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:

- Unify the fields EARN1 and EARN2 in one single field, adding an extra field named INEXACT that captures whether we use the precise answer from EARN1 or the approximated answer in EARN2.
- The resulting variable is called EARNT and is measured in thousands of dollars (e.g. $10,000 has EARNT = 10).
- 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 (some have 990+ values for answer-codification reasons).
- We suppress individuals with a reported height greater than 8 feet (some have 990+ values for answer-codification reasons).
- We cut off the individual with highest income, as he reports an income that doubles the second highest income.

Part (b)

The regression run is the following:

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

Residuals:
    Min      1Q  Median      3Q     Max 
-36.235 -18.811  -8.736   5.193 167.621 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -57.6099    10.8670  -5.301 1.28e-07 ***
HEIGHT_I      1.2856     0.1629   7.892 4.87e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28.05 on 1982 degrees of freedom
Multiple R-squared:  0.03047,	Adjusted R-squared:  0.02998 
F-statistic: 62.29 on 1 and 1982 DF,  p-value: 4.865e-15
```

The resulting parameter for HEIGHT_I is imperceptibly sensitive to the inclusion of the control variable for exactness of income, although including it does boost the R^2 considerably and reduces the residual standard error.

The transformation needed to interpret the intercept as average earnings for people with average height is substracting the mean fro 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 
-36.235 -18.811  -8.736   5.193 167.621 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  28.0113     0.6297  44.485  < 2e-16 ***
HEIGHT_I_C    1.2856     0.1629   7.892 4.87e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28.05 on 1982 degrees of freedom
Multiple R-squared:  0.03047,	Adjusted R-squared:  0.02998 
F-statistic: 62.29 on 1 and 1982 DF,  p-value: 4.865e-15
```

This states that a person with average height will earn on average an income of $28,011 approximately.

Part (c)

We have run the following models:

```
> # Linear-linear
> m03 <- lm(EARNT ~ HEIGHT_I + WEIGHT + MEN + INEXACT, data = dta)
> m04 <- lm(EARNT ~ HEIGHT_I + WEIGHT + MEN, data = dta)
> m05 <- lm(EARNT ~ HEIGHT_I + WEIGHT + MEN + WEIGHT * MEN + HEIGHT_I * MEN + INEXACT, data = dta)
> m06 <- lm(EARNT ~ HEIGHT_I + WEIGHT + MEN + WEIGHT * MEN + HEIGHT_I * MEN, data = dta)
>
> # Log-linear
> m07 <- lm(log(EARNT) ~ HEIGHT_I + WEIGHT + MEN + WEIGHT * MEN, data = dta2)
> m08 <- lm(log(EARNT) ~ HEIGHT_I + WEIGHT + MEN + WEIGHT * MEN + INEXACT, data = dta2)
> m09 <- lm(log(EARNT) ~ HEIGHT_I + WEIGHT + MEN + WEIGHT * MEN + HEIGHT * MEN, data = dta2)
> m10 <- lm(log(EARNT) ~ HEIGHT_I + WEIGHT + MEN + WEIGHT * MEN, data = dta2)
> m11 <- lm(log(EARNT) ~ HEIGHT_I_ST + WEIGHT_ST + MEN + WEIGHT_ST * MEN + INEXACT, data = dta2)
> m12 <- lm(log(EARNT) ~ HEIGHT_I_ST + WEIGHT_ST + MEN + WEIGHT_ST * MEN + HEIGHT_I_ST * MEN + INEXACT, data = dta2)
> m13 <- lm(log(EARNT) ~ HEIGHT_I_ST + WEIGHT_ST + MEN + WEIGHT_ST * MEN + HEIGHT_I_ST * MEN + INEXACT * MEN, data = dta2)
> m14 <- lm(log(EARNT) ~ HEIGHT_I_ST + WEIGHT_ST + MEN + WEIGHT_ST * MEN + HEIGHT_I_ST * MEN, data = dta2)
> m15 <- lm(log(EARNT) ~ HEIGHT_I + WEIGHT + MEN + WEIGHT * MEN + HEIGHT_I * MEN + INEXACT, data = dta2)
> m16 <- lm(log(EARNT) ~ HEIGHT_I + WEIGHT + MEN + WEIGHT * MEN + INEXACT, data = dta2)
> m17 <- lm(log(EARNT) ~ HEIGHT_I + WEIGHT + MEN + INEXACT, data = dta2)
> m18 <- lm(log(EARNT) ~ HEIGHT_I + WEIGHT + MEN, data = dta2)
> m19 <- lm(log(EARNT) ~ HEIGHT_I + WEIGHT + MEN + WEIGHT * MEN + HEIGHT_I * MEN + INEXACT, data = dta2)
> m20 <- lm(log(EARNT) ~ HEIGHT_I + WEIGHT + MEN + WEIGHT * MEN + HEIGHT_I * MEN, data = dta2)
>
> # Log-log
> m21 <- lm(log(EARNT) ~ log(HEIGHT_I) + log(WEIGHT) + MEN + INEXACT, data = dta2)
```

In the models that use logarithms we only use the observations with positive earnings.

