Statistical Modeling and Inference – Problem Set #1

Niti Mishra
Miquel Torrens
Bálint Ván

October 11th, 2015

Solution to proposed exercises.

Exercise 1

A matrix $\mathbf{A}$ is positive semidefinite iff $\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0$, $\forall \mathbf{x}$. Using this definition and setting $\mathbf{A} = \mathbf{X}^T\mathbf{X}$, we derive:

$$\begin{aligned} \mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{x}^T\left(\mathbf{X}^T\mathbf{X}\right)\mathbf{x} &= \left(\mathbf{x}^T\mathbf{X}^T\right)\left(\mathbf{X}\mathbf{x}\right) \\ &= \left(\mathbf{X}\mathbf{x}\right)^T\left(\mathbf{X}\mathbf{x}\right) \\ &= \mathbf{b}^T\mathbf{b} \\ &= ||\mathbf{b}||^2 \\ &\geq 0. \end{aligned}$$

Thus, $\mathbf{X}^T\mathbf{X}$ is positive semidefinite under such conditions.

Exercise 2

We carry out the prostate cancer example. Correlations of predictors in the data:

```
##            lcavol    lweight       age         lbph         svi
## lcavol  1.0000000 0.28052138 0.2249999  0.027349703  0.53884500
## lweight 0.2805214 1.00000000 0.3479691  0.442264399  0.15538490
## age     0.2249999 0.34796911 1.0000000  0.350185896  0.11765804
## lbph    0.0273497 0.44226440 0.3501859  1.000000000 -0.08584324
## svi     0.5388450 0.15538490 0.1176580 -0.085843238  1.00000000
## lcp     0.6753105 0.16453714 0.1276678 -0.006999431  0.67311118
## gleason 0.4324171 0.05688209 0.2688916  0.077820447  0.32041222
## pgg45   0.4336522 0.10735379 0.2761124  0.078460018  0.45764762
##                 lcp    gleason      pgg45
## lcavol   0.675310484 0.43241706 0.43365225
## lweight  0.164537142 0.05688209 0.10735379
## age      0.127667752 0.26889160 0.27611245
## lbph    -0.006999431 0.07782045 0.07846002
## svi      0.673111185 0.32041222 0.45764762
## lcp      1.000000000 0.51483006 0.63152825
## gleason  0.514830063 1.00000000 0.75190451
## pgg45    0.631528246 0.75190451 1.00000000
```

Linear model fit to the prostate cancer (training) data.

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi + lcp +
##     gleason + pgg45, data = prostate[prostate$train == TRUE,
##     ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.64870 -0.34147 -0.05424  0.44941  1.48675
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.46493    0.08931  27.598  < 2e-16 ***
## lcavol       0.67953    0.12663   5.366 1.47e-06 ***
## lweight      0.26305    0.09563   2.751  0.00792 **
## age         -0.14146    0.10134  -1.396  0.16806
## lbph         0.21015    0.10222   2.056  0.04431 *
## svi          0.30520    0.12360   2.469  0.01651 *
## lcp         -0.28849    0.15453  -1.867  0.06697 .
## gleason     -0.02131    0.14525  -0.147  0.88389
## pgg45        0.26696    0.15361   1.738  0.08755 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7123 on 58 degrees of freedom
## Multiple R-squared:  0.6944, Adjusted R-squared:  0.6522
## F-statistic: 16.47 on 8 and 58 DF,  p-value: 2.042e-12
```

The F-value is 1.6698, with the corresponding p-value of 0.1693. The mean prediction error on the test data is 0.5140. In contrast, prediction using the mean training value of `lpsa` ha a test error of 0.7409.

Having seen the correlations in Table 3.1 do you anticipate the opposite signs for `lcp` and `lcavol` on Table 3.2?

We can anticipate that one of them will not be significant, i.e. that its sign will be unsignificantly positive or negative, possibly opposite to the sign of the other covariate.

