Biostatistics 201A, Midterm Practice Problems With Solutions

Inference, Including Estimzation, CIs, Tests and Power

(1) Charity Donations: A charity suspects one of its volunteers of stealing donations. The volunteer is going door to door asking for contributions. Past experience shows that honest volunteers collect on average \$3 per household. However, obviously, the actual amount varies for each household. In the last week the volunteer visited 500 households and the end of the week reported his total donations as \$1149 with a standard deviation of the amount collected from each person of s = 5.

(a) What is your best estimate of the volunteers average donation per household during this period?

Solution: The best average of the mean is the sample mean. Here we have a total of \$1149 in donations from n = 500 houses so we have $\bar{X} = 1149/500 = \$2.298$ per household. This is certainly a fair percentage lower than previous volunteers. The question is whether it is so much lower as to be statistically significant (not likely to have happened by chance.)

(b) Find a 95% confidence interval for this volunteer's average donation based on the data from the past week. Based on this interval does the volunteer appear to be typical? Discuss.

Solution: A confidence interval for a single mean is given by

$$\bar{X} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$$

Here we have n = 500 and s = 5. A t-distribution with 500 degrees of freedom is essentially a Z distribution so we have $t_{.-25,499}1.96$ and the resulting confidence interval is

$$2.298 \pm (1.96)\frac{5}{\sqrt{500}} = [1.860, 2.736]$$

This interval tells us that if we follow this volunteer around for a long time their mean donation per household should be somewhere between \$1.86 and \$2.74. Since the interval lies entirely below the value of \$3.00 per household that is typical for this organization the volunteer does NOT appear to be typical. He is taking in less money epr household.

(c) Conduct the hypothesis test an investigator would use to prove the volunteer is stealing. Be careful to state the null and alternative hypotheses, both mathematically and in English, with a justification and explain your final conclusion. Use $\alpha = .025$. and compare your result to that of part (b).

Solution: Since we specifically want to show the volunteer is stealing we would want to prove that they are giving us LESS money than they should. (There is a presumption here that the volunteer collects a normal amount of money, keeps some for himself, and gives us the rest hoping we won't

notice.) Since what we want to prove is the alternative hypothesis we have

$H_0 : \mu \geq 3$–our volunteer turns in as much per household or more than a typical volunteer for this charity.
$H_A : \mu < 3$–this volunteer is turning in less money epr household than a typical volunteer.

Our test statistic is

$$t_{obs} = \frac{2.298 - 3}{5/\sqrt{500}} = -3.14$$

Since this is a 1-sided test, our p-value is

$$P(t_{499} < -3.14) = P(t_{.499} > 3.14) = .000895$$

I got the exact p-value from STATA. During an exam you could approximate it using a t (or for that matter Z) table as approximately .001. In any case it is smaller than $\alpha = .025$ so we reject the null hypothesis and conclude the volunteer is turning in less money than is typical of our volunteers.

(d) Why did I say to use $\alpha = .025$ in part (c) when you computed a 95% CI in part (b)?

Solution: A confidence interval is a 2-sided quantity. If you have a 95% confidence interval there is a 2.5% that it misses the true mean by being too high and a 2.5% chance it misses the true mean by being too low. Thus if we want to do a 1-sided test where we care only about low values and want it to have a comparable value to a 95% CI we need to use $\alpha = . - 25 = 2.5\%$.

(e) Do the results of parts (b) and (c) prove the volunteer is stealing? Give two possible alternative explanations.

Solution: The results of (c) and (d) suggest the volunteer is giving us less money than our typical volunteers. However this is NOT the same as proof of stealing!! One possibility is that this is simply an unlucky sample and the volunteers's meant ake per household is really higher than this. (This doesn't seem too likely since our p-value was so small but remember that the p-value isn't the probability the null is true, it is just the probability of our data ASSUMING the null is true which is a slightly different thing.) Another possibility is that this volunteer is just incompetent and can't convince people to make donations. Not all volunteers are created equal and average!

(f) What is the effect size for the difference between the volunteer being studied and the agencies typical volunteers? How big would you consider this effect to be?

Solution: The effect size for a one-sample t-test is simply the difference between the estimated mean and the hypothesized mean divided by the standard deviation:

$$d = \frac{\bar{X} - \mu_{H_0}}{s} = \frac{2.298 - 3}{5} = -.14$$

This is a very small effect size by the standards we learned in class. (Note that as a percentage of the average the effect is pretty big but the standard deviation is very large compared to the

mean.) The only reason we have been able to detect this effect so convincingly is that the sample size (number of houses) is extremely large.

(g) What is the minimum detectable effect size for this study for a two-sided test with $\alpha = .05$ and 80% power? (Note–you need to use STATA to do this as stated. On an exam I would have to give you a printout to interpret.) In general do you consider the power here to be high? Discuss.

Solution: I did the calculation on STATA (see results below). The minimum detectable effect size is d = .13 SDs I also got the minimum detectable effect size for a 1-sided test which is d = .11, slightly smaller. Since the n is so large we could actually do all calculations using Z and could compute the power analytically (though I would consider that a hard bonus problem for this class.) The smallest effect size we can detect is very small indeed meaning that our power is excellent. The smaller the effect you can see the stronger the information you have is.

```
Two-sided test:

. sampsi 0 .13, sd(1) n(500) onesample

Estimated power for one-sample comparison of mean
  to hypothesized value

Test Ho: m =      0, where m is the mean in the population

Assumptions:

         alpha =   0.0500  (two-sided)
 alternative m =      .13
            sd =        1
 sample size n =      500

Estimated power:

         power =   0.8282


One-sided test:
. sampsi 0 .11, sd(1) n(500) onesample onesided

Estimated power for one-sample comparison of mean
  to hypothesized value

Test Ho: m =      0, where m is the mean in the population

Assumptions:
```

```
         alpha =   0.0500  (one-sided)
 alternative m =      .11
            sd =        1
 sample size n =      500

Estimated power:

         power =   0.7924
```

(h) Suppose instead of comparing the volunteer from the previous parts of this problem to previous population values we instead had two volunteers, one of whom was the the one above and the second a volunteer who had averaged $\bar{X} = \$3$ per donation with s = 4 for 25 donations. (i) Find a confidence interval for the difference in means between the two volunteers. (ii) Perform an appropriate hypothesis test to see if the amounts they collect are different (iii) Give an estimate of the effect size for the difference in their average collections.

Solution: The first thing we need to do is find the pooled estimate of the standard deviation. This is given by

$$s_{pooled} = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} = \sqrt{\frac{499(5^2)+24(4^2)}{523}} = 4.959$$

Since the bigger part of the sample had a standard deviation of 5 this is not too surprising. The confidence interval is

$$\bar{X}_2 - \bar{X}_1 \pm t_{\alpha/2,n_1+n_2-2} s_{pooled}\sqrt{1/n1_1 + 1/n_2}$$

or

$$3 - 2.298 \pm (1.96)(4.959)\sqrt{1/500 + 1/25} = [-1.29, 2.69]$$

This says that the true difference in mean donation per household between the two volunteers could be anywhere from the first colunteer making \$1.29 per house more to the second volunteer making \$2.69 per house more. The interval includes 0 or equivalently has values that correspond to either of the volunteers making more money so we can't be sure there is a difference between them. This is a result of the relatively large standard deviation and small sample size for the second voluteer. Note that when we were treating this number as a population value it had no variance at all!

The equivalent test has

$H_0 : \mu_1 = \mu_2$–there is no difference in average donation between the two volunteers.
$H_A : \mu_1 \neq \mu_2$–there is a difference in average donation between the two volunteers.

The test statistic

$$t_{obs} = \frac{\bar{X}_2 - \bar{X}_1}{s_{pooled}\sqrt{1/n_1 + 1/n_2}} = \frac{.702}{1.016} = .69$$

Under the null hypothesis this has a t-distribution with 523 degrees of freedom (essentially a Z distribution) and the corresponding p-value is

$$2P(t_{523} \geq .69) = .490$$

Again I got the exact p-value off of STATA. However even just with a t-table it would be very obvious that the p-value is large and we fail to reject the null hypothesis. We do not have enough evidence to say there is a difference in average donation per household between the volunteers.

Finally, we are asked for the effect size corresponding to this comparison which is just the difference in means divided by the pooled estimate of the standard deviation. We have

$$d = \frac{\bar{X}_2 - \bar{X}_1}{s_{pooled}} = \frac{.702}{4.959} = .142$$

This is actually if anything a slightly bigger effect than what we saw abiove (though it is still very small) but the sample size for the second group is so small that it does not appear significant.

Classical ANOVA

(1) Honey I Shrunk the Tumor: Dr. Clever is studying a rare form of cancer. People who have this form of cancer have traditionally received either standard chemotherapy (C) or aggressive chemotherapy (AC). Dr. Clever has himself developed a new medication (I) that is injected directly into the tumor site. Dr. Clever is interested in knowing how much the different treatments shrink people's cancer tumors. He has collected data on tumor size before and after treatment for 13 subjects who have received standard chemotherapy, 25 subjects who have received agressive chemotherapy and 25 subjects who have received his new injectable treatment. For each person he has calculated the percentage reduction in tumor size, Y. For instance, Y = 50 corresponds to a 50% reduction in tumor size. An ANOVA printout for his data is shown below.

```
. oneway PercentShrunk Treatment, tabulate

           |     Summary of PercentShrunk
 Treatment |     Mean   Std. Dev.    Freq.
-----------+------------------------------------
         C |      37       20         13
        AC |      38       20         25
         I |      50       10         25
-----------+------------------------------------
     Total |    42.56    17.55        63
```

```
Number of obs =     63     R-squared     =  0.121
Root MSE      =  16.73     Adj R-squared =  0.091

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups       2305.56        2   1152.78        4.12      0.021
 Within groups      16800.00       60    280.00
------------------------------------------------------------------------
    Total           19105.56       62    308.15
```

(a) Based on this data is there evidence the different treatments shrink the tumors by different average amounts? Justify your answer by performing an appropriate **overall** hypothesis test. State the mathematical hypotheses in the classical ANOVA framework, give the p-value, and your real-world conclusions.

