STAT 423/523 Statistical Methods for Engineers and Scientists

Lecture 11: Multiple Linear Regression

Chencheng Cai

Washington State University

Multiple Linear Regression

In cases when we have more than one predictor variable, we can extend the simple linear regression model to a **multiple linear regression model**:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots \beta_p x_{ki} + \epsilon_i,$$

where

- y_i is the response variable,
- x_{ji} is the jth predictor variable for the ith observation
- $\epsilon_i \sim N(0, \sigma^2)$ is the error term.

Multiple Linear Regression

In cases when we have more than one predictor variable, we can extend the simple linear regression model to a **multiple linear regression model**:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots \beta_p x_{ki} + \epsilon_i,$$

where

- y_i is the response variable,
- x_{ji} is the jth predictor variable for the ith observation
- $\epsilon_i \sim N(0, \sigma^2)$ is the error term.

The predictors could be:

- additional covariates in the dataset
- interactions between predictors
- nonlinear functions of predictors

Ordinary Least Squares

We follow the same principle as in simple linear regression and minimize the residual sum of squares (RSS):

$$\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k = \underset{\beta_0, \beta_1, \ldots, \beta_k}{\arg\min} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \cdots - \beta_k x_{ki})^2$$

Ordinary Least Squares

We follow the same principle as in simple linear regression and minimize the residual sum of squares (RSS):

$$\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k = \underset{\beta_0, \beta_1, \ldots, \beta_k}{\arg\min} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \cdots - \beta_k x_{ki})^2$$

We compute the partial derivatives of the RSS with respect to each β_j:

$$\frac{\partial \mathrm{RSS}}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \cdots - \beta_k x_{ki})$$

$$\frac{\partial \mathrm{RSS}}{\partial \beta_j} = -2 \sum_{i=1}^{n} x_{ji} (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \cdots - \beta_k x_{ki}), \; j = 1, \ldots, k$$

Ordinary Least Squares

The OLS estimators can be obtained by setting the partial derivatives to zero:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \cdots - \beta_k x_{ki}) = 0$$

$$\sum_{i=1}^{n} x_{1i}(y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \cdots - \beta_k x_{ki}) = 0$$

$$\sum_{i=1}^{n} x_{2i}(y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \cdots - \beta_k x_{ki}) = 0$$

$$\vdots$$

$$\sum_{i=1}^{n} x_{ki}(y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \cdots - \beta_k x_{ki}) = 0$$

Ordinary Least Squares

This is a linear system of equations in the unknowns $\beta_0, \beta_1, \ldots, \beta_k$.

$$\sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_{1i} + \beta_2 \sum_{i=1}^{n} x_{2i} + \cdots + \beta_k \sum_{i=1}^{n} x_{ki}$$

$$\sum_{i=1}^{n} x_{1i}y_i = \beta_0 \sum_{i=1}^{n} x_{1i} + \beta_1 \sum_{i=1}^{n} x_{1i}^2 + \beta_2 \sum_{i=1}^{n} x_{1i}x_{2i} + \cdots + \beta_k \sum_{i=1}^{n} x_{1i}x_{ki}$$

$$\sum_{i=1}^{n} x_{2i}y_i = \beta_0 \sum_{i=1}^{n} x_{2i} + \beta_1 \sum_{i=1}^{n} x_{2i}x_{1i} + \beta_2 \sum_{i=1}^{n} x_{2i}^2 + \cdots + \beta_k \sum_{i=1}^{n} x_{2i}x_{ki}$$

$$\vdots$$

$$\sum_{i=1}^{n} x_{ki}y_i = \beta_0 \sum_{i=1}^{n} x_{ki} + \beta_1 \sum_{i=1}^{n} x_{ki}x_{1i} + \beta_2 \sum_{i=1}^{n} x_{ki}x_{2i} + \cdots + \beta_k \sum_{i=1}^{n} x_{ki}^2$$

Ordinary Least Squares

We can write it in matrix form:

$$\begin{bmatrix} n & \sum_{i=1}^n x_{1i} & \sum_{i=1}^n x_{2i} & \cdots & \sum_{i=1}^n x_{ki} \\ \sum_{i=1}^n x_{1i} & \sum_{i=1}^n x_{1i}^2 & \sum_{i=1}^n x_{1i}x_{2i} & \cdots & \sum_{i=1}^n x_{1i}x_{ki} \\ \sum_{i=1}^n x_{2i} & \sum_{i=1}^n x_{2i}x_{1i} & \sum_{i=1}^n x_{2i}^2 & \cdots & \sum_{i=1}^n x_{2i}x_{ki} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{ki} & \sum_{i=1}^n x_{ki}x_{1i} & \sum_{i=1}^n x_{ki}x_{2i} & \cdots & \sum_{i=1}^n x_{ki}^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{1i}y_i \\ \sum_{i=1}^n x_{2i}y_i \\ \vdots \\ \sum_{i=1}^n x_{ki}y_i \end{bmatrix}$$

Ordinary Least Squares

We can write it in matrix form:

$$\begin{bmatrix} n & \sum_{i=1}^n x_{1i} & \sum_{i=1}^n x_{2i} & \cdots & \sum_{i=1}^n x_{ki} \\ \sum_{i=1}^n x_{1i} & \sum_{i=1}^n x_{1i}^2 & \sum_{i=1}^n x_{1i}x_{2i} & \cdots & \sum_{i=1}^n x_{1i}x_{ki} \\ \sum_{i=1}^n x_{2i} & \sum_{i=1}^n x_{2i}x_{1i} & \sum_{i=1}^n x_{2i}^2 & \cdots & \sum_{i=1}^n x_{2i}x_{ki} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{ki} & \sum_{i=1}^n x_{ki}x_{1i} & \sum_{i=1}^n x_{ki}x_{2i} & \cdots & \sum_{i=1}^n x_{ki}^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{1i}y_i \\ \sum_{i=1}^n x_{2i}y_i \\ \vdots \\ \sum_{i=1}^n x_{ki}y_i \end{bmatrix}$$

more compactly, we can write it as:

$$\begin{bmatrix} S_{x_0x_0} & S_{x_0x_1} & S_{x_0x_2} & \cdots & S_{x_0x_k} \\ S_{x_1x_0} & S_{x_1x_1} & S_{x_1x_2} & \cdots & S_{x_1x_k} \\ S_{x_2x_0} & S_{x_2x_1} & S_{x_2x_2} & \cdots & S_{x_2x_k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ S_{x_kx_0} & S_{x_kx_1} & S_{x_kx_2} & \cdots & S_{x_kx_k} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} = \begin{bmatrix} S_{x_0y} \\ S_{x_1y} \\ S_{x_2y} \\ \vdots \\ S_{x_ky} \end{bmatrix}$$

where $S_{x_jx_l} = \sum_{i=1}^n x_{ji}x_{li}$ and $S_{x_jy} = \sum_{i=1}^n x_{ji}y_i$ with $x_{0i} = 1$.

