Lab 1e. Linear Regression

Multiple Regression and Validation of Assumptions

Yerim Oh

Country Data

We have data on 193 countries, with the following variables recorded (the data were assembled in 2012, so all values are for the countries as of that year):

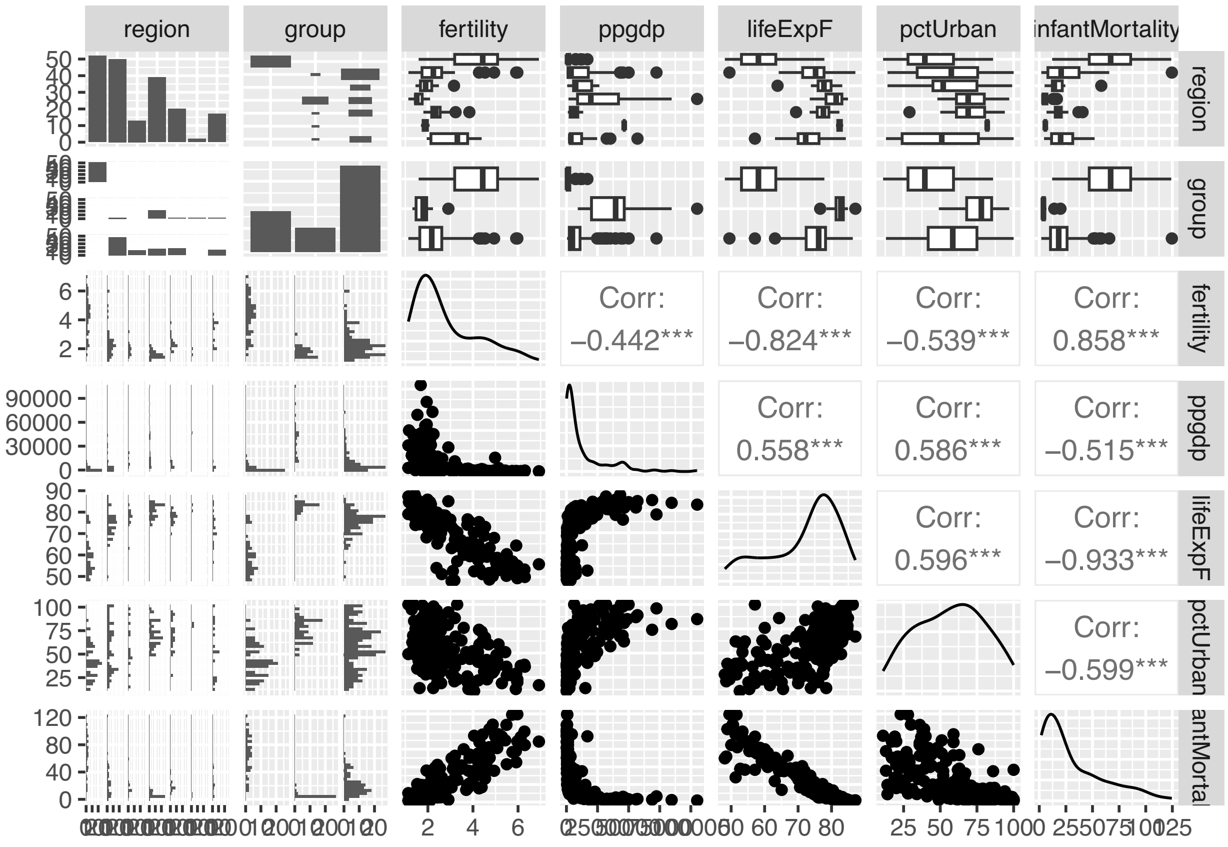
- `region`: Region of the world: Africa, Asia, Caribbean, Europe, Latin Amer, North America, NorthAtlantic, Oceania.
- `group`: A factor with levels oecd for countries that are members of the OECD, the Organization for Economic Co-operation and Development, as of May 2012, africa for countries on the African continent, and other for all other countries. No OECD countries are located in Africa.
- `fertility`: Total fertility rate, number of children per woman.
- `ppgdp`: Per capita gross domestic product in US dollars.
- `lifeExpF`: Female life expectancy, years.
- `pctUrban`: Percent urban.
- `infantMortality`: Infant deaths by age 1 year per 1000 live births

Let's develop models to predict `infantMortality` as a function of the other variables.

```
countries <- read.csv("http://www.cknudson.com/data/UN.csv")
countries <- countries[,-1]
countries <- countries %>% drop_na()
```

1. Make some exploratory plots of the data.

```
ggpairs(countries)
```



2. Fit a model that includes `region` and `group` as explanatory variables and take a look at the summary output. What is going on?

```
model_fit1 <- lm(infantMortality ~ region + group, countries)
summary(model_fit1)
```

```
##
## Call:
## lm(formula = infantMortality ~ region + group, data = countries)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.209  -7.993  -0.767   5.968  98.356
##
## Coefficients: (1 not defined because of singularities)
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)            65.321      2.669  24.476  < 2e-16 ***
## regionAsia            -39.141      3.823 -10.239  < 2e-16 ***
## regionCaribbean       -46.998      5.968  -7.876 2.79e-13 ***
## regionEurope          -53.734      4.899 -10.969  < 2e-16 ***
## regionLatin Amer      -45.020      5.086  -8.851 6.95e-16 ***
## regionNorth America  -49.940     14.680  -3.402  0.00082 ***
## regionOceania         -41.764      5.406  -7.725 6.86e-13 ***
## groupoecd              -9.688      4.815  -2.012  0.04566 *
## groupother                 NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.24 on 185 degrees of freedom
## Multiple R-squared:  0.5827, Adjusted R-squared:  0.5669
## F-statistic: 36.91 on 7 and 185 DF,  p-value: < 2.2e-16
```

The NA estimate is a sign of multicollinearity. Looking back at the variable descriptions, we see that there is an "Africa" level for region and an "africa" level for the group. Those variables contain duplicate information in that case.

