MSG500/MVE190
Linear Models - Lecture 2

Rebecka Jörnsten
Mathematical Statistics
University of Gothenburg/Chalmers University of Technology

October 31, 2012

1 RECAP

The regression model is used to summarize the relationship or association between the dependent variable y and the independent variable x:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

We fit this model to the data $(x_i, y_i)_{i=1}^n$ using Least Squares, i.e. we find β_0 and β_1 that minimize

$$Q(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$$

Fitting a model using least squares makes sense if the 5 basic assumptions hold for the data:

1. The models is sufficient for the data. Does it look like a line fits through the mass of the $y - x$ data cloud? If not, try transforming x and/or y.
2. The random scatter appears to be symmetric around the model (mean 0 errors, symmetric distribution around 0). If not, try data transformations or consider a different fitting criterion.
3. Uncorrelated errors. Thinks about the study design. Is time involved? Are data generated by following individuals over time? Are observations clustered or grouped (same school, same hospital, different cities)? If so, you need to use time series or mixed effects models, not standard regression modeling.
4. Constant error variance. If the scatter appears to be larger for some ranges of x and/or y the error variance is non-constant. If so, the Least Squares criterion is not optimal for fitting since you let poor quality observations influence the fit as much as better quality ones. Try data transformations to suppress the non-constant variance or use a Weighted Least Squares criterion instead (more later).
5. No outliers. Since all observations contribute to the fit, look for observations whose $x-y$ relationship seems to differ from the bulk of the data. These observations may severely hamper finding a good model for the rest of the data. Look for extremes in x and/or y and remove these from the data. Note, a data transformation may make these outlying observations fit more with the rest of the data. Try that first.

Some other problems to look out for: groups in the data, missing values, omitted variables, extrapolation, causal interpretation.

2 More about Least Squares

$$Q = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2 \longrightarrow \text{ minimize w.r.t. } \beta_0, \beta_1$$

We take the derivative w.r.t. β_0 and set it equal to 0:

$$\frac{dQ}{d\beta_0} = \sum_{i=1}^{n} -2(y_i - (\beta_0 + \beta_1 x_i)) = 0$$

and the same with β_1:

$$\frac{dQ}{d\beta_1} = \sum_{i=1}^{n} -2x_i(y_i - (\beta_0 + \beta_1 x_i)) = 0$$

We manipulate the above expressions into the so-called *normal equations*:

$$\sum_{i=1}^{n} y_i = n\beta_0 + (\sum_{i=1}^{n} x_i)\beta_1$$

$$\sum_{i=1}^{n} x_i y_i = (\sum_{i=1}^{n} x_i)\beta_0 + (\sum_{i=1}^{n} x_i^2)\beta_1$$

We first solve for β_0:

$$\hat{\beta}_0 = \bar{y} - \beta_1\bar{x}, \quad \bar{y} = \sum_{i=1}^{n} y_i/n, \bar{x} = \sum_{i=1}^{n} x_i/n$$

and then for β_1, plugging in the above solution for β_0:

$$\sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} x_i(\bar{y} - \beta_1\bar{x}) + \beta_1 \sum_{i=1}^{n} x_i^2$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2}$$

which after some manipulation we can write as

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \simeq \frac{COV(x,y)}{COV(x,x)} = Corr(x,y)\sqrt{\frac{V(y)}{V(x)}}$$

Take a look at this last expression. The solution for the slope β_1 consists of two parts: the relationship summary between x and y as captures by the *correlation* between x and y, and a scale factor related to the spreads of x and y as captured by their variances. If we first standardize the data so that the standard deviation1 of x and y are both 1, then the slope of the regression model is equal to the correlation. What does this tell us?

- regression is a *linear* relationship summary (since that's what the correlation measures)
- If y and x are standardized the slope is between -1 and 1

This last point is often misinterpreted as follows: Let's say you let a group of students take a standardized test and then you let them re-take it a few months later. You plot the standardized first test scores against the standardized second test scores in a scatter plot and fit a regression model to the data. The slope will be less than 1 by definition (see above). This is one interpretation: "students that did poorly the first time around did better the second time around so they must have gotten scared about the first result and studied more. students that did well the first time around go lazy and did worse the second time around". Now, this interpretation is wrong. What we see here is what's called the "regression toward the mean" effect. If you are the best there's nowhere to go but down, and if you were the worst there's nowhere to go but up! Mathematically, we see this effect reflected in the slope less than 1.

