Linear Regression Models with Categorical Predictors

Cheng Peng

STA200 Statistics II

Contents

1 Introduction

We have briefly introduced the simple linear regression model earlier with an explicit expression.

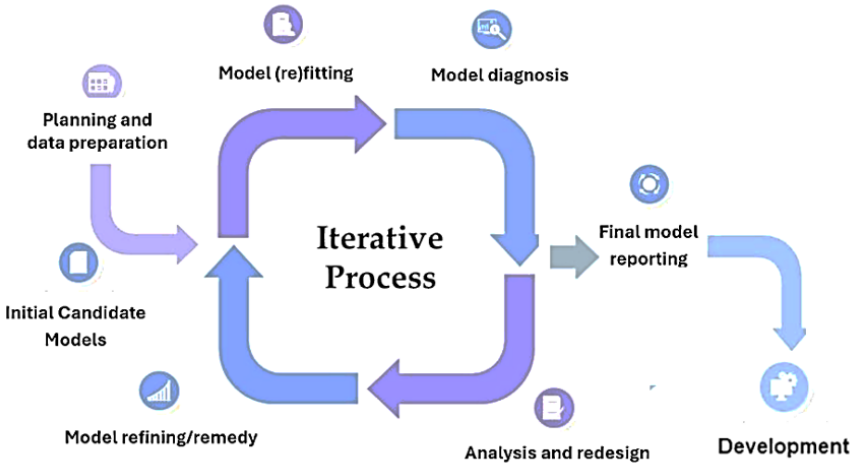
$$y = \beta_0 + \beta_1 x + \epsilon,$$

Where y is a continuous random numerical variable, x is either a numerical variable or a binary categorical variable, and $\epsilon \to N(0, \sigma)$, the basic applications of a simple linear regression model are twofold:

(1) **Assess** the linear correlation between the response variable (y) and the predictor variable (x).

(2) **Predict** the response for given new values of the predictor variable x.

In general, regression modeling is an iterative process in real-world applications, as illustrated in the following chart.



In the course project and data analysis, we focus on the iterative loop of model identification and reporting on the final model. To be more specific, assuming we have an analytic data set that is ready for modeling,

- **Select appropriate candidate models to address the practical questions** - This includes checking model assumptions and ensuring the analytic dataset contains sufficient information for the candidate models. *We have highlighted all assumptions when introducing new models in the previous notes.*
- **Fit candidate models to the data** - This step involves verifying the outputs to ensure the model parameters (e.g., coefficients) are appropriately estimated.

  Note: *the fitted models in this step may not be valid or optimal, we should not interpret the model at this point.* **We only report the final model obtained from the iterative model identification process.**
- **Model diagnostics** - This is the most crucial step in any regression analysis, where we assess whether all assumptions of the candidate models are satisfied. For example, we apply residual diagnostic methods (commonly used in linear regression) to evaluate linearity, normality, homoscedasticity (constant variance), and other key assumptions. If no violations are detected, we then use application-specific performance measures to select the best model for reporting and implementation.
- **Model remedy and refinement** - If one or more assumptions are violated, we need to use various methods to fix the problems. *We will not discuss this major topic in this class, but will cover this critical topic in detail in subsequent courses.*
- **Enter the refined/modified model into the loop for the next iteration** - The refined or modified model will be refitted, analyzed, and diagnosed to determine whether any violations remain. This process continues iteratively until the final valid and optimal model is identified.
- **Final model reporting** - This topic will not be emphasized in this class but will be highlighted in subsequent statistical modeling classes.

In this note, we will extend simple linear regression to multiple linear regression by incorporating two or more predictor variables into the model. We begin by revisiting the simple linear regression model discussed earlier this semester. Next, we will generalize the binary categorical predictor to multi-category predictors through the use of binary indicator variables – effectively creating a special case of multiple linear regression with multiple binary categorical variables. This approach integrates ANOVA analysis into the regression framework.

2 Simple Linear Regression Revisited

The term **simple linear regression** (SLR) simply means that the regression equation has the form $y = \beta_0 + \beta_1 x + \epsilon$. The predictor variable is either a continuous variable or a binary categorical variable that takes only two possible distinct values, such as *success* vs *failure*, *disease* vs *disease-free*, etc. **The response must be a continuous normal variable whose mean may be influenced by the predictor x but not the variance.**

The key assumptions for SLR are

- **linearity**: linear relationship between y and x;
  - diagnostic check: scatter plot of x and y
- **deterministic predictor** x: the predictor variable is not a random variable.
  - No direct information, check this assumption
- **normality**: the only random variable y is a normal random variable with density function $N(\beta + 0 + \beta_1 x, \sigma)$, or $\epsilon \to N(0, \sigma)$.
  - diagnostic check: Q-Q plot
- **constant variance**: this assumption is contained in the above **normality** assumption.
  - diagnostic check: residual plot, studentized residual plot
- **Influential observations**: This is also related to the **normality assumption**
  - diagnostic check: leverage plot (Cook's distance)

2.1 SLR with A Continuous Predictor

When the predictor variable is continuous, the interpretation of the slope β_1 reflects the change in y when x increases by one unit. The sign of β_1 reflects the direction of linear association between x and y. Next, we use a numerical example to illustrate the regression modeling process. The data set can be found at https://pengdsci.github.io/STA200/dataset/EduWage.csv. The R function `lm()` will be used to perform regression analysis.

The variables in the data set are defined as

- **wage**: average hourly earnings
- **educ**: years of education
- **exper**: years potential experience
- **tenure**: years with current employer
- **lwage**: log(wage) - in economics, it is called log wage.
- **region**: the geographic region of the respondent
- **smsa**: A Standard Metropolitan Statistical Area (SMSA) is a geographical region defined by the U.S. Office of Management and Budget that consists of a core urban area with a substantial population, along with adjacent communities that have a high degree of economic and social integration with that core. SMSAs are used for statistical purposes to analyze urbanization, population density, and economic activities in metropolitan areas, providing insights into urban data trends.

Step 1: Load data and explore the dataset.

```
## Load the dataset
edu.wage <- read.csv("https://pengdsci.github.io/STA200/dataset/EduWage.csv")
## summary of variables in the dataset
summary(edu.wage)
```

```
|       wage            educ          exper          tenure
|  Min.   : 0.530   Min.   : 0.00   Min.   : 1.00   Min.   : 0.000
|  1st Qu.: 3.330   1st Qu.:12.00   1st Qu.: 5.00   1st Qu.: 0.000
|  Median : 4.650   Median :12.00   Median :13.50   Median : 2.000
|  Mean   : 5.896   Mean   :12.56   Mean   :17.02   Mean   : 5.105
|  3rd Qu.: 6.880   3rd Qu.:14.00   3rd Qu.:26.00   3rd Qu.: 7.000
|  Max.   :24.980   Max.   :18.00   Max.   :51.00   Max.   :44.000
```

```
|      lwage           region              smsa
|  Min.   :-0.6349   Length:526         Min.   :0.0000
|  1st Qu.: 1.2030   Class :character   1st Qu.:0.0000
|  Median : 1.5369   Mode  :character   Median :1.0000
|  Mean   : 1.6233                      Mean   :0.7224
|  3rd Qu.: 1.9286                      3rd Qu.:1.0000
|  Max.   : 3.2181                      Max.   :1.0000
```

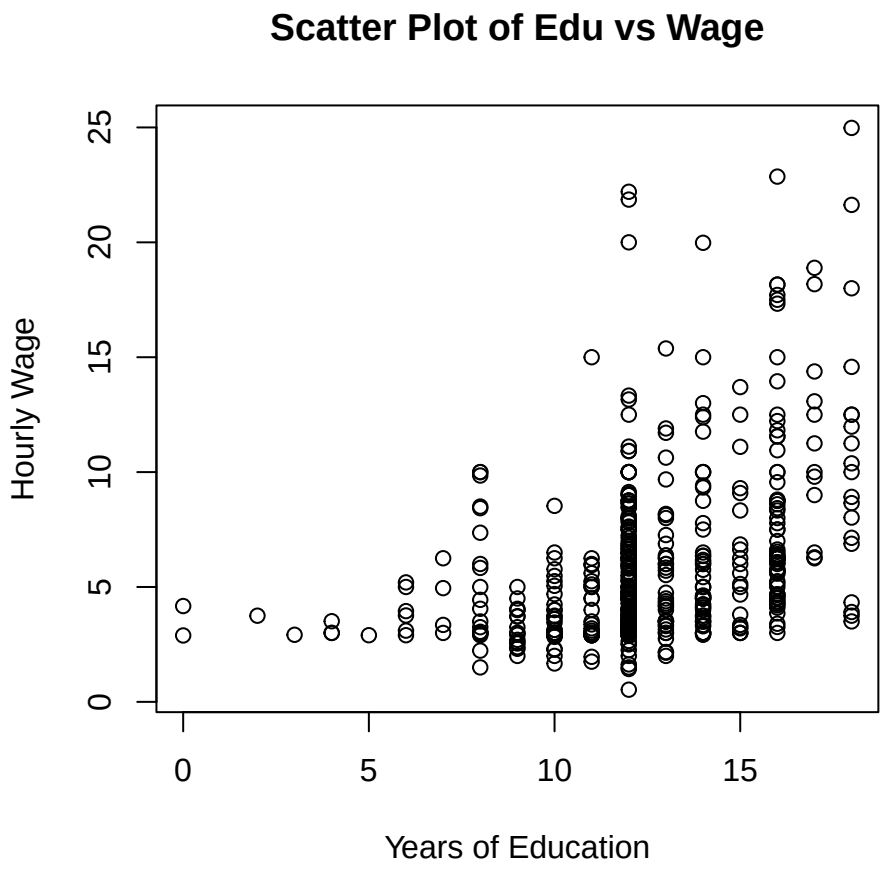
The R function `summary(dataset.name)` returns a five-number summary of all numerical variables in the dataset. There is a categorical variable `region` in the dataset. To see the distribution of **categorical variables**, we use the R function `table(variable.name)` to see the frequency distribution of the categorical variables.

```
table(edu.wage$region)
```

```
|
| northcen    other    south     west
|      132      118      187       89
```

