Advanced Topics in Linear Regression

Multi-linear Regression with Two Features

Concept:

Multi-linear regression (often called *multiple linear regression*) is an extension of simple linear regression where we predict a target variable using **two or more features (independent variables)**.

With **two features**, the model tries to fit a plane (instead of a line) in 3D space.

Mathematical Form:

If we have two features x_1 and x_2, the regression model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

Where:

- y = dependent (target) variable
- x_1, x_2 = independent (input) features
- β_0 = intercept (bias term)
- β_1, β_2 = regression coefficients (weights) for the features
- ϵ = error term (residuals)

Geometric Interpretation:

- **Simple linear regression (1 feature):** fits a straight line in 2D space.
- **Multiple regression with 2 features:** fits a **plane** in 3D space.
- **More than 2 features:** fits a hyperplane in higher dimensions (not visually possible beyond 3D).

Example:

Suppose we want to predict **house price (y)** based on:

- x_1: square footage
- x_2: number of bedrooms

The model might look like:

$$\text{Price} = 50{,}000 + 200 \cdot (\text{sqft}) + 10{,}000 \cdot (\text{bedrooms})$$

- Intercept = 50,000 (base price)
- Each extra square foot adds $200
- Each additional bedroom adds $10,000

Key Points:

1. **Coefficients interpretation:** Each β tells us how much y changes when the corresponding x increases by 1 unit, holding the other feature constant.
2. **Assumptions:** Linear relationship, no multicollinearity, normally distributed errors, homoscedasticity.
3. **Visualization:** In 3D, the regression plane tries to minimize the squared distance of all data points from the plane.

F-Statistic in Regression

Core Idea

When we build a **multi-linear regression model**, we're essentially asking:

👉 *Do the features, taken together, explain a significant amount of variation in the target variable compared to a model with no features (just the mean)?*

The **F-statistic** measures this by comparing:

- **Full model (with predictors)** vs. **Restricted model (intercept-only model, i.e., mean of y)**.

If the predictors add *significant explanatory power*, the F-statistic will be large.

Formula:

For a regression with k predictors and n observations:

$$F = \frac{\mathrm{MSR}}{\mathrm{MSE}} = \frac{(SSR/k)}{(SSE/(n-k-1))}$$

Where:

- **SSR (Regression Sum of Squares):** Variation explained by the model
- **SSE (Error Sum of Squares):** Variation left unexplained (residuals)
- **MSR = SSR / k** = Mean square due to regression
- **MSE = SSE / (n-k-1)** = Mean square error

How it Works Mathematically

Let's say we have **two features** x_1 and x_2:

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$

We compute:

- **SST (Total Sum of Squares):** total variation in y
- **SSR (Regression Sum of Squares):** variation explained by x_1,x_2
- **SSE (Error Sum of Squares):** variation left unexplained

SST = SSR + SSE

Why Not Just Use t-tests for Each Coefficient?

- **t-test** checks each predictor *individually*.
- But predictors can be correlated (multicollinearity). In that case:
    - Individually, each predictor might not look significant.
    - But **together, they might explain a lot of variation**.

- The **F-test captures this “joint significance”**.

So in multi-linear regression, the F-test is more **global**, while t-tests are **local**.

Interpretation:

- **High F-statistic (with low p-value):** At least one predictor significantly improves the model fit.
- **Low F-statistic (with high p-value):** The model does not explain variation in the target variable better than a baseline (just using the mean).

Example:

Suppose we’re predicting **exam scores (y)** based on:

- Hours studied (x_1)
- Sleep hours (x_2)

If the regression output gives:

- F-statistic = **25.4**
- p-value < 0.001

This means: the model with "hours studied" and "sleep hours" explains exam scores significantly better than just using the mean exam score.

Key Points:

1. **t-test vs F-test:**
    - t-test → checks if a single predictor is useful.
    - F-test → checks if the whole model (all predictors together) is useful.
2. **Degrees of freedom:**
    - Numerator = k (number of predictors)
    - Denominator = n–k–1 (sample size minus predictors minus intercept).

3. **Software:** In regression output (like in statsmodels, R, sklearn), you'll usually see the F-statistic along with its p-value at the top of the summary.

✅ So, the **F-statistic is like a global test of usefulness for your regression model**.

Interaction in Regression

Concept:

An **interaction** occurs when the effect of one predictor on the target variable **depends on the level of another predictor**.

In other words, predictors don't just add up independently; they can *modify each other's influence*.

Mathematical Form:

For two predictors x_1 and x_2:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 \cdot x_2) + \epsilon$$

- β_3 is the **interaction term coefficient**.
- If $\beta_3 \neq 0$, the effect of x_1 on y depends on x_2 (and vice versa).

Example:

Suppose we study **salary (y)** as a function of:

- **Education (x1, in years)**
- **Work experience (x2, in years)**

Model without interaction:

$$\text{Salary} = \beta_0 + \beta_1(\text{Education}) + \beta_2(\text{Experience})$$

Model with interaction:

$$\text{Salary} = \beta_0 + \beta_1(\text{Education}) + \beta_2(\text{Experience}) + \beta_3(\text{Education} \times \text{Experience})$$

Interpretation: the return to experience might be higher for more educated people — the two variables **amplify each other**.

Qualitative Predictors (Categorical Variables)

Concept:

A **qualitative predictor** is a variable that represents categories instead of numeric values (e.g., gender, region, car type).