2.1 Looking at the slope estimate

We have

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} =$$
$$= \frac{\sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} y_i = \sum_{i=1}^n k_i y_i.$$

That is, the slope estimate is just a weighted average of y values. The weights $k_i = \frac{(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$ are large for observations y_i whose x-values are far from the average $\bar{x}$, i.e. the extreme x-values contribute most to the regression slope estimate. This makes sense if you think about eye-balling an estimate of a slope.

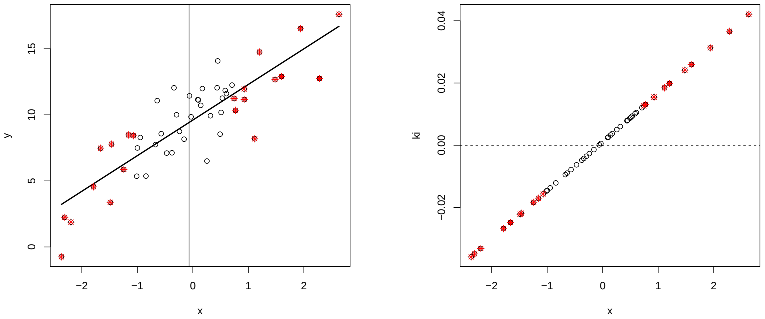


Figure 1: Left: regression line, extreme x-values are highlighted. Right: the weights k_i as a function of x

In Figure 1 you see the regression model fit and the observations that contribute most to the fit (extreme x-values). Note that the weights k_i can be "pre-computed", i.e. I don't even need to know the value of y_i to determine its weight in the estimate of β_1. This is another definition of a *linear* model, linear refering to the weights not depending on y. In fact, in statistics we think of models where the regression can be computed like this (weights that are not depending on y) as linear.

Some properties

We have $\hat{\beta}_1 = \sum_i k_i y_i$. It follows that

$$E[\hat{\beta}_1] = E[\sum_i k_i y_i] \underset{\text{only } y \text{ is random}}{=} \sum k_i E[y_i] = \sum_i k_i(\beta_0 + \beta_1 x_i)$$

Now, $\sum_i k_i = 0$ and $\sum_i k_i x_i = 1$ (prove this to yourself) so

$$E[\hat{\beta}_1] = \beta_1$$

That is, $\hat{\beta}_1$ is an *unbiased* estimate of the slope parameter β_1. (We can prove this for $\hat{\beta}_0$ as well - try it yourself.)

How about the variance of the slope estimate? We assume that the error has variance $V(\epsilon) = \sigma^2$. Since $y = \beta_0 + \beta_1 x + \epsilon$ and only ϵ is random, it follows that $V(y) = \sigma^2$ as well.

$$V[\hat{\beta}_1] = V[\sum_i k_i y_i] \underset{\text{uncorrelated errors}}{=} \sum_i k_i^2 V[y_i] \underset{\text{constant variance}}{=} \sum_i k_i^2 \sigma^2$$

