

SLR

```
le <- read.csv("../datasets/life_expectancy.csv")
y <- le$X2015Life.expectancy
x1 <- le$Medical.doctors
x2 <- le$Nurses
x3 <- le$Pharmacists
model_slr <- lm(y ~ x1)
anova_slr <- anova(model_slr)
model_mlr <- lm(y ~ x1 + x2 + x3)
anova_mlr <- anova(model_mlr)
```

Sum of squares terms

Shouldn't need unless calculating SLR coefficients manually

```
ss_xy <- sum((x1 - mean(x1)) * (y - mean(y)))
ss_xx <- sum((x1 - mean(x1))^2)
```

SLR model summary

```
summary(model_slr)

##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.1102  -3.5062   0.4287   4.0203  11.7057
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  64.86623    0.61511  105.45  <2e-16 ***
## x1           0.34852    0.02367   14.72  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.446 on 176 degrees of freedom
## Multiple R-squared:  0.5519, Adjusted R-squared:  0.5494
## F-statistic: 216.8 on 1 and 176 DF, p-value: < 2.2e-16
```

- Standard errors of the coefficients: estimates standard deviation of the sampling distribution of b_1 . In other words, it is the “spread” of the values we would get when repeatedly sampling b_1 while holding the level of X constant.
- t-values: the test statistic for the distribution that holds under the null hypothesis that X has no impact on Y , which is to say that $\beta_i = 0$.
- p-values: probability that the t-distribution is greater than or equal to the t-value in question. Tells us whether the test statistic is significant or not.
- **Residual standard error** is the standard error of the residuals and equals \sqrt{MSE} . It estimates the standard deviation of the error, σ , and MSE estimates the variance of the error, σ^2 . MSE can be found from the SLR ANOVA table below.
- **Multiple R-squared** is the coefficient of determination, which measures the proportion of the total variation in Y accounted for by the inputs of the model. In other words, it is SSR/SST . r is just the square root of this.

- Adjusted R-squared is adjusted for the bias in R^2 that causes it to become larger as more input variables are added, regardless of whether they actually make the model better.
- F-statistic is the two-sided Global F test (ANOVA test) statistic, $MSR/MSE = b_1^2/s^2[b_1]$, distributed on $F(df_R - df_F, df_F)$ and it is equivalent to the two-sided t-test for β_1 in SLR: $F^* = (t^*)^2$ and the p-values are equal. It is the same statistic as the GLT.

SLR ANOVA table

```
anova_slr
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x1           1 6428.4   6428.4   216.78 < 2.2e-16 ***
## Residuals 176 5219.2     29.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Df shows the degrees of freedom of regression and the degrees of freedom of error
- Sum Sq shows the SSR $\sum(\hat{Y}_i - \bar{Y})^2$ and the SSE $\sum(Y_i - \hat{Y}_i)^2$, which sum to the SSTO.
- Mean Sq shows the MSR and MSE, which are just the Sum Sq column divided elementwise by the Df column.
- F value and Pr(>F) shows the same global F test as in the model summary, testing whether $\beta_1 = 0$.

MLR model summary

```
summary(model_mlr)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.7701  -3.0922  -0.1913   3.8118  10.9957
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  64.47982    0.60561  106.471 < 2e-16 ***
## x1           0.23418    0.03936   5.950 1.44e-08 ***
## x2           0.02574    0.01467   1.755  0.0810 .
## x3           0.32116    0.12417   2.586  0.0105 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.27 on 174 degrees of freedom
## Multiple R-squared:  0.5851, Adjusted R-squared:  0.578
## F-statistic: 81.8 on 3 and 174 DF, p-value: < 2.2e-16
```

- stderr, t-value, p-value, residual stderr are all as in SLR
- Multiple R-squared is still the coefficient of determination, which measures the proportion of the total variation in Y accounted for by the inputs of the model. In other words, it is SSR/SST . r is just the square root of this.

- Adjusted R-squared is adjusted for the bias in R^2 that causes it to become larger as more input variables are added, regardless of whether they actually make the model better.
- F-statistic is still MSR/MSE and tests whether *all* non-intercept parameters are zero or not.

MLR ANOVA table

```
anova_mlr
```

Type I

```
## Analysis of Variance Table
##
## Response: y
##          Df Sum Sq Mean Sq  F value    Pr(>F)
## x1         1 6428.4   6428.4  231.4630 < 2.2e-16 ***
## x2         1  200.9    200.9   7.2325  0.007856 **
## x3         1  185.8    185.8   6.6897  0.010515 *
## Residuals 174 4832.5     27.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Type I table is sequential, so the order matters.

The rows are as follows:

$$\begin{aligned} &SSR(X_1) \\ &SSR(X_2|X_1) \\ &SSR(X_3|X_1, X_2) \end{aligned}$$

So we observe the *marginal* amount of variation accounted for by (or the amount of variation of error reduced by) the regression terms given that all *previous* terms are accounted for.

The Sum Sq rows always sum to SSTO.

The F-values are GLTs that test the marginal effect of input variables. For example, in the x3 row, we test the significance of the marginal reduction in error variance attributed to X_3 after X_1, X_2 have already been considered, which is testing whether $\beta_3 = 0$ given all other predictors have been considered. To test β_2 , we need to change the order to place x2 last.

The F-statistics are calculated as follows:

$$\begin{aligned} F^* &= \frac{SSR(X_1)/(df_R - df_F)}{SSE/df_F} \\ F^* &= \frac{SSR(X_2|X_1)/(df_R - df_F)}{SSE/df_F} \\ F^* &= \frac{SSR(X_3|X_1, X_2)/(df_R - df_F)}{SSE/df_F} \end{aligned}$$

```
library(car)
Anova(model_mlr)
```

Type II

```
## Anova Table (Type II tests)
##
```

```
## Response: y
##           Sum Sq Df F value    Pr(>F)
## x1          983.1  1 35.3989 1.441e-08 ***
## x2           85.6  1  3.0805  0.08100 .
## x3          185.8  1  6.6897  0.01051 *
## Residuals 4832.5 174
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The rows are as follows:

$$SSR(X_1|X_2, X_3)$$

$$SSR(X_2|X_1, X_3)$$

$$SSR(X_3|X_1, X_2)$$

So only the last row of the Type I table is equivalent to the corresponding row of the Type II table.

So we observe the *marginal* amount of variation accounted for by (or the amount of variation of error reduced by) the regression terms given that *all other* terms are accounted for.

The **Sum Sq** rows *do not* sum to SSTO.

The F-tests are equivalent to the t-tests in the model summary. This is to say that the t-tests test the marginal effect of a single predictor given that all other terms have been included in the model.