WilloughbyHW1

Bryant Willoughby

2025-09-10

Problem 1 Instructions

The dataset `teengamb` concerns a study of teenage gambling in Britain. After you install R and the `faraway` package, you can load the faraway package and teengamb data.

1. Make a numerical and graphical summary of the data, commenting on any features that you find interesting. For the variable "sex", assign the label "male" and "female" and ask R to treat it as a categorical variable before you compute the summary.

Solution to this question should be no longer than 1.5 pages.

2. In your summary, report the means and medians of variables "income" and "gamble". Please comment on the relative locations of their means and medians, and explain why the means are larger than the medians.
3. How many different values are there for the variable "verbal"? (hint: help(unique)) Based on the boxplot (and any other ways you can define and explain) of the variable "verbal", what could be the possible values of outlying verbal scores?
4. Suppose you are interested in how variables, such as "verbal", "income" and "gamble" differ for different "sex". Use numerical and/or graphical tool(s) to summarize the data for this purpose, commenting on any features that you find interesting. Limit the output you present to a quantity that a busy reader would find sufficient to get a basic understanding of your answer.

Hints: Some useful R functions for this homework can be found in the course documents that are available on Canvas. You can always type help(subject) to get detailed help on the subject, e.g. help(plot). Or you can type help.start() to get interactive help with a search engine.

Problem 1: Data loading and factor variable

```r
# Load faraway package and teengamb data
library(faraway)
```

```
## Warning: package 'faraway' was built under R version 4.4.3
```

```r
data(teengamb)

# Create factor variable for sex with labels
teengamb$sex <- factor(teengamb$sex, levels = c(0, 1), labels = c("male", "female"))
str(teengamb)  # Check structure
```

```
## 'data.frame':    47 obs. of  5 variables:
##  $ sex   : Factor w/ 2 levels "male","female": 2 2 2 2 2 2 2 2 2 2 ...
##  $ status: int  51 28 37 28 65 61 28 27 43 18 ...
##  $ income: num  2 2.5 2 7 2 3.47 5.5 6.42 2 6 ...
##  $ verbal: int  8 8 6 4 8 6 7 5 6 7 ...
##  $ gamble: num  0 0 0 7.3 19.6 0.1 1.45 6.6 1.7 0.1 ...
```

(1) Numerical and graphical summary

The summary measures and histograms are provided for the numeric variables. `Status` appears approximately uniform, `verbal` approximately normal, and `income` and `gamble` both right-skewed. There are more male counts (28) versus female counts (19). A pairwise correlation plot among the numeric variables is also provided. `Status` and `verbal` appear most strongly correlated.

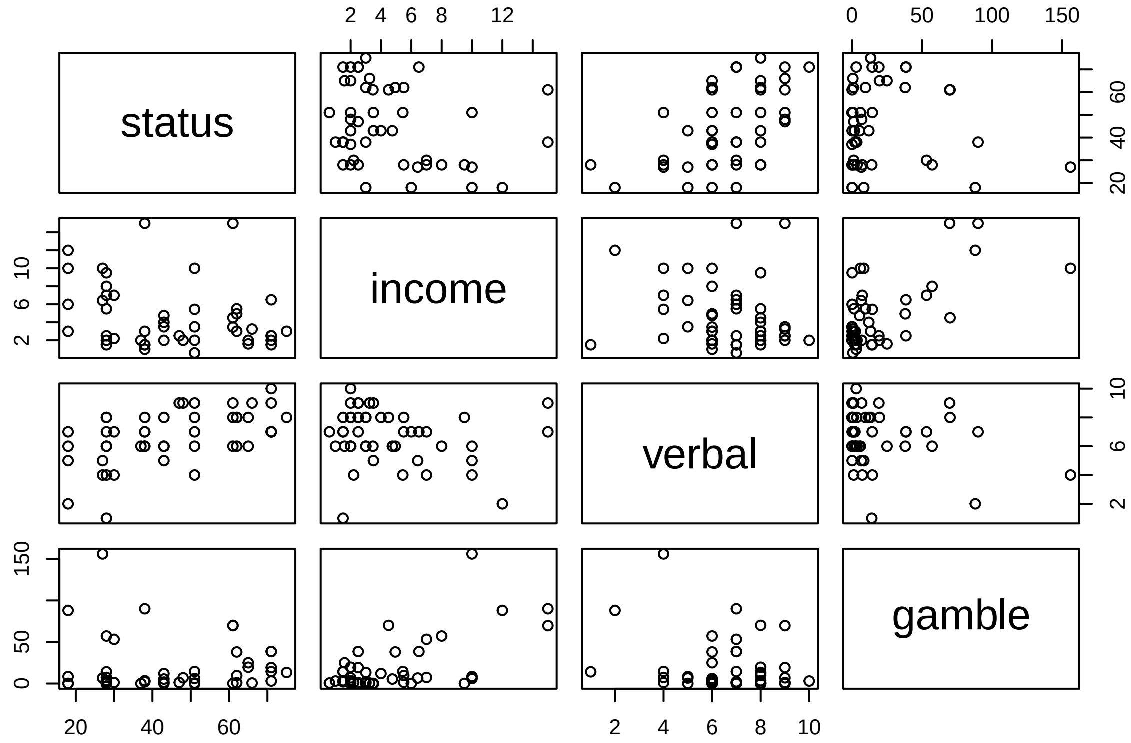
```r
# Numerical summary for numeric variables
summary(teengamb[c("status", "income", "verbal", "gamble")])
```

```
##      status          income           verbal          gamble     
##  Min.   :18.00   Min.   : 0.600   Min.   : 1.00   Min.   :  0.0  
##  1st Qu.:28.00   1st Qu.: 2.000   1st Qu.: 6.00   1st Qu.:  1.1  
##  Median :43.00   Median : 3.250   Median : 7.00   Median :  6.0  
##  Mean   :45.23   Mean   : 4.642   Mean   : 6.66   Mean   : 19.3  
##  3rd Qu.:61.50   3rd Qu.: 6.210   3rd Qu.: 8.00   3rd Qu.: 19.4  
##  Max.   :75.00   Max.   :15.000   Max.   :10.00   Max.   :156.0
```

