Introduction to Statistical Modeling

Multiple Linear Regression

Joris Vankerschaver

Example: Mineral composition versus growth

- Case study: 26 8-year-old Japanese larches
- **Goal**: study association between height Y of tree (in cm) and mineral composition of needles:
  - Nitrogen X_N
  - Phosphorus X_P
  - Potassium X_K
  - Residual ash X_r
- Independent variables expressed in percentage observed in dried needles.

Image source: Wikipedia.

Simple linear regression not possible

- Separate linear regression models

$$E(Y|X_P) = \alpha + \beta_P X_P$$

  only allow to predict based on 1 mineral.
- More accurate predictions based on all minerals simultaneously.
- Separate models might not show **pure effect**
  - β_P is mean difference in length between trees that differ 1 unit in proportion phosphorus.
  - **Confounding**: even if phosphorus would not have influence on length, trees with higher level of phosphorus might be taller because they contain, for example, more potassium.
  - Necessity to compare trees with different level of phosphorus, but same level of potassium.

Multiple linear regression

- Assume that

$$E(Y|X_N, X_P, X_K, X_r) = \alpha + \beta_N X_N + \beta_P X_P + \beta_K X_K + \beta_r X_r$$

for unknown **intercept** α and **slopes** $\beta_N, \beta_P, \beta_K, \beta_r$.

- Now prediction based on multiple minerals possible.
- Confounding partially circumvented:

$$\begin{aligned} & E(Y|X_N, X_P = x_P + \delta, X_K, X_r) \\ & \qquad - E(Y|X_N, X_P = x_P, X_K, X_r) \\ & = \alpha + \beta_N X_N + \beta_P (x_P + \delta) + \beta_K X_K + \beta_r X_r \\ & \qquad - \alpha - \beta_N X_N - \beta_P x_P - \beta_K X_K - \beta_r X_r = \beta_P \delta \end{aligned}$$

$\beta_P =$ difference in mean length between trees that differ 1 unit in proportion phosphorus, but have **same value for other explaining variables**.

Analysis of larches

```
model_l <- lm(length ~ phosphor)
summary(model_l)$coefficients
```

```
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept)  -69.10736   45.99195 -1.502597 1.45988e-01
phosphor    1060.29029  177.07503  5.987802 3.51108e-06
```

We call in this case the association between phosphorus and length **unadjusted**.

Analysis of larches

Parameters estimated using least squares method:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-185.32987	36.29808	-5.105776	4.670943e-05
nitrogen	97.76404	24.57181	3.978708	6.836171e-04
phosphor	256.97496	169.90534	1.512460	1.453213e-01
potassium	126.57293	46.42886	2.726169	1.265285e-02
residu	40.27678	36.61454	1.100021	2.837734e-01

We say in this case that the association between phosphorus and length is **adjusted** for nitrogen, potassium, and residual ash.

Tests and confidence intervals

- Tests and confidence intervals for parameter β based on

$$\frac{\hat{\beta} - \beta}{SE(\hat{\beta})} \sim t_{n-p}$$

  with p number of unknown parameters in model.
- Or directly using

```
confint(model_l2)
```

Analysis of larches

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-185.32987	36.29808	-5.105776	4.670943e-05
nitrogen	97.76404	24.57181	3.978708	6.836171e-04
phosphor	256.97496	169.90534	1.512460	1.453213e-01
potassium	126.57293	46.42886	2.726169	1.265285e-02
residu	40.27678	36.61454	1.100021	2.837734e-01

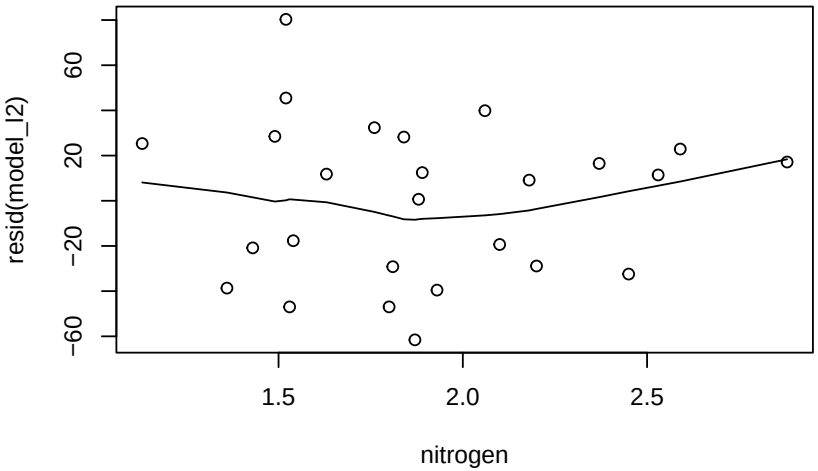
- **95% CI** for β_N needs $t_{21,0.975} = 2.08$

