Solutions to exercises

Agresti 13.1

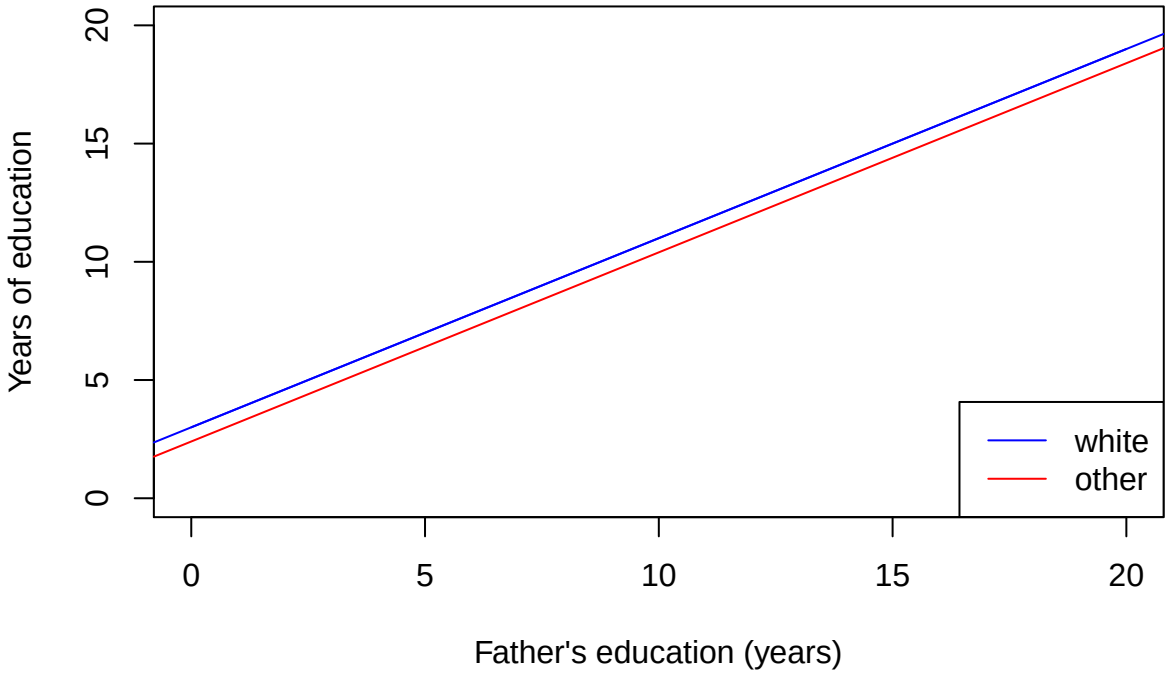
(a) The mean for white people (z=0) is

$$E(y|z=0) = 11 + 2 * 0 = 11.$$

Otherwise (z=1), the mean is

$$E(y|z=1) = 11 + 2 * 1 = 13.$$

(b) We plot the regression lines for the association between education and father's education for the two `race` groups:



(c) Fixing father's education to be x, the expected education is $3 + 0.8 * x$ for whites and $3 + 0.8 * x - 0.6$. That is, the difference is -0.6. For instance for $x = 12$, the expected education is

```
3+0.8*12
```

```
## [1] 12.6
```

for whites and

```
3+0.8*12-0.6
```

```
## [1] 12
```

for others, so the difference is -0.6.

Agresti exercise 13.5

(a) We get the prediction equation:

$$\hat{y} = 8.3 + 9.8 * f - 5.3 * s + 7 * m_1 + 2 * m_2 + 1.2 * m_3 + 0.501 * x.$$

