Lesson: Motivating model selection

Module: Model Selection

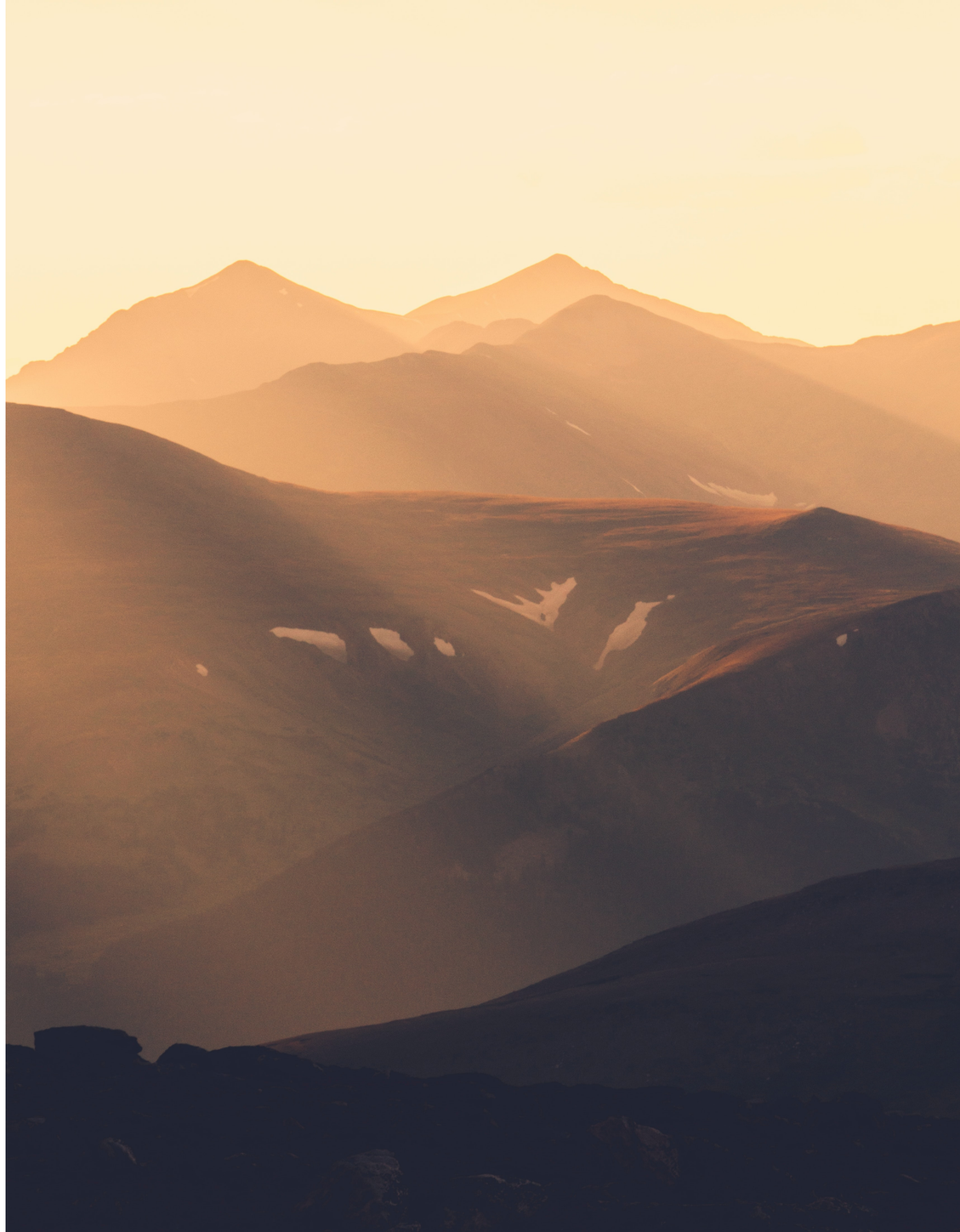


Photo by Calvin Weibel on Unsplash

Photo by Bank Phrom on Unsplash

Two different extremes:

1. Models that underfit
2. Models that overfit

Photo by Bank Phrom on Unsplash

Can R^2 help us?

Photo by Bank Phrom on Unsplash

In simple cases, t-tests and F-tests can help us choose a model. But what about more complicated cases?

Example: Consider a model with 100 potential predictors. The goal is to find the best model with some subset of predictors. The possible combination of predictors is *huge!*

Photo by Bank Phrom on Unsplash

Two options for model selection:

1. Hypothesis testing methods: backward elimination, forward selection, and stepwise regression.

2. Criterion-based methods: AIC, BIC, adjusted R^2.

Lesson: Testing based procedures and their shortfalls

Module: Model Selection

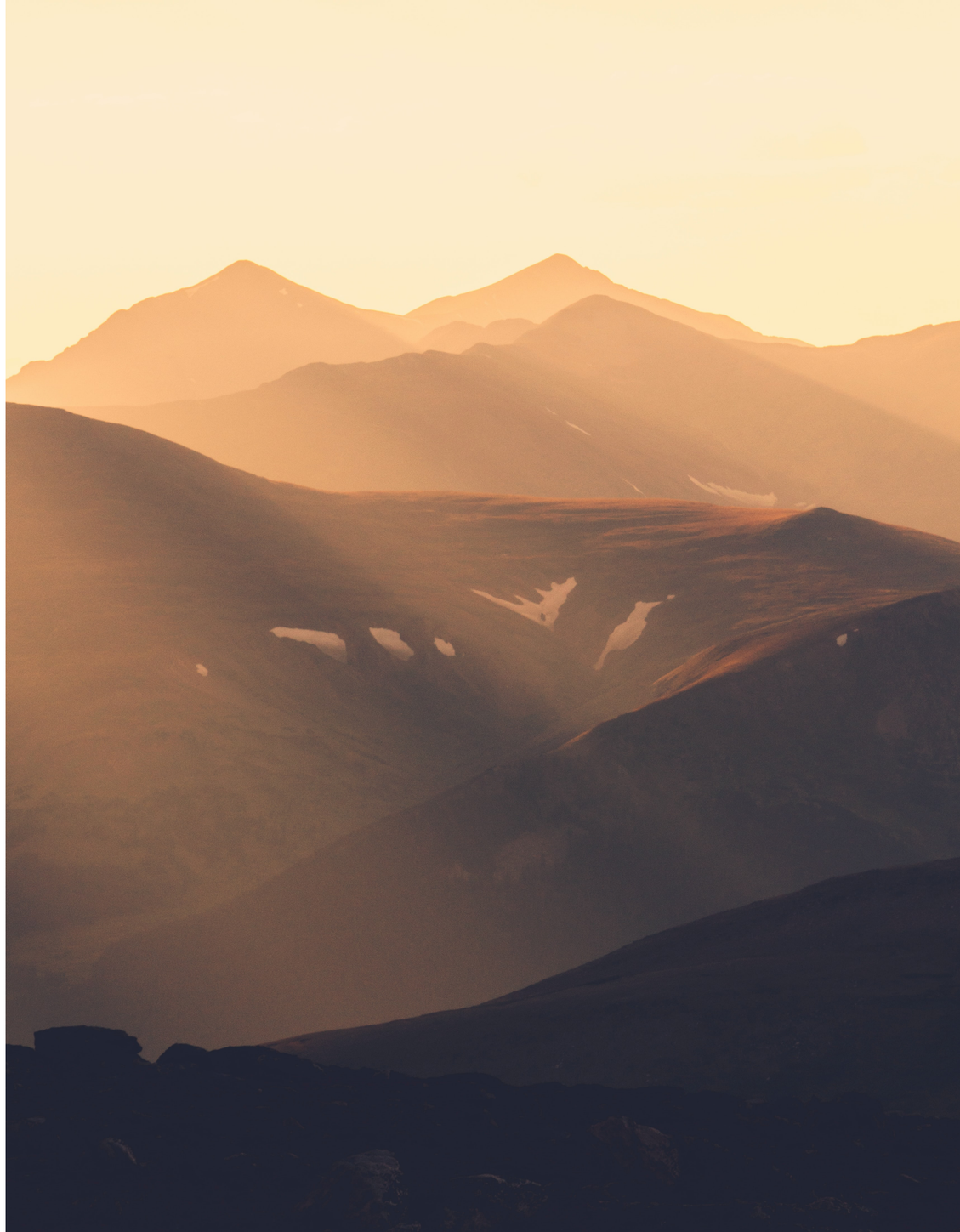


Photo by Calvin Weibel on Unsplash

Photo by Bank Phrom on Unsplash

Backward Elimination.

