ASTA

The ASTA team

Contents

1 The regression problem

1.1 Example

- We will study the dataset in Agresti Table 13.1 available as `Income.txt` on the course website. We read in data in RStudio

```
Income <- read.delim("https://asta.math.aau.dk/datasets?file=Income.txt")
```

- We have a sample with measurements of 3 variables:

  - `y=income`: Quantitative variable, which is yearly income. This will be our response.
  - `x=education`: Quantitative predictor, which is the number of years of education.
  - `z=race`: Explanatory factor with levels `b`(black), `h`(hispanic) and `w`(white).

- We always start with some graphics (remember the function `gf_point` for plotting points and `gf_lm` for adding a regression line).

```
library(mosaic)
gf_point(inc ~ educ, color = ~race, data = Income) %>% gf_lm()
```
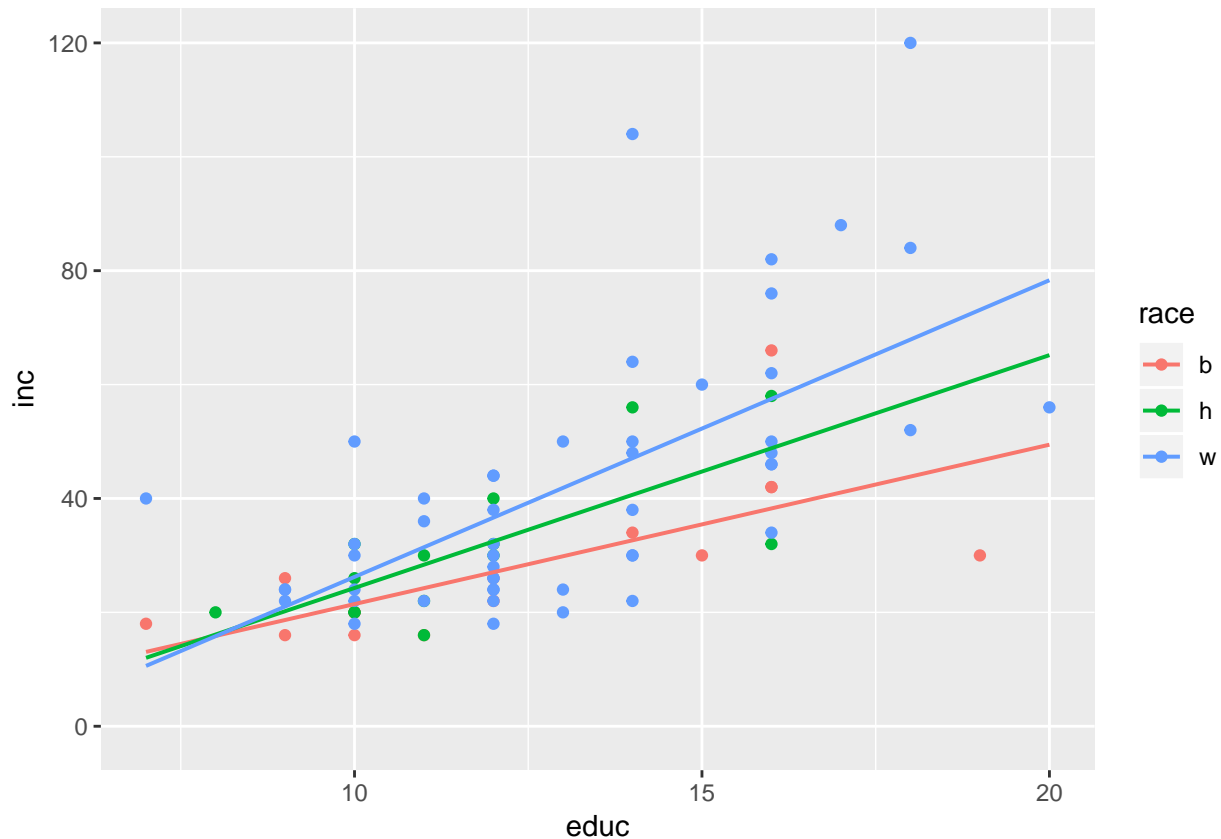
- An unclear picture, but a tendency to increasing income with increasing education.
- The trend lines for the three races are different. But is the difference significant? Or can the difference be explained by sampling variation?
- Such a regression with both qualitative and quantitative predictors is called an analysis of covariance (ANCOVA). When the model only contains qualitative predictors, the problem is known as analysis of variance (ANOVA) which is the topic of the next lecture.

2 Dummy coding

2.1 Dummy coding

- First, we will look at the model **without interaction**, i.e. the effect of `education` is the same for all races, which corresponds to parallel lines.

- We also have to introduce dummy coding of the factor z:

    - $z_1 = 1$ if `race=b` and zero otherwise
    - $z_2 = 1$ if `race=h` and zero otherwise

- This determines the regression model:

$$E(y|x, z) = \alpha + \beta x + \beta_1 z_1 + \beta_2 z_2$$

which corresponds to **parallel** regressions lines for each race.

- `w`: $(z_1 = 0, z_2 = 0)$ $E(y|x) = \alpha + \beta x$

- b: $(z_1 = 1, z_2 = 0)$ $E(y|x) = \alpha + \beta_1 + \beta x$.

- h: $(z_1 = 0, z_2 = 1)$ $E(y|x) = \alpha + \beta_2 + \beta x$.

- β_1 is the difference in `Intercept` between black and white.

- β_2 is the difference in `Intercept` between Hispanic and white.

2.2 Example

- We want to tell `R` that we want `w` as reference for race (default is lexicographical ordering, i.e. (`b, h, w`) and `b` would then be the reference):

```
Income$race <- relevel(Income$race, "w")
```

- Then we use `+` in the model formula to only have additive effects of `educ` and `race`, i.e. a model without interaction:

```
model1 <- lm(inc ~ educ + race, data = Income)
summary(model1)
```

```
##
## Call:
## lm(formula = inc ~ educ + race, data = Income)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.664  -9.622  -1.642   6.552  57.620
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -15.6635     8.4121  -1.862   0.0665 .
## educ          4.4317     0.6191   7.158 4.42e-10 ***
## raceb       -10.8744     4.4730  -2.431   0.0174 *
## raceh        -4.9338     4.7632  -1.036   0.3036
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.46 on 76 degrees of freedom
## Multiple R-squared:  0.462,  Adjusted R-squared:  0.4408
## F-statistic: 21.75 on 3 and 76 DF,  p-value: 2.853e-10
```