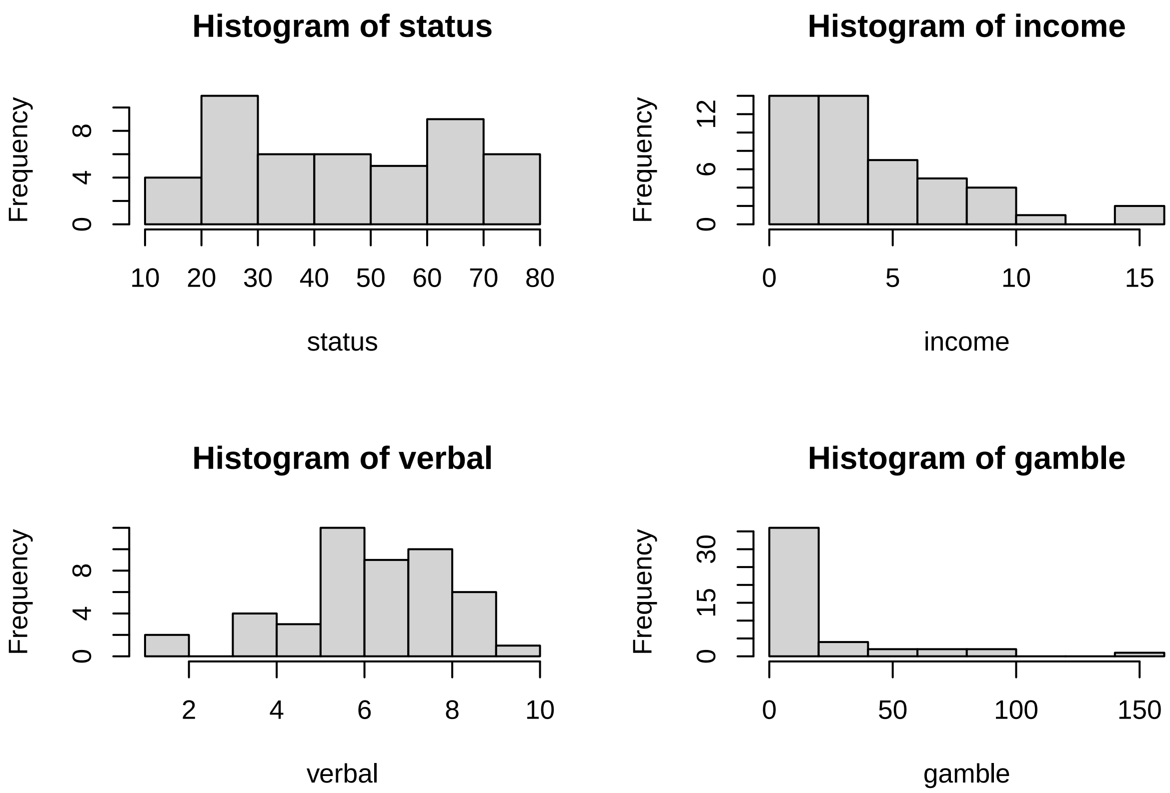
```r
# Frequency table for sex
table(teengamb$sex)
```

```
## 
##   male female 
##     28     19
```

```r
# Pairs plot for numeric variables
pairs(teengamb[c("status", "income", "verbal", "gamble")])
```



```r
# Histograms for numeric variables
par(mfrow = c(2, 2))  # Layout for 4 variables
for (v in c("status", "income", "verbal", "gamble")) {
  hist(teengamb[[v]], main = paste("Histogram of", v), xlab = v)
}
```



```r
par(mfrow = c(1, 1))  # Reset layout
```

