A Brief Introduction to Multiple Linear Regression Models

Cheng Peng

STA200 Statistics II

Contents

1 Introduction

We have introduced **simple linear regression (SLR)** with only one numerical or binary categorical predictor variable, as well as **multiple linear regression** with categorical predictor variables. We also demonstrated how to use the simple linear regression model to perform one-sample and two-sample tests, and how to apply multiple linear regression with categorical predictors to conduct one-way and two-way ANOVA analyses.

This note will outline the **general multiple linear regression models**. More detailed diagnostics and remedies will be covered in-depth in subsequent statistical modeling courses.

Before delving into details, we first classify **linear regression** models into several categories based on the **types of predictor variables**. Keep in mind that the response variable in linear regression **must** be a **continuous normal random variable** with **constant variance** (in this and subsequent linear regression courses). Its meaning, however, can be influenced by the predictor variables.

Classification of Linear Regression Models by Predictor Types

- **General Linear Regression Model**: All predictor variables are **continuous** or **numerical** (including frequency counts).
- **Analysis of Variance (ANOVA) Models**: All predictor variables are **categorical**.
- **Analysis of Covariance (ANCOVA) Models**: The model includes **both numerical and categorical** predictor variables.

The term “linear” in **linear regression** does **not** refer to a linear relationship between the response variable y and the predictor variable x. Instead, it describes the **linearity in the regression coefficients**. Below are some examples to clarify this often-misunderstood concept.

- The model $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$ is considered a linear regression model (specifically a quadratic polynomial regression model) because the relationship between the coefficients β_0, β_1, and β_2 is linear. This should be clear since y and x represent observed values in the dataset, while the β terms are the parameters being estimated.
- $y = \beta_0 + \frac{1}{\beta_1 x_1 + \beta_2 x_2} + \beta_3 x_3 + \epsilon$ is **not** a **linear** regression since $\beta_0, \beta_1, \beta_2$, and β_3 are **not** linearly related. In other words, substituting values for x and Y **does not** result in a linear equation!

In the following sections, we will briefly introduce some commonly used regression models that were not covered in previous notes. Our focus will be on:

- Model structures
- Key assumptions
- Data analysis implementation using R (including model diagnostics)

Note that remedial methods for addressing violations of model assumptions will not be covered in detail in this course.

2 Polynomial Regression Models

The polynomial regression is an extension of linear regression that allows for capturing nonlinear relationships between the response y and x. The basic model equations have the following form.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p + \epsilon, \quad \text{where} \;\; \epsilon \to N(0, \sigma)$$

Degrees of polynomial determined the highest power og x in the model, e.g., linear = 1, quadratic = 2, cubic = 3, etc.

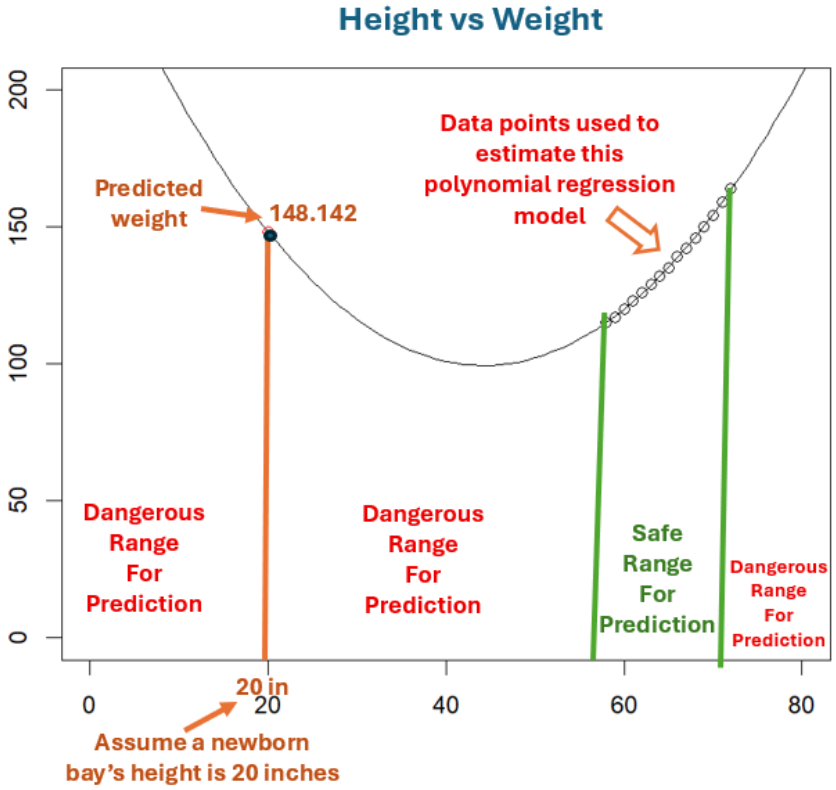
2.1 Model Assumptions and Challenges

In addition to the general assumptions of standard linear regression models apply to polynnomial regression models, polynomial regression models have additional assumptions due to their model structure. For convenience, we list all related assumptions, including those required in the standard linear regression models.

