

Predictive Analytics (ISE529)

Linear Regression (I)

Dr. Tao Ma

ma.tao@usc.edu

Tue/Thu, May 22 - July 1, 2025, Summer

USC
Viterbi

School of Engineering

Daniel J. Epstein

*Department of Industrial
and Systems Engineering*

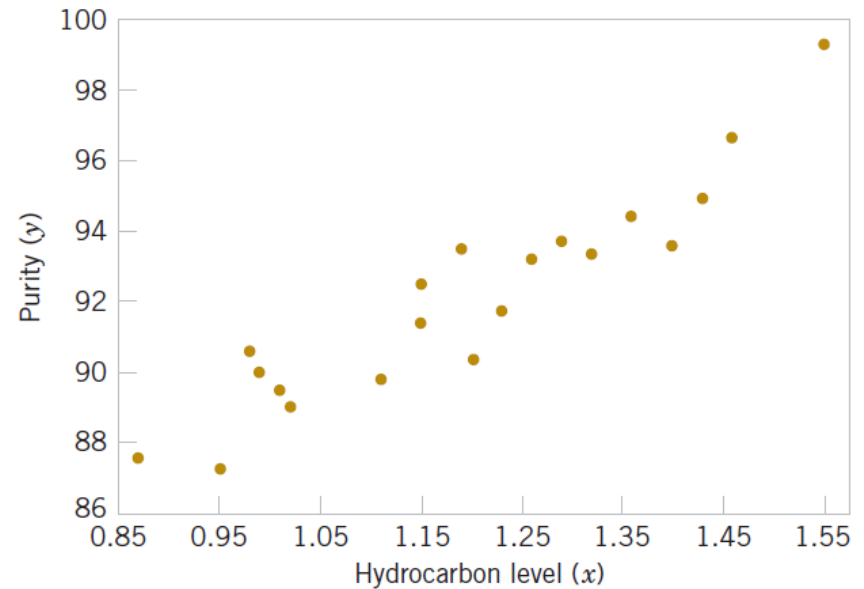


LEAST SQUARE METHOD

- Linear Regression Model
- Hypothesis Tests
- Confidence Intervals
- Prediction
- Model Adequacy Checking
- Correlation

Simple Linear Regression

Obs	Hydrocarbon	Oxygen Purity
#	x (%)	y (%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.4	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.2	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33



Scatter diagram of oxygen purity versus hydrocarbon

Simple Linear Regression

Suppose that the true relationship between Y and x is a straight line and that the observation Y at each level of x is a random variable. We assume that each observation, Y , can be described by the model

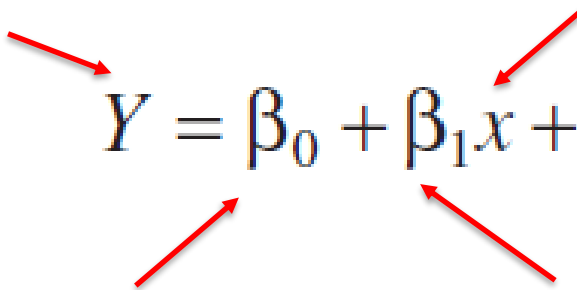
dependent or response variable

regressor or predictor or independent

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

intercept

slope



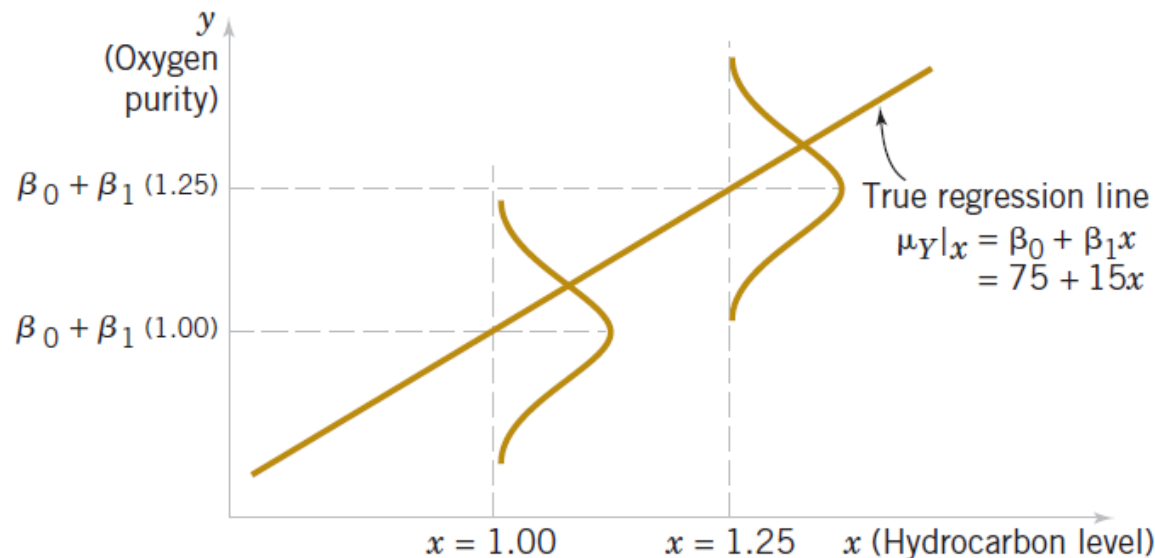
where the intercept β_0 and the slope β_1 are unknown regression coefficients. ϵ is a random error with mean zero and (unknown) variance σ^2 .

