16-17. Simple Linear Regression

Readings: Kleinbaum, Kupper, Nizam, and Rosenberg (KKNR): Chs. 1-5, 7

SAS: PROC REG

Homework: Homework 7 due by midnight on October 29
Homework 8 due by midnight on November 5

Overview

A) Preview of Topics and Motivating Example
B) Continuous Outcome and Different Types of Covariates
C) Simple Linear Regression
D) Simple Linear Regression Assumptions
E) Variability of the Regression Parameters
F) Inference for Least Squares Estimators
G) Measuring Goodness of Fit
H) F Test for Simple Linear Regression
I) Prediction and Estimation in SLR

A. Preview of Topics and Motivating Example

- Motivate fitting a line to data: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

- Need to find $\hat{\beta}_0, \hat{\beta}_1, Var(\hat{\beta}_0), Var(\hat{\beta}_1)$ by minimizing:

$$SS_{Error} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

- Want to test H_0: $\beta_1 = 0$ (i.e. no linear association between X and Y)

- Test if the slope is 0 using $t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \sim t_{n-2}$ and $\hat{\beta}_1 \pm t_{n-2,1-\frac{\alpha}{2}} \times SE(\hat{\beta}_1)$

- Partitioning out the Regression and Residual Components: $SS_{Total} = SS_{Model} + SS_{Error}$

- ANOVA vs parameter table for simple linear regression

- R^2: proportion of the variance of *Y* that can be explained by the variable *X*

- Prediction Intervals (larger variance) vs. Confidence Intervals (smaller variance)

- **Lecture 18**: Simple Linear Regression Example

Motivating Example for Examining the Relationship Between Two Variables

Example: Lung function in children (FEV data) [Am J Epidemiology, 110(1): 15-26, 1980.]

Study Objective: To describe how lung function develops in children, and how smoking affects development.

Study Design: Cross-sectional survey. A random sample of children ages 3 to 19 from the East Boston area from which 654 had usable data.

Variables Measured: FEV (forced expiratory volume), age, sex, height, current smoking status. (FEV) measures how much air a person can exhale during a forced breath. *Higher* FEV indicates *better* lung function.

Outcome Variable (Y): FEV

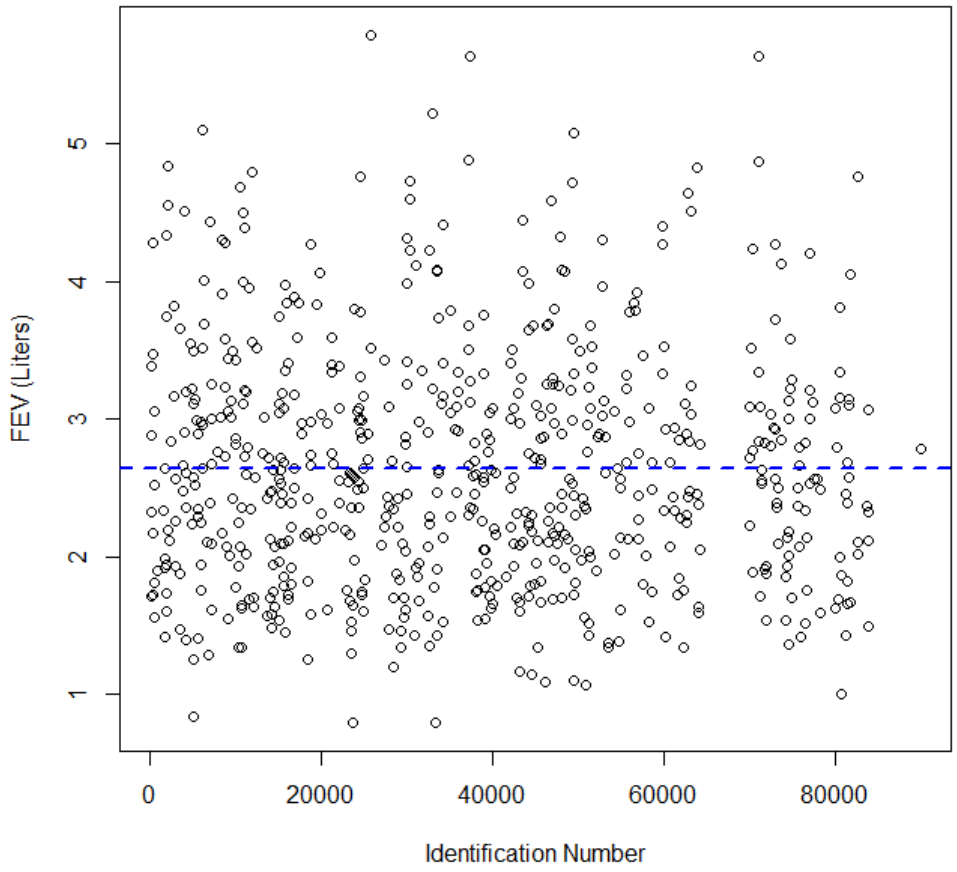
Primary Explanatory Variable (X): age, sex, height, smoking status (depending on the question of interest)

Covariates (C): age, sex, height, smoking status (depending on the question of interest)

B1. Continuous Outcome and No Covariates

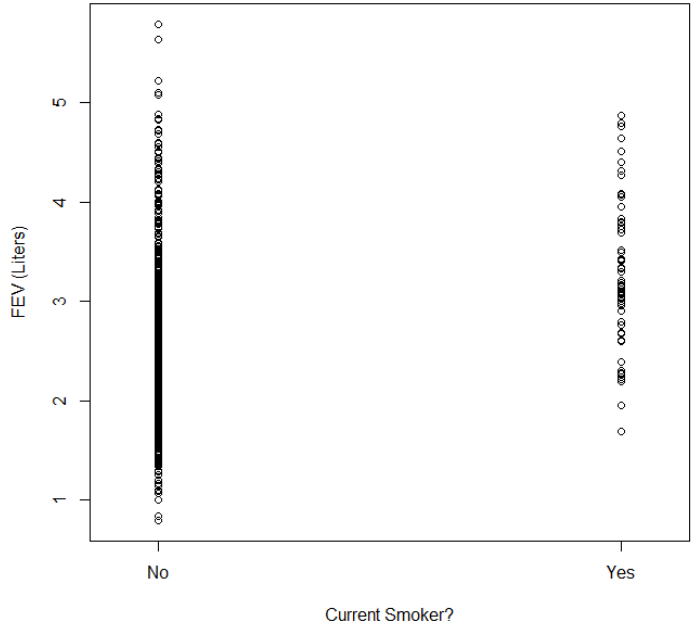
If the only information you have about a child is their identification number and FEV measurement, what would be your "best guess" for the FEV of this individual ("best guess" for Y)?

$$E(Y) = E(\text{FEV}) = \frac{\sum_{i=1}^{n} Y_i}{n} = \bar{Y} = 2.64 \text{ liters}$$



B2. Continuous Outcome and Binary Covariate

If you knew the child is a non-smoker, what would be your "best guess" for the FEV of this individual? Let's call this the expected value of Y given X, or $E(Y|X)$.



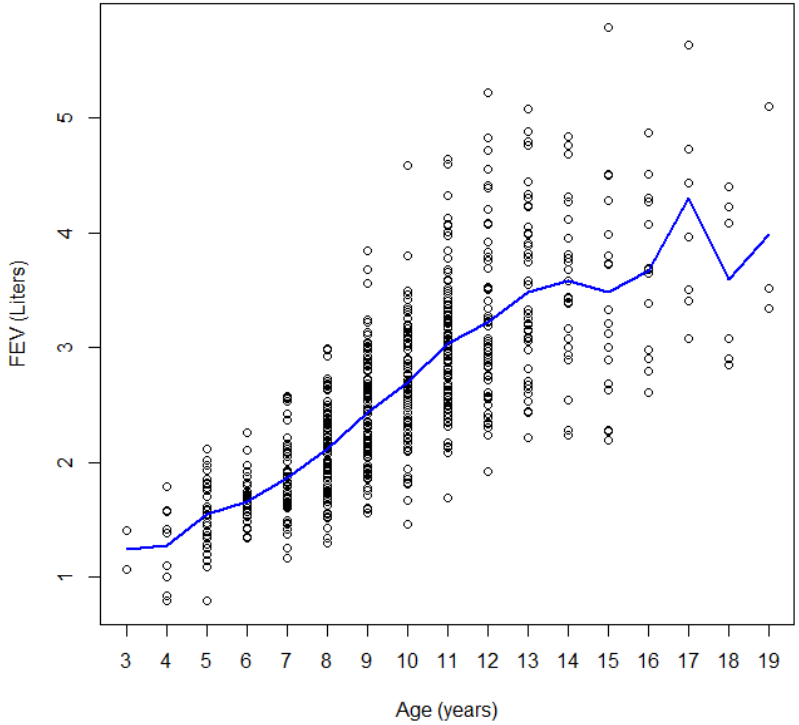
$$E(Y|X=0) = E(\text{FEV}|\text{Non-smoker}) = \frac{\sum_{i=1}^{n_{ns}} Y_{i,ns}}{n_{ns}} = \bar{Y}_{ns} = 2.57 \text{ liters}$$

$$E(Y|X=1) = E(\text{FEV}|\text{Smoker}) = \frac{\sum_{j=1}^{n_s} Y_{j,s}}{n_s} = \bar{Y}_s = 3.28 \text{ liters}$$

How could we test whether FEV differed for smokers and nonsmokers?