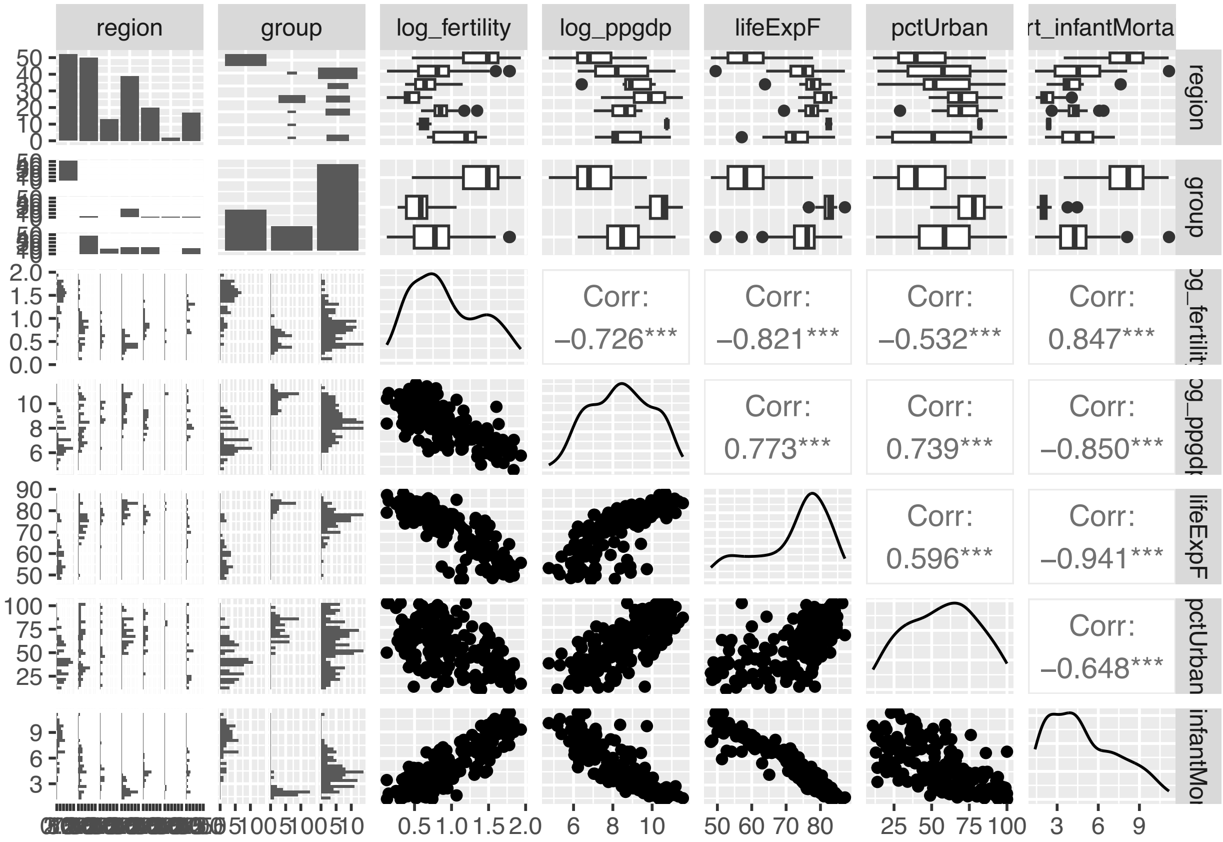
```
group_africa_inds <- which(countries$group == "africa")
region_africa_inds <- which(countries$region == "Africa")
identical(group_africa_inds, region_africa_inds)
```

```
## [1] TRUE
```

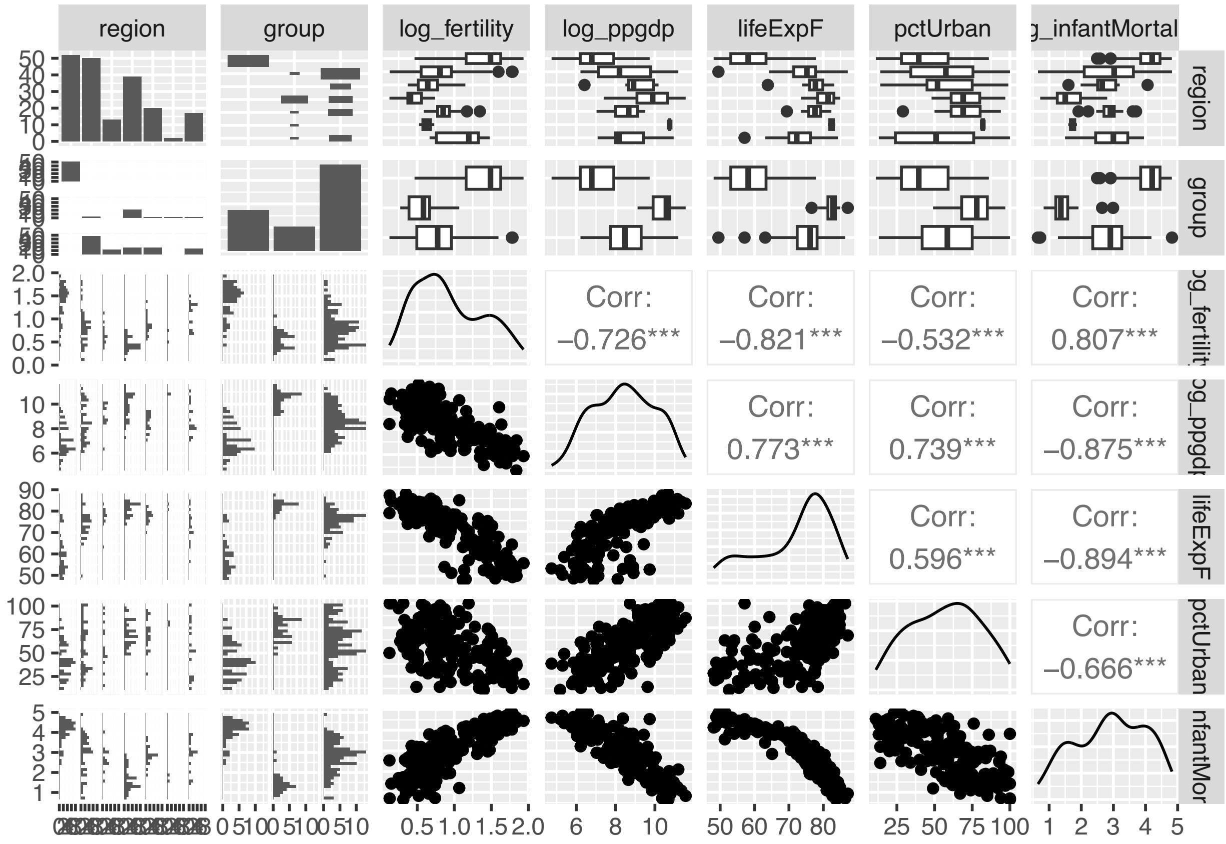
3. Develop a predictive model by transforming the response and/or quantitative explanatory variables so that the associations between the transformed variables are approximately linear. Include just one of region and group in your model.

```

countries_transformed <- countries %>%
  transmute(region = region,
            group = group,
            log_fertility = log(fertility),
            log_ppgdp = log(ppgdp),
            lifeExpF = lifeExpF,
            pctUrban = pctUrban,
            sqrt_infantMortality = sqrt(infantMortality))

ggpairs(countries_transformed)
```

