MSG500/MVE190
Linear Models - Lecture 1

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

October 29, 2012

1 Introduction

Simple linear regression is about summarizing the *relationship* between the *dependent* variable y and the *independent* or *explanatory* variable x.

y: dependent, outcome

x: independent, covariate, explantory

In regression we model y *given* x. That is, we view x as fixed, non-random.
However, in practice, usually both x and y are measured and subject to error or random scatter.
We sometimes try to put the variable with the smallest error in x, *but* more likely there is a natural choice for x and y when we think of the model as being "generative", that x causes y. It is important to realize that we may not be able to prove this causative statement - regression models the *association* between variables.

Model: $y = f(x) + \epsilon,$ where

- $f(x)$ is the explainable variability in y through x, and
- ϵ is the "unexplainable" part, random error or scatter.

The form of the model $f(x)$ is usually assumed to be known. Examples:

- linear: $f(x) = \beta_0 + \beta_1 x$
- polynomial: $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$
- ... more and more complex and flexible
- local smoother: $f(x)$ = local average of y-values in a neighborhood of x-values

	Simple/rigid	added complexity	Flexible
Model structure	Linear	polynomial/nonlinear	Local average/smoother
Estimation properties	Large bias, low variance	...	Small bias, high variance

This *Bias-Variance* is key in statistical modeling.

- Simple models make rigid assumptions on f .... which leaves little flexibility for the data to influence the fit .... which gives us low variance of estimation (not much allowed to vary from fit to fit)... BUT because we don't have much flexibility we may fail to describe the data well, i.e. we may make consistent under- or over-estimates, not capturing the true relationship = Bias
- Flexible models are very responsive to the data... which leads to high variance in estimation (lots can change from fit to fit)...
  BUT we can capture quite complex relationships and match the data structure well. i.e. the models have low bias.

Let's look at some concrete examples. The code that generates these results and figures are available on the course homepage.

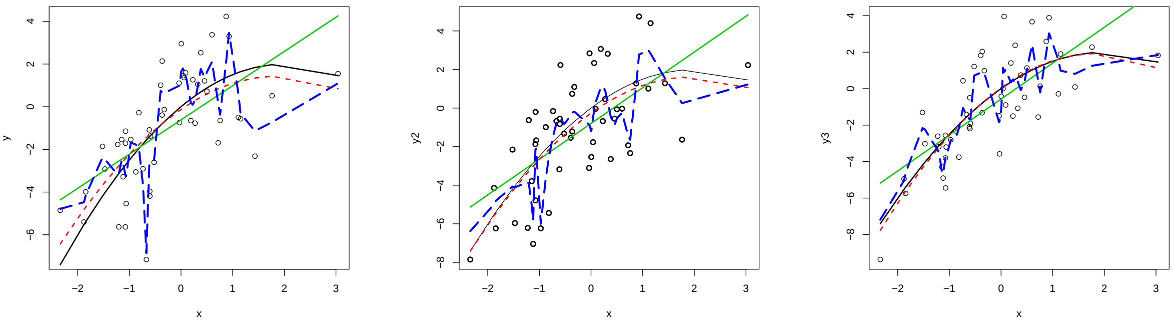


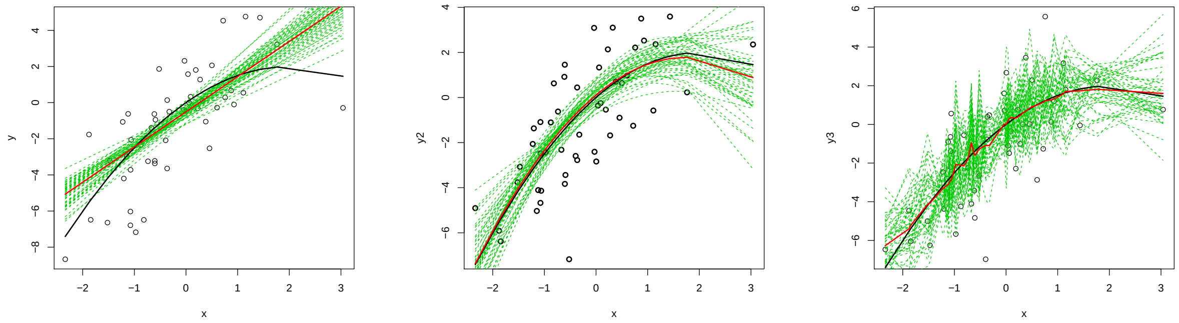
Figure 1: 3 data sets from the same true underlying model (black line): linear (green), quadratic (red) and locally average model (blue) fits

