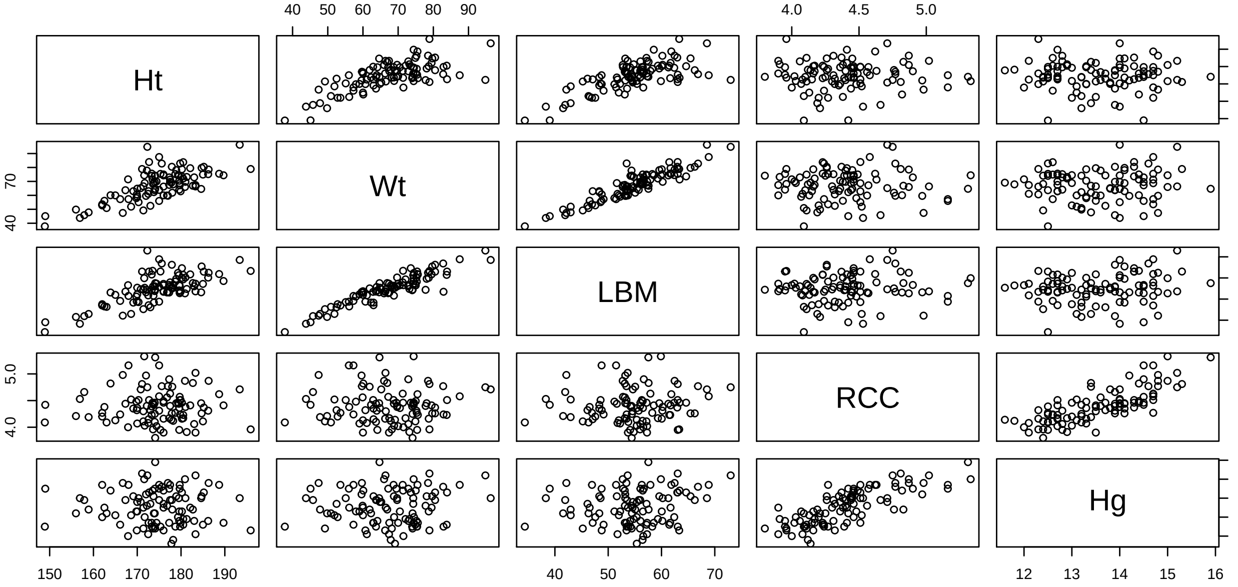
STAT2170 and STAT6180
Applied Statistics

Tutorial Week 9 Solution: Multiple Regression 2021

Question 1

a. Loading the data and checking the scatterplot,

```
aisfemales = read.csv("aisfemales.csv", header = TRUE)
plot(aisfemales)
```



LBM is highly correlated with Wt, less so but still highly correlated with Ht, with a high correlation between Ht and Wt. Only slight correlation between LBM and RCC and Hg. High correlation between Hg and RCC but not with other predictors

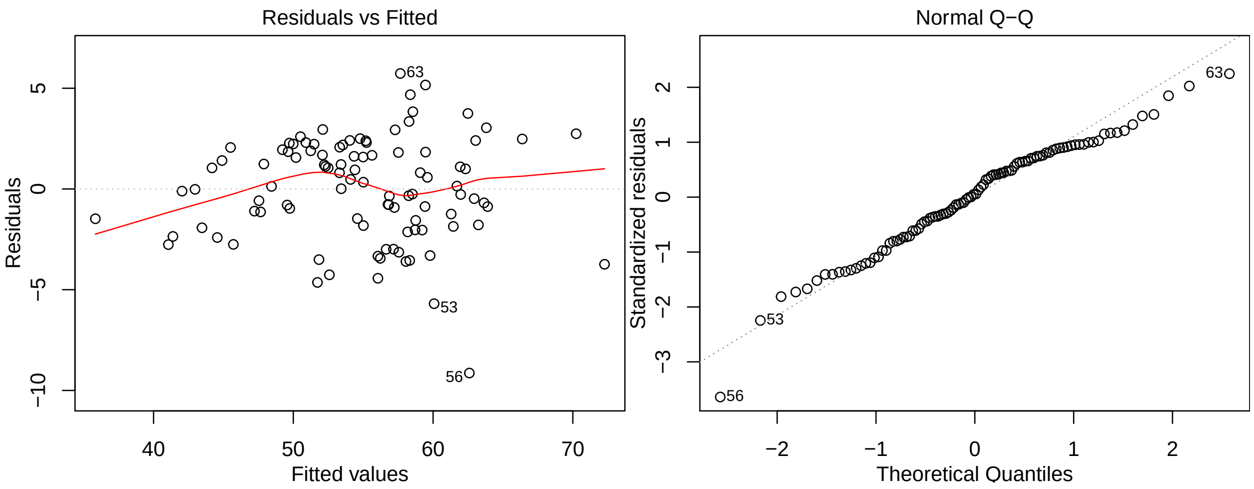
b. Fitting the regression using all covariates.

```
ais.1 = lm(LBM ~ Ht + Wt + RCC + Hg, data = aisfemales)
summary(ais.1)
```