The backward elimination procedure is as follows.

1. Start with the full model, and remove the predictor with highest p-value greater than α_0.

2. Refit the model with the remaining predictors and repeat until all predictors are significant.

Photo by Bank Phrom on Unsplash

Forward Selection.

The forward selection procedure is as follows:

1. Start with all possible simple linear regression models. Pick the one that is best (e.g., smallest F-test p-value).

2. Add a second predictor and repeat.

Photo by Bank Phrom on Unsplash

Bi-directional Stepwise Regression.

In stepwise regression, at each stage we allow for any variable to be added or removed. There are different variations on how this may be done.

Photo by Bank Phrom on Unsplash

Bi-directional Stepwise Regression.

In stepwise regression, at each stage we allow for any variable to be added or removed. There are different variations on how this may be done.

Photo by Bank Phrom on Unsplash

In R, we can use the `update()` function to carry out these procedures.

1. Fit the full model as `lmod = lm()`.

2. Run `mod = update(lmod, .~. -InSigVariable)`. This will refit the model without the insignificant variable.

Photo by Bank Phrom on Unsplash

We'll walk through these procedures in R in more detail in a future lesson.

Photo by Bank Phrom on Unsplash

Testing-based procedures are cheap and easy, but there are some disadvantages:

1. Final models are not necessarily “optimal” in any mathematical or statistical sense.

2. Selection methods overstate of the importance of the remaining predictors.

3. Selection methods aren’t directly related to prediction or explanation!

4. Models are sometimes smaller than optimal for prediction.

Lesson: Criterion-based selection procedures: AIC

Module: Model Selection

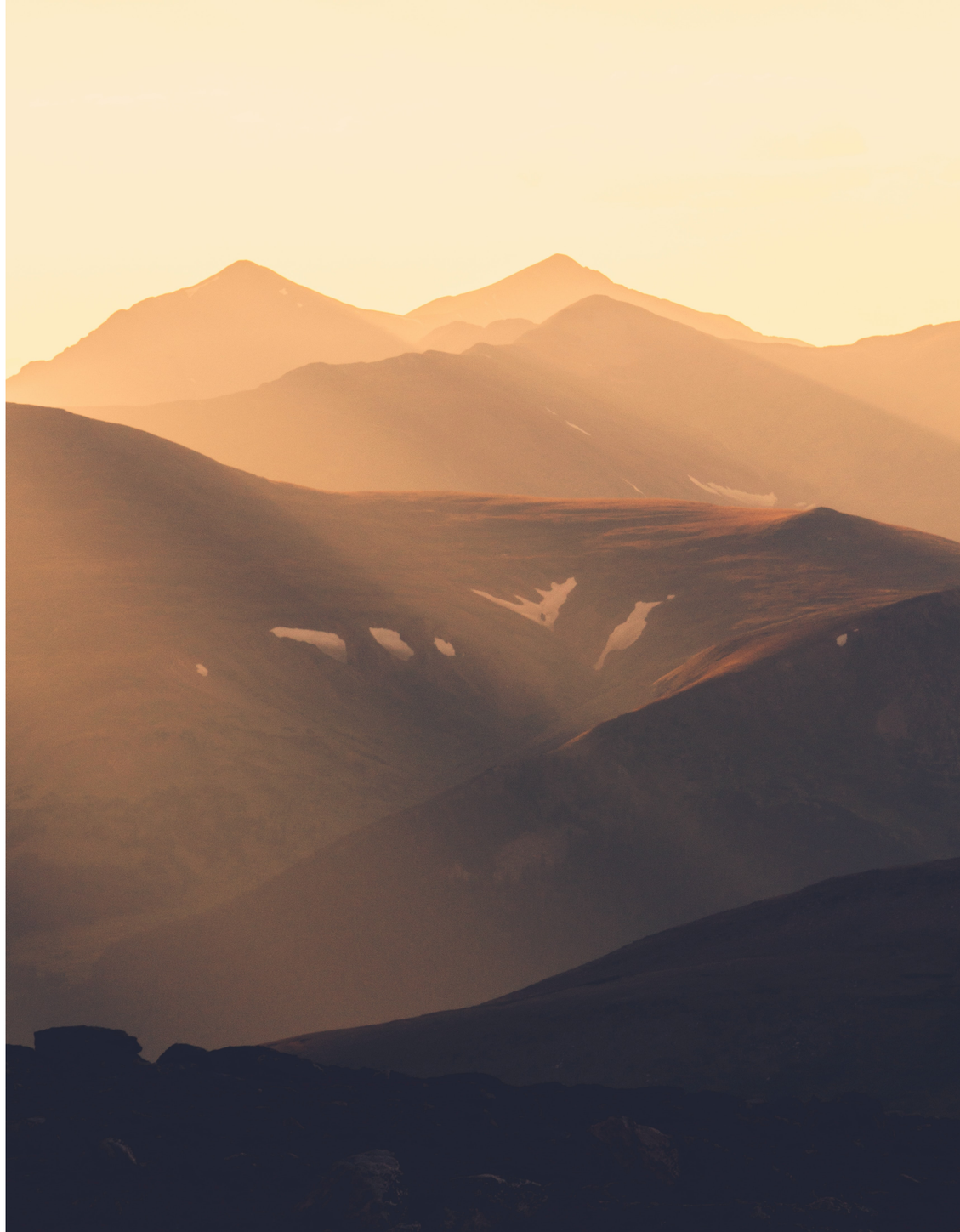


Photo by Calvin Weibel on Unsplash

Photo by Bank Phrom on Unsplash

Sometimes, we can compare models according to a given criterion.

Some possibilities:

1. Akaike Information Criterion (AIC).

2. Bayes Information Criterion (BIC).

3. Adjusted R^2.

Photo by Bank Phrom on Unsplash

$$
\begin{tikzcd}
g_7(\mathbf{x}) & & g_6(\mathbf{x}) \\
& \boxed{f(\mathbf{x})} \arrow[ul] \arrow[ur] \arrow[dl] \arrow[dr] \arrow[dll] \arrow[drr] \arrow[dll] & \\
g_1(\mathbf{x}) & & g_3(\mathbf{x}^\dagger) \\
g_5(\mathbf{x}^\dagger) & & g_4(\mathbf{x}^*) \\
& g_2(\mathbf{x}^*) &
\end{tikzcd}
$$

Photo by Bank Phrom on Unsplash

Select a model g, parametrized by β, that is ***close*** to the true model f. ***Close*** is measured by the Kullback-Leibler distance:

$$D_{KL}(f, g)$$

$$= \int f(\mathbf{x}) \log\left(\frac{f(\mathbf{x})}{g(\mathbf{x}; \beta)}\right) d\mathbf{x}$$

