STAT 425 - Homework 2

Shashi Roshan (sroshan2@illinois.edu)

05 May, 2020

Answer 1.a

```
library(faraway)
```

```
##
## Attaching package: 'faraway'

## The following object is masked from 'package:rpart':
##
##     solder
```

```
data(teengamb)

lm_model = lm(gamble ~ ., data = teengamb)

# Percentage of variation explained
summary(lm_model)$r.squared * 100
```

```
## [1] 52.67234
```

Percentage of variation explained 52.6723413

Answer 1.b

```
residuals = residuals(lm_model)
# Observation which has highest positive residual
max_residual_position = residuals[which.max(residuals)]
max_residual_position
```

```
##       24
## 94.25222
```

```
# Observation which has lowest negative residual
min_residual_position = residuals[which.min(residuals)]
min_residual_position
```

```
##        39
## -51.08241
```

```r
# Mean
mean_of_residuals = mean(residuals)
mean_of_residuals
```

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

```r
# Median
median_of_residuals = median(residuals)
median_of_residuals
```

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

Answer 1.c

```r
summary(lm_model)$coefficients[2,1]
```

```
## [1] -22.11833
```

When all other predictors are held constant, the difference in predicted expenditure between male and females will be :

(β_{sex} * (0)) - (β_{sex} * (1)) = 0 - (β_{sex}) = 0-(-22.11) = 22.11. This means that males will have 22.12 currency units expenditure more than females.

Answer 1.d

```r
avg_status = mean(teengamb$status)
avg_income = mean(teengamb$income)
avg_verbal = mean(teengamb$verbal)

predict(lm_model, data.frame( status = avg_status, income = avg_income, verbal = avg_verbal, sex = 0),
```

```
##        fit       lwr      upr
## 1 28.24252 -18.51536 75.00039
```

