

Solutions To Homework Assignment 4

Warmup Problems

(1) CIs and PIs

(a) A **confidence interval** gives you a range of values which you are (reasonably) sure includes the AVERAGE value of Y associated with a given X. In other words, if you found many data points with that X, then the average value of the corresponding Y's would lie in the interval. A **prediction interval** gives you a range of values which you are (reasonably) certain includes the value of Y associated with a SINGLE data point or the NEXT data point at a particular value of X. To emphasize: you use a CI if you are interested in the AVERAGE Y for a given value of X and you use a PI if you are interested in a SINGLE or PARTICULAR Y for a given value of X. As an example, suppose I want to predict peoples' heights (Y) based on their ages (X). If I want to predict the AVERAGE height of all 10 year olds I use a confidence interval for Y when X=10. However, suppose my cousin has just had a baby girl named Susan and I want to predict what Susan's height will be when she is 10. Then I want a prediction interval because I am trying to predict the height of a SINGLE 10 year old. The prediction interval will always be wider than the confidence interval because it is harder to predict for a single person than for lots of people. In the example above, think of it this way. Any single child could be really tall or short making it hard to guess in advance. However, if I want the average for all children, some will be tall, some short and these will balance each other out in the average. Similarly, it is much easier for me to guess what the average score for the class will be on a midterm than it is for me to guess what any individual student's score will be. The formulas for the CI and PI reflect this difference. Look at the formulas for the variances or standard deviations—the extra 1 in the PI formula makes it wider. I have written the formulas for the variances for the CIs and PIs below because they are a little easier to explain. You of course take the square root to get the standard deviation which you use in the confidence interval formula.

$$CI: \quad s_{Y_0}^2 = MSE \left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{SSX} \right)$$

$$PI: \quad s_{Y_0}^2 = MSE \left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{SSX} \right) = MSE + s_{Y_0}^2$$

Both formulas contain the variance of the points about the line, MSE . This makes sense. The more variable the data is, the harder it will be to make predictions, so you will need the CIs and PIs to be wider. Both formulas contain $1/n$. This also makes sense. The more data you have, the more accurate your estimates, the better your predictions will be, and the narrower your CIs and PIs can be. Finally, both intervals contain a piece which says how far the value you are predicting at, X_0 is from the middle of your data set, \bar{X} . The further you get from your data, the less reliable your predictions will be. As we have discussed many times, predicting too far outside the range of the data is risky. Thus the further X_0 is from \bar{X} the wider the CIs and PIs should be. These factors all affect how accurate the value ON THE ESTIMATED REGRESSION LINE IS. If you got the line EXACTLY right then you would make NO error in predicting the average value of Y for a given X because that is what the line represents. However, you would still make errors in the predictions of INDIVIDUAL points because individual points vary about the population regression line. Thus for individual predictions (PIs) you should have an extra factor measuring the variability of points about the regression line. But this is just MSE again. This is why we have the extra 1 in the PI formula. It gives us the extra variability of individual points about the regression line which is the added difficulty in

predicting a single point as opposed to the average value (what is on the line).

(b) The 95% confidence intervals for b_0 and b_1 are given in the STATA/SAS printouts below. The interval for b_0 tells us that we are 95% sure that the average sedation level of people with no blood cortisol is between -29.05 and 12.67 points. As noted in the HW3 solutions negative values for sedation do not make sense and in fact, b_0 may not be a very meaningful quantity in this problem. Since the interval includes 0 it is possible that 0 cortisol corresponds to 0 sedation. The interval for b_1 tells us we are 95% sure that a 1 $\mu\text{g}/\text{dL}$ increase in cortisol level is associated with between a 4.48-8.87 increase in sedation score. This is a fairly wide range, presumably because we only have 7 data points, but it is entirely above 0, telling us we are sure the relationship between cortisol level and sedation is a positive one.

(c) There is a significant relationship between cortisol and sedation if $\beta_1 \neq 0$. Based on the confidence interval from part (b), which is not even close to including 0 we are at least 95% sure that there is a significant (in fact positive) relationship between cortisol level and sedation. We can test this formally using either an overall F test (which asks if the model as a whole is useful as we did on homework 3) or a t-test for cortisol (which asks if that particular variable makes a useful contribution to the model). In a simple linear regression, asking whether one variable is useful is the same as asking if the model as a whole is useful so the F test and the t-test have the same hypotheses, p-value and conclusions. All that differs is the test statistic. I have included the two tests together here for reference.

$H_0 : \beta_1 = 0$ —there is no (linear) relationship between cortisol level and sedation. The model is not useful for explaining variation in sedation.

$H_A : \beta_1 \neq 0$ —there is a relationship between blood cortisol and sedation; the model is useful overall or is a significant improvement over not using cortisol level to help predict sedation.

Viewed as an F-test, our test statistic is $F_{obs} = MSR/MSE = 61.27$, the ratio of explained to unexplained variability. I got the F statistic from the STATA/SAS printouts below. The corresponding p-value is $P(F_{1,5} \geq 61.27) = .0005$. (Note that STATA gives the p-value to more decimal places for the F test in the ANOVA table than for the t-test—I have no idea why but it is good to look there first!) Since this p-value is much smaller than $\alpha = .05$ we reject the null hypothesis and conclude that there is a significant relationship between blood cortisol level and sedation.

Viewed as a t-test, our test statistic is $t_{obs} = b_1/s_{b_1} = 7.83$ and the corresponding 2-sided p-value is $2P(t_5 \geq 7.83) = .0005$ (or .001 if you use the rounded version from the coefficients table in STATA). The conclusions are the same as for the F test. Note also that the F statistic in this case is just the square of the t statistic: $7.83^2 = 61.27$. This is always the case in simple linear regression.

(d) The STATA and SAS printouts are given below. Note that in STATA I had to fit thre regression model and then immediately after type follow-up commands. In SAS I had to include the X values for my new points as part of the data set. Then when I do the commands for printing out the intervals as part of the regression command I get predictions and intervals for them as well as the points used to fit the model. We see that we are 95% sure that the average sedation score of people with a blood cortisol level of 12 micrograms/deciliter is between 65 and 78.9. From the PI we are 95% sure that the sedation score of an individual person with a cortisol level of 12 $\mu/\text{deciliter}$ is between 59.7 and 84.1, a much broader range. Some individuals may have values this extreme but the average will not be.

It is important to understand the distinction between CIs for the intercept and slope and CIs and PIs for Y at a particular value of X. The CI for β_0 , the incercept is actually a CI for Y at a particular X, namely $X = 0$. However the CI for β_1 is a CI for the cverage CHANGE in Y associated with a one unit CHANGE IN X. It is not associated with a particular value of X at all.

