Prediction intervals with R

```
> sat = read.table("http://www.utstat.utoronto.ca/~brunner/302f13/code_n_data/
                    lecture/sat.data")
> apply(sat,2,mean)
VERBAL   MATH    GPA
595.65 649.53   2.63
> mod1 = lm(GPA ~ VERBAL+MATH, data=sat); summary(mod1)

Call:
lm(formula = GPA ~ VERBAL + MATH, data = sat)

Residuals:
     Min       1Q   Median       3Q      Max
-2.24875 -0.35113  0.04659  0.38745  1.03527

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.6062975  0.4414062   1.374    0.171
VERBAL      0.0023072  0.0005522   4.178 4.42e-05 ***
MATH        0.0009999  0.0006093   1.641    0.102
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5484 on 197 degrees of freedom
Multiple R-squared: 0.1161, Adjusted R-squared: 0.1071
F-statistic: 12.93 on 2 and 197 DF,  p-value: 5.284e-06

> betahat = mod1$coefficients; betahat
 (Intercept)       VERBAL         MATH
0.6062974824 0.0023071729 0.0009998537
>
> apply(sat,2,mean)
VERBAL   MATH    GPA
595.65 649.53   2.63
> # Predict GPA for an "average" student, with Verbal=596, Math=650
> yhat0 = betahat[1] + 596*betahat[2] + 650*betahat[3]; yhat0
(Intercept)
   2.631277
>
> # You knew it all along. Use the predict function.
> # Need to create a data frame corresponding to the vector x0.
>
> avstudent = data.frame(VERBAL=596,MATH=650) # Remember R is case sensitive.
> avstudent
  VERBAL MATH
1    596  650
> predict(mod1,newdata=avstudent)
       1
2.631277
> # How about a prediction INTERVAL?
```

$$\mathbf{x}_0'\widehat{\boldsymbol{\beta}} \pm t_{\alpha/2}\, s \sqrt{1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}$$

```
predict(mod1,newdata=avstudent, interval='prediction') # 95% interval by default
       fit      lwr      upr
1 2.631277 1.547127 3.715428
>
```

```
> # Look at a variety of student performance on the SAT.
> student1 = data.frame(VERBAL=400,MATH=400) # Bad on both
> student2 = data.frame(VERBAL=400,MATH=800) # Bad verbal, great math
> student3 = data.frame(VERBAL=800,MATH=400) # Great verbal, bad math
> student4 = data.frame(VERBAL=800,MATH=800) # Legendary
> students = rbind(student1,student2,student3,student4)
> predict(mod1,newdata=students, interval='prediction')
       fit       lwr      upr
1 1.929108 0.7996838 3.058532
2 2.329050 1.2000523 3.458047
3 2.851977 1.6894492 4.014505
4 3.251919 2.1403888 4.363449
>
> # It's not a bad idea to "predict" the observed data.
> # Just look at the first 10 rows, for example
> cbind(sat$GPA,predict(mod1,interval='prediction'))[1:10,]
            fit      lwr      upr
1   2.6 2.552592 1.453707 3.651477
2   2.3 2.124685 1.016973 3.232397
3   2.4 2.789707 1.703429 3.875986
4   3.0 2.674888 1.587223 3.762554
5   3.1 2.975051 1.882648 4.067454
6   2.9 2.848074 1.758515 3.937634
7   3.1 2.588596 1.503828 3.673364
8   3.3 2.815626 1.728040 3.903212
9   2.3 2.713493 1.628171 3.798816
10  3.3 2.818393 1.731568 3.905218
Warning message:
In predict.lm(mod1, interval = "prediction") :
  Predictions on current data refer to _future_ responses
```