# MSG500/MVE190
# Linear Models - Lecture 4

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

November 14, 2014

## 1 RECAP

- The variance of the fitted values increase with the leverage, $V(\hat{y}_i) = \sigma^2 h_{ii}$. This means that the estimated regression line almost pivots around the center $\bar{x}, \bar{y}$ and thus the position of the line can differ a lot in extreme locations of $x$ from data to data set
- The variance of the residuals exhibit the opposite pattern, $V(e_i) = \sigma^2(1 - h_{ii})$
- The residuals have non-constant variance and are correlated! Compare with the true errors that have constant variance and are uncorrelated.
- The noise level, $\sigma^2$ is a nuisance parameter that we estimate as $\hat{\sigma^2} = MSE = RSS/(n-2)$ where $RSS$ is the residual sum of squares.
- By comparing the $RSS$ to the total sum of squares, $SS_T = \sum_i (y_i - \bar{y})^2$ we see how much of the variability in $y$ can be explained through $x$.
- We summarize this with the so-called $R^2$ (R-squared) defined as $R^2 = (SS_T - RSS)/SS_T$. If $R^2$ is near 0, $y$ and $x$ do not have a strong linear relationship, whereas $R^2$ indicate that $y$ and $x$ are closely related in a linear sense. (It doesn't prove that the linear model is true or false, you have to check the residual plots to gauge the model adequacy.)

## 2 Variance Decomposition

If $y$ and $x$ are linearly unrelated, the true $\beta_1 = 0$ and so the marginal and conditional variance are almost equal. If $y = \beta_0 + \beta_1 x$ exactly, then the conditional variance $V(y|x) =:$ knowing $x$ explains everything about $y$.

We used the R-squared to quantify this, where

$$R^2 = \frac{SS_T - RSS}{SS_T} = \text{ the \% of variability in } y \text{ explained by the regression.}$$

Let's take a closer look at the numerator of the R-squared:

$$SS_T - RSS = \sum_i (y_i - \bar{y})^2 - \sum_i (y_i - \hat{y}_i)^2 \underset{\text{expanding the squares}}{=} \sum_i y_i^2 + \sum_i \bar{y}^2 - 2\bar{y}\sum_i y_i - \sum_i y_i^2 - \sum_i \hat{y}_i^2 + 2\sum_i y_i \hat{y}_i =$$

$$= -n\bar{y}^2 - \sum_i \hat{y}_i^2 + 2\sum_i \hat{y}_i^2 + 2\sum_i e_i \hat{y}_i \underset{e_i \text{ uncorrelated } \hat{y}_i}{=} \sum_i \hat{y}_i^2 - n\bar{y}^2 = \sum_i (\hat{y}_i - \bar{y})^2$$

We define this last expression as $SS_{reg} = \sum_i (\hat{y}_i - \bar{y})^2$, or the *regression sum of squares*. This is the *spread among the fitted values around the horizontal line at* $\bar{y}$. In Figure 1 the variance decomposition

is illustrated. We have concluded that the total variance is made up of two parts: the residual sum of

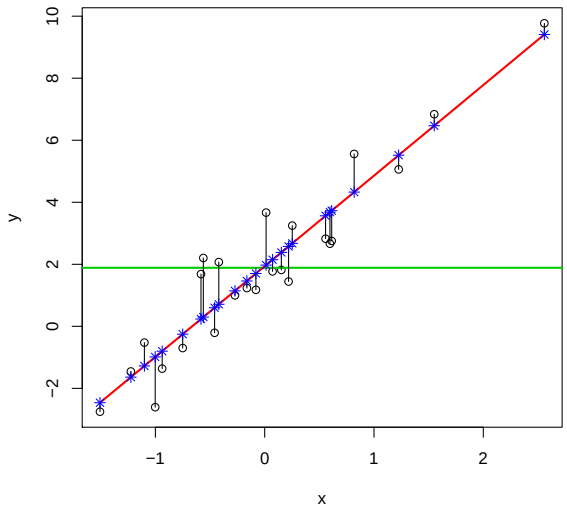


Figure 1: Variance decomposition. $SS_T$ is the spread among the black circles (data) around the mean of $y$ (green line). $RSS$ is the spread of the data around the regression line (red line). $SS_{reg}$ is the spread among the fitted values (blue stars) around the mean of $y$ (green line).

squared (deviations around the regression line) and the regression sum of squares (deviations on the line around the mean of $y$),

$$SS_T = RSS + SS_{reg}.$$

This means we have an alternative definition of $R^2 = SS_{reg}/SS_T$, the percent of total variability accounted for by the model.

