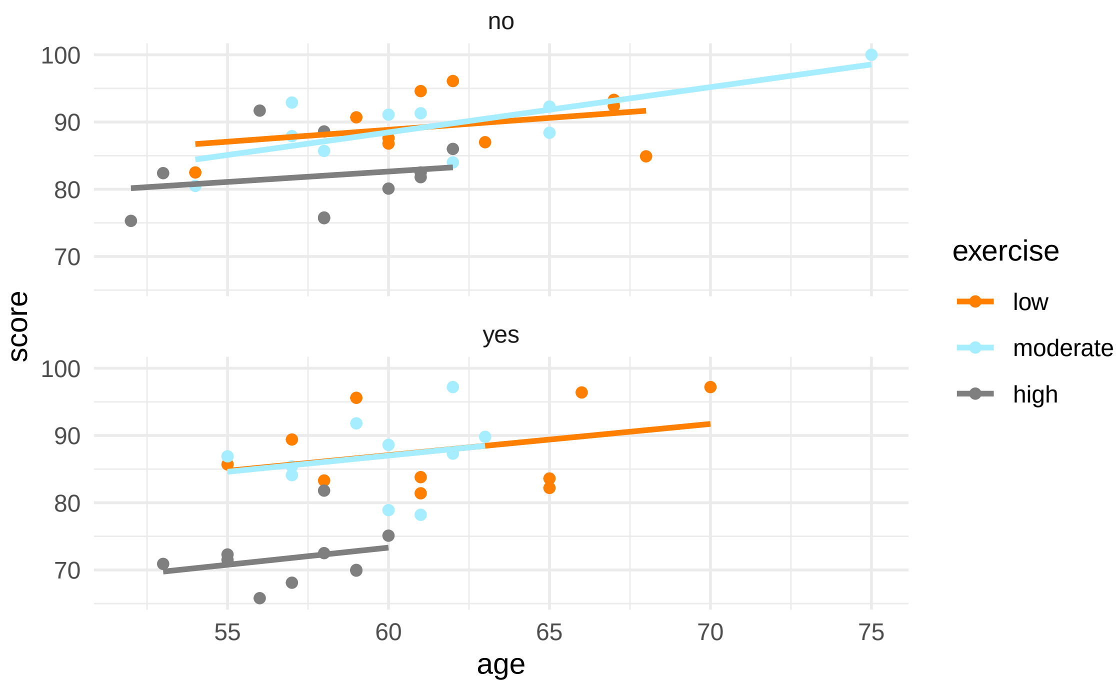
What to consider

But wait, there's more! Remember we have the `age` covariate. It is perfectly reasonable to believe there is a relationship between age and stress, as well as some interaction with the treatment and exercise regime. Things get a little more hairy in two-way ANCOVA, so we will start big, and then select the best model. We will consider the model

$$\texttt{score} \sim \texttt{age * treatment * exercise}$$

Is this supported by EDA

```
stress %>%
  ggplot(aes(x = age, y = score, colour = exercise)) +
  geom_point() +
  geom_smooth(se = FALSE, method = lm, formula = y~x) +
  facet_wrap(~treatment,
             nrow = 2) +
  scale_colour_manual(values = cols) + # This won't work for you
  theme_minimal()
```



Fit it in R

We fit using the `lm` command:

```
stress_2way_ancova <- lm(score ~ age * treatment * exercise, data = stress)
summary(stress_2way_ancova)
```

```
##
## Call:
## lm(formula = score ~ age * treatment * exercise, data = stress)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.2940 -3.4969  0.3342  2.2295 10.2988
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      67.59947   24.92019   2.713  0.00924 **
## age                               0.35411    0.40042   0.884  0.38091
## treatmentyes                     -8.41860   33.82489  -0.249  0.80451
## exercisemoderate                -19.50719   30.73047  -0.635  0.52858
## exercisehigh                     -3.55340   38.80691  -0.092  0.92742
```

```
## age:treatmentyes                      0.11070    0.54501   0.203  0.83990
## age:exercisemoderate                  0.31881    0.49536   0.644  0.52290
## age:exercisehigh                     -0.04420    0.65077  -0.068  0.94613
## treatmentyes:exercisemoderate        18.45149   55.34236   0.333  0.74028
## treatmentyes:exercisehigh           -12.85565   63.49342  -0.202  0.84040
## age:treatmentyes:exercisemoderate   -0.30217    0.91128  -0.332  0.74164
## age:treatmentyes:exercisehigh        0.08848    1.08428   0.082  0.93530
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.204 on 48 degrees of freedom
## Multiple R-squared:  0.6657, Adjusted R-squared:  0.5891
## F-statistic: 8.689 on 11 and 48 DF,  p-value: 3.358e-08
```

