S631 HW8

Erik Parker

October 21, 2017

1. After viewing the *Robey* data, it seems most reasonable to use the variable *tfr* as the response and *region* and *contraceptors* as the two predictors.

a) Perform a one-factor design analysis with the appropriate variables and determine if there are significant differences on the response when *region = Africa vs region = Near.East* and when *region = Asia vs region = Latin.Amer.*

```
rm(list = ls())

repro <- read.table("./Robey.txt")

m1 <- lm(tfr ~ region, data = repro)

summary(m1)
```

```
##
## Call:
## lm(formula = tfr ~ region, data = repro)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6556 -0.7875 -0.0028  0.6444  2.2000
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        5.8556     0.2674  21.897  < 2e-16 ***
## regionAsia        -2.3156     0.4475  -5.175 4.88e-06 ***
## regionLatin.Amer  -1.8056     0.3898  -4.632 2.99e-05 ***
## regionNear.East   -1.0556     0.5348  -1.974   0.0544 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.135 on 46 degrees of freedom
## Multiple R-squared:  0.428,  Adjusted R-squared:  0.3907
## F-statistic: 11.47 on 3 and 46 DF,  p-value: 9.719e-06
```

From this first model, we can easily see that the estimated change in total fertility rate (*tfr*) when moving from *Africa* to the *Near.East* region is a decrease of 1.0556 units. This is easy to see because *regionAfrica* is our first category, so is dropped from the model and becomes the intercept coefficient, $\hat{\beta}_0$. So the change when moving from Africa to the near east is shown by the esimated coefficient for *regionNear.East*, $\hat{\beta}_4$. The accompanying p-value of 0.0544 for the hypothesis test that $\beta_4 = 0$ means that we can't reject the hypothesis that there is no change in *tfr* when moving from *Africa* to *Near.East* when tested at an alpha of 0.05 though.

```
a = c(0, 1, -1, 0)
se_b2b3 = sqrt(t(a) %*% vcov(m1) %*% a)
se_b2b3
```

```
##           [,1]
```

```
## [1,] 0.4573404
```

```
b2b3 = as.numeric(coef(m1)[2] - coef(m1)[3])
t_val = b2b3/se_b2b3
p_val = 2 * (1 - pt(abs(t_val), m1$df))
c(`b2-b3` = b2b3, SE = se_b2b3, `t-Value` = t_val, `p-Value` = p_val)
```

```
##      b2-b3         SE    t-Value    p-Value
## -0.5100000  0.4573404 -1.1151431  0.2705813
```

From the above test, we can see that when we move from *Asia* to *Latin.Amer.*, we cannot reject the null hypothesis that there is no change in the response variable. So, in other terms: there is no significant difference on the expected response when we move from the asian to latin american region.

b) Explain what each regression coefficient means and write out the mean function for each category of the factor.

As briefly mentioned earlier, the first regression coefficient for the intercept actually corresponds to the expected value of our response (*tfr*) when the region = *Africa*. The corresponding mean function for this category is $E(Y|U_1 = 1) = \beta_0$

