Week 11
Linear models II

Today

- Review homework
- Factors and dummy variables
- ANCOVA
- ANOVA
- Checking assumptions
- Data limitations
- In-class exercise

Extending the linear model to categorical variables

- Dummy variables

$$d_i = \begin{cases} 0 & \text{is not level i} \\ 1 & \text{is level i} \end{cases} \qquad (15.1)$$

Extending the linear model to categorical variables

- Dummy variables
- Created automatically when a factor is included in lm() or can be specified using factor()

$$d_i = \begin{cases} 0 & \text{is not level i} \\ 1 & \text{is level i} \end{cases} \qquad (15.1)$$

```
# Fit a linear model treating 'race' as a
factor
model <- lm(write ~ factor(race), data =
my_data)
```

Extending the linear model to categorical variables

- Dummy variables
- Created automatically when a factor is included in lm() or can be specified using factor()
- Compared to a reference level in lm() summary
- Mean of reference level: 114
- Mean of treatment level: 124
- Significant treatment effect

$$d_i = \begin{cases} 0 & \text{is not level i} \\ 1 & \text{is level i} \end{cases} \qquad (15.1)$$

```
# Fit a linear model treating 'race' as a
factor
model <- lm(write ~ factor(race), data =
my_data)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   114.00       1.95   58.37  < 2e-16
grptreat       10.00       2.33    4.29  5.4e-05

Residual standard error: 9.16 on 72 degrees of freedom
Multiple R-squared:  0.204,        Adjusted R-squared:  0.193
F-statistic: 18.4 on 1 and 72 DF,  p-value: 5.43e-05
```

Combining categorical and continuous variables Analysis of Covariance (ANCOVA)

- Understand the effect of treatment (categorical) while controlling for nuisance covariate that adds variation

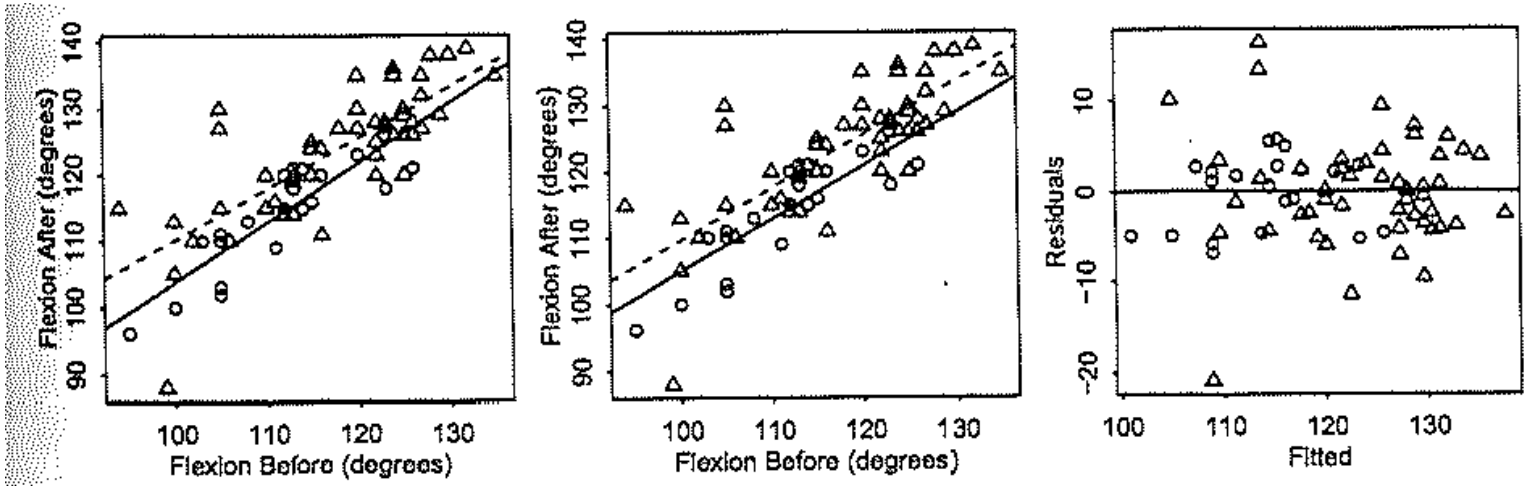
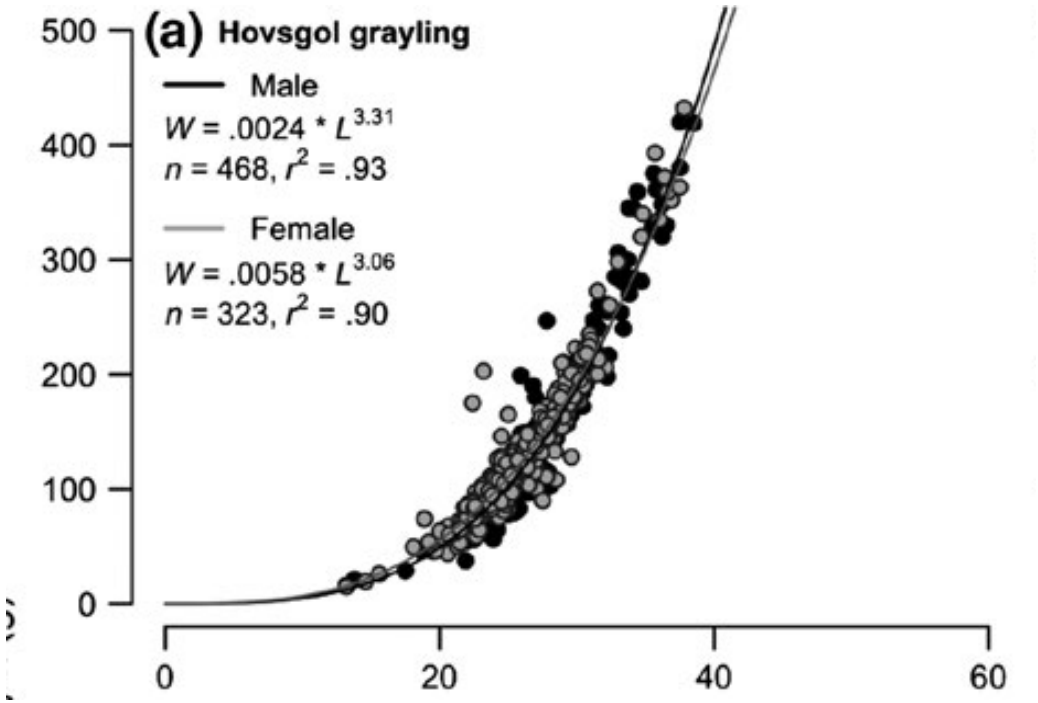


