Potential Hurdles in Multiple Linear Regression

I truly believe that we can overcome any hurdle that lies before us.
- Gillian Anderson

Outline

Learning Objectives

In this video, we will discuss about problems with:

- Multicollinearity.
- Variable Selection.
- Model Misspecification.
  - Outliers and influential leverage points.
  - Non-constant variance of error terms.

Multicollinearity

Multicollinearity

- Multicollinearity occurs when there are two or more predictors that are highly correlated with each other.
- Multicollinearity is less of a problem when the regression model is solely used for prediction.
- When multicollinearity is present, the estimated regression coefficients (the βs) become unstable.
- Recall from the advertising dataset, we have three predictors: TV, radio and newspaper.
- Adding one more predictor: Social media.

Describing and Exploring the Advertising Data

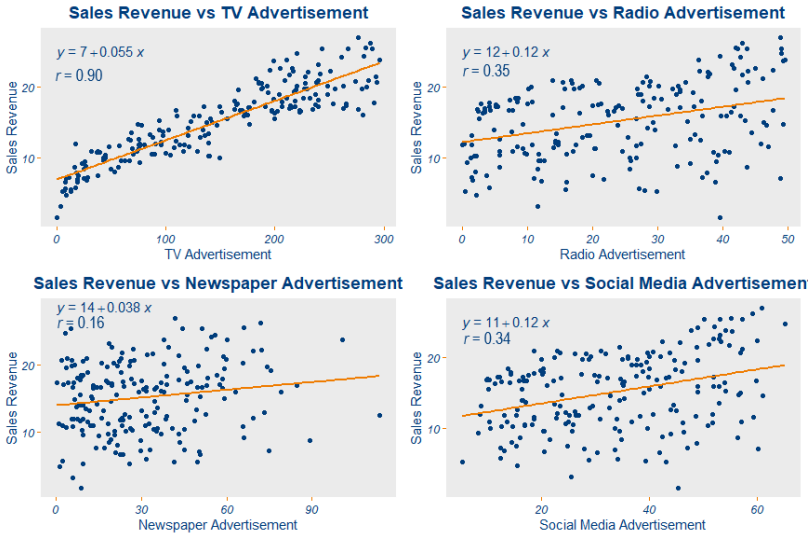
A quick check of the first few rows of our dataset:

- `head(df.advert)`

```
     TV Radio Newspaper Sales.Revenue SocialMedia
1 230.1  37.8      69.2          22.1    45.68304
2  44.5  39.3      45.1          10.4    50.77493
3  17.2  45.9      69.3          12.0    51.60575
4 151.5  41.3      58.5          16.5    53.95697
5 180.8  10.8      58.4          17.9    20.27567
6   8.7  48.9      75.0           7.2    60.04582
```

Describing and Exploring the Advertising Data

cont'd



Detecting Multicollinearity

Checking for pairwise correlation using a correlation matrix.

- From R code:

```
correlation <- cor(df.advert)
correlation
```

```
                    TV      Radio  Newspaper Sales.Revenue SocialMedia
TV           1.00000000 0.05480866 0.05664787     0.9012079  0.04561157
Radio        0.05480866 1.00000000 0.35410375     0.3496311  0.97975682
Newspaper    0.05664787 0.35410375 1.00000000     0.1579600  0.33281724
Sales.Revenue 0.90120791 0.34963110 0.15796003     1.0000000  0.34070451
SocialMedia  0.04561157 0.97975682 0.33281724     0.3407045  1.00000000
```

Detecting Multicollinearity

cont'd

Variance Inflation Factor (VIF)

Quantifies the severity of multicollinearity in a regression analysis.

- Formula to calculate VIF for a predictor i:

$$\text{VIF}_i = \frac{1}{1 - R_i^2}$$

- where R_i^2 is the R^2 value obtained by regressing the variables on predictor i using all other explanatory variables in the model.
- For example, if we want to find $\text{VIF}_{\text{TV}} = \frac{1}{1-R_{\text{TV}}^2}$,
- then R_{TV}^2 is obtained from the equation, $\text{TV} = \beta_0 + \beta_1 \text{Newspaper} + \beta_2 \text{Radio} + \beta_3 \text{SocialMedia}$.

Detecting Multicollinearity

cont'd

Variance Inflation Factor (VIF)

Quantifies the severity of multicollinearity in a regression analysis.

