STAT-S631

Assignment 8

John Koo

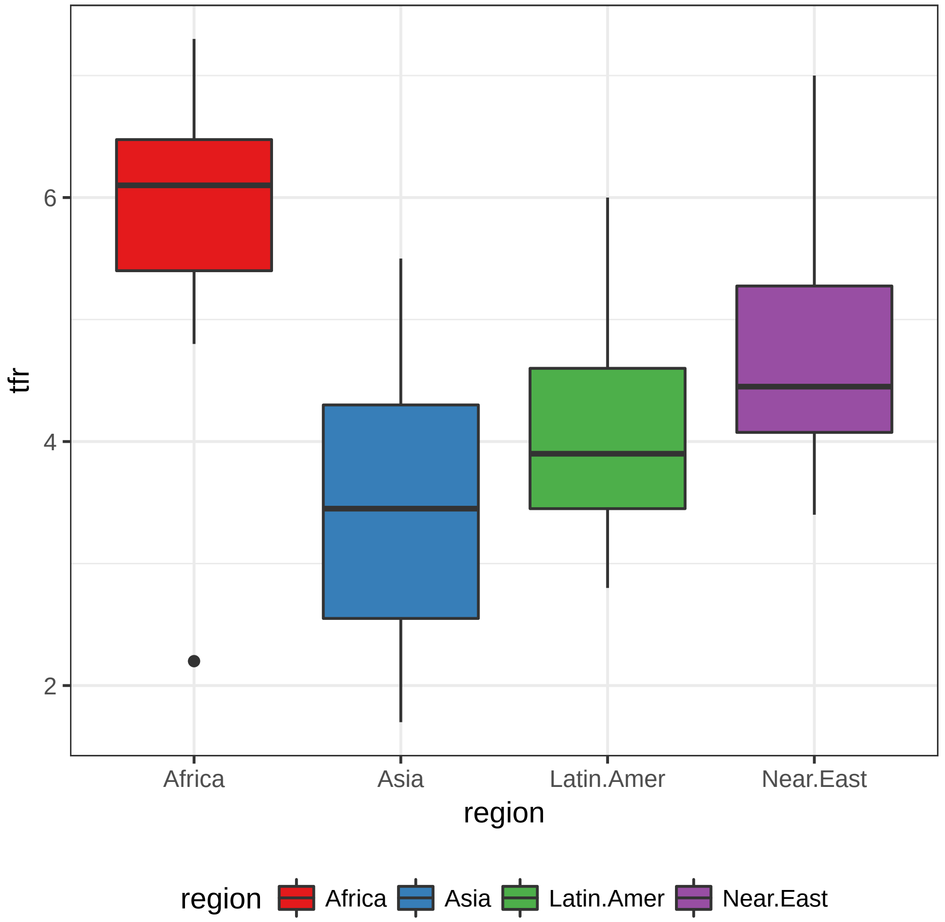
```r
dp <- loadNamespace('dplyr')
import::from(magrittr, `%>%`, `%<>%`)
import::from(readr, read_table2)
import::from(xtable, xtable)
library(ggplot2)
theme_set(theme_bw())

robey.df <- read.table('~/dev/stats-hw/stat-s631/Robey.txt') %>%
  dp$mutate(country = rownames(.))

head(robey.df) %>%
  xtable() %>%
  print(include.rownames = FALSE)
```

region	tfr	contraceptors	country
Africa	4.80	35	Botswana
Africa	6.50	9	Burundi
Africa	5.90	16	Cameroon
Africa	6.10	13	Ghana
Africa	6.50	27	Kenya
Africa	6.40	6	Liberia

`tfr` (total fertility rate) makes the most sense as a response variable since `contraceptors` (percent of married women using contraception) would directly affect this.

