Inference from linear models

Analysis of Ecological and Environmental Data

QERM 514

Mark Scheuerell

10 April 2020

Goals for today

- Understand the concept and practice of partitioning sums-of-squares
- Understand the uses of R^2 and adjusted-R^2 for linear models
- Understand the use of *F*-tests for hypothesis testing
- Understand how to estimate confidence intervals

Partitioning variance

In general, we have something like

$$DATA = MODEL + ERRORS$$

and hence

$$\mathrm{Var}(DATA) = \mathrm{Var}(MODEL) + \mathrm{Var}(ERRORS)$$

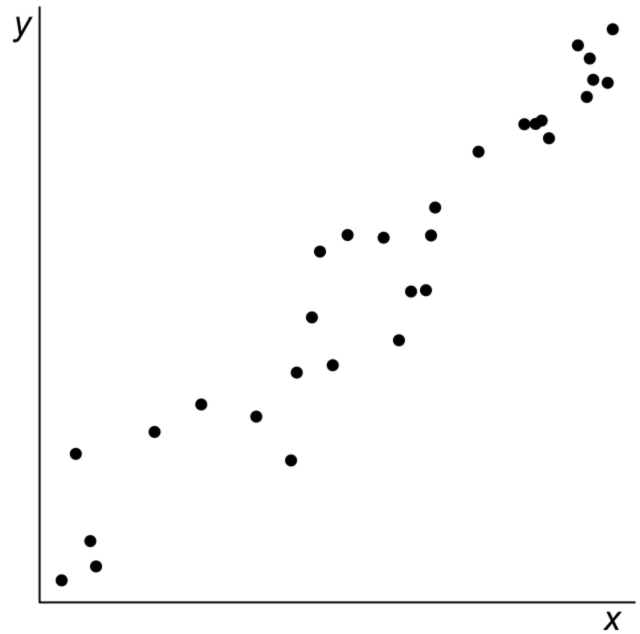
Partitioning total deviations

The total deviations in the data equal the sum of those for the model and errors

$$\underbrace{y_i - \bar{y}}_{\text{Total}} = \underbrace{\hat{y}_i - \bar{y}}_{\text{Model}} + \underbrace{y_i - \hat{y}_i}_{\text{Error}}$$