- **Linearity in parameters**: Model is linear in coefficients (though nonlinear in predictors)
- **Independent errors (observations)**: Residuals should **not** be autocorrelated
- **Homoscedasticity**: Constant variance of residuals
- **Normality of errors**: Residuals approximately normally distributed
- **No multicollinearity**: High-order terms (such as x, x^2, x^3) are often highly correlated, which can lead to multicollinearity issues. In practice, we can center the predictor variable to reduce collinearity between these terms. Specifically, instead of using x, x^2, x^3, we can use $x - \bar{x}$, $(x - \bar{x})^2$, and $(x - \bar{x})^3$ in the regression model.
- **Meaningful relationship**: The polynomial form should reflect the true underlying relationship between y and x. This relationship can be **identified** using a scatter plot. To ensure the final model remains interpretable, we **should** avoid high-degree polynomial regression models. A practical guideline is to **limit the degree to no more than 3**.

Addressing violations of model assumptions can be challenging in some cases. Below, we outline these potential practical challenges.

- **Overfitting**: This is a common challenge in modern data analysis. More details on this challenge will be addressed in the subsequent stats courses.
- **Extrapolation danger**: Polynomials can behave erratically outside the observed x-range, which is a common pitfall in practical applications. The following figure illustrates how predictions outside the safe range (extrapolation) can yield erroneous values. For example, this polynomial regression model was trained on a dataset of adult heights and weights. When we input a newborn's height of 20 inches to predict weight, the model outputs **148.142 pounds* - an obviously unrealistic result!



- **Interpretability**: Coefficients become harder to interpret as the degree increases. Keep the order polynomial regression low whenever possible.
- **Multicollinearity**: Polynomial terms are often highly correlated. As mentioned earlier, one can center the x-values
- **Sensitivity to outliers**: More pronounced than in linear regression. This is also an issue for any regression models.

2.2 Numerical Exam and Best Practices

We use an R built-in data set (which is also available at https://pengdsci.github.io/STA200/dataset/mtcars.csv) to illustrate how to build polynomial regression models. Three polynomial models (including a first-order linear regression model).

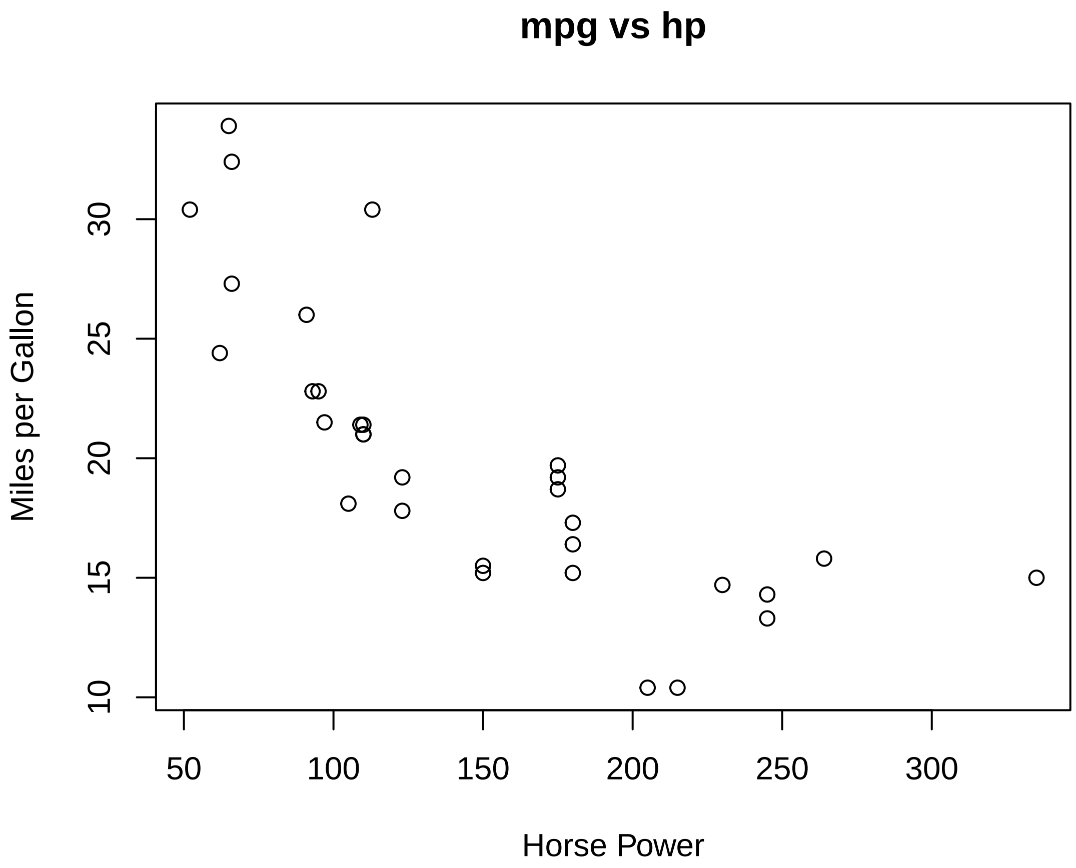
The **mtcars dataset** is a built-in R dataset containing performance and design characteristics for 32 automobiles (1973-74 models). It's commonly used for regression analysis.