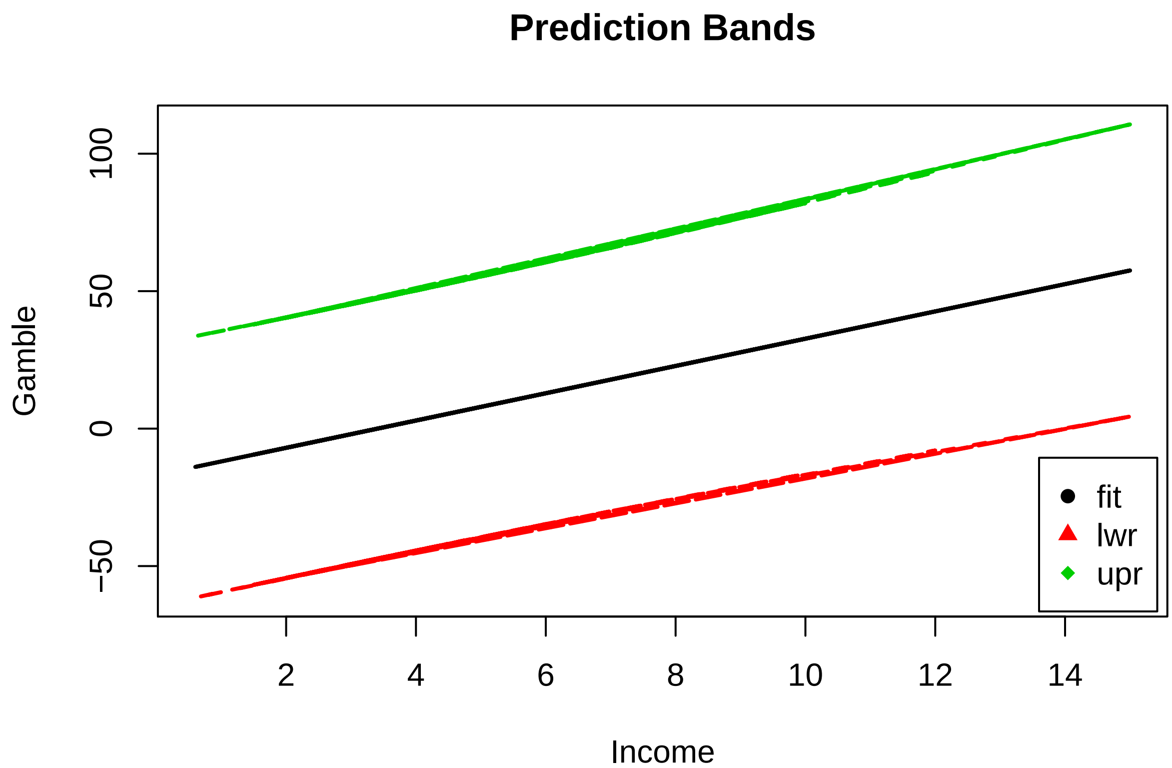
Answer 1.e

```r
pred_intervals = predict(lm_model, data.frame( status = avg_status, income = teengamb$income, verbal =
pred_intervals
```

```
##             fit        lwr       upr
## 1   -6.9849405 -54.283616  40.31373
## 2   -4.5039509 -51.757558  42.74966
## 3   -6.9849405 -54.283616  40.31373
## 4   17.8249556 -30.038377  65.68829
## 5   -6.9849405 -54.283616  40.31373
## 6    0.3091690 -46.921564  47.53990
## 7   10.3819868 -37.076388  57.84036
## 8   14.9470077 -32.736196  62.63021
## 9   -6.9849405 -54.283616  40.31373
## 10  12.8629764 -34.708269  60.43422
## 11  -2.0229612 -49.254127  45.20820
## 12   6.6605024 -40.670490  53.99149
## 13  -5.9925446 -53.270481  41.28539
## 14  -6.9849405 -54.283616  40.31373
## 15  -2.0229612 -49.254127  45.20820
## 16  -9.4659301 -56.832238  37.90038
## 17  30.2299037 -18.740037  79.19984
## 18  32.7108933 -16.542630  81.96442
## 19   2.9390180 -44.315248  50.19328
## 20   0.4580284 -46.773357  47.68941
## 21  -2.0229612 -49.254127  45.20820
## 22  -4.5039509 -51.757558  42.74966
## 23   0.4580284 -46.773357  47.68941
## 24  32.7108933 -16.542630  81.96442
## 25  15.3439660 -32.362328  63.05026
## 26  -9.4659301 -56.832238  37.90038
## 27  10.0842680 -37.362061  57.53060
## 28 -11.9469197 -59.403327  35.50949
## 29 -13.9317114 -61.476289  33.61287
## 30  10.3819868 -37.076388  57.84036
## 31  42.6348517  -7.949118  93.21882
## 32  17.8249556 -30.038377  65.68829
## 33  57.5207893   4.396293 110.64529
## 34  -6.9849405 -54.283616  40.31373
## 35  -9.4659301 -56.832238  37.90038
## 36   5.4200076 -41.879765  52.71978
## 37  -4.5039509 -51.757558  42.74966
## 38  22.7869348 -25.455557  71.02943
## 39  32.7108933 -16.542630  81.96442
## 40  -8.9697322 -56.320712  38.38125
## 41  -6.9849405 -54.283616  40.31373
## 42  57.5207893   4.396293 110.64529
## 43  -2.0229612 -49.254127  45.20820
## 44  -0.7824664 -48.010909  46.44598
## 45   7.6032784 -39.755206  54.96176
## 46  -9.4659301 -56.832238  37.90038
## 47  -4.5039509 -51.757558  42.74966
```

```r
matplot(x = teengamb$income, y = pred_intervals, type = "l", lty = c(1,2,2), lwd = c(2,2,2), main = "Pr

legend("bottomright", inset=0.01, legend=colnames(pred_intervals), col=c(1:3), pch=16:20,bg= ("white"),
```



Answer 1.f

```
lm_model_2 = lm(gamble ~ sex + income, data = teengamb)

# percentage of variation in the response that is explained
summary(lm_model_2)$r.squared * 100
```

```
## [1] 50.13882
```

```
# F-test to formally compare it to the full model
anova(lm_model, lm_model_2)
```

```
## Analysis of Variance Table
##
## Model 1: gamble ~ sex + status + income + verbal
## Model 2: gamble ~ sex + income
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     42 21624
## 2     44 22781 -2   -1157.5 1.1242 0.3345
```

From the F-test, we can see that the F-test value is small Hence we accept the null hypothesis, and conclude that the smaller model (predictor sex and income) is as good as the full model.

Answer 1.g

```
lm_model_sex = lm(gamble ~ sex, data = teengamb)
summary(lm_model_sex)$r.squared
```

```
## [1] 0.1663052
```

```
lm_model_income = lm(gamble ~ income, data = teengamb)
summary(lm_model_income)$r.squared
```

```
## [1] 0.3869797
```

```
lm_model_status = lm(gamble ~ status, data = teengamb)
summary(lm_model_status)$r.squared
```

```
## [1] 0.002542258
```

```
lm_model_verbal = lm(gamble ~ verbal, data = teengamb)
summary(lm_model_verbal)$r.squared
```

```
## [1] 0.04842473
```

```
# Best model uses predictor income
anova(lm_model, lm_model_income)
```

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

Predictor “income” gives the highest R2 score. By comparing the full model with the smaller model using anova, we can see that the p-value of F test is less than the threshold, hence we reject the null hypothesis and conclude that the Full model is better than the smaller model.

