Lecture 13
Linear Regression: Model Diagnosis

Fall 2013

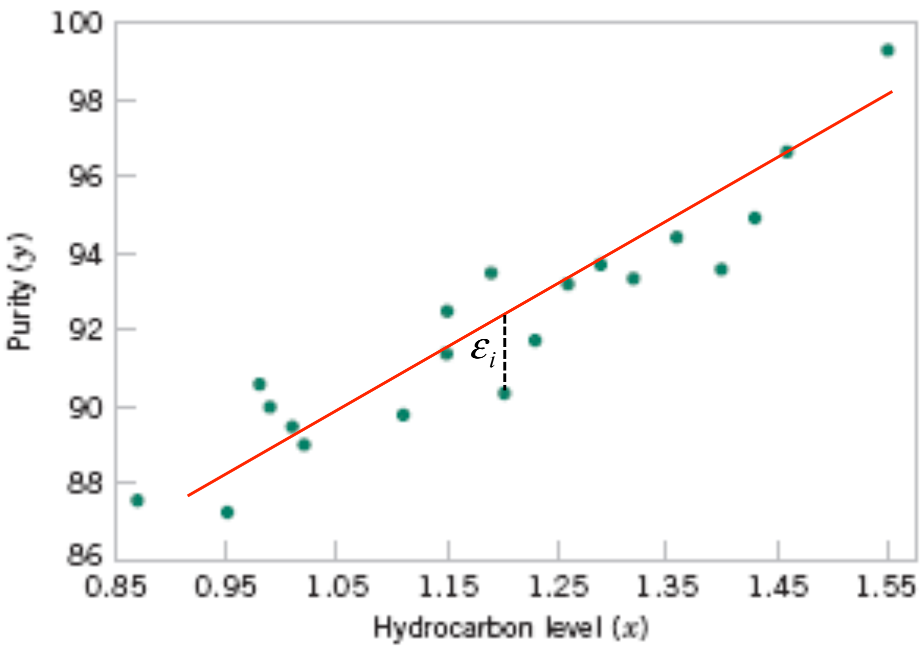
Prof. Yao Xie, yao.xie@isye.gatech.edu
H. Milton Stewart School of Industrial Systems & Engineering
Georgia Tech

Outline

- ANOVA to test $\beta_1 = 0$?
- Mean response and confidence interval
- Prediction of new observations
- Diagnosis of regression model

Simple linear regression

Based on the scatter diagram, it is probably reasonable to assume that the mean of the random variable Y is related to X by the following simple linear regression model:



Response

Regressor or Predictor

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad i = 1, 2, \cdots, n$$

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

Intercept

Slope

Random error

where the slope and intercept of the line are called regression coefficients.

•The case of simple linear regression considers a single regressor or predictor x and a dependent or response variable Y.

Regression coefficients

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} \qquad (11\text{-}10)$$

$$S_{xy} = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n} \qquad (11\text{-}11)$$

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

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

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

Fitted (estimated) regression model

Caveat: regression relationship are valid only for values of the regressor variable within the range the original data. Be careful with extrapolation.

Test for slope - method 1: Use t-test for slope

Under H_0

slope parameter β_1
$$E(\hat{\beta}_1) = \beta_{1,0}$$
$$V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

$$\hat{\beta}_1 \sim N\left(\beta_{1,0}, \quad \sigma^2 / S_{xx} \right)$$

- Under H_0, test statistic

$$T_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2 / S_{xx}}}$$

~ t distribution with n-2 degree of freedom

- Reject H_0 if

$$|t_0| > t_{\alpha/2, n-2}$$

(two-sided test)

Test for slope - method 2: Analysis of variance (ANOVA)

- ANOVA can be used to test for significance of regression
- Partition the total variability in the response variable into meaningful
- Analysis of variance identity

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

error sum of squares
SS_R

regression sum of squares
SS_E

$$SS_T = SS_R + SS_E$$

ANOVA continues …

- Intuition: if the null hypothesis H_0: $\beta_1 = 0$ is true

$\beta_1 = 0$ $\quad Y = \beta_0 + \epsilon$

$$SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$ "small"

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

$$SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$ "large"

- Test statistic

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

reject H_0 if $f_0 > f_{\alpha,1,n-2}$

…doing calculation

$$SS_T = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

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

$$SS_E = SS_T - \hat{\beta}_1 S_{xy}$$

$$SS_R = \hat{\beta}_1 S_{xy}$$

Example: oxygen purity

- Test whether or not **purity** is related to **carbonhydron concentration**

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = SS_T = 173.38$$

$$\hat{\beta}_1 = 14.947$$

$$SS_R = \hat{\beta}_1 S_{xy} = (14.947)10.17744 = 152.13$$

$$SS_E = SS_T - SS_R = 173.38 - 152.13 = 21.25$$

- value of test statistic

$$f_0 = MS_R/MS_E = 152.13/1.18 = 128.86,$$

- p-value $P \simeq 1.23 \times 10^{-9}$

ANOVA will lead to the same conclusion as t-test.

