STAT 5385: Lab 3

Willliam Ofosu Agyapong

02/11/2022

0.1 Problem 3.9: Electricity Consumption

An economist studying the relation between household electricity consumption (Y) and number of rooms in the home (X) employed linear regression model and obtained the following residuals:

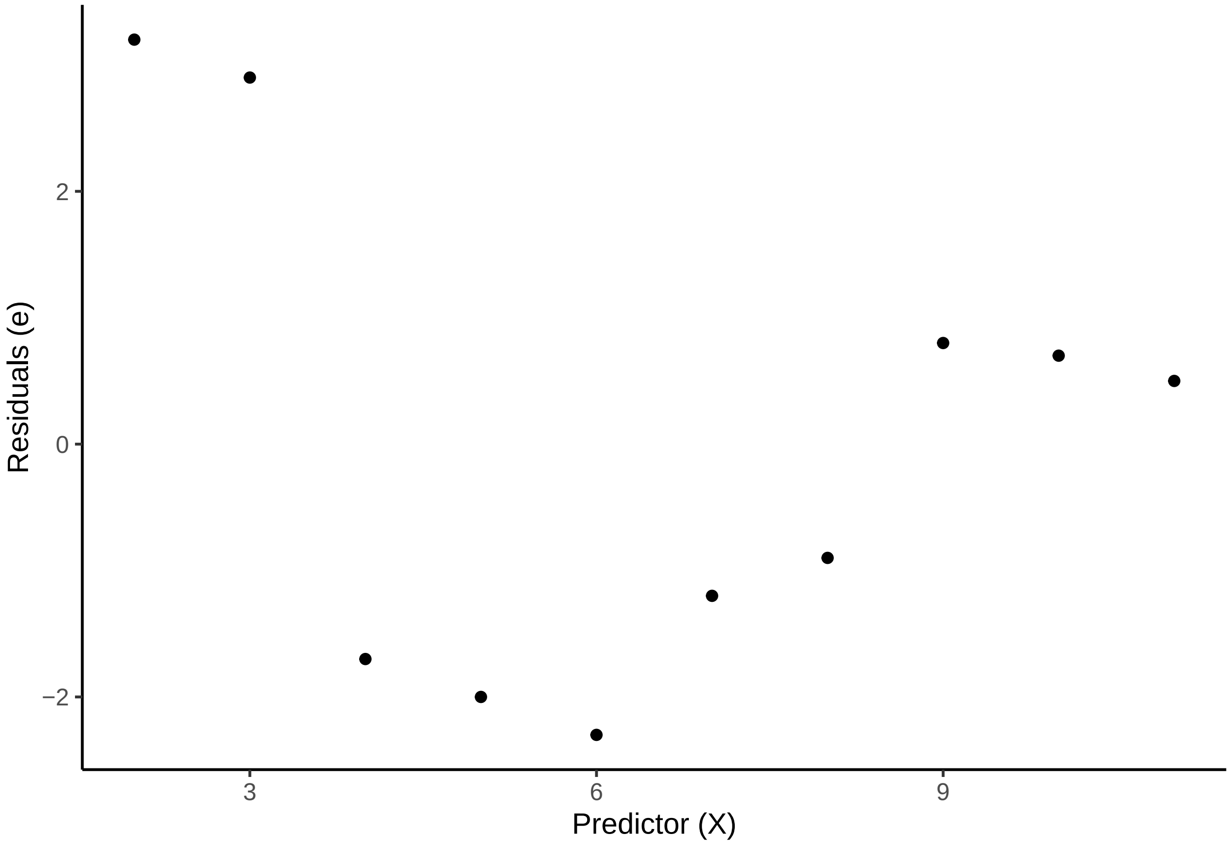
```
# Import the dataset from local drive
electric_resid <- read.table(file = "../Data Sets/Chapter  3 Data Sets/CH03PR09.txt",
                    header = FALSE)

# View first few observations; Everything looks good.
electric_resid <- electric_resid %>%
  mutate(i = as.character(1:n()), .before = "V1")
colnames(electric_resid) <- c("i", "X", "e")
electric_resid2 <- electric_resid
colnames(electric_resid2) <- c("i", "$X_i$", "$e_i$")
kable(t(electric_resid2))
```

i	1	2	3	4	5	6	7	8	9	10
$X_i$	2	3	4	5	6	7	8	9	10	11
$e_i$	3.2	2.9	-1.7	-2.0	-2.3	-1.2	-0.9	0.8	0.7	0.5

We then create a scatter plot of the residuals versus the predictor variable to assess some of the assumptions of the underlying regression model.

```
ggplot(electric_resid, aes(x = X, y = e)) +
  geom_point() +
  labs(x="Predictor (X)", y="Residuals (e)")
```



From the plot, the following problems should be of concern:

- We see that the regression relationship is clearly not linear.
- The error terms appear not to have constant variance. The errors seem to vary in a systematic fashion; the residual is negative for predictor values between 3 and 9, and positive otherwise.

