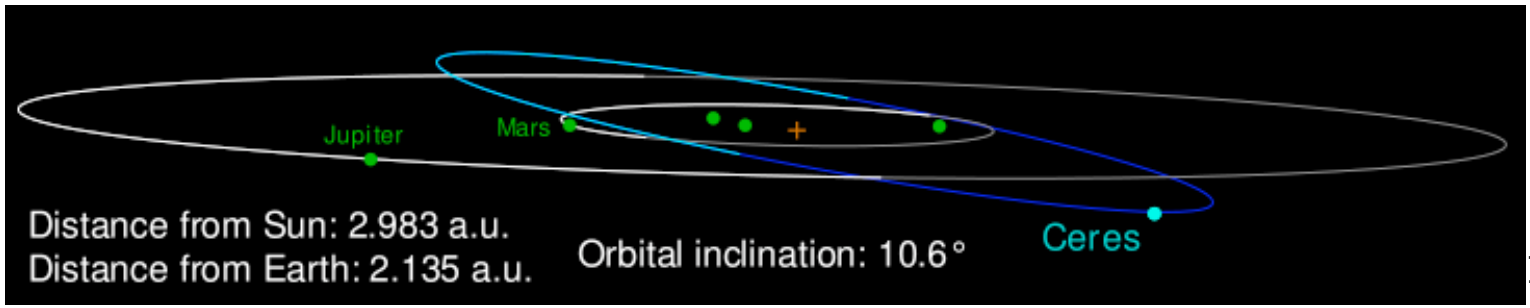
Introduction to Linear Regression

Timothy DelSole

George Mason University, Fairfax, Va and
Center for Ocean-Land-Atmosphere Studies, Calverton, MD

July 24, 2010

1801: The Discovery (and Loss and Re-Discovery) of Ceres



[1]

1/1801 Ceres discovered by Giuseppe Piazzi.

2/1801 Piazzi stopped tracking it due to illness.

9/1801 Piazzi published his observations.

10/1801 Ceres too close to the sun to observe.

11/1801 Gauss (24 years old) determined orbit statistically.

12/1801 von Zach found Ceres, where Gauss predicted it.

[1] figure from http://www.newuniverse.co.uk/Ceres_28dwarf_planet29.html

The Method of Least Squares

Figure: Louis Legendre
Nouvelles methodes pour la determination des orbites des cometes (1805)

Figure: Carl Friedrich Gauss
Theoria Motus Corporum Coelestium in Sectionibus Connicis Solum Ambientium (1809)

Method of Least Squares (univariate)

- Consider data pairs (x_n, y_n) for $n = 1, \ldots, N$.
- Consider function $f(x, \beta_1, \ldots, \beta_M)$.
- Adjust $\beta_1, \ldots, \beta_M$ to "best" fit the data; $y \approx f(x, \beta_1, \ldots, \beta_M)$

Method of Least Squares: determine $\beta_1, \ldots, \beta_M$ that minimizes the sum square difference

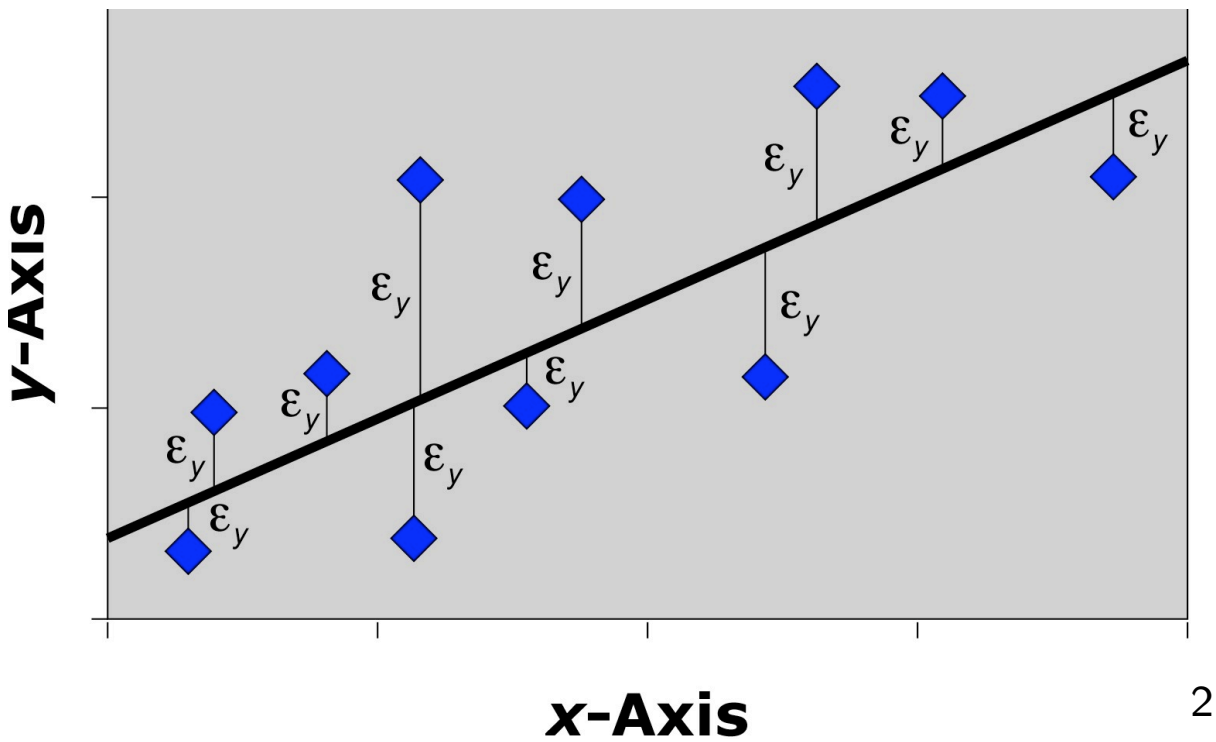
$$\sum_n (y_n - f(x_n, \beta_1, \ldots, \beta_M))^2$$

Note: if there were a combination of parameters $\beta_1, \ldots, \beta_M$ that fit the data exactly, this method would find it.

Method of Least Squares (Linear Case)

To fit $y \approx ax + b$, find a and b that minimizes

$$\sum_n \epsilon_n^2 = \sum_n (y_n - ax_n - b)^2$$



[2]

[2] figure http://www.palass.org/modules.php?name=palaeo_math&page=7

Terminology

$$\underset{\text{predictand}}{y_n} = \underset{\text{slope}}{a}\ \underset{\text{predictor}}{x_n} + \underset{\text{intercept}}{b} + \underset{\text{noise}}{\epsilon}$$

Regression Analysis: set of techniques (e.g., least squares method) for modeling and analyzing variable relations.