Can we simplify?

```r
drop1(stress_2way_ancova, test = "F")
```

```
## Single term deletions
##
## Model:
## score ~ age * treatment * exercise
##                        Df Sum of Sq    RSS    AIC F value Pr(>F)
## <none>                              1299.8 208.54
## age:treatment:exercise  2    3.9559 1303.8 204.72   0.073 0.9297
```

Yes we can!

```r
stress_2way_ancova <- update(stress_2way_ancova, . ~ . - age:treatment:exercise)
summary(stress_2way_ancova)
```

```
##
## Call:
## lm(formula = score ~ age + treatment + exercise + age:treatment +
##     age:exercise + treatment:exercise, data = stress)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5637 -3.3982  0.4173  2.3827 10.3907
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   65.15197   21.12371   3.084  0.00332 **
## age                            0.39353    0.33916   1.160  0.25144
## treatmentyes                  -3.89781   24.16324  -0.161  0.87250
## exercisemoderate             -14.81619   24.97471  -0.593  0.55569
## exercisehigh                  -3.90658   30.22926  -0.129  0.89769
## age:treatmentyes               0.03769    0.38851   0.097  0.92311
## age:exercisemoderate           0.24286    0.40215   0.604  0.54864
## age:exercisehigh              -0.03524    0.50722  -0.069  0.94488
## treatmentyes:exercisemoderate  0.20723    3.35949   0.062  0.95106
## treatmentyes:exercisehigh     -8.12783    3.72077  -2.184  0.03365 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 5.106 on 50 degrees of freedom
## Multiple R-squared:  0.6647, Adjusted R-squared:  0.6043
## F-statistic: 11.01 on 9 and 50 DF,  p-value: 3.181e-09
```

Your turn

What to do

1. Finish the model selection process. What is the final model?

Solutions:
Continuing on:

```
drop1(stress_2way_ancova, test = "F")
```

```
## Single term deletions
##
## Model:
## score ~ age + treatment + exercise + age:treatment + age:exercise +
##     treatment:exercise
##                    Df Sum of Sq    RSS    AIC F value  Pr(>F)
## <none>                          1303.8 204.72
## age:treatment       1     0.245 1304.0 202.73  0.0094 0.92311
## age:exercise        2    12.753 1316.6 201.31  0.2445 0.78401
## treatment:exercise  2   178.891 1482.7 208.44  3.4302 0.04018 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
stress_2way_ancova <- update(stress_2way_ancova, . ~ . - age:treatment)
```

```
drop1(stress_2way_ancova, test = "F")
```

```
## Single term deletions
##
## Model:
## score ~ age + treatment + exercise + age:exercise + treatment:exercise
##                    Df Sum of Sq    RSS    AIC F value  Pr(>F)
## <none>                          1304.0 202.73
## age:exercise        2    12.891 1316.9 199.32  0.2521 0.77814
## treatment:exercise  2   227.543 1531.6 208.38  4.4495 0.01655 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
stress_2way_ancova <- update(stress_2way_ancova, . ~ . - age:exercise)
drop1(stress_2way_ancova, test = "F")
```

```
## Single term deletions
##
## Model:
## score ~ age + treatment + exercise + treatment:exercise
##                    Df Sum of Sq    RSS    AIC F value   Pr(>F)
## <none>                          1316.9 199.32
## age                 1    226.36 1543.3 206.84  9.1097 0.003903 **
```

```
## treatment:exercise  2    220.94 1537.9 204.63  4.4458 0.016409 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(stress_2way_ancova)
```

```
##
## Call:
## lm(formula = score ~ age + treatment + exercise + treatment:exercise,
##     data = stress)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3250 -3.0192  0.2745  2.4650 10.6667
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   58.3195    10.4798   5.565 8.83e-07 ***
## age                            0.5036     0.1668   3.018   0.0039 **
## treatmentyes                  -1.5286     2.2303  -0.685   0.4961
## exercisemoderate               0.1725     2.2323   0.077   0.9387
## exercisehigh                  -5.4851     2.3368  -2.347   0.0227 *
## treatmentyes:exercisemoderate -0.1550     3.1613  -0.049   0.9611
## treatmentyes:exercisehigh     -8.2182     3.1537  -2.606   0.0119 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.985 on 53 degrees of freedom
## Multiple R-squared:  0.6613, Adjusted R-squared:  0.623
## F-statistic: 17.25 on 6 and 53 DF,  p-value: 6.167e-11
```

Solutions:
Everything is now significant, so our final model is:

$$\texttt{score} \sim \texttt{age + treatment + exercise + treatment:exercise}$$

---

2. Interpret the coefficient of age in your final model.

Solutions:
If we consider a subject not in the treatment group who does low exercise, then if we increase their age by 1 year we expect their stress to increase by 0.50355 points on average.

ALTERNATIVELY

Irrespective of subject group or treatment, we expect that if we increase a subjects age by 1 year then their stress will increase by 0.50355 points on average.
