Introduction to Statistical Modeling

Predictivity and variability

Joris Vankerschaver

Prediction

Example

Use model to predict length of larch based on mineral composition of needles.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	160.66283	175.61424	0.9148622	0.370649894
nitrogen	-76.49677	92.34000	-0.8284250	0.416746264
phosphor	-1120.70470	711.42841	-1.5752881	0.130135986
potassium	138.06170	41.29966	3.3429260	0.003084272
nitrogen:phosphor	724.38231	353.05353	2.0517634	0.052870451

- Percentages: nitrogen $= 1.9$, phosphorus $= 0.2$, potassium $= 0.7$.
- Predicted **average** length:

$$160.66 - 76.5 \times 1.9 - 1120.7 \times 0.2 + 138.06 \times 0.7 + 724.38 \times 1.9 \times 0.2 = 163.1.$$

Accuracy of prediction

To determine the accuracy of a prediction, we need to take into account the

- **variability** of the observations around the regression line
- **precision** of the estimated regression line.

Estimating variability via the residual standard error

Residual standard error (RSE):

- CWD basal area: 1.01 on 13 degrees of freedom
- Larches: 35.55 on 21 degrees of freedom.

Residual standard deviation tells that 95% of lengths, given nitrogen, phosphorus and potassium percentages of 1.9, 0.2 and 0.7, are expected to lie within a distance

$$2 \times 35.55 = 71.1$$

of the mean.

Residual standard error in R

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

Residuals:
    Min      1Q  Median      3Q     Max
-54.051 -24.544   5.934  21.866  69.243

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         160.66     175.61   0.915  0.37065
nitrogen            -76.50      92.34  -0.828  0.41675
phosphor          -1120.70     711.43  -1.575  0.13014
potassium           138.06      41.30   3.343  0.00308 **
nitrogen:phosphor   724.38     353.05   2.052  0.05287 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 35.55 on 21 degrees of freedom
Multiple R-squared:  0.8836,    Adjusted R-squared:  0.8614
F-statistic: 39.85 on 4 and 21 DF,  p-value: 1.603e-09
```

Residual standard error (by hand)

RSE can be calculated as

$$RSE = \sqrt{\frac{SSE}{n-p}} = \sqrt{MSE}$$

with SSE, the sum-squared of the residuals, given by

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2.$$

and p the number of parameters in the model.

For example (larches, $p = 5$):

```
SSE <- sum(model_l8$residuals^2)
RSE <- sqrt(SSE/(26 - 5))
RSE
```

```
[1] 35.54858
```

Prediction/confidence intervals

- **Prediction intervals** combine both inaccuracies
  - Variability around the regression line
  - Precision of the regression line.
- Designed to contain, with 95% confidence, a random observation (e.g. CWD basal area or tree length) for given predictor values (e.g. tree density or given proportions of nitrogen, phosphorus, and potassium)
- **Confidence intervals** incorporate only the precision of the regression line.
- Designed to contain, with 95% confidence, the **average** of random observations for given predictor values.

Prediction intervals in R: CWD basal area

```
p <- predict(model3, newdata = data.frame(RIP.DENS=800:2200),
             interval = "confidence")
p[1:3,]  # print first 3 predictions
```

```
        fit        lwr      upr
1 0.9474953 -0.1563700 2.051361
2 0.9568141 -0.1433865 2.057015
3 0.9661229 -0.1304244 2.062670
```

```
p <- predict(model3, newdata = data.frame(RIP.DENS=800:2200),
             interval = "prediction")
p[1:3,]  # print first 3 predictions
```

```
        fit       lwr      upr
1 0.9474953 -1.497766 3.392757
2 0.9568141 -1.486795 3.400423
3 0.9661229 -1.475844 3.408090
```

Prediction intervals in R: Larches

```
newdata <- data.frame(nitrogen = 1.9, phosphor = 0.2,
                      potassium = 0.7)
newdata
```