Part a

```r
robey.df %>%
  ggplot() +
  geom_boxplot(aes(x = region, group = region, y = tfr, fill = region)) +
  scale_fill_brewer(palette = 'Set1') +
  theme(legend.position = 'bottom')
```



From a plot of fertility rate vs. region, we might expect that there is a somewhat significant difference between Africa and the Near East while there isn't much of a significant difference between Asia and Latin America.

```r
region.mod <- lm(tfr ~ region, data = robey.df)
summary(region.mod)
```

```

Call:
lm(formula = tfr ~ region, data = robey.df)

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 the model summary, we can see that the significance value between Africa (the baseline in the model) vs the Near East is 0.0544. If we set our $\alpha = 0.05$ or 0.01, then we would fail to reject the null hypothesis, although the value is still very close to 0.05 and below 0.1, another common significance level.

Comparing Asia and Latin America is a bit more difficult. One thing we can do is perform the following hypothesis test:

$H_0 : \beta_1 = \beta_2$
$H_A : \beta_1 \neq \beta_2$

Then our test statistic is a t value with degrees of freedom $n - p - 1$:

$$t_{n-p-1} = \frac{\hat{\beta}_2 - \hat{\beta}_3}{\sqrt{a^T \hat{\sigma}^2 (X^T X)^{-1} a}}$$

where $a = [0, 1, -1, 0]^T$

```r
coefs <- summary(region.mod)$coefficients
a <- c(0, 1, -1, 0)
t.stat <-
  (coefs['regionAsia', 'Estimate'] - coefs['regionLatin.Amer', 'Estimate']) /
  sqrt(t(a) %*% vcov(region.mod) %*% a)

