Robust regression - Solutions

Numbers in parentheses indicate the number of points for each question.

Total: 12 points.

Data

This exercise is based on the *gapminder* dataset from the package of the same name.

> Jennifer Bryan (2017). gapminder: Data from Gapminder. R package version 0.3.0. https://CRAN.R-project.org/package=gapminder

This dataset includes the life expectancy (*lifeExp*), population (*pop*) and GDP per capita (*gdpPercap*) for 142 countries and 12 years (every 5 years between 1952 and 2007).

```
library(gapminder)
str(gapminder)
```

```
## tibble [1,704 x 6] (S3: tbl_df/tbl/data.frame)
##  $ country  : Factor w/ 142 levels "Afghanistan",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ continent: Factor w/ 5 levels "Africa","Americas",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ year     : int [1:1704] 1952 1957 1962 1967 1972 1977 1982 1987 1992 1997 ...
##  $ lifeExp  : num [1:1704] 28.8 30.3 32 34 36.1 ...
##  $ pop      : int [1:1704] 8425333 9240934 10267083 11537966 13079460 14880372 12881816 13867957 163
##  $ gdpPercap: num [1:1704] 779 821 853 836 740 ...
```

1. Effect of GDP and time on life expectancy

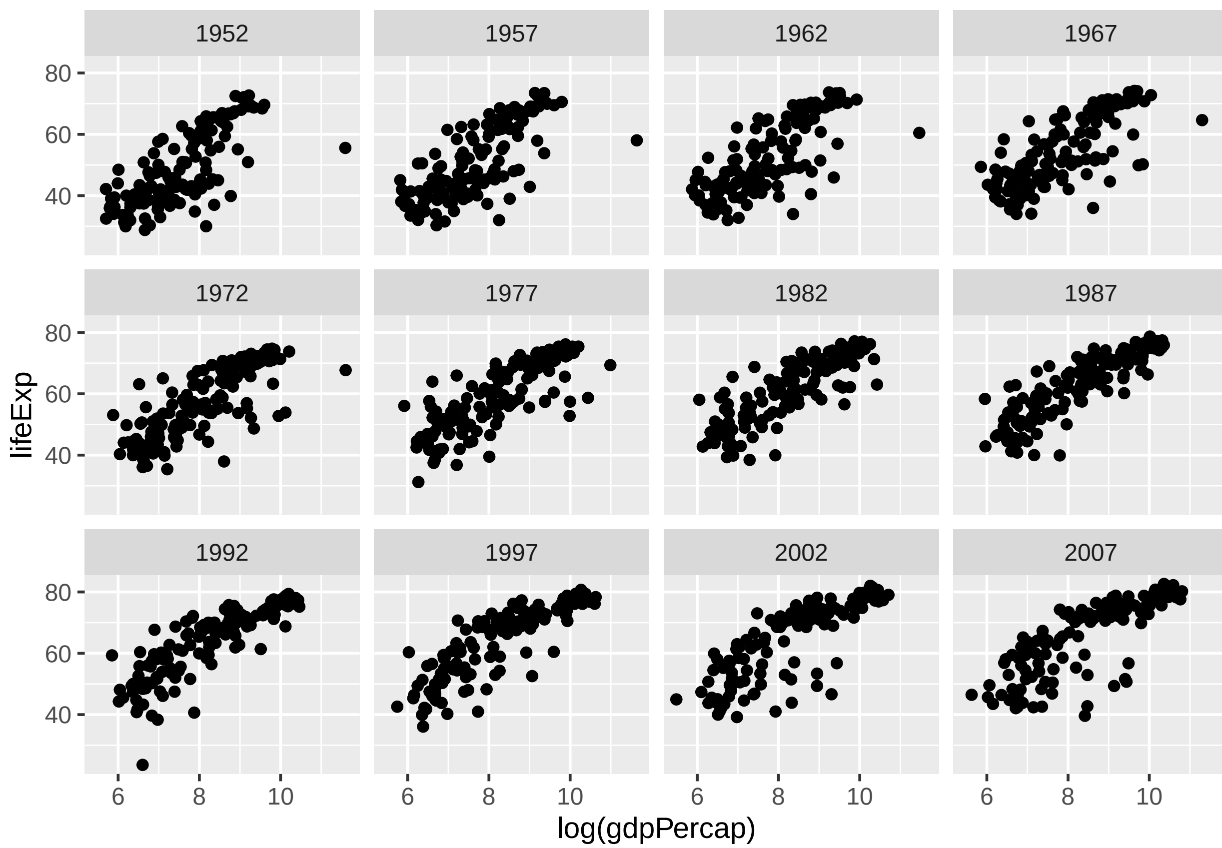
a) First, visualize the life expectancy as a function of GDP per capita and year. It is suggested to represent the logarithm of *gdpPercap* and to separate the different years, for example with facets in *ggplot2*: `... + facet_wrap(~year)`.