- A VIF value of 1 indicates that a particular variable is uncorrelated with all the other predictors.
- A VIF value > 5 indicates a high multicollinearity.
- `vif()` function from the `car` R package:
-
```
vif(lm1)
```

```
        TV       Radio   Newspaper SocialMedia
  1.009408   26.256106    1.165566   25.736587
```

- `Radio` and `SocialMedia` have high multicollinearity.

How Does Multicollinearity Affect a Regression Model?

For our dataset, consider the multilinear equation:

$$\text{Sales Revenue} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{Radio} + \beta_3 \times \text{Newspaper} + \beta_4 \times \text{SocialMedia}$$

```
Call:
lm(formula = Sales.Revenue ~ TV + Radio + Newspaper + SocialMedia,
    data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-6.8779 -0.8197  0.0462  0.8116  3.7563

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.942875   0.594006   6.638 5.06e-10 ***
TV          0.055421   0.001571  35.276  < 2e-16 ***
Radio       0.046745   0.045616   1.025    0.307
Newspaper   0.002932   0.006839   0.429    0.669
SocialMedia 0.057842   0.044980   1.286    0.200
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.682 on 155 degrees of freedom
Multiple R-squared:  0.904,     Adjusted R-squared:  0.9015
F-statistic: 364.9 on 4 and 155 DF,  p-value: < 2.2e-16
```

How Does Multicollinearity Affect a Regression Model?

For our dataset, consider the multilinear equation:

$$\text{Sales Revenue} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{Radio} + \beta_3 \times \text{Newspaper} + \beta_4 \times \text{SocialMedia}$$

```
Call:
lm(formula = Sales.Revenue ~ TV + Radio + Newspaper + SocialMedia,
    data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-6.8779 -0.8197  0.0462  0.8116  3.7563

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.942875   0.594006   6.638 5.06e-10 ***
TV          0.055421   0.001571  35.276  < 2e-16 ***
Radio       0.046745   0.045616   1.025    0.307
Newspaper   0.002932   0.006839   0.429    0.669
SocialMedia 0.057842   0.044980   1.286    0.200
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.682 on 155 degrees of freedom
Multiple R-squared:  0.904,     Adjusted R-squared:  0.9015
F-statistic: 364.9 on 4 and 155 DF,  p-value: < 2.2e-16
```

How Does Multicollinearity Affect a Regression Model?

For our dataset, consider the multilinear equation:

$$\text{Sales Revenue} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{Radio} + \beta_3 \times \text{Newspaper} + \beta_4 \times \text{SocialMedia}$$

```
Call:
lm(formula = Sales.Revenue ~ TV + Radio + Newspaper + SocialMedia,
    data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-6.8779 -0.8197  0.0462  0.8116  3.7563

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.942875   0.594006   6.638 5.06e-10 ***
TV          0.055421   0.001571  35.276  < 2e-16 ***
Radio       0.046745   0.045616   1.025    0.307
Newspaper   0.002932   0.006839   0.429    0.669
SocialMedia 0.057842   0.044980   1.286    0.200
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.682 on 155 degrees of freedom
Multiple R-squared:  0.904,     Adjusted R-squared:  0.9015
F-statistic: 364.9 on 4 and 155 DF,  p-value: < 2.2e-16
```

How to Deal with Multicollinearity?

- Possible solution #1: Principal Component Analysis (PCA) – combines highly correlated variables into unique Principal Components (more on this in DLP advanced modules)
- Possible solution #2: Omit one of the multicollinearity predictors.
  - Decision: Omit the radio advertising budget as they cannot display any pictures of their chairs.
  - Another way: Looking at customer media preference obtained from customer survey data to make decision.

```
Call:
lm(formula = Sales.Revenue ~ TV + Newspaper + SocialMedia, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-6.6225 -0.8592  0.0235  0.8359  3.9731

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.488252   0.395058   8.830 2.03e-15 ***
TV          0.055520   0.001568  35.400  < 2e-16 ***
Newspaper   0.004041   0.006753   0.598     0.55
SocialMedia 0.102901   0.009476  10.859  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.682 on 156 degrees of freedom
Multiple R-squared:  0.9033,    Adjusted R-squared:  0.9015
F-statistic:   486 on 3 and 156 DF,  p-value: < 2.2e-16
```

How to Deal with Multicollinearity?