```
  nitrogen phosphor potassium
1      1.9      0.2       0.7
```

```
predict.lm(model_l8, newdata, interval = "confidence")
```

```
       fit      lwr      upr
1 163.0865 140.6258 185.5472
```

```
predict.lm(model_l8, newdata, interval = "prediction")
```

```
       fit      lwr      upr
1 163.0865 85.82246 240.3505
```

Variability in regression models

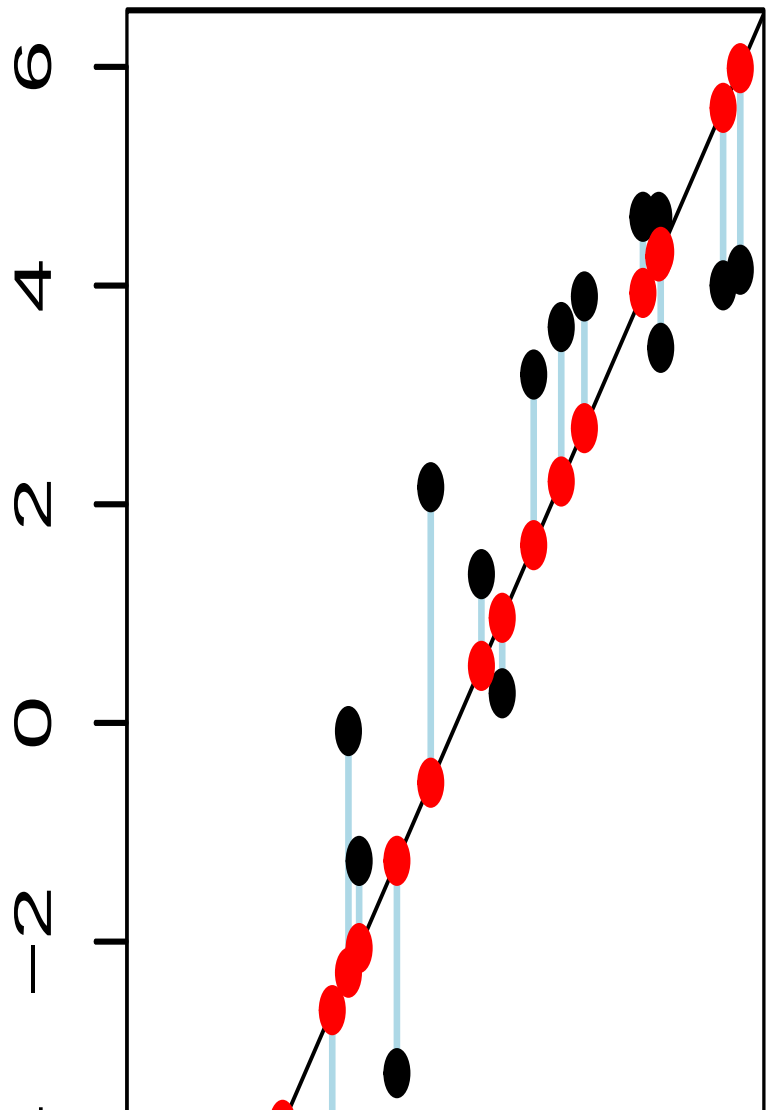
