Homework 4 - Statistical modelling and inference Group 7

November 9, 2015

Students	**Student IDs**
Max van Esso	73539
Marco Fayet	125593
Felix Gutmann	125604

1 Homework lecture 6

To show: Show that $y(\mu + \sigma)$ is 1 standard deviation away from the marginal distribution of t

Prerquisite information:

(a) $t = x + \epsilon$

(b) $x \sim \mathcal{N}(\mu, \sigma^2)$

(c) $\epsilon \sim \mathcal{N}(0, \tau^2)$

Solution:

From (a) we notice that the coefficient of x is 1 and we conclude:

$$\begin{aligned} y(x) &= \mathbb{E}(t|x) \\ &= x \end{aligned}$$

Subsequently we can assume the same distribution as x. In the next step we compute the mean and the variance of t:

$$\begin{aligned}
\mathbb{E}(t) &= \mathbb{E}(x+\epsilon) \\
&= \mathbb{E}(x) + \mathbb{E}(\epsilon) \\
&= \mathbb{E}(x) \\
&= \mu \\
\mathrm{Var}(t) &= \mathrm{Var}(x+\epsilon) \\
&= \mathrm{Var}(x) + \mathrm{Var}(\epsilon) + 2\mathrm{Cov}(x,\epsilon)
\end{aligned}$$

Under the Gauss/Markow assumptions we assume x and ϵ to not be correlated

$$\begin{aligned}
&= \mathrm{Var}(x) + \mathrm{Var}(\epsilon) \\
&= \sigma^2 + \tau^2
\end{aligned}$$

As a result the distribution of t has mean μ and standard deviation $\sigma^2+\tau^2$. After combining results the following holds:

$$\begin{aligned}
\sqrt{\sigma^2+\tau^2} &> \sigma \\
\sqrt{\sigma^2+\tau^2} + \mu &> \sigma + \mu
\end{aligned}$$

Since $y(x) = x$, we can replace $\sigma + \mu$ by $y(\sigma + \mu)$

$$\sqrt{\sigma^2+\tau^2} + \mu > y(\sigma + \mu)$$

Finally, by subtracting μ we see that the difference between μ and $y(\sigma+\mu)$ is less than the standard deviation of t

$$\sqrt{\sigma^2+\tau^2} > y(\sigma + \mu) - \mu$$

2 Exercises in R

2.1

a) Data preparation

Table one shows a brief summary of questionable values for each variable. In total we **excluded 850 rows** from the data set (working data set has 1181 observations).

Table 1: Regression data

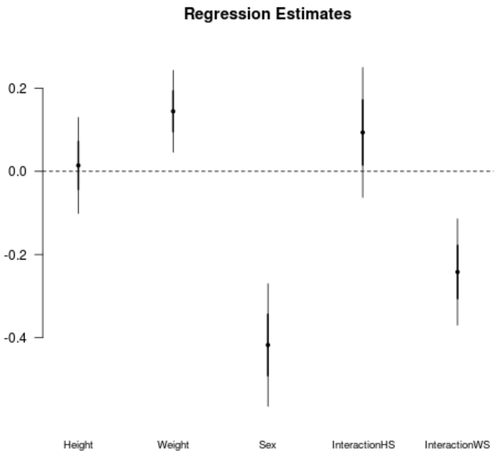
Variable	NA's	Outlier	Zero values
Earnings	651	1 ind. > \$400*k*	187
Height	0	8 ind. > 100 inches	0
Weight	0	44 ind. > 300 Pound	0

b) Model selection and result

Listing one shows the model we chose. We standarized height and weight to make the intercept interpretable. Furthermore, we take the log of earnings to increase interpretability. In addition we introduce interaction variables to reflect the varying effects of height and weight for each gender.

