

1. (20 pts) What about the effect of substitute teachers on student test performance. Enter your regression results as Table 1 and your call your scatterplot Figure 1. I'm really only interested in a few particulars.

- a) What is the "theoretical model" that underlies this regression analysis. I'll give you a hint by getting this started:

$$iqscore_i = b_0 + b_1 subs_i + e_i$$

Identify the "parameters". Define any symbols you use. If you don't include an error term I'm going to be really sad :<(

b_0, b_1 are unknown constants, the intercept and slope of a regression line.

$subs_i$ is the score on the substitute teacher variable for the i 'th case

e_i is a random disturbance (error) term that, at minimum, satisfies the requirement that, for all i , $E(e_i) = 0$, and for an "ordinary" least squares model, it is also necessary that $E(e_i^2) = \sigma_e^2$. Alternatively, we can also impose a stronger requirement that

$$e_i \sim N(0, \sigma_e^2) \text{ for all } i$$

- b) After you calculate your regression estimates, you should be able to fill in the values in this expression for the "linear predictor".

$$\widehat{iqscore}_i = 130.8118 - 1.26 subs_i$$

The regression results are presented in Table 1

- c) I wonder how confident you are about your estimate of the slope? How could you summarize your confidence for me?

The estimate of b_1 is 1.26. The standard error of \hat{b}_1 is 0.1097, which implies a t value of -11.51. If $b_1 = 0$ (referring to the "true" value of the slope), then that standard error indicates that we would expect 95% of all estimated slopes to be within the range from $[0 - 1.96 \cdot 0.1097, 0 + 1.96 \cdot 0.1097]$. The estimate we got is far outside that range, so, by the usual hypothesis testing procedure, we are easily able to reject the null hypothesis that $b_1 = 0$.

A different way of gauging our uncertainty is to construct a confidence interval. That interval is the same width as the one mentioned in the previous paragraph, but it is centered on the estimated value: $[1.26 - 1.96 \cdot 0.1097, 1.26 + 1.96 \cdot 0.1097]$. The interpretation of this interval is that we believe with probability 0.95 that the "true value" of b_1 is within that interval.

- d) What is your estimate of the error term's standard deviation?

In the summary output in R, the error term's standard deviation is represented by the value labeled "Residual standard error." In my output, that value is 27.61. In my regression table, that value is labeled RMSE, an abbreviation for "Root Mean Squared Error".

- e) (5pts Extra Credit) Oh My God. Something went horribly wrong. I ran this:

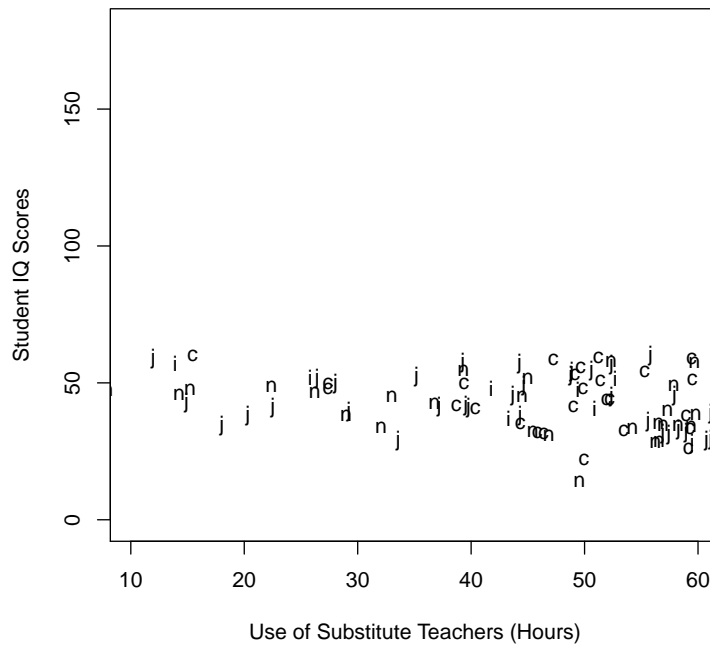
```
plot( iqscore~subs, data=dat, type="n",
      xlab="Use of Substitute Teachers (Hours)",
      ylab="Student IQ Scores")
text( dat$iqscore, dat$subs, labels=abbreviate(dat$religionFactor,1))
```

Table 1: Regression of Student Scores on Substitute Teacher Usage

Variable	<i>OLS.Estimate</i> (<i>std.err.</i>)
Intercept	130.812*** (4.135)
Substitute Usage	-1.262*** (0.110)
R-squared	0.223
adj. R-squared	0.222
RMSE	27.614
F	132.489 * **
N	463

*** $p \leq 0.001$

And something went horribly wrong. Look at this disaster:



What did I do wrong?

The text command has the x and y values reversed. This would fix the problem:

```
text(dat$subs, dat$iqscore, labels=abbreviate(dat$religionFactor,1))
```

- (10 pts) I wonder how your regression changes when you introduce “religion” as a factor variable. Attach a regression table that uses both “subs” and “religionFactor” as predictors.

Table 2: Regression including Religion as a Factor

Variable	<i>OLS.Estimate</i> (<i>std.err.</i>)
Intercept	133.699*** (4.630)
Substitute Usage	-1.274*** (0.105)
Religion Intercept Shifts	
Islamic	9.797** (3.485)
Jewish	-11.627*** (3.478)
None	-7.918* (3.480)
R-squared	0.292
adj. R-squared	0.286
RMSE	26.447
F	47.251 * **
N	463

* $p \leq 0.05$ ** $p \leq 0.01$ *** $p \leq 0.001$

a) Discuss the estimates for religion.