The high correlation between `lcp` and `lcavol` draws us think that there is high colinearity between them. Thus, when we run the regression one of them will absorbe the largest part of the variance, leaving the highly-correlated covariate statistically insignificant.

Exercise 3

(a) Show that $\mathbf{H} = \mathbf{H}^T$

We have:

$$\mathbf{H} = \mathbf{\Phi} \left(\mathbf{\Phi}^T \mathbf{\Phi} \right)^{-1} \mathbf{\Phi}^T$$

Its transpose is:

$$\begin{aligned}
\mathbf{H}^T &= \left(\mathbf{\Phi}\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^T\right)^T \\
&= \left(\mathbf{\Phi}\mathbf{\Phi}^{-1}\left(\mathbf{\Phi}^T\right)^{-1}\mathbf{\Phi}^T\right)^T \\
&= \left(\mathbf{\Phi}^T\right)^T\left(\left(\mathbf{\Phi}^{-1}\right)^T\right)^T\left(\mathbf{\Phi}^{-1}\right)^T\mathbf{\Phi}^T \\
&= \mathbf{\Phi}\mathbf{\Phi}^{-1}\left(\mathbf{\Phi}^T\right)^{-1}\mathbf{\Phi}^T \\
&= \mathbf{\Phi}\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^T \\
&= \mathbf{H}.
\end{aligned}$$

(b) Show that $\mathbf{H} = \mathbf{H}^2$

Given the above definition of $\mathbf{H}$:

