# STAT 5000

## Statistical Methods I

Week 10

Fall 2024

Dr. Danica Ommen

Unit 3

# Introduction to Simple Linear Regression (SLR)

## Research Question

- Study the relationship of two or more quantitative variables
  - quantitative: numbers, usually continuous
  - qualitative: classes, identify groups
- Is there a significant linear relationship between the response variable and the explanatory variable?
- What mean value of response would we predict for a given value of the explanatory variable?
- What value of response would we predict for a given value of the explanatory variable?

## SLR Model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n$$

- $i = 1, \ldots, n$ is the number of observations
- $Y_i$ is the *response* or dependent variable
- $X_i$ is the predictor, *explanatory variable*, or independent variable, treated as known and fixed
- $\epsilon_i$ is the *random error* term representing individual variation and measurement error

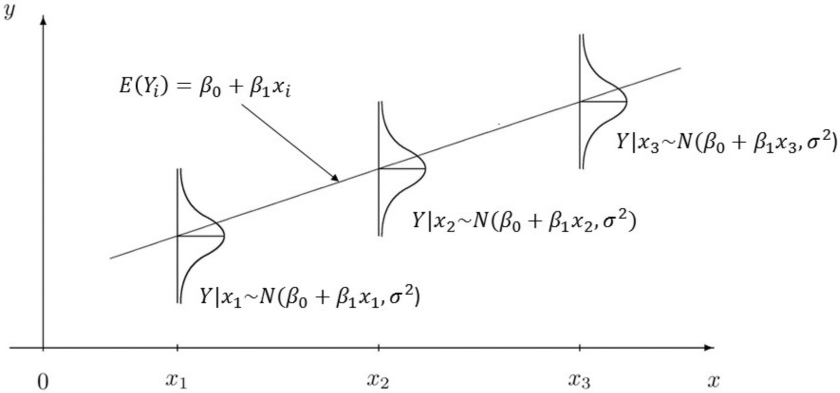
Write SLR model as a linear model $\mathbf{Y} = \mathbf{X}\beta + \epsilon$ where

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ 1 & X_3 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

**Model Assumptions**

- $x$'s are fixed (or conditioned upon)
- The expected response is a linear function of the explanatory variable : $E(Y_i|X_i = x_i) = \beta_0 + \beta_1 x_i$
- additive random errors: $Y_i = E(Y_i|X_i = x_i) + \epsilon_i$
- independent (uncorrelated) random errors
- homogeneous error variance: $Var(\epsilon_i) = \sigma^2$
- normally distributed random errors: $\epsilon_i \sim N(0, \sigma^2)$

## Model and Assumptions



## Model and Assumptions

The conditional distribution of $Y$ given that $X = x$ is

$$N(\beta_0 + \beta_1 x, \sigma^2)$$

- $\beta_1 =$ slope, is the change in the conditional mean of $Y$ for a one unit increase in $x$
- $\beta_0$ is the conditional mean of $Y$ when $X = 0$
- If we replace $x$ by $x - x_0$ to obtain $Y = \beta_0 + \beta_1(x - x_0) + \epsilon$, then $\beta_0$ is the conditional mean of $Y$ when $X = x_0$
- $\sigma^2$ is the variation of responses about the conditional mean for any specific value of the explanatory variable

## Relationship to ANOVA

- ANOVA: each group (each level of explanatory variable) has its own mean
- Each $x_i$ in regression defines its own group, but...
  - too many groups with too few observations per group
  - Linear regression analysis makes stronger assumption about the means (linear structure)

### A bit of history

Sir Francis Galton coined the term “regression”

- biometrician, geneticist, 1870-1920s
- compared the heights of children to their parents
- parents and children had similar means
- short parents had short children, tall parents had tall children
- children were closer to average than their parents
- “regression” to the mean

## Least Squares Estimation

Use data $(Y_i, x_i), i = 1, 2, \cdots, n$ to estimate the regression coefficients in the model $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$

- Choose estimates $b_o$ and $b_1$ to minimize

$$g(b_o, b_1) = \sum_{i=1}^{n} [Y_i - (b_o + b_1 x_i)]^2$$

- Why squared errors?
  - Tradition (Gauss invented least squares estimation)
  - Equivalent to maximum likelihood estimation when errors are independent and normally distributed with constant variance

## Least Squares Estimates

$$
\begin{aligned}
b_o &= \bar{Y} - b_1\bar{x} \\
b_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}
\end{aligned}
$$