In Figure 1 I show you three data sets of size $n = 50$ observations drawn from a true quadratic model $y = 2 * x - .5 * x^2 + \epsilon$, with the noise terms ϵ distributed as $N(0, \sigma = 2)$. As you can see, each data set y (black circles) deviates from the true $f(x) = 2 * x - .5 * x^2$ in a random fashion because of the error term. This randomness is then reflected in the different fits in each of the three figures (comparing the same colored lines between the left, middle and right panels). Now, you can see in Figure 1 that the linear model (green line) doesn't change much from data set to data set. On the other hand the local smoother (blue line) changes a lot.
To Try: Look at the codes for this lecture - you can play around with different models and noise levels and examine how much the fits vary from instance to instance.

Let's see what happens on average. I generate $K = 50$ data sets from the true model and plot the average model fit as well as each individual fit for linear, quadratic and local smoother models.

In Figure 2 you see each model fit (50 different fits) as a green line. The average fit is colored in red, and the true model is the black line. Now, the deviation between the average fit (red) and the true model (black) is the BIAS. In this example the true model is quadratic so the quadratic model (b) and the local smoother (c) exhibit very low bias. That is, these models are able to capture the true $y \sim x$ relationship. Looking at (a) on the other hand you see that the linear model on average consistently deviates from the true model for certain values of x, overestimating y for very small and very large values of x and underestimating y for the middle range of x.

The second thing to note from the figure is that each linear model fit does not deviate much from its average value. That is, in (a) you see that each green line is fairly close to the the red average fit. In (b) we see a slight increase in spread among the fits (the green lines), and in (c) we note that each fit can vary substantially from the average fit. This demonstrates the impact of model flexibility on estimation



(a) (b) (c)

Figure 2: True model=black line. Each model fit is green, average model fit is red. (a) Linear model, (b) Quadratic model, (c) local smoother

VARIANCE. The more rigid or simple a model is, the less each fit will vary from the average fit and vice versa.

The average fit, here computed from 50 different fits to data, tells us something about the *best case scenario* for this type of model. As the number of data sets we average over increases, this average approaches the *expected value* of the model fit - that is, averaging out over the random part of the fit, the error term ϵ in the model. So, the average fit (red line) is the best we can do with each model since the random error contribution to fitting is eliminated. What is left is the model adequacy. Here, we can see that the linear model is not adequate for capturing this particular data structure (which in truth is quadratic).

Now, it might seem logical to "play it safe" by always using as complicated a model as possible. This way, the average fit will be almost guaranteed to agree with the true model (as seen in Figure 2 (c) above) (LOW BIAS). However, it's not quite that simple. As you can see from the figures, what will happen for an individual data set can be quite far from the best case scenario (average fit). Using a complex model like the local smoother in (c) means we are very sensitive to the random error in y. Each fit can be quite far from the average fit (look at the spread among the green lines) (HIGH VARIANCE). The point is that we won't know the truth and so won't know how far from it we are in each particular instance. The statistical way of "playing it safe" is to take both bias and variance into account. We may be willing to risk some bias if we can suppress the variance of the fit. That is, we risk making some consistent over- or underestimations on average but make sure the fit can't be too wild.
To Try: Highly flexible models can be risky to use because of their large estimation variance. However, when we have a lot of data to work with, the estimation variance becomes less of a concern. Try out the code for the lecture with a larger data set, say 100, 250 or 500 observations - what do you see?

2 Least Squares estimation

How do we fit a model to the data? The most commonly used approach is called *Ordinary Least Squares* (OLS). First some notation:

$$(x_i, y_i)_{i=1}^n$$

are the paired $x - y$ observations. The subscript i denotes one of the n paired observations. The model assumption can be written as

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

where we call β_0 the *intercept* and β_1 the *slope*.
We define a cost-function, or modeling criterion

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

Each term $(y_i - (\beta_0 + \beta_1 x_i))$ constitutes a deviation between the true observation y_i and the model value $\beta_0 + \beta_1 x_i$. Q is the squared sum of those deviations. Our goal is to pick β_0 and β_1 to minimize this sum. That is,

$$\text{Best model}=argmin_{\beta_0,\beta_1} Q(\beta_0, \beta_1).$$

The values of β_0 and β_1 that minimize Q are called are *parameter estimates.* We usually denote these with either latin letters or with a "hat" like this: $\hat{\beta}_0, \hat{\beta}_1$ or b_0, b_1. I will use the "hat" notation.