(2) Means and medians of income and gamble

The respective mean and median for `income` and `gamble`variables are provided below. Notice that the mean is pulled towards the right skewness, whereas the median measure is not influenced by this skewness and is therefore smaller in comparison.

```r
# Means and medians of income and gamble with descriptive labels
paste("Mean income:", mean(teengamb$income))
```

```
## [1] "Mean income: 4.64191489361702"
```

```r
paste("Median income:", median(teengamb$income))
```

```
## [1] "Median income: 3.25"
```

```r
paste("Mean gamble:", mean(teengamb$gamble))
```

```
## [1] "Mean gamble: 19.3010638297872"
```

```
paste("Median gamble:", median(teengamb$gamble))
```

```
## [1] "Median gamble: 6"
```

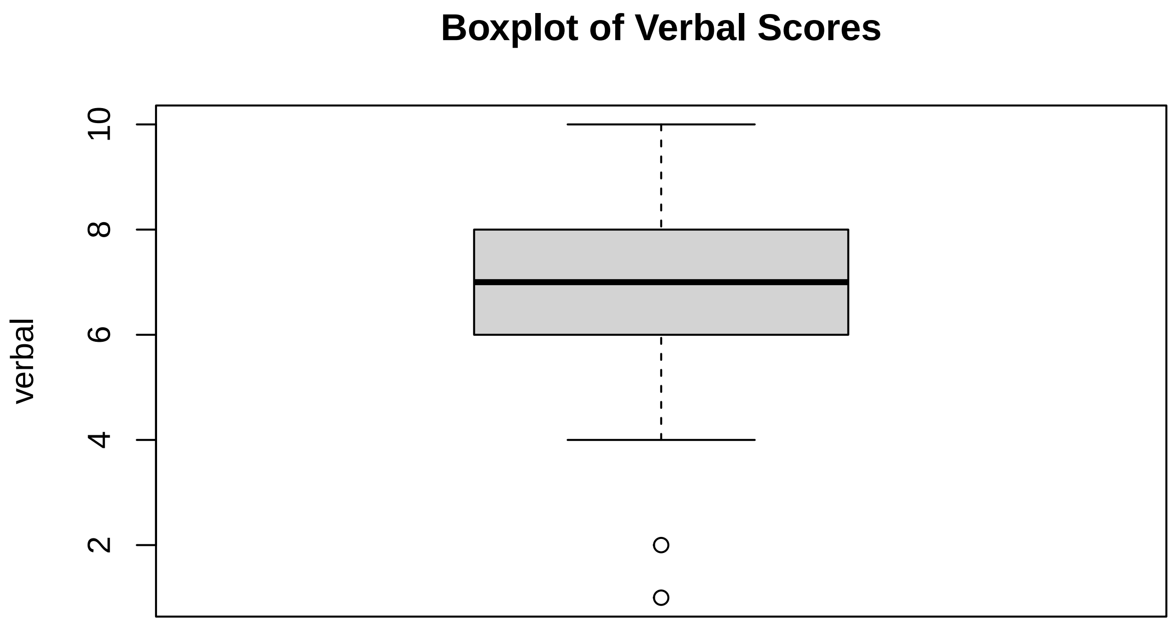
(3) Unique values and boxplot for verbal

There are nine unique values for the `verbal` variable. The two possible outlying verbal scores are one and two, as illustrated in the below boxplot.

```
# Number of unique values for verbal
length(unique(teengamb$verbal))
```

```
## [1] 9
```

```
# Boxplot for verbal
boxplot(teengamb$verbal, main = "Boxplot of Verbal Scores", ylab = "verbal")
```