# p-value
pt(t.stat, region.mod$df.residual) * 2
```

```
          [,1]
[1,] 0.2705813
```

So for typical values of α we would fail to reject the null hypothesis.

Part b

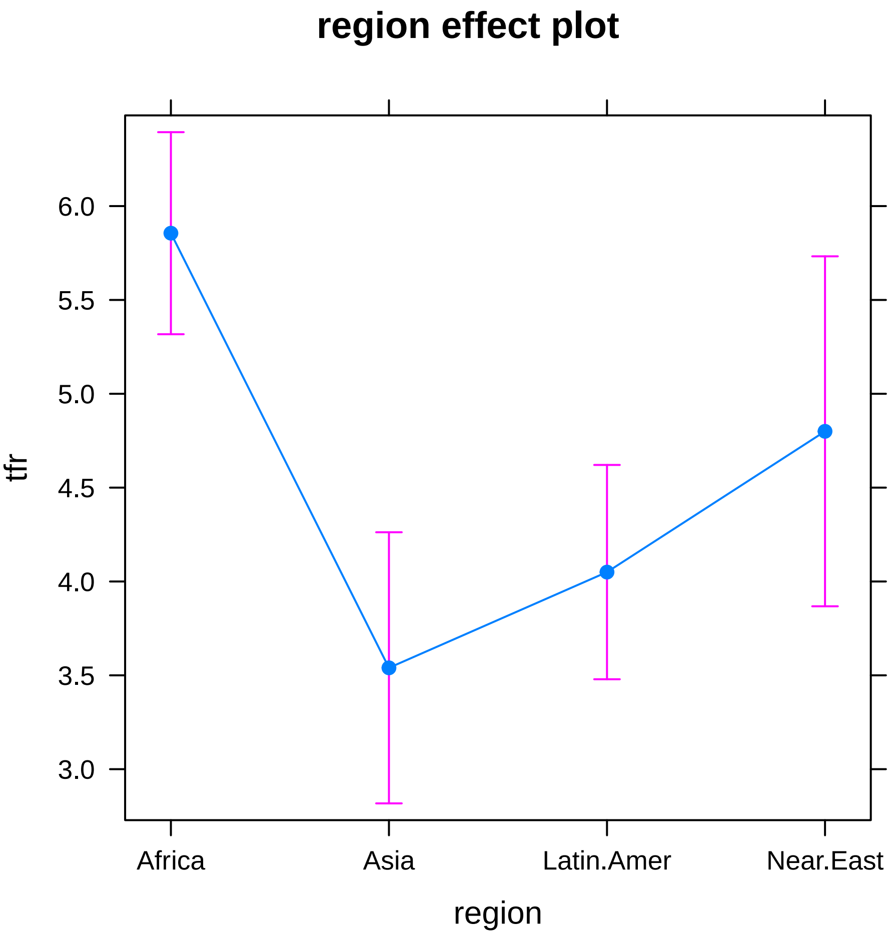
The intercept term $\hat{\beta}_0$ is the average fertility rate of the baseline region, Africa. So on average, the fertility rate of African countries is around 5.856.

The rest of the coefficients are relative to the baseline. So Asia's average fertility rate is 2.316 less than Africa's, Latin America's average fertility rate is 1.806 less than Africa's, and the Near East's average fertility rate is 1.056 less than Africa's.

$E[Y|Africa] = \hat{\beta}_0 = 5.856$
$E[Y|Asia] = \hat{\beta}_0 + \hat{\beta}_1 = 5.856 - 2.316 = 3.54$
$E[Y|LatinAmerica] = \hat{\beta}_0 + \hat{\beta}_2 = 5.856 - 1.806 = 4.05$
$E[Y|NearEast] = \hat{\beta}_0 + \hat{\beta}_3 = 5.856 - 1.056 = 4.8$

Part c

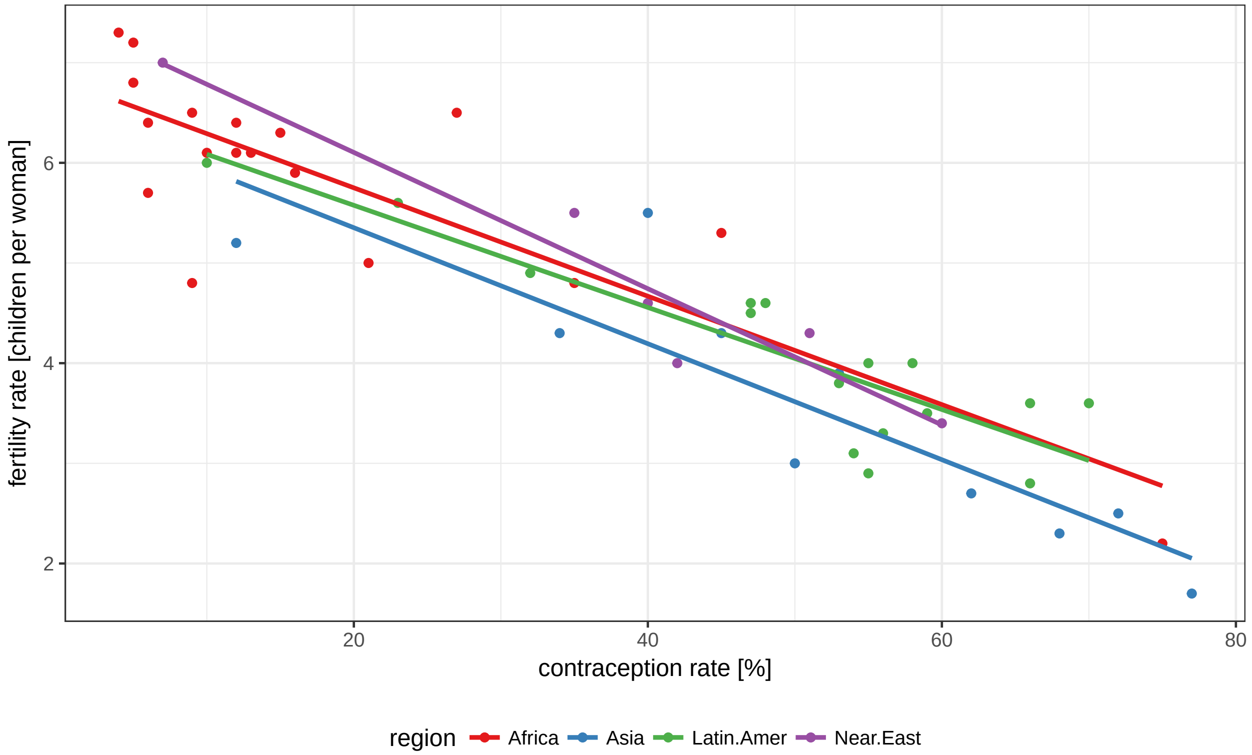