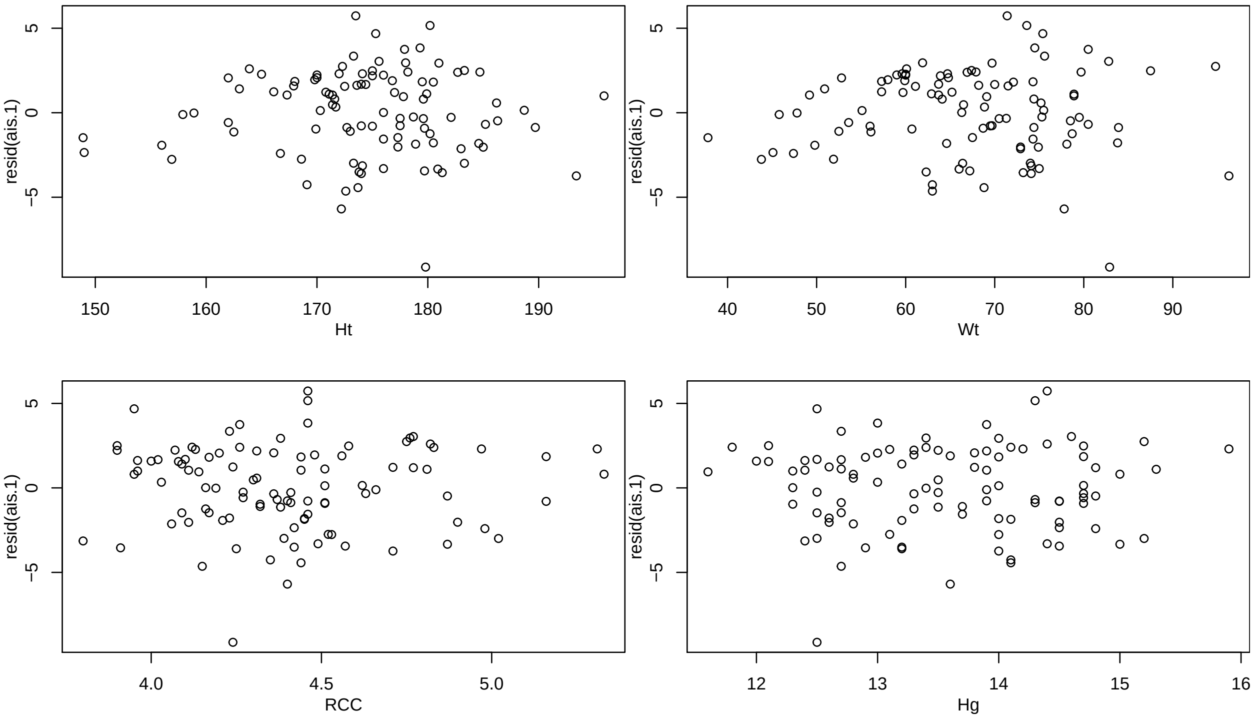
```
#
# Call:
# lm(formula = LBM ~ Ht + Wt + RCC + Hg, data = aisfemales)
#
# Residuals:
#     Min      1Q  Median      3Q     Max
# -9.1360 -1.7924  0.1358  1.9082  5.7336
#
# Coefficients:
```

```
#               Estimate Std. Error t value Pr(>|t|)
# (Intercept) -10.49080    7.84826  -1.337   0.1845
# Ht            0.11370    0.04570   2.488   0.0146 *
# Wt            0.51732    0.03458  14.958   <2e-16 ***
# RCC          -0.57172    1.30736  -0.437   0.6629
# Hg            0.97457    0.45926   2.122   0.0364 *
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 2.587 on 95 degrees of freedom
# Multiple R-squared:  0.866,   Adjusted R-squared:  0.8603
# F-statistic: 153.5 on 4 and 95 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(1, 2))
plot(ais.1, which = 1:2)
```



```
par(mfrow = c(2, 2))
plot(resid(ais.1) ~ Ht + Wt + RCC + Hg, data = aisfemales)
```



