Homework1

Subhankar Ghosh

September 2, 2017

Question 1

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

for all $i = 1, 2, ..., n$ Least square estimator is found by minimizing

$$RSS(\beta_0, \beta_1) = \sum_{i=0}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

$$\frac{\partial RSS(\beta_0, \beta_1)}{\partial \beta_0} = 0$$

$$\sum_{i=0}^{n} 2[y_i - (\beta_0 + \beta_1 x_i)](-1) = 0$$

$$\beta_0 = \frac{\sum_{i=0}^{n} y_i + \beta_1 \sum_{i=0}^{n} x_i}{n}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$\frac{\partial RSS(\beta_0, \beta_1)}{\partial \beta_1} = 0$$

Substituting the value of β_0 in the equation of $RSS(\beta_0, \beta_1)$ we get:

$$\sum_{i=0}^{n} [y_i - \bar{y} - \beta_1(x_i - \bar{x})](x_i - \bar{x}) = 0$$

$$\beta_1 = \frac{\sum_{i=0}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=0}^{n} (x_i - \bar{x})^2}$$

$$\beta_1 = \frac{COV(x, y)}{Var(x)}$$

$$\beta_0 = \bar{y} - \frac{COV(x, y)}{Var(x)} \bar{x}$$

Question 2

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

for all $i = 1, 2, ..., n$ Assumption: $e_i follows N(0, \sigma^2)$

$$N(e_i; 0, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2})$$

$$L = \prod_{i=0}^{n} N(e_i; 0, \sigma^2) = (2\pi\sigma^2)^{n/2} exp(-\frac{1}{2\sigma^2} \sum_{i=0}^{n} (y_i - \beta_0 - \beta_1 x_i)^2)$$

Taking log of both sides

$$log(L) \;=\; log((2\pi\sigma^2)^{n/2}) - \frac{1}{2\sigma^2}\sum_{i=0}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

taking partial derivative of $log(L)$ with respect to $\beta_0 and \beta_1$ and equating to 0 we get:

$$\sum_{i=0}^{n} 2[y_i - (\beta_0 + \beta_1 x_i)](-1) \;=\; 0$$

$$\beta_0 \;=\; \frac{\sum_{i=0}^{n} y_i + \beta_1 \sum_{i=0}^{n} x_i}{n}$$

$$\beta_0 \;=\; \bar{y} - \beta_1 \bar{x}$$

$$\frac{\partial RSS(\beta_0, \beta_1)}{\partial \beta_1} \;=\; 0$$

Substituting the value of β_0 in the equation of $RSS(\beta_0, \beta_1)$ we get:

$$\sum_{i=0}^{n}[y_i - \bar{y} - \beta_1(x_i - \bar{x})](x_i - \bar{x}) \;=\; 0$$

$$\beta_1 \;=\; \frac{\sum_{i=0}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=0}^{n}(x_i - \bar{x})^2}$$

$$\beta_1 \;=\; \frac{COV(x,y)}{Var(x)}$$

$$\beta_0 \;=\; \bar{y} - \frac{COV(x,y)}{Var(x)}\bar{x}$$

Question 3

Given:

$$y_i \;=\; \beta x_i + e_i$$

we have to minimize $RSS(\beta_1)$

$$\frac{\partial RSS(\beta_1)}{\partial \beta_1} = \frac{\partial \sum_{i=0}^{n}[y_i - \beta_1 x_i]^2}{\partial \beta_1} \;=\; 0$$

$$\sum_{i=0}^{n} 2[y_i - \beta_1 x_i](-x_i) = 0$$

$$\beta_1 \;=\; \sum_{i=0}^{n} y_i x_i / \sum_{i=0}^{n} x_i^2$$

$$Var(\beta_1) \;=\; Var(\sum_{i=0}^{n} y_i x_i / \sum_{i=0}^{n} x_i^2)$$

We know that $Var(y_i) = var(e) = \sigma^2$

$$Var(\beta_1) \;=\; \frac{Var(\sum_{i=0}^{n} x_i e)}{[\sum_{i=0}^{n} x_i^2]^2}$$