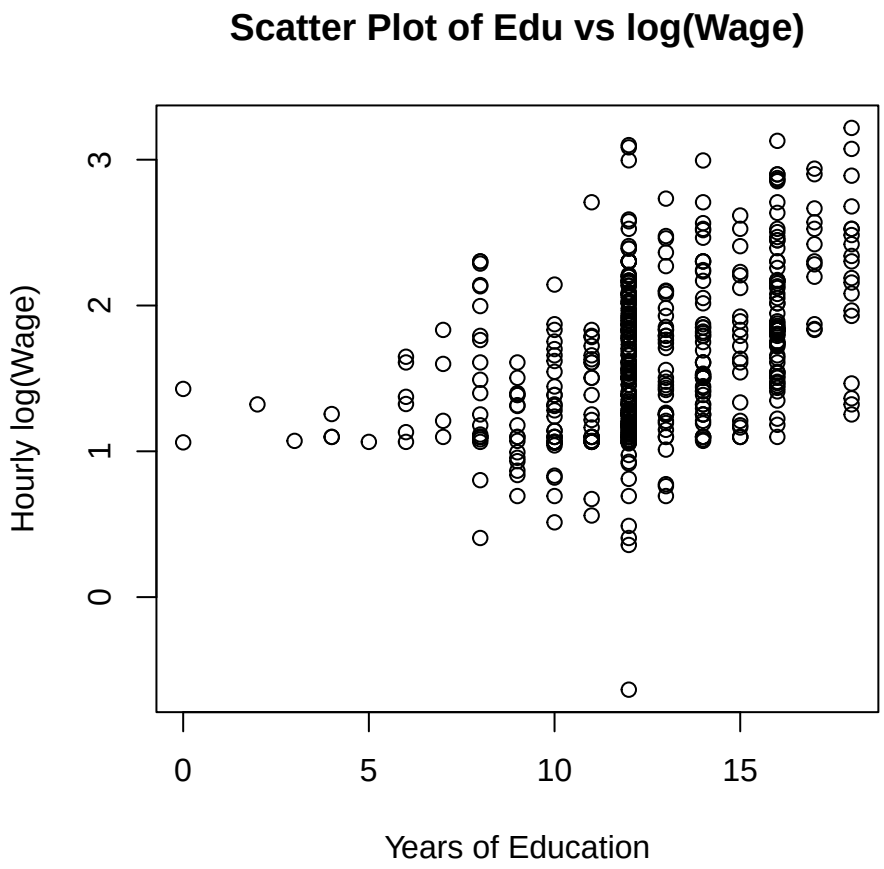
Step 2: objective and candidate model - we examine whether education affects the wage. The candidate model to address this objective will be the simple linear regression model. We first make a scatter plot of `educ` (horizontal axis) and `wage` (vertical axis).

```
plot(edu.wage$educ,                 # horizontal axis
     edu.wage$wage,                 # vertical axis
     xlab = "Years of Education",   # horizontal label
     ylab = "Hourly Wage",          # vertical label
     main = "Scatter Plot of Edu vs Wage"    # title of the plot
     )
```



The above plot shows a linear relationship but also shows non-constant variance. This violates the SLR assumption. Economic studies show that logarithmic wage has less variation. We next try to log wages in the SLR. Before building the model, we make a scatter plot of log wage vs years of education.

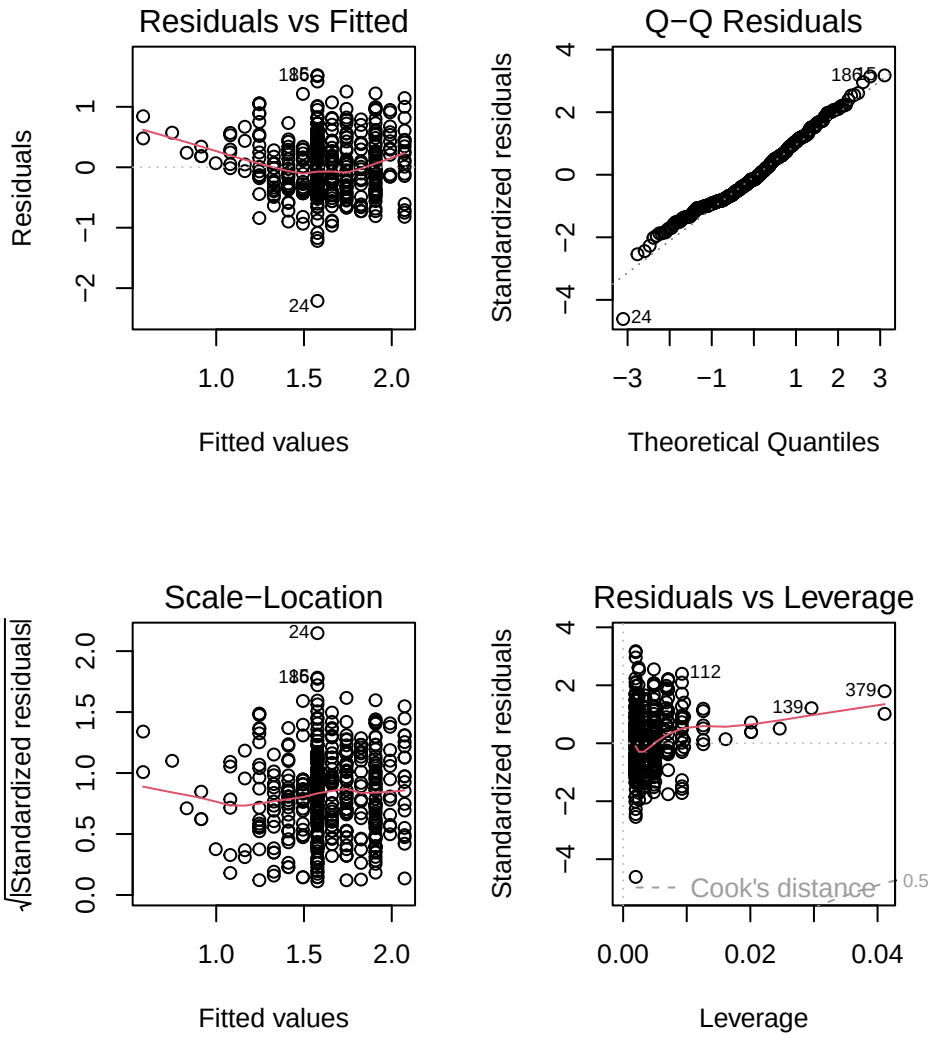
```
plot((edu.wage$educ)[-1],                  # horizontal axis
     (edu.wage$lwage)[-1],                  # vertical axis
     xlab = "Years of Education",   # horizontal label
     ylab = "Hourly log(Wage)",           # vertical label
     main = "Scatter Plot of Edu vs log(Wage)"    # title of the plot
     )
```



The above scatter indicates that the log wage is a better response for SLR.

Step 3: Fitting SLR: log wage $= \beta_0 + \beta_1$educ and perform residual diagnostics.

```
## linear model
lgw.model <- lm(lwage ~ educ, data = edu.wage)
## residual plot
par(mfrow=c(2,2))    # create a graphic grid with 4 graphical cells
plot(lgw.model)
```



Interpretations of Residual Diagnostic Plots:

- **Residual vs Fitted** - shows a minor violation of the assumption of constant variance (the variance increases as the fitted value increases)
- **Q__Q Residual Plot** - There is no significant violation of the normal distribution
- **Scale-Location Plot** - no serious violation except a few outliers (not influential)
- **Residuals-Leverage** - shows no influential points.

