STAT 500: HW8

Jasmine Mou

11/21/2017

Q: Using the `teengamb` dataset with `gamble` as the response and the other variables as predictors. Find your optimal model.

```
data(teengamb, package="faraway")
n <- dim(teengamb)[1]
p <- dim(teengamb)[2] - 1
```

I. Model selection

1. Full regression model without and with transformation.

At the first attempt, fit a simple full regression model `lm_full` without transformation with `gamble` as the response and the other variables as predictors. At the second attempt, create full models with the same predictors and square root transformation `lm_full_sqrt` and log-transformation `lm_full_log`.

```
lm_full <- lm(gamble~., data=teengamb)
lm_full_sqrt <- lm(sqrt(gamble) ~ ., data = teengamb)
lm_full_log <- lm(log(1+gamble) ~ ., data = teengamb)
```

1) Check goodness of fit. *From the `summary` result, we can see with model `lm_full_sqrt` the highest percentage of variance in the response explainable by predictors is achieved, which is about 56.46%; `lm_full_log` has the lowest figure of about 52.06%.*

```
check_r2 <- function(lm){
  r2 <- summary(lm)$r.squared
  return(r2)
}
c(check_r2(lm_full), check_r2(lm_full_sqrt), check_r2(lm_full_log))
```

```
## [1] 0.5267234 0.5645605 0.5206486
```

2) Check significant predictors. *At the 5% level, statistically significant variables for `lm_full` are `sex` and `income`; for `lm_full_sqrt` are `sex` and `income` and `verbal`; and for `lm_full_log` are all predictors.*

```
check_sig <- function(lm){
  coef = summary(lm)$coefficients[,4]
  return(coef[coef<0.05])
}
check_sig(lm_full)
```

```
##          sex       income
## 1.011184e-02 1.791882e-05
```

```
check_sig(lm_full_sqrt)
```

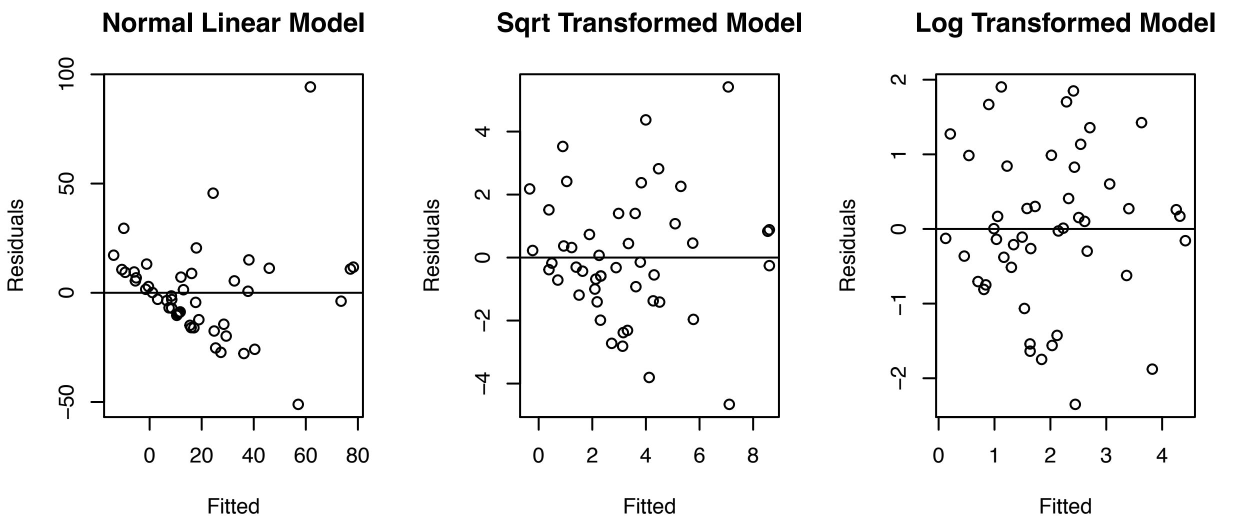
```
##          sex       income       verbal
## 9.676112e-03 7.942336e-06 3.966628e-02
```

```
check_sig(lm_full_log)
```

```
##  (Intercept)          sex       status       income       verbal
## 4.301374e-02 3.197461e-02 3.195076e-02 7.325311e-05 1.567253e-02
```