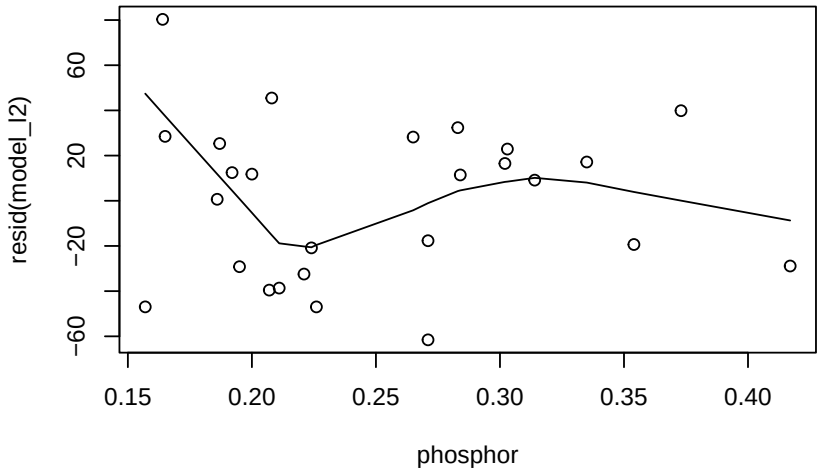
$$[97.76 \pm 2.08 \times 24.57] = [46.66, 148.86].$$

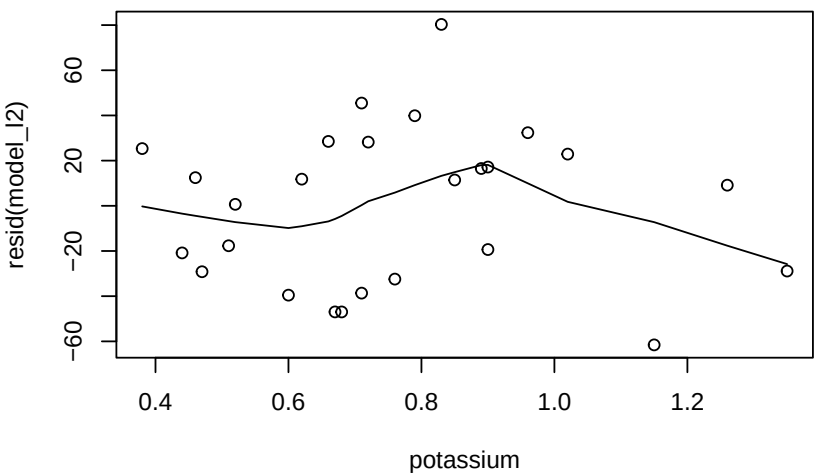
- **95% CI** for β_P

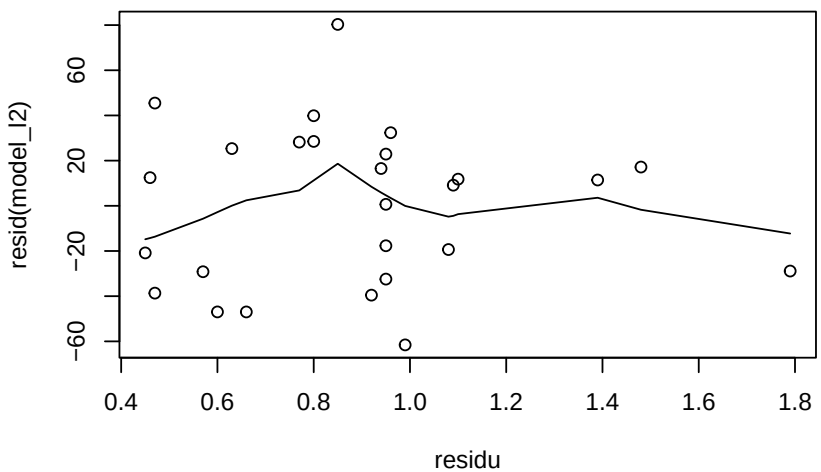
$$[256.97 \pm 2.08 \times 169.91] = [-96.36, 610.31].$$

