

Chapters 3 and 4 Homework Answers

John Fox

Soc. 740, Winter 2012

Exercise D3.1

I restricted the data set to those countries with valid data for all five variables:

```
> UN <- read.table("http://socserv.socsci.mcmaster.ca/jfox/Books/
  Applied-Regression-2E/datasets/UnitedNations.txt", header=TRUE)
> dim(UN)
[1] 207 13
> UN <- na.omit(UN[,c("tfr", "GDPperCapita", "illiteracyFemale", "contraception", "region")])
> dim(UN)
[1] 118 5
```

Note: I broke the the `read.table` command across two lines to fit in this document, not in the actual command that was executed.

Only 118 of the 207 countries remain, suggesting that analyzing the complete cases is a problematic strategy. (Nevertheless, I proceed.)

Nonparametric density estimates (with rug plot), normal quantile-comparison plots, and boxplots appear in the graphs below.

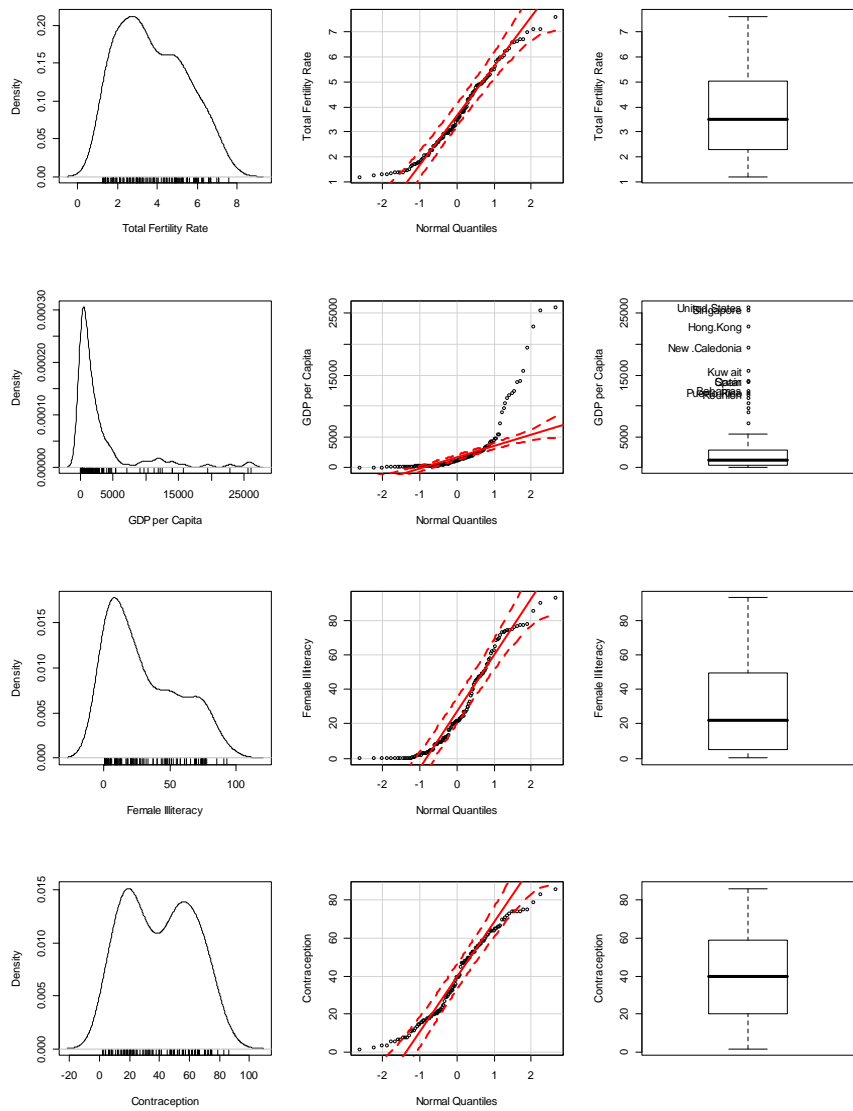
- The distribution of the total fertility rate is slightly positively skewed, and possibly bimodal, with modes near 3 and 5.
- The distribution of GDP per capita is extremely positively skewed, with many outliers in the direction of the skew.
- The distribution of female illiteracy is moderately positively skewed, with two or possibly three modes.