(4) Compare verbal, income, and gamble by sex

The measures below group the `verbal`, `income` and `gamble` variables according to sex. The distributions appear different across sex for all three variables, with a difference in numeric measures (consult summary statistics). The boxplots provide a visual assessment of this information.

```r
# Numerical summaries of verbal scores by sex
print("Summary of verbal scores by sex:")
```

```
## [1] "Summary of verbal scores by sex:"
```

```r
tapply(teengamb$verbal, teengamb$sex, summary)
```

```
## $male
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   1.000   6.000   7.000   6.821   8.250  10.000 
## 
## $female
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   4.000   6.000   6.000   6.421   8.000   8.000
```

```r
# Numerical summaries of income by sex
print("Summary of income by sex:")
```

```
## [1] "Summary of income by sex:"
```

```r
tapply(teengamb$income, teengamb$sex, summary)
```

```
## $male
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.600   2.000   3.375   4.976   6.625  15.000 
## 
## $female
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   1.500   2.000   3.000   4.149   5.750  10.000
```

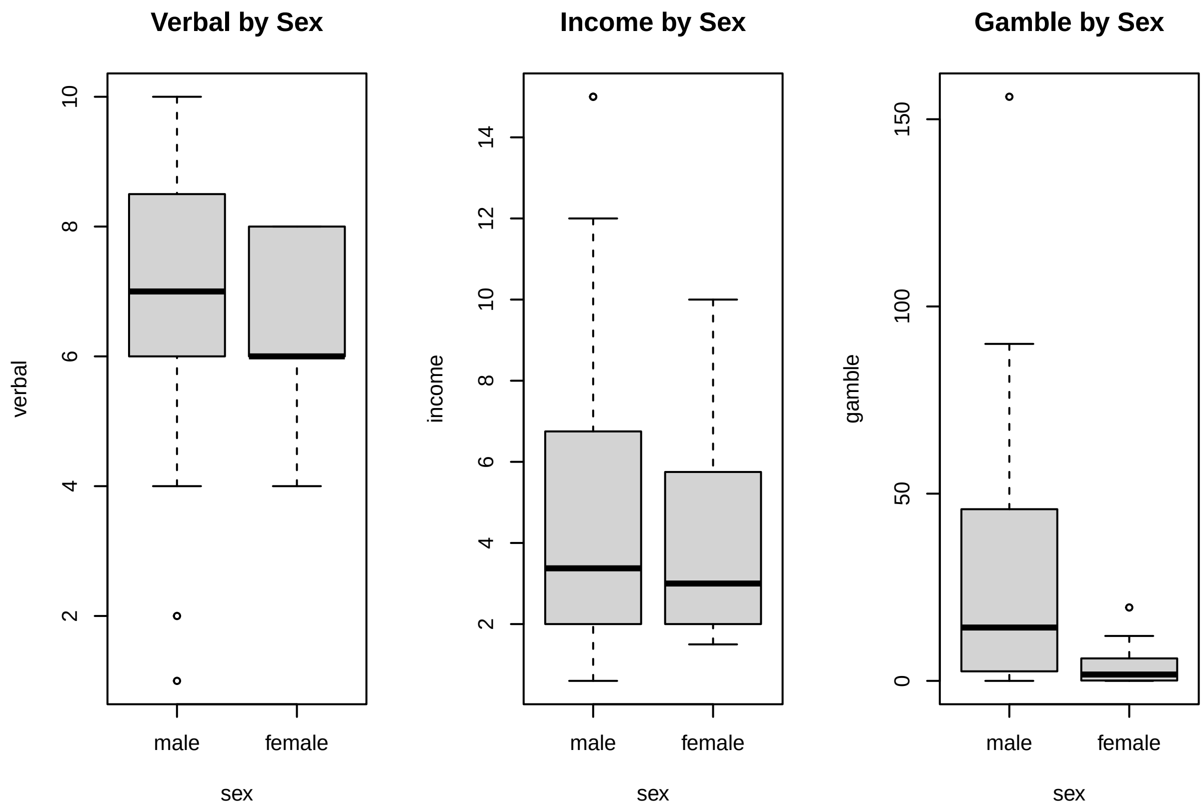
```r
# Numerical summaries of gambling expenditure by sex
print("Summary of gambling expenditure by sex:")
```

```
## [1] "Summary of gambling expenditure by sex:"
```

```r
tapply(teengamb$gamble, teengamb$sex, summary)
```

```
## $male
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.000   2.775  14.250  29.775  42.175 156.000 
## 
## $female
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.000   0.100   1.700   3.866   6.000  19.600
```