Simple Linear Regression

Model structure $Y = \beta_0 + \beta_1 x + \epsilon$

$$E(Y|x) = E(\beta_0 + \beta_1 x + \epsilon) = \beta_0 + \beta_1 x + E(\epsilon) = \beta_0 + \beta_1 x$$

$$V(Y|x) = V(\beta_0 + \beta_1 x + \epsilon) = V(\beta_0 + \beta_1 x) + V(\epsilon) = 0 + \sigma^2 = \sigma^2$$



The distribution of Y for a given value of x
for the oxygen purity-hydrocarbon data

Least Square Method

We call the method for estimating the regression coefficients the **least squares**.

Suppose that we have n pairs of observations $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. The sum of the squares of the deviations of the observations from the true regression line is

$$L = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

Normal equations

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

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

Least Square Estimate

The least squares estimates of the intercept and slope in the simple linear regression model are

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

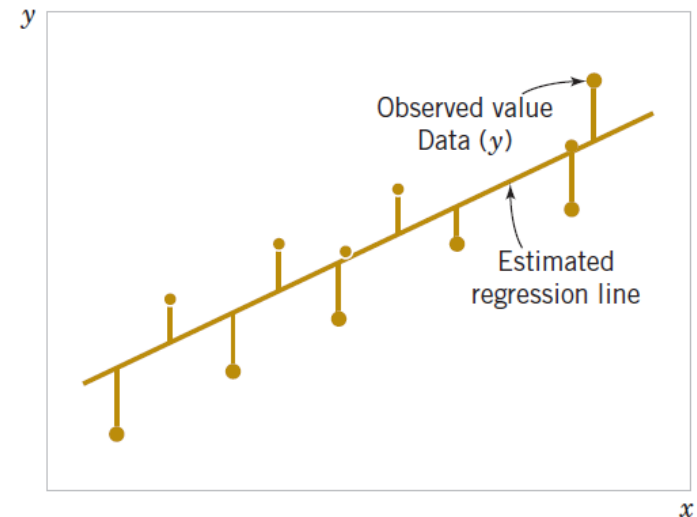
$$\hat{\beta}_1 = \frac{\sum_{i=1}^n y_i x_i - \frac{1}{n} \left(\sum_{i=1}^n y_i \right) \left(\sum_{i=1}^n x_i \right)}{\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2}$$

where $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ and $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$

The fitted or estimated regression line is therefore

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

where $\varepsilon_i = y_i - \hat{y}_i$ is called the **residual**.



Alternative Formula

Let

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}$$

$$S_{xy} = \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) = \sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n}$$

Thus

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

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

Example

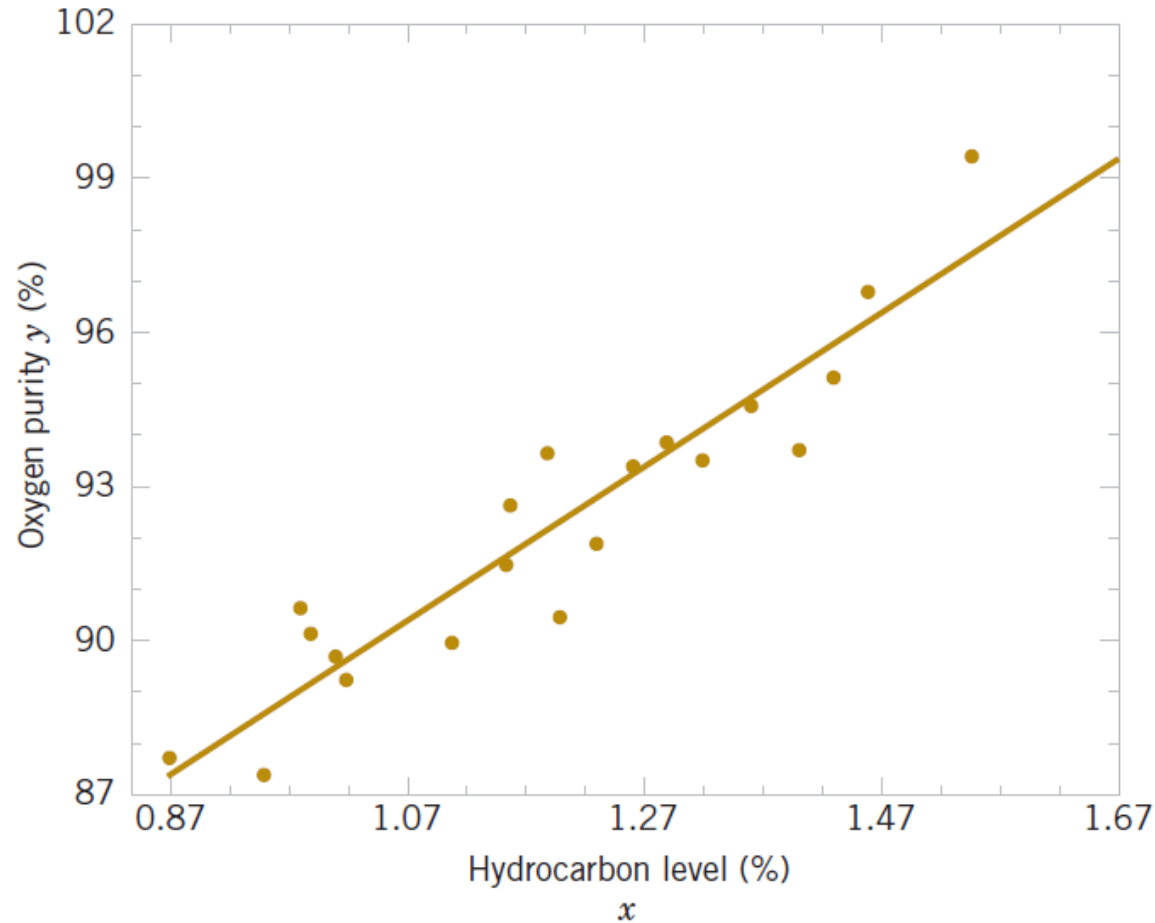
$$n = 20 \quad \sum_{i=1}^{20} x_i = 23.92 \quad \sum_{i=1}^{20} y_i = 1,843.21$$

