§2 The Analysis of Variance Approach to SLR

Objectives: Students will be able to

1. Define the following terms: Total Sum of Squares (SSY), Corrected Total Sum of Squares, Sum of Squares Regression (SSR), Sum of Squared Errors (SSE), Mean Squares Regression, Mean Squared Error.
2. Compute each of the above manually.
3. Explain what is meant by 'Analysis of Variance' (ANOVA)
4. Manually construct/complete an ANOVA table for an SLR model.
5. Define R^2, compute its value, and describe what the value means.
6. Perform the overall F-test manually.
7. Describe the relationship between the following tests: (i) t-test for the correlation, (ii) t-test for the slope coefficient in SLR, and (iii) the overall F-test in SLR.
8. Explain why SLR can be thought of as another way of estimating the mean value of the dependent variable.

§2.1 Sums of Squares

So far we have thought of simple linear regression in terms of finding the straight line that best describes the linear relationship between X and Y. We can also think of the SLR problem as follows:

- There is a certain amount of variation in the dependent variable, Y. We need to quantify the amount of variation that is being observed in Y, and then figure out *why* Y varies by as much as it does.
- Our hypothesis is that Y is linearly associated with X, and that this association is responsible for at least part of the observed variation in Y.
- If we can estimate the linear association, we will be able estimate just how much of the variation in Y can be explained by its association with X (i.e, attributed to the relationship that Y has with X).
- Of course, in real life, we will not able to explain away *all of* the variation we observe in Y --some of that variation will inevitably be left unexplained.
- Putting the ideas above into a formula, we have:

  Total variation observed in Y = Variation in Y that is explained by its relationship with X + Variation in Y that is not explained by its relationship with X

In the linear regression ANOVA Table, each of these three components quantified as follows:

SSY = SSR + SSE, where

SSY = 'Sum of Squares, Y' or 'Total Sum of Squares' (a measure of total variation in Y)
=

SSE = 'Sum of Squares, Error' or 'Sum of Squared Errors' (a measure of the variation unexplained by the linear regression)
=

SSR = 'Sum of Squares, Regression' (a measure of the variation explained by the lin. regression)
= SSY - SSE

The graph below shows this idea for a single data point.

§2.2 The ANOVA Table for Simple Linear Regression

The simple linear regression Analysis of Variance (ANOVA) Table provides a summary of the sum of squares information, and also leads to a useful hypothesis test:

Source of Variation	Degrees of Freedom	Sum of Squares	MS	F
Regression (X)	1	SSR=SSY - SSE	MSR=SSR/1	MSR/MSE
Residual	n-2	SSE	MSE=SSE/n-2	
Total	n-1	SSY		

where

MSR= mean-square regression (the average variation in Y explained or accounted for by the regression, per degree of freedom), and

MSE= mean -square error (the average variation in Y left unexplained, per degree of freedom)

Two additional statistics are usually derived from the sums of squares and the mean squares:

1) "R-square" or "Coefficient of Determination" or just R^2 .
   R^2 = SSR/SSY = proportion of the total variation (SSY) that was explained by the regression.
   Higher R^2 values mean that more of the variation in Y is explained by its relationship with X (people

often say that large R^2 means that the model "fits" the data well).
Two notes:

- In SLR, R^2 actually turns out to be the square of the sample correlation coefficient, r. Verify this for the sodium intake-blood pressure example!
- Be careful--R^2 can be artificially inflated in multiple linear regression models (models with more than 1 predictor) simply by adding lots of predictors to the model, even when the predictors are not strongly related to Y. In multiple linear regression models, we report an 'adjusted R^2', which includes a penalty term that shrinks the R^2 value if predictors are included which do not have a good linear relationship with Y. (More on this later).

2) The value in the F column is the test statistic for the "overall test" for the regression. The null hypothesis in the overall test is "there is no linear association between X and Y" (alternative hypothesis: "there *is* a linear relationship"). In simple linear regression, the overall test is equivalent to a test of H_0: $\beta_1 = 0$ vs. H_A: $\beta_1 \neq 0$. The F statistic follows an $F_{1, n-2}$ distribution under the null hypothesis. Here is the general form of the test:

   H_0: No linear association between X and Y (i.e. regression is not significant)
   H_A: There *is a* linear association between X and Y (i.e. regression *is* significant)

   or, equivalently,

   H_0: $\beta_1 = 0$
   H_A: $\beta_1 \neq 0$

   Test Statistic: $F = MSR/MSE = (SSR/1) / (SSE/(n-2))$, $F \sim F_{1, n-2}$

   P-value = $Pr(F > MSR/MSE)$ where $F \sim F_{1, n-2}$. (Note, the P-value is not doubled. Why not?)