- Possible solution #1: Principal Component Analysis (PCA) – combines highly correlated variables into unique Principal Components (more on this in DLP advanced modules)
- Possible solution #2: Omit one of the multicollinearity predictors.
  - Decision: Omit the radio advertising budget as they cannot display any pictures of their chairs.
  - Another way: Looking at customer media preference obtained from customer survey data to make decision.

```
Call:
lm(formula = Sales.Revenue ~ TV + Newspaper + SocialMedia, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-6.6225 -0.8592  0.0235  0.8359  3.9731

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.488252   0.395058   8.830 2.03e-15 ***
TV          0.055520   0.001568  35.400  < 2e-16 ***
Newspaper   0.004041   0.006753   0.598     0.55
SocialMedia 0.102901   0.009476  10.859  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.682 on 156 degrees of freedom
Multiple R-squared:  0.9033,    Adjusted R-squared:  0.9015
F-statistic:   486 on 3 and 156 DF,  p-value: < 2.2e-16
```

How to Deal with Multicollinearity?

cont'd

Final outcome: Drop Newspaper variable as it is not a significant predictor in the model.

```
Call:
lm(formula = Sales.Revenue ~ TV + SocialMedia, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-6.7423 -0.8525  0.0153  0.8638  3.9493

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.554253   0.378569   9.389   <2e-16 ***
TV          0.055496   0.001565  35.469   <2e-16 ***
SocialMedia 0.104865   0.008871  11.821   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.679 on 157 degrees of freedom
Multiple R-squared:  0.9031,    Adjusted R-squared:  0.9019
F-statistic: 731.8 on 2 and 157 DF,  p-value: < 2.2e-16
```

Variable Selection

Variable Selection

It is not wise to include all or too many predictor variables as **irrelevant** variables may reduce the Adjusted R^2 value or Multicollinearity will exist among the many predictor variables.

- What happens when we include irrelevant predictors?
  - Adding irrelevant predictors will slightly increase the R^2 values as Y can be explained by the extra predictors.
  - Take note that the Adjusted R^2 values takes into account the number of predictors, may end up with an Adjusted R^2 value that is lower than a model without those irrelevant predictors.
  - May also have trouble finding the significance of the regression model, the F-test.
- What happens when we omit relevant predictors?
  - Prediction will not be good as essential information will be missing from the model.
  - R^2 and Adjusted R^2 value will not be optimal.
  - Might not even get a significant F-test result for our regression model.

What is the Variable Selection Process?

Aim: Include just the right number of X variables as predictors, not too many and not too few variables.

1. Select the **response variable, Y**, which we would like to explain or predict.
2. Select the **most important predictor, X**, in determining or explaining the response variable, Y.
3. Select the next most important predictor, **based on the consideration that if the first most important predictor has been considered**, which other predictor will contribute the newest information in explaining Y.
4. Continue selecting the most important remaining predictor until you can **conclude that all the crucial predictors have been included**.

How to Obtain a Good Model Selection?

cont'd

- Although manual variable selection is subjective, it offers several advantages:
  1. When we have **two predictors that are equally good in predicting Y**, we will have control over our selection.
  2. By carefully going through the process of selecting the important predictors, we can gain **further insights from our data and clarify our thoughts why these predictors are important**.
- Variable selection can also be done automatically.
  1. **Forward step-wise selection**: Begin with a model that **contains no predictor** and then start adding the **most significant predictor variables one by one** until a pre-determined stop rule is met.
  2. **Backward step-wise selection**: Begin with a model that **contains all the predictor variables** and then start removing the **non-significant predictor variables one at a time** based on the highest p-value greater than 0.05.
- Note that step-wise selection does not guarantee that the best possible model is produced.

Model Misspecification

Model Misspecification

Model Misspecification

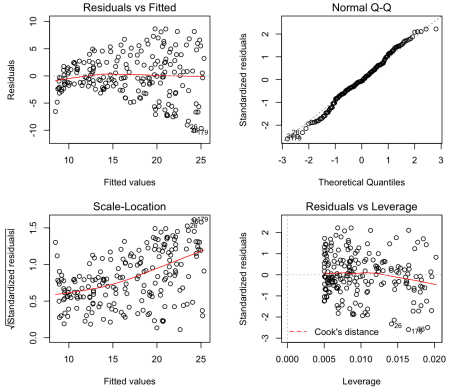
Happens when our regression model fails to represent the situation we are looking at.

- Examples of model misspecification:
  1. The relationship between the response variable and the predictors are nonlinear.
  2. The variability in the response variable, Y, is unequal, which results in the error terms being unequal across the predicted values.
  3. There might be some outliers which could distort the regression estimates.