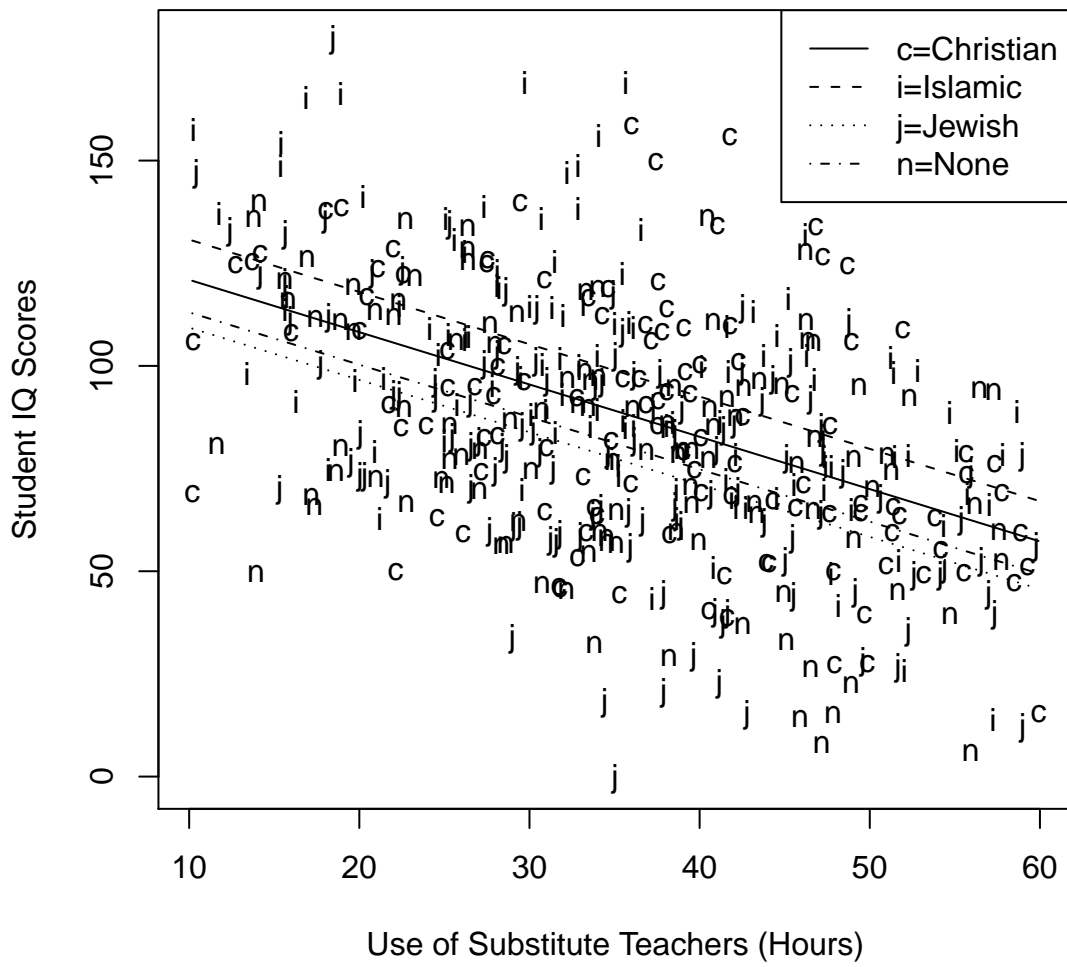
The estimates are presented in Table 2. Religion is coded in four categories, “None”, “Christian”, “Islamic” and “Jewish”. For a factor variable with 4 possible values, R creates contrasts (variables coded 0 or 1) for the last (alphabetically) 3 categories. The estimates represent shifts in the intercept of the regression line for each of the named religions. That is to say, the estimate of 9.797 for Islamic students indicates that our predicted value for children of that faith is 9.797 units higher than our prediction for the group that is represented by the regression intercept, which in this case is Christians. That difference is statistically significant, in the sense that we would be very unlikely to observe such a large effect if the true difference were 0. The parameter estimates for Jewish and None are -11.627 and -7.918, both of which are statistically significantly different from 0. This indicates that the “true” parameters which govern the creation of data in this theory are very unlikely to be 0. Note that we have not conducted a test to tell whether the Jewish students are statistically significantly different from None, but it might be interesting to make that test.

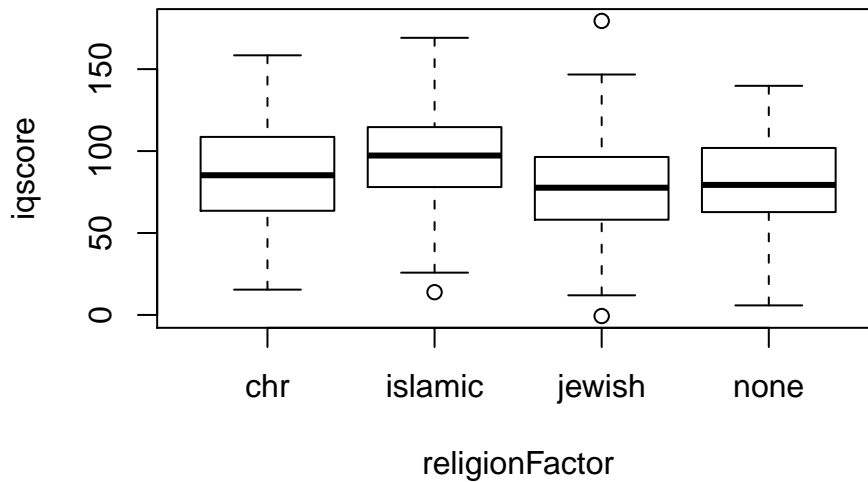
b) I’m having trouble visualizing the effect of religion on student test scores. This command

```
plot( iqscore ~ religionFactor, data=dat)
```

surprised me with this:

Figure 1: Regression with Religion as a Factor





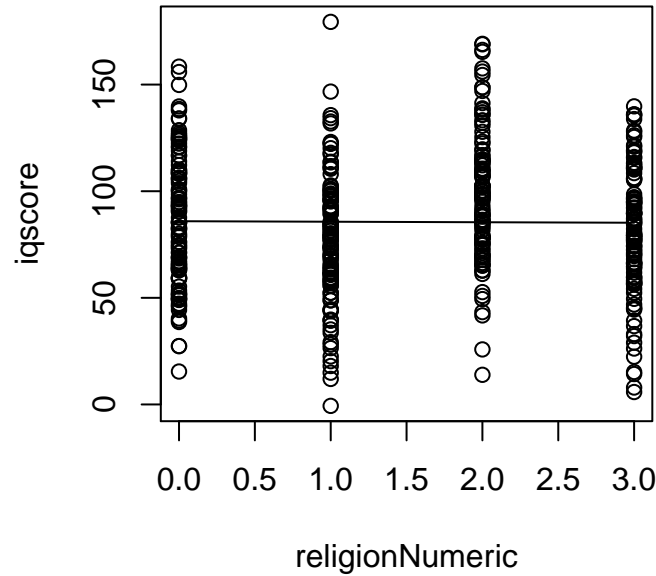
A box plot? Really? Discuss the strengths/weaknesses of this plot. If you have a better plot with which I can see the effect of religion, you should attach it and discuss its strengths.

