SMML Class 7 Lab

John Kubale

2024-10-08

```r
library(ISLR2)
library(dplyr)
library(gridExtra)
library(grid)
library(tidyverse)
```

We'll start with the Wage data from the ISLR2 package.

```r
data(Wage)
head(Wage)
```

```
##        year age            maritl     race       education             region
## 231655 2006  18 1. Never Married 1. White    1. < HS Grad 2. Middle Atlantic
## 86582  2004  24 1. Never Married 1. White 4. College Grad 2. Middle Atlantic
## 161300 2003  45       2. Married 1. White 3. Some College 2. Middle Atlantic
## 155159 2003  43       2. Married 3. Asian 4. College Grad 2. Middle Atlantic
## 11443  2005  50      4. Divorced 1. White      2. HS Grad 2. Middle Atlantic
## 376662 2008  54       2. Married 1. White 4. College Grad 2. Middle Atlantic
##              jobclass         health health_ins  logwage      wage
## 231655  1. Industrial      1. <=Good      2. No 4.318063  75.04315
## 86582  2. Information 2. >=Very Good      2. No 4.255273  70.47602
## 161300  1. Industrial      1. <=Good     1. Yes 4.875061 130.98218
## 155159 2. Information 2. >=Very Good     1. Yes 5.041393 154.68529
## 11443  2. Information      1. <=Good     1. Yes 4.318063  75.04315
## 376662 2. Information 2. >=Very Good     1. Yes 4.845098 127.11574
```

1. What is the mean of `wage` by `jobclass`? What is the difference in the mean of `wage` by `jobclass`?

```r
Wage |> group_by(jobclass) |> summarise(mean_est = mean(wage)) |> mutate(diff = mean_es
```

```
## # A tibble: 2 x 3
##   jobclass       mean_est  diff
```

```
##   <fct>              <dbl> <dbl>
## 1 1. Industrial      103.   NA
## 2 2. Information     121.   17.3
```

2. Fit a model with `wage` as a function of `jobclass` with and without an intercept. What do the coefficients mean? How are they related to the results from #1?

- Industrial is the reference group
- Model with intercept, the difference in wage by jobclass is 17.3, the expected average wage for jobclass = industrial is 103.3, while for information is (103.3 + 17.3) 120.6.
- It's more work to interpret the model with the intercept, except this will be problematic for models with more predictors.

```r
fit_no_int <- lm(wage ~ 0 + jobclass, Wage)

fit_int <- lm(wage ~ jobclass, Wage)

summary(fit_no_int)
```

```
##
## Call:
## lm(formula = wage ~ 0 + jobclass, data = Wage)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -100.507  -25.362   -6.117   15.697  197.750
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## jobclass1. Industrial    103.321      1.039   99.43   <2e-16 ***
## jobclass2. Information   120.593      1.070  112.69   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.83 on 2998 degrees of freedom
## Multiple R-squared:  0.8828, Adjusted R-squared:  0.8827
## F-statistic: 1.129e+04 on 2 and 2998 DF,  p-value: < 2.2e-16
```

```r
summary(fit_int)
```

```
##
## Call:
## lm(formula = wage ~ jobclass, data = Wage)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -100.507  -25.362   -6.117   15.697  197.750
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              103.321      1.039   99.43   <2e-16 ***
## jobclass2. Information    17.272      1.492   11.58   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.83 on 2998 degrees of freedom
## Multiple R-squared:  0.04281,    Adjusted R-squared:  0.04249
## F-statistic: 134.1 on 1 and 2998 DF,  p-value: < 2.2e-16
```

3. Examine the ANOVA table of the two models from #2. What do you observe?

- The RSS for the models are the same the DFs are equal, despite different parameters being estimated. The F test is NA.

```
anova(fit_no_int, fit_int)
```

```
## Analysis of Variance Table
##
## Model 1: wage ~ 0 + jobclass
## Model 2: wage ~ jobclass
##   Res.Df     RSS Df  Sum of Sq F Pr(>F)
## 1   2998 4998547
## 2   2998 4998547  0 9.3132e-10
```

4. Calculate the mean of `wage` by `race`; Fit a simple linear regression of `wage` as a function of `race`; Obtain ANOVA table of the regression model