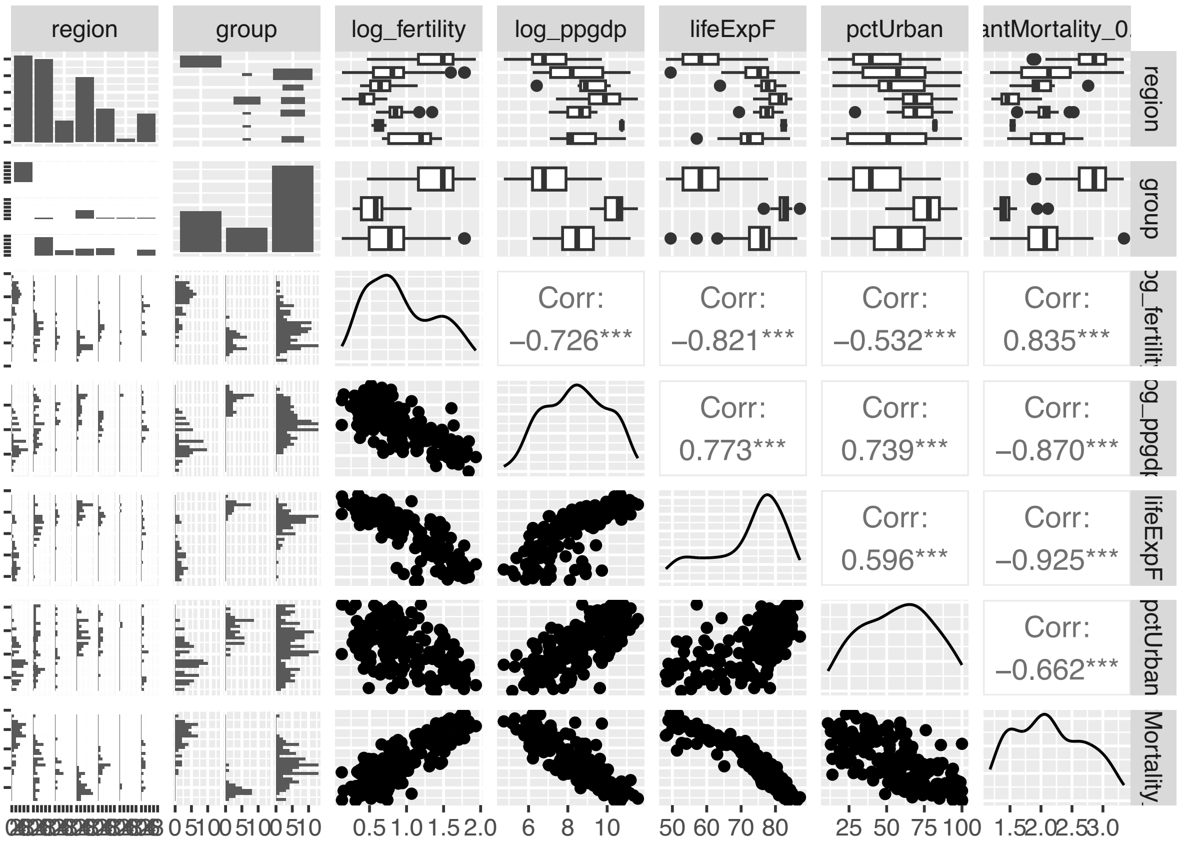
(a) As a challenge, see if you can make a successful guess about the transformations to use on the first try. Make plots of your transformed variables to see how you're doing. You should feel pretty good about your selected transformations before fitting any models. You only need to keep your final selections for transformations, no need to keep any intermediate steps.



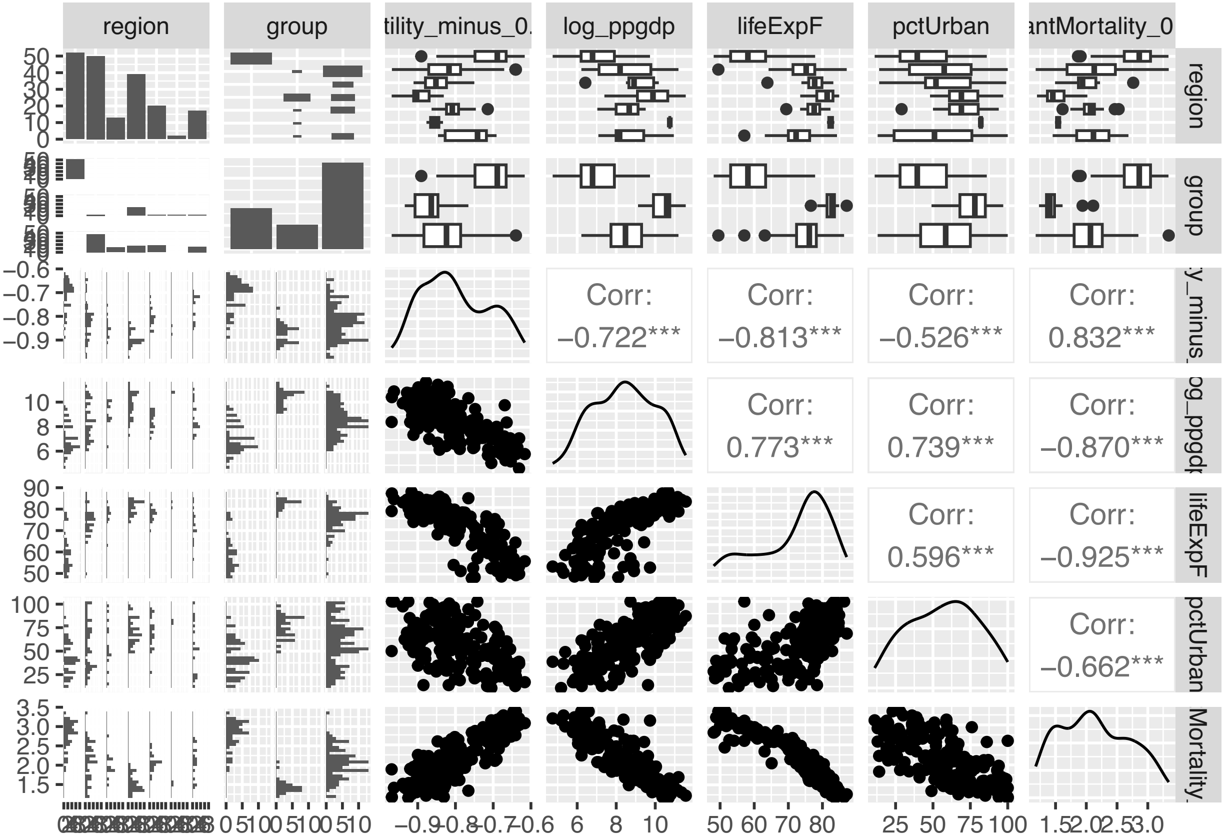
```
countries_transformed <- countries %>%
  transmute(region = region,
            group = group,
            log_fertility = log(fertility),
            log_ppgdp = log(ppgdp),
            lifeExpF = lifeExpF,
            pctUrban = pctUrban,
            log_infantMortality = log(infantMortality))

ggpairs(countries_transformed)
```



```
countries_transformed <- countries %>%
  transmute(region = region,
            group = group,
            log_fertility = log(fertility),
            log_ppgdp = log(ppgdp),
            lifeExpF = lifeExpF,
            pctUrban = pctUrban,
            infantMortality_0.25 = (infantMortality)^0.25)

ggpairs(countries_transformed)
```