If we were interested only in the effect of religion, then it might be best to use that box plot. This reminds us that “religionFactor” is a qualitative variable, and it discourages us from making the mistake of trying to interpolate values between the religions. There is literally no meaning whatsoever in the idea that we can go “halfway between the jewish and none categories” and the box plot reminds of that. Since we are interested also in the effect of substitute teaching, we can take another approach to illustrate the difference of the scores as a function of religion. In Figure 1, four regression lines—one for each religion—are presented to illustrate the effect of the usage of substitute teachers on the outcomes for all four religions. The regression coefficients for the religion contrasts have to be interpreted as “intercept shifters.”

3. (20pts) I recoded religion as a 0-1-2-3 integer-valued variable and estimated a regression.

$$iqscore_i = \hat{b}_0 + \hat{b}_1 \cdot mean(subs_i) + \hat{b}_2 \cdot religionNumeric_i \quad (1)$$

Here's my prediction:



a) From your regression estimates, please re-write my equation 1, inserting the estimated coefficients.

b) I'd like you to discuss your estimate of \hat{b}_2 . What does it mean? How meaningful is it?

The estimate for the effect of religion coded numerically is a very small number, -0.237. It is not statistically significantly different from 0. But let's proceed as if it were significant and interpret it literally. It indicates that the predicted value for religion 0 is

$$iqscore_i = 131.217 - 1.264 \cdot subs_i$$

For the religion coded 1, the predicted value is (assuming my calculator still works):

$$\begin{aligned} iqscore_i &= 131.2 - 1.264 \cdot subs_i - 0.237 \\ &= 130.63 - 1.264 \cdot subs_i \end{aligned}$$

For the religion coded 2, the predicted value is

Table 3: A Regression that Treats Religion Numerically

Variable	<i>OLS.Estimate</i> (<i>std.err.</i>)
Intercept	131.217*** (4.581)
Substitute Usage	-1.264*** (0.110)
religionNumeric	-0.237 (1.150)
R-squared	0.223
adj. R-squared	0.220
RMSE	27.643
F	66.128 ***
N	463

* $p \leq 0.05$

** $p \leq 0.01$

*** $p \leq 0.001$

$$\begin{aligned} \widehat{iqscore}_i &= 131.2 - 1.264 \cdot subs_i - 0.237 \cdot 2 \\ &= 130.40 - 1.264 \cdot subs_i \end{aligned}$$

We'd make the same calculation for religion 3, and then we could create Figure 2, which is similar in nature to Figure 1. The only difference is that all 4 religion lines are basically "on top" of each other, we can't see any difference among the religions.

- c) Compare this model with the "religionFactor" model discussed in the previous question. Is the prediction for any particular religion, say Jewish or Christian, changed significantly?

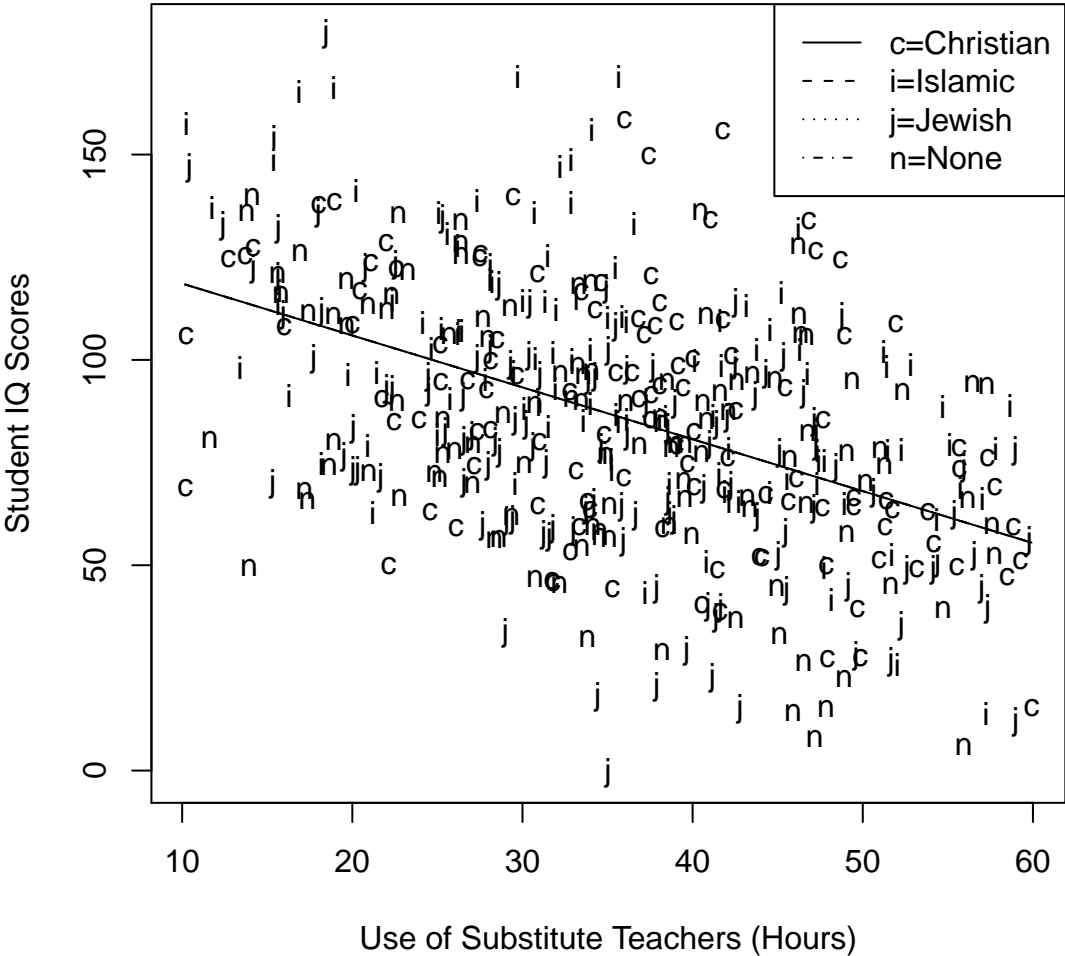
The numeric coding takes the alphabetically ordered categories "c", "i", "j", "n" and scores them 0, 1, 2, 3. This has the effect of enforcing the assumption that the scores of the students (by religious group) will fall in a given order. In this case, the negative coefficient means that we have enforced the requirement that Christians must have the highest score, and "second place" must go to Islamic students, "third place" to Jewish students, and lowest to nonreligious students. The factor coding does not enforce any apriori ordering. It allows the possibility that the Islamic student scores might be higher, which they are in my case.