Photo by Bank Phrom on Unsplash

$$D_{KL}(f, g)$$

$$= \int f(\mathbf{x}) \log(f(\mathbf{x})) d\mathbf{x} - \int f(\mathbf{x}) \log\left(g\left(\mathbf{x}; \hat{\boldsymbol{\beta}}\right)\right) d\mathbf{x}$$

constant wrt g

Photo by Bank Phrom on Unsplash

For linear regression, the AIC reduces to:

Photo by Bank Phrom on Unsplash

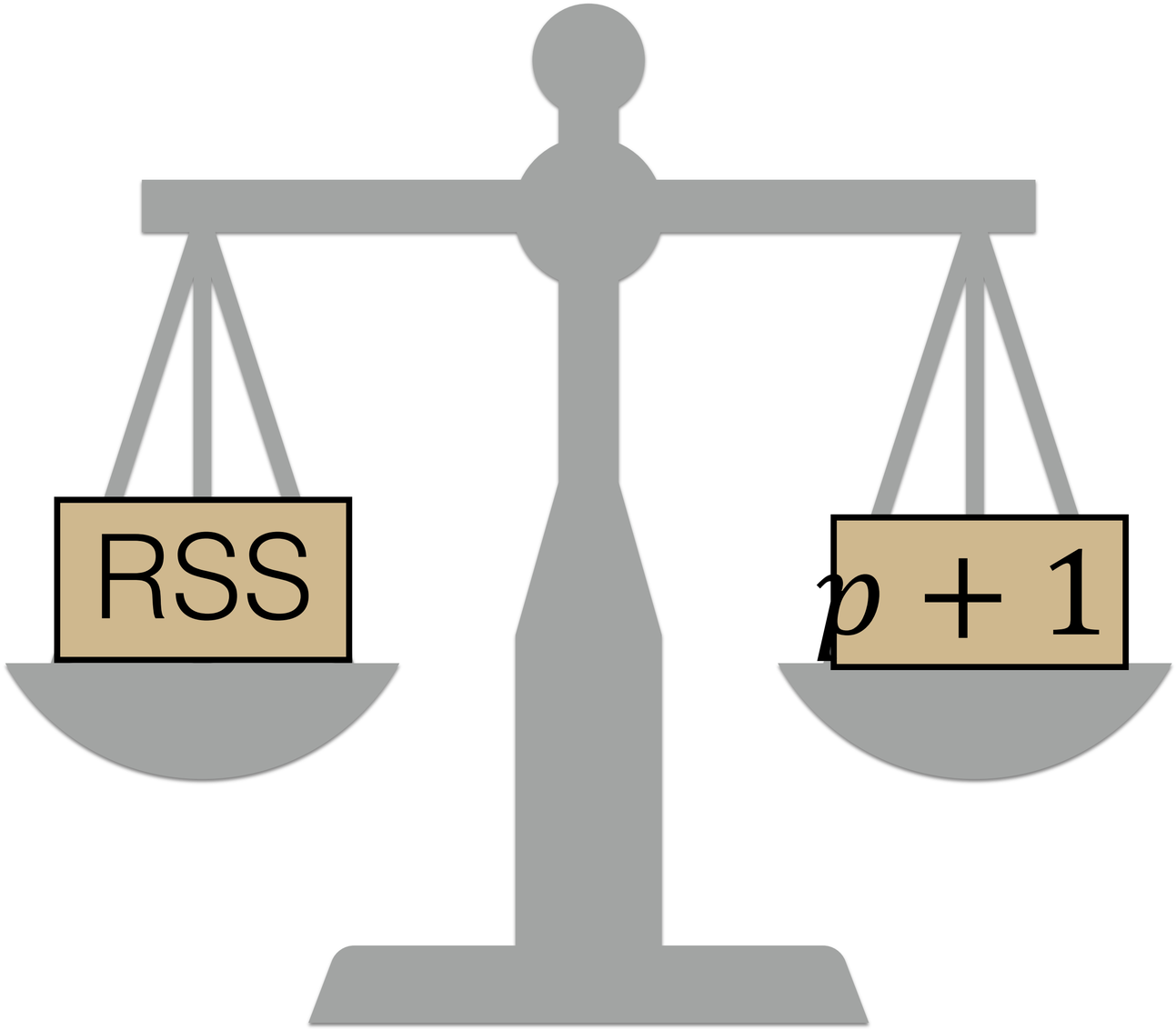
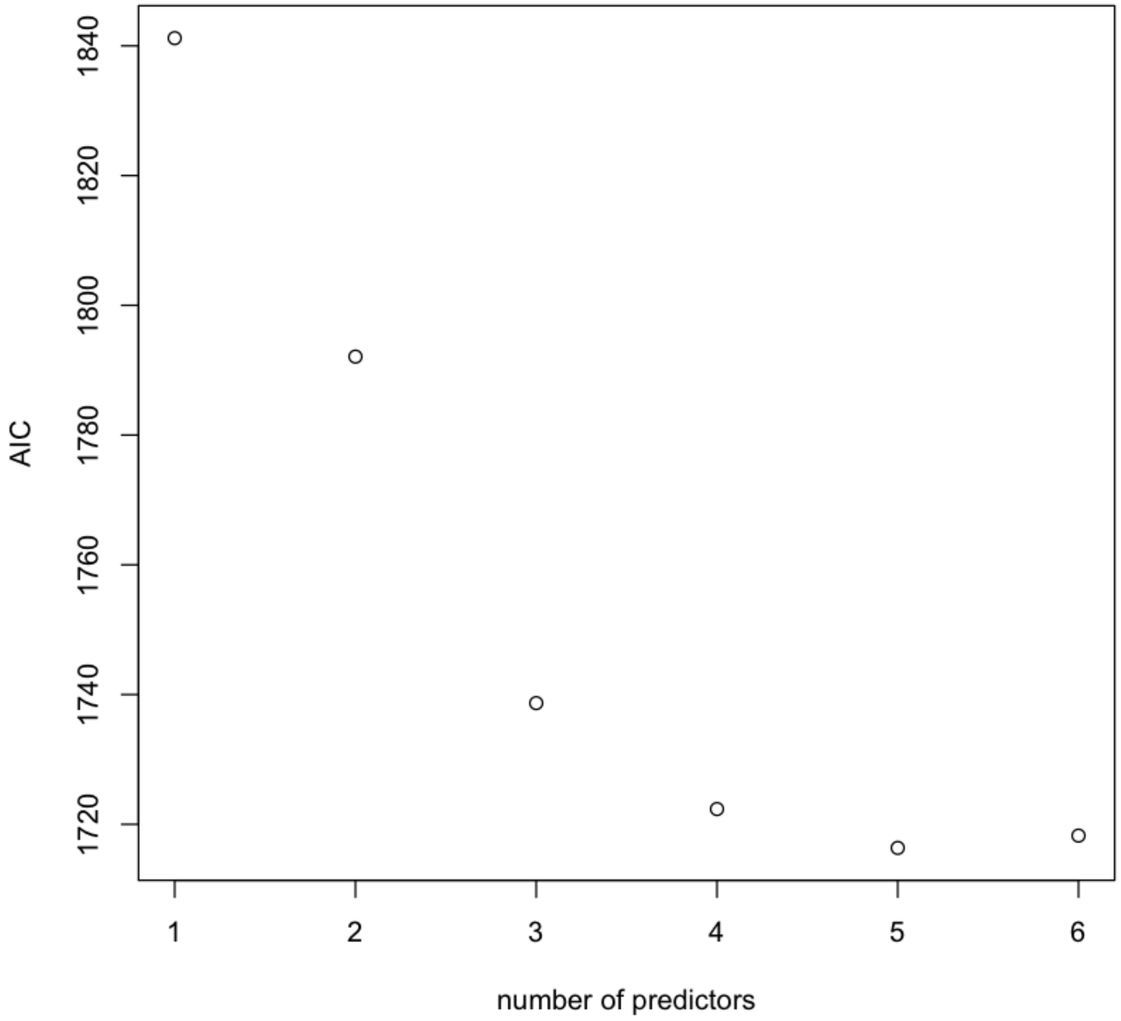


