Inference for Multiple Linear Regression (MLR)

- Three different types of tests
  - Overall / Omnibus F test
  - Partial F test (for a *single* variable in a MLR model)
  - Multiple Partial F test (for a *group* of variables in a MLR model)

- Partial F test (*Variables added in order*)
  - $SS(X_2|X_1)$: "*Extra Sum of Squares*" for adding X_2 to $Y = \beta_0 + \beta_1 X_1$
    $SS(X_2|X_1) = SSR(X_1, X_2) - SSR(X_1)$; difference between *Regression* SS

Source	**df**	**SS**	**MS**	**F**
Regression	*2*	$SSR(x_1,x_2)$	$SSR(x_1,x_2)/2$	$MSR(x_1,x_2)/MSE$
X_1	*1*	SSR_1		
$X_2 \mid X_1$	*1*	SSR_2		
Error(Residual)	*n-(2+1)*	*SSE*		
Total	*n-1*	*SSY*		

  - $F(X_2 \mid X_1) = \dfrac{SS(X_2 \mid X_1)/1}{MSE(X_1, X_2)}$

  - $SS(X_3|X_1, X_2)$: "*Extra Sum of Squares*" for adding X_3 to $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
    $SS(X_3|X_1, X_2) = SSR(X_1, X_2, X_3) - SSR(X_1, X_2)$

Source	**df**	**SS**	**MS**	**F**
Regression	*3*	$SSR(x_1,x_2,x_3)$	$SSR(x_1,x_2,x_3)/3$	$MSR(x_1,x_2,x_3)/MSE$
X_1	*1*	SSR_1		
$X_2 \mid X_1$	*1*	SSR_2		
$X_3 \mid X_2, X_1$	*1*	SSR_3		
Error(Residual)	*n-(k+1)*	*SSE*		
Total	*n-1*	*SSY*		

  - $F(X_3 \mid X_1, X_2) = \dfrac{SS(X_3 \mid X_1, X_2)/1}{MSE(X_1, X_2, X_3)}$

  - Type I SS, "*Variable Added In Order*" (VIO) SS → R: *anova(model)*

```
y.x1x2 <- lm(SBP~QUET+AGE)
summary(y.x1x2)
  anova(y.x1x2)
```

```
          Df Sum Sq Mean Sq F value    Pr(>F)
QUET       1 3537.9  3537.9 44.5048 2.573e-07
AGE        1  582.6   582.6  7.3293   0.01125
Residuals 29 2305.4    79.5
```

```
y.x2x1 <- lm(SBP~AGE+QUET)
summary(y.x2x1) #SAME  as  y.x1x2
  anova(y.x2x1) #DIFF from y.x1x2
```

```
          Df Sum Sq Mean Sq F value    Pr(>F)
AGE        1 3861.6  3861.6 48.5766 1.160e-07
QUET       1  259.0   259.0  3.2576   0.08149
Residuals 29 2305.4    79.5
```

Inference for Multiple Linear Regression (MLR)

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i \;, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2) \quad i = 1, 2, \ldots, n$$

- Partial F test (*Variables added last*)
  - $SS(X^*|X_1,\ldots,X_k)$: "*Extra Sum of Squares*" for adding X^* to $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$
    $SS(X^*|X_1,\ldots,X_k) = SSR(X_1,\ldots,X_k, \boldsymbol{X^*}) - SSR(X_1,\ldots,X_k)$
    SS_{Reg}(*Full Model*) – SS_{Reg}(*Reduced Model*)
  - Type III SS, Type III Test, "*Variable Added Last*" (VAL) Test → R: *summary(model)*
  - $F(X^* \mid X_1,\cdots,X_k) = \dfrac{SS(X^* \mid X_1,\cdots,X_k)/1}{MSE(X_1,\cdots,X_k, X^*)} \sim F_{1,n-k-2}$ under $H_0 : \beta_1^* = 0$
    - "*Variable Added Last*" tests have df(numerator) = 1 → Equivalent to a T-test

$$t_s = \frac{\hat{\beta}_1^* - \beta_1^{*(0)}}{SE(\hat{\beta}_1^*)} \quad \text{when } \beta_1^{*(0)} = 0$$

- Multiple Partial F test (for a group of variables in a MLR model)
  - $SS(X^*_1,\ldots,X^*_s|X_1,\ldots,X_q) = SSR(X_1,\ldots,X_q, X^*_1,\ldots,X^*_s) - SSR(X_1,\ldots,X_q)$
    $= SS\boldsymbol{E}(X_1,\ldots,X_q) - SS\boldsymbol{E}(X_1,\ldots,X_q, X^*_1,\ldots,X^*_s)$
  - $H_0 : \beta_1^* = \beta_2^* = \cdots = \beta_s^* = 0$

$$F(X_1^*,\cdots,X_s^* \mid X_1,\cdots,X_q) = \frac{SS(X_1^*,\cdots,X_s^* \mid X_1,\cdots,X_q)/s}{MSE(X_1^*,\cdots,X_s^*,\ X_1,\cdots,X_q)}$$

Inference for MLR – Confidence Intervals

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i \;, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2) \quad i = 1, 2, \ldots, n$$