Outline

- ANOVA to test $\beta_1 = 0$?
- Mean response and confidence interval
- Prediction of new observations
- Diagnosis of regression model

Mean response

- Observation

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

Y: response

x: predictor

- Mean **response** $E(Y|x) = \mu_{Y|x} = \beta_0 + \beta_1 x$

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

- Variance of **response**

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

Example of response

- Oxygen purity vs carbon hydrolevel

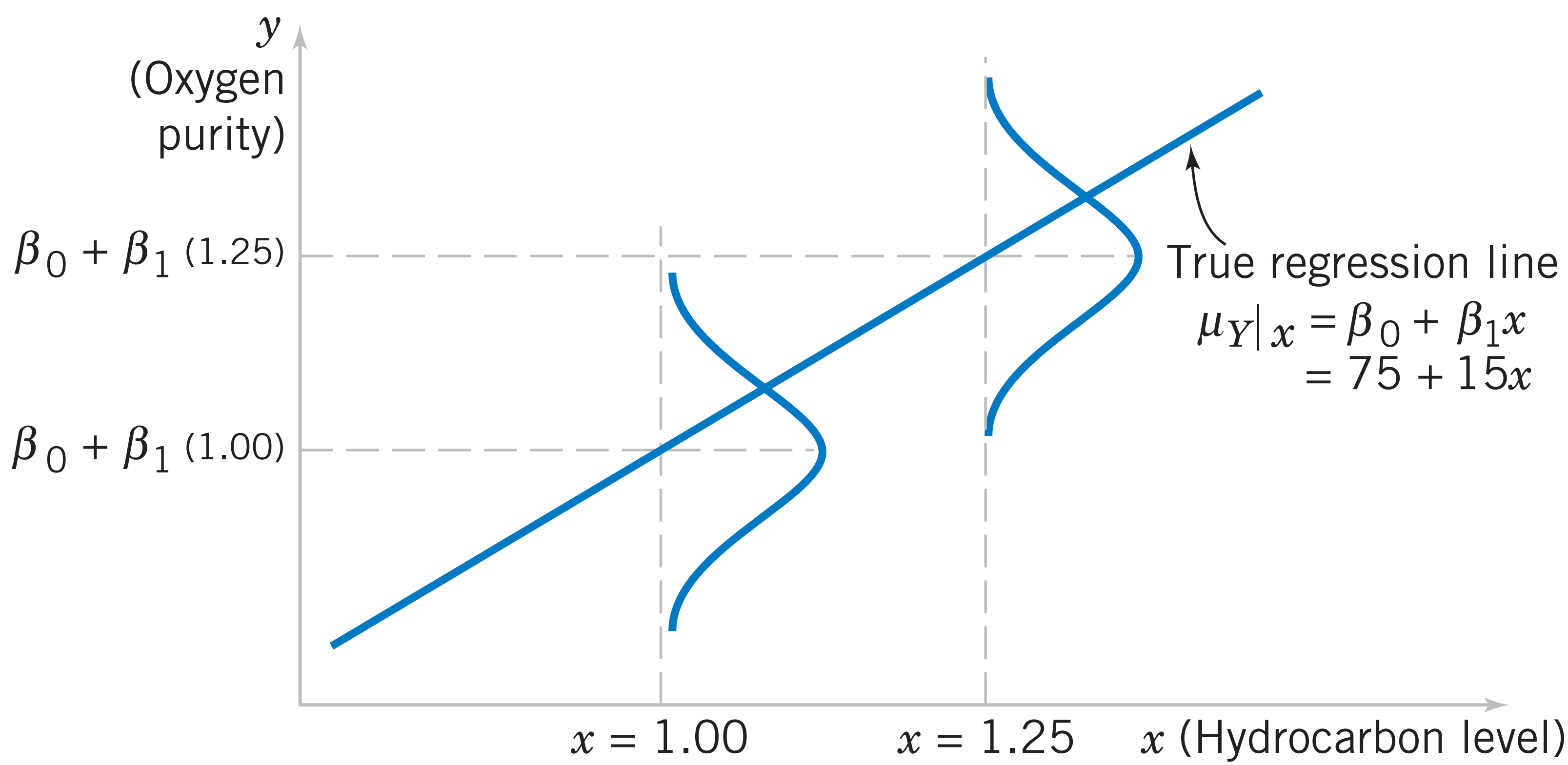


Figure 11-2 The distribution of Y for a given value of x for the oxygen purity–hydrocarbon data.

if $x = 1.25$ $\quad Y$ has mean value $\mu_{Y|x} = 75 + 15(1.25) = 93.75$ and variance 2

Confidence interval of mean response

- A confidence interval can be constructed on mean response of a specified value of x

$$E(Y|x_0) = \mu_{Y|x_0}$$

- Also called the confidence interval about regression line
- Step 1: point estimator for mean response

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

- Step 2: variance of mean response

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

- to construct confidence interval, replace

$\hat{\sigma}^2$ use as an estimate of σ^2

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

- constructed confidence interval for mean response

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

$$\hat{\mu}_{Y|x_0} - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right]} \le \mu_{Y|x_0} \le \hat{\mu}_{Y|x_0} + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right]} \qquad (11\text{-}31)$$

