STAT 423/523 Statistical Methods for Engineers and Scientists

Lecture 10: Simple Linear Regression

Chencheng Cai

Washington State University

Simple Linear Regression

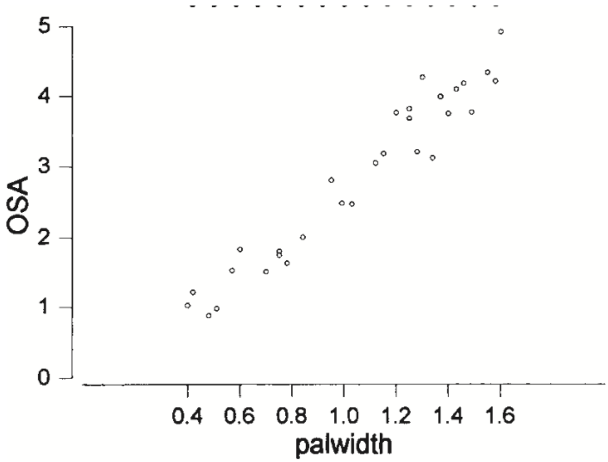
Regression is a statistical method for estimating the relationships among variables. THe simpest form of regression is **simple linear regression**:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

- y_i is the response variable (dependent variable).
- x_i is the predictor variable (independent variable).
- β_0 is the intercept.
- β_1 is the slope.
- ϵ_i is the error term.

Example

- y: ocular surface area
- x: width of the palprebal fissure



Assumptions

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

- Linearity: The relationship between x and y is linear.
- Independence: The errors are independent.
- Normality: The errors are normally distributed.
- Equal variance: The errors have constant variance.

For short, the LINE assumptions give:

$$y_i = \beta_0 + \beta_1 x_i + N(0, \sigma^2) \quad \forall i$$

Violations of Assumptions

- Linearity: Nonliear regression model.
- Independence: Structural equation model (SEM) in econometrics.
- Normality: ϵ_i could have a heavy-tailed distribution.
- Equal variance: Heteroscedasticity.

Some Statistics

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

We assume all x_i's are fixed and known. (not random variables!)

- $E(y_i) = \beta_0 + \beta_1 x_i$ is the mean response for a given x_i.
- $Var(y_i) = Var(\epsilon_i) = \sigma^2$ is the variance of the response.
- $Cov(y_i, y_j) = Cov(\epsilon_i, \epsilon_j)$ for $i \neq j$. (Independence Assumption).

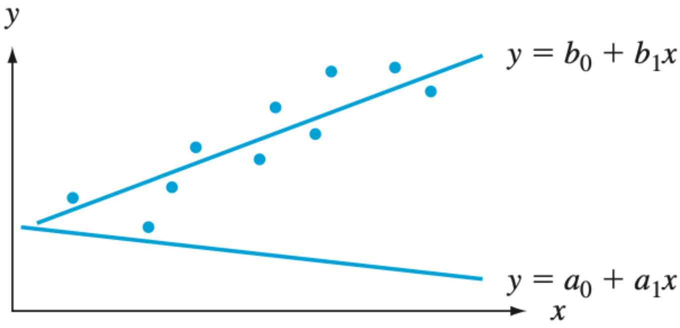
If we get the estimated coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$,

- The **fitted value** for y_i is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.
- The **residual** for y_i is $\hat{\epsilon}_i = y_i - \hat{y}_i$.

Estimation

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

Given the data points



we want to find the line that **best fits** the data points.

Ordinary Least Squares

The first approach is **Ordinary Least Squares** (OLS).

- For each possible parameter values β_0 and β_1, we can calculate the **residual sum of squares** (RSS):

$$\text{RSS}(\beta_0, \beta_1) = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)^2$$

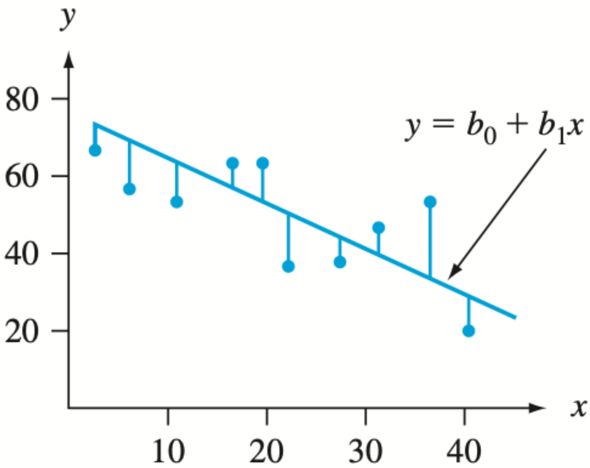
- The OLS estimates are the values of β_0 and β_1 that minimize the RSS:

$$\hat{\beta}_0, \hat{\beta}_1 = \arg\min_{\beta_0, \beta_1} \ \text{RSS}(\beta_0, \beta_1)$$

Residual Sum of Squares

The residual sum of squares is the sum of the squared distance between the data points and the fitted line.