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \cdots + \hat{\beta}_k X_{ik} \qquad i = 1, 2, \ldots, n$$

- 100(1–α)% CI for a Regression Coefficient
  - $\hat{\beta}_i \pm t_{\frac{\alpha}{2}, n-(k+1)} SE(\hat{\beta}_i)$

- *Estimated* Variance for the regression model vs. Variance of a Fitted Value
  - **SLR**: $\hat{\sigma}^2 = s^2_{Y|X} = MSE$
  - **MLR**: $\hat{\sigma}^2 = s^2_{Y|x_1, x_2, \cdots, x_k} = s^2_{Y|\underset{\sim}{X}=(1, x_1, x_2, \cdots, x_k)} = MSE$
  - $s^2_{\hat{Y}|x_1, x_2, \cdots, x_k} = s^2_{\hat{Y}|\underset{\sim}{X}=(1, x_1, x_2, \cdots, x_k)} = \text{var}\left(\hat{Y} \mid \underset{\sim}{X} = (1, x_1, \cdots, x_k)\right)$ is the variance of a fitted value

- Conditional Mean of Y at $\boldsymbol{X}=(1, X_{1*}, X_{2*}, \ldots, X_{k*})$
  - CI for $\mu_{Y|\underset{\sim}{X}=(1, x_{1*}, x_{2*}, \cdots, x_{k*})}$: $\hat{\mu}_{Y|\underset{\sim}{X}=(1, x_{1*}, x_{2*}, \cdots, x_{k*})} \pm t_{\frac{\alpha}{2}, n-(k+1)} SE(\hat{\mu}_{Y|\underset{\sim}{X}=(1, x_{1*}, x_{2*}, \cdots, x_{k*})})$
    - $SE(\hat{\mu}_{Y|\underset{\sim}{X}=(1, x_{1*}, x_{2*}, \cdots, x_{k*})}) = \sqrt{s^2_{\hat{Y}|\underset{\sim}{X}=(1, x_{1*}, x_{2*}, \cdots, x_{k*})}} = s_{\hat{Y}|\underset{\sim}{X}=(1, x_{1*}, x_{2*}, \cdots, x_{k*})}$
    - $\hat{Y} \mid \underset{\sim}{X} = (1, x_{1*}, x_{2*}, \cdots, x_{k*})$ is a linear combination of the $\hat{\beta}_i$ → $s^2_{\hat{Y}|\underset{\sim}{X}=(1, x_{1*}, x_{2*}, \cdots, x_{k*})}$ involves variances & covariances among the $\hat{\beta}_i$
    - ***R***: predict(model, x.new.df, **se.fit**=TRUE) returns $s_{\hat{Y}|\underset{\sim}{X}=(1, x_{1*}, x_{2*}, \cdots, x_{k*})}$ as `$se.fit`

- Prediction Interval for a new Y value at $\boldsymbol{X}=(1, X_{1*}, X_{2*}, \ldots, X_{k*})$
  - CI: $\hat{Y} \mid \underset{\sim}{X} = (1, x_{1*}, \cdots, x_{k*}) \pm t_{\frac{\alpha}{2}, n-(k+1)} SE(\hat{Y} \mid \underset{\sim}{X} = (1, x_{1*}, \cdots, x_{k*}))$
    - $SE(\hat{Y} \mid \underset{\sim}{X} = (1, x_{1*}, \cdots, x_{k*})) = \sqrt{s^2_{Y|\underset{\sim}{X}=(1, x_1, x_2, \cdots, x_k)} + s^2_{\hat{Y}|\underset{\sim}{X}=(1, x_{1*}, x_{2*}, \cdots, x_{k*})}} = \sqrt{MSE + s^2_{\hat{Y}|\underset{\sim}{X}=(1, x_{1*}, x_{2*}, \cdots, x_{k*})}}$

Inference for MLR – Confidence Intervals

```
> y.x1x2 <- lm(SBP~QUET+AGE)                    Df Sum Sq Mean Sq F value    Pr(>F)
  anova(y.x1x2)                    QUET         1 3537.9  3537.9 44.5048 2.573e-07
                                   AGE          1  582.6   582.6  7.3293   0.01125
                                   Residuals   29 2305.4    79.5
```

```
> predict(y.x1x2, x.new.df, interval="confidence", se.fit=T)
  $fit
           fit      lwr     upr
    1 131.6077 126.5475 136.668
  $se.fit
    [1] 2.474176
  $df
    [1] 29
  $residual.scale
    [1] 8.916038  # sqrt(MSE)

> 131.6077 - qt(1-.025,29)*sqrt(2.474176^2)  #[1] 126.5481
> 131.6077 + qt(1-.025,29)*sqrt(2.474176^2)  #[1] 136.6679
```

```
> predict(y.x1x2, x.new.df, interval="prediction", se.fit=T)
  $fit
           fit      lwr      upr
    1 131.6077 112.6833 150.5321
  $se.fit
    [1] 2.474176
  $df
    [1] 29
  $residual.scale
    [1] 8.916038

> 131.6077 - qt(1-.025,29)*sqrt(79.5 + 2.474176^2)  #[1] 112.6828
> 131.6077 + qt(1-.025,29)*sqrt(79.5 + 2.474176^2)  #[1] 150.5326
```