What general trends do you observe? Are there extreme values that could strongly influence a regression model? If so, try to identify these data in the table based on the position of the points in the graph. **(2)**

Solution

```
library(ggplot2)

ggplot(gapminder, aes(x = log(gdpPercap), y = lifeExp)) +
    geom_point() +
    facet_wrap(~year)
```



- Life expectancy increases with *log(gdpPercap)* for each year.
- For the early years, one of the values with a very high GDP deviates from the trend, so this extreme value could strongly influence a regression. If we look for the extreme values of *log(gdpPercap)*, we see that it is Kuwait between 1952 and 1972.

```
library(dplyr)

filter(gapminder, log(gdpPercap) > 11)
```

```
## # A tibble: 5 x 6
##   country continent  year lifeExp    pop gdpPercap
##   <fct>   <fct>     <int>   <dbl>  <int>     <dbl>
## 1 Kuwait  Asia       1952    55.6 160000   108382.
## 2 Kuwait  Asia       1957    58.0 212846   113523.
## 3 Kuwait  Asia       1962    60.5 358266    95458.
## 4 Kuwait  Asia       1967    64.6 575003    80895.
## 5 Kuwait  Asia       1972    67.7 841934   109348.
```

b) Perform a linear regression (`lm`) to determine the effect of GDP per capita, year and their interaction on life expectancy. To help interpret the coefficients, perform the following transformations on the predictors:

- Take the logarithm of *gdpPercap* and standardize it with the function `scale`. *Reminder*: `scale(x)` subtracts each value of `x` from its mean and divides by its standard deviation, so the resulting variable has a mean of 0 and a standard deviation of 1; it represents the number of standard deviations above or below the mean.
- Replace *year* with the number of years since 1952.

Interpret the meaning of each of the coefficients in the model and then refer to the diagnostic graphs. Are the assumptions of the linear model met? **(2)**

Solution

```
gapminder <- mutate(gapminder, gdp_norm = scale(log(gdpPercap)), dyear = year - 1952)

mod_lm <- lm(lifeExp ~ gdp_norm * dyear, gapminder)
summary(mod_lm)
```