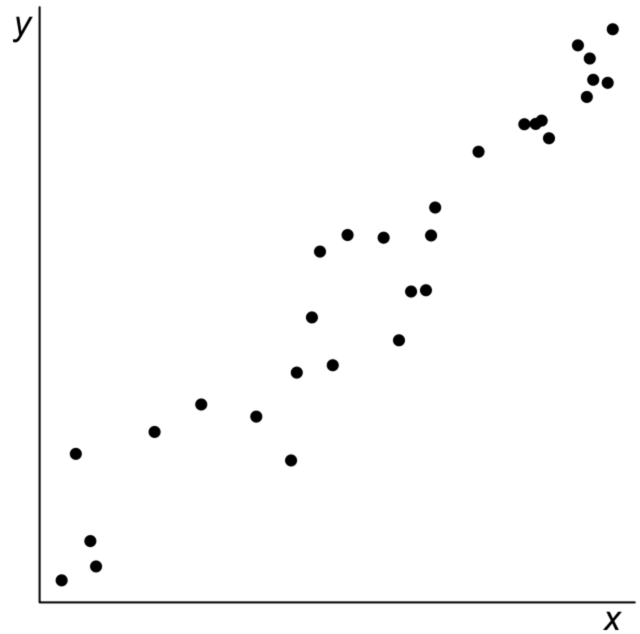
Partitioning total deviations

Here is a plot of some data y and a predictor x

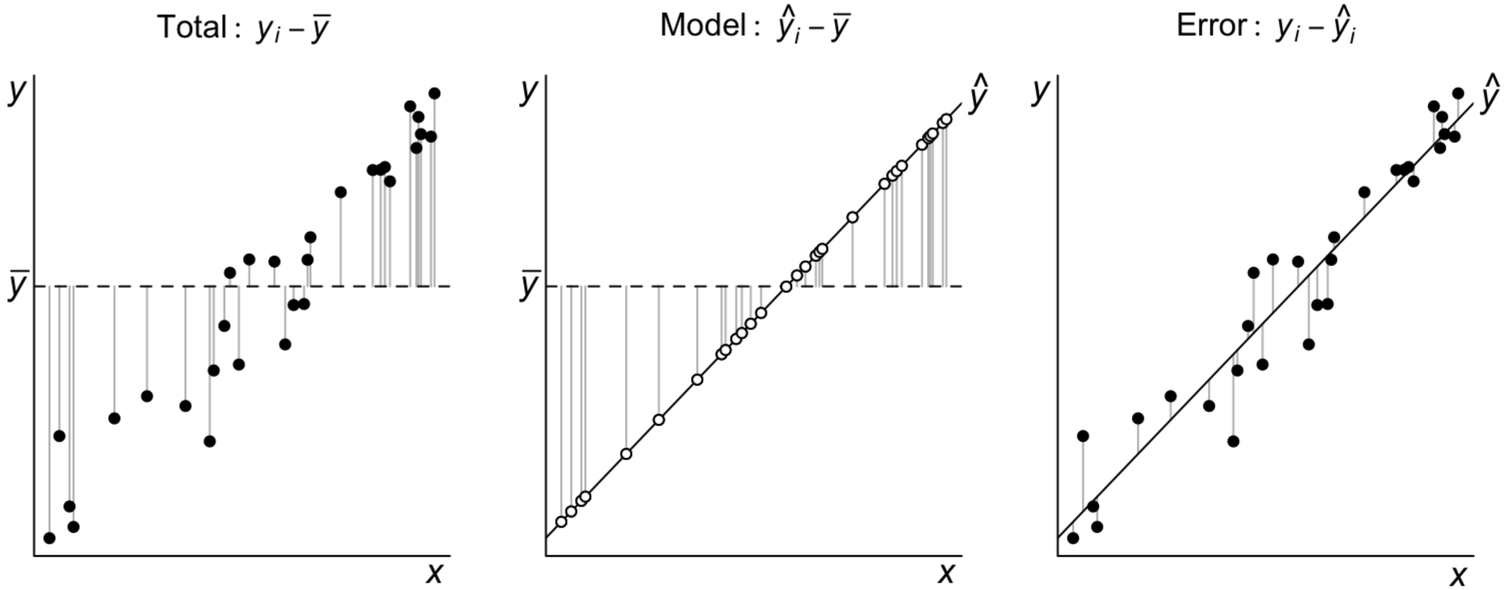


Partitioning total deviations

And let's consider this model: $y_i = \alpha + \beta x_i + e_i$



Partitioning total deviations



Sum-of-squares: Total

The total sum-of-squares $(SSTO)$ measures the total variation in the data as the differences between the data and their mean

$$SSTO = \sum (y_i - \bar{y})^2$$

Sum-of-squares: Model

The model (regression) sum-of-squares (SSR) measures the variation between the model fits and the mean of the data

$$SSR = \sum \left(\hat{y}_i - \bar{y}\right)^2$$

Sum-of-squares: Error

The error sum-of-squares (SSE) measures the variation between the data and the model fits

$$SSE = \sum \left(y_i - \hat{y}_i\right)^2$$

Partitioning sums-of-squares

The sums-of-squares have the same additive property as the deviations

$$\underbrace{\sum (y_i - \bar{y})^2}_{SSTO} = \underbrace{\sum (\hat{y}_i - \bar{y})^2}_{SSR} + \underbrace{\sum (y_i - \hat{y}_i)^2}_{SSE}$$

Goodness-of-fit

How about a measure of how well a model fits the data?

