Saarland University

The Elements of Stastical Learning

Assignement 2

Due Date: 15.11.2017

Thibault Schowing
Sarah Mcleod

November 14, 2017

Problem 1

Derive the variance formula:

$$Var(\frac{1}{k}\sum_{i=1}^{k} X_i) = \rho\sigma^2 + \frac{1-\rho}{k}\sigma^2$$

where $X_i, i = 1, ..., k$, are identically distributed random variables with positive pairwise correlation ρ and $Var(X_i) = \sigma^2$ for $i = 1, ..., k$.

Can use the property $Var(aX) = a^2 Var(X)$ to have

$$\frac{1}{k^2} Var(\sum_{i=1}^{k} X_i)$$

And use the property:

$$Var(a_1X_1 + a_2X_2 + ... + a_nX_n) = \sum_{i=1}^{n} a_i^2 Var(X_i) + \sum_{i=1}^{n} \sum_{j=1,j\neq i}^{n} a_i a_j Cov(X_i, X_j)$$

to expand $Var(\frac{1}{k}\sum_{i=1}^{k} X_i)$ which makes

$$Var(\frac{1}{k}\sum_{i=1}^{k} X_i) = \frac{1}{k^2}[\sum_{i=1}^{k} Var(X_i) + \sum_{i=1}^{k} \sum_{j=1,j\neq i}^{k} Cov(X_i, X_j)]$$

Because $Cov(X, Y) = Cov(Y, X)$, we can simplify the double sum

$$Var(\frac{1}{k}\sum_{i=1}^{k} X_i) = \frac{1}{k^2}[\sum_{i=1}^{k} Var(X_i) + k(k-1)Cov(X_i, X_j)]$$

We can now replace the variance with σ^2 and reduce the sum to k:

$$Var(\frac{1}{k}\sum_{i=1}^{k} X_i) = \frac{1}{k^2}[k\sigma^2 + k(k-1)Cov(X_i, X_j)]$$

And for the covariance we have:

$$\rho(X, Y) = \frac{Cov(X, Y)}{sd(X)sd(Y)} = \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

from which we have: $Cov(X_i, X_j) = \rho\sigma^2$ and so

$$Var(\frac{1}{k}\sum_{i=1}^{k} X_i) = \frac{1}{k^2}[k\sigma^2 + k(k-1)\rho\sigma^2]$$

$$Var(\frac{1}{k}\sum_{i=1}^{k}X_i) = \frac{1}{k^2}[k\sigma^2 + k^2\rho\sigma^2 - k\rho\sigma^2]$$

$$Var(\frac{1}{k}\sum_{i=1}^{k}X_i) = \frac{k\sigma^2 + k^2\rho\sigma^2 - k\rho\sigma^2}{k^2}$$

Simplifying we get

$$Var(\frac{1}{k}\sum_{i=1}^{k}X_i) = \frac{k^2\rho\sigma^2}{k^2} + \frac{k\sigma^2(1-\rho)}{k^2}$$

$$Var(\frac{1}{k}\sum_{i=1}^{k}X_i) = \rho\sigma^2 + \frac{(1-\rho)}{k}\sigma^2$$

which was the first statement.

Problem 2

We consider $y = X\beta + \epsilon$ and the least square estimator $\hat{\beta} = (X^T X)^{-1} X^T y$. We assume some arbitrary linear estimator $\tilde{\theta} = c^T y$ is unbiased for $\theta = a^T \beta$. To show that the least square estimate is the best linear unbiased estimate in terms of variance, we need to show

$$Var(\hat{\theta}) \leq Var(\tilde{\theta})$$

We know

$$Var(\tilde{\theta}) = Var(\hat{\theta} - (\tilde{\theta} - \hat{\theta}))$$

By the rules of variance this becomes

$$Var(\tilde{\theta}) = Var(\hat{\theta}) + Var(\tilde{\theta} - \hat{\theta}) + 2Cov(\hat{\theta}, (\tilde{\theta} - \hat{\theta})))$$

Solving for $Var(\tilde{\theta} - \hat{\theta})$ with $\hat{\theta} = a^T M y$, where $M = (X^T X)^{-1} X^T$

$$Var(\tilde{\theta} - \hat{\theta}) = Var(c^T y - a^T M y)$$

$$= Var(c^T - a^T M)y$$

$$= Var(c - aM)^T y$$

From matrix variance properties this becomes

$$= (c - aM)Var(y)(c - aM)^T$$

$$= \sigma^2 (c - aM)(c^T - aM)^T$$

This is > 0 because .... something about positive semidefinite matrices?

