6-1 Regression Diagnostics for Multiple Regression MAT 374

Diagnostics play an important role in the development and evaluation of multiple regression models. Most of the diagnostic procedures for simple linear regression that we described previously carry over directly to multiple regression. We will revisit these diagnostic procedures and also introduce some specialized diagnostics and remedial procedures for multiple regression.

The Four Assumptions of the Multiple Linear Regression Model

The four conditions ("LINE") that comprise the multiple linear regression model generalize the simple linear regression model assumptions to account for the presence of multiple predictors:

1. **Linear Function:** The mean of the response, $E(Y|X_1, X_2, \ldots, X_p)$, at each set of values of the predictors, is a linear function of the predictors.

2. **Independent:** The errors, ε_i, are independent.

3. **Normally Distributed:** The errors, ε_i, at each set of predictor values are normally distributed.

4. **Equal Variances:** The errors, ε_i, at each set of predictor values have equal variances, denoted σ^2.

– An equivalent way to think of the first (linearity) condition is that the mean of the error, $E(\varepsilon_i|X_1, X_2, \ldots, X_p)$, is zero.

– An alternative way to describe all four assumptions together is that the errors, ε_i, are independent normal random variables with mean zero and constant variance σ^2:

$$\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$$

– As in simple linear regression, we can assess whether these conditions seem to hold for a multiple linear regression model applied to a particular sample dataset by examining the estimated errors (the residuals), e_i.

