Psyc 2111, Jones

Summary of Lab Week 10

Regression

Regression tells us how one variable or set of variables predicts another. We call the variables that do the predicting the predictors, and we call the variable being predicted the outcome.

Load the data in ‘lab10-1.txt’. As always, start by getting a `summary()` of the data to see what the variables and their ranges are. Our goal with this dataset will be to see how well high school GPA and AP scores can predict a student’s college GPA. We do this using regression, specifically by regressing the outcome (college GPA) onto the three predictors. This will give us the best regression coefficient for each predictor.

Finding the best values of the regression coefficients mathematically involves some basic calculus and matrix algebra, but R has a built-in function that does this. The function is the linear-model function, abbreviated `lm()`. The input to `lm()` uses a special format: we write the outcome, followed by a tilde (~), and then the predictors separated by + signs.

```
> lm(d$collegeGPA ~ d$highschoolGPA + d$APcalc + d$APpsyc)
```

The output of `lm()` gives the regression coefficients. The first one is the intercept, which we call `b0`. This is the predicted value of the outcome when the predictors are all zero. The other regression coefficients each correspond to a predictor.

The regression coefficient for each predictor can be interpreted similarly to a correlation. If it is positive, then the predictor has a positive effect on the outcome. If the regression coefficient is negative, the predictor has a negative effect on the outcome. The size of the coefficient tells you the size of the effect. When it’s positive, the regression coefficient tells you how much the outcome is expected to increase for every unit increase in the predictor. When it’s negative, it tells you how much the outcome is expected to decrease for every unit increase in the predictor.

Once you have the regression coefficients, you can use them to make predictions about new subjects. Create scores for high school GPA, AP calc, and AP psych for a hypothetical new student. Use the summary of the data you got above to choose sensible values for all three predictors. Now use these scores to predict the new person’s college GPA. Your prediction comes from the regression equation, which multiplies each predictor by its regression coefficient (`b`) and then adds everything up including the intercept.

```
> collegeGPA.hat = b0 + b_hsGPA*hsGPA + b_APcalc*APcalc + b_APpysc*APpsyc
```

Now increase your new subject’s AP calc score by 1 and get a new prediction. Notice how the prediction increased by exactly the regression coefficient for AP calc. That’s the best way to understand what the regression coefficient means. Do the same thing with the other two predictors to see how changing them changes the prediction.

Reliability of regression coefficients

From looking at the regression coefficients, it appears that both AP scores are good predictors of college GPA, but high school GPA is not, because its coefficient is close to zero. Now we will do hypothesis testing to see whether our impressions based on the regression coefficients are correct. That is, we want to know whether the positive effects of the AP scores are indeed reliable, meaning that we can reject the null hypothesis that the true population values of these regression coefficients are zero. In the same way, we want to test whether the weak effect of high school GPA in our data is close enough to zero that it can be explained by chance, meaning that we would not reject the null hypothesis for this predictor.

To do hypothesis testing with regression, we use the `summary()` function. This is a very flexible function that gives different outputs for different kinds of inputs. If you input the results of a regression to `summary()`, it outputs a set of test statistics and p-values for all of the tests you might want to do with that regression.

```
> regression = lm(d$collegeGPA ~ d$highschoolGPA + d$APcalc + d$APpsyc)
> summary(regression)
```

There's a lot here, and we'll go through most of it. First, look at the output under `Coefficients`. There's one row for each regression coefficient, including the intercept. The Estimate column shows the values of the regression coefficients estimated from the data. There are the same values we saw already. The remaining columns test the reliability of each regression coefficient.

The `Std. Error` column gives the standard error of each coefficient. This standard error works in the same way as other standard errors—it's the standard deviation of the sampling distribution for the regression coefficient, and it tells you the typical difference between the true population value and the value estimated from the sample (if you drew hypothetical samples repeatedly). If the estimated value of a regression coefficient is much larger (in absolute value) than its standard error, then it's unlikely that the true value is zero. This is because, if zero were the true value, then you probably wouldn't have gotten a value as extreme as the one you actually got.

The next column shows the value of the `t` statistic, which is equal to the regression coefficient divided by its standard error. Try this out just to verify it—divide one of the regression coefficients by its standard error and check that the result is the `t` value.

Notice that the `t` values for both AP scores are very large. A `t` value this large will almost never happen by chance. We can see this from the final column, which shows the p-value. R gives you the two-tailed p-value, which is the probability of getting a result that is as large in absolute value as the result you got. For the AP scores, the p-value is very small (`e-16` stands for `10` to the power `-16`, i.e. 16 places after the decimal point). This reflects the fact that both `t`s are very far out in the t distribution and are thus very unlikely. Because `p` is so small in both cases, we reject the null hypotheses that the true regression coefficients are zero and conclude that the AP scores both reliably predict college GPA.

