MSG500/MVE190
Linear Models - Lecture 3

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

November 2, 2012

1 RECAP

- Both the slope estimate and the fitted values are weighted averages of the dependent variable observations y_i
- The weights are large (in absolute value) for extreme values of x_i, so the extremes in x are the ones that determine which y_i influence the model fit
- The amount of influence an observation y_i has on its own fitted values is called *leverage* and is defined as $h_{ii} = \frac{(x_i - \bar{x})^2}{\sum_j (x_j - \bar{x})^2}$
- If x is near constant, $h_{ii} \simeq 1/n$ so the fitted values become just the mean of y (a slope of 0). If h_{ii} is very large, then $\hat{y}_i \simeq y_i$, i.e. the fitted value is almost the observation itself - we haven't learnt anything from the pattern among the rest of the observations.
- There are three sources of error in the slope estimation: the noise level, the sample size and the spread in x.
- The residuals $e_i = y_i - \hat{y}_i$ are our primary diagnostic tool. Residual plots of e vs x or e vs $\hat{y}$ can be used to check the basic assumptions and detect outliers.
- We can detect outliers using diagnostic plots like the impact on the slope or on the residual sum of squares in response to dropping an observation i.

2 Properties of Least Squares estimates

We have already discussed the properties of the slope estimate (unbiased and the three-sources of error of the variance). What about the properties of the fitted values and the residuals. We have already determined that both the residuals and the fitted values are uncorrelated with the x (orthogonal projection). What about the mean and variance?

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \rightarrow E[\hat{y}i] \underset{\text{since estimates of } \beta \text{ are unbiased}}{=} E[y_i] = \beta_0 + \beta_1 x_i$$

For the variance of the fitted values (or of the regression line itself) we have

$$V[\hat{y}_i] = V[\hat{\beta}_0 + \hat{\beta}_1 x_i] = V[\bar{y} + \hat{\beta}_1(x_i - \bar{x})] \underset{\hat{y} \text{ and } \hat{\beta}_1 \text{ uncorrelated}}{=} V(\bar{y}) + (x_i - \bar{x})^2 V(\hat{\beta}_1)$$

The expression for $V[\hat{y}]$ simplified because $\bar{y}$ and $\hat{\beta}_1$ are uncorrelated, which we can show as follows:

$$Cov(\bar{y}, \hat{\beta}_1) = Cov(\frac{1}{n}\sum y_i, \sum_i k_i y_i) \underset{y_i, y_j \text{ uncorrelated}}{=} \sum_i \frac{k_i}{n} V(y_i) = \frac{\sigma^2}{n} \sum_i k_i = 0.$$

Therefore, we can write

$$V[\hat{y}_i] = V(\bar{y}) + (x_i - \bar{x})^2 V(\hat{\beta}_1) = \frac{\sigma^2}{n} + \sigma^2 \frac{(x_i - \bar{x})^2}{\sum_j (x_j - \bar{x})^2} = h_{ii}\sigma^2$$

So, the variance of the fitted value is directly proportional to the leverage!!! This makes sense. Points with high leverage have a lot of influence on the fit so of course their fitted value will vary a lot from instance to instance since the fit will react strongly to any small random change at this location. In Figure 1 we see the impact on leverage on the variance of the fitted values. Near the center of the data (mean of x) the different regression line estimates (from 50 different data sets generated from the same underlying true model (solid black line)) vary almost not at all, whereas at locations of extreme x (high leverage) the position of the regression line can vary substantially.

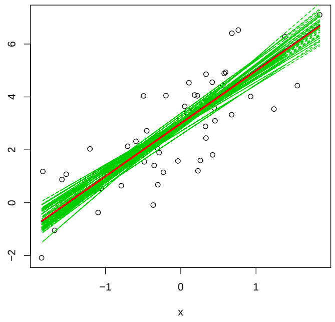

Figure 1: 50 different regression line estimates (green line) estimated from data generated under the true model (black line). The impact of the leverage on the variance of the fitted values. At points with extreme leverage (extreme x) there is a lot of variability in the fitted value (green regression lines) whereas low leverage observations (near the mean of x) exhibit very low variance for the fitted values.