3) Check the constant variance assumption for the errors. *For `lm_full` and `lm_full_sqrt`, the plot suggests an increase in variance along the fitted values. For `lm_full_log` the variance looks constant along the fitted values.*

```
check_cva <- function(lm, name){
  plot(fitted(lm), residuals(lm), xlab="Fitted", ylab="Residuals", main=name)
  abline(h=0)
}
par(mfrow=c(1,3))
check_cva(lm_full, "Normal Linear Model")
check_cva(lm_full_sqrt, "Sqrt Transformed Model")
check_cva(lm_full_log, "Log Transformed Model")
```



4) Check the normality assumption. *Under the model `lm_full`, the residuals have a long tail possibly due to outliers, and look slightly right-skewed. From the Shapiro-Wilk normality test results at $\alpha = 0.05$, we are able to reject the normality of `lm_full`, but fail to reject the normality of `lm_full_sqrt` and `lm_full_log`.*

```
check_normality <- function(lm, name, plot=TRUE){
  res = residuals(lm)
  if(plot){
    qqnorm(res, ylab="Residuals", main=name)
    qqline(res)
  }
  return(shapiro.test(res))
}
par(mfrow=c(1,3))
check_normality(lm_full, "Normal Q-Q Plot")
```

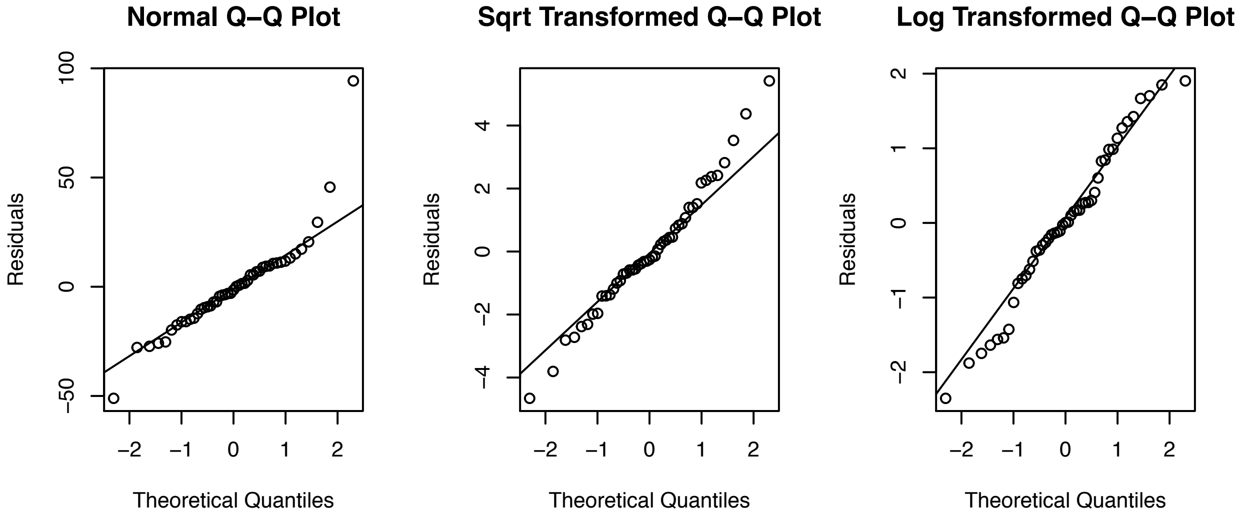
```
##
##  Shapiro-Wilk normality test
##
## data:  res
## W = 0.86839, p-value = 8.16e-05
```

```
check_normality(lm_full_sqrt, "Sqrt Transformed Q-Q Plot")
```

```
##
##  Shapiro-Wilk normality test
##
```

```
## data:  res
## W = 0.98321, p-value = 0.7272
```

```r
check_normality(lm_full_log, "Log Transformed Q-Q Plot")
```




```
##
##  Shapiro-Wilk normality test
##
## data:  res
## W = 0.97609, p-value = 0.4418
```