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

Example: oxygen purity

$$\hat{\mu}_{Y|x_0} = 74.283 + 14.947x_0$$

95% confidence interval is given by

$$\hat{\mu}_{Y|x_0} \pm 2.101\sqrt{1.18\left[\frac{1}{20} + \frac{(x_0 - 1.1960)^2}{0.68088}\right]}$$

To use this:

predicting mean oxygen purity when $x_0 = 1.00\%$

$$\hat{\mu}_{Y|x_{1.00}} = 74.283 + 14.947(1.00) = 89.23$$

the 95% confidence interval is

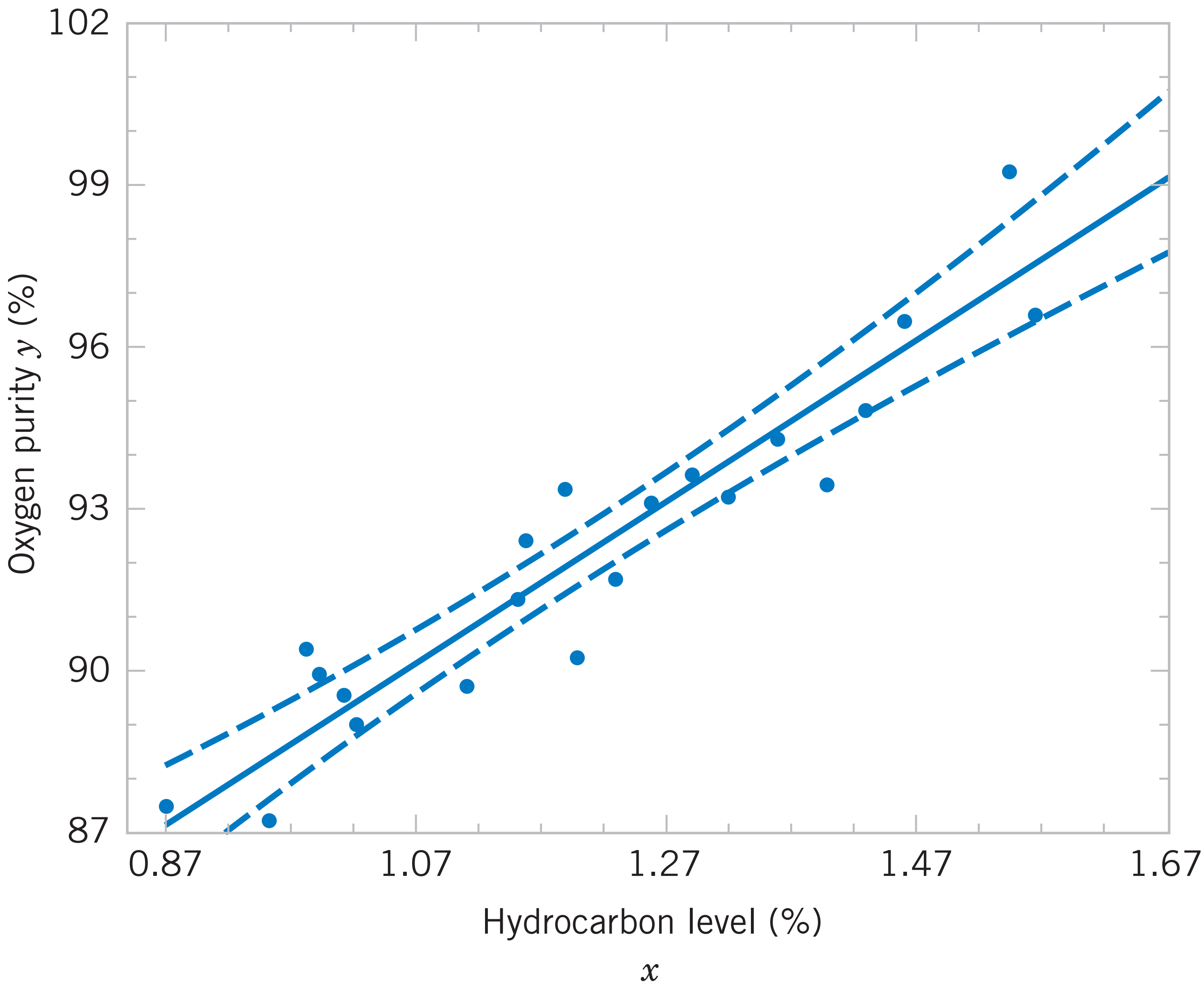
$$89.23 \pm 2.101\sqrt{1.18\left[\frac{1}{20} + \frac{(1.00 - 1.1960)^2}{0.68088}\right]}$$

$$89.23 \pm 0.75$$

the 95% CI on $\mu_{Y|1.00}$ is

$$88.48 \leq \mu_{Y|1.00} \leq 89.98$$

Confidence interval on mean response: plotted



Outline

- ANOVA to test $\beta_1 = 0$?
- Mean response and confidence interval
- Prediction of new observations
- Diagnosis of regression model