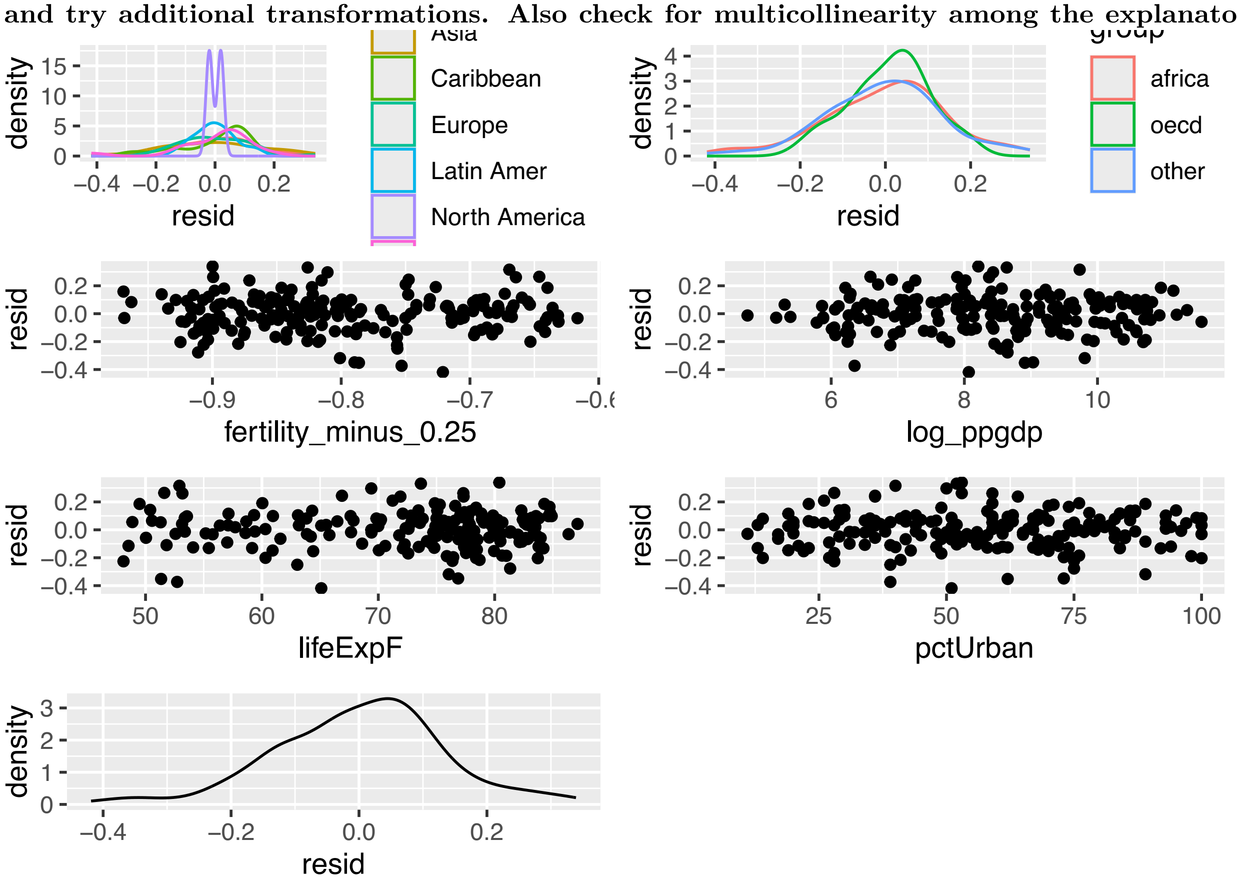
```r
countries_transformed <- countries %>%
  transmute(region = region,
            group = group,
            fertility_minus_0.25 = -1/(fertility^0.25),
            log_ppgdp = log(ppgdp),
            lifeExpF = lifeExpF,
            pctUrban = pctUrban,
            infantMortality_0.25 = (infantMortality)^0.25)

ggpairs(countries_transformed)
```