Regression Model: a proposed equation relating two or more variables (e.g., $y = ax + b + \epsilon$).

Regression Parameters: unknown parameters in a model (e.g., a, b, variance of ϵ) that are estimated from data.

Overdetermined System: more samples than regression parameters.

Underdetermines System: more parameters than samples

Multiple Least Squares

Method generalizes easily to multiple predictors:

$$y_n = x_{n1}\beta_1 + x_{n2}\beta_2 + \cdots + x_{nK}\beta_K + \epsilon_n$$

Least squares estimates are the $\beta_1, \beta_2, \ldots, \beta_K$ that minimize

$$\sum_n \epsilon_n^2 = \sum_n (y_n - (x_{n1}\beta_1 + x_{n2}\beta_2 + \cdots + x_{nK}\beta_K))^2$$

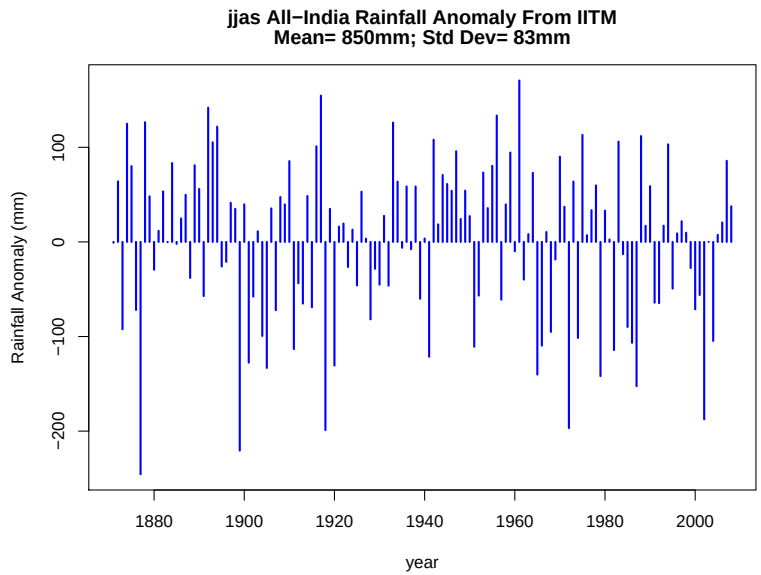
The Least Squares Solution

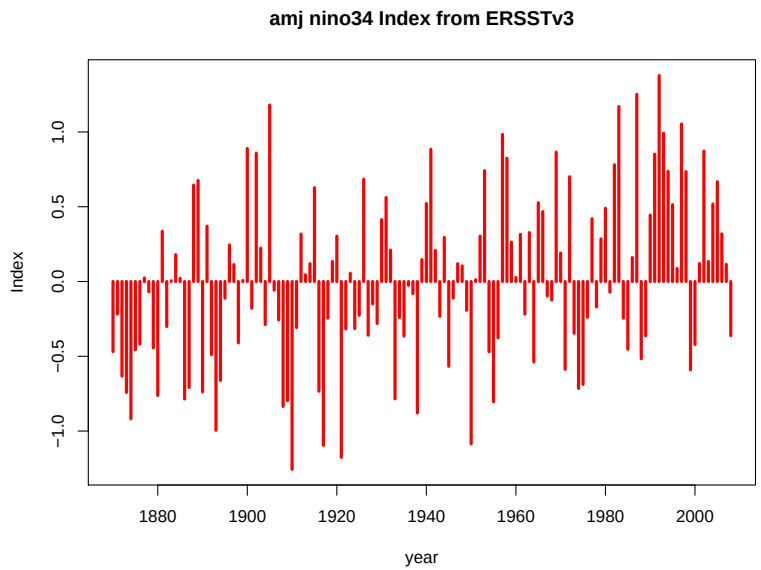
Calculus: Set derivative of sum square residual to zero and solve.

$$\frac{\partial}{\partial \beta_m} \sum_n (y_n - f(x_n, \beta_1, \ldots, \beta_M))^2 = 0$$

Least Squares Prediction

Is All-India Seasonal (JJAS) Monsoon Rainfall (ISMR) related to JFM-NINO4 Index?





Use Linear Regression To Find a Linear Relation

Least squares fit:

$$\text{ISMR} = -48 * \text{NINO4} + 855$$

Does this fit imply that NINO4 and ISMR are "really" related?

Suppose ISMR and NINO3.4 are Independent

Assuming ISMR and NINO3.4 are independent $\Rightarrow$ "true" model is

$$\text{ISMR} = b + \epsilon$$

where b is constant and ϵ is a random variable (NINO4 does not appear).

The most convenient hypothesis, owing to an extensive literature about it, is that the random variable has a Gaussian distribution.

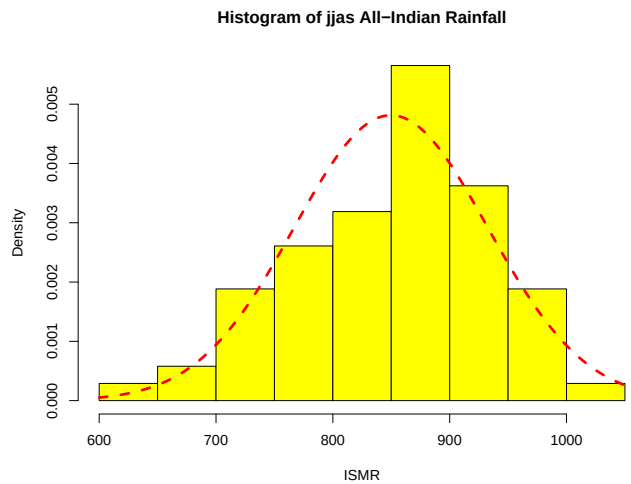


Figure: Histogram of JJAS All-India Rainfall. Red dashed line shows a Gaussian distribution with mean 849cm and standard error 83cm.