- $SSTO$ measures the variation in y *without* considering X
- SSE measures the reduced variation in y *after* considering X
- Let's consider this reduction in variance as a proportion of the total

Goodness-of-fit

A common option is the *coefficient of determination* or (R^2)

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

$$0 < R^2 < 1$$

Degrees of freedom

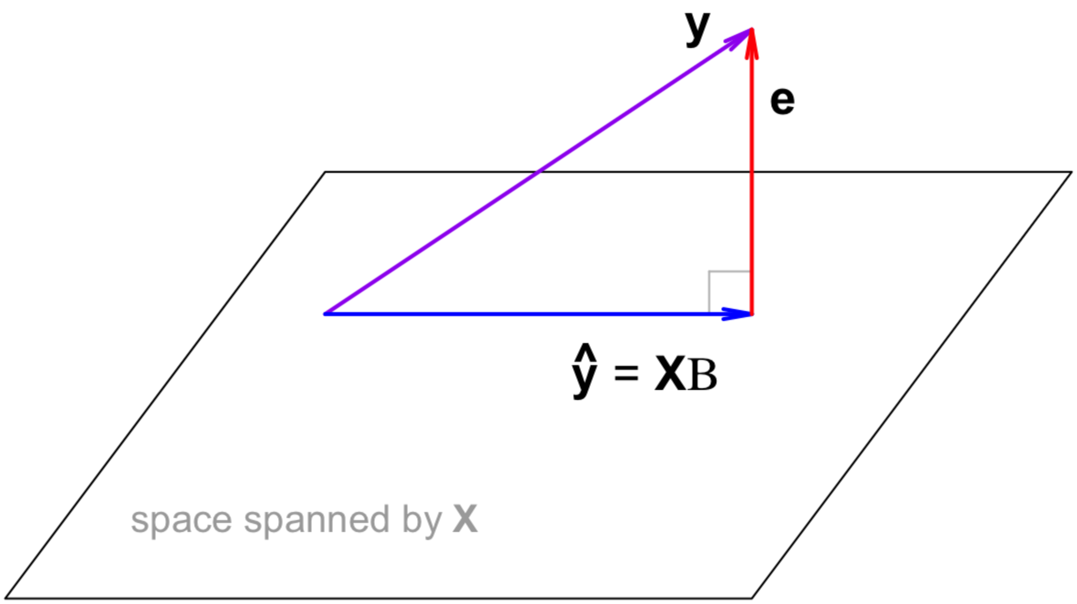
The number of *independent* elements that are free to vary when estimating quantities of interest

Degrees of freedom

An example

- Imagine you have 7 hats and you want to wear a different one on each day of the week.
- On day 1 you can choose any of the 7, on day 2 any of the remaining 6, and so forth
- When day 7 rolls around, however, you are out of choices: there is only one unworn hat
- Thus, you had 7 - 1 = 6 days of freedom to choose your hat

Model in geometric space



$\mathbf{y}$ is n-dim; $\hat{\mathbf{y}}$ is k-dim; $\mathbf{e}$ is $(n-k)$-dim

Degrees of freedom

Linear models

Beginning with $SSTO$, we have

$$SSTO = \sum (y_i - \bar{y})^2$$

The data are unconstrained and lie in an n-dimensional space, but estimating the mean $(\bar{y})$ from the data costs 1 degree of freedom (df), so

$$df_{SSTO} = n - 1$$

Degrees of freedom

Linear models

For the SSR we have

$$SSR = \sum \left(\hat{y}_i - \bar{y}\right)^2$$

We estimate the data $(\hat{y})$ with a k-dimensional model, but we lose 1 df when estimating the mean, so

$$df_{SSR} = k - 1$$

Degrees of freedom

Linear models

The SSE is analogous

$$SSE = \sum \left(y_i - \hat{y}_i\right)^2$$

The data lie in an n-dimensional space and we represent them in a k-dimensional subspace, so

$$df_{SSE} = n - k$$

Mean squares

The expectation of the sum-of-squares or “mean square” gives an indication of the variance for the model and errors