Dealing with Model Misspecification

Diagnostic Plot

Scatterplot that plots the fitted values against the standardized residuals.

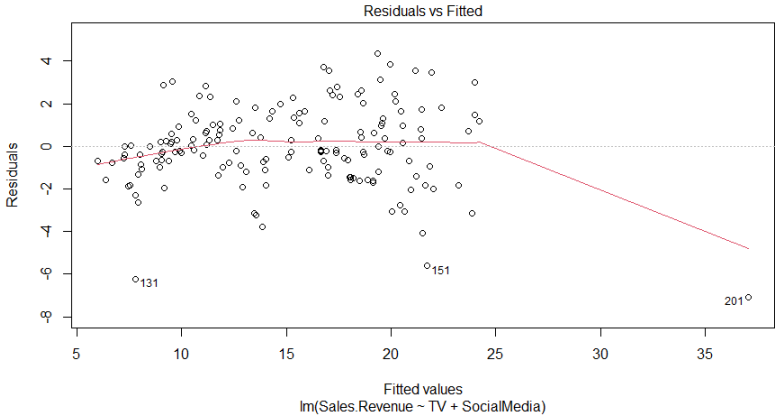


Source: `http://www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/`

Diagnostic Plot in Advertising Dataset

Suppose we had an additional market in our training data point #201 that we suspect is an outlier.

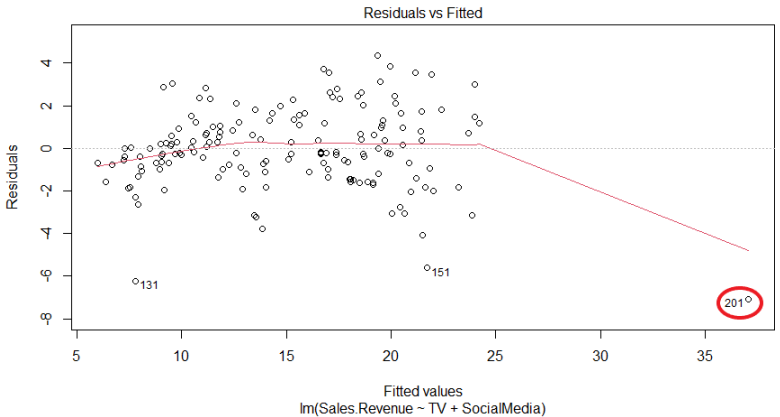
- `plot(lm4, which=1)`



Diagnostic Plot in Advertising Dataset

Suppose we had an additional market in our training data point #201 that we suspect is an outlier.

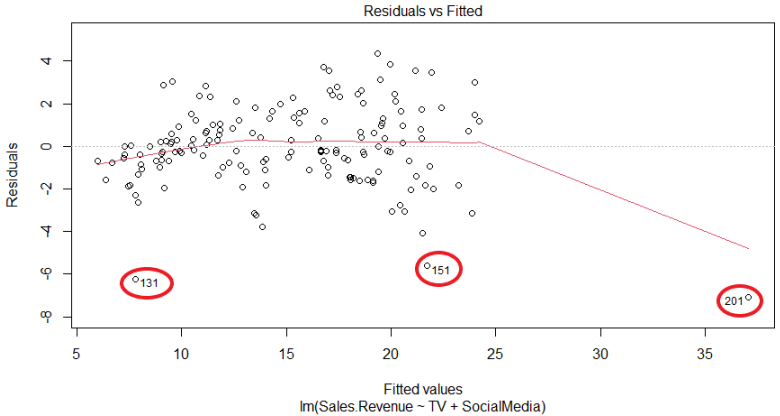
- `plot(lm4, which=1)`



Diagnostic Plot in Advertising Dataset

Suppose we had an additional market in our training data point #201 that we suspect is an outlier.

- `plot(lm4, which=1)`



Dealing with Outliers

Common techniques:

- Using residual vs leverage plot.
- Look at the Cook's distance.

- To find Cook's distance, we first need to know what is an influential point.
- Influential observations are observation points that when being removed, they will drastically change the coefficient estimates of the regression model.
- Influential point is **an outlier that greatly affects the regression model output.**
- Outlier may or may not affect the regression model output, **depending on the values of the predictors of the outliers**.
- Influential points are special type of outliers.