- The distribution of contraception is reasonably symmetric with two modes, near 20 and 60; the tails of the distribution are shorter than for the normal distribution.

```
> windows(width=9, height=12)
> par(mfrow=c(4, 3))
> library(car)
> with(UN, {
+   plot(density(tfr), xlab="Total Fertility Rate", main="")
+   rug(tfr)
+   qqPlot(tfr, ylab="Total Fertility Rate", xlab="Normal Quantiles",
+     labels=rownames(UN))
+   Boxplot(tfr, ylab="Total Fertility Rate", labels=rownames(UN))
+   plot(density(GDPperCapita), xlab="GDP per Capita", main="")
+   rug(GDPperCapita)
+   qqPlot(GDPperCapita, ylab="GDP per Capita", xlab="Normal Quantiles",
+     labels=rownames(UN))
+   Boxplot(GDPperCapita, ylab="GDP per Capita", labels=rownames(UN))
+   plot(density(illiteracyFemale), xlab="Female Illiteracy", main="")
+   rug(illiteracyFemale)
```

```

+ qqPlot(illiteracyFemale, ylab="Female Illiteracy", xlab="Normal Quantiles",
+       labels=rownames(UN))
+ Boxplot(illiteracyFemale, ylab="Female Illiteracy",
+       labels=rownames(UN))
+ plot(density(contraception), xlab="Contraception", main="")
+ rug(contraception)
+ qqPlot(contraception, ylab="Contraception", xlab="Normal Quantiles",
+       labels=rownames(UN))
+ Boxplot(contraception, ylab="Contraception", labels=rownames(UN))
+ })

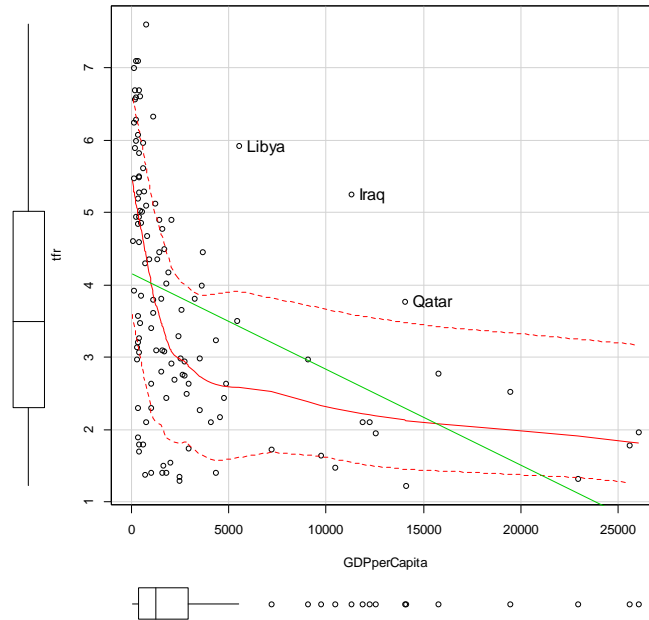
```