$$\bar{x} = 1.1960 \quad \bar{y} = 92.1605$$

$$\sum_{i=1}^{20} y_i^2 = 170,044.5321 \quad \sum_{i=1}^{20} x_i^2 = 29.2892$$

$$\sum_{i=1}^{20} x_i y_i = 2,214.6566$$

Example



Scatter plot of oxygen purity y versus hydrocarbon level x
and regression model $\hat{y} = 74.283 + 14.947x$

Estimating σ^2

the **error sum of squares**

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

unbiased estimator of σ^2

$$\hat{\sigma}^2 = \frac{SS_E}{n-2}$$

Hypothesis test

If X_1, X_2, \dots, X_n is a random sample of size n from a normal distribution with unknown mean μ and **unknown variance** σ^2 and if \bar{X} is the sample mean, The random variable

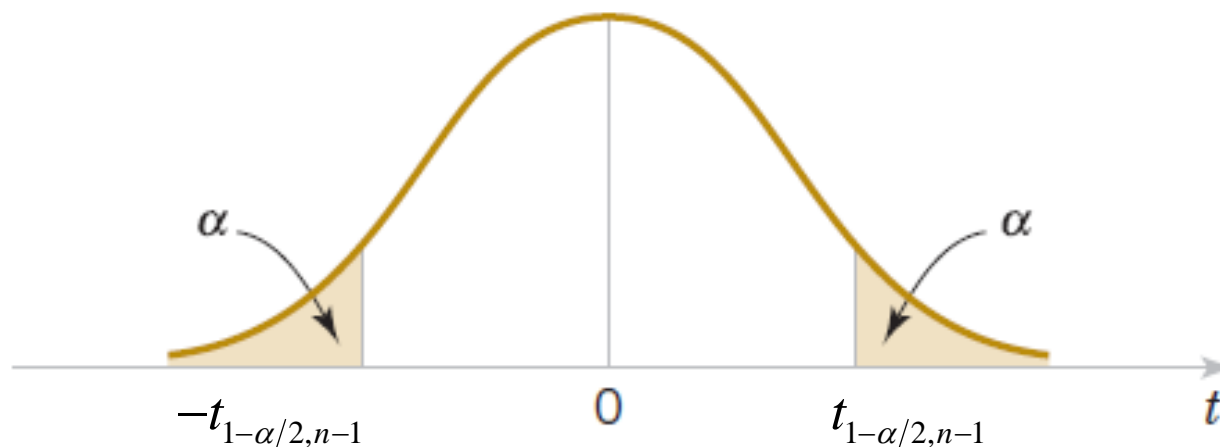
$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

has a **t distribution** with $n - 1$ degrees of freedom.

Hypothesis test

$$P\left(-t_{1-\alpha/2, n-1} \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t_{1-\alpha/2, n-1}\right) = 1 - \alpha$$

where $t_{1-\alpha/2, n-1}$ is the upper $100\alpha / 2$ percentage point of t distribution with $n-1$ degrees of freedom.



Hypothesis test

Make hypothesis:

e.g. $H_0 : \mu = \mu_0$ (null hypothesis regarding the mean)

$H_A : \mu \neq \mu_0$ (alternative hypothesis regarding the means)

$$P\left(-t_{1-\alpha/2, n-1} \leq \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \leq t_{1-\alpha/2, n-1}\right) = 1 - \alpha$$

Hypothesis testing is to test the following probability:

$P(\text{sample data} \mid \text{null hypothesis is true})$

Properties of the Least Squares Estimators

$$E(\hat{\beta}_1) = \beta_1 \quad V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

$$E(\hat{\beta}_0) = \beta_0 \quad \text{and} \quad V(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]$$

In simple linear regression, the estimated standard error of the slope and the estimated standard error of the intercept are

$$se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \quad \text{and} \quad se(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]}$$

Hypothesis

The complete assumptions are that the errors are normally and independently distributed with mean zero and variance σ^2 , abbreviated NID $(0, \sigma^2)$.