Overall, there is a minor violation of the constant variance assumption. We will learn methods in subsequent courses to refine the model. For illustration, we decided to report the above model.

Step 4 - summarize the log wage model

```
summary(lgw.model)
```

```
|
| Call:
| lm(formula = lwage ~ educ, data = edu.wage)
|
```

```
| Residuals:
|      Min       1Q   Median       3Q      Max
| -2.21158 -0.36393 -0.07263  0.29712  1.52339
|
| Coefficients:
|             Estimate Std. Error t value Pr(>|t|)
| (Intercept) 0.583773   0.097336   5.998 3.74e-09 ***
| educ        0.082744   0.007567  10.935  < 2e-16 ***
| ---
| Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
|
| Residual standard error: 0.4801 on 524 degrees of freedom
| Multiple R-squared:  0.1858,  Adjusted R-squared:  0.1843
| F-statistic: 119.6 on 1 and 524 DF,  p-value: < 2.2e-16
```

Brief Report:

The estimated slope coefficient $\hat{\beta}_1 = 0.082744$ indicates that log wage increases by 0.082744 for each additional year of education ($p \approx 0$). The coefficient of determination $R^2 \approx 0.1858$ shows that the log wage and years of education have a weak linear correlation.

Remarks: Here are a few comments on the general principles of regression analysis

- In this example, we use only two variables in the data set with more than 2 variables to illustrate how to implement an SLR. In practice, if you have more information (more variables) available in the data, we should use all relevant information - **this is the general principle of all data analysis**. This means we need to use multiple regression to include two or more predictor variables! We will cover general MLR in the next note.
- **Implication of SLR**: An implication of SLR is that all other variables do not affect the response variable y. This also implies that, in general, an SLR can never be an optimal model if it is invalid.
- **Choose an optimal subset of predictors**: This means we always start with multiple candidate models in multiple linear regression (MLR) models. For example, if we have two variables x_1 and x_2, we can fit three different **first order** (in predictor variables) MLR based on the combinations of x_1, x_2, and $x_1 + x_2$. More on different MLRs will be discussed in the next note.

2.2 SLR with A Binary Predictor

We have discussed the regression approach to the two-sample t-test at the beginning of the semester. Here we use the sample wage dataset and a binary categorical `smsa`(Standard Metropolitan Statistical Area) to see whether the average wage in **rural** and **urban** areas is different.

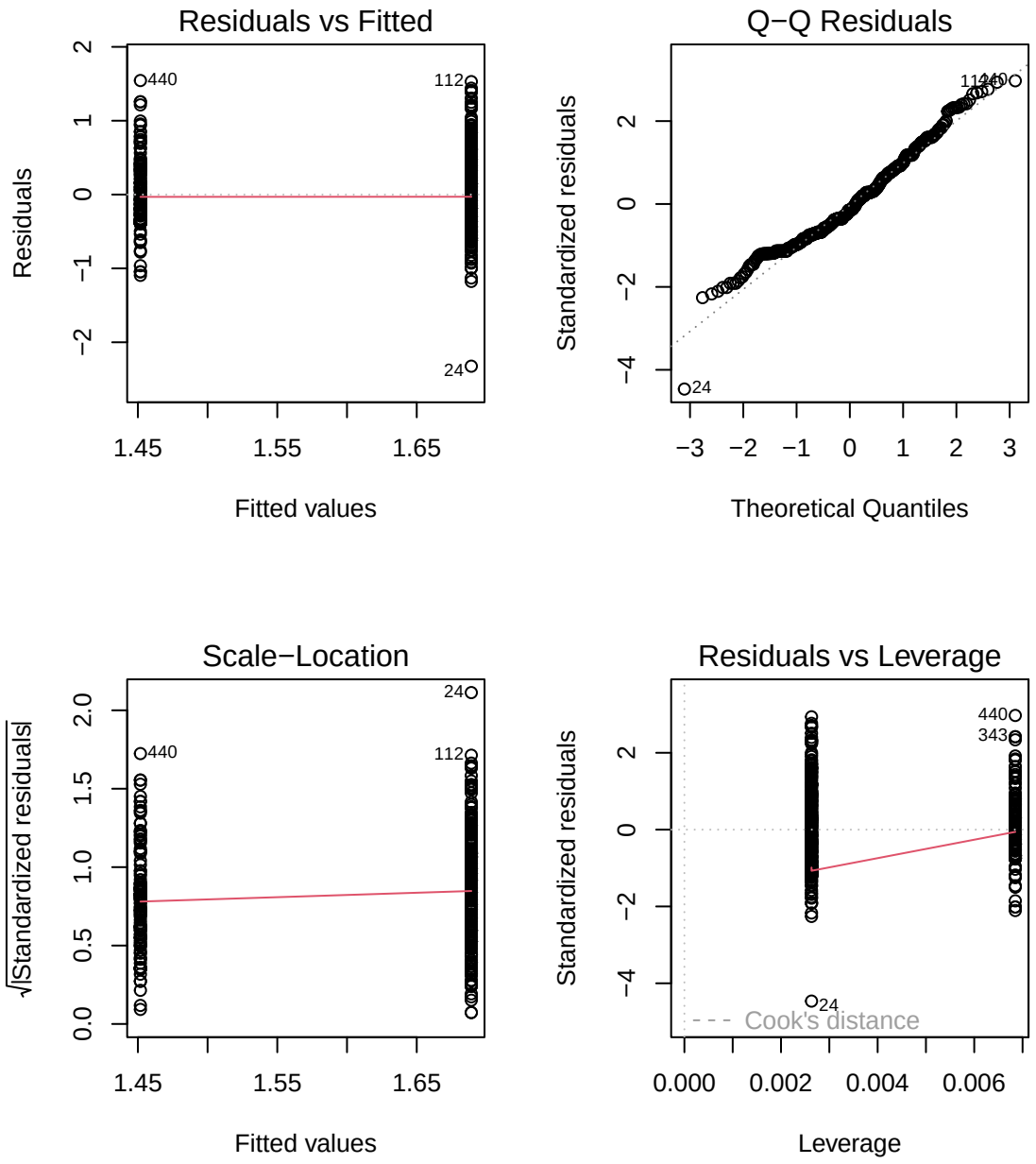
Recall the assumption in the two-sample test:

- Wage distributions in both populations (rural and urban populations) are normal, equivalent to the **normality assumption** on SLR
- The variances of wages in both populations are unknown but equal, equivalent to the **constant variance assumption** in SLR.

When using a software program, we have to specify the predictor variable x to be a factor (categorical variable). In R, the function `factor()` converts a categorical or discrete numerical variable (with finite distinct values) to a factor variable.

We next implement the SLR with a binary factor variable (`smsa`) and use log wage `lwage` as the response.

```r
# lm model
smsa.lm <- lm(lwage ~ factor(smsa), data = edu.wage)
# residual plots
par(mfrow=c(2,2))
plot(smsa.lm)
```



1. Interpretations of Residual Diagnostic Plots

- **Residual-fitted Values**: does not reveal special patterns. The variances of the two groups seem to be similar to each other.
- **Q-Q Plot**: No serious violation of the normality assumption was found from the Q-Q plot. Note that

there is an outlier in the plot.
- **Scale-location Plot**: does not reveal any special pattern except for an outlier (obs N0. 24).
- **Residuals vs Leverage**: The Cook's distance of observation does not have a significant leverage (no influential).

Overall, there is no significant violation of model assumptions. We report the current model in the following:

2. Interpretations of Regression Coefficients

The binary variable`smsa` has two values: 0 = rural area, 1 = urban area. In general, a binary variable (taking values 0 and 1) is also commonly called a **dummy variable** in statistics and regression analysis.

```
summary(smsa.lm)
```

```
|
| Call:
| lm(formula = lwage ~ factor(smsa), data = edu.wage)
|
| Residuals:
|     Min      1Q  Median      3Q     Max
| -2.3240 -0.3700 -0.0797  0.3422  1.5439
|
| Coefficients:
|              Estimate Std. Error t value Pr(>|t|)
| (Intercept)    1.45182    0.04314  33.652  < 2e-16 ***
| factor(smsa)1  0.23732    0.05076   4.676 3.73e-06 ***
| ---
| Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
|
| Residual standard error: 0.5213 on 524 degrees of freedom
| Multiple R-squared:  0.04005, Adjusted R-squared:  0.03822
| F-statistic: 21.86 on 1 and 524 DF,  p-value: 3.733e-06
```

- **(Intercept)**: $\hat{\beta}_0 = 1.41518$ is the average log wage of **rural area** (`smsa = 0` is the baseline of the categorical variable `sasa`).

$$\log(\text{wage}_{\text{rural}}) = 1.41518 \implies \text{wage}_{\text{rural}} = e^{1.41518} = 4.117228$$

The above equations convert the log wage to the original wage. That is, the mean wage in rural areas is 4.117228.