Photo by Bank Phrom on Unsplash



Criterion-Based Procedures

One natural choice would be to pick a fitted model *m*, parameterized by ___, that is close to the true model *M*. We could measure distance by:

This is know n as the Kullback-Leibler information/distance. But, we don't know *M*. We can substitute in the MLE of ____ and rearrange to obtain:

The first term is constant and doesn't depend *m*. Akaike showed that _____ can be estimated by:

Criterion-Based Procedures

```
b = regsubsets(Response ~ ., data = df)
rs = summary(b)
rs$which
```

Produces:

(Intercept)	syct	mmin	mmax	cach	chmin	chmax
TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE
TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE
TRUE	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE
TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE
TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE

Lesson: Criterion-based selection procedures: BIC

Module: Model Selection

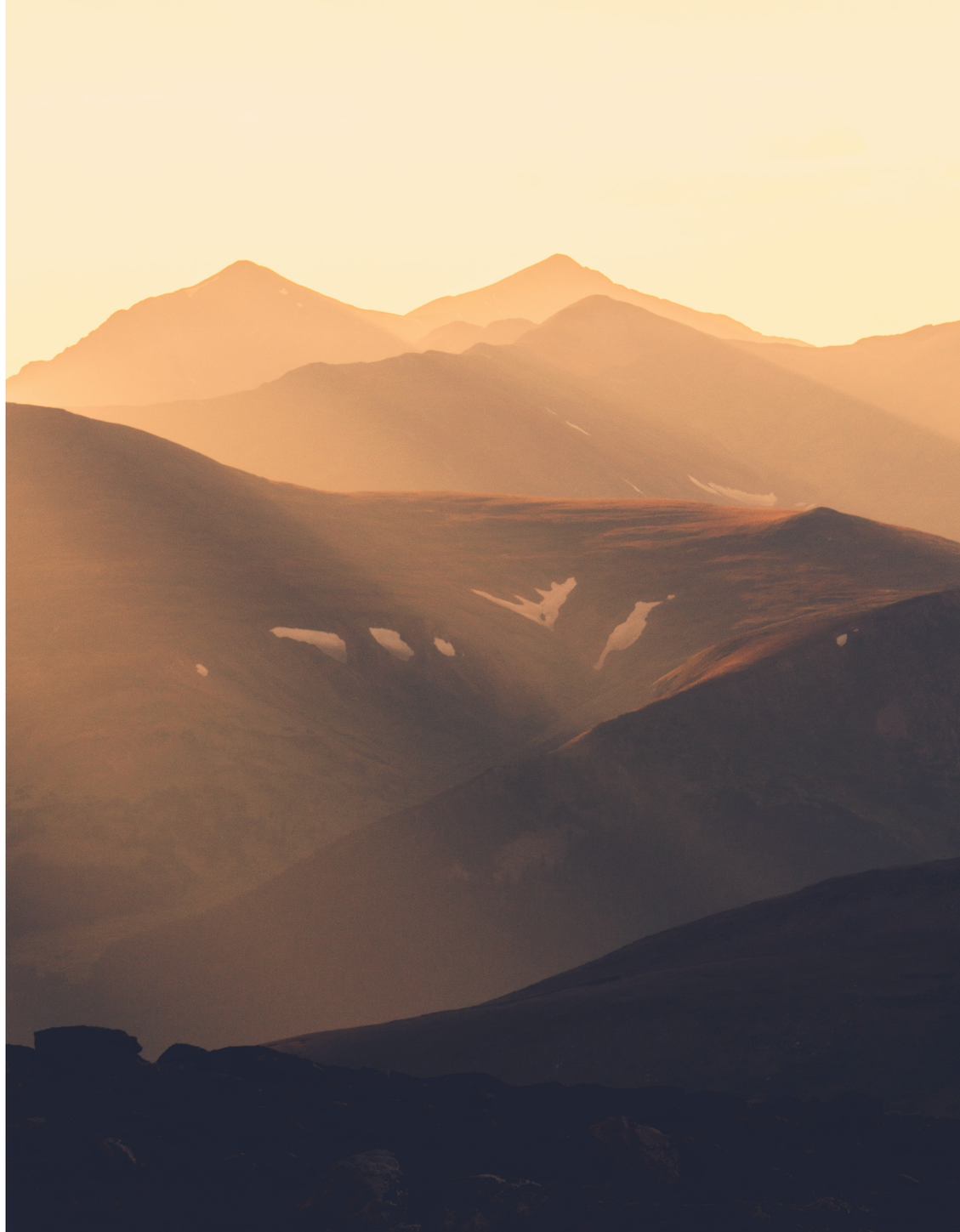


Photo by Calvin Weibel on Unsplash

Photo by Bank Phrom on Unsplash

$$BIC\left(g\left(\mathbf{x};\hat{\boldsymbol{\beta}}\right)\right) = (p+1)\log(n) - 2\log L\left(\hat{\boldsymbol{\beta}}\right)$$

Photo by Bank Phrom on Unsplash

It is often said that:

1. AIC is a better metric when prediction is the goal of the statistical model.

2. BIC is a better metric when explanation is the goal of the statistical model.

But things are a bit more complicated than that!

