Analysis of Covariance

*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()
```
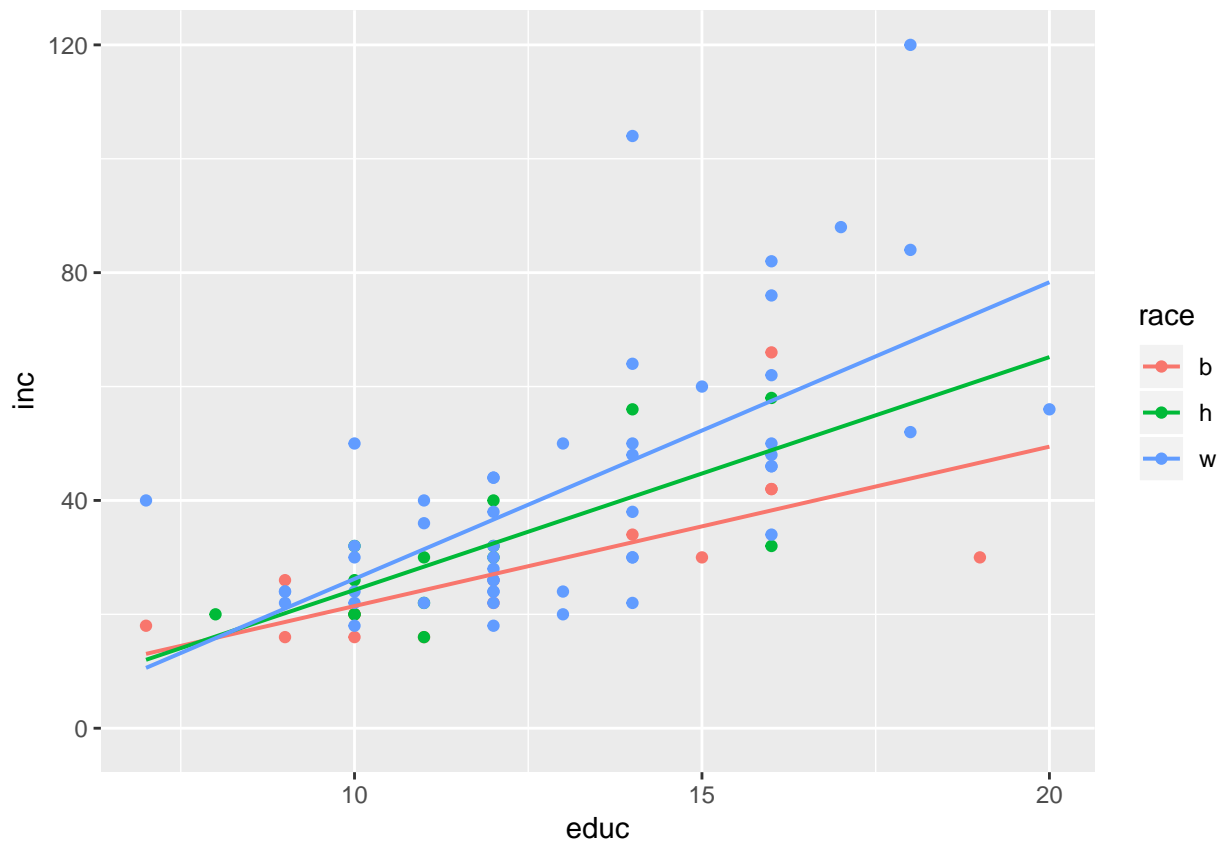

- An unclear picture, but a tendency to increasing income with increasing education.
- The trend lines for the three races are different. But is the difference significant? Or can the difference be explained by sampling variation?
- Such a regression with both qualitative and quantitative predictors is called an analysis of covariance (ANCOVA). When the model only contains qualitative predictors, the problem is known as analysis of variance (ANOVA) which we briefly mention at the end.