- **slope parameter** (`factor(sasa)1` = level 1 of factor variable `smsa =1`: urban area): $\hat{\beta}_1 = 0.23732$ is the difference of the average log wage between the current category (urban area) and the baseline category (rural area). A more practical interpretation of the slope requires some algebra:

$$\log(\text{wage}_{\text{urban}}) - \log(\text{wage}_{\text{rural}}) = 0.23732 \implies \log \frac{\text{wage}_{\text{urban}}}{\text{wage}_{\text{rural}}} = 0.23732$$

Expanding both sides of the last equation in the above, we have

$$\frac{\text{wage}_{\text{urban}}}{\text{wage}_{\text{rural}}} = e^{0.23732} = 1.267847 \implies \text{wage}_{\text{urban}} = 1.267847 \times \text{wage}_{\text{rural}},$$

We re-express the last equation in the above to get

$$\text{wage}_{\text{urban}} - \text{wage}_{\text{rural}} = 1.267847 \times \text{wage}_{\text{rural}} - \text{wage}_{\text{rural}} = 0.267847\text{wage}_{\text{rural}}$$

Which is equivalent to

$$\frac{\text{wage}_{\text{urban}} - \text{wage}_{\text{rural}}}{\text{wage}_{\text{rural}}} = 0.267847.$$

This means that **the wage in urban area is 26.7847% higher than the rural area**..

3. Relationship between SLR and One-way ANOVA

First of all, we can extract a one-way ANOVA table from a linear regression model using the R function `anova(lm.object)`

```
anova(smsa.lm)
```

```
| Analysis of Variance Table
|
| Response: lwage
|               Df  Sum Sq Mean Sq F value    Pr(>F)
| factor(smsa)   1   5.941  5.9406  21.862 3.733e-06 ***
| Residuals    524 142.389  0.2717
| ---
| Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F statistic in the above ANOVA test (with 1 degree of freedom in the numerator and 524 degrees of freedom in the denominator) is

$$F = 21.862.$$

The t-test statistic in the SLR is a t distribution with 524 degrees of freedom, which is close to the standard normal distribution.

$$Z \approx T = 4.676.$$

It can be proven that $t^2_{\text{d.f}} = F_{1,\ \text{d.f}}$. This means the result of the t-test in SLR and the F-test in the ANOVA procedure are identical. Therefore, **the SLR with a binary predictor can the one-way ANOVA is a special SLR!**.

3 Linear Regression Approach to One-Way ANOVA

We claimed that one-way ANOVA is a special linear regression model when there is a binary categorical predictor is included in the model. However, a general CRD allows multiple treatments. In this section, we will illustrate how to use linear regression to perform a general one-way ANOVA with multiple treatment groups.

For ease of presentation, we continue using the wage dataset with log wage as the dependent variable and region as a categorical treatment factor (consisting of four groups: South, West, North-Central, and Other).

When performing linear regression with a categorical variable that has more than two categories, we must introduce a series of dummy variables to properly represent the multi-category variable.

Why Use Dummy Variables?

- **Encode Categorical Data**: They allow categorical variables (nominal or ordinal) to be included in regression models.
- **Avoid Arbitrary Numerical Assignments**: Using numbers like 1, 2, 3 for categories implies an order or magnitude, which may not exist (e.g., `Red = 1, Green = 2, Blue = 3` falsely suggests that Green is "greater" than Red).
- **Interpretability**: Each dummy variable represents the presence (1) or absence (0) of a category, making coefficients easy to interpret. To be more specific, the coefficient of a dummy variable represents the discrepancy between the category associated with the dummy variable and the baseline category.

Steps to Create Dummy Variables from a Multi-Category Variable

- **choose a baseline category**: The baseline category serves as a reference, allowing other categories to be compared against it.
  - Most software programs automatically set the baseline as the category with the smallest value (e.g., the first in alphabetical order for categorical variables).
- **Create Dummy Variables**: For each non-baseline category, define a dummy variable labeling that category. For example, `region = (northcen, other, south, west)`, the three dummy variables are defined as
  - dummy.other = 1 if lived in `other` region, 0 if not in `other` region.
  - dummy.south = 1 if lived in `south` and 0 if not in `south` region;
  - dummy.west = 1 if lived in `west` and 0 if not in `west`.

In general, a categorical variable has k categories, and we need to define $k-1$ dummy variables. For example,

region	dummy.other	dummy.south	dummy.west
south	0	1	0
west	0	0	1
northcen	0	0	0
other	1	0	0
south	0	1	0

- **Interpretation of Dummy Variables in Regression**

The coefficients for `other`, `south`, and `west` represent the difference from the reference category (`northcen`).

- **Dummy variables in R**
  - **Categorical variable with character values**: In this case, R will define dummy variables internally and use the default reference category.
  - **Categorical variable with values in numerical form**: In this case, we must use R's `factor()` function to convert the categorical variable into a factor variable (numeric values will be treated as category labels). R then internally defines dummy variables and selects the baseline category as the one with the smallest numerical label.
  - If the default reference category is not preferred, you can use R's `relevel()` function to manually specify a custom baseline category for easier interpretation.

```r
# Example data
data <- data.frame(color = c("Red", "Green", "Blue", "Green", "Red"))

# Set "Green" as the baseline (reference) category
data$color <- factor(data$color)
data$color <- relevel(data$color, ref = "Green")

# Check levels (first level is the baseline)
levels(data$color)  # Output: "Green" "Red" "Blue"
```

```
| [1] "Green" "Blue"  "Red"
```

On the Significance of the Categorical Variable:

- If **any** of the dummy variables is statistically **significant** (i.e., p-value < 0.05), the original categorical variable is considered **significant**.
- If **all** dummy variables are statistically **insignificant**, the original categorical variable is **insignificant**.

- If the original categorical variable is **insignificant**, it should be excluded from the regression model in real real-world application.

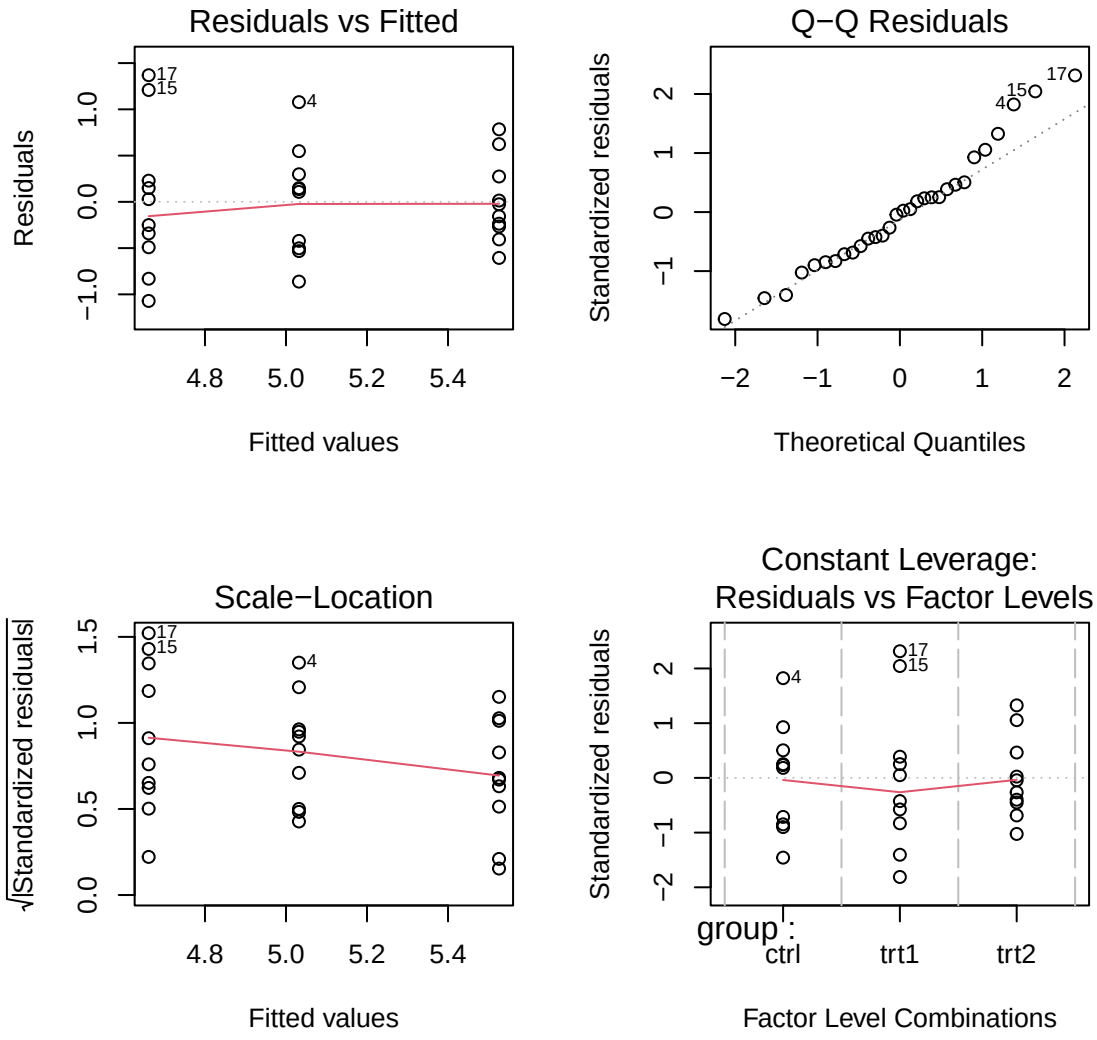
In the output of `lm()`, the F test in the bottom portion provides the significance test of the categorical variable. This F test is the same F test in the one-way ANOVA!!!

An Numerical Example

We use the **plant growth** dataset to build a regression model with the multi-category predictor `group` and compare it with a one-way ANOVA procedure. The data set is at <>. We will follow several logical steps to conduct the analysis.

Step 1: fit candidate model and perform diagnostics