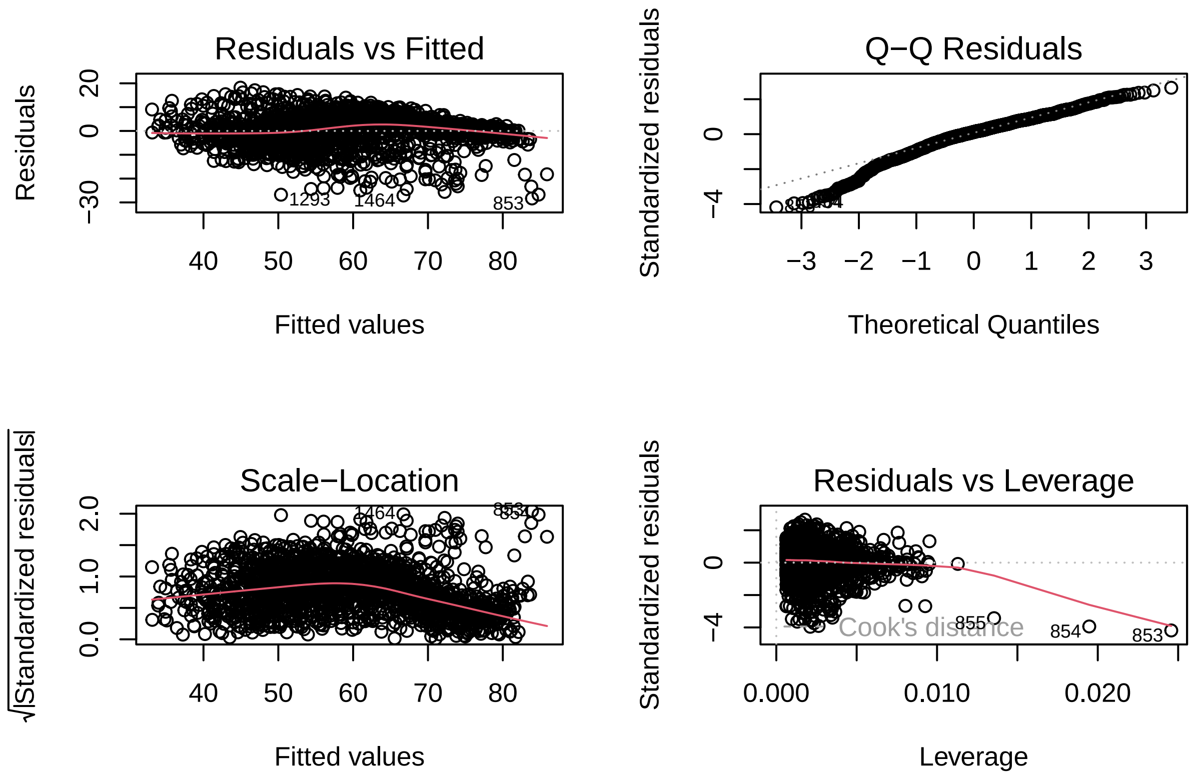
```
##
## Call:
## lm(formula = lifeExp ~ gdp_norm * dyear, data = gapminder)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.340  -3.496   0.802   4.557  18.172
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    54.311695   0.324413 167.415  < 2e-16 ***
## gdp_norm       10.694128   0.340945  31.366  < 2e-16 ***
## dyear           0.192828   0.009923  19.433  < 2e-16 ***
## gdp_norm:dyear -0.034776   0.009774  -3.558 0.000384 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.854 on 1700 degrees of freedom
## Multiple R-squared:  0.719,  Adjusted R-squared:  0.7185
## F-statistic:  1450 on 3 and 1700 DF,  p-value: < 2.2e-16
```

Interpretation of the coefficients:

- (Intercept): `When`gdp_norm`and`dyear‘ are equal to zero, therefore for a country with the mean of *log(gdpPercap)* in 1952, life expectancy is 54.3 years.
- gdp_norm: When `dyear` = 0 (in 1952), each standard deviation above the mean of *log(gdpPercap)* increases life expectancy by 10.7 years.
- dyear: For a country at the mean *log(gdpPercap)*, life expectancy increases by 0.19 years per year.
- gdp_norm:dyear: For each year, the effect of a unit increase in `gdp_norm` on life expectancy decreases by 0.034; *OR* for each standard deviation above the mean of *log(gdpPercap)*, the effect of time on life expectancy decreases by 0.034.

In other words, both predictors have a positive effect on the response, but their interaction is negative, so when one increases, the effect of the other becomes less important.

```
par(mfrow = c(2,2))
plot(mod_lm)
```



On the graph of residuals vs. expected values, we see that the variance decreases when the predicted value is higher (non-homogeneous variance). Also, on the leverage graph, we can see the extreme points, even if Cook's distance does not exceed the threshold of 0.5 or 1 (with many data points, it takes very extreme values to obtain such a large Cook's distance).

c) Compare the result of the model in (b) with two more robust alternatives: robust regression based on Tukey's biweight (function `lmrob` from the *robustbase* package) and median regression (function `rq` from the *quantreg* package, choosing only the median quantile). Explain how the estimates and standard errors of the coefficients differ between the three methods. **(2)**

Note: Use the `showAlgo = FALSE` option when applying the `summary` function to the output of `lmrob`, to simplify the summary.