Analysis of larches: individual residual plots

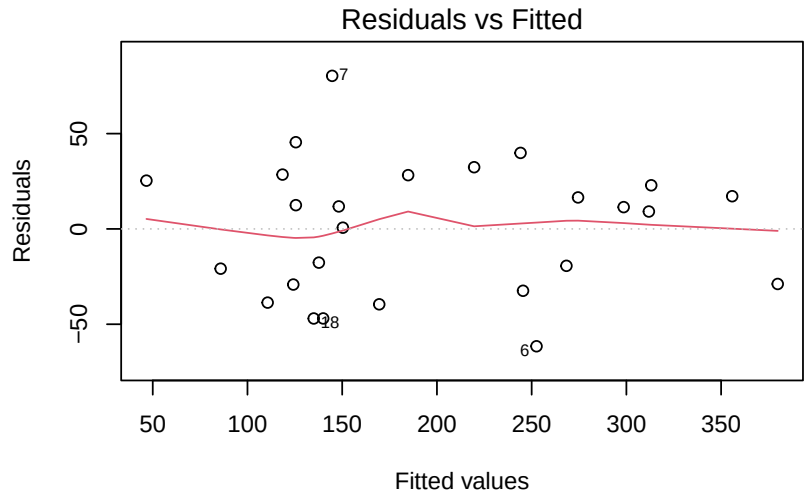


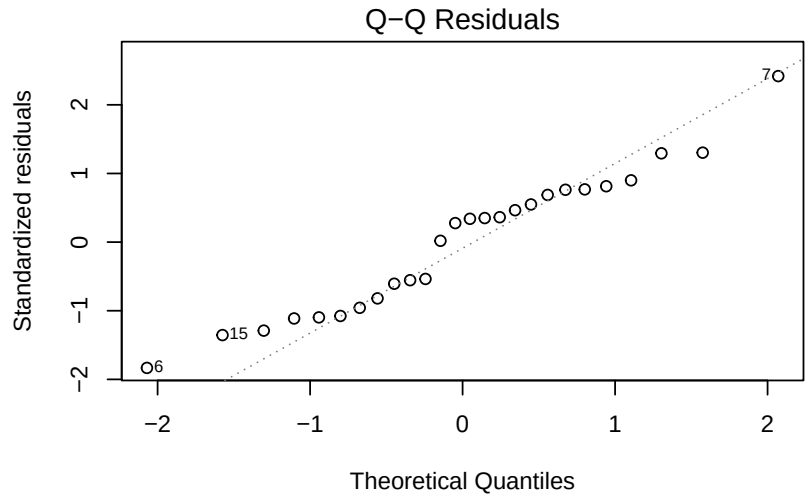


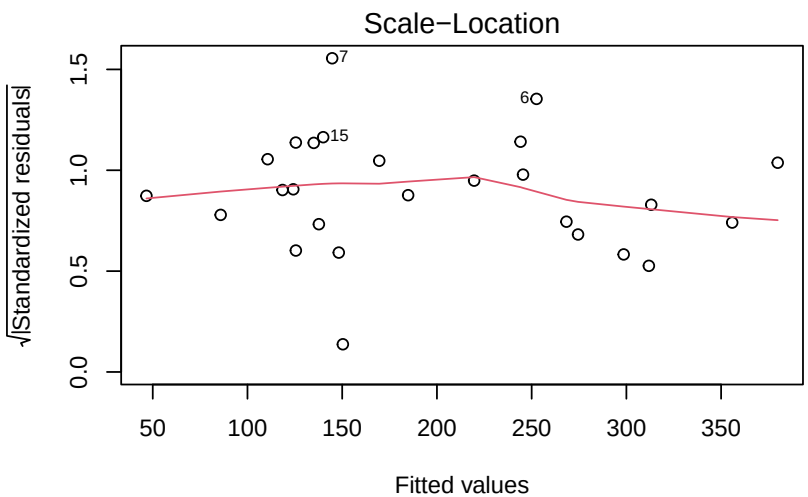


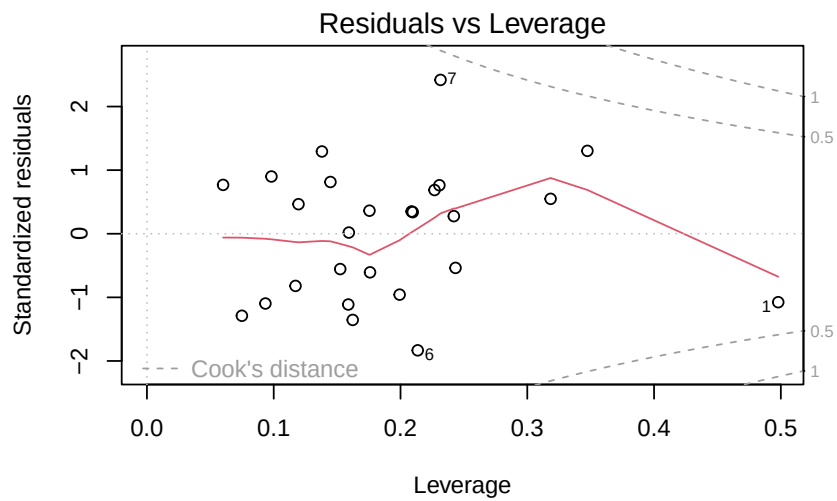


Analysis of larches: residual plots









Interaction or effect modification

- **Interaction** or **effect modification**: effect of variable X on outcome Y depends on third variable Z.
- Examples:
  - Effect of nitrogen on growth depends on proportion phosphorus
  - Pharmacogenetics: effect of steroids for asthma on lung function depends on certain genes.
  - Gene-environment interactions: effect of certain genes on COPD depends on history of smoking.
- Model interactions through **cross-product term**:

