Variance explained and partitioning

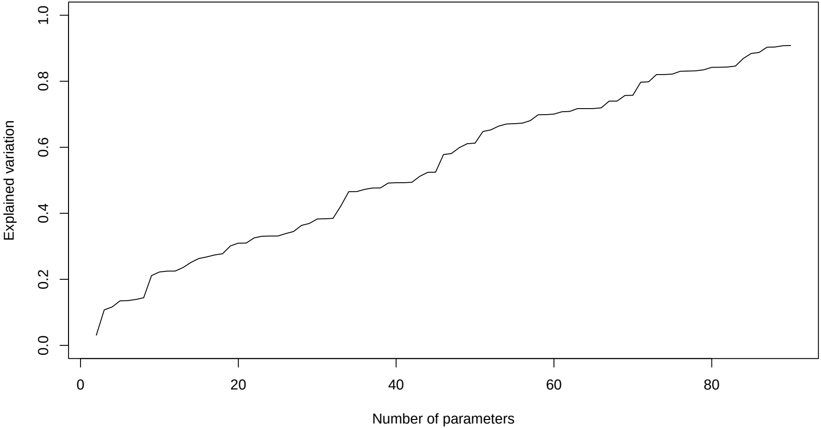
Bert van der Veen

Department of Mathematical Sciences, NTNU

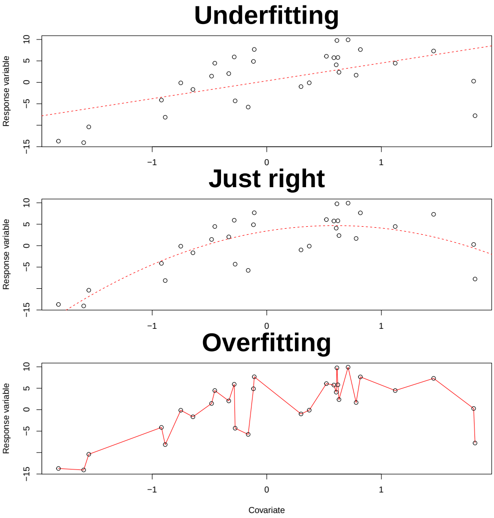
Recap

The problem of model complexity

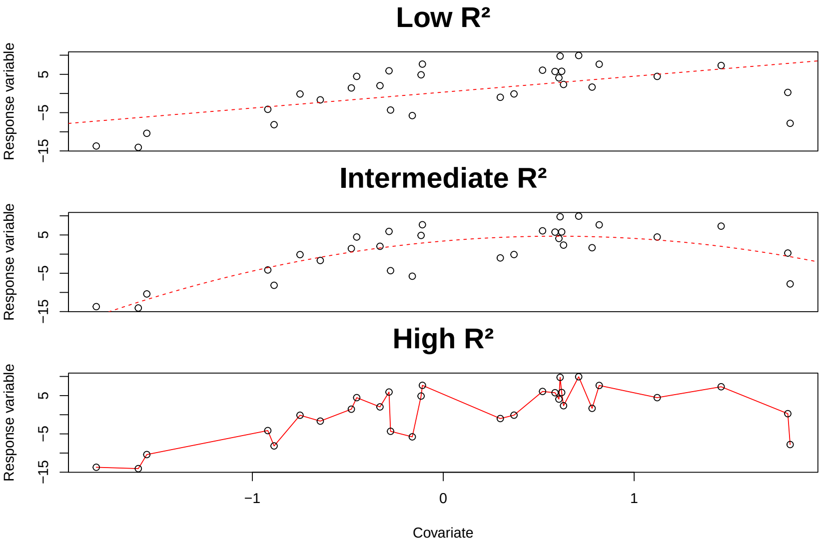
A model with always fit better if you add a parameter



Model complexity



Explained variation



R^2: coefficient of determination

The R^2 statistic helps us to assess how much variability we have explained.

Low R^2: not so much?
High R^2: too much?

