# CHAPTER 5: CALIBRATING A SNOW GAUGE

math 189 : data analysis and inference : winter 2018

Jelena Bradic
http://www.math.ucsd.edu/~jbradic/
Assistant Professor, Department of Mathematics, University of California, San Diego
jbradic@ucsd.edu

# Introduction

## The data

## Background

## Investigations


Jelena Bradic http://www.math.ucsd.edu/~jbradic/ Assistant Professor Department of Mathematics, University of California, San Diego jbradic@ucsd

Project 4

## Introduction:

**Topic:**

Snow Gauge : measuring the density of snow ❄ ???? ⟷ radio active source decay (signal emissions)

⟷ predict flooding

**Goal:** Develop a procedure to calibrate the snow gauge by predicting the snow density.

## Data:

Polyethylene blocks ⇒ simulate snow . 30 measurements ⇒ 10 reported measurements of same thing

⇒ amplified version of the gamma photon count

⇒ predict snow density from the "gain"

* multiple x for one y. (when plotting)
* Use the blue dot to predict everything else
* Use regression to predict rather than making causual relationship.
* Not independent observations ⇒ look at the simulated data

## Background

x snow density
m "gain"
} ⇒ exponentialize them

## Investigation

- Fitting

  Whether the regression model is appropriate
- Predicting
- Cross-Validation

DUE

At The End of

WK 7

# Introduction

* Main source of Water for Northern California comes from the Sierra Nevada mountains.
* To help monitor the water supply, the Forest Service of the United States Department of Agriculture (USDA) operates a gamma transmission snow gauge in the Central Sierra Nevada near Soda Springs, CA. The gauge is used to determine a death profile of snow density.

# Snow gauge

* The snow gauge does not disturb the snow in the measurement process, which means the same snow-pack can be measured over and over again. With this replicate measurements on the same volume of snow, researchers can study snow-pack settlement over the course of the winter season and the dynamics of rain on snow.
* When rain falls on snow, the snow absorbs the water up to a certain point, after which flooding occurs. The denser the snow pack the less water it can absorb.
* Analysis the snow pack profile may help with monitoring the water supply and flood management.

# Snow gauge(cont.)

* The gauge does not directly measure snow density. The density reading is converted from a measurement of gamma ray emissions.
* Due to instrument wear and radioactive source decay, there may be changes over the seasons in the functions used to cover the measured values into density readings.
* To adjust the conversion method, a calibration run is made each year at the beginning of the winter season.
* In this lab we will develop a procedure to calibrate the snow gauge.

Jelena Bradic http://www.math.ucsd.edu/~jbradic/ Assistant Professor Department of Mathematics, University of California, San Diego jbradic@ucsd

# Description

* The data are from a calibration run of the USDA Forest Service's snow gauge located in the Central Sierra Nevada mountain range near Soda Springs.
* The run consists of placing polyethylene blocks of known densities between the two poles of the snow gauge and taking readings on the blocks. The polyethylene blocks are used to simulated snow.
* For each polyethylene blocks, 30 measurements are taken. Only the middle 10 are reported here.
* The measurement reported are amplified version of the gamma photon count made by the detector. We call the gauge measurement the "gain".
* The data available here consists of 10 measurements for each of 9 densities in grams per cubic centimeter of polyethylene.

