Regression Analysis: HW 2

Matthew Tichenor

September 13, 2015

1 Introduction

The data set used features demographics information such as the population per TV, for 40 different countries. Here a simple linear model is analyzed: $log(y_i) = \beta_0 + \beta_1 \cdot log(x_i) + \epsilon_i$, where y_i is the population per TV, x_i is the population per Doctor and ϵ_i is the random error term, for $i = 1, ..., 40$. The random error terms are assumed to be independent and identically distributed with mean, 0, and variance, σ^2.

This model makes the following basic assumptions:

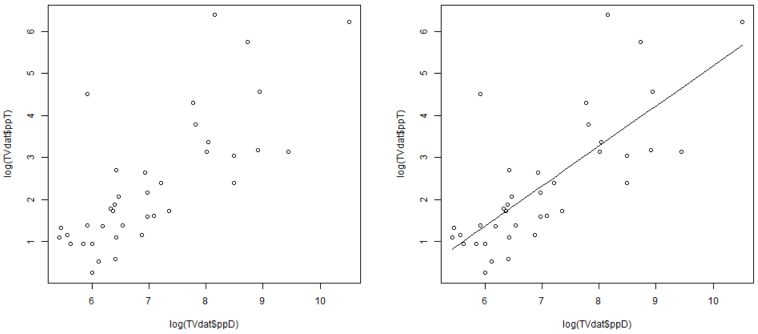
1. The model is sufficient.
2. The errors are symmetric around 0.
3. The errors are uncorrelated.
4. The variance of the errors is constant.
5. There are no outliers.

2 Checking the Assumptions

One part of this homework is to check if these assumptions hold. In the following sub-sections I check each assumption using graphs.

2.1 Checking Model Adequacy

A simple plot of the data can be used to check model sufficiency. Figure 1a plots the data, and it appears that there is a linear relationship between the log of population per Doctor and population per TV.



(a) Basic Plot

(b) Regression Line

Figure 1b displays the plot with the fitted regression line. Here the linear relationship is no longer left to the imagination, as the fitted line is the least squares regression line, which uses estimates, $\hat{\beta}_0$, $\hat{\beta}_1$ that minimize the cost function $Q(\beta_0, \beta_1) = \sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x))^2$.

2.2 Checking the Errors

Three of the five basic assumption, involve the errors. Because the errors are assumed to be independent, identically distributed, have mean 0 and have a constant variance, it follows that errors should be symmetric about 0. Figure 2 shows a plot of the residuals vs the fitted values.

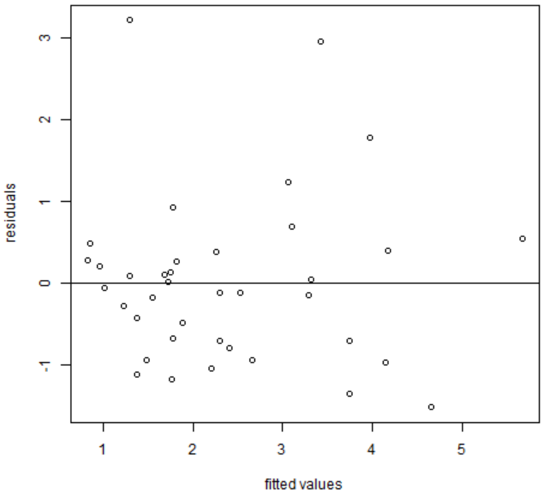


Figure 2: Residual Plot

The fitted values are the population per TV estimates, $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$; and the residuals are the error estimates, $y_i - \hat{y}_i$. If the errors are not symmetric about 0, then the residuals would also not be symmetric about 0. Since the residuals in Figure 2 are fairly symmetric about the line $y - \hat{y} = 0$, there seem to be no problems with this assumption.

In addition to checking the prior assumption, the residual plot can also be used to check if the errors have a constant variance. If the errors have a non-constant variance, then there will be a pattern to the residuals in the residual plot. Examining Figure 2, there doesn't seem to be a pattern in the residual plot, so it's safe to assume the errors have a constant variance.

The third assumption, that the errors are uncorrelated is difficult if not impossible not verify. Realistically, there probably is a relationship between the population per TV for different countries, such as say France and Germany, since these two countries lie in the same geographic area, Western Europe, and may have similar economies since they are both members of the European Economic Union. In order to continue this analysis, I assume that this is not the case, and that the responses, as well as the errors, are uncorrelated.

2.3 Checking for Outliers

The ordinary least squares regression model is sensitive to outliers, so it's important to check that there are no outliers in the data. Figure 2 can be used to find outliers, and it does appear that at least one exists, though there may be more. The residual plot is only one tool of several to find outliers in the data, so the leverage plot is examined in this section.