- These are best linear unbiased estimators (blue)
- Predicted (fitted) values: $\hat{Y}_i = b_o + b_1 x_i$
- Residuals: $e_i = Y_i - \hat{Y}_i$

## Least Squares Estimation

- Choose $b_o, b_1$ to minimize $g(b_o, b_1) = \sum_{i=1}^{n}(Y_i - (b_o + b_1 x_i))^2$
- Taking derivatives and setting them equal to zero yields the normal equations

$$\begin{aligned} b_o n + b_1 \sum x_i &= \sum Y_i \\ b_o \sum x_i + b_1 \sum x_i^2 &= \sum x_i Y_i \end{aligned}$$

- The normal equations can also be written as

$$\begin{aligned} \sum e_i &= \sum (Y_i - (b_o + b_1 x_i)) = 0 \\ \sum x_i e_i &= \sum x_i (Y_i - (b_o + b_1 x_i)) = 0 \end{aligned}$$

## Least Squares Estimation

Normal equations can be written in matrix form

$$\begin{bmatrix} b_o n + b_1 \sum x_i \\ b_o \sum x_i + b_1 \sum x_i^2 \end{bmatrix} = \begin{bmatrix} \sum Y_i \\ \sum x_i Y_i \end{bmatrix}$$

that is equivalent to

$$\begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} b_o \\ b_1 \end{bmatrix} = \begin{bmatrix} \sum Y_i \\ \sum x_i Y_i \end{bmatrix}$$

and can be written as $X^T X \mathbf{b} = X^T \mathbf{Y}$

where $\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$ $\quad \mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$ $\quad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$

**Least Squares Estimation**

Solution to the normal equations

$$\begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = (X^T X)^{-1} X^T Y = \begin{bmatrix} \bar{Y} - b_1 \bar{x} \\ \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \end{bmatrix}$$

## Least Squares Estimation

Variance-covariance matrix of the least squares estimator:

$$\begin{aligned}
Var\begin{bmatrix} b_o \\ b_1 \end{bmatrix} &= \begin{bmatrix} Var(b_o) & Cov(b_o, b_1) \\ Cov(b_o, b_1) & Var(b_1) \end{bmatrix} \\
&= Var\left(\left(X^TX\right)^{-1}X^TY\right) \\
&= \left(X^TX\right)^{-1}X^TVar(Y)\left[\left(X^TX\right)^{-1}X^T\right]^T \\
&= \left(X^TX\right)^{-1}X^T\ \left[\sigma^2 I\right]\ X\left(X^TX\right)^{-1} \\
&= \sigma^2\left(X^TX\right)^{-1}X^TX\left(X^TX\right)^{-1} \\
&= \sigma^2\left(X^TX\right)^{-1}
\end{aligned}$$

**Least Squares Estimation**

The the variance-covariance matrix of the least squares estimator for the regression coefficients is

$$
\begin{aligned}
Var\begin{bmatrix} b_o \\ b_1 \end{bmatrix} &= \begin{bmatrix} Var(b_o) & Cov(b_o, b_1) \\ Cov(b_o, b_1) & Var(b_1) \end{bmatrix} \\
&= \sigma^2 \left(X^T X\right)^{-1} \\
&= \sigma^2 \begin{bmatrix} \frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2} & \frac{-\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \frac{-\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} & \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{bmatrix}
\end{aligned}
$$

## Least Squares Estimation

- Matrix of second partial derivatives of $g(b_o, b_1)$

$$\begin{bmatrix} \frac{\partial^2 g(b_o,b_1)}{\partial b_o^2} & \frac{\partial^2 g(b_o,b_1)}{\partial b_o \partial b_1} \\ \frac{\partial^2 g(b_o,b_1)}{\partial b_o \partial b_1} & \frac{\partial^2 g(b_o,b_1)}{\partial b_1^2} \end{bmatrix} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} = X^T X$$

  Since this matrix is positive definite (if we have at least two different $x_i$ values), it guarantees we have a minimum.
- Note: least squares estimate of the slope is different if there is no intercept in the model

