# STAT 500: HW2

*Jasmine Mou*

*9/26/2017*

1. The dataset `teengamb` concerns a study of teenage gambling in Britain. Fit a regression model with the expenditure on gambling as the response and the `sex`, `status`, `income` and `verbal` score as predictors. Present the output.

```
data(teengamb, package='faraway')
lm_teengamb <- lm(gamble ~ sex + status + income + verbal, data=teengamb)
summary(lm_teengamb)
```

```
##
## Call:
## lm(formula = gamble ~ sex + status + income + verbal, 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) What percentage of variation in the response is explained by these predictors?
   *From the summary result we get 52.67% of variation in the response is explained by these predictors.*

   (b) Which observation has the largest (positive) residual? Give the case number.
   *From the residuals we get the one with the largest (positive) residual is the case 24.*

   (c) Compute the mean and median of the residuals.
   *Mean of the residuals is about 0. Median of the residuals is about* $-1.4514$ *(rounded by 4 digits).*

   (d) Compute the correlation of the residuals with the fitted values.
   *The correlation of the residuals with the fitted values is* $-1.0706588 \times 10^{-16}$.

   (e) Compute the correlation of the residuals with the income.
   *The correlation of the residuals with the income is* $-7.2423817 \times 10^{-17}$.

   (f) For all other predictors held constant, what would be the difference in predicted expenditure on gambling for a male compared to a female?
   *As sex=0 represents male and sex=1 represents female, we can check the difference of interest for a female compared to a male from the estimate of coefficients in sex is -22.1183301.*

2. An experiment was conducted to determine the effect of four factors on the resistivity of a semiconductor `wafer`. The data is found in wafer where each of the four factors is coded as - or + depending on whether

the low or the high setting for that factor was used. Fit the linear model `resist ~ x1 + x2 + x3 + x4`.

```
data(wafer, package='faraway')
lm_wafer <- lm(resist ~ x1 + x2 + x3 + x4, data=wafer)
```

(a) Extract the X matrix using the `model.matrix` function. Examine this to determine how the low and high levels have been coded in the model.

```
X <- model.matrix(~ x1 + x2 + x3 + x4, wafer)
```

*Here the low ("-") becomes 0 and high ("+") becomes 1 in the extracted X matrix.*

(b) Compute the correlation in the X matrix. Why are there some missing values in the matrix?

```
cor(X)
```

```
## Warning in cor(X): the standard deviation is zero
```

```
##             (Intercept) x1+ x2+ x3+ x4+
## (Intercept)           1  NA  NA  NA  NA
## x1+                  NA   1   0   0   0
## x2+                  NA   0   1   0   0
## x3+                  NA   0   0   1   0
## x4+                  NA   0   0   0   1
```

*There are missing values in the matrix as the standard deviations of intercepts, which will be the denominators in the correlation calculation formula, are 0s.*

(c) What difference in resistance is expected when moving from the low to the high level of `x1`?
*From the coefficients values in the linear model, we get the expected difference in resistance when moving from the low to the high of* ***x1*** *is 25.7625.*

(d) Refit the model without `x4` and examine the regression coefficients and standard errors? What stayed the same as the original fit and what changed?

```
lm_wafer_refit <- lm(resist ~ x1 + x2 + x3, data=wafer)
summary(lm_wafer_refit)
```

```
##
## Call:
## lm(formula = resist ~ x1 + x2 + x3, data = wafer)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.137 -20.550   3.575  18.462  41.013
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   229.54      13.32  17.231 7.88e-10 ***
## x1+            25.76      13.32   1.934 0.077047 .
## x2+           -69.89      13.32  -5.246 0.000206 ***
## x3+            43.59      13.32   3.272 0.006677 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.64 on 12 degrees of freedom
## Multiple R-squared:  0.7777, Adjusted R-squared:  0.7221
## F-statistic: 13.99 on 3 and 12 DF,  p-value: 0.0003187
```

*What remains the same: estimated coefficients of* `x1+`, `x2+`, `x3+`. *What changes: the estimated coefficients of the intercept, which increases in the amount of about the half of estimated coefficients of* `x4+` *in the original model; and the standard errors of the intercept,* `x1+`, `x2+`, `x3+`.

(e) Explain how the change in the regression coefficients is related to the correlation matrix of X.
*As there is no changes in the regression coefficients of* `x1+`, `x2+`, `x3+` *before and after adding* `x4+`, *meaning that* `x4+` *has no correlations with the rest predictors being held. This relationship has been verified by the correlation matrix of X, in which the correlation values between* `x4+` *and all other predictors are 0s.*

Attachment: RMarkdown Codes