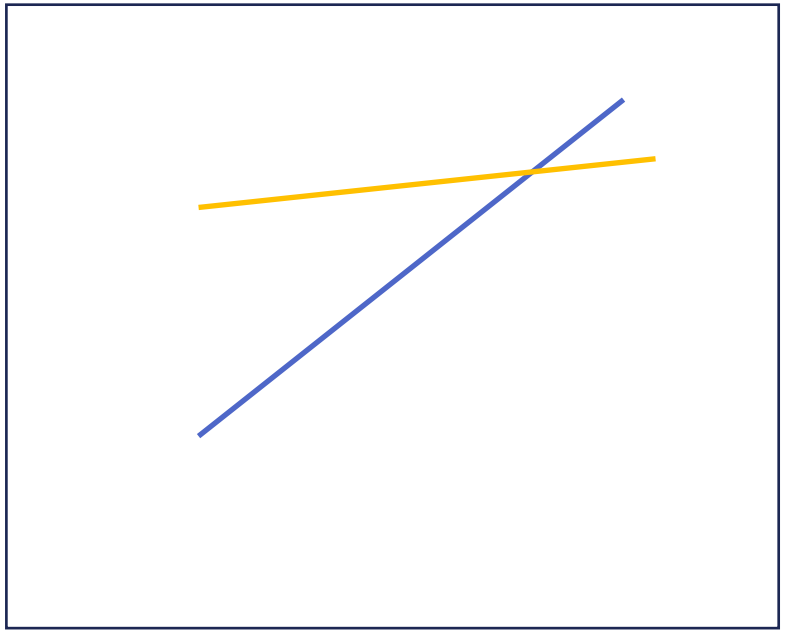
Figure 15.2: In all the three panels, the control cases are marked with a circle and treatment with a triangle. The separate lines model fit is shown in the first panel, the parallel lines model fit seen in the second panel and the residuals versus fitted plot in the third.

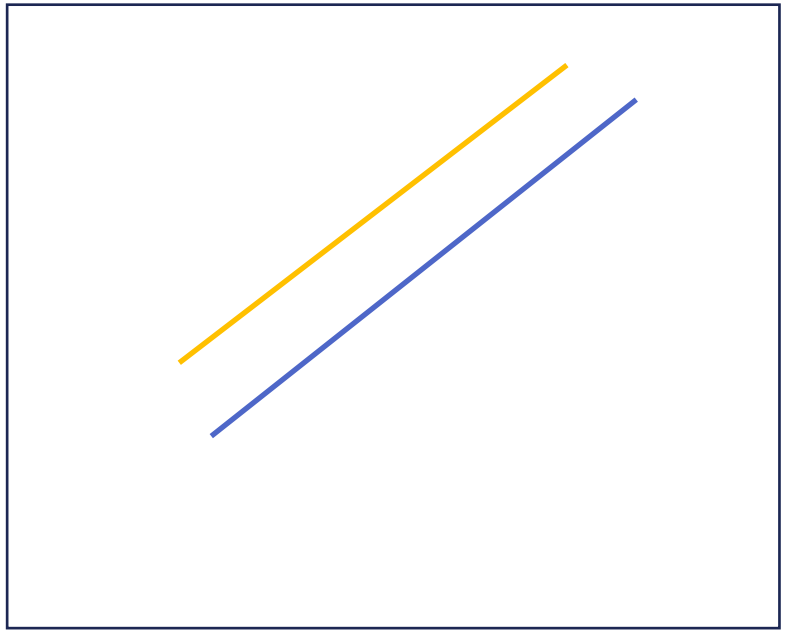
Combining categorical and continuous variables Analysis of Covariance (ANCOVA)

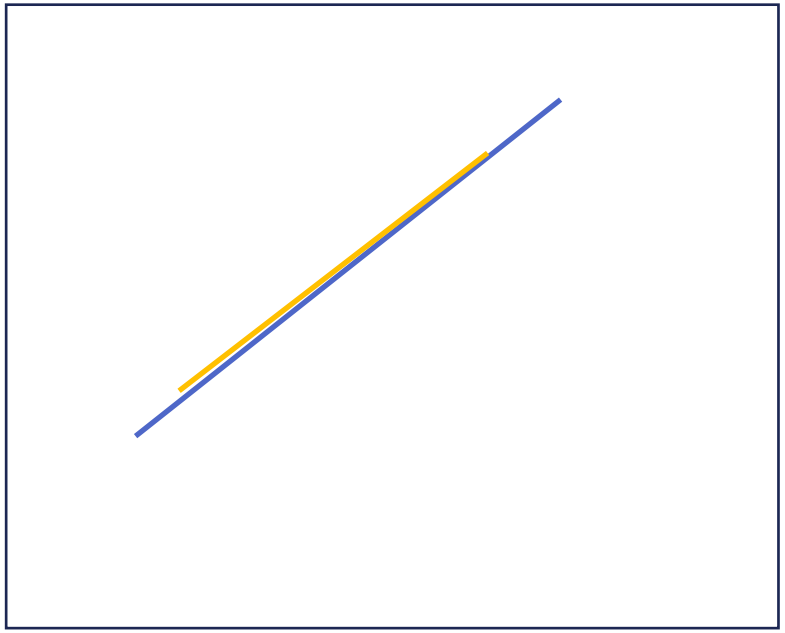
- Understand the relationship between two continuous variables when the relationship may differ by group (categorical)