```r
## Load the dataset
plant <- read.csv("https://pengdsci.github.io/STA200//dataset/oneWayPlantGrowth.csv")
##
plant.lm <- lm(weight ~ group, data = plant)  # factor(group) will also work
## residual plots
par(mfrow=c(2,2))
plot(plant.lm)
```



None of the above residual plots shows significant violations of model assumptions except for a few outliers that are not influential. We report the model.

Step 2: Reporting the linear model

```
summary(plant.lm)
```

```
|
| Call:
| lm(formula = weight ~ group, data = plant)
|
| Residuals:
|     Min      1Q  Median      3Q     Max
| -1.0710 -0.4180 -0.0060  0.2627  1.3690
|
| Coefficients:
|             Estimate Std. Error t value Pr(>|t|)
| (Intercept)   5.0320     0.1971  25.527   <2e-16 ***
```

```
| grouptrt1    -0.3710     0.2788  -1.331   0.1944
| grouptrt2     0.4940     0.2788   1.772   0.0877 .
| ---
| Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
|
| Residual standard error: 0.6234 on 27 degrees of freedom
| Multiple R-squared:  0.2641,  Adjusted R-squared:  0.2096
| F-statistic: 4.846 on 2 and 27 DF,  p-value: 0.01591
```

Interpretation of the output

- The explicit model formula has only the original group categorical variable with three categories. In fact, the implicit model formula uses dummy variables. In other words, the actual model formula used in the model is weight $= \beta_0 + \beta_1$grouptrt1 $+ \beta_2$grouptrt2, **where $\beta_0 = \mu_{\mathbf{ctr}}$, $\beta_1 = \mu_{\mathbf{trt1}} - \mu_{\mathbf{ctr}}$ and $\beta_2 = \mu_{\mathbf{trt2}} - \mu_{\mathbf{ctr}}$**
- The F test in the bottom portion yields a p-value of 0.01591. At a significance level of 0.05, we reject the null hypothesis that **region is significant**. In other words, this F-test tests the **null hypothesis: $H_0 : \beta_1 = \beta_2 = 0$.** This equivalent to $H_0 : \mu_{\mathbf{ctr}} = \mu_{\mathbf{trt1}} = \mu_{\mathbf{trt2}}$. **This is exactly the F test in the one-way ANOVA!** We will extract the one-way ANOVA from the linear model shortly.
- **Interpretation of Regression coefficients**: At the significance level of 0.05, neither of the two dummy variables is statistically significant.
  - $\beta_0 = \mu_{\text{ctr}}$, This represents the mean of the baseline category. The p-value tests the hypothesis $H_0 : \beta_0 = 0$.
  - $\hat{\beta}_1 = \hat{\mu}_{\text{trt1}} - \hat{\mu}_{\text{ctr}} = -0.3710$: The p-value of **0.1944** corresponds to testing the hypothesis $H_0 : \beta_1 = 0$ (or equivalently, $H_0 : \mu_{\text{trt1}} - \mu_{\text{ctr}} = 0$). The p-value indicates **no statistically significant difference** between `trt1` and `ctr`. The negative sign of the estimate suggests that the sample mean of `ctr` is higher than that of `trt1`.
  - $\hat{\beta}_2 = \hat{\mu}_{\text{trt2}} - \hat{\mu}_{\text{ctr}} = 0.4940$, The p-value of **0.0877** corresponds to testing the hypothesis $H_0 : \beta_2 = 0$ (or equivalently, $H_0 : \mu_{\text{trt2}} - \mu_{\text{ctr}} = 0$). The p-value indicates **no statistically significant difference** between `trt2` and `ctr`. The positive sign of the estimate suggests that the sample mean of `trt2` is higher than that of `ctr`.

Next, we extract the one-way ANOVA table directly from the above linear regression model.

```r
# extract ANOVA table directly from linear regression model
anova(plant.lm)
```

```
| Analysis of Variance Table
|
| Response: weight
|           Df  Sum Sq Mean Sq F value  Pr(>F)
| group      2  3.7663  1.8832  4.8461 0.01591 *
| Residuals 27 10.4921  0.3886
| ---
| Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-test from the one-way ANOVA table above is identical to the one reported in the previous regression model. We can also generate the same one-way ANOVA table using R's built-in function`aov()`

The coefficients associated with dummy variables in the regression output represent the difference between the mean of the underlying category and that of the baseline category. That is, $\hat{\beta}_1 = \hat{\mu}_{\text{trt1}} - \hat{\mu}_{\text{ctr}} = -0.3710$ and $\hat{\beta}_2 = \hat{\mu}_{\text{trt2}} - \hat{\mu}_{\text{ctr}} = 0.4940$. This information is given in Tukey's HSD (see the output of `TukeyHSD()`)

```r
plant.aov <- aov(weight ~ group, data = plant)
TukeyHSD(plant.aov)
```

```
|   Tukey multiple comparisons of means
|     95% family-wise confidence level
|
| Fit: aov(formula = weight ~ group, data = plant)
|
| $group
|             diff        lwr       upr     p adj
| trt1-ctrl -0.371 -1.0622161 0.3202161 0.3908711
| trt2-ctrl  0.494 -0.1972161 1.1852161 0.1979960
| trt2-trt1  0.865  0.1737839 1.5562161 0.0120064
```

The first two estimated differences (`trt1 - ctr` and `trt2 - ctr`) correspond to the regression coefficients of the dummy variables in the previous model. The p-values reported in the regression model are unadjusted (based on individual tests), whereas Tukey's HSD is a group test, and its corresponding p-values are adjusted to control the overall Type I error rate (i.e., to prevent inflation of the false positive rate).

Remarks:

- The linear regression model used to perform one-way ANOVA has **multiple dummy predictor variables** in the model formula (although implicitly expressed). In other words, the regression model for the one-way ANOVA is a special **multiple linear Regression** (MLR) model.
- Multiple Linear Regression (MLR) maintains the same assumptions as simple linear regression: independent and identically distributed (i.i.d.) observations, deterministic predictors, linearity between response and predictor, normality of residuals, and homoscedasticity (constant variance). Since MLR involves multiple predictor variables, an additional assumption is the absence of high correlations between predictors. Violation of this assumption is referred to as **multicollinearity** in regression modeling.
- The reason we encode a categorical variable with k categories into (k-1) dummy variables (rather than k) is to avoid multicollinearity issues. For example, suppose we have a categorical variable **Education** with 3 levels: High School (baseline/reference), Bachelor's, and Master's. If we create 3 dummy variables:
  - D1 = 1 if High School, 0 otherwise
  - D2 = 1 if Bachelor's, 0 otherwise
  - D3 = 1 if Master's, 0 otherwise

  This creates perfect **multicollinearity** because for every observation: D1 + D2 + D3 = 1 (always)!!!

To conclude this section, watch the YouTube video (https://www.youtube.com/watch?v=CMAwKuCw5CM&list=PLfQppi3mzF5nQzSm-RWQnvHFg46GQVK_4&index=4) that focuses on the ANOVA table extracted from the linear regression.

4 Multiple Linear Regression Approach to Two-way ANOVA

This section focuses on linear regression with two categorical variables, each with multiple levels. We have previously learned about two-way ANOVA based on a replicated Randomized Block Design

(RBD), which allows for inference on interaction effects. The linear regression model to be discussed will also include an interaction term.