- Q is *additive* over the observations - and all observations contribute equally to Q.
- Q is a function of the *squared* deviations, meaning positive and negative deviations are equally important.

In Figure 3 I depict a small data set of size 25. I plot the model that minimizes Q (red line) and the corresponding deviations (vertical solid lines). For comparison I also include another model (blue line). As you can see, on average, the magnitude of the deviations to the blue line exceeds those to the least squares fit (red line).

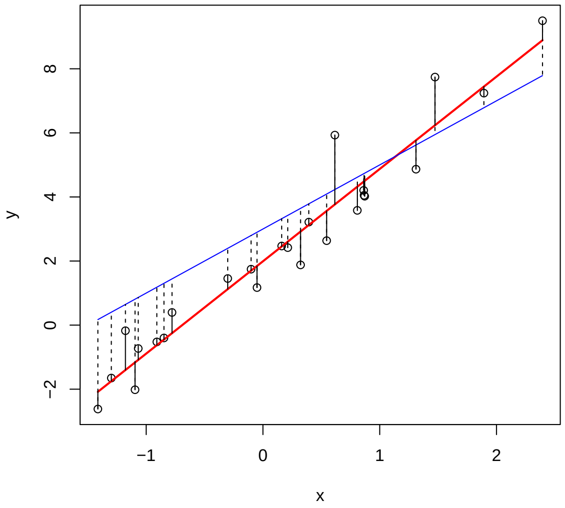


Figure 3: The deviations between the model fit and each observation for two different fits. Red line minimizes the least squares Q

Least Squares is a very popular choice for the estimation objective since it's so easy to work with. We get nice, simple solutions for the parameter estimates $\hat{\beta}_j, j = 0, 1$. However, there are several assumptions we have to make about the data for Least Squares to be a sensible approach to modeling:

2.1 The 5 basic assumptions

1. The model is sufficient

The model you assume for the data should be sufficient to describe the data. In Figure 4 I depict 3 data sets. The first is one where a linear model $\beta_0 + \beta_1 x$ is true, whereas the middle and right plots are examples of violations of this assumption. The middle plot is an example of *groups in data*, i.e. that you may need more than one model to describe the data. The right figure is an example of a violation of the linearity of the model, here the data is actually coming from a quadratic relationship.

How do we deal with violations of the model suffiency?

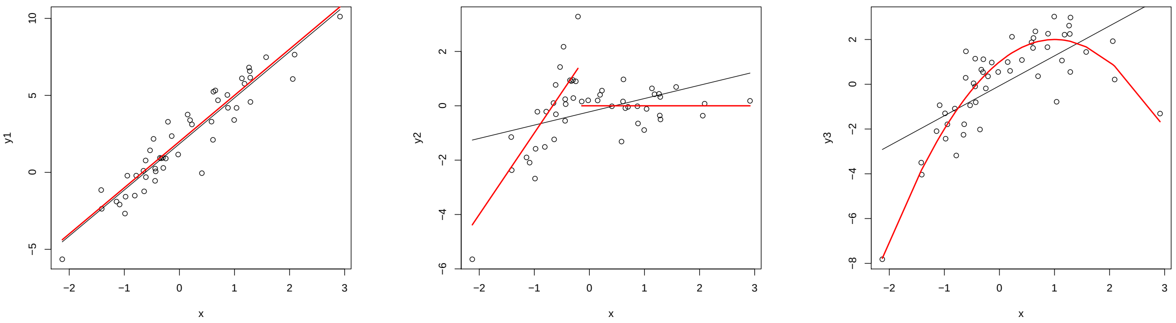


Figure 4: Sufficient Model: Left: sufficient model. Middle,Right: Insufficient. Black line=fitted model. Red=true model.

- Data transformations: You try to transform x and/or y to make the scatter plot of y vs x looks as linear as possible. Some commonly used transformations include log, square-root, inverse, etc.
- Up the complexity: If transformations don't fix the problem you have to use more complex models, e.g. use a quadratic model for the data in the right panel of Figure 4.

2. Symmetric errors around 0

The Least Squares criterion uses square deviations to measure how well the model fits the data. That means, positive and negative errors are valued equally. This would not make sense if your errors have a skewed distribution where e.g. you see many larger positive errors than you do negative ones. Also, it assumes that the errors do not have a drift, i.e. that their expected value is 0 ($E[\epsilon] = 0$).