Ordinary Least Squares

The OLS estimators can be computed using matrix algebra:

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} S_{x_0x_0} & S_{x_0x_1} & S_{x_0x_2} & \cdots & S_{x_0x_k} \\ S_{x_1x_0} & S_{x_1x_1} & S_{x_1x_2} & \cdots & S_{x_1x_k} \\ S_{x_2x_0} & S_{x_2x_1} & S_{x_2x_2} & \cdots & S_{x_2x_k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ S_{x_kx_0} & S_{x_kx_1} & S_{x_kx_2} & \cdots & S_{x_kx_k} \end{bmatrix}^{-1} \begin{bmatrix} S_{x_0y} \\ S_{x_1y} \\ S_{x_2y} \\ \vdots \\ S_{x_ky} \end{bmatrix}$$

Ordinary Least Squares

We can verify the solution is compatible with the simple linear regression case.

Ordinary Least Squares

We can verify the solution is compatible with the simple linear regression case. When $k = 1$, we have:

$$
\begin{aligned}
\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} &= \begin{bmatrix} S_{x_0 x_0} & S_{x_0 x_1} \\ S_{x_1 x_0} & S_{x_1 x_1} \end{bmatrix}^{-1} \begin{bmatrix} S_{x_0 y} \\ S_{x_1 y} \end{bmatrix} \\
&= \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix} \\
&= \frac{1}{n \sum x_i^2 - (\sum x_i)^2} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix} \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix} \\
&= \frac{1}{n \sum x_i^2 - (\sum x_i)^2} \begin{bmatrix} \sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i \\ n \sum x_i y_i - \sum x_i \sum y_i \end{bmatrix} \\
&= S_{xx}^{-1} \begin{bmatrix} \bar{y} S_{xx} - \bar{x} S_{xy} \\ S_{xy} \end{bmatrix}
\end{aligned}
$$

Ordinary Least Squares

For the variance component, we have:

$$\hat{\sigma}^2 = \text{MSE} = \frac{\text{RSS}(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)}{n-k-1} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k-1}$$

where

Ordinary Least Squares

For the variance component, we have:

$$\hat{\sigma}^2 = \text{MSE} = \frac{\text{RSS}(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)}{n-k-1} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k-1}$$

where

- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_k x_{ki}$ is the **predicted** or **fitted** value of y_i
- The degrees of freedom is $n-k-1$ because we have estimated $k+1$ parameters $(\beta_0, \beta_1, \ldots, \beta_k)$ from the data.

Ordinary Least Squares

The OLS estimators are **unbiased**:

$$\mathbb{E}[\hat{\beta}_j] = \beta_j, \;\; j = 0, 1, \ldots, k$$

Ordinary Least Squares

The OLS estimators are **unbiased**:

$$\mathbb{E}[\hat{\beta}_j] = \beta_j, \;\; j = 0, 1, \ldots, k$$

Let $s_{\hat{\beta}_j}$ be the estimated standard error of $\hat{\beta}_j$. Then

$$\frac{\hat{\beta}_j}{s_{\hat{\beta}_j}} \sim t_{n-k-1}$$

which is a t-distribution with $n - k - 1$ degrees of freedom.

Confidence interval and t-test

The $(1-\alpha)$ confidence interval for β_j is given by:

$$\hat{\beta}_j \pm t_{\alpha/2,n-k-1} s_{\hat{\beta}_j}.$$

Confidence interval and t-test

The $(1-\alpha)$ confidence interval for β_j is given by:

$$\hat{\beta}_j \pm t_{\alpha/2,n-k-1} s_{\hat{\beta}_j}.$$

Consider the hypothesis test:

$$H_0 : \beta_j = 0 \text{ vs. } H_a : \beta_j \neq 0$$

We reject H_0 if:

Confidence interval and t-test

The $(1-\alpha)$ confidence interval for β_j is given by:

$$\hat{\beta}_j \pm t_{\alpha/2,n-k-1} s_{\hat{\beta}_j}.$$

Consider the hypothesis test:

$$H_0 : \beta_j = 0 \text{ vs. } H_a : \beta_j \neq 0$$

We reject H_0 if:

- The CI does not contain 0.
- The t-statistic

  $$t = \frac{\hat{\beta}_j}{s_{\hat{\beta}_j}}$$

  has absolute value greater than $t_{\alpha/2,n-k-1}$.
- The p-value

  $$p = 2(1 - F_{t,n-k-1}(|t|))$$

  is less than α.

Confidence interval and t-test

- The standard error of $\hat{\beta}_j$ can be read from the output of the regression models in R and Python.
- Sometimes the p-values are reported in the output as well.

Confidence interval and t-test

- The standard error of $\hat{\beta}_j$ can be read from the output of the regression models in R and Python.
- Sometimes the p-values are reported in the output as well.
- A covariate $x_{ji}, i = 1, \ldots, n$ is **significant** if the null hypothesis $H_0 : \beta_j = 0$ is rejected.
- A covariate $x_{ji}, i = 1, \ldots, n$ is **insignificant** if the null hypothesis $H_0 : \beta_j = 0$ is not rejected.

Confidence interval and t-test

- The standard error of $\hat{\beta}_j$ can be read from the output of the regression models in R and Python.
- Sometimes the p-values are reported in the output as well.
- A covariate $x_{ji}, i = 1, \dots, n$ is **significant** if the null hypothesis $H_0 : \beta_j = 0$ is rejected.
- A covariate $x_{ji}, i = 1, \dots, n$ is **insignificant** if the null hypothesis $H_0 : \beta_j = 0$ is not rejected.
- Insignificant covariates can be removed from the model to simplify the model.

Example

We consider the **mtcars** dataset in R and run a linear regression model of `mpg` (miles per gallon) on `disp` (displacement), `hp` (gross horsepower), and `wt` (weight of car).

```
Call:
lm(formula = mpg ~ disp + hp + wt, data = mtcars)

Residuals:
Min      1Q Median      3Q     Max
-3.891 -1.640 -0.172  1.061  5.861

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.105505   2.110815  17.579  < 2e-16 ***
disp        -0.000937   0.010350  -0.091  0.92851
hp          -0.031157   0.011436  -2.724  0.01097 *
wt          -3.800891   1.066191  -3.565  0.00133 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.639 on 28 degrees of freedom
Multiple R-squared:  0.8268,Adjusted R-squared:  0.8083
F-statistic: 44.57 on 3 and 28 DF,  p-value: 8.65e-11
```

Example

- The estimated intercept is $\hat{\beta}_0 = 37.11$.
- The estimated slope for `disp` is $\hat{\beta}_1 = -0.000937$.
- The estimated slope for `hp` is $\hat{\beta}_2 = -0.03116$.
- The estimated slope for `wt` is $\hat{\beta}_3 = -3.8009$.