```r
plot(effects::Effect('region', region.mod, confidence.level = .95))
```



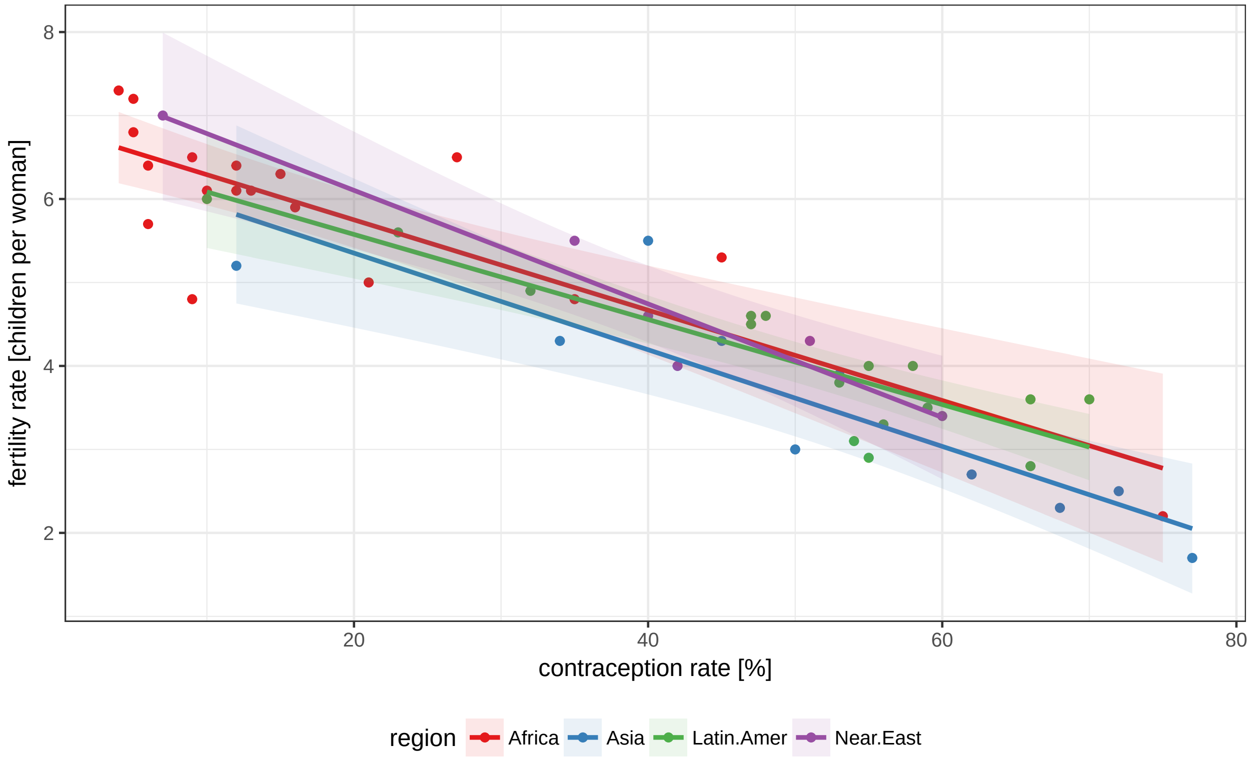
The above plot shows the sample mean and 95% confidence interval of fertility rate for each region. Going back to part a, the sample mean fertility rate of Africa is inside the 95% confidence interval of the fertility rate of the Near East (and vice versa), and the same is true of Asia and Latin America.

Part d

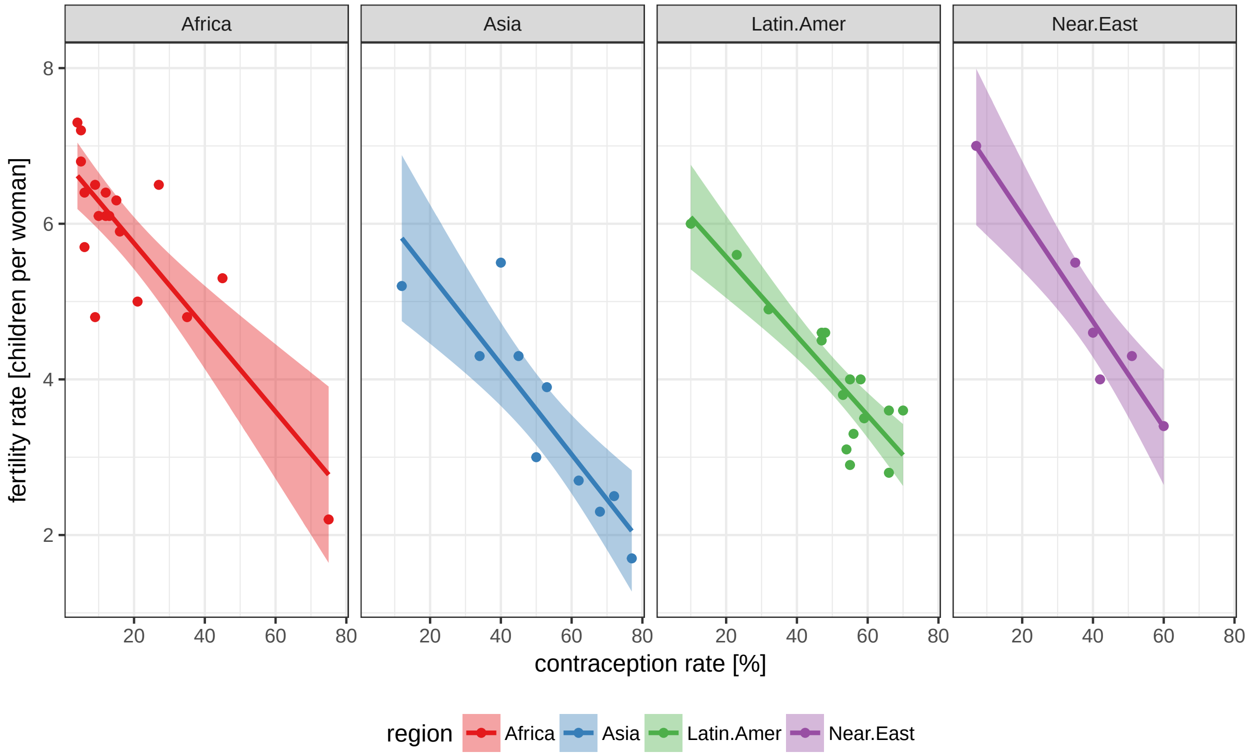
```
ggplot(robey.df) +
  geom_point(aes(x = contraceptors, y = tfr, colour = region)) +
  stat_smooth(method = 'lm',
              aes(x = contraceptors, y = tfr,
                  fill = region, colour = region),
              se = FALSE) +
  scale_colour_brewer(palette = 'Set1') +
  scale_fill_brewer(palette = 'Set1') +
  labs(x = 'contraception rate [%]',
       y = 'fertility rate [children per woman]') +
  theme(legend.position = 'bottom')
```



