MIE237

Neil Montgomery

2016-03-01

Model assumption wrap-up

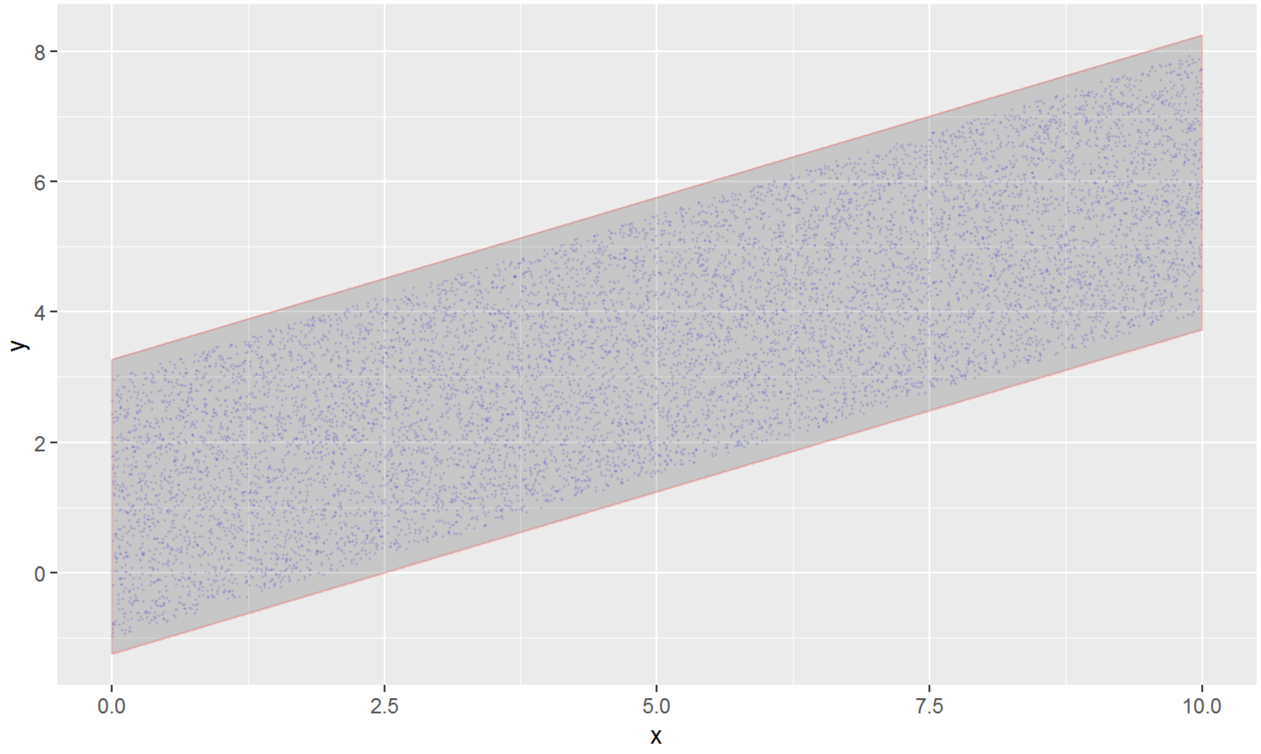
Non-linearity and non-equal variances are fatal flaws. Options are to consider different models possibly including transforming variables.

Non-normality is not necessarily fatal. With a large enough sample, the following calculations are still approximately correct:

1. The p-value for $H_0 : \beta_1 = 0$ versus $H_0 : \beta_1 \neq 0$
2. The confidence interval for β_1
3. The confidence interval for the mean response at x_0.

But non-normality fatal for the prediction interval at x_0.

Non-normality and prediction intervals



In this example the simulated error follows a uniform distribution rather than a normal distribution. The sample size is $n = 10000$. Pointwise prediction intervals are in red.

Textbook notes

We have covered 11.1 to 11.6, 11.8, the diagnostic parts of 11.10, and 11.12.

I wouldn't bother with 11.7. It is a strange little section out of place in Chapter 11.

11.9 concerns formally testing for non-linearity in a specific situation that don't come up all that much in practice, so we skip this section.