$$\begin{aligned}
\mathbf{H}^2 = \mathbf{H}\mathbf{H} &= \left(\mathbf{\Phi}\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^T\right)\left(\mathbf{\Phi}\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^T\right) \\
&= \mathbf{\Phi}\left[\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)\right]\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^T \\
&= \mathbf{\Phi}\mathbf{I}\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^T \\
&= \mathbf{\Phi}\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^T \\
&= \mathbf{H}.
\end{aligned}$$

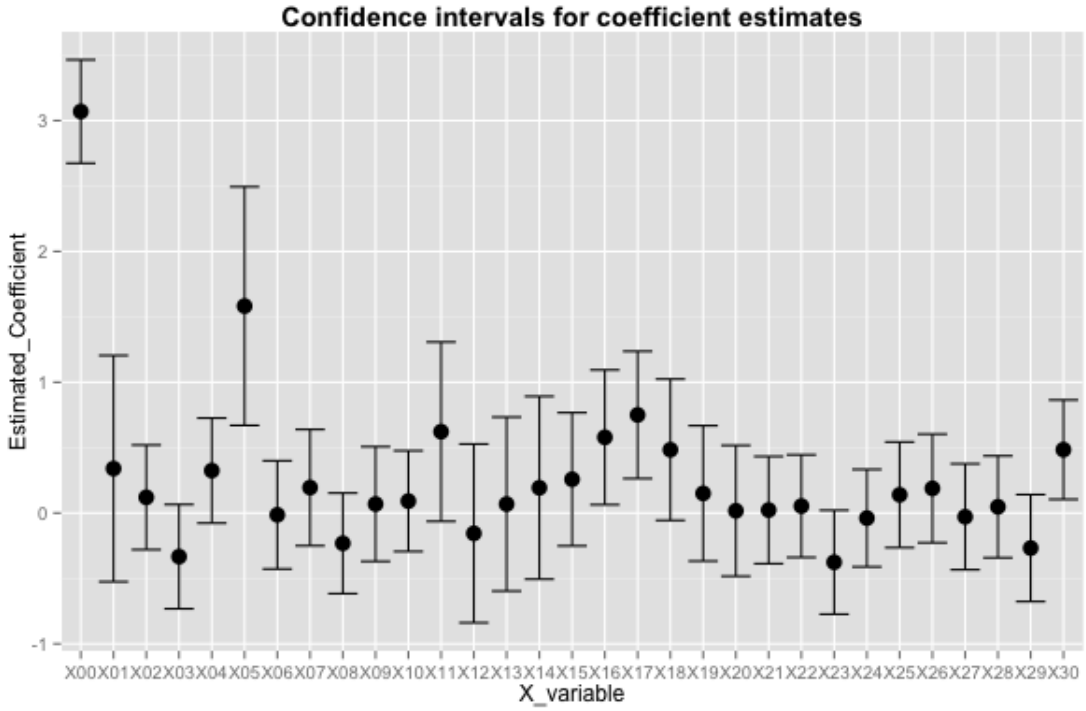
(c) Show that $\mathrm{tr}(\mathbf{H}) = M + 1$ **.**

We use the properties of the trace to solve this:

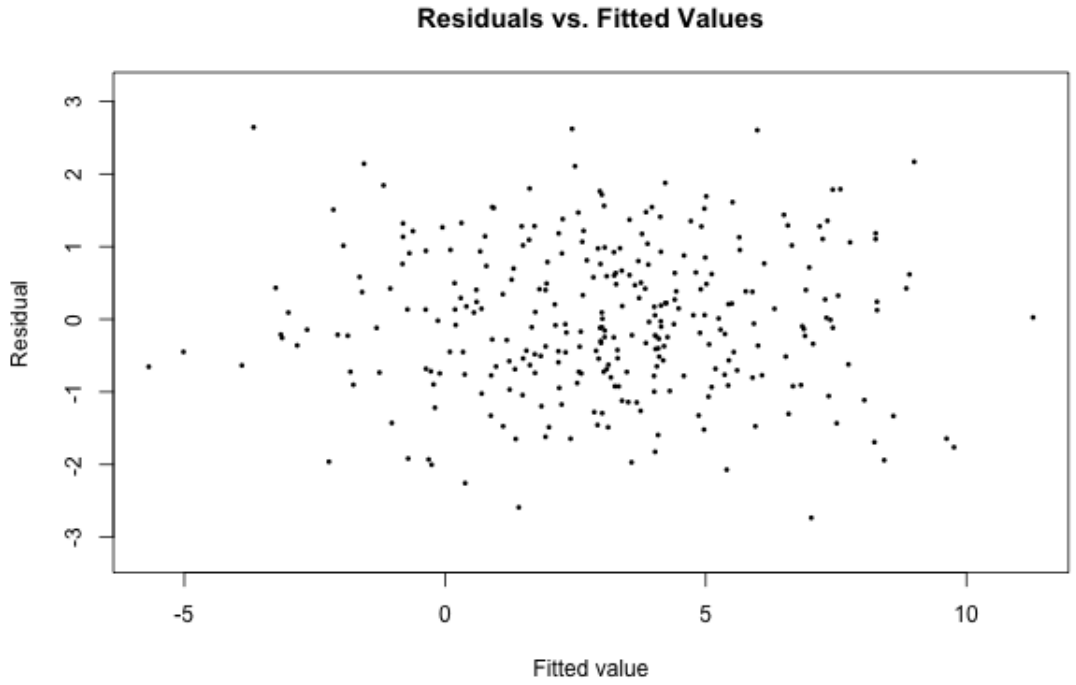
$$\begin{aligned}
\mathrm{tr}(\mathbf{H}) &= \mathrm{tr}\left(\mathbf{\Phi}\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^T\right) \\
&= \mathrm{tr}\left(\mathbf{\Phi}^T\mathbf{\Phi}\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\right) \\
&= \mathrm{tr}\left(\mathbf{I}_{M+1}\right) \\
&= M + 1.