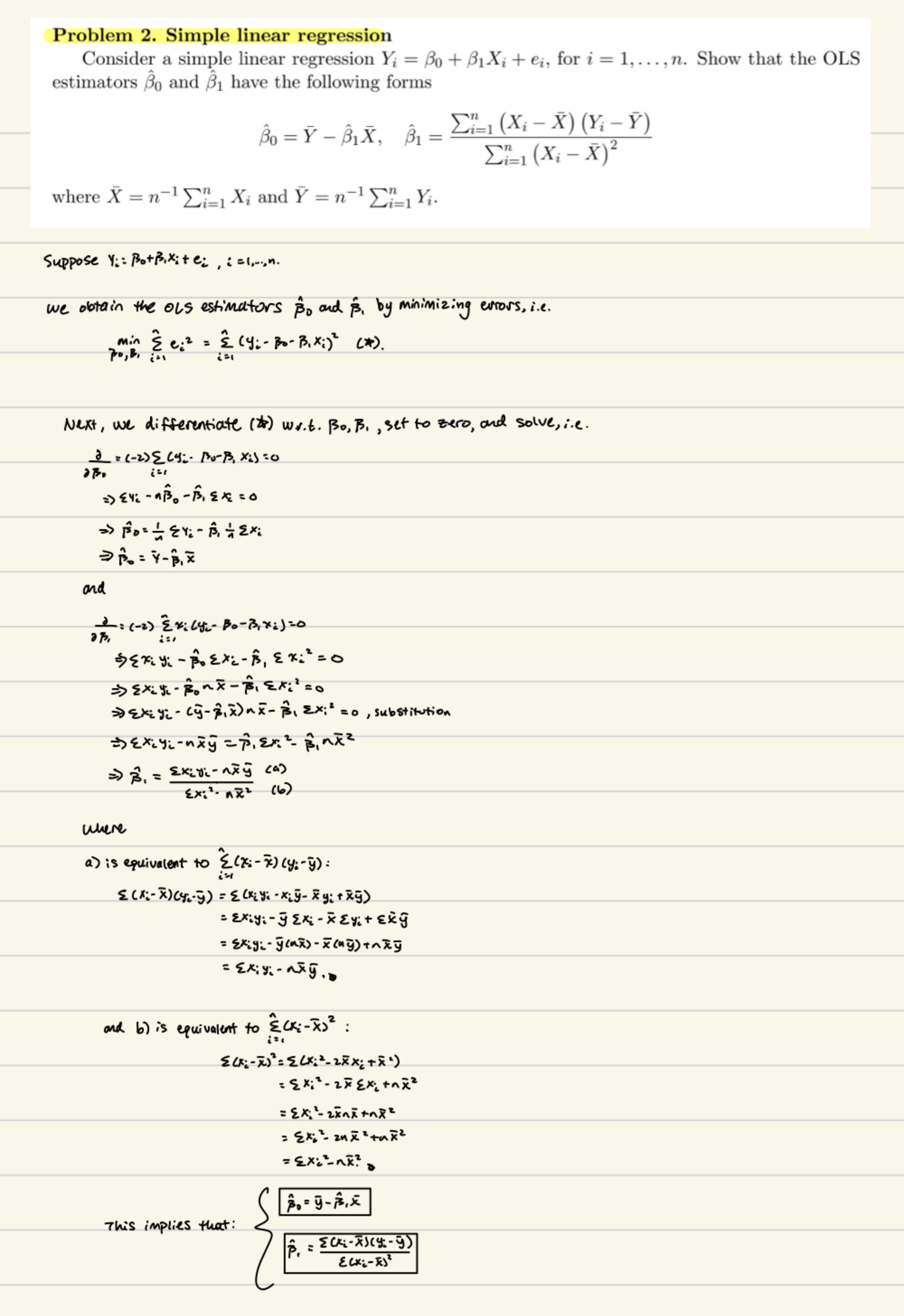
```r
# Boxplots by sex
par(mfrow = c(1, 3))
boxplot(verbal ~ sex, data = teengamb, main = "Verbal by Sex")
boxplot(income ~ sex, data = teengamb, main = "Income by Sex")
boxplot(gamble ~ sex, data = teengamb, main = "Gamble by Sex")
```



```r
par(mfrow = c(1, 1))
```

Problem 2

Problem 3



Problem 2. Simple linear regression

Consider a simple linear regression $Y_i = \beta_0 + \beta_1 X_i + e_i$, for $i = 1, \ldots, n$. Show that the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ have the following forms

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}, \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}$$

where $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$ and $\bar{Y} = n^{-1} \sum_{i=1}^{n} Y_i$.