Now, $R^2$ can be used to gauge if the regression model is "helpful", but it is not a statistical test. In order to decide how large is large enough for $R^2$ to state that the $y - x$ relationship is real we have to do some addition work. We will figure out how large a variant of this quantity can get just by chance when $y$ and $x$ are *unrelated*.

**Null hypothesis:** $\beta_1 = 0$. We will do all our work under this assumption. The test will be based on three different estimates for the noise level.

### Using the MSE to estimate $\sigma^2$

We already know that the MSE is an unbiased estimate of the $\sigma^2$. This is true whether the true $\beta_1 = 0$ or not. We estimate $\hat{\beta}_1$ and compute the residuals and the MSE and obtain

$$\hat{\sigma^2} = MSE = \frac{RSS}{n-2}$$

### Using the total sum of squares, $SS_T$

If $\beta_1 = 0$, then $y_i = \beta_0 + \epsilon_i$, with $V(\epsilon_i) = \sigma^2$. The total sum of squares $SS_T = \sum_i (y_i - \bar{y})$ can now also provide an estimate of $\sigma^2$ since $E[SS_T] = (n-1)\sigma^2$ (do the math for this at home).
That is,

$$\hat{\sigma^2} \underset{\text{if } \beta_1 = 0}{=} \frac{SS_T}{n-1}$$

**Using the regression sum of squares, $SS_{reg}$**

If the null is true, $\beta_1 = 0$, then we can get a third estimate of $\sigma^2$ from the regression sum of squares. Below, all expectations are under the null:

$$E[SS_{reg}] = E\left[\sum_i (\hat{y}_i - \bar{y})^2\right] = E\left[\sum_i (\hat{y}_i^2 + \bar{y}^2 - 2\bar{y}\hat{y}_i)\right] = E\left[\sum_i \hat{y}_i^2\right] + nE[\bar{y}^2] - 2E\left[\bar{y}\sum_i \hat{y}_i\right]$$

Using the fact that $V(Z) = E(Z^2) - (E[Z])^2, Cov(Z,W) = E(ZW) - E(Z)E(W)$ and that $E[y_i] = E[\hat{y}_i] = \beta_0$ under the null we get

$$E[SS_{reg}] = \sum_i \left(V[\hat{y}_i] + E[\hat{y}_i]^2\right) + n\left(V[\bar{y}] + E[\bar{y}]^2\right) - 2\sum_i \left(Cov(\bar{y}, \hat{y}_i) + E[\bar{y}]E[\hat{y}_i]\right) =$$

$$= \sigma^2 \sum_i h_{ii} + n\beta_0^2 + n\frac{\sigma^2}{n} + n\beta_0^2 - 2\sum_i \left(\sum_j Cov\left(\frac{y_j}{n}, \hat{y}_i = \sum_l h_{il} y_l\right)\right) - 2n\beta_0^2 =$$

$$= \sigma^2 \sum_i h_{ii} + \sigma^2 - 2\sum_i (\sum_j \frac{\sigma^2}{n} h_{ij}) = \sigma^2 \sum_i h_{ii} + \sigma^2 - 2\frac{\sigma^2}{n} \sum_i \sum_{j \atop =1} h_{ij} =$$

$$= \sigma^2 \sum_i h_{ii} + \sigma^2 - 2\sigma^2 = \sigma^2 (\sum_i h_{ii} - 1) = \sigma^2$$

(Puh!) This followed from $\sum_i h_{ii} = 2$. We have thus concluded that another estimate for $\sigma^2$, provided that the null is true, $\beta_1 = 0$ is

$$\hat{\sigma^2} \underset{\text{if } \beta_1 = 0}{=} SS_{reg}$$

(Looking ahead to multivariate regression when you have $p$ independent variables (one intercept and $p$ slope parameters), $E[SS_{reg}] = \sigma^2 p$.)