\end{aligned}$$

Exercise 4

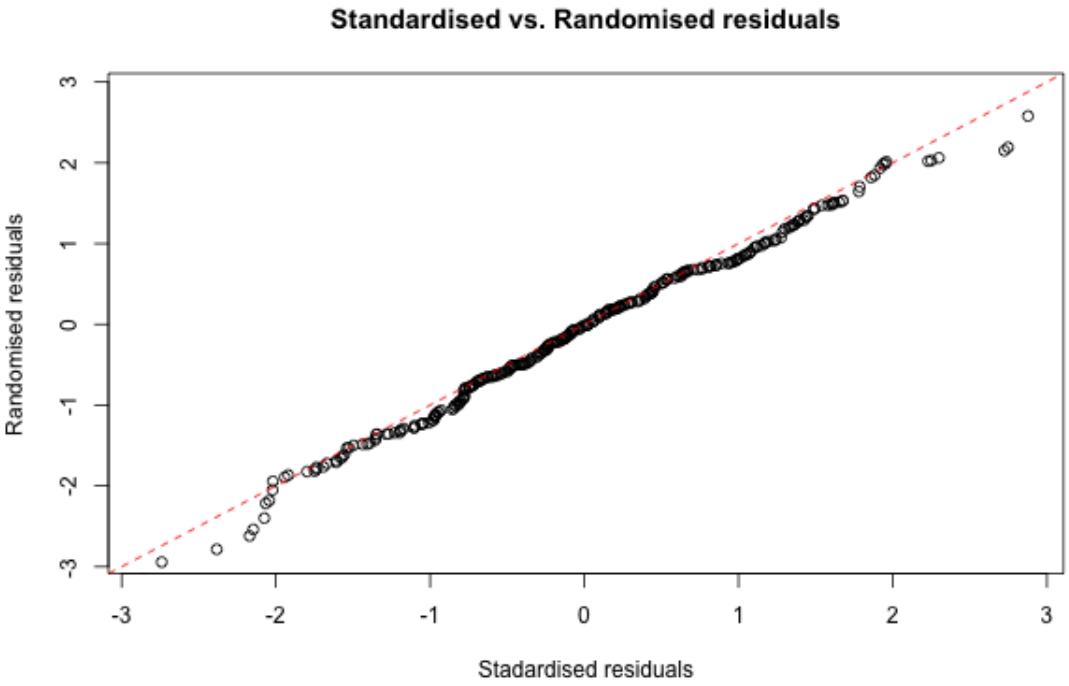
Plot 1:



Plot 2:



Plot 3:


Standardised vs. Randomised residuals
Randomised residuals
Stadardised residuals
-3
-2
-1
0
1
2
3

Exercise 5

As stated in the slides, note that $\mathbf{V}_r\mathbf{V}_r^T$ is the hat matrix in the general case:

$$\mathbf{\Phi w}^* = \mathbf{V}_r\theta_{MLE} = \mathbf{V}_r\mathbf{V}_r^T\mathbf{t} = \mathbf{Ht}$$

So we need to prove that $\mathbf{H}$ is a projection matrix. In mathematical terms, it means it must satisfy the following prperty:

$$\mathbf{H} = \mathbf{H}^2$$

Afterwards, we shall prove that $\text{tr}\left(\mathbf{V}_r\mathbf{V}_r^T\right) = r$.

First of all let's prove it is a projection matrix:

$$\mathbf{H}^2 = \mathbf{HH} = \left(\mathbf{V}_r\mathbf{V}_r^T\right)\left(\mathbf{V}_r\mathbf{V}_r^T\right) = \mathbf{V}_r\left(\mathbf{V}_r^T\mathbf{V}_r\right)\mathbf{V}_r^T = \mathbf{V}_r\mathbf{I}_r\mathbf{V}_r^T = \mathbf{V}_r\mathbf{V}_r^T = \mathbf{H}$$

Thus, this is a projection matrix. Note that we can use $\mathbf{V}_r^T\mathbf{V}_r = \mathbf{I}_r$ because we know that as part of the singular value decomposition $\mathbf{V}_r$ is by definition an orthogonal matrix, and so that its columns are orthogonal to each other. They are actually orthonormal, meaning that they are unit vectors.