Exercise D3.3

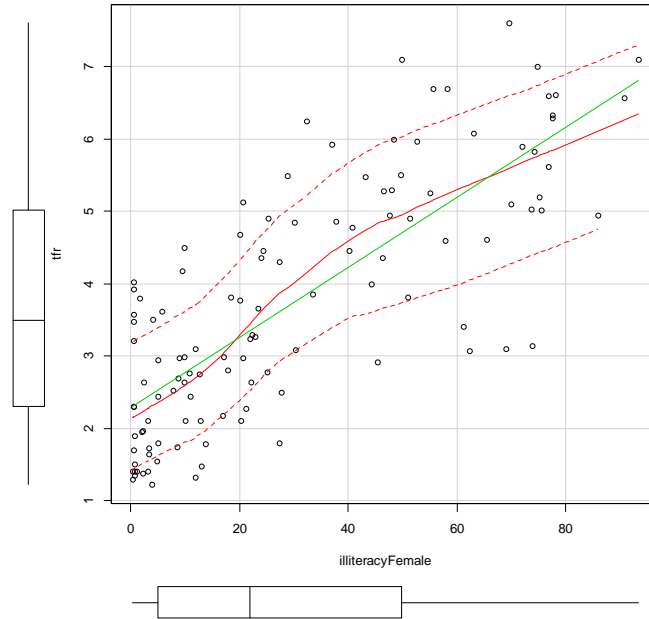
The three scatterplots, of the total fertility rate versus each of GDP per capita, female illiteracy, and contraception, are similar to the graphs created for Exercise D2.3, except now I've eliminated all countries with missing data on any of the variables.

```
> scatterplot(tfr ~ GDPperCapita, span=0.7, data=UN, id.method="identify")  
[1] "Iraq" "Libya" "Qatar"
```



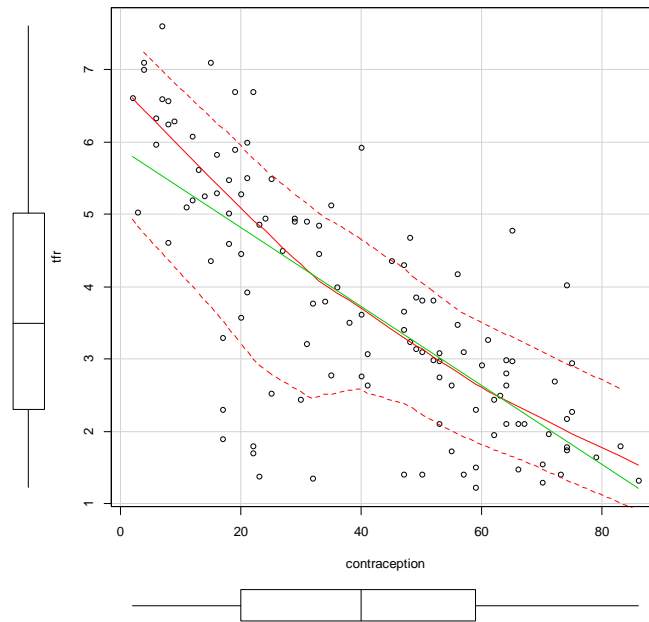
The relationship between the total fertility rate and GDP per capita appears to be negative and monotone but highly nonlinear, with some outliers above the fitted lowess curve. Notice that I opted here to identify points with the mouse.

```
> scatterplot(tfr ~ illiteracyFemale, span=0.6, data=UN)
```



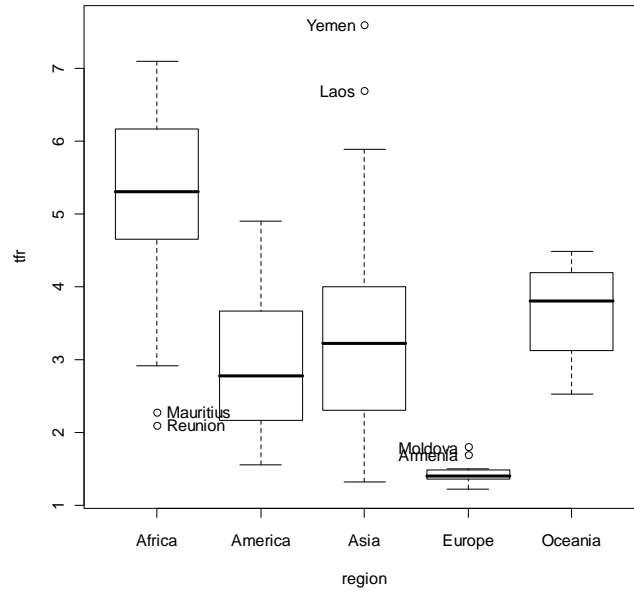
The relationship between the total fertility rate and female illiteracy is positive, monotone, and only slightly nonlinear; there are no obvious outliers.

```
> scatterplot(tfr ~ contraception, span=0.5, data=UN)
```



The relationship between the total fertility rate and contraception is negative, monotone, and nearly linear; there are no obvious outliers.

