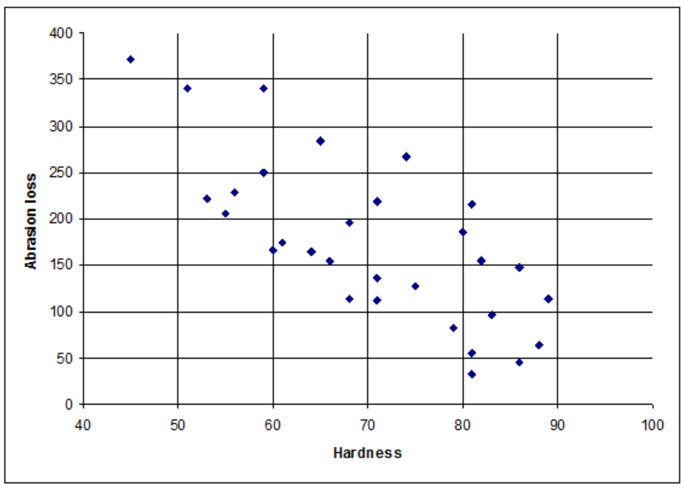
Chapter 11: SIMPLE LINEAR REGRESSION AND CORRELATION

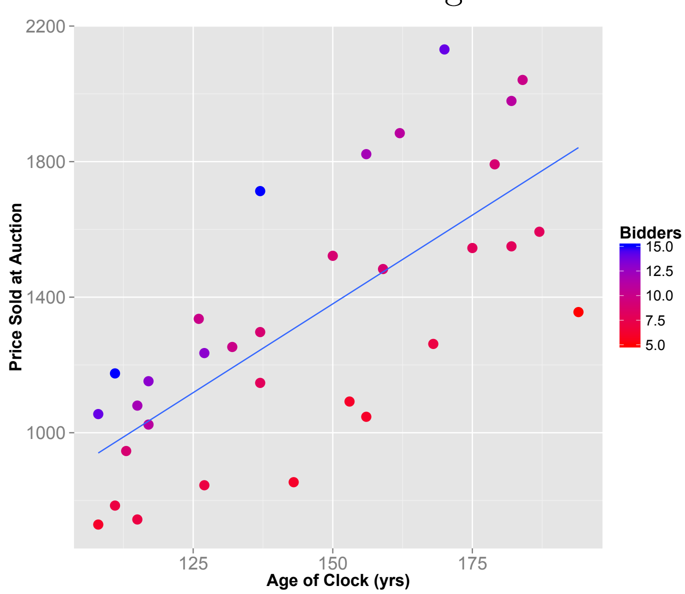
Part 1: Simple Linear Regression (SLR) Introduction

Sections 11-1 and 11-2

Abrasion Loss vs. Hardness



Price of clock vs. Age of clock



- **Regression** is a method for studying the relationship between two or more **quantitative variables**

- **Simple linear regression (SLR)**:

  One quantitative dependent variable
  - response variable
  - dependent variable
  - Y

  One quantitative independent variable
  - explanatory variable
  - predictor variable
  - X

- **Multiple linear regression**:

  One quantitative dependent variable

  Many quantitative independent variables

  – You'll see this in STAT:3200/IE:3760 Applied Linear Regression, if you take it.

- SLR Examples:
  - predict salary from years of experience
  - estimate effect of lead exposure on school testing performance
  - predict force at which a metal alloy rod bends based on iron content