A mean square is a sum-of-squares divided by its degrees of freedom

$$MS = \frac{SS}{df}$$

$$\Downarrow$$

$$MSR = \frac{SSR}{k-1} \quad \& \quad MSE = \frac{SSE}{n-k}$$

Variance estimates

We are typically interested in two variance estimates:

1. The variance of the residuals $\mathbf{e}$
2. The variance of the model parameters $\mathbf{B}$

Variance estimates

Residuals

In a least squares context, we assume that the model errors (residuals) are independent and identically distributed with mean 0 and variance σ^2

The problem is that we don't know σ^2 and therefore we must estimate it

Variance estimates

Residuals

If $z_i \sim \mathrm{N}(0, 1)$ then

$$\sum_{i=1}^{n} z_i^2 = \mathbf{z}^\top \mathbf{z} \sim \chi_n^2$$

Variance estimates

Residuals

If $z_i \sim \mathrm{N}(0, 1)$ then

$$\sum_{i=1}^{n} z_i^2 = \mathbf{z}^\top \mathbf{z} \sim \chi_n^2$$

In our linear model, $e_i \sim \mathrm{N}(0, \sigma^2)$ so

$$\sum_{i=1}^{n} e_i^2 = \mathbf{e}^\top \mathbf{e} \sim \sigma^2 \cdot \chi_{n-k}^2$$

Variance estimates

Residuals

Thus, given

$$\mathbf{e}^\top \mathbf{e} \sim \sigma^2 \cdot \chi^2_{n-k}$$

$$\mathrm{E}(\chi^2_{n-k}) = n - k$$

$$\mathrm{E}(\mathbf{e}^\top \mathbf{e}) = SSE$$

then

$$SSE = \sigma^2(n-k) \;\Rightarrow\; \sigma^2 = \frac{SSE}{n-k} = MSE$$

Variance estimates

Parameters

Recall that our estimate of the model parameters is

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

Variance estimates

Parameters

Estimating the variance of the model parameters $\boldsymbol{\beta}$ requires some linear algebra

For a scalar z, if $\text{Var}(z) = \sigma^2$ then $\text{Var}(az) = a^2\sigma^2$

For a vector $\mathbf{z}$, if $\text{Var}(\mathbf{z}) = \boldsymbol{\Sigma}$ then $\text{Var}(\mathbf{A}\mathbf{z}) = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top$

Variance estimates

Parameters

The variance of the parameters is therefore

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \\ &= \left[(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \right] \mathbf{y}\end{aligned}$$

$$\Downarrow$$

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \left[(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \right] \text{Var}(\mathbf{y}) \left[(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \right]^\top$$

Variance estimates

Parameters

Recall that we can write our model in matrix form as

$$
\begin{gathered}
\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \\
\mathbf{e} \sim \text{MVN}(\mathbf{0}, \sigma^2 \mathbf{I})
\end{gathered}
$$

Variance estimates

Parameters

We can rewrite our model more compactly as

$$
\begin{gathered}
\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \\
\mathbf{e} \sim \text{MVN}(\mathbf{0}, \sigma^2 \mathbf{I}) \\
\Downarrow \\
\mathbf{y} \sim \text{MVN}(\mathbf{X}\boldsymbol{\beta}, \quad \underbrace{\sigma^2 \mathbf{I}}_{\text{Var}(\mathbf{y}|\mathbf{X}\boldsymbol{\beta})} \quad)
\end{gathered}
$$

Variance estimates

Parameters

Our estimate of $\mathrm{Var}(\hat{\boldsymbol{\beta}})$ is then

$$\begin{aligned}
\mathrm{Var}(\hat{\boldsymbol{\beta}}) &= \left[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\right]\mathrm{Var}(\mathbf{y})\left[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\right]^\top \\
&= \left[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\right]\sigma^2\mathbf{I}\left[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\right]^\top \\
&= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\mathbf{X})\left[(\mathbf{X}^\top\mathbf{X})^{-1}\right]^\top \\
&= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}
\end{aligned}$$