- d) I'm determined to fit a model that has religion as a predictor. Should I use "religionFactor" or "religionNumeric" in my model? Is there a test that you would suggest that might help me to decide?

The numeric coding represents a simplification of the coding implied by the model in which religion is a factor. The contrasts estimated by the factor coding would appear as

$$\widehat{iqscore}_i = b_0 + b_1 subs_i + b_2 I_i + b_3 J_i + b_4 N_i$$

Figure 2: Student Performance with Religion as a Numeric Variable



where I_i , J_i , and N_i are the contrasts (0-1 indicator variables) for Islamic, Jewish and nonreligious students. The numeric coding approach enforces the requirement that those b 's are arranged in a particular order. The numeric coding assumes that there is only one "free parameter" b_2 and then the effects for the other religions are just multiples of that estimate:

$$b_3 = 2 \cdot b_2 \text{ and } b_4 = 3 \cdot b_2$$

The anova test indicates whether or not predictive power is lost by that simplification. In this case, the F test is used to compare the effect of the restriction on the error sum of squares from the 2 regressions. The estimated F value is 22.261, which is significant at the 0.001 level (with degrees of freedom 460 and 458.)

4. (10pts) I asked you to conduct t-test of hypothesis that average support for iraq war is different for men than women.

a) State the null hypothesis of your t-test. Does the test accept or reject your null hypothesis?

Let the mean score on the iraq variable for men be represented by $\hat{\mu}_{men}$ and refer to the true value of that mean be μ_{men} . For women, those values would be $\hat{\mu}_{women}$ and μ_{women} . The Null Hypothesis is that the true means are the same for the 2 sexes

$$H_0 : \mu_{men} = \mu_{women}$$

This may be restated in the form of a difference between the two means:

$$H_0 : \mu_{men} - \mu_{women} = 0$$

That representation is more desirable because it is closer in format to the t-test that we will actually conduct. We will conduct a test to see if the estimated difference between the sexes is different from 0 by calculating a t value:

$$t = \frac{\mu_{men} - \hat{\mu}_{women} - 0}{std.err(\mu_{men} - \hat{\mu}_{women})} = \frac{\hat{\mu}_{men} - \hat{\mu}_{women} - 0}{std.err(\mu_{men} - \hat{\mu}_{women})}$$

The difference between the 2 groups $\mu_{men} - \hat{\mu}_{women}$ is estimated by the difference between the 2 means $\hat{\mu}_{men} - \hat{\mu}_{women}$.

I did not know this during the semester, but during the final exam period I learned from one of the students that there is a formula interface for the t.test in R. This command

```
t.test(dat$iraq~dat$sex)
```

produces the same result as we would get if we split the variable iraq into 2 columns, one for men and one for women, and then estimated

```
t.test(dat$iraq.men, dat$iraq.women)
```

Either way, the estimated value of t is -1.6902, which is not statistically significantly different from 0. We are unable to reject the null hypothesis.

b) I was a little surprised that the R t.test output. It gives a t value, a confidence interval, but did not provide the standard error of the test, but if you are reading r-help, you might have noticed I found a way to calculate the standard error. You would need a calculator to actually figure out a value for the standard error, but it should be easy enough for you to write down a formula you could use if you did have a calculator:

If

$$t = \frac{\hat{\mu}_{men} - \hat{\mu}_{women} - 0}{std.err(\mu_{men} - \hat{\mu}_{women})}$$

we can calculate the standard error by

$$std.err(\mu_{men} - \hat{\mu}_{women}) = \frac{\hat{\mu}_{men} - \hat{\mu}_{women} - 0}{t}$$

c) Please attach your figure that compares the “sampling distribution” with the “Confidence Interval”. Write a paragraph to explain the difference in the meaning of those two constructs. How is the interpretation different? What light do they shed on your null hypothesis?

The Confidence Interval is centered on the estimated difference. We take into account $1.96 \cdot std.error$ units above and below. That is

$$CI = [estimate - 1.96 \cdot std.error, estimate + 1.96 \cdot std.error]$$

The R output indicates that it is $[-0.69515266, 0.05207988]$. In words, we believe with probability 0.95 that the actual value of the difference lies somewhere between those two values.

The sampling distribution of the difference in the means represents the abstract idea that, if we repeated this experiment many times, the differences between the two groups would be gathered around the 'true' value within a band that is $1.96 \cdot std.error$ units in radius. That is, supposing the true value is 0, we expect with probability 0.95 that the observed value will lie in this area.

$$= [0 - 1.96 \cdot std.error, 0 + 1.96 \cdot std.error]$$

If the estimate is NOT within this interval, then we conclude that the true value is unlikely to be 0.