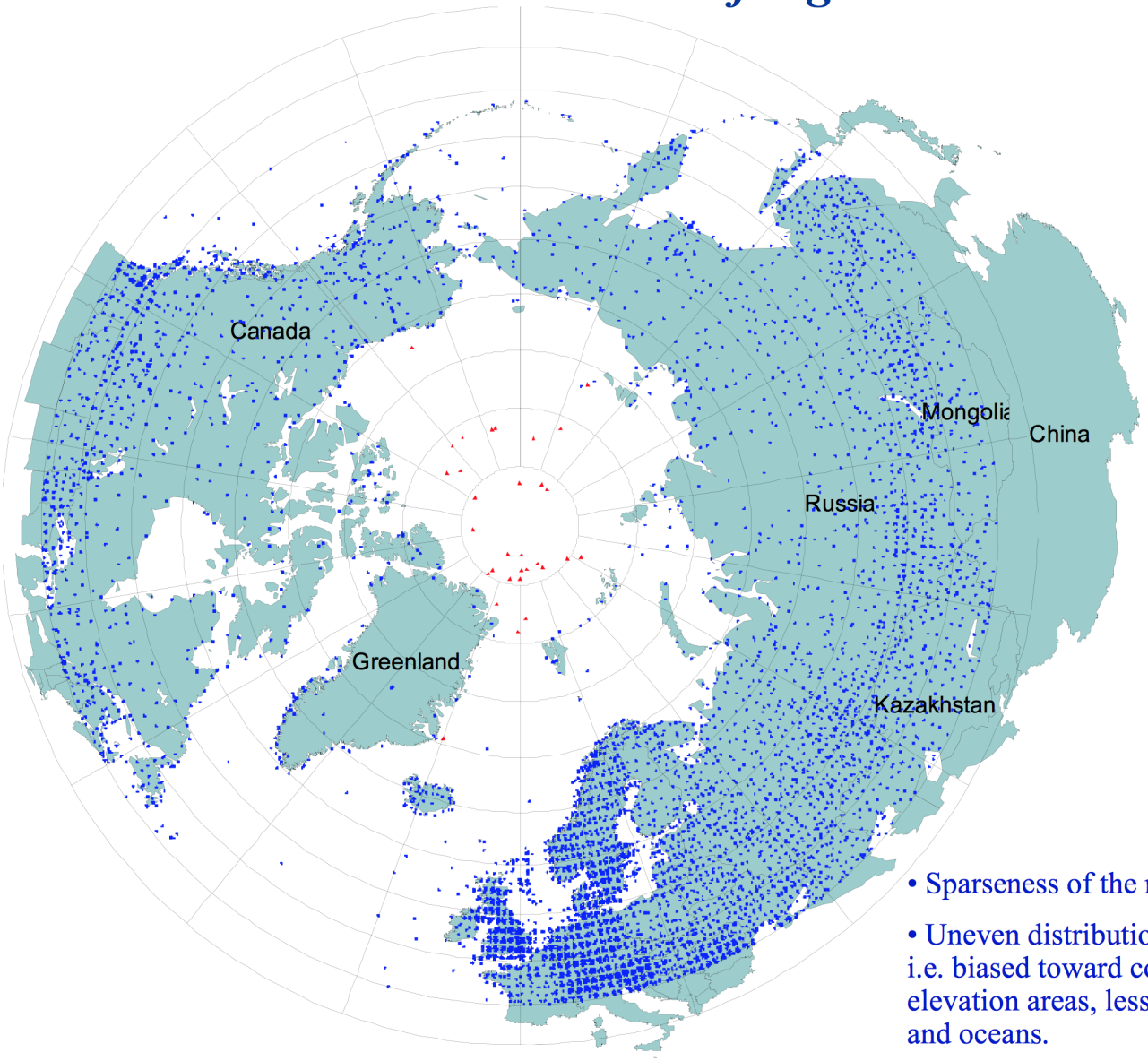
# The Data

| density | gain |
|---|---|
| 0.6860 | 17.60 |
| 0.6860 | 17.30 |
| 0.6860 | 16.90 |
| 0.6860 | 16.20 |
| 0.6860 | 17.10 |
| 0.6860 | 18.50 |
| 0.6860 | 18.70 |
| 0.6860 | 17.40 |
| 0.6860 | 18.60 |
| 0.6860 | 16.80 |
| 0.6040 | 24.80 |
| 0.6040 | 25.90 |
| 0.6040 | 26.30 |
| 0.6040 | 24.80 |
| 0.6040 | 24.80 |
| 0.6040 | 27.60 |
| 0.6040 | 28.50 |
| 0.6040 | 30.50 |
| 0.6040 | 28.40 |
| 0.6040 | 27.70 |
| 0.5080 | 39.40 |
| 0.5080 | 37.60 |
| 0.5080 | 38.10 |
| 0.5080 | 37.70 |
| 0.5080 | 36.30 |
| 0.5080 | 38.70 |
| 0.5080 | 39.40 |
| 0.5080 | 38.80 |
| 0.5080 | 39.20 |
| 0.5080 | 40.30 |
| 0.4120 | 60.00 |

