Exercises from ‘Regression: likelihood estimation and residuals’

1. Show that for any p × q matrix X, $X^T X$ is positive semi- definite. Hint: work with the definition that A is positive semi- definite iff $x^T Ax \geq 0$ for all x. X_{pxq}

Show that $X^T X$ is positive semi-definite.

$v^T X^T Xv = (Xv)^T Xv = ||Xv||^2 \geq 0$

2. Carry out the “Prostate Cancer” example from Section 3.2.1 of Hastie et al. Having seen the correlations in Table 3.1 do you anticipate the opposite signs for lcp and lcavol on Table 3.2? Comment briefly. Lcp and lcavol are highly correlated (cor(lcp, lcavol) = 0.68) suggesting a case of multi-collinearity*. This high correlation suggests that when regressing lpsa on both, the effects of one may already have been captured by the other.

> In statistics, multicollinearity (also collinearity) is a phenomenon in which two or more predictor variables in a multiple regression model are highly correlated, meaning that one can be linearly predicted from the others with a substantial degree of accuracy. (Wikipedia)

3. “Hat” matrix $H = \phi(\phi^T \phi)^{-1}\phi^T$

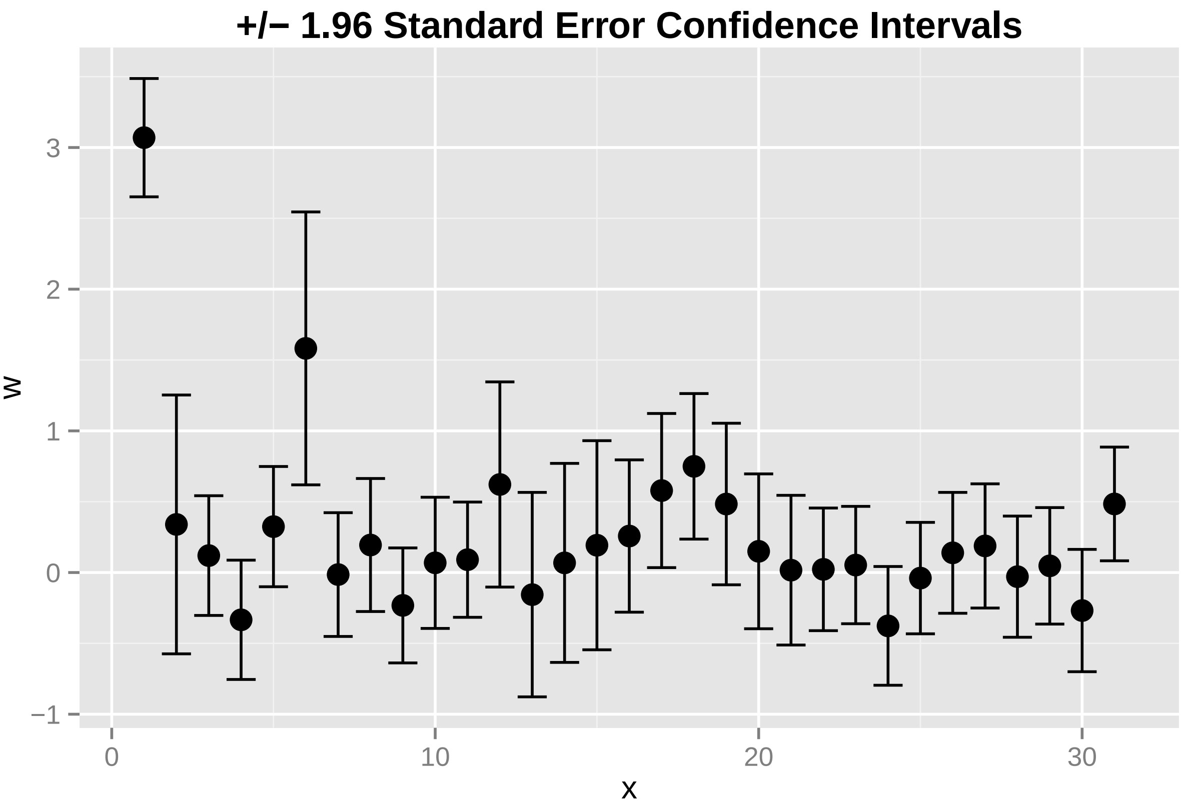
3.1 Show that $H^T = H$. $H^T = (\phi(\phi^T \phi)^{-1}\phi^T)^T = ((\phi)^T)^T((\phi^T \phi)^{-1})^T(\phi)^T = \phi((\phi^T \phi)^T)^{-1}\phi^T = \phi(\phi^T \phi)^{-1}\phi^T$