$$E(Y|X_N, X_P) = \alpha + \beta_N X_N + \beta_P X_P + \beta_{NP} X_N X_P.$$

Interpretation

$$
\begin{aligned}
&E(Y|X_N = x_N + 1, X_P) - E(Y|X_N = x_N, X_P) \\
&\quad = \alpha + \beta_N(x_N + 1) + \beta_P X_P + \beta_{NP}(x_N + 1)X_P \\
&\qquad - \alpha - \beta_N x_N - \beta_P X_P - \beta_{NP} x_N X_P \\
&\quad = \beta_N + \beta_{NP} X_P.
\end{aligned}
$$

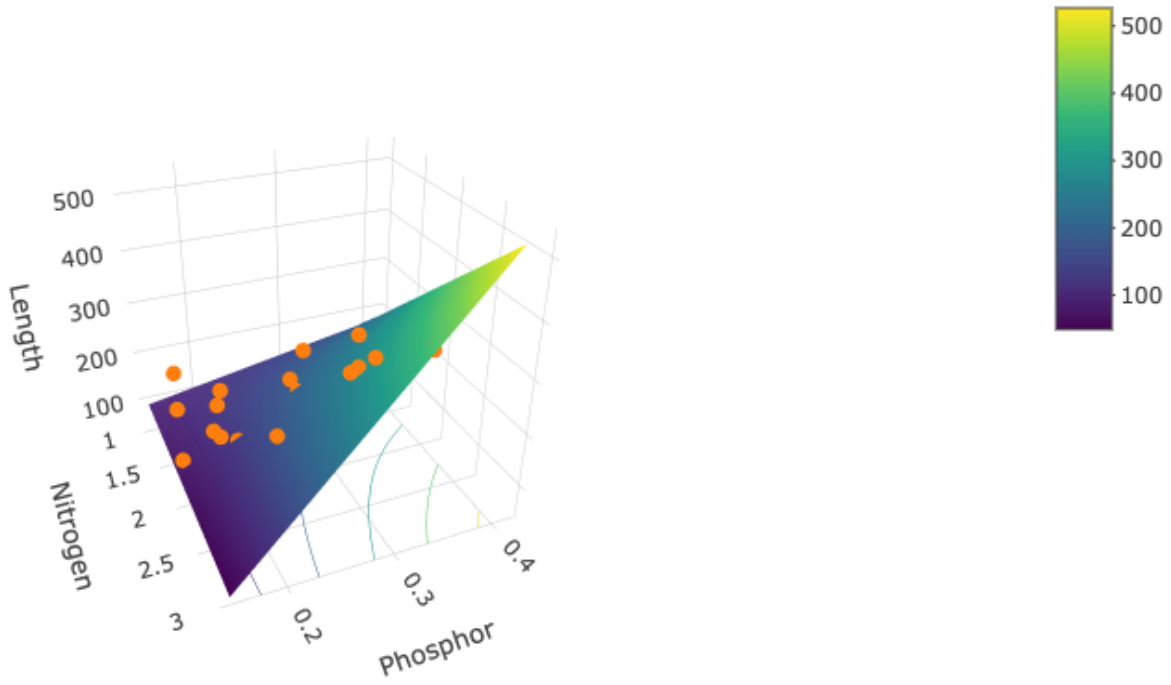
- β_{NP} is difference in nitrogen **effect** between trees that differ 1 percentage in phosphorus.
- Decide whether nitrogen effect depends on quantity phosphorus can be done by testing if $\beta_{NP} = 0$.
- β_N is effect of 1 percentage increase in nitrogen when percentage phosphorus is 0.