- **Example**: Health data

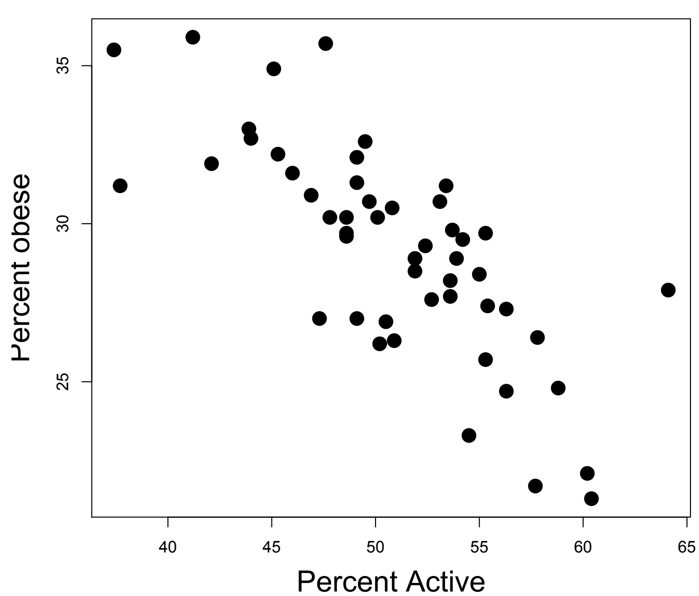
  Variables:

  Percent of Obese Individuals

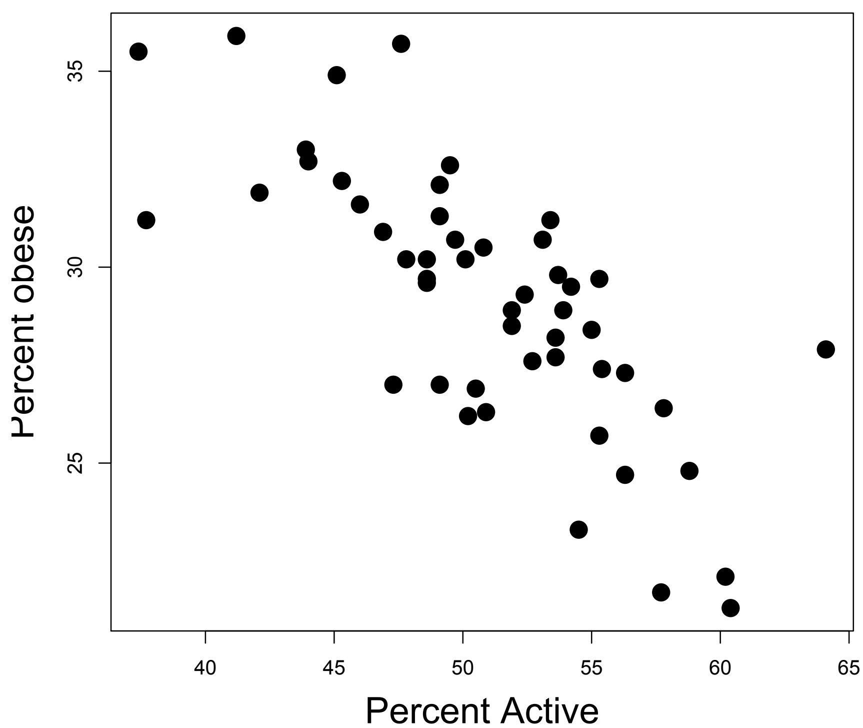
  Percent of Active Individuals

Data from CDC. Units are regions of U.S. in 2014.

```
  PercentObesity PercentActive
1           29.7          55.3
2           28.9          51.9
3           35.9          41.2
4           24.7          56.3
5           21.3          60.4
6           26.3          50.9
.
.
.
```



A scatterplot or scatter diagram can give us a general idea of the relationship between obesity and activity...



The points are plotted as the pairs (x_i, y_i) for $i = 1, \ldots, 25$

Inspection suggests a linear relationship between obesity and activity (i.e. a straight line would go through the bulk of the points, and the points would look randomly scattered around this line).