B3. Continuous Outcome and Continuous Covariate

If we knew the child's age, what would our "best guess" for the FEV of this individual be?



If the child is 15, our "best guess" would be $E(FEV|Age=15)$ = mean of FEV at age 15 = 3.5 L

If the child is 3, our "best guess" would be $E(FEV|Age=3)$ = mean of FEV at age 3 = 1.2 L, however we can note that this only based on 2 observations!

How would you test whether FEV differed for children of different ages?

We could treat age as categorical and compare age groups using ANOVA. However, this might not be the best idea because:

1.

2.

3.

Based on the scatterplot on the previous slide, what do you notice about the relationship between age and FEV?

Instead of taking the average at each age, we could instead **fit a line to the data** (i.e., fit a linear regression model).

C. Simple Linear Regression

Goal: We wish to model the distribution of some continuous response variable (e.g., FEV) across groups defined by a single predictor (e.g., age).

The regression line (i.e. $FEV = \beta_0 + \beta_1 Age$) is a linear approximation to the *graph of averages*, which shows the average value of *Y* (FEV) for each *X* (age).

This model can be used to answer commonly encountered statistical questions:

- Prediction: Estimating a future observation of response *Y*.
- Quantifying distributions: Describing the distribution of response *Y* within groups defined by *X*.
- Comparing distributions across groups defined by *X*.

Motivation for Simple Linear Regression

The regression model allows us to make inferences about groups that have few (if any) subjects by "borrowing" information from other groups.

Interpolation to unobserved groups is less risky than extrapolation outside the range of predictors included in the regression model.

Different *mathematical models* may be appropriate to model the distribution of *Y* across *X*: a straight line, a parabola, a log function, etc.

Ultimately, we are interested in finding the "best" model that describes our data while also being *parsimonious* (i.e., using simpler models and/or as few predictors as possible). Sometimes we choose a model suggested from experience or theory.

We will begin by assuming a straight line model with one predictor:

$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

Note: Even when we do not think a straight line represents the true relationship across groups, we can still estimate the "average rate of change" from the model.

The Regression Line and Its Components

The line $Y = \beta_0 + \beta_1 X_1 + \epsilon$ is known as a ***regression model***. The components and some properties are:

- β_0 is the ***intercept*** of the line
  - The intercept is the expected value of Y when X is zero: *E(Y|X=0)*
  - Oftentimes we do not actually observed values of zero for X (e.g., age of 0)

- β_1 is the ***slope*** of the line
  - The slope is the expected change in *Y* associated with a *one unit* increase in *X*

- ϵ is the **error term**
  - We don't expect the linear relationship to hold exactly for all individuals, so the error term is included to represent this variability
  - Represents the variance of *Y* given a value of *X*
  - Assumed to follow a normal dist. with mean 0 and variance σ_e^2: $\epsilon \sim N(0, \sigma_e^2)$
  - If the model perfectly predicts all individuals, then $\sigma_e^2 = \sigma_{Y|X}^2$, else $\sigma_e^2 > \sigma_{Y|X}^2$

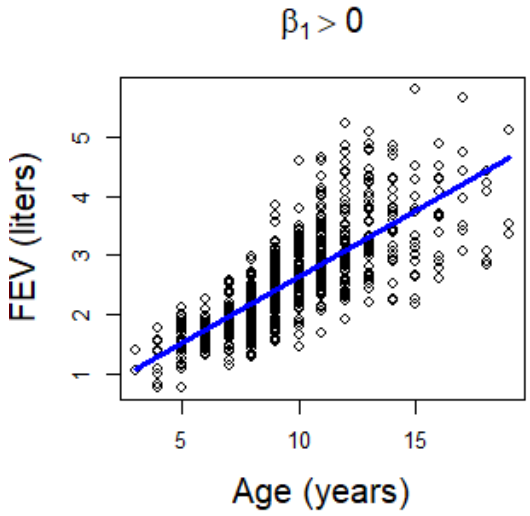
The predicted value of *Y* for a given value of *X* is $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1$, where the hats represent the estimated values for our regression coefficients and the resulting FEV predicted value.

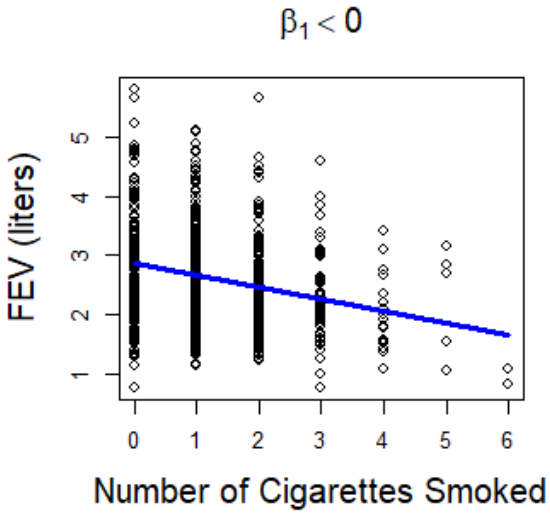
Values of β_1

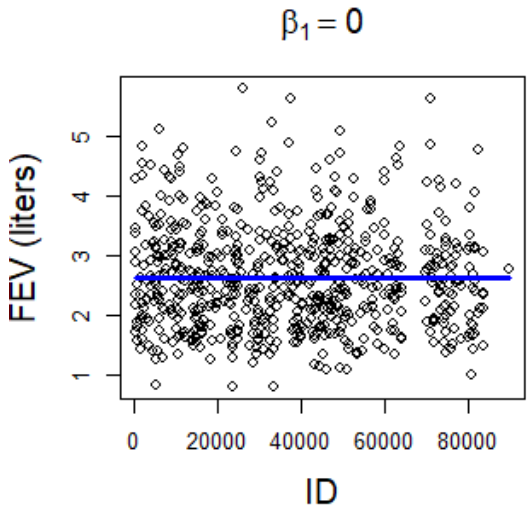
β_1 can take on any value from $-\infty$ to ∞:

- If $\beta_1 > 0$, then as *X* increases, the expected value of *Y* increases.
- If $\beta_1 < 0$, then as *X* increases, the expected value of *Y* decreases.
- If $\beta_1 = 0$, then there is no *linear* relationship between *X* and *Y* (but there could be non-linear relationships)

These interpretations hold for the "continuous" predictors to the right, and also for "categorical" predictors (interpreted with respect to the reference category).







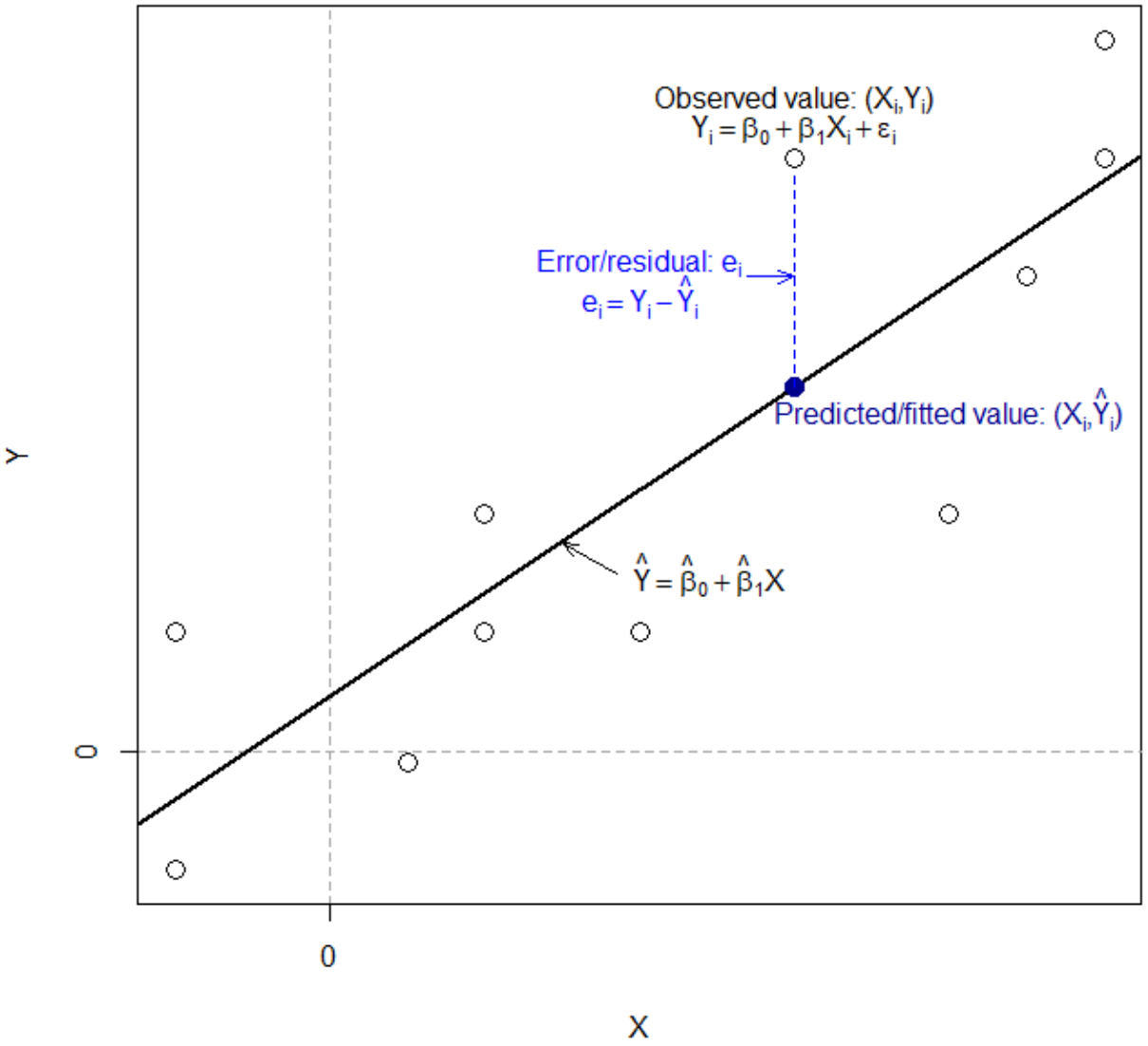
Finding the "Best" Line

Let (X_i,Y_i) be data for the ith individual for $i=1,\ldots,n$.

