Summary of Ch 11

**Multiple Linear Regression:**

For \( Y = X \beta + \epsilon \), where \( \hat{y}_i = \hat{\beta}_0 x_{i0} + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik} \) is our estimator for \( E(Y_i) \)

We have:

\[
\hat{\beta} = (X^T X)^{-1} X^T Y
\]

\[
\text{Cov}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}
\]

\[
SSE = Y^T Y - \hat{\beta}^T X^T Y
\]

\[
S^2 = \frac{SSE}{n-(k+1)}
\]

**Simple Linear Regression:**

For \( y = \beta_0 + \beta_1 x + \epsilon \), where \( \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \) is our estimator for \( E(Y_i) \) (i.e., \( k=1 \) & \( x_{i0} = 1 \))

We have:

\[
\hat{\beta} = \left[ \sum_{i=1}^{n} x_i^2 \right]^{-1} \left[ \sum_{i=1}^{n} x_i y_i \right] = \frac{1}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \left[ \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{bmatrix} \right] \left[ \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix} \right]
\]

\[
\text{Cov}(\hat{\beta}) = \sigma^2 \begin{bmatrix} \frac{\sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} & -\frac{\sum_{i=1}^{n} x_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \\ -\frac{\sum_{i=1}^{n} x_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} & \frac{n}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \end{bmatrix} = \sigma^2 \begin{bmatrix} C_{00} & C_{01} \\ C_{10} & C_{11} \end{bmatrix} = \begin{bmatrix} \text{V}(\hat{\beta}_0) & \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) \\ \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) & \text{V}(\hat{\beta}_1) \end{bmatrix}
\]

\[
SSE = \sum_{i=1}^{n} y_i^2 - \hat{\beta}^T \left[ \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix} \right]
\]

\[
S^2 = \frac{SSE}{n-2}
\]

Hypothesis testing: \( T = \frac{\hat{\beta}_i - \beta_{0i}}{S \sqrt{C_{ii}}} \) has a \( T \) distribution with \( n-2 \) df.
Recall for large samples, we can estimate \( \mu_1 - \mu_2 \) using

\[
Z = \frac{( \bar{Y}_1 - \bar{Y}_2 ) - ( \mu_1 - \mu_2 )}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}}
\]

What if we have small samples instead? Not technically ok to plug in

\[
\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}
\]

because 1) we can plug in \( S^2 \) for \( \sigma^2 \) when \( n \) is large
2) \( T_{n-1} = \frac{\bar{Y}_1 - \mu}{S/\sqrt{n}} \), so if we plug in \( S \), is this \( T \)?

**Solution:** Assume \( \sigma_1^2 = \sigma_2^2 = \sigma^2 \)

Then

\[
Z = \frac{( \bar{Y}_1 - \bar{Y}_2 ) - ( \mu_1 - \mu_2 )}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}
\]

Since \( \sigma^2 \) is unknown, need to find estimator for \( \sigma^2 \). (Perhaps some \( \Phi \) of \( S_1^2 \) or \( S_2^2 \)?)

Let \( Y_{11}, Y_{12}, \ldots, Y_{n_1} \) and \( Y_{21}, Y_{22}, \ldots, Y_{n_2} \) be independent random samples of sizes \( n_1 \) & \( n_2 \) from normal populations.

Let \( \bar{Y}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} Y_{1i} \) & \( \bar{Y}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} Y_{2i} \)

Then the pooled estimator \( S^2 = \frac{\sum_{i=1}^{n_1} (Y_{1i} - \bar{Y}_1)^2 + \sum_{i=1}^{n_2} (Y_{2i} - \bar{Y}_2)^2}{n_1+n_2-2} \)

\( = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} \)

Notice \( S^2 \) is a weighted average of \( S_1^2 \) & \( S_2^2 \)

Then

\[
T = \frac{( \bar{Y}_1 - \bar{Y}_2 ) - ( \mu_1 - \mu_2 )}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{with} \quad n_1 + n_2 - 2 \quad \text{d.f}
\]