Implication: predicting new observations

- Use the fitted linear regression line to predict new observation

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

$$e_{\hat{p}} = Y_0 - \hat{Y}_0$$

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

Compare with variance of mean response:

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

Confidence interval of predicted new values

$$\frac{Y_0 - \hat{Y}_0}{\sqrt{\hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}} \sim T_{n-2}$$

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

$$\hat{y}_0 - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]} \le Y_0 \le \hat{y}_0 + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]} \qquad (11\text{-}33)$$

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

Estimation of variance

- Using the fitted model, we can estimate value of the response variable for given predictor

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

- Residuals: $r_i = y_i - \hat{y}_i$
- Our model: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, $i = 1,\ldots,n$, $Var(\varepsilon_i) = \sigma^2$
- Unbiased estimator (MSE: Mean Square Error)

$$\hat{\sigma}^2 = MSE = \frac{\sum_{i=1}^{n} r_i^2}{n-2}$$

Example: oxygen purity

- 95% confidence interval for prediction at x0 = 1%

$$\hat{y}_0 = 89.23$$

$$89.23 - 2.101\sqrt{1.18\left[1 + \frac{1}{20} + \frac{(1.00 - 1.1960)^2}{0.68088}\right]}$$

$$\leq Y_0 \leq 89.23 + 2.101\sqrt{1.18\left[1 + \frac{1}{20} + \frac{(1.00 - 1.1960)^2}{0.68088}\right]}$$

$$86.83 \leq y_0 \leq 91.63$$

Confidence interval on predicted observations: plotted

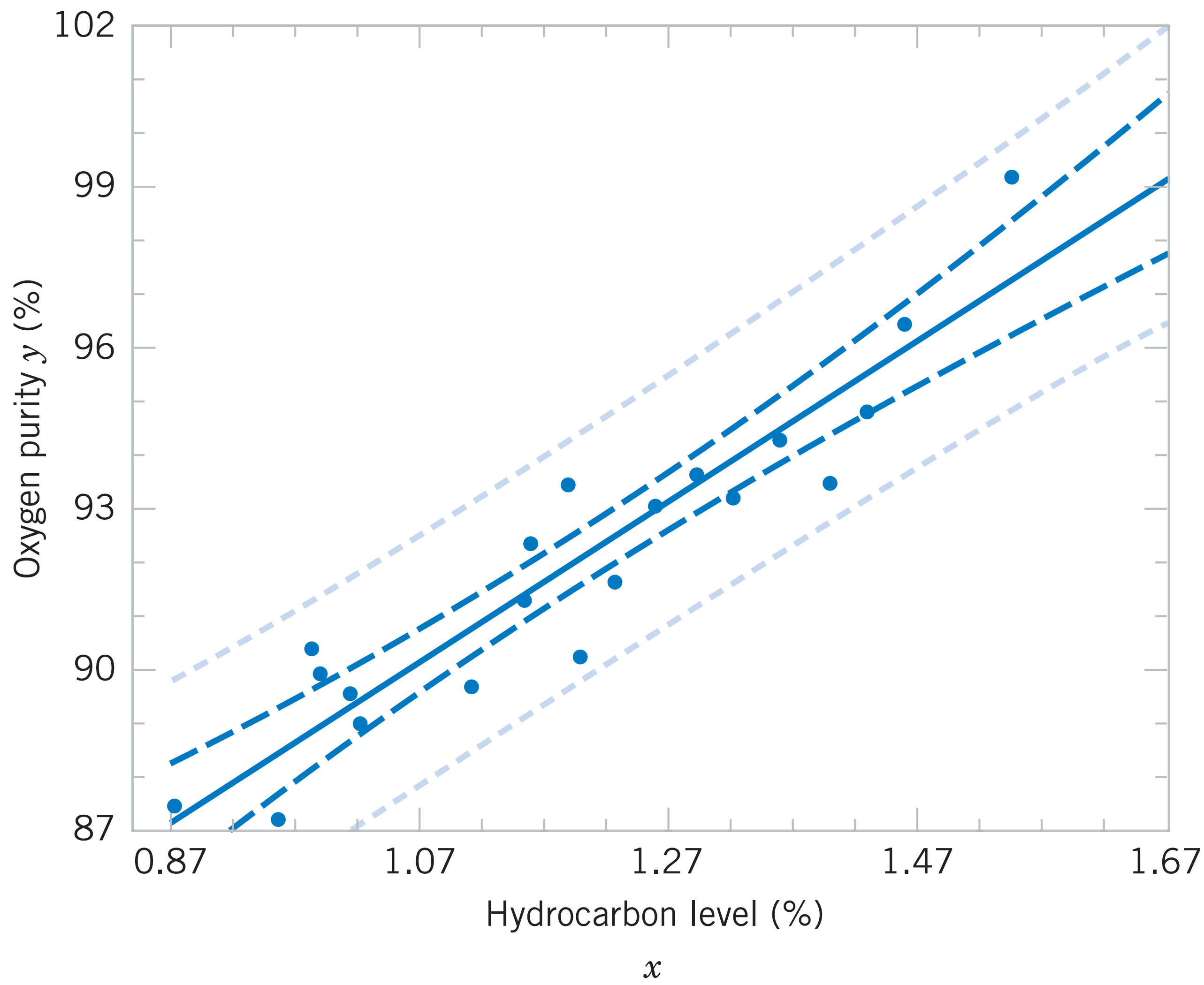


Figure 11-8 Scatter diagram of oxygen purity data from Example 11-1 with fitted regression line, 95% prediction limits (outer lines) and 95% confidence limits on $\mu_{Y|x_0}$.