This is useful when calculating the trace. Since $\mathbf{V}_r$ is orthogonal and made up of unit vectors, all elements in the diagonal of the product $\mathbf{V}_r^T\mathbf{V}_r$ are equal to one and thus $\text{tr}\left(\mathbf{V}_r\mathbf{V}_r^T\right) = \sum_{i=1}^{r} v_{ii} = \sum_{i=1}^{r} 1 = r$, where v_{ii} denotes the diagonal value of the resulting matrix $\mathbf{V}_r^T\mathbf{V}_r$.

Exercise 6

We shall derive the normal equation for a regression model written as:

$$\mathcal{N}\left(\mathbf{t}|\mathbf{\Phi w},(q\mathbf{D})^{-1}\right)=\frac{1}{(2\pi)^{\frac{D}{2}}}(q\mathbf{D})^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\mathbf{t}-\mathbf{\Phi w})^{T}q\mathbf{D}(\mathbf{t}-\mathbf{\Phi w})\right)$$

For the sake of simple notation we denote this function simply as $\mathcal{N}$. We compute its log-likelihood function:

$$\begin{aligned}\prod_{N}\mathcal{N} &= \frac{1}{(2\pi)^{\frac{ND}{2}}}(q\mathbf{D})^{-\frac{N}{2}}\exp\left(-\frac{1}{2}q\sum_{N}(\mathbf{t}-\mathbf{\Phi w})^{T}\mathbf{D}(\mathbf{t}-\mathbf{\Phi w})\right)\\ \log\prod_{N}\mathcal{N} &= C+\frac{N}{2}\log q\mathbf{D}-\frac{1}{2}q(\mathbf{t}-\mathbf{\Phi