## 2.1 The F Goodness-of-fit test

If the null is true, $\beta_1 = 0$, $y$ and $x$ are not related and we have three different estimate for the $\sigma^2$ as seen above. If, on the other hand, the null is not true, both the $SS_T$-based and the $SS_{reg}$-based estimates will be inflated compared with $RSS/(n-2)$. We choose to use the $RSS$ and $SS_{reg}$ to test the null by looking at the ratio $SS_{reg}/RSS$. The reason for this is that the $RSS$ are functions of the residuals, whereas the $SS_{reg}$ is a function of the fitted values $\hat{y}$. Now, we know from before that the residuals, $e$, and the fitted values, $\hat{y}$, are uncorrelated. This will make it easier to work out a distribution for the ratio of the $SS_{reg}$ and $RSS$.

We now make an additional assumption, that $\epsilon$ is *normally distributed*.

- If $\epsilon \sim N(0, \sigma^2)$, $SS_{reg}/\sigma^2 \sim \chi_1^2$. (If we have $p$ model parameters, one intercept and $p-1$ slope parameters the distribution will be $\chi_{p-1}^2$).
- We also have that $RSS/\sigma^2 \sim \chi_{n-2}^2$. (If we have $p$ model parameters, it's $\chi_{n-p}^2$.)
- $\frac{SS_{reg}}{\sigma^2}$ and $\frac{RSS}{\sigma^2}$ are independent.
- Definition of an F-distribution: $F_{\nu_1,\nu_2} \equiv \frac{\chi_{\nu_1}^2}{\chi_{\nu_2}^2}$

Finally, we arrive at our test statistic:

$$F_{observed} = \frac{\left(\frac{SS_{reg}}{p-1}\right)}{\left(\frac{RSS}{n-p}\right)}$$

where $p$ is the number of model parameters, here 2 (intercept $\beta_0$ and slope $\beta_1$). We can also write this as

$$F_{observed} = \frac{\left(\frac{SS_T - RSS}{(n-1)-(n-p)}\right)}{\left(\frac{RSS}{n-p}\right)} = \frac{\left(\frac{\text{reduction in spread from mean-model to regression model}}{\text{number of extra parameters in regression model}}\right)}{\text{Mean Squared Error of regression model}}$$

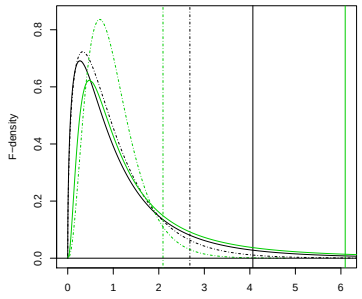


Figure 2: 4 different F-distributions and their 95th percentiles (vertical lines). The black curves correspond to models with $p = 4$ parameters and sample sizes $n = 12$ and $n = 120$ respectively (solid and dashed). The F-distributions are $F_{3,8}$ and $F_{3,116}$. The green curves correspond to $p = 8$ and sample sizes $n = 12$ and $n = 120$ (solid and dashed). The F-distributions are $F_{7,4}$ and $F_{7,112}$. The critical value of the Goodness-of-fit test is decreases with sample size (comparing solid to dashed), and increases with model parameters (comparing green with black).

In Figure 2 you see 4 different F-distributions that would be appropriate to test the goodness-of-fit, or null hypothesis $\beta_1 = 0$, under the scenarios (i) $p = 4, n = 12$, (ii) $p = 4, n = 120$, (iii) $p = 8, n = 12$ and (iv) $p = 8, n = 120$. As you can see from the figure the critical vales ($1 - \alpha$, $\alpha = .05$), 95th percentiles, are smaller for larger sample sizes (meaning we can reject the null with a smaller observed $F$ with a large sample data) and larger for models involving more parameters (we need a larger observed $F$ to reject the null for more complex models).
We can of course also choose to use P-values rather than fixed level test. You can compute the probability $P(F_{3,8} > F_{observed})$, which would tell you the likelihood of a data set where the null is true would generate an $F$-value as extreme or more extreme than the one we observed. If this P-value is small, we reject the null hypothesis that $\beta_1 = 0$.

Call: lm(formula = y x)
Residuals: Min 1Q Median 3Q Max -1.6975 -0.5843 -0.1996 0.5022 2.2361
Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) 1.7157 0.1944 8.825 7.66e-09 *** x 3.0301 0.1827 16.584 2.74e-14 *** — Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.9673 on 23 degrees of freedom Multiple R-squared: 0.9228, Adjusted R-squared: 0.9195 F-statistic: 275 on 1 and 23 DF, p-value: 2.744e-14

| value | df1 | df2 |
|---|---|---|
| 275.018 | 1.000 | 23.000 |

Table 1: F-statistic

In the regression summary output below we see that $F_{observed} = 275.018$. In Table 1 the F value observed and the corresponding degrees of freedom are also recorded in the table. The P-value is $p = 2.742e - 14$.