- The common slope to `educ` is estimated to be $\widehat{\beta} = 4.4316685$, with corresponding p-value=4.42×10^{-10} which is significantly different from zero.
- There is a clear positive effect of `educ` on `income`.
- The estimate for `w`-intercept is $\widehat{\alpha} = -15.6635$, which isn't significantly different from zero if we test at level 5% (this test is not really of interest).
- **The difference** between `b`- and `w`-intercept (`raceb`) is $\widehat{\beta}_1 = -10.8744$, which is significant with p-value=1.74%.
- There is no significant difference between `h`- and `w`-intercept.

```
plotModel(model1)
```
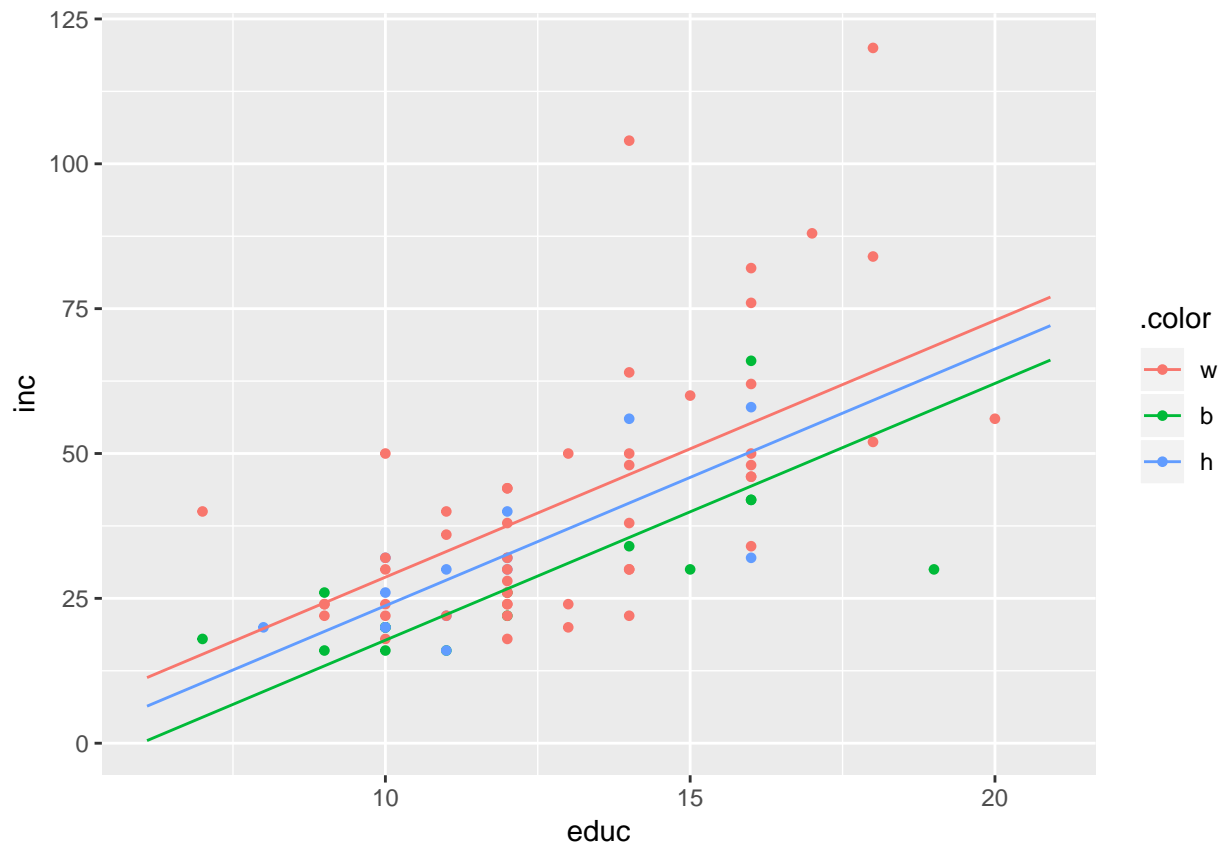

2.3 Example: Prediction equations

```
summary(model1)
```

```
##
## Call:
## lm(formula = inc ~ educ + race, data = Income)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.664  -9.622  -1.642   6.552  57.620
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -15.6635     8.4121  -1.862   0.0665 .
## educ          4.4317     0.6191   7.158 4.42e-10 ***
## raceb       -10.8744     4.4730  -2.431   0.0174 *
## raceh        -4.9338     4.7632  -1.036   0.3036
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.46 on 76 degrees of freedom
```

```
## Multiple R-squared:  0.462,   Adjusted R-squared:  0.4408
## F-statistic: 21.75 on 3 and 76 DF,  p-value: 2.853e-10
```

- Reference/baseline group (white):
$$\widehat{y} = -15.66 + 4.43x$$

- Black:
$$\widehat{y} = -15.66 - 10.87 + 4.43x = -26.54 + 4.43x$$

- Hispanic:
$$\widehat{y} = -15.66 - 4.93 + 4.43x = -20.60 + 4.43x$$

2.4   Agresti – summary

TABLE 13.4: Summary of Regression Equations and Parameters for Model with No Interaction, when Categorical Predictor Has Three Categories

Category	y-Intercept	Slope	Mean $E(y)$ at Fixed x	Difference From Mean of Category 3, Controlling for x
1	$\alpha + \beta_1$	β	$(\alpha + \beta_1) + \beta x$	β_1
2	$\alpha + \beta_2$	β	$(\alpha + \beta_2) + \beta x$	β_2
3	α	β	$\alpha + \beta x$	0
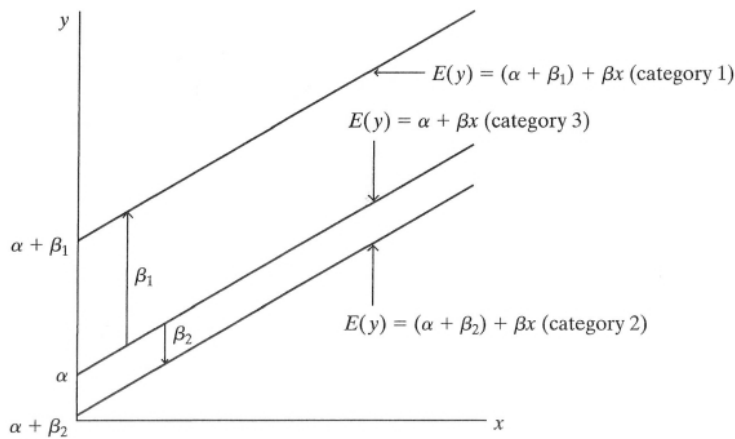
FIGURE 13.5: Graphic Portrayal of a Model with No Interaction, when the Categorical Predictor Has Three Categories

3   Model with interaction

3.1   Interaction

- In the following we will expand the model to include interaction between the effects of race and education on income. Before proceeding, let us recall what interaction means (and doesn't mean) in this context:
- Interaction between the effects of race and education on income does **not** mean that the values of education and race themselves are related or affect each other.

- Interaction between the effects of race and education on income means that the relationship between education and income depends on the value of race. I.e. for each fixed value of race the slope of the line relating education and income may have a different value.
- Often we just refer to this as "interaction between education and race" when it really should read "interaction between the effects of race and education on income".

3.2 Interaction

- We will expand the regression model, so we include interaction between x and z_1 respectively z_2:

$$E(y|x, z) = \alpha + \beta x + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_1 x + \beta_4 z_2 x.$$

- This yields a regression line for each race:
- w ($z_1 = 0, z_2 = 0$): $E(y|x) = \alpha + \beta x$
- b ($z_1 = 1, z_2 = 0$): $E(y|x) = \alpha + \beta_1 + (\beta + \beta_3)x$.
- h ($z_1 = 0, z_2 = 1$): $E(y|x) = \alpha + \beta_2 + (\beta + \beta_4)x$.
- β_1 is **the difference** in `Intercept` between black and white, while β_3 is **the difference** in `slope` between black and white.
- β_2 is **the difference** in`Intercept` between Hispanic and white, while β_4 is the difference in `slope` between Hispanic and white.

3.3 Example: Prediction equations

- When we use `*` in the model formula we include interaction between `educ` and `race`:

```
model2 <- lm(inc ~ educ * race, data = Income)
summary(model2)
```

```
##
## Call:
## lm(formula = inc ~ educ * race, data = Income)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.064  -9.448  -1.453   6.167  56.936
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -25.8688    10.4982  -2.464   0.0161 *
## educ          5.2095     0.7828   6.655  4.3e-09 ***
## raceb        19.3333    18.2928   1.057   0.2940
## raceh         9.2640    24.2797   0.382   0.7039
## educ:raceb   -2.4107     1.4177  -1.700   0.0933 .
## educ:raceh   -1.1208     2.0060  -0.559   0.5781
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.37 on 74 degrees of freedom
## Multiple R-squared:  0.4825, Adjusted R-squared:  0.4475
## F-statistic:  13.8 on 5 and 74 DF,  p-value: 1.618e-09
```

