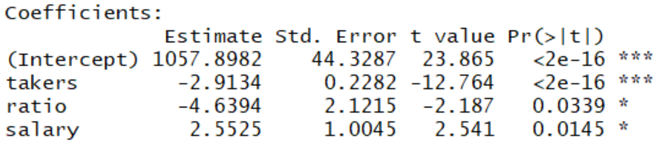
Stats500 Homework 3 *Di Lu, Oct.4, 2017*

1. fit a linear model and test hypothesis

Set the linear regression model as: total ~ takers + ratio + salary



```
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 *  
```

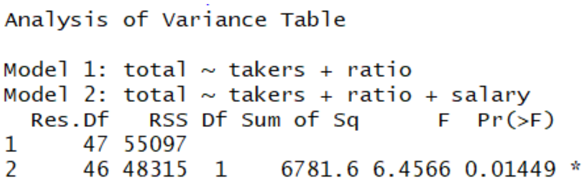
- The coefficient for takers is -2.9134, which means 1 percent higher of eligible students taking SAT, the SAT scores will be lower by 2.9134 on average(significant at 0.1% level)
- The coefficient for ratio is -4.6394, which means 1 percent higher of pupil/teacher ratio in schools, the SAT scores will be lower by 4.6394 on average(significant at 5% level) . This emphasizes teacher's attention at student's studies.
- The coefficient for salary is 2.5522, which means every one dollar increase of teacher's annual salary, the SAT scores will be higher by 2.5525 on average. This means teachers perfo rmances are stimulated by pay(significant at 5% level)
- R-square =0.8239, which means 82.39% variation of response can be explained by model predictors.

To test the hypothesis:

H0: β (salary)=0

Ha: β (salary)≠0

let model_s be lm(total~takers+ratio) and do anova compared to original model. p-value = 0.01449, which means at 5% significance level, we can reject the null hypothesis β(salary)=0



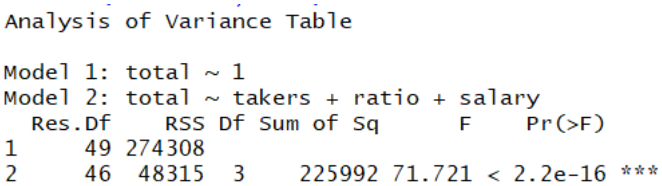
```
Analysis of Variance Table

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 *
```

To test the hypothesis:

H0: β(takers)=β(ratio)=β(salary)=0

Ha: H0 not true

let model_trs be lm(total~1) do anova compared to original model. p-value < 2.2e-16, which means at 5% significance level, we can reject the null hypothesis β(takers)=β(ratio)=β(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 ***
```

2. compute CI and compare with p-value

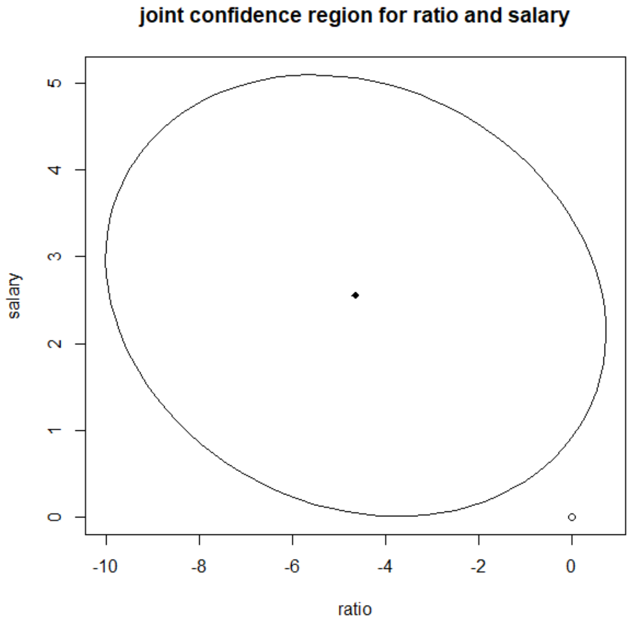
95% CI for salary is [0.5304797, 4.5744605]

99% CI for salary is [-0.146684, 5.251624]

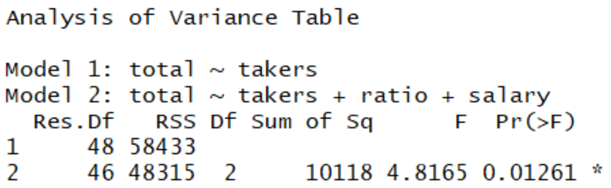
We can see 95% CI does not contain 0, which is equivalent to that at 5% significance level we reject hypothesis β(salary)=0 We can see 99% CI contains 0, which is equivalent to that at 1% significance level we do not have enough evidence to reject hypothesis β(salary)=0 p-value in the regression for salary is 0.0145, which coincides with lying between 95% and 99% CI.

3. compute 95% joint confidence region for ratio and salary

Seen from the graph, the ellipse centers on (-4.6394, 2.5525), and origin lies outside the region, which means at 5% significance level we can reject null hypothesis β(ratio)=β(salary)=0



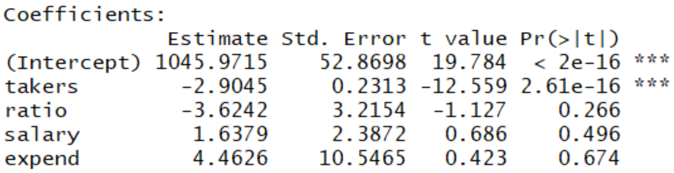
To test the hypothesis β(ratio)=β(salary)=0 let model_rs be lm(total~takers) and do anova compared to original model p-value = 0.01261, which means at 5% significance level, we can reject the hypothesis β(ratio)=β(salary)=0.