## Definition: Multivariate Normal Distribution

Suppose $Z = \begin{bmatrix} Z_1 \\ \vdots \\ Z_m \end{bmatrix}$ is a random vector whose elements are independently distributed standard normal random variables. For any $n \times m$ matrix $A$, we say that

$$\mathbf{Y} = \boldsymbol{\mu} + A\mathbf{Z}$$

has a *multivariate normal distribution* with mean vector

$$E(Y) = E(\boldsymbol{\mu} + A\mathbf{Z}) = \boldsymbol{\mu} + AE(\mathbf{Z}) = \boldsymbol{\mu} + A\mathbf{0} = \boldsymbol{\mu}$$

and variance-covariance matrix

$$Var(\mathbf{Y}) = A[Var(\mathbf{Z})]A^T = AA^T \equiv \Sigma$$

**Multivariate Normal Distribution**

We will use the notation

$$\mathbf{Y} \sim N(\boldsymbol{\mu}, \Sigma)$$

When $\Sigma$ is positive definite, the joint density function is

$$f(\mathbf{y}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{y}-\boldsymbol{\mu})}$$

where

$$\Sigma = \begin{bmatrix} Var(Y_1) & Cov(Y_1, Y_2) & Cov(Y_1, Y_3) & \cdots & Cov(Y_1, Y_n) \\ Cov(Y_2, Y_1) & Var(Y_2) & Cov(Y_2, Y_3) & \cdots & Cov(Y_2, Y_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Cov(Y_n, Y_1) & Cov(Y_n, Y_2) & Cov(Y_n, Y_3) & \cdots & Var(Y_n) \end{bmatrix}$$

**Multivariate Normal Distribution**

The multivariate normal distribution has some useful properties. One is that normality is preserved under linear transformations:

### Multivariate Normal Linear Combinations

If $\mathbf{Y} \sim N(\boldsymbol{\mu}, \Sigma)$, then

$$\mathbf{W} = c + B\mathbf{Y} \sim N(c + B\boldsymbol{\mu}, B\Sigma B^T)$$

for any non-random $c$ and $B$.

**Predicted Values and Residuals**

- Predicted (fitted) values

$$\begin{aligned}\hat{Y}_i &= b_0 + b_1 x_i \\ \hat{\mathbf{Y}} &= X\hat{\beta}\end{aligned}$$

- Residuals

$$e_i = Y_i - \hat{Y}_i$$

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$$

Unit 3

# Inference for Simple Linear Regression (SLR)

## Regression Analysis: ANOVA

- Write the deviation from the overall sample mean as $Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$ where $\hat{Y}_i = b_o + b_1 X_i$
- Partition the corrected total sums of squares

$$\begin{aligned}
SS_{total} &= \sum_i (Y_i - \bar{Y})^2 = \sum_i (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 \\
&= \sum_i (Y_i - \hat{Y}_i)^2 + \sum_i (\hat{Y}_i - \bar{Y})^2 + 2 \sum_i (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) \\
&= \sum_i (Y_i - \hat{Y}_i)^2 + \sum_i (\hat{Y}_i - \bar{Y})^2 \\
&= SS_{residuals} + SS_{model}
\end{aligned}$$

## Regression Analysis: ANOVA

- Cross product term is

$$\begin{aligned} 2\sum_i (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) &= 2\sum_i e_i(b_0 + b_1 x_i - \bar{Y}) \\ &= 2(b_0 - \bar{Y})\sum_i e_i + 2b_1 \sum_i e_i x_i \\ &= 0 \quad \text{because } \sum_i e_i = \sum_i e_i x_i = 0 \end{aligned}$$

- Note that

$$SS_{model} = \sum_i (\hat{Y}_i - \bar{Y})^2 = \sum_i (b_0 + b_1 x_i - \bar{Y})^2 = b_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$$

## Regression Analysis: ANOVA

- $SS_{model}$ = $SS_{total}$ - $SS_{error}$
  $$= \sum_i (\hat{Y}_i - \bar{Y})^2$$
  $$= \sum_i (b_o + b_1 x_i - \bar{Y})^2$$
  $$= b_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$$
- $SS_{model}$ is also denoted by $SS_{regression}$
- $SS_{error}$ is also denoted by $SS_{residuals}$ or $SSE$

**Regression Analysis: ANOVA**

$SS_{error}$ has $n - 2$ degrees of freedom because

- Two parameters must be estimated to calculate $\hat{Y}_i$
- The residuals satisfy two constraints

$$\sum e_i = 0 \qquad \text{and} \qquad \sum e_i x_i = 0$$

**ANOVA Table**

| Source | df | Sums of Squares |
|---|---|---|
| Model | 1 | $SS_{\text{model}} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$ |
| Error | $n-2$ | $SS_{\text{error}} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ |
| Total | $n-1$ | $SS_{\text{total}} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ |

## Mean Squares

- $MS_{\text{error}}$
  - $\hat{\sigma}^2 = MS_{\text{error}} = SS_{\text{error}}/(n-2)$
  - $\hat{\sigma}^2$ is an unbiased estimate of $\sigma^2$

$$E(MS_{\text{error}}) = \sigma^2$$

- $MS_{\text{model}}$
  - $E(MS_{\text{model}}) = \sigma^2 + \beta_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$
  - When $\beta_1$=0, $E(MS_{\text{model}}) = \sigma^2$.
    Otherwise, $E(MS_{\text{model}}) > \sigma^2$.

## F-test for Significance of Model

- $H_0 : \beta_1 = 0 \rightarrow Y_i = \beta_0 + \epsilon_i$
- $H_a : \beta_1 \neq 0 \rightarrow Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
- Test Statistic:
$$F = \frac{MS_{\text{model}}}{MS_{\text{error}}}$$
- Reject $H_0$ if
$$F = \frac{MS_{\text{model}}}{MS_{\text{error}}} > F_{1,n-2,1-\alpha}$$

**Coefficient of Determination ($R^2$)**

$$R^2 = \frac{SS_{\text{model}}}{SS_{\text{total}}}$$

- Fraction of variation in the response variable that can be explained by the linear regression model with the explanatory variable $x$.
- Expressed as percentage: $0\% \leq R^2 \leq 100\%$
- Large values of $R^2$ indicate better model fit.

**Inference for Model Parameters**

- Population Slope - $\beta_1$
- Population Intercept - $\beta_0$
- Conditional Mean - $\mu_{Y|x}$

**Inference for the Slope ($\beta_1$)**

- Discuss inference for $\beta_1$ in detail (then summarize the rest)

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

- $b_1$ is a linear combination of normal random variables (the $Y_i$'s) so $b_1$ is normally distributed with

$$E(b_1) = \beta_1 \qquad \text{Var}(b_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

- $b_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)$

**Inference for the Slope ($\beta_1$)**

Examine

$$\mathrm{Var}(b_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

A more precise estimate of $\beta_1$ can be obtained by :

- Spreading out the $X$ values
- Getting a larger sample i.e. more $(X, Y)$ pairs
- Making the error variance smaller

**Inference for the Slope ($\beta_1$)**

- Use $MS_{error}$ to estimate $\sigma^2$

$$\text{(Note that } MS_{error} \sim \frac{\sigma^2 \chi^2_{n-2}}{n-2} \text{ )}$$

- Standard error of $b_1$ is $S_{b_1} = \sqrt{MS_{error} / \sum_{i=1}^{n} (x_i - \bar{x})^2}$
- $(b_1 - \beta_1)/S_{b_1}$ has a $t$-distribution with $n - 2$ d.f.

## Hypothesis Test for $\beta_1$

- Null and Alternative Hypotheses

$$H_0 : \beta_1 = 0 \qquad H_a : \beta_1 \neq 0$$

- Test Statistic

$$T = \frac{b_1 - 0}{S_{b_1}}$$

- Reject $H_0$ if $|T| > t_{n-2,1-\alpha/2}$
- Note that $T^2 = F$, this $t$-test for $\beta_1$ is the same as the F-test for significance of model from ANOVA Table.
- One-sided alternative hypothesis is possible for the $t$-test: $H_a : \beta_1 > 0$ or $H_a : \beta_1 < 0$

**Confidence Interval for $\beta_1$**

- $100(1-\alpha)\%$ confidence interval for $\beta_1$:

$$b_1 \pm t_{n-2,1-\alpha/2} S_{b_1}$$

**Inference for the Intercept ($\beta_0$)**

- $b_o = \bar{Y} - b_1\bar{x} \sim N(\beta_o, \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}))$
- $b_o$ has standard error $S_{b_o} = \sqrt{MS_{error}\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}$
- Reject $H_o : \beta_o = 0$ if $|t| = \left|\frac{b_o - 0}{S_{b_o}}\right| > t_{n-2,1-\alpha/2}$
- $100(1-\alpha)\%$ confidence interval for $\beta_o$ is

$$b_0 \pm t_{n-2,1-\alpha/2}\, S_{b_0}$$

**Inference for the Intercept ($\beta_0$)**

- Rarely considered
- Values of $x$ must be near 0 for meaningful interpretations
- Would be most likely to use confidence interval