Predictivity

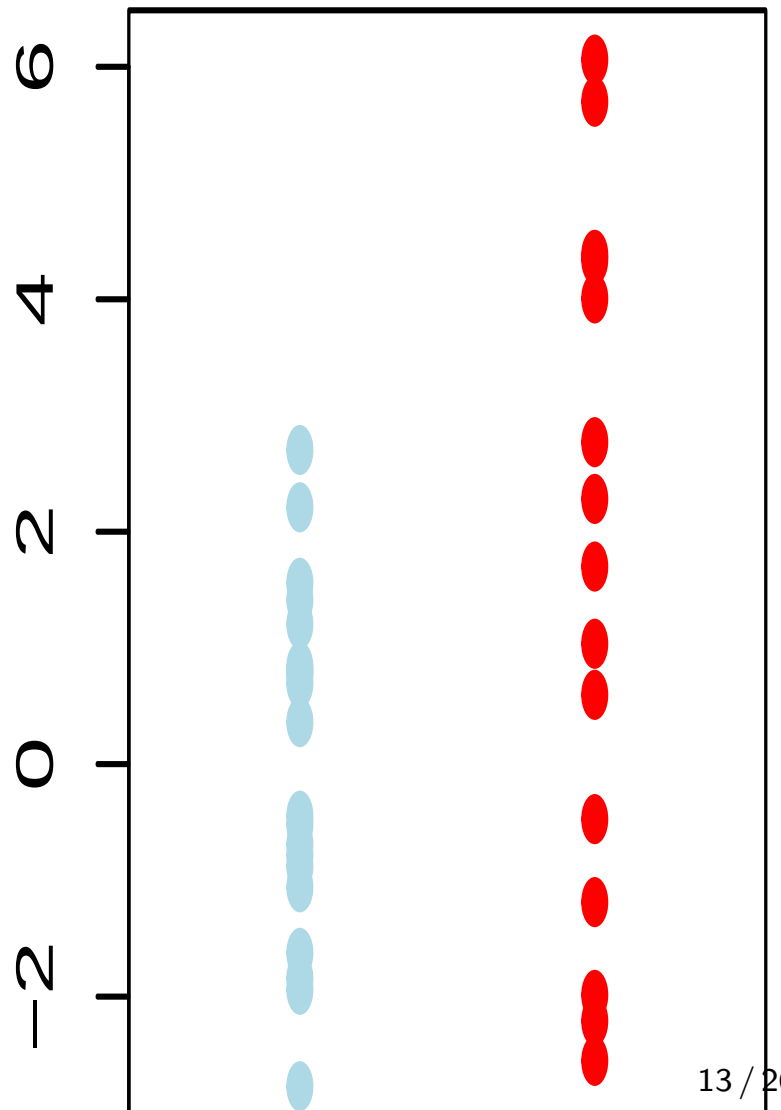
Another way to gain insight in predictivity compares

- variability **around** regression line
- with variability **on** the regression line, explained by the regression line.

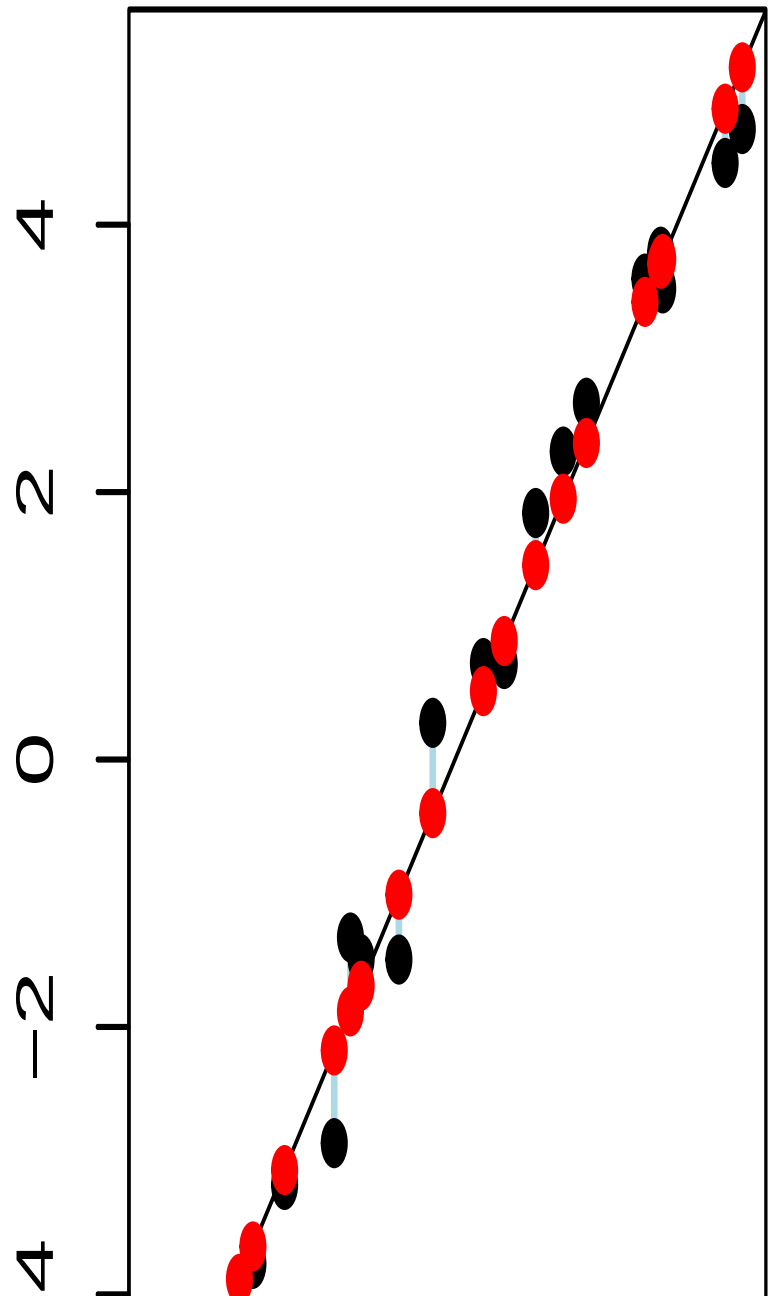
Total and residual variability

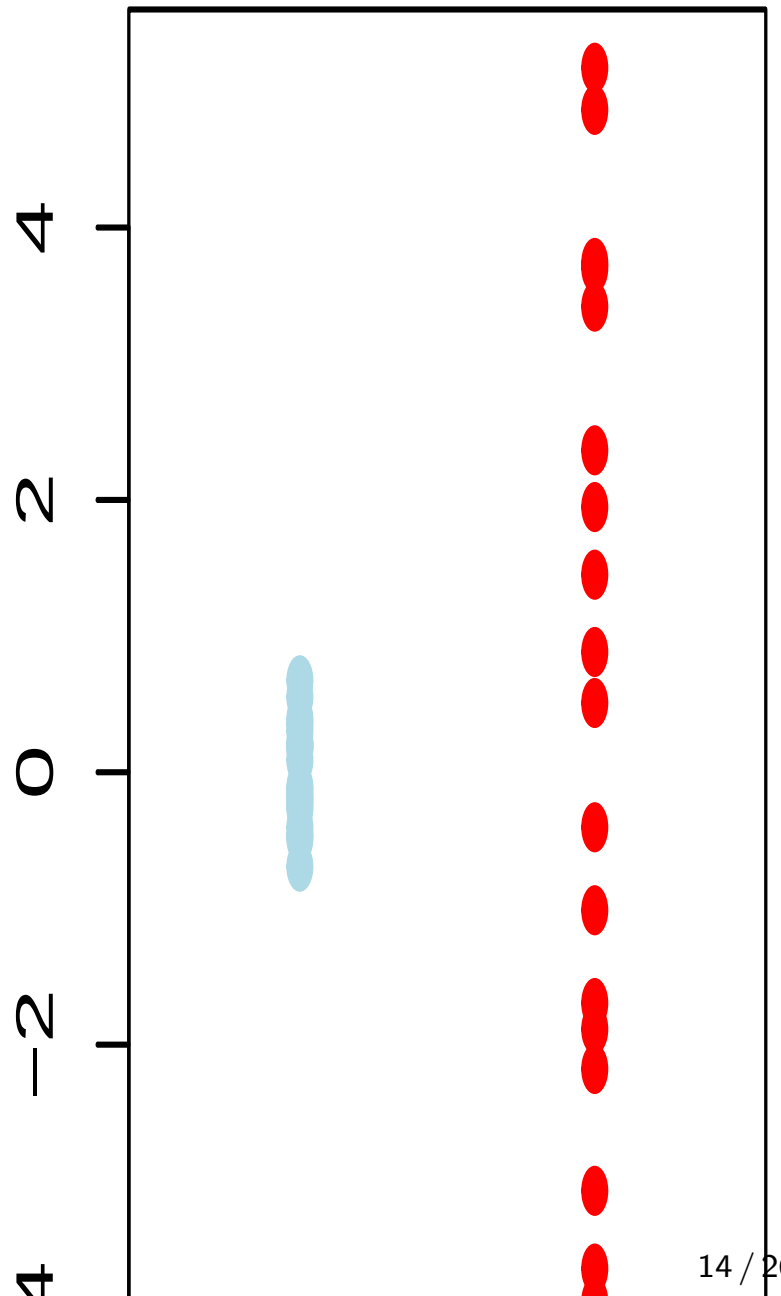