$$\begin{array}{ll} H_0: \beta_1 = \beta_{1,0} & \text{Test Statistic for the Slope} \\ H_1: \beta_1 \neq \beta_{1,0} & T_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{se(\hat{\beta}_1)} \end{array}$$

follows the t distribution with $n - 2$ degrees of freedom under H_0

Test Statistic for the Intercept

$$\begin{array}{ll} H_0: \beta_0 = \beta_{0,0} & \\ H_1: \beta_0 \neq \beta_{0,0} & T_0 = \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]}} = \frac{\hat{\beta}_0 - \beta_{0,0}}{se(\hat{\beta}_0)} \end{array}$$

Oxygen Purity Tests of Coefficients

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

$$\hat{\beta}_1 = 14.947 \quad n = 20, \quad S_{xx} = 0.68088, \quad \hat{\sigma}^2 = 1.18$$

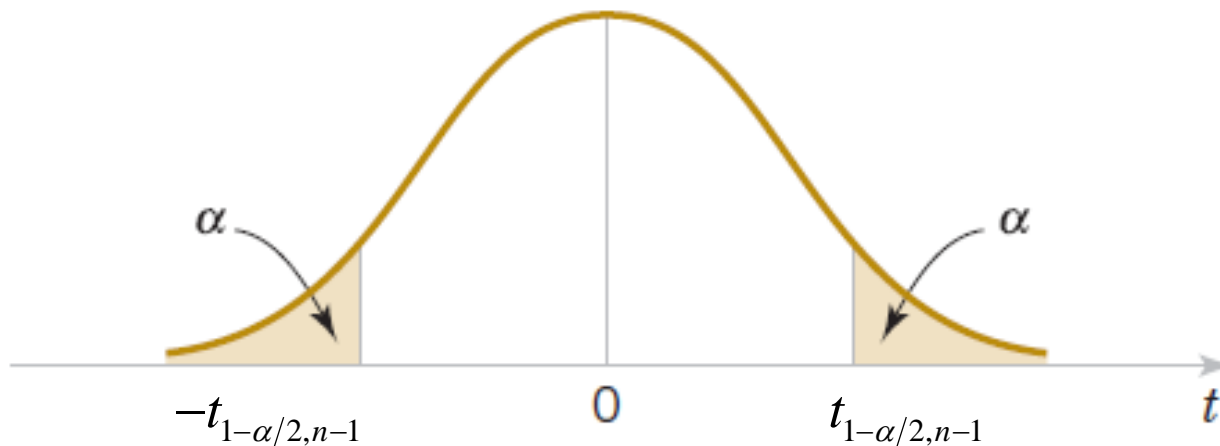
Confidence Interval

Letting $t_{1-\alpha/2, n-1}$ be the upper 100(1- α /2) percentage point of the t distribution with $n - 1$ degrees of freedom, we may write

$$P\left(-t_{1-\alpha/2, n-1} \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t_{1-\alpha/2, n-1}\right) = 1 - \alpha$$

Rearranging this equation yields

$$P\left(\bar{X} - t_{1-\alpha/2, n-1} S/\sqrt{n} \leq \mu \leq \bar{X} + t_{1-\alpha/2, n-1} S/\sqrt{n}\right) = 1 - \alpha$$



Confidence Intervals on **Parameters**

Under the assumption that the observations are normally and independently distributed, a $100(1 - \alpha)\%$ confidence interval on the slope β_1 in simple linear regression is

$$\hat{\beta}_1 \pm t_{\alpha/2, n-2} \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

Similarly, a $100(1 - \alpha)\%$ confidence interval on the intercept β_0 is

$$\hat{\beta}_0 \pm t_{\alpha/2, n-2} \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]}$$

Confidence Interval on the **Mean Response**

$$\hat{\mu}_{Y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

$$V(\hat{\mu}_{Y|x_0}) = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]$$

$$\frac{\hat{\mu}_{Y|x_0} - \mu_{Y|x_0}}{\sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]}}$$

has a t distribution with $n - 2$ degrees of freedom

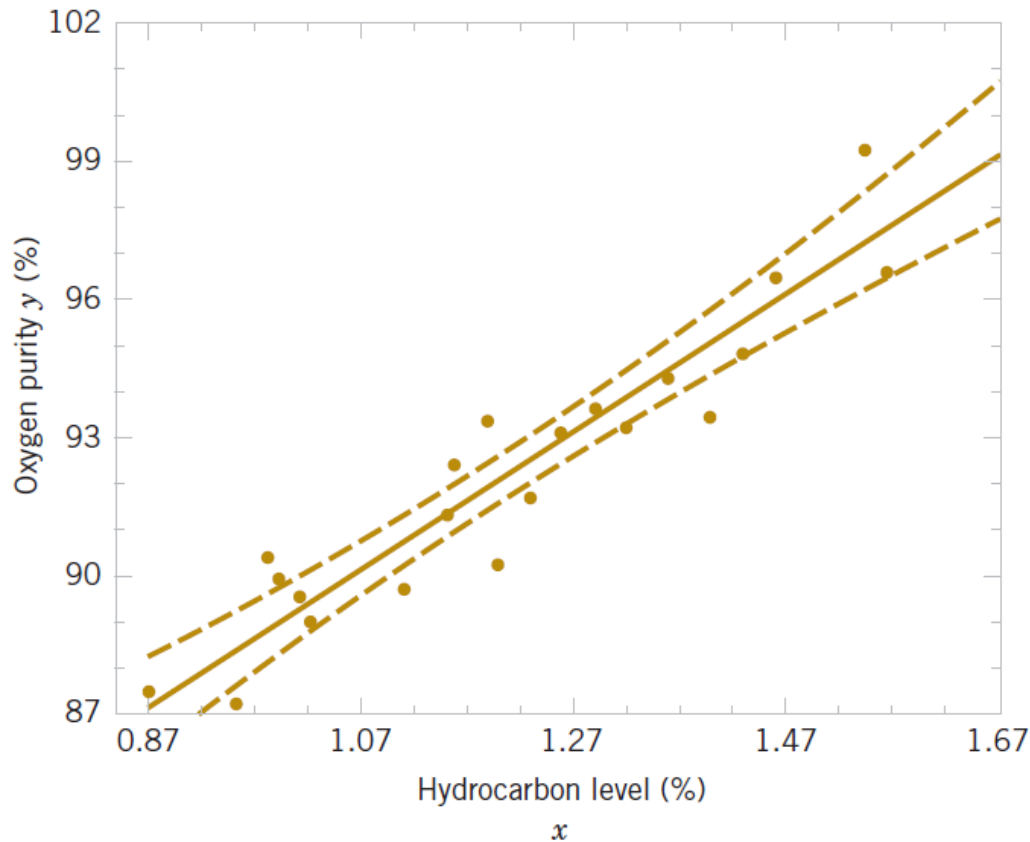
A $100(1 - \alpha)\%$ confidence interval on the mean response at the value of $x = x_0$, say $\mu_{Y|x_0}$, is given by