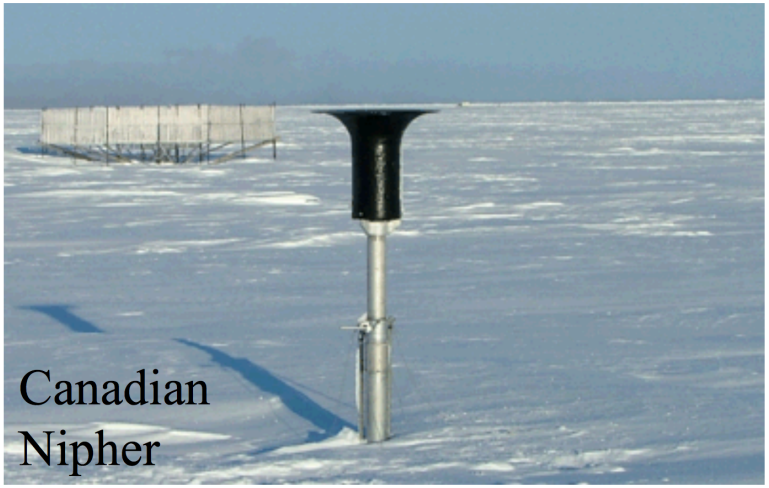
# Challenges with the Data

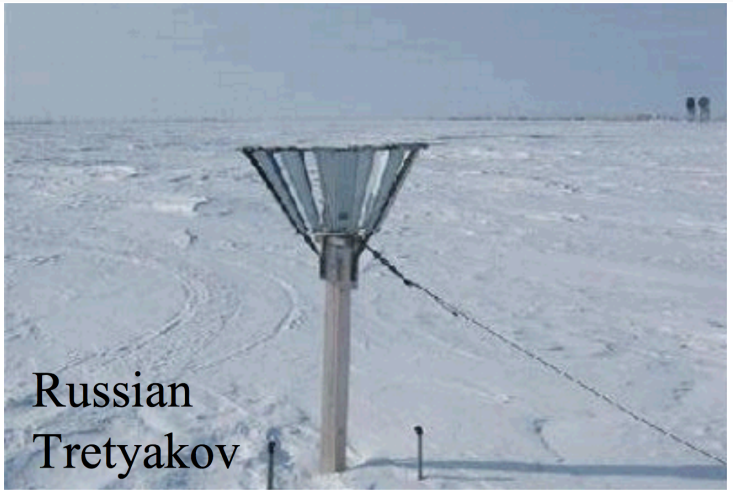
* Operational networks - our knowledge base
  - Decline of the networks in the northern regions, including Siberia, Alaska and N. Canada
  - Few stations in the mountain regions
  - How to sustain and improve the operational networks
* Data quality and compatibility across national boundaries
  - Large biases in gauge measurements of solid precipitation
  - Incompatibility of precipitation data due to difference in instruments and methods of data processing
  - Difficulties to determine precipitation changes in the arctic regions
* Validation of precipitation data, including satellite and reanalysis products and fused products at high latitudes.

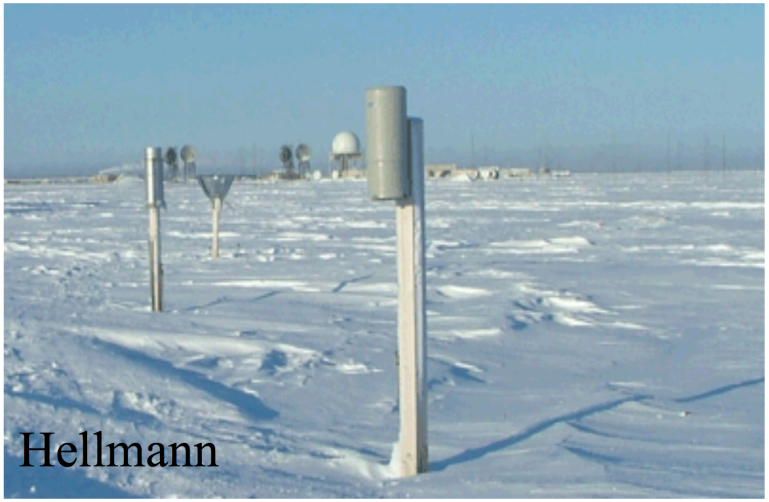
# *Synoptic/climate stations on land above 45°N and the Arctic Ocean drifting stations*