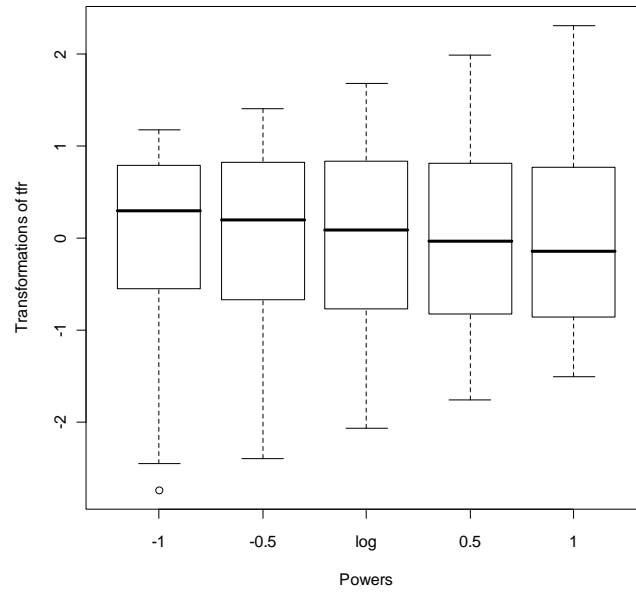
```
> Boxplot(tfr ~ region, data=UN)
[1] "Mauritius" "Reunion" "Laos" "Yemen" "Armenia" "Moldova"
```



Fertility is highest on average in Africa, followed by Oceania, Asia, the Americas, and Europe, which has by far the lowest fertility. The spread in Europe is also much lower than in the other regions. (It would be reasonable, having viewed the graph, to reorder the regions by their medians, but I didn't do that.)

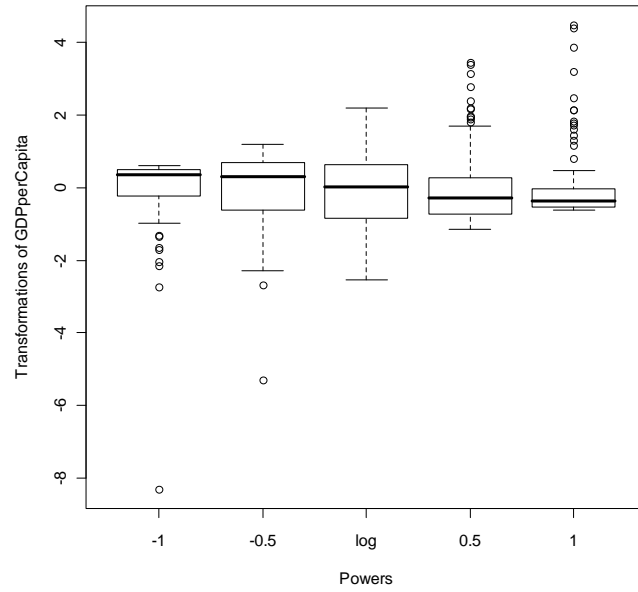
Exercise D4.1

```
> symbox(~ tfr, data=UN)
```



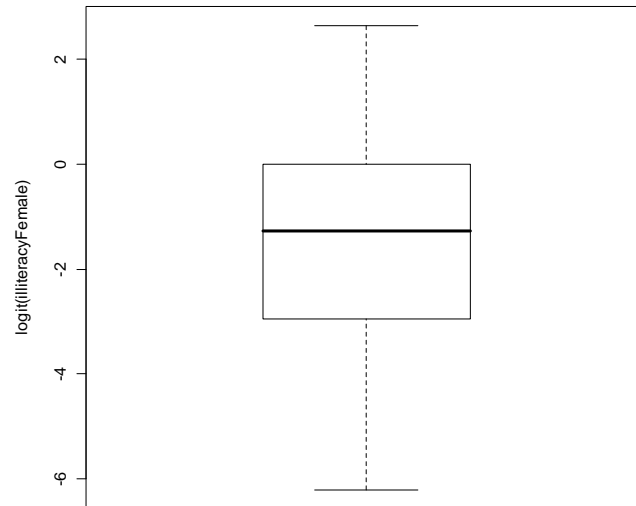
The log or (somewhat better) the square-root transformation makes the distribution of TFR more symmetric.

```
> symbox(~ GDPperCapita, data=UN)
```