Implications

If the model ISMR $= b + \epsilon$ were true, then fitting

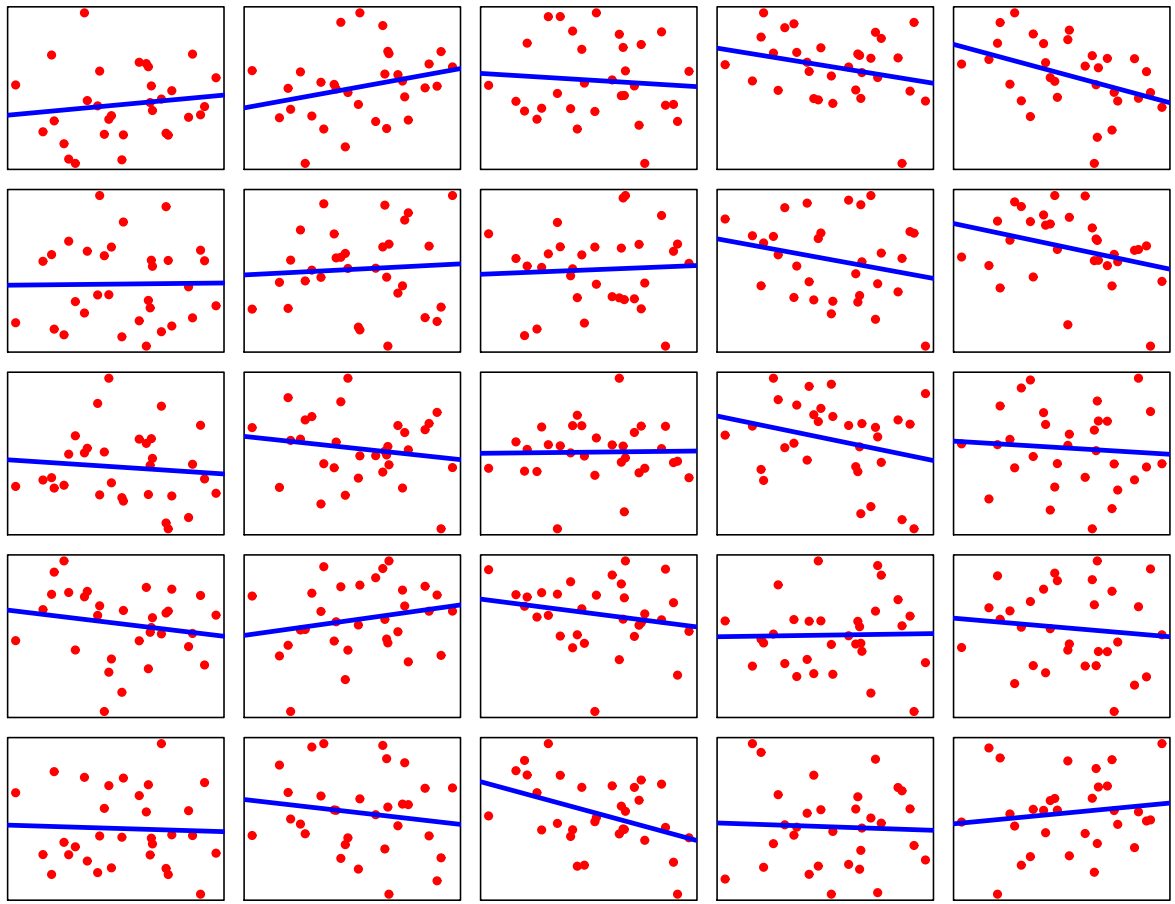
$$\text{ISMR} = a * \text{NINO4} + b + \text{noise}$$

will yield random regression coefficients a and b that depend on the particular realization of the random variable ISMR.

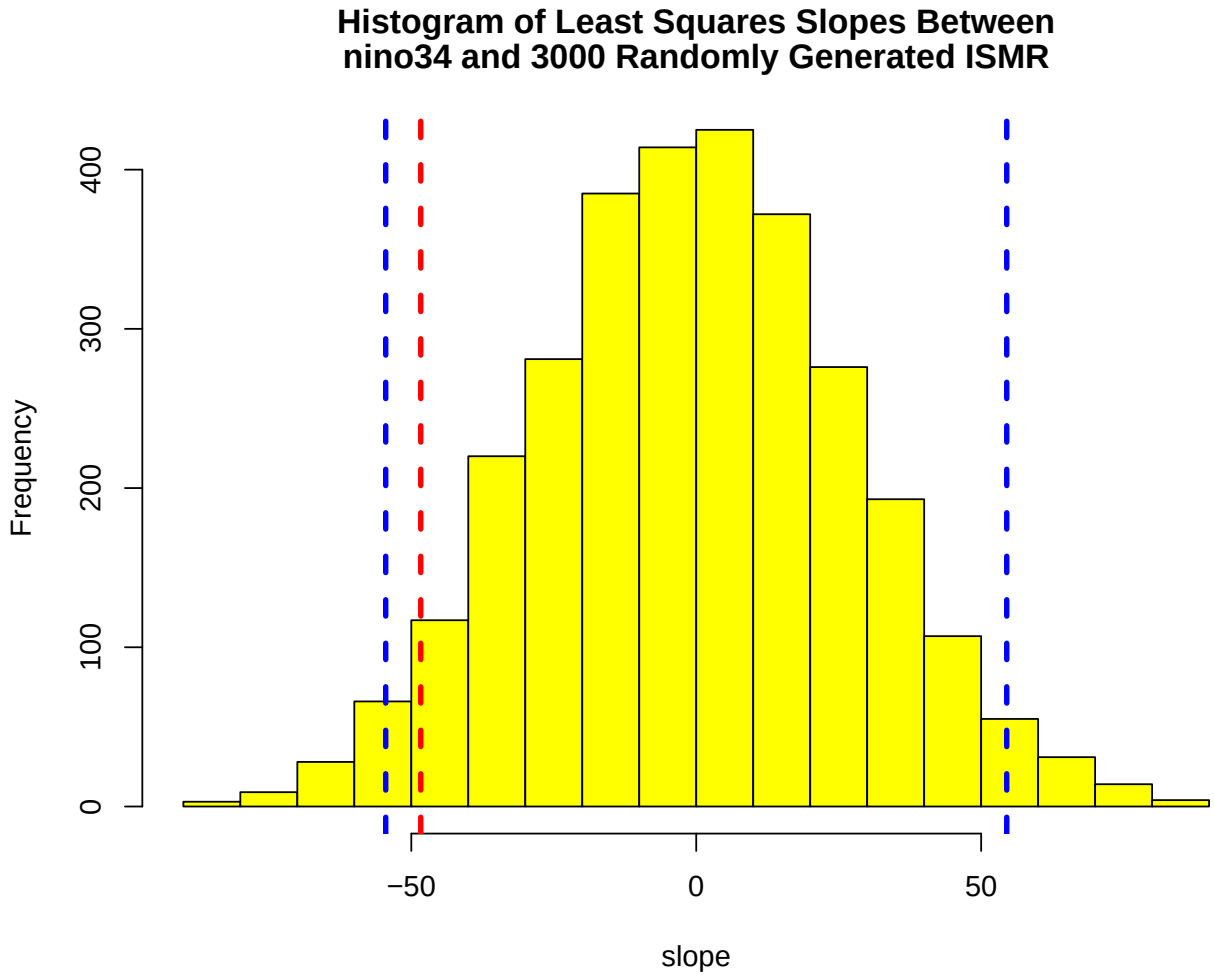
We can estimate the behavior of a by randomly generating ISMR and fitting the above equation. Repeating this many times yields...