# 3 Inference about the slope, $\beta_1$

Again, we assume that the errors are normally distributed, $\epsilon \sim N(0, \sigma^2)$. It follows that $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ and since $\hat{\beta}_1 = \sum_i k_i y_i$ is a linear combination of $y$-values and a linear combination of normally distributed random variables is normal, we have

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{\sum_j (x_j - \bar{x})^2}).$$

Another way of writing this is

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{\sum_j (x_j - \bar{x})^2}}} \sim N(0, 1).$$

There is one problem here. The above is indeed the sampling distribution of $\hat{\beta}_1$ BUT we don't know $\sigma^2$. Is it OK to plug in $\hat{\sigma^2} = RSS/(n-p)$ in the above sampling distribution expression?

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma^2}}{\sum_j (x_j - \bar{x})^2}}} \neq N(0, 1)$$

In the above expression there are two random quantities: $\hat{\beta}_1$ and $\hat{\sigma^2}$ and so the ratio can vary a bit more than the standard normal $N(0,1)$ describes (the ratio has longer tails). What we have instead is

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{\sum_j (x_j - \bar{x})^2}),\ RSS/\sigma^2 \sim \chi^2_{n-p},\ \hat{\sigma^2} = RSS/(n-p)$$

- Definition of a $t$-distribution is the ratio of an independent Normal distributed variable and a $\chi^2$ distributed variable
- Our test statistics $t_{observed} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma^2}}{\sum_j (x_j - \bar{x})^2}}} \sim t_{n-p}$

## 3.1 Hypothesis testing

We fit a model to the data. We formulate the null hypothesis $\beta_1 = 0$ and compute the test statistic under the null:

$$t_{observed} = \frac{\hat{\beta}_1 - 0}{\sqrt{\frac{\hat{\sigma^2}}{\sum_j (x_j - \bar{x})^2}}}$$

We compare $t_{observed}$ to quantiles of the $t$-distribution $t_{n-p}$.
If the absolute value of $t_{observed}$ exceeds the critical value $t_{n-p}(1 - \alpha/2)$ (the $1 - \alpha/2$ percentile, e.g. the 97.5 percentile) we reject the null hypothesis $\beta_1 = 0$ at the level $\alpha$ (two-sided test).

Alternatively, we can compute the P-value. Compute the probability mass of the $t_{n-p}$-distribution for absolute values exceeding $|t_{observed}|$.

In Figure 3 we see two different t-distributions. As you can see, the larger the sample size $(n-p)$ the smaller the critical value so it's "easier" to reject a null hypothesis with more data.

## 3.2 Confidence intervals

There is a duality between hypothesis testing and confidence intervals. We know that

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma^2} \sum_j (x_j - \bar{x})^2}} \sim t_{n-p}.$$

We can write

$$P\left( t_{n-p}(\alpha/2) \leq \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma^2}}{\sum_j (x_j - \bar{x})^2}}} \leq t_{n-p}(1 - \alpha/2) \right) = 1 - \alpha,$$

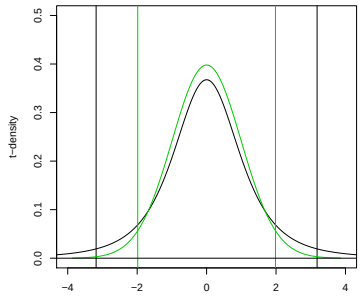


Figure 3: 2 different t-distributions and their 97.5th and 2.5th percentiles (vertical lines). The black curves correspond to models with $n-p=3$ and the green curves correspond to $n-p=100$.

where $t_{n-p}(q)$ denotes the $q$-th quantile of the $t_{n-p}$-distribution. We manipulate the above expression as

$$P\left(\hat{\beta}_1 - t_{n-p}(1-\alpha/2)\sqrt{\frac{\hat{\sigma^2}}{\sum_j (x_j-\bar{x})^2}} \leq \beta_1 \leq \hat{\beta}_1 + t_{n-p}(\alpha/2)\sqrt{\frac{\hat{\sigma^2}}{\sum_j (x_j-\bar{x})^2}}\right) = 1-\alpha.$$

That is, denoting $SE(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma^2}}{\sum_j (x_j-\bar{x})^2}}$, we see that *the random interval*

$$[\hat{\beta}_1 \pm t_{n-p}(1-\alpha/2)SE(\hat{\beta}_1)]$$

covers the true $\beta_1$ with probability $1-\alpha$. We can now form any hypothesis $\beta_1 = \beta_1^*$. If the above random interval does not cover $\beta_1^*$ we can reject this hypothesis at level $\alpha$. We usually use this to test hypothesis $\beta_1 = 0$ of course.