Again, the log transformation works well for GDP per capita.

```
> range(UN$illiteracyFemale)
[1] 0.2 93.4
> Boxplot(~ logit(illiteracyFemale), data=UN)
```

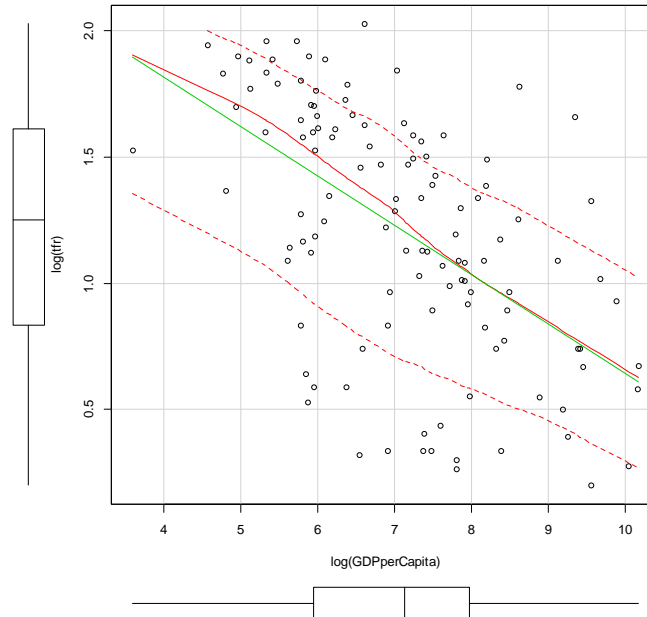


I applied the logit transformation to female illiteracy (which is a percentage that ranges from nearly 0 to about 93).

Contraception is roughly symmetrically distributed, so I left it alone.

Exercise D4.2

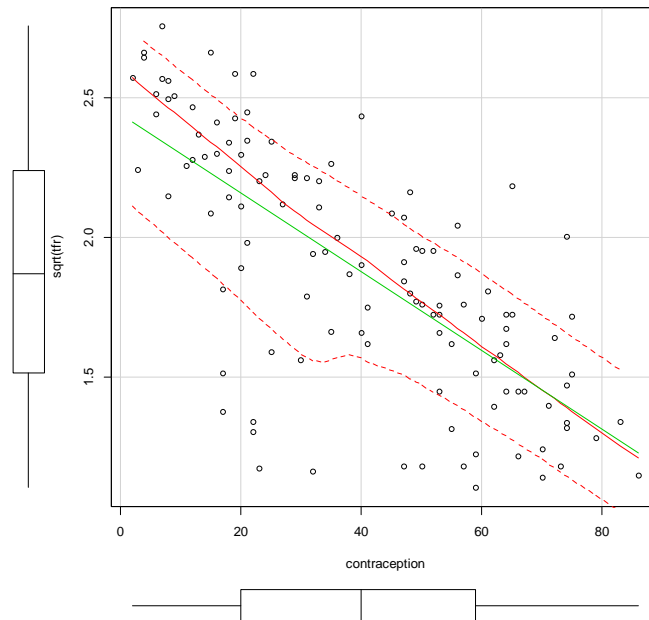
```
> scatterplot(log(tfr) ~ log(GDPperCapita), span=0.7, data=UN)
```



The bulge in the original scatterplot points down and to the left. The log transformation of TFR and GDP per capita, which makes the distributions of both variables more symmetric, also makes their relationship nearly linear.