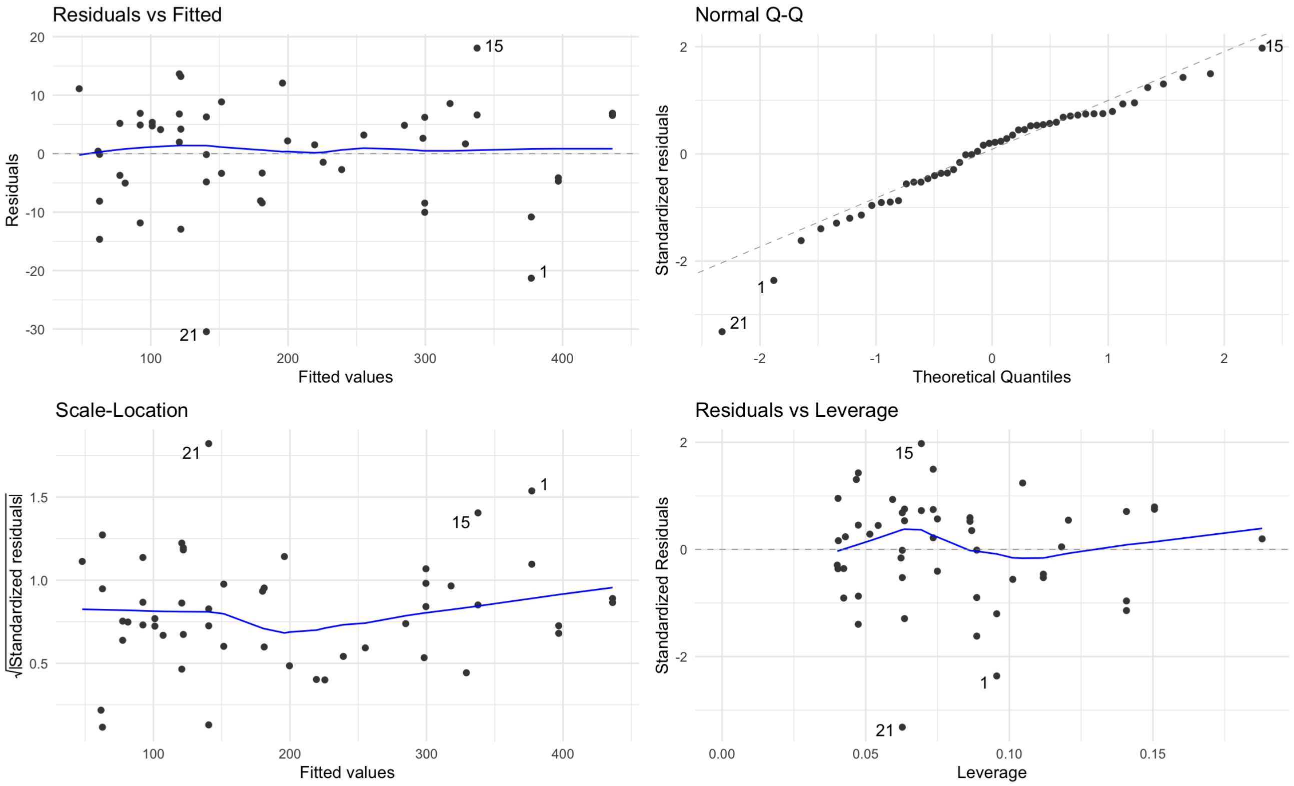

Figure 1: Diagnostic plots for assessing the multiple linear regression model fitted to the Salary dataset from Section 5.3, Example 1.

– For example, Figure 1 shows the four standard diagnostic plots of the model fitted to the Salary dataset from Section 5.3, Example 1. We can see the following:

1. The *Residuals vs Fitted* plot checks the linearity and constant variance assumptions; a random scatter around zero suggests adequacy of the model form. The residuals appear roughly centered around zero with no strong curvature, suggesting the linear model form is reasonable.
2. The *Normal Q–Q* plot assesses normality of the residuals; points lying close to the diagonal line indicate that the error distribution is approximately normal. The points mostly follow the reference line, indicating approximate normality of errors, though slight deviations in the tails may exist.
3. The *Scale–Location* plot evaluates homoscedasticity; a roughly horizontal smooth line implies constant variance across fitted values. The spread of standardized residuals appears relatively constant across fitted values, implying homoscedasticity.
4. The *Residuals vs Leverage* plot identifies influential observations; points with high leverage and large residuals may exert disproportionate influence on the fitted model. No single observation shows both high leverage and large residual magnitude, so there are no clear influential points, though observations 15 and 21 merit some attention.

– When fitting a multiple regression model, it is important to:

1. **Assess model validity.** Determine whether the proposed regression model provides an adequate fit to the data. The main tools used to validate regression assumptions are plots involving standardized residuals and/or fitted values. These plots enable us to visually assess whether assumptions are being violated and, under certain conditions, suggest possible remedies. We will also consider a tool called *marginal model plots*, which have wider application than residual plots.
2. **Identify leverage points and outliers.** Determine which (if any) of the data points have predictor values that exert an unusually large influence on the fitted regression model. Determine which (if any) of the data points are outliers, that is, observations that do not follow the pattern set by the bulk of the data, given the model.

   – **Leverage point**: The sum of the leverages equals p, the number of parameters (regression coefficients including the intercept). A popular rule to classify a point as a leverage point is when the leverage value is greater than double the average leverage.

   – **Outlier**: The common practice of labeling points as outliers in small- to moderate-size data sets if the standardized residual for the point falls outside the interval from –2 to 2 . In very large data sets, we shall change this rule to –4 to 4 . (Otherwise, many points will be flagged as potential outliers.)

   – **Influential point**: A data point is influential if it unduly influences any part of regression analysis, such as the predicted responses, the estimated slope coefficients, or the hypothesis test results. Outliers and high-leverage data points have the potential to be influential, but we generally have to investigate further to determine whether or not they are actually influential.
3. **Assess predictor effects.** Examine the effect of each predictor variable on the response variable, having adjusted for the effects of other predictors, using *added variable plots.*
4. **Evaluate multicollinearity.** Assess the extent of collinearity among predictor variables using *variance inflation factors (VIFs).*
5. **Check constant variance assumption.** Examine whether the assumption of constant error variance is reasonable. If not, determine how to overcome this problem.
6. **Check for autocorrelation.** If the data are collected over time, examine whether the residuals are correlated across time.