1. The Normal Q-Q plot of residuals has slight curvature but close to linear implying errors close to normally distributed.
2. The residuals vs fitted has not discernable pattern.
3. Residuals vs predictor plots no obvious pattern. So linear model seems adequate.

c. RCC is not significant, so remove it from the model. Fit again, and check assumptions. (Both versions of syntax are valid)

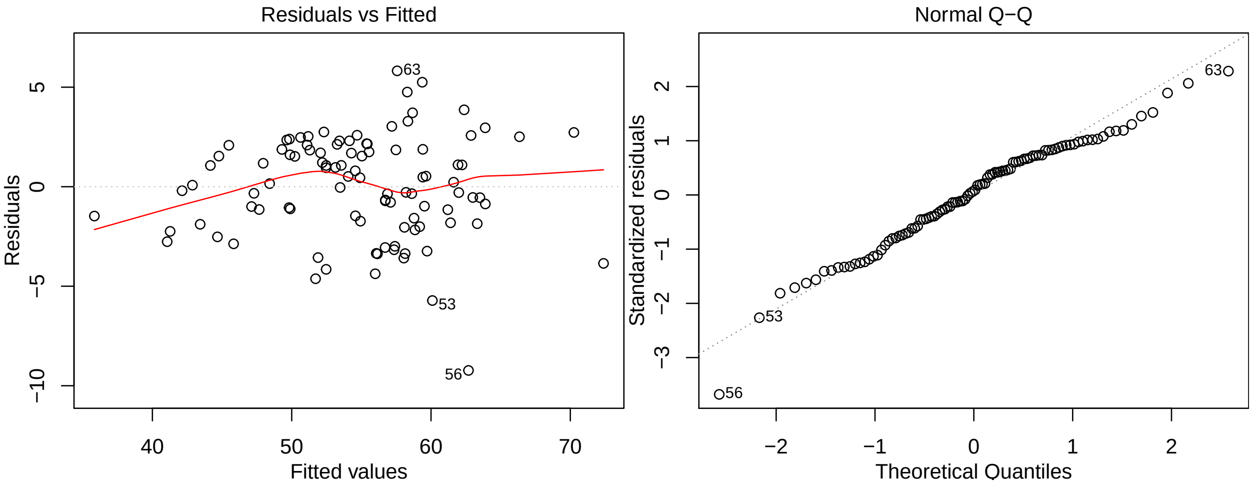
```
ais.2 = lm(LBM ~ Ht + Wt + Hg, data = aisfemales)
ais.2 = update(ais.1, . ~ . - RCC)
summary(ais.2)
```

```
#
# Call:
# lm(formula = LBM ~ Ht + Wt + Hg, data = aisfemales)
#
# Residuals:
#     Min      1Q  Median      3Q     Max
# -9.2310 -1.7536  0.1897  1.8743  5.8204
#
# Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
# (Intercept) -10.46800    7.81496  -1.339  0.18358
# Ht            0.11042    0.04489   2.460  0.01569 *
# Wt            0.51983    0.03396  15.307  < 2e-16 ***
# Hg            0.81696    0.28343   2.882  0.00487 **
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
```