The relationship between TFR and female illiteracy is nearly linear, and the nonlinear pattern, though monotone, is not simple. I decided to leave the variables alone.

```
> scatterplot(sqrt(tfr) ~ contraception, span=0.6, data=UN)
```



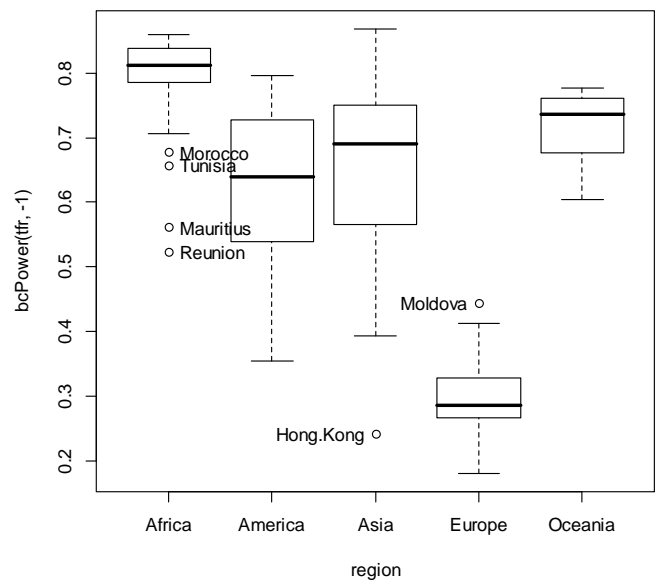
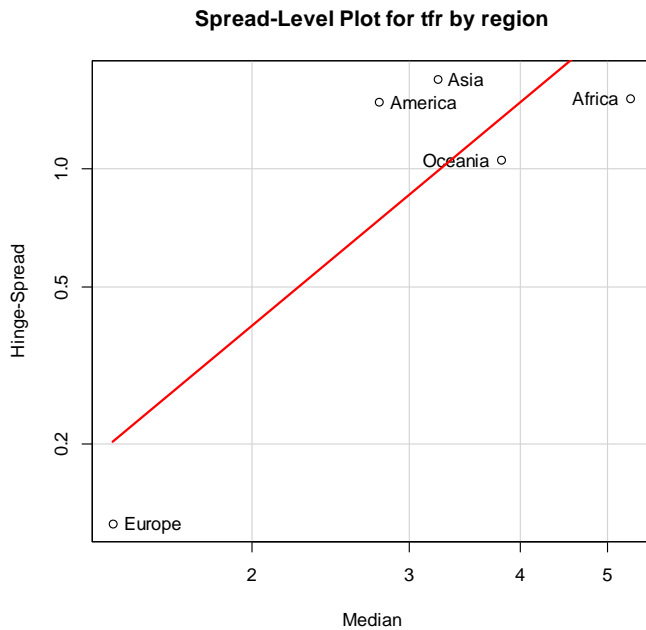
Because TFR is skewed and contraception is not, and because the bulge points down and to the left, I tried transforming TFR down the ladder of powers and roots: A square-root transformation works well. The original relationship is close enough to linear that it would also be reasonable simply to leave the variables alone.

Exercise D4.3

```
> windows(width=11, height=5)
> par(mfrow=c(1,2))
> spreadLevelPlot(tfr ~region, data=UN)
      LowerHinge Median UpperHinge Hinge-Spread
Europe      1.365  1.400      1.49      0.125
America     2.170  2.775      3.66      1.490
Asia        2.300  3.225      4.00      1.700
Oceania     3.130  3.800      4.19      1.060
Africa      4.645  5.300      6.16      1.515
```

Suggested power transformation: -0.9013382

```
> Boxplot(bcPower(tfr, -1) ~ region, data=UN)
[1] "Mauritius" "Morocco" "Reunion" "Tunisia" "Hong.Kong" "Moldova"
```



Tukey's rule suggests the inverse transformation of TFR, but the plot makes clear that this is primarily due to Europe, which has a much lower level and spread than the other regions. Performing this transformation is not especially helpful here, although it does serve to spread out the values for Europe.