Using residuals and standardized residuals for model checking

– In its simplest form, a multiple linear regression model is valid for the data if the conditional mean of Y given X is a linear function of X and the conditional variance of Y given X is constant. In other words,

$$E(Y \mid X = x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

and

$$\mathrm{Var}(Y \mid X = x) = \sigma^2.$$

– When a valid model has been fit, a plot of the standardized residuals r_i against any predictor or any linear combination of the predictors (such as the fitted values) will have the following features:

- A random scatter of points around the horizontal axis, since the mean function of the residuals e_i is zero when a correct model has been fit.
- Constant variability as we look along the horizontal axis.

– For example, figure 2 shows the standardized residual plots against the fitted values and each predictor for the model fitted to the Salary 1 dataset from Section 5.3, Example 1. A random scatter of points around zero without discernible pattern indicates that the linearity assumption is reasonable and the error variance is roughly constant (homoscedasticity).

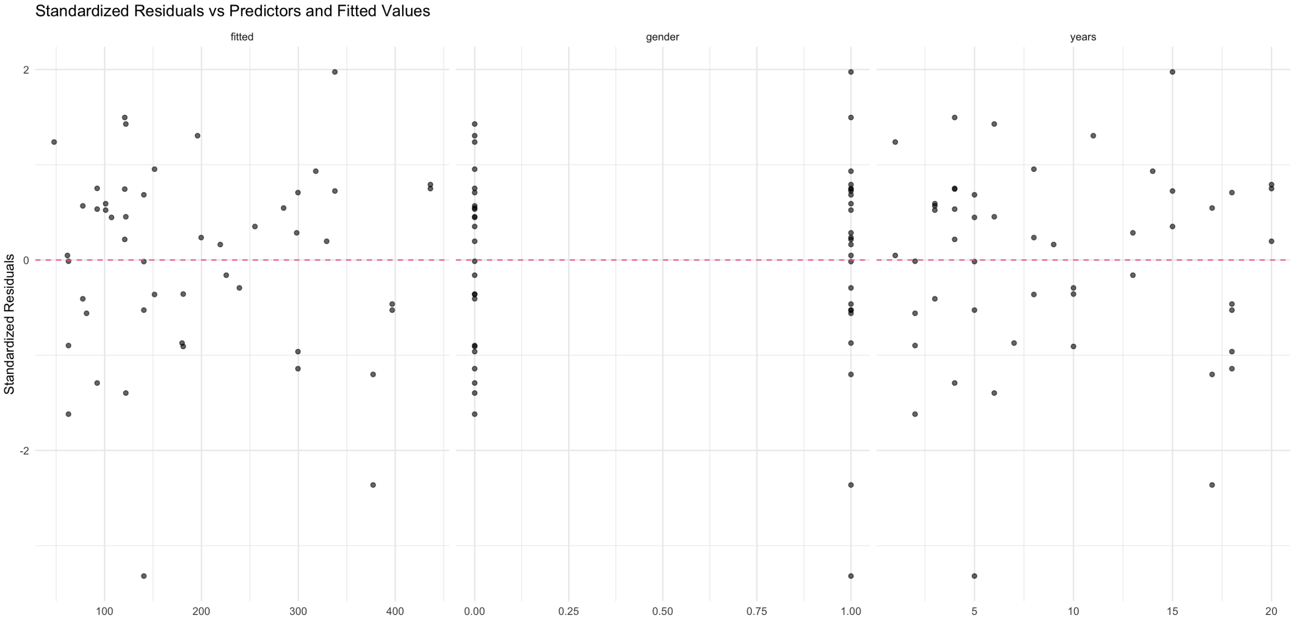

Figure 2: Standardized residual plots against the fitted values and each predictor (gender, years) for assessing multiple linear regression assumptions.