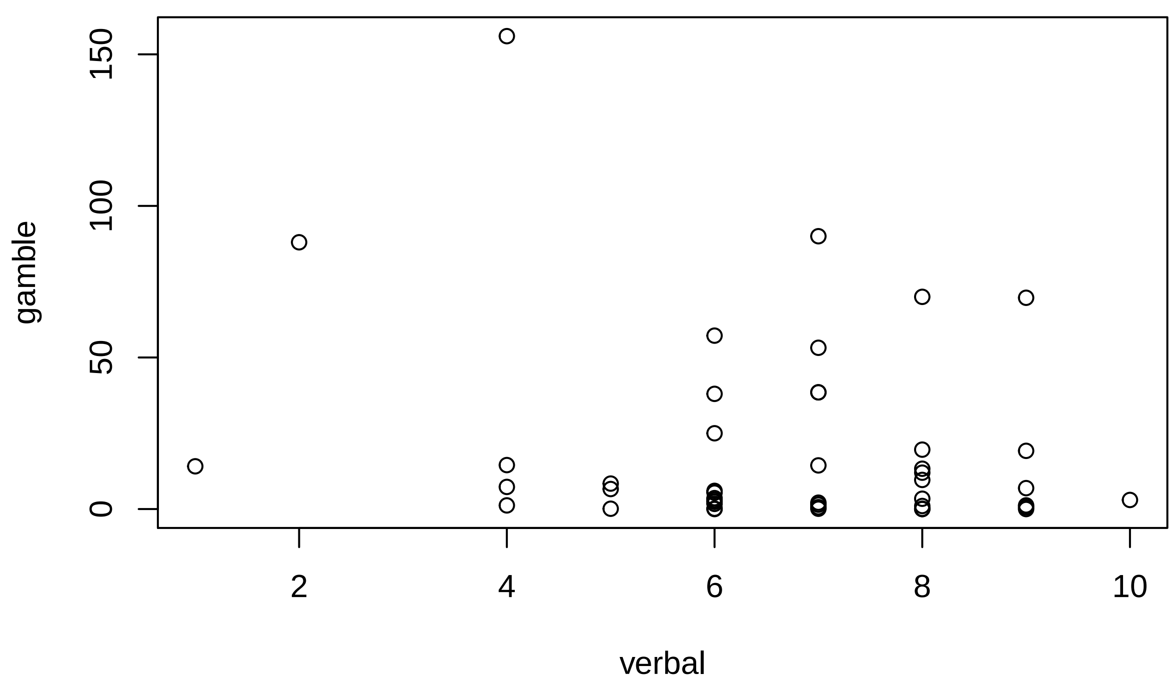
$$
\begin{aligned}
Var(\beta_1) &= \frac{\sum_{i=0}^{n} Var(x_i e)}{[\sum_{i=0}^{n} x_i^2]^2} \\
Var(\beta_1) &= \frac{\sum_{i=0}^{n} x_i^2 Var(e)}{[\sum_{i=0}^{n} x_i^2]^2} \\
Var(\beta_1) &= \frac{\sigma^2}{\sum_{i=0}^{n} x_i^2}
\end{aligned}
$$