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 `race` to be a factor (grouping variable) and we want `w` as reference level for race (default is lexicographical ordering, i.e. `(b, h, w)` and `b` would then be the reference):

```
Income$race <- factor(Income$race)
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)
```
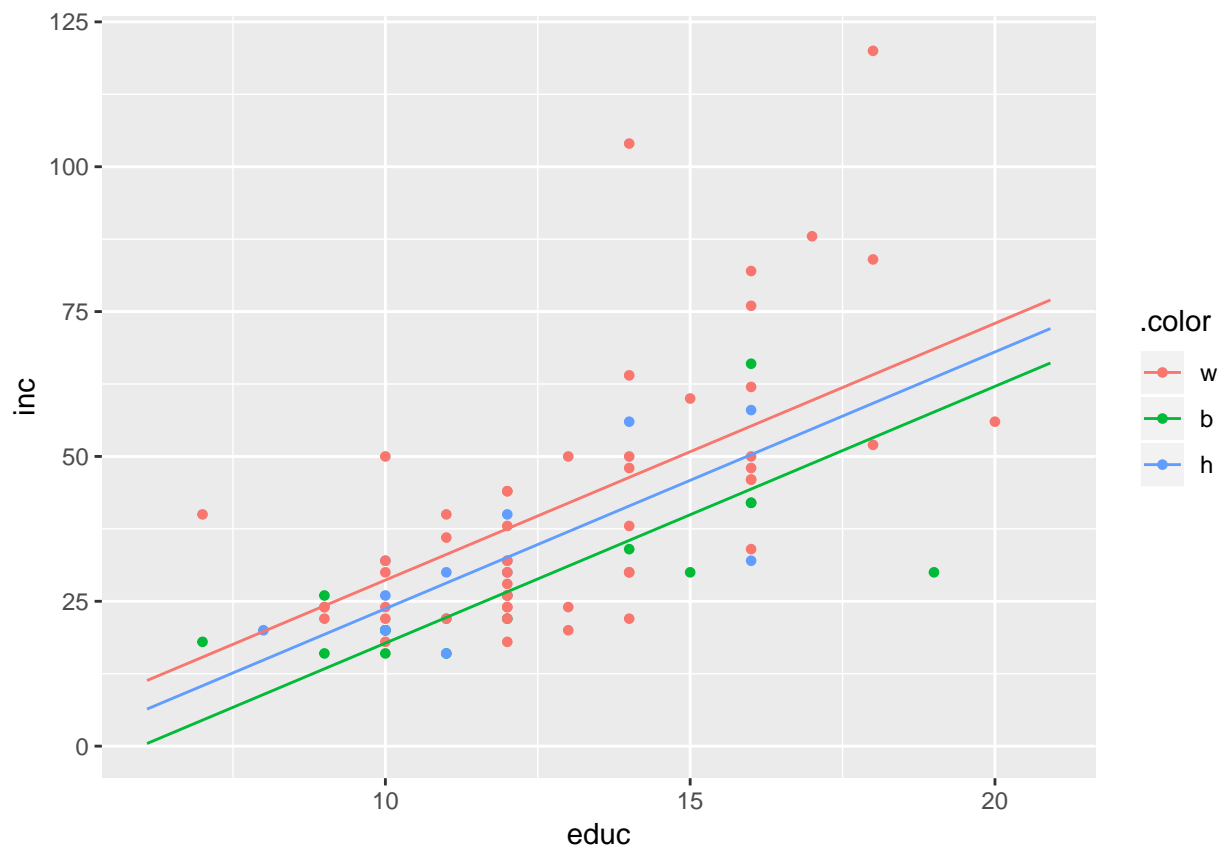
```
##
## Call:
## lm(formula = inc ~ educ + race, data = Income)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.664  -9.622  -1.642   6.552  57.620
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -15.6635     8.4121  -1.862   0.0665 .