Lesson: Criterion-based selection procedures: Adjusted R^2

Module: Model Selection

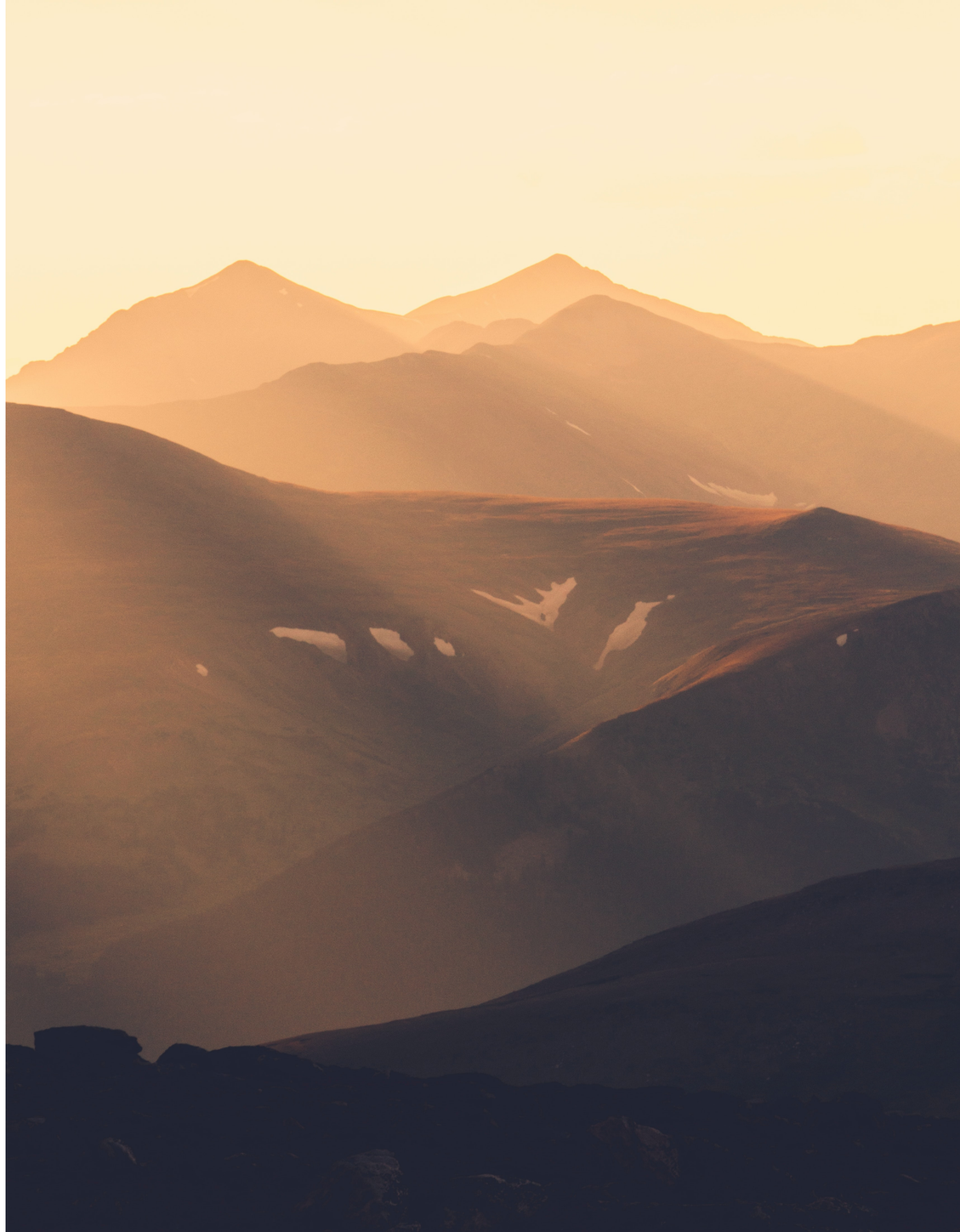


Photo by Calvin Weibel on Unsplash

Photo by Bank Phrom on Unsplash

Consider two models:

1. A "full" model with two predictors:
   $$Y_i = \beta + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i.$$
2. A "reduced" model with only one predictor from the full model: $Y_i = \beta + \beta_1 x_{i,1} + \varepsilon_i$.

Photo by Bank Phrom on Unsplash

$$RSS_{red} = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_{i,1})^2$$

vs

$$RSS_{full} = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_{i,1} - \beta_2 x_{i,2})^2$$

Photo by Bank Phrom on Unsplash

$$RSS_{red} = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_{i,1})^2$$

$$\|$$

$$RSS_{full} = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i,1} - 0 x_{i,2})^2$$

Photo by Bank Phrom on Unsplash

Recall that: $R^2 = 1 - \frac{RSS}{TSS}$.

R^2 should not be used to compare models of different sizes!

Photo by Bank Phrom on Unsplash

The adjusted R^2 is defined as:

$$R_a^2 = 1 - \frac{RSS/(n-(p+1))}{TSS/(n-1)}$$

Lesson: The mean squared prediction error for model selection

Module: Model Selection

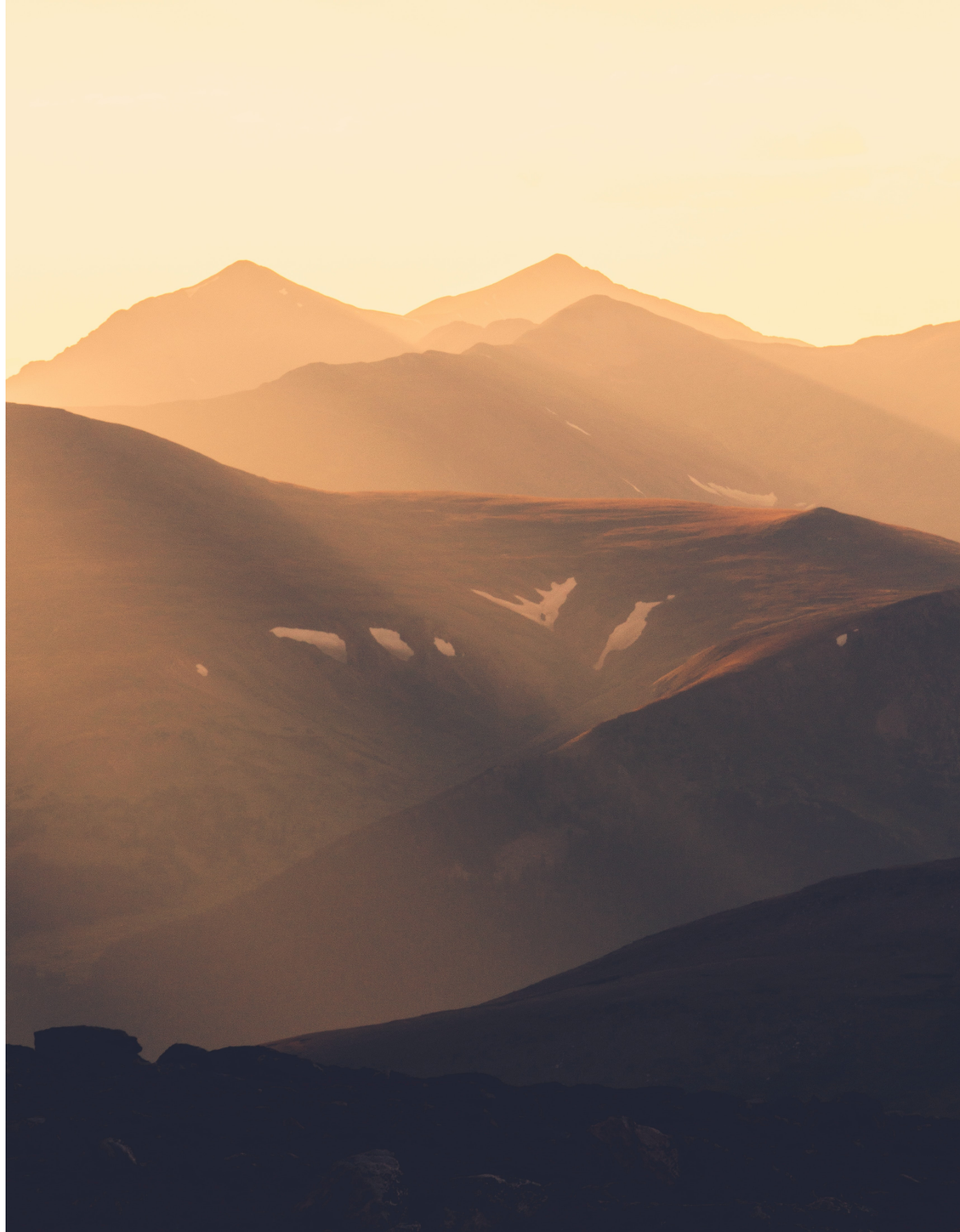


Photo by Calvin Weibel on Unsplash

$$\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}}$$

$$\hat{\mathbf{y}}^{\star} = X^{\star}\hat{\boldsymbol{\beta}}$$

$$\hat{y}_i^\star = \mathbf{x}_i^\star \hat{\boldsymbol{\beta}}$$

$$MSPE = \frac{1}{k}\sum_{i=1}^{k}\left(y_i^\star - \hat{y}_i^\star\right)^2$$

$$= \frac{1}{k}\sum_{i=1}^{k}\left(y_i^\star - \mathbf{x}_i^\star\hat{\boldsymbol{\beta}}\right)^2$$

Photo by Bank Phrom on Unsplash

The MSPE for model selection:

1. Split the data into a training and test set.
2. Fit several competing models on the training set.
3. Calculate the MSPE for each model using the test set.
4. Choose the model with the lowest MSPE.