- Usually larger than 0 but smaller than 1
- Observational studies usually have low R^2
- Experimental studies usually have high R^2

R^2 **cannot be used to assess goodness of fit**

R^2 in linear regression

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \mu_i)^2}{\sum_{i=1}^{n} (y_i - \alpha_i)^2} \tag{1}$$

- the numerator represents the unexplained variation of a model
- the numerator represents the explainable variation of a model
- One minus their fraction is the explained variation

partial R^2

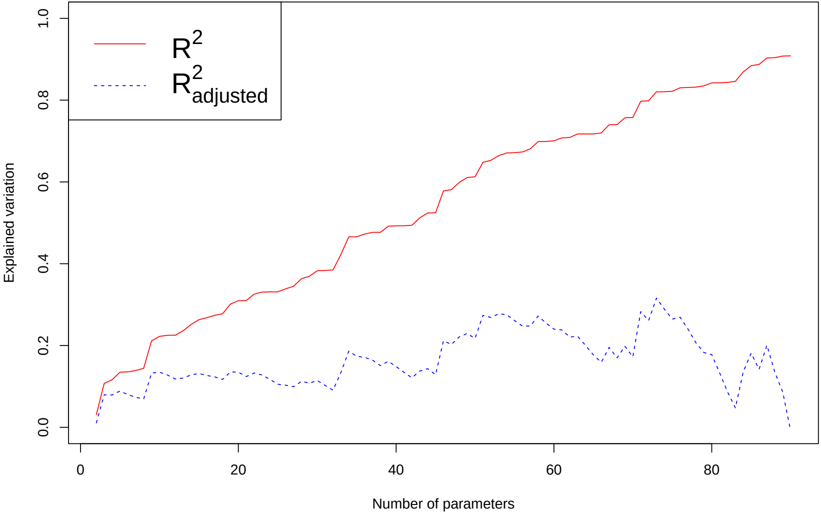
How much variation is explained by a covariate?

$$R_k^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \alpha_i)^2 - \sum_{i=1}^{n}(y_i - \mu_i)^2}{\sum_{i=1}^{n}(y_i - \alpha_i)^2} \tag{2}$$

Allows us to partition the explained variation over the covariates.

- Quantifies the variation additionally explained in a more complex models
- Is sensitive to order of covariates
- Multicollinearity: two covariates can explain similar variation

Adjusted R^2


R2
R2adjusted
Explained variation
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
Number of parameters

Adjusted R^2

$$R^2_{adjusted} = 1 - (1 - R^2)\frac{(n-1)}{n-p-1} \quad (3)$$

- Can be negative, always lower than R^2
- Penalize the statistic for increased complexity
- Penalize for small samples
- Most for small n and large p

Connection to Variance Inflation Factor (VIF)

$$\mathsf{VIF}_k = \frac{1}{1 - R_k^2} \tag{4}$$

Measures increase in uncertainty of a parameter estimator when changing the model.

Psuedo R^2 in Generalised Linear Models

1) GLMs lack the error term of LMs
2) There is often no (clear) residual variance parameter
3) R^2 is thus not clearly defined
4) Many many different R^2 statistics exist

Concepts and name remain the same.

A few GLM R^2's

- Deviance R^2
- Pseudo R^2
- McFadden's R^2
- Cohen's R^2
- Tjur's R^2
- Somer's R^2
- Cox and snell R^2
- Nagelkerke R^2
- Efron's R^2
- Many more (see `performance::R2`)
- which one do we choose?

Deviance R^2

Remember: **deviance** is the GLM equivalent of RSS!

$$\text{deviance } R^2 = 1 - \frac{D(\mathbf{y};\hat{\boldsymbol{\mu}})}{D(\mathbf{y};\hat{\alpha})} \qquad (5)$$
$$\text{Cohen } R^2 = \frac{D(\mathbf{y};\hat{\alpha}) - D(\mathbf{y};\hat{\boldsymbol{\mu}})}{D(\mathbf{y};\hat{\alpha})}$$

$$\text{deviance } R^2_{adjusted} = 1 - \left\{1 - \frac{D(\mathbf{y};\hat{\boldsymbol{\mu}})}{D(\mathbf{y};\hat{\alpha})}\right\}\frac{n-1}{n-p} \qquad (6)$$

Based on reduction in deviance

McFadden R^2

$$\text{McFadden } R^2 = 1 - \frac{\log \mathcal{L}(\mathbf{y}; \Theta_1)}{\log \mathcal{L}(\mathbf{y}; \Theta_1)} \tag{7}$$

Based on reduction in the ratio of log-likelihoods.

