HW 09

Matt Scheffel

2022-10-31

```
library(MASS)
data(birthwt)
head(birthwt)
```

```
##    low age lwt race smoke ptl ht ui ftv  bwt
## 85   0  19 182    2     0   0  0  1   0 2523
## 86   0  33 155    3     0   0  0  0   3 2551
## 87   0  20 105    1     1   0  0  0   1 2557
## 88   0  21 108    1     1   0  0  1   2 2594
## 89   0  18 107    1     1   0  0  1   0 2600
## 91   0  21 124    3     0   0  0  0   0 2622
```

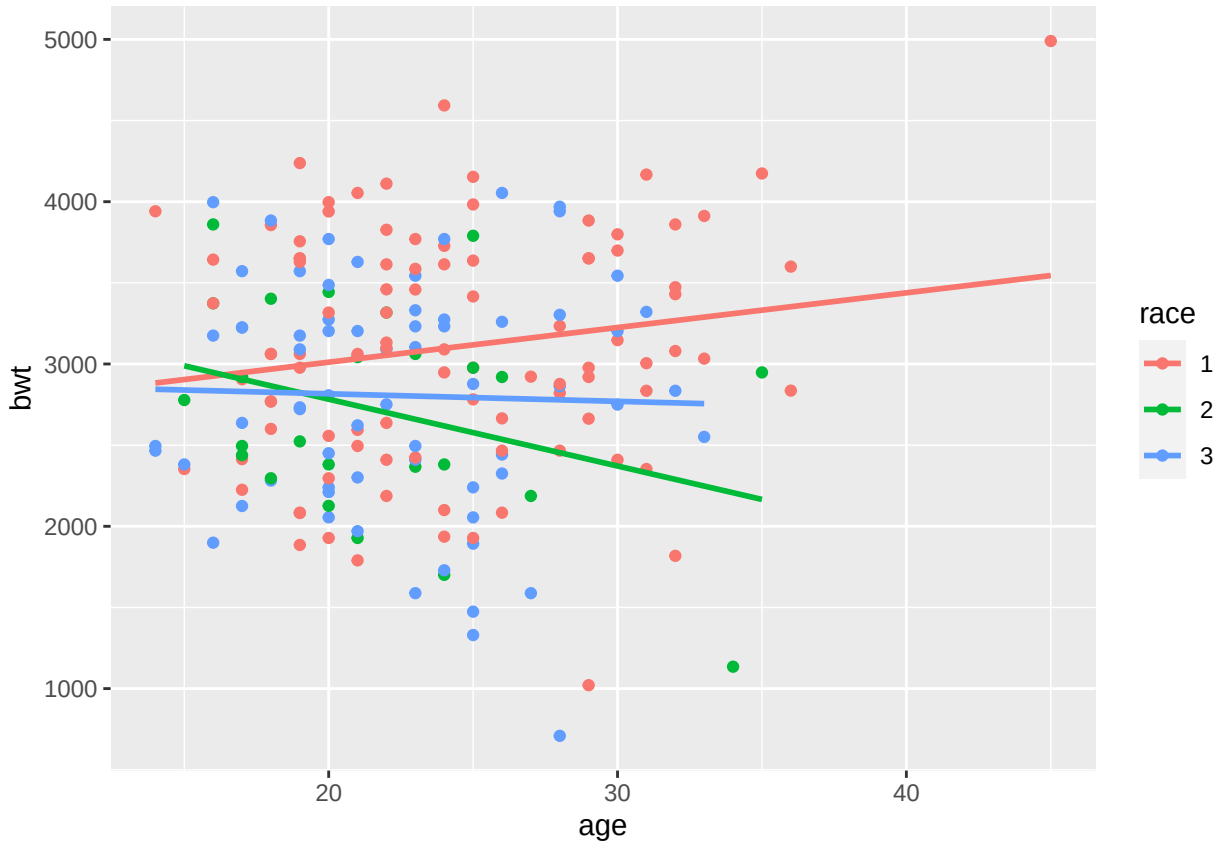
1A

```
df <- MASS::birthwt
df$race <- as.factor(df$race)
df$race <- as.factor(birthwt$race)

library(ggplot2)

ggplot(df, aes(x=age,y=bwt, color=race))+
geom_point()+
#scale_size(range = c(0.1,12))+
geom_smooth(method=lm, fill=NA)
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
labs(x="Age",
y="Birth Weight",
title="Birthweight Against Age",
addRegLine=TRUE, regLineColor="blue")
```

```
## $x
## [1] "Age"
##
## $y
## [1] "Birth Weight"
##
## $addRegLine
## [1] TRUE
##
## $regLineColor
## [1] "blue"
##
## $title
## [1] "Birthweight Against Age"
##
## attr(,"class")
## [1] "labels"
```