Simple Linear Regression

The model

- The basic model

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

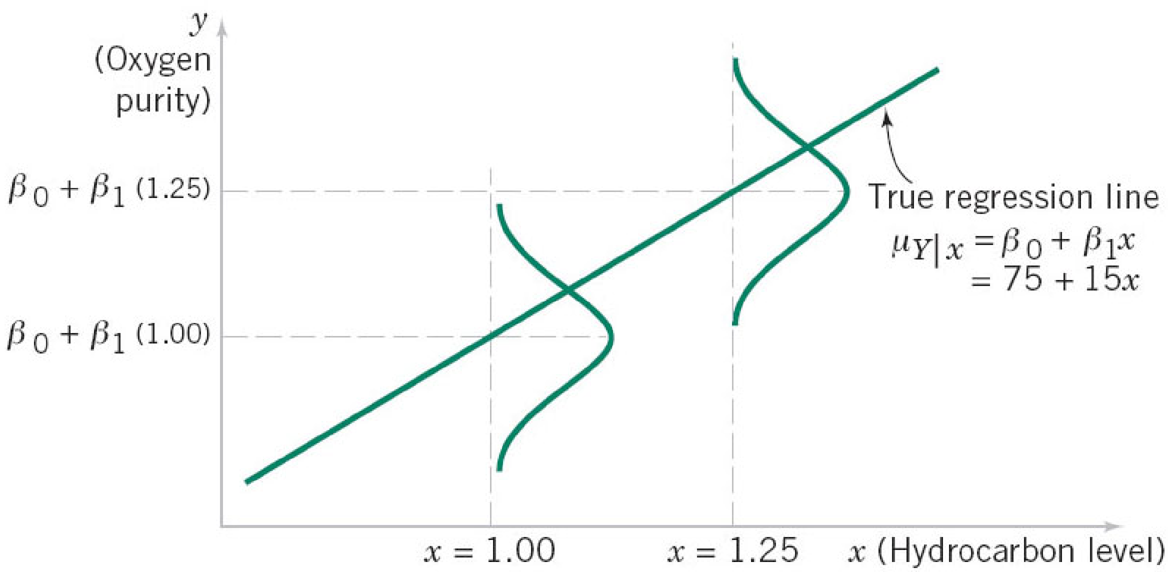
  - Y_i is the observed response or dependent variable for observation i
  - x_i is the observed predictor, regressor, explanatory variable, independent variable, covariate
  - ϵ_i is the error term
  - ϵ_i are iid $N(0, \sigma^2)$
    (iid means independently and identically distributed)

– So, $E[Y_i|x_i] = \beta_0 + \beta_1 x_i + 0 = \beta_0 + \beta_1 x_i$

The conditional mean (i.e. the expected value of Y_i given x_i, or after conditioning on x_i) is "$\beta_0 + \beta_1 x_i$" (a point on the estimated line).

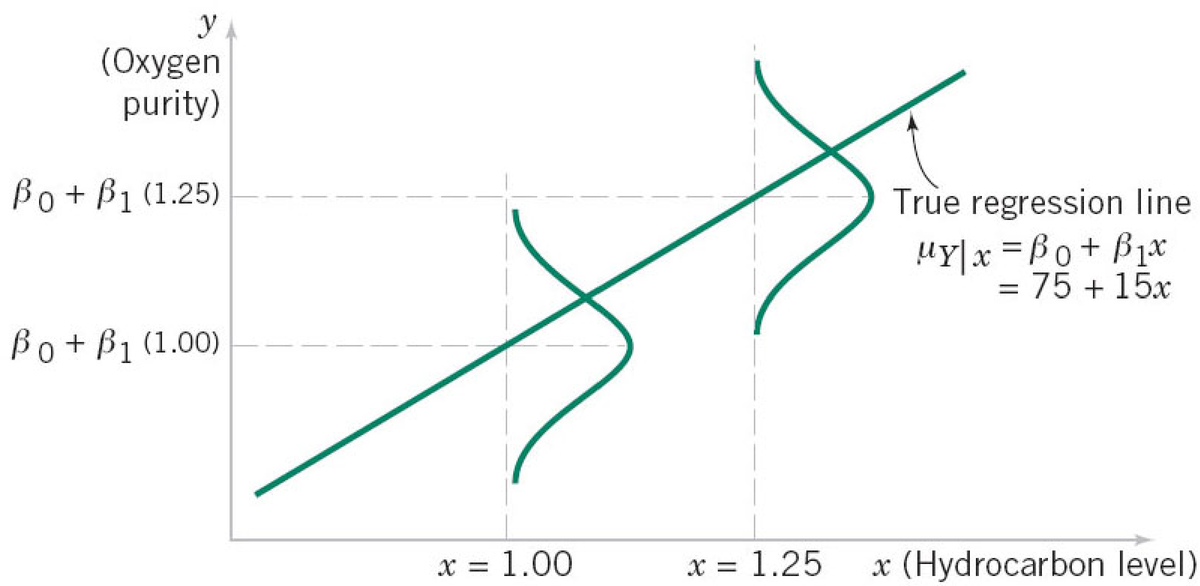
– Or, as another notation, $E[Y|x] = \mu_{Y|x}$