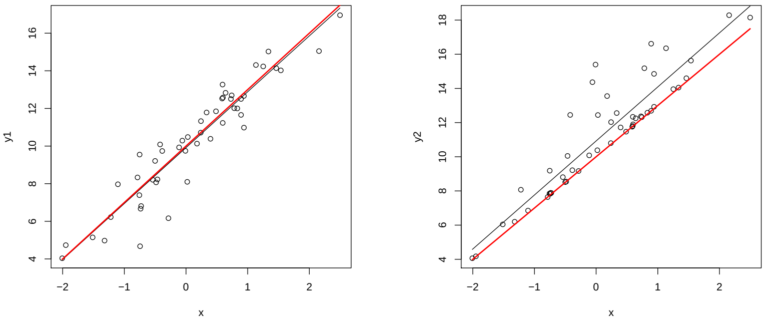


Figure 5: Symmetric Errors around 0. Left: sufficient model. Right: Insufficient. Black line=fitted model, Red line=true model.

In Figure 5, the right panel illustrates what goes wrong when the error distribution is skewed and biased away from zero. Since positive and negative errors count equally in Q, the minimizing model is pulled up toward the observations with large positive errors.

How do we deal with this violation?

- Sometimes a data transformation can suppress the asymmetry of the errors, e.g. logs of y.
- If the transformation doesn't help, you may need to consider a different estimation strategy. Think about this: is the mean a good summary of a skewed distribution? How about the median? You can perhaps come up with a specific distribution of the errors that capture the skewness, e.g. an

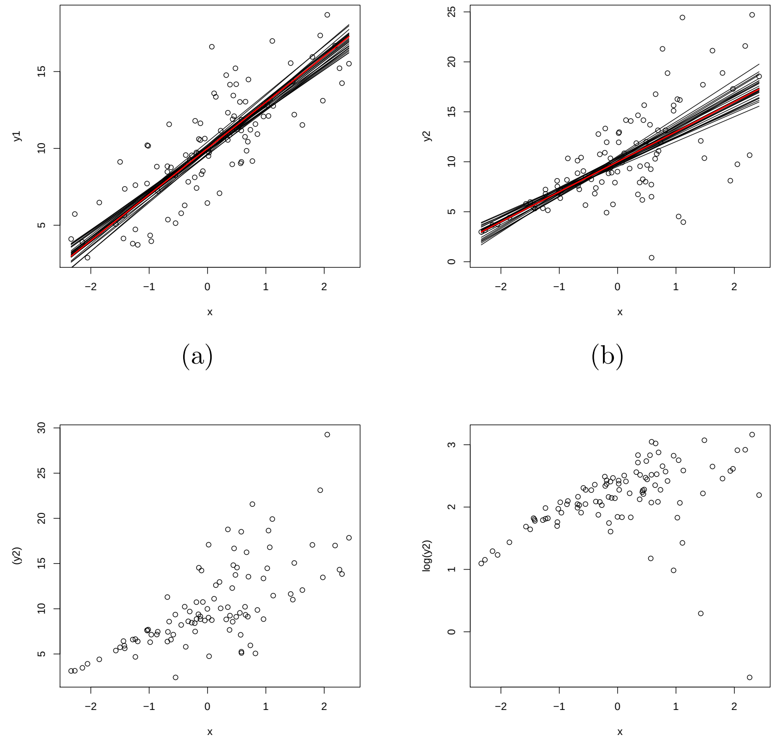
exponential or Poisson distribution for ϵ. Such assumptions do not lead to a Least Squares solution (see Maximum Likelihood methods, or Generalized Linear Models (later in the class)).

3. Uncorrelated errors

Least Squares let all observations contribute equally to the cost function. Now, if some observations are actually just another measurement of the same individual (so-called repeated measures) or observations are grouped together somehow (e.g. students in the same class, classes in the same school, schools in the same city) they share information. In other cases, observations have a natural ordering like time. In all of these situations, the shared information means the errors are correlated.

What's the problem? Well, let's say that we have 10 individuals whose height and weight we have measured, but the 10th individual took these measurements on 3 consecutive days and reported 3 paired height and weight values. We could think of this as a data set of size $n = 12$ BUT in doing so we let the 10th individual contribute 3 times as much to the modeling than anyone else. Compare this with the situation where we had measurements taken from 12 different individuals. Which tells us more about the relationship between height and weight in the population?