Cook's Distance

Cook's distance

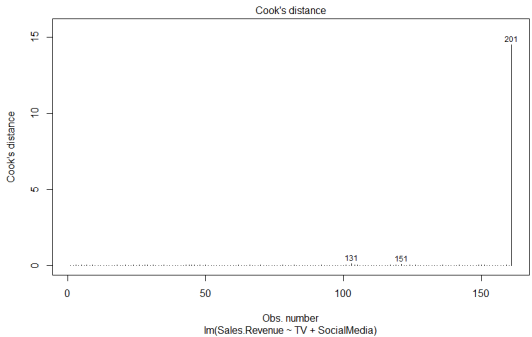
Measures the influence of an observation by quantifying how a regression model changes when an observation is removed, by taking into account both the leverage and residual of each observation.

- Cook's distance helps to identify influential points that negatively affect our regression model.
- General rule of thumb to identify these points (or so-called outliers):
  - The observation with a Cook's Distance of more than 3 times the mean, μ.
  - Alternative: The observation with a Cook's Distance of more than $4/n$, where n is the number of observations.
  - Some books suggest that observation point with a Cook's distance value of more than 1 indicates an influential point. Also, points with values of above 0.5 should be investigated.

Cook's Distance

cont'd

- `plot(lm4, which=4)`



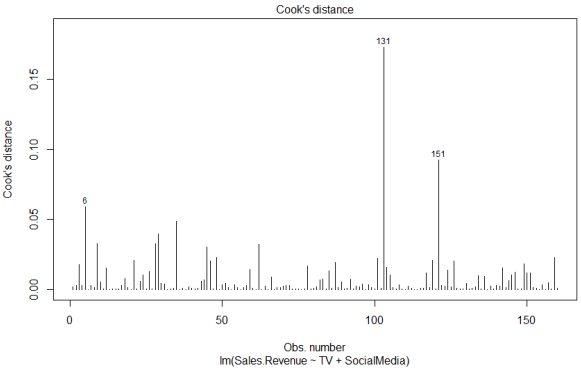
-
```
cooksD <- cooks.distance(lm4)
influential <- cooksD[(cooksD >
    (3 * mean(cooksD, na.rm =
   TRUE)))]
influential
```

-
```
201
14.48272
```

- The Cook's distance of data point #201 in the plot is too high, we may miss out other influential points.

Cook's Distance

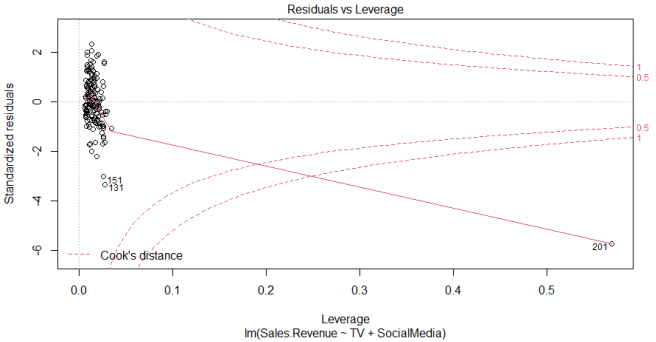
cont'd




- ```
  train2 = train[!(row.names(
     train) %in% c(201)),]
  lm4 <- lm(Sales.Revenue~ TV +
     SocialMedia, train2)
  plot(lm4, which=4)
  ```
- So far, none of the values are above 0.5. Therefore, we only label data point #201 as the influential point/potential outlier.
  ```

Residual vs Leverage Plot

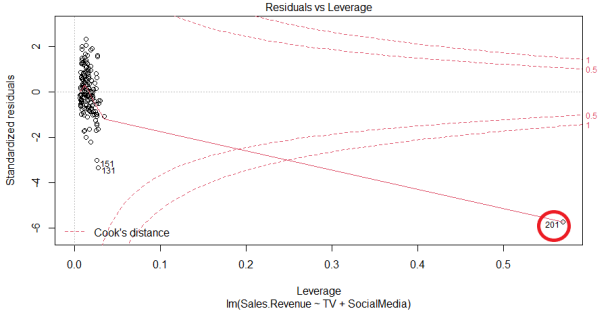
- `plot(lm4, which=5)`



- The leverage of an observation is a measure of its distance from the centre of all observations.

Residual vs Leverage Plot

- `plot(lm4, which=5)`



- Check whether the observation is an error.
- Remove these influential observations and rerun the model to see whether the model fit improves, that is, to see whether the Adjusted R^2 increases.
- The DLP Team decided to remove point #201 and also points #131 & #151 (Standardized residuals were less than -3).

Finalized Model

Before