- To make it easier for us, we need to dummy code race, or use the I function, reference group is white.

```
Wage |> group_by(race) |> summarise(mean_est = mean(wage)) |> mutate(diff = mean_est -
```

```
## # A tibble: 4 x 3
##   race     mean_est  diff
##   <fct>       <dbl> <dbl>
## 1 1. White    113.    NA
## 2 2. Black    102.  -11.0
## 3 3. Asian    120.   18.7
## 4 4. Other     90.0 -30.3
```

```
fit_race <- lm(wage ~ race, Wage)
```

```r
summary(fit_race)
```

```
##
## Call:
## lm(formula = wage ~ race, data = Wage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -92.478 -24.708  -6.251  17.283 216.741
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  112.5637     0.8333 135.088  < 2e-16 ***
## race2. Black -10.9625     2.5634  -4.276 1.96e-05 ***
## race3. Asian   7.7246     3.1236   2.473  0.01345 *
## race4. Other -22.5903     6.8726  -3.287  0.00102 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41.5 on 2996 degrees of freedom
## Multiple R-squared:  0.0121, Adjusted R-squared:  0.01112
## F-statistic: 12.24 on 3 and 2996 DF,  p-value: 5.89e-08
```

```r
anova(fit_race)
```

```
## Analysis of Variance Table
##
## Response: wage
##             Df  Sum Sq Mean Sq F value   Pr(>F)
## race         3   63212 21070.6  12.237 5.89e-08 ***
## Residuals 2996 5158874  1721.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6. How about year? Try `year` as a continuous predictor as well as a categorical predictor in a regression model. Observe F values and df's.

```r
fit_age <- lm(wage ~ year, Wage)
summary(fit_age)
```

```
##
## Call:
## lm(formula = wage ~ year, data = Wage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -90.550 -26.606  -6.415  17.830 206.393
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2595.8616   752.8243  -3.448 0.000572 ***
## year            1.3499     0.3753   3.597 0.000328 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41.65 on 2998 degrees of freedom
## Multiple R-squared:  0.004296,   Adjusted R-squared:  0.003964
## F-statistic: 12.94 on 1 and 2998 DF,  p-value: 0.0003277
```

```r
fit_age_c <- lm(wage ~ as.factor(year), Wage)
summary(fit_age_c)
```

```
##
## Call:
## lm(formula = wage ~ as.factor(year), data = Wage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -90.226 -26.044  -6.238  17.414 208.131
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          106.198      1.839  57.742  < 2e-16 ***
## as.factor(year)2004    4.962      2.638   1.881  0.06011 .
## as.factor(year)2005    3.840      2.695   1.425  0.15439
## as.factor(year)2006    8.044      2.795   2.879  0.00402 **
## as.factor(year)2007    6.696      2.807   2.386  0.01711 *
## as.factor(year)2008    7.354      2.803   2.624  0.00874 **
## as.factor(year)2009    9.773      2.801   3.490  0.00049 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41.66 on 2993 degrees of freedom
## Multiple R-squared:  0.005442,	Adjusted R-squared:  0.003448
## F-statistic: 2.729 on 6 and 2993 DF,  p-value: 0.01203
```

7. wage as a function of jobclass and year (first as a categorical variable and then a continous variable), and their interaction.

A. year as a categorical variable.

```r
fit_wage_job_year_c <- lm(wage ~ jobclass + factor(year), Wage)
summary(fit_wage_job_year_c)
```

```
##
## Call:
## lm(formula = wage ~ jobclass + factor(year), data = Wage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -98.727 -25.078  -6.269  17.351 198.666
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)               98.384      1.922  51.186  < 2e-16 ***
## jobclass2. Information    17.280      1.493  11.576  < 2e-16 ***
## factor(year)2004           3.655      2.584   1.415 0.157271
## factor(year)2005           3.150      2.638   1.194 0.232595
## factor(year)2006           7.528      2.735   2.753 0.005949 **
## factor(year)2007           5.379      2.749   1.957 0.050474 .
## factor(year)2008           7.865      2.743   2.867 0.004168 **
## factor(year)2009           9.104      2.741   3.321 0.000907 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.76 on 2992 degrees of freedom
## Multiple R-squared:  0.04807,	Adjusted R-squared:  0.04585
## F-statistic: 21.59 on 7 and 2992 DF,  p-value: < 2.2e-16
```

B. year as a continous variable.