IN STATA:

```
. reg sedation cortisol
```

Source	SS	df	MS	Number of obs =	7
Model	933.271821	1	933.271821	F(1, 5) =	61.27
Residual	76.1567506	5	15.2313501	Prob > F =	0.0005
				R-squared =	0.9246
				Adj R-squared =	0.9095
Total	1009.42857	6	168.238095	Root MSE =	3.9027

sedation	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
cortisol	6.676446	.8529243	7.83	0.001	4.483935	8.868958
_cons	-8.186748	8.116109	-1.01	0.359	-29.04987	12.67637

For the CI there are two choices:

```
. lincom _cons + 12*cortisol
```

```
( 1) 12 cortisol + _cons = 0
```

sedation	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
(1)	71.93061	2.693906	26.70	0.000	65.0057	78.85552

```
. adjust cortisol = 12, se ci
```

```
Dependent variable: sedation      Command: regress
Covariate set to value: cortisol = 12
```

All	xb	stdp	lb	ub
	71.9306	(2.69391)	[65.0057	78.8555]

```
Key:  xb      = Linear Prediction
      stdp     = Standard Error
      [lb , ub] = [95% Confidence Interval]
```

For the prediction interval:

```
. adjust cortisol = 12, stdf ci
```

Dependent variable: sedation Command: regress
 Covariate set to value: cortisol = 12

All	xb	stdf	lb	ub
	71.9306	(4.7422)	[59.7404	84.1208]

Key: xb = Linear Prediction
 stdf = Standard Error (forecast)
 [lb , ub] = [95% Prediction Interval]

IN SAS:

Note: Here I added a cortisol value of 12 to the data set and asked SAS to get CIs and PIs for all the cortisol values in the set. Point 8 on the list below is the one with the CI and PI we are looking for.

```
proc reg data = work.hw4;
model sedation = cortisol/clb clm cli;
run;
```

The REG Procedure
 Model: MODEL1
 Dependent Variable: sedation sedation

Number of Observations Read	29
Number of Observations Used	7
Number of Observations with Missing Values	22

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	933.27184	933.27184	61.27	0.0005
Error	5	76.15673	15.23135		
Corrected Total	6	1009.42857			

Root MSE	3.90274	R-Square	0.9246
Dependent Mean	54.28571	Adj R-Sq	0.9095
Coeff Var	7.18925		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t	95% Confidence Limits
----------	-------	----	--------------------	----------------	---------	---------	-----------------------

Intercept	Intercept	1	-8.18675	8.11611	-1.01	0.3594	-29.04987	12.67637
cortisol	cortisol	1	6.67645	0.85292	7.83	0.0005	4.48394	8.86896

The REG Procedure
Model: MODEL1
Dependent Variable: sedation sedation

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean	95% CL Predict	Residual
1	32.0000	35.8778	2.7760	28.7419 43.0137	23.5665 48.1891	-3.8778
2	42.0000	41.2190	2.2277	35.4926 46.9453	29.6674 52.7705	0.7810
3	52.0000	50.5660	1.5497	46.5822 54.5497	39.7717 61.3603	1.4340
4	61.0000	56.5748	1.5038	52.7091 60.4404	45.8235 67.3261	4.4252
5	62.0000	61.9159	1.7681	57.3710 66.4609	50.9021 72.9297	0.0841
6	65.0000	70.5953	2.5529	64.0329 77.1577	58.6073 82.5833	-5.5953
7	66.0000	63.2512	1.8676	58.4505 68.0519	52.1295 74.3730	2.7488
8	.	71.9306	2.6939	65.0057 78.8555	59.7404 84.1208	.

(2) Interpreting A Multiple Regression Equation:

(a) **False.** It is never safe to say that a change in X **causes** a change in Y. Just because X and Y are related does not mean one causes the other. The correct statement would be that a change of one unit in X_1 is **associated** with a 14 unit increase in Y, **assuming X_2 is held fixed**. Remember that you do not even know whether X_2 can be held fixed when X_1 is increased!

(b) **False:** The sign of the coefficient has nothing to do with the strength of the relationship between X and Y. It simply tells you the direction of the relationship. If the coefficient is positive, increases in X are associated with increases in Y. If the coefficient is negative, increases in X are associated with decreases in Y. Even the size of the coefficient does not really tell you the strength of the relationship. Suppose X_1 is measured in inches. If I change the units to feet, I will multiply the coefficient by 12 but nothing will have changed. Beware of comparing magnitudes of coefficients!

(c) **True (maybe):** Suppose X_1 is held fixed, say at 0. Then if X_2 is large enough (say 3 or greater), \hat{Y} will be negative. This is not necessarily bad—Y may be a variable that takes on negative values. Note however that this does assume that X_1 can take on the value 0 when X_2 is 3 which need not be possible. Technically you would need to be sure it was realistically possible to get a pair of X_1 and X_2 values that would make Y negative.

(3) Condo Prices:

(a) From the STATA and SAS printouts below the estimated regression equation is

$$\hat{Y} = 21.987 + 3.2053X_1 + 2.866X_2$$

IN STATA:

```
. regress price area pool
```

Source	SS	df	MS	Number of obs =	20
Model	695.594133	2	347.797066	F(2, 17) =	43.44
Residual	136.124019	17	8.00729521	Prob > F =	0.0000
				R-squared =	0.8363
				Adj R-squared =	0.8171
Total	831.718151	19	43.7746395	Root MSE =	2.8297

price	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
area	3.205273	.3698423	8.67	0.000	2.424974	3.985572
pool	2.865825	1.28352	2.23	0.039	.1578343	5.573816
_cons	21.98716	4.301662	5.11	0.000	12.91145	31.06287

IN SAS:

Note: I used proc reg with the clb command to get the CIs for B0 and B1

```
proc reg data=tmp1.hw4;
model price = area pool/clb;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: price

Number of Observations Read	27
Number of Observations Used	20
Number of Observations with Missing Values	7

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	695.59399	347.79699	43.44	<.0001
Error	17	136.12401	8.00729		
Corrected Total	19	831.71800			

Root MSE	2.82972	R-Square	0.8363
Dependent Mean	60.49000	Adj R-Sq	0.8171
Coeff Var	4.67799		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	95% Confidence Limits	
Intercept	1	21.98716	4.30166	5.11	<.0001	12.91145	31.06288

area	1	3.20527	0.36984	8.67	<.0001	2.42497	3.98557
pool	1	2.86582	1.28352	2.23	0.0393	0.15783	5.57381

(b) If the condo is 1200 square feet, then $X_1 = 12$ since area is given in hundreds of square feet. The predicted value when $X_1 = 12$ and $X_2 = 1$ (i.e. the condominium has a pool) is obtained by plugging in to the estimated regression equation. We get

$$\hat{Y} = 21.987 + 3.2053(12) + 2.866(1) = 63.316$$

In other words, a condo which is 1200 square feet and has a pool costs \$63,316 on average.

(c) All other things being equal—that is if two condominiums are the same size—then if one has a pool and the other doesn't, the one with a pool is on average worth \$2,866 more than the one without a pool. This is because $\beta_2 = 2.866$ and Y is in units of thousands of dollars.

(d) A 95% confidence interval for β_2 has the form

$$b_2 \pm t_{.025, n-k-1} s_{b_2}$$

There are $n=20$ condos in this data set, and $k=2$ predictor variables, so $n-k-1 = 17$. From the t-table, $t_{.025, 17} = 2.110$. From the STATA or SAS printout, $b_2 = 2.866$ and $s_{b_2} = 1.28352$. The resulting confidence interval is

$$2.866 \pm (2.110)(1.28352) \quad \text{or} \quad [.158, 5.574]$$

Of course we can read the same answer directly off the STATA printout—you should know how to get it both ways! This means we are 95% certain that β_2 is between .158 and 5.574. In other words, all other things being equal, a condo with a pool is on average worth anywhere between \$158 and \$5574 more than a condo of the same size without a pool. Since the confidence interval does not include 0, we are 95% sure that having a pool increases the value of a condo. Note that I keep using the phrase “all other things being equal.” The interpretation of β_2 has to take into account that we are also using the size of the condo to predict price. If we had only the variable “pool” in the model we would get a different answer! After all, big houses are generally more likely to have pools so there is some relationship between the two X's.

(e) I would expect that having a pool would increase the value of a condo, all other things being equal. Since having a pool is coded as $X_2 = 1$, this would be equivalent to having a positive value of β_2 . Therefore, the appropriate test is

$H_0 : \beta_2 \leq 0$ —having a pool is associated with lower prices or does not affect the price when the size of the condo has been taken into account.

$H_A : \beta_2 > 0$ —having a pool is associated a higher value for a condo even when the size of the condo has been taken into account

Again note that my hypotheses emphasize that this analysis is done when the size of the condo is being used as a predictor! We use a t-test to check our prediction. From STATA/SAS our test statistic is $t_{obs} = 2.23$. The corresponding p-value is $.039/2 = .0195$. Note that we have to divide the p-value from the printouts by two since this is a one-sided test and STATA and SAS automatically give you the p-value for 2-sided tests in regression. Since this p-value is very small, we reject the null hypothesis and conclude that, all other things being equal, having a pool is associated with a higher condo price as expected.

(4) Salary Prediction:

(a) The STATA and SAS printouts of the correlations are given below. We see that the correlation between salary and years with the company is .984, the correlation between salary and age is .895, and the correlation between years and age is .887. Since the correlation between salary and age is very high (close to 1) there is a strong relationship between salary and age and we expect age to be a good predictor of salary.

IN STATA:

```
. cor salary years age
(obs=20)
```

```
-----+-----
      | salary   years   age
-----+-----
salary |  1.0000
years  |  0.9842   1.0000
age    |  0.8950   0.8875   1.0000
```

IN SAS:

```
proc corr data = tmp1.hw4;
var salary age years;
run;
```

Pearson Correlation Coefficients, N = 20
Prob > |r| under H0: Rho=0

	salary	age	years
salary	1.00000	0.89496 <.0001	0.98416 <.0001
age	0.89496 <.0001	1.00000	0.88749 <.0001
years	0.98416 <.0001	0.88749 <.0001	1.00000

(b) The STATA and SAS multiple regression printouts are shown below. We want to test whether the single variable age is useful. Therefore we perform a t test. Our hypotheses are:

$H_0 : \beta_2 = 0$ —age explains no additional variability in salary beyond what is explained by years with the company and gender. Equivalently we could say that age tells us nothing about salary that we didn't already know by knowing the person's years with the company and gender, or that age is not worth adding to the model if years with the company and gender are already included.

$H_A : \beta_2 \neq 0$ —age does explain additional variability in salary beyond what is explained by years with the company and gender. In other words, it is useful to add age to the model even when we have already taken into account years with the company and gender.

The test statistic is $t_{obs} = 1.36$ and the corresponding p-value is .194. Since the p-value is greater than $\alpha = .05$, we fail to reject the null hypothesis. We do not have enough evidence to conclude that age has a significant linear relationship with salary when years with the company and gender are in the model. In other words, age may not tell us anything about salary that we didn't already know from the person's experience

and gender.

In part (a) we concluded that age ought to be a useful predictor of salary. Our hypothesis test seems to contradict that conclusion. This is the result of the relationship between age and experience, a phenomenon called **multicollinearity**. Note that the correct interpretation of the test is that age is not useful WHEN YEARS AND GENDER ARE TAKEN INTO ACCOUNT. This is not the same thing as saying that age BY ITSELF wouldn't be useful. From part (a) we know that age by itself would be useful. From part (a) we can also see that age and years with the company are highly correlated which is the definition of multicollinearity. This is not really that surprising. The longer you have been with the company, the older you will be. People don't really get paid more for being old—they get paid more for having more experience—so age doesn't have anything more to tell us once we know the employee's tenure with the company. The solution to a multicollinearity problem is usually to remove one of the two correlated variables. Here I am inclined to remove age for three reasons. First, intuitively, it seems more naturally that a person's length of stay with the company or experience would determine the pay rather than their age. Secondly, the correlation between salary and years is higher than the correlation between salary and age suggesting that years is individually the better predictor. Finally, if we do the t test for years it is significant EVEN when age is in the model suggesting years is a useful predictor no matter what. However as we will see later that last conclusion is a little shake since the correlation between the X variables tends to make their p-values a little unstable.

IN STATA:

```
. regress salary years age gende
```

Source	SS	df	MS			
Model	11451.1065	3	3817.03549	Number of obs =	20	
Residual	221.44354	16	13.8402212	F(3, 16) =	275.79	
Total	11672.55	19	614.344737	Prob > F =	0.0000	
				R-squared =	0.9810	
				Adj R-squared =	0.9775	
				Root MSE =	3.7202	

salary	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
years	4.266129	.3564989	11.97	0.000	3.510386	5.021873
age	.3242912	.2390162	1.36	0.194	-.1824006	.8309829
gender	-4.9	1.663744	-2.95	0.010	-8.42698	-1.37302
_cons	10.48123	5.882108	1.78	0.094	-1.988279	22.95075

IN SAS:

```
The REG Procedure
      Model: MODEL1
      Dependent Variable: salary
```

```
Number of Observations Read      27
Number of Observations Used      20
Number of Observations with Missing Values  7
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	11451	3817.03549	275.79	<.0001
Error	16	221.44354	13.84022		
Corrected Total	19	11673			

Root MSE	3.72024	R-Square	0.9810
Dependent Mean	51.35000	Adj R-Sq	0.9775
Coeff Var	7.24488		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	10.48123	5.88211	1.78	0.0938
age	1	0.32429	0.23902	1.36	0.1937
years	1	4.26613	0.35650	11.97	<.0001
gender	1	-4.90000	1.66374	-2.95	0.0095

(c) Now we are asked to test whether $\beta_3 = 0$ which is the same as asking whether gender is a useful predictor when age and years are in the model. Our hypotheses are

$H_0 : \beta_3 = 0$ —gender explains no additional variability in salary beyond what is explained by years with the company and age. Another way of saying this is that there is no difference in salary between men and women of equal age and experience.

$H_A : \beta_3 \neq 0$ —gender does explain additional variability in salary beyond what is explained by years with the company and gender. In other words, it is useful to add gender to the model even when we have already taken into account years with the company and age. There is a difference in salary between men and women with the same qualifications.

The test statistic is $t_{obs} = -2.95$ and the corresponding p-value is .010. Since the p-value is less than $\alpha = .05$, we reject the null hypothesis. Gender DOES have a significant linear relationship with salary even when years with the company and age are in the model. In other words, men and women with the same qualifications are paid differing amounts. Note that we have not specifically tested whether women make less than men or vice versa. We have simply checked whether there is a difference. The negative sign of b_3 suggests that women make less than men as we will test in part (d).

(d) Now we are asked to specifically check whether women make less money than men with the same qualifications. Since the alternative hypothesis is the thing we want to prove our alternative will be that women are underpaid. Since women are coded as $X_3 = 1$ this is equivalent to showing that $\beta_3 < 0$. We write the hypotheses as

$H_0 : \beta_3 \geq 0$ —in other words women make the same amount or more as men when age and experience have been taken into account. $H_A : \beta_3 < 0$ —in other words, women make less money than men with the same age and experience.

The test statistic is $t_{obs} = -2.95$ just as before and the corresponding p-value is .005. Note that we have to di-

vide STATA/SAS's p-value by 2 since this is a one-sided test. Since the p-value is less than $\alpha = .05$, we reject the null hypothesis. Gender DOES have a significant linear relationship with salary even when years with the company and age are in the model and in particular, women make less than men with the same qualifications.

(e) Now we are asked to do an overall F test for the regression. The goal is to see whether our three variables taken together are useful for predicting a person's salary. Note that we expect the answer to this question to be yes, both by common sense and because two of our individual predictors have proven significant. Really we should have done this test first! Our hypotheses are

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$: None of years, age, and gender explains a significant amount of the variability in peoples' salaries. The regression of salary on these three variables is not useful.

H_A : At least one of $\beta_1, \beta_2, \beta_3$ is not equal to 0. In other words, at least one of years, age, and gender does help explain the variability in salaries. Overall the regression of salary on years, age, and gender is useful.

From the printouts our test statistic is $F = 275.79$ and our p-value is .000. Since the p-value is much less than $\alpha = .05$ we reject the null hypothesis and conclude that at least one of years, age, and gender is useful for predicting salary. Note however that the F test does not specifically tell us WHICH variable(s) are useful!

(f) From part (e) we know that at least one of our variables is useful. From parts (c) and (d) we know that gender is useful even when both years and age are in the model so it is definitely worth including. From (b) we know that age is NOT useful when years and gender are in the model, but years is useful even when age and gender are in the model. Therefore we want to keep years in the model and remove age. Our final model should consist of the two variables years and gender. This avoids the multicollinearity problem posed by age and leaves us with the two best predictors. Technically we should rerun the regression with just these variables to make sure they are both significant when age is removed and to see how the coefficients are changed.

Problems To Turn In

(5) Parenteral Nutrition Continued:

(a) The STATA/SAS printouts with the confidence intervals are repeated below for your convenience. Note the use of the CLB option after the model command to obtain these from SAS. The intercept is the average value of Y when $X = 0$. In this case, that corresponds to the average urinary calcium per 12 hours when the person takes in NO calcium in their diet. The confidence interval for β_0 is [.84, 48.66] which means that even if the person is taking in no calcium in their diet there is still between .84 to 48.66 mg going out in their urine per 12 hour period. This interval is entirely above 0 meaning there is some calcium loss even if there is no calcium intake. This fits with the original problem statement which indicated that the body takes calcium out of the bones to try to compensate for the low intake on a parenteral diet and presumably some of that would show up in the urine. The slope gives the average change in Y associated with a 1 unit change in X . Here it indicates how much additional calcium comes out in the urine for every additional mg of dietary calcium per 12 hours. The interval is [.19, .39] which means we are 95% sure that for each additional mg of dietary calcium the subject takes in an additional .19 to .30 mg of calcium will come out per 12 hours. In other words, 19-39% of what is taken in comes out.

IN STATA:

```
. regress urinarycalcium dietarycalcium
```

Source	SS	df	MS	Number of obs =	27
Model	55866.4585	1	55866.4585	F(1, 25) =	33.84
Residual	41273.5415	25	1650.94166	Prob > F =	0.0000
				R-squared =	0.5751
				Adj R-squared =	0.5581
Total	97140	26	3736.15385	Root MSE =	40.632

urinarycal~m	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
dietarycal~m	.2884765	.0495908	5.82	0.000	.1863424 .3906106
_cons	24.74954	11.60949	2.13	0.043	.8393576 48.65973