- Reference/baseline group (white):
$$\hat{y} = -25.87 + 5.21x$$

- Black:
$$\hat{y} = -25.87 + 19.33 + (5.21 - 2.41)x = -6.54 + 2.80x$$

- Hispanic:
$$\hat{y} = -25.87 + 9.26 + (5.21 - 1.12)x = -16.60 + 4.09x$$

3.4 Example: Individual tests

```
summary(model2)
```

```
##
## Call:
## lm(formula = inc ~ educ * race, data = Income)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.064  -9.448  -1.453   6.167  56.936
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -25.8688    10.4982  -2.464   0.0161 *
## educ          5.2095     0.7828   6.655  4.3e-09 ***
## raceb        19.3333    18.2928   1.057   0.2940
## raceh         9.2640    24.2797   0.382   0.7039
## educ:raceb   -2.4107     1.4177  -1.700   0.0933 .
## educ:raceh   -1.1208     2.0060  -0.559   0.5781
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.37 on 74 degrees of freedom
## Multiple R-squared:  0.4825, Adjusted R-squared:  0.4475
## F-statistic:  13.8 on 5 and 74 DF,  p-value: 1.618e-09
```

- **The difference** in slope between `b` and `w` (`educ:raceb`) is estimated to $\hat{\beta}_3 = -2.4107$. With p-value=9.33% there is no significant difference.
- Furthermore, there isn't any significant difference of slope between `h` and `w`. In other words there is probably not interaction between `educ` and `race`.
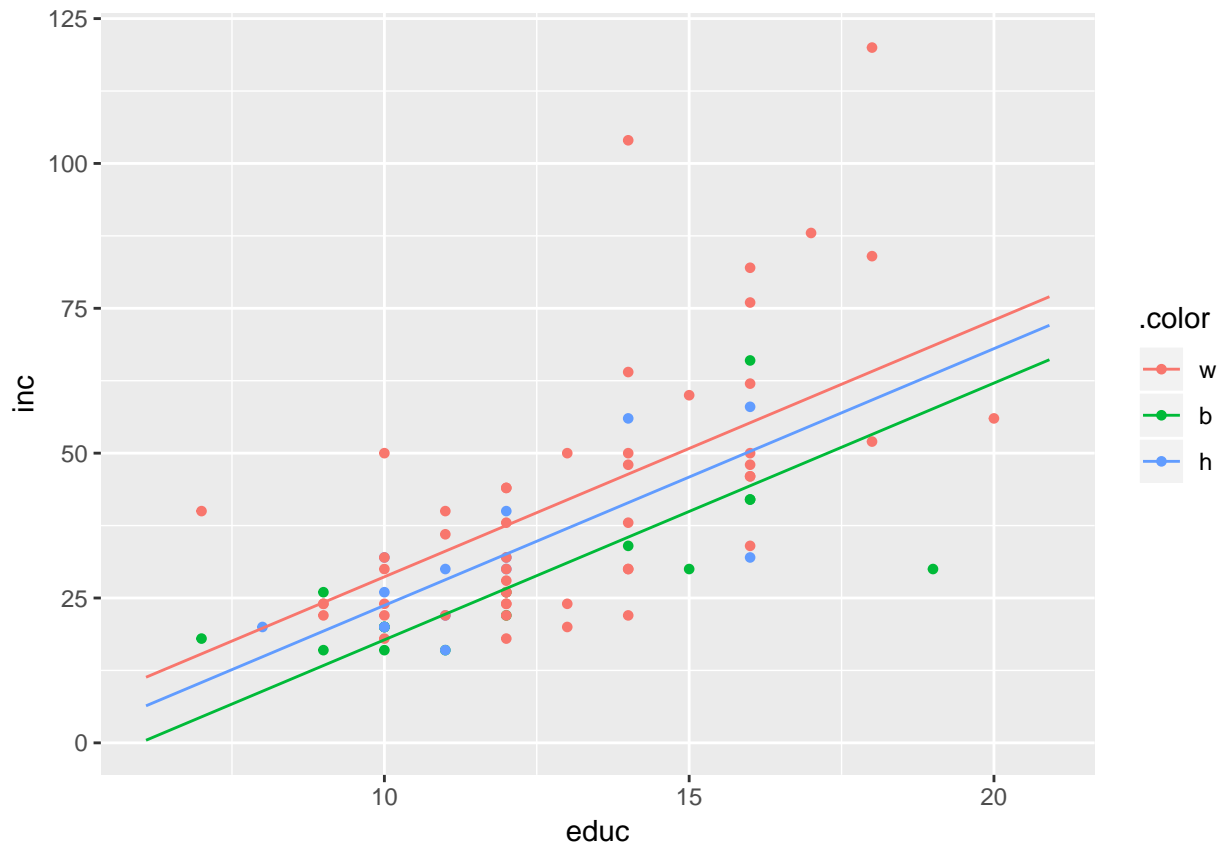
```
gf_point(inc ~ educ, color = ~ race, data = Income) %>% gf_lm()
```
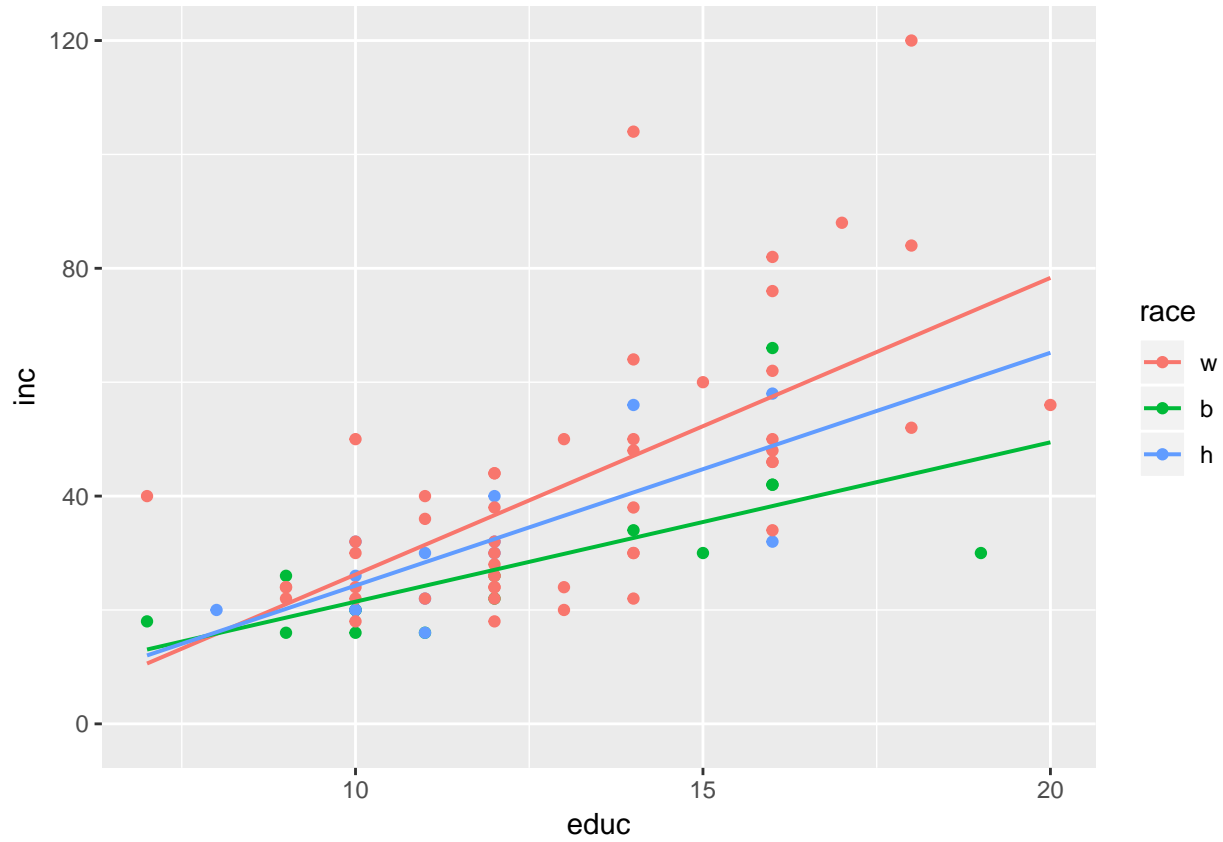