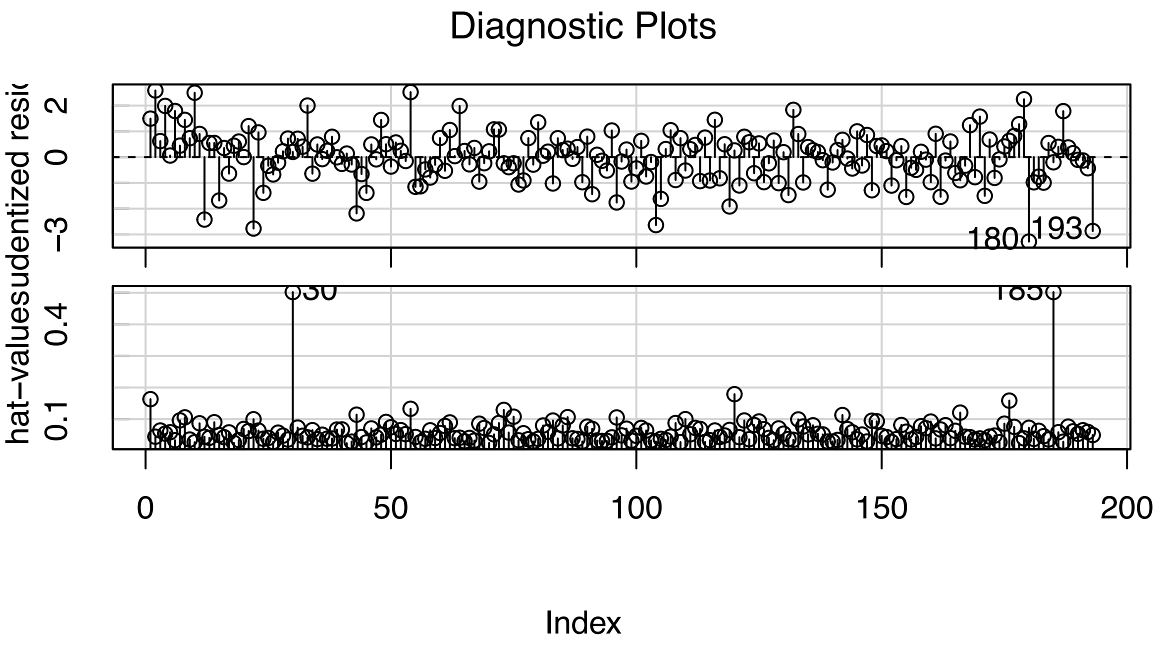
- I don't think additional transformations of infant mortality can simultaneously improve the relationships with fertility and pctUrban; the heteroskedasticity is kind of going in opposite directions in those two plots. But it's not too severe in either plot at this point.
- There is a non-linear relationship between lifeExpF and my transformed infant mortality variable, but the variance of the response is fairly constant across the range of values for lifeExpF. I will handle that non-linearity with a polynomial term in lifeExpF.

```
lm_fit <- lm(infantMortality_0.25 ~ region +
               fertility_minus_0.25 +
               log_ppgdp +
               poly(lifeExpF, 2, raw = TRUE)  +
               pctUrban,
             data = countries_transformed)
countries_transformed <- countries_transformed %>%
  mutate(resid = residuals(lm_fit))

# scatter plots of the residuals vs. each quantitative explanatory variable
sp1 <- ggplot(data = countries_transformed,
              mapping = (aes(x = resid, color = region)))+
  geom_density()
sp2 <- ggplot(data = countries_transformed,
              mapping = (aes(x = resid, color = group)))+
  geom_density()
sp3 <- ggplot(data = countries_transformed,
              mapping = (aes(x = fertility_minus_0.25, y = resid)))+
  geom_point()
sp4 <- ggplot(data = countries_transformed,
              mapping = (aes(x = log_ppgdp, y = resid)))+
  geom_point()
sp5 <- ggplot(data = countries_transformed,
              mapping = (aes(x = lifeExpF, y = resid)))+
  geom_point()
sp6 <- ggplot(data = countries_transformed,
              mapping = (aes(x = pctUrban, y = resid)))+
  geom_point()
sp7 <- ggplot(data = countries_transformed, mapping = (aes(x = resid)))+
  geom_density()

grid.arrange(sp1, sp2, sp3, sp4, sp5, sp6, sp7, ncol=2)
```

(b) Fit a model to your transformed data and create a set of diagnostic plots. These should include (i) scatter plots of the residuals vs. each quantitative explanatory variable in the data set (whether or not you included it in your model); (ii) a density plot or histogram of the residuals; and (iii) diagnostic plots of studentized residuals, and leverage. If you see any serious issues, go back to step (a) and try additional transformations. Also check for multicollinearity among the explanatory variables.



```
# diagnostic plots of studentized residuals, and leverage
car::influenceIndexPlot(lm_fit,
                        vars = c("Studentized", "hat"))
```

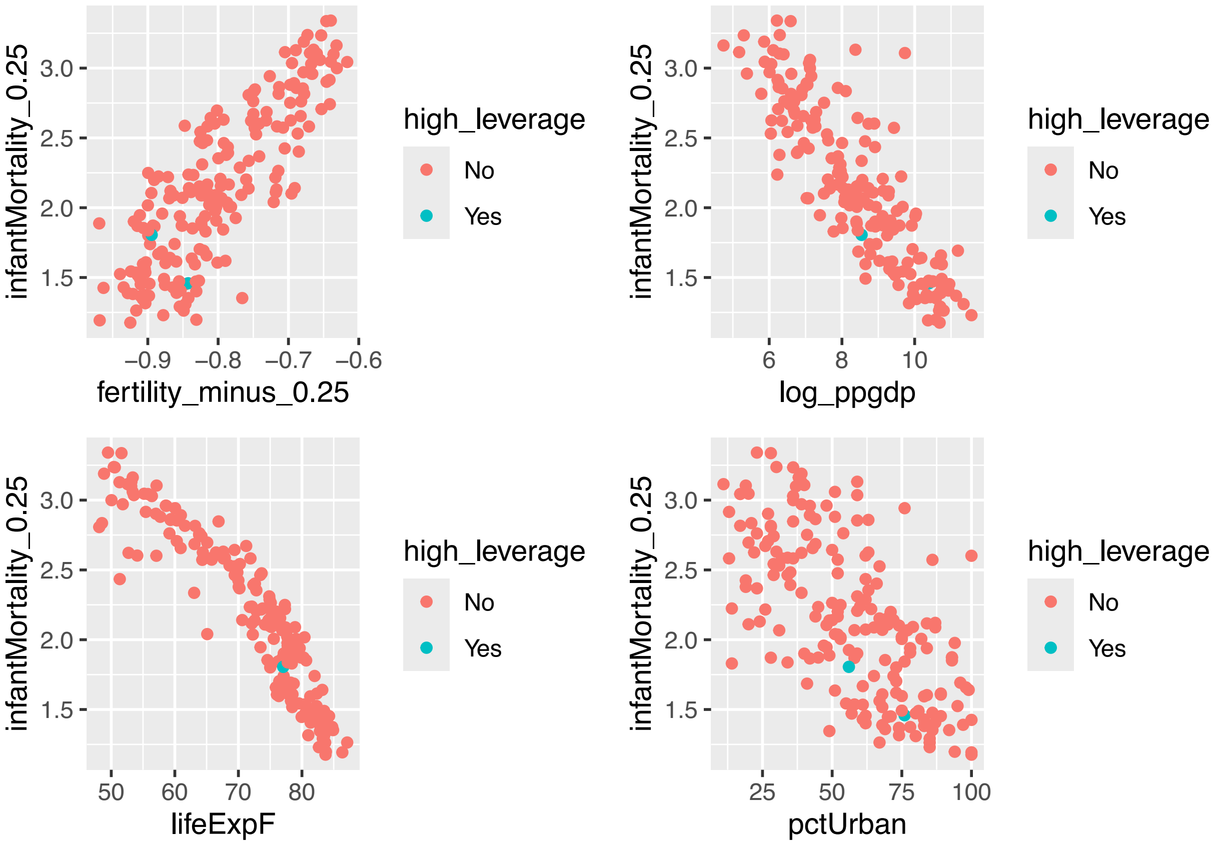


```
2 * length(coef(lm_fit)) / nrow(countries_transformed)# leverage ("hat-values")
```

```
## [1] 0.1243523
```