Next we calculate $Cov(\hat{\theta}, (\tilde{\theta} - \hat{\theta}))$

$$Cov(\hat{\theta}, (\tilde{\theta} - \hat{\theta})) = Cov(a^T, (c - aM)^T y)$$

Intuitively, the "best" estimator would be the one with the smallest covariance between $\hat{\theta}$ and $\tilde{\theta}$. Showing this is where we got stuck.
Assuming you've proved this, then

$$Var(\tilde{\theta}) = Var(\hat{\theta}) + Var(\tilde{\theta} - \hat{\theta})$$

$$Var(\hat{\theta}) \leq Var(\hat{\theta}) + Var(\tilde{\theta} - \hat{\theta})$$

$$Var(\hat{\theta}) \leq Var(\tilde{\theta})$$

Problem 3

TheR^2 statistic is a common measure of model fit corresponding to the fraction of variance in the data that is explained by the model. In general, R^2 is given by the formula

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

The objective here is to show that for univariate regression, $R^2 = Cor(X,Y)^2$

Let's take first the RSS and TSS formula

$RSS = \sum\limits_{i=1}^{n}(y_i - \hat{y}_i)^2$ =amount of variability left unaccounted after the regression

$TSS = \sum\limits_{i=1}^{n}(y_i - \bar{y}_i)^2$ = total variance in Y

$TSS - RSS$ = amount of variance removed/explained by the regression

So we have

$$R^2 = 1 - \frac{\sum\limits_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum\limits_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

And the correlation formula

$$\widehat{Cor}(X,Y) = \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum\limits_{i=1}^{n}(y_i - \bar{y})^2}}$$

For an univariate linear regression, the approximation has the form $(af+b)$ and we'll suppose an exact approximation with $(y-f)^2 = 0$. We will also assume that f is the model that minimize the square-error and that there is no shift of scaling that improve f.

We can write it like this

$$\sum_{i=1}^{n}(f_i - y_i)^2 \leq \sum_{i=1}^{n}(af_i + b - y_i)^2$$

Consider the second part as $g(a,b) = \sum\limits_{i=1}^{n}(af_i + b - y_i)^2$. As f minimizes the loss, g is optimized at $g(1,0)$ and so it is the optimum. We can derive it

$$\frac{d}{da}g(a,b)_{(a=1,b=0)} = \sum_{i=1}^{n}2(af_i + b - y_i)f_i = \sum_{i=1}^{n}2(f_i - y_i)f_i = 0$$

So $ff - yf = 0 \rightarrow yf = ff$

And

$$\frac{d}{db}g(a,b)_{(a=1,b=0)} = \sum_{i=1}^{n} 2(af_i + b - y_i) = \sum_{i=1}^{n} 2(f_i - y_i) = 0$$

So $\bar{y} = \bar{f} \rightarrow$ Mean is the normalized sum.

We can simplify R^2. From the derivative we have $yf = ff$ and $\bar{f} = \bar{y}$. To simplify we assume $\bar{y}$ and $\bar{f} = 0$

$$R^2 = 1 - \frac{\sum\limits_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum\limits_{i=1}^{n}(y_i)^2} = 1 - \frac{yy - 2yf + ff}{yy}$$

Since we have $yf = ff$

$$R^2 = 1 - \frac{yy - ff}{yy} = \frac{ff}{yy}$$

For the correlation we have

$$\rho = \frac{\sum\limits_{i=1}^{n}(f_i y_i)}{\sqrt{\sum\limits_{i=1}^{n} f_i^2}\sqrt{\sum\limits_{i=1}^{n} y_i^2}} = \frac{fy}{\sqrt{(ff)(yy)}} = \frac{ff}{\sqrt{(ff)(yy)}} = \sqrt{\frac{ff}{yy}}$$

