# Contents



**Introduction to Multiple Linear Regression (Ch.6)**

Data for Multiple Linear Regression:

- $Y_i$ is the response variable.
- $X_{i,1}, X_{i,2}, \ldots, X_{i,p-1}$ are $p$-1 predictor (i.e., explanatory, independent) variables.
- Cases denoted by $i$ = 1, 2, …, $n$.

Statistical Model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \,.$$

- $Y_i$ is the value of the response variable for the $i$-th case.
- $X_{i,k}$ is the value of the *k-th* predictor (i.e., explanatory) variable for the *i-th* case.
- $\beta_0$ is the intercept, $\beta_1, \beta_2, \ldots, \beta_{p-1}$ are the regression coefficients for the explanatory variables.
- $\varepsilon_i$'s are independent, normally distributed random errors with mean 0 and variance $\sigma^2$.
- In simple linear regression, p=2.

The predictors can be:

- Separate variables.
- Dummy codes for categorical (i.e., qualitative) variables (more in Ch.8).
- Polynomial terms.
- Transformed variables.
- Interaction terms.
- A combination of the above.

Parameters are:

- $\beta_0, \beta_1, \beta_2, \ldots, \beta_{p-1}$, and $\sigma^2$.
- $\beta_0, \beta_1, \beta_2, \ldots, \beta_{p-1}$ are estimated using OLS (also MLE under the Normal assumption).
- The OLS and MLE lead to the same estimates when $\varepsilon_i$'s are i.i.d. Normal).

The computational algorithms have to be expressed in matrix notation.

Q: What does the "**linear**" mean?

## Statistical Inference in Multiple Linear Regression (Ch. 6, cont.)

The basic approach to statistical inference in multiple linear regression is the same as in simple linear regression. The **main differences** are:

1. Degree of freedoms will change: $df_{Reg} = p - 1$, $df_{Error} = n - p$. This will change the degree of freedoms used in the t-statistic (CI, test for parameters, etc.) and the F-statistic (ANOVA).
2. More computationally intensive (yet, it will be taken care of by the software).

### Inferences concerning regression coefficients

- **Sampling distribution of $b_k$ (i.e., $\hat{\boldsymbol{\beta}}_k$, the estimate of the k-th slope)**
  Under the Normal assumption, the OLS (same as MLE) estimator $b_1$ has distribution
  $$\frac{(b_k - \beta_k)}{se(b_k\ )} \sim t(df = df_E = n - p)$$
  *where:* $\beta_k$ *is the unknown true value of the slope;*
  $se(b_k)$ (or $s(b_k)$) is the standard error of $b_k$;

- **Confidence Interval for $\beta_k$** is constructed as (k=0,1,…,p-1):
  $$b_k \pm t_{(1-\alpha/2,n-p)} se(b_k),$$
  where: $b_k$ is the estimate of $\beta_k$;
  $t_{(1-\alpha/2,n-p)}$ is the critical value for df = n-p at (1- α)100% confidence level;
  $se(b_k)$ is the standard error of $b_k$.

- **Test of significance for $\beta_k$** is constructed as:
  $H_0: \beta_k = \beta_{k0}$ vs $H_a: \beta_k \neq \beta_{k0}$
  $$t_{obs} = \frac{(b_k - \beta_{k0})}{se(b_k)}$$
  $p - value = 2 * P(\ t > |t_{obs}|)$, where $t \sim t(n - \text{p})$
  Reject $H_0$ if p-value < $\alpha$.
  Or, use critical value, reject $H_0$ if $|\ t_{obs}| \geq t_{crit}, t_{crit} = t(1 - \alpha/2, n - p)$
  where: $\beta_{k0}$ is the "hypothesized" value for the slope (often, $\beta_{10} = 0$).
  If $H_a$ is one-sided, adjust the p-value computation is one-sided as well.
  - If Ha: beta > 0, then p-value = P( $t_{(df = n-p)} > t_{obx}$ )
  - If Ha: beta < 0, then p-value = P( $t_{(df = n-p)} < t_{obx}$ )

- If the distribution of $\varepsilon_i$ is not normal but is relatively symmetric, then the CIs and significance tests are reasonable approximations.

### Confidence Interval and Prediction Interval

- The difference between mean response ($E(Y_h) = \mu_h$, or $Y_{mean}$) and a single response ($Y_{h(new)}$) at ($X_{h,1}$, $X_{h,2}$, … , $X_{h,p-1}$).