Now, plug in the expression for the weight k_i above and we get

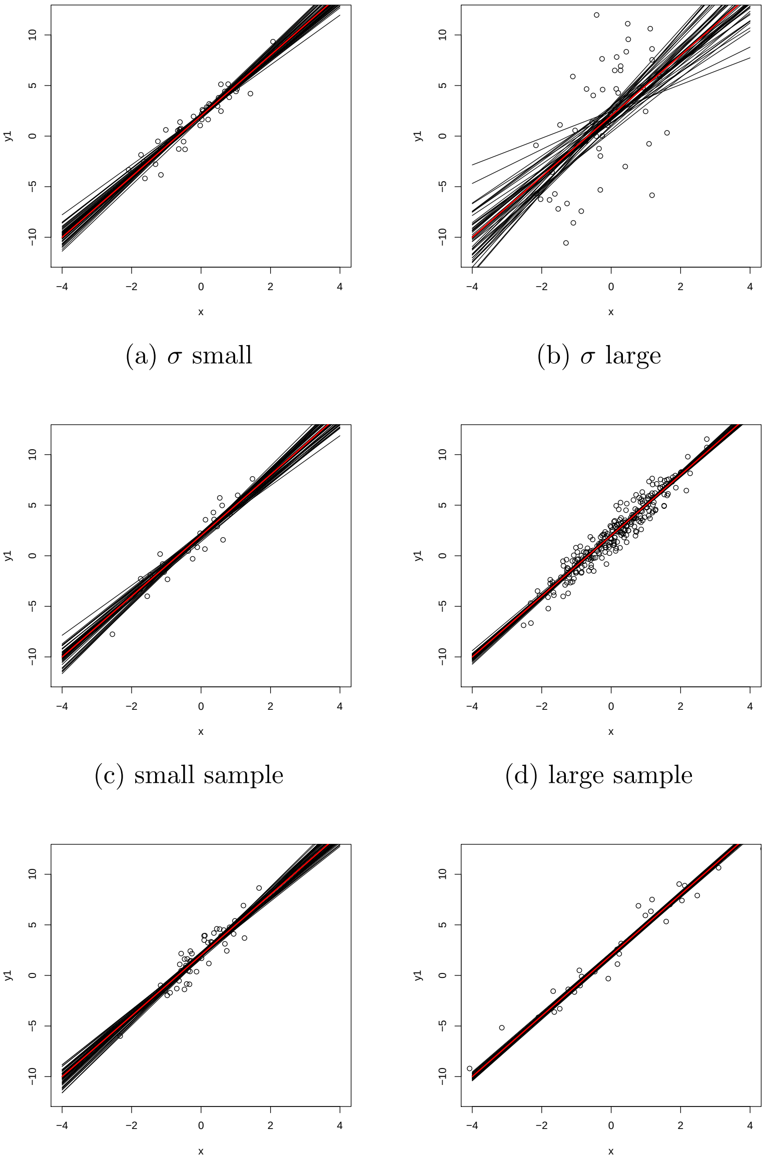
$$V[\hat{\beta}_1] = \sum_i \frac{(x_i - \bar{x})^2}{(\sum_j (x_j - \bar{x})^2)^2} \sigma^2 = \frac{\sigma^2}{\sum_j (x_j - \bar{x})^2}$$

The expression for the estimation variance of $\hat{\beta}_1$ can be interpreted this way: there are 3 sources of error in the slope estimate:

1. $V[\hat{\beta}_1]$ increases with the noise level in the data σ^2. The less noise we have, the easier it is to estimate the relationship.

2. $V[\hat{\beta}_1]$ decreases with the sample size n since $V[\hat{\beta}_1] = \frac{\sigma^2}{(n-1)V(x)}$. The more data we have, the easier it is to estimate the model parameters.

3. $V[\hat{\beta}_1]$ decreases with increasing spread in x ($V(x)$). The more spread we have in x, the easier it is to detect the $y - x$ relationship (we get a chance to see how y is affected by x - if x doesn't vary, how could we see the effect of it on y?).



(a) σ small

(b) σ large

(c) small sample

(d) large sample

(e) $V(x)$ small

(f) $V(x)$ large

Figure 2: Sources of error variance of $\hat{\beta}_1$. Red line is the true model. Black lines are estimated models from 50 different data sets. (a)-(b) the noise level σ. (c)-(d) the sample size. (e)-(f) the spread in x.

As you can see in Figure 2, the impact of these three sources of error in the estimation is easily demonstrated using simulated data. **Check the source code for the lecture notes and play around**

with sample size, noise level and spread in x and see what happens to the slope estimates.

2.2 More properties and diagnostics

- The *fitted values* are defined as $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. These are values that lie on the regression line.
- The deviation between the observation y_i and its fitted value $\hat{y}_i$ is called the *residual* and is defined as $e_i = y_i - \hat{y}_i$.
- If your model includes an intercept (β_0), then the residuals sum to 0: $\sum_i e_i = 0$.
- In addition, $\sum_i x_i e_i = 0$. What this means is that the residuals and the explanatory variable x are uncorrelated, meaning that we "used up" all the linear information in x about y by fitting the Least Squares regression model to the data. This is also referred to as an *orthogonal projection* (see Figure 3).