Example

- The estimated intercept is $\hat{\beta}_0 = 37.11$.
- The estimated slope for `disp` is $\hat{\beta}_1 = -0.000937$.
- The estimated slope for `hp` is $\hat{\beta}_2 = -0.03116$.
- The estimated slope for `wt` is $\hat{\beta}_3 = -3.8009$.
- The intercept, `hp`, and `wt` are significant at $\alpha = 0.05$ level.
- The `disp` is insignificant at $\alpha = 0.05$ level.

Example

- The estimated intercept is $\hat{\beta}_0 = 37.11$.
- The estimated slope for `disp` is $\hat{\beta}_1 = -0.000937$.
- The estimated slope for `hp` is $\hat{\beta}_2 = -0.03116$.
- The estimated slope for `wt` is $\hat{\beta}_3 = -3.8009$.
- The intercept, `hp`, and `wt` are significant at $\alpha = 0.05$ level.
- The `disp` is insignificant at $\alpha = 0.05$ level.
- fitted model is

$$\text{mpg} = 37.11 - 0.0009 \times \text{disp} - 0.0312 \times \text{hp} - 3.801 \times \text{wt} + \epsilon \quad \text{with } \epsilon \sim N(0, 2.639^2)$$

Example

A direct improvement of the model is to remove disp from the model and refit the model:

```
Call:
lm(formula = mpg ~ hp + wt, data = mtcars)

Residuals:
Min      1Q Median      3Q     Max
-3.941 -1.600 -0.182  1.050  5.854

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.22727    1.59879  23.285  < 2e-16 ***
hp          -0.03177    0.00903  -3.519  0.00145 **
wt          -3.87783    0.63273  -6.129 1.12e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.593 on 29 degrees of freedom
Multiple R-squared:  0.8268,Adjusted R-squared:  0.8148
F-statistic: 69.21 on 2 and 29 DF,  p-value: 9.109e-12
```

Model Comparison

Consider two **nested** models:

- The **full model**: (all subscript i are removed for simplicity)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q + \beta_{q+1} x_{q+1} + \cdots + \beta_k x_k + \epsilon$$

- The **reduced model**: (all subscript i are removed for simplicity)

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q + \epsilon$$

Model Comparison

Consider two **nested** models:

- The **full model**: (all subscript i are removed for simplicity)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q + \beta_{q+1} x_{q+1} + \cdots + \beta_k x_k + \epsilon$$

- The **reduced model**: (all subscript i are removed for simplicity)

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q + \epsilon$$

- The reduced model is a special case of the full model with $\beta_{q+1} = \cdots = \beta_k = 0$.
- Comparing the two models is equivalent to testing the null hypothesis:

$$H_0 : \beta_{q+1} = \cdots = \beta_k = 0$$

Model Comparison

$$H_0 : \beta_{q+1} = \cdots = \beta_k = 0$$

In order to compare the nested models, we can use the **F-test**:

$$F = \frac{(SSE_{reduced} - SSE_{full})/(k-q)}{SSE_{full}/(n-k-1)}$$

Model Comparison

$$H_0 : \beta_{q+1} = \cdots = \beta_k = 0$$

In order to compare the nested models, we can use the **F-test**:

$$F = \frac{(SSE_{reduced} - SSE_{full})/(k-q)}{SSE_{full}/(n-k-1)}$$

reject null if

- $F > F_{\alpha,k-q,n-k-1}$
- The p-value:

$$1 - F_{F,k-q,n-k-1}(F)$$

  is less than α.

Example

Recall the **mtcars** dataset, we compare the following two models:

```
> model1 = lm(mpg~disp+hp+wt, mtcars)
> model2 = lm(mpg~disp, mtcars)
```

Example

Recall the **mtcars** dataset, we compare the following two models:

```
> model1 = lm(mpg~disp+hp+wt, mtcars)
> model2 = lm(mpg~disp, mtcars)
```

The F-test result can be read from anova function:

```
> anova(model2, model1)
Analysis of Variance Table

Model 1: mpg ~ disp
Model 2: mpg ~ disp + hp + wt
Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     30 317.16
2     28 194.99  2    122.17 8.7715 0.001102 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model Comparison

- R^2 is a metric for the goodness of fit of the model.
- But we **cannot** use R^2 to compare two models with different number of predictors, because **adding more predictors will always increase** R^2.

Model Comparison

- R^2 is a metric for the goodness of fit of the model.
- But we **cannot** use R^2 to compare two models with different number of predictors, because **adding more predictors will always increase** R^2.
- We can use the **adjusted** R^2:

$$R^2_{adj} = 1 - \frac{n-1}{n-k-1}\frac{\text{SSE}}{\text{SST}}$$

Model Comparison

- R^2 is a metric for the goodness of fit of the model.
- But we **cannot** use R^2 to compare two models with different number of predictors, because **adding more predictors will always increase** R^2.
- We can use the **adjusted** R^2:

$$R^2_{adj} = 1 - \frac{n-1}{n-k-1} \frac{\text{SSE}}{\text{SST}}$$

- The adjusted R^2 adds a penalty for the number of predictors in the model.
- The adjusted R^2 is always less than or equal to R^2.

Example

Recall part of the output of the `mtcars` example:

```
Residual standard error: 2.639 on 28 degrees of freedom
Multiple R-squared:  0.8268,Adjusted R-squared:  0.8083
F-statistic: 44.57 on 3 and 28 DF,  p-value: 8.65e-11
```

Example

Recall part of the output of the `mtcars` example:

```
Residual standard error: 2.639 on 28 degrees of freedom
Multiple R-squared:  0.8268,Adjusted R-squared:  0.8083
F-statistic: 44.57 on 3 and 28 DF,  p-value: 8.65e-11
```

- The R^2 is 0.8268, which means 82.68% of the variability in `mpg` can be explained by the model.
- The adjusted R^2 is 0.8083.

Example

Recall part of the output of the `mtcars` example:

```
Residual standard error: 2.639 on 28 degrees of freedom
Multiple R-squared:  0.8268,Adjusted R-squared:  0.8083
F-statistic: 44.57 on 3 and 28 DF,  p-value: 8.65e-11
```

- The R^2 is 0.8268, which means 82.68% of the variability in `mpg` can be explained by the model.
- The adjusted R^2 is 0.8083.
- The F-statistic and the p-value are for the following hypothesis test:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0.$$

- The p-value is very small, which means at least one of the predictors is significant in the model or the model is significant.

Example

However, if we consider a linear regression model of `mpg` on `disp`, `hp`, and `cyl`.