$$\hat{\mu}_{Y|x_0} \pm t_{\alpha/2, n-2} \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]}$$

where $\hat{\mu}_{Y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$ is computed from the fitted regression model.

Confidence interval

Confidence Interval on the Mean Response



Scatter diagram of oxygen purity data with fitted regression line and 95 percent confidence limits on

Prediction

Note that the error in prediction $e_{\hat{p}} = Y_0 - \hat{Y}_0$

is a normally distributed random variable with mean zero and variance

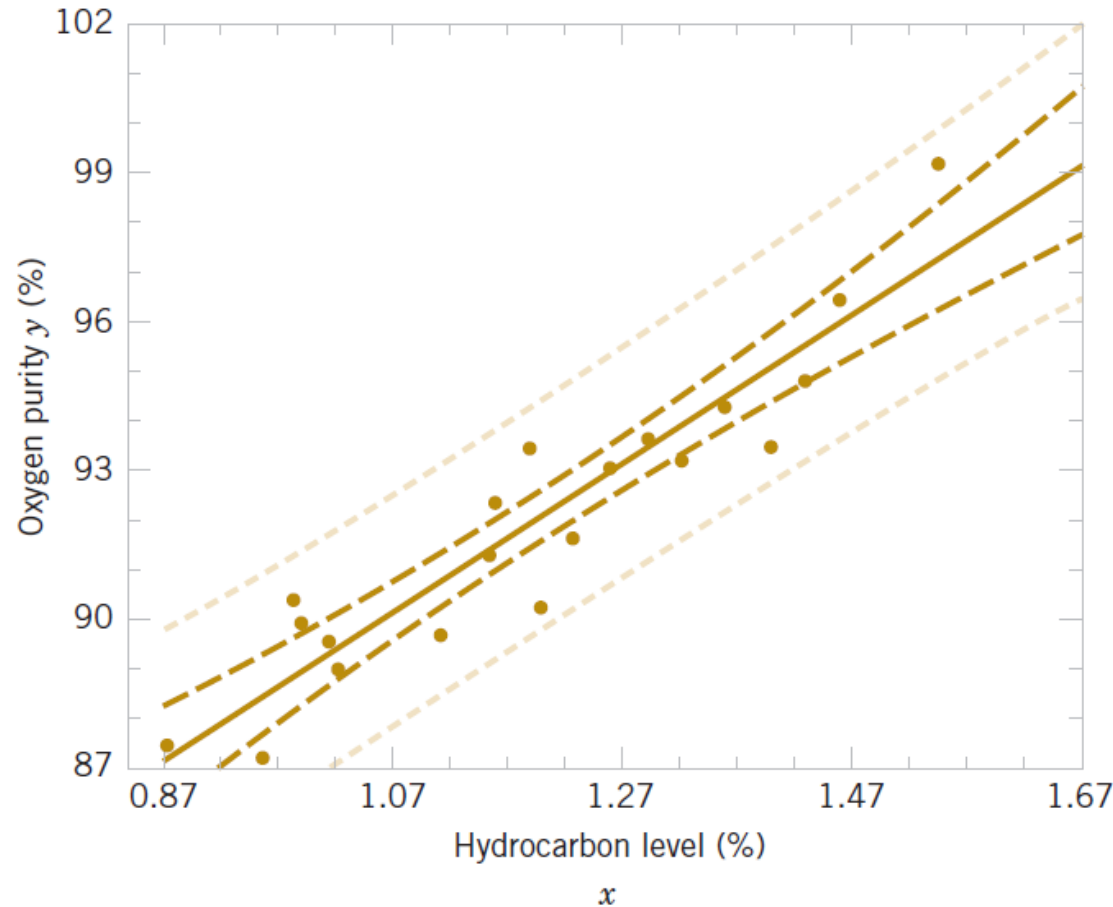
$$V(e_{\hat{p}}) = V(Y_0 - \hat{Y}_0) = \alpha^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]$$

If we use $\hat{\sigma}^2$ to estimate α^2 , we can show that

A $100(1 - \alpha)\%$ prediction interval on a future observation Y_0 at the value x_0 is given by

$$\hat{y}_0 \pm t_{\alpha/2, n-2} \sqrt{\hat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]}$$

The value \hat{y}_0 is computed from the regression model $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$



the **prediction** interval at the point x_0 is always wider than the confidence interval at x_0 . because the prediction interval depends on both the error from the fitted model and the error associated with future observations.