# 4 Prediction Intervals

We can also construct confidence intervals and test for the regression line, fitted values and predictions. We know from before that

$$\hat{y}_i \longrightarrow E[\hat{y}_i] = \beta_0 + \beta_1 x_i,\ V[\hat{y}_i] = \sigma^2 h_{ii}$$

Following the same line of thought as in setting up confidence intervals for $\beta_1$ we find that at each location $x_i$ the random interval

$$[\hat{y}_i \pm t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{h_{ii}}]$$

covers the true line with probability $1-\alpha$. Note that the width of this interval is non-constant, it depends on the location $x_i$ through the leverage. This makes sense, we have already seen that the regression line pivots around the center of mass of the data and can differ substantially at points of high leverage subject to small changes in the data. This is illustrated in Figure 4 for a data set of size $n=25$.

Things look a bit different when we use the regression model for prediction. Let's say we want to predict the outcome $y^{new}$ at location $x^{new}$. Now, the prediction itself we obtain from the regression line: $\hat{y}^{new} = \hat{\beta}_0 + \hat{\beta}_1 x^{new}$, but what about the prediction variance.
The prediction has two sources of errors associated with it. If we *knew* the true regression model we would estimate $y^{new}$ by $\beta_0 + \beta_1 x^{new}$ and our prediction error would be $\sigma$, the standard deviation of the random scatter about the true model. Here, we don't know the true model so our prediction inherits the estimation variance from $\hat{\beta}_0, \hat{\beta}_1$ as well. The prediction has thus the following properties:

$$\hat{y}^{new} \longrightarrow E[\hat{y}^{new}] = \beta_0 + \beta_1 x^{new},\ V[\hat{y}^{new}] = \sigma^2(1+h(x^{new})).$$

Since we have to estimate $\sigma^2$ the t-distribution $t_{n-p}$ has to be used to construct the prediction interval. In Figure 5 I illustrate the difference between a prediction interval and a confidence interval for the line.

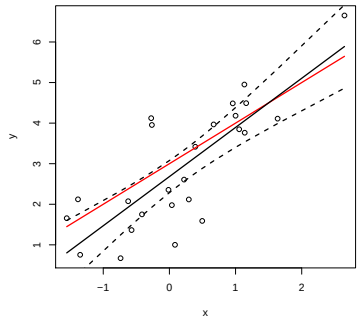


Figure 4: The confidence interval for $\hat{y}$ (the regression line). Notice the width of the interval depends on the leverage. The true line is marked in red, the estimated on in black.

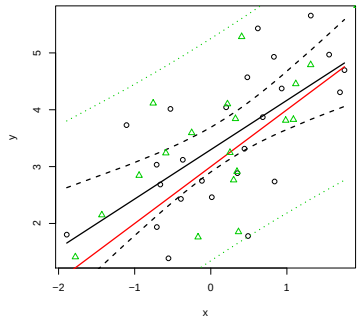


Figure 5: The prediction interval (green dotted lines) and the confidence interval for the line (black dashed lines). A set of new observations are drawn from the true model (red line) and depicted as green triangles. Are these covered by the prediction interval?

The distinction between the confidence interval and prediction interval can be remembered like this. The confidence interval is used to draw inference about the average outcome at a location $x$, whereas the prediction interval is used to draw inference about a single, new occurrence at a location.