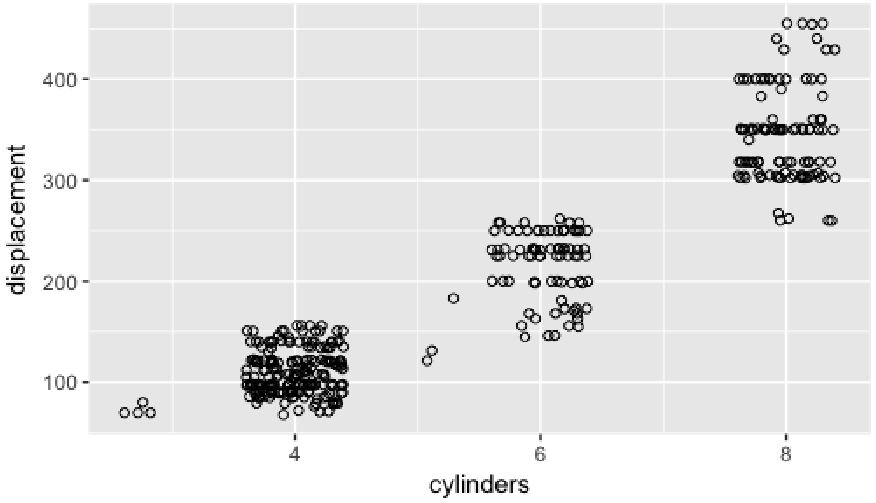
So we have $R^2 = \frac{ff}{yy}$ and $\rho = \sqrt{\frac{ff}{yy}}$

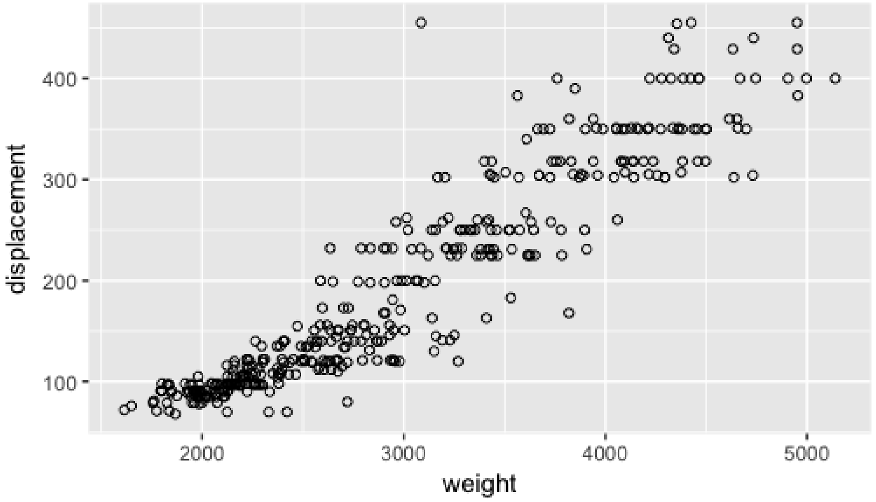
So $R^2 = \rho^2$

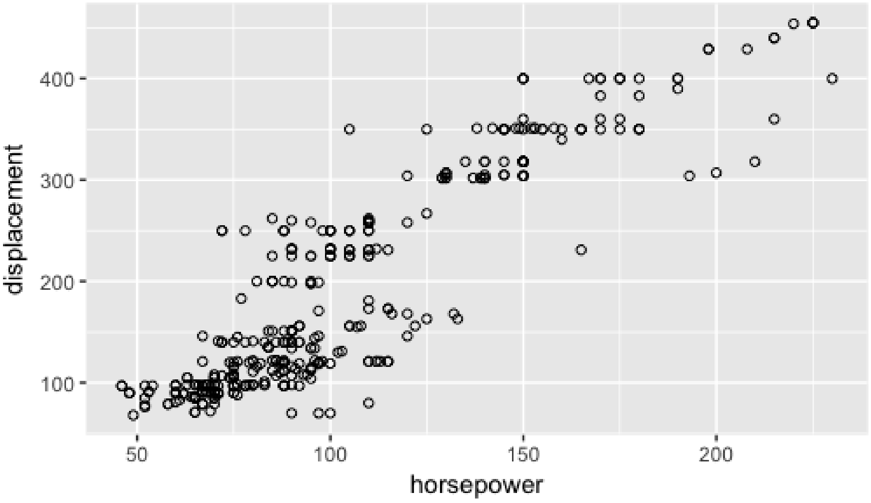
Problem 4

1. From the correlation matrix it looks like displacement, weight, and horsepower all have a high positive correlation with cylinders. Given what we know about cars this makes intuitive sense. Horsepower is most positively correlated with displacement and mpg is most negatively correlated with weight, which, again, makes intuitive sense.