Model Adequacy Checking

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

the **regression sum of squares** $SS_R = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$

the **total corrected sum of squares** of y $SS_T = \sum_{i=1}^n (y_i - \bar{y})^2$

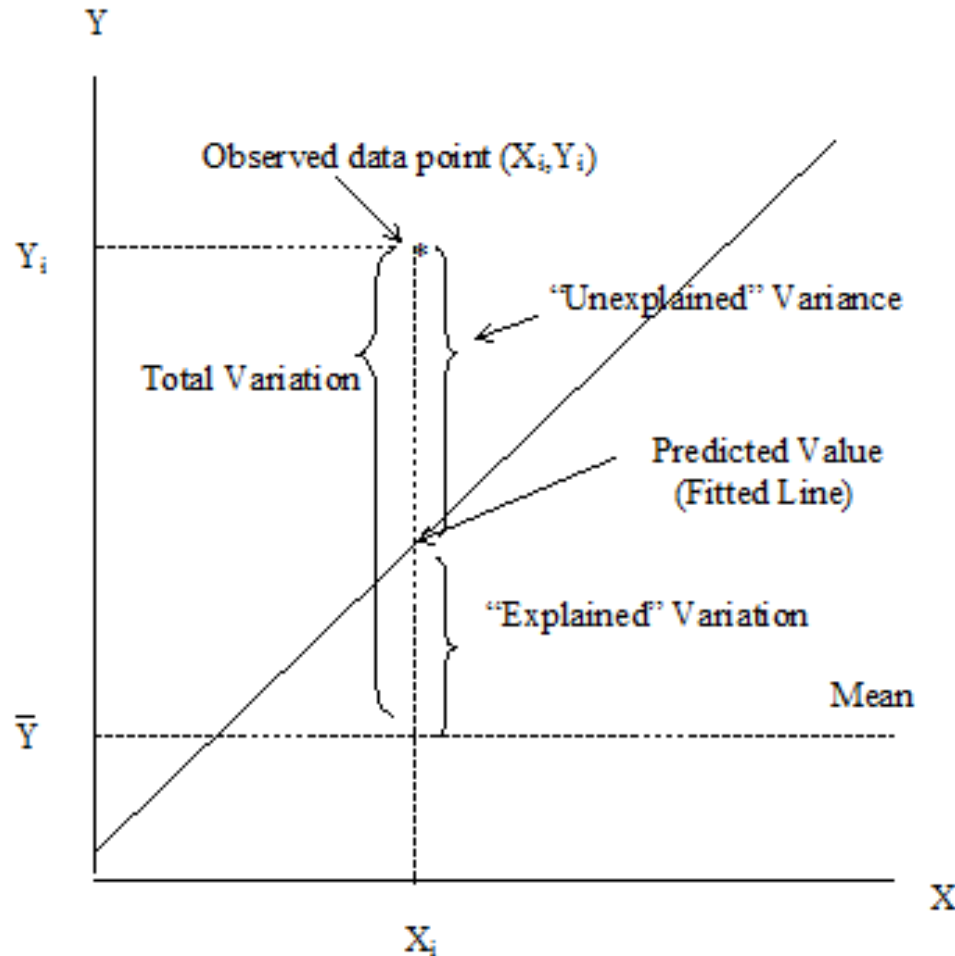
Analysis of Variance (ANOVA)

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

$$SS_T = SS_R + SS_E$$

Model Adequacy Checking

The **coefficient of determination** is $R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$



Model Adequacy Checking

Test for Significance of Regression

If the null hypothesis $H_0: \beta_1 = 0$ is true, the statistic

$$F_0 = \frac{SS_R / 1}{SS_E / (n - 2)} = \frac{MS_R}{MS_E}$$

follows the $F_{1,n-2}$ distribution, and we would reject H_0 if $f_0 > f_{\alpha,1,n-2}$.