For high school GPA, the estimated regression coefficient is actually smaller than its standard error. This means that the estimate is consistent with what would be expected just by chance, i.e. if the true coefficient were zero. This is reflected in the `t` value, which is small. The small `t` leads to a large p-value, which indicates there was a high probability of getting a result like the one we got, just by chance. Therefore we retain the null hypothesis that high school GPA doesn't give any information about college GPA, once we know the AP scores.

Reliability of entire regression

In addition to testing individual predictors, we can test whether the whole regression equation tells us anything meaningful about the outcome. We know this is true because we already know that two of the three predictors explain something meaningful about the outcome, but let's pretend we didn't know this.

The first step in evaluating the regression is to partition the variability in the sample outcome into variability that's explained and variability that is not. First, compute the total variability. This is the sum of squares for the outcome, or the total squared deviations of all the college GPA scores from the mean college GPA.

```
> deviations = d$collegeGPA - mean(d$collegeGPA)
> squared.deviations = deviations^2
> SS.collegeGPA = sum(squared.deviations)
```

The next step is to find the residual variability, meaning the uncertainty in the outcome that's left over after we use the regression equation to get our best predictions of the outcome. To do this, you could use the regression equation to find the prediction for every subject, just like you did above for a single hypothetical new subject. Then you would subtract each prediction from the corresponding true college GPA to find the residuals, meaning the prediction error for all subjects in your sample. Fortunately, `lm()` computes the residuals automatically and saves them as part of the regression you created above.

```
> summary(regression$residuals)
```

There's one residual for every subject, and it's equal to $Y - \hat{Y}$, meaning the prediction error of the regression. To find the total residual variability, square the residuals and add them up. This computes $SS_{\text{residual}} = \sum\left(Y - \hat{Y}\right)^2$.

```
> SS.residual = sum(regression$residuals^2)
```

`SS.residual` is the amount of variability that the regression cannot explain. The amount that the regression can explain is the difference, or reduction, from `SS.collegeGPA` to `SS.residual`. This reduction is the amount of uncertainty that went away when we changed from using the mean to predict the outcome for each subject (in `SS.collegeGPA`) to using $\hat{Y}$ (in `SS.residual`).

```
> SS.regression = SS.collegeGPA - SS.residual
```

The question now is whether `SS.regression` represents a large portion of `SS.total`. If it does, then the regression is explaining a significant amount of the variability in the outcome. To test this, we divide `SS.regression` by

SS.total, to find the fraction of the variability that the regression explains. This fraction is R-squared.

```
> R2 = SS.regression/SS.collegeGPA
```

Compare your result for R-squared to the result you got before from the summary of the regression, which is shown as "Multiple R-squared." The numbers should be the same, because summary() did exactly the same calculations you just did.

R-squared is fairly large for these data—over 89% of the variability in the outcome is explained by the regression. It's very unlikely that so much variability could be explained just by chance, but let's do the hypothesis test to be sure. To do the hypothesis test, we convert SS.regression and SS.residual to mean squares, by dividing by their degrees of freedom. The residual degrees of freedom is the number of subjects minus the number of regression coefficients (which is the number of predictors plus one for the intercept). The degrees of freedom for the regression is the number of predictors (not including the intercept). We use n for the number of subjects (which you can get from dim(d)) and m for the number of predictors (which is 3 in this case).

```
> df.residual = n – (m+1)
> MS.residual = SS.residual/df.residual
> df.regression = m
> MS.regression = SS.regression/df.regression
```

According to the null hypothesis that the regression doesn't explain anything meaningful, MS.regression should be about the same as the population variance of the outcome, σ^2. We estimate σ^2 using MS.residual. By comparing MS.regression to MS.residual, we can see whether MS.regression is larger than it would be by chance. The ratio of MS.regression and MS.residual gives the F statistic, which is our test statistic for this hypothesis test.

```
> F = MS.regression/MS.residual
```

Verify that your result matches the F given by summary(). Notice that F is very large. F statistics are always positive, but by chance they tend to be around 1. The probability of an F as large as 2760 is extremely small, as shown by the p-value given by summary(). You can also get the critical value for F. The qf() function works just like qt() and qnorm(), except that you have to tell it both degrees of freedom.

```
> F.crit = qf(.05,df.regression,df.residual,lower.tail=FALSE)
```

We always do a 1-tailed test with F statistics, because a particularly small value of F never means anything interesting. So we test whether F is bigger than the critical value. In this case, the critical value should be much less than the actual F from the data. This means that we reject the null hypothesis and conclude that the regression is telling us something meaningful about the outcome. In other words, knowing someone's high school GPA and calculus and psychology AP scores gives you real information about what their college GPA is likely to be.