## educ          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)`

4

```
## 
## 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

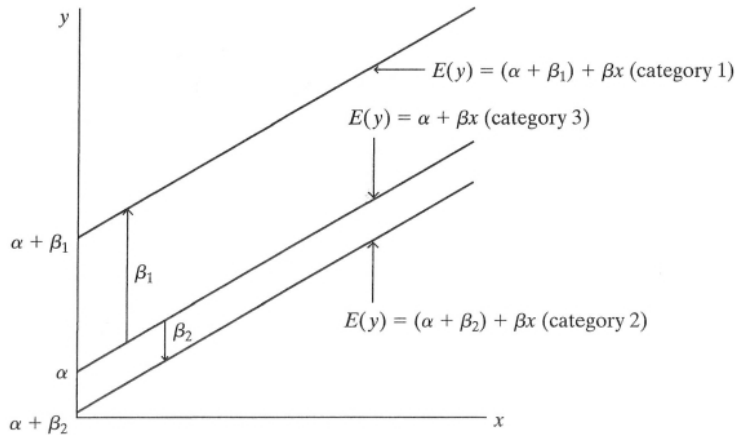| Category	y-Intercept	Slope	Mean $E(y)$ at Fixed x	Difference From Mean of Category 3, Controlling for x
1	$\alpha + \beta_1$	β	$(\alpha + \beta_1) + \beta x$	β_1
2	$\alpha + \beta_2$	β	$(\alpha + \beta_2) + \beta x$	β_2
3	α	β	$\alpha + \beta x$	0

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):
$$\widehat{y} = -25.87 + 5.21x$$

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