IN SAS:

```
proc reg data = work.hw4;
model urinarycalcium = dietarycalcium/clb;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: urinarycalcium urinarycalcium

Number of Observations Read 29
Number of Observations Used 27
Number of Observations with Missing Values 2

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	55866	55866	33.84	<.0001
Error	25	41274	1650.94166		
Corrected Total	26	97140			

Root MSE 40.63178 R-Square 0.5751
Dependent Mean 74.66667 Adj R-Sq 0.5581
Coeff Var 54.41756

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	24.74954	11.60949	2.13	0.0430
dietarycalcium	dietarycalcium	1	0.28848	0.04959	5.82	<.0001

Parameter Estimates

Variable	Label	DF	95% Confidence Limits	
Intercept	Intercept	1	0.83936	48.65973
dietarycalcium	dietarycalcium	1	0.18634	0.39061

(b) The dietary calcium measures and urinary calcium measures are both in mg per 12 hours. A person should not have more calcium come out of the body than goes in over a reasonable period of time so in that sense it makes sense for the coefficient to be less than 1. This is not absolutely precise as it depends on when the person has eaten and when they urinate and so forth, and we do know from the problem descriptions that some extra calcium can be lost from the bones as people on parenteral nutrition try to compensate for the low calcium intake so it would be perhaps possible for the value to be slightly above 1. However one would also imagine that some of the calcium is being used up by the body and in fact the coefficient is significantly below 1 (the whole confidence interval for β_1 lies below 1.)

(c) The STATA/SAS printouts for the calculation are shown below. Note the use of the `clm` and `cli` options after the model statement in SAS. These produce CIs and PIs for all the values in the data set. I included X values but no Y values for the CI and PI values needed for parts (c)-(e) and have starred the corresponding lines in the output. I have not repeated the main regression printout to save space. The resulting interval is [66.11,98.78] which means we are 95% sure that parenteral nutrition patients who are getting 200 mg of calcium per 12 hours put out an AVERAGE of between 66.11 to 98.78 mg of calcium per 12 hour period in their urine. Individual people may put out more or less than this but we are 95% sure that this range includes the average of all such subjects.

IN STATA:

```
. adjust dietarycalcium = 200, se ci
```

```
-----
Dependent variable: urinarycalcium      Command: regress
Covariate set to value: dietarycalcium = 200
-----
```

```
-----
All |          xb          stdp          lb          ub
-----+-----
    |  82.4448  (7.93309)  [66.1064  98.7833]
-----
```

```
Key:  xb          = Linear Prediction
      stdp        = Standard Error
      [lb , ub]   = [95% Confidence Interval]
```

IN SAS:

```
proc reg data = work.hw4;
model urinarycalcium = dietarycalcium/clm cli;
run;
```

The REG Procedure
 Model: MODEL1
 Dependent Variable: urinarycalcium urinarycalcium

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean	95% CL Predict	Residual
1	220.0000	184.5655	20.4466	142.4550 226.6761	90.8847 278.2463	35.4345
2	182.0000	112.1579	10.1333	91.2880 133.0278	25.9121 198.4038	69.8421
3	166.0000	107.5423	9.6481	87.6717 127.4129	21.5328 193.5518	58.4577
4	162.0000	174.4689	18.8545	135.6372 213.3005	82.2154 266.7223	-12.4689
5	137.0000	96.5802	8.6797	78.7041 114.4563	11.0094 182.1509	40.4198
6	136.0000	65.7132	7.9696	49.2995 82.1270	-19.5640 150.9905	70.2868
7	128.0000	77.8292	7.8385	61.6856 93.9728	-7.3964 163.0549	50.1708
8	113.0000	68.0210	7.9026	51.7453 84.2967	-17.2298 153.2718	44.9790
9	100.0000	135.2361	13.0215	108.4177 162.0544	47.3610 223.1111	-35.2361
10	90.0000	137.5439	13.3409	110.0678 165.0199	49.4659 225.6219	-47.5439
11	75.0000	50.7124	8.8376	32.5111 68.9138	-34.9268 136.3517	24.2876
12	71.0000	24.7495	11.6095	0.8394 48.6597	-62.2820 111.7811	46.2505
13	60.0000	105.2345	9.4212	85.8312 124.6378	19.3317 191.1373	-45.2345
14	60.0000	96.5802	8.6797	78.7041 114.4563	11.0094 182.1509	-36.5802
15	43.0000	84.7527	8.0095	68.2568 101.2485	-0.5404 170.0458	-41.7527
16	42.0000	77.2523	7.8322	61.1215 93.3830	-7.9710 162.4755	-35.2523
17	37.0000	24.7495	11.6095	0.8394 48.6597	-62.2820 111.7811	12.2505
18	29.0000	60.8091	8.1744	43.9736 77.6446	-24.5503 146.1685	-31.8091
19	24.0000	24.7495	11.6095	0.8394 48.6597	-62.2820 111.7811	-0.7495
20	22.0000	73.7906	7.8210	57.6828 89.8983	-11.4283 159.0094	-51.7906
21	15.0000	24.7495	11.6095	0.8394 48.6597	-62.2820 111.7811	-9.7495
22	3.0000	24.7495	11.6095	0.8394 48.6597	-62.2820 111.7811	-21.7495
23	1.0000	48.4046	9.0293	29.8085 67.0007	-37.3194 134.1287	-47.4046
24	36.0000	24.7495	11.6095	0.8394 48.6597	-62.2820 111.7811	11.2505
25	31.0000	60.8091	8.1744	43.9736 77.6446	-24.5503 146.1685	-29.8091
26	21.0000	24.7495	11.6095	0.8394 48.6597	-62.2820 111.7811	-3.7495
27	12.0000	24.7495	11.6095	0.8394 48.6597	-62.2820 111.7811	-12.7495
part c/d 28	.	82.4448	7.9331	66.1064 98.7833	-2.8179 167.7076	.
part e 29	.	313.2261	41.7486	227.2433 399.2089	193.2434 433.2087	.

Sum of Residuals	0
Sum of Squared Residuals	41274
Predicted Residual SS (PRESS)	47131

(d) To get the urinary calcium output for the individual patient, Johnny Skims, we need a prediction interval. The STATA printout is below and the SAS printout is the same one we used in part (c). The corresponding interval is [-2.82, 167.71]. The negative value of course doesn't make sense—you can't have a negative amount of calcium in your urine. The best way to interpret this is that Johnny will have somewhere between 0-167.71 mg in his urine per 12 hour period. Note that this interval is much wider than the CI for the

average 200 mg intake patient as it should be. This is because the urinary calcium output for an individual with a given dietary intake can vary about the average output. So for instance, we need to include possible values about 98.78 in our interval because if that is the true average (as is possible from part (c)) then there would be individuals about that average value. A similar thing can be said at the low end of the interval.

IN STATA:

```
. adjust dietarycalcium = 200, stdf ci
```

```
-----
Dependent variable: urinarycalc~m      Command: regress
Covariate set to value: dietarycalc~m = 200
-----
```