w})^{T}\mathbf{D}(\mathbf{t}-\mathbf{\Phi w})\\ -2\log\prod_{N}\mathcal{N} &= C-N\log q\mathbf{D}+q(\mathbf{t}-\mathbf{\Phi w})^{T}\mathbf{D}(\mathbf{t}-\mathbf{\Phi w})\end{aligned}$$

Where C denotes all constant terms that will disappear during the maximization step. To find the normal equation we will only need to maximize with respect to $\mathbf{w}$. The problem therefore is the following:

$$\max_{\mathbf{w}}-2\log\prod_{N}\mathcal{N}$$

Before differentiating, we shall develop a little bit the terms involving $\mathbf{w}$:

$$\begin{aligned}(\mathbf{t}-\mathbf{\Phi w})^{T}\mathbf{D}(\mathbf{t}-\mathbf{\Phi w}) &= \mathbf{t}^{T}\mathbf{Dt}-(\mathbf{\Phi w})^{T}\mathbf{Dt}-\mathbf{t}^{T}\mathbf{D}(\mathbf{\Phi w})-(\mathbf{\Phi w})^{T}\mathbf{D}(\mathbf{\Phi w})\\ &= \mathbf{t}^{T}\mathbf{Dt}-2\mathbf{t}^{T}\mathbf{D\Phi w}-\mathbf{w}^{T}\mathbf{\Phi}^{T}\mathbf{D\Phi w}\end{aligned}$$

To maximize the objective function note that we can just differentiate this term, as the rest will cross out once set to zero:

$$\frac{\partial(-2\log\prod_{N}\mathcal{N})}{\partial\mathbf{w}}=0\Leftrightarrow\frac{\partial\left(\mathbf{t}^{T}\mathbf{Dt}-2\mathbf{t}^{T}\mathbf{D\Phi w}-\mathbf{w}^{T}\mathbf{\Phi}^{T}\mathbf{D\Phi w}\right)}{\partial\mathbf{w}}=0$$