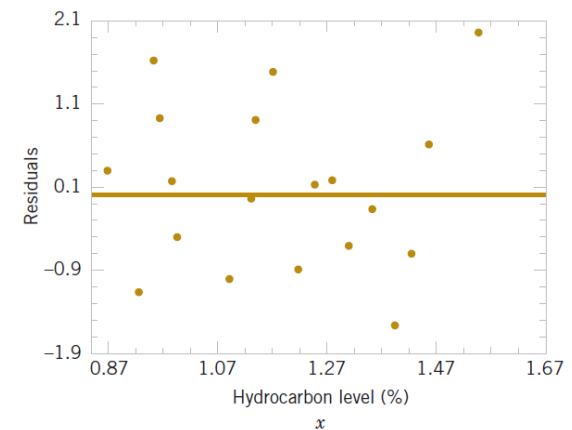
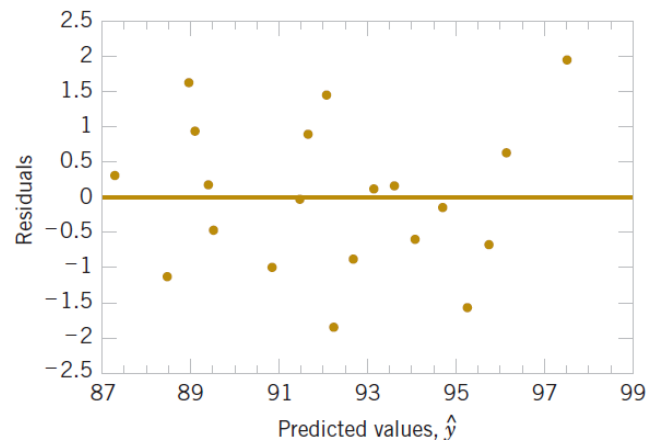
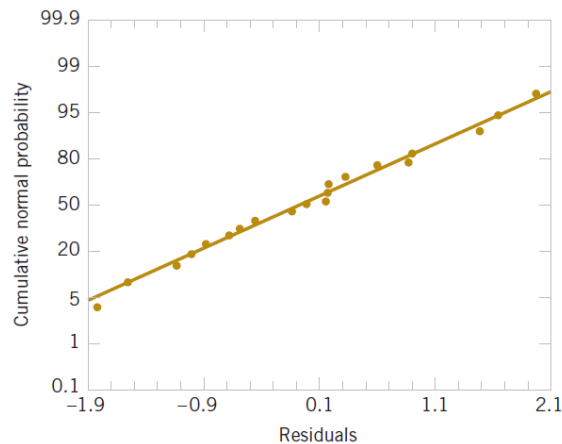
Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	$SS_R = \hat{\beta}_1 S_{xy}$	1	MS_R	MS_R / MS_E
Error	$SS_E = SS_T - \hat{\beta}_1 S_{xy}$	$n - 2$	MS_E	
Total	SS_T	$n - 1$		

Note that $MS_E = \hat{\sigma}^2$.

Model Adequacy Checking

Residual analysis, checking assumptions

1. The errors are uncorrelated random variables with mean zero and constant variance
2. Tests of hypotheses and interval estimation require that the errors be normally distributed.
 - a normal probability plot of residuals
 - plot the residuals against the \hat{y}_i and against the independent variable x .



Both X and Y are random variables, assumed that the observations (X_i, Y_i) , $i = 1, 2, \dots, n$ are jointly distributed random variables obtained from the distribution $f(x, y)$

μ_Y and σ_Y^2 are the mean and variance of Y , μ_X , σ_X^2 are the mean and variance of X , The **correlation coefficient** between Y and X is defined as

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

where σ_{XY} is the covariance between Y and X

Regression model estimators can be written as

$$\beta_0 = \mu_Y - \mu_X \rho \frac{\sigma_Y}{\sigma_X}$$

$$\beta_1 = \frac{\sigma_Y}{\sigma_X} \rho$$

Example

Regression methods were used to analyze the data from a study investigating the relationship between roadway surface temperature (x) and pavement deflection (y). Summary quantities were

$$n = 20, \Sigma y_i = 12.75, \Sigma y_i^2 = 8.86, \Sigma x_i = 1478, \Sigma x_i^2 = 143,215.8, \text{ and } \Sigma x_i y_i = 1083.67.$$

- Calculate the least squares estimates of the slope and intercept. Estimate σ^2 .
- What is the mean pavement deflection when the surface temperature is 90°F?
- What change in mean pavement deflection would be expected for a 1°F change in surface temperature?
- Find a 99% confidence interval on slope.

$$\hat{y} = 0.32999 + 0.00416x \quad \hat{\sigma}^2 = MS_E = \frac{SS_E}{n-2} = \frac{0.143275}{18} = 0.00796$$

$$\hat{y} = 0.32999 + 0.00416(90) = 0.7044$$

$$\hat{\beta}_1 \pm (t_{0.005,18})se(\hat{\beta}_1) \quad t_{\alpha/2, n-2} = t_{0.005,18} = 2.878$$

$$0.0041612 \pm (2.878)(0.000484)$$

$$0.0027682 \leq \beta_1 \leq 0.0055542$$

Example

The regression equation is

$$Y = 12.9 + 2.34 x$$

Predictor	Coef	SE Coef	T
Constant	12.857	1.032	?
X	2.3445	0.1150	?
S = 1.48111 R-sq = 98.1% R-sq(adj) = 97.9%			

Analysis of Variance

Source	DF	SS	MS	F
Regression	1	912.43	912.43	?
Residual error	?	17.55	?	
Total	9	929.98		

- (a) Fill in the missing information.
- (b) Use 3 ways to check that the model defines a useful linear relationship, $\alpha = 0.05$?
- (c) What is your estimate of σ^2 ?