Fits to Random ISMR

$$\mathrm{ISMR}_{ran} = a * \mathrm{NINO4} + b$$



Slopes From The Null Hypothesis



The observed slope (i.e., -48) is within a range of plausible values that would be obtained if ISMR were independent of NINO3.4.

Hypothesis Test

We never know the "real" relation (that's like being God), so the best we can do is to test a hypothesis about reality.

An hypothesis about reality introduced for the purpose of disproving it is called a null hypothesis.

Our null hypothesis is "ISMR is not related to NINO3.4 in nature."

Observed slope is similar to slopes expected from random ISMRs, suggesting that observed slope is indistinguishable from zero.

Comments About Hypothesis Tests

- Above hypothesis test generates random samples from a fitted Gaussian, but uncertainty in the fit itself was ignored.
- Instead of fitting two parameters (e.g., slope and intercept), we could be interested in a model with many predictors.

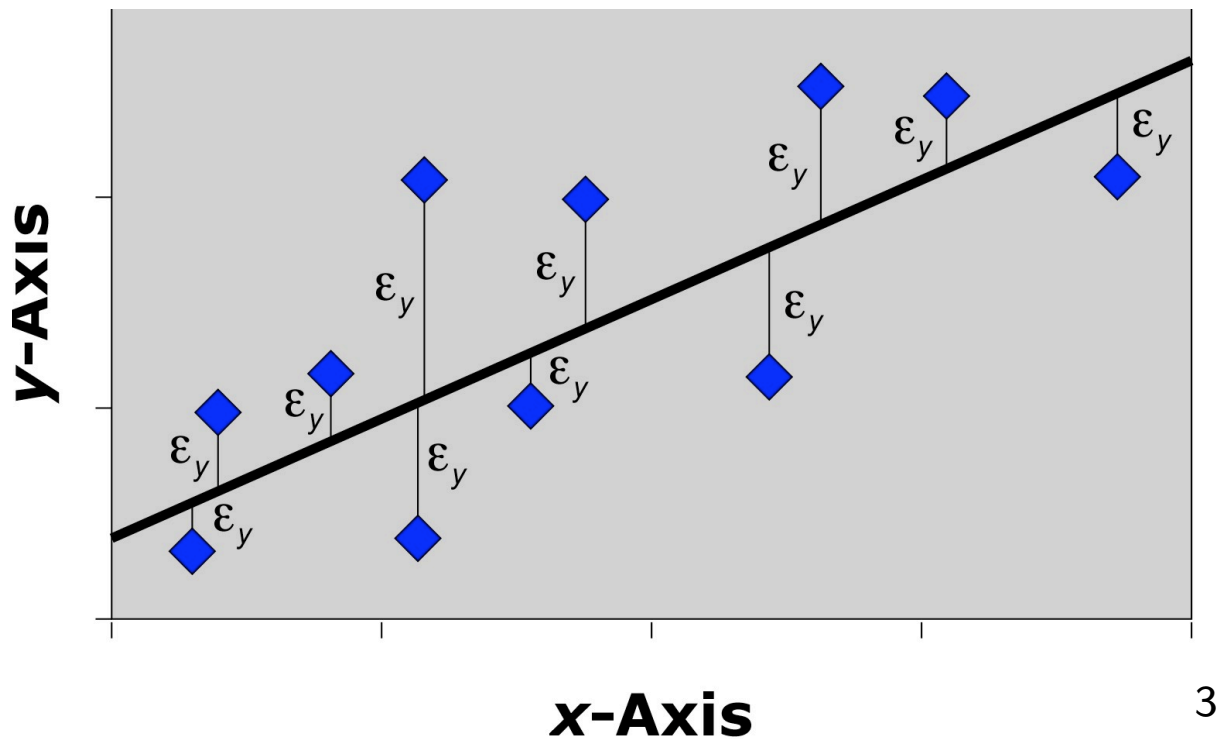
$$y_n = x_{n1}\beta_1 + x_{n2}\beta_2 + \cdots + x_{nK}\beta_K + \epsilon_n.$$

- How do you test the hypothesis $\beta_1 = \beta_2 = \cdots = \beta_K = 0$?
- How do you test the hypothesis $\beta_1 = 0$?

Sum Square Error (SSE): Measure of "Goodness of Fit"

To fit $y \approx ax + b$, find a and b that minimizes

$$SSE = \sum_n \epsilon_n^2 = \sum_n (y_n - ax_n - b)^2$$



[3]

[3] figure http://www.palass.org/modules.php?name=palaeo_math&page=7

Hypothesis Test as a Comparison of Two Models

Consider the linear model

$$y_n = ax_n + b + \epsilon_n.$$

Testing the hypothesis $a = 0$ is equivalent to comparing the models

$$\begin{aligned} \text{Full} \quad & y_n = ax_n + b + \epsilon_n \\ \text{Restricted} \quad & y_n = \qquad\quad b + \epsilon_n \end{aligned}$$

Note: the "restricted" model is a special case of the "full" model.