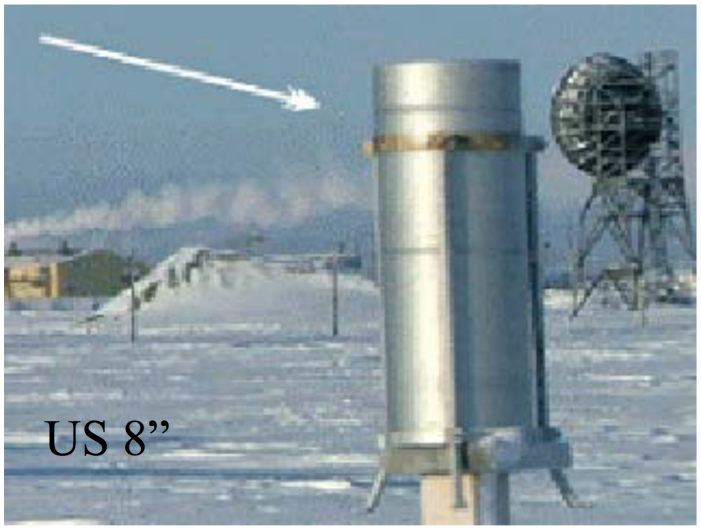



- Sparseness of the networks.
- Uneven distribution of measurement sites, i.e. biased toward coastal and the low-elevation areas, less stations over mountains and oceans.


Canadian
Nipher


Russian
Tretyakov


Hellmann


US 8”


Barrow, CRN- modified DFIR, at 2.5m, Mar 3/03


Barrow, CRN- DFIR, Mar 3/03


Barrow, UAF Wyoming snow fence, Mar/03


Barrow, UAF DFIR, Mar 03

Jelena Bradic http://www.math.ucsd.edu/~jbradic/ Assistant Professor Department of Mathematics, University of California, San Diego jbradic@ucsd

# Location

- The snow gauge is a complex and expensive instrument. It is not feasible to establish a broad network of gauges in the watershed area in order to monitor the water supply. Instead the gauge is primarily used as a research tool.

- The snow gauge has helped to study snow-pack settling, snow-melt runoff, avalanches and rain-on-snow dynamics.

- Gauges exist in Idaho, Colorado, Alaska, Russia, Mongolia, China, Japan .....

- Gauge in California is located in the center of a forest opening that is roughly 62 meters in diameter. The laboratory site is at 2099 meters elevation and is subject to all major high attitude storms which regularly deposit 5-20 cm of wet snow. The snow pack reaches an average depth of 4m each winter.

# Gauge

The snow gauge consists of a cesium-137 radioactive source and an anergy detector mounted on separate vertical poles approximately 70cm apart.

- The lift mechanism at the top of the poles raises and lowers the source and detector together. The radioactive source emits gamma photons also called gamma rays at 662 kilo-electron-volts (keV) in all directions. The detector contains a scintillation crystal which counts those photons eating through the 70-cm gap from the source to the detector crystal.

- The pulses generated by the photons that reach the deter crystal are transmitted by a cable to a preamplifier and then further amplified and transmitted via a buried coaxial cable to the lab. There the signal is stabilized, corrected for temperature drift, and converted to a measurement we have termed the ”gain”. It should be directly proportional to the emission rate.

- The snow pack density typically ranges between 0.1 and 0.6 $g/cm^3$.

# A Physical Model

- The gamma rays that are emitted from the radioactive source are sent out in all directions. Those that are sent in the direction of the detector may be scattered or absorbed by the polyethylene molecules between the source and the detector. With denser polyethelene, fewer gamma rats will reach the detector.

- There are complex physical models for the relationship between the polyethylene density and the detector readings.

- A simplified version of the model that may be workable for the calibration problem of interest is described here. A gamma ray on route to the detector passes a number of polyethylene molecules.The number of molecules depends on the density of the polyethylene. A molecule may either absorb the gamma photon bounce it out of the path to the detector or allow it to pass.