Figure 3: Confidence Interval and Sampling Distribution

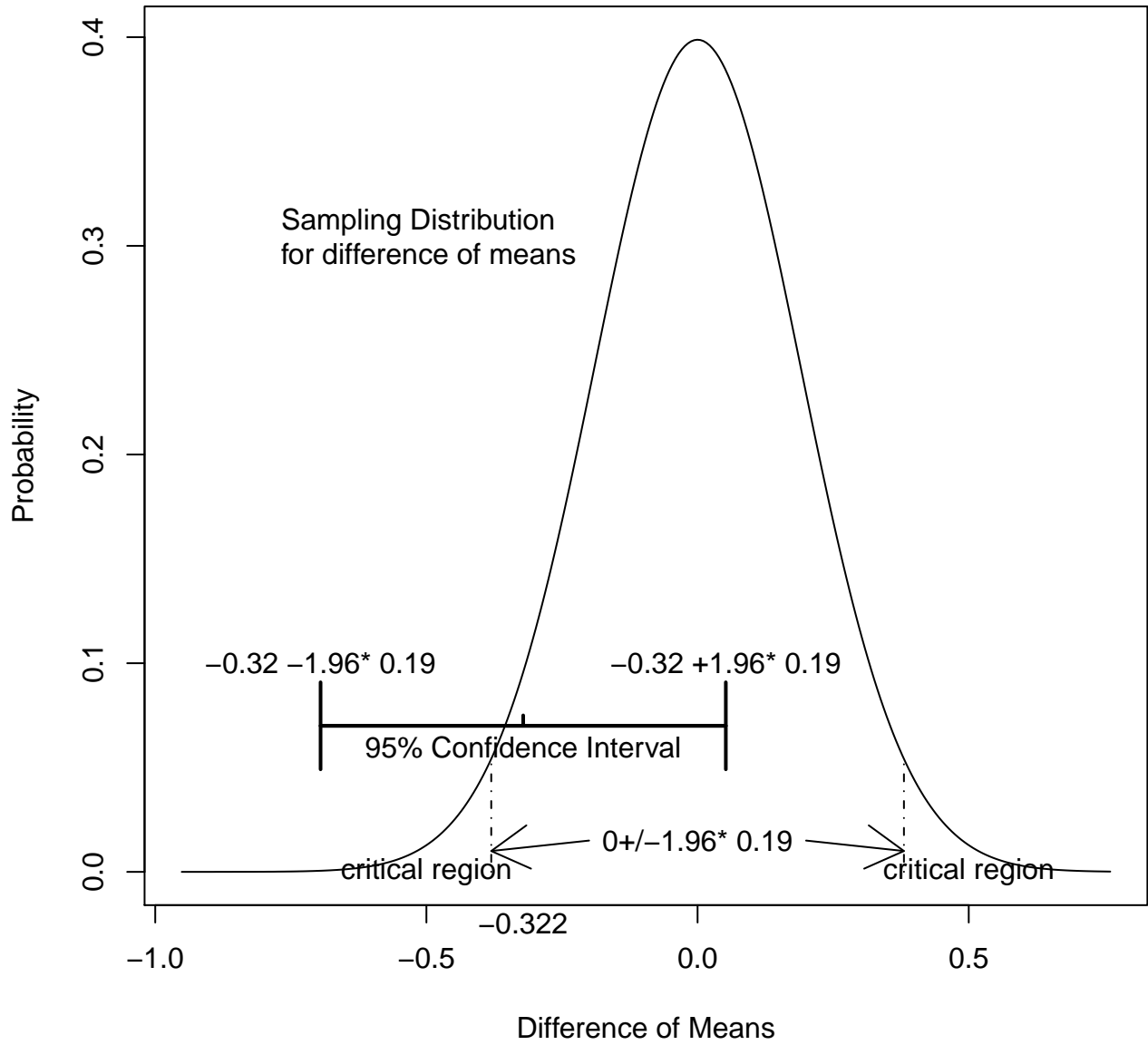


Table 4: Regression of Iraq War Support on Sex

Variable	<i>OLS.Estimate</i> (<i>std.err.</i>)
Intercept	12.612*** (0.135)
Sex:	
Male	0.322 (0.190)
R-squared	0.005
adj. R-squared	0.003
RMSE	2.330
F	2.857
N	600

*** $p \leq 0.001$

5. (15pts) I asked you to make a scatterplot to demonstrate this relationship and conduct a regression. Attach the table & figure.

a) Write down your theoretical model and an estimated model that represents your regression analysis:

$$IraqSupport_i = b_0 + b_1 \cdot male_i + e_i$$

$$Iraq\widehat{Support}_i = 12.61 + 0.32 \cdot male_i$$

b) How different is the predicted support for the war among men different from support among women? How can we read your table to find out?

My regression table is reported in Table 4. The variable "male_i" is coded 1 if a respondent is male, 0 otherwise. The estimated model indicates that the predicted value for respondents for whom male_i = 0 is 12.61. For the male respondents, the predicted value is 12.61+0.32=12.93.

c) Consider your plot. Do you think it is meaningful to draw a "regression line" on that figure? How else could you illustrate your predictions from the model.

I think that the best figure you can get is a box plot, which has sex on the horizontal axis, as in Figure 4.