Yes, I believe an appropriate transformation, such as including a quadratic term, might help remedy the problem.

0.2 Using the Prestige data available in the `Car` package.

In this example, we model prestige as a function of education.

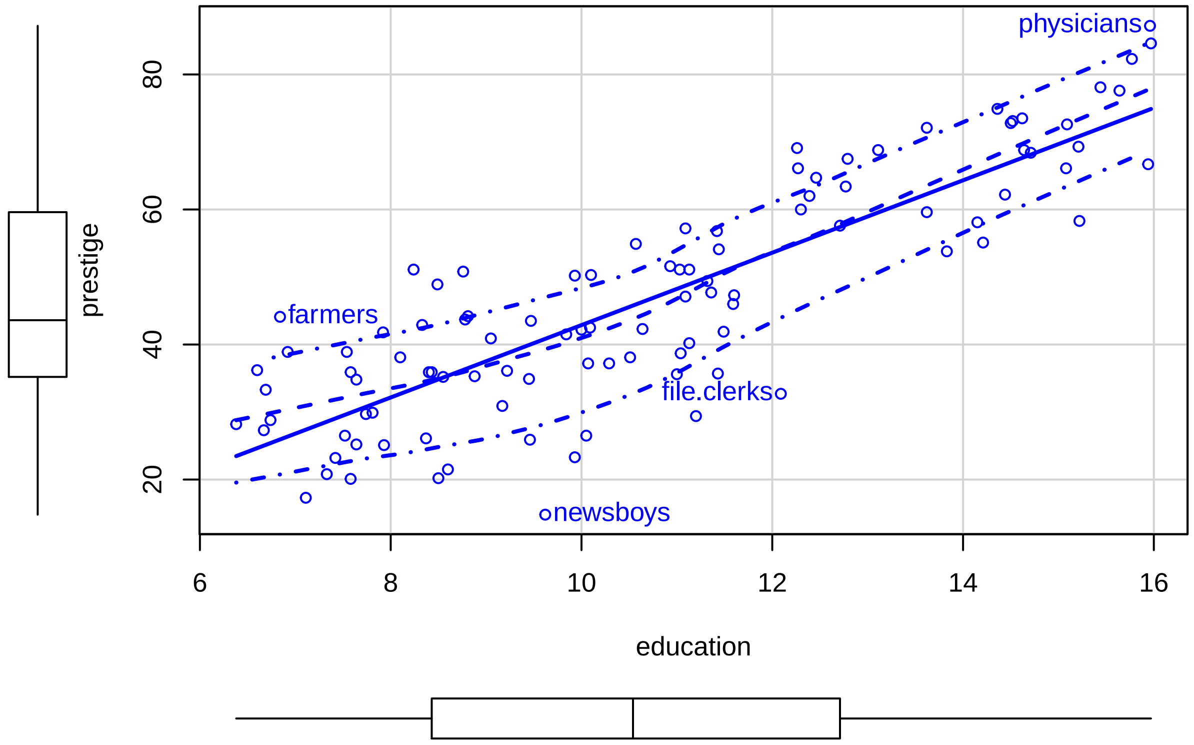
```r
library(car)

# Inspect the data
kable(sample_n(Prestige, 6))
```

	education	income	women	prestige	census	type
nursing.aides	9.45	3485	76.14	34.9	3135	bc
insurance.agents	11.60	8131	13.09	47.3	5171	wc
farm.workers	8.60	1656	27.75	21.5	7182	bc
radio.tv.announcers	12.71	7562	11.15	57.6	3337	wc
physicists	15.64	11030	5.13	77.6	2113	prof
draughtsmen	12.30	7059	7.83	60.0	2163	prof

0.2.1 Initial Explaration of the variables of interest.

```r
scatterplot(prestige ~ education, data=Prestige, id=list(n=4))
```



```
##  physicians file.clerks    newsboys     farmers
##          24          41          53          67
```

The above graph provides useful information about the distributions of prestige and education as well as the relationship between the two variables. For instance, the boxplots tell us that the distribution of education is roughly symmetric while the distribution of prestige appears skewed to the right. Also, there appears to be a positive linear relationship between prestige and education.