Solution: We need to perform an overall F test. The hypotheses in the classical ANOVA (as opposed to regression) framework are

$H_0 : \mu_C = \mu_{AC} = \mu_I$–the three treatment groups have the same average tumor shrinkage.
H_A : At least two of the $\mu's$ differ–the three groups do not all have the same average tumor shrinkage.

You only had to give the mathematical hypotheses but I included the statement in words for completeness. From the printout the p-value for the overall F test is .021 which is less than our default significance level of $\alpha = .05$ so we reject the null hypothesis and conclude that the amount of tumor shrinkage differs by treatment group.

The test statistics and p-values for pairwise comparisons of the group means in this ANOVA are given below. Use them to answer parts (b)-(d).

Pair	t_{obs}	p-value
C vs AC	.175	0.862
C vs I		
AC vs I	2.536	0.0138

(b) One of the rows is blank. Write down the mathematical null and alternative hypothesis corresponding to the missing test, compute the test statistic and give as close an approximation as you can to the p-value.

Solution: Here we are performing a t-test for whether two specific treatment groups have the same mean tumor shrinkage. The hypotheses are

$H_0 : \mu_C = \mu_I$ or $\mu_I - \mu_C = 0$–the standard chemotherapy and the injection groups have the same average tumor shrinkage.

$H_A : \mu_C \neq \mu_I$ or $\mu_I - \mu_C \neq 0$–the two groups do not have the same average tumor shrinkage.

Note that as in (a) you only had to give the mathematical hypotheses for credit. In an ANOVA the test statistic for a comparison of means is

$$t_{obs} = \frac{\bar{Y}_I - \bar{Y}_C}{\sqrt{MSW(\frac{1}{n_I} + \frac{1}{n_C})}} = \frac{50 - 37}{\sqrt{280(\frac{1}{25} + \frac{1}{13})}} = 2.272$$

We can't get the exact p-value from the t-table. However we can get a good approxximation as follows. There are 60 degrees of freedom associated with MSW and therefore with our t-statistic in this ANOVA. From the t-table we see that t_{obs} is between $t_{60,.025} = 2$ and $t_{60,.01} = 2.39$. Thus the one-sided p-value associated with this test statistic must be between .01 and .025. However, since we want a two-sided p-value we have to double this so the p-value for our test must be between .02 and .05. (Note that this is means we would be able to reject the null hypothesis and conclude that these groups are different using our standard significance level of $\alpha = .05$.)

(c) Based on the ANOVA printout, the table and your answer to part (b), order the three treatments from most to least effective (in terms of percentage shrinkage of the tumor), clearly indicating which differences are significant and which are not at $\alpha = .05$.
Solution: From the ANOVA printout we see that the injection treatment has the greatest average tumor shrinkage (is most effective) at 50%, followed by the agressive chemotherapy group which has an average shrinkgae of 38%. The standard chemotherapy treatment has the smallest average shrinkage at 37%. From part (b) and the table preceeding it we see that the differences between the injection group and each of the other two groups are significant at $\alpha = .05$ (p-value .0138 vs agressive chemotherapy and somewhere between .02 and .05 versus standard chemotherapy). However the difference between the two chemotherapy treatments is not statistically significant (p-value .862). Note that you can still give your BEST ESTIMATE of the ordering of the three categories even if you can't be 95% sure of that ordering based on significant differences.

(d) Suppose you used the Bonferroni correction to adjust for the multiple comparisons in part (c). Say what the new significance level, α^*, for the individual tests would be and indicate how your answer to (c) would change (if at all).

Solution: The Bonferroni method says to divide the desired overall significance level α by the number of tests being performed. Thus our new significance level for the individual pairwise tests is $\alpha^* = .05/3 = .0167$. The difference between the two chemotherapy treatments was not significant at $\alpha = .05$ so it certainly isn't at the new significance level. The difference between the injection and agressive chemotherapy groups is still significant since its p-value of .0138 is less than our new α^*. However the difference between the injection and standard chemotherapy groups is no longer significant since its p-value is at least .02 which is greater than α^*. Thus after adjusting for multiple comparisons we no longer have enough evidence to be sure that the injection group and the standard chemotherapy group have different average tumor shrinkage.

(e) Dr. Clever is interested in proving that his new injection therapy works **better** than either of the chemotherapy treatments.

(i) Write down an appropriate linear combination, LC, for the comparison he wishes to do. (You may assume that the two chemotherapy treatments are equally common in the population of interest.)

Solution: The linear combination is

$$LC = \mu_I - \frac{1}{2}(\mu_C + \mu_{AC})$$

This represents the difference between the injection group and the average of the other two groups– what we would get as the mean if we put all chemotherapy subjects together and assumed that these treatments were equally common. We could reverse the order as well:

$$LC = \frac{1}{2}(\mu_C + \mu_{AC}) - \mu_I$$

All this would do is change the sign.

(ii) Give your best estimate of LC and the corresponding standard error.

Solution: We get our best estimate of the linear combination by substituting the sample group means for the population values in the equation:

$$\hat{LC} = \bar{Y}_I - \frac{1}{2}(\bar{Y}_C + \bar{Y_{AC}}) = 50 - \frac{1}{2}(37 + 38) = 12.5$$

This says that on average tumors of people in the injection group shrink 12.5% MORE than those of people who receive chemotherapy. This sounds good but we need to know how much uncertainty there is in this estimate. Of course if we reversed the order of the linear combination we would get -12.5 meaning that the tumors in the chemotyherapy group shrink 12.5% LESS than those of people in the injection group which is equivalent.

The standard error of the linear combination is given by

$$se(\hat{LC}) = \sqrt{MSW(\sum \frac{c_j^2}{n_j})} = \sqrt{MSW(\frac{1}{n_I} + \frac{(.5)^2}{n_C} + \frac{(.5)^2}{n_{AC}})} = \sqrt{280(\frac{1}{25} + \frac{.25}{13} + \frac{.25}{25})} = 4.40$$

(iii) Use these numbers to compute a **90% confidence interval** for the linear combination and give a brief interpretation of it. Your interpretation should incorporate the numerical values of both ends of the interval and also address the question of interest to Dr. Clever.

Solution: For a 90% confidence interval we need the critical value $t_{60,.05} = 1.671$ from the t-table. The resulting confidence interval is

$$12.5 \pm (1.671)(4.4) \rightarrow [5.14, 19.86]$$

This means that the injection treatment shrinks the tumors between 5.14% and 19.86% more on average than chemotherapy. Since this interval is entirely above 0, Dr. Clever can be 90% (really 95%) sure that his injection therapy is more effective than the average chemotherapy treatment.

(f) Optional Bonus Dr. Clever's graduate student, Denny Dull, suggests that instead of looking at the percentage shrinkage of the tumors Dr. Clever should do an ANOVA with 6 groups giving the mean tumor sizes before and after treatment for each of the 3 treatment groups. (i) Explain what regression assumption would be violated if Dr. Clever performed the analysis this way and (ii) give one other potential problem with this approach.

Solution: If we have a before and after measurement for each subject in the ANOVA then the assumption of independence will be violated. People who start with larger tumors probably still have larger tumors after treatment than people who start with smaller tumors. Another issue is that if the tumors in the different groups did not start out at the same time these raw sizes will be hard to compare. By looking at percent shrinkage we have a number that is comparable across both individuals and groups. This may well be an issue since Dr. Clever didn't randomize people to the different treatment groups and it seems likely that sicker people (i.e. people with larger/more serious tumors) would have been given the more aggressive treatments.

(g) Optional Bonus) Suppose, continuing part (e) that Dr. Clever wants to prove his injection therapy shrinks the cancer tumors by **at least 5% more** than the average of the chemotherapy treatments. Write down the hypotheses for this test mathematically and in words, compute the test statistic, and give an approximate p-value and real-world conclusions using $\alpha = .05$.

Solution: There were two reasonable ways to interpret this problem. Since the units of the outcome measure are percentage tumor shrinkage what I had in mind was the following:

$H_0 : \mu_I - .5(\mu_C + \mu_{AC}) \leq 5$–the injection treatment does not shirnk the tumors by at least 5% more than the chemotherapy treatments.
$H_A : \mu_I - .5(\mu_C + \mu_{AC}) > 5$–the injection treatment does shrink the tumors by at least 5% more on average (what Dr. Clever is trying to prove so it is the alternative hypothesis).

Using these hypotheses we are testing the same linear combination as in in part (f) but rather than testing whether the linear combination is 0 we are testing whether it is greater than 5. The resulting test statistic is

$$t_{obs} = \frac{12.5 - 5}{4.40} = 1.70$$

We can use the t-table to get an approcimate p-value. Looking in the row for 60 degrees of freedom we see that the one-sided p-value, which is what we want, is between .025 and .05, probably closer to .05 since $t_{60,.05} = 1.671$. Thus we are able to reject the null hypothesis and conclude that Dr. Clever's treatment has the desired level of superiority though it is fairly close.

Some people instead of interpreting the 5% improvement in treatment shriankage as an absolute number interpreted it as a fraction of the actual chemotherapy treatment effect. This is perfectly reasonable from the way the problem is stated and results in the following hypotheses:

$H_0 : mu_I \leq 1.05 * (.5(\mu_C + \mu_{AC}))$
$H_A : \mu_I > 1.05 * (.5(\mu_C + \mu_{AC}))$

If you do the problem this way you have to re-estimate the linear combination and its standard error since the constants have changed slightly but the procedure is the same as before.

(2) Fun with Genetic Testing: You are studying whether a group of genes is associated with an elevated risk of breast cancer. For each person you record X, and indicator of whether or not the person got breast cancer and Y_j, the expression level (a continuous measure) associated with gene j for a large number of different genes.

(a) Explain what methods we have learned could be used to tell with expression level of a given gene is associated with breast cancer.

Solution: Since we are comparing the mean of a continuous variable (expression level) for two groups (cancer vs no cancer) we could use either a two-sample t-test or an ANOVA. Actually we could also use a simple linear regression with Y = expression level and X as an indicator of whether you have cancer.

(b) Imagine that you wanted to identify the subset of genes tested that were actually related to cancer. Suppose you were testing 10 genes (so $j = 1, 2, \ldots, 10$), what approaches could you use to ensure the overall accuracy of your answers?

Solution: Here most of the multiple comparison methods we learned would work reasonably. Even Bonferroni would only require us to use a p-value of .005 for the individual tests which is not too bad.