5) Check for large leverage points. *Observations #31, #33, #35, #42 are large leverage points for all models so far.*

```r
check_leverage <- function(lm){
  hatv <- hatvalues(lm)
  hatv[which(hatv>2*p/n)]
}
check_leverage(lm_full)
```

```
##        31        33        35        42
## 0.2395031 0.2213439 0.3118029 0.3016088
```

```r
check_leverage(lm_full_sqrt)
```

```
##        31        33        35        42
## 0.2395031 0.2213439 0.3118029 0.3016088
```

```r
check_leverage(lm_full_log)
```

```
##        31        33        35        42
## 0.2395031 0.2213439 0.3118029 0.3016088
```

6) Check for outliers. *Under the model `lm_full`, observation #24 is the outlier. There are no outliers for `lm_full_sqrt` and `lm_full_log`.*

```r
check_outlier <- function(lm){
  stud <- rstudent(lm)
  stud[which(abs(stud) > abs(qt(0.05/(n*2), n-1-p-1)))]
}
check_outlier(lm_full)
```

```
##       24
## 6.016116
```

```r
check_outlier(lm_full_sqrt)
```

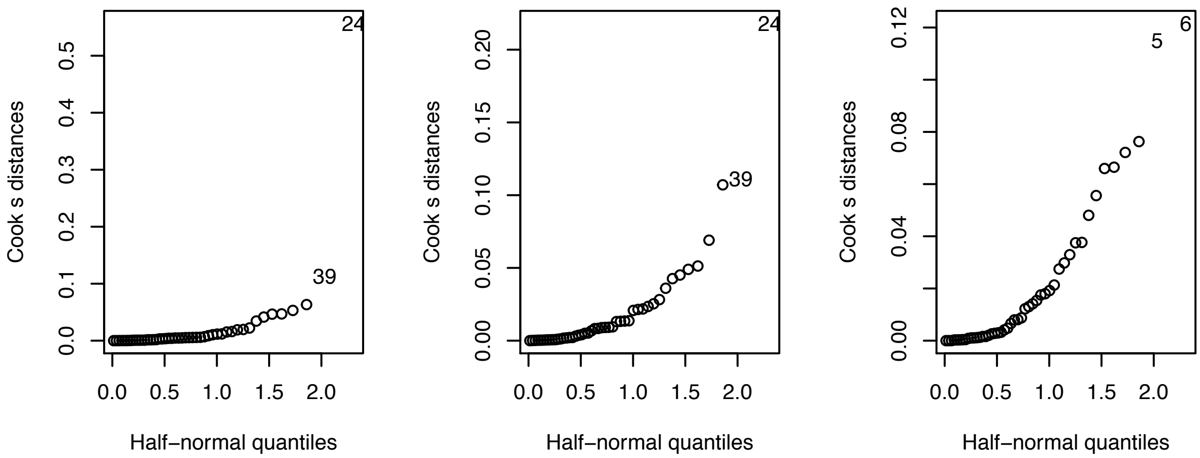
```
## named numeric(0)
```

```r
check_outlier(lm_full_log)
```

```
## named numeric(0)
```

7) Check for influential points with Cook's distance. *Observations #24 and #39 are influential points for `lm_full` and `lm_full_sqrt`. Observations #5 and #6 are influential points for `lm_full_log`.*

```r
check_influential <- function(lm){
  cook <- cooks.distance(lm)
  cook[which(cook>4/(n-p-1))]
  faraway::halfnorm(cook,2, ylab="Cook s distances")
}
par(mfrow=c(1,3))
check_influential(lm_full)
check_influential(lm_full_sqrt)
check_influential(lm_full_log)
```



Step 1 Summary: *`lm_full_log` outperforms `lm_full` and `lm_full_sqrt` after considering all these test above.*

2. Reduced regression model without and with transformation. 8) Check AIC. *For the model without transformation, AIC is minimized by choosing 3 predictors, which are **`income`**, **`sex`**, and **`verbal`**. For the model with square root and log transformation, AIC is minimized by keeping all 4 predictors.*

```r
require(leaps)
sub_lm <- regsubsets(gamble~., data=teengamb)
sub_sqrt <- regsubsets(sqrt(gamble) ~ ., data = teengamb)
sub_log <- regsubsets(log(1 +gamble) ~ ., data = teengamb)
check_AIC <- function(sub){
  rs <- summary(sub)
  AIC <- n * log(rs$rss/n) + (2:(p+1))*2
  np <- which.min(AIC)
  row <- rs$which[np, ]
  row[row==TRUE]
  # plot(AIC ~ I(1:p), ylab="AIC", xlab="Number of Predictors")
}
```

```r
check_AIC(sub_lm)
```

```
## (Intercept)         sex      income      verbal 
##        TRUE        TRUE        TRUE        TRUE 
```

```r
check_AIC(sub_sqrt)
```

```
## (Intercept)         sex      status      income      verbal 
##        TRUE        TRUE        TRUE        TRUE        TRUE 
```

```r
check_AIC(sub_log)
```

```
## (Intercept)         sex      status      income      verbal 
##        TRUE        TRUE        TRUE        TRUE        TRUE 
```