Solution

```r
library(robustbase)
mod_lmrob <- lmrob(lifeExp ~ gdp_norm * dyear, gapminder)
print(summary(mod_lmrob), showAlgo = FALSE)
```

```
##
## Call:
## lmrob(formula = lifeExp ~ gdp_norm * dyear, data = gapminder)
##  \--> method = "MM"
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.838  -3.912   0.133   3.658  18.625
##
## Coefficients:
```

```
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    55.562627   0.352056 157.823   <2e-16 ***
## gdp_norm       12.590304   0.311367  40.436   <2e-16 ***
## dyear           0.178728   0.010456  17.094   <2e-16 ***
## gdp_norm:dyear -0.075843   0.008496  -8.927   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Robust residual standard error: 5.467
## Multiple R-squared:  0.797,  Adjusted R-squared:  0.7967
## Convergence in 9 IRWLS iterations
##
## Robustness weights:
##  8 observations c(37,40,167,853,854,855,1293,1464)
##   are outliers with |weight| = 0 ( < 5.9e-05);
##  147 weights are ~= 1. The remaining 1549 ones are summarized as
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0004047 0.8444000 0.9451000 0.8735000 0.9859000 0.9990000
```

```r
library(quantreg)
mod_rq <- rq(lifeExp ~ gdp_norm * dyear, tau = 0.5, data = gapminder)
summary(mod_rq)
```

```
##
## Call: rq(formula = lifeExp ~ gdp_norm * dyear, tau = 0.5, data = gapminder)
##
## tau: [1] 0.5
##
## Coefficients:
##                Value     Std. Error t value   Pr(>|t|)
## (Intercept)     55.57793   0.37879  146.72594   0.00000
## gdp_norm        12.51016   0.36768   34.02426   0.00000
## dyear            0.18736   0.01142   16.41102   0.00000
## gdp_norm:dyear  -0.07982   0.00869   -9.18818   0.00000
```

Comparison of results:

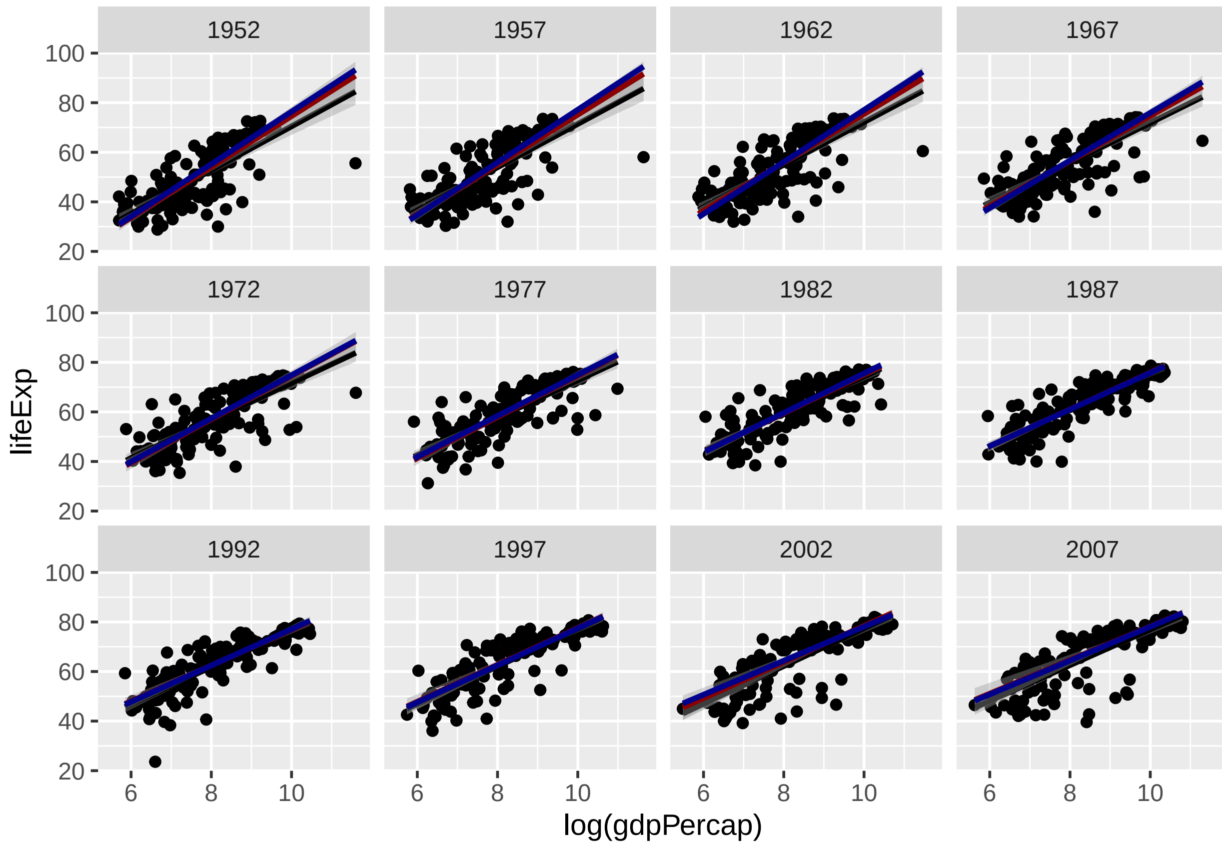
- The coefficients obtained by robust regression and median regression are similar, but robust regression has lower standard errors, confirming the idea that this method is more efficient than median regression while being almost as robust to extreme values.
- For both methods, the effect of `gdp_norm` is greater and the interaction is more negative in contrast to the result obtained with `lm`. Recall that Kuwait had a very high GDP for the early years without having such a high life expectancy. In this case, this extreme value led to underestimating the general trend of the GDP effect and therefore also underestimating how this trend changed over time (interaction).

(d) Superimpose the regression lines of the three models on the graph in (a). With `ggplot` you can use the `geom_smooth` function with `method = "lm"` for linear regression and `method = "lmrob"` for robust regression. For median regression you can use `geom_quantile` as seen in the notes. **(1)**