The difference between a fitted value and an observed value is called the ***residual*** or the "***error***."

The **residual** is a measure of the error we make when predicting Y using our regression equation (i.e., β_0 for the intercept and β_1 for the slope).

We can choose the "best" line as the line that minimizes the error.



There are multiple approaches to quantify the total error:

$$1)\ S = \sum_{i=1}^{n} e_i \qquad 2)\ S = \sum_{i=1}^{n} |e_i| \qquad 3)\ S = \sum_{i=1}^{n} e_i^2$$

1. An infinite number of lines can minimize Approach 1.
2. Approach 2 is analytically difficult to work with.
3. Approach 3 is "easy" to use *and* has theoretical justification. This has become the standard method used to minimize the error/residuals.

Approach 3 is called the "Method of Least Squares" or "Least Squares Regression" because it minimizes the sum of squares due to error (SS_{Error}):

$$SS_{Error} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2$$

$\hat{\beta}_0$ and $\hat{\beta}_1$ from Approach 3 are also the best linear unbiased estimator (BLUE) according to the Gauss-Markov Theorem. This means that (1) $E(e_i) = 0$; (2) e_i is independent of e_j $\forall i, j$; and (3) the errors have equal variance that is the smallest possible variance of the estimate (as compared to other unbiased, linear estimators).

Mathematically stated, Approach 3 identifies estimates for β_0 and β_1, $\hat{\beta}_0$ and $\hat{\beta}_1$, such that for any other possible estimators, $\hat{\beta}_0^*$ and $\hat{\beta}_1^*$, it must be true that:

$$SS_{Error} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2 < \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0^* - \hat{\beta}_1^* X_i\right)^2$$

The ***sum of squares error*** is sometimes called the ***residual sum of squares***.

How can we arrive at these optimal estimates for $\hat{\beta}_0$ and $\hat{\beta}_1$?

One approach is to treat our SS_{Error} as a loss function and minimize if over all choices for β_0 and β_1. To obtain the minimum (or maximum) of a function we find values such that the first (partial) derivatives are equal to zero…

First we will determine $\hat{\beta}_0$:

$$\frac{\partial}{\partial \beta_0} SS_{Error} = \frac{\partial}{\partial \beta_0}\left(\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2\right) = \sum_{i=1}^{n} -2(Y_i - \beta_0 - \beta_1 X_i)$$

Setting this partial derivative equal to 0, we can then solve for the $\hat{\beta}_0$ that minimizes SS_{Error}:

$$\begin{aligned}
\sum_{i=1}^{n} -2\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right) &= 0 \\
\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right) &= 0 \\
\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n}\hat{\beta}_0 - \sum_{i=1}^{n}\hat{\beta}_1 X_i &= 0 \\
\frac{\sum_{i=1}^{n} Y_i}{n} - \frac{\sum_{i=1}^{n}\hat{\beta}_0}{n} - \frac{\hat{\beta}_1 \sum_{i=1}^{n} X_i}{n} &= \frac{0}{n} \\
\bar{Y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{X} &= 0
\end{aligned}$$

Finally, we can solve for $\hat{\beta}_0$: $\bar{Y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{X} = 0 \rightarrow \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

However, we see at this point we have the unknown $\hat{\beta}_1$ in our solution.

Let's now solve for $\hat{\beta}_1$ (which follows similar steps):

$$\frac{\partial}{\partial \beta_1} SS_{Error} = \frac{\partial}{\partial \beta_1}\left(\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2\right) = \sum_{i=1}^{n} -2X_i(Y_i - \beta_0 - \beta_1 X_i)$$

Setting this partial derivative equal to 0 and noting that we can substitute our solution for $\hat{\beta}_0$ into the problem (colored green) we can solve for the $\hat{\beta}_1$ that minimizes SS_{Error}:

$$\begin{aligned}
\sum_{i=1}^{n} -2X_i\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right) &= 0 \\
\sum_{i=1}^{n} -X_i\left(Y_i - [\bar{Y} - \hat{\beta}_1 \bar{X}] - \hat{\beta}_1 X_i\right) &= 0 \\
\sum_{i=1}^{n} -X_i Y_i + \bar{Y}\sum_{i=1}^{n} X_i - \hat{\beta}_1 \bar{X}\sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 &= 0 \\
\hat{\beta}_1\left(\sum_{i=1}^{n} X_i^2 - \bar{X}\sum_{i=1}^{n} X_i\right) &= \sum_{i=1}^{n} X_i Y_i - \bar{Y}\sum_{i=1}^{n} X_i
\end{aligned}$$

Finally, solve for $\hat{\beta}_1$: $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \bar{Y}\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} X_i^2 - \bar{X}\sum_{i=1}^{n} X_i}$. However, this isn't super nice to work with…

Some algebraic acrobatics can help us transform the numerator and denominator into a nicer representation for $\hat{\beta}_1 = \frac{\sum_{i=1}^n X_i Y_i - \bar{Y}\sum_{i=1}^n X_i}{\sum_{i=1}^n X_i^2 - \bar{X}\sum_{i=1}^n X_i} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$.

Numerator (working backwards):

$$\begin{aligned}\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) &= \sum_{i=1}^n X_i Y_i - \bar{Y}\sum_{i=1}^n X_i - \bar{X}\sum_{i=1}^n Y_i + \sum_{i=1}^n \bar{X}\bar{Y} \\ &= \sum_{i=1}^n X_i Y_i - \bar{Y}\sum_{i=1}^n X_i + (n\bar{X}\bar{Y} - n\bar{X}\bar{Y}) \\ &= \sum_{i=1}^n X_i Y_i - \bar{Y}\sum_{i=1}^n X_i\end{aligned}$$

Denominator (working backwards):

$$\begin{aligned}\sum_{i=1}^n (X_i - \bar{X})^2 &= \sum_{i=1}^n (X_i^2 - 2X_i\bar{X} + \bar{X}^2) \\ &= \sum_{i=1}^n X_i^2 - \bar{X}\sum_{i=1}^n X_i - \bar{X}\sum_{i=1}^n X_i + n\bar{X}^2 \\ &= \sum_{i=1}^n X_i^2 - \bar{X}\sum_{i=1}^n X_i - n\bar{X}^2 + n\bar{X}^2 \\ &= \sum_{i=1}^n X_i^2 - \bar{X}\sum_{i=1}^n X_i\end{aligned}$$

To review, we obtained the minimum of our SS_{Error} by taking the first (partial) derivative, setting it equal to 0, and solving for β_0 and β_1. This resulted in optimal estimates which minimize the residuals of

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

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{S_{XY}}{S_{XX}}$$

All simple linear regression parameters can be estimated from 5 summary statistics:

1. n
2. $\sum X_i$ (can determine $\bar{X}$ by dividing by n)
3. $\sum Y_i$ (can determine $\bar{Y}$ by dividing by n)
4. $\sum\left(X_i^2\right)$
5. $\sum(X_i)(Y_i)$

We can also note some interesting connections for the sums of squares making up $\hat{\beta}_1$:

- $\frac{S_{XY}}{S_{XX}} \times \frac{n-1}{n-1} = \frac{\frac{S_{XY}}{n-1}}{\frac{S_{XX}}{n-1}} = \frac{Cov(X,Y)}{Var(X)} = r_{XY}\frac{s_y}{s_x}$, where s_y and s_x are the uncorrected sample standard deviations (i.e., they use n instead of n-1)

D. Simple Linear Regression Assumptions

1. **Existence**: For any fixed value of the variable *X*, *Y* is a random variable with a certain probability distribution having finite mean and variance.

2. **Linearity**: The mean value of *Y* (or a transformation of *Y*), $\mu_{Y|X} = E(Y)$, is a straight-line function of *X* (or a transformation of *X*).

3. **Independence**: The errors, e_i, are independent (i.e., *Y*-values are statistically independent of one another).

4. **Homoscedasticity**: The errors, e_i, at each value of the predictor, x_i, have equal variance (i.e., the variance of *Y* is the same for any *X*). That is,
$$\sigma^2_{Y|X} = \sigma^2_{Y|X=1} = \sigma^2_{Y|X=2} = \ldots = \sigma^2_{Y|X=x}$$

5. **Normal Distribution**: The errors, e_i, at each value of the predictor, x_i, are normally distributed (i.e., for any fixed value of *X*, *Y* has a normal distribution). (Note this assumption does ***not*** state that *Y* is normally distributed.)