- **mpg** Miles/(US) gallon, Continuous

- **cyl** Number of cylinders, 4, 6, 8
- **disp** Displacement (engine size), Cubic inches
- **hp** Gross horsepower, Continuous
- **drat** Rear axle ratio, Continuous
- **wt** Weight, 1000 lbs (tonnes)
- **qsec** 1/4 mile time, Seconds
- **vs** Engine shape, 0 = V-shaped, 1 = Straight
- **am** Transmission, 0 = Automatic, 1 = Manual
- **gear** Number of forward gears, 3, 4, 5
- **carb** Number of carburetors, 1-8

We look at the relationship between `mpg` and `hp`. As pointed out earlier, we make a scatter plot to view the relationship between the response `mpg` and `hp`.

```
mtcars <- read.csv("https://pengdsci.github.io/STA200/dataset/mtcars.csv")
plot(mtcars$hp, mtcars$mpg,
     xlab = "Horse Power",
     ylab = "Miles per Gallon",
     main ="mpg vs hp")
```



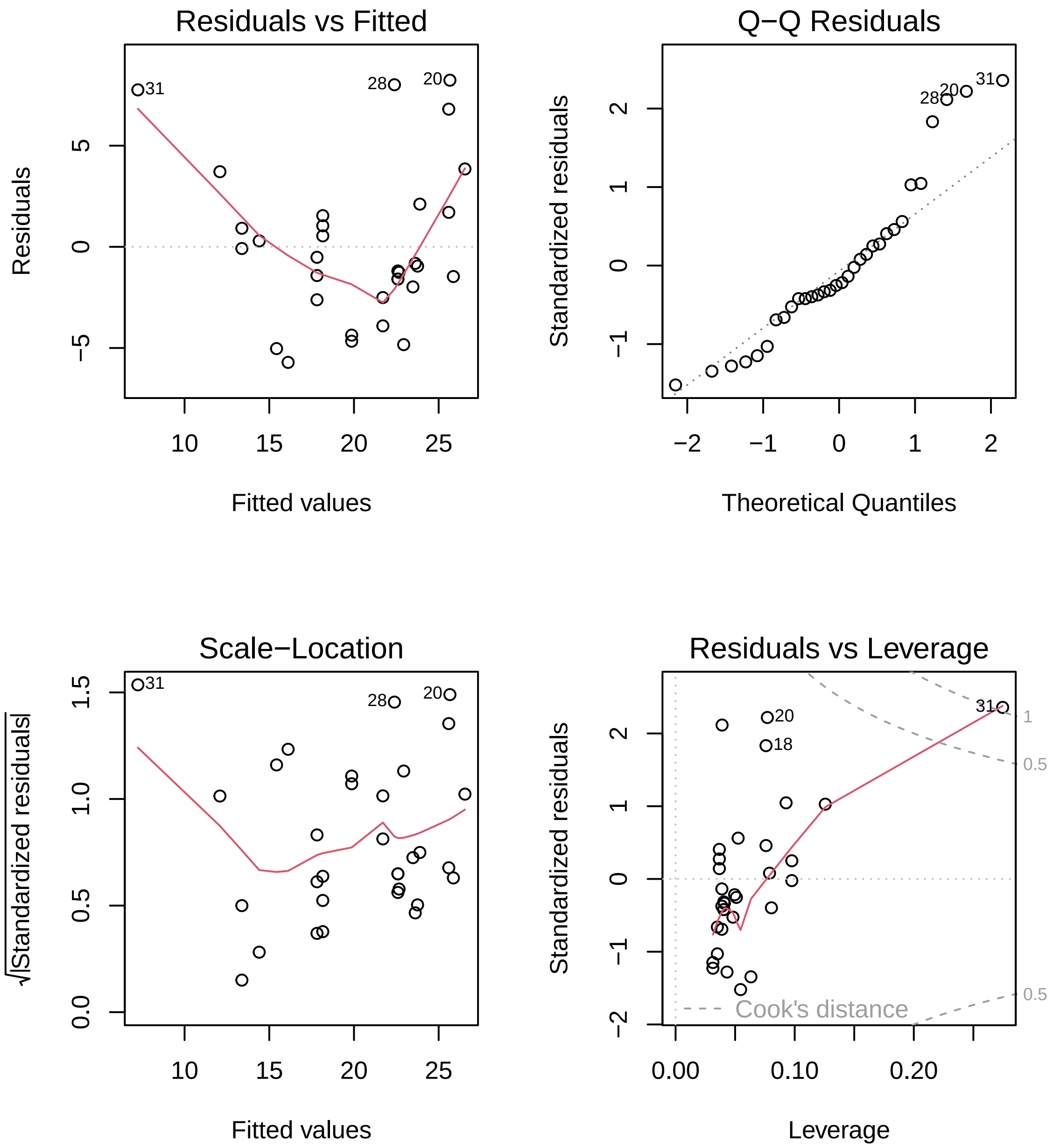
The scatter plot shows a weak curve pattern. We will fit three candidate polynomial regression models and select the best model based on various performance measures.

```
# Fit polynomial models of different degrees
linear_model <- lm(mpg ~ hp, data = mtcars)
```

```
quadratic_model <- lm(mpg ~ hp + I(hp^2), data = mtcars)
cubic_model <- lm(mpg ~ hp + I(hp^2) + I(hp^3), data = mtcars)
```

First-order Linear Regression Model

```
## diagnostic plots
par(mfrow = c(2,2))
plot(linear_model)
```



The `residual vs fitted` plot reveals a curve pattern. The Q-Q plot also indicates a violation of the normality assumption. Apparently, the first-order linear regression model is not the best. Before examining the quadratic regression, we look at the global goodness-of-fit measure.