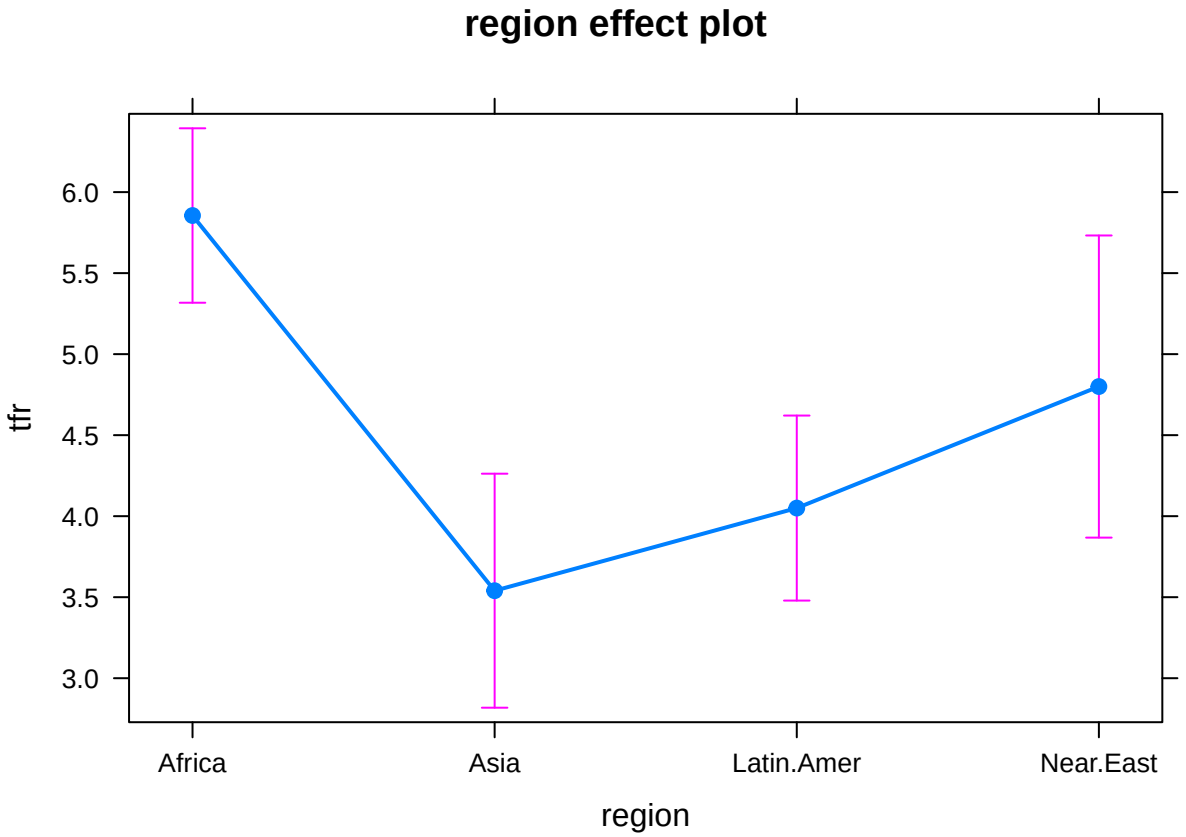
The second regression coefficient of -2.3156, for *regionAsia* corresponds to $\hat{\beta}_1$ and is the estimated change in the response when we move from the first category of the factor (*regionAfrica*) to the second. This means that as we move from Africa to Asia, the estimated total fertilty rate drops 2.32 from 5.86, to 3.54. The corresponding mean function here is $E(Y|U_2 = 1) = \beta_0 + \beta_2$

The third coefficient of -1.8056 from our model corresponds to $\hat{\beta}_2$ and represents the estimated change in the response as we move from *regionAfrica* to *regionLatin.America*. So, as we move from Africa to Latin America, the estimated fertility rate drops to 4.05. The mean function for this category is given by $E(Y|U_3 = 1) = \beta_0 + \beta_3$.

The fourth and final coefficient of our model is -1.0556 and is non-significant at $\alpha = 0.05$. This means that as we move from *regionAfrica* to *regionNear.East* we can't reject the possibility that there is no estimated change in the total fertility rate. This category's mean function is given by $E(Y|U_4 = 1) = \beta_0 + \beta_4$.