Illustration of the linearity, homoscedasticity, and normality assumptions:

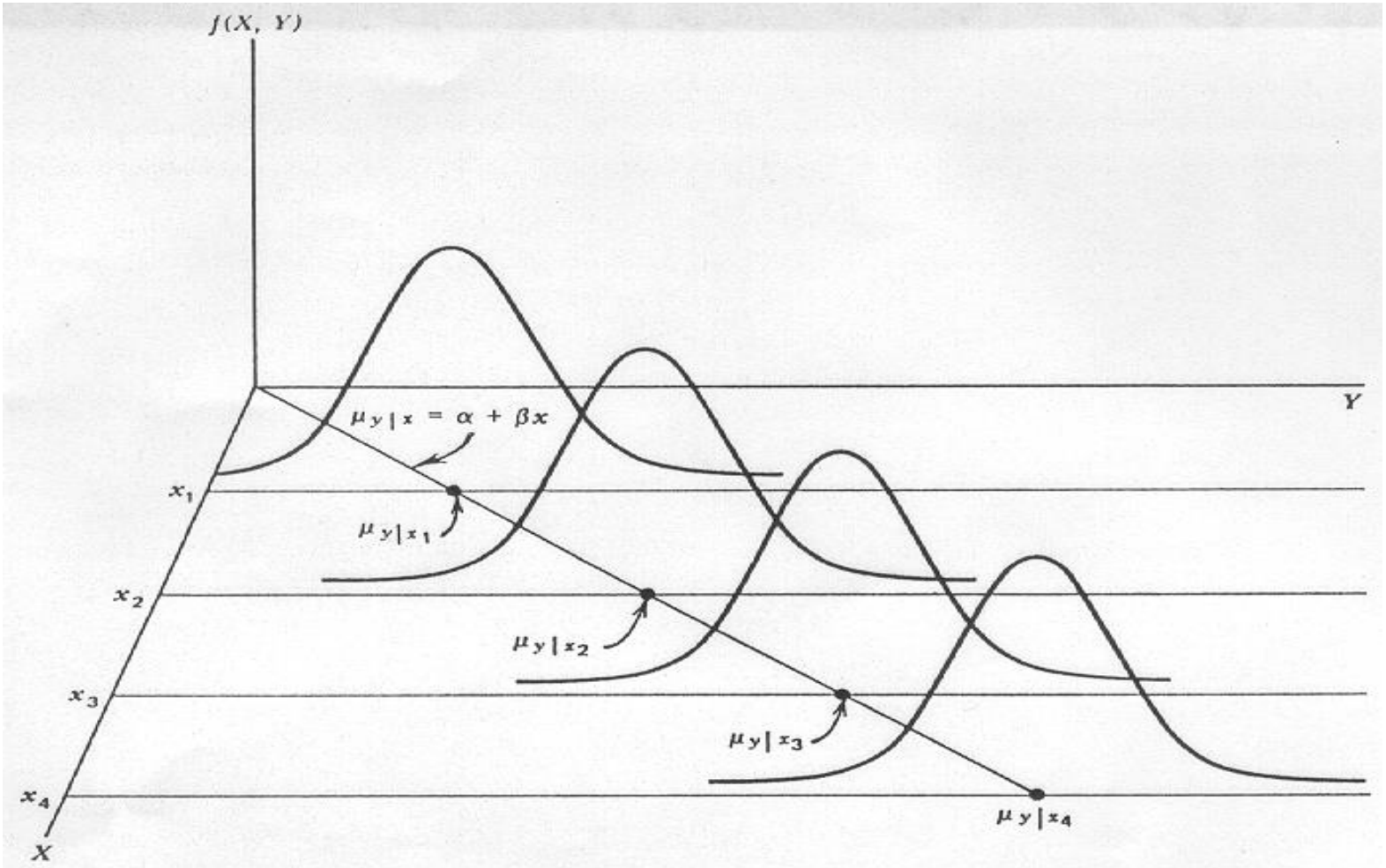


Figure 8.2.1. Representation of the simple linear regression model.

E. Variability of the Regression Parameters

To estimate the variability (standard error) of the regression parameters:

First let's further rearrange $\hat{\beta}_1$:

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\\ &= \frac{\sum_{i=1}^{n}(X_i-\bar{X})Y_i - \bar{Y}\sum_{i=1}^{n}(X_i-\bar{X})}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\\ &= \frac{\sum_{i=1}^{n}(X_i-\bar{X})Y_i}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\\ &= \frac{(X_1-\bar{X})Y_1+\cdots+(X_n-\bar{X})Y_n}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\end{aligned}$$

where $\bar{Y}\sum_{i=1}^{n}(X_i-\bar{X})=0$ because

$$\sum_{i=1}^{n}(X_i-\bar{X})=\sum_{i=1}^{n}X_i-\sum_{i=1}^{n}\bar{X}=\frac{n}{n}\times\sum_{i=1}^{n}X_i-n\bar{X}=n\bar{X}-n\bar{X}=0$$

Recall that we have the following properties/equations for variances which will be useful in determining the variability of our regression parameters:

$$Var(cY) = c^2 \times Var(Y)$$
$$SE(cY) = c \times SE(Y)$$

$$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$$
$$Var(X - Y) = Var(X) + Var(Y) - 2Cov(X, Y)$$

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

$$Cov(X, Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

Correlation coefficient:

$$r_{x,y} = \frac{Cov(X,Y)}{SD(X)SD(Y)} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Returning to our estimate of the variability for $\hat{\beta}_1$:

$$Var[\hat{\beta}_1] = Var\left[\frac{(X_1 - \bar{X})Y_1 + \cdots + (X_n - \bar{X})Y_n}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right] = \frac{Var[(X_1 - \bar{X})Y_1 + \cdots + (X_n - \bar{X})Y_n]}{\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right]^2}$$

Since X is assumed "fixed" it can be treated as a constant, and using $Var(cY) = c^2Var(Y)$, the above expression becomes:

$$\frac{[(X_1 - \bar{X})^2 Var(Y_1) + \cdots + (X_n - \bar{X})^2 Var(Y_n)] + [2(X_1 - \bar{X})(X_2 - \bar{X})Cov(Y_1, Y_2) + \cdots + (X_{n-1} - \bar{X})(X_n - \bar{X})Cov(Y_{n-1}, Y_n)]}{[\sum_{i=1}^{n}(X_i - \bar{X})^2]^2}$$

Note, by the assumption of independence: $Cov(Y_j, Y_k) = 0 \ \ \forall j \neq k$

And by the assumption of homoscedasticity: $Var(Y_1) = Var(Y_2) = \ldots = Var(Y_n) = \sigma^2_{Y|X}$

Therefore, we can simplify $Var[\hat{\beta}_1]$ to

$$Var\left[\hat{\beta}_1\right] = \frac{\sigma^2_{Y|X}\sum_{i=1}^{n}(X_i - \bar{X})^2}{\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right]^2} = \frac{\sigma^2_{Y|X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sigma^2_{Y|X}}{S_{XX}}$$

Now, let's determine the variability of $\hat{\beta}_0$:

$$\begin{aligned}
Var(\hat{\beta}_0) &= Var(\bar{Y} - \hat{\beta}_1\bar{X}) \\
&= Var(\bar{Y}) + Var\left(\hat{\beta}_1\bar{X}\right) - 2Cov(\bar{Y}, \hat{\beta}_1\bar{X}) \\
&= Var\left(\frac{\sum_{i=1}^{n} Y_i}{n}\right) + \bar{X}^2 Var(\hat{\beta}_1) \\
&= \frac{1}{n^2} Var\left(\sum_{i=1}^{n} Y_i\right) + \bar{X}^2 Var(\hat{\beta}_1) \\
&= \frac{1}{n^2} \sum_{i=1}^{n} Var(Y_i) + \bar{X}^2 Var(\hat{\beta}_1) \\
&= \frac{\sigma^2_{Y|X}}{n} + \bar{X}^2 \left(\frac{\sigma^2_{Y|X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right) \\
&= \sigma^2_{Y|X} \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)
\end{aligned}$$

On the next slide we will show that $Cov\left(\bar{Y}, \hat{\beta}_1\bar{X}\right) = 0$.

$$
\begin{aligned}
Cov\left(\bar{Y}, \hat{\beta}_1 \bar{X}\right) &= Cov\left(\bar{Y}, \frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right) \\
&= \frac{Cov\left(\sum_{i=1}^{n} Y_i, \sum_{i=1}^{n}(X_i - \bar{X})Y_i\right)}{n\sum_{i=1}^{n}(X_i - \bar{X})^2} \\
&= \frac{E\left(\sum_{i=1}^{n} Y_i \sum_{i=1}^{n}(X_i - \bar{X})Y_i\right) - E\left(\sum_{i=1}^{n} Y_i\right)E\left(\sum_{i=1}^{n}(X_i - \bar{X})Y_i\right)}{n\sum_{i=1}^{n}(X_i - \bar{X})^2} \\
&= \frac{E\left(\sum_{i=1}^{n}(X_i - \bar{X})Y_i^2 + \sum_{i\neq j}(X_i - \bar{X})Y_iY_j\right) - \left(\sum_{i=1}^{n} E[Y_i]\right)\left(\sum_{i=1}^{n}(X_i - \bar{X})E[Y_i]\right)}{n\sum_{i=1}^{n}(X_i - \bar{X})^2} \\
&= \frac{\sum_{i=1}^{n}(X_i - \bar{X})E\left[Y_i^2\right] + \sum_{i\neq j}(X_i - \bar{X})E[Y_i]E\left[Y_j\right] - n\mu^2\sum_{i=1}^{n}(X_i - \bar{X})}{n\sum_{i=1}^{n}(X_i - \bar{X})^2} \\
&= \frac{(\sigma^2 + \mu^2)\sum_{i=1}^{n}(X_i - \bar{X}) + \mu^2\sum_{i\neq j}(X_i - \bar{X}) - n\mu^2\sum_{i=1}^{n}(X_i - \bar{X})}{n\sum_{i=1}^{n}(X_i - \bar{X})^2} \\
&= 0
\end{aligned}
$$

Note that $E\left(Y_iY_j\right) = E(Y_i)E\left(Y_j\right)\ \forall i \neq j$ by independence and that $\sum_{i=1}^{n}(X_i - \bar{X}) = 0$.