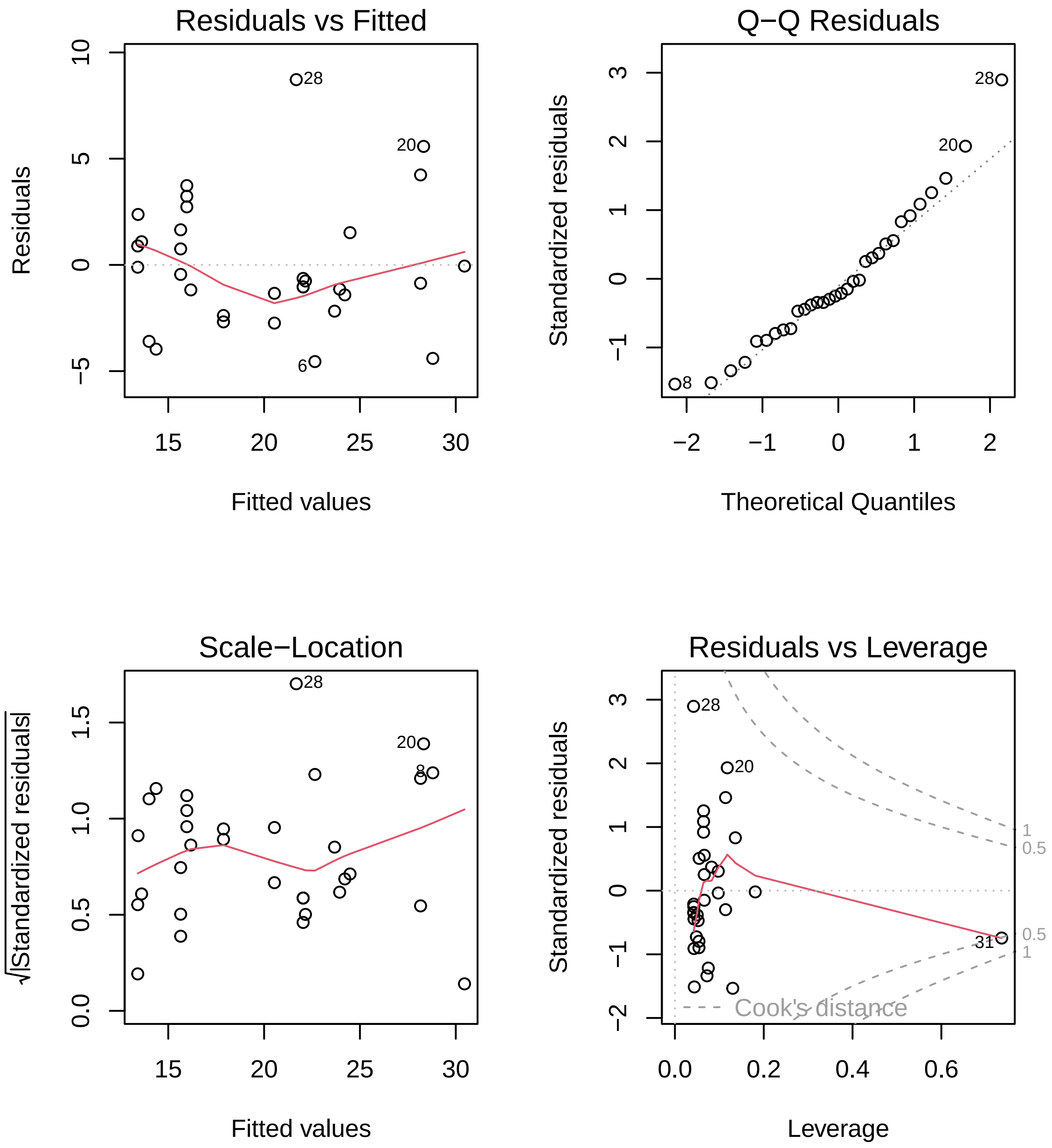
```
summary(linear_model)
```

```
|
| Call:
| lm(formula = mpg ~ hp, data = mtcars)
|
| Residuals:
|     Min      1Q  Median      3Q     Max
| -5.7121 -2.1122 -0.8854  1.5819  8.2360
|
| Coefficients:
|             Estimate Std. Error t value Pr(>|t|)
| (Intercept) 30.09886    1.63392  18.421  < 2e-16 ***
| hp          -0.06823    0.01012  -6.742 1.79e-07 ***
| ---
| Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
|
| Residual standard error: 3.863 on 30 degrees of freedom
| Multiple R-squared:  0.6024,  Adjusted R-squared:  0.5892
| F-statistic: 45.46 on 1 and 30 DF,  p-value: 1.788e-07
```

The coefficient of determination $R^2 = 0.6024$.

Quadratic Regression Model

```
## diagnostic plots
par(mfrow = c(2,2))
plot(quadratic_model)
```



All residual plots of the quadratic regression model look fine. No obvious violations of model assumptions. Next, we look at the coefficient of determination.

```
summary(quadratic_model)
```