Answer 1.h

```
lm_model_income = lm(gamble ~ income, data = teengamb)
summary(lm_model_income)$r.squared
```

```
## [1] 0.3869797
```

```r
lm_model_income_sex = lm(gamble ~ income + sex, data = teengamb)
summary(lm_model_income_sex)$r.squared
```

```
## [1] 0.5013882
```

```r
lm_model_income_sex_verbal = lm(gamble ~ income + sex + verbal, data = teengamb)
summary(lm_model_income_sex_verbal)$r.squared
```

```
## [1] 0.5263344
```

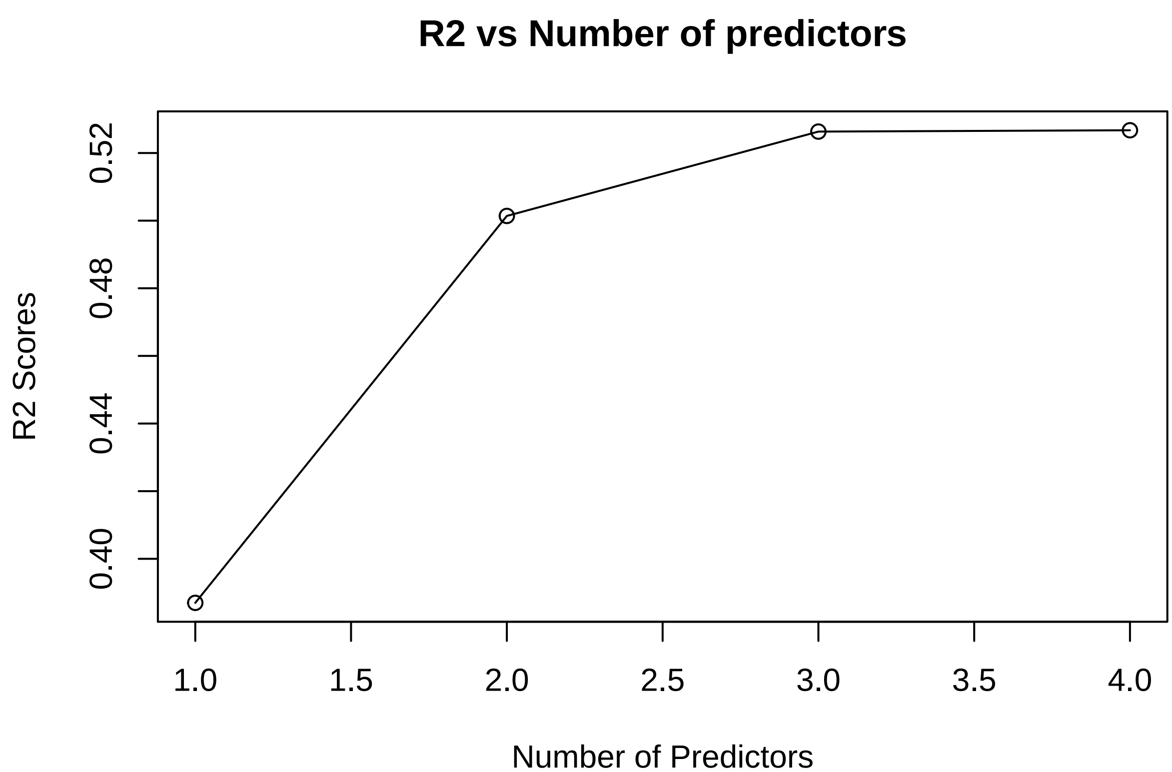
```r
lm_model_income_sex_verbal_status = lm(gamble ~ income + sex + verbal + status, data = teengamb)
summary(lm_model_income_sex_verbal_status)$r.squared
```

```
## [1] 0.5267234
```

```r
r2_scores = c(summary(lm_model_income)$r.squared,
              summary(lm_model_income_sex)$r.squared,
              summary(lm_model_income_sex_verbal)$r.squared,
              summary(lm_model_income_sex_verbal_status)$r.squared)

number_of_predictors = c(1, 2, 3, 4)

plot(number_of_predictors, r2_scores, main = "R2 vs Number of predictors", xlab = "Number of Predictors
lines(r2_scores)
```



The trend line shows that the R2 score increases as the number of predictors increase. The increase is fast initially, and then it slows down.