Summarizing it together:

Standard error of the slope:

$$SE(\hat{\beta}_1) = \sqrt{\frac{\sigma^2_{Y|X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

Standard error of the intercept:

$$SE(\hat{\beta}_0) = \sqrt{\sigma^2_{Y|X}\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)}$$

For *Y* overall the estimated variance is $\hat{\sigma}^2_Y = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}$, but for each value of *X* there is a subpopulation of values of *Y*. The variances of the subpopulations are $\sigma^2_{Y|X}$ and its estimate is:

$$\hat{\sigma}^2_{Y|X} = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n-2} = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2}{n-2} = \frac{SS_{Error}}{n-2} = MSE$$

Note: $\hat{\sigma}^2_{Y|X}$ is only an unbiased estimate of $\sigma^2_{Y|X}$ if the model is correct (if a straight-line model is appropriate), else $\hat{\sigma}^2_{Y|X} > \sigma^2_{Y|X}$.

$\hat{\sigma}^2_{Y|X}$ Behavior:

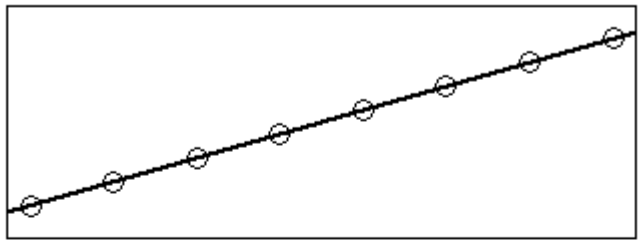
1. If $\hat{\sigma}^2_{Y|X}$ is zero then the points will fall exactly on the regression line.

2. If $\hat{\sigma}^2_{Y|X}$ is small then the points will lie close to the regression line.

3. If $\hat{\sigma}^2_{Y|X}$ is large then the points will not fall close to the regression line. (Due to true variability in Y|X and/or lack of fit)

4. The larger $\hat{\sigma}^2_{Y|X}$ the more scatter there will be in the data about the regression line.

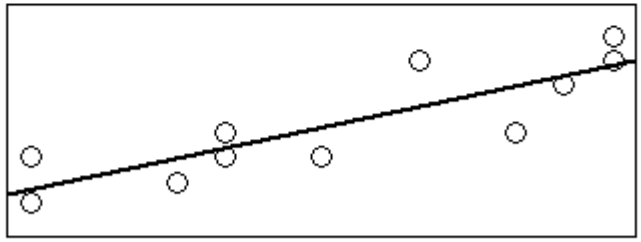
Perfect Fit: $\hat{\sigma}^2_{Y|X} = 0$



Y

X

Imperfect Fit: $\hat{\sigma}^2_{Y|X} > 0$



Y

X

F. Inference for Least Squares Estimators

We can make inferences about the parameters with the additional assumption of normality:

$$\varepsilon_i \sim N(0, \sigma_e^2)$$

$$Y_i | X_i \sim N(\mu_{Y|X}, \sigma_{Y|X}^2)$$

where $\mu_{Y|X}$ is allowed to change (linearly) with the explanatory variable. That is, $\mu_{Y|X} = \beta_0 + \beta_1 X$.

Thus, if we assume normality of the errors, then the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are normally distributed since linear functions of independent normally distributed random variables are themselves normally distributed (See Corollary 4.6.10 in Casella & Berger).

Alternatively:

- If we have a large sample size, asymptotic normality may be assumed for the estimators.

- If asymptotic normality does not hold, bootstrap or Monte Carlo methods may be appropriate.

Testing for Significant Associations

If there is no linear association between the explanatory and response variables, then the slope is zero:

$$H_0: \beta_1 = 0 \quad vs. \quad H_A: \beta_1 \neq 0.$$

We can test if the slope is 0 using a *t*-statistic with *n*-2 degrees of freedom because when H_0 is true our test statistic, *t*, follows a *t*-distribution:

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \sim t_{n-2}$$

We can also calculate a 95% confidence interval for the slope coefficient:

$$\hat{\beta}_1 \pm t_{n-2,1-\frac{\alpha}{2}} \times SE(\hat{\beta}_1)$$

What does it mean if we fail to reject H_0?

1) There is no association
2) There is no linear association
3) We've made a Type II error

Example Regression Output for FEV data

SAS Code:

```
proc reg data=fev;
    model fev = age;
run;
```

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	280.91916	280.91916	872.18	<.0001
Error	652	210.00068	0.32209		
Corrected Total	653	490.91984			

Root MSE	0.56753	**R-Square**	0.5722
Dependent Mean	2.63678	**Adj R-Sq**	0.5716
Coeff Var	21.52349		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	**1**	$\hat{\beta}_0 = 0.43165$	0.07790	5.54	<.0001
age	**1**	$\hat{\beta}_1 = 0.22204$	0.00752	29.53	<.0001

Interpreting and Utilizing the Regression Output

What is the regression equation?

What is the interpretation of the slope parameter?

What is the interpretation of the intercept?

Is there a significant linear relationship between age and FEV?

Calculate the 95% confidence interval for age and interpret.

What is the predicted FEV for an 11-year old child?

What is the predicted difference in FEV for 16-year old children versus 11-year old children?

What is the 95% confidence interval around this predicted difference?

Quality of the Straight-Line Fit

Once the least-squares line is determined, we may wish to know how well the least-squares regression line 'fits' the data.

- Does the fitted line help us predict Y? That is, is least-squares line better than no line at all for predicting Y?
- And if so, to what extent?

We can examine the fit of the regression line by partitioning the **total variability** of *Y* into two components:

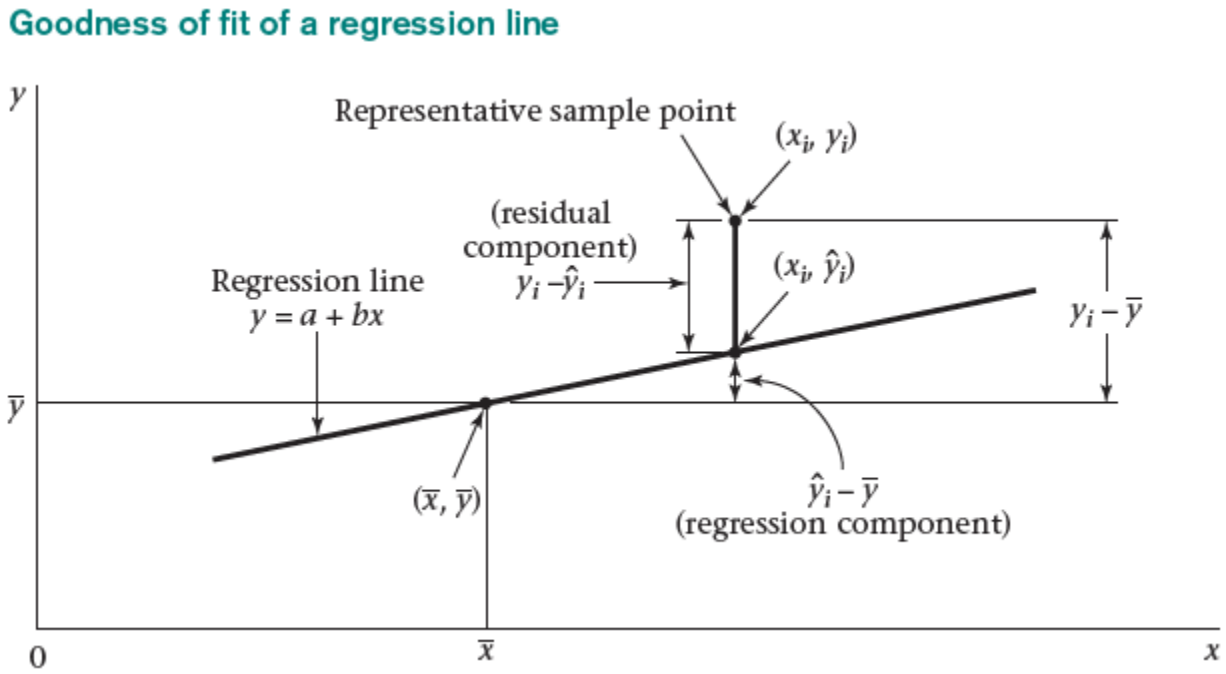
Regression component: The variability in *Y* due to the regression of *Y* on *X*. The regression component is the difference between the predicted *Y* and the mean of the *Y*'s:

$$\hat{Y}_i - \bar{Y}$$

Residual component (***error***): The variability in *Y* "left-over" after the regression of *Y* on *X*. The residual component is the difference between the observed *Y* and predicted *Y*:

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

Partitioning Out the Regression and Residual Components



Rosner 7th Ed., pg. 435

The simplest regression estimate for Y_i is $\bar{Y}$ (an intercept-only model). The difference between the observed *Y*'s and the mean of the *Y*'s , $Y_i - \bar{Y}$, is the ***total error***. The total error can be broken down further as the sum of the ***regression component*** and the ***residual component***:

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

The ***fundamental equation of regression analysis***: with a little algebra (similar to the one-way ANOVA lecture breakdown of total sum of squares on slide 5) we can show that:

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

$$\mathrm{SS}_{\mathrm{Total}} = \mathrm{SS}_{\mathrm{Model}} + \mathrm{SS}_{\mathrm{Error}}$$

Total Sum of Squares (SS_{Total}). The total sum of squares is the sum of squares of the deviations of the individual sample points from the sample mean (Note the relationship between $\mathrm{SS}_{\mathrm{Total}}$ and the variance of Y, $\hat{\sigma}_Y^2$):

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \; ; \;\; \hat{\sigma}_Y{}^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1} = \frac{\mathrm{SS}_{\mathrm{Total}}}{n-1}$$

Error Sum of Squares (SS_{Error}). The error sum of squares is the sum of squares of the residual components (note the relationship between $\mathrm{SS}_{\mathrm{Error}}$ and the variance of Y given X, $\hat{\sigma}_{Y|X}^2$):

$$\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \; ; \;\; \hat{\sigma}_{Y|X}^2 = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n-2} = \frac{\mathrm{SS}_{\mathrm{Error}}}{n-2}$$

Model Sum of Squares (SS_{Model}). The model sum of squares is the sum of squares of the regression components:

$$\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 = \mathrm{SS}_{\mathrm{Model}} = \mathrm{SS}_{\mathrm{Total}} - \mathrm{SS}_{\mathrm{Error}}$$

G. Measuring Goodness of Fit

We often want to know how well our regression model fits the data. Measuring the "goodness of fit" involves quantifying how much scatter there is around the regression line.

Sum of Squares Error (SS_{Error}): The variation in the data after fitting the line (that is, the "left-over" variation) is quantified using the SS_{Error}.

- Large values indicate a lot of left-over variation.
- Small values indicate little left-over variation.

The "R-squared" value, also known as the ***coefficient of determination***, is the proportion of total variation in the data (about the average $\bar{Y}$) that is removed by fitting the regression line. That is, the proportion of the variance of *Y* that can be explained by the variable *X*.

$$R^2 = \frac{\mathrm{SS}_{\mathrm{Total}} - \mathrm{SS}_{\mathrm{Error}}}{\mathrm{SS}_{\mathrm{Total}}} = \frac{\mathrm{SS}_{\mathrm{Model}}}{\mathrm{SS}_{\mathrm{Total}}}$$

R^2 is often multiplied by 100 and is interpreted as the percent of the total variation in the dependent variable *Y* that is explained by the independent variable *X (using a linear model)*.

Connection Between R^2 and the Correlation Coefficient between X and Y

In **simple linear regression**, the square root of R^2 is equivalent to the correlation coefficient between *X* and *Y.*

$$
\begin{aligned}
R^2 = \frac{\mathrm{SS}_{\mathrm{Model}}}{\mathrm{SS}_{\mathrm{Total}}} &= \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \\
&= \frac{\sum_{i=1}^{n}\left(\hat{\beta}_0 + \hat{\beta}_1 X_i - \bar{Y}\right)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \\
&= \frac{\sum_{i=1}^{n}\left([\bar{Y} - \hat{\beta}_1\bar{X}] + \hat{\beta}_1 X_i - \bar{Y}\right)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \\
&= \frac{\hat{\beta}_1^2 \sum_{i=1}^{n}(X_i - \bar{X})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \\
&= \left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)^2 \left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}\right) \\
&= \left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \sum_{i=1}^{n}(X_i - \bar{X})^2}}\right)^2 = r^2
\end{aligned}
$$

Properties of R^2

R^2 can only be between 0 and 1: $0 \leq R^2 \leq 1$

If $R^2 = 0$ then the regression line explains none of the variability in *Y* and the regression line is not better than no line at all for predicting *Y* (no better than using $\bar{Y}$ as our predictor of *Y*):

If $SS_{Model}=0$, then $SS_{Total}=SS_{Model}+SS_{Error}=SS_{Error}$ and

$$R^2 = \frac{SS_{Total} - SS_{Error}}{SS_{total}} = \frac{SS_{Error} - SS_{Error}}{SS_{Total}} = 0$$

If $R^2 = 1$ then there is a perfect fit and the regression line explains all of the variability. In this case every data point falls exactly on the regression line and there is no residual variation.

If $SS_{Error}=0$, then $R^2 = \frac{(SS_{Total} - SS_{Error})}{SS_{Total}} = \frac{SS_{Total} - 0}{SS_{Total}} = 1$

Note that R^2 does <u>not</u> measure the magnitude of the slope or measure the appropriateness of the straight-line model (i.e., a large R^2 does not necessarily imply an "adequate" model).

Example: FEV and age cont.: $R^2 = \frac{SS_{Model}}{SS_{Total}} = \frac{280.91916}{490.91984} = 0.5722$

Interpretation: 57.22% of the variability in FEV can be explained by a linear relationship with age.

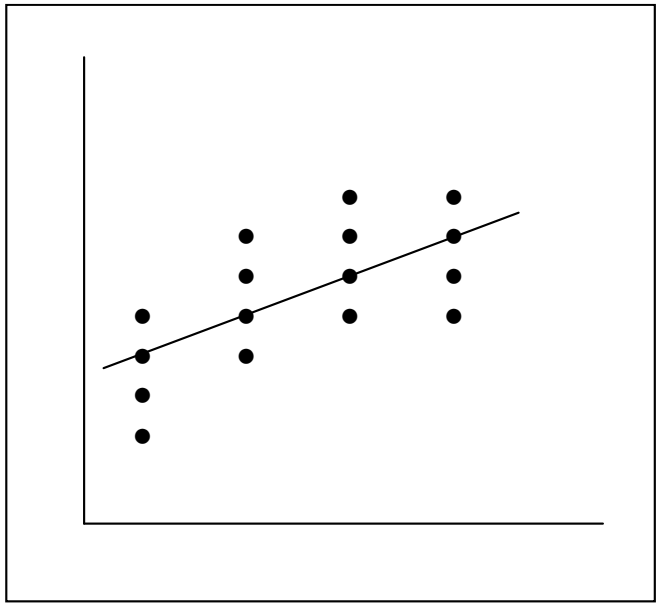
Estimate of $\sigma^2_{Y|X}$ and Lack of Fit

As stated previously, we use **SS_{Error}** to estimate $\sigma^2_{Y|X}$. This estimate is given by:

$$\hat{\sigma}^2_{Y|X} = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n-2} = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2}{n-2} = \frac{\mathrm{SS}_{\mathrm{Error}}}{n-2} = \mathrm{MS}_{\mathrm{Error}}$$

The Mean Square Error (MS_{Error}) will only provide an unbiased estimate of the error variance when the hypothesized model is correct (in this case, if a straight-line model is appropriate), otherwise the MS_{Error} will estimate a quantity larger than the error variance. $\hat{\sigma}^2_{Y|X} > \sigma^2_{Y|X}$.

- If the model is incorrect, then two factors contribute to the inflation of SSE. The true variability in Y|X ("***pure error***") and the error due to fitting an incorrect model ("***lack of fit***").
- With replicate observations (i.e., multiple observations with the same values of the predictor(s)), we can test for lack of fit in the assumed model by obtaining an estimate of $\sigma^2_{Y|X}$ that does not assume the correctness of the straight-line model (i.e., a model-free estimate of the residual variance or "pure error"). ***More in future lectures.***



H. *F* Test for Simple Linear Regression

The ***regression mean square*** *or* ***model mean square*** is the regression (model) sum of squares divided by the number of predictor variables, *p*, in the model (*p* =1 for simple linear regression).

$$MS_{Model} = SS_{Model} / (p)$$

$$E(MS_{Model}) = \sigma^2_{Y|X} + \beta_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2$$

The ***residual mean square*** is the residual sum of squares divided by its degrees of freedom (*n*-2 in the case of simple linear regression).

$$MS_{Error} = SS_{Error} / (n\text{-}p\text{-}1)$$

$$E(MS_{Error}) = E(s^2_{Y|X}) = \sigma^2_{Y|X}$$

Recall from our discussion of ANOVA that the ratio of two variances (s_1^2/s_2^2) follows an *F* distribution under the null hypothesis that the two variances are equal (σ_1^2=σ_2^2):

$$\frac{s_1^2 / \sigma_1^2}{s_2^2 / \sigma_2^2} \sim F_{n_1-1, n_2-1}$$

Under the null hypothesis that the true slope of the regression line is zero (H_0: $\beta_1 = 0$), both MS_{Model} and MS_{Error} are independent estimates of $\sigma^2_{Y|X}$. Thus, the ratio of the regression mean square to the residual mean square will have an *F* distribution with *p* and *n*-*p*-1 degrees of freedom:

$$F = \frac{MS_{Model}}{MS_{Error}} \sim F_{p,n-p-1} \ (F_{1,n-2} \text{ for simple linear regression})$$

The *F* test is used to test if the model including covariate(s) results in a significant reduction of the residual sum of squares compared to a model containing only an intercept.