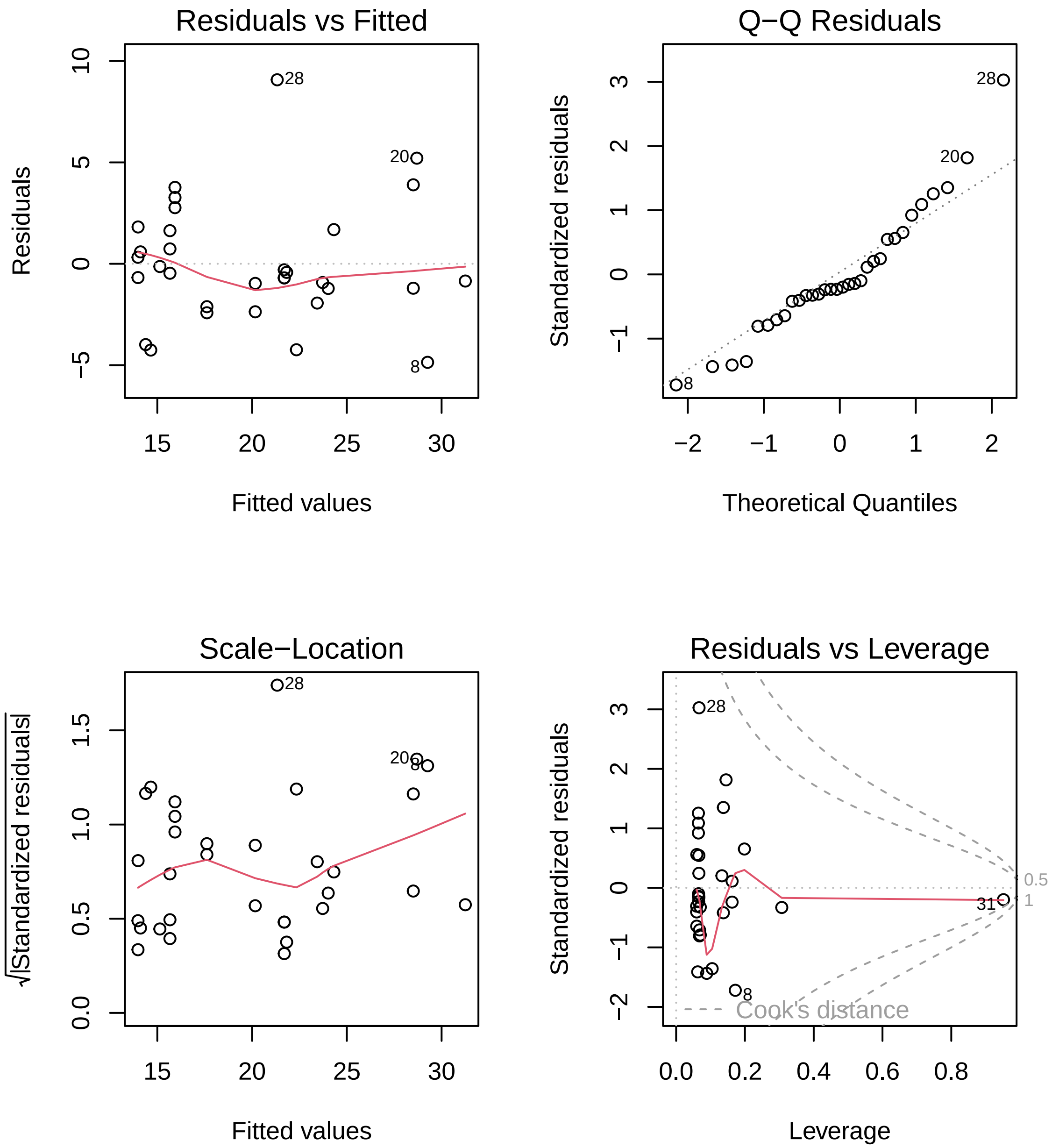
```
|
| Call:
| lm(formula = mpg ~ hp + I(hp^2), data = mtcars)
|
| Residuals:
|     Min      1Q  Median      3Q     Max
| -4.5512 -1.6027 -0.6977  1.5509  8.7213
```

```
|
| Coefficients:
|               Estimate Std. Error t value Pr(>|t|)
| (Intercept)  4.041e+01  2.741e+00  14.744 5.23e-15 ***
| hp          -2.133e-01  3.488e-02  -6.115 1.16e-06 ***
| I(hp^2)      4.208e-04  9.844e-05   4.275 0.000189 ***
| ---
| Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
|
| Residual standard error: 3.077 on 29 degrees of freedom
| Multiple R-squared:  0.7561,  Adjusted R-squared:  0.7393
| F-statistic: 44.95 on 2 and 29 DF,  p-value: 1.301e-09
```

The quadratic model is much better than the first-order linear regression, The $R^2 = 0.756$. Next, we look at cubic regression to see whether quadratic regression can be further improved.

Cubic Polynomial Regression

```
## diagnostic plots
par(mfrow = c(2,2))
plot(cubic_model)
```



The residual plot looks similar to the quadratic regression model. No obvious violation of the model assumption was found. We look at R^2 and compare it with that of the quadratic regression model.

```
summary(cubic_model)
```

```
|
| Call:
| lm(formula = mpg ~ hp + I(hp^2) + I(hp^3), data = mtcars)
|
| Residuals:
|     Min      1Q  Median      3Q     Max
| -4.8605 -1.3972 -0.5736  1.6461  9.0738
```

```
|
| Coefficients:
|               Estimate Std. Error t value Pr(>|t|)
| (Intercept)  4.422e+01  5.961e+00   7.419 4.43e-08 ***
| hp          -2.945e-01  1.178e-01  -2.500   0.0185 *
| I(hp^2)      9.115e-04  6.863e-04   1.328   0.1949
| I(hp^3)     -8.701e-07  1.204e-06  -0.722   0.4760
| ---
| Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
|
| Residual standard error: 3.103 on 28 degrees of freedom
| Multiple R-squared:  0.7606,  Adjusted R-squared:  0.7349
| F-statistic: 29.65 on 3 and 28 DF,  p-value: 7.769e-09
```

The coefficient of determination is 0.7606. Recall that R^2 of the quadratic regression model is 0.7561. The improvement is

$$\frac{0.7606 - 0.7561}{0.7561} = 0.59\%.$$

The improvement is insignificant. We stay with the quadratic regression.