– The random scatter around the mean (i.e. around the line) follows a $N(0, \sigma^2)$ distribution.



Example: Consider the model that regresses Oxygen purity on Hydrocarbon level in a distillation process with...

$$\beta_0 = 75 \text{ and } \beta_1 = 15$$



For each x_i there is a different Oxygen purity mean (which is the center of a normal distribution of Oxygen purity values).

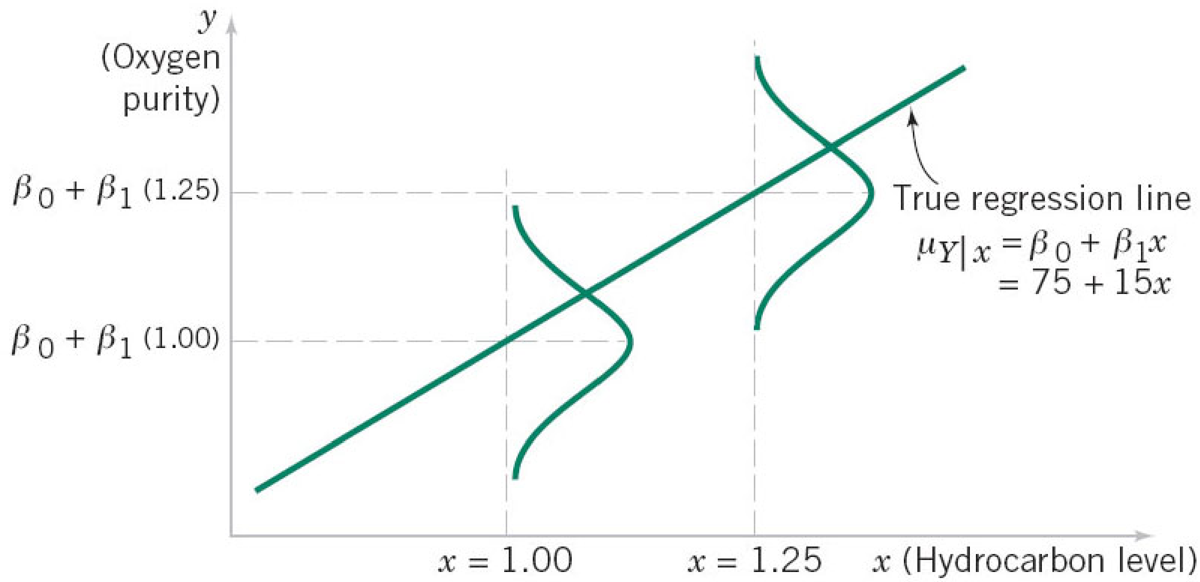
Plugging in x_i to $(75+15x_i)$ gives you the conditional mean at x_i.

The conditional mean for $x = 1$:

$$E[Y|x] = 75 + 15 \cdot 1 = 90$$

The conditional mean for $x = 1.25$:

$$E[Y|x] = 75 + 15 \cdot 1.25 = 93.75$$



These values that randomly scatter around a conditional mean are called **errors**.