Idea: compare variability of residuals and variability of (centered) predictions.



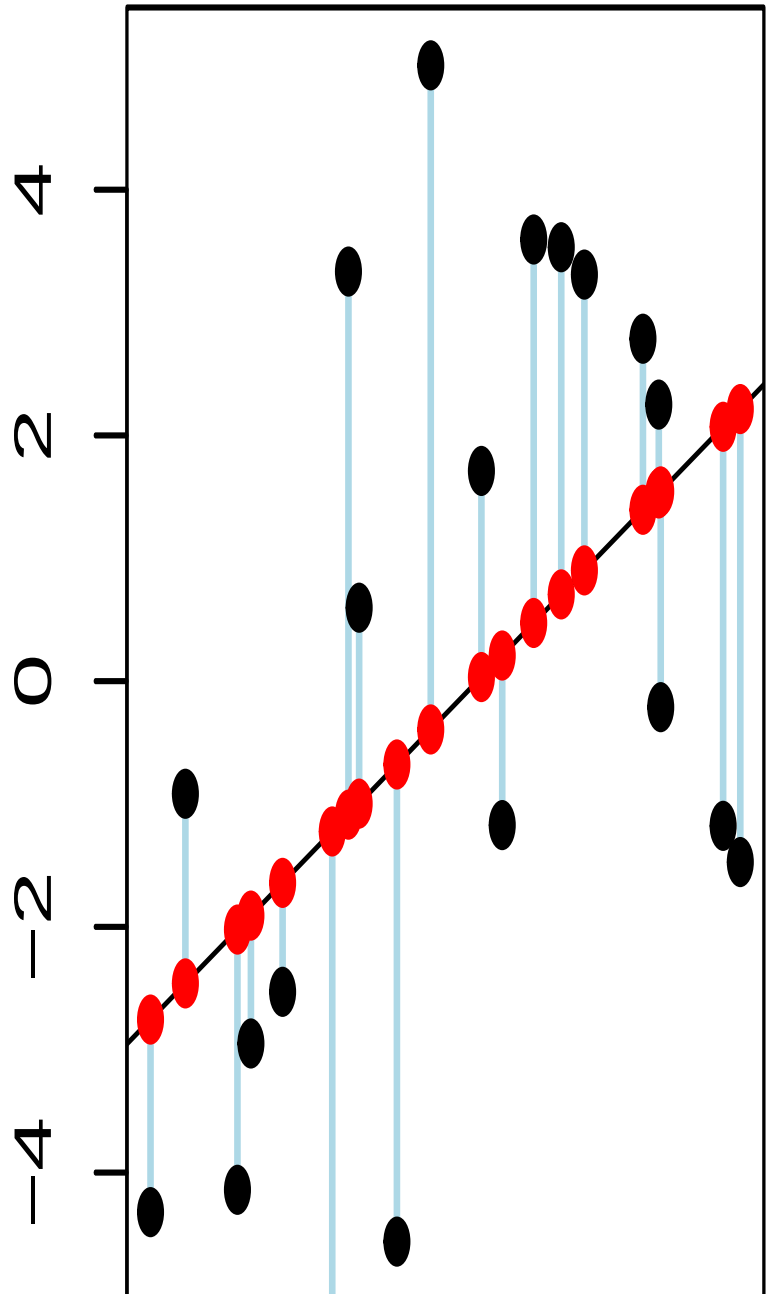


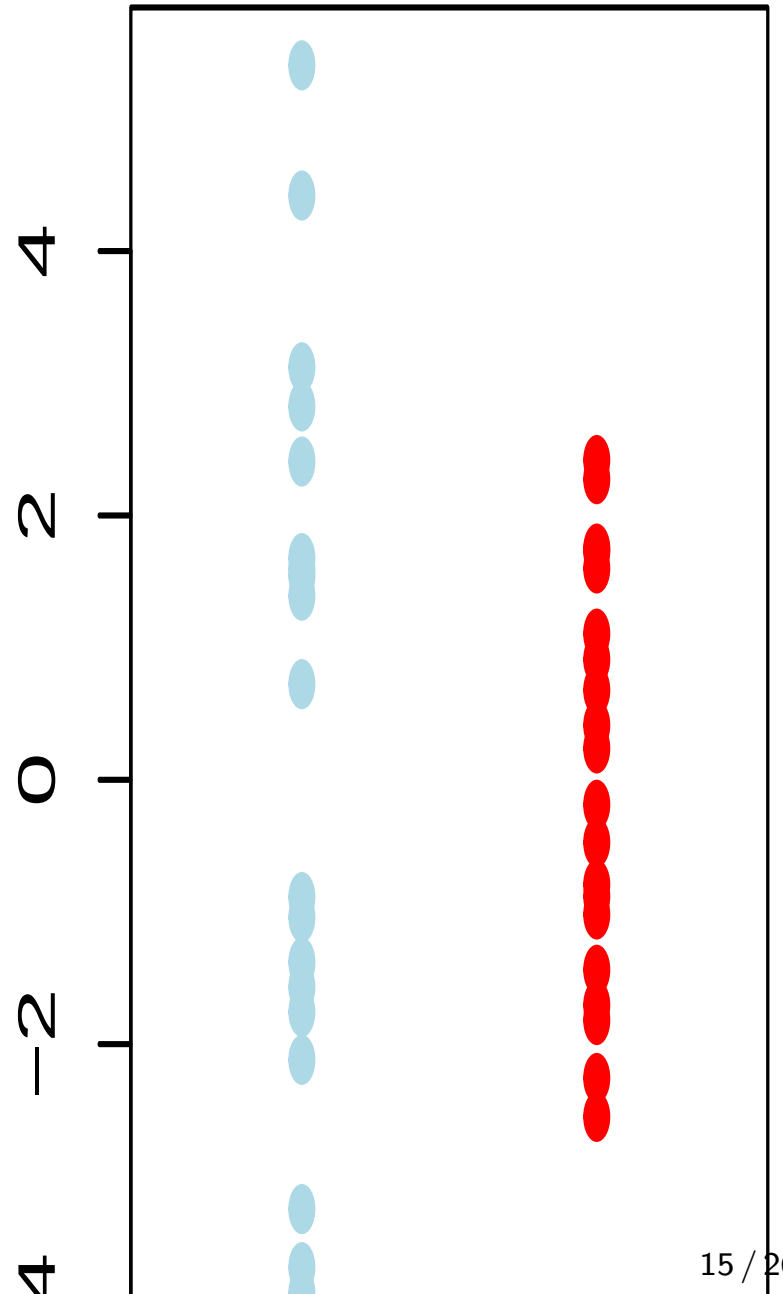
High predictivity: low variability around line





Low predictivity: small variability on line





Sum of squares

- Let $\hat{y}_i$ be the prediction for observation i, then

$$\begin{aligned} SST &= \sum_{i=1}^{n}(y_i - \bar{y})^2 \\ &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} e_i^2 \\ &= SSR + SSE. \end{aligned}$$

- Total sum of squares (SST) = Regression sum of squares (SSR) + Residual sum of squares (SSE).
- Total variability = Variability captured by regression + Variability in residuals.

Multiple correlation coefficient

- **Multiple correlation coefficient** or coefficient of determination:
$$R^2 = \frac{SSR}{SST}.$$
- Expresses the proportion of variability in data is captured by their association with explanatory variable.
- Measure for **predictive value** of explanatory variable.
- Always between 0 and 1.
- Simple linear regression: the square of the correlation between X and Y.

Multiple correlation coefficient

Look at the R `summary` output:

- CWD basal area:

```
Multiple R-squared:  0.7159,    Adjusted R-squared:  0.672
```

71.59% of variability on CWD basal area is explained by tree density.

- Larches:

```
Multiple R-squared:  0.8836,    Adjusted R-squared:  0.861
```

88.36% of variability on tree length is explained by mineral composition of needles.

Note: High R^2 only demanded for prediction, not to estimate effect of X on Y

Aside: adjusted multiple correlation coefficient

- R^2 always increases (gets closer to 1) when model becomes more complex
- To "punish" complexity, use adjusted R^2:

$$R^2_{\text{adj}} = 1 - \frac{n-1}{n-p}(1-R^2).$$

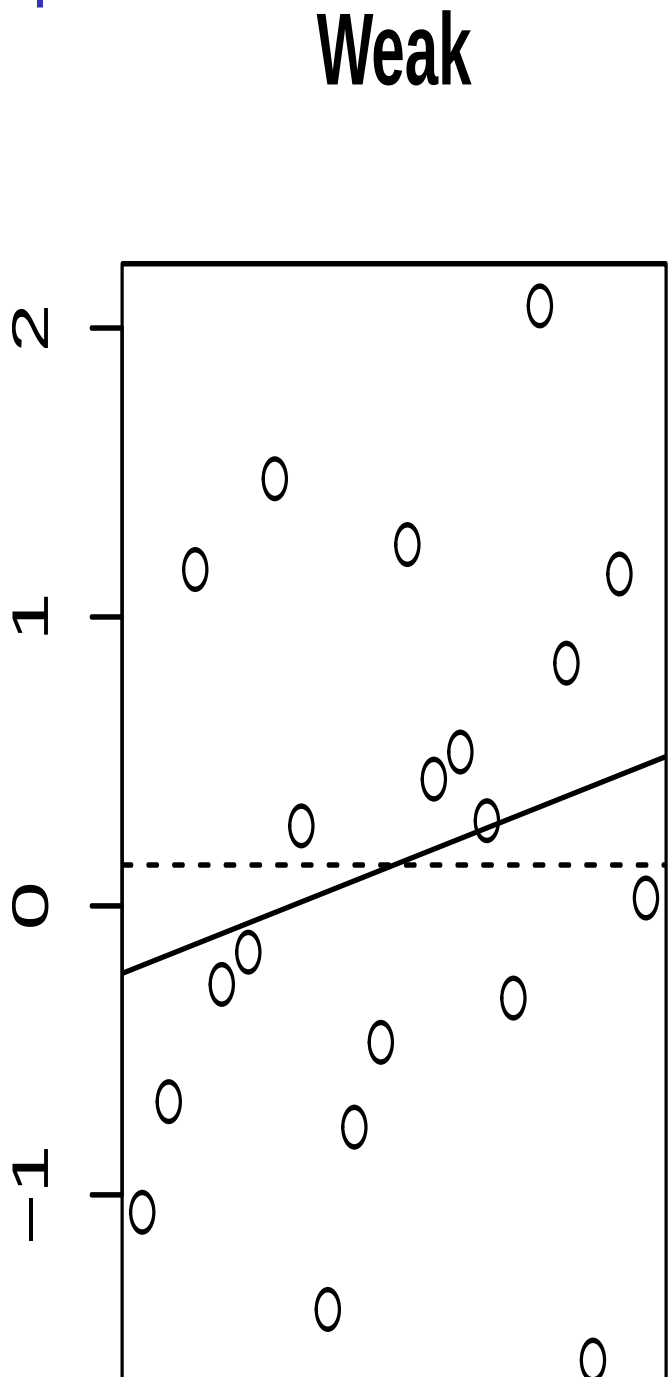
- Adjusted R^2 is always lower than R^2.
- Interpretation not so straightforward, used mainly for **model comparison**.

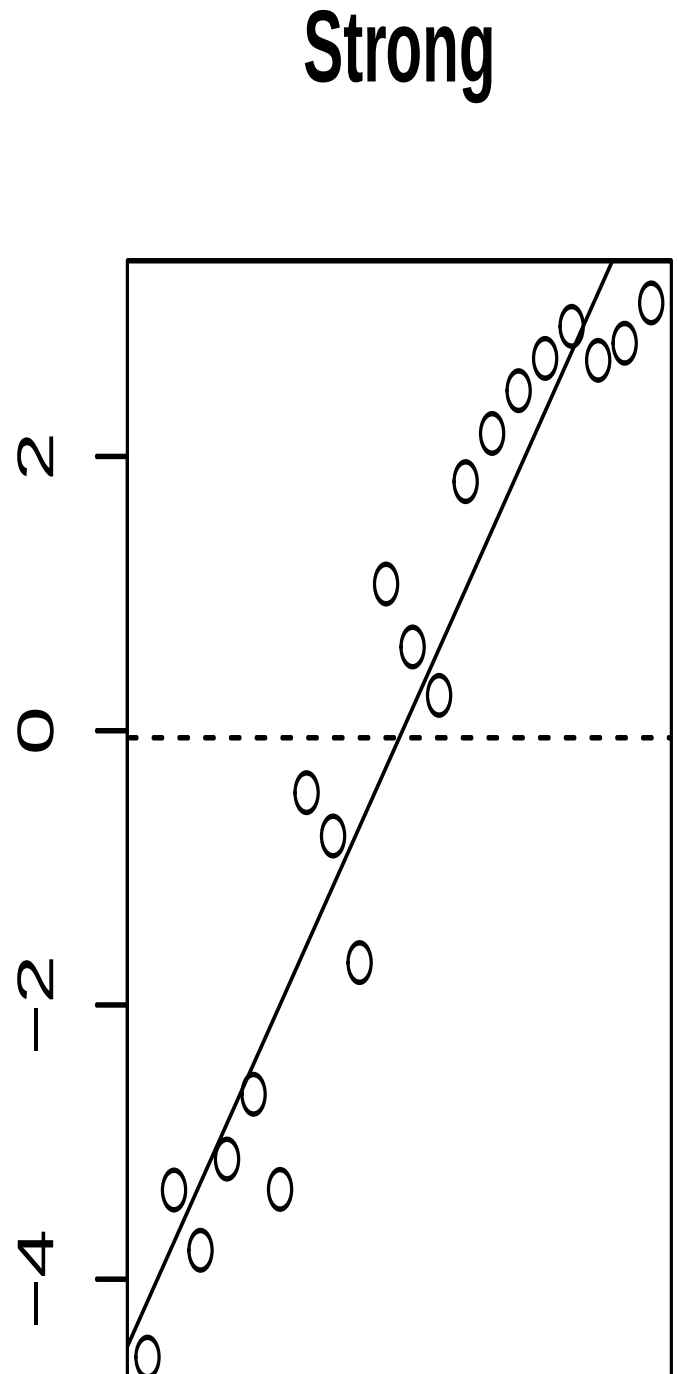
Larches: $n = 26$, $p = 5$, $R^2 = 0.8836$, so

$$R^2_{\text{adj}} = 1 - \frac{25}{21}(1-0.8836) = 0.8614.$$

Comparing simple vs. complex models

Example





Nested models

Nested models:

- Complex model: with many predictors.
- Simple model: like complex, but some predictors have been removed.

Example (larches):

- Complex: $E(Y|X) = \alpha + \beta_N X_N + \beta_P X_P + \beta_K X_K + \beta_r X_r$
- Simple: $E(Y|X) = \alpha + \beta_P X_P$

How do we quantify which model is better?

- Single regression: hypothesis test for β.
- Multiple regression: need to compare effect of **all coefficients at once**.

Intuition: comparing variance

Idea: compare residual variability (SSE) to assess model fit.

- SSE_{complex} *always* lower than SSE_{simple}.