Variance estimates

Parameters

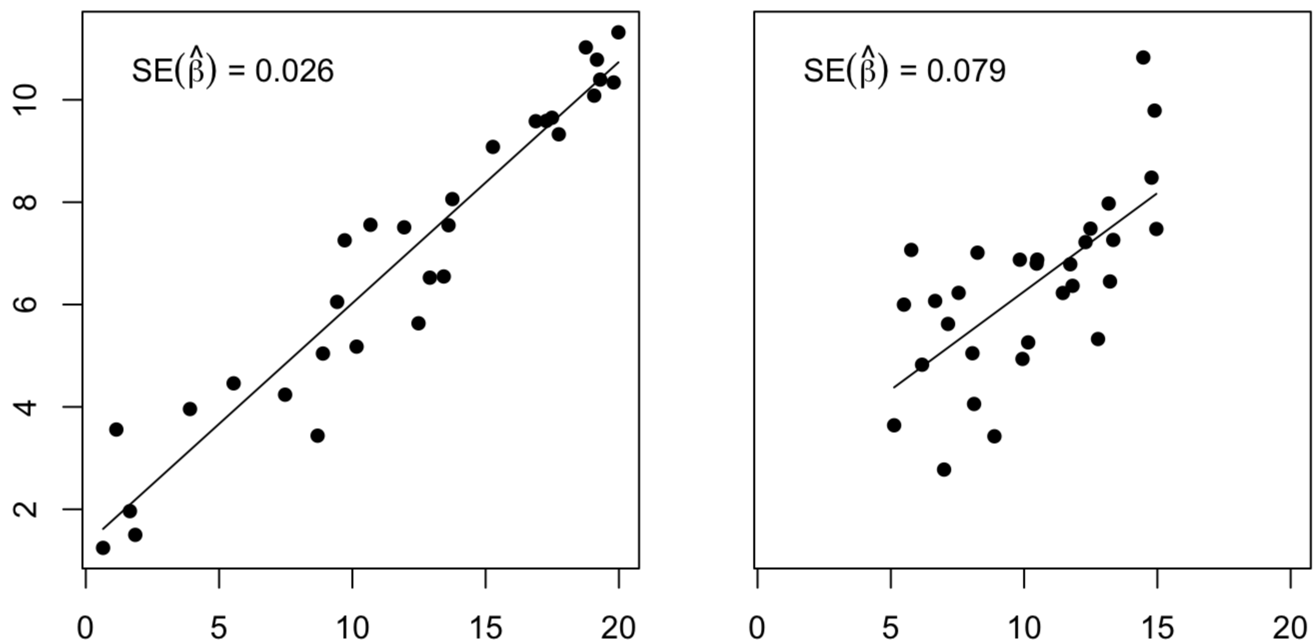
Let's think about the variance of $\hat{\boldsymbol{\beta}}$

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$$

This suggests that our confidence in our estimate increases with the spread in $\mathbf{X}$

Effect of $\mathbf{X}$ on parameter precision

Consider these two scenarios where the slope of the relationship is identical



QUESTIONS?

Inferential methods

Once we’ve estimated the model parameters and their variance, we might want to draw conclusions from our analysis

Comparing models

Imagine we had 2 linear models of varying complexity:

1. a model with one predictor
2. a model with five predictors

It would seem logical to ask whether the complexity of (2) is necessary?