## Inference for Conditional Means

Inference for $\mu_{Y|x} = E(Y|X = x) = \beta_o + \beta_1 x$

- Estimate is $\hat{\mu}_{Y|x} = b_o + b_1 x$
- $\hat{\mu}_{Y|x}$ is a linear function of two normally distributed random variables ($b_0$ and $b_1$, not independent)
- $\hat{\mu}_{Y|x}$ is $N\left(\beta_o + \beta_1 x, \sigma^2 \left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)\right)$
- Note: value of $x$ does not need to be present in sample.

### Inference for Conditional Means

- standard error is

$$S_{\hat{\mu}_{Y|x}} = S_{b_0+b_1x} = \sqrt{MS_{error}\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)}$$

- $100(1-\alpha)\%$ confidence interval for $\beta_0 + \beta_1 x$ is

$$(b_0 + b_1x) \pm t_{n-2,1-\alpha/2}\, S_{\hat{\mu}_{Y|x}}$$

### Confidence Region for a Line Segment

Use the Scheffe' procedure to get simultaneous confidence intervals for every x in an entire line segment:

$$(b_o + b_1 x) \pm \sqrt{2F_{2,n-2,1-\alpha}}\, S_{b_o+b_1 x}$$

for $a \leq x \leq b$

## Prediction

Predict the value for $Y$ at given $x$:

$$Y_{new} = \beta_0 + \beta_1 x + \epsilon$$

- Estimate is still $\hat{Y} = b_0 + b_1 x$
- Standard error is

$$S_{pred} = \sqrt{MS_{error}\left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}$$

- $100(1 - \alpha)\%$ prediction interval:

$$(b_0 + b_1 x) \pm t_{n-2,1-\alpha/2}\, S_{pred}$$

## Comparison

- Confidence Interval for Condition Mean $\mu_{Y|x}$
  - Inference for a point on the population regression line given value of $x$
  - Source of inference is estimating regression line
- Prediction Interval for $Y$
  - Inference for a point in the scatterplot of all population values given value of $x$.
  - Sources of inference are estimating regression line AND predicting $Y$ given the regression line.

Unit 3

# SLR: Forbes Example

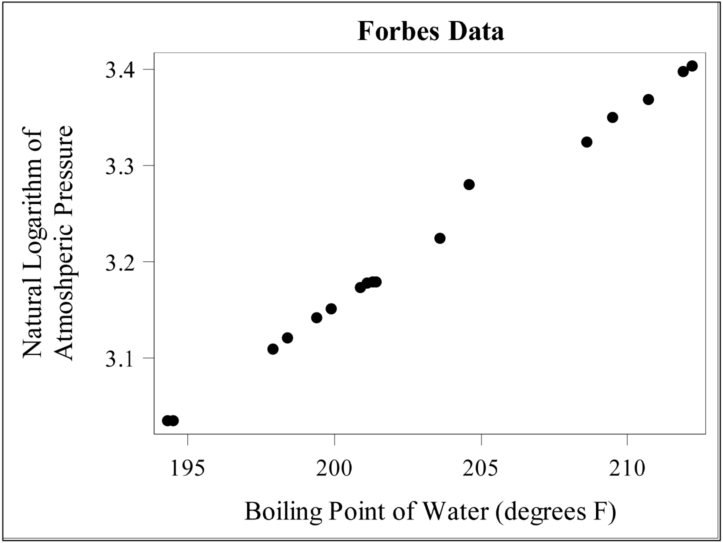
**Forbes Data**

Weisberg, Sanford, *Applied Linear Regression*, Wiley, 1980.

- James D. Forbes collected data in the mountains of Scotland
- n=17 locations (at different altitudes)
- Objective: Predict barometric pressure (in inches of mercury) from boiling point of water (X) in $^o$F.
- Use Y=log(barometric pressure)
- Motivation: Fragile barometers were difficult to transport