Remarks:

- **Prediction Focus in Polynomial Regression**: Our primary objective is prediction rather than inference about regression coefficients. To ensure meaningful predictions, avoid extrapolation beyond the observed data range.
- **Centering Predictors**: While we did not center predictors in the three candidate models presented above, note that centering is recommended in practice when collinearity between polynomial terms leads to estimation challenges in regression coefficients.
- **Multiple Predictor Extension**: Polynomial regression can be extended to models with multiple predictor variables, following the same fundamental principles.

The following YouTube video (https://www.youtube.com/watch?v=ZYN0YD7UfK4) explains a polynomial example using a real-world data set. We will use the same dataset in the next section when discussing the ANCOVA model.

3 Analysis of Covariance Regression Model

An **Analysis of Covariance (ANCOVA)** model incorporates both categorical and continuous predictor variables. It is essential to convert categorical variables to factor variables before including them in the model formula. This step becomes particularly critical when categorical variables use numeric labels. If numerically labeled categorical variables are not properly converted to factors, R will incorrectly treat them as continuous numeric variables, leading to erroneous analysis results.

3.1 The Model Structure

Linear regression models the relationship between a continuous response variable Y and one or more predictor variables (numerical or categorical). For ease of illustration, we discuss the case of one numerical predictor variable and a categorical variable with k levels. That is, the model has the following predictor variables.

- A numerical predictor X
- A categorical predictor G (with k levels, requiring $k-1$ dummy variables)
- An interaction term $X \times G$.

Recall that, for a k level factor variable, we need to define $k-1$ dummy variables to represent the original categorical variable. Let $G_2, G_3, \cdots, G_k$ be the $k-1$ dummy variables (baseline category does not show up in the model formula). Wth the above notation, the model formula is explicitly given by

$$Y = \beta_0 + \beta_1 X + \sum_{i=2}^{k} \beta_i G_i + \sum_{i=2}^{k} \gamma_i (X \times G_i) + \epsilon$$

Where $\epsilon \to N(0, \sigma)$. $X \times G_i$ $(i = 2, 3, \cdots, k)$ are interaction terms.

- β_0: Intercept (baseline mean when $X = 0$ and G is at reference level)
- β_1: Slope of X for the reference group
- β_i: Shift in intercept for group i
- γ_i : Change in slope of X for group i
- ϵ: Random error (assumed $\epsilon \to N(0, \sigma)$)

R defines the above model formula internally. The explicit formula used in the code has the form `y ~ X + G + X*G`.

Key Assumptions Similar assumptions are required in ANCOVA.

- **Linearity**: The Relationship between predictors and response is linear.
- **Independence**: Observations are uncorrelated.
- **Homoscedasticity**: Constant variance of residuals.
- **Normality**: Residuals are normally distributed.

Interpretation of the Interaction Term

An interaction between X and G means the effect of X on Y depends on the level of G.

- If interaction is significant, the slopes for X differ across groups.
- If interaction is not significant, the effect of X is the same across groups (parallel but intercepts may differ).

3.2 ANCOVA Model By Example

For illustration, we use the **depression dataset**, which is available at https://pengdsci.github.io/STA200/dataset/depression.csv.

Some researchers (Daniel, 1999) were interested in comparing the effectiveness of three treatments for severe depression. For the sake of simplicity, we denote the three treatments A, B, and C. The researchers collected the following data (Depression Data) on a random sample of $n = 36$ severely depressed individuals:

- Y_i = measure of the effectiveness of the treatment for individual i;
- X_i = age (in years) of individual i;
- Z_i = treatment for individul i;

Three treatments were used in this study: A, B, and C. Let A be the reference level. Two dummy variables are defined as

- $Z_i^B = 1$ if treatment B was used on individual i, otherwise, $Z_i^B = 0$
- $Z_i^C = 1$ if treatment C was used on individual i, otherwise, $Z_i^C = 0$

The explicit expression of this special ANCOVA is given by

$$Y = \beta_0 + \beta_1 X + \beta_2 Z^B + \beta_3 Z^C + \gamma_1 X \times Z^B + \gamma_2 X \times Z^C + \epsilon$$

Next, let's examine how the interaction effect influences the regression equation.

Case I: The regression line for **treat A** (baseline treatment group). In this case, $Z^B = Z^C = 0$, the initial regression equation is reduced to

$$Y = \beta_0 + \beta_1 X.$$

Case II: The regression line for **treat B**: In this case, $Z^B = 1$ and $Z^C = 0$. The initial regression equation is reduced to the following form

$$Y = \beta_0 + \beta_1 X + \beta_2 + \gamma_1 X = (\beta_0 + \beta_2) + (\beta_1 + \gamma_1)X$$

Case III: The regression line for **treat C**. In this case, $Z^B = 0$ and $Z^C = 1$

