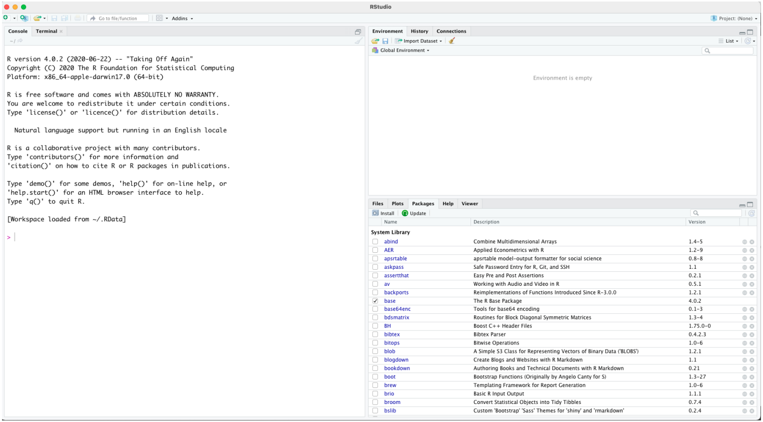
6.5 Practical Econometrics from Chapter 6

Regressions from Chapter 6, Table 6.1

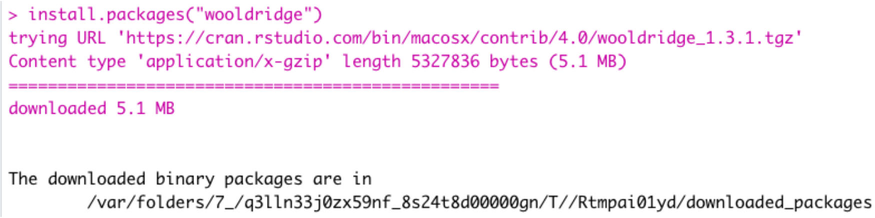
- Open R through R-studio. Your console should look like this



```
R version 4.0.2 (2020-06-22) -- "Taking Off Again"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin17.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]

>
```

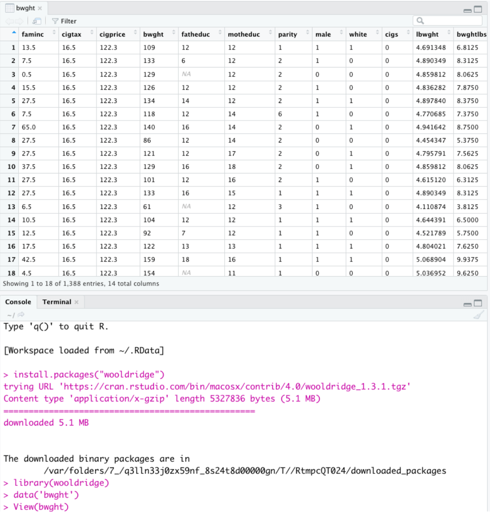
	Name	Description	Version
System Library			
□	abind	Combine Multidimensional Arrays	1.4-5
□	AER	Applied Econometrics with R	1.2-9
□	apsrtable	apsrtable model-output formatter for social science	0.8-8
□	askpass	Safe Password Entry for R, Git, and SSH	1.1
□	assertthat	Easy Pre and Post Assertions	0.2.1
□	av	Working with Audio and Video in R	0.5.1
□	backports	Reimplementations of Functions Introduced Since R-3.0.0	1.2.1
☑	base	The R Base Package	4.0.2
□	base64enc	Tools for base64 encoding	0.1-3
□	bdsmatrix	Routines for Block Diagonal Symmetric Matrices	1.3-4
□	BH	Boost C++ Header Files	1.75.0-0
□	bibtex	Bibtex Parser	0.4.2.3
□	bitops	Bitwise Operations	1.0-6
□	blob	A Simple S3 Class for Representing Vectors of Binary Data ('BLOBS')	1.2.1
□	blogdown	Create Blogs and Websites with R Markdown	1.1
□	bookdown	Authoring Books and Technical Documents with R Markdown	0.21
□	boot	Bootstrap Functions (Originally by Angelo Canty for S)	1.3-27
□	brew	Templating Framework for Report Generation	1.0-6
□	brio	Basic R Input Output	1.1.1
□	broom	Convert Statistical Objects into Tidy Tibbles	0.7.4
□	bslib	Custom 'Bootstrap' 'Sass' Themes for 'shiny' and 'rmarkdown'	0.2.4

- Install data from the package wooldridge. It loads all the datasets we require in R.



```
> install.packages("wooldridge")
trying URL 'https://cran.rstudio.com/bin/macosx/contrib/4.0/wooldridge_1.3.1.tgz'
Content type 'application/x-gzip' length 5327836 bytes (5.1 MB)
==================================================
downloaded 5.1 MB


The downloaded binary packages are in
	/var/folders/7_/q3lln33j0zx59nf_8s24t8d00000gn/T//Rtmpai01yd/downloaded_packages
```

- ▷ Load the *BWGHT* dataset from the package called wooldridge and then have a look at the data. You should get something like this

