Statistical Modeling and Inference – Problem Set #4

Niti Mishra · Miquel Torrens · Bálint Ván

November 9th, 2015

Solution to proposed exercises.

Exercise 1

We need to show that $y(\mu+\sigma)$ is a point less than one standard deviation away from the mean of the marginal distribution of t, that is:

$$y(\mu+\sigma) \leq \bar{t} + \sqrt{\mathbb{V}[t]}$$

Given that $x \sim \mathcal{N}(\mu, \sigma^2)$ and $\varepsilon \sim \mathcal{N}(0, \tau^2)$ are assumed to be uncorrelated:

$$\begin{aligned}
\mathbb{E}[t] &= \mathbb{E}[x+\varepsilon] = \mathbb{E}[x] + \mathbb{E}[\varepsilon] = \mu = \bar{t} \\
\mathbb{V}[t] &= \mathbb{V}[x+\varepsilon] = \mathbb{V}[x] + \mathbb{V}[\varepsilon] = \sigma^2 + \tau^2
\end{aligned}$$

We see that $\mu = \bar{t}$ because given the distribution of its components t is a normally (and thus symmetrically) distributed around its mean, and has the same expected value as x. On the other hand:

$$y(\mu+\sigma) = \mathbb{E}[t|x = \mu+\sigma] = \mu+\sigma$$

And so we would need to show that:

$$\begin{aligned}
y(\mu+\sigma) &\leq \mu + \sqrt{\mathbb{V}[t]} \\
\mu+\sigma &\leq \mu + \sqrt{\sigma^2+\tau^2} \\
\sigma^2 &\leq \sigma^2+\tau^2 \\
\tau^2 &\geq 0
\end{aligned}$$

We know that $\tau^2 \geq 0$ is indeed non-negative, thus proved.

Exercise 2

Part (a)

We perform the following transformations to the raw data:

- Convert the field `HEIGHT` to total amount of inches and name it `HEIGHT_I`.
- Reescale the variable `SEX` to variable `MEN`, which takes value 1 if the individual is a man and 0 otherwise.
- We suppress individuals with a reported weight greater than 500, that is, those with values 998 and 999 which are not valid.
- We suppress individuals with a reported height greater than 8 feet, that is, those with values 998 and 999 which are not valid.
- We preventively cut off the individual with highest income, as he reports an income that doubles the second highest income (potentially an outlier).
- We cut off individuals with no income or with income only declared approximately, i.e. we contemplate valid answers in the variable `EARN1`.

Additionally, a first approach was to unify the fields `EARN1` and `EARN2` in one single field, adding an extra dummy field named `INEXACT` that captured whether we use the precise answer from `EARN1` or the approximated answer in `EARN2`. The following results were slightly better and we had a bigger sample, but the interpretability of the model became much less intuitive and prediction lost sense, thus we finally decided to work with only precise declared outcomes. This is also the approach taken by the book.

Part (b)

The regression run is the following:

```
Call:
lm(formula = EARNT ~ HEIGHT_I, data = dta)

Residuals:
   Min     1Q Median     3Q    Max
-30297 -11309  -3489   6508 172873

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -61946       9585  -6.463  1.5e-10 ***
HEIGHT_I        1272        143   8.900  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18910 on 1179 degrees of freedom
Multiple R-squared:  0.06295,   Adjusted R-squared:  0.06216
F-statistic: 79.21 on 1 and 1179 DF,  p-value: < 2.2e-16
```

The transformation needed to interpret the intercept as average earnings for people with average height is substracting the mean from the `HEIGHT_I` variable. If we do so the resulting intercept is the following:

```
Call:
lm(formula = EARNT ~ HEIGHT_I_C, data = dta)

Residuals:
   Min     1Q Median     3Q    Max
-30297 -11309  -3489   6508 172873

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  23218.9      550.3   42.19   <2e-16 ***
HEIGHT_I_C    1272.5      143.0    8.90   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18910 on 1179 degrees of freedom
Multiple R-squared:  0.06295,	Adjusted R-squared:  0.06216
F-statistic: 79.21 on 1 and 1179 DF,  p-value: < 2.2e-16
```

This states that a person with average height will earn on average an income of approximately $23,219.