```
-----
All |          xb          stdf          lb          ub
-----+-----
    |      82.4448      (41.399)      [-2.81795      167.708]
-----
```

```
Key:  xb          = Linear Prediction
      stdf         = Standard Error (forecast)
      [lb , ub]   = [95% Prediction Interval]
```

IN SAS: See printout for part (c)

(e) Once again we need a prediction interval, this time for an individual who is getting 1000 mg of dietary calcium. The printouts are shown below. We see that Johnny Cream's urinary calcium should be between 193.24 to 433.21 mg per 12 hours. This interval is even wider (nearly 240 mg) than the one for Johnny Skims. The reason is that the average value of X (dietary calcium) is about 173 (I obtained this using the summarize command in STATA, also shown below). Thus the PI in part (d) was at a value right near the center of our data where predictions will be more accurate. As you extrapolate further and further from the center of your data, there is more and more uncertainty about the predictions and hence wider intervals. Johnny Cream's value of 1000 is well beyond the maximum of 554 mg/12 hours in our data set and is a very risky extrapolation.

IN STATA:

```
. adjust dietarycalcium = 1000, stdf ci
```

```
-----
Dependent variable: urinarycalc~m      Command: regress
Covariate set to value: dietarycalc~m = 1000
-----
```

```
-----
All |          xb          stdf          lb          ub
-----+-----
    |      313.226      (58.2571)      [193.243      433.209]
-----
```

```
Key:  xb          = Linear Prediction
      stdf         = Standard Error (forecast)
```

[lb , ub] = [95% Prediction Interval]

. summarize dietarycalcium

Variable	Obs	Mean	Std. Dev.	Min	Max
-----+-----					
dietarycal~m	27	173.037	160.6861	0	554

IN SAS: See printout for part (c)

(f) The multiple regression printouts are shown below. Our hypotheses for the overall F test are

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ —none of dietary calcium, urinary sodium, gfr or protein intake helps to explain urinary calcium levels. The model as a whole is not useful for predicting urinary calcium.

$H_A : \text{At least one } \beta_j \neq 0$ —at least one of the four predictors is related to urinary calcium levels. The model as a whole is useful for explaining the variation in urinary calcium output.

Our test statistic is $F_{obs} = 15.27$. Under the null hypothesis, since we have $p = 4$ variables and $n = 27$ data points, F_{obs} has $p=4$ and $n-p-1=22$ degrees of freedom for its numerator and denominator respectively. According to the printouts the corresponding p-value is $P(F_{4,22} \geq 15.27) = 0.0000$ which is smaller than any reasonable significance level α . We therefore reject the null hypothesis and conclude that at least one of our lab measurements is useful for explaining urinary calcium output. However this test does not tell us WHICH of the variables are useful.

IN STATA:

. regress urinarycalcium dietarycalcium gfr urinarysodium protein

Source	SS	df	MS	Number of obs =	27
-----+-----				F(4, 22) =	15.27
Model	71422.1666	4	17855.5416	Prob > F =	0.0000
Residual	25717.8334	22	1168.99243	R-squared =	0.7352
-----+-----				Adj R-squared =	0.6871
Total	97140	26	3736.15385	Root MSE =	34.191

-----+-----						
urinarycal~m	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
-----+-----						
dietarycal~m	.3450265	.0889042	3.88	0.001	.1606506	.5294025
gfr	.426536	.4871694	0.88	0.391	-.5837915	1.436863
urinarysod~m	.4980064	.2212951	2.25	0.035	.0390685	.9569442
protein	-.5112718	.483096	-1.06	0.301	-1.513152	.490608
_cons	-5.731599	21.8887	-0.26	0.796	-51.12597	39.66278
-----+-----						

IN SAS:

The SAS System

The REG Procedure

Model: MODEL1

Dependent Variable: urinarycalcium urinarycalcium

Number of Observations Read	29
Number of Observations Used	27
Number of Observations with Missing Values	2

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	71422	17856	15.27	<.0001
Error	22	25718	1168.99243		
Corrected Total	26	97140			

Root MSE	34.19053	R-Square	0.7352
Dependent Mean	74.66667	Adj R-Sq	0.6871
Coeff Var	45.79089		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	-5.73160	21.88870	-0.26	0.7959
dietarycalcium	dietarycalcium	1	0.34503	0.08890	3.88	0.0008
urinarysodium	urinarysodium	1	0.49801	0.22130	2.25	0.0348
gfr	gfr	1	0.42654	0.48717	0.88	0.3907
protein	protein	1	-0.51127	0.48310	-1.06	0.3014

Parameter Estimates

Variable	Label	DF	95% Confidence Limits	
Intercept	Intercept	1	-51.12597	39.66278
dietarycalcium	dietarycalcium	1	0.16065	0.52940
urinarysodium	urinarysodium	1	0.03907	0.95694
gfr	gfr	1	-0.58379	1.43686
protein	protein	1	-1.51315	0.49061

(g) The confidence interval for the dietary calcium variable is [.16, .53] which means that assuming urinary sodium, gfr and protein intake are held constant, the urinary calcium output goes up between .16 mg/12 hr to .53 mg/12 hr for every additional mg of dietary calcium. Another way of saying this is that if we had two patients with the same urinary sodium, gfr and protein levels but one was getting 1 mg/12 hours more of dietary calcium then that patient would have a higher urinary calcium level by between .16 to .53 mg/12 hr. This interval is NOT the same as the one in part (a). This is because it tells you the impact of dietary calcium AFTER accounting for the other variables. Apparently adjusting for the levels of these other factors makes a difference which suggests one or more of them are related to dietary calcium. This is not surprising. For instance if you were getting your calcium through milk then you would be getting protein along with the calcium and maybe some sodium as well. I don't know exactly what the IV formula is but

a subject who is getting more nutrition of one sort through it is probably getting more of other things as well.

(h) To check whether an individual variable is a useful part of the model we need to perform an individual t-test. Our hypotheses are

$H_0 : \beta_4 = 0$ —dietary protein provides no information about urinary calcium output beyond what is explained by dietary calcium, urinary sodium and gfr. The variable is not useful once the other three variables have been included in the model.

$H_A : \beta_4 \neq 0$ —dietary protein explains additional variability in urinary calcium beyond what is explained by the other three variables and therefore is worth adding to the model.

Our test statistic is $t_{obs} = -1.06$ and the corresponding p-value is .301 which is much larger than our significance level of $\alpha = .05$. We fail to reject the null hypothesis. There is insufficient evidence to show that dietary protein provides any information about urinary calcium output beyond what is explained by the other three factors. It's inclusion does not improve the model. Among the other predictors, dietary calcium and urinary sodium have p-values of .001 and .035 and therefore do contribute unique explanatory power to the model. However the p-value for gfr is .391 meaning it does not add any explanatory power to the model.

(i) In a simple linear regression you are determining whether a particular variable BY ITSELF is related to our outcome variable. In a multiple regression you are asking if the variable provides additional information about Y BEYOND what is provided by the other variables. If some of the predictors SHARE information about Y then it is possible for them both to be related to Y but NOT both be useful as part of the same multiple regression model. That seems to be what is happening here. Protein and GFR are individually related to urinary calcium but all of the information they provide overlaps with that provided by dietary calcium and urinary sodium so they do not need to be added to the model. Actually we need to be a little careful about this. It is possible that gfr and protein are related to EACH OTHER rather than to dietary calcium or urinary sodium and that is why they are not significant but if we took one out the other would be useful. We can check this by seeing if the X variables are in fact related to each other—that is have shared information—by computing their correlations. The printouts are shown below. We see that protein is highly correlated with dietary calcium and gfr is highly correlated with urinary sodium which explains the observed results. We don't need protein if we know dietary calcium (e.g. milk and cheese consumption—though I guess these patients aren't really getting cheese!) and we don't need gfr if we know urinary sodium.

IN STATA:

```
. corr dietarycalcium urinarysodium gfr protein
(obs=27)
```

	dietarycalcium	urinarysodium	gfr	protein
dietarycalcium	1.0000			
urinarysodium	0.1640	1.0000		
gfr	0.1615	0.6155	1.0000	
protein	0.8823	0.1827	0.2122	1.0000

IN SAS:

```

The SAS System
The REG Procedure
Model: MODEL1
Dependent Variable: urinarycalcium urinarycalcium

```

Number of Observations Read	29
Number of Observations Used	27
Number of Observations with Missing Values	2

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	71422	17856	15.27	<.0001
Error	22	25718	1168.99243		
Corrected Total	26	97140			

Root MSE	34.19053	R-Square	0.7352
Dependent Mean	74.66667	Adj R-Sq	0.6871
Coeff Var	45.79089		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	-5.73160	21.88870	-0.26	0.7959
dietarycalcium	dietarycalcium	1	0.34503	0.08890	3.88	0.0008
urinarysodium	urinarysodium	1	0.49801	0.22130	2.25	0.0348
gfr	gfr	1	0.42654	0.48717	0.88	0.3907
protein	protein	1	-0.51127	0.48310	-1.06	0.3014

Parameter Estimates

Variable	Label	DF	95% Confidence Limits	
Intercept	Intercept	1	-51.12597	39.66278
dietarycalcium	dietarycalcium	1	0.16065	0.52940
urinarysodium	urinarysodium	1	0.03907	0.95694
gfr	gfr	1	-0.58379	1.43686
protein	protein	1	-1.51315	0.49061

(j) We see that the R^2 , R^2_{adj} and RMSE values for the multiple regression are 73.5%, 68.7% and 34.19 mg/12 hours respectively for the multiple regression and 57.5%, 55.8% and 40.63 mg/12 hours for the simple linear regression model. That the two R^2 values are higher means the multiple regression explains a higher proportion of the variability in urinary calcium output than the simple regression—by nearly 15% which seems quite a big improvement. Similarly, the RMSE for the multiple regression is over 5 mg/12 hours smaller, a reduction in our prediction error of over 10%, again a substantial improvement. It does seem based on these numbers that the multiple regression is a big improvement. Another way to see this is to note that the SLR just contained dietary calcium as a predictor. In the multiple regression dietary calcium is significant but so is urinary calcium, meaning that that variable added significant additional explanatory power. However we

have to be a little careful because the interpretations in the multiple regression also include the presence of *gfr* and *protein* which are not significant but may still be poverlapping a bit with the other variables.

(6) College Tuition:

(a) The intercept, b_0 , gives the average tuition value, Y , for a public school ($X_3 = 0$) with no students ($X_2 = 0$) and a score of 0 ($X_1 = 0$) on the rating scale. Note that this does not really make sense since you would not have school with no students. According to the printout, such a school would in fact a charge tuition of -\$2,410 which equally doesn't make sense!

The coefficient of rating, b_1 , tells you how much increase (or decrease) in tuition is associated with a one unit increase in rating, ASSUMING size and type of school are held FIXED. A one point increase in rating is associated with an extra \$96.70 in tuition.

The coefficient of type of school, b_3 has a different interpretation because type of school is an indicator variable. It makes no sense to talk about a "one unit increase in type of school." A school is either public or private. What b_3 gives you is the DIFFERENCE in tuition between a public and private school with the SAME size and rating. The private school costs \$16,900 more, all else equal.

(b) We check the overall usefulness of the model using an **F test**. The null and alternative hypotheses are

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ (i.e. none of the variables rating, size and type is useful for explaining the variability in schools' tuitions.)

H_A : At least one of the β 's is not 0 (i.e. overall the model is useful—at least one of the predictors contains information about tuition costs)

Our test statistic is $F = MSR/MSE=1686.01$ which seems very large. In this problem, $\alpha = .05$, we have data on $n=20$ schools, and there are $k = 3$ predictor variables in the model. The p-value for this test, from the printout, is .000 which is much less than $\alpha = .05$. This reinforces our decision to reject the null hypothesis. We conclude that at least one of rating, size and type is useful in explaining the variability in tuition costs. This is hardly a surprise. Note that you did not need to write out all the details—just indicate which test you were using, give the p-value and your conclusions. I wrote out the details for completeness.

(c) The two numbers that give the percentage of variability explained by a regression model are R^2 and R^2_{adj} . These values are 99.7% and 99.6% respectively. In the multiple regression setting it is more appropriate to use R^2_{adj} . Some people said this was because R^2_{adj} takes into account the degrees of freedom. This is true, and an important part of the answer but it is not the whole picture. R^2_{adj} also takes degrees of freedom into account in simple linear regression—it's just that with only 1 predictor unless the sample size is small the adjustment will not make very much difference. The more important point is that R^2_{adj} helps prevent you from **overfitting** by penalizing you for using extra predictors that provide no real information. R^2 will always go up as you add new predictors, whether they have anything to do with Y or not. This may make your model look much better than it really is. However, R^2_{adj} can actually decrease if the predictors you are adding are not worthwhile. This is because if the number of predictors is large relative to the number of data points the adjustment factor to R^2 can be quite large and unless the SSE has correspondingly gone down a lot when the extra predictors are added the adjustment will penalize you.

(d) To decide whether a model makes good predictions we need to look at the root mean squared error (RMSE), which is the average distance from points to the regression surface of more intuitively the typical error we make when using the X variables to predict Y , and compare this to the Y values we are trying to predict. Here $RMSE = .5869$ which means on average we are off by \$586.90 in our predictions of tuition based on size, rating and type. We don't know the average Y values in this data set. However as noted

above the model indicates that private schools are about \$17,000 more expensive than public schools. Even if public schools were still pretty cheap at this time we must have been talking about at least \$3000 for a public school and \$20000 for a private school. Based on these numbers our error might be quite big for a public school but it is fairly small (on the order of 2-3%) for a private school. We accepted any reasonable statement you made about what you thought typical tuition was—this is largely about common sense. In today’s terms this error seems even smaller.