Observations 24 and 153 show high leverage. Let's look at plots to see if we're worried:

```
countries_transformed$high_leverage <- "No"
countries_transformed$high_leverage[c(24, 153)] <- "Yes"
p1 <- ggplot(data = countries_transformed,
             mapping = aes(x = fertility_minus_0.25,
                           y = infantMortality_0.25,
                           color = high_leverage)) +
  geom_point()
p2 <- ggplot(data = countries_transformed,
             mapping = aes(x = log_ppgdp,
                           y = infantMortality_0.25,
                           color = high_leverage)) +
  geom_point()
p3 <- ggplot(data = countries_transformed,
             mapping = aes(x = lifeExpF,
                           y = infantMortality_0.25,
                           color = high_leverage)) +
  geom_point()
p4 <- ggplot(data = countries_transformed,
             mapping = aes(x = pctUrban,
                           y = infantMortality_0.25,
                           color = high_leverage)) +
  geom_point()
grid.arrange(p1, p2, p3, p4)
```



Not particularly worried.

```
lm_fit_no_high_leverage <- lm(infantMortality_0.25 ~ region +
                                fertility_minus_0.25 +
                                log_ppgdp +
                                poly(lifeExpF, 2, raw = TRUE) +
                                pctUrban,
                              data = countries_transformed %>%
                                filter(high_leverage == "No"))

summary(lm_fit)
```

```
##
## Call:
## lm(formula = infantMortality_0.25 ~ region + fertility_minus_0.25 +
##     log_ppgdp + poly(lifeExpF, 2, raw = TRUE) + pctUrban, data = countries_transformed)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41890 -0.08463  0.00808  0.08117  0.33922
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    3.1147529  0.5930445   5.252 4.19e-07 ***
## regionAsia                     0.0152343  0.0376301   0.405  0.68607
## regionCaribbean                0.0893622  0.0524395   1.704  0.09008 .
## regionEurope                  -0.1137908  0.0472060  -2.411  0.01693 *
## regionLatin Amer               0.1123198  0.0466192   2.409  0.01699 *
## regionNorth America            0.0487761  0.1055323   0.462  0.64450
## regionOceania                 -0.1046066  0.0454322  -2.302  0.02245 *
## fertility_minus_0.25           0.5931509  0.2244648   2.643  0.00895 **
## log_ppgdp                     -0.0734660  0.0144954  -5.068 9.86e-07 ***
## poly(lifeExpF, 2, raw = TRUE)1  0.0491169  0.0167966   2.924  0.00389 **
## poly(lifeExpF, 2, raw = TRUE)2 -0.0006309  0.0001279  -4.933 1.82e-06 ***
## pctUrban                       -0.0010690  0.0006673  -1.602  0.11088
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1362 on 181 degrees of freedom
## Multiple R-squared:  0.9459, Adjusted R-squared:  0.9426
## F-statistic: 287.8 on 11 and 181 DF,  p-value: < 2.2e-16
```

```r
summary(lm_fit_no_high_leverage)
```

```
##
## Call:
## lm(formula = infantMortality_0.25 ~ region + fertility_minus_0.25 +
##     log_ppgdp + poly(lifeExpF, 2, raw = TRUE) + pctUrban, data = countries_transformed %>%
##     filter(high_leverage == "No"))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41783 -0.08486  0.01060  0.07824  0.34129
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     3.1194968  0.5939367   5.252 4.23e-07 ***
## regionAsia                      0.0176982  0.0376805   0.470  0.63915
## regionCaribbean                 0.0869888  0.0524829   1.657  0.09918 .
## regionEurope                   -0.1165407  0.0474976  -2.454  0.01510 *
## regionLatin Amer                0.1106502  0.0466513   2.372  0.01876 *
## regionNorth America             0.0436466  0.1056548   0.413  0.68002
## regionOceania                  -0.1071205  0.0454874  -2.355  0.01961 *
## fertility_minus_0.25            0.6051678  0.2250528   2.689  0.00784 **
## log_ppgdp                      -0.0717393  0.0145672  -4.925 1.91e-06 ***
## poly(lifeExpF, 2, raw = TRUE)1  0.0488508  0.0168218   2.904  0.00415 **
## poly(lifeExpF, 2, raw = TRUE)2 -0.0006285  0.0001281  -4.907 2.07e-06 ***
## pctUrban                       -0.0010935  0.0006677  -1.638  0.10320
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1363 on 179 degrees of freedom
## Multiple R-squared:  0.946,  Adjusted R-squared:  0.9426
## F-statistic: 284.9 on 11 and 179 DF,  p-value: < 2.2e-16
```

There are essentially no differences between the model fits with and without the high leverage observations. Not worried at all.

```r
summary(lm_fit)
```