(c) Now imagine that instead of the 10 genes in part (a) you were testing 10,000 genes. Now what would be a reasonable approach to ensuring overalla ccuracy? Discuss.

Solution: In this scenario, Bonferroni is no longer practical. We would need to use $\alpha = .000005$ which is getting to be an extremely small p-value. Instead we would probably use either the false discovery rate procedure which limits the number of false positive results to a certain fraction rather than insisting we have a high probability of having NO false positives. Alternatively we could use the omninibus approach were we do each test at the original significance level but then count how many false positives we would expect to have with this many tests (here .05*10000 = 500). If we had more significant tests than that we could be fairly sure that at least some of our results were correct.

Simple Linear Regression

(1) Mostly Mozart: Dr. Smart believes that the mother's drinking during pregancy will have a long term negative effect on child's mental development while listening to classical music will have a positive effect. She has therefore conducted a study in which she followed 102 pregnant woman and recorded both X_1, the average number of drinks they had each day during the pregnancy and X_2, the number of minutes they listened to classical music each day. She then went back when the children were 7 years old and recorded their scores, Y, on a standard IQ test. (For reference, an average IQ score is around 100 while scores below 70 correspond to mental retardation and scores around 160 are thought to represent genius.) Dr. Smart planned to perform two simple linear

regressions, one of IQ on mother's alcohol consumption and one of IQ on mother's music listening, along with corresponding correlation and covariance calculations. The STATA printouts for the music analysis are given below. However those for the alcohol analysis seem to have been lost and all that is available are some summary statistics. Use this information to answer the questions on the following pages.

Overall: $n = 102$, $\bar{Y} = 95$

For the Alcohol Analysis: $\bar{X}_1 = 1$, $SCP = -2000$, $SSX = 400$, $SST = SSY = 20000$

For the Music Analysis:

Correlation:

```
. corr IQ Music
(obs = 1-2)
             |       IQ    Music
-------------+------------------
          IQ |   1.0000
       Music |   0.3000   1.0000
```

Covariance

```
.corr IQ Music, c
             |       IQ    Music
-------------+------------------
          IQ |   198.02
       Music |   356.40    7128.7
```

Regression:

```
. reg IQ Music

      Source |       SS       df       MS              Number of obs =     102
-------------+------------------------------           F(  1,   100) =    9.89
       Model |     1800         1     1800             Prob > F      =   0.002
    Residual |    18200       100      182             R-squared     =   0.090
-------------+------------------------------           Adj R-squared =   0.081
       Total |    20000       101   198.02             Root MSE      =   13.49

------------------------------------------------------------------------------
          IQ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       Music |      0.05     0.016      3.145   0.001        0.0185     0.0815
       _cons |     93.50     1.418     65.924   0.000       90.6861    96.3139
------------------------------------------------------------------------------
```

(a) Which is stronger, the relationship between IQ score and maternal alcohol consumption or the relationsip between IQ score and maternal music listening? Explain your reasoning and show any necessary calculations.

Solution: To tell the strength of a relationship we look at the correlation. From the STATA printout the correlation of IQ with maternal music listening is $r = \hat{\rho} = .3$ (marker with a * on the printout above). To get the correlation for of IQ and alcohol consumption we do the following calculation using the numbers given above:

$$Corr(X,Y) = \frac{Cov(X,Y)}{SD(X)SD(Y)} = \frac{SCP}{\sqrt{SSX * SSY}} = \frac{-2000}{\sqrt{400 * 20000}} = -.707$$

The strength of the correlation is determined by how close it is to ± 1. Here the correlation of IQ with alcohol use has the higher absolute magnitude (.7 vs .3) so it is the stronger correlation. Note that the sign has nothing to do with the strength of the relationship. It is better to use the correlation than the covariance here because the covariance has units attached and the two X variables are not comparable. In fact, if you calculate the covariances, the one between IQ and music is MUCH bigger because the listening is measured in minutes of which a person may do many dozen per day while the alcohol consumption is in drinks and shouldn't be more than a few a day (really ANY during a pregnancy!) The fact that alcohol has a stronger relationship is not surprising. It can have a severe chemical effect on the baby's developing brain, going through the placenta, while the music effect is more indirect. Some people said the music correlation was very weak. However in this context, considering the many things that affect mental development, I think it's pretty impressive the correlation with music is even that high.

(b) Find b_0 and b_1, the estimates for the intercept and slope, of the simple linear regression of IQ (Y) on maternal alcohol consumption (X_1) based on these data.

Part b (6 points)

Find b_0 and b_1, the estimates for the intercept and slope, of the simple linear regression of IQ (Y) on maternal alcohol consumption (X_1) based on these data.

Solution: We use our basic hand formulas for the slope and intercept:

$$b_1 = \frac{SCP}{SSX} = \frac{-2000}{400} = -5$$

$$b_0 = \bar{Y} - b_1\bar{X} = 95 - (-5)(1) = 95 + 5 = 100$$

Don't forgetthe double negative sign in calculating the intercept. You weren't asked for it, but for completeness, these values tell us that on average every extra drink the mom has (per day) is associated with a 5 point drop in her child's IQ score. This seems rational–a drink per day during pregancy is quite a bit and alcohol is expected to have negative effects. The intercept tells us that babies whose moms do not drink at all during pregnancy have an average IQ of 100–which is exactly a normal IQ. This also makes sense.

(c) Give the units and real-world interpretations of b_0 and b_1 for the regression of IQ (Y) on maternal music listening (X_2) and say briefly whether they make real-world sense. Your answer should incorporate the actual numerical values of the coefficients.

Solution: From the STATA printout, for the music model we have $b_0 = 93.5$ and $b_1 = .05$. As noted above for the alcohol coefficients, the intercept is the average value of Y when X=0. Here X is music listening, so $b_0 = 93.5$ means that the average IQ of children whose mothers spent no time listening to classical music during their pregnancies is 93.5, a little below the average IQ of 100

but well within the normal range. This seems plausible. Music may heop so maybe children whose mothers didn't listen at all would be a little worse off on average but the effect shouldn't be too big for the reasons discussed above. The slope gives the average change in Y associated with a 1 unit change in X. Here this means every extra minute per day of music listening is associated with an average increase in IQ of .05 points, or to put it on a more interpretable scale, every additional hour is associated with a 60*.05 = 3 point gain. Again this seems reasonable. The music is having the expected positive effect but it's magnitude is not that large. Many people forgot to comment on whether the values made real-world sense. Do make sure you answer all parts of the questions!

(d) Silly Sally, a graduate student at the University of the Statistically Challenged, who is pregnant with her first child, decides on the basis of this study that she will have classical music playing in her house 24 hours a day, 7 days a week. According to the IQ-music regression model what is the predicted IQ for her child? Do you think this prediction is reliable? Explain.

Solution: To make the prediction we simply have to plug the correct value of X_2 into the regression equation. We know that X_2 is measured in **minutes per day**. Since an hour has 60 minutes, if Sally is listening 24-7 for the whole pregnancy she is listening 24*60 = 1440 minutes per day. You need to be careful about the units. You do not multiply this number by 7 since X_2 is not measured in minutes per week. The 7 days a week was just there to tell you she was doing it every day. Plugging in this value we get

$$\hat{Y} = 93.5 + (.05)(1440) = 165.5$$

According to the problem statement an IQ of 165.5 puts you in the genius range. It seems intuitively unlikely that just listening to classical music constantly can make into a genius. Models are only reliable in or near the range of values in the data set used to develop them. It seems highly unlikely that many women in Dr. Smart's study were listening anything like this much since among other things they should have been getting lots of sleep! Therefore we are probably extrapolating too far beyond the range of the data and this prediction is not reliable. (In point of fact, whether the value seemed realistic or not the prediction would not be reliable if the X value were too far outside the range of the data.)

(e) What is Sally assuming when she makes the decision to play non-stop classical music? Is her assumption justified? If so, explain why and if not give an appropriate example (in the context of this problem!) to back up your argument.

Solution: People had more trouble than I expected with this part of the question. In particular, many just repeated the answer to part (d), that Sally was assuming the model would be reliable for all values of X but that realistically the linearity would not continue forever–there would be some limit to how much benefit you could get from the music and after that point the relationship would level off (or even get worse since listening too much could take away from sleep!) This is true but it's not actually the main point (and it's very unlikely I would ask the same exact question twice.) The reason Sally is deciding to listen a lot is she believes the study shows listening to classical music is **good** for her baby's mental development, i.e. that the more she listens the higher her child's IQ will be. This assumes a **causal** relationship–that she can actually **make** the IQ go up by listening more. All the data shows is that there is an **association** between maternal music listening and IQ,

i.e. that **on average** children whose mom's listened a lot had higher IQs than those who didn't. But **correlation is not causation**. It could be that mom's who listened a lot to classical music were more highly educated, of higher socioeconomic status, or otherwise paying more attention to having the best possible prenatal environment for the baby (e.g. not drinking, nutrition, doctor's visits, etc.) and that it was those factors that were driving the higher IQ scores, not the music. Note that regardless of what the shape of the relationship is, this argument about causality applies. We gave a fair amount of partial credit for the extrapolation argument but to get full credit you did need to talk about causality/confounders. There is another point that it is important to make here, which is what is meant by a confounder. As mentioned in Problem 1, to be a confounder a variable here would have to be related to BOTH how much the mom listens to music AND to her child's IQ.

