MODEL ASSESSMENT

OLS Model Assessment and Improvement

- How good is a linear model?
- Which coefficients of the linear model are significant (Identify important variables)
- Can we improve quality of linear model?
  - Are assumptions made about errors reasonable?
    - Normality: Errors are normality distributed
    - Homoscedasticity: Errors in different samples have same variance

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2), \;\; i = 1, 2, \ldots, n$$

  - Are there bad measurements in the data (outliers)

OLS: Properties of Estimates

- Both $\hat{\beta}_0$ and $\hat{\beta}_1$ estimates are unbiased

$$E[\hat{\beta}_0] = \beta_0, \quad E[\hat{\beta}_1] = \beta_1$$

- Variance of the estimates

$$\text{var}[\hat{\beta}_1] = \frac{\sigma^2}{S_{xx}}, \quad \text{var}[\hat{\beta}_0] = \sigma^2 \frac{\sum x_i^2}{n\, S_{xx}}$$

- Estimate of σ^2

$$\hat{\sigma}^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} = \frac{\text{SSE}}{n-2}$$

- Distribution of slope estimate $\quad \hat{\beta}_1 \sim \mathcal{N}(\beta_1, \frac{\sigma^2}{S_{xx}})$

OLS: Confidence Intervals on regression coefficients

- 95% two-sided confidence intervals (CI) for $\hat{\beta}_0$ and $\hat{\beta}_1$

$$\beta_1 \in [\hat{\beta}_1 - 2.18\, s_{\hat{\beta}_1}, \hat{\beta}_1 + \underset{t_{0.025,12}}{2.18}\, s_{\hat{\beta}_1}], \quad s_{\hat{\beta}_1} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{(n-2) S_{xx}}}$$

$$\beta_0 \in [\hat{\beta}_0 - 2.18\, s_{\hat{\beta}_0}, \hat{\beta}_0 + 2.18\, s_{\hat{\beta}_0}], \quad s_{\hat{\beta}_0} = s_e \sqrt{\frac{\sum x_i^2}{n\, S_{xx}}}$$

$$s_e = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{(n-2)}}$$

OLS: Hypotheses test on regression coefficients

- In order to check if linear model fit is good or not we can test whether estimate $\hat{\beta}_1$ is significant (different from zero) or not
- Null hypothesis $H_0 : \beta_1 = 0$
- Alternative hypothesis $H_1 : \beta_1 \neq 0$
- Null hypothesis implies $\hat{y}_i = \hat{\beta}_0 + \epsilon_i$ ⟵ Reduced Model
- Alternative hypothesis implies $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \epsilon_i$ ⟵ Full Model
- Do not Reject null hypothesis if CI for β_1 includes 0
- Similarly if CI for $\hat{\beta}_0$ includes 0, then intercept term is insignificant

OLS: Confidence Intervals on regression coefficients

- 95% two-sided confidence intervals (CI) for $\hat{\beta}_0$ and $\hat{\beta}_1$

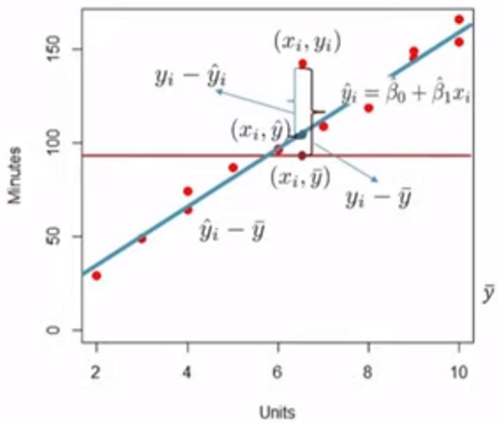
$t_{0.025,12}$

$$\beta_1 \in [\hat{\beta}_1 - 2.18\, s_{\hat{\beta}_1}, \hat{\beta}_1 + 2.18\, s_{\hat{\beta}_1}], \quad s_{\hat{\beta}_1} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{(n-2) S_{xx}}}$$

$$\beta_0 \in [\hat{\beta}_0 - 2.18\, s_{\hat{\beta}_0}, \hat{\beta}_0 + 2.18\, s_{\hat{\beta}_0}], \quad s_{\hat{\beta}_0} = s_e \sqrt{\frac{\sum x_i^2}{n\, S_{xx}}}$$

$$s_e = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{(n-2)}}$$

OLS: Sum Squared Quantities - Definitions



$$SSE = \sum (y_i - \hat{y}_i)^2$$
$$SSR = \sum (\hat{y}_i - \bar{y})^2$$
$$SST = \sum (y_i - \bar{y})^2$$

OLS: F-Test for choosing between models

- F-test for rejecting reduced model
- SST is goodness of fit for reduced model (null hypothesis)
- SSE is goodness of fit for full model (alternative hypothesis)
- F-statistic $F_o = \frac{SST-SSE}{SSE/(n-2)} = \frac{SSR}{SSE/(n-2)}$
- At 5% level of significance reject null hypothesis if $F_o \geq F_{(1,n-2;0.05)}$ (upper critical value of F distribution with 1 and n-2 dfs)
  - Note that the numerator has 1 df

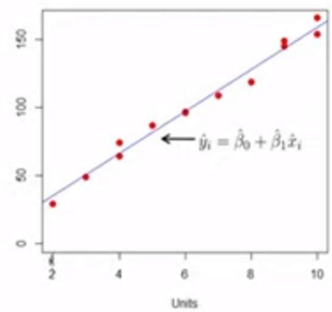
OLS: Example using R

```
Call:
lm(formula = Minutes ~ Units)

Residuals:
    Min      1Q  Median      3Q     Max
-9.2318 -3.3415 -0.7143  4.7769  7.8033

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.162      3.355    1.24    0.239
Units         15.509      0.505   30.71 8.92e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.392 on 12 degrees of freedom
Multiple R-squared:  0.9874, Adjusted R-squared:  0.9864
F-statistic: 943.2 on 1 and 12 DF,  p-value: 8.916e-13
```



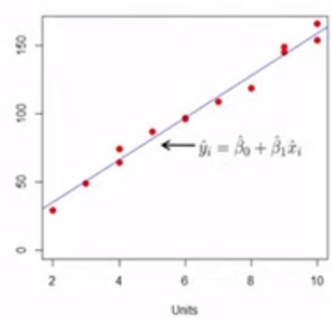
OLS: Example

```
Call:
lm(formula = Minutes ~ Units)

Residuals:
    Min      1Q  Median      3Q     Max
-9.2318 -3.3415 -0.7143  4.7769  7.8033

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.162      3.355    1.24    0.239
Units         15.509      0.505   30.71 8.92e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.392 on 12 degrees of freedom
Multiple R-squared:  0.9874, Adjusted R-squared:  0.9864
F-statistic: 943.2 on 1 and 12 DF,  p-value: 8.916e-13
```




```
              2.5 % 97.5 %
(Intercept) -3.148 11.472
Units       14.409 16.609
```