Residuals

The definition of the residuals is $e_i = y_i - \hat{y}_i$. From the properties of the fitted values it follows that $E[e_i] = 0$. Moreover

$$V[e_i] = Cov(y_i - \hat{y}_i, y_i - \hat{y}_i) = Cov(y_i, y_i) + Cov(\hat{y}_i, \hat{y}_i) - 2Cov(y_i, \hat{y}_i) = \sigma^2 + \sigma^2 h_{ii} - 2Cov(\hat{y}_i + e_i, \hat{y}_i)$$

Using that the residuals are uncorrelated with the fitted values we finally obtain

$$V[e_i] = \sigma^2(1 - h_{ii})$$

This means that the residual variance is *smaller* at locations with high leverage. This makes sense as at those locations the regression line is pulled toward the data and so there is more control over the residual variability.

Another interesting fact is that $Cov(e_i, e_j) = -\sigma^2 h_{ij}$. That is, even though we assume that the random errors (ϵ) are uncorrelated, the residuals from the fit are *not* uncorrelated. That makes sense as well. The fit is obtained by allowing observations y_j influence the fit for y_i (that's what h_{ij}) measures remember). In Table 1 you can compare the properties of the true random scatter and the residuals.

True errors, ϵ_i	Residuals from fit, e_i
$E[\epsilon_i] = 0$	$E[e_i] = 0$
$V[\epsilon_i] = \sigma^2$	$V[e_i] = \sigma^2(1 - h_{ii})$
$Cov(\epsilon_i, \epsilon_j) = 0$	$Cov(e_i, e_j) = -\sigma^2 h_{ij}$

Table 1: Some facts about errors and residuals

One thing to be especially aware of is the non-constant variance of the residuals. That means that you can't really compare residuals directly - what constitute as large value at one location may not be so surprisingly large at another location. It is therefore often better to plot the *standardized residuals* defined as

$$\text{standardized residuals } = \tilde{e}_i = \frac{e_i}{\sqrt{1-h_{ii}}}$$

The standardized residuals all have the same variance, σ^2 are are thus directly comparable in diagnostic plots (see Figure 2).

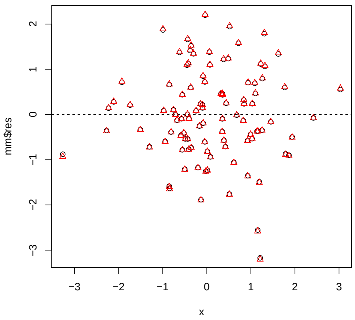

Figure 2: The residuals and standardized residuals (red triangles) from a regression fit

3 The nuisance parameter, σ^2

So far we have discussed the slope and intercept parameter - these are of primary interest to us to interpret the data. However, there is one more parameter in our model, the noise level σ^2. This parameter is not of direct interest to us, but we will need to estimate it in order to draw inference. σ^2 is therefore referred to as a nuisance parameter.

We have $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where $V(\epsilon_i) = \sigma^2$. σ^2 is thus the variance of the true random scatter. We can't get the true errors, but via least squares fitting we can get the residuals. Now, if we had known the true errors we could have estimated the σ^2 as

$$\hat{\sigma^2} = \frac{\sum_{i=1}^n \epsilon_i - \bar{\epsilon}}{n-1}.$$

However, we don't know ϵ so we estimate σ^2 from

$$\hat{\sigma^2} = MSE = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^n e_i^2}{n-2} = \frac{RSS}{n-2}$$

where RSS refers to the *Residual Sum of Squares* and MSE is the *Mean Squared Error*. Notice that we are using $n-2$ in the expression instead of $n-1$ as in the standard variance estimation. That is because we needed to estimate two parameter to get the residuals - both the intercept and the slope. Had we known the errors, ϵ we only needed to estimate their mean $\bar{\epsilon}$.