(f) Suppose instead of fitting two simple linear regressions we fit a single model that used both alcohol consumption and music listening to predict IQ. Relative to their values in the simple linear regression, indicate (by circling your choice) whether you would expect SST, SSR and SSE in the new model to be larger, smaller or stay the same. Briefly explain your reasoning.

```
SST          Stay the Same
SSR          Increase
SSE          Decrease
```

Solution: SST represents the total variability in Y, our outcome variable which here is the children's IQ score. Since we are using the exact same data (just using both X variables at the same time), SST will not change. Note that SST has NOTHING to do with the X variables so it is not effected by how many are in the model. Now we have seen from the earlier parts of the problem that both alcohol consumption and music listening provide information about or have a relationship with IQ (albeit the alcohol relationship is stronger). This means that by including them both in the model I should be able to make BETTER predictions than if I use just one of them. This means that SSR, the amount of variabilty explained, should go up. Similarly, SSE is the amount of variability that is not explained by the model. It can be thought of as the variability attributable to factors not included in the model. Since there are now fewer factors unaccounted for, the predictions should be better and SSE lower. Alternatively you can argue that since SST is the same, SSR has gone up and SST=SSR+SSE that SSE must correspondingly have dropped. Some people tried to argue that because the relationship between music and IQ was positive and that between alcohol and IQ was negative the two variables would cancel each other out and so SSR would go down and SSE up. You need to distinguish between the "variability explained" which is what SSR and SSE represent (and which is not affected by the sign of the coefficients) and the Y value which is effected by the signs of the slopes. It is perfectly true that if you have a mother who listens to music and drinks, the positive effects of the music will be cancelled out by the negative effects of the drinks. But this still leads to a more ACCURATE prediction of the child's IQ than just using the music listening and ignoring the drinking (which results in too high a predicted IQ) or than just using the alcohol use and ignoring the music (which leads to too low a prediction).

(2) Fast Stats on Fast Food: Dr. Nutts has selected n=62 children from urban neighborhoods in the city of Los Seraphim. For each child she has recorded Y, the average amount the child eats each day in **hundreds** of calories, and X, the number of fast food restaurants within 3 miles of the child's house. Some data and a simple linear regression printout from her study are given below, although a few values seem to be missing. If you need a missing value and can't figure out how to compute it, simply explain how you would use it to answer the question if you had it.

$\bar{Y} = 21.2$ $\bar{X} = 5$ $SSX = 202$ $n = 62$

```
. regress Calories Restaurants

      Source |       SS       df       MS            Number of obs   =       62
-------------+------------------------------         F(  1,    60)   =    37.10
       Model |     70.2        1      70.2           Prob > F      = 0.00000004
    Residual |    113.6       60       1.9           R-squared       =   ?
-------------+------------------------------         Adj R-squared   =   ?
       Total |    183.8       61       3.0           Root MSE        =   1.376

------------------------------------------------------------------------------
    Calories |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 Restaurants |      .590       .097      6.09  0.00000004       .396     .783
       _cons |   18.251       .518     35.26  0.00000000     17.215   19.286
------------------------------------------------------------------------------
```

(a) Is there a significant **positive** linear relationship between the number of fast food restaurants in a child's neighborhood and the number of calories they consume? Give the the p-value and your real world conclusions for an appropriate test using $\alpha = .05$. (You do NOT need to write out all the details of the test.)

Solution: Since we are asked to show that there is a **positive** relationship this is a 1-sided test. Though you didn't need to I include the hypotheses for completeness:

$H_0 : \beta_1 \leq 0$–there is a negative or no relationship between number of fast food retaurants in a child's neighborhood and the number of calories the child consumes.
$H_A : \beta_1 > 0$–there is a postive relationship between number of retaurants and calorie consumption. This is what we are trying to prove so it is the alternative hypothesis.

Since this is a simple linear regression we can use either the F test or the t-test to answer the question but STATA gives the two sided p-values for these tests and we want the 1-sided one. Since the data (sample slope of +.590) supports the alternative hypothesis we get the p-value by dividing STATA's two-sided p-value in half, getting .00000002. Since this is much smaller than $\alpha = .05$ we reject the null hypothesis and conclude that there is a significant positive linear relationship between number of fast food retaurants and calorie consumption.

Note: There seems to be ongoing confusion about the difference between the **hypotheses** for the F test in a simple linear regression, which are 2-sided or non-directional, and the **test statistic** which is one sided with only large values of F making you reject. It is the hypotheses that matter as the p-value is given for a particular set of hypotheses, not for the test statistic. The test statistic is just an intermediate step to get the p-value. Here with F we've squared everything to express the model's performance in terms of variability so F can only be positive but that doesn't make the hypotheses directional. Thus you really do have to divide the p-value for the F test in half. Another way to see that is to note that the p-value for the F test is the same (in SLR) as the p-value for the t-test for β_1 which you know STATA gives as 2-sided.

(b) Does number of fast food retaurants explain a high **percentage** of the variability in number of calories consumed? Explain briefly, showing any necessary calculations. (You do NOT need to use more than one number to justify your interpretation.)

Solution: The percentage of variability explained is R^2 or R^2_{adj}. Technically R^2_{adj} is better since it takes degrees of freedom into account and therefore gives an unbiased estimate of the explanatory power of the model in the population. However in simple linear regression, with only one predictor, there is very little difference between the two unless the sample size is tiny. We therefore accepted either value. Since these quantities are missing from the printout we have to compute them using the values in the regression ANOVA table:

$$R^2 = \frac{SSR}{SST} = \frac{70.2}{183.8} = .3819 = 38.19\%$$

$$R^2_{adj} = 1 - \frac{MSE}{MST} = 1 - \frac{1.9}{3.0} = .3667 = 36.67\%$$

We explain a bit over a third of the variability in calorie consumption using just the number of neighborhood fast food restaurants as a predictor. While this may not seem that high (compared to a maximum value of 100%) there are many factors that would seem to be much more important for predicting calorie intake such as what the child actually eats, how large they are, how much energy they expend and so on. I therefore think it's pretty good that just the number of area retaurants explains this much of the variability. However if you said this was not a particularly large R^2 value and explained your reasoning we gave credit for it as it is a judgement call.

(c) Does the number of fast food retaurants in their neighborhood do a good job of **predicting** the number of calories a child consumes? Carefully justify your answer.

Solution: To judge whether a model makes good predictions (on average) we look at the RMSE, our typical error in predicting the values in our sample, and compare it to $\bar{Y}$, the typical Y value we are trying to predict. Here RMSE = 1.376. Since Y is in hundreds of calories this means our typical error (or if you prefer average distance from the true value to the value predicted by the regression line) is 137.6 calories. According to the problem the average calorie consumption for kids in this sample is $\bar{Y} = 21.2 = 2120$ calories. In percentage terms we typically make about a $1.367/21.2 = 6.45\%$ error. Again, considering all the other important factors in a child's diet, this seems fairly impressive to me. An extra hundred or so calories is roughly equivalent to a glass of skim milk. However if you argued that an extra 137 calories a day could cumulatively be a big deal

in terms of gaining or losing weight or something similar we accepted that answer.

(d) Find an interval which you can be 95% sure contains the average calorie consumption of children in a neighborhood with 10 fast food restaurants. Show your work.

Solution: Since we want the average value of Y (calorie consumption) at a given X (number of restaurants) rather than a particular person's Y value we need a **confidence interval for Y**. The basic formula is

$$\hat{Y}_0 \pm t_{\alpha/2,n-2} s_{\hat{Y}_0}$$

We have $\hat{Y}_0 = b_0 + b_1 X_0 = 18.251 + .590(10) = 24.151$. Since we want a 95% interval the value of t is $t_{.025,60} = 2$. Finally the standard error is

$$s_{\hat{Y}_0} = RMSE\sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{SSX}} = 1.376\sqrt{\frac{1}{62} + \frac{(10-5)^2}{202}} = .511$$

Putting this all together we get $24.151 \pm 2(.511)$ or [23.122, 25.180] hundreds of calories or [2312, 2518] calories as the range of possible values for the average caloric intake of children in a neighborhood with 10 fast food retaurants.

(e) Standard guidelines suggest that very active children aged 9-13 need approximately 2200 calories per day, while less active children need fewer calories. Based on this and your answer to part (d) does it seem that children in neighboorhoods with 10 fast food restaurants have unhealthy diets on average? Explain briefly.

Solution: The interval from part (d) lies entirely above the recommended daily calorie count for very active children and these are the children who need the most calories. This implies that the average calorie consumption of kids in such neighborhoods is too high for any sort of child. If by an unhealthy diet you mean the children eating more calories than are good for them (which is certainly one important factor) we are more than 95% sure that the children in these neighborhoods have unhealthy diets.

(f) Optional Bonus Suppose you had measured Y in calories and X in dozens of restaurants. What would the resulting regression equation have been? Show your work.

Solution: First let's convert the units for the coefficient of the restaurant variable. Let X be the original variable, number of restaurants, and let X^* be the new variable, dozens of restaurants. The slope, $b_1 = .590$ means for every extra retaurant (change of 1 in X) the calorie consumption goes up by .590 hundreds (or 59 calories). Thus for every extra dozen restaurants (1 unit change in X^*) the number of calories consumed must go up 12*.590 = 7.08 hundreds of calories. Basically, when we plug in a change of 1 in dozens of restaurants, it's like having a change of 12 in the number of restuarants, and the slope which gives the change in Y associated with a 1 unit change in the predictor must reflect this. Thus the equation at this step becomes

$$\hat{Y} = b_0 + b_1^* X^* = 18.251 + 7.08X^*$$

Now this equation gives everything in hundreds of calories. If we want to convert to just plain calories we need to multiply Y (and thus also the right hand side of the equation which is equal to Y) by 100. For instance an old Y = 10 means 10 hundred calories which corresponds to $Y^* = 1000$ calories in the new units. Thus our final equation is

$$Y^* = 1825.1 + 708X^*$$

(3) Television Ads: You own a chain of stores that sells television sets and you want to know whether your advertising is increasing your sales. Let Y be the number of TVs you sell in a given month, and let X be the amount of money you spend on advertising in a given month in thousands of dollars. You have data on advertising expenditures and sales for n=42 months and have fit a simple linear regression of Y on X. The printout for this regression is given below along with a few useful summary statistics. Use it to answer the following questions:

```
The regression equation is
TV Sales = 48.4 + 10.2 Ad-Spending

Parameter Estimates

Predictor       Coef       Stdev    t-ratio        p
Constant       48.40       17.61       2.75    0.009
Ad Spending  10.2457      0.5224      19.61    0.000

Root MSE = 38.54       R-sq = 90.6%     R-sq(adj) = 90.3%

Analysis of Variance

SOURCE       DF          SS          MS         F        p
Regression    1      571411      571411    384.70    0.000
Error        40       59413        1485
Total        41      630824
**********
Summary Statistics For Number of Televisions Sold Per Month

                N     MEAN   MEDIAN    STDEV      MIN      MAX
TV Sales       42    373.5    363.0    124.0    146.0    609.0
```

(a) What **percentage of variability** in television sales is explained by advertising expenditures? Does the model do a good job in this respect? Explain.