Differentiating:

$$\begin{aligned}-2\mathbf{t}^{T}\mathbf{D\Phi}+\mathbf{w}^{T}\left(\mathbf{\Phi}^{T}\mathbf{D\Phi}+\left(\mathbf{\Phi}^{T}\mathbf{D\Phi}\right)^{T}\right) &= 0\\ 2\mathbf{t}^{T}\mathbf{D\Phi} &= \mathbf{w}^{T}\left(\mathbf{\Phi}^{T}\mathbf{D\Phi}+\left(\mathbf{\Phi}^{T}\mathbf{D\Phi}\right)^{T}\right)\\ \mathbf{t}^{T}\mathbf{D\Phi} &= \mathbf{w}^{T}\mathbf{\Phi}^{T}\mathbf{D\Phi}\\ \mathbf{w}^{T} &= \mathbf{t}^{T}\mathbf{D\Phi}\left(\mathbf{\Phi}^{T}\mathbf{D\Phi}\right)^{-1}\\ \mathbf{w} &= \left(\mathbf{t}^{T}\mathbf{D\Phi}\left(\mathbf{\Phi}^{T}\mathbf{D\Phi}\right)^{-1}\right)^{T}\\ \mathbf{w} &= \left(\mathbf{\Phi}^{T}\mathbf{D\Phi}\right)^{-1}\mathbf{\Phi}^{T}\mathbf{D}^{T}\mathbf{t}\end{aligned}$$

From this we can obtain the normal equation:

$$\mathbf{\Phi}^{T}\mathbf{D\Phi w}=\mathbf{\Phi}^{T}\mathbf{D}^{T}\mathbf{t}$$

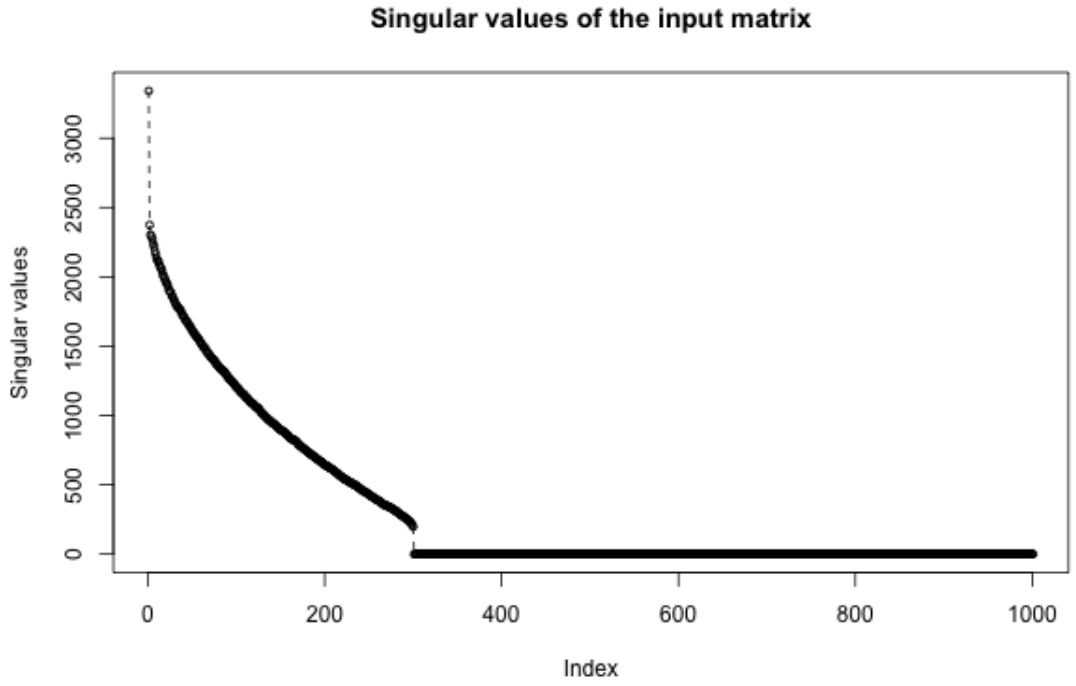
Given that in practice $\mathbf{D}$ is a diagonal matrix, $\mathbf{D}=\mathbf{D}^{T}$ and so:

$$\mathbf{\Phi}^{T}\mathbf{D\Phi w}=\mathbf{\Phi}^{T}\mathbf{Dt}$$