(b) The predicted alcohol consumption for divorced males whose father died in the past three years and with alcohol consumption three years previously equal to

i) 0 drinks:

```
8.3 + 9.8  + 7
```

```
## [1] 25.1
```

ii) 10 drinks:

```
8.3 + 9.8  + 7 + 0.501*10
```

```
## [1] 30.11
```

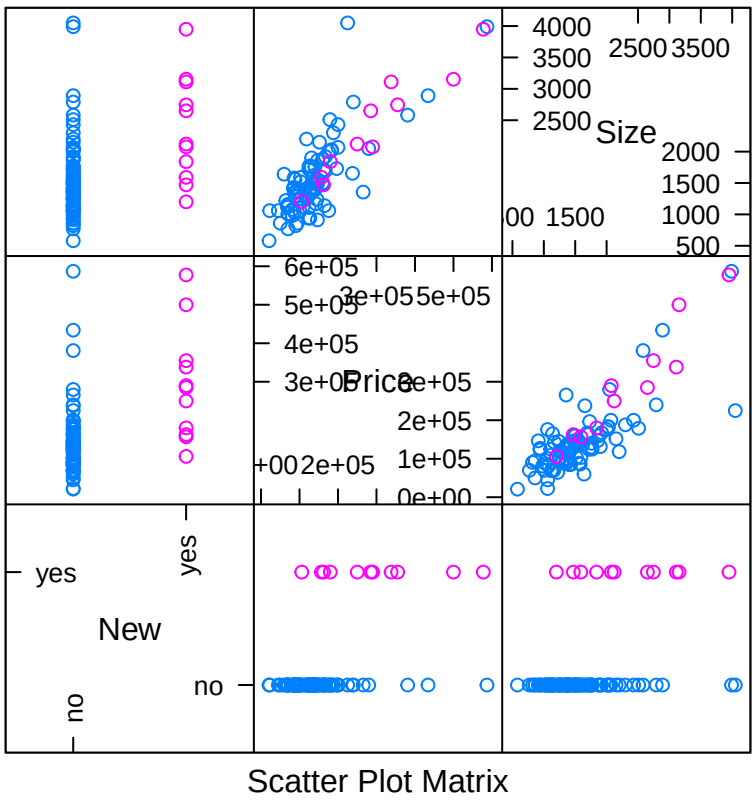
Agresti exercise 13.7

Import data (this data set includes the variable `new`):

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

First interpret the following scatterplot matrix:

```
splom(~HousePriceFull[,c(5,6,7)], groups = New, data = HousePriceFull)
```



- The house price seems to increase with size and be higher for new houses.

Fit the linear model corresponding to Table 13.17:

```
model <- lm( Price ~ Size + New, data = HousePriceFull )
summary(model)
```

```
##
## Call:
## lm(formula = Price ~ Size + New, data = HousePriceFull)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -205102  -34374   -5778   18929  163866
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -40230.867  14696.140  -2.738  0.00737 **
## Size           116.132      8.795  13.204  < 2e-16 ***
## Newyes       57736.283  18653.041   3.095  0.00257 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 53880 on 97 degrees of freedom
## Multiple R-squared:  0.7226, Adjusted R-squared:  0.7169
## F-statistic: 126.3 on 2 and 97 DF,  p-value: < 2.2e-16
```