# 5 Demo 4

We continue with the television data.

```
> 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)
```

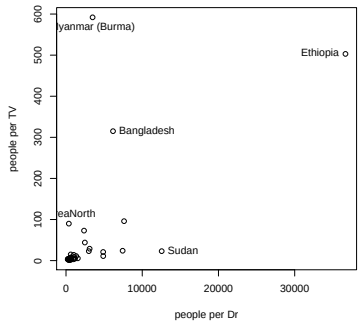


Figure 6: People per TV vs People per Dr

```
> 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])])
```

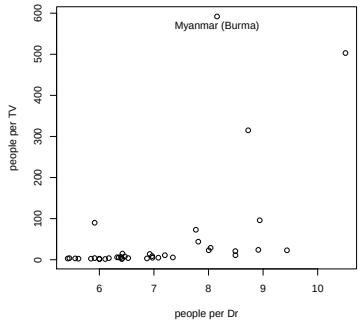


Figure 7: People per TV vs People per Dr: logs on ppDr to even out the spread in $x$

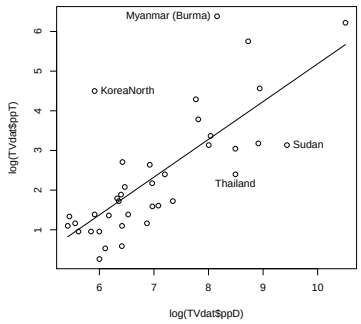


Figure 8: People per TV vs People per Dr: logs on ppTV to suppress non-constant variance. Regression line

```
> 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 2: Regression summary

## 5.1 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)
```

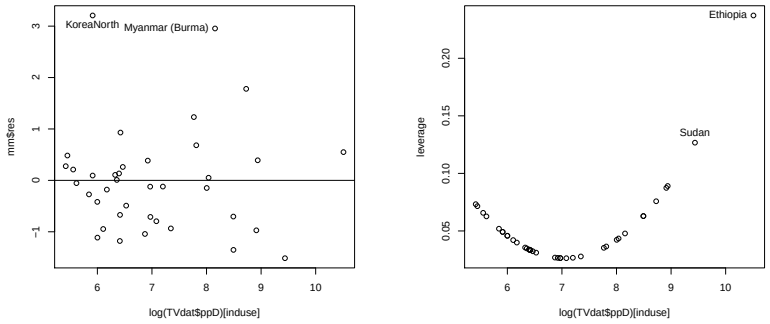


Figure 9: Residual and Leverage plot for the Television regression model

```
> 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)
```

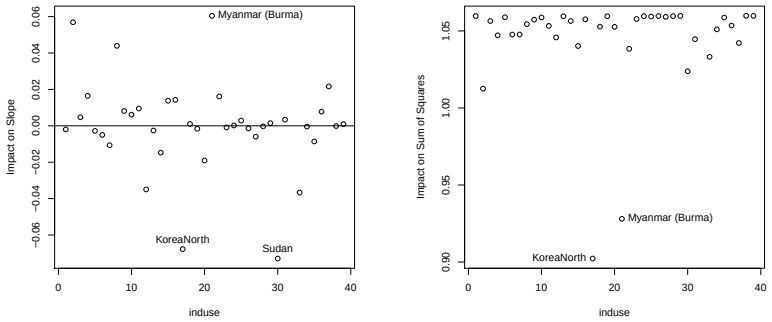


Figure 10: Impact on Slope (left) and Residual sum of squares (right) when dropping observation $i$

```
> print(summary(mm))

Call:
lm(formula = log(TVdat$ppT) ~ log(TVdat$ppD))

Residuals:
    Min      1Q  Median      3Q     Max
-1.5139 -0.7092 -0.0871  0.3575  3.2077

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     -4.3417     0.9933  -4.371 0.000101 ***
log(TVdat$ppD)   0.9527     0.1388   6.864 4.95e-08 ***
---
Signif. codes:  0 Ś***Š 0.001 Ś**Š 0.01 Ś*Š 0.05 Ś.Š 0.1 Ś Š 1

Residual standard error: 1.045 on 36 degrees of freedom
  (2 observations deleted due to missingness)
```

```
Multiple R-squared: 0.5669,        Adjusted R-squared: 0.5549
F-statistic: 47.12 on 1 and 36 DF,  p-value: 4.949e-08
```

The $R^2$ for the model regression people per TV on people per Dr is $R^2 = 0.57$. That means that $100 * 0.57$ percent of the variability in people per TV is explained by people per Dr.