The MSE is a key quantity in regression analysis. Let's first check that the MSE is an estimate of σ^2.

$$E[RSS] = E[\sum_i e_i^2] = E[\sum_i (y_i - \hat{y}_i)^2] = \sum_i (E[y_i^2] + E[\hat{y}_i^2] - 2E[y_i \hat{y}_i])$$

I will now use that $V[Z] = E[Z^2] - (E[Z])^2$ for a random variable Z and rewrite the above expression as

$$E[RSS] = \sum_i V[y_i] + (E[y_i])^2 + V[\hat{y}_i] + (E[\hat{y}_i])^2 - 2E[(\hat{y}_i + e_i)\hat{y}_i]$$

Using that $Cov(Z, W) = E[ZW] - E[Z]E[W]$ we can write

$$E[RSS] = \sum_i V[y_i] + V[\hat{y}_i] + 2(E[y_i])^2 - 2Cov(\hat{y}_i + e_i, \hat{y}_i) - 2E[y_i]E[\hat{y}_i]$$

Now, $Cov(e_i, \hat{y}_i) = 0$ since residuals and fitted values are uncorrelated and so we have

$$E[RSS] = \sum_i V[y_i] + V[\hat{y}_i] - 2Cov(\hat{y}_i, \hat{y}_i) = \sum_i V[y_i] - V[\hat{y}_i] = n\sigma^2 - \sigma^2 \sum_i h_{ii}$$

Finally, we can show that $\sum_i h_{ii} = \sum_i (\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_j (x_j - \bar{x})^2}) = 2$ and so

$$E[RSS] = \sigma^2(n-2), \quad E[MSE] = \sigma^2$$

4 Variance Decomposition

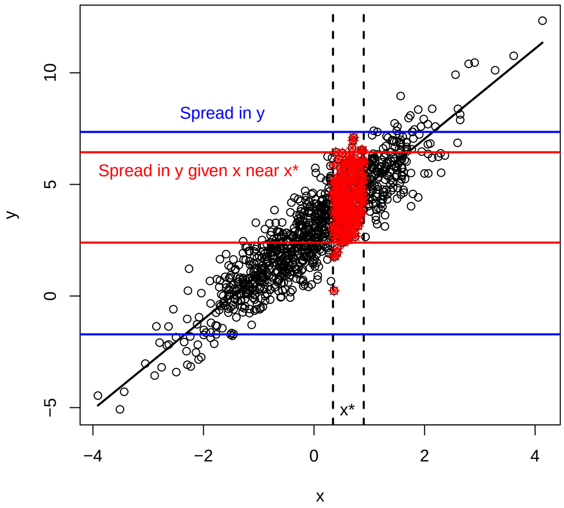

Figure 3: Variance Decomposition. y and x closely related. The blue lines delimit the spread in y, whereas the red lines delimit the spread among y-values whose corresponding x values are close to x^*.

Regression is all about modeling y *given* x. That is, without x our best guess for a future value of y is the average of the observed y-values: $\hat{y}_i = \bar{y}$. This is what we get from regression modeling if all the x's are constant (leverage is equal to $1/n$ for all observations). y varies around its mean according to $SD(y) = \sqrt{V(y)}$. We now bring x into the equation. If we know that the x-value for the observation is x^P, can we improve on the guess $\bar{y}$ for the corresponding y-value. Regression modeling gives us the estimate $y(\hat{x^*}) = \hat{\beta}_0 + \hat{\beta}_1 x^*$. The spread among the y-values for those y with x near x^* we denote by $V(y|x = x^*)$. If y and x are strongly dependent on each other, the spread $V(y|x = x^*)$ (conditional variance) is much less than the total spread among y-values, $V(y)$ (marginal variance). If y and x are unrelated, the regression model doesn't add much information about y and so the conditional and marginal variance are almost the same. In Figure 3 I illustrate this discussion. The scatter plot of x and y values show the data set as a clould. The value x^* is marked between two vertical bars that identify the observations with x-values near x^*. These observations are marked in red. The blue horizontal lines are located 2.5% and 97.5% quantiles of all the y's. The red horizontal lines are the corresponding quantiles for the y's marked in red. As you can see, the marginal spread in y, illustrated by the distance between the two blue horizontal bars, far exceeds the conditional spread, illustrated by the distance between the red bars.