```
Call:
lm(formula = mpg ~ disp + hp + cyl, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.0889 -2.0845 -0.7745  1.3972  6.9183 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 34.18492    2.59078  13.195 1.54e-13 ***
disp        -0.01884    0.01040  -1.811   0.0809 .  
hp          -0.01468    0.01465  -1.002   0.3250    
cyl         -1.22742    0.79728  -1.540   0.1349    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.055 on 28 degrees of freedom
Multiple R-squared:  0.7679,Adjusted R-squared:  0.743 
F-statistic: 30.88 on 3 and 28 DF,  p-value: 5.054e-09
```

Example

However, if we consider a linear regression model of `mpg` on `disp`, `hp`, and `cyl`.

```
Call:
lm(formula = mpg ~ disp + hp + cyl, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0889 -2.0845 -0.7745  1.3972  6.9183

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 34.18492    2.59078  13.195 1.54e-13 ***
disp        -0.01884    0.01040  -1.811   0.0809 .
hp          -0.01468    0.01465  -1.002   0.3250
cyl         -1.22742    0.79728  -1.540   0.1349
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.055 on 28 degrees of freedom
Multiple R-squared:  0.7679,Adjusted R-squared:  0.743
F-statistic: 30.88 on 3 and 28 DF,  p-value: 5.054e-09
```

None of the covariates are significant at $\alpha = 0.05$ level. But they are jointly significant.

Multicollinearity

The **multicollinearity** is a problem when two or more predictors are highly correlated with each other.

Multicollinearity

The **multicollinearity** is a problem when two or more predictors are highly correlated with each other.

- It can cause the estimated coefficients to be unstable and have large standard errors.
- Individual covariates may not be significant, but the model is significant.

Multicollinearity

The **multicollinearity** is a problem when two or more predictors are highly correlated with each other.

- It can cause the estimated coefficients to be unstable and have large standard errors.
- Individual covariates may not be significant, but the model is significant.

To verify it, we can check the correlation matrix of the predictors in prevous example:

```
> cor(mtcars[,c("disp", "hp", 'cyl')])
          disp        hp       cyl
disp 1.0000000 0.7909486 0.9020329
hp   0.7909486 1.0000000 0.8324475
cyl  0.9020329 0.8324475 1.0000000
```

Multicollinearity

To measure the multicollinearity, we can use the **variance inflation factor** (VIF):

$$VIF_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the R^2 of the regression of x_j on all other predictors.

Multicollinearity

To measure the multicollinearity, we can use the **variance inflation factor** (VIF):

$$VIF_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the R^2 of the regression of x_j on all other predictors.

- If $VIF_j > 10$, we consider x_j is highly correlated with other predictors.
- If $5 < VIF_j < 10$, we consider x_j is correlated with other predictors.
- If $1 < VIF_j < 5$, we consider x_j is lightly correlated with other predictors.
- If $VIF_j = 1$, we consider x_j is not correlated with other predictors.

Example

We can use the `vif` function in R to compute the VIF for each predictor:

```
> library(car)
> model = lm(mpg~disp+hp+cyl, mtcars)
> vif(model)
disp      hp        cyl
5.521460 3.350964 6.732984
```

Example

We can use the `vif` function in R to compute the VIF for each predictor:

```
> library(car)
> model = lm(mpg~disp+hp+cyl, mtcars)
> vif(model)
disp       hp      cyl
5.521460 3.350964 6.732984
```

We should consider removing `cyl` from the model.

Higher Order Predictor

In many cases, the dependence between the response and the predictors is not linear:

- The response is a nonlinear function of the predictors.
- The response depends on an interaction between two or more predictors.

Higher Order Predictor

In many cases, the dependence between the response and the predictors is not linear:

- The response is a nonlinear function of the predictors.
- The response depends on an interaction between two or more predictors.
- A linear regression with two predictors x_1 and x_2 can be written as:

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

Higher Order Predictor

In many cases, the dependence between the response and the predictors is not linear:

- The response is a nonlinear function of the predictors.
- The response depends on an interaction between two or more predictors.
- A linear regression with two predictors x_1 and x_2 can be written as:

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

- A linear regression with two predictors x_1 and x_2 and their interaction can be written as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon$$

Higher Order Predictor

In many cases, the dependence between the response and the predictors is not linear:

- The response is a nonlinear function of the predictors.
- The response depends on an interaction between two or more predictors.
- A linear regression with two predictors x_1 and x_2 can be written as:

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

- A linear regression with two predictors x_1 and x_2 and their interaction can be written as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon$$

- A linear regression with two predictors x_1 and x_2 and their quadratic terms can be written as:

$$y =$$

Higher Order Predictor

In many cases, the dependence between the response and the predictors is not linear:

- The response is a nonlinear function of the predictors.
- The response depends on an interaction between two or more predictors.
- A linear regression with two predictors x_1 and x_2 can be written as:

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

- A linear regression with two predictors x_1 and x_2 and their interaction can be written as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon$$

- A linear regression with two predictors x_1 and x_2 and their quadratic terms can be written as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \epsilon$$

Higher Order Predictor

- A linear regression with two predictors x_1 and x_2 and their interaction and quadratic terms can be written as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \epsilon$$

Higher Order Predictor

- A linear regression with two predictors x_1 and x_2 and their interaction and quadratic terms can be written as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \epsilon$$

- A linear regression with two predictors x_1 and x_2 in a nonlinear function can be written as:

$$y = \beta_0 + \beta_1 f_1(x_1) + \beta_2 f_2(x_2) + \epsilon$$

  for some known nonlinear functions f_1 and f_2.

Higher Order Predictor

- A linear regression with two predictors x_1 and x_2 and their interaction and quadratic terms can be written as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \epsilon$$

- A linear regression with two predictors x_1 and x_2 in a nonlinear function can be written as:

$$y = \beta_0 + \beta_1 f_1(x_1) + \beta_2 f_2(x_2) + \epsilon$$

for some known nonlinear functions f_1 and f_2.

Drawbacks:

- It can easily overkill the problem if we add too many higher order terms.
- A natural collinearity between the predictors and the higher order terms.
- Need variable selection to find the best model.

Example

The `trees` dataset in R contains the measurements of the girth, height, and volume of black cherry trees.

Example

The `trees` dataset in R contains the measurements of the girth, height, and volume of black cherry trees.
We can fit a linear regression model of `Volume` on `Girth` and `Height`:

```
Call:
lm(formula = Volume ~ Girth + Height, data = trees)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4065 -2.6493 -0.2876  2.2003  8.4847

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
Girth         4.7082     0.2643  17.816  < 2e-16 ***
Height        0.3393     0.1302   2.607   0.0145 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-squared:  0.948,Adjusted R-squared:  0.9442
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

Example

- All coefficients are significant at $\alpha = 0.05$ level.
- The R^2 is 0.948.