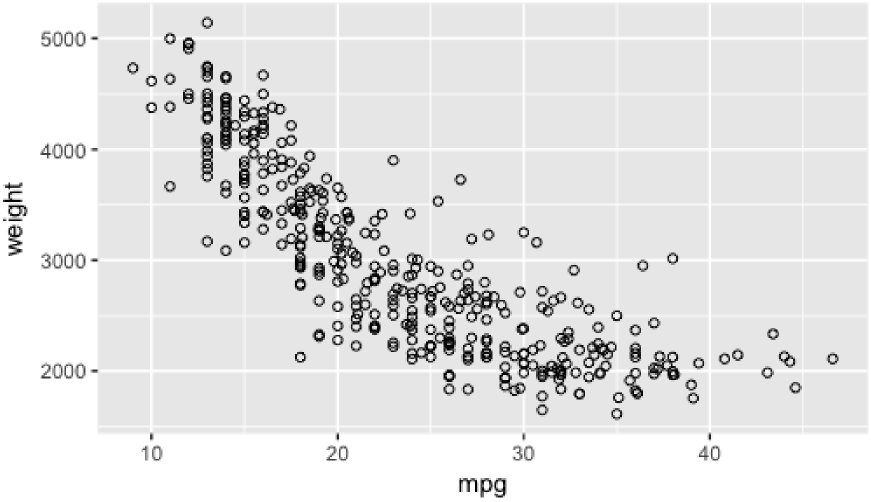
2. The scatter plots for three highly correlated variables are below.

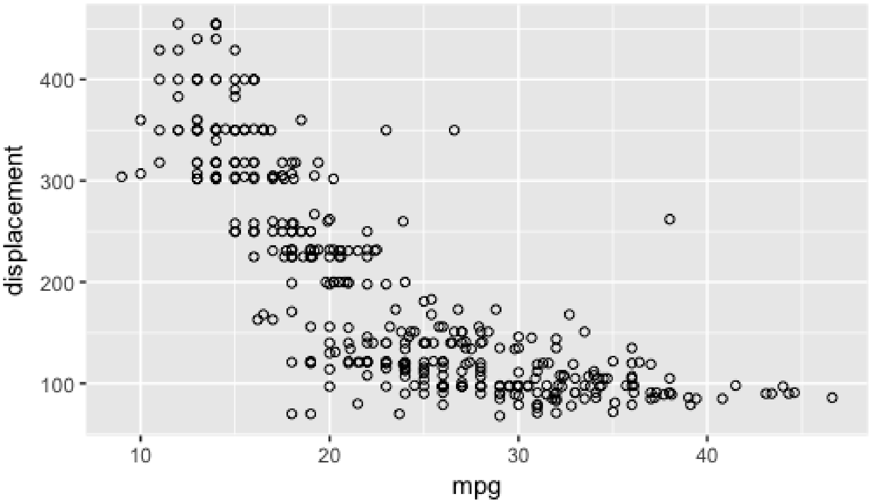


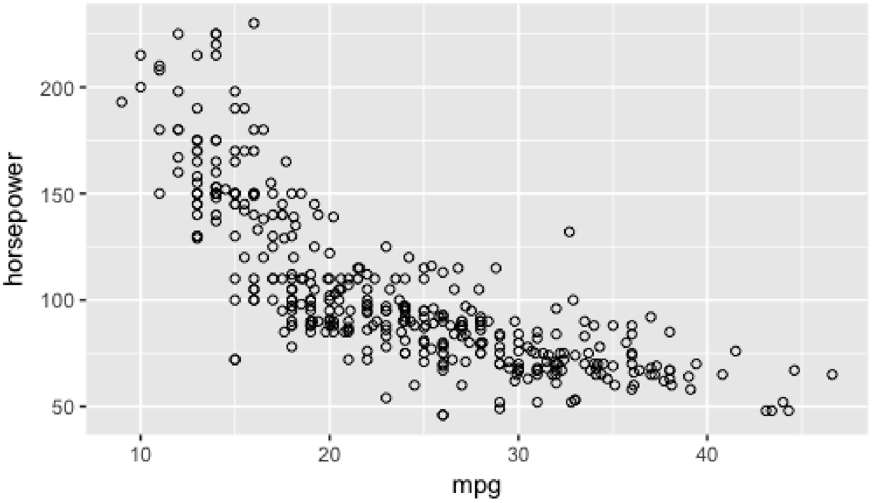




Scatterplots for three highly anti-correlated variables are below.