(e) A 95% confidence interval for β_2 has the form

$$b_2 \pm t_{\alpha/2, n-k-1} s_{b_2}$$

Here $\alpha = .05$, $n=20$, and $k=3$. From the t table, $t_{.025, 16} = 2.120$. From the printout, $b_2 = -.01923$ and $s_{b_2} = .01606$. The resulting confidence interval is

$$-.01923 \pm (2.120)(.01606) \text{ or } [-.05327, .01482]$$

Note that you must keep the negative sign on the value of b_2 !! The confidence interval contains the value 0. This means that β_2 MIGHT equal 0. It does not mean $\beta_2 = 0$ for certain—it simply means, in essence, that we cannot reject the null hypothesis that b_2 is 0. Thus we conclude that size of school may not be useful for predicting tuition costs, ASSUMING THAT RATING AND TYPE OF SCHOOL ARE IN THE MODEL. It is also quite possible that by itself size is a useful predictor.

(f) We need to perform a t test. The appropriate null and alternative hypotheses are

$H_0 : \beta_3 \leq 0$ —a private school has the same or lower tuition than an otherwise equivalent (same size/same rating) public school

$H_A : \beta_3 > 0$ —even after taking into account size and rating a private school has higher tuition than a public school.

Note that we were asked to show private schools are more expensive so we need a one-sided test with higher tuition at the private school as the alternative. Since the private schools are coded as 1 this corresponds to $\beta_3 > 0$. People sometimes try to tell me in this problem that a public school has no impact because its X value is 0 but that a private school does have an impact on tuition. This makes no sense. For there to be a difference, there have to be two types of schools. It matters which type you are, public or private. We wrote the model as “you pay this much more for a private school than a public one” but could just as easily have written “you pay this much less for a public school than for a private one.”

The test statistic is $t_{obs} = 50.21$ and the corresponding p-value is $.001/2 = .05$. Note that the printout gives a two-sided p-value for testing whether there is a difference between public and private schools—we have to divide by 2 to get the p-value for our 1-sided test. We reject the null hypothesis (just barely—the convention is actually that you reject as long as the pvalue is less than or equal to .05!) and conclude that all else equal (or after accounting for size and rating) private schools are more expensive than public schools.

(g) The University of Southern North Dakota at Hoople was invented by Peter Shickele, a composer who writes humorous music purporting to be the work of P.D.Q. Bach, an illegitimate son of the famous Baroque composer J.S. Bach. SND at Hoople is the infamous P.D.Q. Bach’s alma mater. We can find an estimate of the tuition at SND by plugging into the estimated regression equation. To plug into the regression equation we need to know X_1 , X_2 and X_3 . We know $X_3 = 1$ since this was said to be a private school. We are also given that $X_1 = 60$ since the school has a rating of 60. Some people try to use a rating of .6. The ratings are NOT percentages. It is true that the maximum rating is 100, but I told you the scores were between 0 and 100, not between 0 and 1! Just because a variable lies between 0 and 100 does not mean you should automatically convert it to a percentage! The most confusion was caused by the value of X_2 . In the problem

statement, I told you that X_2 was measured in thousands of students. This means that if the school has 1000 students, $X_2 = 1$.

$$\hat{Y} = -2.41 + .0967(60) - .0192(1) + 16.9(1) = 20.236$$

Thus the tuition is \$20,236.

Since we are talking about an **individual** school we want a **prediction** interval for the tuition. (If we were interested in the tuition at the **average** private school with a ranking of 60 and 1000 students we would use a confidence interval.) From the printout we are 95% sure that the tuition at SND at Hoople will be between \$18,728 and \$21,744.

(h) Optional Bonus: The difference in uncertainty between predicting the average value of Y at a given set of X's X and a particular value of Y at that set of X's corresponds to the variation of the individual points about the regression line/surface which represents the average Y at a set of X's. Specifically $s_{\hat{Y}_0}^2 = s_{\hat{Y}_0}^2 + MSE$ —the addition occurs on the level of the variances. We are given $s_{\hat{Y}_0} = .402$ and $MSE = .34$ from the printout so we have

$$s_{Y_0} = \sqrt{(.402)^2 + .34} = .708$$

(i) According to the printout $b_3 = 16.858$ meaning a private school on average costs \$16,858 more than a public school with the same rating, while $b_1 = .0967$ which means that every extra point of rating is worth about \$96.7. Putting these together, a public school would need to have a rating of $16,858/96.7 = 174.3$ points higher to make up the difference. This is not possible since the rating scale only goes to 100 points and in fact this is a good example correlation not necessarily being causation. Public school tuition is usually somewhat capped by the state regardless of how good the school is—better ones may on average cost more but raising your quality indefinitely is not going to mean your tuition can go up in an unlimited way.

(j) Dr. Dimwit is more than a little dim. First, although b_1 is negative meaning our best estimate is that smaller schools have higher tuition, if we look at the p-value for the t-test we note that the coefficient of the size variable is NOT significantly different from 0 so we can not be certain the relationship is really negative. This is one problem. Secondly, even if decreasing the size of the school would increase the tuition it is NOT clear that this would increase total REVENUE. If the tuition goes up only a little (which the extremely small size of the coefficient suggests) but we reduce the number of students a lot, our total tuition take will go down severely. Plus with so few students we won't be able to offer a lot of classes and so on. This definitely doesn't seem like a good idea. Thirdly, correlation is NOT causation. You could argue that smaller schools generally have higher tuition because they tend to be liberal arts schools which (a) are private and (b) have smaller student to teacher ratios and (c) high ratings from students because of their focus on teaching. Just cutting the number of students admitted without changing other factors will not necessarily make your tuition go up!