3.2 Show that $H^2 = H$. $H^2 = \phi(\phi^T \phi)^{-1}\phi^T \phi(\phi^T \phi)^{-1}\phi^T = \phi(\phi^T \phi)^{-1}\phi^T$

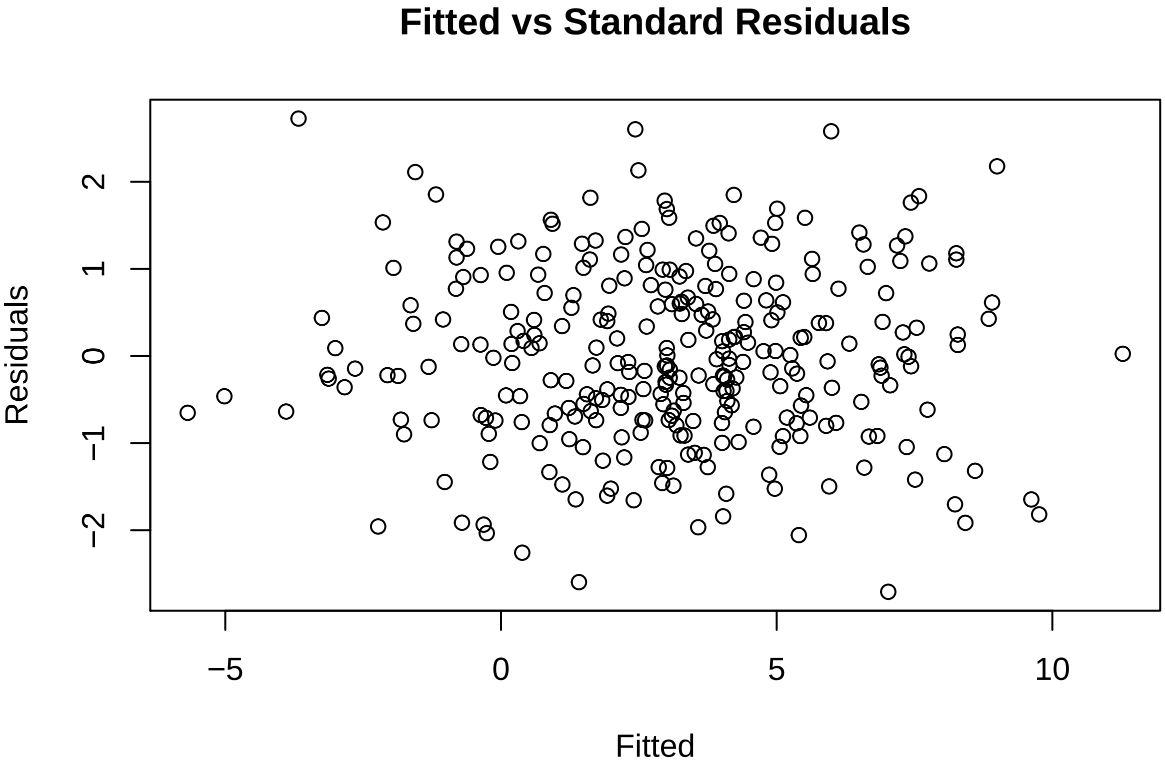
3.3 Show that $tr(H) = M + 1$. $tr(H) = tr(\phi(\phi^T \phi)^{-1}\phi^T) = tr(\phi^T \phi(\phi^T \phi)^{-1}) = tr(I_{(m+1)}) = M + 1.$

4. Extract the first 300 rows and 31 columns from the synthetic_regression.txt dataset; hence you have 300 replications and 30 input variables. Fit a linear regression using MLE in R. Note that a small challenge here is how to set this up in R in presence of a moderately large number of input variables.