Outline

- ANOVA to test $\beta_1 = 0$?
- Mean response and confidence interval
- Prediction of new observations
- Diagnosis of regression model

Diagnosis for linear regression

- We have made various assumptions for linear regression models
- Diagnosis of linear regression model: examine these assumptions using various statistical tools
- Assumptions:
  - Estimation: errors are **uncorrelated**
  - Test hypothesis: errors are **normally distributed**
  - Input - output variables are related **linearly**

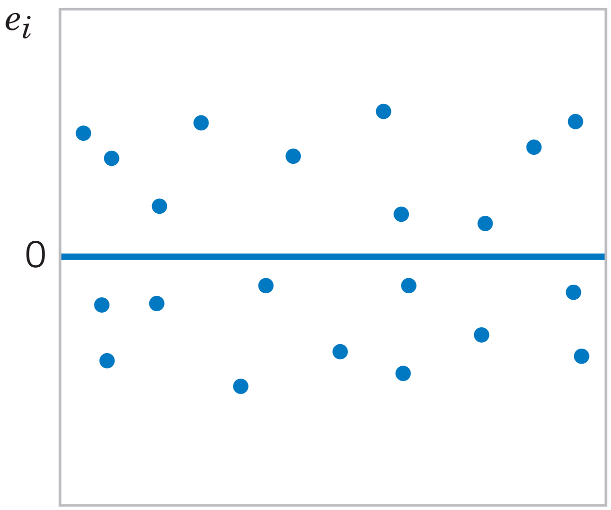
Residual analysis

- Residuals

$$e_i = y_i - \hat{y}_i, i = 1, 2, \ldots, n,$$

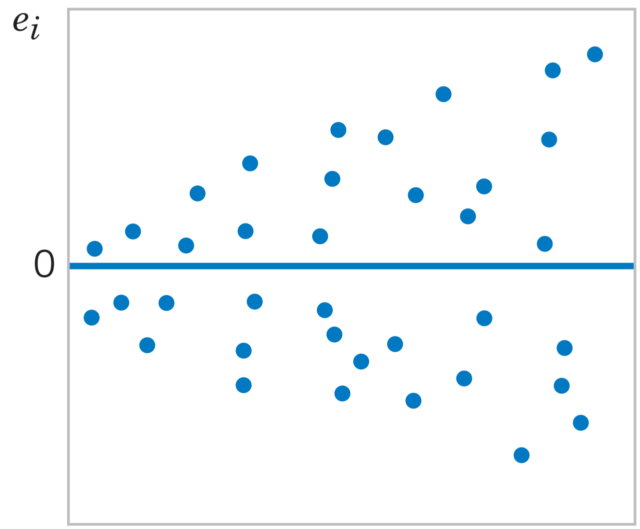
- Residual analysis
  - check whether errors are normally distributed with constant variance
  - whether should include additional (non-linear) terms

What should residuals look like



(a)

OK



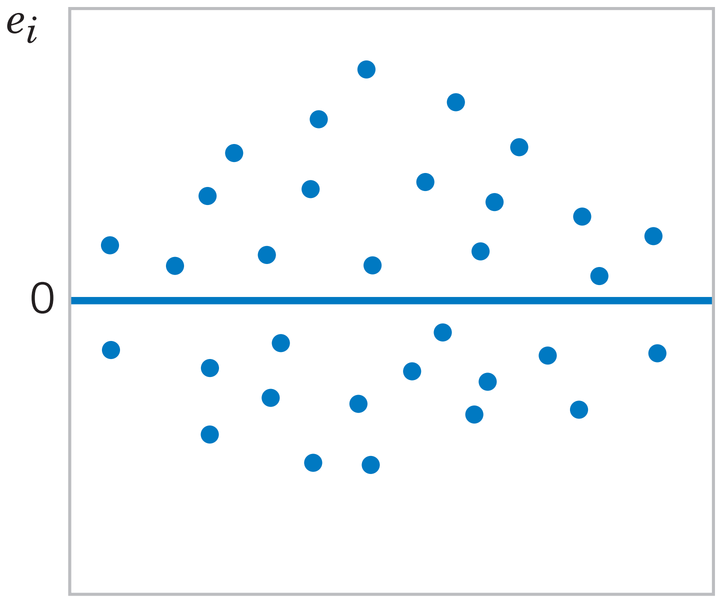
(b)