4 Test for no interaction

4.1 Test for no interaction

```
plotModel(model1)
```

```
plotModel(model2)
```
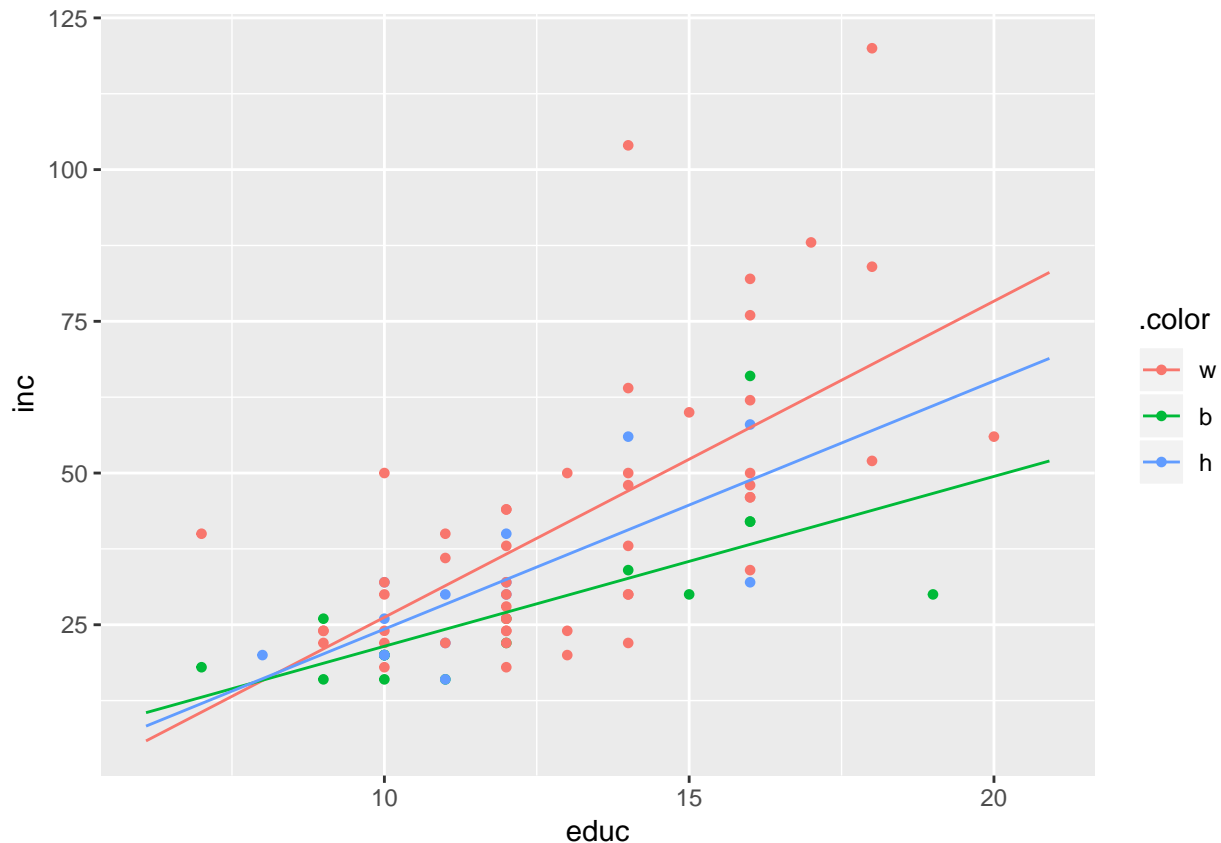
```
summary(model1)$r.squared
```

```
## [1] 0.4619906
```

```
summary(model2)$r.squared
```

```
## [1] 0.4824822
```

- Is `model2` significantly better than `model1`? I.e. is R^2 significantly higher for `model2`?

4.2 Hypothesis and test statistic

- The simpler `model1` is obtained from the more complicated `model2` by setting $\beta_3 = 0$ and $\beta_4 = 0$, so the null hypothesis "the simpler additive model describes data sufficiently well compared to the complicated interaction model" is really the simple mathematical hypothesis:

$$H_0 : \beta_3 = 0, \beta_4 = 0.$$

- We will look at the difference between R^2 for the two models, but as before (for multiple linear regression) we have to convert this to an F statistic which we can then calculate a p-value for.
- Formula for F_{obs} (no need to learn this by heart):

$$F_{obs} = \frac{(R_2^2 - R_1^2)/(\mathrm{df}_1 - \mathrm{df}_2)}{(1 - R_2^2)/\mathrm{df}_2}$$

where df_1 and df_2 are n minus the number of model parameters for the two models (i.e. 80-4=76 and 80-6=74 in our case).

- The formula for F_{obs} can be rewritten in terms of sums of squared errors (SSE) for each model (no need to memorize it):

$$F_{obs} = \frac{(SSE_1 - SSE_2)/(\mathrm{df}_1 - \mathrm{df}_2)}{(SSE_2)/\mathrm{df}_2}.$$

- In the literature SSE is sometimes denoted by RSS for **Residual Sums of Squares**; i.e SSE = RSS.

4.3 Test for no interaction in R

- In R the calculations are done using `anova`:

```
anova(model1, model2)
```

```
## Analysis of Variance Table
##
## Model 1: inc ~ educ + race
## Model 2: inc ~ educ * race
##   Res.Df   RSS Df Sum of Sq     F Pr(>F)
## 1     76 18164
## 2     74 17472  2    691.84 1.465 0.2377
```

- The F-test for dropping the interaction `educ:race` has F-value=1.465, which in no way is significant with p-value=23.77%.

5 Hierarchy of models

5.1 Hierarchy of models

- `Interaction`: The most general model with main effects `educ` and `race` and interaction `educ:race`:

```
Interaction <- lm(inc ~ educ * race, data = Income)
```

- `MainEffects`: The model where there are additive effects of `educ` and `race`.

```
MainEffects <- lm(inc ~ educ + race, data = Income)
```

- `educEff`: Model where there only is an effect of `educ` (simple lin. reg.).

```
educEff <- lm(inc ~ educ, data = Income)
```

- `raceEff`: Model where there only is an effect of `race` (a different mean for each group – more on this in the ANOVA lecture).

```
raceEff <- lm(inc ~ race, data = Income)
```

- We can, corresponding to Agresti Table 13.10, make F-tests for 3 pairwise comparisons of models.

5.2 Example

- Comparing `MainEffects` and `Interaction` is what we have already done.

```
anova(MainEffects, Interaction)
```

```
## Analysis of Variance Table
##
## Model 1: inc ~ educ + race
## Model 2: inc ~ educ * race
##   Res.Df   RSS Df Sum of Sq     F Pr(>F)
## 1     76 18164
## 2     74 17472  2    691.84 1.465 0.2377
```

- We recognize $F = 1.465$ with p-value=23.77%, i.e. `model2` isn't significantly better than `model1`. So no `educ:race` interaction.

- In the same manner we can compare `educEff` and `MainEffects`. I.e. we investigate whether the effect of `race` can be left out.

```
anova(educEff, MainEffects)
```

```
## Analysis of Variance Table
##
## Model 1: inc ~ educ
## Model 2: inc ~ educ + race
##   Res.Df   RSS Df Sum of Sq     F  Pr(>F)
## 1     78 19625
## 2     76 18164  2    1460.6 3.0556 0.05292 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- If any, the effect of `race` is weak with p-value=5.292%.

- Finally, we compare `raceEff` and `MainEffects`. Clearly `educ` cannot be left out (P-value=4.422×10^{-10}).