Much of 11.10 concerns solving model violation problems, which can be summarized simply as "try taking logarithms or square roots of one or both variables until the situation improves" and I won't cover it directly. The diagnostic plots were covered in this course.

11.11 is a case study, which is fine to look at.

Regression in general

Models

We've analyzed the special case $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ with $\varepsilon_i \sim N(\mu, \sigma^2)$.

The linear regression model in general ("multiple regression") is:

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \varepsilon_i$$

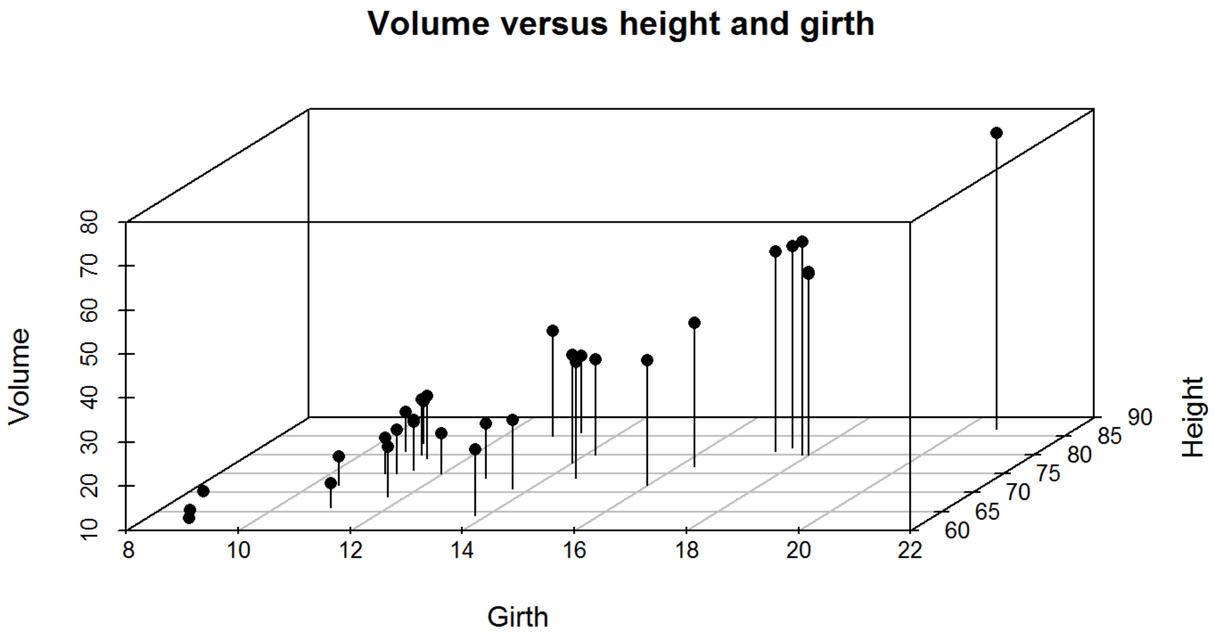
Matrix notation is more efficient—will revisit this momentarily.

y is random. The inputs are not random. They can be whatever you like, even functions of one another, with one technical limitation, which we'll see momentarily. So these are valid linear models:

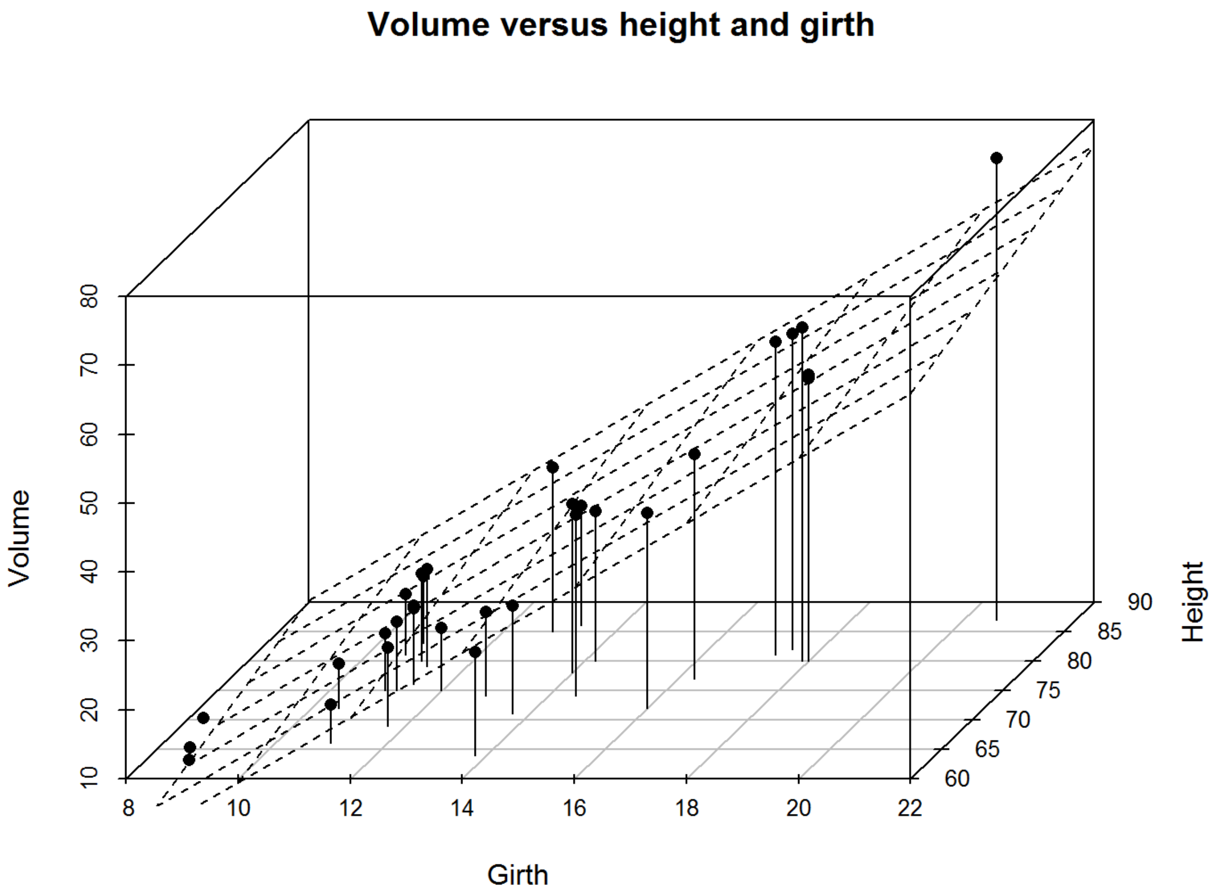
$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i, \quad \text{both dummy variables} \\ y_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \varepsilon_i \\ y_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 \log x_{1i} + \varepsilon_i \end{aligned}$$

What is being done?

R comes with some sample datasets. One is called `trees` and has variables Girth, Height, and Volume. Here's a 3d plot:



Fitting a surface to the points



General notation

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

The ε_i are independent $N(0, \sigma^2)$ random variables.

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}$$

The Fundamental Issues

- Familiar issues with similar answers
    - Parameter testing and estimation
    - Mean response and prediction
    - Model assumptions
- New issues:
    - Parameter interpretation
    - Model selection: which variables?
    - "Multicollinearity" (correlated inputs)

Multiple regression parameter estimation

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \varepsilon_i$$

β_i is:

- the change in y
- given an increase of one unit of x_i
- **given [values of] all other variables in the model**.

Multiple Regression Parameter Hypothesis Testing (Interpretation)

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \varepsilon_i$$

The canonical hypothesis test for a single parameter:

$$\begin{aligned} H_0 &: \beta_i = 0 \\ H_1 &: \beta_i \neq 0 \end{aligned}$$

If H_0 is true, it means the ith variable (x_i) isnot significantly related to y…

…**given all the other x's in the model**

Multiple Regression Parameter Hypothesis Testing (Interpretation)

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \varepsilon_i$$

"Is there any linear relationship between y and the input variables?"

