Stat 500 – Homework 3 (Solutions)

1. The model is fit and the tests are performed below:

```
> data(sat)
> g<-lm(total ~ takers + ratio + salary, data = sat)
> summary(g)

Coefficients:
              Estimate   Std. Error  t value   Pr(>|t|)
(Intercept)   1057.8982   44.3287    23.865    <2e-16 ***
takers          -2.9134    0.2282   -12.764    <2e-16 ***
ratio           -4.6394    2.1215    -2.187    0.0339 *
salary           2.5525    1.0045     2.541    0.0145 *

Residual standard error: 32.41 on 46 degrees of freedom
Multiple R-Squared: 0.8239,     Adjusted R-squared: 0.8124
F-statistic: 71.72 on 3 and 46 DF,  p-value: < 2.2e-16
```

We can see that the coefficient for *takers* is highly significant (p-value <2e-16) and the coefficients for *ratio* (p-value =0.0339) and *salary* (p-value =0.0145) are marginally significant . Since the multiple R-squared is large (0.8239), one can say that the model fits the data well.

Since the p-value for the t-statistic corresponding to the coefficient of *salary*, i.e., β_{salary} is 0.0145 , we reject the null hypothesis $\beta_{salary} = 0$, when takers and ratio are included in the model.

The p-value (<2.2e-16) for the F-statistic in the above summary indicates that the second hypothesis $\beta_{takers} = \beta_{salary} = \beta_{ratio} = 0$ is rejected . Thus the above regression is significant. In other words, at least one of these predictors has a significant effect on the response.

2. The confidence intervals are obtained using the following commands:

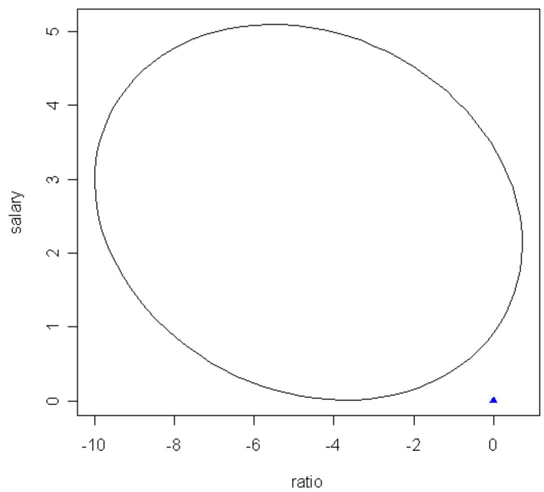
```
> confint(g,"salary",level=.95)
            2.5 %     97.5 %
salary 0.5304797 4.574461

> confint(g,"salary",level=.99)
            0.5 %        99.5 %
salary -0.1466840   5.251624
```

The above confidence intervals show that for $\alpha = 0.05$ we would reject the hypothesis that the coefficient of *salary* is zero, but for $\alpha = 0.01$ we fail to reject it . Hence we can conclude that $0.01 < \text{p-value} < 0.05$.

3. The joint confidence interval plot is generated by the following commands:

```
> library(ellipse)
>plot(ellipse(g,c('ratio','salary')),type="l")
>points(0,0,col=4,pch=17)
```



It is used for testing the hypothesis :

$H_0 : \beta_{salary} = \beta_{ratio} = 0$ vs. H_1 : They are not both equal to zero.
From the plot, we see that (0,0) is not in the ellipse, therefore we reject H_0 at $\alpha = 0.05$.

4. The variable expend is now added to the model :

```
> g1<-lm(total ~ takers + ratio + salary + expend, data = sat)
> summary(g1)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1045.9715	52.8698	19.784	< 2e-16 ***
takers	-2.9045	0.2313	-12.559	2.61e-16 ***
ratio	-3.6242	3.2154	-1.127	0.266
salary	1.6379	2.3872	0.686	0.496
expend	4.4626	10.5465	0.423	0.674

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 coefficients of *ratio* and *salary* change somewhat from before and surprisingly both of them become insignificant . But the coefficient of *takers* is unaltered and its still significant . Introduction of the *expend* variable does not improve the multiple R-squared value a lot (from 0.8239 to 0.8246) . So the addition of covariate *expend* does not seem to improve the fit very much .