If the null hypothesis is true, then the expected value of the *F* ratio should be 1. If the null hypothesis is false, then the expected value of the *F* ratio is greater than 1.

The t-test and the F-test are equivalent for testing H_0: $\beta_1 = 0$ in simple linear regression:

$$\text{If } X \sim t_n, \text{ then } X^2 \sim F_{1,n}.$$

$$\text{Recall, } t = \frac{\widehat{\beta}_1}{SE(\widehat{\beta}_1)} \text{ where } t \sim t_{n-p-1} \text{ under } H_0.$$

ANOVA table for displaying regression results

The Analysis of Variance (ANOVA) table is typically used to summarize regression results, where *n* is the sample size and *p* is the number of variables included in the model:

Source	Sum of Squares	Degrees of Freedom	Mean Square	Variance Ratio (F)	p-value
Model	SS_{Model}	p	MS_{Model}	$F=\frac{MS_{Model}}{MS_{Error}}$	$Pr(F_{p,n-p-1} > F)$
Error	SS_{Error}	n-p-1	$MS_{Error}=\sigma^2_{Y\|X}$		
Total	SS_{Total}	n-1			

***Example*:** FEV and age continued from output

$H_0: \beta_1 = 0$ *vs.* $H_A: \beta_1 \neq 0$

$F = 872.18$

$Pr(F_{1,653} > 872.18) < 0.0001$

Conclusion: Reject the null hypothesis that $\beta_1 = 0$ and conclude that there is a significant association between age and FEV ($p < 0.0001$).

Annotated SAS Output: FEV and Age (reproducing same results from earlier)

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	p = 1	280.91916	SS_{Model}/p = 280.91916	MS_{Model}/MS_{Error} = 872.18	<.0001
Error	n-p-1 = 652	210.00068	$SS_{Error}/(n-p-1)$ = 0.32209		
Corrected Total	n-1 = 653	490.91984			

Root MSE	0.56753	**R-Square**	SS_{Model}/SS_{Total} = 0.5722
Dependent Mean	2.63678	**Adj R-Sq**	$1-(1-R^2)*((n-1)/(n-p-1)$ = 0.5716
Coeff Var	Root MSE/Dep Mean = 21.52349		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	**1**	$\hat{\beta}_0 = 0.43165$	0.07790	5.54	<.0001
age	**1**	$\hat{\beta}_1 = 0.22204$	$\hat{\beta}_1/t_1$ = 0.00752	$\sqrt{F} = \sqrt{872.18}$ = 29.53	<.0001

I. Prediction and Estimation in Simple Linear Regression

Reference Range (Population Confidence Interval): description of the variability in the underlying population (usually the central 95% of the population) and is usually estimated from *large* samples of individuals representative of the population.

Confidence Interval: description of the variability in our sample estimate of the true underlying mean (or any other population parameter).

***Example*:** FEV in children cont. ($\bar{y}$= 2.64, $\hat{\sigma}_Y$ = 0.867, n = 654)
What is the expected FEV for a single child ($\hat{Y}$)?

What is the 95% Reference Range? (Note: Need to assume FEV is normally distributed.)

What is our estimate of the true underlying mean FEV in children ($\hat{\mu}$)?

What is the 95% Confidence Interval? (Note: FEV doesn't need to be normal if CLT applies.)

In regression, we assume that the underlying mean changes according to the level of an explanatory variable(s).

Estimation: The expected mean (average) μ for a given value of *X,* say X_0, in the underlying population is:

$$\hat{\mu}_{Y|X_0} = \hat{\beta}_0 + \hat{\beta}_1 X_0$$

The standard error of the estimate (*as shown on the next slide*) is given by:

$$SE(\hat{\mu}_{Y|X_0}) = \sqrt{\frac{\hat{\sigma}^2_{Y|X}}{n} + \frac{\hat{\sigma}^2_{Y|X}}{n-1}\left(\frac{(X_0 - \bar{X})^2}{\hat{\sigma}^2_X}\right)}$$

Prediction: The predicted value of *Y* for a given value of *X,* say X_0, for a randomly selected *individual* from the underlying population is:

$$\hat{Y}|X_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0$$

Its standard error is given by:

$$SE(\hat{Y}|X_0) = \sqrt{\hat{\sigma}^2_{Y|X} + \frac{\hat{\sigma}^2_{Y|X}}{n} + \frac{\hat{\sigma}^2_{Y|X}}{n-1}\left(\frac{(X_0 - \bar{X})^2}{\hat{\sigma}^2_X}\right)}$$

This is broken down as the variance of the individual around $\mu_{Y|X}$ and the variance in estimating $\mu_{Y|X}$

Calculating the standard error of $\hat{\mu}_{Y|X_0} = \hat{\beta}_0 + \hat{\beta}_1 X_0$

First note that we can further manipulate our equation:

$$\hat{\mu}_{Y|X_0} = \hat{\beta}_0 + \hat{\beta}_1 X_0 = \left(\bar{Y} - \hat{\beta}_1 \bar{X}\right) + \hat{\beta}_1 X_0 = \bar{Y} + \hat{\beta}_1 (X_0 - \bar{X})$$

Then the variance is

$$\begin{aligned} Var\left(\hat{\mu}_{Y|X_0}\right) &= Var\left(\bar{Y} + \hat{\beta}_1 (X_0 - \bar{X})\right) \\ &= Var(\bar{Y}) + (X_0 - \bar{X})^2 Var\left(\hat{\beta}_1\right) + 2(X_0 - \bar{X}) Cov(\bar{Y}, \hat{\beta}_1) \\ &= Var(\bar{Y}) + (X_0 - \bar{X})^2 \frac{\hat{\sigma}^2_{Y|X}}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ &= \frac{\hat{\sigma}^2_{Y|X}}{n} + \frac{\hat{\sigma}^2_{Y|X}}{n-1}\left(\frac{(X_0 - \bar{X})^2}{\hat{\sigma}^2_X}\right) \end{aligned}$$

because $\hat{\sigma}^2_X = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n-1}$ and $Var\left(\hat{\beta}_1\right) = \frac{\hat{\sigma}^2_{Y|X}}{\sum_{i=1}^n (X_i - \bar{X})^2}$.

Note that when $X_0 = \bar{X}$, $SE\left(\hat{\mu}_{Y|X_0 = \bar{X}}\right) = \sqrt{\frac{\hat{\sigma}^2_{Y|X}}{n}}$.

Confidence Intervals and Prediction Intervals

Confidence Interval: A 95% confidence interval describes the variability in the estimate of the underlying mean. It can be calculated (without assuming normality of the errors) as:

$$\hat{\mu}_{Y|X} \pm t_{n-2,0.975} SE(\hat{\mu}_{Y|X})$$

Prediction Interval: A 95% prediction interval (like a reference range) describes the variability in the underlying population. It can be calculated (assuming normality of the errors) as:

$$(\hat{Y}|X) \pm t_{n-2,0.975} SE(\hat{Y}|X)$$

Prediction intervals will be wider than confidence intervals since there is more variability around estimating an individual point as compared to the mean (i.e., prediction intervals take the true error term into account).

***Example*:** FEV in children (NOTE: $\sum x_i = 6495$, $\sum x_i^2 = 70201$, $\bar{X}=9.931$). Note that MSE=$s^2_{Y|X}$.

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	280.91916	280.91916	872.18	<.0001
Error	652	210.00068	MSE = $s^2_{Y\|X}$ = 0.32209		
Corrected Total	653	490.91984			

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	**1**	$\hat{\beta}_0 = 0.43165$	0.07790	5.54	<.0001
age	**1**	$\hat{\beta}_1 = 0.22204$	0.00752	29.53	<.0001

95% Confidence Interval for the underlying mean FEV among all children aged 16.

$\hat{\mu}_{Y|X=16} = 0.43165 + 0.22204(16) = 3.98\ L$

$$SE\left(\hat{\mu}_{Y|X=16}\right) = \sqrt{\frac{\hat{\sigma}^2_{Y|X}}{n} + \frac{\hat{\sigma}^2_{Y|X}}{n-1}\left(\frac{(X_0 - \bar{X})^2}{\hat{\sigma}^2_X}\right)}$$

$$= \sqrt{\frac{0.32209}{654} + \frac{0.32209}{653}\left[\frac{\left(16 - \left(\frac{6495}{654}\right)\right)^2}{\frac{70201 - \left(\frac{6495^2}{654}\right)}{653}}\right]}$$