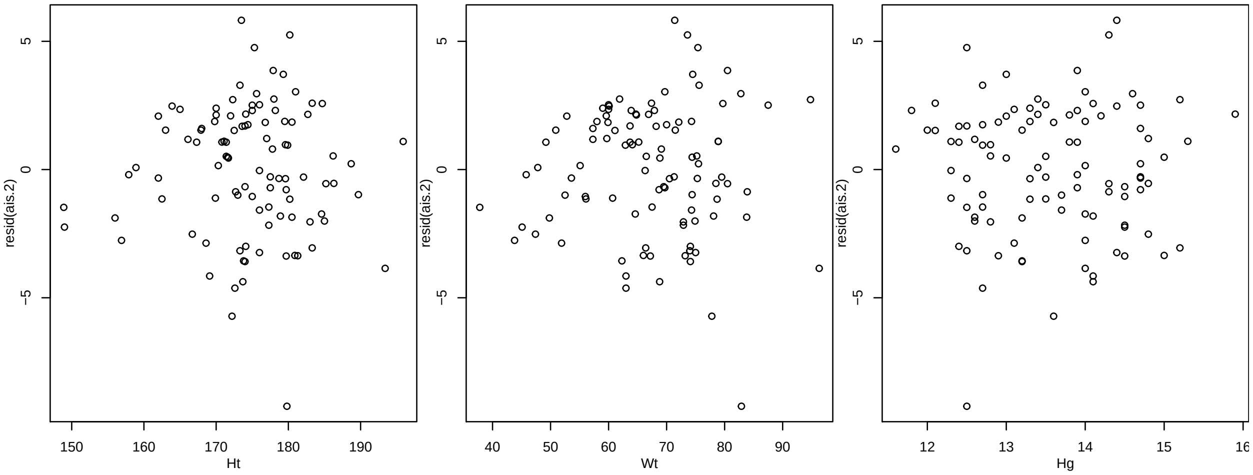
```
# Residual standard error: 2.576 on 96 degrees of freedom
# Multiple R-squared:  0.8657,  Adjusted R-squared:  0.8615
# F-statistic: 206.3 on 3 and 96 DF,  p-value: < 2.2e-16
```

From output we can see `Hg` has increased in significance from a P-Value from 0.0364 to 0.00487. There was correlated information in `RCC` that has been removed that `Hg` is now providing (hence more significant)

```
par(mfrow = c(1, 2))
plot(ais.2, which = 1:2)
```



```
par(mfrow = c(1, 3))
plot(resid(ais.2) ~ Ht + Wt + Hg, data = aisfemales)
```



Quantile plot looks linear and residual plots have no pattern, suggesting linear model still adequate. The loss in R^2 is only a percent, suggesting reduced model is better.

d. Constructing the ANOVA tables for the full and fitted model we have.

```
anova(ais.1)
```

```
# Analysis of Variance Table
```

```
#
# Response: LBM
#           Df  Sum Sq Mean Sq  F value  Pr(>F)
# Ht         1 2379.85 2379.85 355.6421 < 2e-16 ***
# Wt         1 1671.79 1671.79 249.8314 < 2e-16 ***
# RCC        1   26.27   26.27   3.9264 0.05043 .
# Hg         1   30.13   30.13   4.5032 0.03643 *
# Residuals 95  635.71    6.69
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(ais.2)
```

```
# Analysis of Variance Table
#
# Response: LBM
#           Df  Sum Sq Mean Sq  F value  Pr(>F)
# Ht         1 2379.85 2379.85 358.6637 < 2e-16 ***
# Wt         1 1671.79 1671.79 251.9540 < 2e-16 ***
# Hg         1   55.13   55.13   8.3083 0.00487 **
# Residuals 96  636.99    6.64
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Full Model Regression SS $= 2379.85 + 1671.79 + 26.27 + 30.13 = 4108.04$. Reduced model Regression SS = $= 2379.85 + 1671.79 + 55.13 = 4106.77$. RCC Regression SS $= 4108.04 - 4106.77 = 1.27$.

NB: Alternatively, we can find RCC Regression SS but subtracting the Residual SS from each model (RCC SS is added to the Residual SS in the reduced model). RCC Regression SS $= 636.99 - 635.71 = 1.28$. Note that due to 2 decimal place rounding in the displayed table, the computed values are off by 0.01.

Source	*df*	*SS*	*MS*	F-value
RCC|Ht + Wt + Hg	1	1.27	1.27	0.1898
Ht + Wt + Hg	3	4106.77		
Error	95	635.71	6.692	
Total	99	4734.75		

- The F-statistic has $1, 95$ degrees of freedom and P-Value is given by,
    - P-value $= P(F_{1,95} \geq 0.1898) = 0.6641$

NB: We can reconcile this with the ANOVA table where `RCC` is added last to demonstrate the Sequential ANOVA structure.