```r
fit_wage_job_year <- lm(wage ~ jobclass + year + jobclass*year, Wage)
summary(fit_wage_job_year)
```

```
##
## Call:
## lm(formula = wage ~ jobclass + year + jobclass * year, data = Wage)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -98.926 -25.281  -6.222  17.300 199.381 
## 
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)  
## (Intercept)                 -1470.6489  1015.5276  -1.448   0.1477  
## jobclass2. Information      -2483.8832  1474.5589  -1.684   0.0922 .
## year                            0.7847     0.5063   1.550   0.1213  
## jobclass2. Information:year     1.2470     0.7352   1.696   0.0899 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 40.73 on 2996 degrees of freedom
## Multiple R-squared:  0.04819,    Adjusted R-squared:  0.04723 
## F-statistic: 50.56 on 3 and 2996 DF,  p-value: < 2.2e-16
```

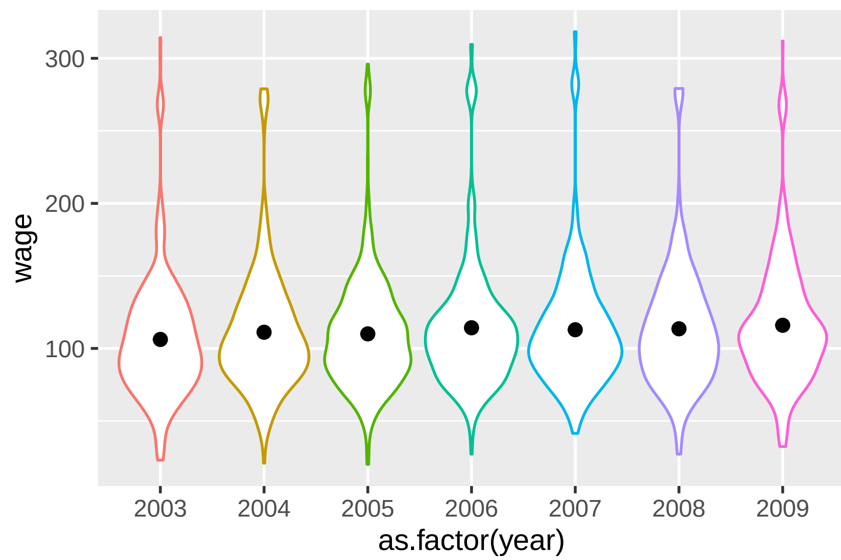
$y_i = \gamma_0 + \gamma_2 d_{2i} + \gamma_3 x_i + \gamma_4 d_{2i} x_i + \epsilon_i$ * What are γ_0, etc.?

8. `wage` as a function of `year` (as a continuous predictor), `race` and their interaction. Also, try to set `race`="3. Asian" as the reference category.

9. Visual examination of #8

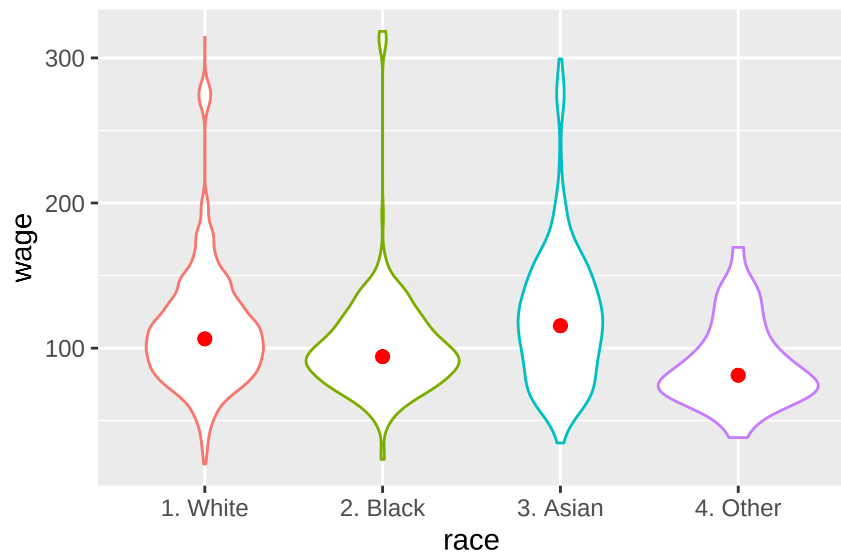
A. Violin plot of `wage` ~ `year` with the subgroup means displayed on the violins