9) Check Adjusted R^2. *The choice of predictors to maximize R^2 is the same as that to minimize AIC.*

```r
check_adj_r2 <- function(sub){
  rs <- summary(sub)
  adj_r2 <- rs$adjr2
  np <- which.max(adj_r2)
  row <- rs$which[np, ]
  row[row==TRUE]
  # plot(rs$adjr2 ~ I(1:p), xlab="Number of Predictors", ylab=expression(paste("Adjusted ", R^2)))
}
check_adj_r2(sub_lm)
```

```
## (Intercept)         sex      income      verbal 
##        TRUE        TRUE        TRUE        TRUE 
```

```r
check_adj_r2(sub_sqrt)
```

```
## (Intercept)         sex      status      income      verbal 
##        TRUE        TRUE        TRUE        TRUE        TRUE 
```

```r
check_adj_r2(sub_log)
```

```
## (Intercept)         sex      status      income      verbal 
##        TRUE        TRUE        TRUE        TRUE        TRUE 
```

10) Check Mallows C_p. *The choices of predictors to minimize Mallows C_p are the same for all models:* ***income****,* ***sex****, and* ***verbal****.*

```r
check_cp <- function(sub){
  rs <- summary(sub)
  cp <- rs$cp
  np <- which.min(cp)
  row <- rs$which[np,]
  row[row==TRUE]
  # plot(rs$cp ~ I(1:p), xlab="Number of Predictors", ylab=expression(paste(C[p], " Statistic")))
  # abline(1,1)
}
check_cp(sub_lm)
```

```
## (Intercept)         sex      income      verbal 
##        TRUE        TRUE        TRUE        TRUE 
```

```r
check_cp(sub_sqrt)
```

```
## (Intercept)         sex      status      income      verbal 
```

```
##        TRUE        TRUE        TRUE        TRUE        TRUE
```

```r
check_cp(sub_log)
```

```
## (Intercept)         sex      status      income      verbal
##        TRUE        TRUE        TRUE        TRUE        TRUE
```

Step 2 Summary: *we can see variables* `sex`, `income` *and* `verbal` *are useful predictors to all 3 models. Thus combined with* Step 1 Summary, *candidate models are now A) log transformed model with full predictors, B) log transformed model with predictors* `sex`, `income` *and* `verbal`, *and C) the model without transformation and with predictors* `sex`, `income` *and* `verbal` *and with the outlier #24 removed. Model C fails the normality check. With adjusted R^2 as the principle for RSS, then model A is the optimal model.*

```r
lm_C <- lm(gamble~sex + income + verbal, data=teengamb[-c(24:24),])
check_normality(lm_C, "lm_C normality", FALSE)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res
## W = 0.98274, p-value = 0.7193
```

```r
lm_A <- lm_full_log
summary(lm_A)
```

```
##
## Call:
## lm(formula = log(1 + gamble) ~ ., data = teengamb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.35012 -0.56865  0.00413  0.71512  1.90319
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.71620    0.82240   2.087   0.0430 *
## sex         -0.87120    0.39268  -2.219   0.0320 *
## status       0.02983    0.01344   2.219   0.0320 *
## income       0.21565    0.04904   4.398 7.33e-05 ***
## verbal      -0.26165    0.10388  -2.519   0.0157 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.085 on 42 degrees of freedom
## Multiple R-squared:  0.5206, Adjusted R-squared:  0.475
## F-statistic:  11.4 on 4 and 42 DF,  p-value: 2.347e-06
```