## Forbes Data

| Obs | Boil. Point of Water (°F) | Barametric Pressure (in Hg) | Nat.Log Barametric Pressure | Obs | Boil. Point of Water (°F) | Barametric Pressure (in Hg) | Nat.Log of Barametric Pressure |
|---|---|---|---|---|---|---|---|
| 1 | 194.3 | 20.79 | 3.03447 | 10 | 201.4 | 24.02 | 3.17889 |
| 2 | 194.5 | 20.79 | 3.03447 | 11 | 203.6 | 25.14 | 3.22446 |
| 3 | 197.9 | 22.40 | 3.10906 | 12 | 204.6 | 26.57 | 3.27978 |
| 4 | 198.4 | 22.67 | 3.12104 | 13 | 208.6 | 27.76 | 3.32360 |
| 5 | 199.4 | 23.15 | 3.14199 | 14 | 209.5 | 28.49 | 3.34955 |
| 6 | 199.9 | 23.35 | 3.15060 | 15 | 210.7 | 29.04 | 3.36867 |
| 7 | 200.9 | 23.89 | 3.17346 | 16 | 211.9 | 29.88 | 3.39719 |
| 8 | 201.1 | 23.99 | 3.17764 | 17 | 212.2 | 30.06 | 3.40320 |
| 9 | 201.3 | 24.01 | 3.17847 | | | | |


Forbes Data
Natural Logarithm of
Atmoshperic Pressure
3.1
3.2
3.3
3.4
195
200
205
210
Boiling Point of Water (degrees F)

## Analysis of the Forbes Data

- Proposed regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

  where $\epsilon_i \sim N(0,\ \sigma^2)$, $i = 1, 2, ..., 17$
- $Y_i$=log(pressure)
- $X_i$=boiling point ($^o$F)
- $\beta_1$ is the increase in mean log(pressure) when boiling point of water increases by 1 $^o$F
- $\beta_0$ is the mean log(pressure) when boiling point of water is 0 $^o$F (Is this extrapolation realistic?)

## Analysis of the Forbes Data

- Estimated regression model

$$\hat{Y} = b_0 + b_1 x = -0.97097 + 0.020623x$$

- Could have subtracted 212 $^o$F from each boiling point. Then the estimated model is

$$\begin{aligned} \hat{Y} &= b_0 + 212b_1 + b_1(x - 212) \\ &= 3.401106 + 0.020623(x - 212) \end{aligned}$$

- Then 3.401106 is an estimate of the mean log(pressure) at 212 $^o$F.

## Predicted Values

$$\hat{Y}_i = -0.97097 + 0.020623x$$

- Values on the estimated regression line.
- Predict values of $Y_i$ for a given value of $x_i$
  - $x_i = 201.1\ ^o$F:

$$\hat{Y}_i = -0.97097 + 0.020623(201.1) = 3.176315$$

  - $x_i = 210.7\ ^o$F:

$$\hat{Y}_i = -0.97097 + 0.020623(210.7) = 3.374296$$

## Residuals

$$e_i = Y_i - \hat{Y}_i$$

- Vertical distance between observed value of $Y$ and predicted value of $Y$.
- Residuals:
  - $x_i = 201.1$ $^o$F and $Y_i = 3.17764$:

$$e_i = 3.17764 - 3.176315 = 0.001325$$

  - $x_i = 210.7$ $^o$F and $Y_i = 3.36867$:

$$e_i = 3.36867 - 3.374296 = -0.005626$$

## ANOVA Table

| Source | df | SS | MS | F | p-value |
|---|---|---|---|---|---|
| Model | 1 | 0.22573 | 0.22573 | 2961.55 | $< 0.0001$ |
| Error | 15 | 0.00114 | 0.00007622 | | |
| Total | 16 | 0.22688 | | | |

**ANVOVA F-test**

- $H_0 : \beta_1 = 0$
- $H_a : \beta_1 \neq 0$
- $F = 2961.55$ with p-value $< 0.0001$
- Reject $H_0 \implies$ There is a significant linear relationship between boiling point of water and log of barometric pressure.

**Coefficient of Determination**

$$R^2 = \frac{SS_{\text{model}}}{SS_{\text{Total}}} = \frac{0.22573}{0.22688} = 0.9950$$

99.50% of the variation in log(barometric pressure) can be explained by the linear regression model with boiling point of water.

## Inference for Slope

- Test $H_o : \beta_1 = 0$ ($Y_i = \beta_0 + \epsilon_i$)
  versus $H_a : \beta_1 \neq 0$ ($Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$)
- Evaluate
$$t = \frac{b_1 - 0}{S_{b_1}} = \frac{.020623 - 0}{0.000379} = 54.42$$
- The least squares estimate of the slope is 54 standard errors away from zero (p-value $<<$ .0001).
  - It is extremely unlikely that an estimate that far from zero could occur simply because of random errors when $\beta_1$ is actually zero.
  - Consequently, reject the null hypothesis and conclude that the slope is positive.

