Math 423 Assignment 2

Jonathan Pearce 260672004

10/27/2017

```
library(MASS)
file1<-"http://www.math.mcgill.ca/yyang/regression/data/salary.csv"
#code from assginment document
salary<-read.csv(file1,header=TRUE)
x1<-salary$SPENDING/1000
y<-salary$SALARY
fit.Salary<-lm(y~x1)
sum<-summary(fit.Salary)
#sum
```

Question a.

Intercept Estimate Calculation

```
n <- length(x1)
p <- 2

intercept = 0
intercept = (1/(sum(x1^2) - (1/n)*(sum(x1)^2))) * (mean(y)*(sum(x1^2)) - mean(x1)*sum(x1*y))
intercept
```

```
## [1] 12129.37
```

Slope Estimate Calculation

```
slope = 0
slope = (1/(sum(x1^2) - (1/n)*(sum(x1)^2))) * (sum(x1*y) - sum(x1)*sum(y)/n)
slope
```

```
## [1] 3307.585
```

We have estimates $\beta_0 = 12129.371$ and $\beta_1 = 3307.585$. These values match the estimates column in the summary output.

Question b.

Residual Standard Error Calculation

```
ssresidual = sum((y - (intercept + slope*x1))^2)
meanSqError = ssresidual/(n-p)
resStandError = sqrt(meanSqError)
resStandError
```

```
## [1] 2324.779
```

We compute the Residual standard error on line 20 to be 2324.779

Question c.

Intercept Standard Error Calculation (values from summary table)

```
stdError = sum$coefficients[[1]]/sum$coefficients[[5]]
stdError
```

```
## [1] 1197.351
```

Intercept Standard Error Calculation (data)

```
var = ssresidual/(n-2)
sxx = sum((x1 - mean(x1))^2)

stdErrorData = sqrt(var * (1/n + (mean(x1)^2)/sxx))
stdErrorData
```

```
## [1] 1197.351
```

Using both values from the summary table and the data directly, we compute the Standard Error on line 15 to be 1197.351

Question d.

R-Squared Calculation

```
ssr = sum((mean(y) - (intercept + slope*x1))^2)
sst = sum((y - mean(y))^2)

rsq = ssr/sst
rsq
```

```
## [1] 0.6967813
```

Our calculation finds an R^2 value of 0.6968, this matches the value in the summary table

Question e.

Theoretical Calculation

$$SS_{Reg} = \sum_{i=1}^{n} (\widehat{y}_i - \bar{y})^2$$

$$= \sum_{i=1}^{n} ((\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1}) - (\widehat{\beta}_0 + \widehat{\beta}_1 \bar{x}_1))^2$$

$$= \widehat{\beta}_1^2 \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2$$

$$
\begin{aligned}
&= \widehat{\beta}_1^2 S_{xx} \\
&= \widehat{\beta}_1 \frac{S_{xy}}{S_{xx}} S_{xx} \\
&= \widehat{\beta}_1 S_{xy}
\end{aligned}
$$

Therefore we have verified theoretically that this equation is correct

Numerical Calculation

SS_{Reg} calculated from a previous question

```
#from previous question
print(ssr, digits = 14)
```

```
## [1] 608555014.63283
```

Product of slope estimate and S_{xy}

```
#verification: product of slope estimate and s_xy
sxy = sum((y - mean(y))*(x1 - mean(x1)))
ssr_2 = sxy * slope
print(ssr_2, digits = 14)
```

```
## [1] 608555014.63283
```

Therefore we have verified numerically that this equation is correct

Question f.

F-statistic Calculation

```
F = (ssr/(p-1))/(ssresidual/(n-p))
F
```

```
## [1] 112.5995
```

We compute the F statistic on line 22 to be 112.5995

Question g.

Theoretical Calculations

$$
\begin{aligned}
trace(I_n - H_1) &= trace(I_n) - trace(H_1) \\
&= n - trace(1_n(1_n^T 1_n)^{-1} 1_n^T) \\
&= n - trace(1_n(n)^{-1} 1_n^T) \\
&= n - \frac{1}{n} trace(1_n 1_n^T) \\
&= n - \frac{1}{n} n \\
&= n - 1
\end{aligned}
$$

$$
trace(H - H_1) = trace(H) - trace(H_1)
$$

$$= trace(X(X^TX)^{-1}X^T) - 1$$
$$= trace(X^TX(X^TX)^{-1}) - 1$$
$$= trace(I_p) - 1$$
$$= p - 1$$

Therefore we have verified theoretically that these equations are correct

Numerical Calculation

```
one = matrix( rep(1, len=n), ncol = 1)
H1 = (1/n)*(one %*% t(one))

sum(diag(diag(n)-H1))
```

```
## [1] 50
```

Given that $n = 51$ we have $n - 1 = 50$.

```
colOne = matrix( rep(1, len=n), ncol = 1)
colTwo = data.matrix(x1, rownames.force = NA)
xMatrix = cbind(colOne, x1)

H = xMatrix %*% ginv(t(xMatrix) %*% xMatrix) %*% t(xMatrix)

sum(diag(H-H1))
```