The impact of ignoring the correlated errors is that we delude ourselves into thinking we have more informative data that we actually do. The grouped or clustered or repeated observations provide some redundant information and so the "effective" sample size is not 12 in the case of the 3 repeated measures. Treating the observations as independent when they are not can have a substantial impact on inference (testing hypotheses, setting up confidence intervals for parameters, etc). There is no easy "test" to spot this violation. You have to think about the design of the study - are observations clustered or grouped? If so, you have to use so-called *Mixed effects models* or *Time series models.*



(a)

(b)

(c)

(d)

Figure 6: Constant Error Variance. Top panel. 25 data sets (a) Sufficient model. (b) Insufficient model. Black lines=the 25 model fits, Red line=true model. Lower panel. (c) non-constant error variance. (d) trying to suppress the violation by taking logs

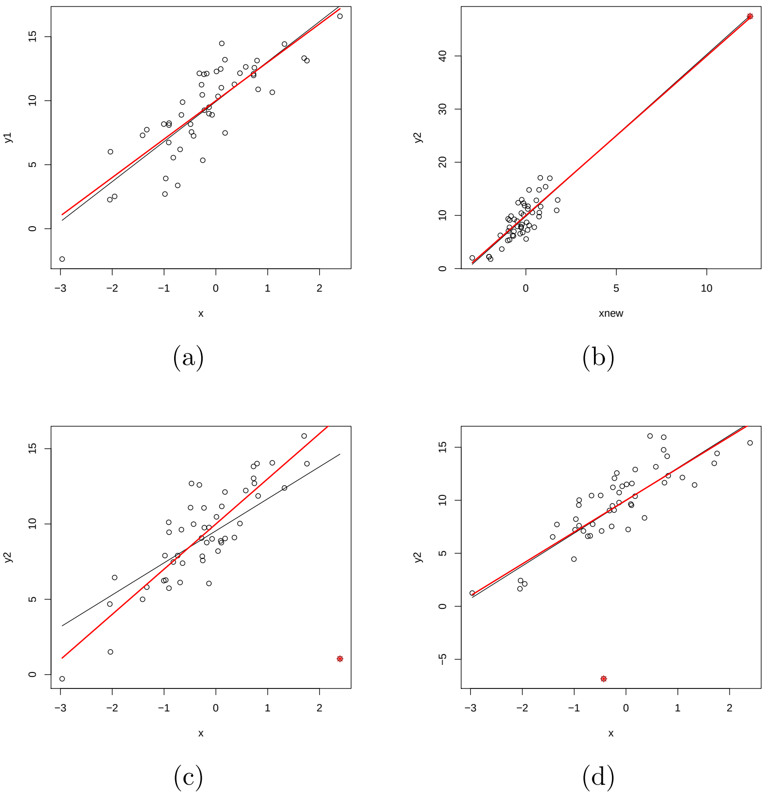
4. Constant error variance

Since all observations contribute equally to the Least Squares criterion we are making an implicit assumption that the noise level (or data quality) is about the same. In Figure 6 I provide an example where this is true (a) and when this assumption is violated (b). The impact of non-constant error variance (assuming the mean of the error is still 0) is on the variance of the estimation. This is seen comparing the left and right panels. In panel (c) I depict one particular instance of this violation (increasing variance with x). In panel (d) I take log of y - as you can see this can suppress non-contant variance a little.

How do we deal with non-constant error variance?

- Try data transformations that suppresses the non-constant error variance. Logs and square root works quite well in many cases.
- Weighted Least Squares. We can use prior knowledge about the data quality to weight the observations in the cost-function. There are also methods where we can learn the appropriate data weights from the data itself (later in the class).

5. No outliers



(a) (b) (c) (d)

Figure 7: No outliers. (a) No outliers present, (b) Isolated observation in x, (c) outlier in y, (d) outlier in y for a non-extreme x-value

Since all observations contribute to the fit of the model using Least Squares, all observations are assumed to come from the same underlying model. If some observations violate this assumption (a measurement error, an unusual outcome, etc), this can have a huge impact on the fit.

In Figure 7 I depict some examples. In (b) we have an outlier in the x - the independent variable. This might not seem like such a big deal at first, but as you will see later in the class, such violations can cause you to be overly confident about the fit of your model. Case (c) is the worst kind of outlier. The presence of an unusual observation in y for an extreme value of x will have a huge impact on the fit.