Since the underlying concept is similar to the one-way ANOVA covered in the previous section, the explicit model formula remains relatively simple. However, the implicit model formula for linear regression with two categorical variables (including an interaction effect) is more complex than that of one-way ANOVA. We will use an example to illustrate some key features of multiple linear regression (MLR) involving only categorical variables with an interaction term.

The dataset recorded the yield of different crops under different fertilizers. The two-way data table is given below.

	Crop			
Fertilizer	Wheat	Corn	Soy	Rice
Blend X	123	128	166	151
	156	150	178	125
	112	174	187	117
	100	116	153	155
	168	109	195	158
Blend Y	135	175	140	167
	130	132	145	183
	176	120	159	142
	120	187	131	167
	155	184	126	168
Blend Z	156	186	185	175
	180	138	206	173
	147	178	188	154
	146	176	165	191
	193	190	188	169

We have used various R commands earlier to convert a two-way data table to an R dataframe for regression modeling. We next use similar code to convert the above table to a dataframe.

```
## To avoid typing errors, we input the table cell by cell from the data table
crop.data <- data.frame(
  Fertilizer = rep(c("Blend X", "Blend Y", "Blend Z"), each = 20),
  Crop = rep(rep(c("Wheat", "Corn", "Soy", "Rice"), each = 5), times = 3),
  Yield = c(123, 156, 112, 100, 168,   # Blend X - Wheat
            128, 150, 174, 116, 109,   # Blend X - Corn
            166, 178, 187, 133, 195,   # Blend X - Soy
            151, 125, 117, 155, 138,   # Blend X - Rice
            ##
            135, 130, 176, 120, 155,   # Blend Y - Wheat
            175, 132, 120, 187, 184,   # Blend Y - Corn
            140, 145, 159, 131, 126,   # Blend Y - Soy
            167, 188, 142, 167, 168,   # Blend Y - Rice
            ##
            156, 180, 147, 146, 193,   # Blend Z - Wheat
            186, 138, 178, 176, 190,   # Blend Z - Corn
            185, 206, 188, 165, 188,   # Blend Z - Soy
            175, 173, 154, 191, 169)   # Blend Z - Rice
)

# Convert factors to factors (important for modeling)
crop.data$Fertilizer <- as.factor(crop.data$Fertilizer)
crop.data$Crop <- as.factor(crop.data$Crop)
```

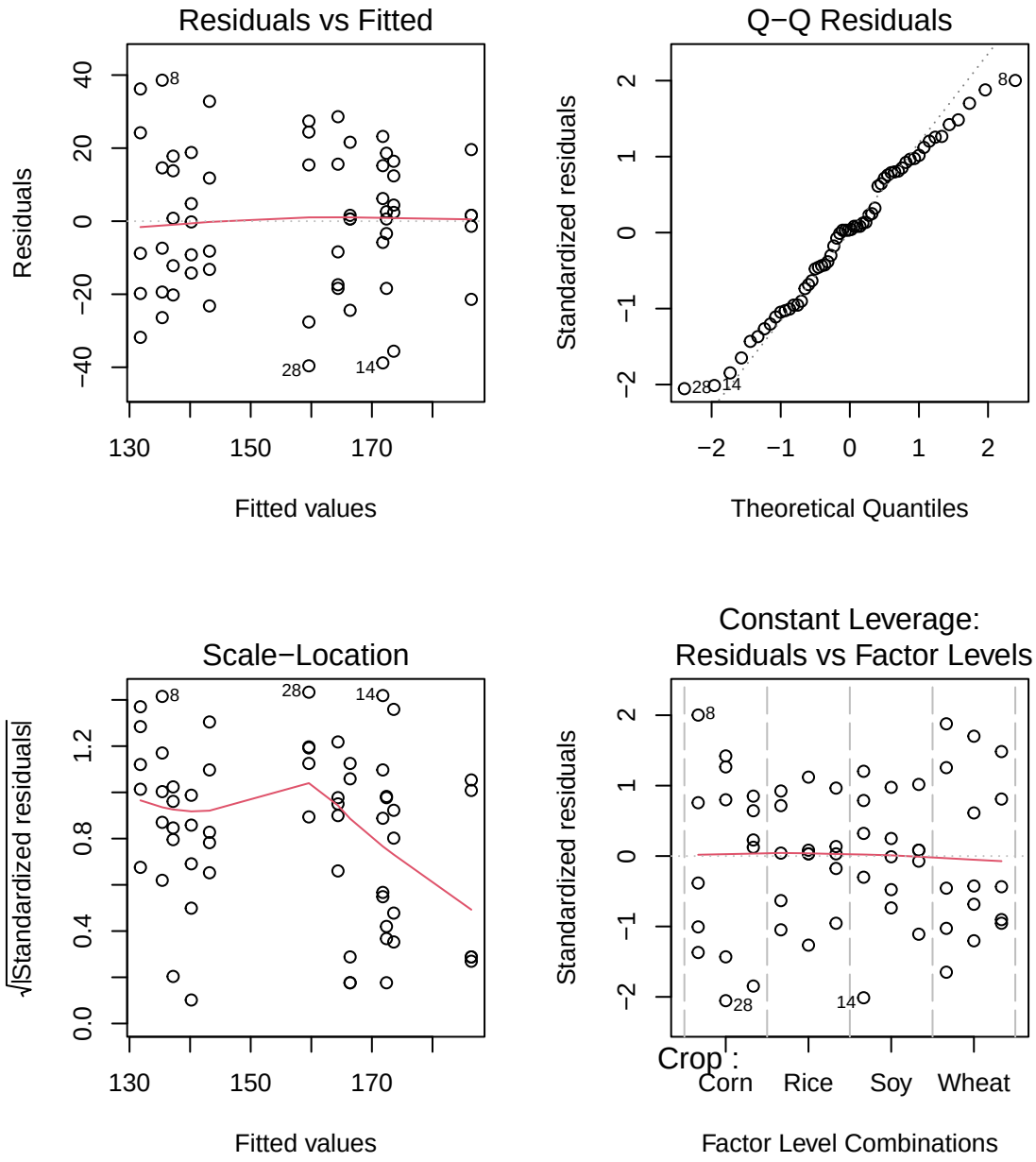
Next, we fit a linear regression model with an interaction term and list the terms (dummy variables) in the implicit internal model formula by using two R functions `model.matrix(model.name)` and `colnames(model.matrix.name)`. See how these R functions are called in the following code chunk.

```
## Explicit form is still a simple
crop.lm <- lm(Yield ~ Crop * Fertilizer, data = crop.data)
## model matrix provides all dummy variables implicitly defined in the R
model.mtx <- model.matrix(crop.lm)
## extract the column names of the model matrix ==> list of dummy variables in the
## implicit internal model formula
colnames(model.mtx)
```

```
|  [1] "(Intercept)"                "CropRice"
|  [3] "CropSoy"                    "CropWheat"
|  [5] "FertilizerBlend Y"          "FertilizerBlend Z"
|  [7] "CropRice:FertilizerBlend Y"  "CropSoy:FertilizerBlend Y"
|  [9] "CropWheat:FertilizerBlend Y" "CropRice:FertilizerBlend Z"
| [11] "CropSoy:FertilizerBlend Z"   "CropWheat:FertilizerBlend Z"
```

These dummy variables will appear in the output of the regression model in the subsequent model output. Before summarizing the model, we check the model assumptions to see whether significant violations exist using residual diagnostic plots.

```
par(mfrow = c(2,2))  # define the layout with 2x2 graphical panels
plot(crop.lm)
```



The residual plots look fine except for a pattern in the Q-Q plot, which indicates the distribution of the residuals is slightly skewed to the left. We will not explore methods of remedies in this class, but they will be discussed in the subsequent classes. Next, we report the results of the models.