We choose model `m13` for several reasons. First, the residual standard error is the smallest between within log-linear models; second, it has a relatively high R^2 and, third, the regressors are all significant and have an intuitive sign. The results are the following:

```
Call:
lm(formula = log(EARNT) ~ HEIGHT_I_ST + WEIGHT_ST + MEN + WEIGHT_ST *
    MEN + HEIGHT_I_ST * MEN + INEXACT * MEN, data = dta2)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2748 -0.5289  0.1158  0.6890  2.2998

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.59563    0.04366  59.447  < 2e-16 ***
HEIGHT_I_ST      0.08166    0.04314   1.893  0.05850 .
WEIGHT_ST       -0.07483    0.03565  -2.099  0.03597 *
MEN              0.39671    0.07505   5.286  1.4e-07 ***
INEXACT          0.69900    0.06110  11.440  < 2e-16 ***
WEIGHT_ST:MEN    0.18072    0.05795   3.118  0.00185 **
HEIGHT_I_ST:MEN -0.01897    0.06916  -0.274  0.78391
MEN:INEXACT     -0.38219    0.09862  -3.875  0.00011 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9653 on 1794 degrees of freedom
Multiple R-squared:  0.1144,    Adjusted R-squared:  0.1109
F-statistic:  33.1 on 7 and 1794 DF,  p-value: < 2.2e-16
```

Part (d)

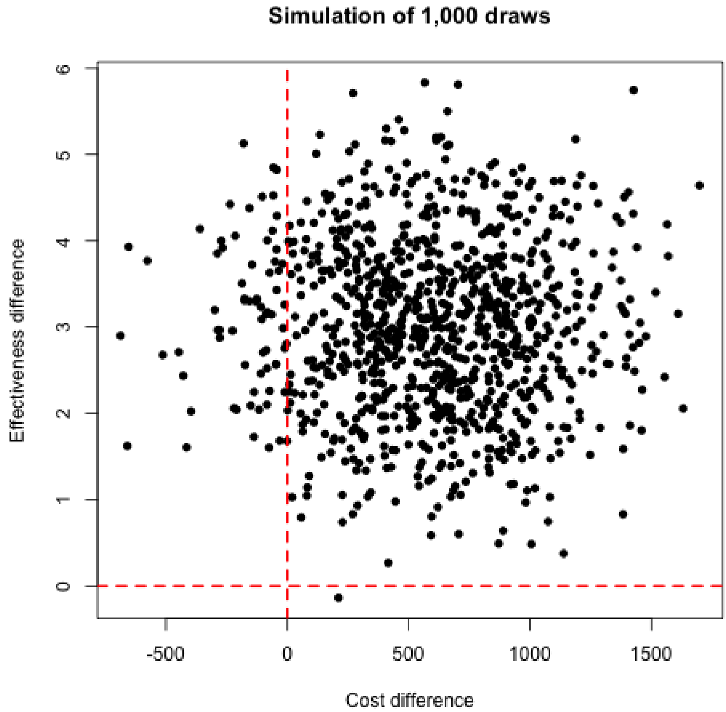
The results are only for strictly positive income individuals. The interpretation of the coefficients is the following:

- The intercept says that a woman of average height and weight (who would report the exact amount of salary) is expected to earn $1,000 \times \exp(2.5956) = \$13,405$.
- The coefficient for `HEIGHT_I_ST` says that exceeding the average height by one standard deviation increases the expected income by 8.1%. Exceeding by two standard deviations would increase the expected salary by 16.2%, and so on. In the case of men, this is diminished by the coefficient of `HEIGHT_I_ST:MEN`, which sets the total increase for men to 5.7% when one standard deviation away, although this coefficient is not significant.
- The coefficient for `WEIGHT_ST` says that exceeding the average weight by one standard deviation decreases the expected income by 7.6%. In the case of men, this is actually overturned by the coefficient on `WEIGHT_ST:MEN`, and the aggregated effect is of +10.9% on income, so the data shows a negative effect on women and positive on men.
- The coefficient on `MEN` suggests that an man of average height and weight earns on average 27.3% more than a woman of average height and weight.
- The coefficient on `INEXACT` means that a woman who reports inexact answers is expected to earn 69.9% more. In the case of men, we add the value of the interaction term to expect an increase of 31.7% for the average man.