The random error of observation i is denoted as ϵ_i. The errors around a conditional mean are normally distributed, centered at 0, and have a variance of σ^2 or $\boldsymbol{\epsilon_i \sim N(0, \sigma^2)}$.

Here, we assume all the conditional distributions of the errors are the same, so we're using a constant variance model.

$$V[Y_i|x_i] = V(\beta_0 + \beta_1 x_i + \epsilon_i) = V(\epsilon_i) = \sigma^2$$

- The model can also be written as:

$$Y_i|x_i \sim N(\underbrace{\beta_0 + \beta_1 x_i}_{\substack{\text{Conditional}\\ \text{mean}}}, \sigma^2)$$

  - mean of Y given x is $\beta_0 + \beta_1 x$ (known as conditional mean)

  - $\beta_0 + \beta_1 x_i$ is the **mean value** of all the Y's for the given value of x_i

The regression line itself represents all the conditional means.

All the observed points will <u>not</u> fall on the line, there is some random noise around the mean (we model this part with an error term).

Usually, we will not know β_0, β_1, or σ^2 so we will estimate them from the data.

- Some interpretation of parameters:
  - β_0 is conditional mean when x=0
  - β_1 is the slope, also stated as the change in mean of Y per 1 unit change in x
  - σ^2 is the variability of responses about the conditional mean

Simple Linear Regression

Assumptions

- Key assumptions

  – linear relationship exists between Y and x

    *we say the relationship between Y and x is linear if the means of the conditional distributions of $Y|x$ lie on a straight line

  – independent errors
    (this essentially equates to independent observations in the case of SLR)

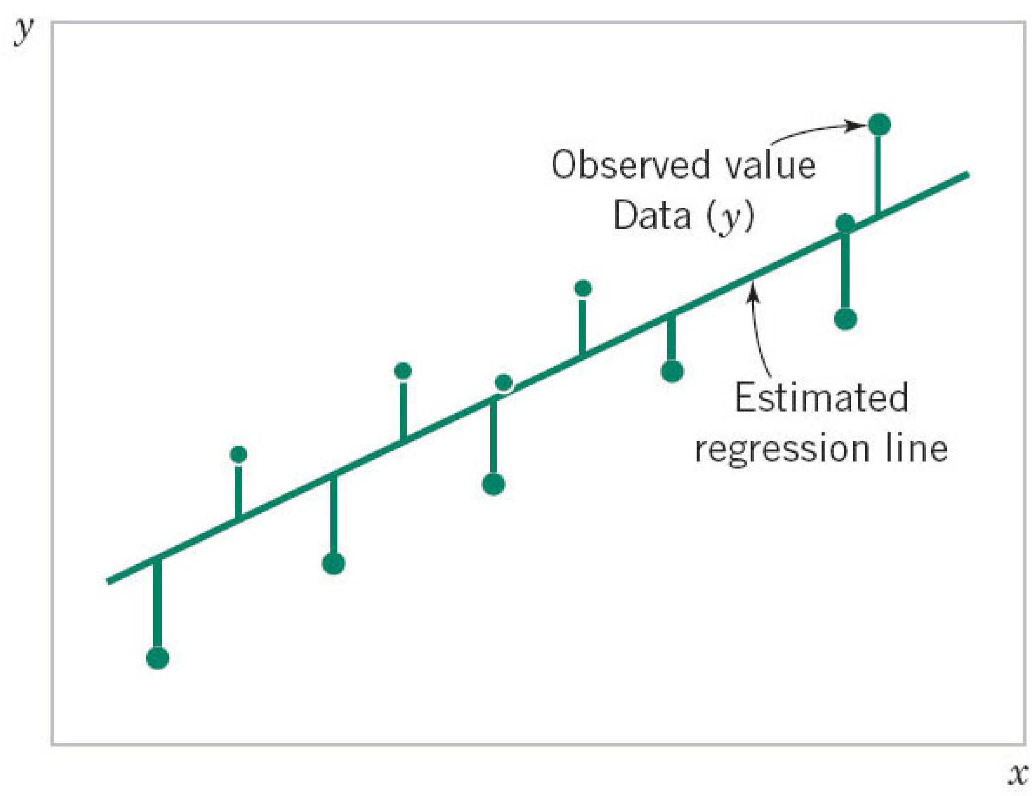
  – constant variance of errors

  – normally distributed errors

Simple Linear Regression
Estimation

We wish to use the sample data to *estimate* the population parameters: the slope β_1 and the intercept β_0

