HW3 Solutions - Stats 500

Problem 1.

```r
library(faraway)
data(teengamb)
gambmod = lm(gamble ~ ., data = teengamb)
RSS = sum(residuals(gambmod)**2)
DF = df.residual(gambmod)
summary(gambmod)
```

```
##
## 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).

The variables sex and income are significant, with p-values equal to 0.0101 and 1.79e-05, respectively.

(b).

Assuming all other variables are held constant, a female teenager is expected to spend $22.12 less on gambling than a male teenager.

(c).

```r
gambmod2 = lm(gamble ~ income, teengamb)
rss = sum(residuals(gambmod2)**2)
df = df.residual(gambmod2)
F = ((rss - RSS) /(df - DF)) / (RSS / DF)
```

```r
pvalue = pf(F, df - DF, DF, lower.tail = FALSE)
pvalue
```

```
## [1] 0.01177211
```

We conclude that at the 5% level, including the other predicotrs significantly improves the fit of the model.

Problem 2.

(a).

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

```
##
## 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
```

Based on the summary, the p-value for $H_0 : \beta_{salary} = 0$ is 0.0667, so at the 5% level we fail to reject $\beta_{salary} = 0$. On the other hand, the p-value for the F-test with null hypothesis: $\{\beta_{expend} = \beta_{ratio} = \beta_{salary} = 0\}$ is 0.01209 so we can say that at least one of these variables is significant.

(b).

```r
satmod2 = lm(total ~ expend + ratio + salary + takers, data = sat)
summary(satmod2)$coef['takers',]
```

```
##      Estimate    Std. Error       t value      Pr(>|t|)
## -2.904481e+00  2.312600e-01 -1.255937e+01  2.606559e-16
```

The t-test suggests that 'takers' is significant. (p-value $= 2.606559e - 16 \approx 0$.) Now we compute the F-statistic:

```r
RSS = sum(satmod2$residuals**2); DF = df.residual(satmod2)
rss = sum(satmod$residuals**2); df = df.residual(satmod)
F = ((rss - RSS) /(df - DF)) / (RSS / DF)
pvalue = pf(F, df - DF, DF, lower.tail = FALSE)
pvalue
```

```
## [1] 2.606559e-16
```

As expected, the F-test yields the same result.

Problem 3:

(a).

```r
x0 = data.frame(t(c(0, apply(teengamb[2:4], 2, mean))))
colnames(x0)[1] <- "sex"
predict(gambmod, new = x0, interval = 'prediction')
```

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

(b).

```r
x1 = data.frame(t(c(0, apply(teengamb[2:4], 2, max))))
colnames(x1)[1] <- "sex"
predict(gambmod, new = x1, interval = 'prediction')
```

```
##        fit      lwr    upr
## 1 71.30794 17.06588 125.55
```

This confidence interval is wider because of extrapolation.

(c).

```r
transformed.model = lm(sqrt(gamble) ~ ., data = teengamb)
transformed.prediction = predict(transformed.model, new = x0, interval = 'prediction')
transformed.prediction
```

```
##        fit       lwr      upr
## 1 4.049523 -0.245035 8.344082
```

Note that we always have $\sqrt{\text{gamble}} \geq 0$, so the negarive part of the interval is irrelevant, i.e. $\sqrt{Y_{X_0}} \in [0, 8.344082]$ with 95% confidnence. Now if we are 95% confident that $\sqrt{Y_{X_0}} \in [0, 8.344082]$, then we're also equally confident that $Y_{X_0} \in [0, 8.3441^2] = [0, 69.624]$. Hence a 95% (prediction) C.I for for an average male's gambling expenditure is $[0, 69.624]$.

(d).

```r
x2 = x0
x2[1,] <- c(1, 20, 1, 10)
predict(transformed.model, new = x2, interval = 'prediction')
```

```
##         fit       lwr      upr
## 1 -2.08648 -6.908863 2.735903
```

A negative fit for $\sqrt{Y_{X_2}}$ is meaningless, and as before, the negative part of the interval is not informative. So we choose 0 as our fit (point estimator) and $[0, 2.736]$ as the 95% C.I. for $\sqrt{Y_{X_2}}$. Equivalently, $Y_{X_2} \in [0, 7.49]$ with 95% confidence.

We now consider why the predicted value is negative. According to the scatter plot, we can see that female verbal is between 4 and 8. It cannot cover 10 which is a new observation value. Therefore, the fitted model may not be valid for the new observation, or it generates a huge error in terms of prediction. Note also that the lowest income for a female is 1.5.

```r
Income = teengamb$income[which(teengamb$sex == 1)]
Verbal = teengamb$verbal[which(teengamb$sex == 1)]
plot(Income, Verbal, main = 'Verbal vs. Income for Females')
```