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

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 $\widehat{\beta_3} = -2.4107$. 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()
```
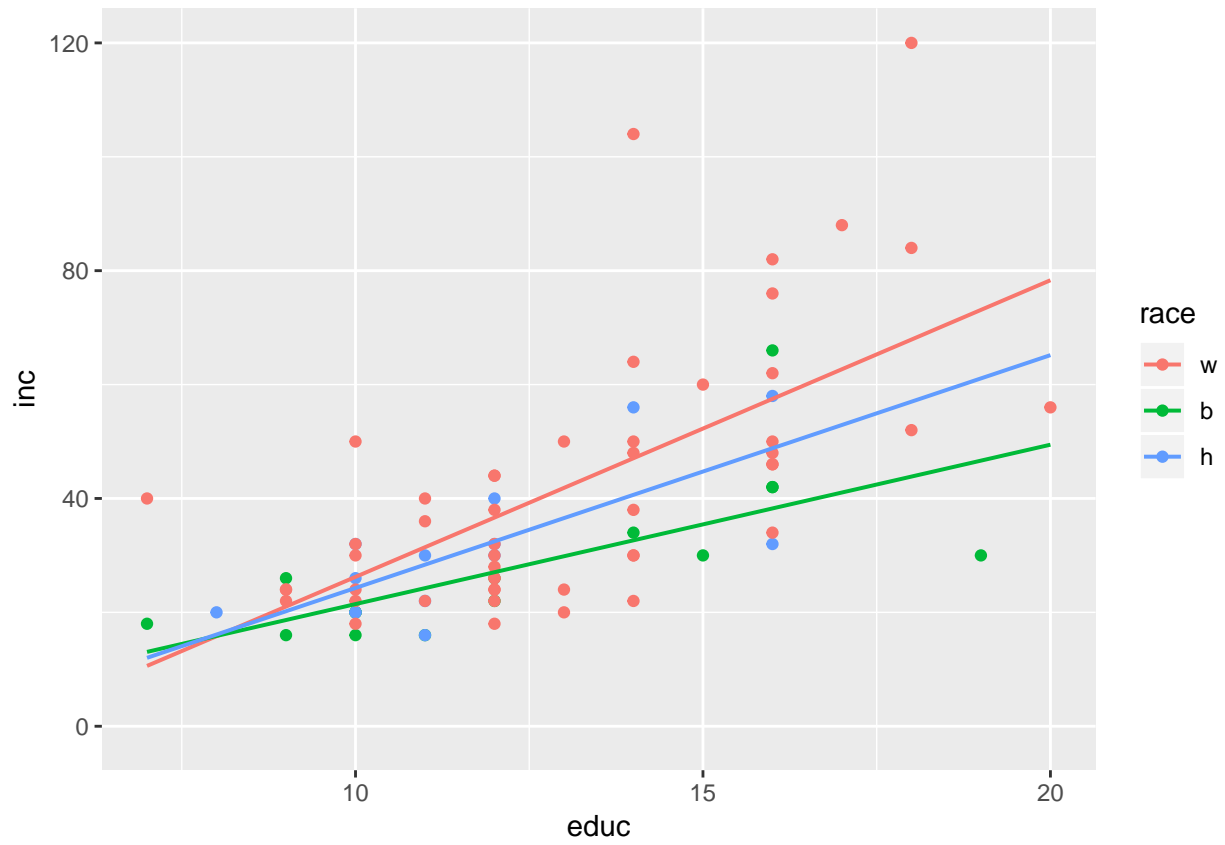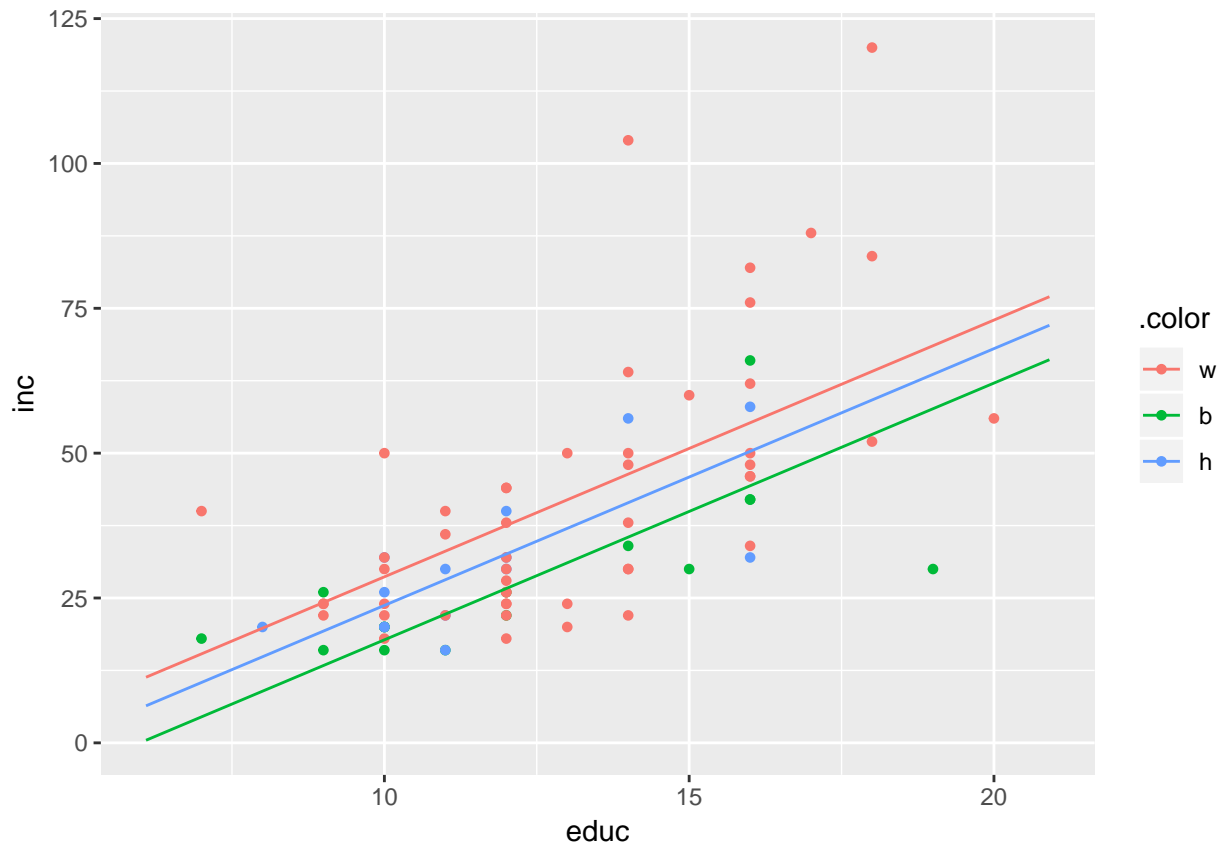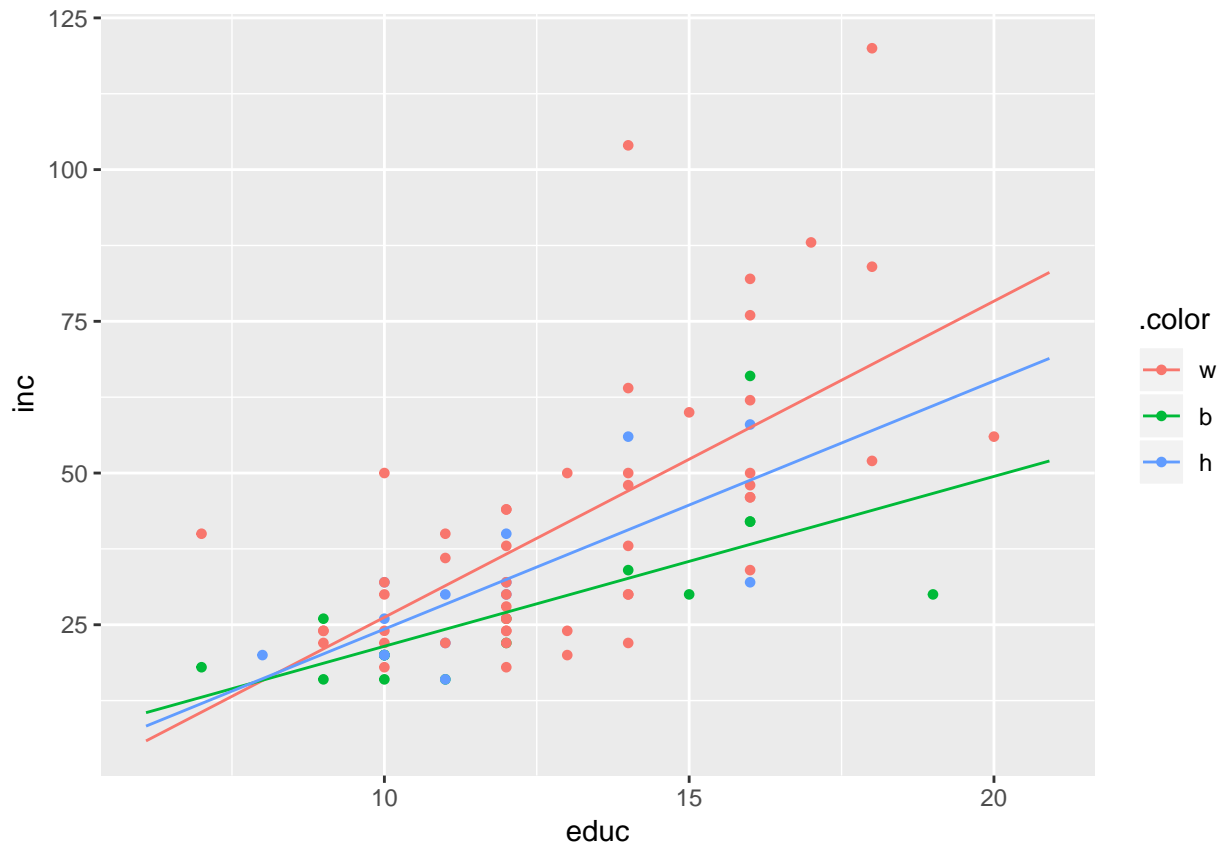
4 Test for no interaction

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).

11

- 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 later).

```
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   Model with only a single categorical preditor

- The model called `raceEff` above has a different mean income for each race.

  - α is the mean income for the reference grouph (`w`)
  - β_1 is the difference between the `b` and `w` in mean income
  - β_2 is the difference between the `h` and `w` in mean income

- Mathematically

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

  - `w`: $(z_1 = 0, z_2 = 0)$ $E(y|x) = \alpha$
  - `b`: $(z_1 = 1, z_2 = 0)$ $E(y|x) = \alpha + \beta_1$.
  - `h`: $(z_1 = 0, z_2 = 1)$ $E(y|x) = \alpha + \beta_2$.