Example

- All coefficients are significant at $\alpha = 0.05$ level.
- The R^2 is 0.948.

In the mean time, we can check the correlation between the variables:

```
> cor(trees)
          Girth    Height    Volume
Girth  1.0000000 0.5192801 0.9671194
Height 0.5192801 1.0000000 0.5982497
Volume 0.9671194 0.5982497 1.0000000
```

Example

- All coefficients are significant at $\alpha = 0.05$ level.
- The R^2 is 0.948.

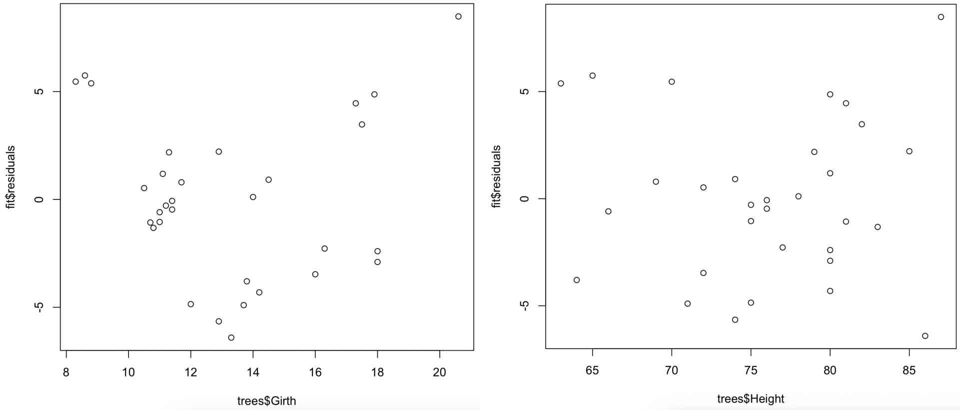
In the mean time, we can check the correlation between the variables:

```
> cor(trees)
          Girth    Height    Volume
Girth  1.0000000 0.5192801 0.9671194
Height 0.5192801 1.0000000 0.5982497
Volume 0.9671194 0.5982497 1.0000000
```

- A moderate correlation between `Girth` and `Height`.
- Multicollinearity is not a serious problem here.

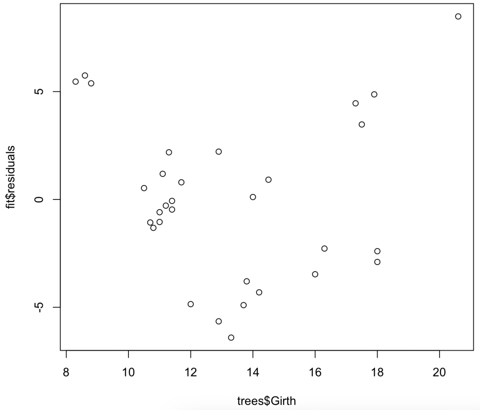
Example

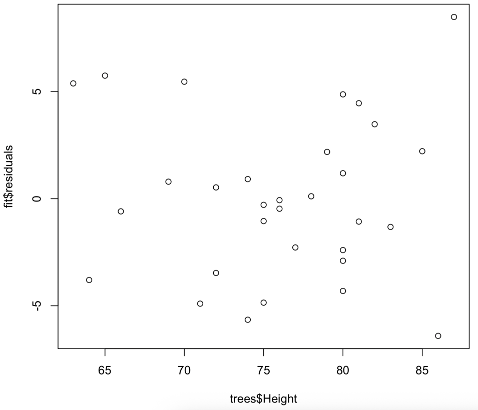
We check the residual of the previous model against the covariates:



Example

We check the residual of the previous model against the covariates:

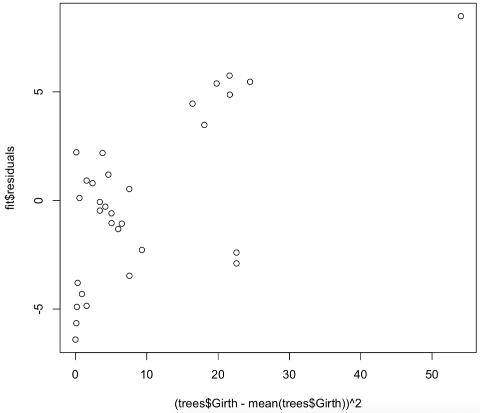


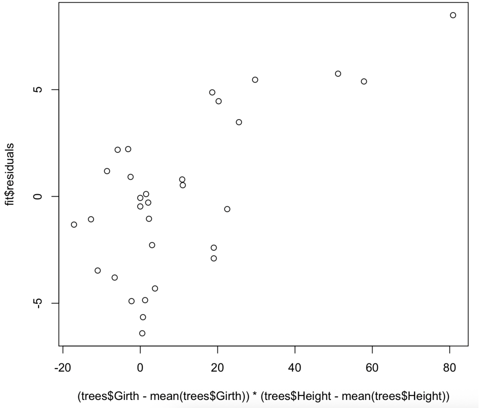


Some quadratic patterns can be observed from the first plot.

Example

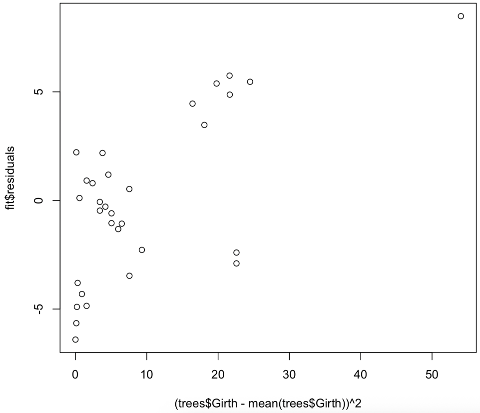
To confirm the second-order polynomial model, we plot the residual against (1) the quadratic term of `Girth` and (2) the interaction term of `Girth` and `Height`:

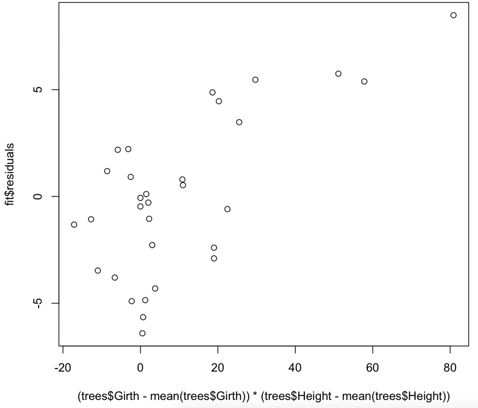




Example

To confirm the second-order polynomial model, we plot the residual against (1) the quadratic term of `Girth` and (2) the interaction term of `Girth` and `Height`:





Both shows an increasing pattern.