In Figure 4 I show you marginal and conditional spreads for a weak $y - x$ relationship data. Here, the spread in the vertical wedge (red observations) are almost the same as the marginal spread.

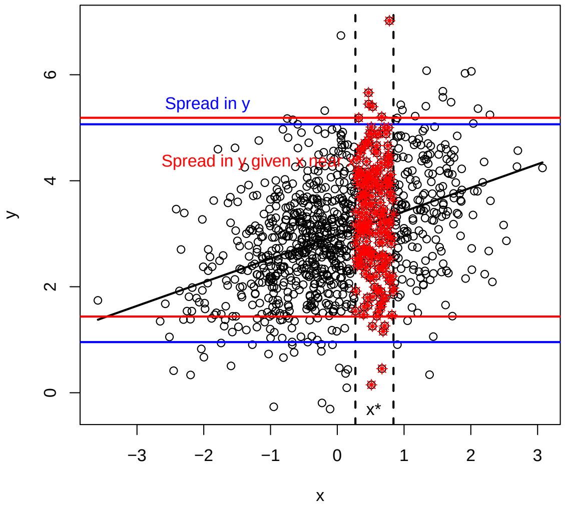

Figure 4: Variance Decomposition. y and x weakly related. The blue lines delimit the spread in y, whereas the red lines delimit the spread among y-values whose corresponding x values are close to x^*.

This variance decomposition is a useful tool for summarizing the efficacy of regression modeling: does x "help" describing y or can we just use the mean of y to summarize the data? We compare the following two quantities:

1. $RSS = SS_E$, the residual sum of squares, also called the error sum of squares. It's defined as $RSS = SS_E = \sum_i (y_i - \hat{y}_i)^2$
2. SS_T, the total sum of squares. It's defined as $SS_T = \sum_i (y_i - \bar{y})^2$

Note, since the regression model has two parameters to match to the data, the RSS is *always* smaller than the total sum of squares, SS_T.

4.1 The multiple R-squared, R^2

We summarize the regression fit with a quantity called the R-squared, or R^2, or sometimes the *coefficient of determination.*

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

In Figure 5 I plot three different data sets and their fitted regression line. The R^2-values are also shown in the plot. **Use the source code for the lecture to simulate some other data sets and compare the resulting R^2.**

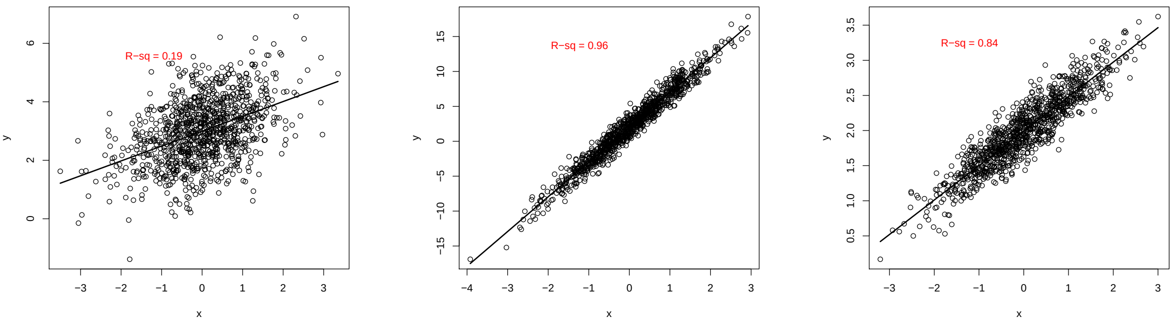

Figure 5: R-squared: R-squared values computed for three different regression fits to data.