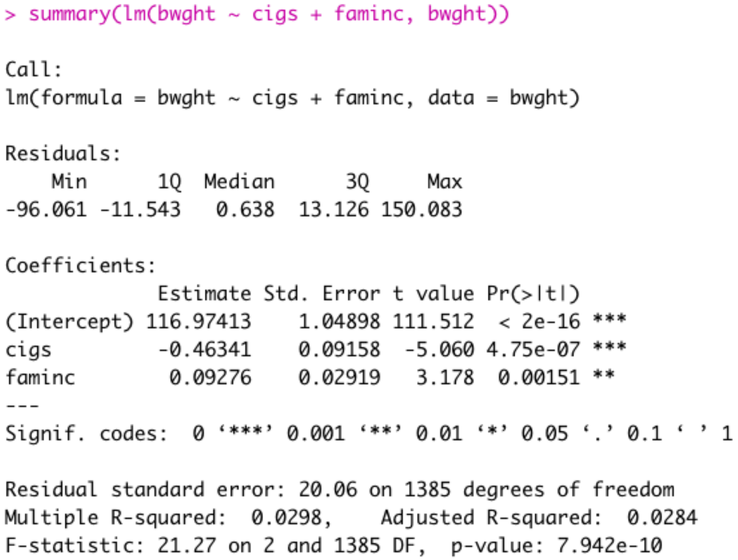


	faminc	cigtax	cigprice	bwght	fatheduc	motheduc	parity	male	white	cigs	lbwght	bwghtlbs
1	13.5	16.5	122.3	109	12	12	1	1	1	0	4.691348	6.8125
2	7.5	16.5	122.3	133	6	12	2	1	0	0	4.890349	8.3125
3	0.5	16.5	122.3	129	NA	12	2	0	0	0	4.859812	8.0625
4	15.5	16.5	122.3	126	12	12	2	1	0	0	4.836282	7.8750
5	27.5	16.5	122.3	134	14	12	2	1	1	0	4.897840	8.3750
6	7.5	16.5	122.3	118	12	14	6	1	0	0	4.770685	7.3750
7	65.0	16.5	122.3	140	16	14	2	0	1	0	4.941642	8.7500
8	27.5	16.5	122.3	86	12	14	2	0	0	0	4.454347	5.3750
9	27.5	16.5	122.3	121	12	17	2	0	1	0	4.795791	7.5625
10	37.5	16.5	122.3	129	16	18	2	0	1	0	4.859812	8.0625
11	27.5	16.5	122.3	101	12	16	2	1	0	0	4.615120	6.3125
12	27.5	16.5	122.3	133	16	15	1	1	1	0	4.890349	8.3125
13	6.5	16.5	122.3	61	NA	12	3	1	0	0	4.110874	3.8125
14	10.5	16.5	122.3	104	12	12	1	1	1	0	4.644391	6.5000
15	12.5	16.5	122.3	92	7	12	1	1	0	0	4.521789	5.7500
16	17.5	16.5	122.3	122	13	13	1	1	1	0	4.804021	7.6250
17	42.5	16.5	122.3	159	18	16	1	1	1	0	5.068904	9.9375
18	4.5	16.5	122.3	154	NA	11	1	0	0	0	5.036952	9.6250

Showing 1 to 18 of 1,388 entries, 14 total columns

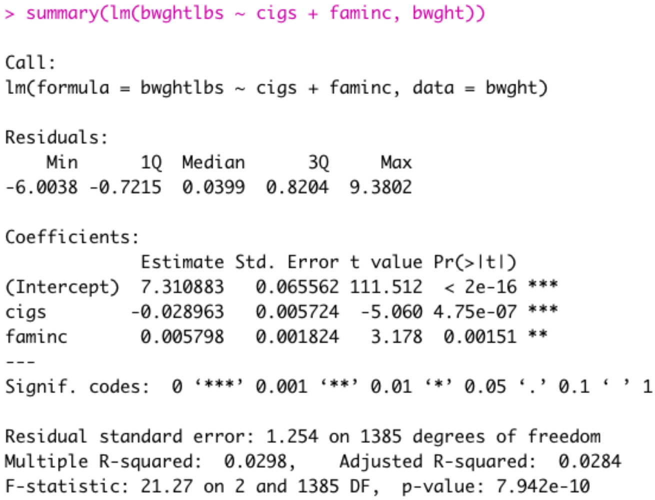
```
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]

> install.packages("wooldridge")
trying URL 'https://cran.rstudio.com/bin/macosx/contrib/4.0/wooldridge_1.3.1.tgz'
Content type 'application/x-gzip' length 5327836 bytes (5.1 MB)
==================================================
downloaded 5.1 MB


The downloaded binary packages are in
        /var/folders/7_/q3lln33j0zx59nf_8s24t8d00000gn/T//RtmpcQT024/downloaded_packages
> library(wooldridge)
> data('bwght')
> View(bwght)
```

- Lets try and recreate Table [6.1]



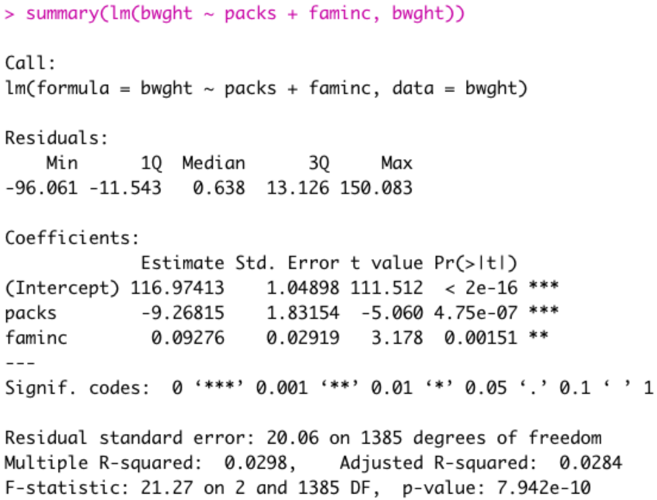
```
> summary(lm(bwght ~ cigs + faminc, bwght))

Call:
lm(formula = bwght ~ cigs + faminc, data = bwght)

Residuals:
    Min      1Q  Median      3Q     Max 
-96.061 -11.543   0.638  13.126 150.083 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 116.97413    1.04898 111.512  < 2e-16 ***
cigs         -0.46341    0.09158  -5.060 4.75e-07 ***
faminc        0.09276    0.02919   3.178  0.00151 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.06 on 1385 degrees of freedom
Multiple R-squared:  0.0298,	Adjusted R-squared:  0.0284 
F-statistic: 21.27 on 2 and 1385 DF,  p-value: 7.942e-10
```

- Changing the birth weight from ounces to lbs



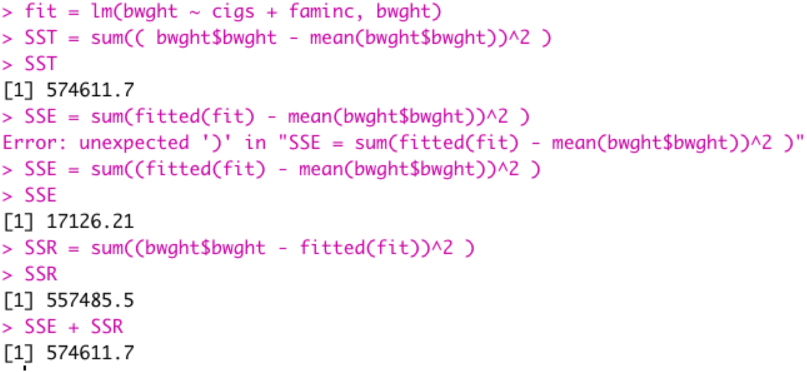
```
> summary(lm(bwghtlbs ~ cigs + faminc, bwght))

Call:
lm(formula = bwghtlbs ~ cigs + faminc, data = bwght)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.0038 -0.7215  0.0399  0.8204  9.3802 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.310883   0.065562 111.512  < 2e-16 ***
cigs        -0.028963   0.005724  -5.060 4.75e-07 ***
faminc       0.005798   0.001824   3.178  0.00151 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.254 on 1385 degrees of freedom
Multiple R-squared:  0.0298,	Adjusted R-squared:  0.0284 
F-statistic: 21.27 on 2 and 1385 DF,  p-value: 7.942e-10
```