Example

Candidate model 1: now we add the quadratic term of `Girth` to the model:

```
Call:
lm(formula = Volume ~ Girth + Height + I(Girth^2), data = trees)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2928 -1.6693 -0.1018  1.7851  4.3489

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.92041   10.07911  -0.984 0.333729
Girth       -2.88508    1.30985  -2.203 0.036343 *
Height       0.37639    0.08823   4.266 0.000218 ***
I(Girth^2)   0.26862    0.04590   5.852 3.13e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.625 on 27 degrees of freedom
Multiple R-squared:  0.9771,Adjusted R-squared:  0.9745
F-statistic: 383.2 on 3 and 27 DF,  p-value: < 2.2e-16
```

Example

Candidate model 2: we add the interaction term of `Girth` and `Height` to the model:

```
Call:
lm(formula = Volume ~ Girth + Height + Girth * Height, data = trees)

Residuals:
    Min      1Q  Median      3Q     Max
-6.5821 -1.0673  0.3026  1.5641  4.6649

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  69.39632   23.83575   2.911  0.00713 **
Girth        -5.85585    1.92134  -3.048  0.00511 **
Height       -1.29708    0.30984  -4.186  0.00027 ***
Girth:Height  0.13465    0.02438   5.524 7.48e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.709 on 27 degrees of freedom
Multiple R-squared:  0.9756,Adjusted R-squared:  0.9728
F-statistic: 359.3 on 3 and 27 DF,  p-value: < 2.2e-16
```

Example

Candidate model 3: we add both terms to the model:

```
Call:
lm(formula = Volume ~ Girth + Height + Girth * Height + I(Girth^2),
    data = trees)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0748 -0.8494  0.0051  1.8396  4.0604

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  26.48906   33.61492   0.788   0.4378
Girth        -4.58977    1.98854  -2.308   0.0292 *
Height       -0.32992    0.62857  -0.525   0.6041
I(Girth^2)    0.17071    0.09762   1.749   0.0921 .
Girth:Height  0.05701    0.05024   1.135   0.2668
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.611 on 26 degrees of freedom
Multiple R-squared:  0.9781,Adjusted R-squared:  0.9748
F-statistic: 290.8 on 4 and 26 DF,  p-value: < 2.2e-16
```

Example

Which model should be choose?

- Model 1: $Volume \sim Girth + Height + I(Girth^2)$
- Model 2: $Volume \sim Girth + Height + Girth * Height$
- Model 3: $Volume \sim Girth + Height + I(Girth^2) + Girth * Height$

Example

Which model should be choose?

- Model 1: $Volume \sim Girth + Height + I(Girth^2)$
- Model 2: $Volume \sim Girth + Height + Girth * Height$
- Model 3: $Volume \sim Girth + Height + I(Girth^2) + Girth * Height$

- Model 1 v.s. Model 2:
  They have the same number of predictors. So we compare their R^2, which suggests Model 1 is better.
- Model 1 v.s. Model 3 and Model 2 v.s. Model 3:
  Both are nested models. We can use the F-test to compare them.

Example

Model 1 v.s. Model 3:

```
> anova(fit1, fit3)
Analysis of Variance Table

Model 1: Volume ~ Girth + Height + I(Girth^2)
Model 2: Volume ~ Girth + Height + Girth * Height + I(Girth^2)
Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     27 186.01
2     26 177.23  1    8.7781 1.2877 0.2668
```

Example

Model 1 v.s. Model 3:

```
> anova(fit1, fit3)
Analysis of Variance Table

Model 1: Volume ~ Girth + Height + I(Girth^2)
Model 2: Volume ~ Girth + Height + Girth * Height + I(Girth^2)
Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     27 186.01
2     26 177.23  1    8.7781 1.2877 0.2668
```

Model 2 v.s. Model 3:

```
> anova(fit2, fit3)
Analysis of Variance Table

Model 1: Volume ~ Girth + Height + Girth * Height
Model 2: Volume ~ Girth + Height + Girth * Height + I(Girth^2)
Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     27 198.08
2     26 177.23  1    20.845 3.0579 0.09214 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Example

After all, we can choose Model 1 as the best model:

$$\text{Volume} = -9.92 - 2.89 \times \text{Girth} + 0.376 \times \text{Height} + 0.269 \times \text{Girth}^2 + \epsilon$$

Example

After all, we can choose Model 1 as the best model:

$$\text{Volume} = -9.92 - 2.89 \times \text{Girth} + 0.376 \times \text{Height} + 0.269 \times \text{Girth}^2 + \epsilon$$

Can we do better than this?

Example

The volume of a cyclindrical tree is given by:

$$V = \pi r^2 h$$

where r is the radius of the tree and h is the height of the tree.

Example

The volume of a cyclindrical tree is given by:

$$V = \pi r^2 h$$

where r is the radius of the tree and h is the height of the tree. Therefore, we conjecture that

$$\text{Volume} \propto \text{Girth}^2 \times \text{Height}$$

Example

The volume of a cyclindrical tree is given by:

$$V = \pi r^2 h$$

where r is the radius of the tree and h is the height of the tree. Therefore, we conjecture that

$$\text{Volume} \propto \text{Girth}^2 \times \text{Height}$$

Hence, we fit a simple linear regression model of `Volume` on $\texttt{Girth}^2 \times \texttt{Height}$ **without intercept**:

$$\text{Volume} = \beta_1 \times \text{Girth}^2 \times \text{Height} + \epsilon$$

Example

```
Call:
lm(formula = Volume ~ 0 + I(Girth^2 * Height), data = trees)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6696 -1.0832 -0.3341  1.6045  4.2944

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
I(Girth^2 * Height) 2.108e-03  2.722e-05   77.44   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.455 on 30 degrees of freedom
Multiple R-squared:  0.995,Adjusted R-squared:  0.9949
F-statistic:  5996 on 1 and 30 DF,  p-value: < 2.2e-16
```

Example