Compare SSEs of Two Models

$$\text{Full } SSE_F = \sum_n \left(y_n - \hat{a}x_n - \hat{b}\right)^2$$

$$\text{Restricted } SSE_R = \sum_n \left(y_n - \hat{b}'\right)^2$$

- If difference between SSE_F and SSE_R is small, then we prefer the restricted model because it is more parsimonious (i.e., it is the least complex model to explain the variability).
- If full model fits data "better," then we prefer the full model.
- This suggests that a comparison of models can be based on

$$SSE_R - SSE_F.$$

Compare SSEs

- If $SSE_R - SSE_F$ is small, then the two models have similar errors and we prefer the restricted model.
- If $SSE_R - SSE_F$ is large, then the full model has smaller errors than the restricted, so we prefer the full model.
- But what determines "small" or "large?"
- Normalize by the SSE of one of the models:

$$\frac{SSE_R - SSE_F}{SSE_F}$$

A Fundamental Theorem in Linear Regression

Consider the models

$$\begin{aligned} \text{Full } \mathbf{y} &= \mathbf{x}_1\boldsymbol{\beta}_1 + \cdots + \mathbf{x}_{M_R}\boldsymbol{\beta}_{M_R} + \cdots + \mathbf{x}_{M_F}\boldsymbol{\beta}_{M_F} + \boldsymbol{\epsilon} \\ \text{Reduced } \mathbf{y} &= \mathbf{x}_1\boldsymbol{\beta}_1 + \cdots + \mathbf{x}_{M_R}\boldsymbol{\beta}_{M_R} + \boldsymbol{\epsilon} \end{aligned}$$

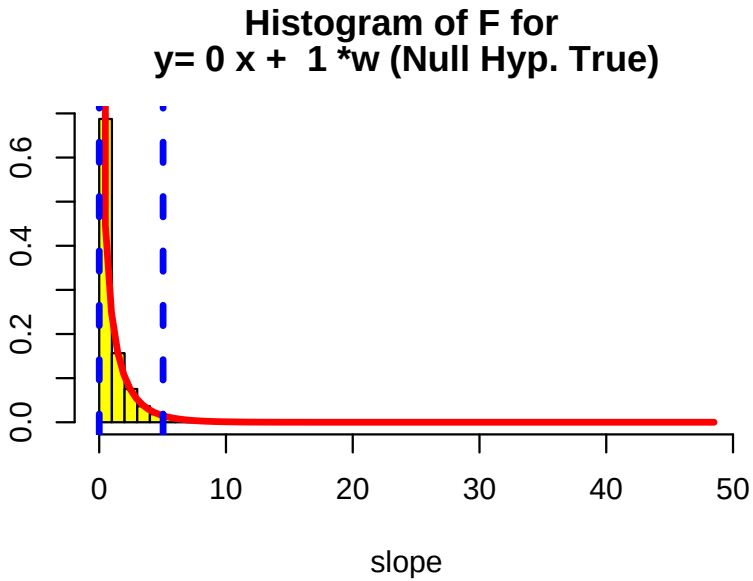
where all vectors are N-dimensional, and the elements of $\boldsymbol{\epsilon}$ are independent Gaussian variables with zero mean and variance σ^2.

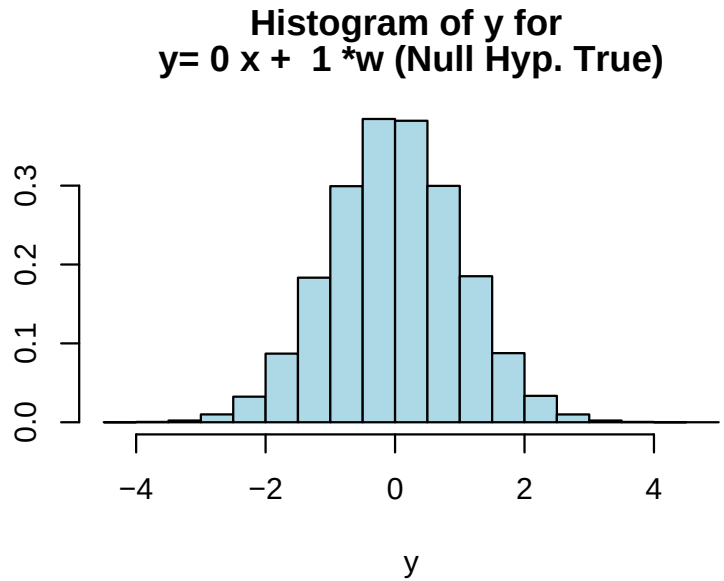
- $\mathbf{X}^F$: predictors of the full model $\{\mathbf{x}_1, \ldots, \mathbf{x}_{M_F}\}$.
- $\mathbf{X}^R$: predictors of the reduced model $\{\mathbf{x}_1, \ldots, \mathbf{x}_{M_R}\}$.
- M_F: Number of predictors in full model.
- M_R: Number of predictors in reduced model.
- N: Sample size.

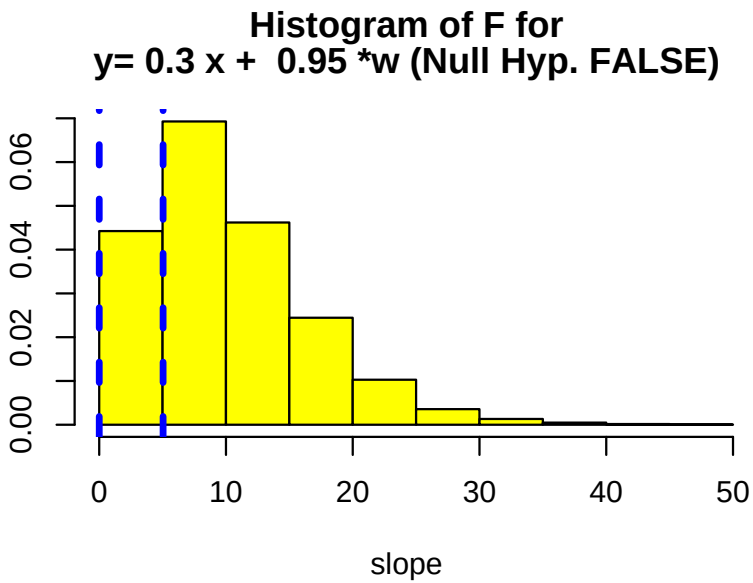
If $\beta_{M_R+1} = \cdots = \beta_{M_F} = 0$, then