```
anova(aov(LBM ~ Ht + Wt + Hg + RCC, data = aisfemales))
```

```
# Analysis of Variance Table
#
# Response: LBM
#           Df  Sum Sq Mean Sq  F value    Pr(>F)
# Ht         1 2379.85 2379.85 355.6421 < 2.2e-16 ***
# Wt         1 1671.79 1671.79 249.8314 < 2.2e-16 ***
# Hg         1   55.13   55.13   8.2383  0.005056 **
# RCC        1    1.28    1.28   0.1912  0.662879
# Residuals 95  635.71    6.69
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Note that due to rounding again, the values are slightly off from the manual calculation from the displayed table. E.g. You can check the precise SS values with `anova(ais.1)[['Sum Sq']]` and `anova(ais.2)[['Sum Sq']]`.

e. From the output we have the regression coefficients.

```
summary(ais.2)$coefficients
```

```
#                Estimate Std. Error   t value     Pr(>|t|)
# (Intercept) -10.4680045 7.81495887 -1.339483 1.835760e-01
# Ht            0.1104182 0.04488781  2.459870 1.568733e-02
# Wt            0.5198296 0.03395967 15.307263 1.671372e-27
# Hg            0.8169568 0.28342847  2.882409 4.869868e-03
```

The terms of interest are: $b_{Wt} = 0.5198$, $s.e.(b_{Wt}) = 0.034$, the quantile is $t_{96,1-0.05/2} = 1.985$ For a unit increase in weight we have the expected change in LBM of,

$$\begin{aligned} b_{Wt} \pm t_{n-p,1-\alpha/2} s.e.(b_{Wt}) &= 0.5198 \pm 1.985 \times 0.034 \\ &= 0.5198 \pm 0.0675 \\ &= (0.4523, 0.5873) \end{aligned}$$

From the above we can see that 0.5 is contained in the interval and thus we have data consistent with the claim of the 0.5 kg increase in LBM for each 1kg increase in weight.

Question 2

The file **lifeexp.txt** contains data on life expectancy in years for a number of countries and data on the population per doctor and TV.