$$T_0 = \frac{\hat{\beta}_0 - \beta_0}{se(\beta_0)} = \frac{12.857}{1.032} = 12.4583$$

$$F_0 = \frac{MS_R}{MS_E} = \frac{912.43}{2.1938} = 415.913$$

$$T_1 = \frac{\hat{\beta}_1 - \beta_1}{se(\beta_1)} = \frac{2.3445}{0.115} = 20.387$$

$$MS_E = \frac{SS_E}{n - 2} = \frac{17.55}{8} = 2.1938$$

Example Code

- import standard libraries at this top level.
- import only a few items from a given module, help keep the "namespace" clean.
- inserted a line break `\` to break a long line into multi-lines to ease readability.
- use `dir()` to show the attributes of the object as well as any methods associated with it.

```
import numpy as np
import pandas as pd
from matplotlib.pyplot import subplots
import statsmodels.api as sm
from statsmodels.stats.outliers_influence \
    import variance_inflation_factor as VIF
from statsmodels.stats.anova import anova_lm
```

```
A = np.array([3,5,11])
dir(A)
```

```
A[0:2]
```

```
A.sum?
```

```
A.sum()
```

```
matrix1 = [[12,7,3,2],
            [4 ,5,6,4],
            [7 ,8,9,6]]
matrix1
```

```
matrix2 = [[5,8,1],
            [6,7,3],
            [4,5,9],
            [3,7,13]]
matrix2
```

```
type(matrix1)
```

```
res = [[0 for x in range(3)] for y in range(3)]
res
```

```
len(matrix1)
```

```
matrix2[0]
```

```
len(matrix2[0])
```

```
len(matrix2)
```

```
for i in range(len(matrix1)):
    for j in range(len(matrix2[0])):
        for k in range(len(matrix2)):

            # resulted matrix
            res[i][j] += matrix1[i][k] * matrix2[k][j]

print (res)
res = [[0 for x in range(3)] for y in range(3)]
```

Use numpy

```
res = np.dot(matrix1,matrix2)
print(res)
```


Example Code

Simple Linear Regression

- response variable = `medv` (median house value)
- predictor = `lstat` (percent of households with low socioeconomic status)
- predict `medv`
- use `statsmodels` to implement regression methods.

```
Boston = pd.read_csv("Boston.csv")
Boston.columns
Boston['medv'][1:5]
```

```
type(Boston)
```

```
Boston[1:5]['medv']
```

```
Boston.iloc[5:8,2:6]
```

```
X = pd.DataFrame({'intercept': np.ones(Boston.shape[0]),
                  'lstat': Boston['lstat']})
X.iloc[5:8,0:3]
```

```
X["lstat"][1:5]
```

```
y = Boston['medv']
y[0:5]
```

```
model = sm.OLS(y, X)
results = model.fit()
```

```
results.summary()
```

```
results.summary2()
```

```
dir(results)
```

```
newdata = pd.DataFrame({'intercept': np.ones(3),
                        'lstat': [5, 10, 15]})
newdata.iloc[0:8,0:3]
```

```
results.predict(newdata)
```

```
new_predictions = results.get_prediction(newdata);
```

```
new_predictions.predicted_mean
```

Do matrix multiplication using Numpy

```
B = results.params
B
```

```
type(B)
```

```
predictions = np.dot(newdata,B)
predictions
```

use `@` for matrix multiplication

```
newdata@B
```

```
new_predictions.conf_int(alpha=0.05)
```

```
new_predictions.conf_int(obs=True, alpha=0.05)
```

MAXIMUM LIKELIHOOD METHOD

Method of Maximum Likelihood

- Suppose that X is a random variable with probability mass or density function $f(x / \theta)$, where θ is unknown parameters. Let x_1, x_2, \dots, x_n be the observed values in a random sample of size n . Then the **likelihood function** of the sample is:

$$L(\theta) = f(x_1 | \theta) \cdot f(x_2 | \theta) \cdot \dots \cdot f(x_n | \theta)$$

- Note that the likelihood function is now a function of only the unknown parameters θ . The **maximum likelihood estimator** (MLE) of θ is the value of θ that maximizes the likelihood function $L(\theta)$. Intuitively, it is the value of θ that makes the observed data “most probable” or “most likely”.

Optimization – the 1st order partial derivative

$$\frac{\partial L(x_1, x_2, \dots, x_n \mid \theta)}{\partial \theta} = 0$$

Because the likelihood function is a product function, it is more convenient to maximize the logarithm of the likelihood function; i.e.,

$$\frac{\partial \log L(x_1, x_2, \dots, x_n \mid \theta)}{\partial \theta} = 0$$

Example

Let $y_1, y_2, \dots, y_n \sim N(\mu, \sigma^2)$, i.e., the density function is

$$f(y_i | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_i - \mu)^2}{2\sigma^2}\right]$$

Also recall, $\mu = E(y_i | x_i) = f(x_i) = \beta_0 + \beta_1 x_i$.

Then the likelihood function of a random sample of size n is

$$\begin{aligned} L(y_i | \mu, \sigma^2) &= f(y_1 | \mu, \sigma^2) f(y_2 | \mu, \sigma^2) \cdots f(y_n | \mu, \sigma^2) \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2\right] \end{aligned}$$

Therefore, the log-likelihood is

$$l(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

Example

Take the 1st order derivative and set it to zero, we get

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 = 0$$

$$\frac{\partial l}{\partial \beta_0} = \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial l}{\partial \beta_1} = \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) x_i = 0$$

Example

Take the 1st order derivative and set it to zero, we get

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2 = \frac{1}{n} \sum_{i=1}^n \varepsilon_i^2$$

$$\hat{\beta}_1 = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}$$

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