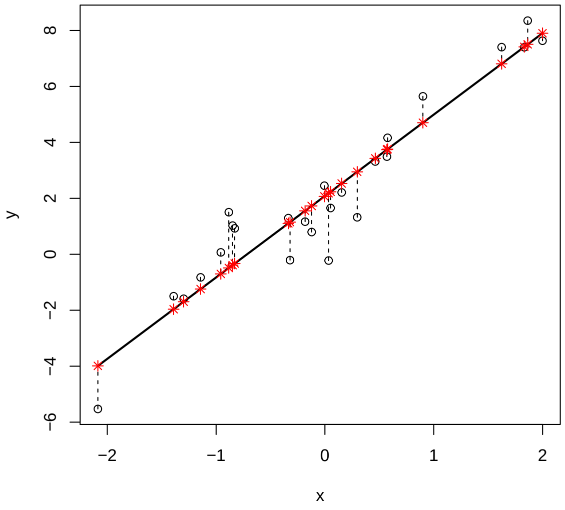


Figure 3: The estimated regression model (solid black) and the residuals (vertical, dashed). The fitted values as red asterisks. Note that the residuals are ortogonal to the x-axis (orthongonal projection)

While the scatter plot (the $x - y$ points in Figure 3) is a useful tool for data exploration, the *residual plot* is a better diagnostic tool (used to spot problems with the fit, checking the 5 basic assumptions). We thus plot just the residuals versus x in Figure 4. We use the residual plot to check the basic assumptions.

1. Check that there are no trends or patterns in the residual plot (checking basic assumption 1)
2. Check that the residuals are evenly distributed symmetrically around the horizontal line at 0 (checking basic assumption 2)
3. You *cannot* check basic assumption 3 (uncorrelated errors) with residual plots in general
4. Check that the residuals have about the same spread around the horizontal line at 0 (contant error variance). Another plot that's useful here is to plot the residuals against the fitted values to see if the error variance is a function of the expected value of y. (checking basic assumption 4).
5. Look for extreme values in the residual plot - observations that were not captured by the model (outliers). You won't catch all outliers this way - observations that are so extreme they drive the fit (extreme x AND y values) may actually have small residual values (more later). (checking basic assumption 5).

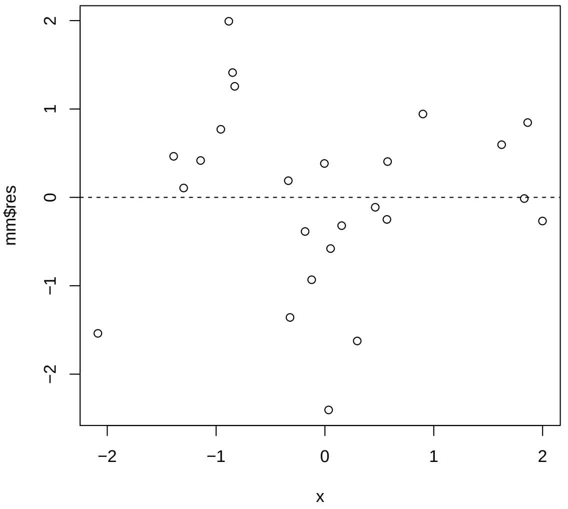


Figure 4: Residual plot. The residuals in Figure 3 plotted against x. The horizontal line at 0 can be used to gauge the goodness of fit of the model.

Fitted values and leverage

The fitted values $\hat{y}_i$ are uncorrelated with the residuals: $\sum_i \hat{y}_i e_i = 0$ (do the math to check this). It follows from the fact that the fitted values are just a function of the x_i's and x and the residuals e are uncorrelated. What does this mean? This means that the fitted values $\hat{y}$ is the part of y explainable from x, whereas the residuals (orthogonal to x) is the unexplainable part.

Now, we can write the fitted values as follows:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \bar{y} + \hat{\beta}_1(x_i - \bar{x}) = \bar{y} + (\sum_j k_j x_j)(x_i - \bar{x}),$$

where I used the definition of the weight $k_j = \frac{x_j - \bar{x}}{\sum_i (x_i - \bar{x})^2}$ from before. We manipulate this expression further to obtain

$$\hat{y}_i = \sum_j (\frac{1}{n} y_j + k_j(x_i - \bar{x}) y_j) = \sum_j (\frac{1}{n} + k_j(x_i - \bar{x})) y_j = \sum_j h_{ij} y_j.$$

That is, the fitted values are *also* a weighted average of y-values just like the slope estimate was. Here, the weights h_{ij} tell us how much observation y_j contribute to the fitted value at observation i: $\hat{y}_i$. Not, if x is constant all $h_{ij} = \frac{1}{n}$. That means, if all the x values are equal the fitted value is simply $\hat{y}_i = \bar{y}$, the mean of y. That makes sense. If x is constant there is no additional information in x about y so the best guess for y is the mean value.