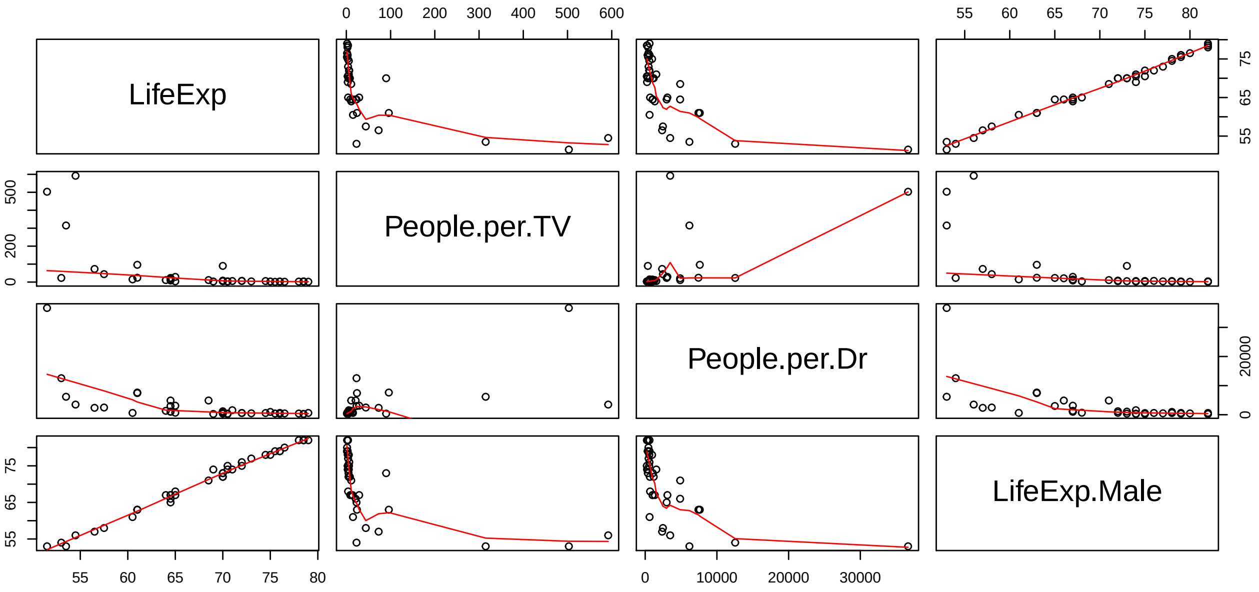
LifeExp	The life expectancy in years
People.per.TV	The average number of people per TV (this is a measure of affluence rather than directly affecting life expectancy)
People.per.Dr	The average number of people per physician

Read the data using a read.table command

```
lifeexp = read.table("lifeexp.txt", header = TRUE)
```

a. Can use either the `plot` function or the `pairs` with smoother, `pairs` version below:

```
pairs(lifeexp, panel = panel.smooth)
```

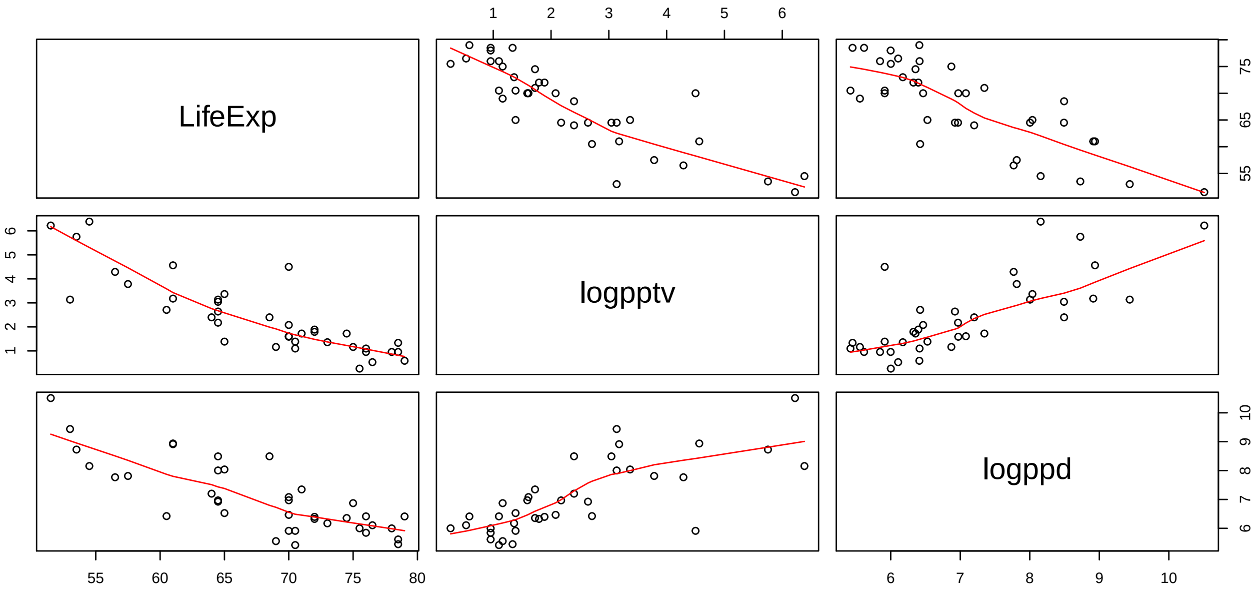


Appears to be nonlinear, People per TV and Doctor both highly skewed.