(c) Take a look at the summary output for your chosen model. Which variables would hypothesis tests suggest have a strong relationship with infant mortality rates?

```
##
## Call:
## lm(formula = infantMortality_0.25 ~ region + fertility_minus_0.25 +
##     log_ppgdp + poly(lifeExpF, 2, raw = TRUE) + pctUrban, data = countries_transformed)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41890 -0.08463  0.00808  0.08117  0.33922
##
```

```
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   3.1147529  0.5930445   5.252 4.19e-07 ***
## regionAsia                    0.0152343  0.0376301   0.405  0.68607
## regionCaribbean               0.0893622  0.0524395   1.704  0.09008 .
## regionEurope                 -0.1137908  0.0472060  -2.411  0.01693 *
## regionLatin Amer              0.1123198  0.0466192   2.409  0.01699 *
## regionNorth America           0.0487761  0.1055323   0.462  0.64450
## regionOceania                -0.1046066  0.0454322  -2.302  0.02245 *
## fertility_minus_0.25          0.5931509  0.2244648   2.643  0.00895 **
## log_ppgdp                    -0.0734660  0.0144954  -5.068 9.86e-07 ***
## poly(lifeExpF, 2, raw = TRUE)1  0.0491169  0.0167966   2.924  0.00389 **
## poly(lifeExpF, 2, raw = TRUE)2 -0.0006309  0.0001279  -4.933 1.82e-06 ***
## pctUrban                     -0.0010690  0.0006673  -1.602  0.11088
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1362 on 181 degrees of freedom
## Multiple R-squared:  0.9459, Adjusted R-squared:  0.9426
## F-statistic: 287.8 on 11 and 181 DF,  p-value: < 2.2e-16
```

There is fairly strong evidence of an association between fertility, ppgdp, and lifeExpF and infant mortality rates. We have to conduct an F test to investigate region:

```
reduced_fit <- lm(infantMortality_0.25 ~ fertility_minus_0.25 +
                    log_ppgdp +
                    poly(lifeExpF, 2, raw = TRUE) +
                    pctUrban,
                  data = countries_transformed)

anova(reduced_fit, lm_fit)
```

```
## Analysis of Variance Table
##
## Model 1: infantMortality_0.25 ~ fertility_minus_0.25 + log_ppgdp + poly(lifeExpF,
##     2, raw = TRUE) + pctUrban
## Model 2: infantMortality_0.25 ~ region + fertility_minus_0.25 + log_ppgdp +
##     poly(lifeExpF, 2, raw = TRUE) + pctUrban
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    187 4.2820
## 2    181 3.3592  6   0.92277 8.2866 6.147e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

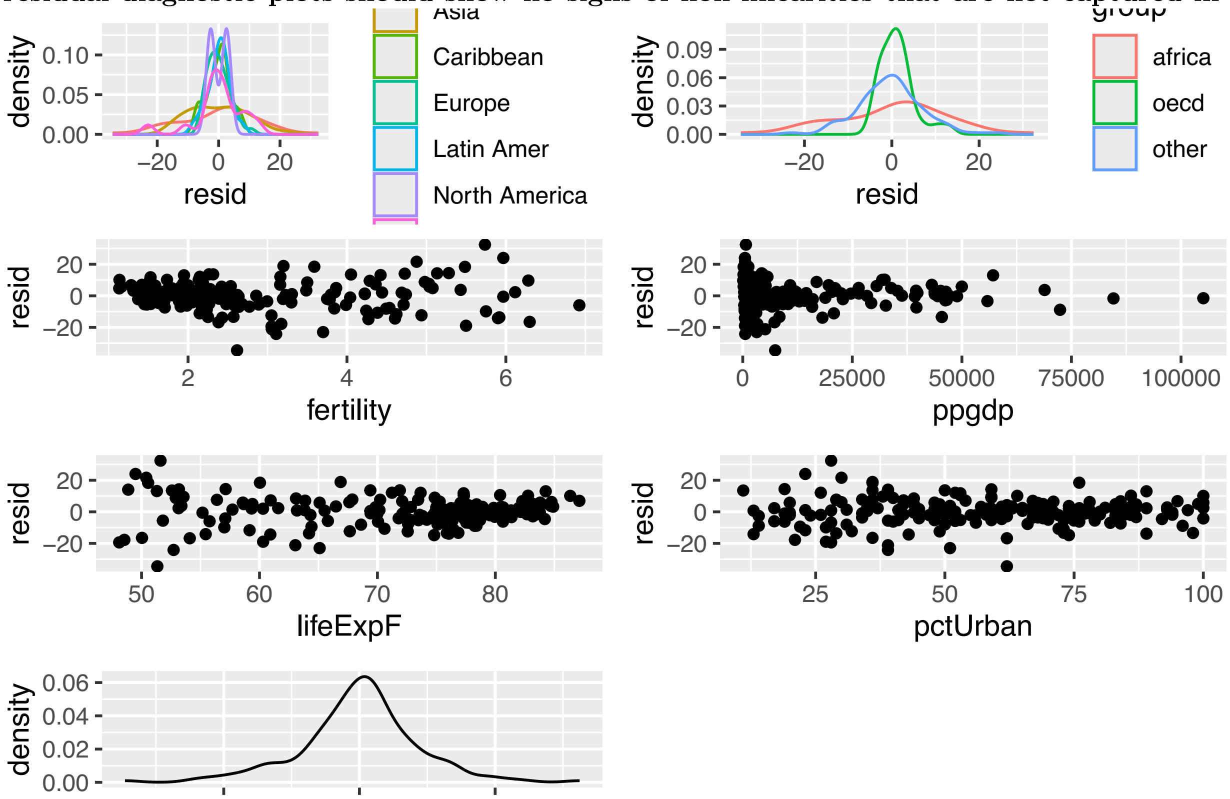
A hypothesis test says region is important too. Basically the only variable a hypothesis test says we could drop is `pctUrban`.

4. Develop a predictive model by including polynomial terms in any explanatory variables that have a non-linear relationship with the response. You should not use any transformations of the explanatory or response variables for this model.

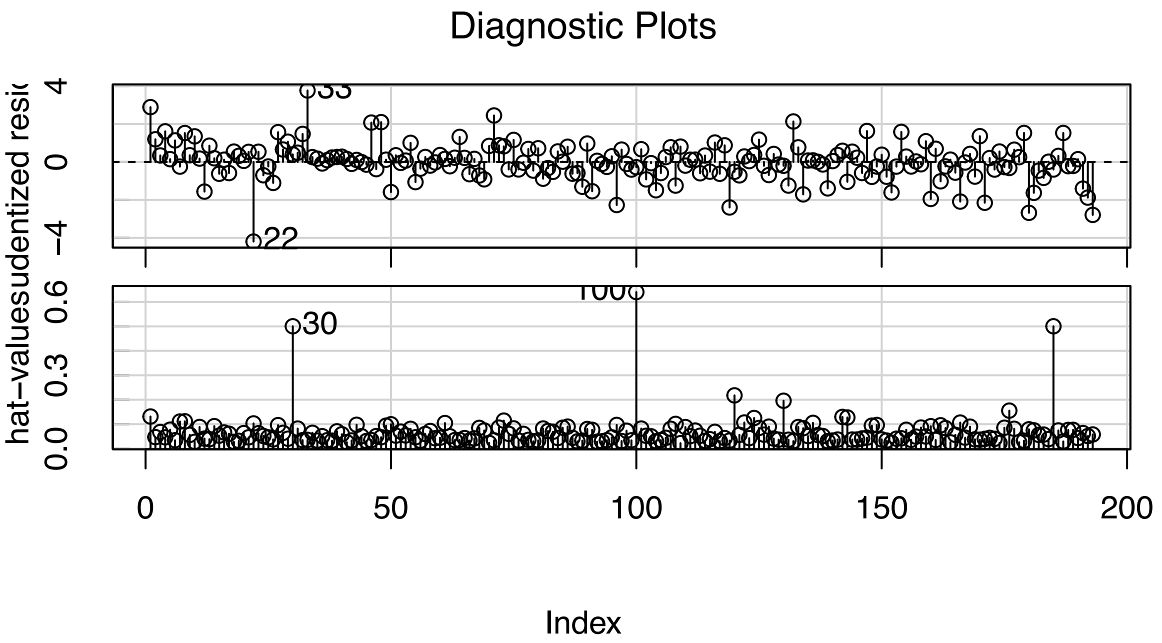
```