Part (c)

We have run the following models:

```
> # Linear-linear
> m03 <- lm(EARNT ~ HEIGHT_I + WEIGHT + MEN, data = dta)
> m04 <- lm(EARNT ~ HEIGHT_I + WEIGHT + MEN + WEIGHT * MEN + HEIGHT_I * MEN, data = dta)
>
> # Log-linear
> m05 <- lm(log(EARNT) ~ HEIGHT_I + WEIGHT + MEN + WEIGHT * MEN, data = dta)
> m06 <- lm(log(EARNT) ~ HEIGHT_I + WEIGHT + MEN + WEIGHT * MEN + HEIGHT * MEN, data = dta)
> m07 <- lm(log(EARNT) ~ HEIGHT_I + WEIGHT + MEN + WEIGHT * MEN, data = dta)
> m08 <- lm(log(EARNT) ~ HEIGHT_I_ST + WEIGHT_ST + MEN + WEIGHT_ST * MEN + HEIGHT_I_ST * MEN, data = dta)
> m09 <- lm(log(EARNT) ~ HEIGHT_I + WEIGHT + MEN, data = dta)
> m10 <- lm(log(EARNT) ~ HEIGHT_I + WEIGHT + MEN + WEIGHT * MEN + HEIGHT_I * MEN, data = dta)
>
> # Log-log
> m11 <- lm(log(EARNT) ~ log(HEIGHT_I) + log(WEIGHT) + MEN, data = dta)
```

We choose model m08 for several reasons. First, the residual standard error is the smallest between within log-linear models; second, it has a relatively high R^2; third, the regressors are mostly significant and have an intuitive sign, and, fourth, the use of centered variables will help the interpretation of the coefficients. The results are the following:

```
Call:
lm(formula = log(EARNT) ~ HEIGHT_I_ST + WEIGHT_ST + MEN + WEIGHT_ST *
    MEN + HEIGHT_I_ST * MEN, data = dta)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3010 -0.3889  0.1484  0.5681  2.3223

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      9.51288    0.04653 204.454  < 2e-16 ***
HEIGHT_I_ST      0.10792    0.05248   2.056 0.039955 *
WEIGHT_ST       -0.09748    0.04088  -2.385 0.017259 *
MEN              0.41765    0.07377   5.662 1.88e-08 ***
WEIGHT_ST:MEN    0.24201    0.06394   3.785 0.000162 ***
HEIGHT_I_ST:MEN -0.09361    0.07806  -1.199 0.230692
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8771 on 1175 degrees of freedom
Multiple R-squared:  0.09793,	Adjusted R-squared:  0.09409
F-statistic: 25.51 on 5 and 1175 DF,  p-value: < 2.2e-16
```

Part (d)

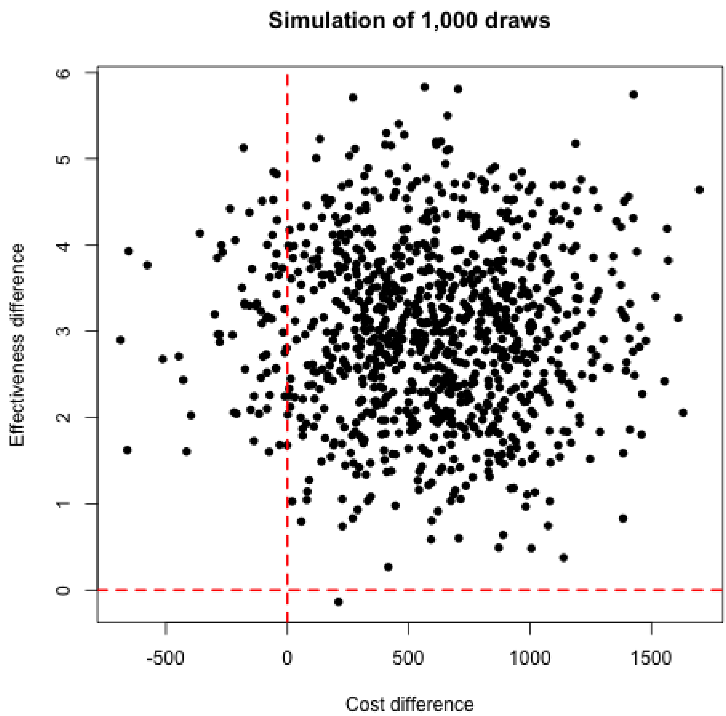
The results are only for strictly positive income individuals. The interpretation of the coefficients is the following:

- The intercept says that a woman of average height and weight is expected to earn $\exp(9.5129) \approx \$13,533$.
- The coefficient for `HEIGHT_I_ST` says that for an average weight woman exceeding the average height by one standard deviation increases the expected income by 10.8%. Exceeding by two standard deviations would increase the expected salary by 21.6%, and so on. In the case of the average weight men, this is diminished by the coefficient of `HEIGHT_I_ST:MEN`, which sets the total increase for men to 1.4% when one standard deviation away, although this coefficient is not significant, so in fact it is likely that there is no different effect for men and women whatsoever.
- The coefficient for `WEIGHT_ST` says that for an average height woman exceeding the average weight by one standard deviation decreases the expected income by 9.7%. In the case of average height men, this is actually overturned by the coefficient on `WEIGHT_ST:MEN`, and the aggregated effect is of +14.4% on income, so the data shows a negative effect on women and positive one on men.
- The coefficient on `MEN` suggests that an man of average height and weight earns on average 41.7% more than a woman of average height and weight.

Exercise 3

Part (a)

We generate random draws and plot the result:



Part (b)

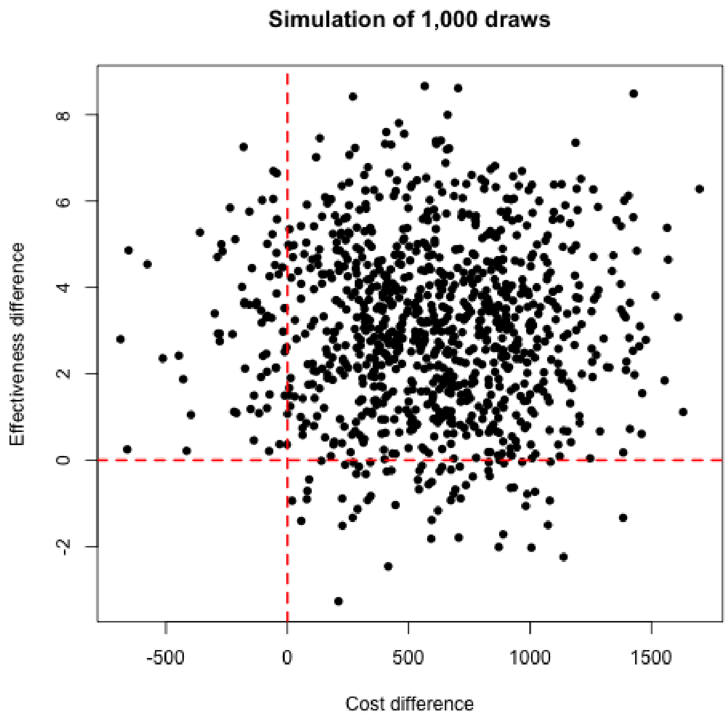
We take the sample ratio of the simulations we have drawn and we compute the quantiles of this sample ratio distribution, keeping those that concentrate the confidence percentage that we desire.

For the 50% case, we take quantiles 25% and 75%, leaving 50% of the sample inside (50% confidence interval). In our sample this interval is $(108.7, 313.7)$.