**Proof:** Let \( W = (n_1+n_2-2)S^2 \)

\[
W = \frac{\sum_{i=1}^{n_1} (Y_{1i} - \bar{Y}_1)^2 + \sum_{i=1}^{n_2} (Y_{2i} - \bar{Y}_2)^2}{\sigma^2} = \chi^2_{n_1-1} + \chi^2_{n_2-1}
\]

Recall MGF for \( \chi^2_n = \frac{1}{(1-2\psi)^{n/2}} \) & when you add RVs, the MGFs multiply (from Theorem 6.2)

So we have

\[
W = \frac{1}{(1-2t)^{n_1+n_2-2}} = \chi^2_{n_1+n_2-2}
\]

So \( W \) has a \( \chi^2 \) dist with \( n_1-1 + n_2-1 = n_1 + n_2 - 2 \) d.f.

Then

\[
T = \frac{Z}{\sqrt{\frac{W}{n_1+n_2-2}}} = \frac{( \bar{Y}_1 - \bar{Y}_2 ) - ( \mu_1 - \mu_2 )}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{with} \quad n_1 + n_2 - 2 \quad \text{d.f}
\]
13.1-13.2 The Analysis of Variance Procedure

ANOVA: analysis of variance

* attempts to analyze the variation in a set of responses and assigns portions of this variation to each variable in a set of independent variables
* objective: identify important independent variables and determine how they affect the response
* Partitions \( \sum_{i=1}^{k} (y_{i} - \bar{y})^2 \) (called total sum of squares) into parts, each of which is attributed to one of the independent variables in the experiment, plus a remainder associated with random error
* Each of the pieces of the total sum of squares (divided by appropriate constant) provides an independent and unbiased estimator of \( \sigma^2 \)
* When a variable is highly related to the response, its portion of the total sum of squares will be inflated
* can be detected by comparing the sum of squares for the variable with the sum of squares for error (SSE)

Method: Suppose we want to compare the means of two normally distributed populations (of sample size \( n_1 = n_2 \)) with means \( \mu_1 \) and \( \mu_2 \) and with equal variances \( \sigma_1^2 = \sigma_2^2 = \sigma^2 \)

* Previously: used T test for \( \bar{Y}_1 - \bar{Y}_2 \)
* Now: Look at SS:

\[
\text{Total } \text{SS} = \sum_{i=1}^{2} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2 = \sum_{i=1}^{2} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2
\]

where \( Y_{ij} \) is \( j \)th observation in \( i \)th sample and \( \bar{Y} \) is mean of all \( n = 2n_1 \) observations

Can rewrite this as (proof in Sec 13.6):

\[
\underbrace{n_1 \sum_{i=1}^{2} (\bar{Y}_i - \bar{Y})^2}_{\text{call this } \text{SST} \text{ treatment}} \quad \quad \underbrace{\sum_{i=1}^{2} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2}_{\text{call this } \text{SSE} \text{ error}}
\]

where \( \bar{Y}_i \) = average of observations in \( i \)th sample for \( i = 1, 2 \)

First examine \( \text{SSE} = \sum_{i=1}^{2} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 = \sum_{i=1}^{2} (n_i - 1) S_i^2 = (n_1 - 1) S_1^2 + (n_2 - 1) S_2^2 \)

where \( S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 \)

Recall pooled estimator \( S_p^2 = \frac{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2}{n_1 + n_2 - 2} = \frac{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2}{n_1 + n_2 - 2} = \frac{\text{SSE}}{n_1 + n_2 - 2} \)
Next examine \( \text{SST} = n_1 \sum_{i=1}^{n_1} (\bar{Y}_{i} - \bar{Y})^2 = \frac{n_1}{2} (\bar{Y}_1 - \bar{Y}_2)^2 \)

This will be large if \( |\bar{Y}_1 - \bar{Y}_2| \) is large, hence, the larger \( \text{SST} \) is, the greater will be the weight of evidence to indicate a significant difference between \( M_1 \) & \( M_2 \).

Q: When will \( \text{SST} \) be large enough to indicate a significant difference between \( M_1 \) & \( M_2 \)?

