STAT 500: HW3

Jasmine Mou

10/03/2017

1. Using the `teengamb` data, fit a model with `gamble` as the response and the other variables as predictors.

```
data(teengamb, package='faraway')
# teengamb$sex <- factor(teengamb$sex, levels=c(0,1))
lm_teengamb <- lm(gamble~., data=teengamb)
summary(lm_teengamb)
```

```
##
## Call:
## lm(formula = gamble ~ ., data = teengamb)
##
## 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
```

   (a) Which variables are statistically significant at the 5% level?
   *Variables **sex** and **verbal** are statistically significant at the 5% level.*

   (b) What interpretation should be given to the coefficient for `sex`?
   A female is expected to spend 22.12 less on gambling in pounds per year compared to a male, given their performance on 'status', 'income', 'verbal' are all the same.

   (c) Fit a model with just `income` as a predictor and use an F-test to compare it to the full model.

```
lm_teengamb_income <- lm(gamble~income, data=teengamb)
anova(lm_teengamb_income, lm_teengamb)
```

```
## 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
```

Suppose the null hypothesis is it's enough to just use variable ***income*** *as the predictor, and alternative hypothesis is predictors should include* ***sex***, ***status***, ***income***, *and* ***verbal*** *in the prediction model. The p-value for F-test is 0.01177 < 0.05, thus we have enough evidence to reject the null hypothesis.*

2. Using the `sat` data:

```
data(sat, package='faraway')
```

(a) Fit a model with `total` sat score as the response and `expend`, `ratio` and `salary` as predictors. Test the hypothesis that salary = 0. Test the hypothesis that salary = ratio = expend = 0. Do any of these predictors have an effect on the response?

```
lm_sat <- lm(total~expend + ratio + salary, data=sat)
lm_sat_null0 <- lm(total~expend+ratio, data=sat)
anova(lm_sat_null0, lm_sat)
```

```
## 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
```

The p-value of F-test is 0.06667, greater than the significant level of 0.05. Thus we don't have enough evidence to reject the hypothesis that there is no difference in the effect on the response between the models with and without the predictor ***salary****, given both models have predictors* ***expend*** *and* ***ratio****.*

```
lm_sat_null1 <- lm(total~1, data=sat)
anova(lm_sat_null1, lm_sat)
```

```
## 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
```

The p-value of F-test is 0.01209, less than the significant level of 0.05. Thus we have enough evidence to reject the null hypothesis that there is no difference in the model without any predictors predictors and the model with predictors ***salary***, ***ratio***, ***expend****. These predictors do have effects over the response when they are considered together. Yet we are not sure if they are all important predictors in affecting the response.*

(b) Now add `takers` to the model. Test the hypothesis that takers = 0. Compare this model to the previous one using an F-test. Demonstrate that the F-test and t-test here are equivalent.

```
lm_sat_adding_takers <- lm(total~expend + ratio + salary + takers, data=sat)
anova(lm_sat, lm_sat_adding_takers)
```

```
## 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
```

The p-value of F-test is 2.607e-16, far smaller than level 0.05. It means we have strong evidence to reject the null hypothesis that there is no difference between the models without and with the predictor ***takers****, given predictors* ***expend****,* ***ratio****,* ***salary*** *are available in both models.*

```
summary(lm_sat_adding_takers)
```

```
##
## 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
```

The p-value of t-test is 2.61e-16, which is very close to the p-value of F-test 2.607e-16. And since there is only one predictor ***takers*** *dropped for the null hypothesis compared to the alternative hypothesis, F-test and T-test are equivalent here, with the same conclusion that we have strong evidence to reject the null hypothesis the predictor* ***takers*** *make no difference in the model's effect to the response* ***total****.*

3. Using the `teengamb` data, fit a model with `gamble` as the response and the other variables as predictors.

   (a) Predict the amount that a male with average (given these data) `status`, `income` and `verbal` score would gamble along with an appropriate 95% CI.

```
x <- data.frame(model.matrix(lm_teengamb))
x <- subset(x, sex==0) # male
x_mean <- apply(x,2,mean)
pmean <- predict(lm_teengamb, new=data.frame(t(x_mean)), interval="confidence", level=0.95)
```

The amount that a male with average status, income and verbal score would gamble is 29.775 in pounds per year, with 95% C.I. of (21.1213214, 38.4286786).