```
ggplot(robey.df) +
  geom_point(aes(x = contraceptors, y = tfr, colour = region)) +
  stat_smooth(method = 'lm',
              aes(x = contraceptors, y = tfr,
                  fill = region, colour = region),
              alpha = .1) +
  scale_colour_brewer(palette = 'Set1') +
  scale_fill_brewer(palette = 'Set1') +
  labs(x = 'contraception rate [%]',
       y = 'fertility rate [children per woman]') +
  theme(legend.position = 'bottom')
```



```
ggplot(robey.df) +
  geom_point(aes(x = contraceptors, y = tfr, colour = region)) +
  stat_smooth(method = 'lm',
              aes(x = contraceptors, y = tfr,
                  fill = region, colour = region)) +
  scale_colour_brewer(palette = 'Set1') +
  scale_fill_brewer(palette = 'Set1') +
  labs(x = 'contraception rate [%]',
       y = 'fertility rate [children per woman]') +
  theme(legend.position = 'bottom') +
  facet_wrap(~ region, nrow = 1)
```



We can nitpick a bit, but overall, it appears that there's no reason to believe that there is a significant difference in slopes or intercepts among the regions. Especially once we add the standard errors, they all overlap.

Part e

```
full.mod <- lm(tfr ~ contraceptors * region, data = robey.df)
summary(full.mod)
```

```

Call:
lm(formula = tfr ~ contraceptors * region, data = robey.df)