- Changing the birth weight back to ounces, but now we change independent variable to packs of cigs



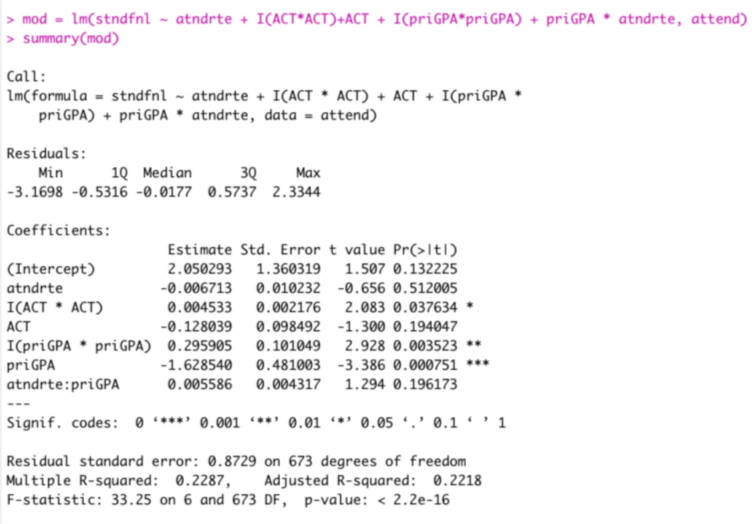
```
> summary(lm(bwght ~ packs + faminc, bwght))

Call:
lm(formula = bwght ~ packs + faminc, data = bwght)

Residuals:
    Min      1Q  Median      3Q     Max 
-96.061 -11.543   0.638  13.126 150.083 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 116.97413    1.04898 111.512  < 2e-16 ***
packs        -9.26815    1.83154  -5.060 4.75e-07 ***
faminc        0.09276    0.02919   3.178  0.00151 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.06 on 1385 degrees of freedom
Multiple R-squared:  0.0298,	Adjusted R-squared:  0.0284 
F-statistic: 21.27 on 2 and 1385 DF,  p-value: 7.942e-10
```

- Obtaining SST, SSE and SSR from the regressions



```
> fit = lm(bwght ~ cigs + faminc, bwght)
> SST = sum(( bwght$bwght - mean(bwght$bwght))^2 )
> SST
[1] 574611.7
> SSE = sum(fitted(fit) - mean(bwght$bwght))^2 )
Error: unexpected ')' in "SSE = sum(fitted(fit) - mean(bwght$bwght))^2 )"
> SSE = sum((fitted(fit) - mean(bwght$bwght))^2 )
> SSE
[1] 17126.21
> SSR = sum((bwght$bwght - fitted(fit))^2 )
> SSR
[1] 557485.5
> SSE + SSR
[1] 574611.7
```

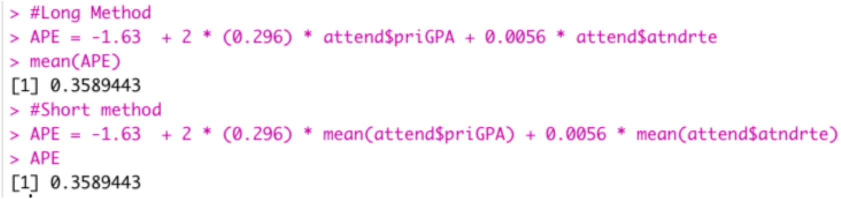
- Lets try and recreate example [6.19]



```
> mod = lm(stndfnl ~ atndrte + I(ACT*ACT)+ACT + I(priGPA*priGPA) + priGPA * atndrte, attend)
> summary(mod)

Call:
lm(formula = stndfnl ~ atndrte + I(ACT * ACT) + ACT + I(priGPA *
    priGPA) + priGPA * atndrte, data = attend)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1698 -0.5316 -0.0177  0.5737  2.3344

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         2.050293   1.360319   1.507 0.132225
atndrte            -0.006713   0.010232  -0.656 0.512005
I(ACT * ACT)        0.004533   0.002176   2.083 0.037634 *
ACT                -0.128039   0.098492  -1.300 0.194047
I(priGPA * priGPA)  0.295905   0.101049   2.928 0.003523 **
priGPA             -1.628540   0.481003  -3.386 0.000751 ***
atndrte:priGPA      0.005586   0.004317   1.294 0.196173
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8729 on 673 degrees of freedom
Multiple R-squared:  0.2287,	Adjusted R-squared:  0.2218
F-statistic: 33.25 on 6 and 673 DF,  p-value: < 2.2e-16
```

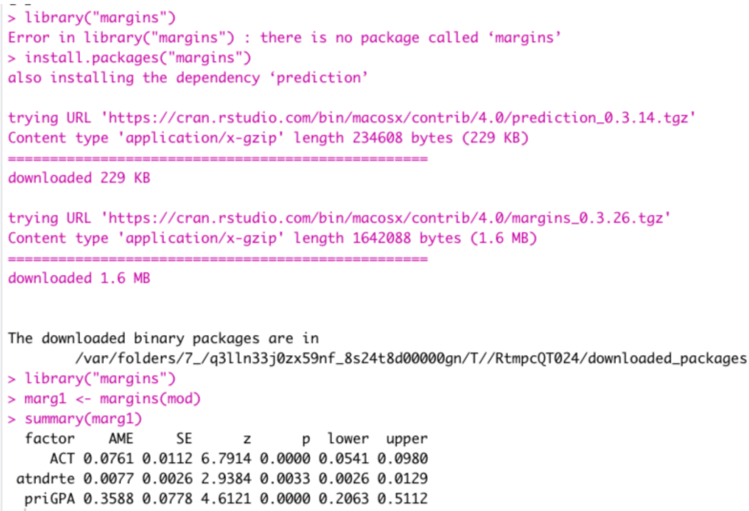
- Calculating APE for priGPA using two manual methods.