   (b) Repeat the prediction for a male with maximal values (for this data) of `status`, `income` and `verbal` score. Which CI is wider and why is this result expected?

```
x_max <- apply(x,2,max)
pmax <- predict(lm_teengamb, new=data.frame(t(x_max)), interval="confidence", level=0.95)
pmax
```

```
##        fit      lwr      upr
## 1 71.30794 42.23237 100.3835
```

The amount that a male with maximal status, income and verbal score would gamble is 71.3079422 in pounds per year, with 95% C.I. of (42.2323674, 100.3835169). The CI for maximal values is wider for mean values, as when the data point move further from the middle of data, there will be a considerable increase in uncertainty, which makes the C.I. wider.

(c) Fit a model with `sqrt(gamble)` as the response but with the same predictors. Now predict the response and give a 95% prediction interval for the individual in (a). Take care to give your answer in the original units of the response.

```
lm_teengamb_sqrt <- lm(sqrt(gamble)~., data=teengamb)
psqr <- predict(lm_teengamb_sqrt, new=data.frame(t(x_mean)), interval="confidence", level=0.95)^2
psqr
```

```
##        fit      lwr      upr
## 1 19.27805 12.93025 26.88931
```

The amount that a male with maximal status, income and verbal score would gamble is 19.2780529 in pounds per year, with 95% C.I. of (12.9302509, 26.8893074).

(d) Repeat the prediction for the model in (c) for a female with `status=20, income=1, verbal = 10`. Comment on the credibility of the result.

```
y <- c(1, 1, 20, 1, 10)
names(y) <- c("X.Intercept", "sex", "status", "income", "verbal")
psqrt_female <- predict(lm_teengamb_sqrt, new=data.frame(t(y)), interval="confidence", level=0.95)^2
psqrt_female
```

```
##        fit      lwr        upr
## 1 4.353398 19.76636 0.07451699
```

The amount that a female with `status=20, income=1, verbal = 10` would gamble is 4.3533977 in pounds per year, with 95% C.I. of (0.074517, 19.76636).

Attachment: RMarkdown Codes

````
---
title: 'STAT 500: HW3'
author: "Jasmine Mou"
date: "10/03/2017"
output:
  pdf_document:
    latex_engine: xelatex
---

1. Using the `teengamb` data, fit a model with `gamble` as the response and the other variables as predictors.

```{r 1}
data(teengamb, package='faraway')
# teengamb$sex <- factor(teengamb$sex, levels=c(0,1))
lm_teengamb <- lm(gamble~., data=teengamb)
summary(lm_teengamb)
```

(a) Which variables are statistically significant at the 5% level?
*Variables `sex` and `verbal` are statistically significant at the 5% level.*

(b) What interpretation should be given to the coefficient for `sex`?
*A female is expected to spend 22.12 less on gambling in pounds per year compared to a male, given their performance on 'status', 'income', 'verbal' are all the same.*

(c) Fit a model with just `income` as a predictor and use an F-test to compare it to the full model.

```{r}
lm_teengamb_income <- lm(gamble~income, data=teengamb)
anova(lm_teengamb_income, lm_teengamb)
```
*Suppose the null hypothesis is it's enough to just use variable `income` as the predictor, and alternative hypothesis is predictors should include `sex`, `status`, `income`, and `verbal` in the prediction model. The p-value for F-test is 0.01177 < 0.05, thus we have enough evidence to reject the null hypothesis.*

2. Using the `sat` data:

```{r 2}
data(sat, package='faraway')
```

(a) Fit a model with `total` sat score as the response and `expend`, `ratio` and `salary` as predictors. Test the hypothesis that β<sub>salary</sub> = 0. Test the hypothesis that β<sub>salary</sub> = β<sub>ratio</sub> = β<sub>expend</sub> = 0. Do any of these predictors have an effect on the response?

