

Stat 505 Assignment 5 Solutions

The Manchester Guardian's DataBlog website includes data on cancer rates for 50 countries (drawn, they say from the 2008 World Cancer Research Fund, where it's credited to GloboCan <http://globocan.iarc.fr/>). They also provide some tips on how to avoid cancer, like: control your weight, limit alcohol, limiting sweets and red meat.

To try to explain the cancer rate outcome, I downloaded data from the GapMinder site. First I had to clean up country names in the cancer rates file to match those of gapminder (See appendix).

I used the following variables:

smoking the proportion of adults who smoke. I expect higher smoking rates to coincide with more cancer.

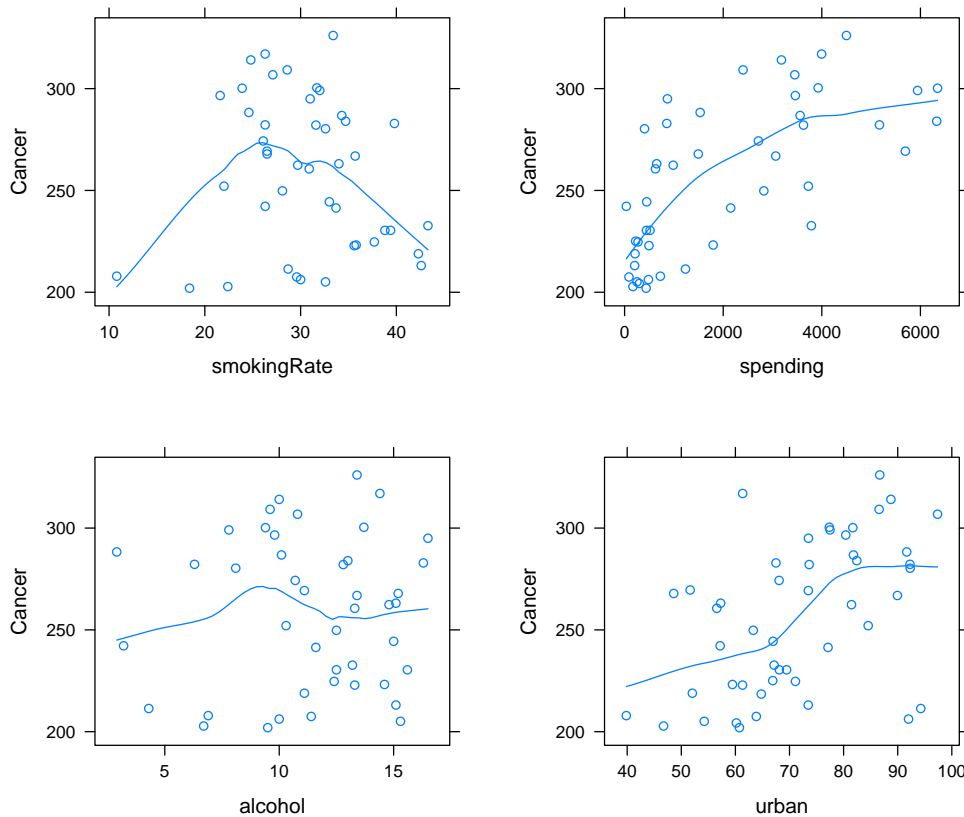
alcohol consumption. Again, I expect a positive association.

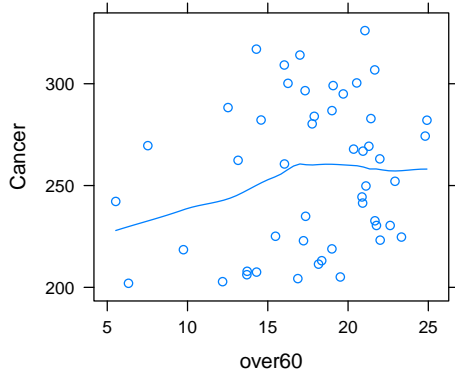
spending cost of the health care system per capita.

over60 proportion of population over age 60. I expect a positive association.

urban proportion of urban dwellers – again a positive association is expected.