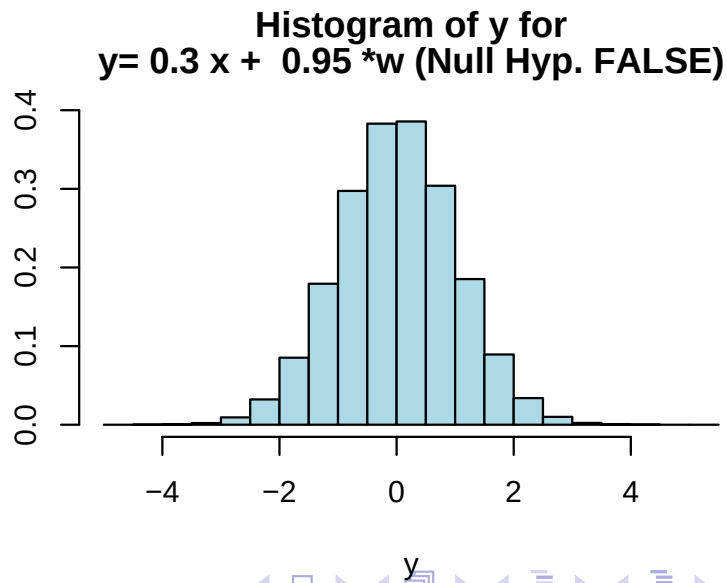
$$\frac{SSE_R - SSE_F}{SSE_F} \frac{N - M}{(N - M_R) - (N - M_F)} \sim F_{M-M_R, N-M_F}$$

Example F-Test









Formal Hypothesis Test

Figure: Ronald Fischer

In considering the appropriateness of any proposed experimental design, it is always needful to forecast all possible results of the experiment, and to have decided without ambiguity what interpretation shall be placed upon each one of them. *Ronald Fisher*

Formal Hypothesis Test Procedure

Decision rule– a rule that completely describes our decision to accept or reject the null hypothesis for every possible observation.

Acceptance and Rejection Regions

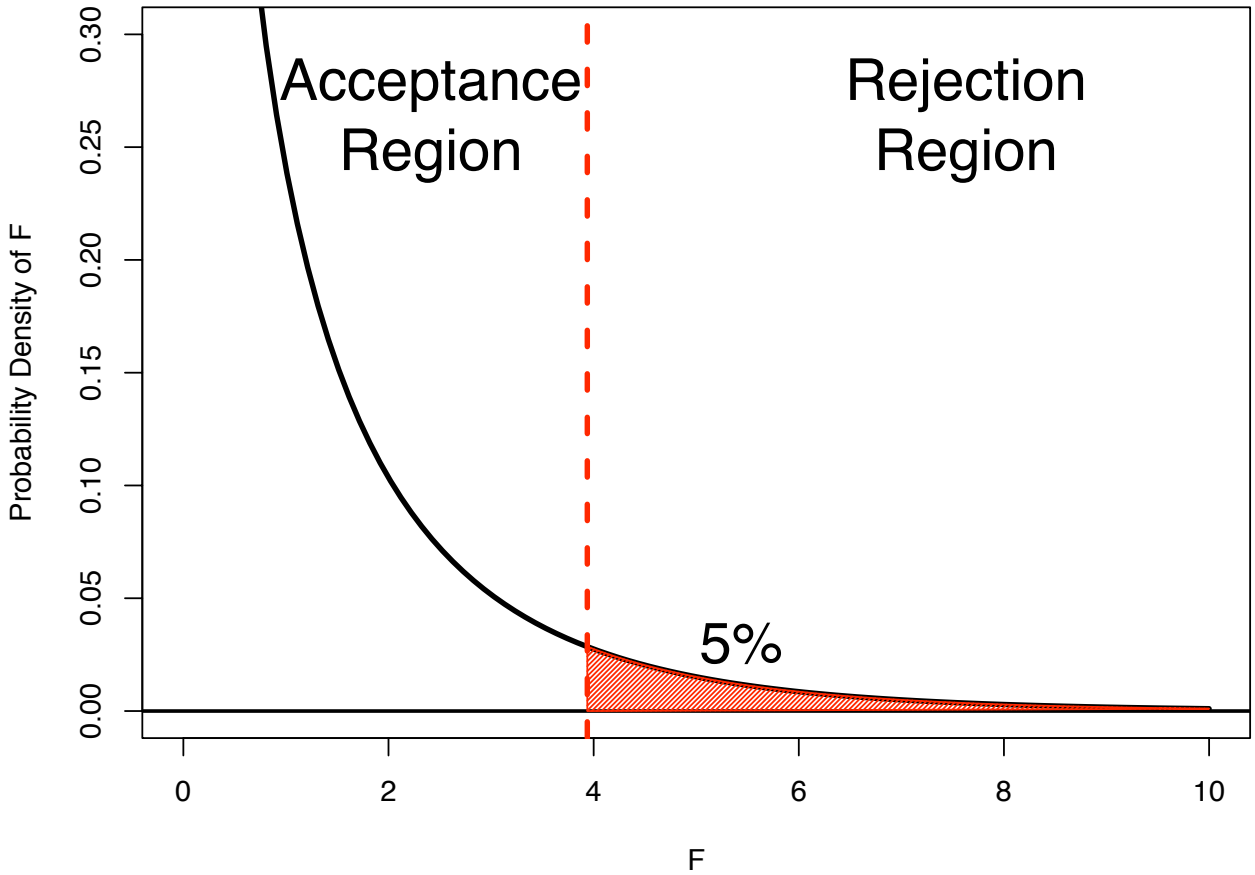
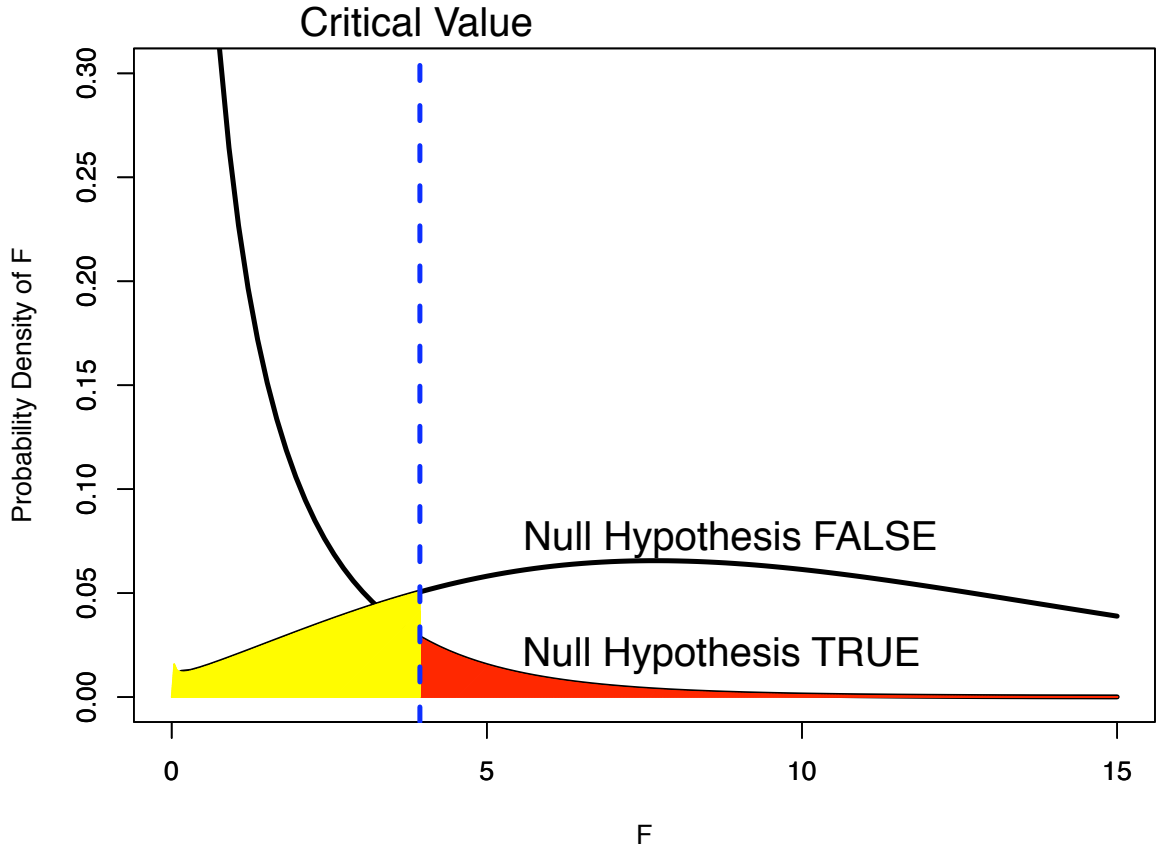