```
violin_y<-ggplot(aes(x=as.factor(year), y=wage,
                     color=as.factor(year)), data=Wage)+
  geom_violin(trim=T)+
  stat_summary(fun=mean, geom="point", size=2, color="black")+
  theme(legend.position="none")
violin_y
```



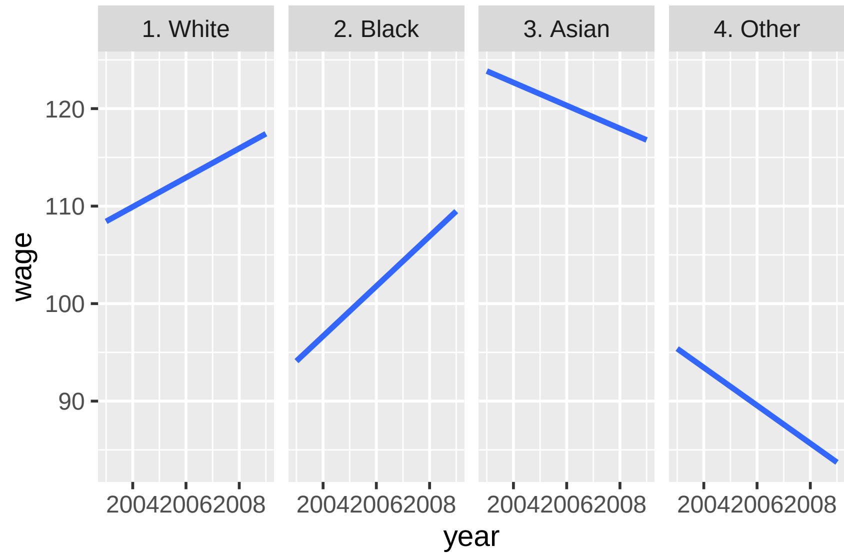
B. Violin plot of `wage` ~ `race` with the subgroup medians displayed on the violins

```
violin_r<-ggplot(aes(x=race, y=wage, color=race),data=Wage)+
  geom_violin(trim=T)+
  stat_summary(fun=median, geom="point", size=2, color="red")+
  theme(legend.position="none")
violin_r
```



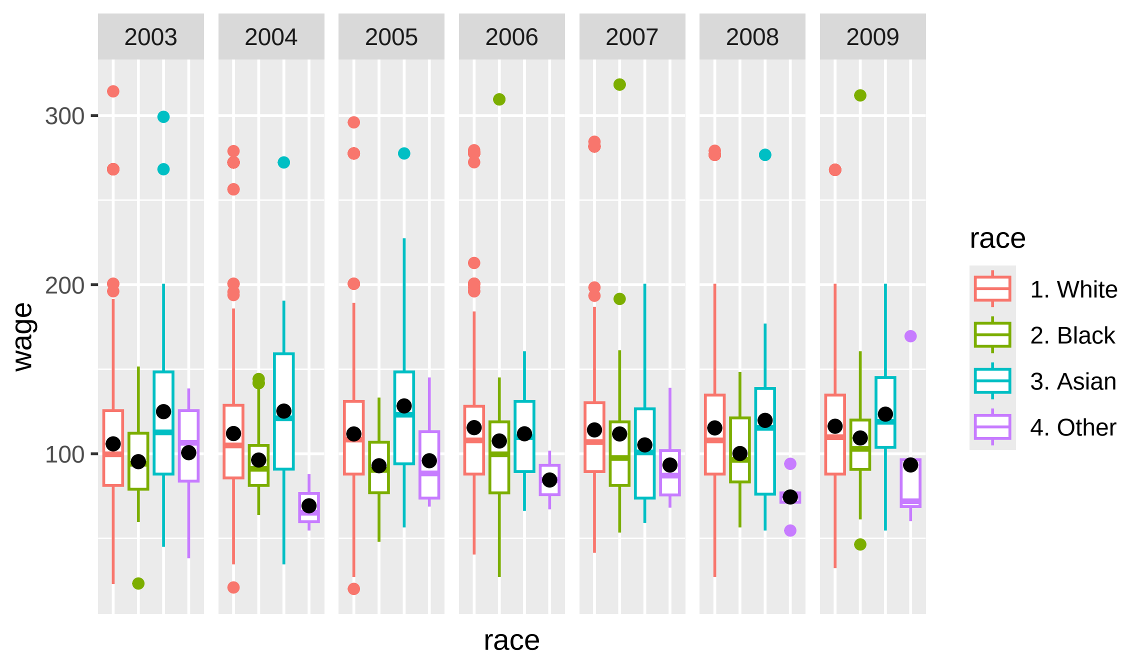
C. Regression lines of `wage` ~ `year` by `race`

```
int<-ggplot(aes(x=year,y=wage),data=Wage)+
  facet_grid(~race)+
  geom_smooth(method="lm", se=FALSE)
int
```



D. Boxplot of `wage ~ race` by `year`

```
ggplot(aes(x=race,y=wage, color=race),data=Wage)+
  facet_grid(~year)+
  geom_boxplot()+
   stat_summary(fun=mean, geom="point", size=2, color="black")+
   theme(axis.text.x=element_blank(),
         axis.ticks.x=element_blank())
```



E. Putting plots into one

```r
layout <- rbind(c(1,2),
                c(3,3))

grid.arrange(violin_y, violin_r,
             int, layout_matrix=layout)
```