II. Model Inference

The optimal model is

$$gamble = e^{1.72+-0.87*sex+0.03*status+0.22*income+-0.26*verbal} - 1$$

Thus assuming all other variables are held constant, a female is expected to spend 0.42 times on (gambling+1) compared to a male in pounds/year.

````
---
title: 'STAT 500: HW8'
author: "Jasmine Mou"
date: "11/21/2017"
output: pdf_document
---
Q: Using the `teengamb` dataset with `gamble` as the response and the
other variables as predictors. Find your optimal model.
```{r}
data(teengamb, package="faraway")
n <- dim(teengamb)[1]
p <- dim(teengamb)[2] - 1
```

I. Model selection

**1. Full regression model without and with transformation.**
*At the first attempt, fit a simple full regression model `lm_full`
without transformation with `gamble` as the response and the other
variables as predictors. At the second attempt, create full models with
the same predictors and square root transformation `lm_full_sqrt` and
log-transformation `lm_full_log`.*
```{r}
lm_full <- lm(gamble~., data=teengamb)
lm_full_sqrt <- lm(sqrt(gamble) ~ ., data = teengamb)
lm_full_log <- lm(log(1+gamble) ~ ., data = teengamb)
```
1) Check goodness of fit.
*From the `summary` result, we can see with model `lm_full_sqrt` the
highest percentage of variance in the response explainable by predictors
is achieved, which is about 56.46%; `lm_full_log` has the lowest figure
of about 52.06%.*
```{r}
check_r2 <- function(lm){
  r2 <- summary(lm)$r.squared
  return(r2)
}
c(check_r2(lm_full), check_r2(lm_full_sqrt), check_r2(lm_full_log))
```

2) Check significant predictors.
*At the 5% level, statistically significant variables for `lm_full` are
`sex` and `income`; for `lm_full_sqrt` are `sex` and `income` and
`verbal`; and for `lm_full_log` are all predictors. *
```{r}
check_sig <- function(lm){
  coef = summary(lm)$coefficients[,4]
  return(coef[coef<0.05])
}
check_sig(lm_full)
````
````

check_sig(lm_full_sqrt)
check_sig(lm_full_log)
```

3) Check the constant variance assumption for the errors.
*For `lm_full` and `lm_full_sqrt`, the plot suggests an increase in variance along the fitted values. For `lm_full_log` the variance looks constant along the fitted values. *
```{r, fig.height=3}
check_cva <- function(lm, name){
  plot(fitted(lm), residuals(lm), xlab="Fitted", ylab="Residuals",
main=name)
  abline(h=0)
}
par(mfrow=c(1,3))
check_cva(lm_full, "Normal Linear Model")
check_cva(lm_full_sqrt, "Sqrt Transformed Model")
check_cva(lm_full_log, "Log Transformed Model")
```

4) Check the normality assumption.
Under the model `lm_full`, the residuals have a long tail possibly due to outliers, and look slightly right-skewed. From the Shapiro-Wilk normality test results at α = 0.05, we are able to reject the normality of `lm_full`, but fail to reject the normality of `lm_full_sqrt` and `lm_full_log`.
```{r, fig.height=3}
check_normality <- function(lm, name, plot=TRUE){
  res = residuals(lm)
  if(plot){
    qqnorm(res, ylab="Residuals", main=name)
    qqline(res)
  }
  return(shapiro.test(res))
}
par(mfrow=c(1,3))
check_normality(lm_full, "Normal Q-Q Plot")
check_normality(lm_full_sqrt, "Sqrt Transformed Q-Q Plot")
check_normality(lm_full_log, "Log Transformed Q-Q Plot")
```
5) Check for large leverage points.
Observations #31, #33, #35, #42 are large leverage points for all models so far.
```{r}
check_leverage <- function(lm){
  hatv <- hatvalues(lm)
  hatv[which(hatv>2*p/n)]
}
check_leverage(lm_full)
check_leverage(lm_full_sqrt)
check_leverage(lm_full_log)

````
```
6) Check for outliers.
*Under the model `lm_full`, observation #24 is the outlier. There are no
outliers for `lm_full_sqrt` and `lm_full_log`.*
```{r}
check_outlier <- function(lm){
  stud <- rstudent(lm)
  stud[which(abs(stud) > abs(qt(0.05/(n*2), n-1-p-1)))]
}
check_outlier(lm_full)
check_outlier(lm_full_sqrt)
check_outlier(lm_full_log)
```
7) Check for influential points with Cook's distance.
*Observations #24 and #39 are influential points for `lm_full` and
`lm_full_sqrt`. Observations #5 and #6 are influential points for
`lm_full_log`.*
```{r, fig.height=3}
check_influential <- function(lm){
  cook <- cooks.distance(lm)
  cook[which(cook>4/(n-p-1))]
  faraway::halfnorm(cook,2, ylab="Cook s distances")
}
par(mfrow=c(1,3))
check_influential(lm_full)
check_influential(lm_full_sqrt)
check_influential(lm_full_log)
```