Question 4

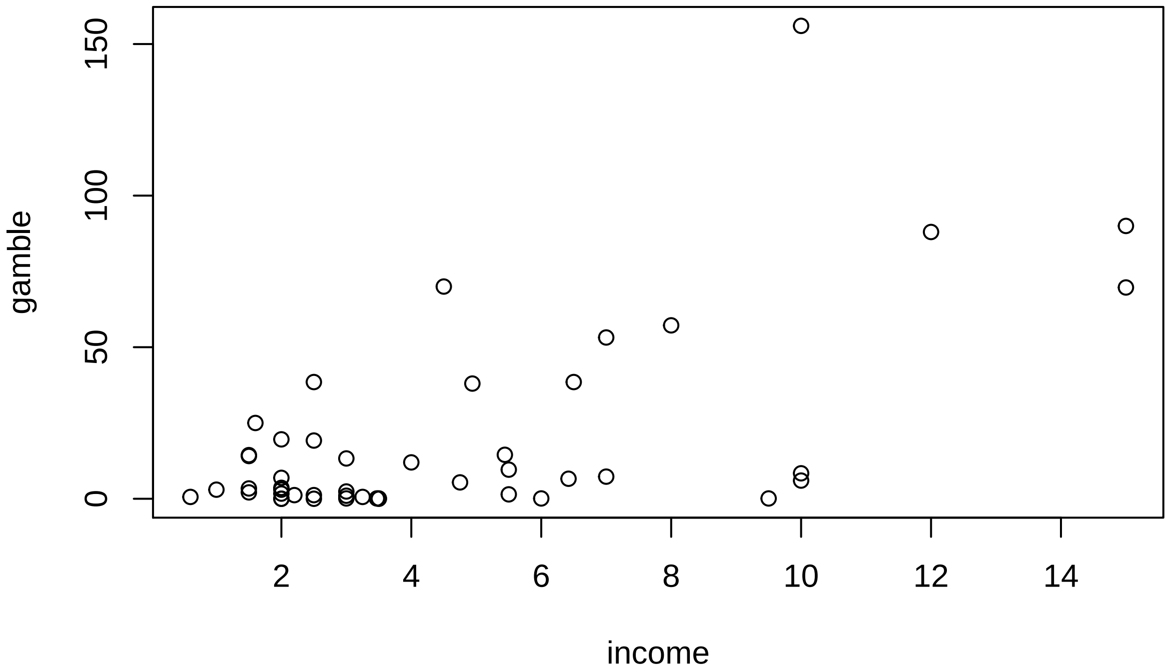
```r
library(faraway)
data(teengamb)
teen = teengamb
head(teen)
```

```
##   sex status income verbal gamble
## 1   1     51   2.00      8    0.0
## 2   1     28   2.50      8    0.0
## 3   1     37   2.00      6    0.0
## 4   1     28   7.00      4    7.3
## 5   1     65   2.00      8   19.6
## 6   1     61   3.47      6    0.1
```

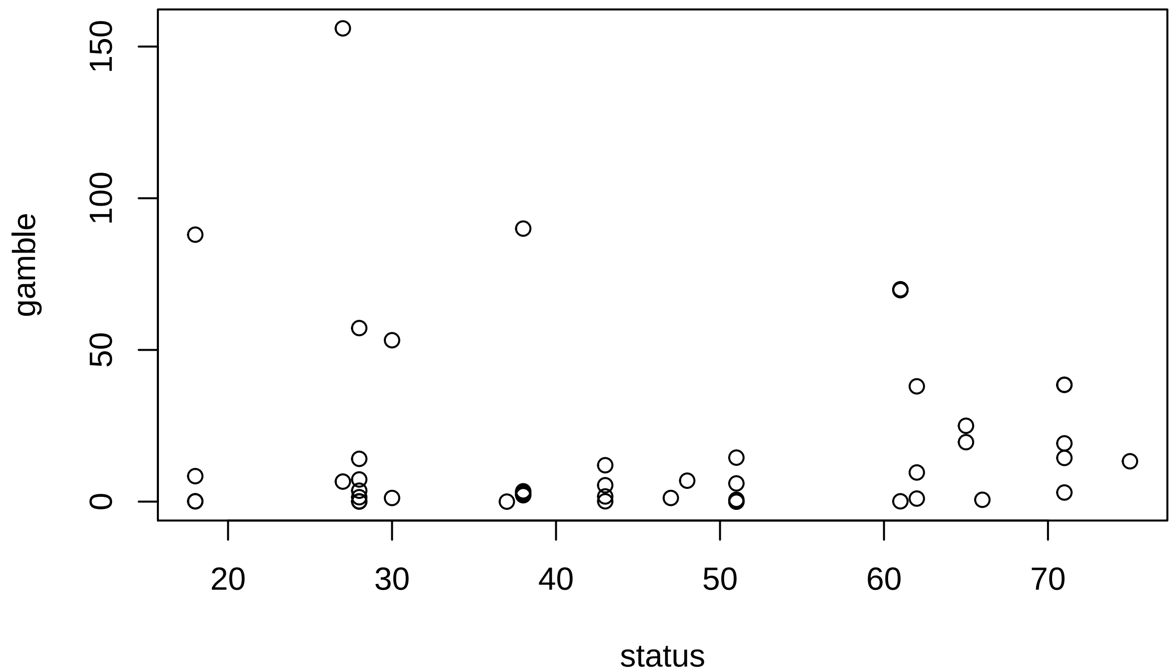
```r
plot(gamble ~ verbal, teen)
```



```r
plot(gamble ~ income, teen)
```




```r
plot(gamble ~ status, teen)
```




```r
model = lm(formula = gamble ~ ., data = teen)
summary(model)
```

```
##
## Call:
## lm(formula = gamble ~ ., data = teen)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.082 -11.320  -1.451   9.452  94.252
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 22.55565   17.19680   1.312   0.1968
## sex        -22.11833    8.21111  -2.694   0.0101 *
## status       0.05223    0.28111   0.186   0.8535
## income       4.96198    1.02539   4.839 1.79e-05 ***
## verbal      -2.95949    2.17215  -1.362   0.1803
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.69 on 42 degrees of freedom
## Multiple R-squared:  0.5267, Adjusted R-squared:  0.4816
## F-statistic: 11.69 on 4 and 42 DF,  p-value: 1.815e-06
```