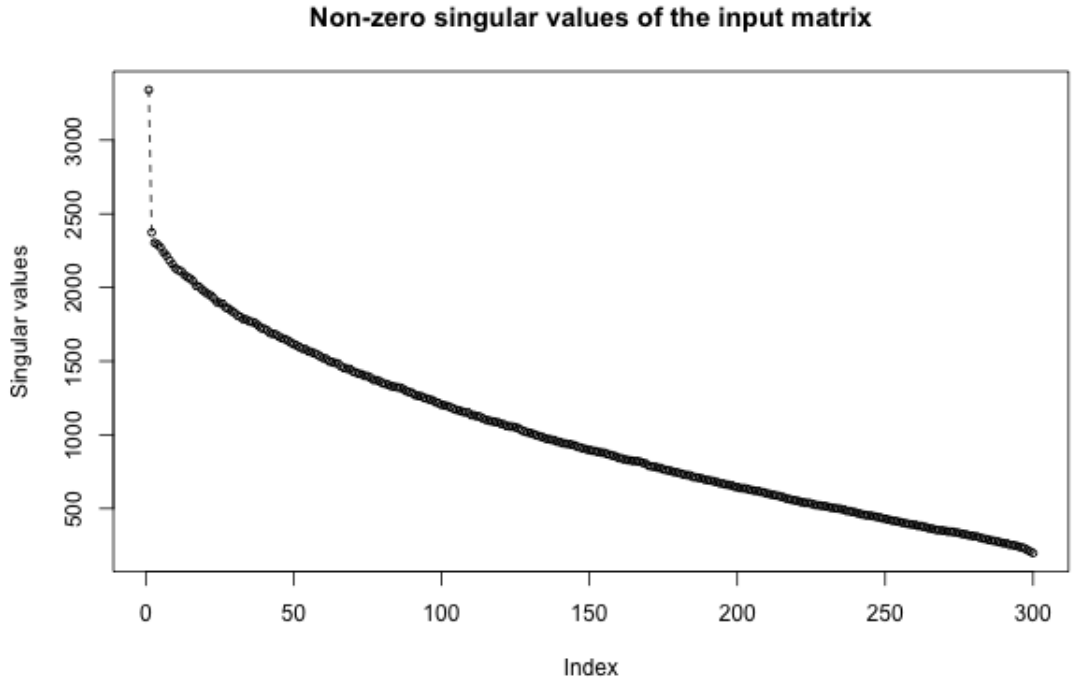
Hence proved.

Exercise 7

We plot the non-zero singular values:



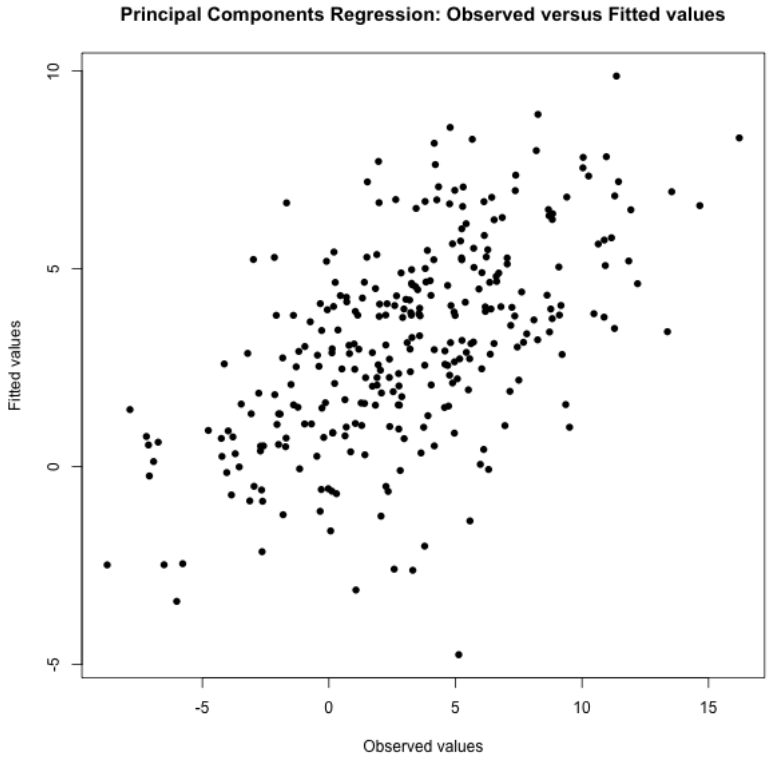
Note that some of these values are barely larger than zero, so they are plotted above but visually lie upon the horizontal axis. If we suppress them, the plot looks like this:



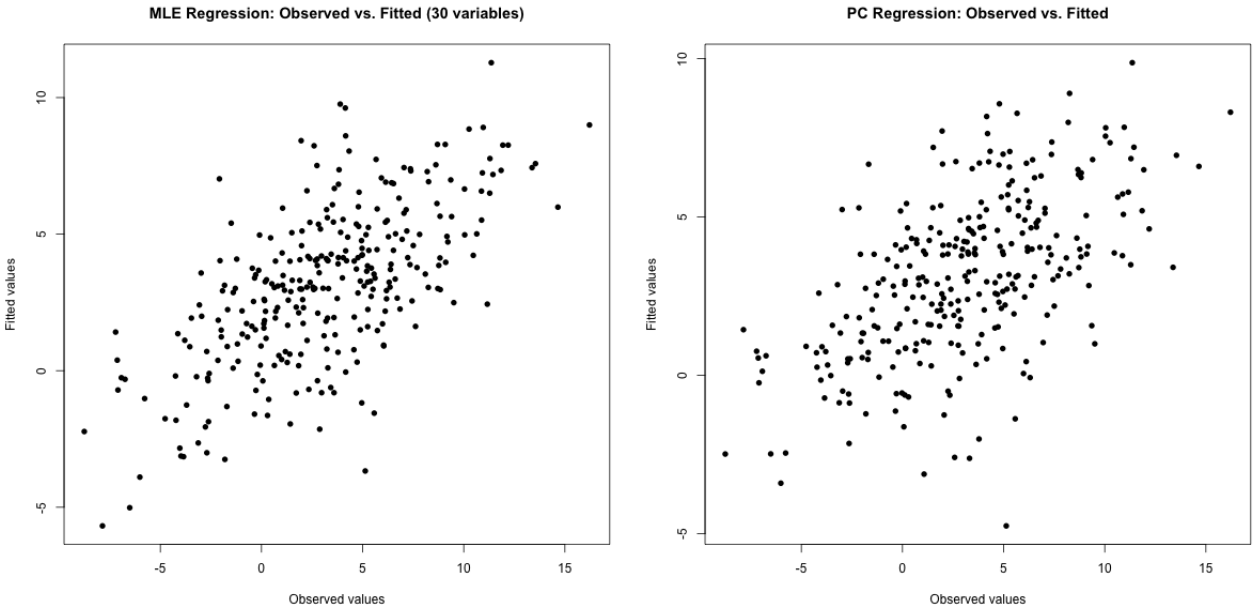
The rank of r of the input matrix is precisely the number of singular values greater than zero, which is 300, as you can see above.

$$r = 300.$$

Now let's perform the principal components regression. We can fit the values using this model to obtain the following results:



We can compare these results with the ones obtained from the model performed in Exercise 3, related to the MLE regression (left picture):



It seems that the results for the Principal Components Regression (PCR) are not as compelling as with the MLE Regression. Visually we can already observe some more dispersion of observed versus fitted, and the PCR R^2 is lower than the one using the MLE, with $R^2_{MLE} \approx 0.4275 > 0.3327 \approx R^2_{PC}$. Also, the correlation between observed versus fitted is consequently lower with PCR.