Recall we assumed \( \bar{Y}_{ij} \) normally dist w/ \( E(\bar{Y}_{ij}) = M_i \) for \( i = 1, 2 \) and \( \text{VC}(\bar{Y}_{ij}) = \sigma^2 \) & \( \text{SSE}/(2n_1-2) = S_p^2 \) (unbiased).

So \( E\left(\frac{\text{SSE}}{2n_1-2}\right) = \sigma^2 \)

Thus \( \frac{\text{SSE}}{\sigma^2} = \sum_{j=1}^{n_1} \frac{(\bar{Y}_{ij} - \bar{Y})^2}{\sigma^2} + \sum_{j=1}^{n_1} \frac{(\bar{Y}_{ij} - \bar{Y}_j)^2}{\sigma^2} \) has a \( \chi^2 \) dist with \( 2n_1-2 \) df.

In Sec 13.6, obtain \( E(\text{SST}) = \sigma^2 + \frac{n_1}{2} (M_1 - M_2)^2 \)

\( \rightarrow \text{SST} \) estimates \( \sigma^2 \) if \( M_1 = M_2 \) and something > \( \sigma^2 \) if \( M_1 \neq M_2 \).

**Hypothesis testing:**

\( H_0: M_1 = M_2 \)

\( H_a: M_1 \neq M_2 \)

Then \( Z = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{2\sigma^2/n_1}} \) has a stand normal dist (Why? \( \bar{Y}_{ij} \) normally dist, \( M_1 - M_2 = 0 \),

and \( \text{VC}(\bar{Y}_1 - \bar{Y}_2) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_1} = 2\frac{\sigma^2}{n_1} \))

Thus \( Z^2 = \left( \frac{n_1}{2} \right) \left( \frac{(\bar{Y}_1 - \bar{Y}_2)^2}{\sigma^2} \right) = \frac{\text{SST}}{\sigma^2} \) has a \( \chi^2 \) dist w/ 1 df.

Note \( \text{SST} \) is a function of only \( \bar{Y}_1 \) & \( \bar{Y}_2 \), whereas \( \text{SSE} \) is a function of only the sample variances \( S_1^2 \) & \( S_2^2 \). Since \( \bar{Y} \) & \( S^2 \) are independent, then \( \text{SST} \) & \( \text{SSE} \) are independent.

Recall \( \frac{\chi^2}{n} = \frac{\chi^2/n_1}{\chi^2/n_2} \) mean squares

Thus \( \frac{\text{SST}}{\sigma^2 / 1} = \frac{\text{SSE}}{\sigma^2 / (2n_1-2)} \)

\( \frac{\text{SSE}}{\text{SSE} / (2n_1-2)} \)

\( \frac{\text{SST}}{\sigma^2 / 1} \)

Thus \( \frac{\text{MST}}{\text{MSE}} \) = \( \frac{\text{SST}}{\text{SSE} / (2n_1-2)} \)

Under null hypothesis \( H_0: M_1 = M_2 \), both \( \text{MST} \) & \( \text{MSE} \) estimate \( \sigma^2 \). However, when \( H_0 \) is false, \( \text{MST} \) estimates something larger than \( \sigma^2 \) & tends to be larger than \( \text{MSE} \).

Thus to test \( H_0: M_1 = M_2 \) versus \( H_a: M_1 \neq M_2 \) use \( F = \frac{\text{MST}}{\text{MSE}} \)

Rejection region: \( F > F_{2, \alpha} \), one-tailed \( F \) test.

**Important note:** This is equivalent to the two tailed \( T \) test of Ch. 10.

Why bother with all this? Because \( F \) test generalizes to allow for comparison of any number of treatments! (Not just 2!) Subject of Sec. 13.3.
13.3-13.4 Comparison of More than Two Means

Now we want to generalize to the following hypothesis testing:

$H_0: \mu_1 = \mu_2 = \ldots = \mu_k$

$H_a$: at least one of these equalities does not hold

Then Total $SS = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2$

This can be rewritten as $SS = \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}^2 - CM$

where $CM = \frac{(\text{total of all observations})^2}{n} = \frac{1}{n} \left( \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij} \right)^2 = n \bar{Y}^2$

Since we consider general case where sample sizes $n_i$ may be unequal, we have:

$SST = \sum_{i=1}^{k} n_i \left( \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} - \bar{Y} \right)^2$

Sample mean of $i^{th}$ pop = $\bar{Y}_i$

$SSE = \text{Total SS} - SST \; \& \; \text{easier to compute}$

Since the actual expression for $SSE$ is $\sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij})^2$

Sample mean of $i^{th}$ pop = $\bar{Y}_i$

Then again $\frac{SSE}{\sigma^2}$ is a $\chi^2$ distribution, but this time

with $n_1 - 1 + n_2 - 1 + \ldots + n_k - 1 \; \text{df} = n_1 + n_2 + \ldots + n_k - k = n - k \; \text{df}$

As in Sec 13.2, can rewrite SST so that it is only a function of sample means of $i^{th}$ population: $\frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} = \bar{Y}_i$

This is because $\bar{Y} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij} = \frac{1}{n} \sum_{i=1}^{k} n_i \bar{Y}_i = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}$

Sample mean of $i^{th}$ pop = $\bar{Y}_i$

This time, $\frac{SST}{\sigma^2}$ can be written in term of $k-1$ quantities $(\bar{Y}_i - \bar{Y})^2$

So $\frac{SST}{\sigma^2}$ has a $\chi^2$ distributions with $k-1 \; \text{df}$

Thus $\frac{SST}{\sigma^2} / (k-1) = \frac{MST}{SSE/(n-k)} = \frac{F}{n-k}$

Again, under null hypothesis, both $\text{MST} \& \text{MSE}$ estimate $\sigma^2$. However, when $H_0$ is false, $\text{MST}$ estimates something larger than $\sigma^2$ & tends to be larger than $\text{MSE}$

Thus to test $H_0 \; \text{vs} \; H_a$, use $F = \frac{\text{MST}}{\text{MSE}}$, and rejection region is $F > F_{k-1, n-k}$, one-tailed $F$ test
Ex 23: Four groups of students were subjected to different teaching techniques and tested at the end of a specified period of time. Do the data present sufficient evidence to indicate a difference in mean achievement for the four teaching techniques?

(This is Ex 13.2 in textbook)

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>65</td>
<td>75</td>
<td>59</td>
<td>94</td>
<td>454</td>
</tr>
<tr>
<td>87</td>
<td>69</td>
<td>78</td>
<td>89</td>
<td>549</td>
</tr>
<tr>
<td>73</td>
<td>83</td>
<td>67</td>
<td>80</td>
<td>425</td>
</tr>
<tr>
<td>79</td>
<td>81</td>
<td>62</td>
<td>88</td>
<td>351</td>
</tr>
<tr>
<td>81</td>
<td>72</td>
<td>83</td>
<td>76</td>
<td>6</td>
</tr>
<tr>
<td>69</td>
<td>79</td>
<td>76</td>
<td>90</td>
<td>67</td>
</tr>
</tbody>
</table>

We have \( H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 \)

\( H_1: \) one of these equalities does not hold

\[ CM = \frac{1}{23} (454 + 549 + 425 + 351)^2 = 137,601.8 \]

\( \bar{\mu} = \frac{1}{23} (454 + 549 + 425 + 351) = 77.35 \)

Total SS = \( \sum_{i=1}^{4} \sum_{j=1}^{n_i} y_{ij}^2 - CM = (65)^2 + (87)^2 + \ldots + (88)^2 = 137,601.8 \)

SST = \( \sum_{i=1}^{4} n_i (\bar{\mu}_i - \bar{\mu})^2 = 6(75.67 - 77.35)^2 + 7(78.43 - 77.35)^2 + 6(70.83 - 77.35)^2 + 4(87.75 - 77.35)^2 = 712.6 \)

SSE = Total SS - SST = 1909.2 - 712.6 = 1196.6

So \( F = \frac{712.6 / (4-1)}{1196.6 / (23-4)} = \frac{237.5}{63.0} = 3.77 \)

For \( F \) at 19, p-value is between .025 and .05

So for any \( \alpha \geq \) p-value, we reject \( H_0 \) and accept \( H_1 \) that there is a difference in mean achievement.