Question a. What percentage of variation in the response is explained by these predictors?

***Solution** From the summary we can see the $R^2 = 0.5267$ therefore 52.67% variation in the response can be explained by these predictors (sex, status, income and verbal).

Question b. Give the case number that corresponds to the highest positive residual.

```
print(which(model$residuals == max(model$residuals)))
```

```
## 24
## 24
```

```
print(max(model$residuals))
```

```
## [1] 94.25222
```

The highest positive residual corresponds to case number 24. The highest positive residual is 94.25.

Question c. Compute the mean and median of the residuals.

```
mean(model$residuals)
```

```
## [1] -3.065293e-17
```

```
median(model$residuals)
```

```
## [1] -1.451392
```

Question d. Compute the correlation of the residuals with the fitted values.

```
cor(model$fitted.values, model$residuals)
```

```
## [1] -1.070659e-16
```

Question e. Compute the correlation of the residuals with income.

```
cor(model$residuals, teen$income)
```

```
## [1] -7.242382e-17
```

Question f. When all other predictors are held constant, what would be the difference in the predicted expenditure on gambling for a male compared to a female?

Approach one is to fit a model with only sex as the predictor and then find the mean of the fitted values of male and female and find the difference

```
#Finding all male cases
maleIndex = which(teen$sex == 1)

modelsex = lm(formula = gamble ~ sex, data = teen)
maleGamble = mean(modelsex$fitted.values[maleIndex])
femaleGamble = mean(modelsex$fitted.values[-maleIndex])
abs(maleGamble - femaleGamble)
```

```
## [1] 25.90921
```

Approach two is that we take the mean of the response values of male and female separately and find the difference. Since sex is a categorical variable the regression model will essentially fit by finding the mean of the y variable(gamble) at different values of sex.

```
abs(mean(teen[maleIndex, 'gamble']) - mean(teen[-maleIndex, 'gamble']))
```

```
## [1] 25.90921
```

As expected we find the same answer for both the approaches.

Question g. Which variables are statistically significant atthe 0.05 significance level?

From Pr(>|t|) column in the summary we can see that **sex** and **income** are statistically significant with a significance level of 0.05 since their P(>|t|) is less that 0.05

Question h. Predict the amount that a male with average status, income and verbal score would gamble along with a 95 percent prediction interval. Repeat the prediction for a male with maximal values of status, income and verbal score. Which prediction interval is wider and why is this result expected?

```
meaninput = data.frame(sex = 1, status = mean(teen$status), income = mean(teen$income), verbal = mean(te
maxinput = data.frame(sex = 1, status = max(teen$status), income = max(teen$income), verbal = max(teen$v
print("Prediction for  a male with average status, income and verbal score")
```

```
## [1] "Prediction for  a male with average status, income and verbal score"
```

```
predict(model, meaninput, interval = "predict")
```

```
##        fit       lwr      upr
## 1 6.124186 -41.19262 53.44099
```

```
print("Prediction for  a male with maximum status, income and verbal score")
```

```
## [1] "Prediction for  a male with maximum status, income and verbal score"
```

```
predict(model, maxinput, interval = "predict")
```

```
##        fit       lwr      upr
## 1 49.18961 -9.250356 107.6296
```

Prediction interval for male with maximal values of status, income and verbal score has the wider prediction interval.