Note: You should try and do these manual methods in excel or R in order to get clarity about what is being done. One representative example might be done in class.



```
> #Long Method
> APE = -1.63  + 2 * (0.296) * attend$priGPA + 0.0056 * attend$atndrte
> mean(APE)
[1] 0.3589443
> #Short method
> APE = -1.63  + 2 * (0.296) * mean(attend$priGPA) + 0.0056 * mean(attend$atndrte)
> APE
[1] 0.3589443
```

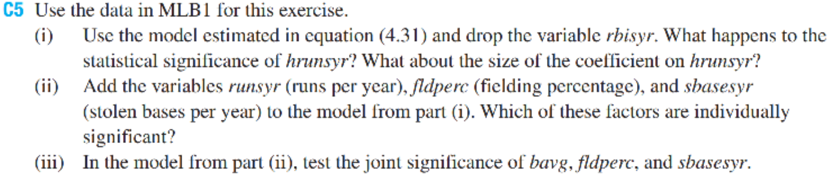
- Another method using the margins package.



```
> library("margins")
Error in library("margins") : there is no package called ‘margins’
> install.packages("margins")
also installing the dependency ‘prediction’

trying URL 'https://cran.rstudio.com/bin/macosx/contrib/4.0/prediction_0.3.14.tgz'
Content type 'application/x-gzip' length 234608 bytes (229 KB)
==================================================
downloaded 229 KB

trying URL 'https://cran.rstudio.com/bin/macosx/contrib/4.0/margins_0.3.26.tgz'
Content type 'application/x-gzip' length 1642088 bytes (1.6 MB)
==================================================
downloaded 1.6 MB


The downloaded binary packages are in
        /var/folders/7_/q3lln33j0zx59nf_8s24t8d00000gn/T//RtmpcQT024/downloaded_packages
> library("margins")
> marg1 <- margins(mod)
> summary(marg1)
  factor    AME     SE      z      p  lower  upper
     ACT 0.0761 0.0112 6.7914 0.0000 0.0541 0.0980
 atndrte 0.0077 0.0026 2.9384 0.0033 0.0026 0.0129
  priGPA 0.3588 0.0778 4.6121 0.0000 0.2063 0.5112
```

C5 Chapter 4 using data on MLB1

This question is tackled here to introduce you to the MLB dataset.



C5 Use the data in MLB1 for this exercise.

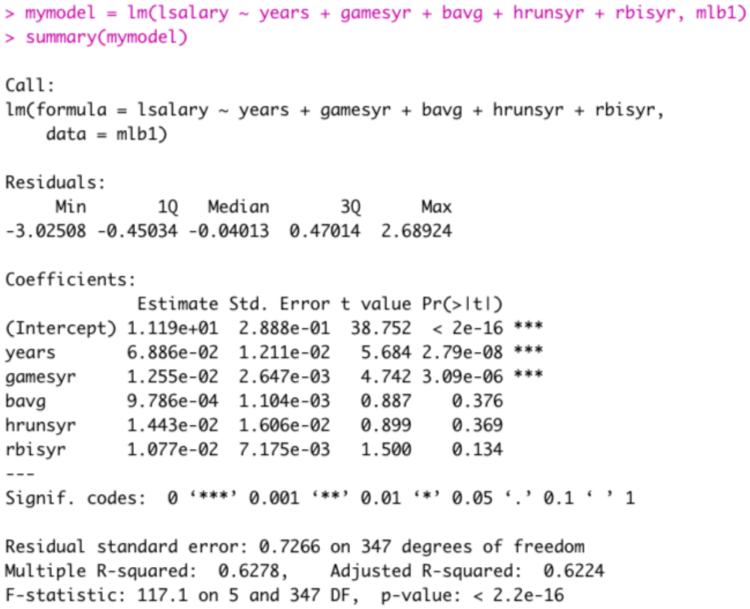
(i) Use the model estimated in equation (4.31) and drop the variable *rbisyr*. What happens to the statistical significance of *hrunsyr*? What about the size of the coefficient on *hrunsyr*?

(ii) Add the variables *runsyr* (runs per year), *fldperc* (fielding percentage), and *sbasesyr* (stolen bases per year) to the model from part (i). Which of these factors are individually significant?

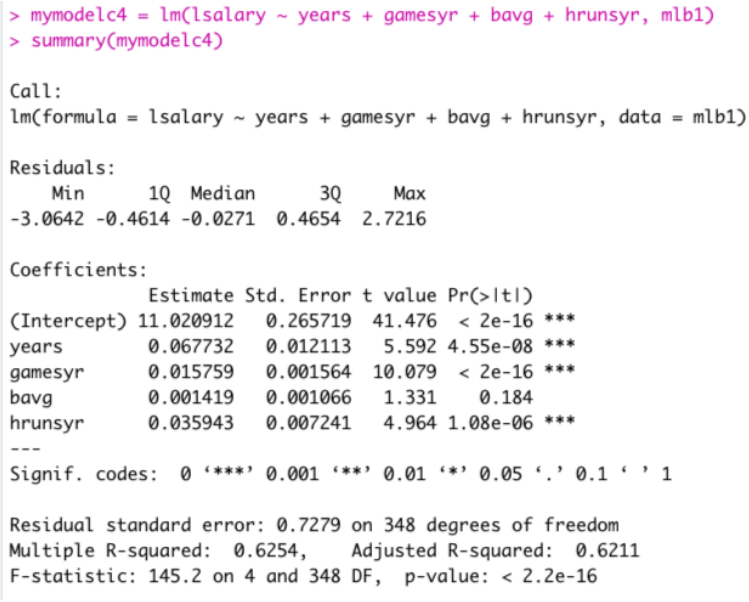
(iii) In the model from part (ii), test the joint significance of *bavg*, *fldperc*, and *sbasesyr*.