```
anova(raceEff, MainEffects)
```

```
## Analysis of Variance Table
##
## Model 1: inc ~ race
## Model 2: inc ~ educ + race
##   Res.Df   RSS Df Sum of Sq     F     Pr(>F)
## 1     77 30410
## 2     76 18164  1    12245 51.235 4.422e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5.3 Example

- The methods generalize to models with more than 2 predictors.
- We return to the dataset `Ericksen`, where we study the response `crime`:

```
Ericksen <- read.delim("https://asta.math.aau.dk/datasets?file=Ericksen.txt")
model <- lm(crime ~ city * highschool + city * poverty, data = Ericksen)
```

- The variables are:
  - `crime`: Quantitative variable
  - `city`: `city` or `state`
  - `highschool`: Quantitative variable
  - `poverty`: Quantitative variable

- The model has 3 predictors with main effects and includes
  - interaction between `city` and `highschool`
  - interaction between `city` and `poverty`.

```
summary(model)
```

```
##
## Call:
## lm(formula = crime ~ city * highschool + city * poverty, data = Ericksen)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.741  -8.745  -1.557   7.820  47.470
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          61.1456    18.1254   3.373 0.001305 **
## citystate            18.1526    20.4131   0.889 0.377413
## highschool           -1.5711     0.6062  -2.592 0.011979 *
## poverty               5.3105     1.4333   3.705 0.000463 ***
## citystate:highschool  0.7025     0.7327   0.959 0.341523
## citystate:poverty    -5.1862     1.6619  -3.121 0.002773 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.16 on 60 degrees of freedom
## Multiple R-squared:  0.6577, Adjusted R-squared:  0.6292
## F-statistic: 23.06 on 5 and 60 DF,  p-value: 7.748e-13
```

- There isn't significant (p-value=34.1523%) interaction between `city` and `highschool`.
- I.e. the effect of `highschool` on crime is the same in metropolitan areas (`city=city`) and the non-metropolitan areas (`city=state`).
- There is clearly (p-value=0.2773%) interaction between `city` and `poverty`.
- I.e. the effect of `poverty` on crime is different in metropolitan and non-metropolitan areas.
- For `city=state`, the effect of `poverty` (on crime) is smaller than in the major cities.
- Hence, poverty has larger effect on crime in the major cities than in the states outside the major cites.

5.4 Multicollinearity and variance inflation factors

- Ideally the predictors in linear regression should be **uncorrelated**, which is almost never the case.
- The consequence of the two predictors being correlated (**collinear**), is that the uncertainty of the parameter estimates increase (because the squared standard error increases) by a factor commonly called the variance inflation factor (VIF).

- If multiple pairs of predictors are collinear, we say that the model suffers from **multicollinearity**.
- If we have a model with p predictors, then the VIF of x_j is:

$$\text{VIF}_j = \frac{1}{1 - R_j^2},$$

  where R_j^2 is the multiple R^2 value of a model using x_j as a response and the remaining $p-1$ predictors as explanatory variables.
- The larger VIF_j is, the higher the collinearity between x_j and the remaining predictors is.
- **Rule of thumb:** If a VIF is larger than 10 the collinearity is too high.

6 One way analysis of variance

6.1 Example

- The data set `chickwts` is available in `R`, and on the course webpage.
- Newly hatched chicks were randomly allocated into six groups, and each group was given a different feed supplement.
- Their weights in grams after six weeks are given along with feed types, i.e. we have a sample with corresponding measurements of 2 variables:
  - `weight`: a numeric variable giving the chick weight.
  - `feed`: a factor giving the feed type.
- Always start with some graphics:

```
library(mosaic)
gf_boxplot(weight ~ feed, data = chickwts)
```
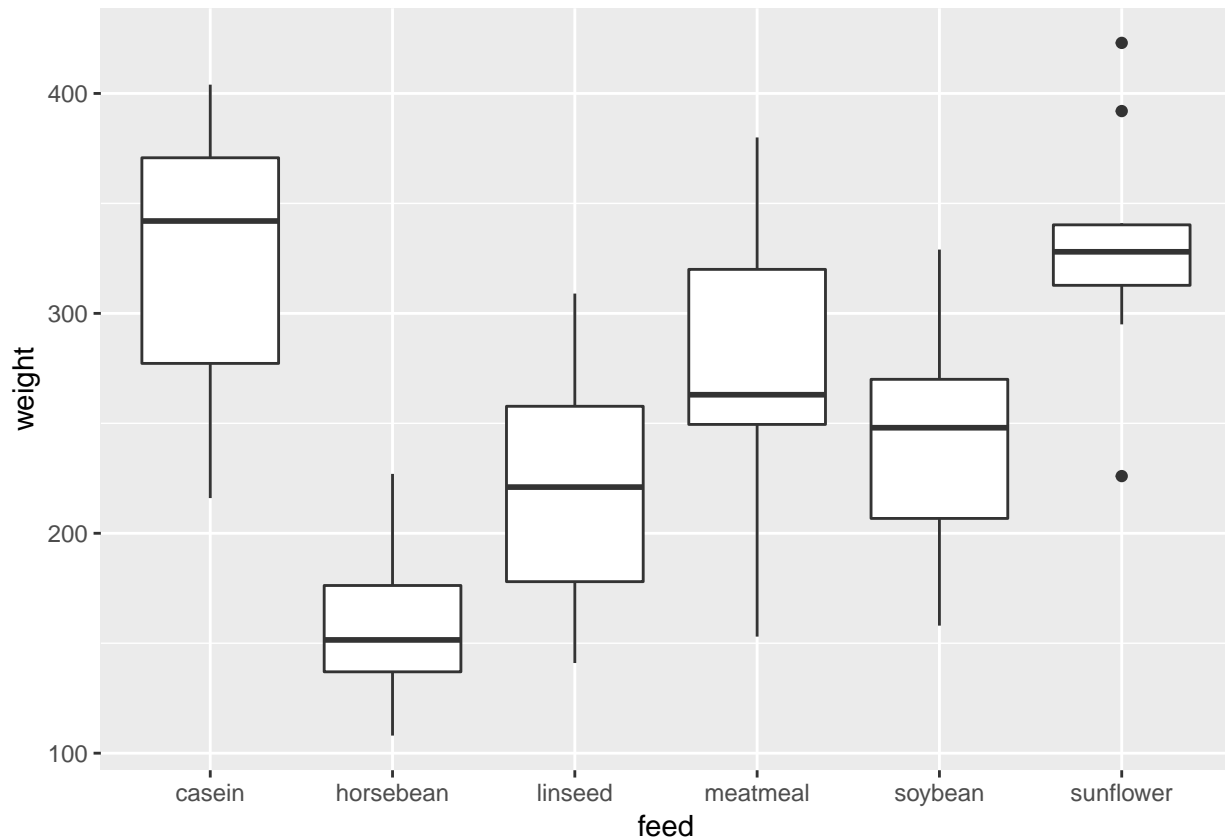
6.2 The ANOVA Model

- We measure the response y which in this case is `weight`.
- We want to study the effect of the factor x on y. In this case $x =$ `feed` and divides the sample in $g = 6$ groups.
- The mean responses within the groups are denoted $\mu_1, \mu_2, \ldots, \mu_g$.
- We will assume that

  - $y = \mu_x + \epsilon$, when y is a response in group x
  - ϵ are a sample from a population with mean zero and standard deviation σ.
  - The standard deviation for the population in each group is the same and equals σ
  - The response variable, y, is normal distributed within each group.

- The ANOVA test is a *test of independence* between the quantitative response variable and the qualitative explanatory variables.

7 Estimation of mean values

7.1 Estimates

- Least squares estimates for population means $\widehat{\mu}_x$ is given by the average of the response measurements in group x.
- For a given measured response y we let $\widehat{y}$ denote the model's prediction of y, i.e.

$$\widehat{y} = \widehat{\mu}_x$$

  if y is a response for an observation in group x.
- We use `mean` to find the mean, for each group:

```
mean(weight ~ feed, data = chickwts)
```

```
##    casein horsebean   linseed  meatmeal   soybean sunflower
##  323.5833  160.2000  218.7500  276.9091  246.4286  328.9167
```

- We can e.g. see that $\widehat{y} = 323.6$, when `feed=casein` but $\widehat{y} = 160.2$, when `feed=horsebean`.
- Is it a significant difference?