$$= 0.05074$$

95% confidence interval: $3.98 \pm 1.96(0.05074) = (3.885, 4.089)$

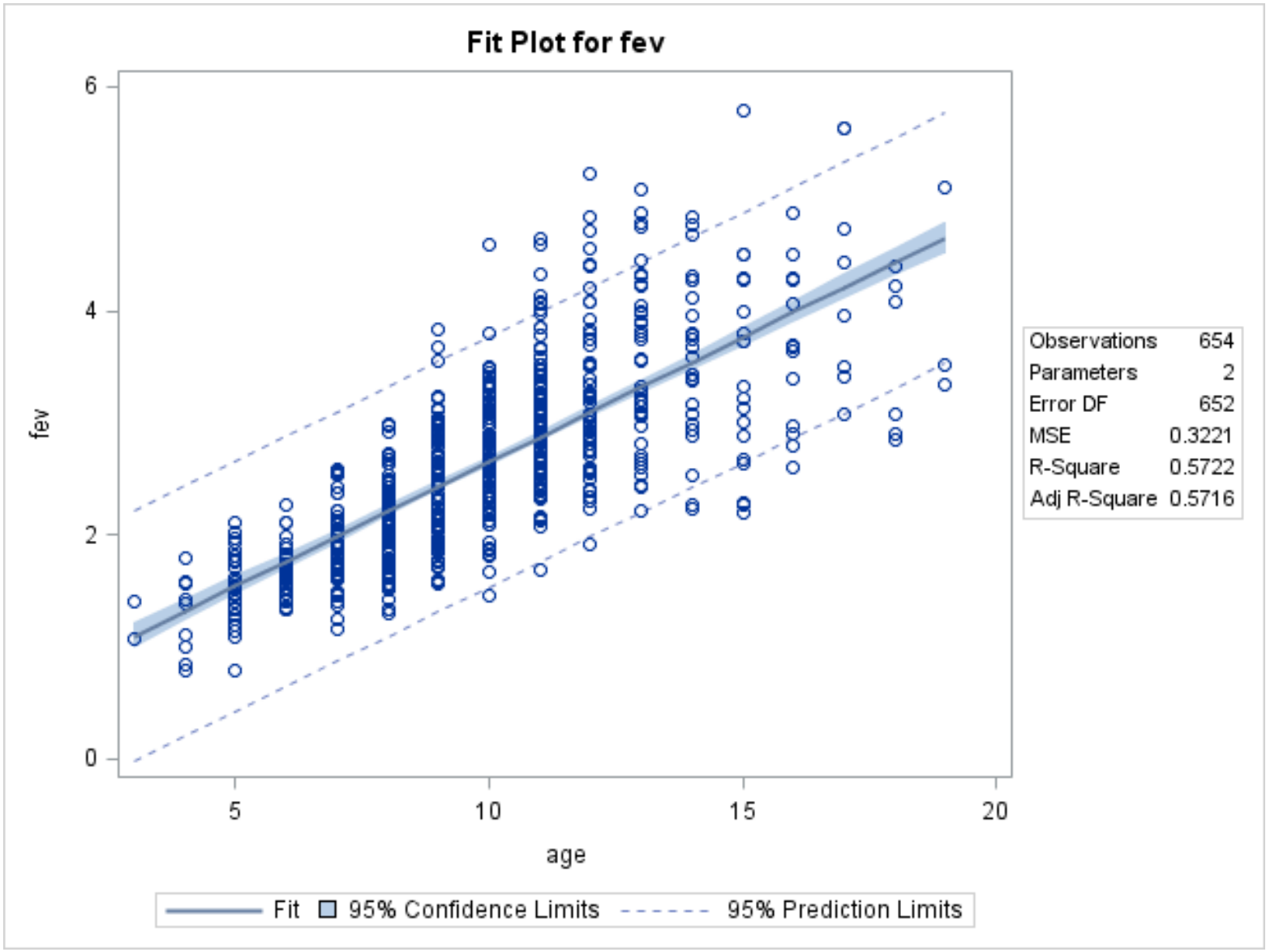
Note: $t_{654-2,0.975} = 1.963609$

95% Prediction Interval for FEV in a randomly selected child aged 16.

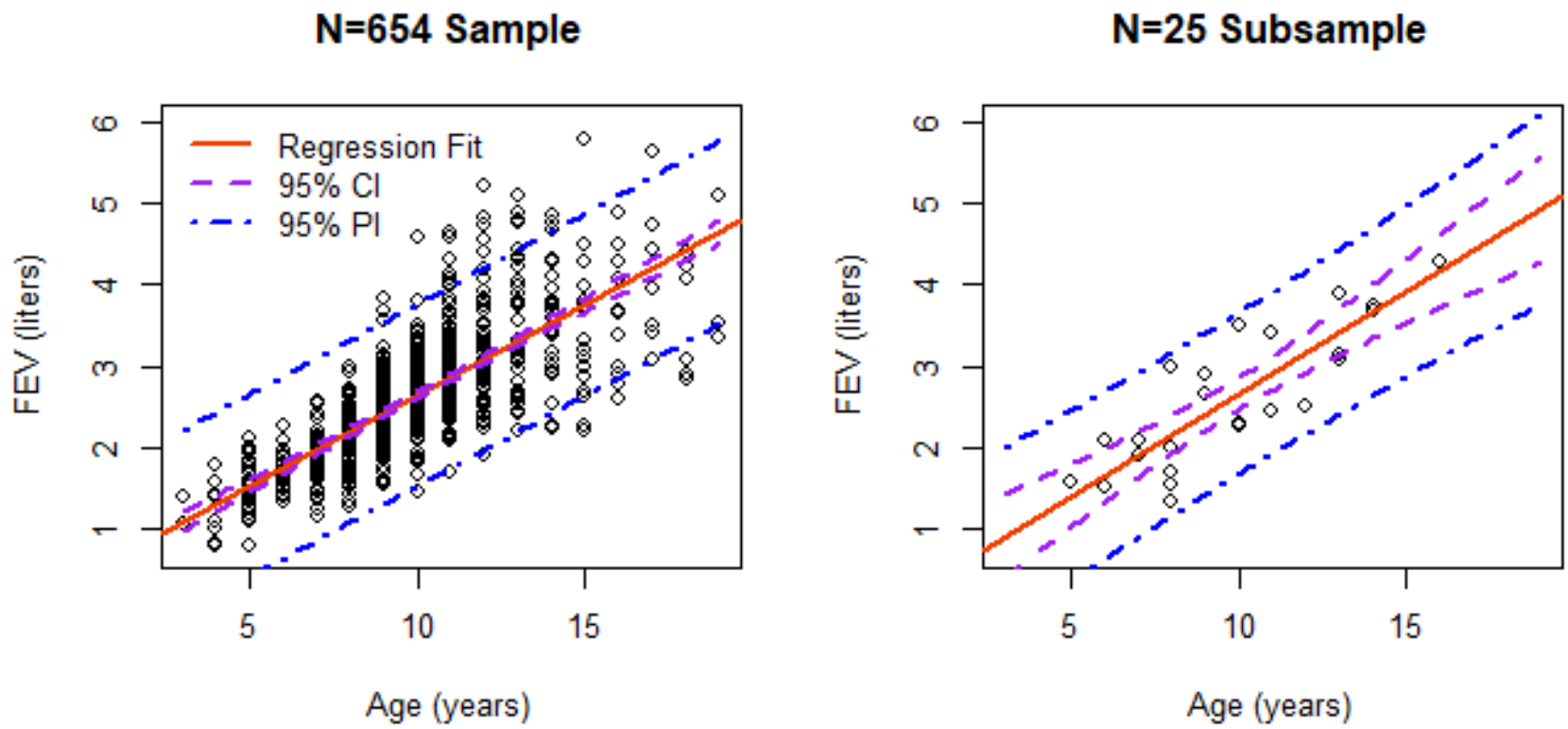
$$SE\left(\hat{Y}\middle|X_0\right) = \sqrt{\hat{\sigma}^2_{Y|X} + \frac{\hat{\sigma}^2_{Y|X}}{n} + \frac{\hat{\sigma}^2_{Y|X}}{n-1}\left(\frac{(X_0 - \bar{X})^2}{\hat{\sigma}^2_X}\right)} = 0.569793$$

95% prediction interval: $3.98 \pm 1.96(0.56793) = (2.867, 5.101)$

FEV Data: Comparing Prediction Intervals and Confidence Intervals



FEV Data: Comparing PI and CI for Different Sample Sizes



From these two figures of the entire data set and a small subsample, we can note a few interesting observations:

1. The change in slope has a larger impact on the intervals, especially CI, at the extremes. This is because when $X_0 = \bar{X}$ the standard error is minimized.
2. A smaller sample size more greatly impacts the confidence interval width than the prediction interval width.

Properties of Prediction Intervals and Confidence Intervals

Prediction intervals are wider than confidence intervals.

Both prediction intervals and confidence intervals are wider at the ends of the data (the SEs are at a minimum at $\bar{X}$).

Confidence intervals shrink considerably as the sample size grows. Prediction intervals stay about the same as the sample size grows.

Large samples are generally required if prediction intervals are to be used as reference ranges.

Notation Summary

Right Notation:

Truth: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

Expected: $E[Y_i] = \beta_0 + \beta_1 X_i$ because $[\varepsilon_i] = 0$

Estimate: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

Wrong Notation:

$Y_i \neq \beta_0 + \beta_1 X_i$

Implies Y vs X is a perfect line

$E[Y_i] \neq \hat{\beta}_0 + \hat{\beta}_1 X_i$

$E[Y_i] = E[\beta_0 + \beta_1 X_i + \varepsilon_i] = \beta_0 + \beta_1 X_i$

$E[\hat{Y}_i] = E[\hat{\beta}_0 + \hat{\beta}_1 X_i] = \beta_0 + \beta_1 X_i$

because $E[\hat{\beta}] = \beta$

Truth vs. estimate:

$\hat{Y}_i \neq \beta_0 + \beta_1 X_i$

$Y_i \neq \hat{\beta}_0 + \hat{\beta}_1 X_i$

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