Now, the weight h_{ii}, the amount observation y_i contributes to its own fitted value, is called the *leverage*. It is defined as

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_k (x_k - \bar{x})^2}.$$

Some things to note about the leverage:

- h_{ii} is on the order $1/n$
- h_{ii} is large is x_i is extreme, i.e. far from $\bar{x}$.
- if $h_{ii} >> h_{ij}$, the fitted value $\hat{y}_i$ is dominated by its own value y_i.

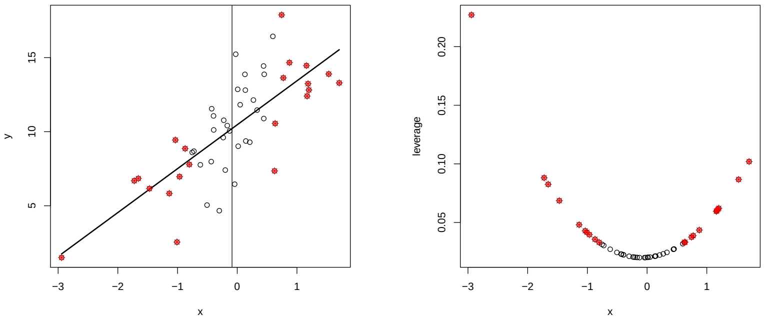


Figure 5: Left: regression line, extreme x-values are highlighted. Right: the leverage h_{ii} function of x_i

In Figure 5 I show a data set with extreme x-values highlighted. The corresponding leverages is shown in the right panel.

Leverage can be used to identify potentially problematic observations. In Figure 6 we revisit the outlier discussion from Lecture 1. In panel (b) we see an observation with high leverage, but little influence on the fit since the y-value is not extreme. Panel (c) shows a high-leverage observation with an extreme y-value. Since high-leverage observations influence the fitted values a lot, the regression line is drawn toward this extreme y-value. Panel (d) is an outlier with low leverage since its x-value is not extreme. Outliers of type (b) do not show up as large residuals, but can be detected with a leverage plot like Figure 5 (b). Outliers of type (c) may not show up in the residual plot either, but outliers of type (d) will.

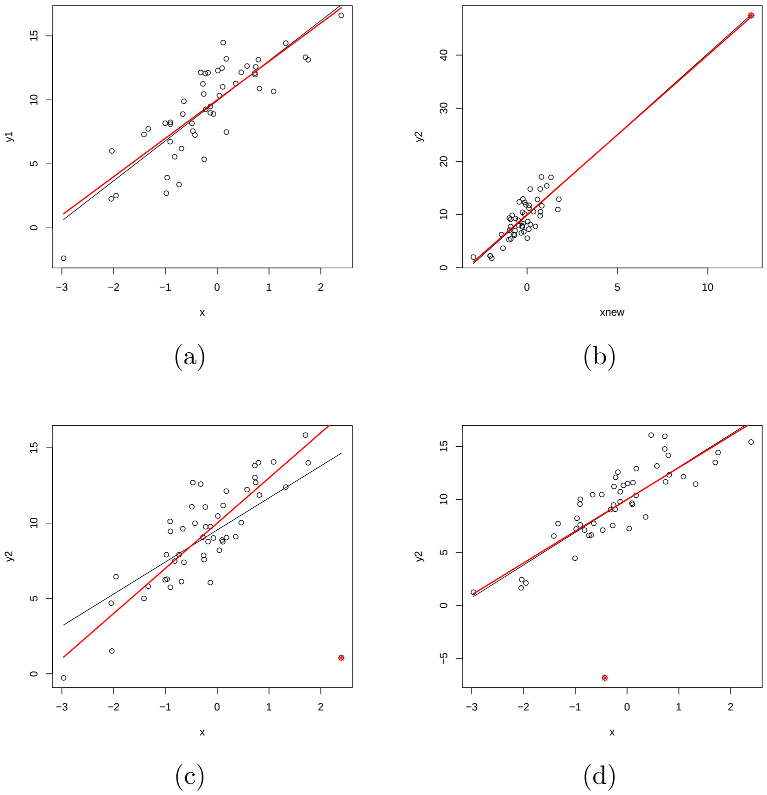
We can conclude that the leverage and residual plot are not sufficient to diagnose problems with the fit. Some other diagnostic plots are needed.