Solution

```r
ggplot(gapminder, aes(x = log(gdpPercap), y = lifeExp)) +
    geom_point() +
    geom_smooth(method = "lm", color = "black") +
    geom_smooth(method = "lmrob", color = "darkred") +
    geom_quantile(quantiles = 0.5, color = "darkblue", size = 1) +
    facet_wrap(~year)
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



2. Variation of effects by quantile

a) Based on your observation of the data in 1(a), would it be useful to model different quantiles of life expectancy based on the predictors? Justify your answer. **(1)**

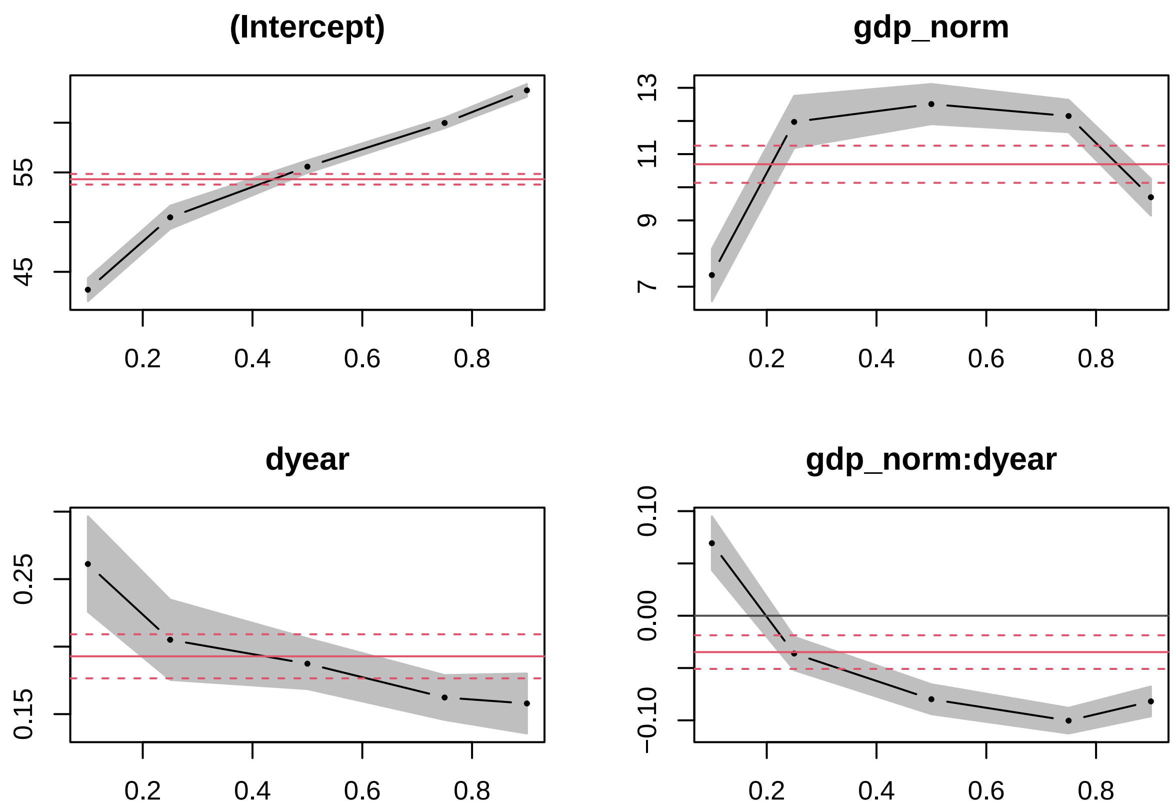
Solution

Yes, because the variance is not homogeneous, it seems to be lower when GDP is high, so this predictor will have a different effect on the different quantiles of life expectancy.

b) Perform a quantile regression with the same predictors as in 1(b), with the following quantiles: `(0.1, 0.25, 0.5, 0.75, 0.9)`. Use the `plot` function on the quantile regression summary and describe how the effect of the predictors varies between quantiles. **(2)**

Solution