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 ***
contraceptors                  -0.054099   0.007718  -7.009 1.41e-08 ***
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:regionAsia       -0.003795   0.012389  -0.306    0.761    
contraceptors:regionLatin.Amer  0.003136   0.012044   0.260    0.796    
contraceptors:regionNear.East  -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
```

As expected, the `region` terms (differences in the intercept among regions) and the interaction terms (differences in the slope among regions) are all not significant, although we can't say this just from the individual t-tests.

Again, the baseline region is Africa. On average, if no country in Africa used contraception, a fertility rate of 6.833% is expected (although we can't exactly say this since contraception rate of 0 isn't in the data for any country). Similarly, the average fertility rates of Asia, Latin America, and the Near East are 6.51%, 6.595%, and 7.464% respectively, given that the contraceptive use rate is 0.

Moving onto the slopes, the average fertility rate in Africa decreases by 0.054 per 1% increase in contraception use. For Asia, this value is -0.058, for Latin America, it's -0.051 and for the Near East, it's -0.068.

Part f

First, we can build a few more models:

```
contraceptor.mod <- lm(tfr ~ contraceptors, data = robey.df)
summary(contraceptor.mod)
```

```

Call:
lm(formula = tfr ~ contraceptors, data = robey.df)

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
```

```
no.interact.mod <- lm(tfr ~ contraceptors + region, data = robey.df)
summary(no.interact.mod)
```

```

Call:
lm(formula = tfr ~ contraceptors + region, data = robey.df)

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 ***
contraceptors    -0.05575    0.00466 -11.963 1.42e-15 ***
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    
---
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
```

```r
slopes.mod <- lm(tfr ~ contraceptors:region, data = robey.df)
summary(slopes.mod)
```

```

Call:
lm(formula = tfr ~ contraceptors:region, data = robey.df)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.53692 -0.26526 -0.05922  0.35752  1.20887 

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     6.820867   0.161691  42.185  < 2e-16 ***
contraceptors:regionAfrica     -0.053772   0.007024  -7.655  1.1e-09 ***
contraceptors:regionAsia       -0.063243   0.004294 -14.727  < 2e-16 ***
contraceptors:regionLatin.Amer -0.055089   0.004002 -13.765  < 2e-16 ***
contraceptors:regionNear.East  -0.054074   0.006460  -8.371  1.0e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5649 on 45 degrees of freedom
Multiple R-squared:  0.8613,    Adjusted R-squared:  0.849 
F-statistic: 69.85 on 4 and 45 DF,  p-value: < 2.2e-16
```

```r
anova(contraceptor.mod, no.interact.mod, full.mod)
```

```
Analysis of Variance Table

Model 1: tfr ~ contraceptors
Model 2: tfr ~ contraceptors + region
Model 3: tfr ~ contraceptors * region
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     48 15.840                           
2     45 14.163  3   1.67724 1.7018 0.1812
3     42 13.798  3   0.36524 0.3706 0.7746
```

```r
anova(region.mod, no.interact.mod, full.mod)
```

```
Analysis of Variance Table

Model 1: tfr ~ region
Model 2: tfr ~ contraceptors + region
Model 3: tfr ~ contraceptors * region
  Res.Df    RSS Df Sum of Sq        F    Pr(>F)    
1     46 59.208                                    
2     45 14.163  1    45.045 137.1158 8.226e-15 ***
3     42 13.798  3     0.365   0.3706    0.7746    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(slopes.mod, full.mod)
```