Question i. Fit a model with just income as a predictor and use an F-test to compare it to the full model.

```
modelIncome = lm(formula = gamble ~ income, data = teen)
anova(modelIncome, model)
```

```
## Analysis of Variance Table
##
## Model 1: gamble ~ income
## Model 2: gamble ~ sex + status + income + verbal
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1     45 28009
## 2     42 21624  3    6384.8 4.1338 0.01177 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the Pr(>F) value is smaller that $\alpha(0.05)$ we can reject the null hypothesis $H_0 : \beta_1 = \beta_2 = \beta_4 = 0$.

Question 5

```
data(sat)
sat = sat
head(sat)
```

```
##            expend ratio salary takers verbal math total
## Alabama     4.405  17.2 31.144      8    491  538  1029
## Alaska      8.963  17.6 47.951     47    445  489   934
## Arizona     4.778  19.3 32.175     27    448  496   944
## Arkansas    4.459  17.1 28.934      6    482  523  1005
## California  4.992  24.0 41.078     45    417  485   902
```

```
## Colorado    5.443  18.4 34.571     29    462  518   980
```

Question a

```r
sat_model = lm(total ~ expend + ratio + salary, data = sat)
summary(sat_model)
```

```
##
## Call:
## lm(formula = total ~ expend + ratio + salary, data = sat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -140.911  -46.740   -7.535   47.966  123.329
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1069.234    110.925   9.639 1.29e-12 ***
## expend        16.469     22.050   0.747   0.4589
## ratio          6.330      6.542   0.968   0.3383
## salary        -8.823      4.697  -1.878   0.0667 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68.65 on 46 degrees of freedom
## Multiple R-squared:  0.2096, Adjusted R-squared:  0.1581
## F-statistic: 4.066 on 3 and 46 DF,  p-value: 0.01209
```

```r
model_without_salary = lm(total ~ expend + ratio, data = sat)
anova(model_without_salary, sat_model)
```

```
## Analysis of Variance Table
##
## Model 1: total ~ expend + ratio
## Model 2: total ~ expend + ratio + salary
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     47 233443
## 2     46 216812  1     16631 3.5285 0.06667 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null Hypothesis $H_0 : \beta_salary = 0$. But the P value of the F test is greater than 0.05 thus we can reject the null hypothesis and conclude $\beta_salary \neq 0$.

```r
model_without_predictor = lm(total ~ 1, data = sat)
anova(model_without_predictor, sat_model)
```

```
## Analysis of Variance Table
##
## Model 1: total ~ 1
## Model 2: total ~ expend + ratio + salary
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     49 274308
## 2     46 216812  3     57496 4.0662 0.01209 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null Hypothesis $H_0 : \beta_s alary = \beta_e xpend = \beta_r atio = 0$. The P value of the F test is less than 0.05 thus we cannot reject the null hypothesis. We agree to the null hypothesis that $\beta_s alary = \beta_e xpend = \beta_r atio = 0$.

Question b

```r
full_model = lm(total ~ expend + ratio + salary + takers, data = sat)
```

```r
summary(full_model)
```

```
##
## Call:
## lm(formula = total ~ expend + ratio + salary + takers, data = sat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -90.531 -20.855  -1.746  15.979  66.571
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1045.9715    52.8698  19.784  < 2e-16 ***
## expend         4.4626    10.5465   0.423    0.674
## ratio         -3.6242     3.2154  -1.127    0.266
## salary         1.6379     2.3872   0.686    0.496
## takers        -2.9045     0.2313 -12.559 2.61e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.7 on 45 degrees of freedom
## Multiple R-squared:  0.8246, Adjusted R-squared:  0.809
## F-statistic: 52.88 on 4 and 45 DF,  p-value: < 2.2e-16
```

t -test for takers : P value obtained after performing t test on takers shows it is very significant with a P value of 2.61e-16

```r
anova(sat_model, full_model)
```

```
## Analysis of Variance Table
##
## Model 1: total ~ expend + ratio + salary
## Model 2: total ~ expend + ratio + salary + takers
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     46 216812
## 2     45  48124  1    168688 157.74 2.607e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After performing F test for takers $H_0 : \beta_t akers = 0$ We get a P value of 2.6e-16 (<0.05) thus we reject the null hypothesis. We observe that P value obtained from t-test and F-test for a single variable are equivalent.