## 5.2 Testing and Confidence intervals

Looking at the regression model summary above we see that the slope estimate $\hat{\beta}_1 = 0.95$ with standard error $SE(\hat{\beta}_1) = 0.14$. From this we can construct confidence intervals as follows:

```
> library(xtable)
> z <- data.frame(matrix(confint(mm), 2, 2))
> dimnames(z) <- list(c("intercept", "slope"), c("2.5%", "97.5%"))
> xtable(z, digits = c(0, 3, 3), caption = "95 percent confidence intervals",
+     label = "tab:ci")
```

| | 2.5% | 97.5% |
|---|---|---|
| intercept | -6.356 | -2.327 |
| slope | 0.671 | 1.234 |

Table 3: 95 percent confidence intervals

```
> p <- locator()
```

In Table 3 we see that the confidence interval for the slope does not cover 0 so we reject this hypothesis at the 5% level.

```
> z <- data.frame(matrix(ms$fstat, 1, 3))
> dimnames(z) <- list(c(""), c("value", "df1", "df2"))
> xtable(z, digits = c(0, 3, 3, 3), caption = "F-statistic", label = "tab:fsumTV")
```

| value | df1 | df2 |
|---|---|---|
| 47.119 | 1.000 | 36.000 |

Table 4: F-statistic

```
> pval <- 1 - pf(ms$fs[1], ms$fs[2], ms$fs[3])
> ord <- 3 + abs(floor(log10(pval)))
> p <- locator()
```

From the regression summary we also see that the $F_{observed} = 47.119$. In Table 4 the F value observed and the corresponding degrees of freedom are also recorded in the table. The P-value is $p = 4.949e - 08$.

We can thus reject the hypothesis that people per TV and people per doctor are unrelated as the association is highly significant.

## 5.3 Prediction intervals

The following code withholds 10 countries from the regression analysis and checks to see if the prediction intervals will cover their people per TV values.

```
> ii <- sample(seq(1, dim(TVdat)[1]), 10)
> x <- sort(log(TVdat$ppD[-ii]))
> y <- log(TVdat$ppT[-ii])[sort.list(log(TVdat$ppD[-ii]))]
> plot(x, y, xlab = "log-ppDr", ylab = "log-ppTV")
> mm <- lm(y ~ x)
> lines(x[is.na(y) == F], mm$fitted, lwd = 2)
> xp <- seq(min(log(TVdat$ppD)), max(log(TVdat$ppD)), by = 0.1)
> xuse <- x[is.na(y) == F]
> hats <- 1/length(xuse) + (xp - mean(xuse))^2/sum((xuse - mean(xuse))^2)
> lines(xp, mm$coef[1] + mm$coef[2] * xp + qt(0.975, length(xuse) -
+     2) * sqrt(ms$sigma^2 * hats), lwd = 2, lty = 2)
> lines(xp, mm$coef[1] + mm$coef[2] * xp - qt(0.975, length(xuse) -
+     2) * sqrt(ms$sigma^2 * hats), lwd = 2, lty = 2)
> lines(xp, mm$coef[1] + mm$coef[2] * xp + qt(0.975, length(xuse) -
+     2) * sqrt(ms$sigma^2 * (1 + hats)), lwd = 2, col = 3, lty = 3)
> lines(xp, mm$coef[1] + mm$coef[2] * xp - qt(0.975, length(xuse) -
+     2) * sqrt(ms$sigma^2 * (1 + hats)), lwd = 2, col = 3, lty = 3)
> points(log(TVdat$ppD)[ii], log(TVdat$ppT)[ii], col = 3, pch = 2)
> points(log(TVdat$ppD)[ii], mm$coef[1] + mm$coef[2] * log(TVdat$ppD)[ii],
+     col = 3, pch = 8)
> id <- identify(log(TVdat$ppD)[ii], log(TVdat$ppT)[ii], row.names(TVdat)[ii],
+     pos = T)
```

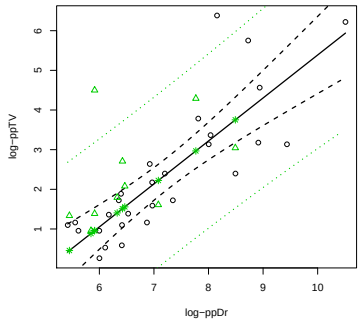


Figure 11: The regression line and confidence interval for 30 countries. 10 countries (green triangles) were not used as part of the estimation. Their prediction are shown as green asterisks and the prediction interval as green dotted lines.

In Figure 11 I show the estimated regression line and confidence interval for a subset of 30 countries (black lines). I withheld 10 countries to test the model's prediction capacity. The green triangles are the true observations for these 10 countries, whereas the green asterisks are the model predictions. The prediction interval (green dotted lines) does a good job in covering the true values.