Example 1

Revisit the dataset `NYC.csv`, where the primary objective is to examine how the predictors influence `Price`. Earlier analyses suggested that `Food`, `Decor`, and `East` have significant effects on `Price`, while `Service` has little effect, and interactions with `East` are negligible. The final fitted multiple linear regression model can written as

$$\widehat{\text{Price}} = \hat{\beta}_0 + \hat{\beta}_1 \,\text{Food} + \hat{\beta}_2 \,\text{Decor} + \hat{\beta}_3 \,\text{East}.$$

a. Construct the four standard diagnostic plots and briefly comment on linearity, normality, constant variance, and influence based on these plots.

b. Plot standardized residuals vs. each predictor. Note any curvature (nonlinearity), funneling (heteroscedasticity), or outliers.

Added variable plots

– An added-variable plot (also called a partial regression plot) is a graphical tool used to visualize the relationship between the response variable and a particular predictor in a multiple linear regression model, after adjusting for the effects of the other predictors.

– For a given predictor X_j in the model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon,$$

the added-variable plot for X_j is constructed by:

1. Regressing Y on all predictors except $X_j$and saving the residuals, which represent the variation in Ynot explained by the other predictors.

2. Regressing $X_j$on all the other predictors and saving those residuals, which represent the part of $X_j$not linearly associated with the others.

3. Plotting the residuals from Step 1 against those from Step 2.

– The slope of the fitted line in this plot equals the least squares estimate $\hat{\beta}_j$. Thus, an added-variable plot allows us to visualize the unique contribution of each predictor to the response, detect potential nonlinearities, and identify outliers or influential observations.

– For example, Figure 3 shows the added-variable plots for the model fitted to the Salary 1 dataset. Each panel shows the relationship between the residuals of the response (after removing the effects of the other predictors) and the residuals of the predictor of interest (after adjusting for the others). The fitted blue line in each panel reflects the partial slope for that variable in the multiple regression model. We can observe the following:

- the plot for `years` shows a strong positive linear relationship, indicating that salary increases with years of experience even after adjusting for gender and interaction effects.
- the plot for `gender` shows only a mild upward trend, suggesting a relatively small effect of gender on salary once years are controlled.