Analysis of larches with interaction

```
model_l3 <- lm(length ~ nitrogen * phosphor)
summary(model_l3)
```

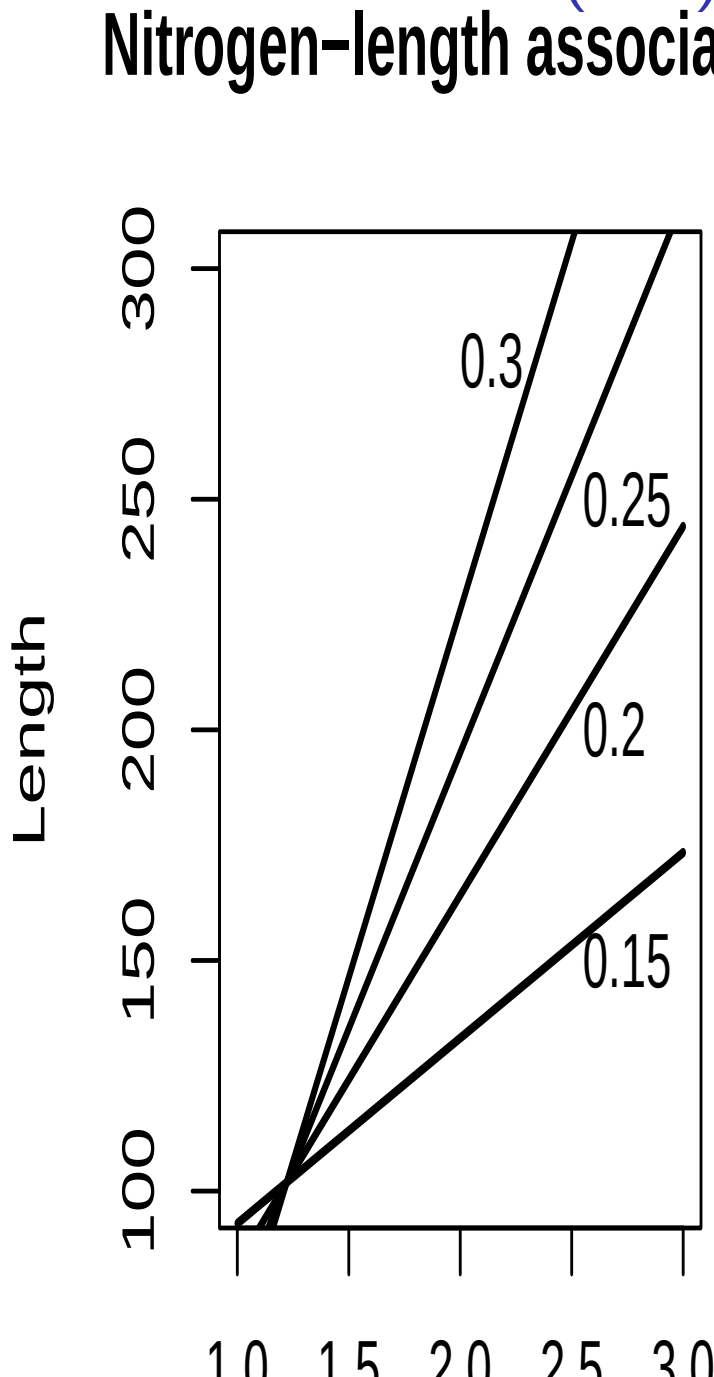
```
Call:
lm(formula = length ~ nitrogen * phosphor)

Residuals:
    Min      1Q  Median      3Q     Max
-57.533 -32.025   0.205  23.121 107.795

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         198.42     211.94   0.936   0.3593
nitrogen            -79.04     111.67  -0.708   0.4865
phosphor           -971.01     858.65  -1.131   0.2703
nitrogen:phosphor   794.97     426.20   1.865   0.0755 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 42.99 on 22 degrees of freedom
Multiple R-squared:  0.8216,    Adjusted R-squared:  0.7973
F-statistic: 33.78 on 3 and 22 DF,  p-value: 2.057e-08
```