Solution: The percentage of variability explained is given by $R^2 = 90.6\%$ or $R^2_{adj} = 90.3\%$ from the printout above. The second number is more accurate since it takes the degrees of freedom into account. However, we accepted either one. They are actually very similar in this case–the regression has explained just over 90% of the variability in television sales. Since the maximum is 100% this seems like a pretty high amount. I would say the regression is doing a good job of explaining the variability in sales.

(b) Do advertising expenditures give good **predictions** for the number of television sales? Briefly justify your answer using appropriate numbers from the printouts.

Solution: The key to answering this question is to compare the average error we are making in our predictions with the values we are trying to predict. The average error is given by RMSE = 38.54. Since Y is television sales, this means that when we try to predict the number of TVs sold each month we are off by about 38 sets. On average (see the "Summary Statistics Table" on the printout) we sell around 373.5 TVs per month with a low one month of 146 and a high another month of 609. So it looks like typically we make an error of around 10% (38.54/373.5) in our predictions. Personally I think this is a fairly big error–I am off by several dozen TVs–but we gave credit for saying it wasn't a big error as long as you made it clear you understood that you needed to compare RMSE to the Y values. We gave a small amount of partial credit for certain other answers (such as p-values). However, remember that R^2, SSR, etc. do NOT tell you whether the regression gives good predictions–they simply say whether you have explained a lot. Even if you explain a lot the remaining error may be significant.

(c) Is there a **significant linear relationship** between sales, Y, and advertising, X? Justify your answer by performing either a T test or an F test (your choice), making sure to give the null and alternative hypotheses both mathematically and in words. Also give the test statistic, the p-value, say whether or not you reject the null hypothesis and why, and state your real world conclusions. (Use $\alpha = .005$)

Solution: To answer this question we need to test whether or not $\beta_1 = 0$. This is because a slope of 0 corresponds to X being useless for predicting Y. Our hypotheses, regardless of whether we are doing a t or an F test, are

$H_0 : \beta_1 = 0$ There isn't a significant linear relationship between advertising expenditures and television sales. (Or advertising costs do not help explain television sales or any of the other ways we had of writing this.)
$H_A : \beta_1 \neq 0$ There is a significant linear relationship between advertising expenditures and televisions sales. How much you spend on advertising does help explain how many TVs you sell, etc.

Our t statistic would be $t_{obs} = \frac{b_1 - 0}{s_{b_1}} = 19.61$ from the printout. The F statistic would be $F_{obs} = \frac{MSR}{MSE} = 384.70$. The corresponding p-value can be read off the printout either from the predictor table or the ANOVA table as .000. This is certainly smaller than $\alpha = .005$ so we reject the null hypothesis. We conclude that there is a significant linear relationship between advertising expenditures and television sales. Knowing how much you spend on advertising does tell you something about what your sales will be like. In fact, the relationship is positive so spending more on ads is associated with higher sales, just as we would hope.

(d) Find a 99% confidence interval for β_1, the slope of the regression line, and briefly explain what it tells you about the relationship between advertising and television sales.

Solution: The general formula for a confidence interval for β_1 in a simple linear regression is

$$b_1 \pm t_{\alpha/2,n-2} s_{b_1}$$

From the printout $b_1 = 10.2457$ and $s_{b_1} = .5224$. We have n=42 months worth of data and $\alpha/2 = .005$ for a 99% interval so we need $t_{.005,40} = 2.704$. The resulting confidence interval is [8.83, 11.66]. This means that we are 99% sure that β_1 is between 8.83 and 11.66. What does that mean? It means for every extra $1000 we spend on advertising we sell between 8.83 and 11.66 more TVs on average. This uses the definition that β_1 gives the change in Y (here TV sales) associated with a one unit change in X (here $1000 more spent on advertising.)

(e Suppose your company makes a $100 profit per television sold BEFORE taking advertising costs into account. According to your **best estimate**, do the ads appear to be paying for themselves? Can you be 99% (really 99.5%) sure? Explain briefly.

Solution: Our best estimate is that $\beta_1 = 10.2457$. In other words, for every extra $1000 spent on advertising we sell an extra 10.2457 TVs. Since we make a profit of $100 per TV before advertising expenses this means every $1000 we spend on advertising results in an extra $1024.57 in TV sales. The sales exceed the costs of the ads–barely!–so our best guess is that the ads are paying for themselves. However, we are NOT 99% sure the ads are paying for themselves. From part (e) all we can say is that we are 99% sure that we have increased our sales between $883 and $1151 for each $1000 spent on ads. The values at the low end of the interval do NOT cover the advertising costs. In fact we might lose over $100 for every $1000 spent on ads! For those of you keeping track of the exact percentages, we can be 99.5% sure of generating AT LEAST $883 in additional sales (there is a .5% chance of above $1166) which is why I worded the question as I did.

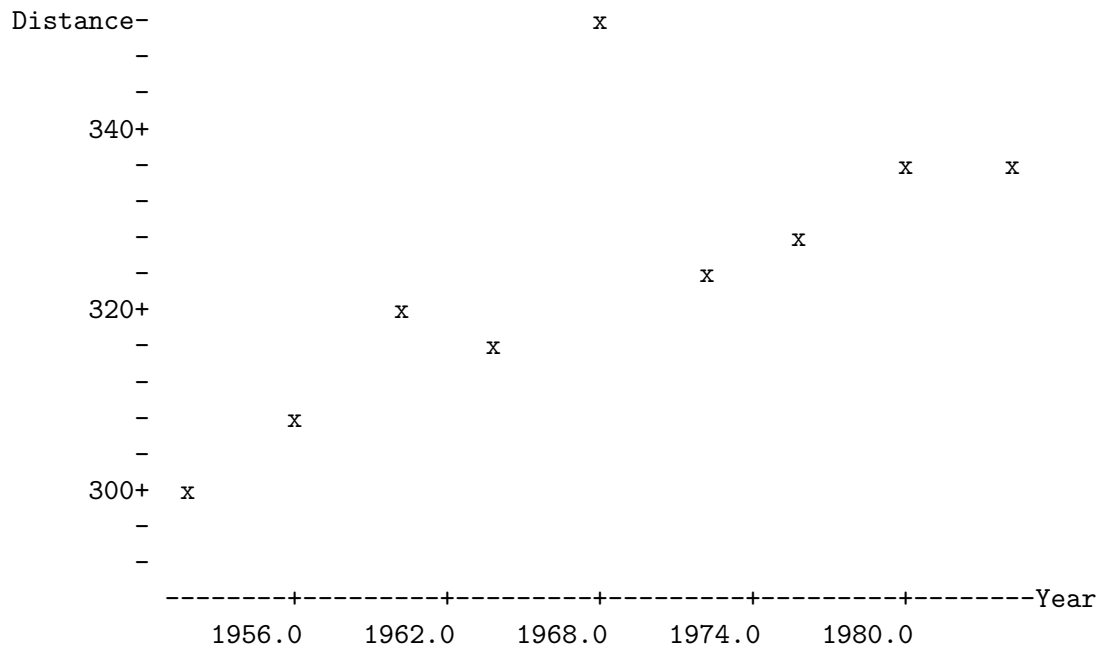
(4) Leaping Into the Future

In the modern Olympic era, performances in track and field have been steadily improving. The table below gives the winning distance (in inches) for the Olympic long jump from 1952 to 1984. Below is a regression printout for a simple regression of distance on year. Use the printout to answer the following questions.

Year	Distance
1952	298
1956	308.25
1960	319.75
1964	317.75
1968	350.5
1972	324.5
1976	328.5
1980	336.25
1984	336.25

```
Scatterplot

        -
```



Regression Analysis

```
. reg Distance Year

      Source |       SS       df       MS              Number of obs =       9
-------------+------------------------------           F(  1,     7) =    9.21
       Model |  1137.52604     1  1137.52604           Prob > F      =  0.0190
    Residual |  864.973958     7  123.567708           R-squared     =  0.5681
-------------+------------------------------           Adj R-squared =  0.5063
       Total |      2002.5     8    250.3125           Root MSE      =  11.116

------------------------------------------------------------------------------
    Distance |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        Year |   1.088542   .3587706     3.03   0.019     .2401839    1.936899
       _cons |  -1817.833   706.0703    -2.57   0.037    -3487.424   -148.2423
------------------------------------------------------------------------------
```

(a) Give the units and interpretation of b_1 in the simple regression model.

Solution: The regression coefficient b_1 always gives than change in Y associated with a one unit change in X. Since b_1 must convert from X units to Y units, the units of b_1 are the units of Y divided by the units of X. In this problem, X is in years and Y is distance in inches, so the units of b_1 are inches per year. Since $b_1 = 1.08854$, a one unit change in year is associated with a 1.08854 inch change in distance, i.e. the winning long jump distance increases by 1.08854 inches per year.

Naturally, since the Olympics are only held every four years, this really means that the winning distance increases by about 4.35 inches every Olympiad.

(b) What proportion of the variability in distance is explained by year using the simple linear regression model? Does the model do a good job in this respect?

Solution: The proportion or percentage of variability explained by the regression is given by $R^2 = 56.8\%$, or, if we want an unbiased estimate, by $R^2_{adj} = 50.6$. Whichever number you use, the regression is explaining barely over half the variability and leaving nearly half the variability unexplained. This is not very good, though it is certainly better than nothing.

(c) Does the simple linear regression model do a good job of predicting the Y values? Make sure you justify your answer.

Solution: This was one of the most frequently missed questions on the exam on which it appeared. In order to tell whether a regression makes good predictions, you need to know how big the errors made by the regression are. One way of evaluating this is to look at the typical distance from the points to the regression line. This number is estimated by $s_{Y|X} = \sqrt{MSE}$. This number can be found as Root MSE on the printout, or by taking the square root of MSE from the ANOVA table. Here $RMSE = \sqrt{123.57} = 11.1161$. To tell whether this means the errors are large, we must compare RMSE to the Y values we are trying to predict. The Y values in this problem range from 298 to 336. Thus we are making an error of roughly 3-4%. This seems pretty good. However, we really should consider the context of the problem. The errors we are making are on the order of 11 inches–nearly a foot. Long jump competitions are usually decided by much less than this so our errors, in context, are still rather large. Note: Many people tried to use R^2 or an F test to say whether the model is a good predictor. These values try to get at whether the model explains a lot of variability. You can explain quite a lot of variability and still have bad predictions.