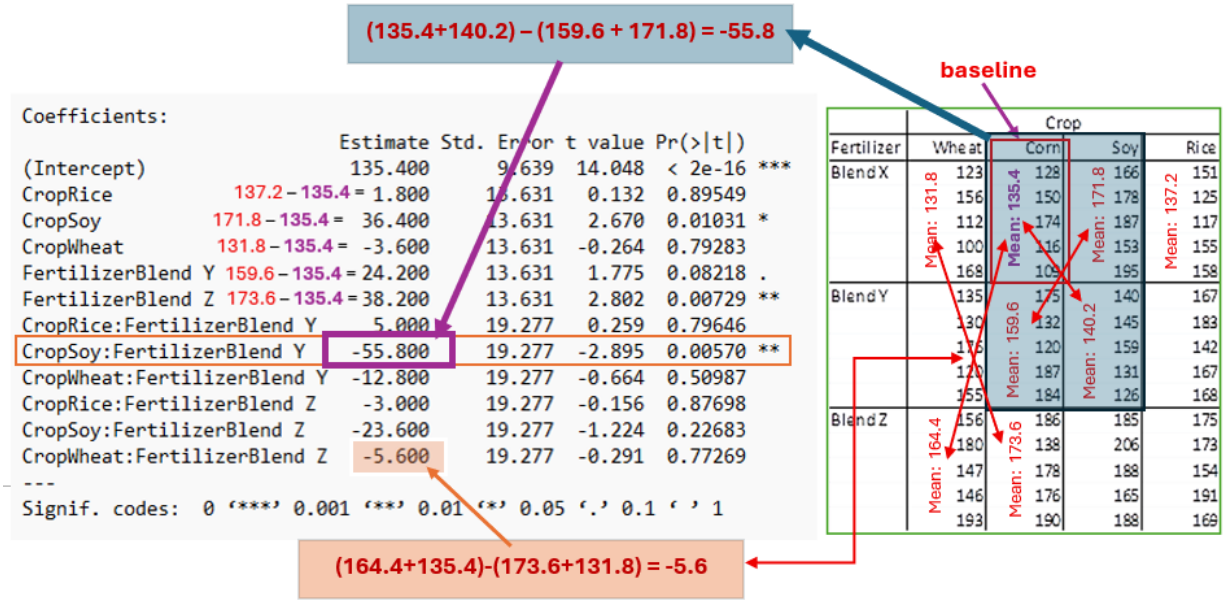
```
summary(crop.lm)
```

```
|
| Call:
| lm(formula = Yield ~ Crop * Fertilizer, data = crop.data)
|
| Residuals:
|     Min      1Q  Median      3Q     Max
```

```
| -39.60 -15.00   0.70  15.45  38.60
|
| Coefficients:
|                              Estimate Std. Error t value Pr(>|t|)
| (Intercept)                   135.400      9.639  14.048  < 2e-16 ***
| CropRice                        1.800     13.631   0.132  0.89549
| CropSoy                        36.400     13.631   2.670  0.01031 *
| CropWheat                      -3.600     13.631  -0.264  0.79283
| FertilizerBlend Y              24.200     13.631   1.775  0.08218 .
| FertilizerBlend Z              38.200     13.631   2.802  0.00729 **
| CropRice:FertilizerBlend Y      5.000     19.277   0.259  0.79646
| CropSoy:FertilizerBlend Y     -55.800     19.277  -2.895  0.00570 **
| CropWheat:FertilizerBlend Y   -12.800     19.277  -0.664  0.50987
| CropRice:FertilizerBlend Z     -3.000     19.277  -0.156  0.87698
| CropSoy:FertilizerBlend Z     -23.600     19.277  -1.224  0.22683
| CropWheat:FertilizerBlend Z    -5.600     19.277  -0.291  0.77269
| ---
| Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
|
| Residual standard error: 21.55 on 48 degrees of freedom
| Multiple R-squared:  0.4543,  Adjusted R-squared:  0.3292
| F-statistic: 3.632 on 11 and 48 DF,  p-value: 0.0008828
```

Output Interpretations

- The F-test in the bottom portion of the output indicates that at least one factor (`crop` or `fertilizer` or both) is significant ()
- **Intercept** - the mean of yields of corn using fertilizer X. i.e., **(128 + 150 + 174 + 116 + 109)/5 = 135.4.** This is the baseline category, defined by `Corn` and `Blend X`. All other categories will be compared to this baseline either directly or indirectly.
  - The coefficients for `CropRice`, `CropSoy`, and `CropWheat` represent the difference in **mean yield** between these crops and `Corn` when using `Blend X` fertilizer.
  - The coefficients for `FertilizerBlend Y` and `FertilizerBlend Z` represent the difference in **mean yield** for `Corn` when comparing `Blend Y` and `Blend Z` to the baseline (`Blend X`).
  - The coefficients for the interaction terms are not straightforward to interpret because they represent combined effects. Refer to the annotated output and original data table below to understand how these interaction term coefficients were calculated.



As we did with the one-way ANOVA, we can also extract the two-way ANOVA table directly from the linear regression model above.

```
anova(crop.lm)
```

```
| Analysis of Variance Table
|
| Response: Yield
|                 Df  Sum Sq Mean Sq F value    Pr(>F)
| Crop             3  2965.7   988.6  2.1282 0.1089382
| Fertilizer       2  9702.2  4851.1 10.4436 0.0001716 ***
| Crop:Fertilizer  6  5892.6   982.1  2.1143 0.0687635 .
| Residuals       48 22296.4   464.5
| ---
| Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The two-way ANOVA table provides three F-tests for assessing the two individual factors and their interaction effect, whereas the F-test in the regression output only tests whether all factors are jointly insignificant—a global goodness-of-fit measure. From this perspective, the two-way ANOVA F-tests offer more granular insights. However, regression coefficient inference allows for more detailed testing at individual factor levels, as well as for interaction effects between factor levels. Although Tukey's HSD procedure provides confidence intervals (and family-wise tests for simultaneous comparisons), it does not explicitly account for interaction effects in the multiple comparisons.

```
## fit an anova model since TukeyHSD() requires the aov() object
crop.aov <- aov(Yield ~ Crop * Fertilizer, data = crop.data)
## HSD calls
TukeyHSD(crop.aov)
```

```
|   Tukey multiple comparisons of means
|     95% family-wise confidence level
|
| Fit: aov(formula = Yield ~ Crop * Fertilizer, data = crop.data)
|
| $Crop
|                   diff       lwr       upr     p adj
```

