SUMMARY: Properties of R^2

- R^2 falls between 0 and 1. The larger the value, the better the explanatory variables collectively predict y.
- $R^2 = 1$ only when all residuals are 0, that is, when all regression predictions are perfect (each $y = \hat{y}$), so residual SS $= \Sigma(y - \hat{y})^2 = 0$.
- $R^2 = 0$ when each $\hat{y} = \bar{y}$. In that case, the estimated slopes all equal 0, and the correlation between y and each explanatory variable equals 0.
- R^2 gets larger, or at worst stays the same, whenever an explanatory variable is added to the multiple regression model.
- The value of R^2 does not depend on the units of measurement.

SUMMARY: F Test That All the Multiple Regression β Parameters $= 0$

1. **Assumptions:**
   - Multiple regression equation holds
   - Data gathered using randomization
   - Normal distribution for y with same standard deviation at each combination of predictors.
2. **Hypotheses:**

$$H_0: \beta_1 = \beta_2 = \ldots = 0 \text{ (all the beta parameters in the model } = 0)$$

$$H_a: \text{At least one } \beta \text{ parameter differs from 0.}$$

3. **Test statistic:** $F =$ (mean square for regression)/(mean square error)
4. **P-value:** Right-tail probability above observed F test statistic value from F distribution with

$$df_1 = \text{number of explanatory variables,}$$

$$df_2 = n - (\text{number of parameters in regression equation}).$$

5. **Conclusion:** The smaller the P-value, the stronger the evidence that at least one explanatory variable has an effect on y. If a decision is needed, reject H_0 if P-value $\leq$ significance level, such as 0.05. Interpret in context.

SUMMARY: Significance Test About a Multiple Regression Parameter (such as β_1)

1. **Assumptions:**
   - Each explanatory variable has a straight-line relation with μ_y, with the same slope for all combinations of values of other predictors in model
   - Data gathered with randomization (such as a random sample or a randomized experiment)
   - Normal distribution for y with same standard deviation at each combination of values of other predictors in model
2. **Hypotheses:**

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

When H_0 is true, y is independent of x_1, controlling for the other predictors.

3. **Test statistic:** $t = (b_1 - 0)/se$. Software supplies the slope estimate b_1, its se, and the value of t.
4. **P-value:** Two-tail probability from t distribution of values larger than observed t test statistic (in absolute value). The t distribution has

$$df = n - \text{number of parameters in regression equation}$$

(such as $df = n - 3$ when $\mu_y = \alpha + \beta_1 x_1 + \beta_2 x_2$, which has three parameters).

5. **Conclusion:** Interpret P-value in context; compare to significance level if decision needed.

SUMMARY: The Process of Multiple Regression

Steps should include:

1. Identify response and potential explanatory variables
2. Create a multiple regression model; perform appropriate tests (F and t) to see if and which explanatory variables have a statistically significant effect in predicting y
3. Plot y versus $\hat{y}$ for resulting models and find R and R^2 values
4. Check assumptions (residual plot, randomization, residuals histogram)
5. Choose appropriate model
6. Create confidence intervals for slope
7. Make predictions at specified levels of explanatory variables
8. Create prediction intervals