b. Adding in the log transformed variables.

```
lifeexp$logpptv = log(lifeexp$People.per.TV)
lifeexp$logppd = log(lifeexp$People.per.Dr)
```

```
pairs(LifeExp ~ logpptv + logppd, data = lifeexp,
      panel = panel.smooth)
```



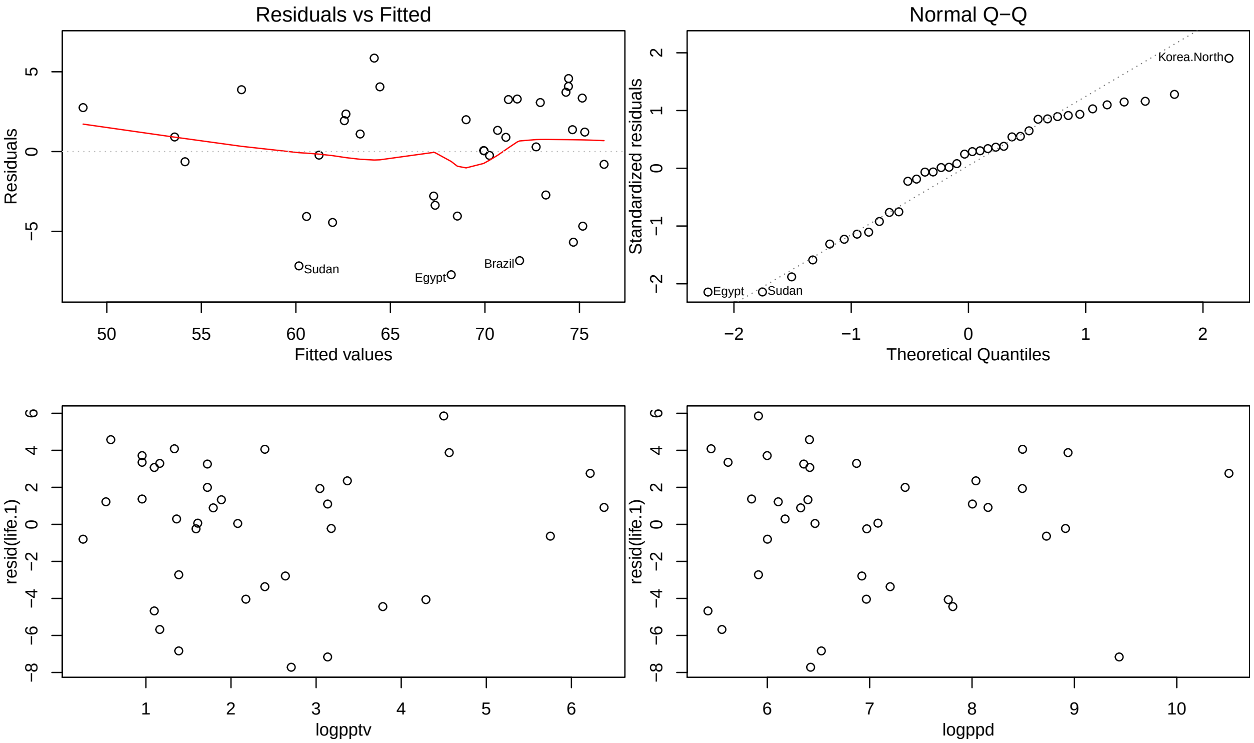
Structure in the data between the response and log predictors looks much more linear. We should be able to fit a linear model now to this data.

c. Fitting the regression on the log transformed predictors.

```
life.1 = lm(LifeExp ~ logpptv + logppd, data = lifeexp)
```