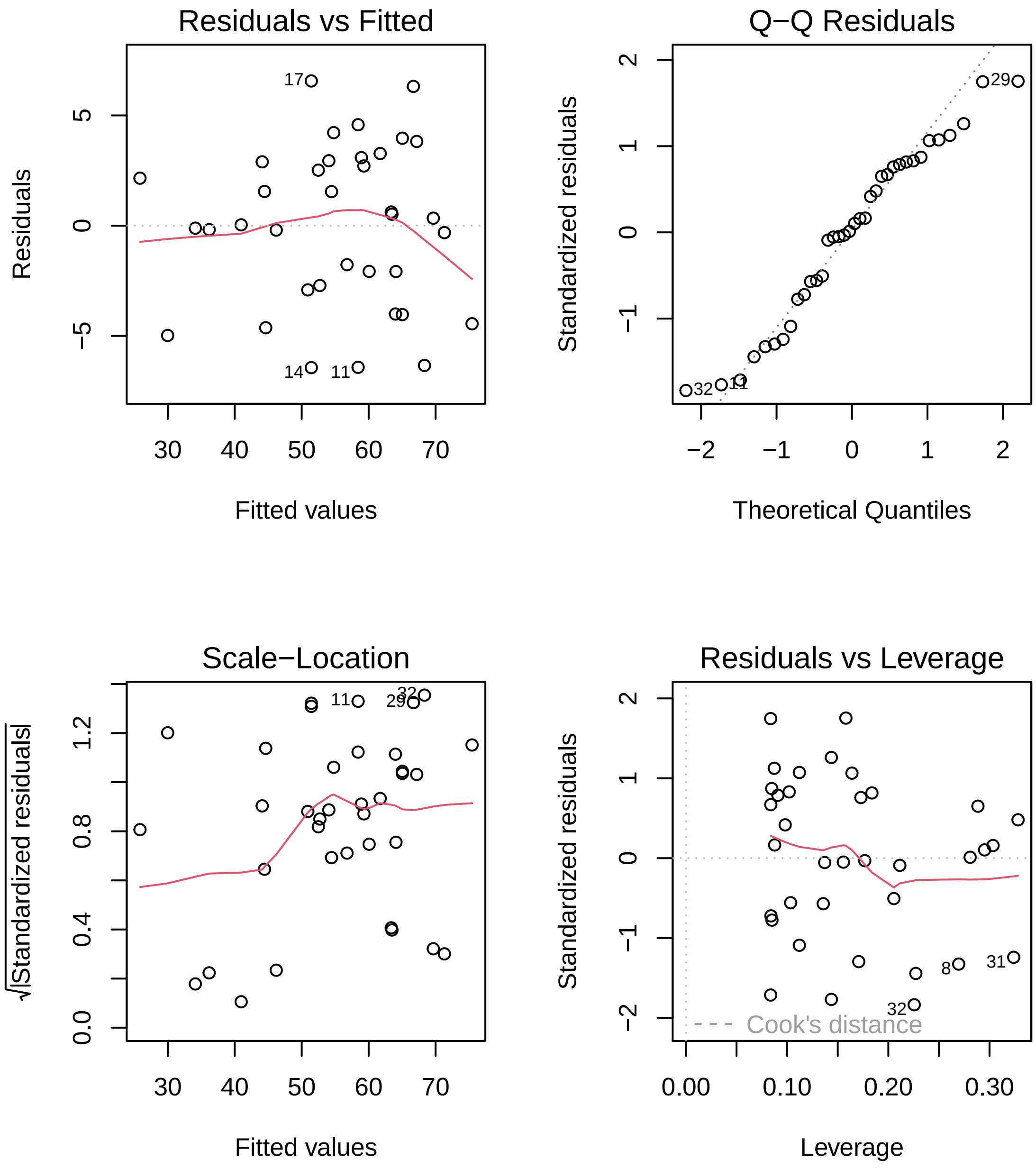
$$Y = \beta_0 + \beta_1 X + \beta_3 + \gamma_2 X = (\beta_0 + \beta_3) + (\beta_1 + \gamma_2)X$$

We can see from the above three regression equations associated with the three treatment groups that when the interaction effect is significant, the following hold:

- Each treatment group has a distinct regression line intercept.
- Each treatment group has a distinct regression line slope.

We now implement ANCOVA on the depression data. The object is to assess how treatments influence the effectiveness of the treatments **adjusted by age**.

```r
## load data to R
depression <- read.csv("https://pengdsci.github.io/STA200/dataset/depression.csv")
## ANCOVA with interaction
dep.lm <- lm(y ~ x * z, data = depression)
## Residual diagnostics
par(mfrow = c(2,2))    # create a lattice with four graph panels
plot(dep.lm)
```



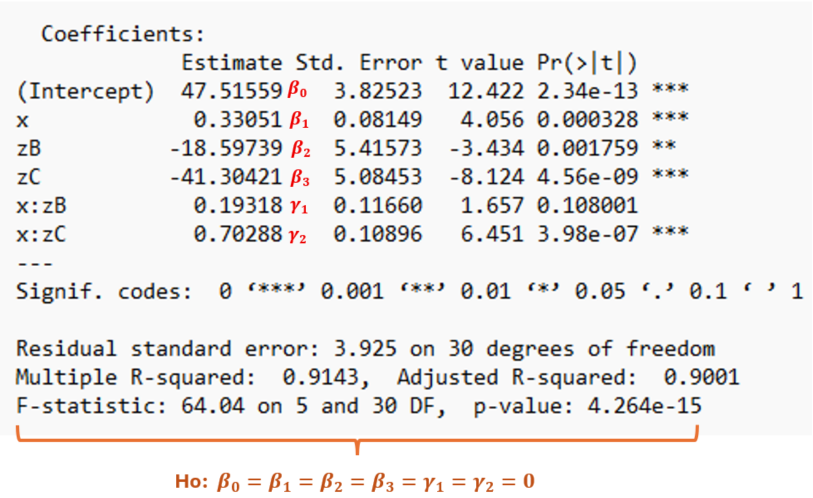
The residual plots did not reveal obvious violations of the model assumptions. The initial model is valid. We next explicitly write the three regression lines associated with the three treatment groups.

```
summary( dep.lm)
```

```
|
| Call:
| lm(formula = y ~ x * z, data = depression)
|
| Residuals:
|     Min      1Q  Median      3Q     Max
| -6.4366 -2.7637  0.1887  2.9075  6.5634
```

```
|
| Coefficients:
|              Estimate Std. Error t value Pr(>|t|)
| (Intercept)  47.51559    3.82523  12.422 2.34e-13 ***
| x             0.33051    0.08149   4.056 0.000328 ***
| zB          -18.59739    5.41573  -3.434 0.001759 **
| zC          -41.30421    5.08453  -8.124 4.56e-09 ***
| x:zB          0.19318    0.11660   1.657 0.108001
| x:zC          0.70288    0.10896   6.451 3.98e-07 ***
| ---
| Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
|
| Residual standard error: 3.925 on 30 degrees of freedom
| Multiple R-squared:  0.9143,  Adjusted R-squared:  0.9001
| F-statistic: 64.04 on 5 and 30 DF,  p-value: 4.264e-15
```