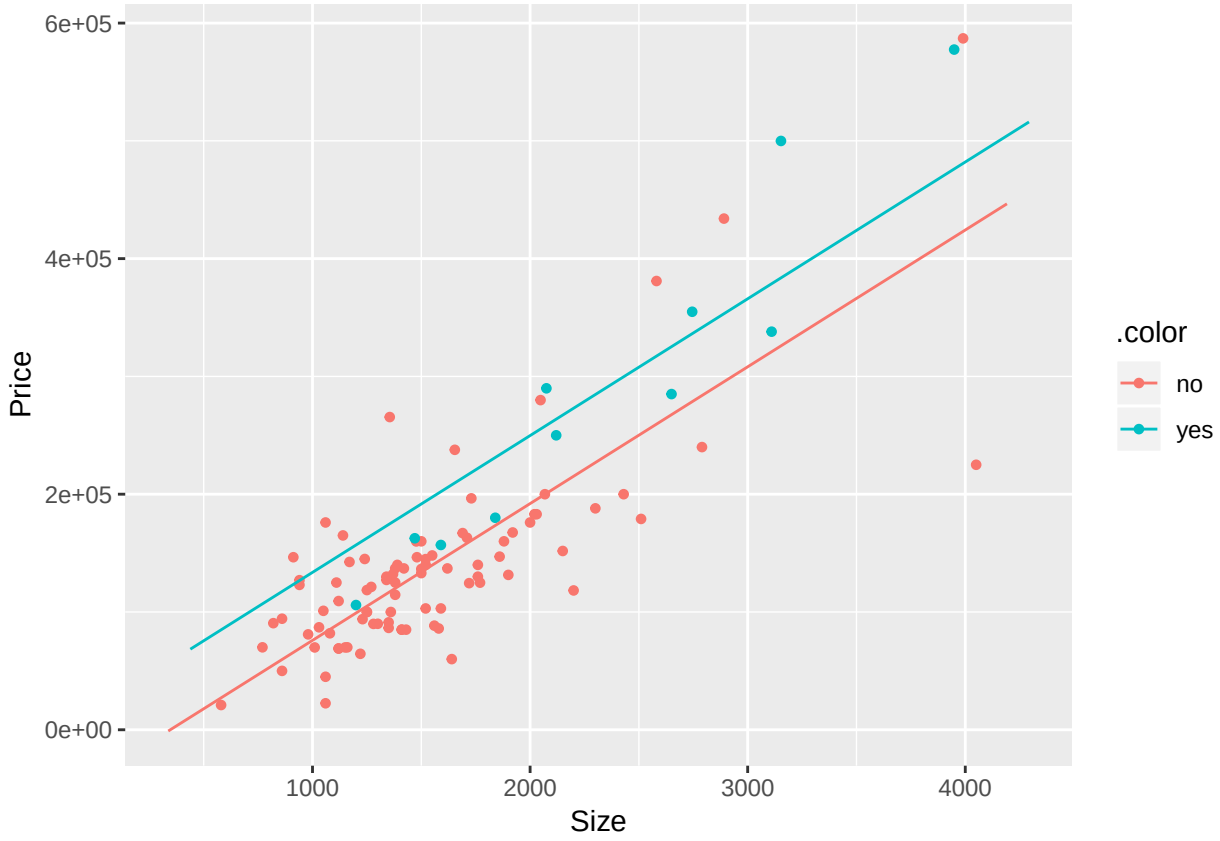
Write the prediction equation with appropiate notation:

$$\hat{y} = -40230.867 + 116.132 * size + 57736.283 * z,$$

where z is the dummy variable for new.

Plot the two regression lines:

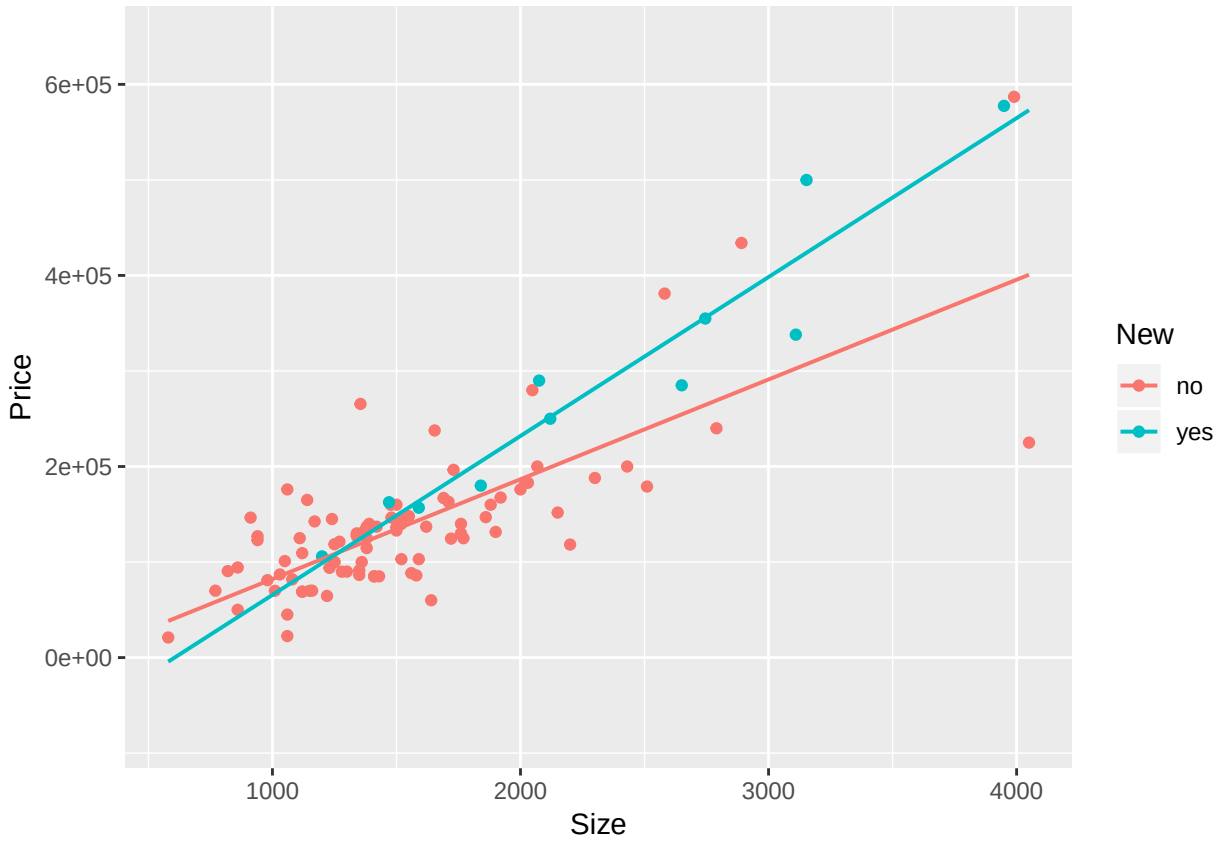
```
plotModel( model )
```



Agresti exercise 13.8

Make the relevant plot(s) using `gf_point`:

```
gf_point(Price ~ Size,  col = ~New, data = HousePriceFull) %>% gf_lm()
```



Fit the linear model corresponding to Table 13.18 in Agresti:

```
model1 <- lm( Price ~ Size*New, data = HousePriceFull )
 summary(model1)
```

```
##
## Call:
## lm(formula = Price ~ Size * New, data = HousePriceFull)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -175748  -28979   -6260   14693  192519
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -22227.808  15521.110  -1.432  0.15536
## Size           104.438      9.424  11.082  < 2e-16 ***
## Newyes      -78527.502  51007.642  -1.540  0.12697
## Size:Newyes     61.916     21.686   2.855  0.00527 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 52000 on 96 degrees of freedom
## Multiple R-squared:  0.7443, Adjusted R-squared:  0.7363
## F-statistic: 93.15 on 3 and 96 DF,  p-value: < 2.2e-16
```

Write the prediction equations for old and new houses:

$$\begin{aligned}\hat{y}_{old} &= -22227.808 + 104.438 * size \\ \hat{y}_{new} &= (-22227.808 - 78527.502) + (104.438 + 61.916) * size \\ &= -100755.3 + 166.354 * size\end{aligned}$$

Is the interaction significant?

- Vi apply the `anova` function to the models with and without interaction:

```
anova(model,model1)
```

```
## Analysis of Variance Table
##
## Model 1: Price ~ Size + New
## Model 2: Price ~ Size * New
##   Res.Df        RSS Df  Sum of Sq      F   Pr(>F)
## 1     97 2.8161e+11
## 2     96 2.5957e+11  1 2.2041e+10 8.1519 0.005272 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This shows that the interaction is significant with a p-value of 0.005272. Alternatively, the test for interaction could be found in the summary of `model1` in the `Size:Newyes` line. This only works when the categorical variable has two levels, because in this case the model with interaction only contains one extra parameter.

Agresti exercise 13.20

(a) The least permissive people seem to be older (because the slope for age is negative) white (because the parameter corresponding to race is positive and white is the reference group, white are least permissive) females (because the parameter corresponding to sex is negative and male is the reference group) with a low level of education (slope of education is positive) coming from the south (difference is positive, south is reference) who are fundamentalist Protestants (has the higest negative difference to reference group), frequently attend church (slope is negative), and do not tolerate freedom of speech (slope is negative).

(b) Similarly, the most permissive people seem to be younger black males with a high level of education, coming from the "non-south", who are Jewish, rarely go to church, and tolerate freedom of speech.