Lesson: Introducing the problem of collinearity

Module: Model Selection

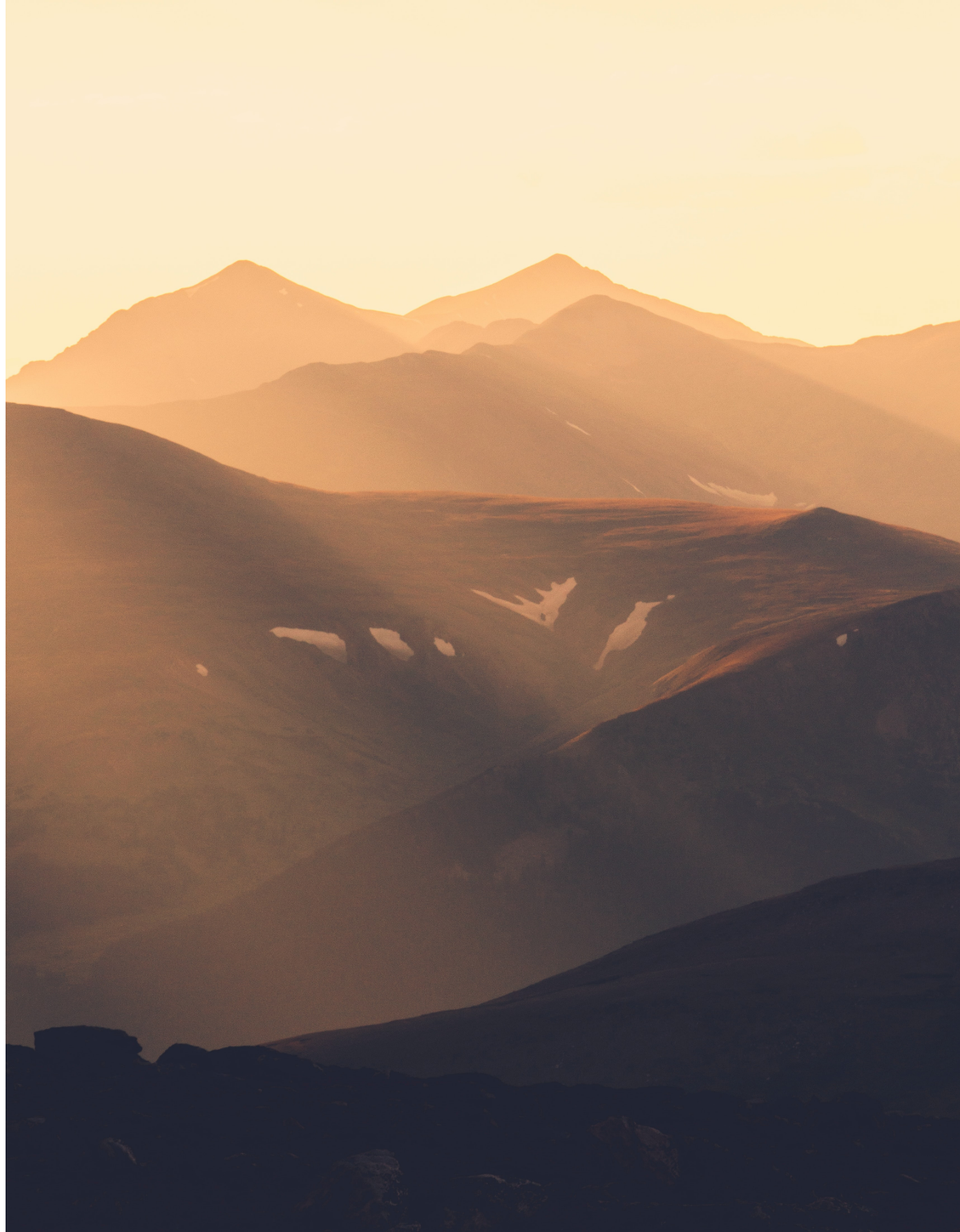


Photo by Calvin Weibel on Unsplash

$$(X^T X)\boldsymbol{\beta} = X^T \mathbf{Y}$$

Multicollinearity occurs when one predictor variable in a dataset is a near linear combination of others.

A matrix is **ill-conditioned** when the inverse cannot be computed with reliable accuracy.

Statistical misspecification vs lack of information

If $\mathbf{x}_j$ and $\mathbf{x}_k$ are highly correlated, then our interpretation of linear regression is not necessarily valid.

Lesson: Diagnosing multicollinearity

Module: Model Selection

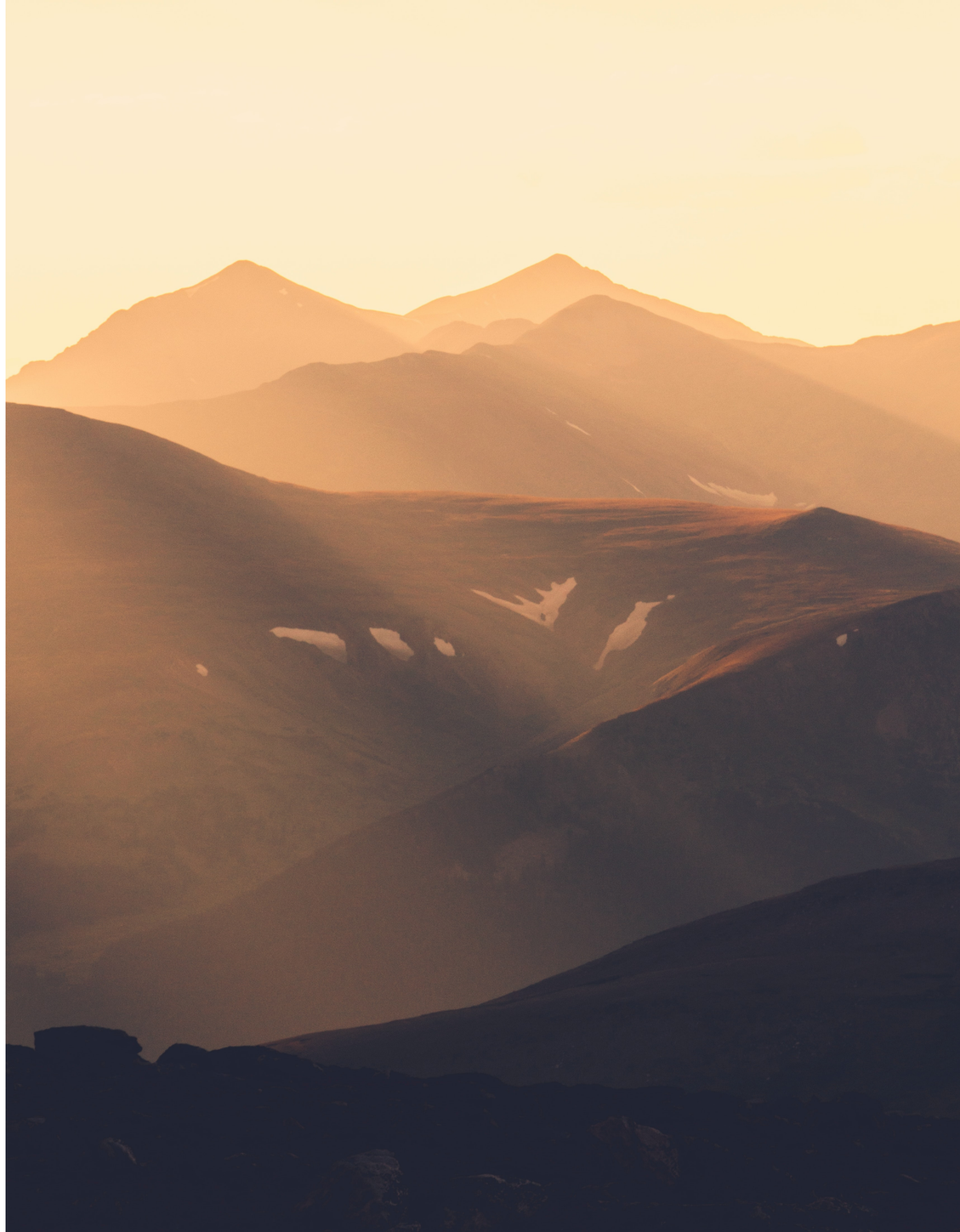


Photo by Calvin Weibel on Unsplash

Diagnosing multicollinearity:

1. Examine the correlation matrix for the predictors This is a check of *pairwise* collinearity, but nothing more complicated!