**Inference for Slope**

- A 95% confidence interval for the slope indicates that the slope is "very well" estimated from these data

$$
\begin{aligned}
& b_1 \pm t_{15,.975} S_{b_1} \\
\Rightarrow \quad & 0.020623 \pm (2.131)(0.00037895) \\
\Rightarrow \quad & (0.0198, 0.0214)
\end{aligned}
$$

## Inference for Intercept

- Test $H_o : \beta_0 = 0 \; (Y_i = \beta_1 x_i + \epsilon_i)$

  versus $H_a : \beta_0 \neq 0 \; (Y_i = \beta_0 + \beta_1 x_i + \epsilon_i)$
- Evaluate $t = \dfrac{b_0 - 0}{S_{b_0}} = \dfrac{-0.971 - 0}{0.0769} = -12.6$
- The least squares estimate of the intercept is 12.6 standard errors away from zero (p-value $<<$ .0001).
  Reject the null hypothesis and conclude that the intercept is negative. (No practical motivation )
- A 95% confidence interval for the intercept is

  $$b_0 \pm t_{15,.975} S_{b_0} \;\Rightarrow\; -0.971 \pm (2.131)(0.0769) \;\Rightarrow\; (-1.135, -0.807)$$

## Inference for Conditional Mean

- Construct a 95% confidence interval for the mean of possible log-pressure measurements when the boiling point of water is x=209 $^o$F
- Estimated mean is
$$\hat{\mu}_{Y|x} = b_0 + b_1 x = -0.9710 + (.0206)(209) = 3.339$$
- Evaluate the standard error of this estimate
$$S_{\hat{\mu}_{Y|x}} = \sqrt{.0000762\left(\frac{1}{17} + \frac{(209 - 202.953)^2}{530.78}\right)} = 0.00312$$
- A 95% confidence interval is
$\hat{\mu}_{Y|x} \pm t_{15,.975} S_{\hat{\mu}_{Y|x}} \;\Rightarrow\; 3.339 \pm (2.131)(0.00312) \;\Rightarrow\; (3.333, 3.346)$

### Inference for Conditional Mean

- Apply the exponential function to the end points to get an *approximate* confidence interval for the mean pressure

$$(28.02, 28.39) \text{ inches of Hg}$$

- This could be computed with either the REG procedure or the GLM procedure in SAS by adding an additional line to the data file with X=209 and a missing value for Y

**Simultaneous Confidence Region**

Scheffe procedure for constructing a 95% confidence region for a segment of the true regression line

Evaluate $(b_0 + b_1 x) \pm \sqrt{2F_{(2,n-2),1-\alpha}} S_{b_0+b_1 x}$

$$\Rightarrow (b_0 + b_1 x) \pm \sqrt{2F_{(2,15),0.95}} S_{b_0+b_1 x}$$

$$\Rightarrow (b_0 + b_1 x) \pm (2.713) \sqrt{.0000762 \left( \frac{1}{17} + \frac{(x - 202.953)^2}{530.78} \right)}$$

## Prediction Interval

- Construct a 95% prediction interval for a log-pressure value when the boiling point of water is x=209 $^{o}$F
- Prediction is the estimated mean

$$\hat{Y} = b_0 + b_1 x + error = -0.9710 + (.0206)(209) + 0 = 3.339$$

- Evaluate the standard error of the prediction (include the variation of the associated random error, estimated as $MS_{error} = .0000762$ )

$$S_{pred} = \sqrt{.0000762\left(1 + \frac{1}{17} + \frac{(209 - 202.953)^2}{530.78}\right)} = 0.00927$$

## Prediction Interval

- A 95% prediction interval is

$$\hat{y} \pm t_{15,.975} S_{pred} \quad \Rightarrow \quad 3.339 \pm (2.131)(0.00927)$$
$$\Rightarrow \quad (3.319, 3.359)$$

- Apply the exponential function to the end points to get an *approximate* prediction interval for barometric pressure:
  (27.63, 28.76) inches of Hg
- This could be computed with either the REG or GLM procedure in SAS by adding an additional line to the data file with X=209 and a missing value for Y

# Questions?

**Contact me:**

Email: dmommen@iastate.edu
Student Office Hours: Thursdays @ 10-11 am