- Same point estimation: $\hat{\mu}_h(ie, \hat{Y}_{mean}) = \hat{Y}_{h(new)} = b_0 + b_1 X_{h,1} + \ldots + b_{p-1} X_{h,p-1}$

- Different standard errors: $se(\hat{Y}_{h(new)}) = \sqrt{\left[se(\hat{Y}_{mean})\right]^2 + MSE}$

- The CI (for mean response) and PI (for single response) are:

  (point est.) ± (critical value)(std.err of the point est.)

  where the critical value is $t_{(1-\alpha/2,n-p)}$. (Recall that $df_E = n - p$.)

**The ANOVA table**

- Partitioning **sums of squares:**

  *SSTO = SSR + SSE*

  Total variation (measured by Sum of Squares) in Y: $SSTO = \sum\left(Y_i - \bar{Y}\right)^2$

  Variation in Y that can be explained by X: $SSR = \sum\left(\hat{Y}_i - \bar{Y}\right)^2$

  Variation due to randomness (unexplained variation): $SSE = \sum\left(Y_i - \hat{Y}_i\right)^2$

- Partitioning **Degrees of Freedom:**

  $df_{Total}$ (n–1) = $df_{Reg}$ (p–1) + $df_{Error}$ (n–p)

- The ANOVA table (note the change in df)

| Source | Sum of Squares | df | Mean Squares | F | P-value |
|---|---|---|---|---|---|
| Regression(Model) | SSR | p-1 | $MSR=SSR/df_R$ | F=MSR/MSE | $P(F_{(dfR,\ dfE)} > F)$ |
| Error(Residual) | SSE | n-p | $MSE=SSE/df_E$ | | |
| Total | SSTO | n-1 | | | |

- The "Global" F-test in the ANOVA table

  $H_0$: $\beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$

  $H_a$: $\beta_k \neq 0$, for at least one β, k=1,., p-1 (ALOI)

  Under $H_0$, the F-ratio follows F-distribution with degree of freedoms ($df_R$, $df_E$).

**Coefficient of multiple determination**

- $R^2$ = SSR/SSTO is the proportion of the variation in Y (measured by the sum of squares) that can be determined/explained by the current multiple linear regression model using $X_{i,1}, \ldots, X_{i,p-1}$ .
- $R^2$ is between (0, 1).
- $R^2$ is meaningful when the current model is appropriate. Be sure to check the model assumptions regardless of $R^2$ value.
- Models with small $R^2$ can still provide meaningful insight about the data.
- For models with the same number of predictors, a larger value of $R^2$ is preferred. However, note that $R^2$ increases when more predictors are added to the model. (Adjusted-$R^2$ will be introduced in multiple linear regression.)
- Adjusted-$R^2$: $adj-R^2 = R_a^2 = 1 - \dfrac{n-1}{n-p}(1-R^2) = 1-(n-1)\dfrac{MSE}{SST}$

**Simultaneous CI**

- Change the critical value to adjust for the family confidence.
- The Bonferroni method:
  - Split the family α-level to g members in the family. I.e., use $\alpha^* = \alpha/g$ for each CI or test in the family.
  - $B = t_{(1-\alpha/(2g), n-p)}$ .
  - Still use the t-distribution, but with adjusted level.
  - Can be applied to simultaneous/joint CI's for $\beta_k$'s, simultaneous/joint CI's for mean predictions, and simultaneous/joint PI's for individual predictions.
- Working-Hotelling for simultaneous/joint CI's for mean predictions:
$$\hat{\mu}_h \pm W \times se(\hat{Y}_{mean}),\ where\ W = \sqrt{pF_{(1-\alpha;\, p,\ n-p)}}\ .$$
- Scheffé's method for simultaneous/joint PI's for individual predictions:
$$\hat{Y}_{h(new)} \pm S \times se(\hat{Y}_{h(new)}),\ where\ S = \sqrt{gF_{(1-\alpha;\ g,\ n-p)}}\ .$$

**Be aware of hidden extrapolation!**

- Why "hidden?"
- We will introduce a numerical measure later.

**General Linear Tests: Extra Sum of Squares and Partial F-test (Ch. 7)**

Consider a linear regression with 5 predictors (*$X_1$, $X_2$, $X_3$, $X_4$, $X_5$*):

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3} + \beta_4 X_{i,4} + \beta_5 X_{i,5} + \varepsilon_i.$$

**The hypotheses:**

$H_0$: $\beta_4 = \beta_5 = 0$
$H_1$: $\beta_4$ and $\beta_5$ are not both 0

**The Full and Reduced Model:**

To test the above hypothesis, consider the following 2 models:

- F: The Full Model (Includes all of the predictors)
$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3} + \beta_4 X_{i,4} + \beta_5 X_{i,5} + \varepsilon_i$$
- R: The Reduced Model: (Plug $H_0$ into the Full Model and "reduce")
$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3} + 0 \cdot X_{i,4} + 0 \cdot X_{i,5} + \varepsilon_i$$
After cleaning up the equation, we have the following model:
$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3} + \varepsilon_i$$

Look at the difference between the reduced model and the full model
- in SSE (reduce unexplained SS), or
- in SSR (increase explained SS)
- Since SSR+SSE=SST, the change in SSE and the change in SSR are equivalent, as long as the Y-values are NOT changed.
- However, in some cases (see Examples in the Lab note), only SSE should be used.