Interpreting ANCOVA output

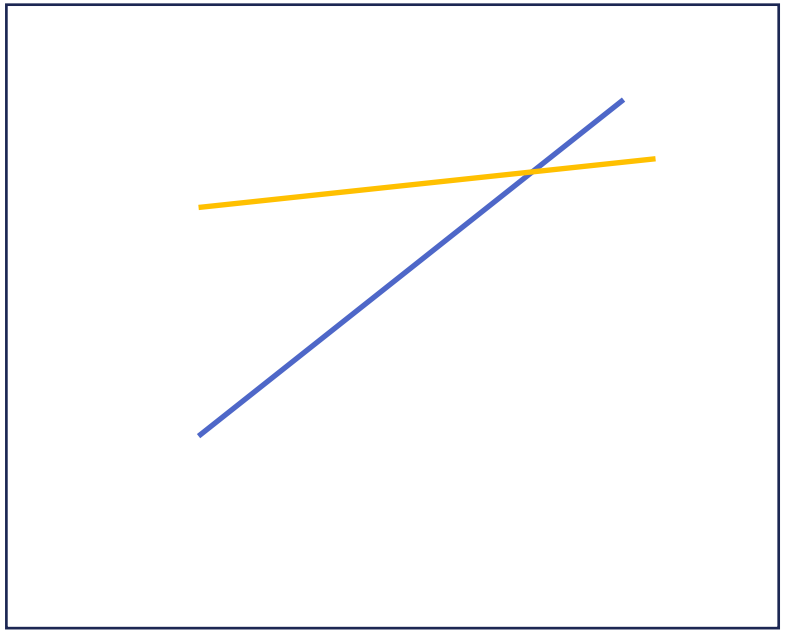


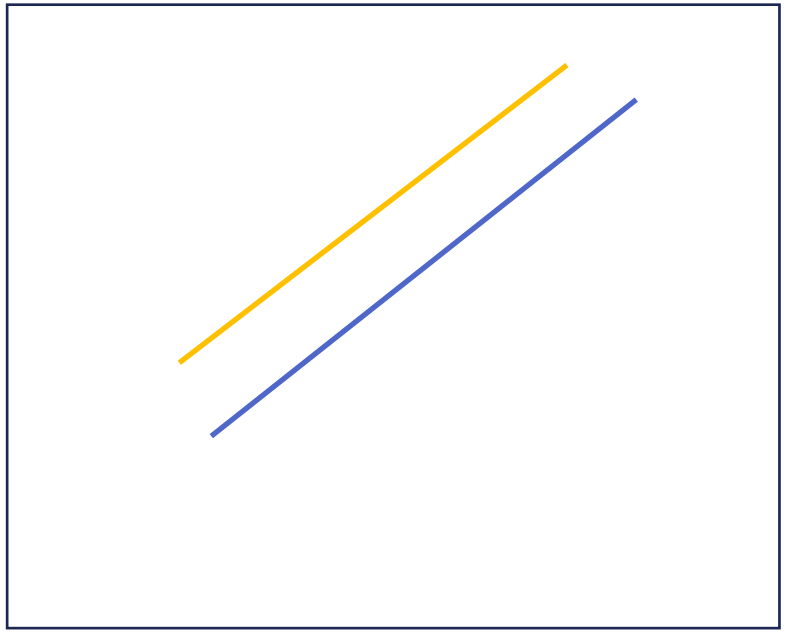


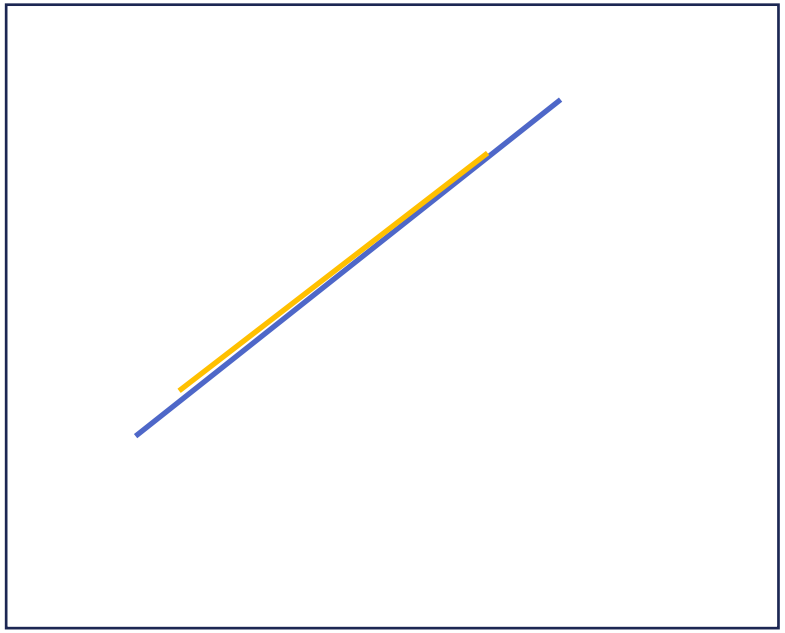


Interpreting ANCOVA output

Mod1 <- lm(Y ~ Xcon*Xfac)
or
Mod1 <- lm(Y ~ Xcon + Xfac + Xcon:Xfac)