(d) Is there a significant linear relationship between years and distance? Justify your answer using an appropriate test.

Solution: We could use either a t test or an F test since they are the same for simple linear regression. Our null and alternative hypotheses are

$H_0 : \beta_1 = 0$–i.e. there is not a significant relationship
$H_A : \beta_1 \neq 0$–i.e. there is a significant relationship between the year and the distance of the winning long jump.

From the printout, the test statistics are $t_{obs} = 3.03$ for the t test, and $F = 9.2057$ for the F test. In both cases, the p-value for the test is .0190 which is much less than $\alpha = .05$. Therefore, we reject the null hypothesis and conclude that there is a significant linear relationship between year and the winning long jump distance. To get full credit, you only needed to quote the p-value and explain your conclusions.

(e) In 1968, the Olympics were held in Mexico City, and many records were set, probably due to

the high altitude. A point like this is called an outlier. Explain what would happen to your answers to (b)-(d) if this point were removed.

Solution: If the point is removed, the regression line will go right through the middle of the rest of the points. Thus the amount of unexplained variability will be smaller and the amount of explained variability will be higher. This will cause R^2 to go up, $s_{Y|X}$ to go down (and hence we will get better predictions), and F to increase (resulting in a lower p-value for our test).

(5) Computer Chaos:

You have been hired as a statistical consultant by a large hardware store. They are interested in knowing how their sales of fans depend on the weather. They have presented you with data from the previous summer. Their data consists of two variables, Y, the number of fans sold in each week, and X the hottest temperature during that week. They have given you data for n=12 weeks. During those weeks the average temperature was found to be $\bar{X} = 80$ and the average number of fans sold per week was $\bar{Y} = 160$. You have further managed to calculate from your data that $SCP = 200, SSX = 100$, and $RMSE = 4$. You have gotten sick of doing the calculations by hand and decided to use a computer. Unfortunately (what a shock) the program is malfunctioning and your printout has a lot of blanks. In this problem you will fill in the blanks and answer some questions for the hardware store. **Note:** It is possible to completely answer parts (b)-(g) even if you can't fill in a single number in the printout, so don't give up on them!!

(a) Below is a printout given by your computer. Fill in the blanks (____) with the appropriate numbers using the information given above. I have left a blank page after this one on which to show your work and a suggested order for doing the calculations. Give at least a brief indication, either in formulas or words, of how you got the numbers. If there is a number you can't figure out, put in a symbol for it and show how you would get all the other numbers using the symbol.

```
The regression equation is

Fans = _____ + ______Temperature


Predictor         Coef    SE Coef        T             P

Constant          ____         15      _______       1.000

Temperature       ____       ____      _______        .000


RMSE = _____   R-sq =  ____    R-sq(adj) = .6857


Analysis of Variance
```

```
SOURCE       DF    Sum Squares  Mean Squares      F        P

Regression   ___         ____        _____     _____    _____

Error        ___         ____        _____

Total        ___          560
```

(1) Find the estimated regression equation.
(2) Fill in the table below the regression equation.
(3) Fill in RMSE and MSE.
(4) Fill in the degrees of freedom in the ANOVA table, and then the rest of the ANOVA table. (Note: No calculations are needed for the p-value!)
(5) Fill in R^2.

Solution: I give the filled in printout below. I got the numbers in this order. First,

$$b_1 = \frac{SCP}{SSX} = \frac{200}{100} = 2$$

Second,

$$b_0 = \bar{Y} - b_1\bar{X} = 160 - (2)(80) = 0$$

This lets us fill in the regression equation and the Coef column of the table below it. To get the missing entry in the SE Coef column we need to compute

$$s_{b_1} = \frac{RMSE}{\sqrt{SSX}} = \frac{4}{\sqrt{100}} = .4$$

Then we get the t ratios by dividing the Coef column values by the SE Coef column values to get 0 and 5.

We are given that RMSE= 4. Also, $MSE = s^2_{Y|X} = 4^2 = 16$ in the ANOVA table. Then we can fill in the degrees of freedom. In a simple regression we have 1 degree of freedom for regression, n-2 = 12-2 = 10 for error, and n-1=12-1 = 11 for total.

Now we can fill in the various sums of squares. We know MSE = 16 and has 10 degrees of freedom. Since MSE = SSE/n-2 we must have SSE = 16*10 = 160. Next, we note that SSR + SSE = SST and we know SST = 560 from the table. Thus SSR = 400. MSR = SSR/1 for simple regression so MSR = 400 also. F = MSR/MSE = 400/16 = 25. The p-value for the F test in a simple linear regression is the same as that for the t-test of β_1 so we fill in 0. This completes the ANOVA table.

Finally we must obtain R^2. From the ANOVA table, $R^2 = SSR/SST = 400/560 = 71.43\%$. The complete printout is below.

```
The regression equation is

Fans = 0 + 2Temperature

Predictor     Coef   SE Coef       T      P
Constant         0        15    0.00   1.00
Temperature      2        .4    5.00   0.00

RMSE = 4      R-sq =  71.43%  R-sq(adj) = 68.57%

Analysis of Variance

SOURCE       DF          SS          MS         F        p
Regression    1          400         400        25     .000
Error        10          160          16
Total        11          560
```

(b) Give the units and real-world interpretations of the regression coefficients β_0 and β_1. (Note: You do not need to quote the numbers to do this, though it may be helpful to do so if you know them.)

Solution: This question, especially the part about units, created some difficulties so take careful note of my answers. First, β_0 is the average value of Y when X=0. In real world terms this means that β_0 represents the number of fans sold by the store WHEN THE TEMPERATURE IS 0 DEGREES. Since β_0 is in essence a Y value it must have the same units as Y, in this case, fans per week. It was necessary to say this explicitly! (Note that we found $b_0 = 0$ meaning we estimate that on average the store sells 0 fans when the temperature is 0 degrees. This makes perfect sense!) β_1 is the slope of the regression line and represents the change in Y associated with a one unit change in X. Here that means that β_1 tells us how many extra fans we sell for each 1 degree that the temperature increases. Our best estimate is $b_1 = 2$ meaning that we sell, on average, 2 extra fans for every additional degree of temperature. The units of b_1 are always the units of Y divided by the units of X. This is because we must multiply X by b_1 and come out with a Y value. Here those units are fans per week per degree.

(c) Is there a significant linear relationship between temperature and the number of fans sold by the store? Answer this question by performing a t test. You must state the null and alternative hypotheses, both mathematically and in words, quote the p-value, and give your conclusions. You do not need to quote the test statistic and no calculations are required. Use $\alpha = .05$.

Solution: Testing whether there is a significant relationship between fan sales and temperature is equivalent to testing whether $\beta_1 = 0$. Our hypotheses are

$H_0 : \beta_1 = 0$–there is not a significant relationship between fan sales and temperature, or equivalently, temperature does not help explain the variability in fan sales.
$H_A : \beta_1 \neq 0$–i.e. there is a significant linear relationship between temperature and fan sales, or equivalently temperature does help explain the variability in fan sales.

From the printout in (a) the p-value for the t-test is .000. This is much smaller than $\alpha = .05$ so we reject the null hypothesis and conclude that there is a significant linear relationship between fan sales and temperature. Knowing how warm it is does tell the store something about how many fans they will sell. This is hardly a surprise. However, knowing b_1 does give them an idea of how many fans they should stock at any given time. Note that we could also have used an F test for this problem if I hadn't specified the t test!

(d) Calculate a 95% confidence interval for β_0. Based on your interval, is β_0 different from 0? Explain. (Note: If you couldn't get b_0 in part (a), you may assume it is 1 for this part of the problem.) What does this interval tell you?

Solution: The formula for a confidence interval for β_0 is

$$b_0 \pm t_{\alpha/2,n-2} s_{b_0}$$

Here we found $b_0 = 0$ in part (a), and $s_{b_0} = 15$ from the printout in part (a). We have n=12 data points and want a 95% confidence interval so we use $t_{.025,10} = 2.228$. The resulting confidence interval is $0 \pm (2.228)(15)$ or [-33.42, 33.42]. Since this interval includes 0 we cannot conclude that β_0 is different from 0. In fact, since b_0 was exactly 0 it is obvious that our data are consistent with β_0 being 0! This makes perfect sense. β_0 is the number of fans the store sells when the temperature outside is 0. Obviously if it is below freezing the store won't be selling any fans!

(e) Is temperature a good **predictor** of fan sales? Quote the number that you use to determine this and briefly explain your reasoning.

Solution: This was one of the most frequently missed questions on the exam. Take careful note!! To tell if the regression is making good predictions we must look at RMSE. This quantity tells us the average distance from the data points to the regression line and can be roughly interpreted as the average error we are making in guessing Y. If this value is small we are doing a good job and if it is large we are doing a bad job. Here $RMSE = 4$ which means our predictions are typically off by about 4 fans per week. We are told that the store sells $\hat{Y} = 160$ fans in an average week, so only being off by 4 fans is very good–an error of about 2.5%. Note that we MUST compare $RMSE$ to the Y values to tell if our predictions are good! An error of 4 is very small compared to sales of 160 but would be very large if we were only selling 5 fans a week! Also note that a high R^2 does NOT prove our predictions are good. It says we have explained a high percentage of the variability in Y but even a small percentage of unexplained variability can result in large errors from a practical point of view. Similarly, a very small p-value does not prove our predictions are good. It says our X variable is useful–that our predictions are much BETTER than if we didn't use X–but it doesn't prove they are right. For an example simply see the homework warmup problem on electricity usage. We had a p-value of 0 but horrible predictions! Of course in that case we also had the wrong model, but the basic idea still holds....

(f) What **percentage of the variability** in fan sales is explained by the regression on temperature? Quote the number that you use too determine this and say whether you think the regression is doing a good job in this respect.

Solution: We use R^2 or R^2_{adj} to find the percentage of variability explained by the regression. They have the same intuitive meaning but R^2_{adj} is a little more accurate because it takes degrees of freedom into account. We have $R^2_{adj} = 68.57\%$, so the regression explains roughly two thirds of the variability in fan sales out of a possible 100%. This is pretty good–well over half–but not fabulous–values in the 80's or 90's are usually considered very high.