# A Physical Model

- If each molecule, acts independently, then the chance that a gamma ray successfully arrives at the detector is $p^m$ where p is the chance , a single molecule will neither absorb nor bounce the gamma ray, and m is the number of molecules in a straight line path from the source to the detector.

-Thsi probability can be re-expressed as

$$e^{m \log p} = e^{bx}$$

where x the density, is proportional to m the number of molecules.

Jelena Bradic http://www.math.ucsd.edu/~jbradic/ Assistant Professor Department of Mathematics, University of California, San Diego jbradic@ucsd

The aim of this lab is to provide a simple procedure for converting gain into density when the gauge is in operation. Keep in mind that the experiment was conducted by varying density and measuring the response in gain, but when the gauge is ultimately in use, the snow-pack density is to be estimated from the measured gain.

* [**Fitting**] Use the data to fit the gain, or a transformation of gain, to density. Try sketching the least squares line on a scatter plot.
  ** Do the residuals indicate any problems with the fit ?
  ** If the densities f the polyethylene blocks are not reported exactly, how might this affect the fit ?
  ** What if the blocs of polyethylene were not measured in random order?
* [**Predicting**] Ultimately we are interested in answering questions such as: Given a gain reading of 38.6, what is the density of the snow-pack ? or Given a gain reading of 426.7, what is the density of snow-pack? These two numeric values, 38.6 and 426.7, were chosen because they are the average gains for the 0.508 and 0.001 densities, respectively.
  ** Develop a procedure for adding bands around your least squares line that can be used to make interval estimates for the snow-pack density from gain measurements. Keep in mind how the data were collected: several measurements of gain were taken for polyenthylene blocks of known density.

## How do we find clusters of palindromes? How do we determine whether a cluster is just a chance occurrence or a potential replication site?

* [**Cross-Validation**] To check how well your procedure works, omit the set of measurements corresponding to the block of density 0.508, apply your ”estimation”/calibration procedure to the remaining data, and provide an interval estimate for the density of a block with an average reading of 38.6. Where does the actual density fall in the interval? Try the same test, for the set f measurements at the 0.001 density.

Can you make your code available online ?

# Scenario 1:

① Do the residuals indicate any problems with the fit?
② If densities (y) is not reported exactly, how might this affected the fit?
③ What if the blocks were not measured in a random order?

Regression:

① Draw Scatterplot: ⇒ Identify which one is response variable, which one is explanatory variable.
+ Residual Plot — x-axis ↑ y-axis ↑

⇒ Analyze relationship: sign: positive / negative
shape: linear / quadratic / etc.
level of correlation: strong / moderately strong / weak

② Calculate Correlation: Def: strength of the linear association between two variables. $[-1, 1]$

Population mean: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1 + \cdots + X_n}{n}$

Population variance: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

Population correlation: R (empirical Correlation Coefficient)

$$= \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i \cdot Y_i) - (\bar{X}\cdot\bar{Y})}{S_X \cdot S_Y}$$

slope of regression: $\hat{\beta}_1 = \frac{S_y}{S_x} \cdot R$

Interpretation: For each additional % point in explanatory variable (x), we would expect the % point in response variable (y) increase on average $\hat{\beta}_1$ points.

Intercept: $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

Intuition: A regression line always passed through $(\bar{X}, \bar{Y})$

Interpretation: With no explanatory variable (x), we expect on average to have $\hat{\beta}_0$ of response variable (y).

③ Fitting:

⇒ Residual: distance between observed and predicted

Goal: We want a line with small residuals

Formula: $\hat{\varepsilon}_i = Y_i - \hat{\beta}_1 X_i - \hat{\beta}_0$

$(\varepsilon_i = Y_i - \hat{Y}_i)$

⇒ least square: Def: = Minimizes the sum of the squares of data point to regression line.

formula = $\text{Min}\sum_{i=1}^{n}(\text{Residuals}_i)^2$

$= \min_{\beta_0 \beta_1}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2$

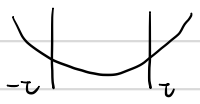
Why Least Square: ① most commonly used
② Easier to compute
③ A residual twice as large as another is usually more than twice as bad.

