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 $b1$. 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 $\beta_1 = 0$.

s_Y can be found via:

```
n <- length(y)
sqrt(sum(anova_slr[, "Sum Sq"]) / (n - 1))
```

```
## [1] 8.112072
```

MLR

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)
```

Type II

```
## Loading required package: carData
```

```
Anova(model_mlr)
```

```
## 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.

Inference

Critical t-values

Two-sided:

```
alpha <- 0.05
dfe <- anova_slr["Residuals", "Df"]
t_crit <- qt(1 - alpha / 2, df = dfe)
```

One-sided (i.e. $H_0 : \beta_1 = 50, H_a : \beta_1 > 50$):

```
dfe <- anova_slr["Residuals", "Df"]
t_crit <- qt(1 - alpha, df = dfe)
```

Test statistics

$$b^* = \frac{b_1 - \beta_1}{s[\beta_1]}$$

p-values

Confidence intervals

```
confint(model_slr)
```

```
##                 2.5 %     97.5 %
## (Intercept) 63.6522927 66.0801698
## x1           0.3018074  0.3952402
```

```
confint(model_mlr)
```

```
##                   2.5 %      97.5 %
## (Intercept) 63.284537614 65.67510062
## x1           0.156496305  0.31186592
## x2          -0.003205879  0.05469436
## x3           0.076085264  0.56623243
```

Mean response $E[\hat{Y}_h]$:

```
library(ALSM)
```

```
## Loading required package: leaps
```

```
## Loading required package: SuppDists
```

```
x_h <- median(x1)
ci.reg(model_slr, newdata = x_h, type = "m")
```

```
##       x1      Fit Lower.Band Upper.Band
## 1 15.715 70.34328   69.51917   71.16739
```

Single new predicted value $\hat{Y}_h$:

```
ci.reg(model_slr, newdata = x_h, type = "n")
```

```
##       x1      Fit Lower.Band Upper.Band
## 1 15.715 70.34328   59.56468   81.12188
```

The mean of 3 new predicted values with the same X_h, $\bar{Y}_{h(new)}$:

```
ci.reg(model_slr, newdata = x_h, type = "nm", m = 3)
```

```
##       x1      Fit Lower.Band Upper.Band
## 1 15.715 70.34328   64.08398   76.60258
```

Inference on Correlation Coefficient

```
cor(x1, y)
```

```
## [1] 0.7429066
```

```
cor(cbind(y, x1, x2, x3))
```

```
##            y        x1        x2        x3
## y  1.0000000 0.7429066 0.6612558 0.6370947
## x1 0.7429066 1.0000000 0.7793268 0.7038282
## x2 0.6612558 0.7793268 1.0000000 0.6864924
## x3 0.6370947 0.7038282 0.6864924 1.0000000
```

```
cor.test(x1, y)
```

```
##
##  Pearson's product-moment correlation
##
## data:  x1 and y
## t = 14.723, df = 176, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6689143 0.8023215
## sample estimates:
##       cor
## 0.7429066
```

This is equivalent to the ANOVA F test or the T test for β_1 for SLR for $\rho = 0, \beta_1 = 0$ but not for values other than 0.

Lack of Fit

Requires replicates or grouping

```r
model_reduced <- model_slr
model_full <- lm(y ~ as.factor(x1))
anova(model_reduced, model_full)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x1
## Model 2: y ~ as.factor(x1)
##   Res.Df    RSS  Df Sum of Sq     F Pr(>F)
## 1    176 5219.2
## 2      3   72.3 173    5146.8 1.234 0.5103
```

F-values

Critical values: Easiest to think in terms of Full and Reduced models of GLT.

```r
# 2 parameters including intercept
p_full <- 2
# 1 parameter; just the intercept
p_reduced <- 1

df_full <- n - p_full
df_reduced <- n - p_reduced

f_star <- qf(1 - alpha, df_reduced - df_full, df_full)

p <- 1 - pf(f_star, df_reduced - df_full, df_full)
```

GLT:

SSE(F) can be pulled from the ANOVA table of the full model. SSE(R) is equal to SSTO for some reason…

Still comes out to be MSR/MSE