   Conclusion: Same as conclusion for test on the slope (since, in SLR, the overall test is the same as the test on the slope).

   As you can see, the test statistic is the ratio of the average 'explained' variation, per degree of freedom, with the average 'unexplained' variation, per degree of freedom. The larger the F-test statistic, the greater the variation in Y that is explained by our model, relative to the variation left unexplained; Therefore, large F-test statistics imply that there *is* a significant linear association between Y and X; how large is 'large enough' to reject the null hypothesis is determined by computing the p-value for the test, and comparing that p-value against the pre-set significance level.

Practice with the F tables:

Example: SAS and R ANOVA tables for the sodium intake - blood pressure example. The SAS code shown on page 9 produces the complete ANOVA table by default (shown below). The R code from page 9 produces certain elements of the table.

SAS:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	3136.87719	3136.87719	79.46	<.0001
Error	10	394.78947	39.47895		
Corrected Total	11	3531.66667			

R:

```
Residual standard error: 6.283 on 10 degrees of freedom
Multiple R-squared:  0.8882,    Adjusted R-squared:  0.877
F-statistic: 79.46 on 1 and 10 DF,  p-value: 4.511e-06
```

Students: practice calculating the statistics in the ANOVA table manually –without the use of any software.

§2.3 SLR: Additional Notes

1) A strategy for SLR analysis:
   i. State the SLR model.
   ii. Perform a descriptive statistical analysis for Y and X.
   iii. Plot Y vs X. Inspect plot for linear pattern.
   iv. Calculate and characterize r. Interpret.
   v. Fit the SLR model.
   vi. Check for validity of SLR assumptions *.
   vii. Perform the overall F-test. If not significant, report this finding and stop working with the current model --it is not significant; do not interpret the slope estimate (the failure to reject the null means that you have not been able to show that the slope is anything other than zero).
   If significant, proceed with additional inferences/interpretations (below).
   viii. Present estimate of the model. Determine a confidence interval for the slope, and present and interpret it.
   ix. Proceed with additional model inferences, if appropriate and wanted: test on Y-intercept, prediction intervals, confidence intervals on μ_Y.

*We don't have all the necessary tools yet to check the assumptions; so we will skip this step *for now*. We will acquire those tools later.

2) The SLR line and the Correlation Coefficient:
   - The magnitude of r does not provide a measure of the slope of the regression line.
   - The magnitude of r does not indicate whether or not the straight line model is the best model to consider.

3) SLR vs $\bar{y}$:
   Last semester, if you had been presented with data on a variable, Y, and had been asked to estimate its average, you would probably have calculated $\bar{y}$ as your estimate. This semester, we are assuming that the average of Y depends linearly on X, and so we fit a SLR model to the sample data in an effort to learn about the average value of Y.

   When we perform the overall test in SLR (i.e. the test for the significance of the slope), we are actually comparing the usefulness of our SLR model in predicting the average value of Y against the crude approach where you simply calculate $\bar{y}$ as an estimate of the average value of Y. If we fail to reject the null hypothesis of a zero slope, that means that the SLR approach is no more useful than the crude approach.

4) At this point, we have studied three ways to test whether or not there is a linear association between Y and X (below):

   <u>T-test for the sample correlation:</u>
   H_0: $\rho_{yx} = 0$ vs. H_A: $\rho_{yx} \neq 0$ where ρ_{yx} is the true correlation between Y and X. (See page 5 of our lecture notes for details).

   <u>T-test for the SLR slope:</u>
   H_0: $\beta_1 = 0$ vs. H_A: $\beta_1 \neq 0$ where β_1 = true slope of straight line that is assumed to relate X and Y. (See page 13 of our lecture notes for details).

Overall F-test for SLR:

$H_0: \beta_1 = 0$ vs. $H_A: \beta_1 \neq 0$, with test Statistic $F = MSR/MSE = (SSR/1) / (SSE/(n-2))$, $(F \sim F_{1, n-2})$

It can be proven, mathematically, that the T-test statistics will be equal to each other, and that squaring them will yield the F-test statistic value. The p-values for all three will be identical, in simple linear regression. The latter fact should make sense on an intuitive level, because in each case the hypotheses are actually the same: H_0: there is no linear association between X and Y vs. H_A: there is a linear association between X and Y.