displacement
400
300
200
100
10
20
30
40
mpg


horsepower
200
150
100
50
10
20
30
40
mpg

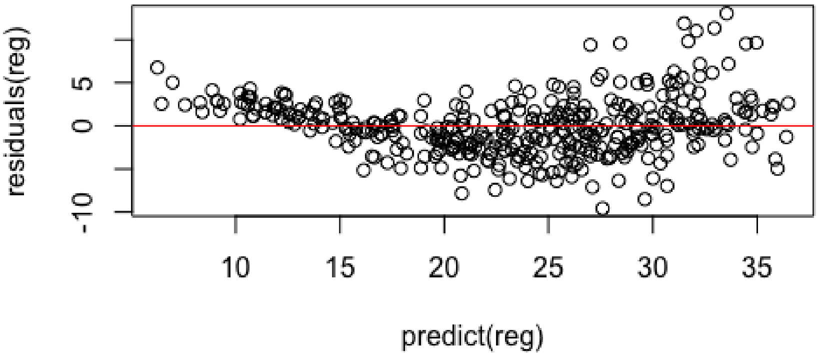
The number of cylinders, the weight of the vehicle, and the horsepower have the highest correlations with displacement, respectively. Some of these numbers make intuitive sense from what we know about how internal combustion engines work. If you want more displacement adding more cylinders is one option to achieve this. Weight, displacement and horsepower are also highly anti-correlated with mpg. So, as the vehicle gets heavier, or gets an increase in power, the efficiency of the engine reduces and you get fewer miles per gallon.

3. The p-values show that all of the predictors have a statistically significant relationship to the outcome. From the R^2 values displacement as a predictor is a more accurate model. Also, year as a predictor is the least accurate model.

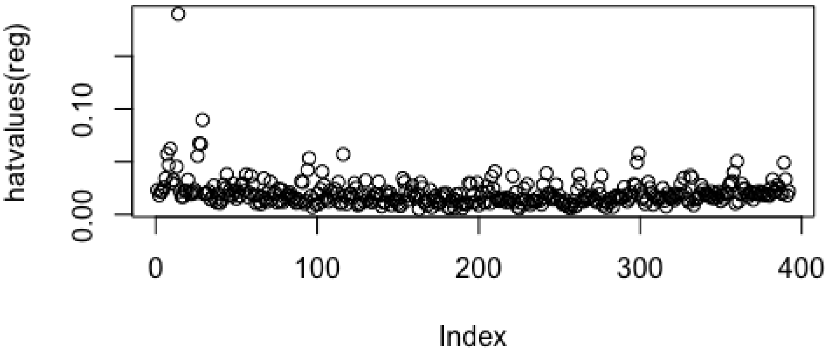
Run	Type	R^2	p-value
mpg	cyl	0.6047	
mpg	disp	0.6482	
mpg	hrsp	0.6059	
mpg	year	0.337	

4. The R^2 is 0.8125, which is close to 1. This suggests that this model is an accurate model. Cylinders and horsepower were both statistically significant predictor alone, but in multivariate regression they are not. The coefficients describe the size of the effect of the predictor on the response variable, therefore a negative coefficient gives the amount of decrease in the response for every increase in the predictor.

5. The residual plot does suggest a non-linearity in the data. The far left side of the plot shows consistent over prediction (i.e. bias) in the data. There don't seem to be any unusually large outliers. There is a high leverage point, point at index 14, as seen in the image below.



6. The transformations with other variables like weight, year or cylinder improve the result (R^2 value) but transforming displacement with itself does not make the model better. Doing transformations on one predictor doesn't give you as much benefits as using more predictors.


hatvalues(reg)
0.00
0.10
0
100
200
300
400
Index