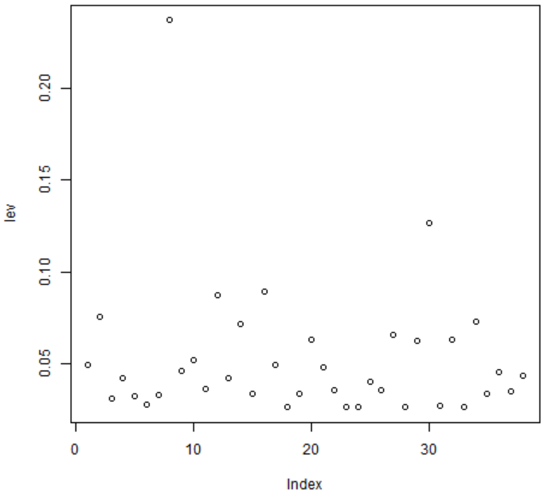


Figure 3: Residual Plot

The leverage plot appears in Figure 3. The leverage in the leverage plot is $h_{ii} = \frac{(x_i-\bar{x})^2}{\sum(x_j-\bar{x})^2}$, which measures how much influence a single observation has on the regression line. Outliers have a large influence on the regression line relative to other observations. In the leverage plot for this analysis, there is a one observation that has a much higher leverage than other observations. This observation is the 8th observation.

To identify the outlier, I examine the data set. The data set looks something like this:

```
life  ppTV  ppDr flife mlife
Argentina       70.5   4.0   370    74    67
Bangladesh      53.5 315.0  6166    53    54
```

```
Brazil          65.0   4.0   684    68    62
Canada          76.5   1.7   449    80    73
China           70.0   8.0   643    72    68
Colombia        71.0   5.6  1551    74    68
Egypt           60.5  15.0   616    61    60
Ethiopia        51.5 503.0 36660    53    50
.
.
.
```

Ethiopia is the 8th observation and seems to have very few doctors compared to the number of TV's. Removing outliers can be tricky. Ethiopia's population per TV and population per Doctor do not seem to be the result of any recording error by researchers. While removing Ethiopia from the dataset would produce a better, simple linear model, this is not a good reason for removing the observation. On the contrary, the model should be removed and replaced with a better model. Since this is the only model known thus far, the outlier is ignored and analysis continues.

3 The Estimates

The parameters estimated in the model are β_0, β_1, and σ^2. These are estimated by $\hat{\beta_0} = \bar{y} - \hat{\beta_1}\bar{x}$, $\hat{\beta_1} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$ and $\hat{\sigma^2} = \frac{\sum(y_i-\hat{y_i})^2}{n-2}$. From our data we have $\hat{\beta_0} = -4.3417$, $\hat{\beta_1} = 0.9527$ and $\hat{\sigma^2} = (1.045)^2$.

Because $\hat{\beta_1}$ and $\hat{\beta_0}$ are normally distributed and $\hat{\sigma^2}$ has a X^2 distribution, we can construct confidence intervals for β_0 and β_1 using a T distribution. We calculate these intervals with $\hat{\beta} \pm std(\hat{\beta}) \cdot t_{\frac{\alpha}{2},n-2}$. With 95% confidence $-6.356 \leq \beta_0 \leq -2.327$ and $0.671 \leq \beta_1 \leq 1.234$. Note that $0 \notin (0.671, 1.234)$.

4 Variance Decomposition

In the previous section, σ^2 was estimated with $\hat{\sigma^2} = \frac{\sum(y_i-\hat{y_i})^2}{n-2}$. This estimate is also called the Mean Square Error (MSE), and the numerator is called the Sum of Squares Error (SSE). The quantity $\sum(y-\bar{y})^2$ is the Sum of Squares Total (SST) and $\sum(\hat{y_i}-\bar{y})^2$ is the Sum of Squares Regression (SSreg). SST describes the total variation in the y_i's, SSE describes the variation in the y_i's due to random error and the SSreg describes the variation in the y_i's that can be explained with the model. Furthermore, SST = SSreg + SSE. We can look at the percentage of total variation that is explained by the model, which is $R^2 = \frac{SSreg}{SST} = \frac{SST-SSE}{SST} = 0.567$.

Because we know R^2 and $\hat{\sigma^2}$, we can calculate the SST and SSreg. Since $MSE = \frac{SSE}{n-2} = \frac{SSE}{36} = (1.045)^2$, we have $SSE = 39.313$. Now that SSE is known we can do some algebra to find SST and SSreg, since $R^2 = \frac{SST-SSE}{SST} = 0.567$, and SSreg = SST - SSE. I found $SST = 90.771$ and $SSreg = 51.458$. Despite rounding errors, we can see $SSreg + SSE = 51.458 + 39.313 = 90.771 = SST$.

5 Code

The following lines of codes were used in this assignment:

```
> par(mfrow=c(2,2))
> TVdat <- read.table("TV.dat", sep="\t")
> plot(log(TVdat$ppD),log(TVdat$ppT))
> plot(log(TVdat$ppD),log(TVdat$ppT))
> mm <- lm(log(TVdat$ppT)~log(TVdat$ppD))
> lines(sort(log(TVdat$ppD)[is.na(TVdat$ppT)==F]),mm$fit[sort.list(log(TVdat$ppD)[is.na(TV
> plot(mm$fit,mm$res, xlab="fitted values", ylab="residuals")
> abline(0,0)
> lev <- hat(model.matrix(mm))
> plot(lev)
> summary(mm)
> qt(0.975,df=36)
```