- **Least squares estimation**
  - To choose the 'best fitting line' using least squares estimation, we minimize the sum of the squared vertical distances of each point to the fitted line.



– We let ‘hats’ denote predicted values or estimates of parameters, so we have:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

where $\hat{y}_i$ is the estimated conditional mean for x_i,

$\hat{\beta}_0$ is the estimator for β_0,

and $\hat{\beta}_1$ is the estimator for β_1

– We wish to choose $\hat{\beta}_0$ and $\hat{\beta}_1$ such that we minimize the sum of the squared vertical distances of each point to the fitted line, i.e. minimize $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

– Or minimize the function g:

$$g(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

– This vertical distance of a point from the fitted line is called a **residual**. The residual for observation i is denoted e_i and

$$e_i = y_i - \hat{y}_i$$

– So, in least squares estimation, we wish to minimize the **sum of the squared residuals** (or error sum of squares SS_E).

– To minimize
$$g(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

we take the derivative of g with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$, set equal to zero, and solve.

$$\frac{\partial g}{\partial \hat{\beta}_0} = -2\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)) \;=\; 0$$
$$\frac{\partial g}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))x_i \;=\; 0$$

Simplifying the above gives:

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$
$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} (x_i^2) = \sum_{i=1}^{n} y_i x_i$$

And these two equations are known as the least squares normal equations.

Solving the normal equations gets us our estimators $\hat{\beta}_0$ and $\hat{\beta}_1$...

Simple Linear Regression

Estimation

– Estimate of the slope:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

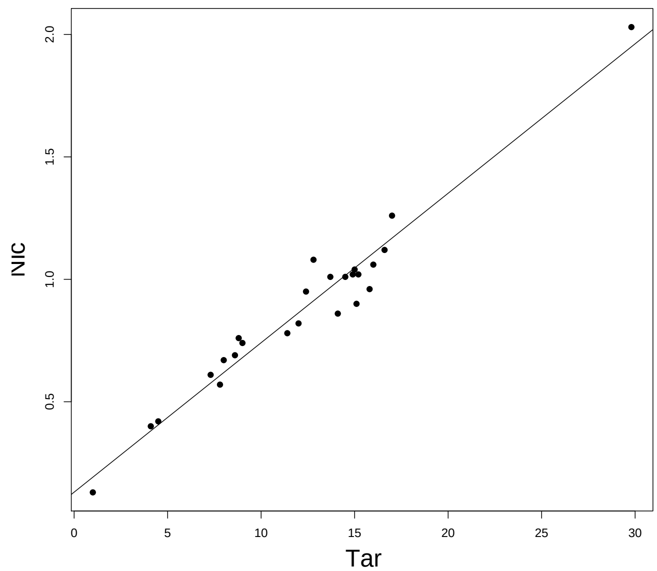
– Estimate of the Y-intercept:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

the point $(\bar{x}, \bar{y})$ will always be on the least squares line

Alternative formulas for $\hat{\beta}_0$ and $\hat{\beta}_1$ are also given in the book.

- **Example**: Cigarette data (Nicotine vs. Tar content)



$n = 25$

Least squares estimates from software:

$\hat{\beta}_0$=0.1309 and $\hat{\beta}_1$=0.0610

Summary statistics:

$\sum_{i=1}^{n} x_i = 305.4 \qquad \bar{x} = 12.216$

$\sum_{i=1}^{n} y_i = 21.91 \qquad \bar{y} = 0.8764$

$\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) = 47.01844$

$\sum_{i=1}^{n}(x_i - \bar{x})^2 = 770.4336$

$\sum_{i=1}^{n} x_i^2 = 4501.2 \quad \sum_{i=1}^{n} y_i^2 = 22.2105$

Using the previous formulas and the summary statistics...