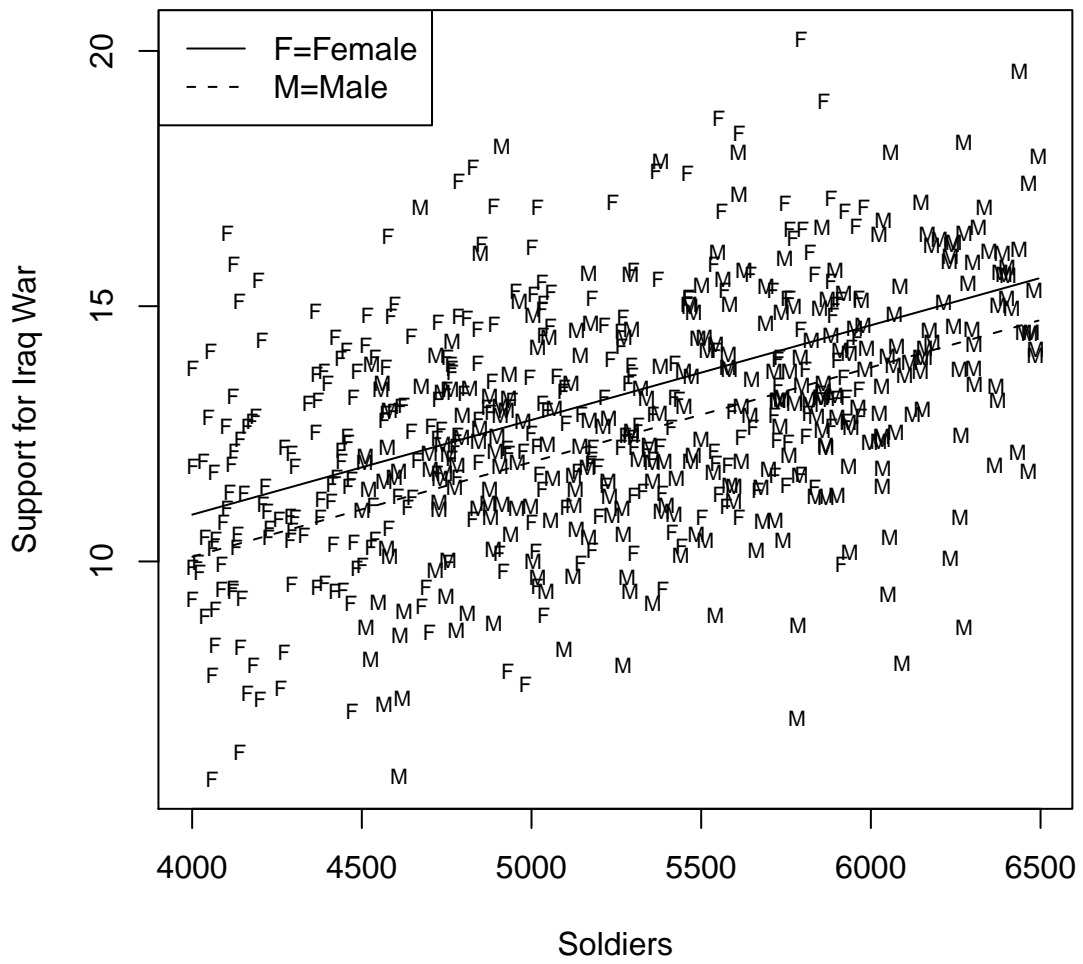
A regression "line" is not meaningful because it implies that the "male" variable could meaningfully take on values that are not exactly 0 or 1. There is no "meaning" in a value of male equal to 0.43, but drawing a line makes it seem as if that were meaningful.

It is better to represent the predicted values of categorical input variables by some other tool, such as a "dot chart" or a simple text marker superimposed on the box plot.

d) I wonder if the t-test you conducted in question 4 reaches a different conclusion from the regression analysis that you have just conducted. Is it different? Do you expect it should be?

The t-test for the difference of means reaches the same conclusion as the t-test for the difference between b_1 and 0 in the regression model. I expected it would

Figure 4: Sex and the Iraq War



because the two models are testing exactly the same estimated difference. Recall that the formula for the t test is generally

$$t = \frac{\text{estimate} - \text{null}}{\text{std.err.}}$$

The numerator in the t test for difference of means is actually identical to the difference observed in the regression equation. Hence, if there is going to be a different result from the 2 t-tests, it will have to be caused by a difference in the way the standard error is calculated in the 2 tests.

I am trying to get the R programmers to include the estimated standard error in the default output from `t.test` in order to facilitate this comparison. In my case, I have calculated the standard error from the difference of means test

$$\text{std.err}(\mu_{\text{men}} - \widehat{\mu}_{\text{women}}) = 0.1902373$$

which agrees almost exactly with the t value in the regression table for \hat{b}_1 .

I'm a bit frustrated by the details here. The estimates of the standard error are not exactly the same because of the way R currently conducts the defaults in `t.test`. Please read `?t.test`, where you see that the Welch approximation method has been used to calculate the standard error. Before 1990 or so, the standard statistical approach was to assume that the variances of the 2 groups being compared were equal. That would not require the Welch approximation. That is equivalent to the regression assumption of “homoskedasticity”, which we expressed as $E(e_i^2) = \sigma_e^2$. I believe that is what you get if you conduct this test:

```
myt <- t.test(dat$iraq~dat$sex,var.equal=T)
```

If you then examine `myt`'s contents, you find out that the degrees of freedom parameter is equal to the sample size minus 2:

```
> myt$parameter
df
598
```

On the other hand, if you allow R to use the Welch approximation to allow for the possibility of different variances in the 2 groups, you will get a different estimate for the standard error and an approximate degrees of freedom value:

```
> myt.welch <- t.test(dat$iraq~dat$sex)
> myt.welch$parameter
df
596.516
```

This non-integer value of the degrees of freedom is, well, hard to understand. It is not equal to “the number of cases minus 2.” Instead, because the Welch approach tries to account for the possible difference in variance between the 2 groups, it forces us to read from the t-table in a different row.

If the variances of the 2 groups really were different, then the regression model would have to be re-done. We would drop the homoskedasticity assumption, and we would allow the 2 variances to differ. The name for that type of analysis is “generalized least squares”.

Table 5: Regression with Sex and Soldiers

Variable	<i>OLS.Estimate</i> (<i>std.err.</i>)
Intercept	3.497*** (0.718)
Sex	
Male	-0.825*** (0.191)
Soldiers	0.00185*** (0.000144)
R-squared	0.221
adj. R-squared	0.219
RMSE	2.063
F	84.773 * **
N	600

* $p \leq 0.05$
** $p \leq 0.01$
*** $p \leq 0.001$

6. (15pts) I asked you to calculate a regression that predicts support for the war as a function of a variable called “soldiers” and “sex”. Attach your table and your figure. Then discuss these points.

a) In your table, there should be an estimated coefficient for “soldiers”. What is the effect of the soldier variable on support for the war? How should we interpret its standard error?