7.2 Contrast coding

- In many cases there is a group corresponding to "no treatment" and we are interested in the effect of different treatments.
- In this example we only have different `feeds`, which are sorted in lexicographical order by R, so `casein` is the reference.
- We can specify the model via:

  - `Intercept` corresponding to the mean response for the reference (`casein`).
  - For each of the other groups we have a **contrast**, which measures **the difference** between the mean value for that group and the reference group.

- For a given contrast we can calculate standard error, t-score and p-value, and thereby investigate whether there is a difference between this group and the reference group.
- In Agresti this is referred to as using **dummy variables**.

7.3 Example

```
model <- lm(weight ~ feed, data = chickwts)
summary(model)
```

```
##
## Call:
## lm(formula = weight ~ feed, data = chickwts)
##
## Residuals:
##      Min      1Q   Median       3Q      Max
## -123.909  -34.413    1.571   38.170  103.091
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    323.583     15.834  20.436  < 2e-16 ***
## feedhorsebean -163.383     23.485  -6.957 2.07e-09 ***
## feedlinseed   -104.833     22.393  -4.682 1.49e-05 ***
## feedmeatmeal   -46.674     22.896  -2.039 0.045567 *
## feedsoybean    -77.155     21.578  -3.576 0.000665 ***
## feedsunflower    5.333     22.393   0.238 0.812495
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.85 on 65 degrees of freedom
## Multiple R-squared:  0.5417, Adjusted R-squared:  0.5064
## F-statistic: 15.36 on 5 and 65 DF,  p-value: 5.936e-10
```

- We get information about contrasts and their significance:
- `Intercept` corresponding to `casein` has `weight` different from zero ($p < 2 \times 10^{-16}$).
- Weight difference between `casein` and `horsebean` is extremely significant (p=2×10^{-9}).
- There is no significant weight difference between `casein` and `sunflower` (p=81%).

8 Overall test for effect

8.1 Graphical representation of models

- We have two alternative explanations of the data.
- Simple model with one parameter (mean): "The feed type doesn't matter. The weight is just random around a common mean value".
- Complex model with six parameters (means): "The feed type is important. For each feed type we get a different mean value and the weights are random around these values."
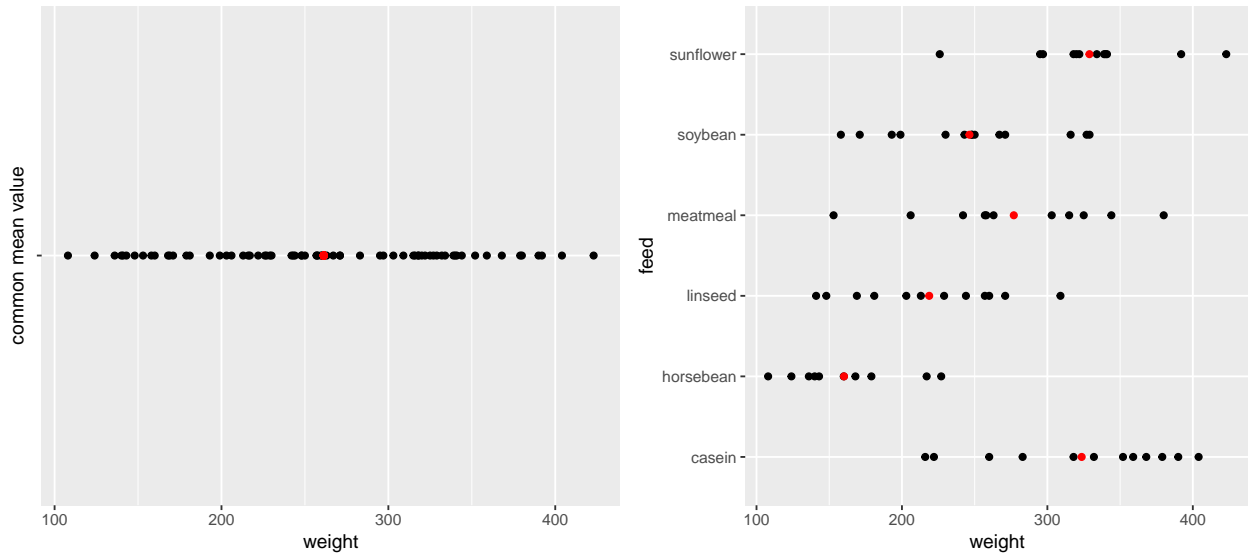
8.2 Hypotheses and test statistic

- Is the complex model significantly better (i.e. is there any effect of the explanatory grouping variable)? We can write the corresponding hypotheses in two different ways

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_g \quad \text{against} \quad H_a : \text{At least 2 of the population means are different}$$

- Alternatively

$$H_0 : \text{ All contrasts are equal to zero.} \quad H_a : \text{ At least one contrast is non-zero.}$$

- We will (indirectly) use R^2 to do the test. If it is large, the complex model has good predictive power compared to the simple model. To judge significance we use

$$F_{obs} = \frac{(n-g)R^2}{(g-1)(1-R^2)} = \frac{(TSS - SSE)/(g-1)}{SSE/(n-g)}.$$

- Large values of R^2 implies large values of F_{obs}, which points to the alternative hypothesis.
- I.e. when we have calculated the observed value F_{obs}, then we have to find the probability that a new experiment would result in a larger value.

8.3 Interpretation of F statistic - Variance between/within groups

- It can be shown that the numerator of F_{obs} is a measure of **the variance between the groups**, i.e. how much "boxes" vary around the total average (the red line).

- Likewise it can be shown the denominator of F_{obs} is a measure for **the variance within groups**, i.e. how "tall" the boxes in the boxplot are.

18
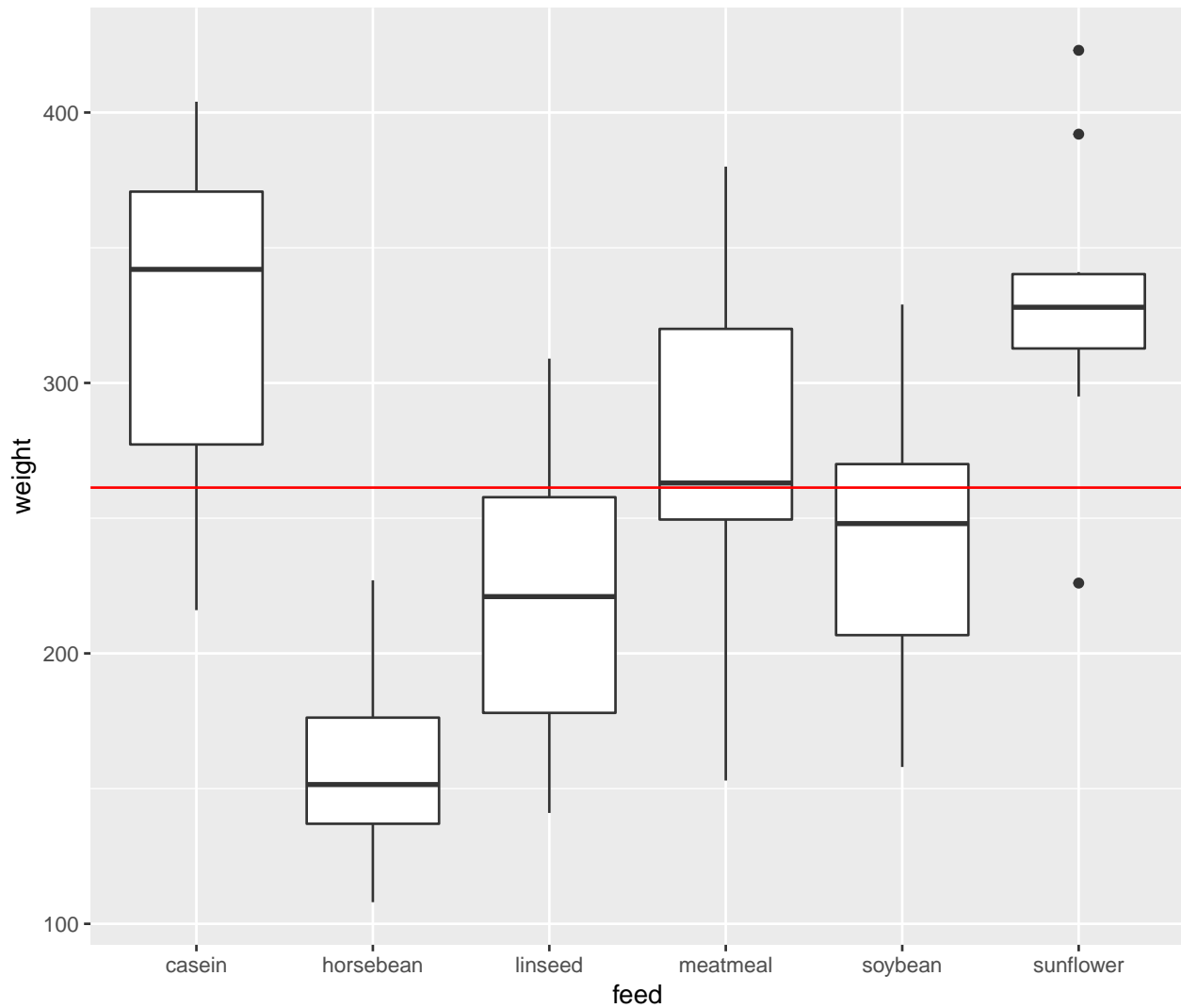
- If the boxes' deviations from the red line are to be explained by randomness, then the two types of variances should be of same magnitude. This is measured by the F-test statistic, which can be stated as