2. Examine the signs and standard error of $\hat{\boldsymbol{\beta}}$.

3. Compute and examine the *variance inflation factors.*

4. Compute the *condition number* of the design matrix.

Variance inflation factors. Note that:

$$Var\left(\hat{\beta}_j\right) = \frac{\sigma^2}{\sum_{i=1}^{n}\left(x_{i,j} - \overline{x_j}\right)^2} \times \frac{1}{1 - R_j^2}$$

Condition number of $(X^T X)$*:*

$$\kappa = \sqrt{\frac{\lambda_{largest}}{\lambda_{smallest}}}$$

Lesson: The problem of multicollinearity: solutions and R Implementation

Module: Model Selection

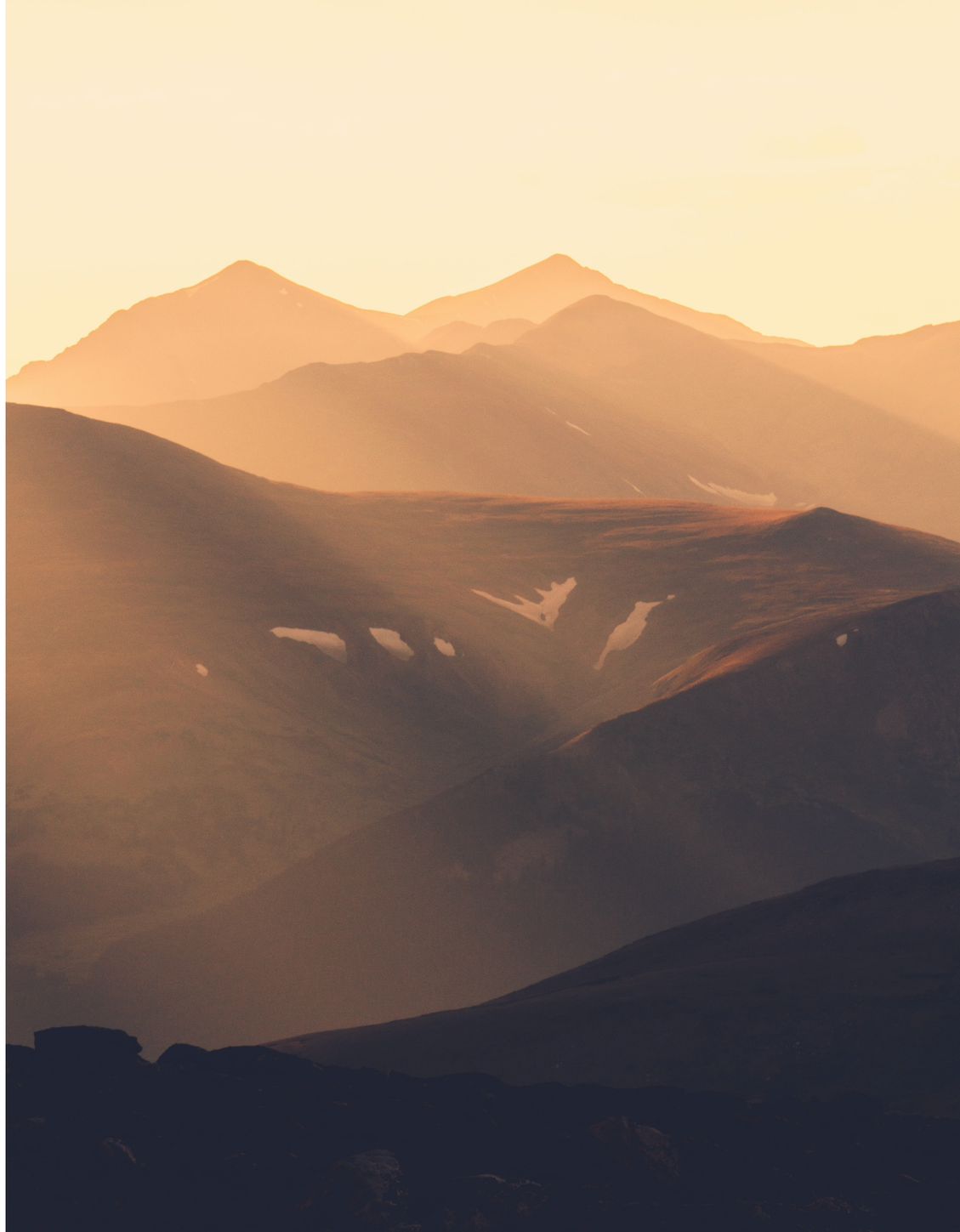


Photo by Calvin Weibel on Unsplash

We can rid our model of multicollinearity by:

1. Carefully removing some of the predictors.

2. Using alternative estimation methods (e.g., ridge regression).

3. *Principal Component Analysis (PCA)* is a dimension reduction tool used to reduce a large set of correlated predictor variables to a smaller, less correlated set.

Photo by Vincent Ledvina on Unsplash

```
Call:
lm(formula = acoustic1 ~ thickness + diameter + perforation +
    weight + stiffness + airPerm, data = fabric)

Residuals:
      Min        1Q    Median        3Q       Max
-0.062713 -0.021894 -0.002938  0.020573  0.065304

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.5858162  0.0896400   6.535 5.09e-06 ***
thickness   -0.0288504  0.1606300  -0.180   0.8596
diameter    -0.1340331  0.1192525  -1.124   0.2767
perforation -0.0215229  0.0128301  -1.678   0.1117
weight       0.0008418  0.0004612   1.825   0.0856 .
stiffness   -0.0002042  0.0001622  -1.258   0.2253
airPerm     -0.0003019  0.0001228  -2.458   0.0250 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03871 on 17 degrees of freedom
Multiple R-squared:  0.8393,    Adjusted R-squared:  0.7826
F-statistic:  14.8 on 6 and 17 DF,  p-value: 6.511e-06
```