not OK: variance of residuals increase with magnitude of xi or yi

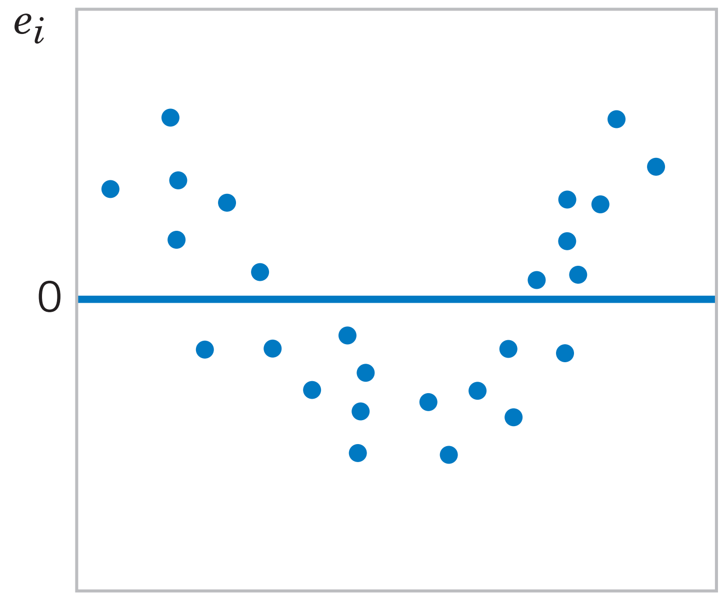
data transformation can solve this problem

$$\sqrt{y},\ \ln y,\ \text{or}\ 1/y$$

Some other abnormal residual plots



(c)



(d)

not OK: variance not constant

not OK: model inadequacy

Check normality

- Step 1: standardize residuals

$$d_i = e_i / \sqrt{\hat{\sigma}^2}, i = 1, 2, \ldots, n$$

- if residuals are normal, 95% of these d_i should be in (-2, 2)
- Step 2: plot normal probability plot of residuals

- probability plot: (sec 6.6) a graphical method to determine whether the samples are from assumed distribution

Construct normal probability plot

- Rank the samples: from smallest to largest

$$x_1, x_2, \ldots, x_n \text{ is arranged as } x_{(1)}, x_{(2)}, \ldots, x_{(n)}.$$

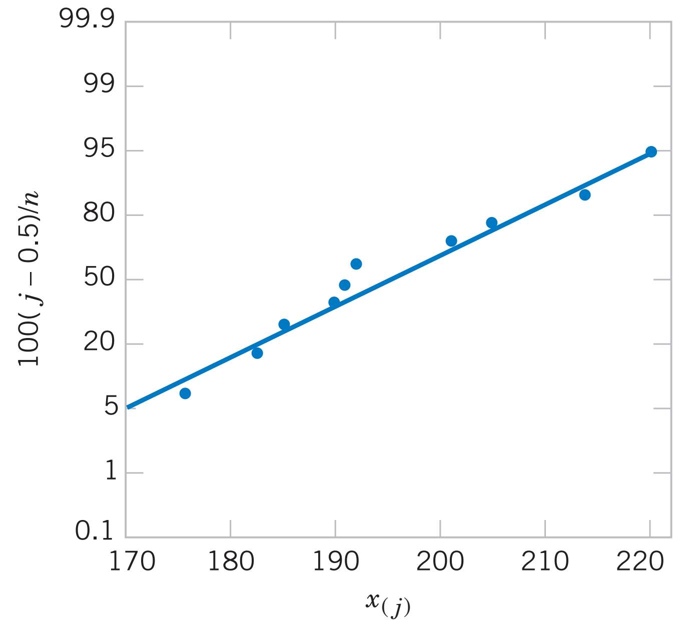
- Ordered observations $x_{(j)}$ are plotted against their assumed frequency $(j - 0.5)/n$

- If the assumed distribution is true, this should approximately follow a straight line