```
Analysis of Variance Table

Model 1: tfr ~ contraceptors:region
Model 2: tfr ~ contraceptors * region
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     45 14.358                           
2     42 13.798  3   0.56068 0.5689 0.6386
```

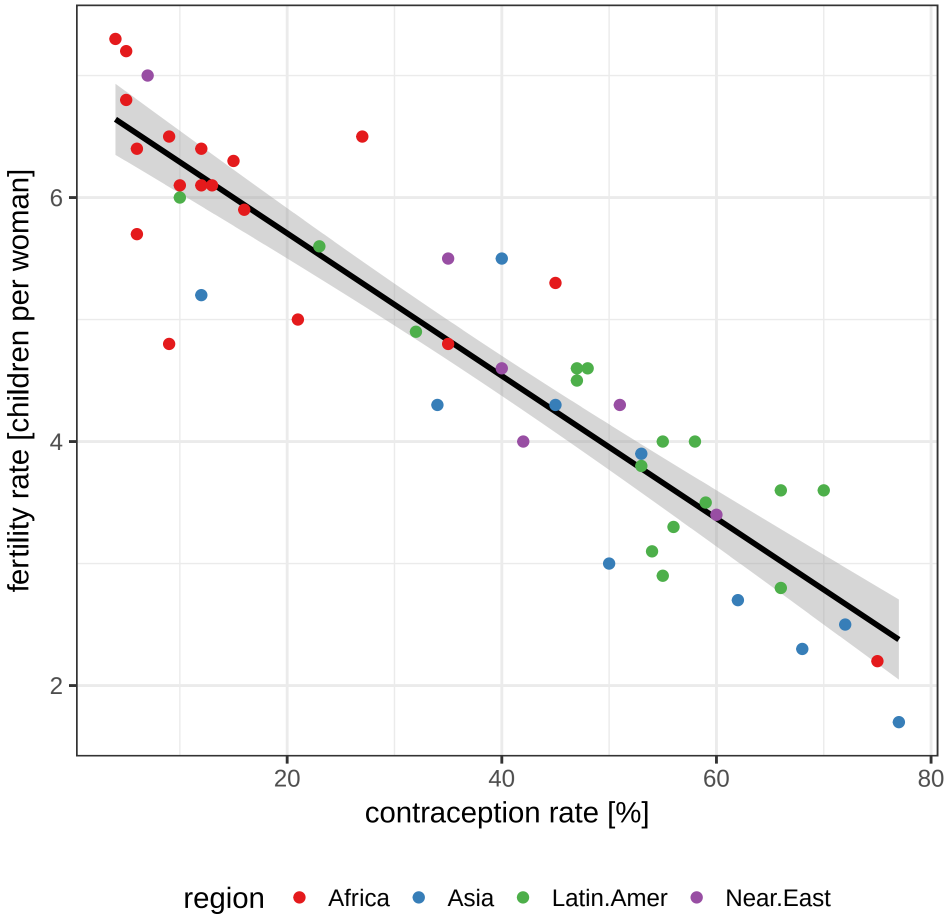
```
anova(contraceptor.mod, slopes.mod)
```

```
Analysis of Variance Table

Model 1: tfr ~ contraceptors
Model 2: tfr ~ contraceptors:region
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1     48 15.840                          
2     45 14.358  3    1.4818 1.548 0.2152
```

Just from the individual model summaries, we might be tempted to say `slopes.mod`, the model for which each region has its own slope but they all share the same intercept, is the best model, since all of the coefficients are significant. However, we can't take all of the individual t-tests together as one summary, and from the F-tests, it appears that this model is not significantly different from the model that disregards region altogether and just regresses on contraception use rate.

Part g

```
ggplot(robey.df) +
  stat_smooth(method = 'lm', colour = 'black',
              aes(x = contraceptors, y = tfr)) +
  geom_point(aes(x = contraceptors, y = tfr, colour = region)) +
  scale_colour_brewer(palette = 'Set1') +
  labs(x = 'contraception rate [%]',
       y = 'fertility rate [children per woman]') +
  theme(legend.position = 'bottom')
```



```
summary(contraceptor.mod)
```

```

Call:
lm(formula = tfr ~ contraceptors, data = robey.df)

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
```

Since there is only one regressor in the model, the interpretation of the plot is straightforward. According to the model, the relationship between `ftr`, the response, and `contraceptors`, the explanatory variable, is on average linear. For a unit increase in `contraceptors`, we expect, on average, that `ftr` decreases by 0.058.

Part h

For this part, we'll set $\alpha = .01$.

```
predict(contraceptor.mod,
        newdata = dp$summarise_all(robey.df, mean),
        interval = 'predict',
        confidence = .99)
```

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

We are 99% confident that this interval will capture the fertility rate of a new country that happens to have a contraception use rate that is the sample mean of all of the other countries.