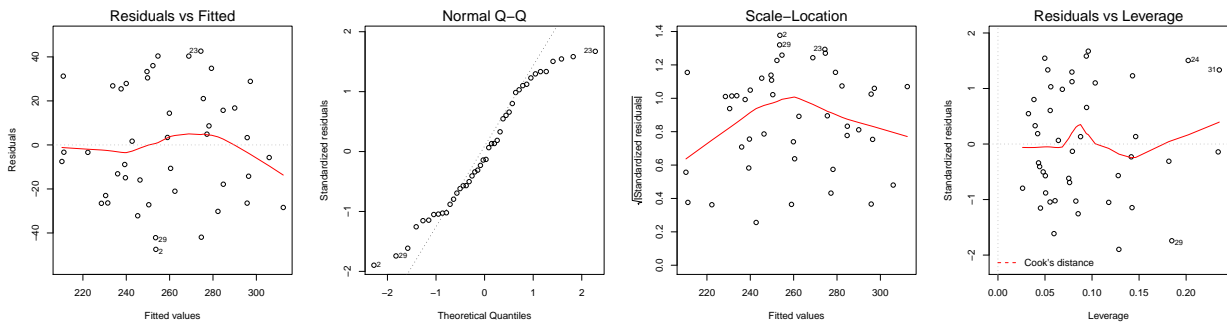
I made scatterplots of each to look for trends.





From the plots, I see that the smoking rate data does not associate with cancer rate well at all, spending might have a quadratic association, and the other three variables have weak positive associations with cancer rate.

Fitting a linear model to all predictors shows that smokingRate and over60 have negative signs (given the other terms are in the model) and t-ratios close to 0 (-.29, -1.3). I dropped them (one at a time) to obtain a model with just 3 predictors which explains 51% of the variation in cancer rates in these 44 countries. I plotted R's default diagnostic plots for this reduced model.



I see no problems with the residual plots in that the first and third plots show no fan, the comparison to normality plot shows short tails (not a problem), and there are no high leverage countries in the last plot.

The coefficient on spending indicates an increase of .01 (SE = .002) in cancer rate for each dollar per person spent on health care. This is a result of better diagnosis and longer life expectancy. As populations get more urban, cancer rates increase at a rate of .76 (SE = .33) for each increase of 1 percent more urban dwellers. This could be as a result of greater exposure to carcinogens in cities, or just an artifact of these data being from more industrialised countries.

Conclusions: This is not a very representative sample of countries in the world in that Asia and Southern Hemisphere countries are under represented. As the original article indicated, countries with higher spending on health care have higher cancer rates for two reasons: better detection, and longer life expectancy. These data are at the country level, and provide little or no advice for an individual to avoid cancer. We're all going to die, so if we avoid sickness for many years, the chances of dying of cancer increase.