Figure: Acceptance and rejection regions for an F distribution.

Significance Level

The probability of rejecting the null hypothesis when it is true is called the significance level. Above, the significance level is 5%.

Errors in Hypothesis Testing



Properties of the Decision Rule

- Reject null hypothesis when it is true: 5% error rate
- Accept null hypothesis when it is false: 12% error rate

In general, it is not possible to reduce both errors simultaneously.

General Hypothesis Test

Assessing whether x and y are related can be interpreting as testing two hypotheses of the model

$$y = ax + b + \epsilon$$

H_0: Null Hypothesis: $a = 0$

H_1: Alternative Hypothesis: $a \neq 0$

For two hypotheses, there are two types of errors:

"False Alarm": Decide H_1 when H_0 is true (prob = significance)

"Miss": Decide H_0 when H_1 is true. (prob = 1 - power)

General Hypothesis Test

Assessing whether x and y are related can be interpreting as testing two hypotheses of the model

$$y = ax + b + \epsilon$$

H_0: Null Hypothesis: $a = 0$

H_1: Alternative Hypothesis: $a \neq 0$

For two hypotheses, there are two types of errors:

"False Alarm": Decide H_1 when H_0 is true (prob = significance)

"Miss": Decide H_0 when H_1 is true. (prob = 1 - power)

In general, it is not possible to reduce both errors simultaneously.

The fundamental theorem of linear regression gives the most powerful decision rule for given significance level.

Consider the model

$$\mathbf{y} = \mathbf{x}_1\boldsymbol{\beta}_1 + \cdots + \mathbf{x}_{M_R}\boldsymbol{\beta}_{M_R} + \cdots + \mathbf{x}_{M_F}\boldsymbol{\beta}_{M_F} + \boldsymbol{\epsilon}$$

To test hypothesis $\beta_{M_R+1} = \cdots = \beta_{M_F} = 0$, use the statistic

$$\frac{SSE_R - SSE_F}{SSE_F} \frac{N - M}{(N - M_R) - (N - M_F)} \sim F_{M-M_R, N-M_F}$$

Practical Advice About Testing Hypotheses

I used to do all statistical analyses with FORTRAN– using codes that I wrote myself.

Vast majority of statistical researchers use numerical packages, e.g., MATLAB, R, S, SAS.

Important life lesson: I have wasted more time trying to do statistics in FORTRAN than I spent learning a new statistical package. Numerical packages simplify statistical analysis so much that the time needed to learn them is well worth the time.

What is R?

R is a free, interactive statistical computing package.

- R is a language: you can program your own methods.
- R is free, in contrast to MATLAB or SAS.
- R is interactive, in contrast to FORTRAN.
- R is popular among researchers. Every major statistical computation is available in packages.
- R has an extensive website (www.r-project.org).
- R has an extensive development community.

But, there are some downsides:

- R requires some time and effort to learn.
- Maps are harder to plot than in other packages (e.g., GrADS).

Manuals and Documentation

- Documentation of an R command, e.g., "mean", can be obtained by typing "help(mean)" or "?mean" .
- Important manuals downloaded free at www.r-project.org
- *Introduction to R* is very good and painless. The first 7 chapters (32 pages) are essential. Download from http://cran.r-project.org/manuals.html
- The appendix of *Introduction to R* has an example session that is very useful for first timers.
- Other important sites:
  - www.r-project.org/search.html.
  - wiki.r-project.org
  - tolstoy.newcastle.edu.au/R/
  - cran.r-project.org
  - www.dangoldstein.com/search_r.html

Testing Independence of ISMR and AMJ-NINO3.4 in "R"

```
> lmodel.out = lm(ismr ~ nino34)
> summary(lmodel.out)

Call:
lm(formula = ISMR ~ NINO34)

Residuals:
     Min       1Q   Median       3Q      Max
-168.201  -51.544    7.734   49.975  180.565

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    855.00      15.66  54.609   <2e-16 ***
nino34         -48.34      29.19  -1.656    0.108
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 86.91 on 29 degrees of freedom
Multiple R-squared: 0.0864, Adjusted R-squared: 0.0549
F-statistic: 2.743 on 1 and 29 DF,  p-value: 0.1085
```

Interpretation of the Summary of lm

Residuals: Useful for checking "outliers."

Coefficients: Estimates of the regression parameters β, their standard errors, t-value and p-value for significance. Statistically significant coefficients ($pval < 0.05$) are indicated by asterisks symbols.

Residual standard error: Standard error of regression equation, equal to `sqrt(deviance(lmodel.out)/df.residual(lmodel.out))`

Multiple R-Squared: Means that 8.6% of the total variability is due to the linear association between the variables.