5 Demo 3

We continue with the television data. 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)
```

```
> 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])])
```

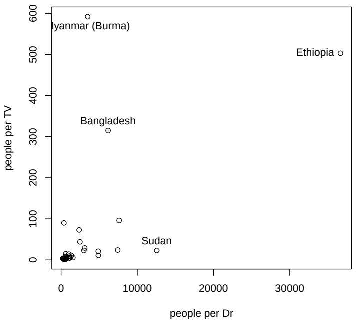

Figure 6: People per TV vs People per Dr

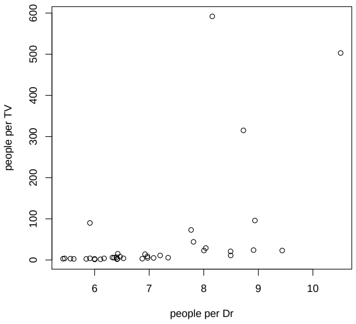

Figure 7: People per TV vs People per Dr: logs on ppDr to even out the spread in x

```
> 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)
```

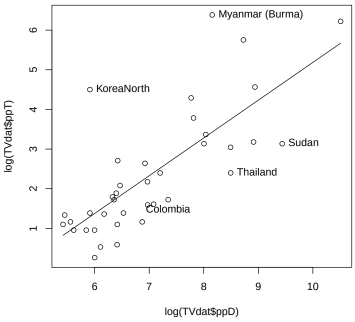

Figure 8: People per TV vs People per Dr: logs on ppTV to suppress non-constant variance. Regression line

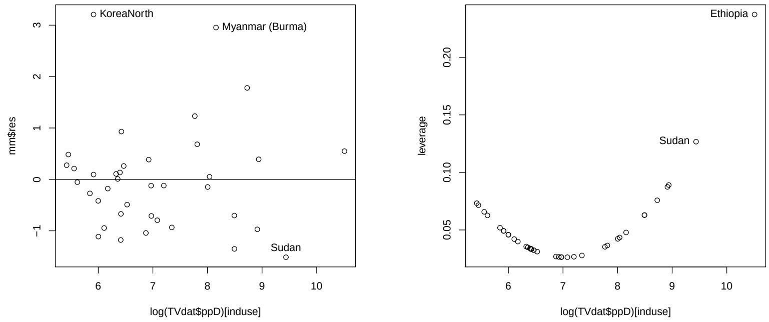

Figure 9: Residual and Leverage plot for the Television regression model

```
> 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)
```

```
> 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)
```

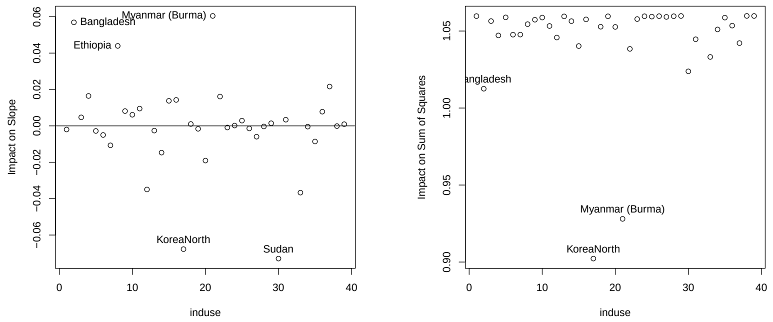

Figure 10: Impact on Slope (left) and Residual sum of squares (right) when dropping observation i

```
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.