For the 95% case, we take quantiles 2.5% and 97.5%, leaving 95% of the sample inside (95% confidence interval). The interval is $(-57.3, 677.9)$.

Part (b)

We change the standard error and we obtain the following results:

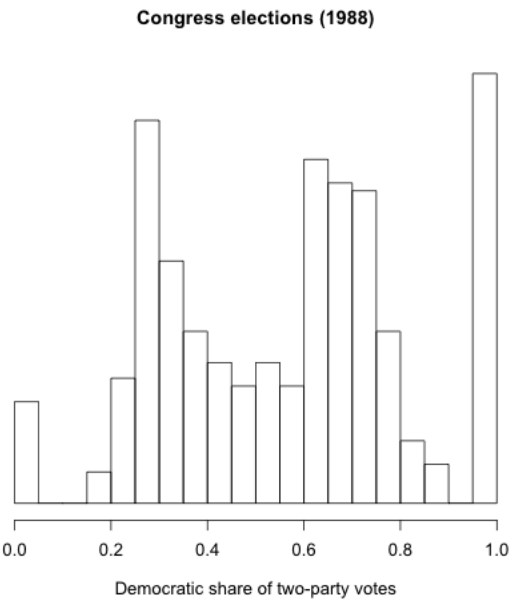


The intervals are computed analogously are the following:

- 50% confidence interval: $(71.2, 326.9)$.
- 95% confidence interval: $(-1386.8, 2122.8)$.

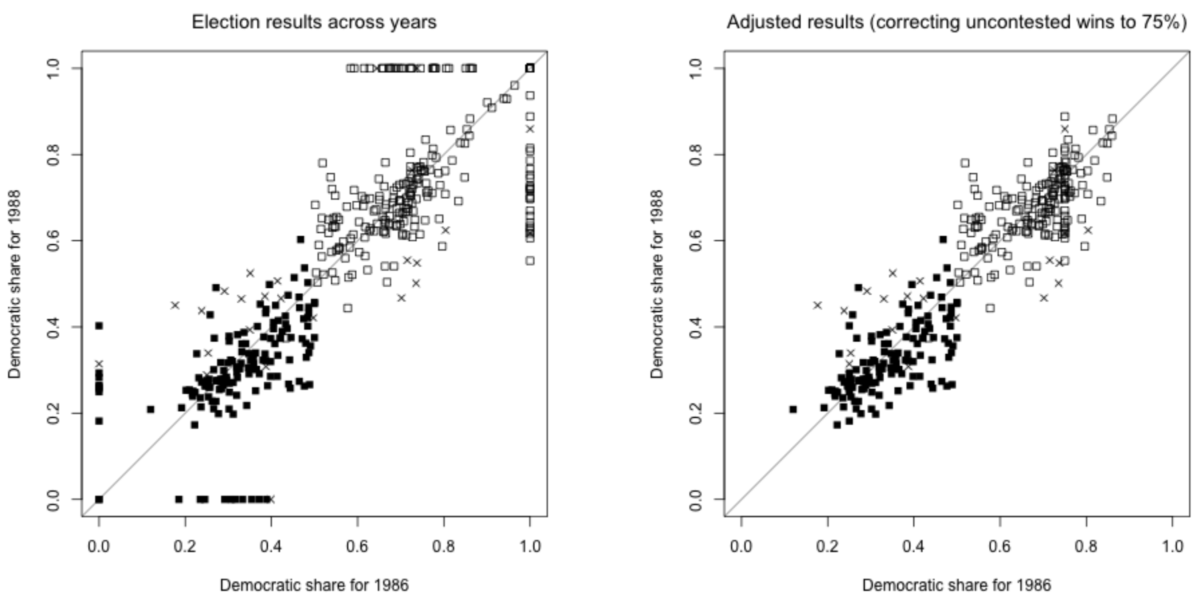
Exercise 4

We plot a histogram the district share of democratic votes data for 1988:



Districts with less than 10% share are condensed towards zero and districts with more than 90% are condensed towards one, as they represent uncontested districts.