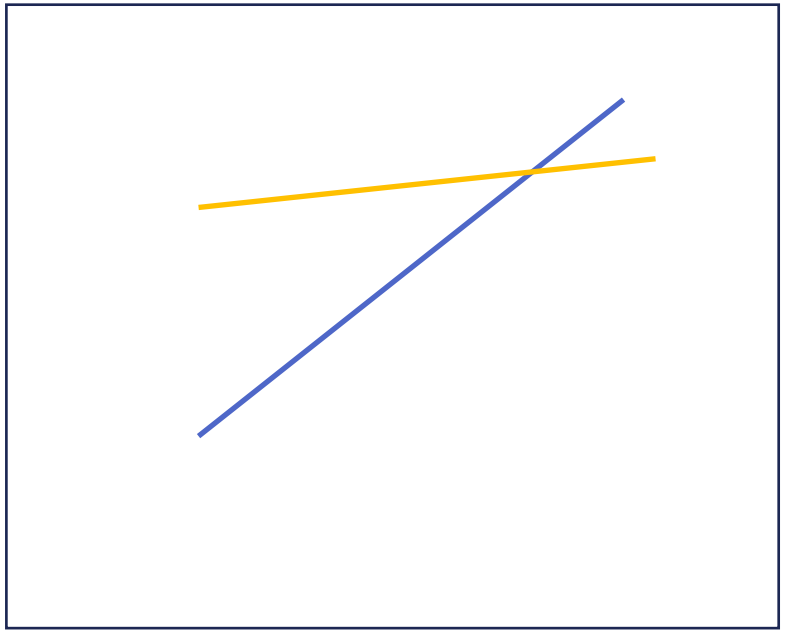


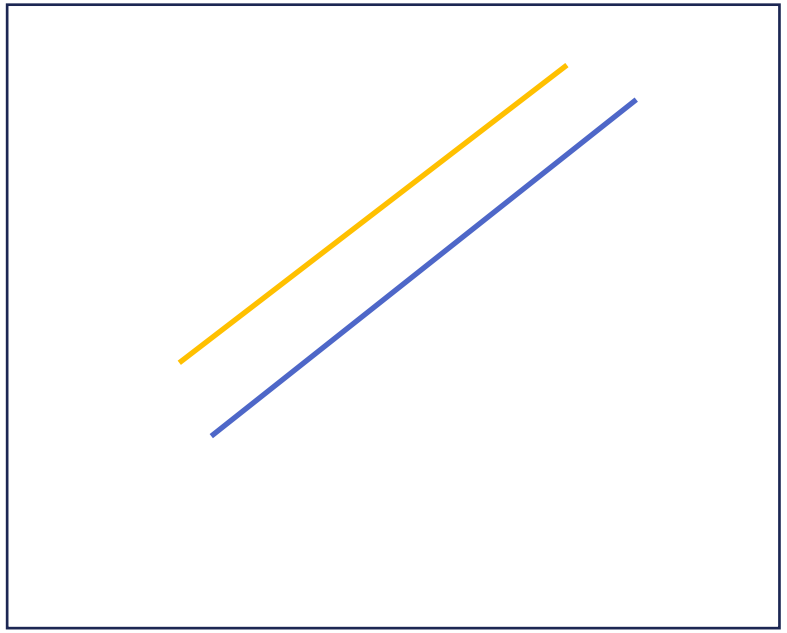


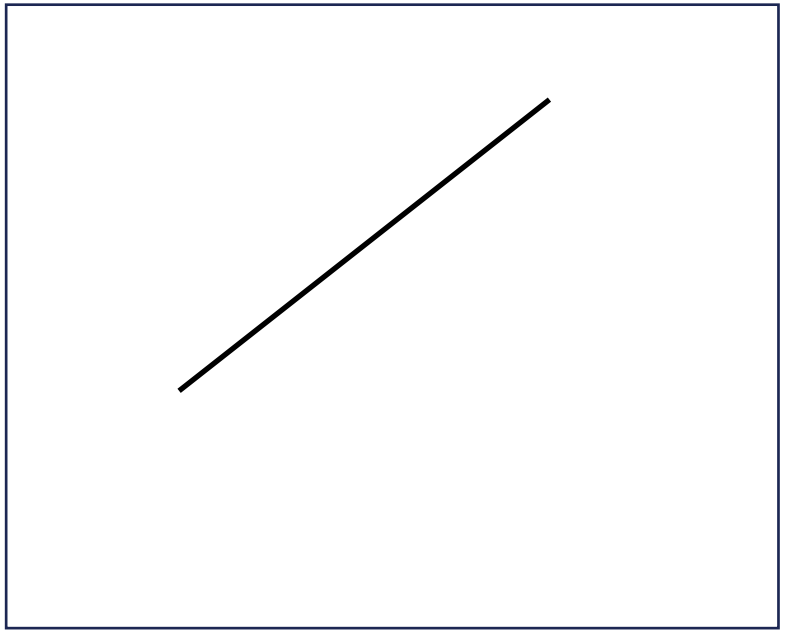
Interpreting ANCOVA output

Mod1 <- lm(Y ~ Xcon*Xfac)
or
Mod1 <- lm(Y ~ Xcon + Xfac + Xcon:Xfac)

Mod2 <- lm(Y ~ Xcon + Xfac)

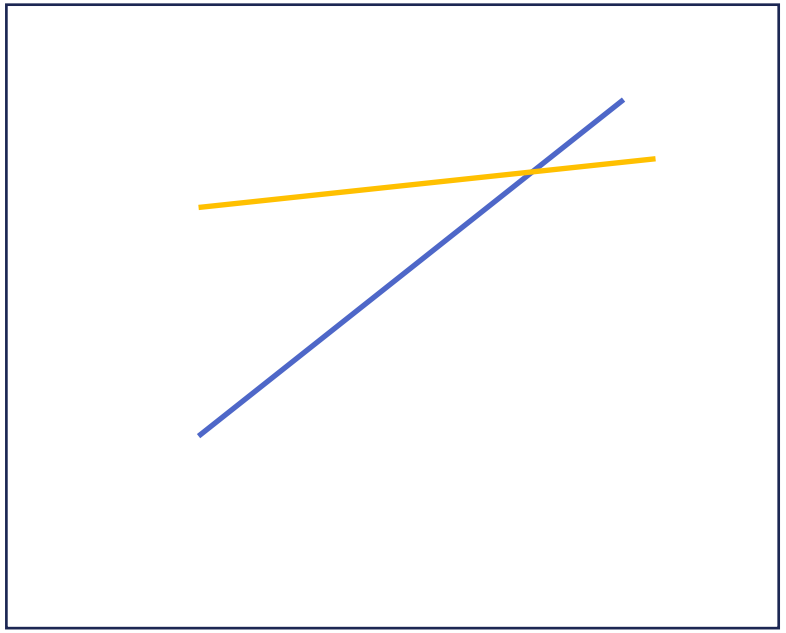




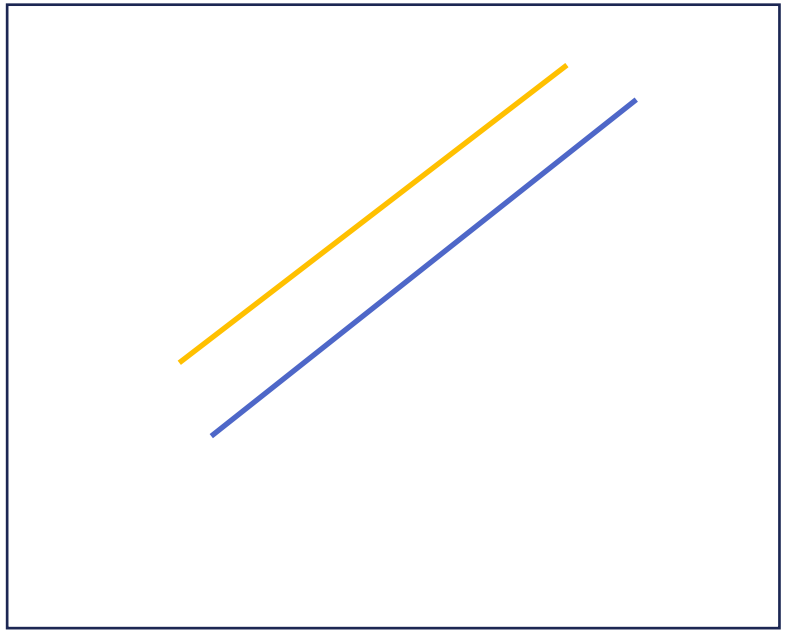


Interpreting ANCOVA output

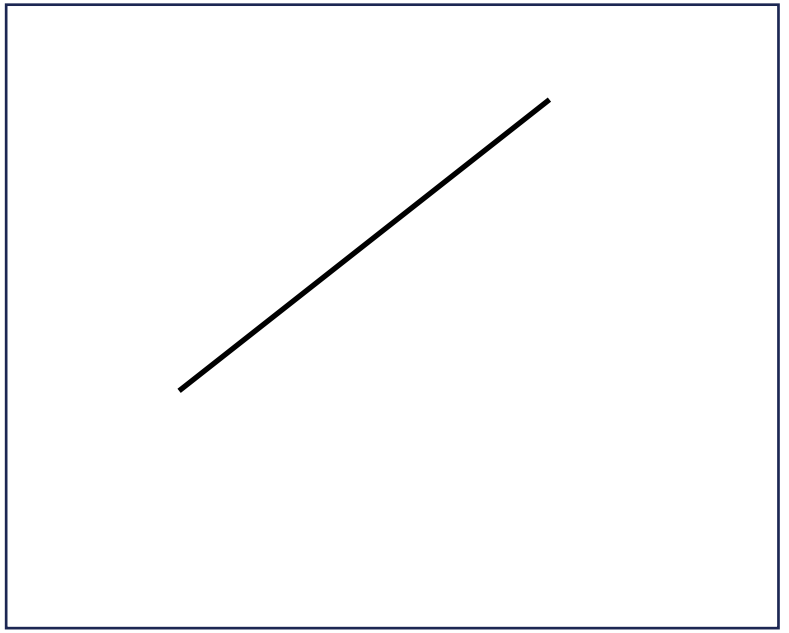
Mod1 <- lm(Y ~ Xcon*Xfac)
or
Mod1 <- lm(Y ~ Xcon + Xfac + Xcon:Xfac)



Mod2 <- lm(Y ~ Xcon + Xfac)