Listing 1: Chosen model

```
lm (log(earnings) ~ height.standardized + weight.standardized + sex + sex*height.standardized
        + sex*weight.standardized
```



c) Interpretation of coefficients

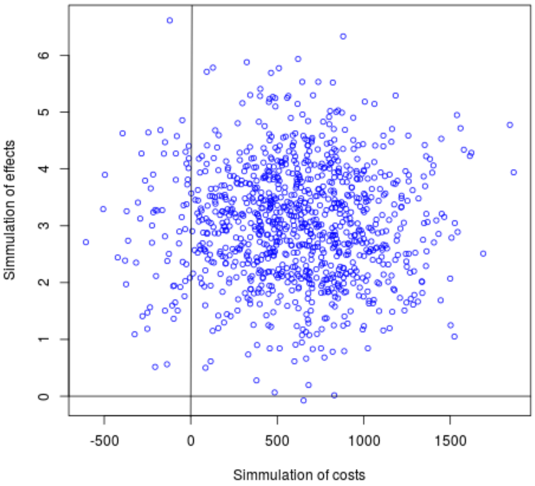
Interpretation:

- The intercept is the predicted log earning for males of average height and weight. Taking the exponent of this value gives us an average predicted earnings of $20,538 for this category.
- The coefficient for height_standardized (0.014) is the predicted difference in log earning corresponding to a 1 standard deviation difference in height for males of average weight.
- The coefficient for weight_standardized (0.14) is the predicted difference in log earning corresponding to a 1 standard deviation difference in weight for males of average height.
- The coefficient of sex (-0.42) corresponds to the predicted difference in log earnings between men and women of average height and weight.
- Given that the sex variable takes a value of 1 for females, the interactive terms add to the total effect of height and weight on earnings for women. We therefore get a measure of the change in predicted log earnings given a 1 standard deviation in height and weight keeping other predictors constant. These correspond to a total of -10% and 70% respectively.
- In conclusion, our model tells us that additional height adds a positive effect to predicted earnings for both genders. Weight will bring down predicted earnings for women but increase them for men. This is consistent with recent scientific literature (http://www.otago.ac.nz/news/news/otago192804.html)

2.2

(a)

The plot shows 1000 random draws of cost differences vs. cost effectiveness for the given parameters:



(b)

- We compute the ratio and the mean of that ratio as a point estimate
- We compute the 50% and 95% as the corresponding quantiles of our simulations

Example of a result:

Parameter	**Results**
Mean	246.15
Quantile (0.25,0.75)	(72.47,331.31)
Quantile (0.025,0.0.095)	(-91.87, 785.00)

(c)

Example of a result (standard error for effectiveness=2):

Parameter	**Results**
Mean	245.29
Quantile (0.25,0.75)	(107.71,323.91)
Quantile (0.025,0.0.095)	(-1314.23,1891.72)

2.3

Objective:

Construct a nonlinear prediction for congressional elections in the US following section 7.3 in Gelman.

First step:

Data from 1986 is used to predict the outcome of the elections in 1988. Results are then applied to predict the 1990 elections outcome.

Predictors of the regression model:

1. Constant term
2. Democratic share of votes in the previous election
3. Indicator to account for incumbency (i.e. whether the sitting candidate is running for re-election)