(a) Use the model estimated in equation (4.31) and drop the variable rbisyr. What happens to the statistical significance of hrunsyr? What about the size of the coefficient on hrunsyr? Recall the model:

$$\begin{aligned} log(salary) =& \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg \\ & + \beta_4 hrunsyr + \beta_5 rbisyr \end{aligned}$$

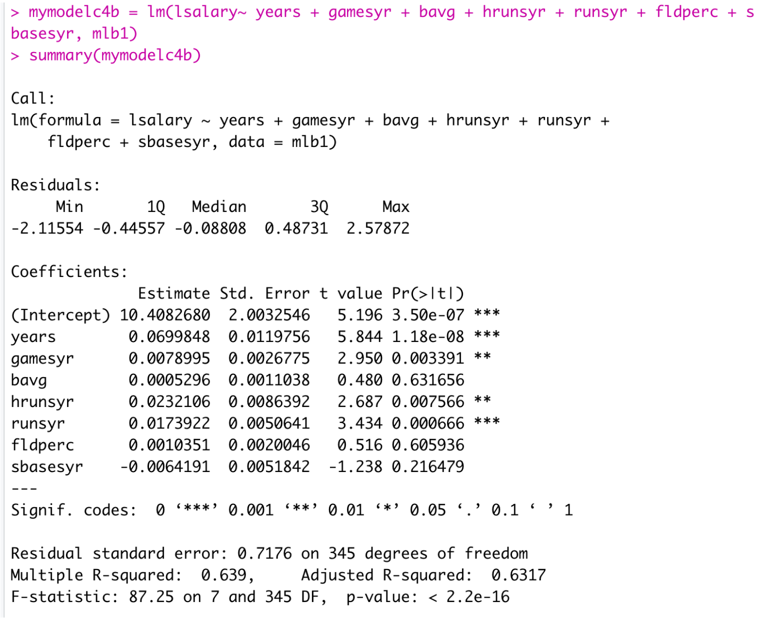


```
> mymodel = lm(lsalary ~ years + gamesyr + bavg + hrunsyr + rbisyr, mlb1)
> summary(mymodel)

Call:
lm(formula = lsalary ~ years + gamesyr + bavg + hrunsyr + rbisyr,
    data = mlb1)

Residuals:
     Min       1Q   Median       3Q      Max 
-3.02508 -0.45034 -0.04013  0.47014  2.68924 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.119e+01  2.888e-01  38.752  < 2e-16 ***
years       6.886e-02  1.211e-02   5.684 2.79e-08 ***
gamesyr     1.255e-02  2.647e-03   4.742 3.09e-06 ***
bavg        9.786e-04  1.104e-03   0.887    0.376    
hrunsyr     1.443e-02  1.606e-02   0.899    0.369    
rbisyr      1.077e-02  7.175e-03   1.500    0.134    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7266 on 347 degrees of freedom
Multiple R-squared:  0.6278,	Adjusted R-squared:  0.6224 
F-statistic: 117.1 on 5 and 347 DF,  p-value: < 2.2e-16
```



```
> mymodelc4 = lm(lsalary ~ years + gamesyr + bavg + hrunsyr, mlb1)
> summary(mymodelc4)

Call:
lm(formula = lsalary ~ years + gamesyr + bavg + hrunsyr, data = mlb1)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.0642 -0.4614 -0.0271  0.4654  2.7216 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 11.020912   0.265719  41.476  < 2e-16 ***
years        0.067732   0.012113   5.592 4.55e-08 ***
gamesyr      0.015759   0.001564  10.079  < 2e-16 ***
bavg         0.001419   0.001066   1.331    0.184    
hrunsyr      0.035943   0.007241   4.964 1.08e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7279 on 348 degrees of freedom
Multiple R-squared:  0.6254,	Adjusted R-squared:  0.6211 
F-statistic: 145.2 on 4 and 348 DF,  p-value: < 2.2e-16
```

(b) Add the variables $runsyr$ (runs per year), $fldperc$ (fielding percentage), and $sbasesyr$ (stolen bases per year) to the model from part (a). Which of these factors are individually significant?



```
> mymodelc4b = lm(lsalary~ years + gamesyr + bavg + hrunsyr + runsyr + fldperc + s
basesyr, mlb1)
> summary(mymodelc4b)

Call:
lm(formula = lsalary ~ years + gamesyr + bavg + hrunsyr + runsyr + 
    fldperc + sbasesyr, data = mlb1)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.11554 -0.44557 -0.08808  0.48731  2.57872 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 10.4082680  2.0032546   5.196 3.50e-07 ***
years        0.0699848  0.0119756   5.844 1.18e-08 ***
gamesyr      0.0078995  0.0026775   2.950 0.003391 ** 
bavg         0.0005296  0.0011038   0.480 0.631656    
hrunsyr      0.0232106  0.0086392   2.687 0.007566 ** 
runsyr       0.0173922  0.0050641   3.434 0.000666 ***
fldperc      0.0010351  0.0020046   0.516 0.605936    
sbasesyr    -0.0064191  0.0051842  -1.238 0.216479    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7176 on 345 degrees of freedom
Multiple R-squared:  0.639,	Adjusted R-squared:  0.6317 
F-statistic: 87.25 on 7 and 345 DF,  p-value: < 2.2e-16
```

(c) In the model from part (b), test the joint significance of bavg, fldperc, and sbasesyr

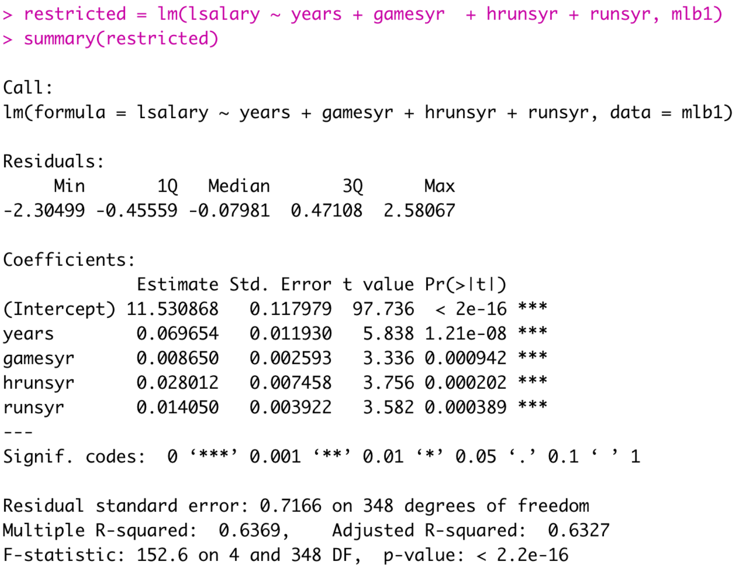
To test the joint significance of the three variables, we can use the F-statisitcs with the following null hypothesis

$$H_0 : \beta_3 = \beta_6 = \beta_7 = 0$$

This null comprises of 3 **exclusion restrictions**. We need to know by how much SSR increases when we drop the variables bavg, fldperc and sbaseyr from the model estimated in part (b). SSR would always increase when we add variables, but does it increase in a statistically significant manner?