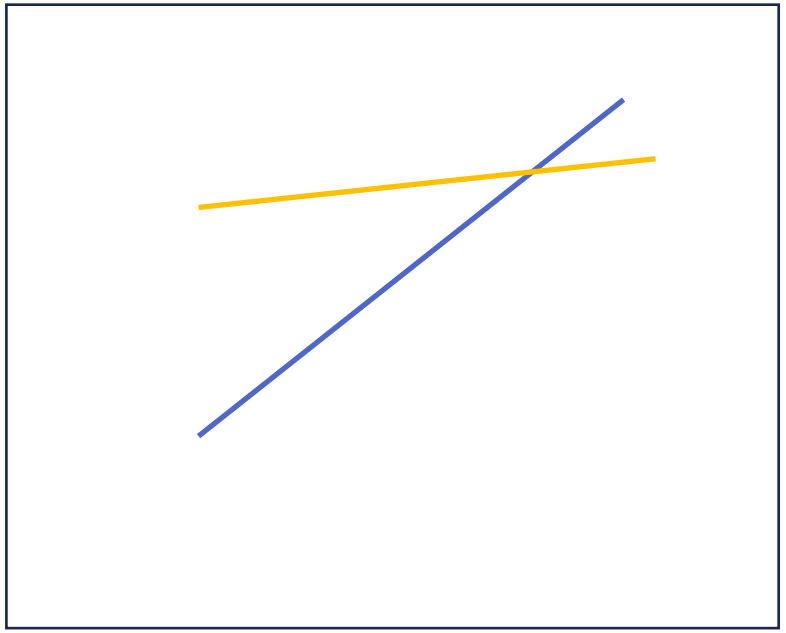


Mod3 <- lm(Y ~ Xcon)

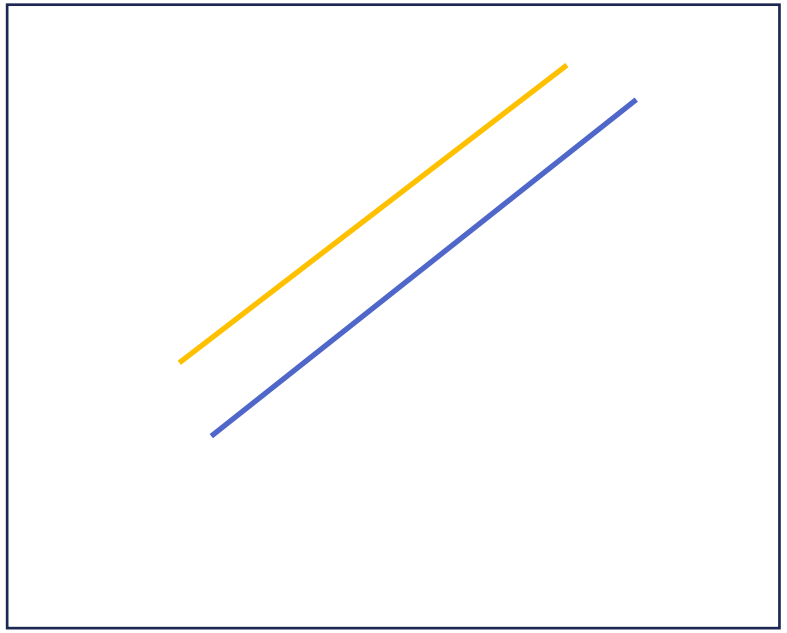


Interpreting ANCOVA output

Mod1 <- lm(Y ~ Xcon*Xfac)
or
Mod1 <- lm(Y ~ Xcon + Xfac + Xcon:Xfac)



Mod2 <- lm(Y ~ Xcon + Xfac)



```
lmod4 = lm(faft ~ fbef+grp+fbef:grp,hips)
summary(lmod4)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    13.558     17.620    0.77     0.44
fbef            0.902      0.158    5.71  2.5e-07
grptreat       19.621     20.055    0.98     0.33
fbef:grptreat  -0.132      0.177   -0.75     0.46
```

One factor, multiple levels – Analysis of Variance (ANOVA)

- Also fits the lm() linear model framework

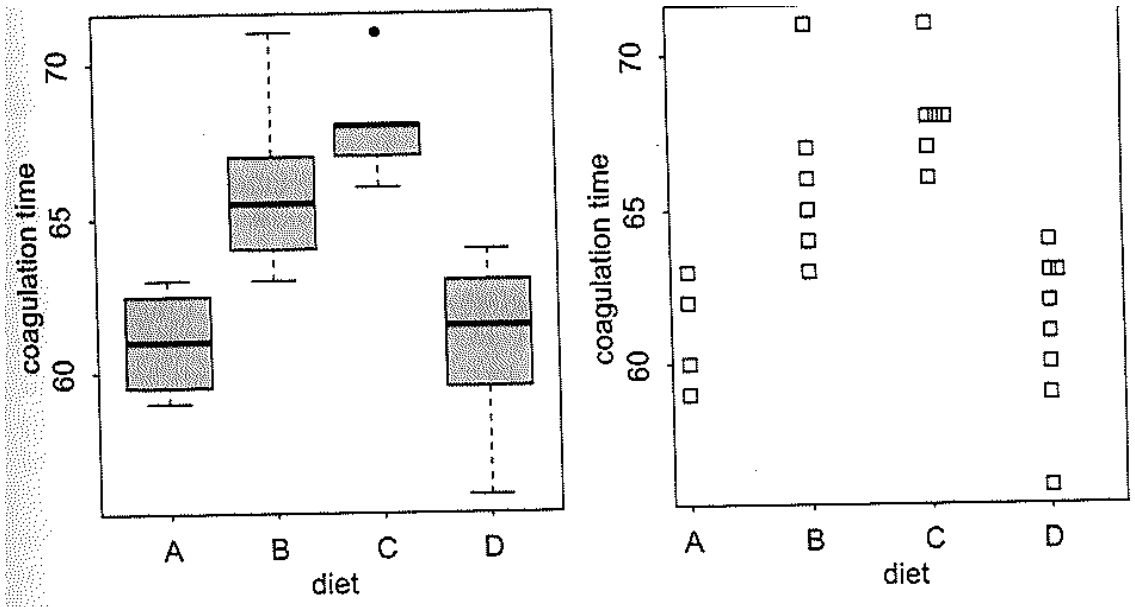


Figure 16.1: A boxplot on the left and a stripchart on the right showing the blood coagulation data.

```
lmod = lm(coag ~ diet, coagulation)
summary(lmod)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.10e+01   1.18e+00   51.55  < 1e-04
dietB       5.00e+00   1.53e+00    3.27  0.00380
dietC       7.00e+00   1.53e+00    4.58  0.00018
dietD       2.72e-15   1.45e+00    0.00  1.00000

Residual standard error: 2.37 on 20 degrees of freedom
Multiple R-squared:  0.671,        Adjusted R-squared:  0.621
F-statistic: 13.6 on 3 and 20 DF,  p-value: <1e-04
```

One factor, multiple levels – Analysis of Variance (ANOVA)

- Estimate the means and CIs for each level

```
library(emmeans)
emmeans(lmod, ~ diet)
 diet emmean    SE df lower.CL upper.CL
 A        61 1.183 20     58.5     63.5
 B        66 0.966 20     64.0     68.0
 C        68 0.966 20     66.0     70.0
 D        61 0.837 20     59.3     62.7

Confidence level used: 0.95
```

Checking ANOVA assumptions

- What are they?