- Impact on the slope: We can plot the value of the slope estimate when an observation i is excluded and compare with the slope estimate when it is included. Outliers of type (c) will show up in this diagnostic plot since their presence alter the slope estimate a lot.

- Impact on the Least Squares criterion: We can plot the value of the Q (sum of the squared deviations between observations and model) when we exclude observation i. If excluding i lowers the least squares criterion by a large amount we have detected an outlier of type (d).

- The Cook's distance: We can combine the above (impact on slope, leverage) into one measure - called the Cook's distance. This will allow us to identify outliers of type (b), (c) or (d) (more later).

In Figure 7 I show the impact on the slope estimate by dropping observation i and the impact on the Least Squares cost-function Q.



(a)

(b)

(c)

(d)

Figure 6: No outliers. (a) No outliers present, (b) Extreme observation in x, (c) outlier in y, (d) outlier in y for a non-extreme x-value

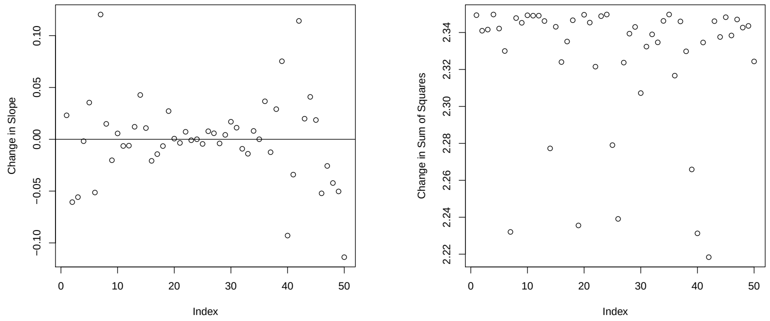


Figure 7: Left: Impact on slope estimate by dropping observation i. Right: Impact on Sum of Squares criterion Q by dropping observation i.

3 Demo 2

Let's revisit the Television data from Lecture 1.

```
> TVdat <- read.table("TV.dat", sep = "\t")
> print(dim(TVdat))
```

```
[1] 40  5
```

```
> print(names(TVdat))
```

```
[1] "life"  "ppTV"  "ppDr"  "flife" "mlife"
```

```
> print(row.names(TVdat))
```

```
 [1] "Argentina"       "Bangladesh"      "Brazil"          "Canada"
 [5] "China"           "Colombia"        "Egypt"           "Ethiopia"
 [9] "France"          "Germany"         "India"           "Indonesia"
[13] "Iran"            "Italy"           "Japan"           "Kenya"
[17] "KoreaNorth"      "KoreaSouth"      "Mexico"          "Morocco"
[21] "Myanmar (Burma)" "Pakistan"        "Peru"            "Philippines"
[25] "Poland"          "Romania"         "Russia"          "South Africa"
[29] "Spain"           "Sudan"           "Taiwan"          "Tanzania"
[33] "Thailand"        "Turkey"          "Ukraine"         "United Kingdom"
[37] "United States"   "Venezuela"       "Vietnam"         "Zaire"
```

```
> plot(TVdat$ppD, TVdat$ppT, xlab = "people per Dr", ylab = "people per TV")
> id <- identify(TVdat$ppD, TVdat$ppT, row.names(TVdat), pos = T)
```

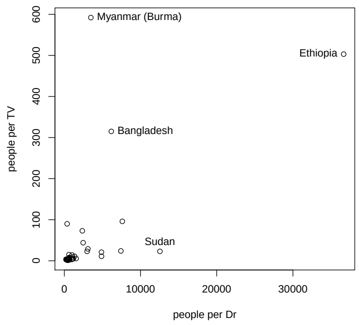


Figure 8: People per TV vs People per Dr

```
> plot(log(TVdat$ppD), TVdat$ppT)
> id <- identify(log(TVdat$ppD), TVdat$ppT, row.names(TVdat), pos = T)
```

```
> plot(log(TVdat$ppD), log(TVdat$ppT))
> id <- identify(log(TVdat$ppD), log(TVdat$ppT), row.names(TVdat),
+     pos = T)
> 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(TVdat$ppT
+     F])])
```