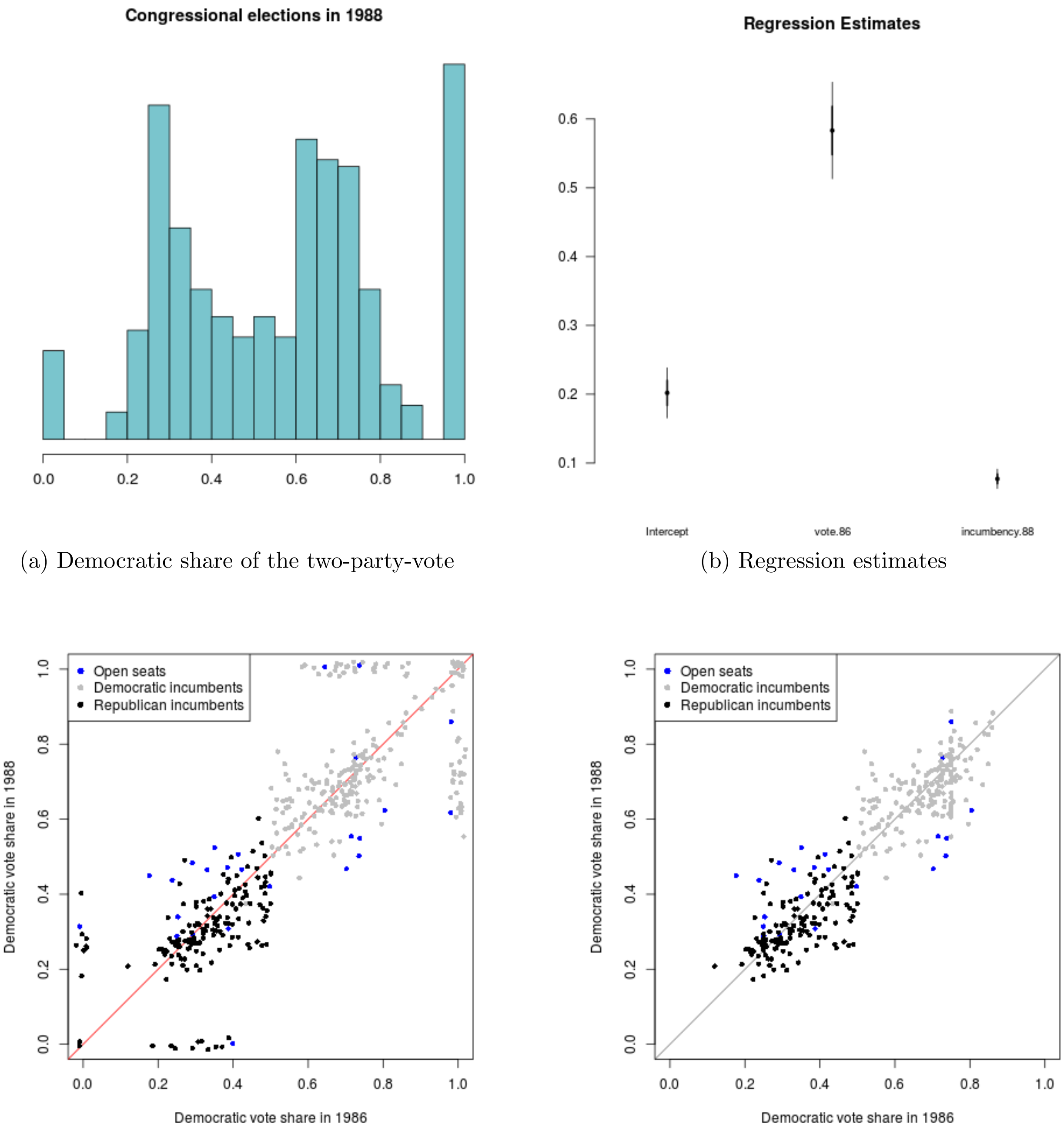
See plot (b) for the results.

Auxilliary remarks on resulting plots:

- Plot (a): Histogram of the 1988 election results for the Democratic party. Spikes at 0 and 1 represent uncontested elections, where the running candidate faced no opposition.
- Plot (b): Regression estimates of the model with first and second standard errors.
- Plot (c): Gaphical representation of the untreated data. The uncontested elections that can be observed along the edges of the plot have been slightly "jittered" meaning that we introduced an error term to avoid stacking them onto one another.
- Plot (d): The same data after removing uncontested elections in 1986 and 1988 following the Gelman procedure.

Simulation results:

Table 2 shows the results of a thousand simulations of the congressional election forecasting model. The last column gives us the results that we are interested in, i.e. the predicted number of congressional districts won by the Democratic party.



(a) Democratic share of the two-party-vote

(b) Regression estimates

(c) Congressional election data (1986 - 1988)

(d) Regression analysis data (data adjusted)

Table 2: Simulation results for the congressional election forecasting model

Sim	σ	β_0	β_1	β_2	$\widetilde{y}_1$	$\widetilde{y}_2$	$\dots$	$\widetilde{y}_{55}$	$\dots$	$\widetilde{y}_{435}$	$\sum_i(\widetilde{y}_i > 0.5)$
1	0.070	0.193	0.605	0.067	0.770	0.722	$\dots$	NA	$\dots$	0.734	244
2	0.066	0.210	0.573	0.079	0.809	0.670	$\dots$	NA	$\dots$	0.811	251
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
1000	0.066	0.184	0.608	0.077	0.727	0.618	$\dots$	NA	$\dots$	0.784	246
Median	0.066	0.203	0.580	0.078	0.727	0.649	$\dots$	NA	$\dots$	0.790	248
Mean	0.067	0.203	0.581	0.078	0.726	0.651	$\dots$	NA	$\dots$	0.787	247.629
SD	0.002	0.018	0.035	0.007	0.066	0.068	$\dots$	NA	$\dots$	0.071	2.861