R code

```
> cancer <- read.csv("CancerRates.csv", head = T)
> cancer <- subset(cancer, !is.na(Overall))[, 1:2]
> names(cancer)[2] <- "Cancer"
> newDF <- read.csv("life_expectancy_at_birth-gapminder.csv")
> names(newDF)[1] <- "country"
> subset(cancer, !(country %in% newDF$country))
      country Cancer
7          USA 300.2
12 The Netherlands 286.8
24 Republic of Korea 262.4
25      Slovakia 260.6
29 Chinese Taipei 244.1
36   FYR Macedonia 225.1
50 South African Republic 202.0
> levels(cancer$country)[which(!(levels(cancer$country) %in% levels(newDF$country)))] <- c("Taiwan",
+       "Macedonia, FYR", "Reunion", "Korea, Rep.", "Slovak Republic",
+       "South Africa", "Netherlands", "United States")
> cancer$country <- factor(cancer$country)
> cancer2 <- merge(cancer, newDF, all.x = T)
> newDF <- read.csv("Smoking-gapminder.csv")
> names(newDF)[1] <- "country"
> levels(cancer2$country)[13] <- levels(newDF$country)[47]
> cancer2 <- merge(cancer2, newDF, all.x = T)
> require(lattice)
> print(xyplot(Cancer ~ smokingRate, cancer2, type = c("p", "smooth")))
> newDF <- read.csv("healthspendingperpersonUSgapminder.csv")
> names(newDF) <- c("country", "spending")
> cancer2 <- merge(cancer2, newDF, all.x = T)
> print(xyplot(Cancer ~ spending, cancer2, type = c("p", "smooth")))
> newDF <- read.csv("carbon_dioxide_total_emissions-gapminder.csv")
> cancer2 <- merge(cancer2, newDF, all.x = T)
> newDF <- read.csv("totalpopulation-gapminder.csv")
> cancer2 <- merge(cancer2, newDF, all.x = T)
> newDF <- read.csv("alcoholconsumption-Gapminder.csv")
> cancer2 <- merge(cancer2, newDF, all.x = T)
> print(xyplot(Cancer ~ alcohol, cancer2, type = c("p", "smooth")))
> newDF <- read.csv("urbanpopulation-gapminder.csv")
> cancer2 <- merge(cancer2, newDF, all.x = T)
> print(xyplot(Cancer ~ urban, cancer2, type = c("p", "smooth")))
> newDF <- read.csv("outOfPocketHealthSpending-WDI.csv")
> cancer2 <- merge(cancer2, newDF, all.x = T)
> newDF <- read.csv("GDPpercapita-gapminder.csv", na.string = "..")
> cancer2 <- merge(cancer2, newDF, all.x = T)
> newDF <- read.csv("totalabove60-gapminder.csv", na.string = "..")
> newDF <- merge(newDF, read.csv("totalpopulation-gapminder.csv"))
> newDF$over60 <- newDF$over60/newDF$population * 100
> cancer2 <- merge(cancer2, newDF[, 1:2], all.x = T)
> write.csv(cancer2, file = "cancer2.csv", row.names = FALSE)
> print(xyplot(Cancer ~ over60, cancer2, type = c("p", "smooth")))
> summary(lm(Cancer ~ over60 + spending + urban + alcohol + smokingRate,
+ cancer2))
Call:
```

```
lm(formula = Cancer ~ over60 + spending + urban + alcohol + smokingRate,
    data = cancer2)
```

Residuals:

Min	1Q	Median	3Q	Max
-51.710	-19.143	-3.325	22.964	38.128

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	173.469324	30.224642	5.739	1.3e-06
over60	-0.352725	1.280611	-0.275	0.784474
spending	0.009782	0.002563	3.817	0.000484
urban	0.917766	0.342787	2.677	0.010896
alcohol	3.152003	1.653069	1.907	0.064134
smokingRate	-1.047224	0.808362	-1.295	0.202966

Residual standard error: 26.73 on 38 degrees of freedom

(6 observations deleted due to missingness)

Multiple R-squared: 0.5374, Adjusted R-squared: 0.4765

F-statistic: 8.828 on 5 and 38 DF, p-value: 1.254e-05

```
> summary(lm(Cancer ~ over60 + spending + urban + alcohol, cancer2))
```

Call:

```
lm(formula = Cancer ~ over60 + spending + urban + alcohol, data = cancer2)
```

Residuals:

Min	1Q	Median	3Q	Max
-50.678	-20.563	-2.766	25.493	38.885

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	166.045707	29.933411	5.547	2.20e-06
over60	-0.871262	1.226984	-0.710	0.482
spending	0.010831	0.002452	4.417	7.72e-05
urban	0.804348	0.334288	2.406	0.021
alcohol	2.352363	1.546784	1.521	0.136

Residual standard error: 26.96 on 39 degrees of freedom

(6 observations deleted due to missingness)

Multiple R-squared: 0.5169, Adjusted R-squared: 0.4674

F-statistic: 10.43 on 4 and 39 DF, p-value: 7.628e-06

```
> summary(lm(Cancer ~ spending + urban + alcohol, cancer2))
```

Call:

```
lm(formula = Cancer ~ spending + urban + alcohol, data = cancer2)
```

Residuals:

Min	1Q	Median	3Q	Max
-47.448	-21.519	-3.347	25.819	42.601

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	161.64983	29.10420	5.554	2.00e-06
spending	0.01043	0.00237	4.399	7.85e-05
urban	0.76083	0.32658	2.330	0.025
alcohol	1.69560	1.23209	1.376	0.176

```
Residual standard error: 26.79 on 40 degrees of freedom
(6 observations deleted due to missingness)
Multiple R-squared: 0.5107,      Adjusted R-squared: 0.474
F-statistic: 13.92 on 3 and 40 DF,  p-value: 2.323e-06
> par(mfrow = c(1, 4))
> plot(lm(Cancer ~ spending + urban + alcohol, cancer2))
```