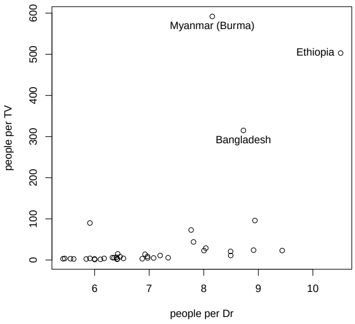


Figure 9: People per TV vs People per Dr: logs on ppDr to even out the spread in x

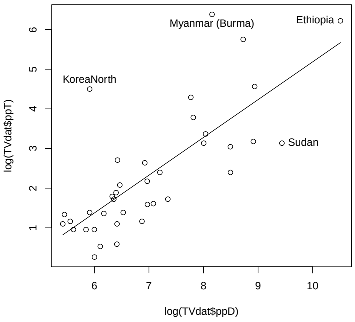


Figure 10: People per TV vs People per Dr: logs on ppTV to suppress non-constant variance. Regression line

```
> library(xtable)
> xtable(summary(mm), caption = "Regression summary", label = "tab:ch4")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-4.3417	0.9933	-4.37	0.0001
log(TVdat$ppD)	0.9527	0.1388	6.86	0.0000

Table 1: Regression summary

3.1 Residuals and Leverage

```
> induse <- seq(1, dim(TVdat)[1])[is.na(TVdat$ppT) == F]
> plot(log(TVdat$ppD)[induse], mm$res)
> abline(h = 0)
> id <- identify(log(TVdat$ppD)[induse], mm$res, row.names(TVdat)[induse],
+     pos = T)
```

```
> lmi <- lm.influence(mm)
> plot(log(TVdat$ppD)[induse], lmi$hat, ylab = "leverage")
> id <- identify(log(TVdat$ppD)[induse], lmi$hat, row.names(TVdat)[induse],
+     pos = T)
```

```
> plot(induse, lmi$coef[, 2], ylab = "Impact on Slope")
> abline(h = 0)
```

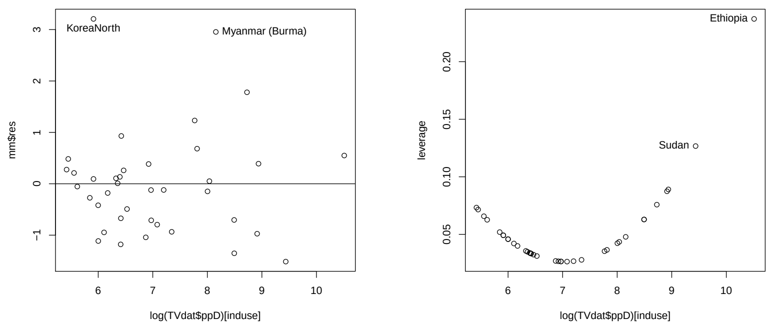


Figure 11: Residual and Leverage plot for the Television regression model

```
> id <- identify(induse, lmi$coef[, 2], label = row.names(TVdat)[induse],
+     pos = T)

> plot(induse, lmi$sig, ylab = "Impact on Sum of Squares")
> id <- identify(induse, lmi$sig, label = row.names(TVdat)[induse],
+     pos = T)
```

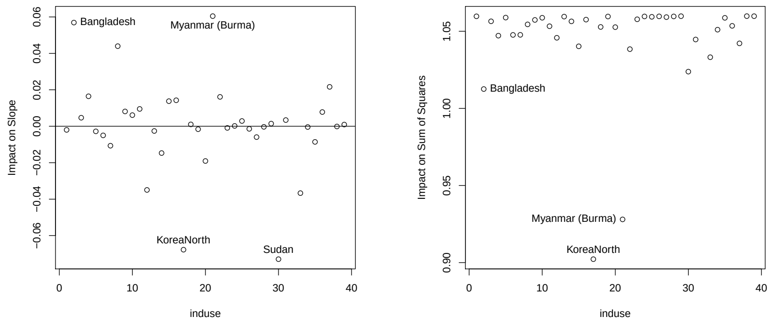


Figure 12: Impact on Slope (left) and Residual sum of squares (right) when dropping observation i