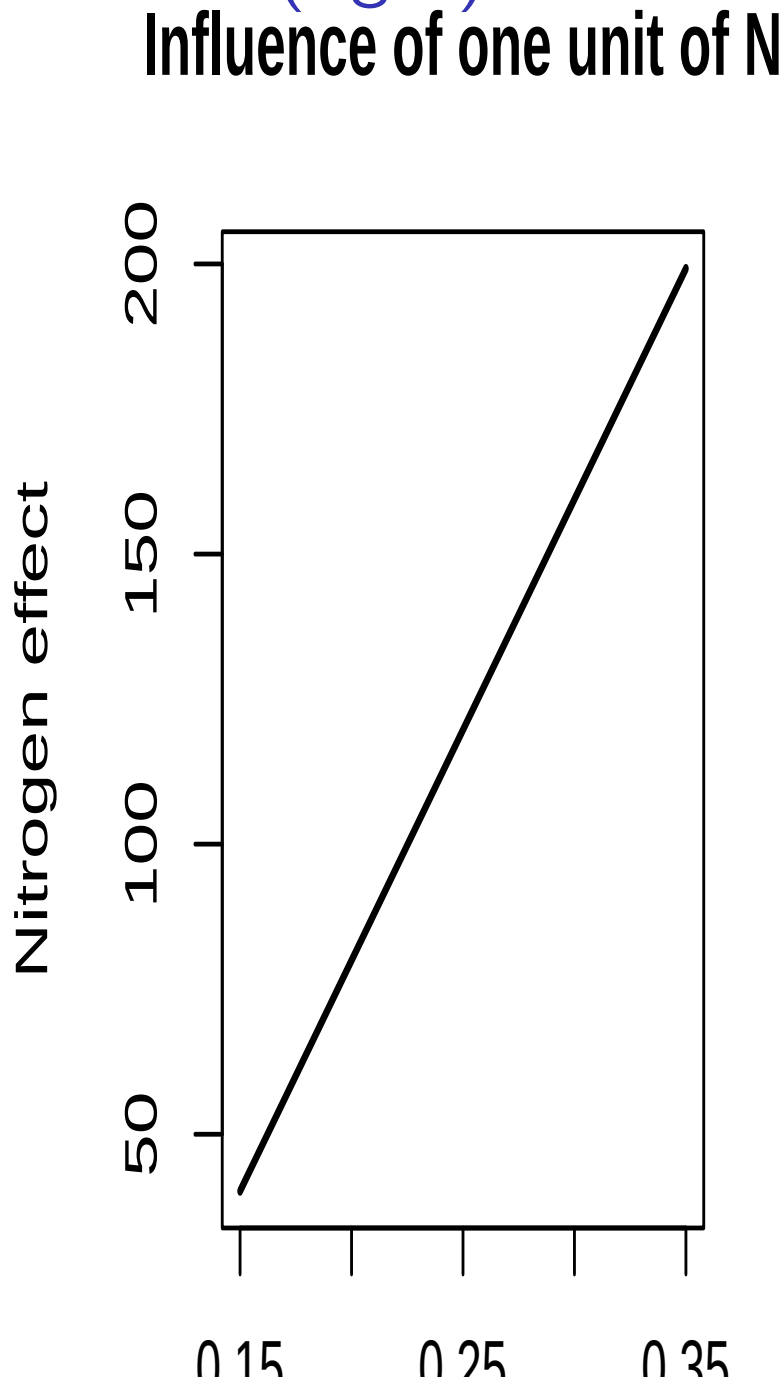
Predicted association (full model, interaction)



Interactive visualization at
https://shiny-stats.fly.dev/multi-regression/

Predicted association (left) and effect (right)