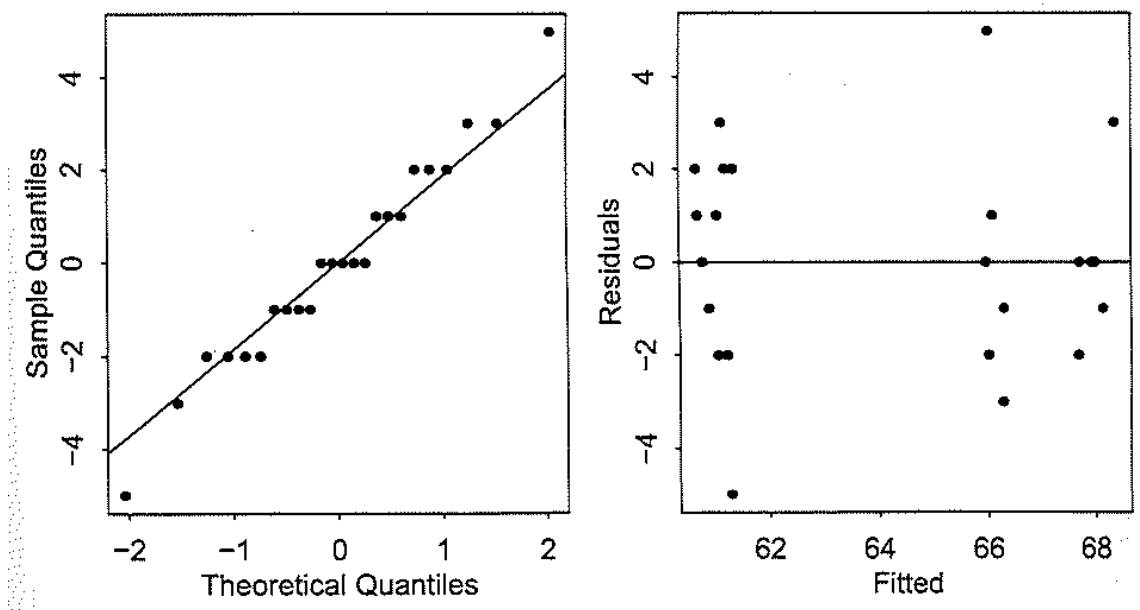


Figure 16.2: Diagnostics for the blood coagulation model.

Checking ANOVA assumptions

- Normality, homogeneity of variance, independence

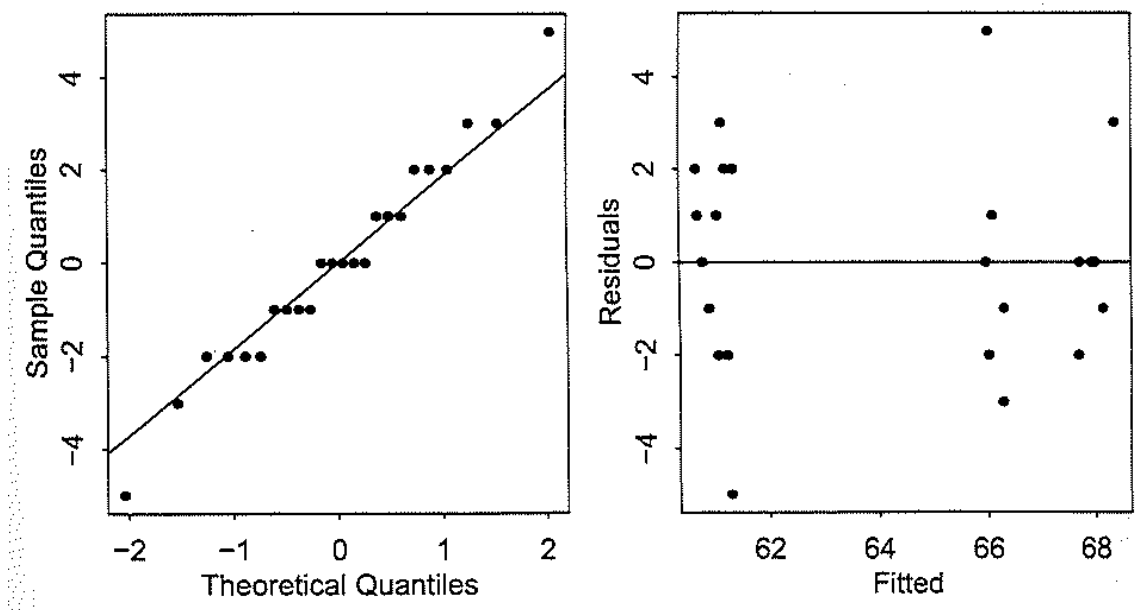


Figure 16.2: Diagnostics for the blood coagulation model.

Checking ANOVA assumptions

- Normality, homogeneity of variance, independence

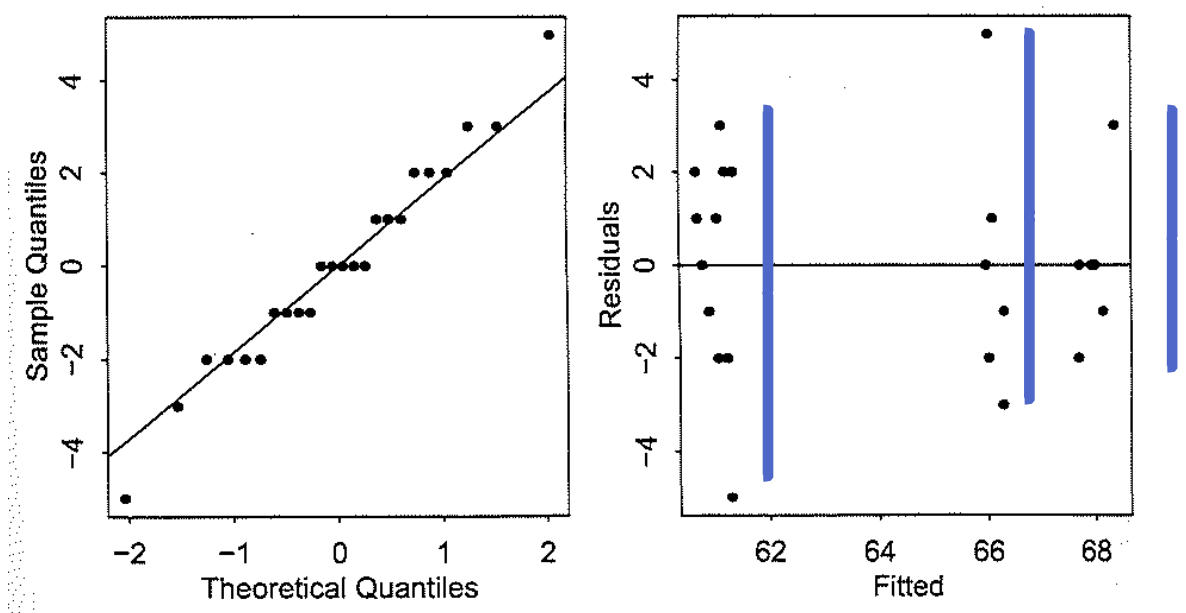


Figure 16.2: Diagnostics for the blood coagulation model.