```
Call:
lm(formula = Volume ~ 0 + I(Girth^2 * Height), data = trees)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.6696 -1.0832 -0.3341  1.6045  4.2944 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
I(Girth^2 * Height) 2.108e-03  2.722e-05   77.44   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.455 on 30 degrees of freedom
Multiple R-squared:  0.995,Adjusted R-squared:  0.9949 
F-statistic:  5996 on 1 and 30 DF,  p-value: < 2.2e-16
```

- The result is way much better than the previous models.
- But the coefficient is not

$$\frac{1}{4\pi \times 12} = 0.0066$$

- The tree is not a solid cylinder.

Example

If the tree is not a solid cylinder, literally speaking, it could be any of the following models:

$$\text{Volume} \propto \text{Girth}^{\alpha} \times \text{Height}^{3-\alpha},$$

for some $\alpha \in (0, 3)$.

Example

If the tree is not a solid cylinder, literally speaking, it could be any of the following models:

$$\text{Volume} \propto \text{Girth}^{\alpha} \times \text{Height}^{3-\alpha},$$

for some $\alpha \in (0, 3)$.

- Reason 1: the unit of `Girth` is in foot, and the unit of `Height` is in inches. The unit of `Volume` is in cubic foot.
- Reason 2: the volume is zero if either `Girth` or `Height` is zero.
- We used $\alpha = 2$ to fit the model.

Example

If the tree is not a solid cylinder, literally speaking, it could be any of the following models:

$$\text{Volume} \propto \text{Girth}^{\alpha} \times \text{Height}^{3-\alpha},$$

for some $\alpha \in (0, 3)$.

- Reason 1: the unit of `Girth` is in foot, and the unit of `Height` is in inches. The unit of `Volume` is in cubic foot.
- Reason 2: the volume is zero if either `Girth` or `Height` is zero.
- We used $\alpha = 2$ to fit the model.

How can we determine α?

Example

If we have

$$\text{Volume} = V_0 \times \text{Girth}^{\alpha} \times \text{Height}^{3-\alpha}$$

for some $V_0 > 0$, we can take the logarithm of both sides:

$$\log(\text{Volume}) = \log(V_0) + \alpha \log(\text{Girth}) + (3 - \alpha) \log(\text{Height})$$

Example

If we have

$$\text{Volume} = V_0 \times \text{Girth}^{\alpha} \times \text{Height}^{3-\alpha}$$

for some $V_0 > 0$, we can take the logarithm of both sides:

$$\log(\text{Volume}) = \log(V_0) + \alpha \log(\text{Girth}) + (3 - \alpha) \log(\text{Height})$$

By moving terms around, we have

$$\log(\text{Volume}) - 3 \log(\text{Height}) = \log(V_0) + \alpha \log(\text{Girth}/\text{Height})$$

Example

If we have

$$\text{Volume} = V_0 \times \text{Girth}^{\alpha} \times \text{Height}^{3-\alpha}$$

for some $V_0 > 0$, we can take the logarithm of both sides:

$$\log(\text{Volume}) = \log(V_0) + \alpha \log(\text{Girth}) + (3 - \alpha) \log(\text{Height})$$

By moving terms around, we have

$$\log(\text{Volume}) - 3 \log(\text{Height}) = \log(V_0) + \alpha \log(\text{Girth}/\text{Height})$$

- We can define $y = \log(\text{Volume}) - 3 \log(\text{Height})$.
- We can define $x = \log(\text{Girth}/\text{Height})$.
- Then $\log(V_0)$ and α are the intercept and slope of the linear regression model of y on x.

Example

```
> y = log(trees$Volume) - 3*log(trees$Height)
> x = log(trees$Girth/trees$Height)
> summary(lm(y~x))

Call:
lm(formula = y ~ x)

Residuals:
     Min        1Q    Median        3Q       Max
-0.169031 -0.046756 -0.002936  0.067338  0.134836