We will try to run a linear regression model that predicts the democratic share of vote in 1988 given the share of vote from the previous election and the sign of the party of the incumbent until the election.



In the first plot, bolded squares are districts with Republican incumbents in 1988, white squares belong to Democratic incumbents and crosses represent open seats. The line represented is the 45 degree line. In the second plot we show the same picture but correcting for uncontested wins in 1986, adjusting the shares to 75%-25% in favor of the 1986 uncontested winner.

The regression is run using only the contested wins in 1988 using the adjusted data. The incumbent variable is coded as follows: -1 for Republican holder, 1 for Democratic holder and 0 for open seat. The summary of the model is the following:

```
lm(formula = vote.88 ~ vote.86 + incumbency.88)
               coef.est coef.se
(Intercept)    0.20     0.02
vote.86        0.58     0.04
incumbency.88  0.08     0.01
---
n = 343, k = 3
residual sd = 0.07, R-Squared = 0.88
```

Now, with these results we would like to estimate the outcome for the following election in 1990. To do this, we use 1,000 predictive simulations drawn with the book's `sim()` function. We multiply the new data (share of votes of 1988 and incumbent as of 1990) and we multiply it with the coefficients of the predictive simulations (that is, $\mathbf{\Phi}_{90}\mathbf{w}$) and we add normally distributed errors.

The results of this simulations show for each district (total of 435) what will the Democratic share of votes be for each simulation (total of 1,000), thatt is, a 1000×435 matrix. The districts won uncontested in 1990 have values `NA` as the model does not predict them. We set these to zero.

The number of elections won by the Democrats in 1990 is predicted using the simple rule of win in case the predicted value is above 50% and lose otherwise. The results are shown in the book's Figure 7.5, which we approximately replicate here:

	sigma	beta0	beta1	beta2	pred_y1	pred_y2	pred_y3	pred_y4	pred_y5	pred_dem_wins
simulation1	0.0690	0.1756	0.6307	0.0717	0.7521	0.7063	0.5519	0.3768	0.2579	248.0
simulation2	0.0702	0.2063	0.5824	0.0791	0.7698	0.6215	0.5652	0.2731	0.2657	252.0
simulation3	0.0701	0.1954	0.5994	0.0716	0.7187	0.7528	0.6222	0.3690	0.3242	251.0
simulation4	0.0724	0.2076	0.5699	0.0808	0.5825	0.7278	0.5636	0.2671	0.2948	245.0
simulation5	0.0667	0.2183	0.5591	0.0779	0.8249	0.6805	0.5785	0.2402	0.3696	253.0
simulation6	0.0656	0.2078	0.5757	0.0806	0.6814	0.6785	0.4860	0.2891	0.1543	246.0
simulation7	0.0651	0.2005	0.5823	0.0705	0.6358	0.4945	0.6133	0.2392	0.4317	248.0
simulation8	0.0674	0.2087	0.5732	0.0821	0.6625	0.6051	0.6600	0.2587	0.3795	244.0
simulation9	0.0657	0.2219	0.5494	0.0807	0.6703	0.5640	0.5351	0.3119	0.2537	245.0
simulation10	0.0698	0.2237	0.5292	0.0882	0.7076	0.5925	0.6057	0.3415	0.3273	252.0
mean	0.0682	0.2066	0.5751	0.0783	0.7006	0.6424	0.5782	0.2966	0.3059	248.4
median	0.0682	0.2077	0.5745	0.0798	0.6945	0.6500	0.5718	0.2811	0.3095	248.0
sd	0.0025	0.0141	0.0276	0.0056	0.0700	0.0807	0.0495	0.0509	0.0788	3.4

Gelman and Hill show that the average total wins predicted was actually way off the final result, which was 260, more than three standard deviations away from the mean, so this model does not really look applicable to 1990.