Interaction effect:

The regression lines for age and birth weight are not parallel and they intersect with each other. This visual shows us that there is a significant interaction effect between age and race.

1B

```
MLRbwt <- lm(bwt~age*race, data = df)
summary(MLRbwt)
```

```
##
## Call:
## lm(formula = bwt ~ age * race, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2182.35  -474.23    13.48   523.86  1496.51
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2583.54     321.52   8.035 1.11e-13 ***
## age            21.37      12.89   1.658   0.0991 .
## race2        1022.79     694.21   1.473   0.1424
## race3         326.05     545.30   0.598   0.5506
## age:race2     -62.54      30.67  -2.039   0.0429 *
## age:race3     -26.03      23.20  -1.122   0.2633
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 710.7 on 183 degrees of freedom
## Multiple R-squared:  0.07541,    Adjusted R-squared:  0.05015
## F-statistic: 2.985 on 5 and 183 DF,  p-value: 0.01291
```

Regression equation:

bwt = 2583.54 + 21.37(age) + 1022.79(race2) + 326.05(race3) - 62.54(age:race2) - 26.03(age:race3)

Race 1 is the reference level. For a one unit increase in age, the weight increases by 21.37 units.

Race 2 (vs. Race 1) shows that for a one unit increase in age, the weight decreases by 41.17 units.

Race 3 (vs. Race 1) shows that for a one unit increase in age, the weight decreases by 4.66 units.

10/27/22

Module 08 Homework

2)

a) The table shows us that there is fairly large disparity amongst the 3 geographic areas regarding mean teacher pay. The West has the highest average pay, coming in at around 7% higher than the North. The North, in turn, come in at around 7% higher than the South (which ranks last at $52,894.) Region seems to play a role in average teacher pay.

b) The table shows a positive correlation between mean public school expenditure (per student) and mean teacher pay. North to West shows a slight increase and South to North shows a significant increase. School regions that spend more on their students on average also appear to better compensate their teachers on average.

c) Using a multiple linear regression model w/ teacher pay as the response variable gives further insight into the relationship between the variables because we have seen from the table that both the region and the mean expenditures can have an impact on teacher compensation. It is best to simultaneously evaluate these variables together in a MLR model so we gain greater insight into their relationship with each other through the recorded output/results.

3) H_0: not significant

H_a: significant

(MS = SS/df , F = MS/MSE)

	Sum. Sq.	Mean Sq.	F	Pr(>F)
Spend:Area	9720281	4860140.5	0.94067	0.397903

P-value > 0.05

Thus, we accept H_0 at the 5% level of significance.

Since we accept H_0, this means the interaction terms are not significant.

b) The reference class for this model is the North Region.

c) β_2 is the coefficient for Area South. The coefficient is 5.294e+02.

An interpretation of β_2 in this regression model:

$$Y = 1.160e+04 + 3.289e+00\,(Spend) + 5.294e+02\,(Area\,South) + 1.671e+03\,(Area\,West)$$

is that with all other variables fixed, a one unit change in Area South will result in a 5.294e+02 unit change in Y, with Y being average public teacher pay.

d) Bonferroni correction = $\frac{\text{p-value } (\alpha)}{n}$

i) North & South:

$N = \frac{2.43e^{-11}}{2}$, $S = \frac{0.4934}{2}$

ii) North & West:

$N = \frac{2.43e^{-11}}{2}$, $W = \frac{0.0422}{2}$

iii) South & West:

$S = \frac{0.4934}{2}$, $W = \frac{0.0422}{2}$

e) The intervals indicate that geographic region is correlated with mean annual salary for teachers.