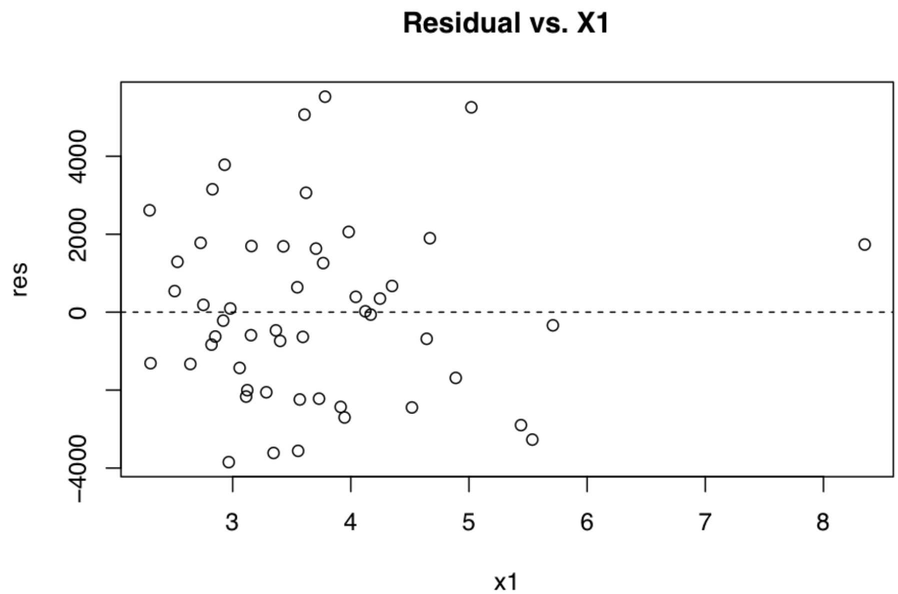
```
## [1] 1
```

Given that $p = 2$ we have $p - 1 = 1$.

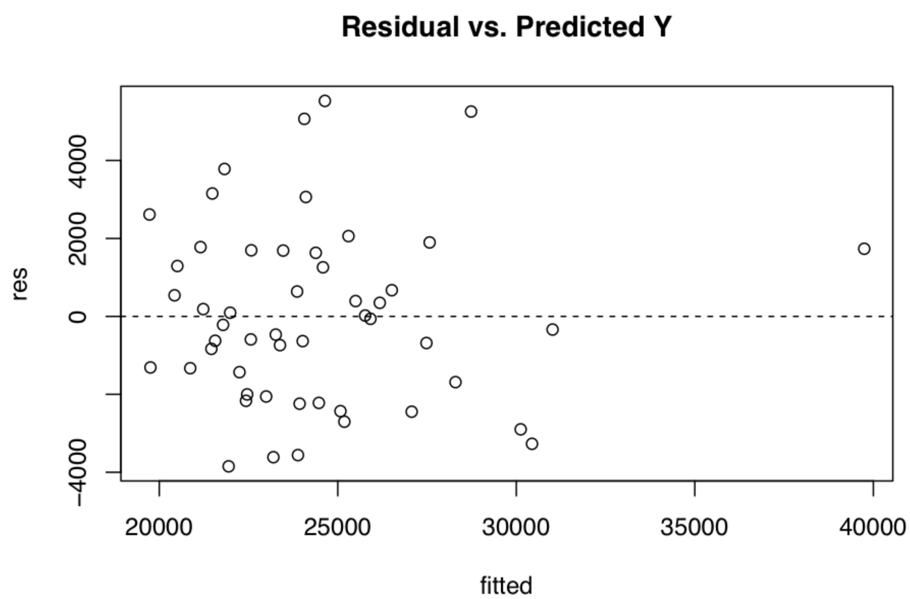
Therefore we have verified numerically that these equations are correct

Question h.

```
fitted = intercept + x1*slope
res = y - fitted
plot(x1,res,abline(h=0,lty=2))
title('Residual vs. X1')
```




```r
plot(fitted,res,abline(h=0,lty=2))
title('Residual vs. Predicted Y')
```



In both plots the residuals seem to have constant variance, mean of zero and do not appear to have any obvious pattern. Therefore the assumptions of the least squares analysis are valid.

Orthogonality

```r
#part 1
t(one)  %*%  res
```

```
##                 [,1]
## [1,] -8.149073e-10
```

```r
#part 2
t(xMatrix) %*% res
```

```
##             [,1]
##     -8.149073e-10
## x1 -2.789420e-09
```

```r
#part 3
t(fitted) %*% res
```

```
##                  [,1]
## [1,] -0.00001928955
```

All three outputs are virtually zero, it's just floating point inaccuracy from the calculations that prevent the computations from returning 0 as the answer. This confirms that the residual values are orthogonal to the 1 matrix, X and the fitted y values

Question i.

```r
yHatOne = intercept + (4800/1000)*slope
yHatOne
```

```
## [1] 28005.78
```

Therefore we would expect this teacher's salary to be 28006 dollars

Question j.

```r
x1New<-matrix(c(1,4.8),nrow=1)

stdErrorYNew = sqrt(var * x1New %*% (ginv(t(xMatrix) %*% xMatrix)) %*% t(x1New))
stdErrorYNew
```

```
##           [,1]
## [1,] 473.5628
```

We compute the estimated standard prediction error for y^{new} to be 473.5628