Step 1 Summary: *`lm_full_log` outperforms `lm_full` and
`lm_full_sqrt` after considering all these test above.*

2. Reduced regression model without and with transformation.
8) Check AIC.
*For the model without transformation, AIC is minimized by choosing 3
predictors, which are `income`, `sex`, and `verbal`. For the model with
square root and log transformation, AIC is minimized by keeping all 4
predictors.*
```{r, message=FALSE, warning=FALSE}
require(leaps)
sub_lm <- regsubsets(gamble~., data=teengamb)
sub_sqrt <- regsubsets(sqrt(gamble) ~ ., data = teengamb)
sub_log <- regsubsets(log(1 +gamble) ~ ., data = teengamb)
check_AIC <- function(sub){
  rs <- summary(sub)
  AIC <- n * log(rs$rss/n) + (2:(p+1))*2
  np <- which.min(AIC)
  row <- rs$which[np, ]
  row[row==TRUE]
  # plot(AIC ~ I(1:p), ylab="AIC", xlab="Number of Predictors")
}
````

check_AIC(sub_lm)
check_AIC(sub_sqrt)
check_AIC(sub_log)
```
9) Check Adjusted $R^2$.
*The choice of predictors to maximize $R^2$ is the same as that to
minimize AIC.*
```{r}
check_adj_r2 <- function(sub){
  rs <- summary(sub)
  adj_r2 <- rs$adjr2
  np <- which.max(adj_r2)
  row <- rs$which[np, ]
  row[row==TRUE]
  # plot(rs$adjr2 ~ I(1:p), xlab="Number of Predictors",
ylab=expression(paste("Adjusted ", R^2)))
}
check_adj_r2(sub_lm)
check_adj_r2(sub_sqrt)
check_adj_r2(sub_log)
```
10) Check Mallows C_p.
*The choices of predictors to minimize Mallows C_p are the same for
all models: `income`, `sex`, and `verbal`.*
```{r}
check_cp <- function(sub){
  rs <- summary(sub)
  cp <- rs$cp
  np <- which.min(cp)
  row <- rs$which[np,]
  row[row==TRUE]
  # plot(rs$cp ~ I(1:p), xlab="Number of Predictors",
ylab=expression(paste(C[p], " Statistic")))
  # abline(1,1)
}
check_cp(sub_lm)
check_cp(sub_sqrt)
check_cp(sub_log)
```
Step 2 Summary: *we can see variables `sex`, `income` and `verbal`
are useful predictors to all 3 models. Thus combined with *Step 1
Summary*, candidate models are now A) log transformed model with full
predictors, B) log transformed model with predictors `sex`, `income` and
`verbal`, and C) the model without transformation and with predictors
`sex`, `income` and `verbal` and with the outlier #24 removed. Model C
fails the normality check. With adjusted R^2 as the principle for RSS,
then model A is the optimal model.*

```{r}
lm_C <- lm(gamble~sex + income + verbal, data=teengamb[-c(24:24),])
check_normality(lm_C, "lm_C normality", FALSE)

```
lm_A <- lm_full_log
summary(lm_A)
```

II. Model Inference
The optimal model is $$gamble = e^{`r round(lm_A$coefficients[1],2)` + `r round(lm_A$coefficients[2],2)`*sex + `r round(lm_A$coefficients[3],2)`*status + `r round(lm_A$coefficients[4],2)`*income + `r round(lm_A$coefficients[5],2)`*verbal}-1$$

Thus assuming all other variables are held constant, a female is expected to spend `r round(exp(lm_A$coefficients[2]),2)` times on (gambling+1) compared to a male in pounds/year.