c) Obtain and describe "effects" plot for this model

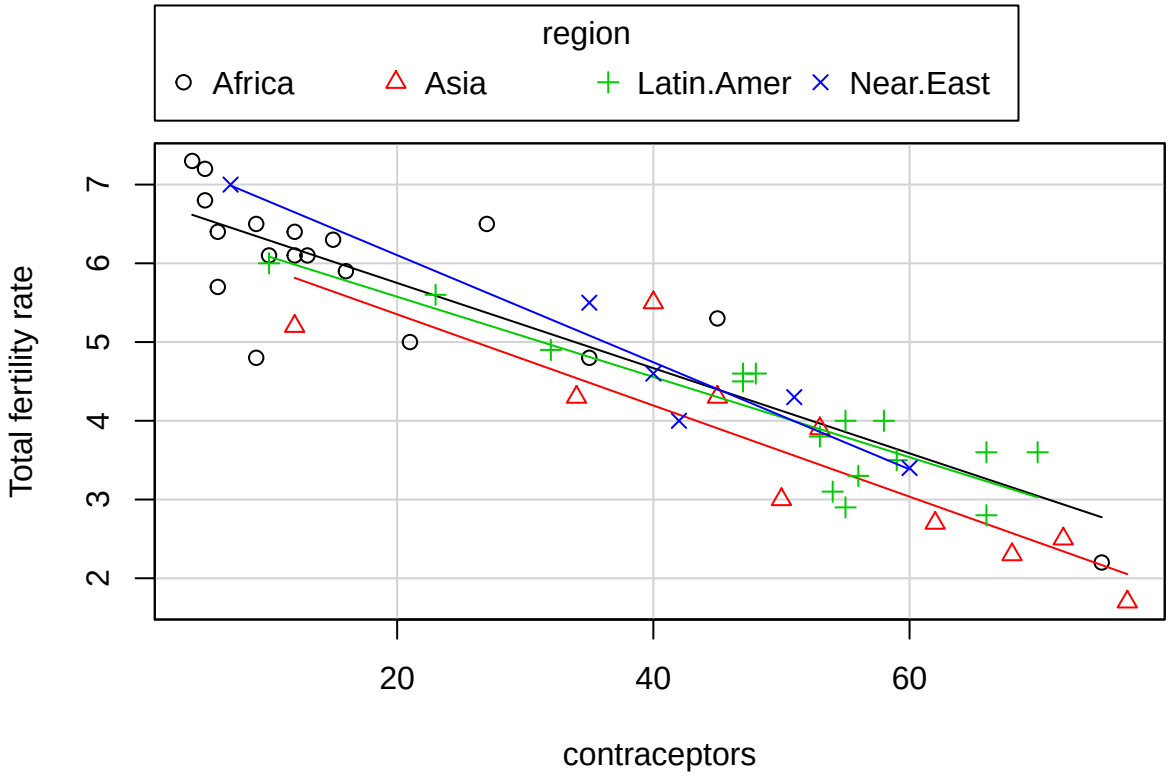
```
library(alr4)

plot(Effect(c("region"), m1))
```



From the above plot we can see that, as shown by the model coefficients, the highest levels of total fertility can be seen in the *Africa* and *Near.East* reions. The lowest levels can be seen in the countries in the *Asia* region. The *Near.East* region shows a very large variance, with a lot of overlap with the *Africa* region, explaining why we could not reject the null hypothesis that $\beta_4 = 0$ for *Near.East* in part a.

d) Obtain a scatterplot for a model with both regressors. Does this plot suggest that changes in the continuous regressor are associated with changes in the response? Does it suggest different slopes should be considered for different levels? Should different intercepts be considered?

```
scatterplot(tfr ~ contraceptors | region, data = repro, smooth = FALSE, boxplots = FALSE,
    ylab = "Total fertility rate")
```



Because the slopes of the four lines in the plot above are not zero, this suggests that changes in the continuous regressor, *contraceptors*, are associated with changes in the expected response, *tfr*. Specifically, because the slopes for each region are decreasing, we can see that, in general, an increase in the percent of contraceptors used by married women of childbearing age is associated with a lower expected total fertility rate.

Furthermore, it is clear from the plot above that the slopes of the lines for the different factor levels are not completely identical, but they are not that different from oneanother either, in fact they seem relatively close to parallel. From this plot on these data, it seems as though there is no obvious need to consider different slopes for the four factor levels.

Similarly, it also appears that, while they are not identical, the intercepts of the lines for the four different factor levels appear to be quite similar and don't seem to have any obvious, significant, differences. Thus it does not appear necessary for us to consider different intercepts for these factor levels.

e) Obtain a model with an interaction and interpret three estimated coefficients, one related to the factor, one related to the continuous regressor, and one related to the interaction.

```
m2 <- lm(tfr ~ region * contraceptors, data = repro)

summary(m2)
```

```
##
## Call:
## lm(formula = tfr ~ region * contraceptors, data = repro)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.54546 -0.26527 -0.04661  0.34689  1.30579
```

```
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     6.832351   0.194090  35.202  < 2e-16 ***
## regionAsia                     -0.322375   0.563627  -0.572    0.570
## regionLatin.Amer               -0.237356   0.520948  -0.456    0.651
## regionNear.East                 0.631733   0.632999   0.998    0.324
## contraceptors                  -0.054099   0.007718  -7.009 1.41e-08 ***
## regionAsia:contraceptors       -0.003795   0.012389  -0.306    0.761
## regionLatin.Amer:contraceptors  0.003136   0.012044   0.260    0.796
## regionNear.East:contraceptors  -0.013920   0.016141  -0.862    0.393
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5732 on 42 degrees of freedom
## Multiple R-squared:  0.8667, Adjusted R-squared:  0.8445
## F-statistic: 39.01 on 7 and 42 DF,  p-value: < 2.2e-16
```