- If it is *much* lower, decide that complex model is better.

Formalized via F-test:

- Null hypothesis: simple and complex model fit data equally well.
- Alternative hypothesis: complex model is better.
- Test statistic:

$$f = \frac{\frac{SSE_{\text{simple}} - SSE_{\text{complex}}}{p_{\text{complex}} - p_{\text{simple}}}}{\frac{SSE_{\text{complex}}}{n - p_{\text{complex}}}} \sim F_{p_{\text{complex}} - p_{\text{simple}}, n - p_{\text{complex}}}.$$

Example: larches

Residual sum of squares:

- $SSE_{\text{simple}} = 91404.49$
- $SSE_{\text{complex}} = 30121.92$

Number of parameters:

- $p_{\text{simple}} = 2$
- $p_{\text{complex}} = 5$

Hypothesis test:

- Test statistic: $f = 14.24139$
- p-value: $p = 0.00002744.$

Conclusion: complex model is significantly better.

Example in R: larches

```
model_l1 <- lm(length ~ phosphor)
model_l2 <- lm(length ~ nitrogen + phosphor + potassium + residu)
anova(model_l1, model_l2)
```

```
Analysis of Variance Table

Model 1: length ~ phosphor
Model 2: length ~ nitrogen + phosphor + potassium + residu
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1     24 91404
2     21 30122  3     61283 14.241 2.744e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

R summary command

```
summary(model_l8)
```

```
(...)
Residual standard error: 35.55 on 21 degrees of freedom
Multiple R-squared:  0.8836,    Adjusted R-squared:  0.8614
F-statistic: 39.85 on 4 and 21 DF,  p-value: 1.603e-09
```

Last line:

- F-statistic: compares model to model with intercept only.
- **"Is my complex model capturing something meaningful?"**