lm_fit2 <- lm(infantMortality ~ region +
                fertility +
                poly(ppgdp, 2, raw = TRUE) +
                lifeExpF +
                poly(pctUrban, 2, raw = TRUE),
              data = countries)
countries <- countries %>%
  mutate(resid = residuals(lm_fit2))

# scatter plots of the residuals vs. each quantitative explanatory variable
```

```r
sp1 <- ggplot(data = countries, mapping = (aes(x = resid, color = region)))+
  geom_density()
sp2 <- ggplot(data = countries, mapping = (aes(x = resid, color = group)))+
  geom_density()
sp3 <- ggplot(data = countries, mapping = (aes(x = fertility, y = resid)))+
  geom_point()
sp4 <- ggplot(data = countries, mapping = (aes(x = ppgdp, y = resid)))+
  geom_point()
sp5 <- ggplot(data = countries, mapping = (aes(x = lifeExpF, y = resid)))+
  geom_point()
sp6 <- ggplot(data = countries, mapping = (aes(x = pctUrban, y = resid)))+
  geom_point()
sp7 <- ggplot(data = countries, mapping = (aes(x = resid)))+
  geom_density()

grid.arrange(sp1, sp2, sp3, sp4, sp5, sp6, sp7, ncol=2)
```

(a) Fit your selected model to the data and create the same diagnostic plots you made for your model in part 3(b). You will not be able to solve problems with non-constant variance of the residuals (heteroskedasticity), but the residual diagnostic plots should show no signs of non-linearities that are not captured in your model.



```r
# diagnostic plots of studentized residuals, and leverage
car::influenceIndexPlot(lm_fit2,
                        vars = c("Studentized", "hat"))
```



```r
2 * length(coef(lm_fit2)) / nrow(countries)# leverage ("hat-values")
```

```
## [1] 0.134715
```

There are clear problems, but this is about as good as you can do without using transformations.

```r
summary(lm_fit2)
```

(b) Take a look at the summary output for your chosen model. Which variables would hypothesis tests suggest have a strong relationship with infant mortality rates?

```
##
## Call:
## lm(formula = infantMortality ~ region + fertility + poly(ppgdp,
##     2, raw = TRUE) + lifeExpF + poly(pctUrban, 2, raw = TRUE),
##     data = countries)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.553  -4.177   0.506   4.728  32.447
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      1.615e+02  1.197e+01  13.492  < 2e-16 ***
## regionAsia                       2.837e+00  2.507e+00   1.131  0.25938
## regionCaribbean                  3.237e+00  3.495e+00   0.926  0.35553
## regionEurope                     1.091e+00  3.088e+00   0.353  0.72429
## regionLatin Amer                 3.872e+00  3.224e+00   1.201  0.23127
## regionNorth America              1.613e+00  7.125e+00   0.226  0.82116
## regionOceania                   -9.204e+00  2.945e+00  -3.125  0.00207 **
## fertility                        6.353e+00  9.220e-01   6.890 9.03e-11 ***
## poly(ppgdp, 2, raw = TRUE)1      1.156e-04  1.378e-04   0.839  0.40261
## poly(ppgdp, 2, raw = TRUE)2     -6.848e-10  1.561e-09  -0.439  0.66140
## lifeExpF                        -2.005e+00  1.496e-01 -13.407  < 2e-16 ***
## poly(pctUrban, 2, raw = TRUE)1 -1.134e-01  1.523e-01  -0.745  0.45731
## poly(pctUrban, 2, raw = TRUE)2  1.887e-04  1.346e-03   0.140  0.88867
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 9.086 on 180 degrees of freedom
## Multiple R-squared:  0.9095, Adjusted R-squared:  0.9035
## F-statistic: 150.7 on 12 and 180 DF,  p-value: < 2.2e-16
```

Neither of the terms for ppgdp are showing up as statistically significant according to individual t tests. Again, we really need an F test:

```
lm_fit2a <- lm(infantMortality ~ region +
                 fertility +
                 lifeExpF +
                 poly(pctUrban, 2, raw = TRUE),
               data = countries)
anova(lm_fit2a, lm_fit2)
```

```
## Analysis of Variance Table
##
## Model 1: infantMortality ~ region + fertility + lifeExpF + poly(pctUrban,
##     2, raw = TRUE)
## Model 2: infantMortality ~ region + fertility + poly(ppgdp, 2, raw = TRUE) +
##     lifeExpF + poly(pctUrban, 2, raw = TRUE)
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1    182 14967
## 2    180 14861  2    106.38 0.6442 0.5263
```

- The F test does not indicate strong evidence of a relationship between `ppgdp` and `infantMortality` rates, after accounting for the effects of `region`, `fertility`, `lifeExpF`, and `pctUrban`.
- But note we can't trust this F test! The conditions are not satisfied and F tests are particularly sensitive to the conditions.

5. Are your hypothesis test results consistent between the models you developed in parts 3 and 4? Which model would you prefer if you had to conduct some hypothesis tests about which variables have a strong association with infant mortality rates? Why?

No. My model with transformations provided strong evidence of a relationship between ppgdp and infant mortality rates after accounting for the other covariates, but the model without transformations did not. The hypothesis tests from the model without transformations cannot be trusted because the conditions for inference were not satisfied.