To answer this question, I will be interpreting the coefficients for *regionAsia*, *contraceptors*, and the interaction *regionAsia:contraceptors*.

First, we can see from the above summary that *regionAsia* shows an estimated slope of -0.322 which is quite non-significant though with a p-value of 0.57. This means that this coefficient is not at all clearly different from zero, suggesting that we can't conclude there is any difference in the total fertility rate when we move from Africa to Asia, specifically when we consider a model that already includes the continuous regressor, and an interaction between it and the factor which are held constant.

Then, for the continuous regressor *contraceptors* we find a low, but significant (at $p = 1.41e^{-}08$), estimated coefficient value of -0.054. This can be interpreted as meaning that every percentage increase in the contraceptive use among married women (holding the other regressor constant), we see an average decrease in the total fertility rate by 0.054 points.

Finally, when we examine an interaction term, specifically *regionAsia:contraceptors*, we see that there is a highly insignificant estimated coefficient, a result that is mirrored across all of the interaction terms. This means that in this model we don't have the evidence to conclude that the slope of the regression tfr~contraceptors changes as we move between different factor levels, and here specifically from *regionAfrica* to *regionAsia*. This is not suprising based on the plot seen in part d, where we saw that the slopes of the four different factor levels appeared to be quite similar.

f) Should we include interactions in the model? How about the continuous regressor? The factor?

```
fmtry1 <- lm(tfr ~ region + contraceptors, data = repro)
summary(fmtry1)
```

```
##
## Call:
## lm(formula = tfr ~ region + contraceptors, data = repro)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.56044 -0.30085 -0.05744  0.39619  1.32998
##
## Coefficients:
```

```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        6.86223    0.15674  43.782  < 2e-16 ***
## regionAsia        -0.46203    0.27012  -1.710   0.0941 .
## regionLatin.Amer -0.02800    0.24338  -0.115   0.9089
## regionNear.East    0.12148    0.28217   0.431   0.6689
## contraceptors     -0.05575    0.00466 -11.963 1.42e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.561 on 45 degrees of freedom
## Multiple R-squared:  0.8632, Adjusted R-squared:  0.851
## F-statistic: 70.97 on 4 and 45 DF,  p-value: < 2.2e-16
```

```
fmtry2 <- lm(tfr ~ contraceptors, data = repro)
summary(fmtry2)
```

```
##
## Call:
## lm(formula = tfr ~ contraceptors, data = repro)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5493 -0.3013  0.0254  0.3957  1.2021
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.875085   0.156860   43.83   <2e-16 ***
## contraceptors -0.058416   0.003584  -16.30   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5745 on 48 degrees of freedom
## Multiple R-squared:  0.847,  Adjusted R-squared:  0.8438
## F-statistic: 265.7 on 1 and 48 DF,  p-value: < 2.2e-16
```