F-statistic p-value: P-value for the test of this model versus the model with only the intercept.

Conclusion Regarding Indian Monsoon Rainfall and ENSO

- "$P(>|t|)$" summarizes F-test that ENSO coefficient vanishes.
- This column shows that $p(F > 2.743) = 10.8\%$.
- In general, probability should be less than 5% to be rejected.
- Thus, we cannot reject hypothesis ENSO coefficient vanishes.

We conclude that the ENSO coefficient for fitting

$$ISMR = a * ENSO + b + \epsilon$$

is not large enough to decide that ENSO and ISMR are related.

There is no detectable ENSO-ISMR relation.

Correlation Coefficient

Another way to quantify the degree to which two variables are related is to consider the alternative statistic

$$\rho^2 = \frac{SSE_R - SSE_F}{SSE_R}$$

Since $SSE_R \geq SSE_F$, this ratio is always between 0-1.

If extra predictors are independent, $\rho = 0$. If the extra predictors completely fit the data, then $\rho = 1$.

This ratio is called the squared correlation coefficient.

Testing Significance of a Correlation Coefficient in "R"

```
> cor.test(ISMR , NINO34)

Pearson's product-moment correlation

data:  ismr and nino34
t = -1.6561, df = 29, p-value = 0.1085
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.58713156  0.06741702
sample estimates:
      cor
-0.293945
```

Equivalent Hypothesis Tests

Note that F and ρ^2 depend only on the ratio

$$SSE_R/SSE_F$$

so testing whether a predictor has vanishing coefficient is equivalent to testing whether the correlation coefficient vanishes.

Multiple Correlation Coefficient

We now consider a more complicated question: is y independent of the **joint** set $x_1, x_2, \ldots, x_K$.

This problem is not fundamentally different than testing the simple correlation. In fact, all we do is consider the two models

$$\begin{aligned} \text{Full } \mathbf{y} &= \mathbf{x}_1\beta_1 + \cdots + \mathbf{x}_K\beta_K + b + \boldsymbol{\epsilon} \\ \text{Reduced } \mathbf{y} &= b' + \boldsymbol{\epsilon} \end{aligned}$$

And then evaluate the statistic in the fundamental theorem. Moreover, we can define the quantity

$$R^2 = \frac{SSE_R - SSE_F}{SSE_R}$$

for this model, which is called the multiple correlation coefficient. R^2 is a natural generalization of correlation to multiple variables.

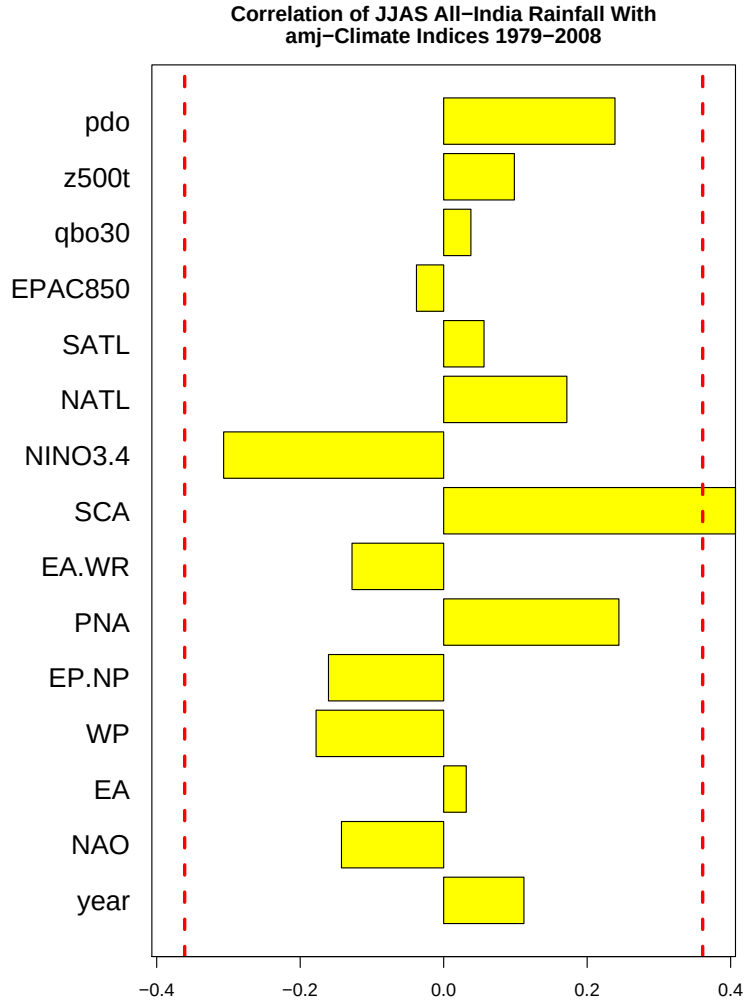


Figure: Correlation between JJAS All-India Monsoon Rainfall and various indices of seasonal variability. The 5% significance level for pair-wise comparison also is shown as the red dashed line.

Test Hypothesis That All Coefficients Vanish

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -16671.897   8192.566  -2.035   0.0612 .
xyear            8.804      4.121   2.137   0.0508 .
xNAO           -15.351     29.796  -0.515   0.6145
xEA            -37.655     30.900  -1.219   0.2431
xWP             56.643     44.861   1.263   0.2273
xEP.NP         -93.421     55.628  -1.679   0.1152
xPNA            18.757     34.156   0.549   0.5915
xEA.WR          -4.303     39.243  -0.110   0.9142
xSCA            28.581     37.915   0.754   0.4634
xNINO3.4       -77.229     38.748  -1.993   0.0661 .
xNATL          -70.167     67.269  -1.043   0.3146
xSATL         -106.613     72.611  -1.468   0.1641
xEPAC850        15.575     27.111   0.574   0.5748
xqbo30          10.908     19.580   0.557   0.5863
xz500t          23.150     19.478   1.189   0.2544
xpdo            41.566     25.116   1.655   0.1202
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 77.64 on 14 degrees of freedom
Multiple R-squared: 0.5292, Adjusted R-squared: 0.0248
F-statistic: 1.049 on 15 and 14 DF,  p-value: 0.4667
```

Conclusion Based on Multiple Regression

- "$P(>|t|)$" tests whether the individual coefficient vanishes.
- None of the column entries is less than 5%.

No detectable relation between ISMR and the other variables.