Exercise 3

Part (a)

We generate random draws and plot the result:



Part (b)

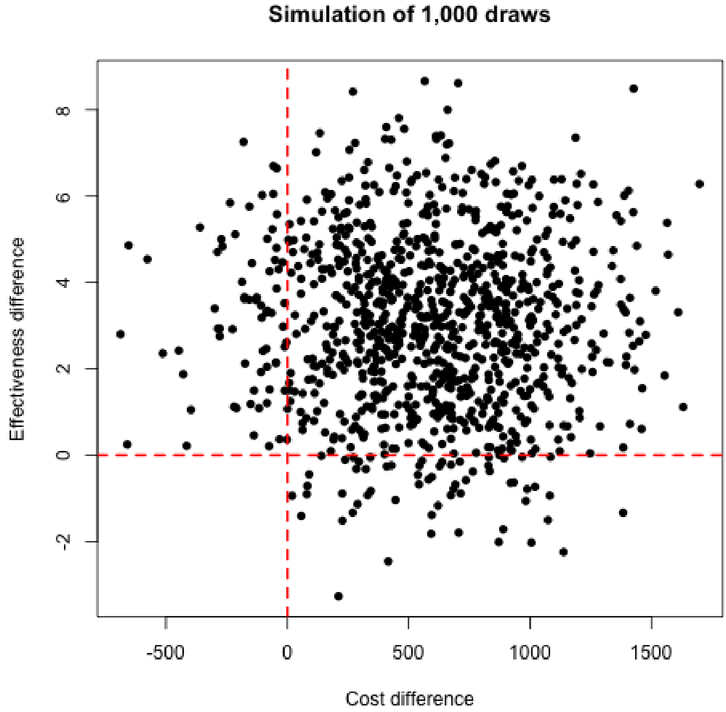
We take the sample ratio of the simulations we have drawn. We take the quantile of this sample ratio and keep the corresponding quantiles.

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

For the 50% 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)

The plot is the following:

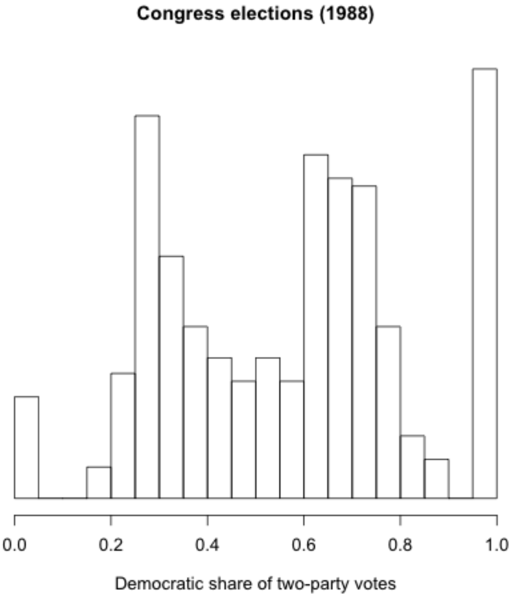


The intervals in this case are the following:

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

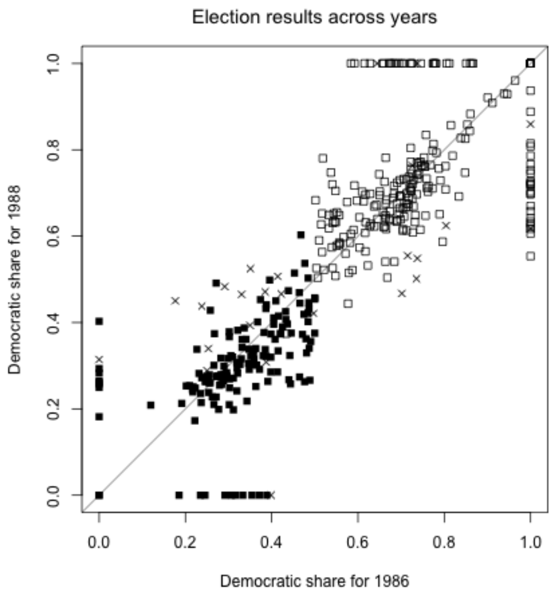
Exercise 4

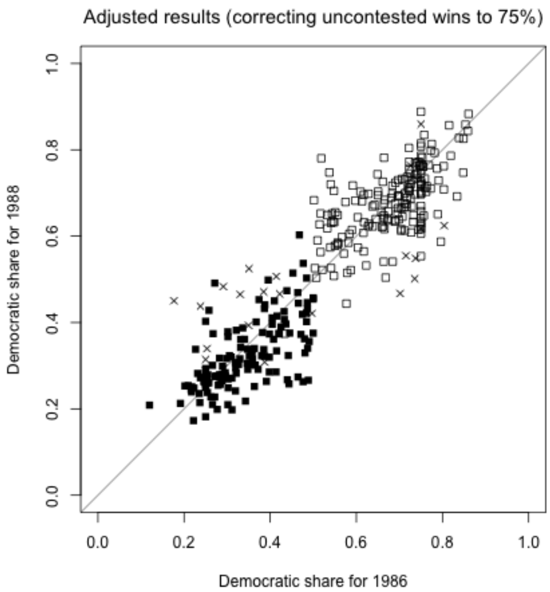
We plot a histogram the district share of democratic votes data for 1988:



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:

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