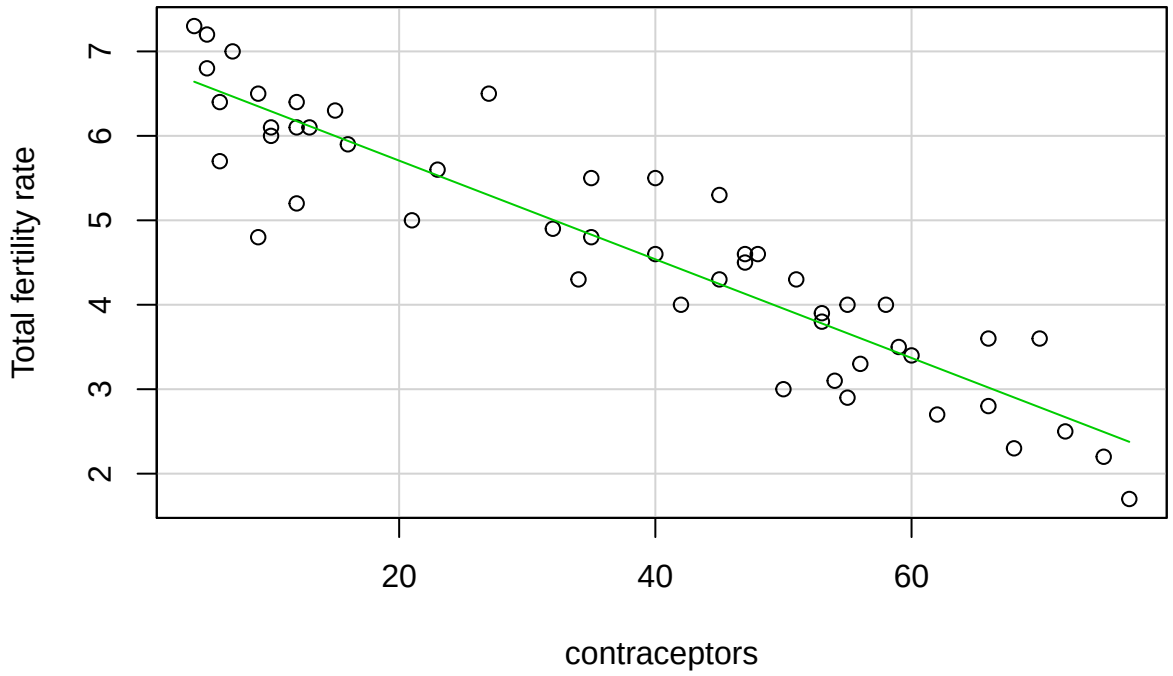
From the previous result, it doesn't seem like it would be wise to include an interaction in this model. We saw above in part e that none of the interaction terms were statistically significant, so it doesn't seem useful to include them. The continuous regressor though seems worthy of inclusion in the model. It was the only regressor which showed significance in the full model, and analysis of the scatterplot in part d suggests that *contraceptors* has useful explanatory potential in understanding changes in total fertility rate. Finally, the two summary tables above, comparing the effects of including *region* in a model already containing *contraceptors* shows us that adding in the region regressor doesn't really add much to the model at all, explaining less than an additional 2% of the variation found in the response. Addditionally, none of the factor levels are signifiacntly different from zero, only *regionAsia* comes close though it still has a relatively large p-value, this suggests that the inclusion of the factor in this model doesn't not have any significant explanatory value and so is not necessary.

So, it seems the final model should include the continuous regressor, but not the factor or an interaction between the two.

g) Obtain a final model, and an effects plot and describe the plot.

```
mfinal <- lm(tfr ~ contraceptors, data = repro)
```

```
scatterplot(tfr ~ contraceptors, data = repro, smooth = FALSE, boxplots = FALSE,
    ylab = "Total fertility rate")
```



The above effects plot shows that in general, as the average rate of contraceptive use increases in a country, the expected total fertility rate decreases relatively linearally.

h) Based on the chosen model, obtain and interpret a prediction interval for a new observation where the new value for the predictor is equal to its sample mean and the region is Asia.

```
summary(repro)
```

```
##         region        tfr        contraceptors
##  Africa    :18   Min.   :1.700   Min.   : 4.00
##  Asia      :10   1st Qu.:3.600   1st Qu.:12.25
##  Latin.Amer:16   Median :4.600   Median :41.00
##  Near.East : 6   Mean   :4.688   Mean   :37.44
##                  3rd Qu.:5.975   3rd Qu.:55.00
##                  Max.   :7.300   Max.   :77.00
```

```
newdata <- data.frame(contraceptors = 37.44, region = "Asia")
predict(mfinal, newdata, interval = "predict", level = 0.95)
```

```
##     fit      lwr      upr
## 1 4.688 3.521472 5.854528
```

So, based on my chosen final model using just the regressor *contraceptors*, we can say that we are 95% confident that the true value for the total fertility rate of a country in the Asian region, and with a contraceptive usage percent of 37.44 is between 3.522 and 5.855 children per woman on average.