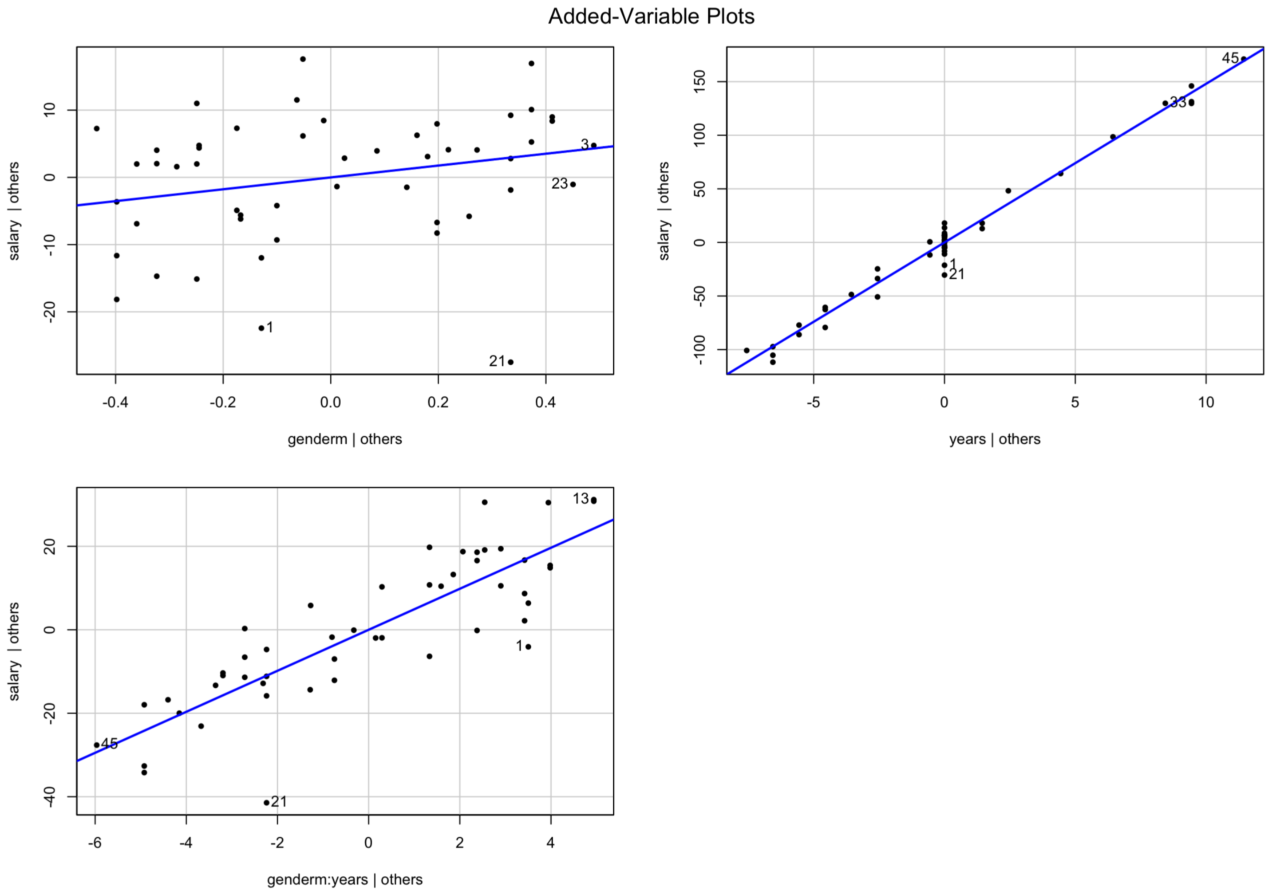

Figure 3: Added-variable plots for the salary model examining the effects of `gender`, `years`, and their interaction (`gender:years`) on `salary`.

- the interaction term (`gender:years`) exhibits a clear positive slope, implying that the rate at which salary increases with years differs between genders.
- no major curvature or outlying points indicate that linearity is reasonable for all predictors.

Example 2

Revisit the `NYC.csv`. Construct the added-variable plot for the final model. Examine each plot and answer the following:

a. What does each plot show about the partial relationship between the predictor and Price?

b. Which predictors show a strong linear trend?

c. Are there any signs of nonlinearity or outliers that may influence the results?

Example 3

The dataset `Fuel.csv` contains information on fuel consumption and related variables for the 50 U.S. states and the District of Columbia (Federal Highway Administration, 2001). The variables are listed in the following table.

Table 1: Variables in the Fuel Consumption Data

Variable	**Description**
`State`	50 states plus DC.
`Drivers`	Number of licensed drivers in the state.
`FuelC`	Gasoline sold for road use (in thousands of gallons).
`Income`	Per person personal income for the year 2000 (in thousands of dollars).
`Miles`	Miles of Federal-aid highways in the state.
`Pop`	2001 population age 16 and over.
`Tax`	State gasoline tax rate (cents per gallon).

The ultimate goal of this analysis is to understand how fuel consumption varies across states. and to practice constructing derived variables (regressors) that remove the effect of population size.

a. To make the variables comparable across states, compute per-capita measures:

$$\text{Dlic} = \frac{\text{Drivers}}{\text{Pop}}, \qquad \text{Fuel} = \frac{\text{FuelC}}{\text{Pop}}.$$

b. Use scatterplots to examine how `Fuel` relates to other variables. Which variables appear to have the strongest linear relationship with fuel consumption?

c. Fit a full regression model using `Dlic`, `Income`, `Miles`, and `Tax` to predict `Fuel`. Interpret the coefficients in context. Which variables appear to be significant predictors of per-capita fuel use?

d. Construct diagnostic plots to assess linearity, constant variance, and normality.

e. Use added-variable (partial regression) plots to visualize the adjusted effect of each predictor. Describe the partial relationship between each predictor and `Fuel`.