```
| Rice-Corn    2.466667 -18.47792 23.411252 0.9891976
| Soy-Corn     9.933333 -11.01125 30.877918 0.5910428
| Wheat-Corn  -9.733333 -30.67792 11.211252 0.6069146
| Soy-Rice     7.466667 -13.47792 28.411252 0.7787573
| Wheat-Rice -12.200000 -33.14459  8.744585 0.4163283
| Wheat-Soy  -19.666667 -40.61125  1.277918 0.0728352
|
| $Fertilizer
|                  diff       lwr      upr     p adj
| Blend Y-Blend X  8.30 -8.183166 24.78317 0.4486138
| Blend Z-Blend X 30.15 13.666834 46.63317 0.0001620
| Blend Z-Blend Y 21.85  5.366834 38.33317 0.0066534
|
| $`Crop:Fertilizer`
|                               diff         lwr        upr     p adj
| Rice:Blend X-Corn:Blend X      1.8 -45.0050873  48.605087 1.0000000
| Soy:Blend X-Corn:Blend X      36.4 -10.4050873  83.205087 0.2718890
| Wheat:Blend X-Corn:Blend X    -3.6 -50.4050873  43.205087 1.0000000
| Corn:Blend Y-Corn:Blend X     24.2 -22.6050873  71.005087 0.8228174
| Rice:Blend Y-Corn:Blend X     31.0 -15.8050873  77.805087 0.5076482
| Soy:Blend Y-Corn:Blend X       4.8 -42.0050873  51.605087 0.9999999
| Wheat:Blend Y-Corn:Blend X     7.8 -39.0050873  54.605087 0.9999854
| Corn:Blend Z-Corn:Blend X     38.2  -8.6050873  85.005087 0.2115608
| Rice:Blend Z-Corn:Blend X     37.0  -9.8050873  83.805087 0.2506320
| Soy:Blend Z-Corn:Blend X      51.0   4.1949127  97.805087 0.0219849
| Wheat:Blend Z-Corn:Blend X    29.0 -17.8050873  75.805087 0.6066374
| Soy:Blend X-Rice:Blend X      34.6 -12.2050873  81.405087 0.3423125
| Wheat:Blend X-Rice:Blend X    -5.4 -52.2050873  41.405087 0.9999997
| Corn:Blend Y-Rice:Blend X     22.4 -24.4050873  69.205087 0.8838055
| Rice:Blend Y-Rice:Blend X     29.2 -17.6050873  76.005087 0.5967325
| Soy:Blend Y-Rice:Blend X       3.0 -43.8050873  49.805087 1.0000000
| Wheat:Blend Y-Rice:Blend X     6.0 -40.8050873  52.805087 0.9999990
| Corn:Blend Z-Rice:Blend X     36.4 -10.4050873  83.205087 0.2718890
| Rice:Blend Z-Rice:Blend X     35.2 -11.6050873  82.005087 0.3177646
| Soy:Blend Z-Rice:Blend X      49.2   2.3949127  96.005087 0.0315296
| Wheat:Blend Z-Rice:Blend X    27.2 -19.6050873  74.005087 0.6939901
| Wheat:Blend X-Soy:Blend X    -40.0 -86.8050873   6.805087 0.1614830
| Corn:Blend Y-Soy:Blend X     -12.2 -59.0050873  34.605087 0.9988789
| Rice:Blend Y-Soy:Blend X      -5.4 -52.2050873  41.405087 0.9999997
| Soy:Blend Y-Soy:Blend X      -31.6 -78.4050873  15.205087 0.4784497
| Wheat:Blend Y-Soy:Blend X    -28.6 -75.4050873  18.205087 0.6263743
| Corn:Blend Z-Soy:Blend X       1.8 -45.0050873  48.605087 1.0000000
| Rice:Blend Z-Soy:Blend X       0.6 -46.2050873  47.405087 1.0000000
| Soy:Blend Z-Soy:Blend X       14.6 -32.2050873  61.405087 0.9946081
| Wheat:Blend Z-Soy:Blend X     -7.4 -54.2050873  39.405087 0.9999915
| Corn:Blend Y-Wheat:Blend X    27.8 -19.0050873  74.605087 0.6653705
| Rice:Blend Y-Wheat:Blend X    34.6 -12.2050873  81.405087 0.3423125
| Soy:Blend Y-Wheat:Blend X      8.4 -38.4050873  55.205087 0.9999691
| Wheat:Blend Y-Wheat:Blend X   11.4 -35.4050873  58.205087 0.9993994
| Corn:Blend Z-Wheat:Blend X    41.8  -5.0050873  88.605087 0.1210713
| Rice:Blend Z-Wheat:Blend X    40.6  -6.2050873  87.405087 0.1469832
| Soy:Blend Z-Wheat:Blend X     54.6   7.7949127 101.405087 0.0103529
| Wheat:Blend Z-Wheat:Blend X   32.6 -14.2050873  79.405087 0.4309559
| Rice:Blend Y-Corn:Blend Y      6.8 -40.0050873  53.605087 0.9999964
```

Soy:Blend Y-Corn:Blend Y	-19.4	-66.2050873	27.405087	0.9529104
Wheat:Blend Y-Corn:Blend Y	-16.4	-63.2050873	30.405087	0.9861932
Corn:Blend Z-Corn:Blend Y	14.0	-32.8050873	60.805087	0.9962200
Rice:Blend Z-Corn:Blend Y	12.8	-34.0050873	59.605087	0.9982727
Soy:Blend Z-Corn:Blend Y	26.8	-20.0050873	73.605087	0.7126723
Wheat:Blend Z-Corn:Blend Y	4.8	-42.0050873	51.605087	0.9999999
Soy:Blend Y-Rice:Blend Y	-26.2	-73.0050873	20.605087	0.7399717
Wheat:Blend Y-Rice:Blend Y	-23.2	-70.0050873	23.605087	0.8584571
Corn:Blend Z-Rice:Blend Y	7.2	-39.6050873	54.005087	0.9999936
Rice:Blend Z-Rice:Blend Y	6.0	-40.8050873	52.805087	0.9999990
Soy:Blend Z-Rice:Blend Y	20.0	-26.8050873	66.805087	0.9422853
Wheat:Blend Z-Rice:Blend Y	-2.0	-48.8050873	44.805087	1.0000000
Wheat:Blend Y-Soy:Blend Y	3.0	-43.8050873	49.805087	1.0000000
Corn:Blend Z-Soy:Blend Y	33.4	-13.4050873	80.205087	0.3943466
Rice:Blend Z-Soy:Blend Y	32.2	-14.6050873	79.005087	0.4497486
Soy:Blend Z-Soy:Blend Y	46.2	-0.6050873	93.005087	0.0559743
Wheat:Blend Z-Soy:Blend Y	24.2	-22.6050873	71.005087	0.8228174
Corn:Blend Z-Wheat:Blend Y	30.4	-16.4050873	77.205087	0.5372002
Rice:Blend Z-Wheat:Blend Y	29.2	-17.6050873	76.005087	0.5967325
Soy:Blend Z-Wheat:Blend Y	43.2	-3.6050873	90.005087	0.0956612
Wheat:Blend Z-Wheat:Blend Y	21.2	-25.6050873	68.005087	0.9163288
Rice:Blend Z-Corn:Blend Z	-1.2	-48.0050873	45.605087	1.0000000
Soy:Blend Z-Corn:Blend Z	12.8	-34.0050873	59.605087	0.9982727
Wheat:Blend Z-Corn:Blend Z	-9.2	-56.0050873	37.605087	0.9999234
Soy:Blend Z-Rice:Blend Z	14.0	-32.8050873	60.805087	0.9962200
Wheat:Blend Z-Rice:Blend Z	-8.0	-54.8050873	38.805087	0.9999811
Wheat:Blend Z-Soy:Blend Z	-22.0	-68.8050873	24.805087	0.8953849

The following annotated output illustrates how to calculate various differences in the results of Tukey's HSD procedure.

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Yield ~ Crop * Fertilizer, data = crop.data)

$Crop
158.6667 – 156.2 = 2.46667 diff       lwr       upr     p adj
Rice-Corn     2.466667 -18.47792 23.411252 0.9891976
Soy-Corn      9.933333 -11.01125 30.877918 0.5910428
Wheat-Corn   -9.733333 -30.67792 11.211252 0.6069146
Soy-Rice      7.466667 -13.47792 28.411252 0.7787573
Wheat-Rice  -12.200000 -33.14459  8.744585 0.4163283
Wheat-Soy   -19.666667 -40.61125  1.277918 0.0728352

$Fertilizer
152.35 – 144.05 = 8.3  diff       lwr      upr     p adj
Blend Y-Blend X  8.30 -8.183166 24.78317 0.4486138
Blend Z-Blend X 30.15 13.666834 46.63317 0.0001620
Blend Z-Blend Y 21.85  5.366834 38.33317 0.0066534

$`Crop:Fertilizer`
   137.2 – 135.4 = 1.8                          diff         lwr        upr     p adj
Rice:Blend X-Corn:Blend X      1.8 -45.0050873  48.605087 1.0000000
Soy:Blend X-Corn:Blend X      36.4 -10.4050873  83.205087 0.2718890
Wheat:Blend X-Corn:Blend X    -3.6 -50.4050873  43.205087 1.0000000
Corn:Blend Y-Corn:Blend X     24.2 -22.6050873  71.005087 0.8228174
Rice:Blend Y-Corn:Blend X     31.0 -15.8050873  77.805087 0.5076482
Soy:Blend Y-Corn:Blend X       4.8 -42.0050873  51.605087 0.9999999
Wheat:Blend Y-Corn:Blend X     7.8 -39.0050873  54.605087 0.9999854
Corn:Blend Z-Corn:Blend X     38.2  -8.6050873  85.005087 0.2115608
Rice:Blend Z-Corn:Blend X     37.0  -9.8050873  83.805087 0.2506320
Soy:Blend Z-Corn:Blend X      51.0   4.1949127  97.805087 0.0219849
Wheat:Blend Z-Corn:Blend X    29.0 -17.8050873  75.805087 0.6066374
Soy:Blend X-Rice:Blend X      34.6 -12.2050873  81.405087 0.3423125
```

135.4 $\bar{Y}_{XC}$ 137.2 $\bar{Y}_{XR}$

Fertilizer	Crop: Wheat	Corn	Soy	Rice
Blend X	123	128	166	151
$\bar{Y}_X$	156	150	178	125
	112	174	187	117
144.05	100	116	153	155
	168	109	195	158
Blend Y	135	175	140	167
$\bar{Y}_Y$	130	132	145	183
	176	120	159	142
152.35	120	187	131	167
	155	184	126	168
Blend Z	156	186	185	175
$\bar{Y}_Z$	180	138	206	173
	147	178	188	154
174.2	146	176	165	191
	193	190	188	169
	$\bar{Y}_W$	$\bar{Y}_C$	$\bar{Y}_S$	$\bar{Y}_R$
	146.4667	156.2	166.1333	158.6667

We can see that Tukey's HSD.

The following YouTube video (28 minutes) is one of the very few that uses a linear regression approach to perform ANOVA analysis (https://www.youtube.com/watch?v=CS5ogBL-MHo). Please pay close attention to the explanation of the linear regression output, particularly the explanation of the two-way ANOVA with interaction term.

Finally, practice two-way ANOVA using regression approaches based on the following data table with two binary factors. Because both factors are binary, the interpretation of the output should be straightforward.

	no caffeine	**caffeine**
no beer	2.24, 1.62, 1.48, 1.70, 1.06, 1.39, 2.69, 0.28, 2.24, 1.15, 1.53, 2.43	0.62, 1.72, 1.75, 1.84, 1.30, 1.52, 1.31, 1.63, 1.91, 1.33, 0.84, 0.45
beer	1.71, 2.19, 2.27, 2.35, 2.47, 2.07, 2.56, 2.35, 1.50, 2.63, 2.48, 1.94	2.05, 1.51, 1.65, 2.68, 2.06, 1.80, 2.68, 1.93, 1.29, 1.93, 1.35, 1.37