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{47.01844}{770.4336} = 0.061029$$

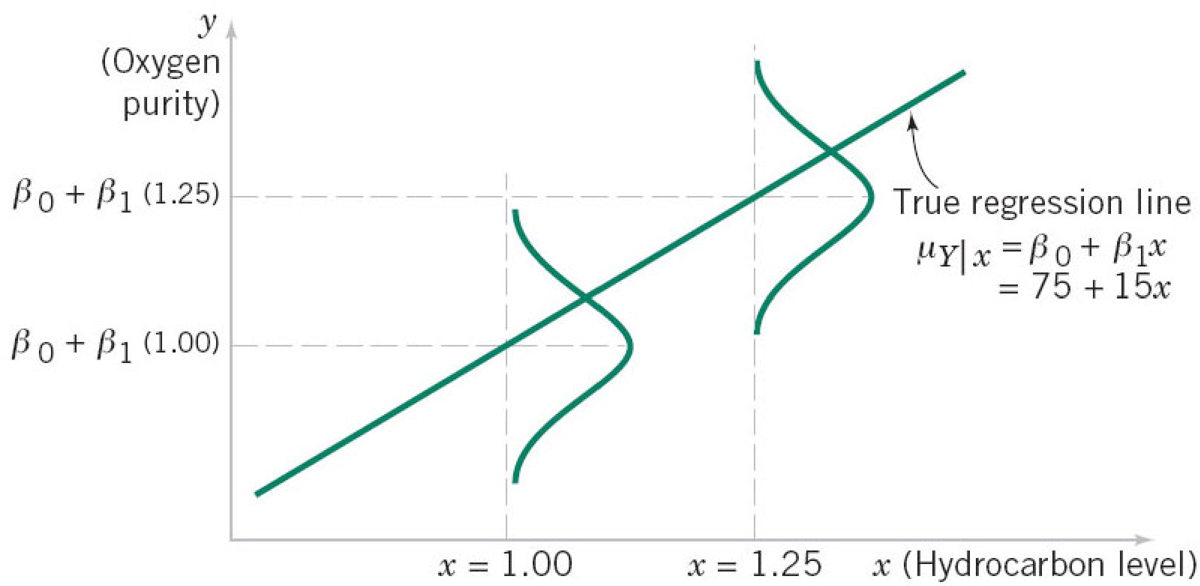
and

$$\begin{aligned}\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ &= 0.8764 - 0.061029(12.216) \\ &= 0.130870\end{aligned}$$

(Same estimates as software)

Simple Linear Regression
Estimating σ^2

- One of the assumptions of simple linear regression is that the variance for each of the conditional distributions of $Y|x$ is the same at all x-values (i.e. constant variance).



- In this case, it makes sense to pool all the observed error information (in the residuals) to come up with a common estimate for σ^2

Recall the model:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{with} \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

– We use the **error sum of squares** (SS_E) to estimate σ^2...

$$\hat{\sigma^2} = \frac{SS_E}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = MSE$$

* SS_E = error sum of squares $=\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

* MSE is the mean squared error

* $E[MSE] = E[\hat{\sigma^2}] = \sigma^2$ (Unbiased estimator)

* $\hat{\sigma} = \sqrt{\hat{\sigma^2}} = \sqrt{MSE}$

* ‘2’ is subtracted from n in the denominator because we’ve used 2 degrees of freedom for estimating the slope and intercept (i.e. there were 2 parameters estimated when modeling the conditional mean)

* When we estimated σ^2 in a single normal population, we divide $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ by $(n-1)$ because we only estimated 1 mean structure parameter which was μ, now we’re estimate two parameters for our mean structure, β_0 and β_1.