Hypothesis test to compare models

Recall our partitioning of sums-of-squares, where

$$SSTO = SSR + SSE$$

We might prefer the more complex model (call it Θ) over the simple model (call it θ) if

$$SSE_{\Theta} < SSE_{\theta}$$

or, more formally, if

$$\frac{SSE_{\theta} - SSE_{\Theta}}{SSE_{\Theta}} > \text{a constant}$$

Hypothesis test to compare models

If Θ has k_Θ parameters and θ has k_θ, we can scale this ratio to arrive at an F-statistic that follows an F distribution

$$F = \frac{\left(SSE_\theta - SSE_\Theta\right)/(k_\Theta - k_\theta)}{SSE_\Theta/(n - k_\Theta)} \sim F_{k_\Theta - k_\theta, n - k_\Theta}$$

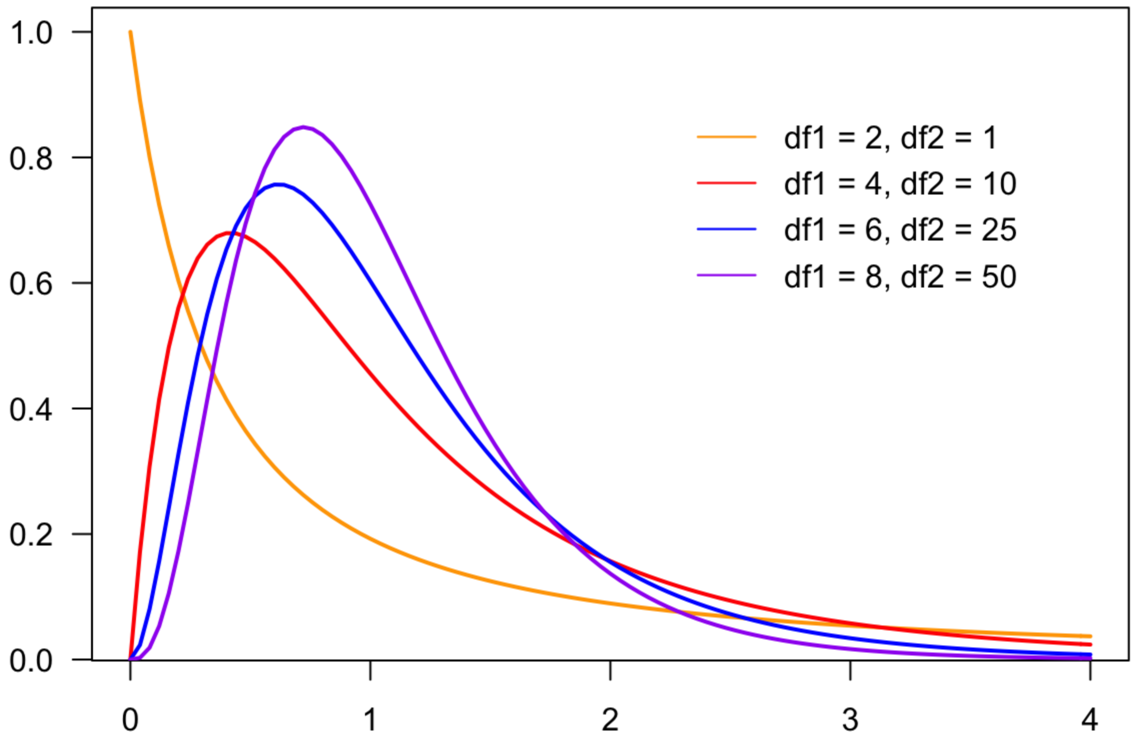
F-distribution

The F-distribution is the ratio of two random variates, each with a χ^2_n distribution

If $A \sim \chi^2_{df_A}$ and $B \sim \chi^2_{df_B}$ are independent, then

$$\frac{\left(\frac{A}{df_A}\right)}{\left(\frac{B}{df_B}\right)} \sim F_{df_A, df_B}$$

F-distribution



Test of *all* predictors in a model

Suppose we wanted to test whether the collection of predictors in a model were better than simply estimating the data by their mean.

$$\Theta : \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$
$$\theta : \mathbf{y} = \boldsymbol{\mu} + \mathbf{e}$$

We write the null hypothesis as

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

and we would reject H_0 if $F > F^{(\alpha)}_{k_\Theta - k_\theta, n - k_\Theta}$