Suppose $Y_i = \beta_0 + \beta_1 X_i + e_i$, $i = 1, \ldots, n$.

we obtain the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ by minimizing errors, i.e.

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \quad (*).$$

Next, we differentiate $(*)$ w.r.t. β_0, β_1 , set to zero, and solve, i.e.

$$\frac{\partial}{\partial \beta_0} = (-2) \sum_{i=1} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\Rightarrow \Sigma y_i - n\hat{\beta}_0 - \hat{\beta}_1 \Sigma x_i = 0$$

$$\Rightarrow \hat{\beta}_0 = \frac{1}{n} \Sigma y_i - \hat{\beta}_1 \frac{1}{n} \Sigma x_i$$

$$\Rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

and

$$\frac{\partial}{\partial \beta_1} = (-2) \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\Rightarrow \Sigma x_i y_i - \hat{\beta}_0 \Sigma x_i - \hat{\beta}_1 \Sigma x_i^2 = 0$$

$$\Rightarrow \Sigma x_i y_i - \hat{\beta}_0 n\bar{x} - \hat{\beta}_1 \Sigma x_i^2 = 0$$

$$\Rightarrow \Sigma x_i y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) n\bar{x} - \hat{\beta}_1 \Sigma x_i^2 = 0 \text{ , substitution}$$

$$\Rightarrow \Sigma x_i y_i - n\bar{x}\bar{y} = \hat{\beta}_1 \Sigma x_i^2 - \hat{\beta}_1 n\bar{x}^2$$

$$\Rightarrow \hat{\beta}_1 = \frac{\Sigma x_i y_i - n\bar{x}\bar{y} \quad (a)}{\Sigma x_i^2 - n\bar{x}^2 \quad (b)}$$

where

a) is equivalent to $\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$:

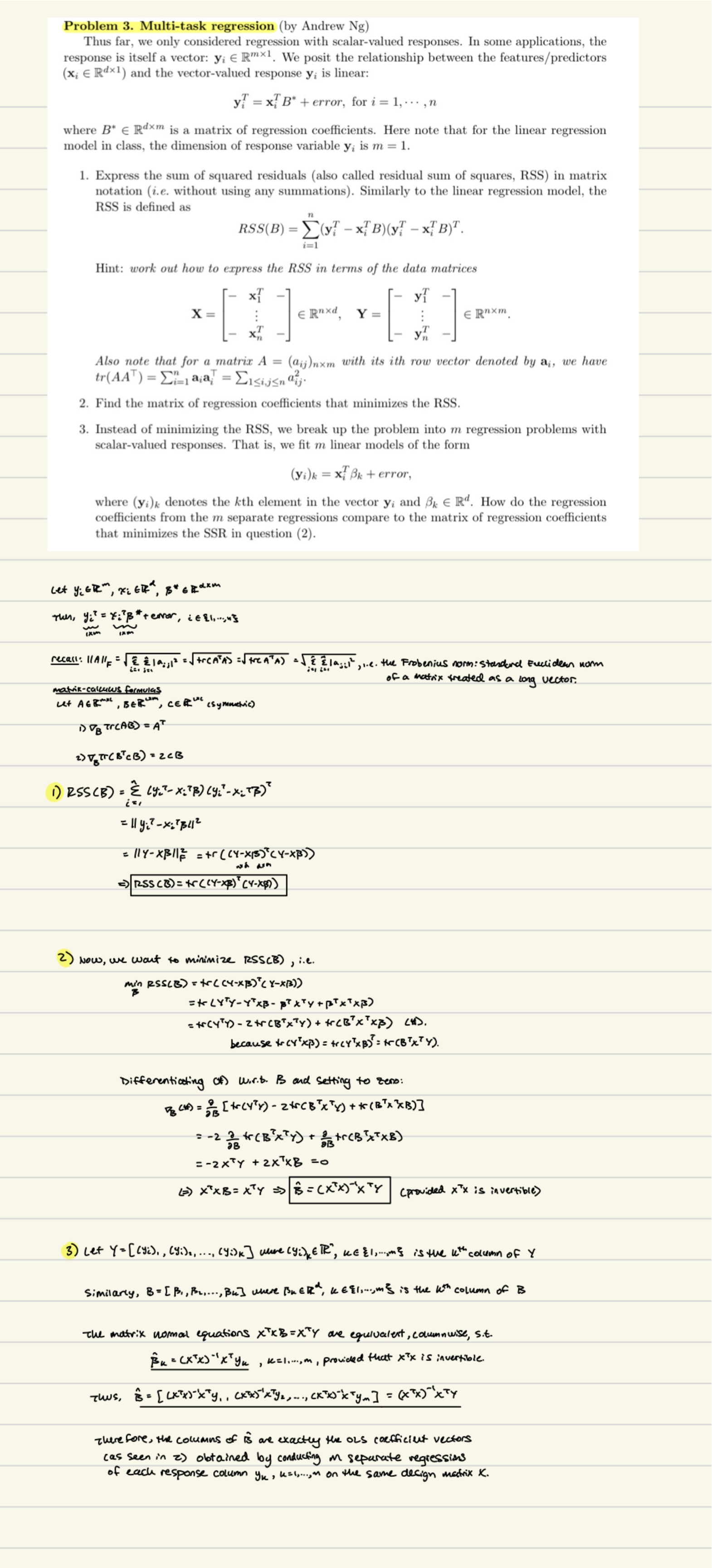
$$\begin{aligned} \Sigma (x_i - \bar{x})(y_i - \bar{y}) &= \Sigma (x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x}\bar{y}) \\ &= \Sigma x_i y_i - \bar{y} \Sigma x_i - \bar{x} \Sigma y_i + \Sigma \bar{x}\bar{y} \\ &= \Sigma x_i y_i - \bar{y}(n\bar{x}) - \bar{x}(n\bar{y}) + n\bar{x}\bar{y} \\ &= \Sigma x_i y_i - n\bar{x}\bar{y}. \end{aligned}$$