````
---
title: 'STAT 500: HW2'
author: "Jasmine Mou"
date: "9/26/2017"
output:
  pdf_document:
    latex_engine: xelatex
---

1. The dataset `teengamb` concerns a study of teenage gambling in
Britain. Fit a regression model with the expenditure on gambling as the
response and the `sex`, `status`, `income` and `verbal` score as
predictors. Present the output.
```{r}
data(teengamb, package='faraway')
lm_teengamb <- lm(gamble ~ sex + status + income + verbal, data=teengamb)
summary(lm_teengamb)
```

(a) What percentage of variation in the response is explained by these
predictors?
*From the summary result we get 52.67% of variation in the response is
explained by these predictors.*

(b) Which observation has the largest (positive) residual? Give the case
number.
*From the residuals we get the one with the largest (positive) residual
is the case `r which.max(lm_teengamb$residuals)[[1]]`.*

(c) Compute the mean and median of the residuals.
*Mean of the residuals is about `r round(mean(lm_teengamb$residuals),
digits=4)`. Median of the residuals is about $`r
round(median(lm_teengamb$residuals), digits=4)`$ (rounded by 4 digits).*

(d) Compute the correlation of the residuals with the fitted values.
*The correlation of the residuals with the fitted values is $`r
cor(lm_teengamb$residuals, lm_teengamb$fitted)`$.*

(e) Compute the correlation of the residuals with the income.
*The correlation of the residuals with the income is $`r
cor(lm_teengamb$residuals, teengamb$income)`$.*

(f) For all other predictors held constant, what would be the difference
in predicted expenditure on gambling for a male compared to a female?
*As sex=0 represents male and sex=1 represents female, we can check the
difference of interest for a male compared to a female from the estimate
of coefficients in sex is `r lm_teengamb$coefficients[["sex"]]`.*
````

2. An experiment was conducted to determine the effect of four factors on the resistivity of a semiconductor `wafer`.
The data is found in wafer where each of the four factors is coded as - or + depending on whether the low or the high setting for that factor was used. Fit the linear model `resist` \textasciitilde{} `x1 + x2 + x3 + x4`.

```{r}
data(wafer, package='faraway')
lm_wafer <- lm(resist ~ x1 + x2 + x3 + x4, data=wafer)
```

(a) Extract the X matrix using the `model.matrix` function. Examine this to determine how the low and high levels have been coded in the model.
```{r}
X <- model.matrix(~ x1 + x2 + x3 + x4, wafer)
```
*Here the low ("-") becomes 0 and high ("+") becomes 1 in the extracted X matrix.*

(b) Compute the correlation in the X matrix. Why are there some missing values in the matrix?
```{r}
cor(X)
```
*There are missing values in the matrix as the standard deviations of intercepts, which will be the denominators in the correlation calculation formula, are 0s.*

(c) What difference in resistance is expected when moving from the low to the high level of `x1`?
*From the coefficients values in the linear model, we get the expected difference in resistance when moving from the low to the high of `x1` is `r lm_wafer$coefficients[["x1+"]]`. *

(d) Refit the model without `x4` and examine the regression coefficients and standard errors? What stayed the same as the original fit and what changed?
```{r}
lm_wafer_refit <- lm(resist ~ x1 + x2 + x3, data=wafer)
summary(lm_wafer_refit)
```
*What remains the same: estimated coefficients of `x1+`, `x2+`, `x3+`. What changes: the estimated coefficients of the intercept, which increases in the amount of about the half of estimated coefficients of `x4+` in the original model; and the standard errors of the intercept, `x1+`, `x2+`, `x3+`. *

(e) Explain how the change in the regression coefficients is related to the correlation matrix of X.
*As there is no changes in the regression coefficients of `x1+`, `x2+`,

`x3+` before and after adding `x4+`, meaning that `x4+` has no correlations with the rest predictors being held. This relationship has been verified by the correlation matrix of X, in which the correlation values between `x4+` and all other predictors are 0s.*