Doing the diagnostic checks we have,

```
par(mfrow = c(2, 2))
plot(life.1, which = 1:2)
plot(resid(life.1) ~ logpptv + logppd, data = lifeexp)
```



Quantile plot looks close to linear, some divergence away from linear. So errors near close to normally distributed. Residual plots have no pattern, suggesting linear model adequate.

d. Finding the coefficient estimates and their standard errors we have,

```
summary(life.1)
```

```
#
# Call:
# lm(formula = LifeExp ~ logpptv + logppd, data = lifeexp)
#
# Residuals:
#     Min      1Q  Median      3Q     Max
# -7.7173 -2.7718  0.9026  2.9923  5.8553
#
# Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
# (Intercept)  90.6222     4.3557  20.806  < 2e-16 ***
# logpptv      -2.9156     0.5907  -4.936 1.95e-05 ***
# logppd       -2.2589     0.7474  -3.022  0.00467 **
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 3.704 on 35 degrees of freedom
# Multiple R-squared:  0.7868,  Adjusted R-squared:  0.7747
# F-statistic:  64.6 on 2 and 35 DF,  p-value: 1.788e-12
```

For both the variables, the quantile of interest is derived from a t-distribution with $n - p = 38 - 3 = 35$ degrees of freedom. This is the quantile $t^* = t_{35,0.975} = 2.0301$

Consider the impact of a 10% increase on the regression equation. The regression equation is,

$$\begin{aligned}\widehat{LifeExp} &= b_0 + b_1 \log(pptv) + b_2 \log(ppd) \\ &= 90.6222 - 2.9156 \log(pptv) - 2.2589 \log(ppd)\end{aligned}$$

So the **change** in Life Expectancy for a 10% increase in *pptv* involves the difference in the above equation between $\log(1.1 \times pptv)$ and $\log(pptv)$. That is,

$$\begin{aligned}\widehat{LifeExp}_{1.1pptv} - \widehat{LifeExp} &= 90.6222 - 2.9156 \log(1.1pptv) - 2.2589 \log(ppd) \\ &\quad - (90.6222 - 2.9156 \log(pptv) - 2.2589 \log(ppd)) \\ &= -2.9156 \log(1.1pptv) - -2.9156 \log(pptv) \\ &= -2.9156 \log(1.1) - 2.9156 \log(pptv) + 2.9156 \log(pptv) \\ &= -2.9156 \log(1.1)\end{aligned}$$

So to compute the confidence interval in this case we just need to multiple the estimate and standard error by $\log(1.1) = 0.0953102$

For the people per TV variable the terms of interest are: $b_{logpptv} = -2.9156$, $s.e.(b_{logpptv}) = 0.5907$,. So, for 10% increase in logpptv (1.1 units) we have the expected change in LifeExp of,

$$\begin{aligned}log(1.1)b_{logpptv} \pm t_{n-p,1-\alpha/2} log(1.1) \times s.e.(b_{logpptv}) &= log(1.1) \times -2.9156 \pm 2.0301 \times log(1.1) \times 0.5907 \\ &= -0.2778864 \pm 0.114296 \\ &= (-0.3922, -0.1636)\end{aligned}$$

We have a similar result for the people per Dr variable. The terms of interest are: $b_{logppd} = -2.2589$, $s.e.(b_{logppd}) = 0.7474$. For a 10% increase in ppd (1.1 units) we have the expected change in LifeExp of,

$$\begin{aligned}log(1.1)b_{logpptv} \pm t_{n-p,1-\alpha/2} log(1.1) \times s.e.(b_{logpptv}) &= log(1.1) \times -2.2589 \pm 2.0301 \times log(1.1) \times 0.7474 \\ &= -0.2152962 \pm 0.1446141 \\ &= (-0.3599, -0.0707)\end{aligned}$$

- **Final interpretation**:
  - For a 10% increase in people per TV we are 95% confident that we would expect a drop in life expectancy between 0.1636 and 0.3922 years.
  - For a 10% increase in people per Dr we are 95% confident that we would expect a drop in life expectancy between 0.0707 and 0.3599 years.