The results of the regression are presented in Table 5. The estimate for the Soldiers variable is 0.00185 and its standard error is 0.000144. Note that these values can be obtained from the summary function in R and that some table-generating programs, such as memisc’s mtable, will truncate the standard error at 0.

The estimate indicates that a 1 unit increase in the number of soldiers in Iraq from a person’s district will increase support for the war by 0.00185 units.

The standard error is very small compared to that estimated effect, so the null hypothesis that $b_2 = 0$ is easily rejected. A confidence interval can be obtained by the R function confint and the 95% interval for that estimate is: [0.001572141, 0.002137826]

b) In your Figure, you should have two lines, one for each sex. I’d like you to discuss the gap between those lines and interpret that gap in light of the numbers reported in your regression table. In particular, how far apart are the two lines and how confident are we about that difference?

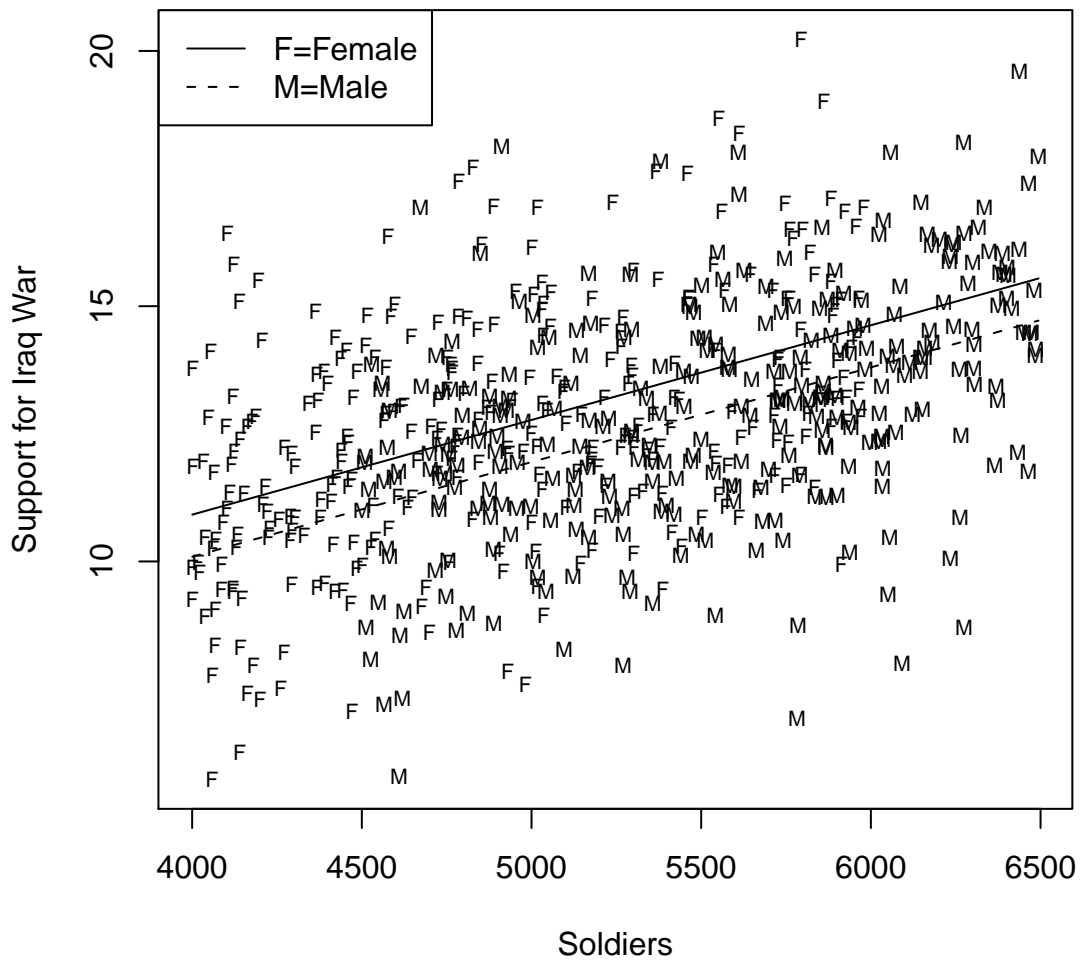
Consider Figure 5. The distance between the two lines is exactly equal to the estimate for the Sex variable in the regression model. Controlling for the “soldiers” variable, this model seems to indicate that the effect of being a male is to actually *reduce* support for the war in Iraq. Consider how peculiar and surprising this result is. When we compare men against women, it appears men

are more supportive of the war. When we include the effect of soldiers from one's own region, it appears men are less supportive.

- c) Your fellow classmates got better results than you did. I've checked and their estimates of R^2 are about 0.10 points higher than yours. Do you think your regression is worse because your R^2 is not as good? Why not?

I don't think it is correct to say one regression is "worse" than another. The R-square is simply an indication of the random error term's variance in comparison to the variance of the input variables. If the error's variance is small, then R-square will be high.

Figure 5: Home State Soldiers and War Support



7. (10pts) There's no doubt about it. Interpreting a logistic regression is a tongue twister. Attach your estimates from the fitted logistic model that predicts support for Obama. Attach whatever figures you think are helpful.
- a) Write anything you want that helps me understand the "meaning" of this model's parameter estimates. Don't worry about "deviance" or "AIC". Just focus on the estimates of the b's and their standard errors.

The theoretical model is that the probability of supporting Obama is

$$Prob(obama = 1 | sex_i, soldiers_i) = \frac{1}{1 + e^{-(b_0 + b_1 Male_i + b_2 Soldiers_i)}}$$

The estimates are presented in Table 6. These are Maximum Likelihood Estimates. They are not unbiased, but they are consistent and asymptotically Normally distributed.