(g) A weather forecast says next week's temperature will soar to 100! Predict the number of fans you will sell next week. Suppose you want a range of possible values for the number of fans you will sell. Calculate the appropriate interval and explain your reasoning. (Use $\alpha = .05$) How many fans should you stock to be sure you have enough on hand?

Solution: For the prediction we simply plug X=100 into the regression equation and find that we expect to sell $\hat{Y} = 0+2(100) = 200$ fans. For the interval, since we are dealing with a single specific Y, namely next week's sales, rather than average sales when it is 100 degrees, we want a prediction interval. The basic formula for a prediciton interval is

$$\hat{Y_0} \pm t_{\alpha/2,n-2} s_{Y|X} \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{SSX}}$$

We have $t_{.025,10} = 2.228, s = 4, n = 12, X_0 = 100, \bar{X} = 80, SSX = 100$. The resulting interval is [194.98, 205.02]. To be sure we have enough fans in stock we should have 106, using the high end of the interval to be safe since this is the maximum number we could sell and rounding up since we can't sell part of a fan and 205 could be slightly too low.

Multiple Regression

(1) Harry Potter and the Sorcerer's Statistic: Polygon Pictures, the film-making branch of Mathematical Media Incorporated, is interested in knowing what factors contribute to the profitability of their movies. For their last n=27 films they have recorded Y, the box-office sales (in millions of dollars), X_1, the production costs for the film (in millions of dollars), X_2, the number of theaters in which the film was shown, X_3 the advertising budget for the film (in millions of dollars), and X_4 which is 1 if the movie featured a big name star and 0 if it didn't. They have also classified the films as action/adventure ($X_5 = 1, X_6 = 0$), comedy ($X_5 = 0, X_6 = 1$), or romance ($X_5 = 0, X_6 = 0$). A multiple regression printout for their data is shown below along with some possibly helpful statistics. Use this information to answer the following questions.

```
Correlations:
            Box Theaters      Ads
Theaters  0.900
Ads       0.925    0.889
Cost      0.950    0.912    0.927
**********************************
Summary         N     MEAN   MEDIAN    STDEV
Box            27    35.00    25.00    14.78
```

```
              MIN       MAX        Q1        Q3
Box          5.00     70.00     20.00     40.00
**********************************
The regression equation is
Box = - 0.842 + 1.84 Cost + 0.0025 Theaters - 0.628 Ads + 5.47 Star
          + 4.59 Action - 5.14 Comedy

Predictor       Coef       Stdev    t-ratio        p
Constant     -0.8423      0.9715      -0.87    0.396
Cost          1.8365      0.1339      13.71    0.000
Theaters      0.0025    0.001455       1.71    0.102
Ads          -0.6282      0.5574      -1.13    0.273
Star          5.4713      0.7191       7.61    0.000
Action        4.5880      0.6523       7.03    0.000
Comedy       -5.1441      0.6412      -8.02    0.000

RMSE = 0.8970      R-sq = 99.7%     R-sq(adj) = 99.6%

Analysis of Variance

SOURCE       DF          SS          MS         F        p
Regression    6     5665.32      944.22   1173.54    0.000
Error        20       16.09        0.80
Total        26     5681.41
```

(a) Is the regression **overall** useful for explaining the box-office take of the movies? Justify your answer with an appropriate hypothesis test using $\alpha = .05$. You do not need to write out all the details here–just explain your basic reasoning–but on the exam you should be prepared to give the full details!

Solution: To check whether the regression as a whole is useful we use an overall F test. Our hypotheses are

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$–None of the variables production costs, advertising expenses, type of film, etc. is useful for explaining box office sales.
H_A : At least one of the $\beta_i's \neq 0$–At least one of the variables listed above is useful for explaining variability in box office sales. The regression as a whole is useful for explaining sales.

For an F test, our test statistic is the F value from next to the ANOVA table, here $F = 1173.54$ and the corresponding p-value is 0.000. (Note that we can't use any of the t statistics or their corresponding p-values. Since this is a multiple regression, testing for a single variable–as the t test does–is not equivalent to testing whether the regression is useful overall.) Since our p-value is less than our significance level $\alpha = .05$ we reject the null hypothesis and conclude that at least one of the variables production cost, advertising expenses, etc. IS useful for explaining variability in box office sales. This is hardly a surprise! All you needed to do was quote the appropriate p-value and state your conclusions.

(b) What percentage of the variability in box-office take is explained by the variables in this model? What value should you use to check this and why?

Solution: From the regression printout $R^2_{adj} = 99.6\%$ so we have explained nearly all of the variability in box office take. Note that we use R^2_{adj} since this is a multiple regression and we want to avoid overfitting. R^2_{adj} will penalize us for adding useless variables to the model because it takes the number of predictor variables into account.

(c) Does the model do a good job of predicting box-office take? Briefly justify your answer.

Solution: To judge whether a model does a good job of prediction we look at RMSE and compare it to the Y values. Here we have $RMSE = .8970$ which means we make an average error of \$897,000 in predicting the box office take of movies. The box office take of movies in our sample ranged from \$5 million to \$70 million with a mean of \$35 million. For the movies at the low end we are making a pretty big error but overall in my judgment we are doing a pretty good job based on how difficult it is to predict the success of films! (Also note, that it is possible the errors the model makes get bigger as the box office take of the film increases. We would need a residual or scatterplot to tell this for sure.)

(d) Find a 95% confidence interval for β_4, the coefficient of the big name star variable, and give a brief real-world interpretation of your interval.

Solution: The general formula for the confidence interval of a multiple regression coefficient is given by

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

From the regression printout, $b_4 = 5.4713$ and $s_{b_4} = .7191$. We have n=27 data points and k=6 predictor variables and we want a 95% confidence interval so we need $t_{.025,20} = 2.086$. The resulting confidence interval is $5.4713 \pm (2.086)(.7191)$ or [3.9713, 6.9713].

Now β_4 is the coefficient of an indicator variable so it gives the difference in box office sales between equivalent films (production costs, ads, etc.) one of which has a big name star and the other of which does not. Thus the confidence interval says that we are 95% sure that having a big name start in our film adds between \$3.97 million and \$6.97 million to the box office sales of a film, all other things being equal.

(e) Suppose Polygon Pictures gets 50% of the box office take. What is the maximum amount they could pay a big name star and be 95% (really 97.5%) sure it was economically worthwhile?

Solution: If Polygon only gets 50% of the box office take, then multiplying the CI from part (f) by .5 we see that having a big name star in a film adds somewhere between \$1.985 million and \$3.485 million to **Polygon's profits**, all other variables being held fixed. To be 95% sure (actually 97.5%) that they make money by using a star, Polygon cannot afford to pay more than the low end of this interval or \$1.985 million. Many people tried to use the high end of the interval. But

$3.485 million is the MOST Polygon can benefit by having the star so they can't possibly make a profit on the deal if they pay the star that much!

(f) Polygon Pictures is about to begin filming Harry Potter and the Sorcerer's Statistic, an adventure film that they plan to release in 3000 theaters with an advertising budget of $2 million, production costs of $30 million, and no big name stars. What is the predicted box office sales for this film?

Solution: All we have to do is plug in the values of the X variables to our estimated regression equation, being careful of our units. X_1 is production costs for the film in millions of dollars, and so from the problem statement $X_1 = 30$. The number of theaters is $X_2 = 3000$. The advertising expenses are recorded in millions of dollars so $X_3 = 2$, the movie has no big name star so $X_4 = 0$ and the movie is an adventure film so $X_5 = 1$ and $X_6 = 0$. Thus our predicted box office sales are

$$\hat{Y} = -.8423 + 1.8365(30) + .0025(3000) - .6282(2) + 5.4713(0) + 4.5880(1) - 5.1441(0) = 65.0843$$

In other words the studio can expect to make a little over $65 million on the Harry Potter film. (Note: If you round and use the coefficients given at the start of the printout you get 65.192. We accepted either answer.)

(g) Suppose the Polygon CEO wants an interval which is 95% certain to contain the true box-office take of the Harry Potter film. What type of interval should she use, a confidence interval or a prediction interval? Explain briefly.

Solution: Since they want an interval which will contain the sales figures for a **single** film, namely Harry Potter, Polygon should use a **prediction interval.** A confidence interval would only be used if they wanted to learn about the **average** sales of all films with the same production costs, advertising expenses, and so on as the Harry Potter film.

(h) What would be the difference in box-office sales if Polygon decreased the overall production costs by $1 million but added a big name star to the cast? Does this seem like a good move? Briefly justify your answer. Note that you should NOT need to make an entirely new prediction of sales.

Solution: We know that the coefficient of production costs is estimated by $b_1 = 1.84$. Therefore, reducing the overall production costs by $1 million will, on average, be associated with $1.84 million LESS in box office sales. On the other hand $b_4 = 5.47$ so on average, adding a big name star is associated with additional revenue of $5.47 million. The expected net difference in box office sales under the proposed change is 5.47 - 1.84 = 3.63 or $3.63 million in **additional** profit. Since the studio increases its box office take without a net increase in expenses (production costs were brought down, even including the new star!) this definitely seems like a good move if it is possible. Note that it is NOT necessary to completely redo the prediction–you can get the answer just by looking at the coefficients of the variables that have changed.

(i) Which type of film is generally most profitable, all other things being equal? An action/adventure film, a comedy, or a romance? Explain briefly.

Solution: Action/adventure films are the most profitable, followed by romance films, and comedy films are the least profitable. We deduce this by looking at the coefficients of the Action and Comedy variables, X_5 and X_6. Note that $X_5 = 1, X_6 = 0$ corresponds to an action film, $X_5 = X_6 = 0$ is a romance film, and $X_5 = 0, X_6 = 1$ is a comedy film. Thus romance films serve as a reference category. Since $b_5 = 4.588$ we see that, all other things being equal, an action film makes \$4.588 million more than a romance film, while $b_6 = -5.1441$ means than on average a comedy makes \$5.1441 million less than an otherwise equivalent romance film. Note that these amounts only give the **difference** in box office sales between the different types of films. It is not accurate to say that a comedy film creates a loss or that a romance film brings in no money. To know the actual box office sales of a film you need to know the values of $X_1 - X_4$ as well.