and b) is equivalent to $\sum_{i=1}^{n} (x_i - \bar{x})^2$:

$$\begin{aligned} \Sigma (x_i - \bar{x})^2 &= \Sigma (x_i^2 - 2\bar{x} x_i + \bar{x}^2) \\ &= \Sigma x_i^2 - 2\bar{x} \Sigma x_i + n\bar{x}^2 \\ &= \Sigma x_i^2 - 2\bar{x} n\bar{x} + n\bar{x}^2 \\ &= \Sigma x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 \\ &= \Sigma x_i^2 - n\bar{x}^2. \end{aligned}$$

This implies that: $\left\{ \begin{array}{l} \boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}} \\ \\ \boxed{\hat{\beta}_1 = \dfrac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\Sigma (x_i - \bar{x})^2}} \end{array} \right.$

Figure 1: Problem 2 Solution



Problem 3. Multi-task regression (by Andrew Ng)

Thus far, we only considered regression with scalar-valued responses. In some applications, the response is itself a vector: $\mathbf{y}_i \in \mathbb{R}^{m\times 1}$. We posit the relationship between the features/predictors ($\mathbf{x}_i \in \mathbb{R}^{d\times 1}$) and the vector-valued response $\mathbf{y}_i$ is linear:

$$\mathbf{y}_i^T = \mathbf{x}_i^T B^* + error, \text{ for } i = 1, \cdots, n$$

where $B^* \in \mathbb{R}^{d\times m}$ is a matrix of regression coefficients. Here note that for the linear regression model in class, the dimension of response variable $\mathbf{y}_i$ is $m = 1$.

1. Express the sum of squared residuals (also called residual sum of squares, RSS) in matrix notation (*i.e.* without using any summations). Similarly to the linear regression model, the RSS is defined as

$$RSS(B) = \sum_{i=1}^{n} (\mathbf{y}_i^T - \mathbf{x}_i^T B)(\mathbf{y}_i^T - \mathbf{x}_i^T B)^T.$$

Hint: *work out how to express the RSS in terms of the data matrices*

$$\mathbf{X} = \begin{bmatrix} - & \mathbf{x}_1^T & - \\ & \vdots & \\ - & \mathbf{x}_n^T & - \end{bmatrix} \in \mathbb{R}^{n\times d}, \quad \mathbf{Y} = \begin{bmatrix} - & \mathbf{y}_1^T & - \\ & \vdots & \\ - & \mathbf{y}_n^T & - \end{bmatrix} \in \mathbb{R}^{n\times m}.$$

Also note that for a matrix $A = (a_{ij})_{n\times m}$ *with its ith row vector denoted by* $\mathbf{a}_i$, *we have* $tr(AA^\top) = \sum_{i=1}^n \mathbf{a}_i\mathbf{a}_i^\top = \sum_{1\le i,j\le n} a_{ij}^2$.