**The Partial F-test**

$$F^* = \frac{(\mathrm{SSE}(R) - \mathrm{SSE}(F))/(df_E(R) - df_E(F))}{\mathrm{SSE}(F)/df_E(F)}$$

Let $df_1 = df_E(R) - df_E(F)$, $df_2 = df_E(F)$. P-value = $P(F_{(df1,\, df2)} > F^*)$.
Reject $H_0$ if p-value < $\alpha$, or $F^* > F_{(1-\alpha,\, df1,\, df2)}$.

Q. Why?
A. Cochran's Theorem. (Stat 616, Generalized Linear Models)

Write down the reduced models for each of the following hypotheses:

- $H_0$: $\beta_4 = \beta_5$, vs., $H_1$: $\beta_4 \neq \beta_5$

- $H_0$: $\beta_4 = 1$, $\beta_5 = 2$, vs., $H_1$: At Least One Inequality

**Notation for Extra Sum of Squares**

- SSE($X_1$,$X_2$,$X_3$,$X_4$,$X_5$) is the SSE for the ***full*** model, SSE(F).
- SSE($X_1$,$X_2$,$X_3$) is the SSE for the ***reduced*** model, SSE(R).
- SSE($X_4$,$X_5$ | $X_1$,$X_2$,$X_3$) is the difference in the SSE: SSE($X_1$,$X_2$,$X_3$) – SSE($X_1$,$X_2$,$X_3$,$X_4$,$X_5$).
- The Extra Sum of Square measures the "marginal" contribution of ($X_4$,$X_5$) when ($X_1$,$X_2$,$X_3$) are already in the regression model.

**Special Cases and other applications of Extra Sum Square**

- Compare models that differ by one predictor variable ($H_0$: $\beta_4 = 0$), $F(1,n-p)=t^2(n-p)$
  - This is equivalent to the t-test.

- Compare the full model against the null model ($Y_i = \beta_0 + \varepsilon_i$)
  - This is testing $H_0$: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$, and is equivalent to the "Global" F-test in ANOVA.

- Add one variable at a time (Type I Sum of Square)
  - SSR ($X_1$)
  - SSR ($X_2$ | $X_1$)
  - SSR ($X_3$ |$X_1$, $X_2$)
  - SSR ($X_4$ |$X_1$, $X_2$, $X_3$)

  SSR ($X_1$) +SSR ($X_2$ | $X_1$) + SSR ($X_3$ | $X_1$, $X_2$) + SSR ($X_4$ | $X_1$, $X_2$, $X_3$) =SSR($X_1$, $X_2$, $X_3$, $X_4$)

- Coefficients of partial determination and coefficients of partial correlation.
  - See text Ch.7.4 (p. 268)
  - We will revisit this topic when we introduce "added variable plots" (aka. "partial regression plot") in Ch.10.

## Standardized Regression Model (Ch. 7, cont.)

**The Procedure:**

1. Standardize Y and each X by subtracting the mean and then dividing by the standard deviation of each variable. I.e., get z-scores for Y and each X, respectively.
   Then divide the results by $\sqrt{n-1}$. (This step can be optional.)
2. The regression coefficients on the above transformed variables are the standardized regression coefficients.

**The result:**

- The standardized regression model does not have intercept. I.e. intercept = 0.

- It put regression coefficients in common units. (In comparison, the units for the usual coefficients are units for Y divided by units for X.)

- Interpretation is that a one standard deviation increase in X corresponds to a (standardized beta)x(standard deviation of y) increase in Y.

## Multicollinearity

What is Multicollinearity?

Why is Multicollinearity problematic?

- The numerical analysis problem is that the matrix $X^TX$ is close to singular and is therefore difficult to invert accurately.

- The statistical problem is that there is too much correlation among the explanatory variables. As a result, it is difficult to determine the association of 1 predictor vs. the response (i.e., the regression coefficient) while other predictors are in the model.

- In data analysis, regression coefficients and their standard errors are not well estimated and may be meaningless.

- $R^2$ and predicted values are usually ok, though.

Solving the statistical problem may solve the numerical problem as well.

- We want to refine a model that currently has redundancy in the explanatory variables.
- The above should be done regardless of whether $X^TX$ can be inverted without difficulty.

We will discuss this topic again in model diagnostics.