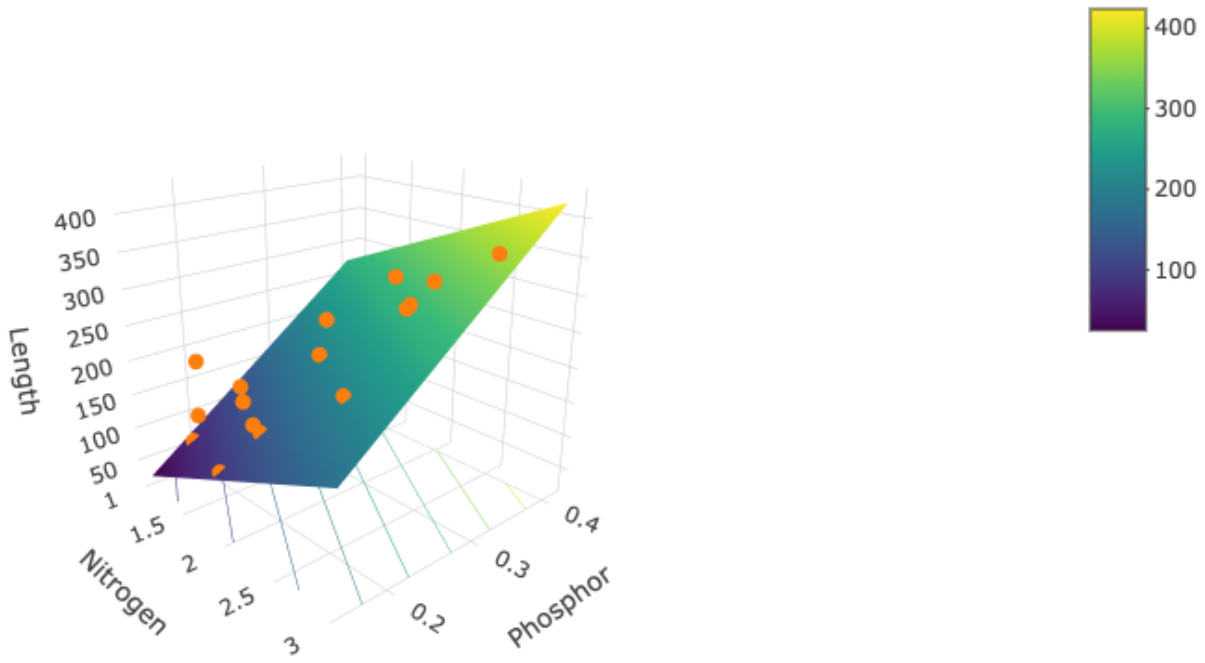


Analysis of larches without interaction

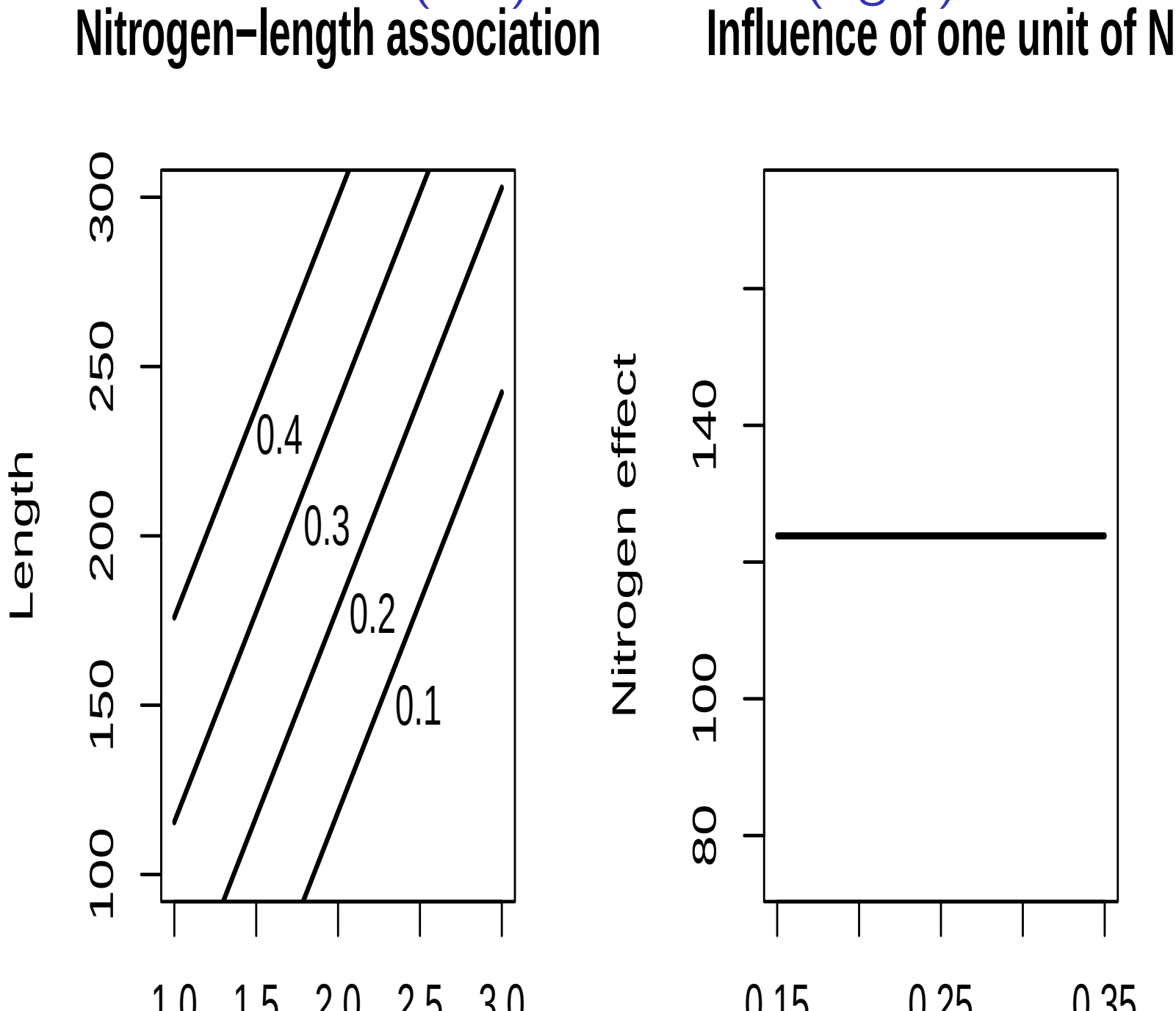
```r
model_l4 <- lm(length ~ nitrogen + phosphor)
summary(model_l4)
```

```
Call:
lm(formula = length ~ nitrogen + phosphor)

Residuals:
    Min      1Q  Median      3Q     Max
-57.834 -34.950  -0.539  20.364 127.287

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -189.64      42.53  -4.460 0.000179 ***
nitrogen      123.83      26.62   4.652 0.000111 ***
phosphor      604.44     162.65   3.716 0.001135 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 45.25 on 23 degrees of freedom
Multiple R-squared:  0.7934,    Adjusted R-squared:  0.7755
F-statistic: 44.17 on 2 and 23 DF,  p-value: 1.329e-08
```