2. Find the matrix of regression coefficients that minimizes the RSS.

3. Instead of minimizing the RSS, we break up the problem into m regression problems with scalar-valued responses. That is, we fit m linear models of the form

$$(\mathbf{y}_i)_k = \mathbf{x}_i^T \beta_k + error,$$

where $(\mathbf{y}_i)_k$ denotes the kth element in the vector $\mathbf{y}_i$ and $\beta_k \in \mathbb{R}^d$. How do the regression coefficients from the m separate regressions compare to the matrix of regression coefficients that minimizes the SSR in question (2).

Let $y_i \in \mathbb{R}^m$, $x_i \in \mathbb{R}^d$, $\beta^* \in \mathbb{R}^{d\times m}$

Thus, $\underbrace{y_i^T}_{1\times m} = \underbrace{x_i^T\beta^*}_{1\times m} + error$, $i \in \{1, \dots, n\}$

recall: $\|A\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}|^2} = \sqrt{tr(A^TA)} = \sqrt{tr(A^TA)} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}|^2}$, i.e. the Frobenius norm: standard Euclidean norm of a matrix treated as a long vector.

matrix-calculus formulas

Let $A \in \mathbb{R}^{m\times c}$, $B \in \mathbb{R}^{c\times m}$, $C \in \mathbb{R}^{c\times c}$ (symmetric)

1) $\nabla_B Tr(AB) = A^T$

2) $\nabla_B Tr(B^TCB) = 2CB$

1) $RSS(B) = \sum_{i=1}^{n} (y_i^T - x_i^T\beta)(y_i^T - x_i^T\beta)^T$

$= \| y_i^T - x_i^T\beta\|^2$

$= \|Y - X\beta\|_F^2 = tr(\underset{m\times n}{(Y-X\beta)^T}\underset{n\times m}{(Y-X\beta)})$

$\Rightarrow \boxed{RSS(B) = tr((Y-X\beta)^T(Y-X\beta))}$

2) Now, we want to minimize $RSS(B)$, i.e.

$\min_{\beta} RSS(B) = tr((Y-X\beta)^T(Y-X\beta))$

$= tr(Y^TY - Y^TX\beta - \beta^TX^TY + \beta^TX^TX\beta)$

$= tr(Y^TY) - 2tr(B^TX^TY) + tr(B^TX^TX\beta)$ (*).

because $tr(Y^TX\beta) = tr(Y^TXB)^T = tr(B^TX^TY)$.

Differentiating (*) w.r.t. B and setting to zero:

$\nabla_B(*) = \frac{\partial}{\partial B}[tr(Y^TY) - 2tr(B^TX^TY) + tr(B^TX^TXB)]$

$= -2\frac{\partial}{\partial B}tr(B^TX^TY) + \frac{\partial}{\partial B}tr(B^TX^TXB)$

$= -2X^TY + 2X^TXB = 0$

$\Leftrightarrow X^TXB = X^TY \Rightarrow \boxed{\hat{B} = (X^TX)^{-1}X^TY}$ (provided X^TX is invertible)

3) Let $Y = [(y_i)_1, (y_i)_2, \dots, (y_i)_K]$ where $(y_i)_k \in \mathbb{R}^n$, $k \in \{1, \dots, m\}$ is the k^{th} column of Y

Similarly, $B = [\beta_1, \beta_2, \dots, \beta_k]$ where $\beta_k \in \mathbb{R}^d$, $k \in \{1, \dots, m\}$ is the k^{th} column of B

The matrix normal equations $X^TXB = X^TY$ are equivalent, columnwise, s.t.

$\underline{\hat{\beta}_k = (X^TX)^{-1}X^Ty_k}$, $k = 1, \dots, m$, provided that X^TX is invertible.

Thus, $\underline{\hat{B} = [(X^TX)^{-1}X^Ty_1, (X^TX)^{-1}X^Ty_2, \dots, (X^TX)^{-1}X^Ty_m] = (X^TX)^{-1}X^TY}$

Therefore, the columns of $\hat{B}$ are exactly the OLS coefficient vectors (as seen in 2) obtained by conducting m separate regressions of each response column y_k, $k = 1, \dots, m$ on the same design matrix X.

Figure 2: Problem 3 Solution