```
Call:
lm(formula = Sales.Revenue ~ TV + SocialMedia, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-7.0641 -1.0018 -0.1618  1.1797  4.3468

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.953493   0.346484  14.296   <2e-16 ***
TV          0.055076   0.001751  31.461   <2e-16 ***
SocialMedia 0.062352   0.006604   9.442   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.88 on 158 degrees of freedom
Multiple R-squared:  0.8834,    Adjusted R-squared:  0.8819
F-statistic: 598.6 on 2 and 158 DF,  p-value: < 2.2e-16
```

After

```
Call:
lm(formula = Sales.Revenue ~ TV + SocialMedia, data = train2)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1552 -0.9131 -0.0620  0.7793  3.8890

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.639259   0.348300   10.45   <2e-16 ***
TV          0.055243   0.001465   37.72   <2e-16 ***
SocialMedia 0.105709   0.008196   12.90   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.543 on 155 degrees of freedom
Multiple R-squared:  0.9159,    Adjusted R-squared:  0.9148
F-statistic: 844.2 on 2 and 155 DF,  p-value: < 2.2e-16
```

Finalized Model

Before

```
Call:
lm(formula = Sales.Revenue ~ TV + SocialMedia, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-7.0641 -1.0018 -0.1618  1.1797  4.3468

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.953493   0.346484  14.296   <2e-16 ***
TV          0.055076   0.001751  31.461   <2e-16 ***
SocialMedia 0.062352   0.006604   9.442   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.88 on 158 degrees of freedom
Multiple R-squared:  0.8834,    Adjusted R-squared:  0.8819
F-statistic: 598.6 on 2 and 158 DF,  p-value: < 2.2e-16
```

After

```
Call:
lm(formula = Sales.Revenue ~ TV + SocialMedia, data = train2)

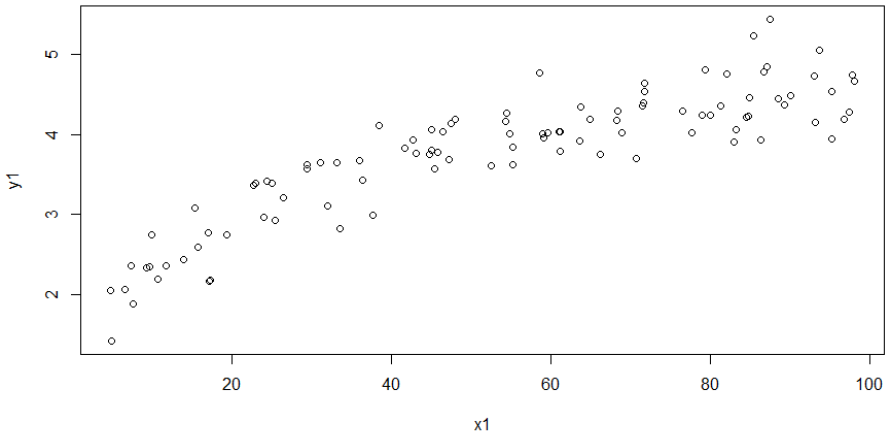
Residuals:
    Min      1Q  Median      3Q     Max
-4.1552 -0.9131 -0.0620  0.7793  3.8890