Hypothesis test to compare models

$$SSE_{\Theta} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\top} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{e}^{\top}\mathbf{e} = SSE$$

$$SSE_{\theta} = (\mathbf{y} - \bar{y})^{\top} (\mathbf{y} - \bar{y}) = SSTO$$

$$\Downarrow$$

$$F = \frac{(SSTO - SSE)/(k-1)}{SSE/(n-k)}$$

Predictors of plant diversity

Later in lab we will work with the `gala` dataset[†] in the **faraway** package, which contains data on the diversity of plant species across 30 Galapagos islands

For now let's hypothesize that

$$\text{diversity} = f(\text{area, elevation, distance to nearest island})$$

[†]From Johnson & Raven (1973) *Science* 179:893-895

Testing one predictor

We might ask whether any one predictor could be dropped from a model

For example, can nearest be dropped from ourf full model?

$$\text{species}_i = \alpha + \beta_1 \text{area}_i + \beta_2 \text{elevation}_i + \beta_3 \text{nearest}_i + \epsilon_i$$

Testing one predictor

One option is to fit these two models and compare them via our F-test with $H_0 : \beta_3 = 0$

$$\text{species}_i = \alpha + \beta_1 \text{area}_i + \beta_2 \text{elevation}_i + \beta_3 \text{nearest}_i + \epsilon_i$$

$$\text{species}_i = \alpha + \beta_1 \text{area}_i + \beta_2 \text{elevation}_i + \epsilon_i$$

Testing one predictor

Another option is to estimate a t-statistic as

$$t_i = \frac{\hat{\beta}_i}{\mathrm{SE}\left(\hat{\beta}_i\right)}$$

and compare it to a t-distribution with $n - k$ degrees of freedom

Testing 2+ predictors

Sometimes we might want to know whether we can drop 2+ predictors from a model

For example, can we drop both elevation and nearest from our full model?

$$\text{species}_i = \alpha + \beta_1 \text{area}_i + \beta_2 \text{elevation}_i + \beta_3 \text{nearest}_i + \epsilon_i$$

$$\text{species}_i = \alpha + \beta_1 \text{area}_i + \epsilon_i$$

$H_0 : \beta_2 = \beta_3 = 0$

Testing a subspace

Some tests cannot be expressed in terms of the inclusion or exclusion of predictors

Consider a test of whether the areas of the current and adjacent island could be added together and used in place of the two separate predictors

$$\text{species}_i = \alpha + \beta_1 \text{area}_i + \beta_2 \text{adjacent}_i + \cdots + \epsilon_i$$

$$\text{species}_i = \alpha + \beta_1 (\text{area} + \text{adjacent})_i + \cdots + \epsilon_i$$

$H_0 : \beta_{\text{area}} = \beta_{\text{adjacent}}$

Testing a subspace

What if we wanted to test whether a predictor had a specific (non-zero) value?

For example, is there a 1:1 relationship between species and elevation after controlling for the other predictors?

$$\text{species}_i = \alpha + \beta_1 \text{area}_i + \underline{1}\text{elevation}_i + \beta_3 \text{nearest}_i + \epsilon_i$$

$H_0 : \beta_2 = 1$

Testing a subspace

We can also modify our t-test from before and use it for our comparison by including the hypothesized β_{H_0} as an offset

$$t_i = \frac{(\hat{\beta}_i - \beta_{H_0})}{\mathrm{SE}\left(\hat{\beta}_i\right)}$$

Caveats about hypothesis testing

Null hypothesis testing (NHT) is a slippery slope

- p-values are simply the probability of obtaining a test statistic as large or greater than that observed
- p-values are **not** weights of evidence
- "Critical" or "threshold" values against which to compare p-values must be chosen *a priori*
- Be aware of "p hacking" where researchers make *many* tests to find significance

QUESTIONS?