Informally expressed as:

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ H_1 &: \text{Any } \beta_i \neq 0 \end{aligned}$$

Parameter estimates in matrix form

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

The matrix $\boldsymbol{X}'\boldsymbol{X}$, and the fact is must be inverted, plays a key role in multiple regression analysis.

For instance, let's divine the technical requirements on the input variables X.

Requirement: no linear dependence (includes "no constants").

Properties of the parameter estimates

Expected values and variances of vectors of random variables works in a non-shocking way.

A summary of the important results:

$$E(\boldsymbol{y}) = \boldsymbol{\beta}$$

So we have $\hat{\boldsymbol{\beta}}$ is unbiased for $\boldsymbol{\beta}$.

Also:

$$\text{Var}\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2 (\boldsymbol{X}' \boldsymbol{X})^{\boldsymbol{-1}}$$

Let's denote the diagonal of $(\boldsymbol{X}' \boldsymbol{X})^{\boldsymbol{-1}}$ by:

$$c_{00}, c_{11}, \ldots, c_{kk}$$

Distribution of components of $\boldsymbol{\beta}$

Same as always:

$$\frac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{c_{ii}}} \sim N(0, 1)$$

We don't know the value of σ^2. So we do the same thing as in simple regression.

Denote by $\hat{y}_i$ the fitted value at $\boldsymbol{x_i}$:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \ldots + \hat{\beta}_k x_{ki}$$

If we denote the i^{th} row of the $\boldsymbol{X}$ matrix by $\boldsymbol{x_i'} = [1 \quad x_{1i} \quad x_{2i} \;\cdots\; x_{ki}]$ this can be re-written as $\hat{y}_i = \boldsymbol{x_i'}\hat{\boldsymbol{\beta}}$.

The residuals $\hat{\varepsilon}_i = y_i - \hat{y}_i$ are also the same as before.

Estimating the error variance

We end up with the sum of squares decompostion, which is also the same as before:

$$\sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2 = \sum_{i=1}^{n} \left(\hat{y}_i - \overline{y}\right)^2 + \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$$

They are all sums of squares of normal distributions.

So they have χ^2 distributions. The degrees of freedom are $n-1$, k, and $n-(k+1)$, respectively. (They add up!)

And we have our estimator for σ^2:

$$MSE = \frac{SSE}{n-(k+1)}$$

Distribution of components of β again

$$\frac{\hat{\beta}_i - \beta_i}{\sqrt{MSE}\sqrt{c_{ii}}} \sim ???$$

Confidence intervals and hypothesis tests proceed as usual.

trees example

```
## 
## Call:
## lm(formula = Volume ~ Girth + Height, data = .)
## 
## 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          0.000000275 ***
## Girth         4.7082     0.2643  17.816 < 0.0000000000000002 ***
## 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: < 0.00000000000000022
```

The "overall" hypothesis test

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ H_1 &: \text{Any } \beta_i \neq 0 \end{aligned}$$

Similar to before. A few slides ago we had $SST = SSR + SSE$ with $n-1$, k, and $n-(k+1)$ degrees of freedom. From this we can define $MSR = SSR/k$ and we use:

$$\frac{MSR}{MSE} \sim F_{k,n-(k+1)}$$

(Note: there is no $T^2 = F$ relationship to be had, unlike in the simple regression case.)

trees example again

```
##
## Call:
## lm(formula = Volume ~ Girth + Height, data = .)
##
## 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          0.000000275 ***
## Girth         4.7082     0.2643  17.816 < 0.0000000000000002 ***
## 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: < 0.00000000000000022
```

trees ANOVA table

```
trees %>%
  lm(Volume ~ Girth + Height, data = .) %>%
  anova
```

```
## Analysis of Variance Table
##
## Response: Volume
##           Df Sum Sq Mean Sq  F value               Pr(>F)
## Girth      1 7581.8  7581.8 503.1503 < 0.0000000000000002 ***
## Height     1  102.4   102.4   6.7943              0.01449 *
## Residuals 28  421.9    15.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

R^2

Same as before:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

Same old meaning, now with even more potential for abuse!

(Note: square root of R^2 is now nothing in particular in multiple regression)