```{r}
lm_sat <- lm(total~expend + ratio + salary, data=sat)
lm_sat_null0 <- lm(total~expend+ratio, data=sat)
anova(lm_sat_null0, lm_sat)
```
*The p-value of F-test is 0.06667, greater than the significant level of 0.05. Thus we don't have enough evidence to reject the hypothesis that there is no difference in the effect on the response between the models with and without the predictor `salary`, given both models have predictors `expend` and `ratio`.*

```{r}
lm_sat_null1 <- lm(total~1, data=sat)
anova(lm_sat_null1, lm_sat)
```
*The p-value of F-test is 0.01209, less than the significant level of 0.05. Thus we have enough evidence to reject the null hypothesis that there is no difference in the model without any predictors predictors and the model with predictors `salary`, `ratio`, `expend`. These predictors do have effects over the response when they are considered together. Yet we are not sure if they are all important predictors in affecting the response.*
````

(b) Now add `takers` to the model. Test the hypothesis that β_{takers} = 0. Compare this model to the previous one using an F-test. Demonstrate that the F-test and t-test here are equivalent.

```{r}
lm_sat_adding_takers <- lm(total~expend + ratio + salary + takers, data=sat)
anova(lm_sat, lm_sat_adding_takers)
```
The p-value of F-test is 2.607e-16, far smaller than level 0.05. It means we have strong evidence to reject the null hypothesis that there is no difference between the models without and with the predictor `takers`, given predictors `expend`, `ratio`, `salary` are available in both models.

```{r}
summary(lm_sat_adding_takers)
```
*The p-value of t-test is 2.61e-16, which is very close to the p-value of F-test 2.607e-16. And since there is only one predictor `takers` dropped for the null hypothesis compared to the alternative hypothesis, F-test and T-test are equivalent here, with the same conclusion that we have strong evidence to reject the null hypothesis the predictor `takers` make no difference in the model's effect to the response `total`. *

```{r echo=FALSE}
# fstat <- ((deviance(lm_sat)-deviance(lm_sat_adding_takers))/(df.residual(lm_sat)-
df.residual(lm_sat_adding_takers)))/(deviance(lm_sat_adding_takers)/df.residual(lm_sat_adding_takers))

# tstat <- coef(summary(lm_sat_adding_takers))["takers",1]/coef(summary(lm_sat_adding_takers))
["takers",2]
```

3. Using the `teengamb` data, fit a model with `gamble` as the response and the other variables as predictors.
(a) Predict the amount that a male with average (given these data) `status`, `income` and `verbal` score would gamble along with an appropriate 95% CI.

```{r}
x <- data.frame(model.matrix(lm_teengamb))
x <- subset(x, sex==0) # male
x_mean <- apply(x,2,mean)
pmean <- predict(lm_teengamb, new=data.frame(t(x_mean)), interval="confidence", level=0.95)
```
The amount that a male with average status, income and verbal score would gamble is `r pmean[1]` in pounds per year, with 95% C.I. of (`r pmean[2]`, `r pmean[3]`).

(b) Repeat the prediction for a male with maximal values (for this data) of `status`, `income` and `verbal` score. Which CI is wider and why is this result expected?

```{r}
x_max <- apply(x,2,max)
pmax <- predict(lm_teengamb, new=data.frame(t(x_max)), interval="confidence", level=0.95)
pmax
```
*The amount that a male with maximal status, income and verbal score would gamble is `r pmax[1]` in pounds per year, with 95% C.I. of (`r pmax[2]`, `r pmax[3]`). The CI for maximal values is wider for mean values, as when the data point move further from the middle of data, there will be a considerable increase in uncertainty, which makes the C.I. wider. *

(c) Fit a model with `sqrt(gamble)` as the response but with the same predictors. Now predict the response and give a 95% prediction interval for the individual in (a). Take care to give your answer in the original units of the response.

```{r}
lm_teengamb_sqrt <- lm(sqrt(gamble)~., data=teengamb)
psqr <- predict(lm_teengamb_sqrt, new=data.frame(t(x_mean)), interval="confidence", level=0.95)^2
psqr
```
The amount that a male with maximal status, income and verbal score would gamble is `r psqr[1]` in pounds per year, with 95% C.I. of (`r psqr[2]`, `r psqr[3]`).

(d) Repeat the prediction for the model in (c) for a female with `status=20, income=1, verbal = 10`.

Comment on the credibility of the result.

```{r}
y <- c(1, 1, 20, 1, 10)
names(y) <- c("X.Intercept", "sex", "status", "income", "verbal")
psqrt_female <- predict(lm_teengamb_sqrt, new=data.frame(t(y)), interval="confidence", level=0.95)^2
psqrt_female
```

The amount that a female with `status=20, income=1, verbal = 10` would gamble is `r psqrt_female[1]` in pounds per year, with 95% C.I. of (`r psqrt_female[3]`, `r psqrt_female[2]`).