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.18569    0.12963  -47.72   <2e-16 ***
x            1.99067    0.07279   27.35   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08043 on 29 degrees of freedom
Multiple R-squared:  0.9627,Adjusted R-squared:  0.9614
F-statistic:   748 on 1 and 29 DF,  p-value: < 2.2e-16
```

Example

- From the output, we have

$$\hat{V}_0 = e^{-6.18569} = 0.000206$$
$$\hat{\alpha} = 1.99067$$

Example

- From the output, we have

$$\hat{V}_0 = e^{-6.18569} = 0.000206$$
$$\hat{\alpha} = 1.99067$$

- The standard error for $\hat{\alpha}$ is 0.07279.
- There is no significant difference between $\hat{\alpha}$ and 2.

Example

- From the output, we have

$$\hat{V}_0 = e^{-6.18569} = 0.000206$$
$$\hat{\alpha} = 1.99067$$

- The standard error for $\hat{\alpha}$ is 0.07279.
- There is no significant difference between $\hat{\alpha}$ and 2.
- Therefore, it is reasonable to use the model:

$$\text{Volume} = 0.000206 \times \text{Girth}^2 \times \text{Height} + \text{error}$$

Catergorical Covariates

A **categorical covariate** is a covariate that takes on a limited number of values.

Catergorical Covariates

A **categorical covariate** is a covariate that takes on a limited number of values.

- The major of the students in a class.
- The gender
- The color of a car.
- Tree species.
- etc..

Catergorical Covariates

A **categorical covariate** is a covariate that takes on a limited number of values.

- The major of the students in a class.
- The gender
- The color of a car.
- Tree species.
- etc..

Catogoritcal variables are known as **factors** as we discussed before in ANOVA.

Catergorical Covariates

A **categorical covariate** is a covariate that takes on a limited number of values.

- The major of the students in a class.
- The gender
- The color of a car.
- Tree species.
- etc..

Catogoritcal variables are known as **factors** as we discussed before in ANOVA. It does not make sense to run a linear regression on a categorical variable directly. Be need to encode them into numerical variables.

One-hot Encoding

Let x_{ji} be the j-th variable (categorical) of the i-th observation such that

$$x_{ji} \in \{\text{cat 1}, \text{cat 2}, \ldots, \text{cat D}\}$$

That is x_{ji} can take on D different values.

One-hot Encoding

Let x_{ji} be the j-th variable (categorical) of the i-th observation such that

$$x_{ji} \in \{\text{cat 1}, \text{cat 2}, \ldots, \text{cat D}\}$$

That is x_{ji} can take on D different values.

The **one-hot encoding** of x_{ji} is to creat d binary variables:

$$x_{j1i}, x_{j2i}, \ldots, x_{jDi}$$

where

$$x_{jdi} = \begin{cases} 1 & \text{if } x_{ji} = \text{cat d} \\ 0 & \text{otherwise} \end{cases}$$

One-hot Encoding

Let x_{ji} be the j-th variable (categorical) of the i-th observation such that

$$x_{ji} \in \{\text{cat 1}, \text{cat 2}, \ldots, \text{cat D}\}$$

That is x_{ji} can take on D different values.

The **one-hot encoding** of x_{ji} is to creat d binary variables:

$$x_{j1i}, x_{j2i}, \ldots, x_{jDi}$$

where

$$x_{jdi} = \begin{cases} 1 & \text{if } x_{ji} = \text{cat d} \\ 0 & \text{otherwise} \end{cases}$$

- The d-th variable is 1 if the j-th variable is in the d-th category, and 0 otherwise.
- The variables are call **dummy variables**.
- For each observation, only one of the d dummy variables is 1, and all others are 0.

One-hot Encoding

One problem of the one-hot encoding is that the D dummy variables are not independent. To see this, we have

$$x_{j1i} + x_{j2i} + \cdots + x_{jDi} = 1.$$

One-hot Encoding

One problem of the one-hot encoding is that the D dummy variables are not independent. To see this, we have

$$x_{j1i} + x_{j2i} + \cdots + x_{jDi} = 1.$$

In practice we only need $D - 1$ dummy variables:

$$x_{j1i}, x_{j3i}, \ldots, x_{j(D-1)i}$$

- When all $D - 1$ dummy variables are 0, the j-th variable is in the last category.
- The last category is called the **reference category**.

Linear Regression on Categorical Variables

Now consider a linear regression model with 2 predictors x_1 and x_2, where x_2 is a categorical variable with D categories.

Linear Regression on Categorical Variables

Now consider a linear regression model with 2 predictors x_1 and x_2, where x_2 is a categorical variable with D categories.

- The model is $y \sim x_1 + x_2$.
- But x_2 is a categorical variable,

Linear Regression on Categorical Variables

Now consider a linear regression model with 2 predictors x_1 and x_2, where x_2 is a categorical variable with D categories.

- The model is $y \sim x_1 + x_2$.
- But x_2 is a categorical variable, we need to use $D-1$ dummy variables:

$$y \sim x_1 + x_{21} + x_{22} + \cdots + x_{2(D-1)}$$

- The model can be written as:

$$y = \beta_0 + \beta_1 x_1 + \beta_{21} x_{21} + \cdots + \beta_{2(D-1)} x_{2(D-1)} + \epsilon$$

Linear Regression on Categorical Variables

Now consider a linear regression model with 2 predictors x_1 and x_2, where x_2 is a categorical variable with D categories.

- The model is $y \sim x_1 + x_2$.
- But x_2 is a categorical variable, we need to use $D-1$ dummy variables:

$$y \sim x_1 + x_{21} + x_{22} + \cdots + x_{2(D-1)}$$

- The model can be written as:

$$y = \beta_0 + \beta_1 x_1 + \beta_{21} x_{21} + \cdots + \beta_{2(D-1)} x_{2(D-1)} + \epsilon$$

- The linear regression becomes a multiple linear regression with D predictors.

Linear Regression on Categorical Variables

$$y = \beta_0 + \beta_1 x_1 + \beta_{21} x_{21} + \cdots + \beta_{2(D-1)} x_{2(D-1)} + \epsilon$$

Linear Regression on Categorical Variables

$$y = \beta_0 + \beta_1 x_1 + \beta_{21} x_{21} + \cdots + \beta_{2(D-1)} x_{2(D-1)} + \epsilon$$

Interpretation of the coefficients:

- β_0 is the intercept of the model for the reference category.
- β_1 is the slope of the model for all categories.
- β_{2d} is the **difference** between the intercept of the d-th category and the reference category.

Example

We use the `mtcars` dataset and treat `cyl` as a categorical variable.

Example

We use the `mtcars` dataset and treat `cyl` as a categorical variable.

```
Call:
lm(formula = mpg ~ disp + factor(cyl), data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-4.8304 -1.5873 -0.5851  0.9753  6.3069

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  29.53477    1.42662  20.703  < 2e-16 ***
disp         -0.02731    0.01061  -2.574  0.01564 *
factor(cyl)6 -4.78585    1.64982  -2.901  0.00717 **
factor(cyl)8 -4.79209    2.88682  -1.660  0.10808
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.95 on 28 degrees of freedom
Multiple R-squared:  0.7837,Adjusted R-squared:  0.7605
F-statistic: 33.81 on 3 and 28 DF,  p-value: 1.906e-09
```

Example

- The `cyl` variable has 3 categories: 4, 6, and 8.
- The reference category is 4.
- Two dummy variables are created: `factor(cyl)6` and `factor(cyl)8`.
- For 4-cylinder cars, the model is:

$$\text{mpg} = 29.53 - 0.02731 \times \texttt{disp} + \epsilon$$

- For 6-cylinder cars, the model is:

$$\text{mpg} = 29.53 - 0.02731 \times \texttt{disp} - 4.79 + \epsilon$$

- For 8-cylinder cars, the model is:

$$\text{mpg} = 29.53 - 0.02731 \times \texttt{disp} - 4.79 + \epsilon$$

How can we make the slope to depend on the cylinder?

Example

We can add the interaction term of disp and cyl to the model:

```
Call:
lm(formula = mpg ~ disp * factor(cyl), data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.4766 -1.8101 -0.2297  1.3523  5.0208 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)       40.87196    3.02012  13.533 2.79e-13 ***
disp              -0.13514    0.02791  -4.842 5.10e-05 ***
factor(cyl)6     -21.78997    5.30660  -4.106 0.000354 ***
factor(cyl)8     -18.83916    4.61166  -4.085 0.000374 ***
disp:factor(cyl)6  0.13875    0.03635   3.817 0.000753 ***
disp:factor(cyl)8  0.11551    0.02955   3.909 0.000592 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.372 on 26 degrees of freedom
Multiple R-squared:  0.8701,Adjusted R-squared:  0.8452 
F-statistic: 34.84 on 5 and 26 DF,  p-value: 9.968e-11
```

Example

- The model for 4-cylinder cars is:

$$\text{mpg} = 40.87 - 0.13514 \times \texttt{disp} + \epsilon$$

- The model for 6-cylinder cars is:

$$\text{mpg} = (40.87 - 21.79) + (-0.13514 + 0.139) \times \texttt{disp} + \epsilon$$

- The model for 8-cylinder cars is:

$$\text{mpg} = (40.87 - 18.84) + (-0.13514 + 0.1155) \times \texttt{disp} + \epsilon$$