Levene's Test
ANOVA on abs(residuals)

CARS package
leveneTest()

Data limitations

- What if we have two factors, each with multiple levels, and only one observation for each combination?

```
data(composite, package="faraway")
composite
  strength laser   tape
1    25.66   40W   slow
2    29.15   50W   slow
3    35.73   60W   slow
4    28.00   40W medium
5    35.09   50W medium
6    39.56   60W medium
7    20.65   40W   fast
8    29.79   50W   fast
9    35.66   60W   fast
```

- Can we fit a model with interactions?

Data limitations

- What if we treat laser as a continuous numeric variable?

```
composite$wattage <-as.numeric(composite$laser)
composite$wattage <- 40 + (composite$wattage-1)*10

#fit a linear model predicting strength from wattage and tape
m1 <- lm(strength ~ wattage + tape + wattage:tape,
data=composite)summary(m1)
```

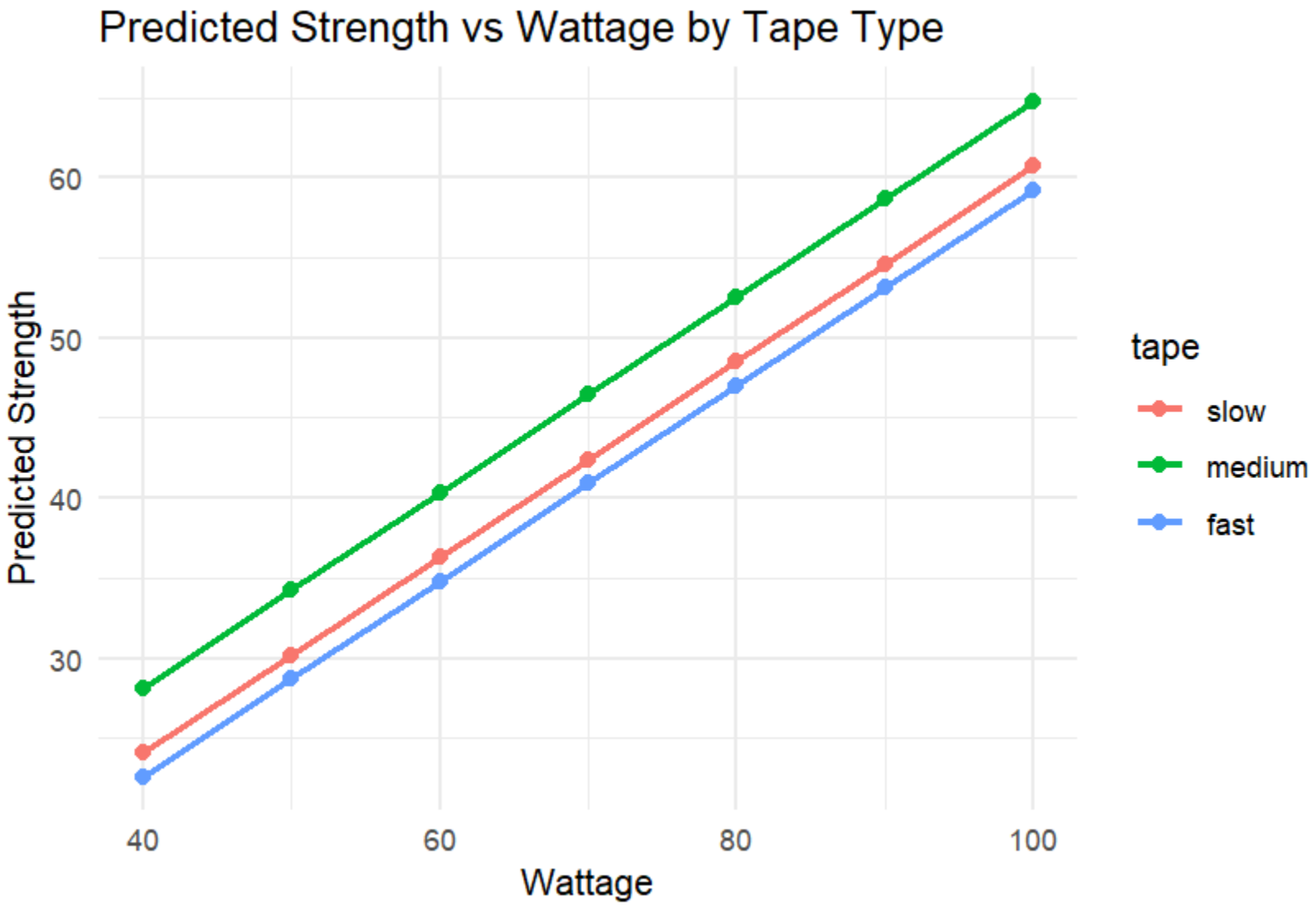
Data limitations

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          5.00500    4.39604   1.139   0.3376
wattage              0.50350    0.08677   5.803   0.0102 *
tapemedium           0.31167    6.21694   0.050   0.9632
tapefast           -13.83000    6.21694  -2.225   0.1126
wattage:tapemedium   0.07450    0.12271   0.607   0.5866
wattage:tapefast     0.24700    0.12271   2.013   0.1376
```

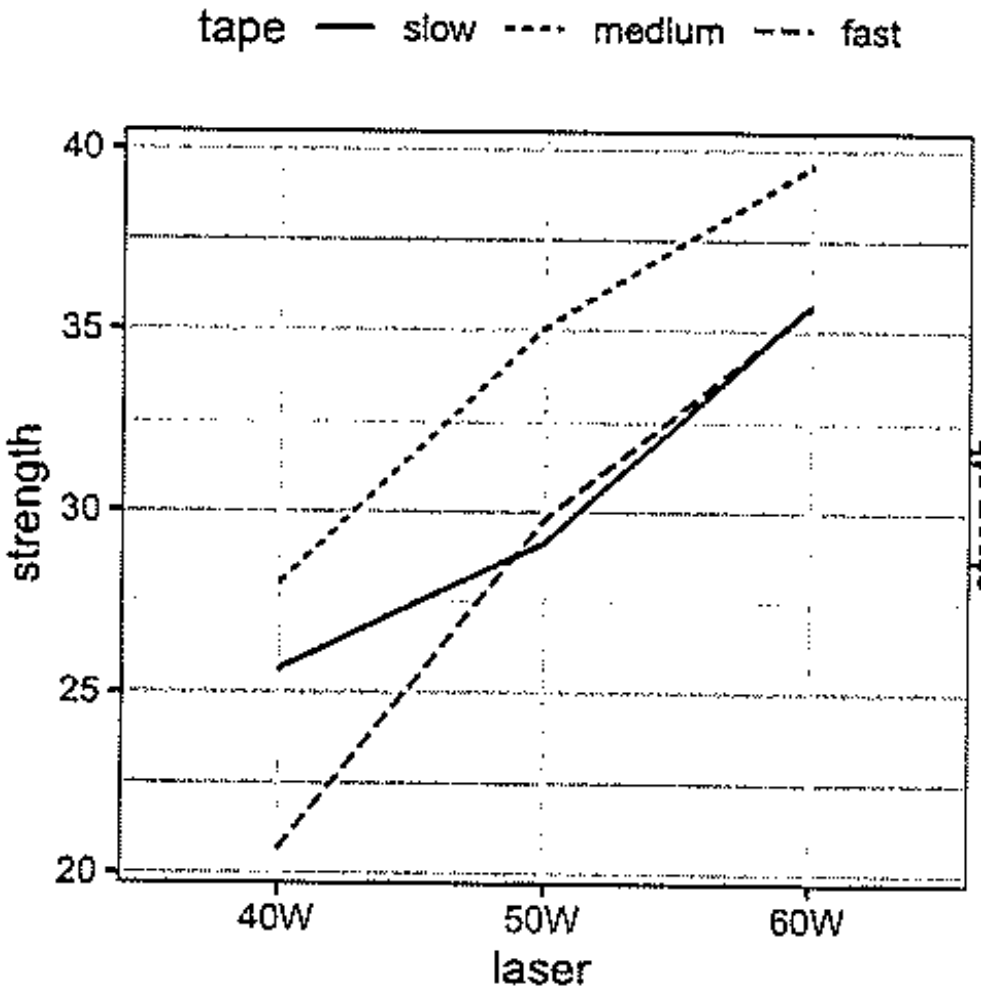
Data limitations

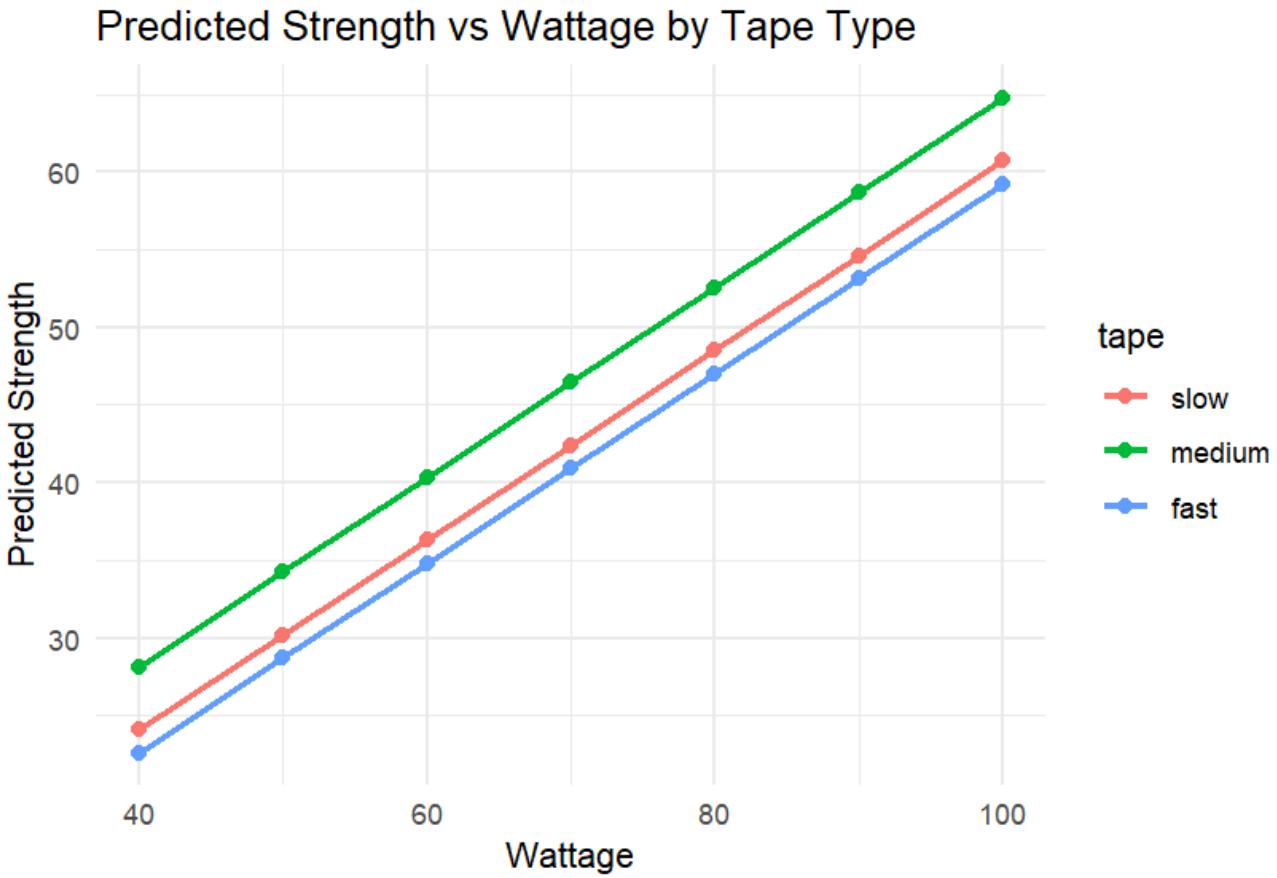
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.35333    3.13763  -0.113 0.914720
wattage      0.61067    0.06038  10.113 0.000162 ***
tapemedium   4.03667    1.20768   3.343 0.020496 *
tapefast    -1.48000    1.20768  -1.225 0.274961
```

Data limitations



Data limitations





Exercise: simulation and estimation of linear models with one categorical and one continuous predictor

- Using your code from last week's homework, add a two-level categorical X value and simulate sufficient data (say 100 observations total, split evenly between the two levels) to estimate the parameters of an ANCOVA model. You can decide whether there is an interaction or not (same slopes but different intercepts vs. different slopes and intercepts) or even whether the categorical variable matters at all.

- Write a brief (2-3 sentence) ecological scenario for these data and give the variables related names and units. Finish with an ecological question that fits the ANCOVA format. For example: "Does the depth distribution of pigmented and unpigmented Daphnia (a crustacean zooplankton) respond similarly to changes in light intensity?" There's a continuous response variable (depth), a continuous predictor (light intensity), and a two-level factor (pigmentation).

Exercise: simulation and estimation of linear models with one categorical and one continuous predictor

- Exchange data (.csv file) with a partner. Share only the data and the ecological background and question. Don't share your simulation code or parameters. That is, don't give them the answer!

- Fit alternative models to the data given to you, starting with the full model (both X variables and their interaction), and reducing down until only statistically significant terms remain.

- Write 2-3 sentences explaining the results in ecological terms and directly answering the question posed.