```
> anova(g,g1)

Analysis of Variance Table
Model 1: total ~ takers + ratio + salary
Model 2: total ~ takers + ratio + salary + expend

 Res.Df    RSS     Df  Sum of Sq   F    Pr(>F)
1   46    48315
2   45    48124    1     191      0.179   0.6742
```

The F-test for comparing g to g1 yield a p-value approximately equal to 0.6742 so we fail to reject the hypothesis $\beta_{expend} = 0$ at the 0.05 level .

5. To test the hypothesis $\beta_{expend} = \beta_{salary} = \beta_{ratio} = 0$:

```
> g2<-lm(total ~ takers , data = sat)
> summary(g2)
Coefficients:
              Estimate  Std. Error   t value   Pr(>|t|)
(Intercept) 1053.3204    8.2112     128.28  <2e-16 ***
takers         -2.4801    0.1862     -13.32  <2e-16 ***

Residual standard error: 34.89 on 48 degrees of freedom
Multiple R-Squared: 0.787,      Adjusted R-squared: 0.7825
F-statistic: 177.3 on 1 and 48 DF,  p-value: < 2.2e-16

> anova(g2,g1)

Analysis of Variance Table
Model 1: total ~ takers
Model 2: total ~ takers + ratio + salary + expend

     Res.Df  RSS   Df  Sum of Sq     F        Pr(>F)
1       48   5843   3
2       45  48124   3     10309    3.2133    0.03165 *
```

The F-test for comparing g2 (model with only takers as coefficient) to g1(full model) yield a p-value approximately equal to 0.03165 so we reject the hypothesis $\beta_{expend} = \beta_{salary} = \beta_{ratio} = 0$ at the 0.05 level . It is only marginally significant , however (p-value > 0.01) .

Based on the entire analysis , it appears that expend , salary and ratio together have a marginal effect on the response , but because they are correlated (see below) , removing one of them from the model while leaving the other two in does not significantly change the fit . Takers , on the other hand , has a highly significant effect on total SAT scores .

```
> cor(sat$salary,sat$expend)
[1] 0.8698015
> cor(sat$ratio,sat$expend)
[1] -0.3710254
```

2. Based on Chapter 3, problem 5 (p. 51).

Note that general F-statistic corresponds to testing

$$H_o : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \qquad \text{vs.} \qquad H_A : \text{not } H_o$$

So,

$$\begin{aligned}
F &= \frac{(RSS_{H_o} - RSS_{H_o \cup H_A}) \,/\, (df_{H_o} - df_{H_o \cup H_A})}{RSS_{H_o \cup H_A} / df_{H_o \cup H_A}} \\
&= \frac{(RSS_{H_o} - RSS_{H_o \cup H_A}) \,/\, (n - 1 - (n - (p+1)))}{RSS_{H_o \cup H_A} / (n - (p+1))} \\
&= \left\{ \frac{n-p-1}{p} \right\} \cdot \frac{RSS_{H_o} - RSS_{H_o \cup H_A}}{RSS_{H_o \cup H_A}}
\end{aligned}$$

and

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} = 1 - \frac{RSS_{H_o \cup H_A}}{RSS_{H_o}}$$

where the last equality comes from the fact that the RSS under H_o corresponds to only having β_o parameter (all of others are assumed to be 0). So,

$$\begin{aligned}
F &= \left\{ \frac{n-p-1}{p} \right\} \cdot \frac{1 - \frac{RSS_{H_o \cup H_A}}{RSS_{H_o}}}{\frac{RSS_{H_o \cup H_A}}{RSS_{H_o}}} \\
&= \left\{ \frac{n-p-1}{p} \right\} \cdot \frac{1 - (1 - R^2)}{1 - R^2} \\
&= \left\{ \frac{n-p-1}{p} \right\} \cdot \frac{R^2}{1 - R^2}
\end{aligned}$$