Photo by Vincent Ledvina on Unsplash

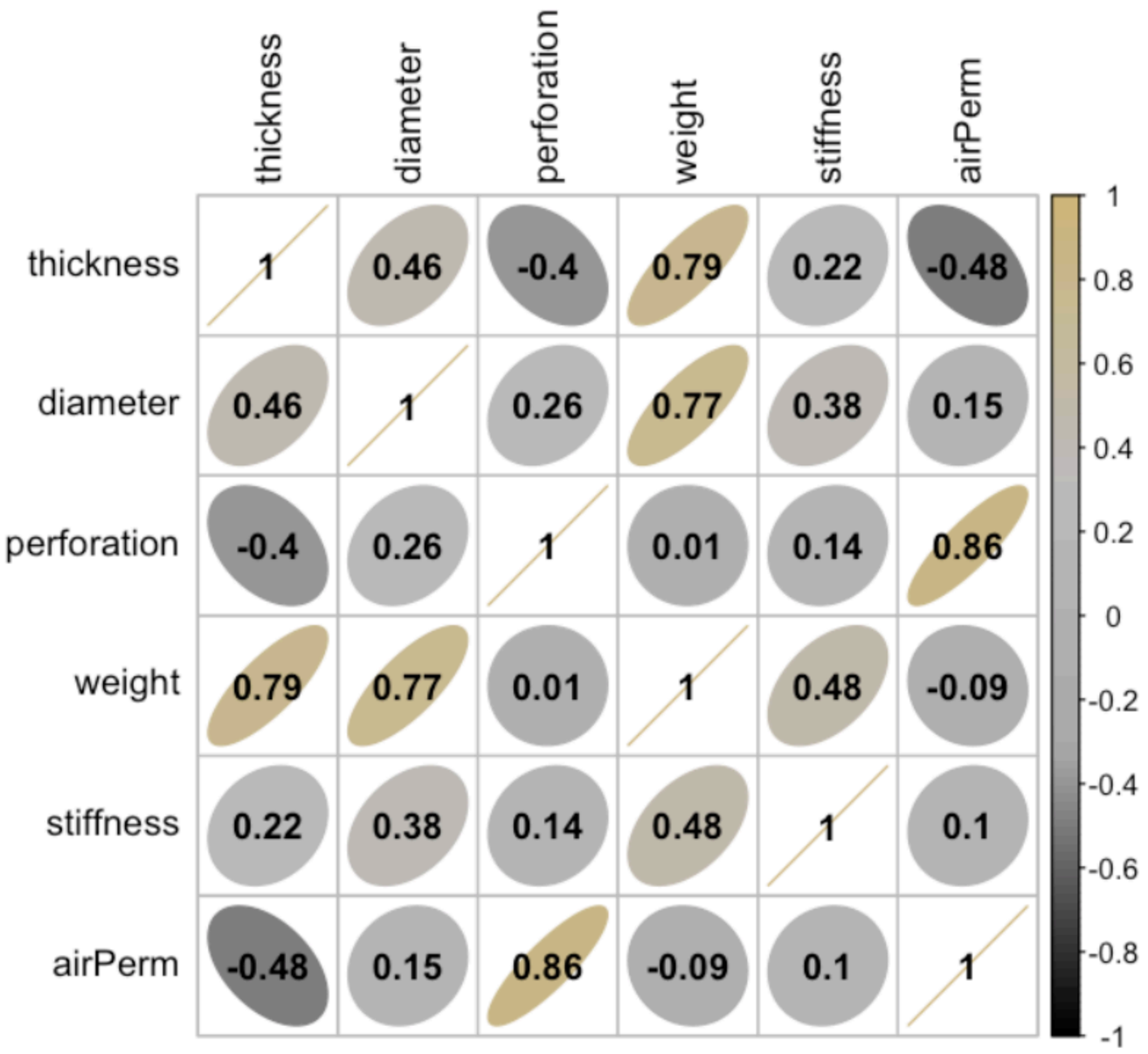


Photo by Vincent Ledvina on Unsplash

$$\kappa = 12046$$

```
library(car)
vif(lm_fabric_full)
```

thickness	5.7177718901297
diameter	2.99024905912487
perforation	4.25314494034739
weight	8.21796900907649
stiffness	1.45859061852712
airPerm	4.2855401058216

```
Call:
lm(formula = acoustic1 ~ perforation + weight + airPerm, data = fabric)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.074211 -0.025324  0.003821  0.020601  0.076329 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.6159414  0.0689289   8.936 2.03e-08 ***
perforation -0.0244787  0.0122470  -1.999   0.0594 .  
weight       0.0004294  0.0001628   2.638   0.0158 *  
airPerm     -0.0003053  0.0001172  -2.605   0.0169 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03845 on 20 degrees of freedom
Multiple R-squared:  0.8135,	Adjusted R-squared:  0.7855 
F-statistic: 29.08 on 3 and 20 DF,  p-value: 1.717e-07
```

Photo by Vincent Ledvina on Unsplash

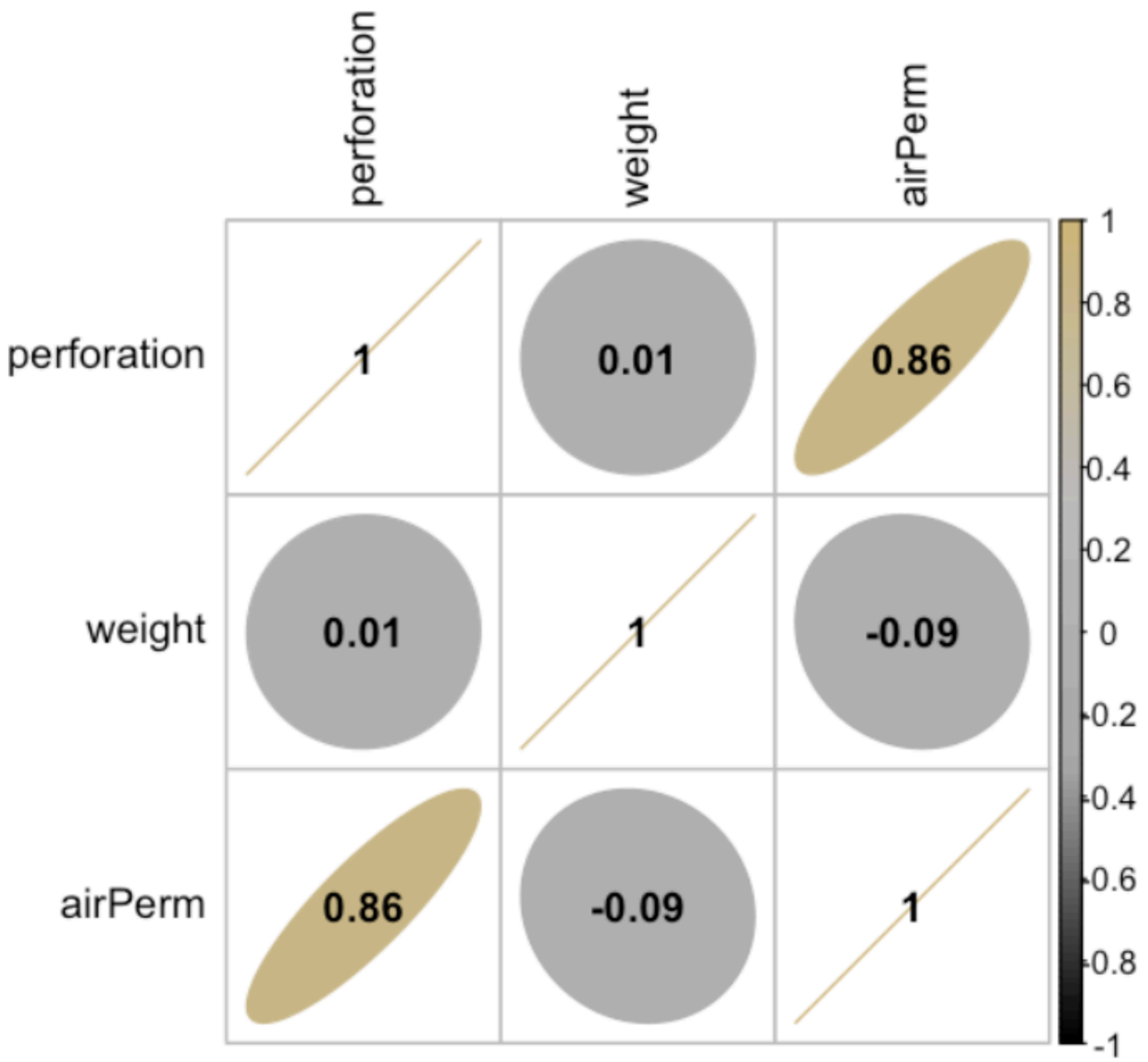


Photo by Vincent Ledvina on Unsplash

$$\kappa = 4531$$

```
vif(lm_fabric)
```

perforation	3.92763492568555
weight	1.03773036013852
airPerm	3.95753365045106