Confidence intervals for β

We can also use confidence intervals (CI's) to express uncertainty in $\hat{\beta}_i$

They take the form

$$100(1-\alpha)\%\ \mathrm{CI} : \hat{\beta}_i \pm t_{n-p}^{(\alpha/2)}\ \mathrm{SE}(\hat{\beta})$$

where here α is our *predetermined* Type-I error rate

Bootstrap confidence intervals

The F- and t-based CI's we have described depend on the assumption of normality

The bootstrap[†] method provides a way to construct CI's without this assumption

[†]Efron (1979) *The Annals of Statistics* 7:1–26

Bootstrap procedure

1. Fit your model to the data
2. Calculate $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$
3. Do the following *many* times:
    - Generate $\mathbf{e}^*$ by sampling *with replacement* from $\mathbf{e}$
    - Calculate $\mathbf{y}^*$ = $\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{e}^*$
    - Estimate $\hat{\boldsymbol{\beta}}^*$ from $\mathbf{X}$ & $\mathbf{y}^*$)
4. Select the $\frac{\alpha}{2}$ and $(1 - \frac{\alpha}{2})$ percentiles from the saved $\hat{\boldsymbol{\beta}}^*$

Confidence interval for new predictions

Given a fitted model $\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{e}$, we might want to know the uncertainty around a new estimate $\mathbf{y}^*$ given some new predictor $\mathbf{X}^*$

CI for the mean response

Suppose we wanted to estimate the uncertainty in the *average* response given by

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

Recall that the general formula for a CI on a quantity z is

$$100(1 - \alpha)\% \text{ CI} : \text{E}(z) \ \pm \ t_{df}^{(\alpha/2)} \text{SD}(z)$$

So we would have

$$\hat{\mathbf{y}}^* \ \pm \ t_{df}^{(\alpha/2)} \sqrt{\text{Var}\left(\hat{\mathbf{y}}^*\right)}$$

CI for the mean response

We can calculate the SD of our expectation as

$$
\begin{aligned}
\text{Var}\left(\hat{\mathbf{y}}^*\right) &= \text{Var}\left(\mathbf{X}^* \hat{\boldsymbol{\beta}}\right) \\
&= \mathbf{X}^{*\top} \text{Var}\left(\hat{\boldsymbol{\beta}}\right) \mathbf{X}^* \\
&= \mathbf{X}^{*\top} \left[\sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}\right] \mathbf{X}^* \\
&\Downarrow \\
\text{SD}\left(\hat{\mathbf{y}}^*\right) &= \sigma \sqrt{\mathbf{X}^{*\top} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^*}
\end{aligned}
$$

CI for the mean response

So our CI on the mean response is given by

$$\hat{\mathbf{y}}^* \pm t_{df}^{(\alpha/2)} \sigma \sqrt{\mathbf{X}^{*\top} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^*}$$

CI for a specific response

What about the uncertainty in a *specific* prediction?

In that case we need to account for our additional uncertainty owing to the error in our relationship, which is given by

$$\hat{\mathbf{y}}^* = \mathbf{X}^* \hat{\boldsymbol{\beta}} + \mathbf{e}$$

CI for a specific response

The SD of the new prediction is given by

$$\begin{aligned}
\text{Var}\left(\hat{\mathbf{y}}^*\right) &= \mathbf{X}^{*\top}\text{Var}\left(\hat{\boldsymbol{\beta}}\right)\mathbf{X}^* + \text{Var}\left(\mathbf{e}\right) \\
&= \mathbf{X}^{*\top}\left[\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\right]\mathbf{X}^* + \sigma^2 \\
&= \sigma^2\left(\mathbf{X}^{*\top}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^* + 1\right) \\
&\Downarrow \\
\text{SD}\left(\hat{\mathbf{y}}^*\right) &= \sigma\sqrt{1 + \mathbf{X}^{*\top}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^*}
\end{aligned}$$

CI for a specific response

So our CI on the new prediction is given by

$$\hat{\mathbf{y}}^* \pm t_{df}^{(\alpha/2)} \sigma \sqrt{1 + \mathbf{X}^{*\top} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^*}$$

This is typically referred to as the *prediction interval*