Recall that logistic regression is thought of as a two step process. The estimated coefficients in the logistic regression are used to place the observations along a scale called "the linear predictor". That's the exponent in the denominator of the probability model. It indicates the tendency of a respondent to say "Yes". That linear predictor is then converted into a probability value between 0 and 1 by a logistic transformation.

Using the values in the table, the linear predictor formula would give us two parallel lines:

$$\text{For Females : } \hat{\eta}_i = -6.56 + 0.00114 \cdot Soldiers_i$$

$$\text{For Males : } \hat{\eta}_i = -6.94 + 0.00114 \cdot Soldiers_i$$

I've incorporated the difference between males and females, even though the standard error on that variable is on the large side and we can't reject the null hypothesis that the true effect is 0.

Those have to be transformed onto the probability scale:

$$Prob(obama_i = 1) = \frac{1}{1 + e^{-\hat{\eta}_i}}$$

The

- a) (5pts) Extra Credit. Present a plot that displays the predicted probabilities for men and women. It is extra extra good if you have "confidence intervals" in these plots, but these are not required to be extra good.

I made a couple of figures to illustrate the effects. This, again, supposes that we are treating males and females separately, even though this particular regression seems to say they are not statistically significantly different.

Consider 6. This shows the predicted probability that Obama is supported as a function of sex and the number of soldiers from the respondent's state.

There has been more emphasis lately on taking into account our uncertainty about those predictions. I've not really been bothered by it too much until lately. I used R's predict function with the se.fit option set to TRUE, so the output includes

Figure 6: Sex, Soldiers, and Obama

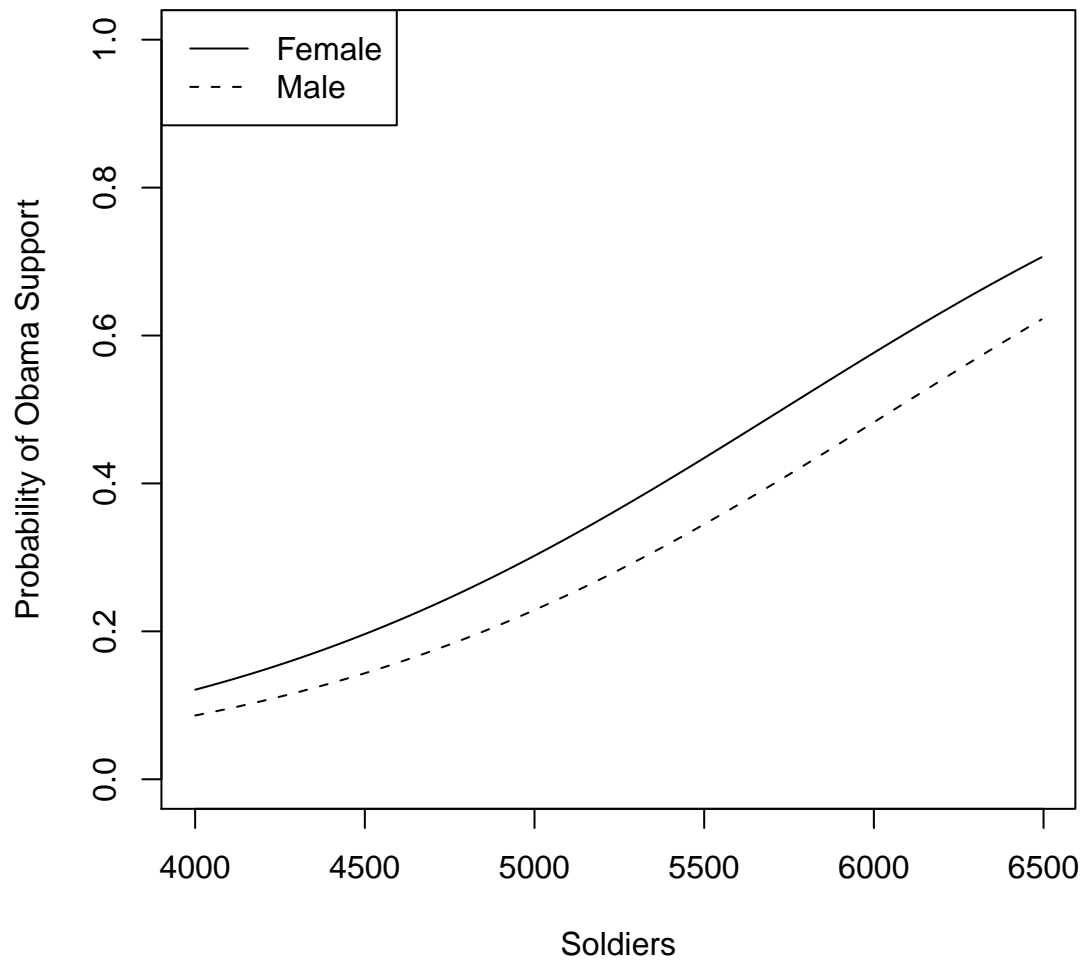


Table 6: Logistic Regression

Variable	<i>Maximum Likelihood Estimate</i> (<i>std.err.</i>)
(Intercept)	-6.566*** (0.836)
Sex:	
Male	-0.378 (0.205)
Soldiers	0.00114*** (0.000163)
Nagelkerke R-sq.	0.126
Likelihood-ratio (χ^2)	57.263 * **
Deviance ($-2LLR$)	706.554
AIC	712.554
N	600

* $p \leq 0.05$ ** $p \leq 0.01$ *** $p \leq 0.001$

both a fitted prediction and its standard error. We can use that standard error to construct a confidence interval around each predictive line.

This particular example is a bit disappointing because the two sexes are so similar. The 95% confidence intervals on the predicted values overlap very much. In Figure , the shaded confidence interval represents the prediction for females, while the dotted lines represent the predictions for makes with the 95% confidence intervals. The two intervals overlap from left to right, meaning that we would not really expect males and females to differentiate themselves. I know you are wondering “how did he get that shading that looks so great” and my answer is ?polygon.

Figure 7: Logistic Predictions with Confidence Intervals