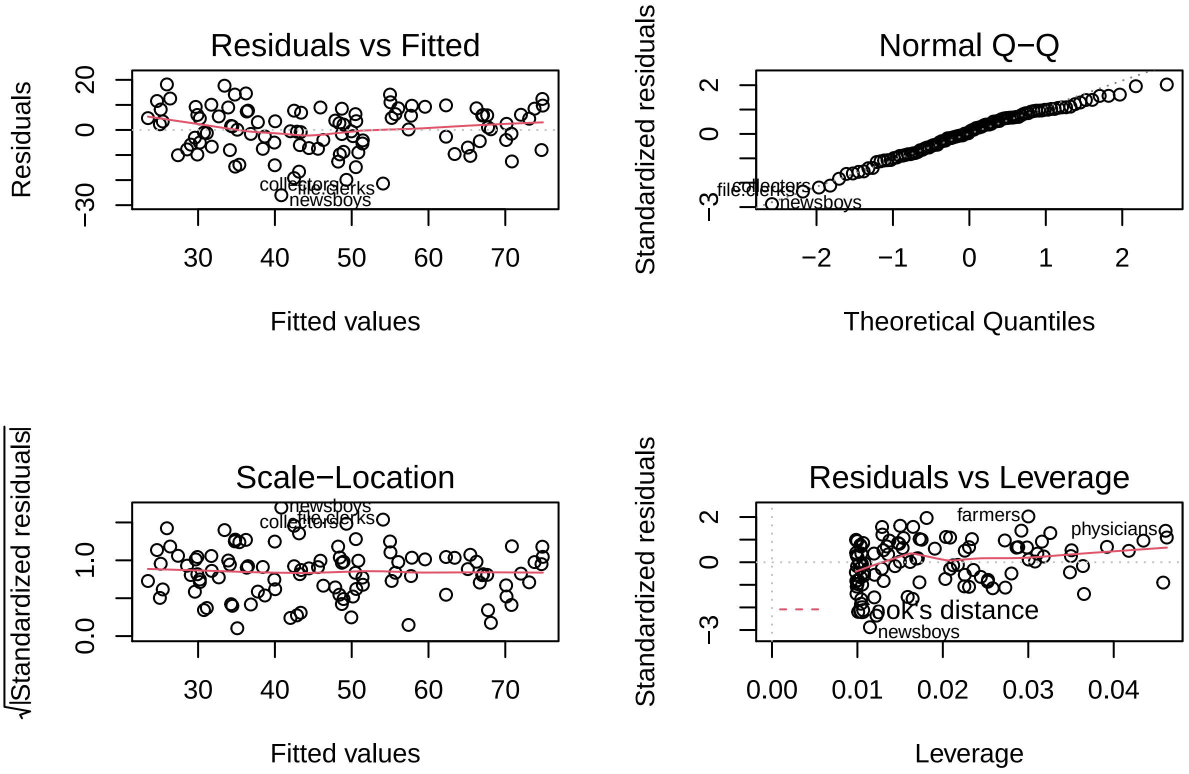
Next, we proceed to model the relationship between prestige and education, which will provide a formal way to help us assess our initial observations, and the appropriateness of our proposed model.

```
# fit a model between prestige and income or education
# and apply appropriate transformation
mdl_prestige_vs_edu <- lm(prestige ~ education, data = Prestige)
summary(mdl_prestige_vs_edu)
```

```
##
## Call:
## lm(formula = prestige ~ education, data = Prestige)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.040  -6.523   0.661   6.743  18.164
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -10.732      3.677   -2.92   0.0043 **
## education      5.361      0.332   16.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.1 on 100 degrees of freedom
## Multiple R-squared:  0.723,  Adjusted R-squared:  0.72
## F-statistic:  261 on 1 and 100 DF,  p-value: <2e-16
```

From the above output, for prestige (Y) and education (X) , the estimated regression function is $\hat{Y} = -10.73 + 5.36X$, where the slope coefficient signifies a positive relationship between education and prestige. That is, as education increases prestige appears to also increase. We defer the type of relationship present to the next section.

```r
# Obtain diagnostic plots
par(mfrow = c(2,2))
plot(mdl_prestige_vs_edu)
```



- Linearity: We observe from the Residuals vs. Fitted plot that there is a linear relationship between prestige and education. This suggests that a linear regression function might be appropriate to model the relationship between prestige and education.
- Constant Error Variance: Again, the Residuals vs. Fitted plot suggests that the error variance are roughly equal.
- Normality of Error Terms: the normal Q-Q plot provides some evidence of a slight departure of the error terms from normality, as seen in the upper part of the plot, probably due to the skewness observed earlier from looking at the boxplot for prestige. I will conclude that this is not too concerning, so the residuals can be assumed to be normally distributed.
- Outliers: From the Residuals vs. Fitted and Scale-location, there appears to be some outliers, but probably not too much concerning.

Overall, a linear regression function appears to provide a better fit for the relationship between prestige and education.