4.1: One that shows the 31 estimated coefficients, as points, and ± 1.96 standard errors around them. Plot:

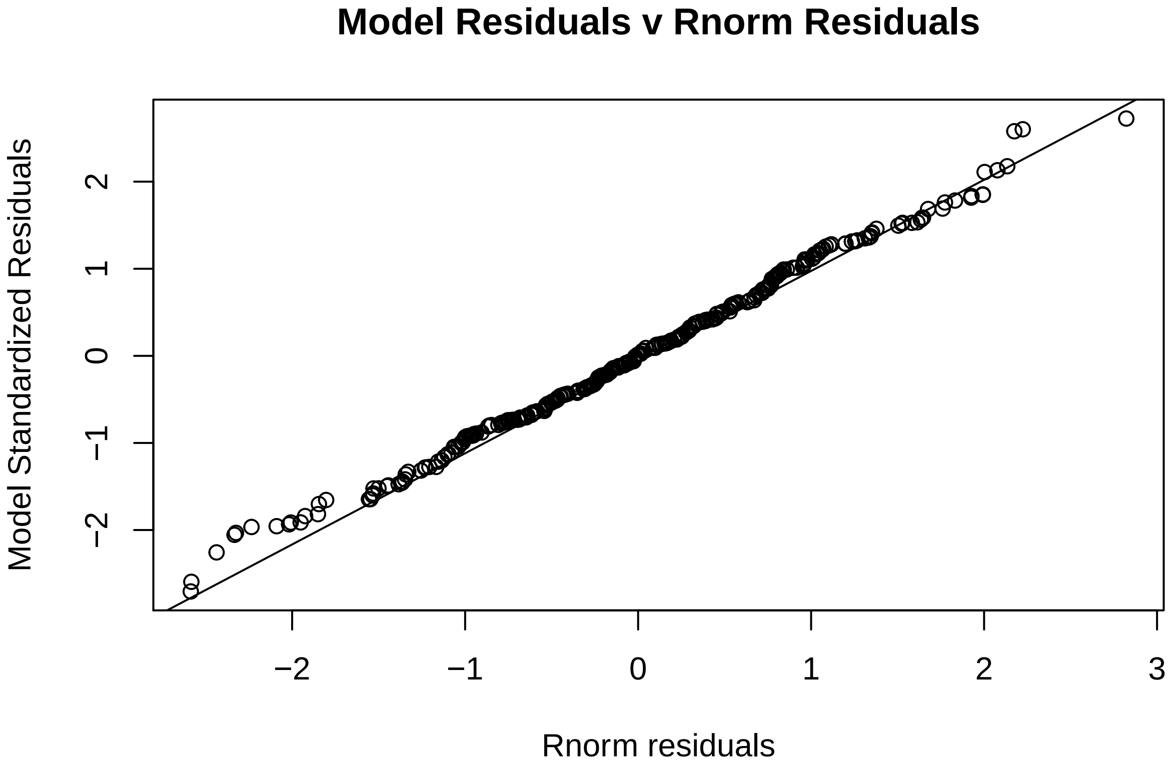


4.2 One that plots standardised residual versus fitted highlighting with red colour the observations with high leverage, bigger that three times 31/300. Plot:



```
## numeric(0)
```

4.3 One that plots quantiles of the standardised residuals vs quantiles of a sample of 300 iid standard Gaussians, with the 45-degree line superimposed. Plot:


Model Residuals v Rnorm Residuals
Model Standardized Residuals
Rnorm residuals