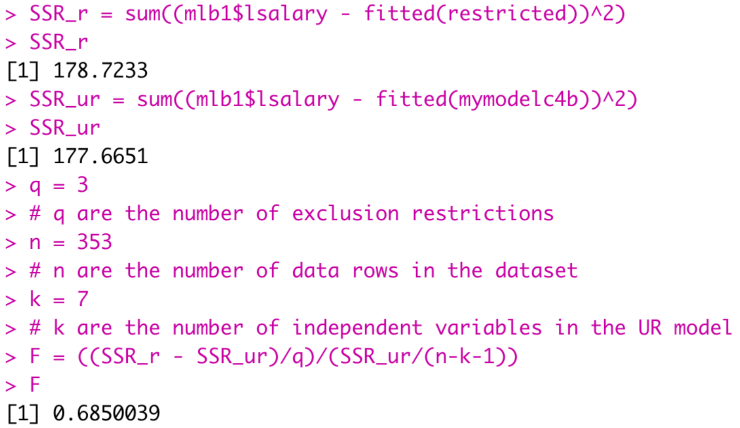
In this context, we can run a **restricted model** where restrictions are dictated by the null. In other words, if the null is true, our model would look like the following:

$$\begin{aligned} +log(salary) =& \beta_0 + \beta_1 years + \beta_2 gamesyr \\ &+ \beta_4 hrunsyr + \beta_5 runsyr \end{aligned}$$



```
> restricted = lm(lsalary ~ years + gamesyr  + hrunsyr + runsyr, mlb1)
> summary(restricted)

Call:
lm(formula = lsalary ~ years + gamesyr + hrunsyr + runsyr, data = mlb1)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.30499 -0.45559 -0.07981  0.47108  2.58067 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 11.530868   0.117979  97.736  < 2e-16 ***
years        0.069654   0.011930   5.838 1.21e-08 ***
gamesyr      0.008650   0.002593   3.336 0.000942 ***
hrunsyr      0.028012   0.007458   3.756 0.000202 ***
runsyr       0.014050   0.003922   3.582 0.000389 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7166 on 348 degrees of freedom
Multiple R-squared:  0.6369,	Adjusted R-squared:  0.6327 
F-statistic: 152.6 on 4 and 348 DF,  p-value: < 2.2e-16
```

Now let us calculate the F-statistic:

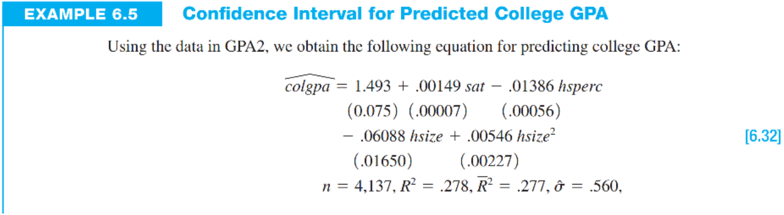


```
> SSR_r = sum((mlb1$lsalary - fitted(restricted))^2)
> SSR_r
[1] 178.7233
> SSR_ur = sum((mlb1$lsalary - fitted(mymodelc4b))^2)
> SSR_ur
[1] 177.6651
> q = 3
> # q are the number of exclusion restrictions
> n = 353
> # n are the number of data rows in the dataset
> k = 7
> # k are the number of independent variables in the UR model
> F = ((SSR_r - SSR_ur)/q)/(SSR_ur/(n-k-1))
> F
[1] 0.6850039
```

A low F-value with q,n-k-1 degree of freedom implies that we fail to reject the null hypothesis which implies that the variables are jointly insignificant which often justifies dropping them from the model. (page 143)

Example 6.5: Prediction Intervals

Consider the following example in the text:

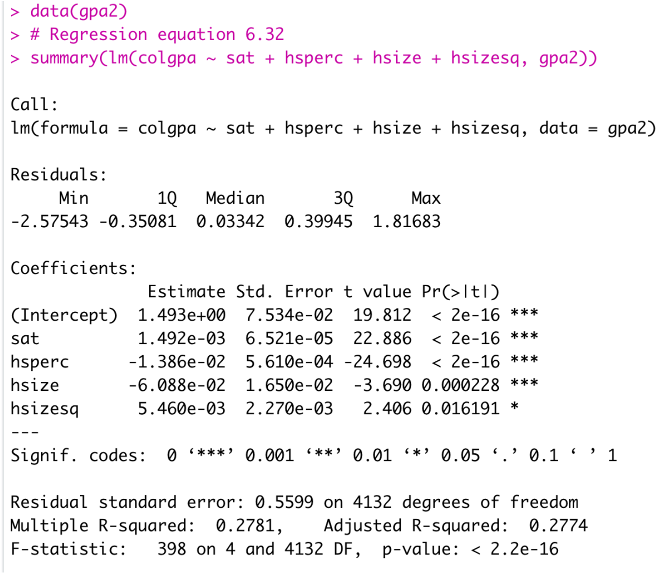


EXAMPLE 6.5 **Confidence Interval for Predicted College GPA**

Using the data in GPA2, we obtain the following equation for predicting college GPA:

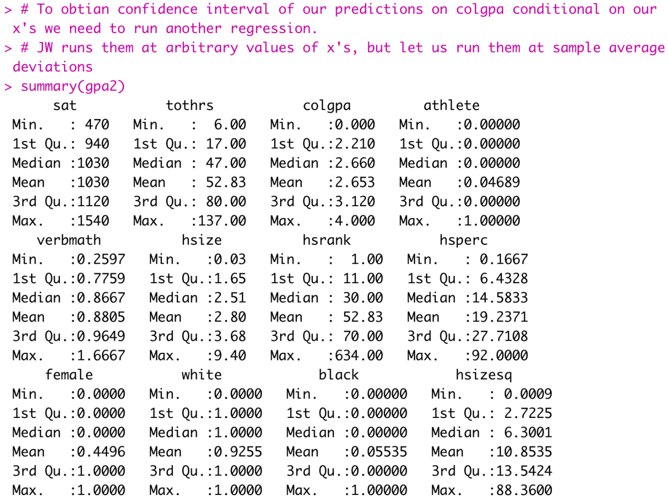
$$\begin{aligned}\widehat{colgpa} = {} & \underset{(0.075)}{1.493} + \underset{(.00007)}{.00149}\, sat - \underset{(.00056)}{.01386}\, hsperc \\ & - \underset{(.01650)}{.06088}\, hsize + \underset{(.00227)}{.00546}\, hsize^2 \qquad [6.32] \\ & n = 4{,}137,\ R^2 = .278,\ \bar{R}^2 = .277,\ \hat{\sigma} = .560,\end{aligned}$$

We estimate it as follows:



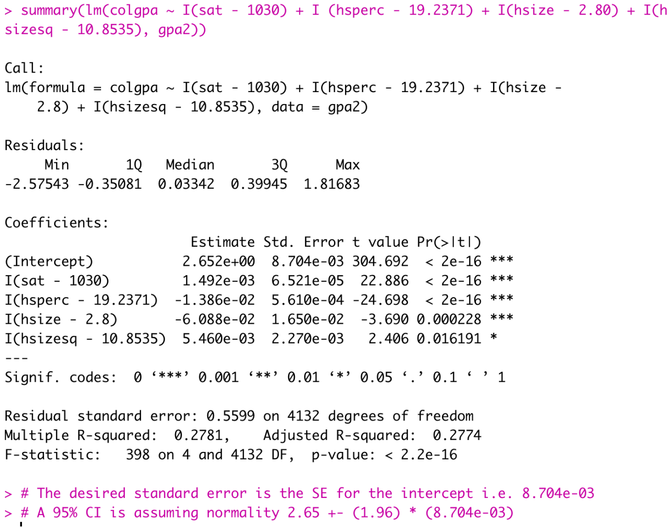
```
> data(gpa2)
> # Regression equation 6.32
> summary(lm(colgpa ~ sat + hsperc + hsize + hsizesq, gpa2))

Call:
lm(formula = colgpa ~ sat + hsperc + hsize + hsizesq, data = gpa2)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.57543 -0.35081  0.03342  0.39945  1.81683 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.493e+00  7.534e-02  19.812  < 2e-16 ***
sat          1.492e-03  6.521e-05  22.886  < 2e-16 ***
hsperc      -1.386e-02  5.610e-04 -24.698  < 2e-16 ***
hsize       -6.088e-02  1.650e-02  -3.690 0.000228 ***
hsizesq      5.460e-03  2.270e-03   2.406 0.016191 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5599 on 4132 degrees of freedom
Multiple R-squared:  0.2781,	Adjusted R-squared:  0.2774 
F-statistic:   398 on 4 and 4132 DF,  p-value: < 2.2e-16
```

Instead of the values which are given in text, we use the sample means for the respective variable to generate in-sample confidence intervals. We begin by finding the sample means.



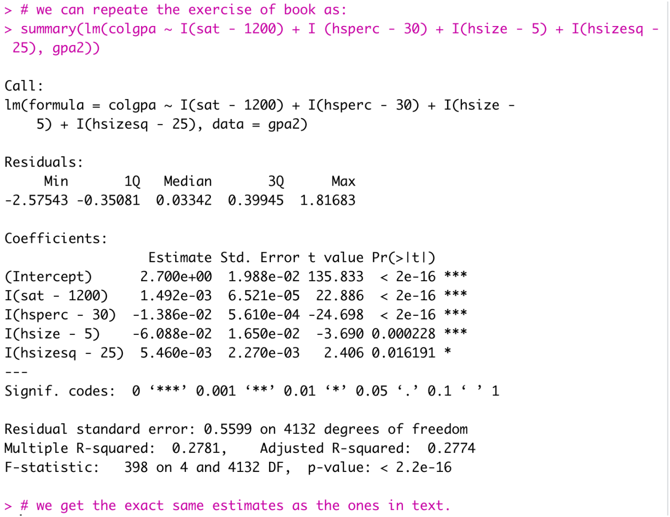
```
> # To obtian confidence interval of our predictions on colgpa conditional on our
 x's we need to run another regression.
> # JW runs them at arbitrary values of x's, but let us run them at sample average
 deviations
> summary(gpa2)
      sat           tothrs           colgpa          athlete       
 Min.   : 470   Min.   :  6.00   Min.   :0.000   Min.   :0.00000  
 1st Qu.: 940   1st Qu.: 17.00   1st Qu.:2.210   1st Qu.:0.00000  
 Median :1030   Median : 47.00   Median :2.660   Median :0.00000  
 Mean   :1030   Mean   : 52.83   Mean   :2.653   Mean   :0.04689  
 3rd Qu.:1120   3rd Qu.: 80.00   3rd Qu.:3.120   3rd Qu.:0.00000  
 Max.   :1540   Max.   :137.00   Max.   :4.000   Max.   :1.00000  
    verbmath          hsize           hsrank           hsperc       
 Min.   :0.2597   Min.   :0.03   Min.   :  1.00   Min.   : 0.1667  
 1st Qu.:0.7759   1st Qu.:1.65   1st Qu.: 11.00   1st Qu.: 6.4328  
 Median :0.8667   Median :2.51   Median : 30.00   Median :14.5833  
 Mean   :0.8805   Mean   :2.80   Mean   : 52.83   Mean   :19.2371  
 3rd Qu.:0.9649   3rd Qu.:3.68   3rd Qu.: 70.00   3rd Qu.:27.7108  
 Max.   :1.6667   Max.   :9.40   Max.   :634.00   Max.   :92.0000  
     female           white            black            hsizesq       
 Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   : 0.0009  
 1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:0.00000   1st Qu.: 2.7225  
 Median :0.0000   Median :1.0000   Median :0.00000   Median : 6.3001  
 Mean   :0.4496   Mean   :0.9255   Mean   :0.05535   Mean   :10.8535  
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:13.5424  
 Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :88.3600  
```

We estimate the model at the de-mean values as:



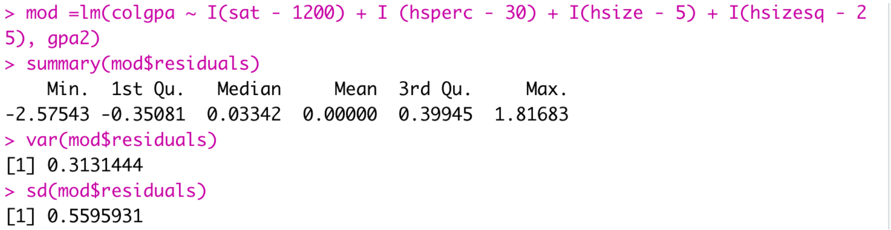
```
> summary(lm(colgpa ~ I(sat - 1030) + I (hsperc - 19.2371) + I(hsize - 2.80) + I(h
sizesq - 10.8535), gpa2))

Call:
lm(formula = colgpa ~ I(sat - 1030) + I(hsperc - 19.2371) + I(hsize -
    2.8) + I(hsizesq - 10.8535), data = gpa2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.57543 -0.35081  0.03342  0.39945  1.81683

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           2.652e+00  8.704e-03 304.692  < 2e-16 ***
I(sat - 1030)         1.492e-03  6.521e-05  22.886  < 2e-16 ***
I(hsperc - 19.2371)  -1.386e-02  5.610e-04 -24.698  < 2e-16 ***
I(hsize - 2.8)       -6.088e-02  1.650e-02  -3.690 0.000228 ***
I(hsizesq - 10.8535)  5.460e-03  2.270e-03   2.406 0.016191 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5599 on 4132 degrees of freedom
Multiple R-squared:  0.2781,    Adjusted R-squared:  0.2774
F-statistic:   398 on 4 and 4132 DF,  p-value: < 2.2e-16

> # The desired standard error is the SE for the intercept i.e. 8.704e-03
> # A 95% CI is assuming normality 2.65 +- (1.96) * (8.704e-03)
```

We can check our method by running the same exercise as is done in JW:



```
> # we can repeate the exercise of book as:
> summary(lm(colgpa ~ I(sat - 1200) + I (hsperc - 30) + I(hsize - 5) + I(hsizesq -
 25), gpa2))

Call:
lm(formula = colgpa ~ I(sat - 1200) + I(hsperc - 30) + I(hsize -
    5) + I(hsizesq - 25), data = gpa2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.57543 -0.35081  0.03342  0.39945  1.81683

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.700e+00  1.988e-02 135.833  < 2e-16 ***
I(sat - 1200)    1.492e-03  6.521e-05  22.886  < 2e-16 ***
I(hsperc - 30)  -1.386e-02  5.610e-04 -24.698  < 2e-16 ***
I(hsize - 5)    -6.088e-02  1.650e-02  -3.690 0.000228 ***
I(hsizesq - 25)  5.460e-03  2.270e-03   2.406 0.016191 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5599 on 4132 degrees of freedom
Multiple R-squared:  0.2781,    Adjusted R-squared:  0.2774
F-statistic:   398 on 4 and 4132 DF,  p-value: < 2.2e-16

> # we get the exact same estimates as the ones in text.
```

In order to make out of sample predictions we need variance of the prediction error as in 6.36.



```
> mod =lm(colgpa ~ I(sat - 1200) + I (hsperc - 30) + I(hsize - 5) + I(hsizesq - 2
5), gpa2)
> summary(mod$residuals)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
-2.57543 -0.35081  0.03342  0.00000  0.39945  1.81683 
> var(mod$residuals)
[1] 0.3131444
> sd(mod$residuals)
[1] 0.5595931
```

The above standard error is $\hat{\sigma}^2$. We know that

$$se(\hat{e}^0) = \{[se(\hat{y}^0)]^2 + \hat{\sigma}^2\}^{1/2}$$

The $se(\hat{y}^0)$ was obtained previously as 1.988e-02. This allows us to build confidence interval as:

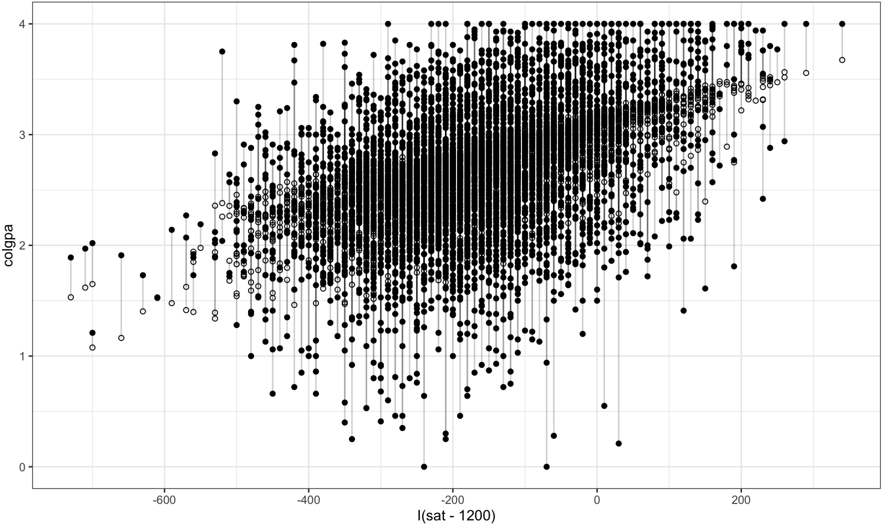
$$\hat{y}^0 \pm t_{0.025} se(\hat{e}^0)$$

To plot fitted values against residuals, when you're plotting one variable.

```
> library(ggplot2)
> mod =lm(colgpa ~ I(sat - 1200) + I (hsperc - 30) + I(hsize - 5) + I(hsizesq - 2
5), gpa2)
> gpa2$predict = predict(mod)
> gpa2$residuals = residuals(mod)
> # Here I have saved the fitted values and residuals as data in the GPA2 file
> ggplot(gpa2, aes(x = I(sat - 1200), y = colgpa)) +
+     geom_segment(aes(xend = I(sat - 1200), yend = predict), alpha = .2) +  # Lin
es to connect points
+     geom_point() +  # Points of actual values
+     geom_point(aes(y = predict), shape = 1) +  # Points of predicted values
+     theme_bw()
```

More on residual v fitted values here

```
> ggsave("fig65.png",width = 9.4,height = 5.64,dpi = 300)
```


colgpa
I(sat - 1200)