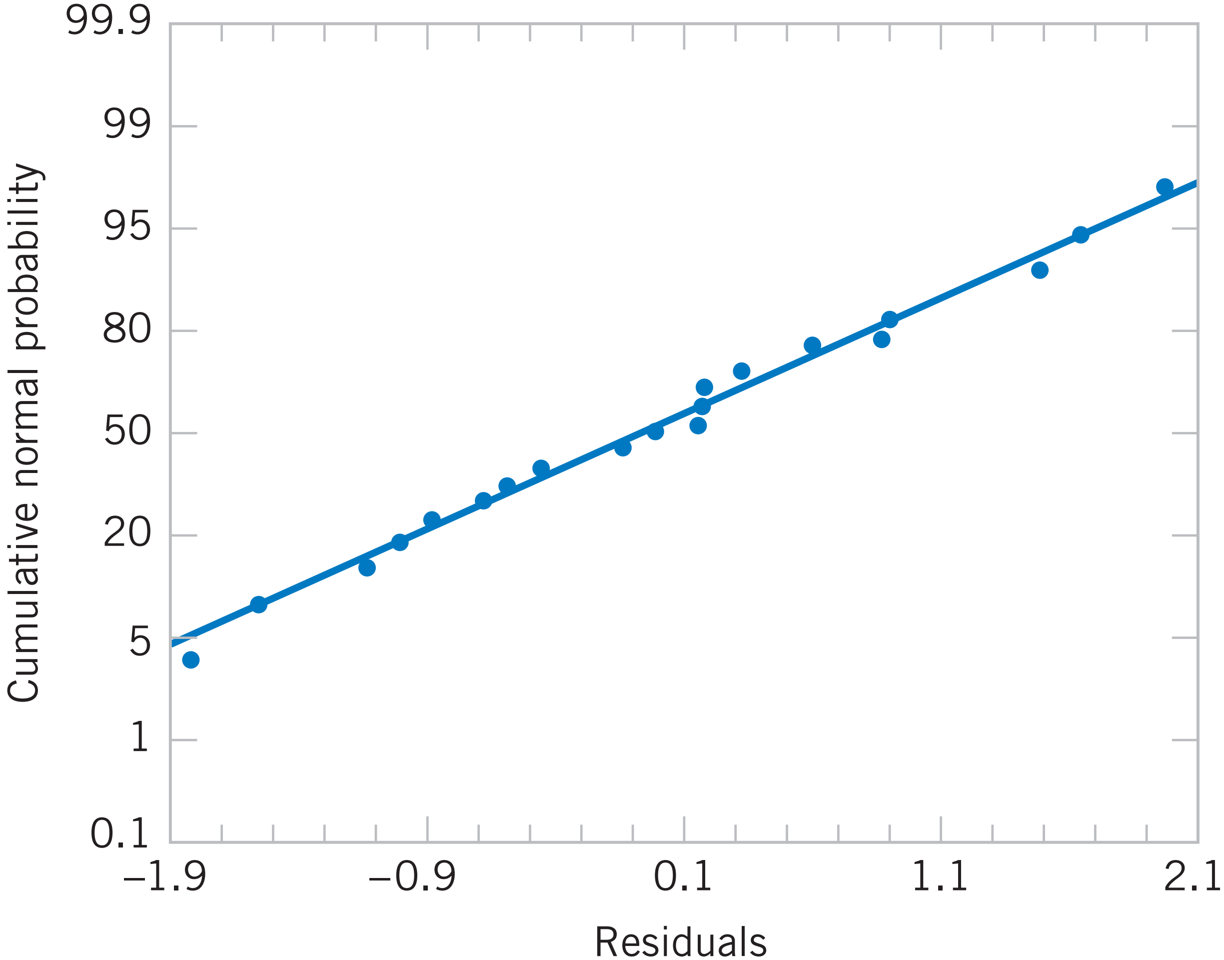
Example: probability plot

- Battery life

j	$x_{(j)}$	$(j - 0.5)/10$	z_j
1	176	0.05	−1.64
2	183	0.15	−1.04
3	185	0.25	−0.67
4	190	0.35	−0.39
5	191	0.45	−0.13
6	192	0.55	0.13
7	201	0.65	0.39
8	205	0.75	0.67
9	214	0.85	1.04
10	220	0.95	1.64



Normal probability plot of oxygen level example



Coefficient of determination

- A widely used measure for a regression model
  - ratio of sum of square

The **coefficient of determination** is

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

$$0 \le R^2 \le 1$$

  - amount of variability in data explained by regression model
  - oxygen level example:

$$R^2 = SS_R/SS_T = 152.13/173.38 = 0.877$$

Test of correlation coefficient

- Can be used to check linearity assumption
- Definition of correlation coefficient $\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$
- Hypothesis test: whether or not there's correlation

$$H_0: \rho = \rho_0$$
$$H_1: \rho \neq \rho_0$$

- estimator

$$R = \frac{\sum_{i=1}^{n} Y_i (X_i - \bar{X})}{\left[\sum_{i=1}^{n} (X_i - \bar{X})^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \right]^{1/2}} = \frac{S_{XY}}{(S_{XX} SS_T)^{1/2}}$$

Test statistic for test correlation or not

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

If H_0 is true

$$T_0 = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \sim T_{n-2}$$

Reject H_0 $\quad |t_0| > t_{\alpha/2, n-2}$

Confidence interval of correlation coefficient

- For large sample size, n larger than 25, the statistic

$$Z = \text{arctanh}\, R = \frac{1}{2} \ln \frac{1 + R}{1 - R} \sim N\left(\frac{1}{2} \ln \frac{1+\rho}{1-\rho}, \frac{1}{n-3} \right)$$