$$F_{obs} = \frac{\text{variance between groups}}{\text{variance within groups}}$$

8.4   Example

```
model <- lm(weight ~ feed, data = chickwts)
summary(model)
```

```
##
## Call:
## lm(formula = weight ~ feed, data = chickwts)
##
## Residuals:
##      Min      1Q   Median      3Q      Max
```

```
## -123.909  -34.413    1.571   38.170  103.091
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     323.583     15.834  20.436  < 2e-16 ***
## feedhorsebean  -163.383     23.485  -6.957 2.07e-09 ***
## feedlinseed    -104.833     22.393  -4.682 1.49e-05 ***
## feedmeatmeal    -46.674     22.896  -2.039 0.045567 *
## feedsoybean     -77.155     21.578  -3.576 0.000665 ***
## feedsunflower     5.333     22.393   0.238 0.812495
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.85 on 65 degrees of freedom
## Multiple R-squared:  0.5417, Adjusted R-squared:  0.5064
## F-statistic: 15.36 on 5 and 65 DF,  p-value: 5.936e-10
```

- The last line gives us the value of $F_{obs} = 15.36$ and the corresponding p-value (5.9×10^{-10}). Clearly there is a significant difference between the types of `feed`.

9 Two way analysis of variance

9.1 Additive effects

- The data set `ToothGrowth` is available in **R** and on the webpage. For more info about this data, use `?ToothGrowth`.

- The data describes the tooth length in guinea pigs where some receive vitamin C treatment and others are given orange juice in different dosage.

- A total of 60 observations on 3 variables.

   - `[,1]` `len` The tooth length
   - `[,2]` `supp` The type of the supplement (`OJ` or `VC`)
   - `[,3]` `dose` The dosage (`LO, ME, HI`)

- We will study the response `len` with the predictors `supp` and `dose`.

- At first we look at the model with additive effects

   - `len`=μ + `"effect of supp"`+ `"effect of dose"` + `error`

- This is also called the main effects model since it does not contain interaction terms.

- The parameter μ corresponds to the `Intercept` and is the mean tooth length in the reference group (supp `OJ`, dose `LO`).

- The effect of `supp` is the difference in mean when changing from `OJ` to `VC`.

- The effect of `dose` is the difference in mean when changing from `LO` to either`ME` or `HI`.

9.2 Dummy coding

- Let us introduce dummy variables:

    - $s_C = 1$ if supp VC and zero otherwise.
    - $d_M = 1$ if dose is ME and zero otherwise.
    - $d_H = 1$ if dose is HI and zero otherwise.

- Then we state the model

$$\text{length} = \mu + \beta_1 s_C + \beta_2 d_M + \beta_3 d_H + \text{error}.$$

- Interpretation:

    - μ is the expected tooth length when supp is OJ and dose is LO ($s_C = d_M = d_H = 0$)).
    - β_1 is the effect of supplement OJ to VC ($s_C = 1$).
    - β_2 is the effect of increasing dosage from LO to ME ($d_M = 1$).
    - β_3 is the effect of increasing dosage from LO to HI ($d_H = 1$).

9.3 Main effect model in R

- The main effects model is fitted by

```
MainEff <- lm(len ~ supp + dose, data = ToothGrowth)
summary(MainEff)
```

```
##
## Call:
## lm(formula = len ~ supp + dose, data = ToothGrowth)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.085  -2.751  -0.800   2.446   9.650
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   12.4550     0.9883  12.603  < 2e-16 ***
## suppVC        -3.7000     0.9883  -3.744 0.000429 ***
## doseME         9.1300     1.2104   7.543 4.38e-10 ***
## doseHI        15.4950     1.2104  12.802  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.828 on 56 degrees of freedom
## Multiple R-squared:  0.7623, Adjusted R-squared:  0.7496
## F-statistic: 59.88 on 3 and 56 DF,  p-value: < 2.2e-16
```

- The model has 4 parameters.
- The F test at the end compares with the (null) model with only one overall mean parameter. Does it seem like supp and dose has an additive effect?

9.4   Testing effect of supp

- Alternative model without effect of supp:

```
doseEff <- lm(len ~ dose, data = ToothGrowth)
```

- We can compare R^2 to see if `doseEff` (Model 1) is sufficient to explain the data compared to `MainEff` (Model 2). This is done by converting to F-statistic:

$$F_{obs} = \frac{(R_2^2 - R_1^2)/(df_1 - df_2)}{(1 - R_2^2)/df_2} = \frac{(SSE_1 - SSE_2)/(df_1 - df_2)}{(SSE_2)/df_2}.$$

- In **R** the calculations are done using `anova`:

```
anova(doseEff, MainEff)
```

```
## Analysis of Variance Table
##
## Model 1: len ~ dose
## Model 2: len ~ supp + dose
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     57 1025.78
## 2     56  820.43  1    205.35 14.017 0.0004293 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

9.5   Testing effect of dose

- Alternative model without effect of dose:

```
suppEff <- lm(len ~ supp, data = ToothGrowth)
anova(suppEff, MainEff)
```

```
## Analysis of Variance Table
##
## Model 1: len ~ supp
## Model 2: len ~ supp + dose
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     58 3246.9
## 2     56  820.4  2    2426.4 82.811 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```
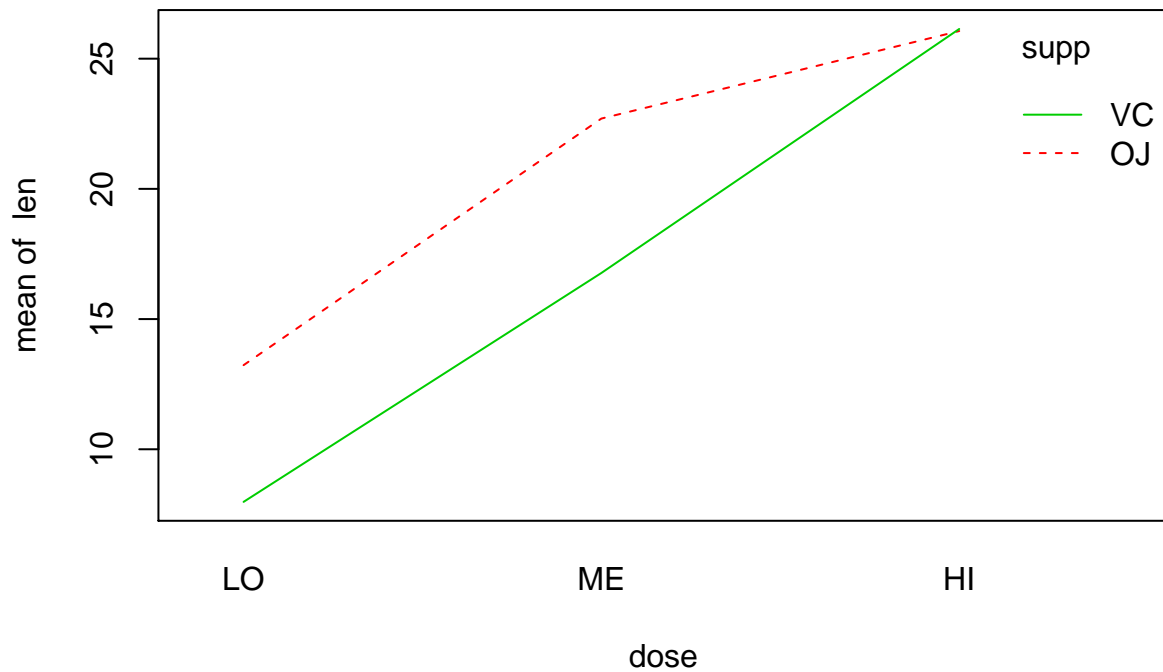
10   Interaction

10.1   Example

- We will extend the model by introducing an interaction between `supp` and `dose`.
- A graphical check for no interaction in the main effects model:
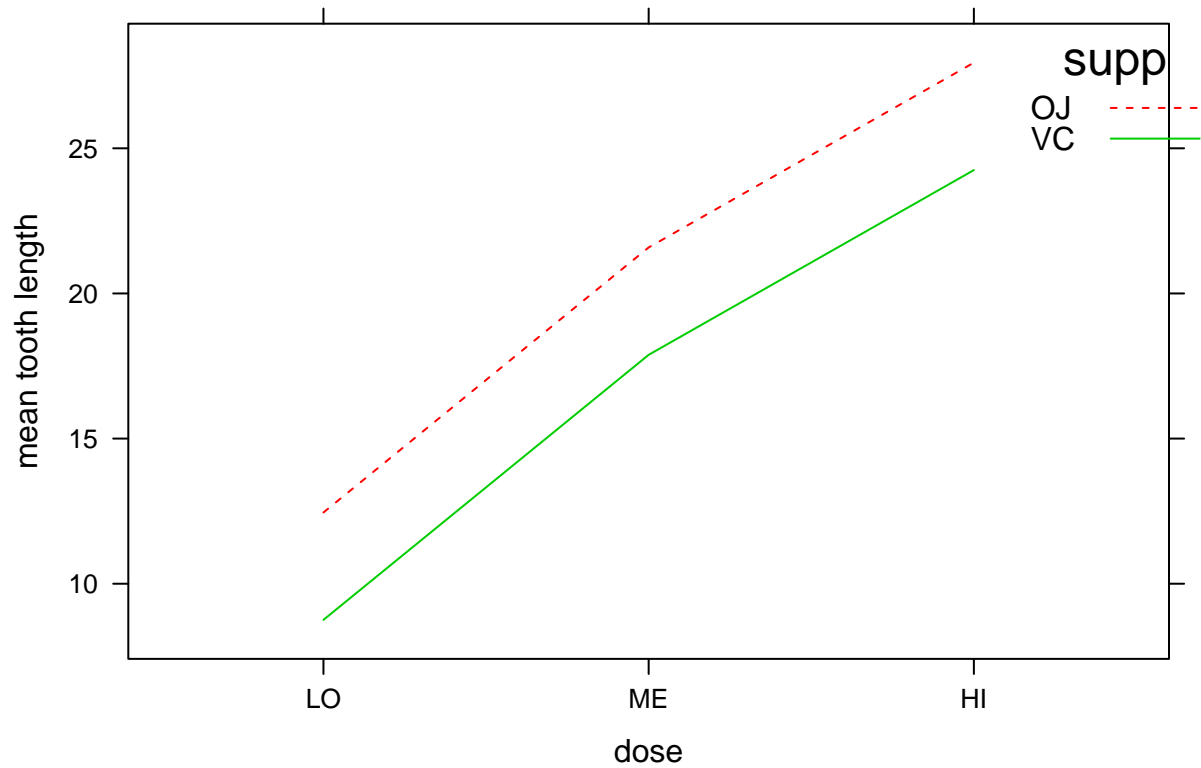
- Interaction plot:

```r
with(ToothGrowth, interaction.plot(dose, supp, len, col = 2:3))
```


- For each of the supplement types we plot the average number of tooth length as a function of dosage.
- If the main effects model is correct then the difference between supplements is the same for all levels of dosage, i.e. the curves should be parallel - except for noise.

- This does not seem to be the case.

10.2 Dummy coding

- The extended model can be formulated as

$$\texttt{length} = \mu + \beta_1 s_C + \beta_2 d_M + \beta_3 d_H + \beta_4 s_C d_M + \beta_5 s_C d_H + \texttt{error}$$

- Interpretation:
    - μ is the expected tooth length for supp OJ and dose LO ($s_C = d_M = d_H = 0$).
    - β_1 is the effect of changing from supp OJ to VC, dose is LO ($s_C = 1, d_M = d_H = 0$).
    - β_2 is the effect of increasing dose from LO to ME, when supp is OJ ($s_C = 0, d_M = 1$).
    - β_3 is the effect of increasing dose from LO to HI, when supp is OJ ($s_C = 0, d_H = 1$).
    - β_4 is an additional effect of both changing from supp OJ to VC and increasing dose from LO to ME ($s_C = 1, d_M = 1$)
    - β_5 is an additional effect of both changing from supp OJ to VC and increasing dose from LO to HI ($s_C = 1, d_H = 1$)

10.3 Example

- We fit the interaction model by changing plus to multiply in the model expression from before:

```
Interaction <- lm(len ~ supp*dose, data = ToothGrowth)
```

- Now we can think of an experiment with 6 groups corresponding to each combination of the predictors.
- Looking at the group averages it looks like, the supplement types behave quite differently depending on dose:

```
mean(len ~ supp + dose, data = ToothGrowth)
```

```
## OJ.LO VC.LO OJ.ME VC.ME OJ.HI VC.HI
## 13.23  7.98 22.70 16.77 26.06 26.14
```

- But is that significant?

```
anova(MainEff, Interaction)
```

```
## Analysis of Variance Table
##
## Model 1: len ~ supp + dose
## Model 2: len ~ supp * dose
##   Res.Df    RSS Df Sum of Sq     F  Pr(>F)
## 1     56 820.43
## 2     54 712.11  2    108.32 4.107 0.02186 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- With a p-value of 2.1860269% there is a significant interaction supp:dose, i.e. the lack of parallel curves in the interaction plot is significant.

```
summary(Interaction)
```

```
##
## Call:
## lm(formula = len ~ supp * dose, data = ToothGrowth)
##
## Residuals:
##     Min     1Q Median     3Q    Max
##   -8.20  -2.72  -0.27   2.65   8.27
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     13.230      1.148  11.521 3.60e-16 ***
## suppVC          -5.250      1.624  -3.233  0.00209 **
## doseME           9.470      1.624   5.831 3.18e-07 ***
## doseHI          12.830      1.624   7.900 1.43e-10 ***
## suppVC:doseME   -0.680      2.297  -0.296  0.76831
## suppVC:doseHI    5.330      2.297   2.321  0.02411 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.631 on 54 degrees of freedom
## Multiple R-squared:  0.7937, Adjusted R-squared:  0.7746
## F-statistic: 41.56 on 5 and 54 DF,  p-value: < 2.2e-16
```

- The additional effect of both changing from `supp` `OJ` to `VC` and increasing `dose` from `LO` to `HI` (β_5=`suppVC:doseHI`) is highly significant.