⇒ least absolute deviation: Def: Minimizes the sum of the data point to regression line.

formula = $\text{Min}\sum_{i=1}^{n}|\text{Residual}_i|$

$= \min_{\beta_0, \beta_1}\sum_{i=1}^{n}|Y_i - \beta_0 - \beta_1 X_i|$

⇒ $\sum_{i=1}^{n} h(\text{residuals}_i)$, $h(\text{residuals}_i) = \begin{cases} |x| & |x| > \tau \\ x^2 & |x| < \tau \end{cases}$




$\tau > 0$

Note: The statements are not causal, unless the study is a randomized controlled experiment.

# Scenario 2 :

① adding bands around least square line ⇒ make estimates (Prediction Interval + Confidence Interval)

Note : several measurement of gain was taken for one density

④ Predicting :

* Fitted Value ⟺ Old Value

  New Value ⟺ New Value

* Definition :

  ⇒ Prediction : Give explanatory variable (x) to predict response variable (y) (Plug in x to the model)

  ⇒ Extrapolation : Apply a model estimate to values outside of the realm of the original data ( $x=0$ ) ← unreachable data

* Least Square : $\varepsilon_i \sim N(0, \sigma^2)$ $\hat{\beta_1}$ is MLE

  Least Absolute Value : $\varepsilon_i \sim$ expl ) $\hat{\beta_1}$ is MLE

* Requirement :

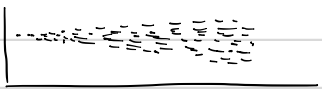
  ① Plot Residuals + Scatterplot, check linearity

  ② Plot QQ-plot + residuals histogram, Check if residuals plot ~ Normal.

  ③ If use scatterplot to plot residuals, should only see scatter plots without clear pattern. ( x ⇒ explanatory variable. y ⇒ residuals )

  [homoscedasticity] ⇒ constant variability

  [13.1] Residual Plot

   ← violation ?

* $R^2$ :

  ⇒ Definition : The strength of the fit of a linear model

  ⇒ formula : $= \left( \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i \cdot y_i) - (\bar{x} \cdot \bar{y})}{s_x \cdot s_y} \right)^2$ ⇒ smaller than R

  ⇒ Interpretation : [what percent of variability in the response variable is explained by the model. The remainder of the variability is explained by variables not included in the model / due to randomness in data]

  $R^2 \times 100\%$ of the variability in the response variable is explained by the model.

  ⇒ Adjusted $R^2$ : ( If we have more than 2 variables ⇒ gives more accurate interpretation)

  * Categorical Data :

    ↳ reference level = (categorical data = 0) = ( y = intercept)

    ↳ Use LS to test if X and Y are independent

    $H_0$ : $\beta_1 = 0$ $H_1$ : $\beta_1 \neq 0$

Which region (northeast, midwest, west, or south) has the lowest poverty percentage?

$H_0: \beta_1 = 0$ $H_1: \beta_1 \neq 0$ Hypothesis testing

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 9.50 | 0.87 | 10.94 | 0.00 |
| region4midwest | 0.03 | 1.15 | 0.02 | 0.98 |
| region4west | 1.79 | 1.13 | 1.59 | 0.12 |
| region4south | 4.16 | 1.07 | 3.87 | 0.00 |

(Intercept) ← 9.50 : $\hat{\beta}_0$

region4midwest, region4west, region4south ← slope based on reference level : $\hat{\beta}_1$

Std. Error of estimator

→ probability of getting non-zero $\beta_1$ value

$$SE_{\hat{\beta}} = \frac{\sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

$$\frac{|\hat{\beta} - \beta|}{SE_{\hat{\beta}}} \sim T_{n-4}$$

$$= \frac{|\text{point estimate} - \text{null}|}{SE}$$

⑤ Outlier:

* Distinguish if that outlier is influential / high leverage

Relationship between biological IQ & foster IQ

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.20760    9.29990   0.990    0.332
bioIQ         0.90144    0.09633   9.358  1.2e-09

Residual standard error: 7.729 on 25 degrees of freedom
Multiple R-squared: 0.7779,Adjusted R-squared: 0.769
F-statistic: 87.56 on 1 and 25 DF,  p-value: 1.204e-09
```

n-2

Confidence Interval: Estimate ± SE × $t_{n-2}$