- The approximate $1-\alpha$ confidence interval

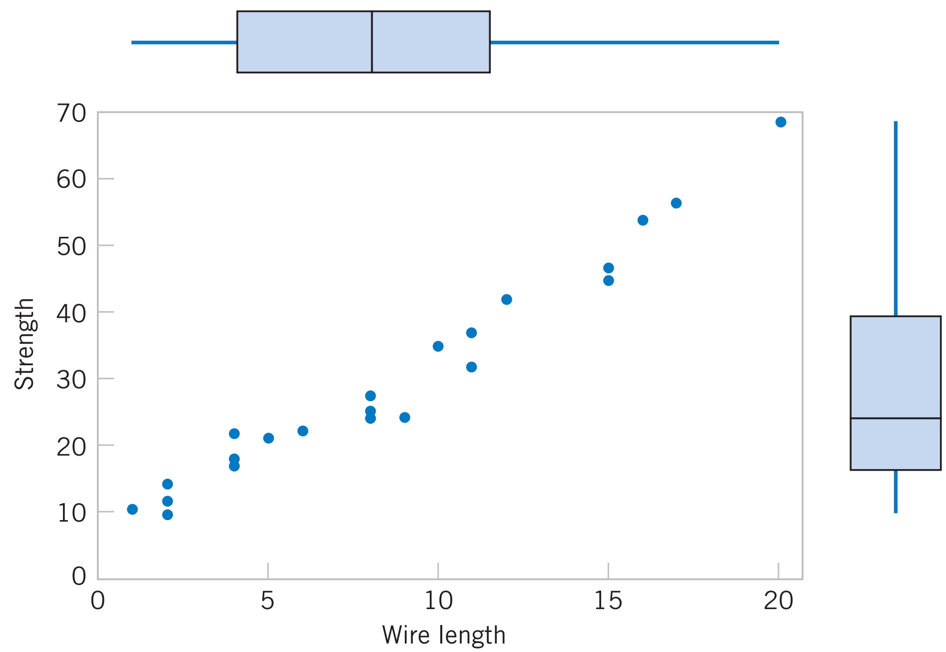
Confidence Interval for a Correlation Coefficient

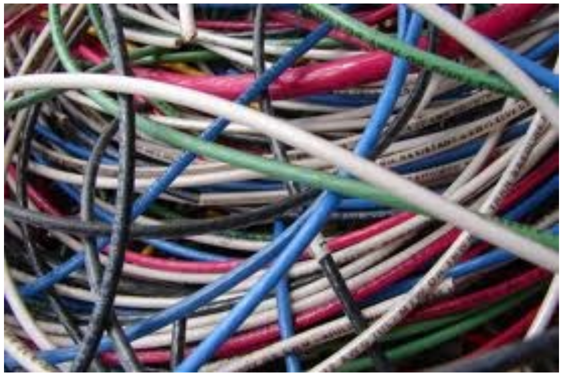
$$\tanh\left(\text{arctanh}\, r - \frac{z_{\alpha/2}}{\sqrt{n-3}}\right) \le \rho \le \tanh\left(\text{arctanh}\, r + \frac{z_{\alpha/2}}{\sqrt{n-3}}\right) \qquad (11\text{-}50)$$

Example: wire bond pull strength

Table 1-2 Wire Bond Pull Strength Data

Observation Number	Pull Strength y	Wire Length x_1
1	9.95	2
2	24.45	8
3	31.75	11
4	35.00	10
5	25.02	8
6	16.86	4
7	14.38	2
8	9.60	2
9	24.35	9
10	27.50	8
11	17.08	4
12	37.00	11
13	41.95	12
14	11.66	2
15	21.65	4
16	17.89	4
17	69.00	20
18	10.30	1
19	34.93	10
20	46.59	15
21	44.88	15
22	54.12	16
23	56.63	17
24	22.13	6
25	21.15	5





Example: wire bond pull strength

$$n = 25$$

$$\alpha = 0.05$$

Sample correlation coefficient

$$r = \frac{\sum_{i=1}^{n} Y_i (X_i - \overline{X})}{\left[\sum_{i=1}^{n} (X_i - \overline{X})^2 \sum_{i=1}^{n} (Y_i - \overline{Y})^2 \right]^{1/2}} = \frac{2027.7132}{[(698.560)(6105.9)]^{1/2}} = 0.9818$$

Value of test statistic

$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.9818\sqrt{23}}{\sqrt{1-0.9640}} = 24.8$$

Compare with threshold $t_{0.025,23} = 2.069$

Example continue

- Approximate 95% confidence interval for true correlation

Confidence Interval for a Correlation Coefficient

$$\tanh\left(\text{arctanh } r - \frac{z_{\alpha/2}}{\sqrt{n-3}}\right) \le \rho \le \tanh\left(\text{arctanh } r + \frac{z_{\alpha/2}}{\sqrt{n-3}}\right) \qquad (11\text{-}50)$$

$$\arctan h(x) = \frac{1}{2}\ln\frac{1+x}{1-x}$$

$$\tanh\left(2.3452 - \frac{1.96}{\sqrt{22}}\right) \le \rho \le \tanh\left(2.3452 + \frac{1.96}{\sqrt{22}}\right)$$

$$0.9585 \le \rho \le 0.9921$$

Summary for model diagnosis

- Check residuals are normally distributed?
  - use normal probability plot
- Check linearity?
  - look at whether residual is stationary
- Check whether or not X and Y are correlated?
  - z-test and ANOVA
  - test and confidence interval for correlation coefficient
- How to use regression model:
  - mean response and its confidence interval
  - predicted y and its confidence interval