The following annotated output labels the estimated regression coefficients and the null hypothesis of the goodness-of-fit F test.



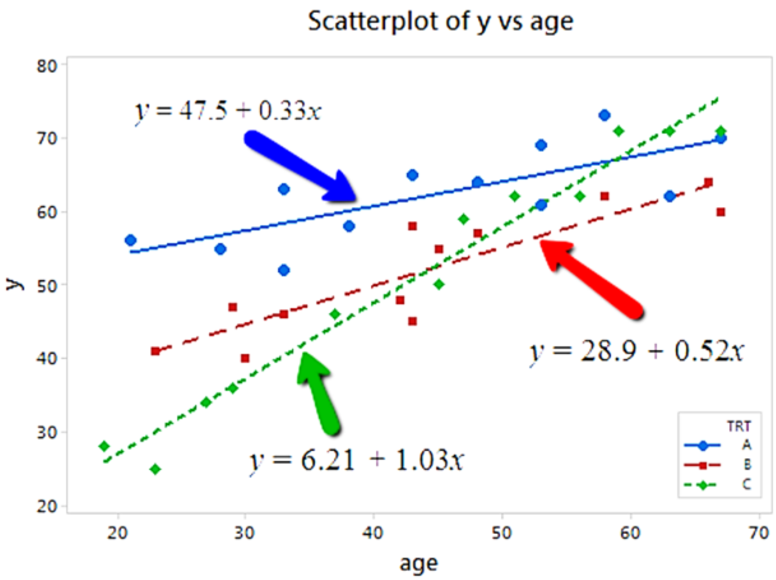
```
  Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  47.51559   3.82523  12.422 2.34e-13 ***
x             0.33051   0.08149   4.056 0.000328 ***
zB          -18.59739   5.41573  -3.434 0.001759 **
zC          -41.30421   5.08453  -8.124 4.56e-09 ***
x:zB          0.19318   0.11660   1.657 0.108001
x:zC          0.70288   0.10896   6.451 3.98e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.925 on 30 degrees of freedom
Multiple R-squared:  0.9143,  Adjusted R-squared:  0.9001
F-statistic: 64.04 on 5 and 30 DF,  p-value: 4.264e-15
```

Ho: $\boldsymbol{\beta_0 = \beta_1 = \beta_2 = \beta_3 = \gamma_1 = \gamma_2 = 0}$

The three regression lines are given by

- **Treatment A**: $y = \beta_0 + \beta_1 X = 47.51559 + 0.33051X$
- **Treatment B**: $y = (\beta_0 + \beta_2) + (\beta_1 + \gamma_1)X = 28.9182 + 0.52369X$
- **Treatment C**: $y = (\beta_0 + \beta_3) + (\beta_1 + \gamma_2)X = 6.21138 + 1.03339X$

The above three regression lines and the scatter plots of the corresponding treatment groups are shown in the following figure.



We can see that the slopes and intercepts of the three regression lines differ. The interaction effect is reflected in the differing slopes, which implies intersections between the regression lines.

The following YouTube video provides an example that is similar to the **depression** data (https://www.youtube.com/watch?v=8YuuIsoYqsg).

4 Concluding Remarks

We briefly introduced the basic multiple linear regression models in this and the previous notes: polynomial, ANOVA, and ANCOVA.

Polynomial regression extends simple linear regression by modeling nonlinear relationships through higher-order terms (e.g., quadratic or cubic). This model is useful when the relationship between predictors and the response variable is curved, allowing for flexibility in fitting complex patterns. However, higher-degree polynomials can lead to overfitting, so careful model selection is required.

ANOVA is a special case of MLR where all predictors are categorical (e.g., group membership). It tests whether the means of different groups are statistically different by comparing between-group and within-group variability. ANOVA is widely used in experimental designs to assess the impact of categorical factors (e.g., treatment effects).

ANCOVA combines ANOVA and regression by including both categorical and continuous predictors. It adjusts for the effect of continuous covariates (confounders) while comparing group means. ANCOVA improves precision by accounting for baseline differences, commonly used in clinical trials and observational studies.

These three models, **polynomial regression** (nonlinear fits), **ANOVA** (categorical group comparisons), and **ANCOVA** (mixed categorical-continuous adjustments), expand the flexibility of MLR to address diverse research questions. Choosing the appropriate model depends on the nature of the predictors and the hypothesized relationships in the data.