```
mod_quant <- rq(lifeExp ~ gdp_norm * dyear, tau = c(0.1, 0.25, 0.5, 0.75, 0.9), data = gapminder)
plot(summary(mod_quant))
```

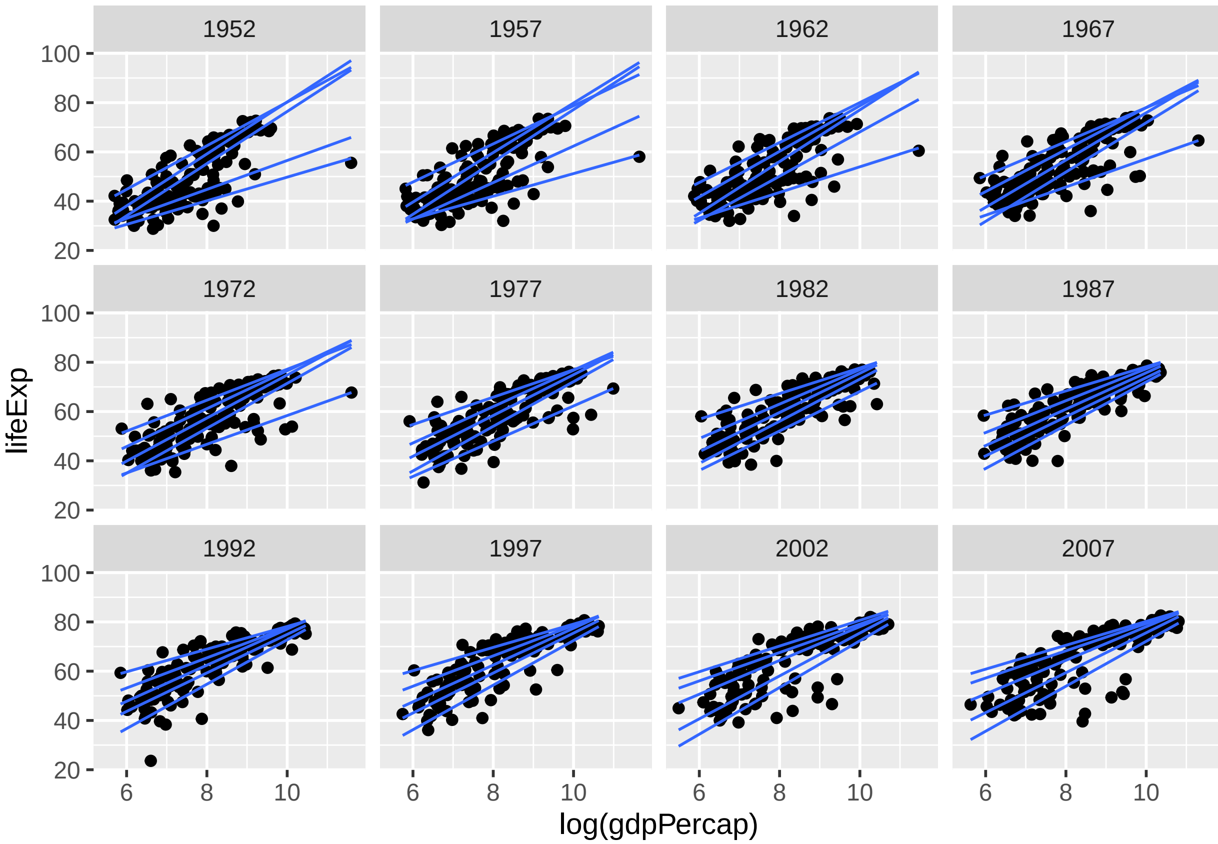


- The effect of GDP on life expectancy is greater for the central quantiles of life expectancy, the predictor has less effect on the 10% and 90% quantiles.
- The effect of the year is greater for the 10% quantile and decreases for the higher quantiles. Thus the increase in life expectancy since 1952 is due more to an increase in the lower values of the life expectancy distribution than to an increase in the higher values.
- The interaction is positive for the 10% quantile (the effect of GDP is greater for more recent years) and negative for the other quantiles (the effect of GDP becomes less important with time).

c) Superimpose the quantile regression lines on the graph of the data. Do the trends for each quantile appear to be affected by extreme values? **(2)**

Solution

```
ggplot(gapminder, aes(x = log(gdpPercap), y = lifeExp)) +
    geom_point() +
    geom_quantile(quantiles = c(0.1, 0.25, 0.5, 0.75, 0.9)) +
    facet_wrap(~year)
```



Yes, it seems that the extreme value between 1952 and 1972 influences the 10% quantile. In this case, the results of the previous part where the 10% quantile differs from the rest of the data (the GDP effect is less pronounced, the interaction is positive) may be the result of this extreme value.

Note on international comparisons

While this dataset is useful for illustrating the concepts of robust regression and quantile regression, it should be noted that this type of statistical analysis comparing variables measured at the national level has several limitations:

- It cannot be assumed that the associations detected apply at a smaller scale (e.g., the relationship between life expectancy and income when comparing national averages is not necessarily the same as the relationship between life expectancy and income at the level of individuals living in each country).
- Averages calculated in different countries are not independent observations, because environmental, social and economic conditions are correlated between nearby countries.
- There are many factors that differentiate countries, so it is difficult to interpret an association as a causal link.

Many articles, particularly in the social sciences, have been published on the methods to be used to make this type of *cross-country comparisons*.