(2) When I Finish Summer School....I'm Going To StatisticsLand?!

Our old friend Professor Sadisticus has gone into the amusement park business. He is currently trying to determine what factors affect attendance at his parks. He has recorded Y, the number of visitors (in millions) to each of his parks each quarter for the past 5 years. Average attendance has been $\bar{Y} = 5$ or 5 million people. He has also recorded data on X_1 time (with time 1 being the first quarter, winter, five years ago), X_2 the price of tickets to the park (in dollars), X_3 the number of rides at the park, X_4 the size of the park (in acres), X_5 the population of the city in which the park is located (in hundreds of thousands of people), and X_6 the average temperature during the quarter (in degrees). He also has indicator variables for whether there were special discounts offered to local residents ($X_7 = 1$ if there was a discount and $X_7 = 0$ if there wasn't) and for the region of the country in which the park was located $X_8 = X_9 = 0$ for the west coast, $X_8 = 1, X_9 = 0$ for the midwest, and $X_8 = 0, X_9 = 1$ for the east coast.) He has fit a multiple regression of Y on these nine variables. Use the regression printout and accompanying summary statistics to answer the questions on the following pages.

```
Correlations:
               Attendance Price Rides Size
Price             -.7
Rides              .8      .5
Size               .7      .4    .9
Population         .7     -.1    .1    .2
**************************************************************************
The regression equation is

Attendance = 2 + .25Time - .2Price + .01Rides - .001Size + .05Population +
.1Temperature + 1Discount  -2Midwest - East

Predictor          Coef      SE Coef         T             P
Constant          2.000         .500      4.00         .0001
Time               .250         .100      2.50         .0142
Price             -.200         .050     -4.00         .0001
Rides              .010         .004      2.50         .0142
Size              -.001         .002     - .50         .6180
Population         .050         .025      2.00         .0480
Temperature        .100         .048      2.08         .0396
Discount          1.000         .400      2.50          0142
Midwest          -2.000        1.000     -2.00         .0480
East             -1.000        1.000     -1.00         .3196

RMSE = .200     R-Sq = 95.6%      R-Sq(adj) = 95.24%

Analysis of Variance

Source            DF          SS          MS            F          P
Regression         9       95.60       10.62        265.5      0.000
Residual Error   110        4.40         .04
Total            119      100.00
```

Note: EXCEPT WHERE INDICATED OTHERWISE YOU SHOULD USE $\alpha = .05$ FOR ALL HYPOTHESIS TESTS ON THIS EXAM

(a) Does the model as a whole do a good job of explaining attendance at StatisticsLand parks? Answer this question by performing an appropriate hypothesis test. State the null and alternative hypotheses both mathematically and in words, give the test statistic and p-value, and explain your real-world conclusions.

Solution: To test whether the model as a whole is useful we need an **overall F test.** Our hypotheses are

$H_0 : \beta_1 = \beta_2 = \cdots \beta_9 = 0$–None of time, price, etc. are helpful for explaining the attendence at StatisticsLand parks.
H_A : At least one $\beta_i \neq 0$–At least one of time, price, etc. is useful for explaining attendence. The model as a whole does a good job of explaining the number of visitors at the theme parks.

Our test statistic is $F_{obs} = 265.5$ and the corresponding p-value from the ANOVA table is 0. Since this is less than our significance level of $\alpha = .05$ we reject the null hypothesis. We conclude (not surprisingly) that at least one of the nine variables does help explain the attendence at the parks.

(b) Does this model do a good job of **predicting** attendance at StatisticsLand parks? Carefully justify your answer.

Solution: To decide whether a model makes good predictions we must look at our typical error, $RMSE$ and compare it to the Y values we are trying to predict. Here $RMSE = .2$, meaning that we typically are off by about 200,000 people when we try to predict the number of visitors to a StatisticsLand park in a given quarter. Since we average $\bar{Y} = 5$ or 5 million visitors
(c) Give the real-world interpretations of b_3 and b_7 in this model. Your answer should include the actual numerical values and units as appropriate.

Solution: The coefficient for the Rides variable is $b_3 = .01$. This means for every additional ride there is at the park, attendance goes up by .01 million = 10,000 people assuming all the other variables are held fixed. The coefficient for the discount variable is $b_7 = 1$. This tells us that on average, when there is a discount being offered 1 million mode people come to the park per quarter than when there isn't a discount, assuming all other variables are held fixed.

(d) Based on the correlation table given above the regression printout, should X_4, the size of the park, be a good predictor of attendance? Should it's coefficient, β_4, be positive or negative? Explain briefly in each case.

Solution: The size variable has a quite strong correlation with attendence, Y, so it should by itself be a good predictor. Since the correlation is positive we would expect b_4 to be positive as well, indicating that as size increases so does attendance. This makes real-world sense. The bigger the park is the more people it can hold.

(e) Are b_4, the estimated regression coefficient for X_4, and its p-value consistent with your expectations from part (e)? Justify your answer, and, if there is a lack of consistency, say what has gone wrong, giving evidence from the data to support your argument.

Solution: No, the coefficient b_4 is negative and it's p-value of .6180 is well above .05 suggesting that this variable is not useful WHEN ALL THE OTHER VARIABLES ARE IN THE MODEL. This is probably a multicollinearity issue. From the correlation table we see that Size is also highly correlated with the number of rides (.9)–not a surprise since to have a lot of rides you need to have a lot of space! Since the rides variable is more highly correlated with attendance this is the one that has stayed significant in the model.

(f) Does it appear that there are statistically significant differences in park attendance in the three regions of the country? If so, in which region(s) is attendance the highest? Briefly justify your answers.

Solution: We need to look at the p-values for the indicator variables for the different regions. The West is our reference region. The Midwest indicator has a p-value of .048 meaning attendance there is significantly different than in the west. From the coefficient of -2 we see that on average attendance there is 2 million people lower per quarter. However, the p-value for the East regio is .3196 so there is insufficient evidence to show that attendence is different in the East than in the West. Thus we conclude that the West and East have the highest attendence. it appears since East's coefficient is negative that West is the highest but we do not have sufficient evidence to say this for sure.

(g) On average, how much would you expect attendance at a theme park to decrease in a quarter if you raised the ticket prices by $5, all other things being equal? Briefly explain your reasoning.

Solution: The coefficient for the price is -.2 meaning that for every extra dollar charged we get a decrease in attendance of 200,000 people, all else being equal. Thus a 5 dollar raise would correspond to a loss of 1 million people.

Part i

Find the predicted attendance for the Los Seraphim theme park this summer (that is, summer of the first year after the recorded data.) Los Seraphim is a west coast city with a population of 5 million people, and an average summer temperature of 70 degrees. You may assume that the park has 50 rides, has 50 acres of space, that admission is $45, and that there are currently no discounts being offered.

Solution: We just plug into the regression equation being careful of our units. We know the data has been measured quarterly for 5 years so there have been 20 time points before this year. Summer is the 3rd quarter of the year (winter, spring, summer, fall) so this must be time point 23. The population is measured in hundreds of thousands of people so we have $X_5 = 50$ for 5 million people. There is no discount offered so $X_7 = 0$ and the park is in the west so $X_8 = X_9 = 0$. The rest of the numbers can be plugged in as they are given yielding

$$\hat{Y} = 2 + .25(23) - .2(45) + .01(50) - .001(50) + .05(50) + .1(70) + 0 - 2(0) - 1(0) = 8.7$$

Thus we predict the park will have 8.7 million customers this summer.

(j) Professor Sadisticus is planning to open a new theme park in the city of Hollybrick. He knows it will take a while for the park to become profitable but would like to be 95% sure that **in total** over the next 10 years attendance will be high enough so that he does not lose money on it. (i) What sort of interval should he use when predicting quarterly attendance at the park to find his projected profits over this period? (ii) Do you see any potential problems with the predictions he

is making? Briefly explain your reasoning in each case. No calculations are required.

Solution: Prof. Sadisticus should use a confidence interval, not a prediction interval. He isn't interested in what happens in one particular quarter–he wants to know on average over the long haul how much money he will make. The problem with his predictions is that he is extrapolating 10 years into the future, way outside the range of his data. There is no guarantee his parks will continue to grow at the same rate for such a long time.

(k) Professor Sadisticus has a theory that as it gets hotter more people come to his theme parks. Because of this he is considering adding a new water ride, the Random Splatter, and setting up ice-cream stands on hot days, but before the does this he wants to be sure that his theory is correct. Perform the appropriate hypothesis test to prove Professor Sadisticus' theory. Be sure you state your null and alternative hypotheses both mathematically and in words with a justification of your choice, give the test statistic and p-value, and explain your real-world conclusions.

Solution: Here we are being asked to do a 1-sided test because he wants to prove as temperature goes up attendance goes up. This must be our alternative hypothesis so we have

$\beta_6 \leq 0$–Once the other variables have been taken into account terperature has a negative or no relationship with attendance.
$\beta_6 > 0$–After accounting for the other factors, temperature has a positive relationship with attendance.

Our test statistric is $t_{obs} = 2.08$. Since the test is 1-sided we have to divide the p-value from the printout in half yielding .0198. Since this is less than our significance level of .05 we reject the null hypothesis and conclude that higher temperatures are associated with higher attendance even after accounting for all the other variables.

(l) Silly Sally, a summer intern at StatisticsLand, is very excited by the results of your test in part (j). She concludes that hotter weather causes people to visit your theme parks and proposes that new parks should be opened in desert areas like Arizona or Saharan Africa. Explain what is wrong with (i) Sally's conclusion and (ii) her proposal. Your answer should include an example of why Sally's conclusion might be wrong.

Solution: Sally is making several mistakes. First, correlation is not necessarily causation. People may be visiting the parks because it is summer and that is when they are on vacation, not because it is hot outside. Her proposal is bad because even if warm weather brought people to the parks, extreme heat as in the desert might not. Models are only useful for making predictions about situations similar to the data on which they were built. Saharan Africa for instance is very different from any of the places in the US that Prof. Sadisticus has his parks and our model will not be reliable there. In fact, probably attendance at parks in such an area would be very low.