Regression needs numbers, so we **encode categories** into **dummy variables** (0/1).

Example:

Suppose "Region" has 3 categories: **North, South, West**.

We create **dummy variables**:

- D_1 = 1 if South, else 0
- D_2 = 1 if West, else 0
- North is the **baseline** (when both D_1 = D_2 = 0).

Model:

$$y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \epsilon$$

- β_0: mean response for **North** (baseline)
- β_1: difference between South and North
- β_2: difference between West and North

Interaction Between Quantitative & Qualitative Predictors

Often, we also include interactions between **categorical** and **numerical** variables.

Example: Predicting salary (y) based on:

- Gender (D = 1 if female, 0 if male)
- Years of Experience (x)

Model:

$$y = \beta_0 + \beta_1 D + \beta_2 x + \beta_3 (D \cdot x) + \epsilon$$

Interpretation:

- β_2: effect of experience for males (baseline)
- β_1: difference in intercept between females and males
- β_3: difference in slope (experience effect) for females compared to males

So this tests whether **experience affects salaries differently by gender**.

✅ **Summary:**

- **Interaction terms** allow predictors to modify each other's effects.
- **Qualitative predictors** are handled via **dummy variables**.
- Together, we can model rich relationships, e.g., how the effect of experience on salary differs by region or gender.

Higher Order (Non-linear) Regression

1. Concept

- Linear regression can be **extended to non-linear relationships** by introducing polynomial (or other basis function) transformations of the input features.
- Instead of fitting a straight line, we fit a **polynomial curve** to capture more complex patterns in data.

---

2. Polynomial Regression Model

For a single input variable x:

$$\hat{y} = w_0 + w_1 x + w_2 x^2 + \cdots + w_d x^d$$

- d = degree of the polynomial
- $w = [w_0, w_1, w_2, \dots, w_d]^T$ = parameter vector (coefficients to be learned)

Example:

$y = a + bx + cx^2 + dx^3$

---

3. Matrix Form

We can rewrite polynomial regression in **matrix notation**:

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^d \\ 1 & x_2 & x_2^2 & \dots & x_2^d \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_N & x_N^2 & \dots & x_N^d \end{bmatrix}$$

- Each row = one data point.
- Each column = one **basis function** (e.g., $x, x^2, x^3, \dots$).

---

4. Basis Functions

- The transformed features ($x, x^2, \dots, x^d$) are called **basis functions**.
- The model predicts $\hat{y}$ as a **linear combination of these basis functions**:

$\hat{y} = Xw$

💡 Even though the function looks non-linear in x, it is **linear in parameters w** → still solvable by linear regression methods.

---

5. Complexity Considerations

- **Model Complexity** → Number of parameters (depends on polynomial degree d)
  - Higher d → more flexibility, but risk of overfitting.
- **Sample Complexity** → Number of data points N needed.
  - Must have enough data to reliably estimate all parameters.

---

6. Extensions

- Instead of just polynomial terms, we can use **other basis functions**:
  - Exponential: e^x
  - Trigonometric: $\sin(x), \cos(x)$
  - A combination of many functions depending on the problem.

✅ Key Takeaways

- Higher-order regression expands linear regression to capture non-linear patterns.
- Polynomial terms are examples of **basis function expansion**.
- Linear in parameters → same fitting approach as linear regression.
- Balance degree dd with data size NN to avoid overfitting.

Polynomial Regression

When we write polynomial regression as:

$\hat{y} = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + \dots$

those terms (x^2, x^3) are just **basis functions of xx**.

So more generally, regression can use **any set of functions of xx**, not just powers:

$\hat{y} = w_0 + w_1 f_1(x) + w_2 f_2(x) + \dots + w_d f_d(x)$

where each $f_i(x)$ could be:

- Polynomial: $x, x^2, x^3, \dots$
- Logarithmic: log(x)
- Exponential: e^x
- Trigonometric: sin(x),cos(x)
- Or even a **combination** of them

👉 The important part is:

- The model is **linear in the parameters w** (so we can still solve it using linear regression techniques).
- It becomes **non-linear in x** because of the chosen basis functions.

This is why polynomial regression is often introduced as a **special case** of *basis function regression*.

Why it's still linear regression:

Take a polynomial regression example:

$$\hat{y} = w_0 + w_1 x + w_2 x^2 + w_3 x^3$$

If we define new features:

$$z_1 = x, \quad z_2 = x^2, \quad z_3 = x^3$$

then the model becomes:

$$\hat{y} = w_0 + w_1 z_1 + w_2 z_2 + w_3 z_3$$

This is **linear in** w_0, w_1, w_2, w_3 → so it's a linear regression model.

We're just using **transformed features** instead of the raw xx.

⚡ When it stops being linear regression:

If the coefficients themselves appear inside nonlinear functions, e.g.:

- $\hat{y} = w_1^2 x$
- $\hat{y} = e^{w_1 x}$
- $\hat{y} = \sin(w_1 x)$

Now the relationship is **nonlinear in parameters** (w_1), so it's a **nonlinear regression model**.

👉 **Rule of thumb:**

- **Linear in coefficients → Linear Regression** (even if predictors are transformed like $x^2, \log(x), e^x$, etc.)
- **Nonlinear in coefficients → Nonlinear Regression**