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.639259   0.348300   10.45   <2e-16 ***
TV          0.055243   0.001465   37.72   <2e-16 ***
SocialMedia 0.105709   0.008196   12.90   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.543 on 155 degrees of freedom
Multiple R-squared:  0.9159,    Adjusted R-squared:  0.9148
F-statistic: 844.2 on 2 and 155 DF,  p-value: < 2.2e-16
```

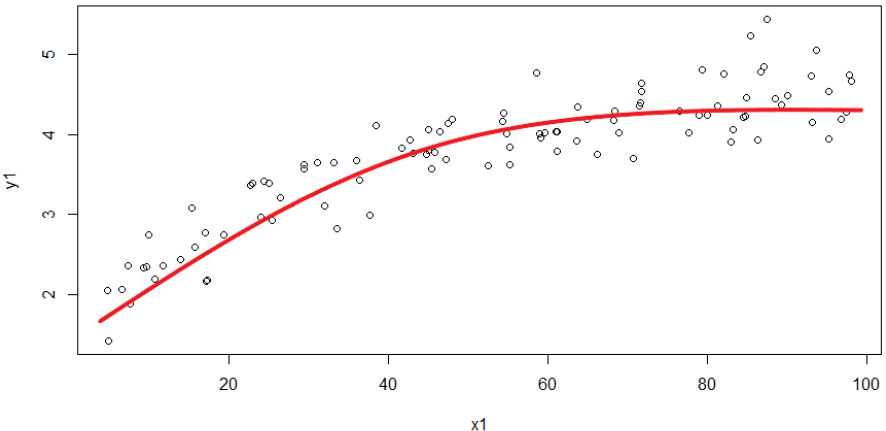
Dealing with Nonlinear Relationships and Unequal Variability

Example:



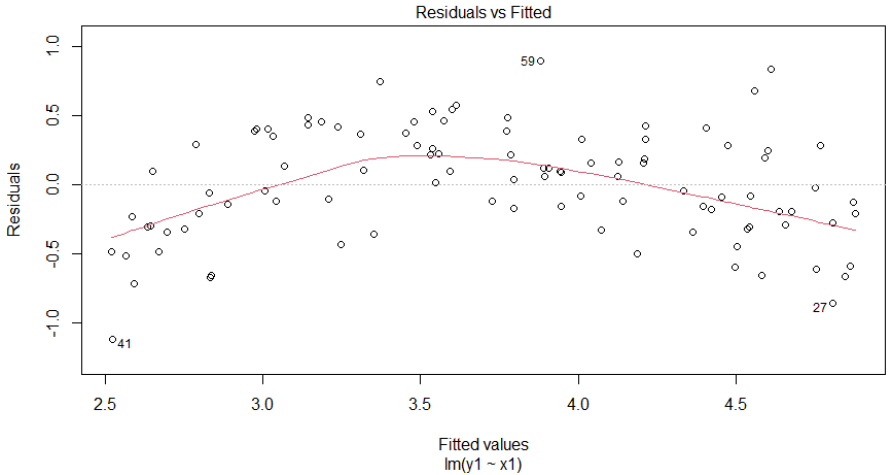
Dealing with Nonlinear Relationships and Unequal Variability

Example:



Dealing with Nonlinear Relationships and Unequal Variability

Diagnostic plot for this example:



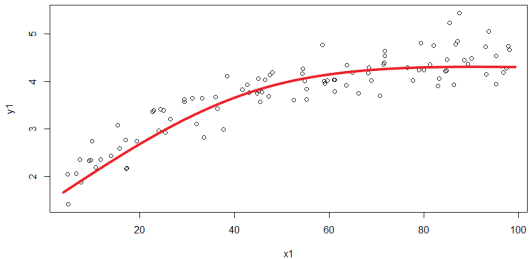
Dealing with Nonlinear Relationships and Unequal Variability

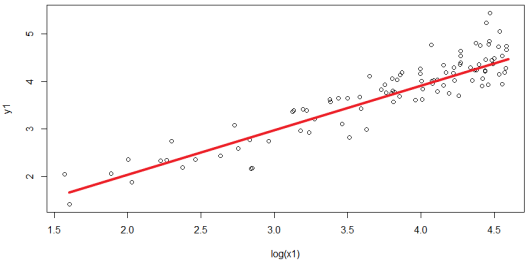
cont'd

Some choices we can consider:

1. Transform all/some of all the predictor variables. For example, you may use logarithms to transform the predictor variables.
2. Improve the model by introducing a new variable, for example, include an interaction term into the model.
3. Use other advanced methods such as nonlinear regression.

(a) Before: $Y_1 = \beta_0 + \beta_1 X_1$ $\Rightarrow$ **(b) After:** $Y_1 = \beta_0 + \beta_1 \mathbf{log}(X_1)$





Summary

In this video, we have discussed the common problems with multiple linear regression models:

- Multicollinearity.
- Variable Selection.
- Model Misspecification.
  - Outliers and influential points.
  - Non-constant variance of error terms.

Up next

- Interaction terms.
- Variable Transformation.
- Inclusion of categorical predictor variables by using dummy variables.

References

- G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*.
  Springer, 2013.
- "Understand forward and backward stepwise regression."
- C. Thieme, "Identifying outliers in linear regression-cook's distance," Jun 2021.
- Stephanie, "Cook's distance / cook's d: Definition, interpretation," Jun 2018.