```
Analysis of Variance Table

Model 1: total ~ takers
Model 2: total ~ takers + ratio + salary
  Res.Df   RSS Df Sum of Sq      F  Pr(>F)
1     48 58433
2     46 48315  2     10118 4.8165 0.01261 *
```

4. add expend to model and comapre to original one

Set the linear regression model as: total ~ takers + ratio + salary + expend



```
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
```

- The coefficient for takers is -2.9045, which means 1 percent higher of eligible students taking SAT, the SAT scores will be lower by 2.9045 on average.(same, significant at 0.1% level)

- The coefficient for ratio is -3.6242, which means 1 percent higher of pupil/teacher ratio in schools, the SAT scores will be lower by 3.6242 on average. (size smaller than in model1, not significant at 10% level)
- The coefficient for salary is 1.6379, which means every one dollar increase of teacher's annual salary, the SAT scores will be higher by 1.6379 on average.(size smaller than in model1, not significant at 10% level)
- The coefficient for expend is 4.4626, which means every one dollar expenditure increase per pupil, the SAT scores will be higher by 4.4626 on average.(not significant at 10% level)
- R-square =0.8246, which means 82.46% variation of response can be explained by model1 predictors. This is just slightly higher than model1. But adjusted R-square drops, so the model's goodness of fit drops considering degree of freedom.

5. test salary=expend=ratio=0, are these predictors effective?

To test the hypothesis:

H0: β(ratio)=β(salary)=β(expend)=0

Ha: H0 not true

let model2_ser be lm(total~takers) and do anova compared to original model. p-value = 0.03165, which means at 5% significance level, we can reject the hypothesis β(ratio)=β(salary)=β(expend)=0

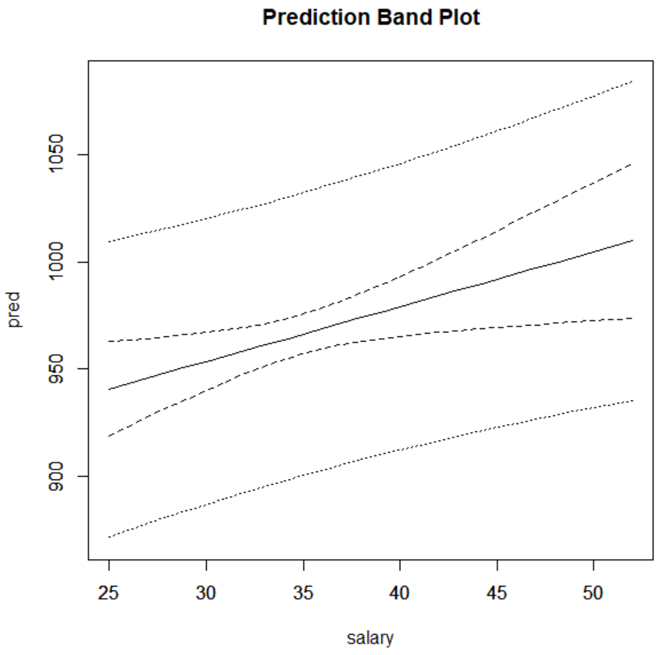
In total, these predictors have an effect on the response, R-square increases from 0.787 to 0.8246. But *expend may have overlapping with salary and ratio, because* (1) add expend variable reduce effect size of salary and ratio and make them insignificant. (2) with only expend, R-square is already 0.8195 and significant at 1%. So we can say expend tells nearly the same thing as ratio and salary. *Meanwhile, expend has no correlation with takers*, whose coefficient and significant level doesn't change much. So in my opinion, we should keep takes and expend.

6. generate predicted value in model1, show CI and PI

The predicted values are ploted as the solid line. It is linear as set by the model and other conditions are takers = 35 and ratio = 17, and range of salary being 25 to 52.

95% CI for mean is shown as the dashed line, which is narrower around the sample mean. 95% PI for actual score is shown as the dotted line, which is broader than the CI band because of the variance of response.

We can say that as for mean, we can predict more precisely, especially around sample mean, but for actual score, errors also include variance of responce, so it is larger.



Appendix: Code in R used in homework 3

```
library(faraway)
attach(sat)
model <- lm(total~takers+ratio+salary) #1
summary(model)

model_s <- lm(total~takers+ratio)
model_trs <- lm(total~1)
anova(model_s,model)
anova(model_trs,model)

conf95 <- confint(model,level = 0.95) #2
conf99 <- confint(model,level = 0.99)

library(ellipse) #3 # plot the confidence region
plot(ellipse(model,c('ratio','salary')),type='l',main="joint confidence region for ratio and salary") #
choose level = 1-0.01261, origin will be on the border
points(model$coefficients['ratio'],model$coefficients['salary'],pch=18) # add the estimates
points(0,0,pch =1) # add the origin


model_rs <- lm(total~takers)
anova(model_rs,model)
model2 <- lm(total~takers+ratio+salary+expend)#4
summary(model2)
model2_ser <- lm(total~takers+expend)#5
summary(model2_ser)
anova(model2_ser,model2)

grid=seq(25,52,1)#6
x0 <- data.frame(takers=35, ratio=17,salary=grid)
pred_mean <- predict(model,x0,interval="confidence")
pred_actual <- predict(model,x0,interval="prediction")
matplot(grid,pred_mean,lty = c(1,2,2),col=1,type =
"l",ylim=c(870,1085),xlab="salary",ylab="pred",main="Prediction Band Plot")
par(new=TRUE)
matplot(grid,pred_actual,lty = c(1,3,3),col=1,type = "l",ylim=c(870,1085),xlab="",ylab="")
rug(salary)
```