STATS 500 Homework 2 YUAN YIN

Problem 1

1. We fit the model with R, the result is as follows:

```
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 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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
```

Figure 1

From the result above, we know that with the increase of one student taking SAT, the average total of test score will decrease around 2.9. what's more, with the increase of average pupil/teacher ratio, the total test score will also decrease, which is reasonable as it means one teacher will teach more students and the education quality will decrease. On the other hand, if the salary of teachers increases, the total SAT score will increase, which is reasonable as teacher will motivate in teaching students. Also as R-squared is 0.8239, which is very closed to 1, we can say that the goodness of fit is quite satisfying.

Now we test the null hypothesis $\beta_{\text{salary}} = 0$,

```
Analysis of Variance Table

Model 1: total ~ takers + ratio
Model 2: total ~ takers + ratio + salary
  Res.Df   RSS Df Sum of Sq      F  Pr(>F)
1     47 55097
2     46 48315  1    6781.6 6.4566 0.01449 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 2

P-value is 0.01449 < 0.05, then we have 95% confidence to reject null hypothesis.

Now we test the null hypothesis $\beta_{\text{takers}} = \beta_{\text{ratio}} = \beta_{\text{salary}} = 0$,

```
Analysis of Variance Table

Model 1: total ~ 1
Model 2: total ~ takers + ratio + salary
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     49 274308
2     46  48315  3    225992 71.721 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 3

P-value is 2.2e-16 < 0.01, then we have 99% confidence to reject null hypothesis.

2. The 95% CI for parameter associated with salary is [0.5304797, 4.574461], and the 99% CI is [-0.146684, 5.251624], we found that value 0 is in the 99% CI but not in the 95% CI, which means we have 95% confidence to reject parameter associated with *salary* is 0 but no more than 99% confidence. In conclusion, we can deduce that the p-value for salary is larger than 0.01 but less than 0.05.

3. The confidence region is as follows:

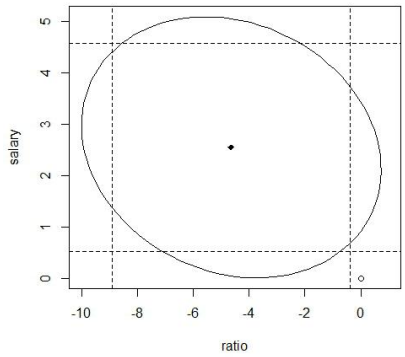

Figure 4

As we see that origin is out of the ellipse, which means if the null hypothesis is $\beta_{\text{salary}} = \beta_{\text{ratio}} = 0$, then we have 95% confidence to reject this null hypothesis.

4. The regression result is as follows:

```
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
---
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
```

Figure 5

We found that parameters of *takers, ratio, salary* also have the same sign with model in question 1, which means their effect on *total* is similar, and the parameter of *expend* is positive, which means with the increase of expenditure of per pupil, the average total SAT score will increase. As we see that the p-value for *ratio* and *salary* is much more larger than in the previous model, which means we won't have enough confidence to reject the null hypothesis that their parameters is 0. This is not what we want to see. R-squared is almost the same with model 1, the goodness of fit is quite the same. In conclusion, the new model is not better than the previous model.

5. If the null hypothesis is $\beta_{\text{salary}} = \beta_{\text{expend}} = \beta_{\text{ratio}} = 0$, we got the p-value is 0.03165, which is less than 0.05. So we have 95% confidence to reject the null hypothesis, comparing with the hypothesis with one of the parameters is 0 (null hypothesis are one of i for $\beta_i = 0$ (i = salary, expend, ratio)), we can't reject $\beta_i = 0$ (i = salary, expend, ratio) singly, so it maybe because these variables have correlations themselves, and they make a correlation effect on the response.

6. The plot is as follows:

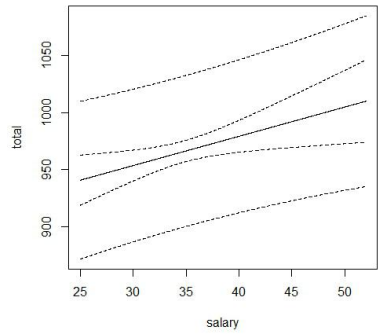

Figure 6

The narrower imaginary line intervals is the CI and the wider one is PI, as we can see that PI almost doesn't curve but CI curves obviously. Also CI is covered by PI from the range above. This is because confidence interval is to estimate the mean of response, which eliminates the effect of error ε , at the same time the prediction to a single data has more uncertainty, so the intervals are more narrow with the same confidence level to PI. Also we found the narrowest intervals is around *salary* equals to 35~40, which means the estimation of mean of *total* has more accuracy when *salary* is in [35,40].

(Remark: H_A is opposite to H_0).

Appendix:

```
library(faraway)
data(sat)
h0a = lm(total ~ takers + ratio + salary, data = sat)
summary(h0a)
h0 = lm(total ~ takers + ratio, data = sat)
anova(h0, h0a)
h00 = lm(total ~ 1, data = sat)
anova(h00, h0a)
conf = confint(h0a, "salary", level = 0.95)
conf
conf = confint(h0a, "salary", level = 0.99)
conf
library(ellipse)
plot(ellipse(h0a, c('ratio', 'salary'), level = 0.95), type = "l", xlim = c(-10,1))
points(0, 0, pch = 1)
points(h0a$coef['ratio'], h0a$coef['salary'], pch = 18)
abline(v = confint(h0a)[3,], lty = 2)
abline(h = confint(h0a)[4,], lty = 2)
ha = lm(total ~ takers + ratio + salary + expend, data = sat)
summary(ha)
h000 = lm(total ~ takers, data = sat)
anova(h000, ha)
grid = seq(25,52,0.01)
pred = predict(h0a, data.frame(salary = grid, takers = 35, ratio = 17), interval = "confidence")
pred1 = predict(h0a, data.frame(salary = grid, takers = 35, ratio = 17), interval = "prediction")
pre = cbind(pred, pred1)
matplot(grid, pre, lty = c(1,2,2), col = 1, type = "l", xlab = "salary", ylab = "total")
```