Case (d) has an impact of a second-order nature. That is, unusual values in y for non-extreme x have little impact on the model fit itself *but* it does impact the estimation of the variance of the fit, which is used to test hypotheses and draw conclusions from your model.

How do we deal with outliers?

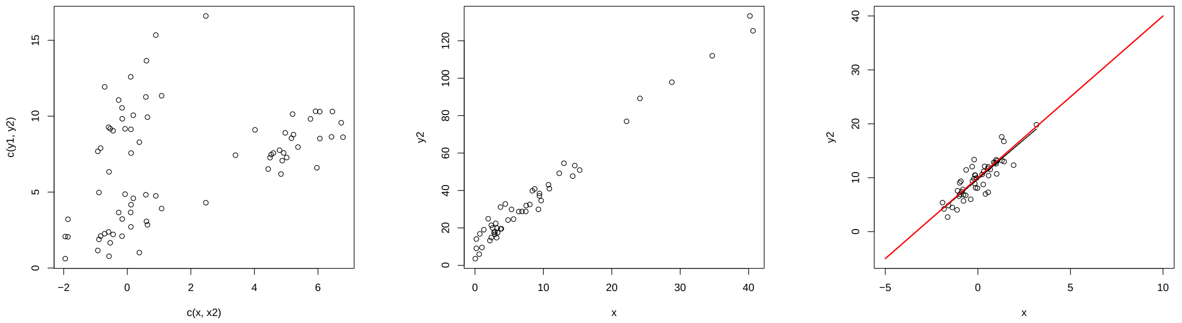
- Drop or down-weigh the observations.
- Caution: Is the observation an "outlier" meaning a poor quality observation OR an important deviation that indicates the model is inadequate? Try to figure out *why* this observation sticks out.
- Caution: Is it OK to drop 1% of the observations? how about 10%? how about 25%? Are the outliers actually a group of data?

We will discuss outliers more in upcoming lectures.

Other things to look out for

Least Squares regression modeling is inappropriate

- if there are groups in the data (Figure 8 (a)) - perhaps need separate models for each group
- if there is uneven spread in the data (Figure 8 (b)) - try transformations of the data
- to use for causal interpretation. If you go on a diet and lose weight, will you also get shorter?
- for extrapolation (Figure 8 (c)). Don't assume the model fits outside the observed range of x.



(a) (b) (c)

Figure 8: Other things to look out for. (a) Groups in the data (b) Uneven spread (c) Extrapolation

3 Summary

- Check the Basic Assumptions
- If violations are detected, try data transformations first, then small model expansions (like quadratic model instead of a linear one).
- How to check: graph! graph! graph!

R-Intro, Demo 1

Let’s look at the ”Television” data set. The data consists of 40 countries. For each country I have data on the expected life span of males and females and on average regardless of gender. I also have data on people per medical doctor as well as people per TV in each contry. Let’s examine the relationship between people per doctor and people per TV for these countries.

```
> 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)
```

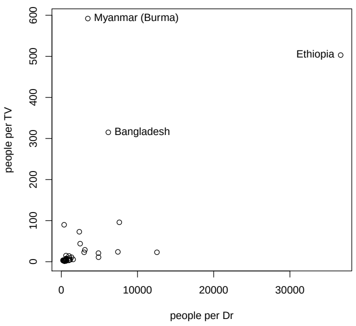


Figure 9: People per TV vs People per Dr

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

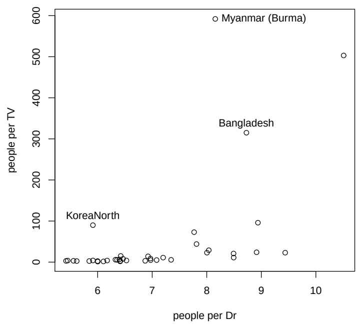


Figure 10: People per TV vs People per Dr: logs on ppDr to even out the spread in x

```
> 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])]])
```

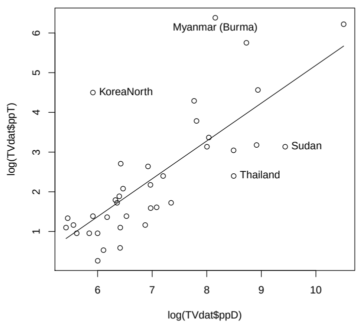


Figure 11: People per TV vs People per Dr: logs on ppTV to suppress non-constant variance. Regression line

```
> 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