GLM R^2's

All quantify change in fit due to increased model complexity, in one way or another.

Which one do we use?

Example: Lizards interaction

First model without interaction of Time and Site

```
model1 <- glm(cbind(grahami, opalinus)~Time+Site,
              data = lizards, family="binomial")
```

Second model with interaction

```
model2 <- glm(cbind(grahami, opalinus)~Time*Site,
              data = lizards, family="binomial")
```

Lizards: R^2 under more complex model

```
nullmodel <- update(model1, formula = .~1)
(devianceR2 <- c(1-deviance(model1)/deviance(nullmodel),1-deviance(model2)
```

```
## [1] 0.2168604 0.2325368
```

```
(adjdevianceR2 <- 1-(1-devianceR2)*(n-1)/(n-c(attr(logLik(model1), "df"),a
```

```
## [1] 0.093206835 0.006812369
```

```
(mcfaddenR2 <- c(1-logLik(model1)/logLik(nullmodel), 1-logLik(model2)/logL
```

```
## [1] 0.1197730 0.1284312
```

Variability of R^2

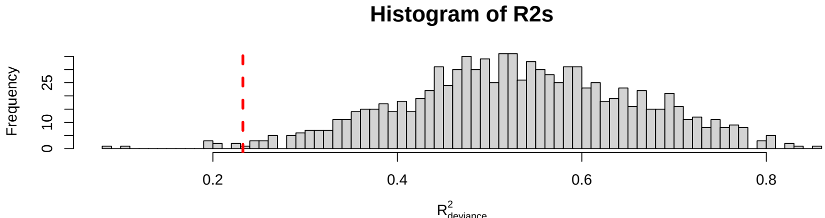
R^2 is a statistic of the data
So it is susceptible to sampling variation!

Veall and Zimmermann (1992)

- Evaluate which R^2 best mimics the original statistic
- Conclude that various popular R^2 severely underestimate the truth

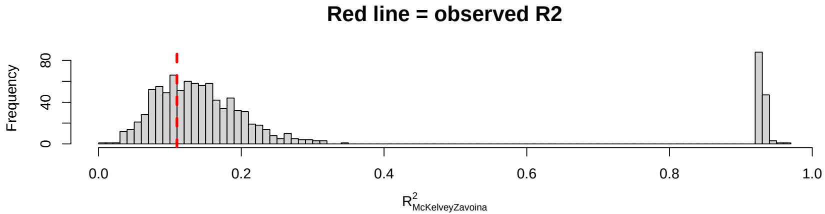
Lizards: deviance R^2

```
R2s <- NULL
for(i in 1:1000){
set.seed(i)
ynew <- as.matrix(stats::simulate(model2))
model2s <- update(model2, formula = ynew~.)
nullmodels <- update(model2s, formula = .~1)
R2s <- c(R2s, 1-deviance(model2s)/deviance(nullmodels))
}
```



Lizards: McKelvey-Zavoina R^2

```
R2s2 <- NULL
for(i in 1:1000){
set.seed(i)
ynew <- as.matrix(stats::simulate(model2))
model2s <- update(model2, formula = ynew~.)
nullmodels <- update(model2s, formula = .~1)
R2s2 <- c(R2s2, DescTools::PseudoR2(model2s, "McKelveyZavoina"))
}
```



Conclusion

- R^2 to quantify improvement of fit in GLMs
- There is no single agreed upon statistic
- Careful with overinterpretation: low R^2 can be just fine