- Graphically (notice we don't have lines anymore since there is no quantitative predictor – predicted means are indicated with crosses on the figure):

```
gf_point(inc ~ race, data = Income, color = ~race) %>%
  gf_jitter(width = 0.1) %>%
  gf_point(mean ~ race, data = favstats(inc ~ race, data = Income),
           size = 6, shape = 4, stroke = 2, show.legend = FALSE)
```
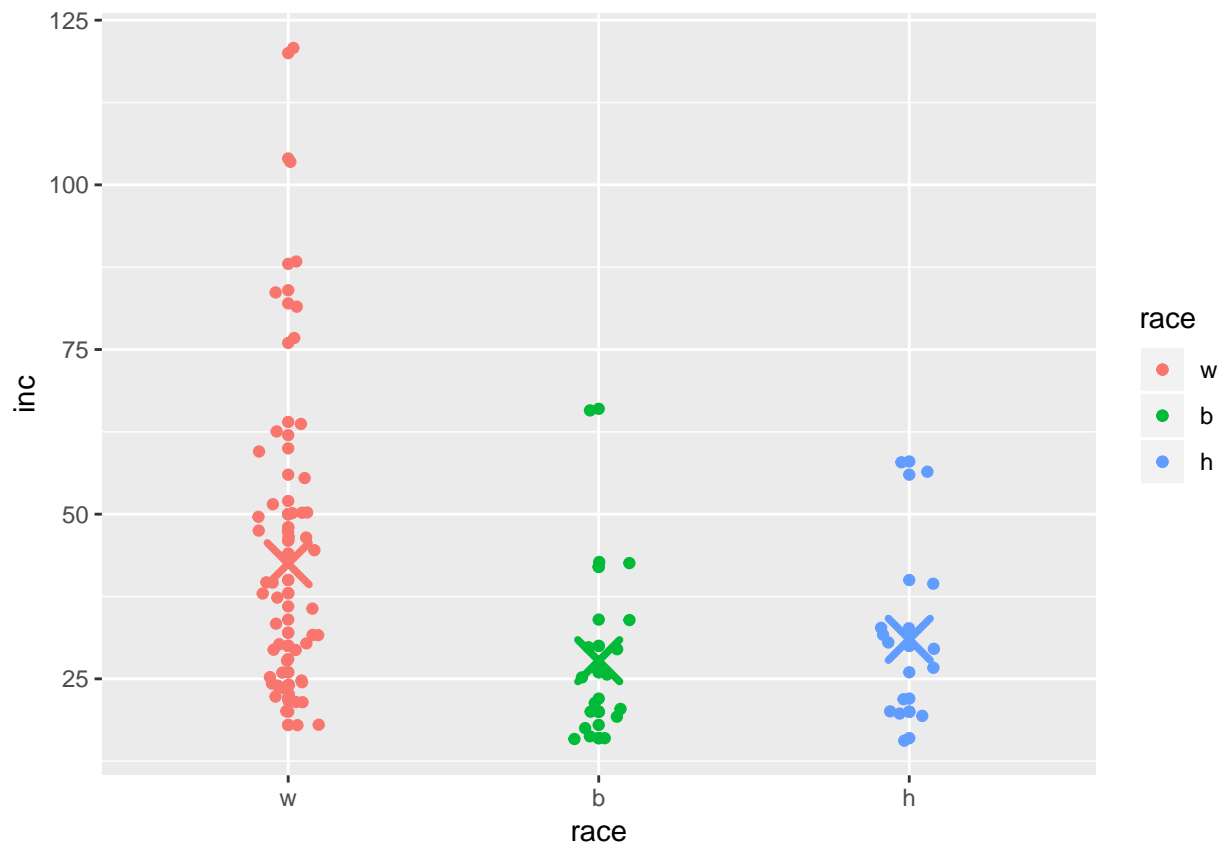

- Summary of the fitted model:

```
summary(raceEff)
```

14

```
## 
## Call:
## lm(formula = inc ~ race, data = Income)
## 
## Residuals:
##    Min    1Q Median    3Q    Max
## -24.48 -12.48  -4.48   7.52  77.52
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   42.480      2.810  15.115   <2e-16 ***
## raceb        -14.730      5.708  -2.581   0.0118 *
## raceh        -11.480      6.009  -1.910   0.0598 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 19.87 on 77 degrees of freedom
## Multiple R-squared:  0.0993, Adjusted R-squared:  0.0759
## F-statistic: 4.244 on 2 and 77 DF,  p-value: 0.01784
```

- Here we see:

    - Significant overall effect of race since $H_0 : \beta_1 = \beta_2 = 0$ is rejected based on p-value of 1.78%
    - Difference between `w` and `b` is significant (p-value is 1.18%)
    - Difference between `w` and `h` is nearly significant (p-value is 5.98%)

5.4 Example with more predictors

- 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.