Predicted association (full model, no interaction)



Predicted association (left) and effect (right)



Constructing statistical models

Statistical models are built hierarchically by adding or removing one predictor at a time:

- **Forward** model construction: starts from empty model.
- **Backward** model construction: starts from full model, i.e., one that contains all available predictors.
- **Stepwise** model construction: combination of both.

Constructing statistical models

Forward model construction:

- Include predictors **1 by 1**
- After each inclusion:
  - Verify if certain predictors in the model are no longer significantly associated with outcome.
  - Remove those **1 by 1** until model contains only significant predictors.
- Repeat until no inclusion is significant

Backward model construction:

- Exclude 1 by 1 the **nonsignificant** predictors.
- Until model only contains significant predictors.

In either case, the included (excluded) predictor is ideally the one that is most (least) strongly associated with outcome, after verification for the other predictors in the model.

Constructing statistical models

- Once all **first order terms** have been verified, investigate **higher order terms** (interactions, quadratic effects, ...)
- This is **not** done in exhaustive way:
  - Because number of higher order terms can get very large.
  - Because problem of multiple testing might sometimes lead relatively easy by mere coincidence to incorrect conclusion that certain higher order term is significant.
- Higher order terms considered for inclusion based on:
  - Biological judgment
  - Insight from residual plots.
- Once final model is obtained, verify through residual plots.

Analysis of larches: provisonally 'final' model

```
                  Estimate Std. Error    t value    Pr(>|t|)
(Intercept)        129.0901  169.32526  0.7623795 0.455194582
nitrogen          -150.7635   97.07997 -1.5529827 0.136925913
phosphor         -1000.0165  682.98384 -1.4641877 0.159492670
potassium          137.9659   41.23717  3.3456675 0.003396905
residu             193.8021   89.10355  2.1750205 0.042462293
nitrogen:phosphor  951.7823  371.56807  2.5615287 0.019086203
phosphor:residu   -598.0778  290.01979 -2.0621964 0.053134178

Residual standard error: 33.43 on 19 degrees of freedom
Multiple R-squared:  0.9069,    Adjusted R-squared:  0.8774
F-statistic: 30.83 on 6 and 19 DF,  p-value: 8.159e-09
```

Constructing statistical models

- How to build statistical models **depends ideally on goal** of these models:
  - Making predictions
  - Determining effect of an exposure on an outcome.
- To make predictions, previous strategy is sensible since it aims at avoiding overfitting.
- To determine effect of an exposure on an outcome, must make sure above all to adjust for all confounders
- Recall: Confounder are factors
  - that are not comparable between exposure groups
  - in particular, that are associated with outcome and exposure, but are not influenced by either.