It is the **vertical** distance, not the orthogonal distance.



OLS

In order to minimize the RSS, we first compute its partial derivatives.

$$\mathrm{RSS}(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\frac{\partial \mathrm{RSS}}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = -2 \sum_{i=1}^{n} y_i + 2N\beta_0 + 2\beta_1 \sum_{i=1}^{n} x_i$$

$$\frac{\partial \mathrm{RSS}}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = -2 \sum_{i=1}^{n} y_i x_i + 2\beta_0 \sum_{i=1}^{n} x_i + 2\beta_1 \sum_{i=1}^{n} x_i^2$$

To find the minimum, we set the partial derivatives to zero.

OLS

The **estimating equations** for OLS are:

$$0 = -2\sum_{i=1}^{n} y_i + 2n\beta_0 + 2\beta_1 \sum_{i=1}^{n} x_i \quad (1)$$

$$0 = -2\sum_{i=1}^{n} y_i x_i + 2\beta_0 \sum_{i=1}^{n} x_i + 2\beta_1 \sum_{i=1}^{n} x_i^2 \quad (2)$$

Compute $(1) \times \sum_i x_i - (2) \times n$:

$$0 = 2n\sum_i x_i y_i - 2\sum_i x_i \sum_i y_i + \left(\left(\sum_i x_i \right)^2 - n \sum_i x_i^2 \right) \beta_1 .$$

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

OLS

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

- The numerator is

$$\sum_i x_i y_i - n^{-1} \sum_i x_i \sum_i y_i = S_{xy} = \sum_i (y_i - \bar{y})(x_i - \bar{x})$$

- The denominator is

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

OLS

Therefore,

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}$$

with

$$S_{xy} = \sum_i (y_i - \bar{y})(x_i - \bar{x}) = \sum_i y_i x_i - n^{-1} \sum_i x_i \sum_i y_i$$

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

From Eq. (1), we can get $\hat{\beta}_0$:

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

OLS

We still have σ^2 to estimate. The easiest way is to estimate it from the residual sum of squares:

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}(\hat{\beta}_0, \hat{\beta}_1)}{n-2}$$

- $n-2$ is the degrees of freedom.

A quick formula in computing $\mathrm{RSS}(\hat{\beta}_0, \hat{\beta}_1)$ is

$$\mathrm{RSS}(\hat{\beta}_0, \hat{\beta}_1) = S_{yy} - \hat{\beta}_1 S_{xy} = S_{yy} - \hat{\beta}_1^2 S_{xx},$$

where

$$S_{yy} = \sum_i (y_i - \bar{y})^2 = \sum_i y_i^2 - n^{-1}\left(\sum_i y_i\right)^2.$$

OLS

Summary for OLS estimators:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}$$

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

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}(\hat{\beta}_0, \hat{\beta}_1)}{n-2} = \frac{S_{yy} - \hat{\beta}_1 S_{xy}}{n-2}$$

Example (Textbook Example 12.8)

x	12	30	36	40	45	57	62	67	71	78	93	94	100	105
y	3.3	3.2	3.4	3.0	2.8	2.9	2.7	2.6	2.5	2.6	2.2	2.0	2.3	2.1

Some statistics:

$$n = 14 \qquad \sum x_i = 890 \qquad \sum x_i^2 = 67182$$

$$\sum y_i = 37.6 \qquad \sum y_i^2 = 103.54 \qquad \sum x_i y_i = 2234.30$$

Example (Textbook Example 12.8)

We can compute the following statistics:

$$S_{xx} = 10603.43, \quad S_{xy} = -155.99, \quad S_{yy} = 2.557$$

The estimators are

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{-155.99}{10603.43} = -0.0147$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{37.6}{14} - (-0.0147) \times \frac{890}{14} = 3.62$$

$$\hat{\sigma}^2 = \frac{S_{yy} - \hat{\beta}_1 S_{xy}}{n-2} = \frac{2.557 - (-0.0147) \times (-155.99)}{14-2} = 0.022$$

Properties of OLS Estimators

- Because x_i's are fixed, S_{xx} is not a random variable.
- S_{xy} can be written as

$$S_{xy} = \sum x_i y_i - n^{-1} \sum x_i \sum y_i = \sum_i \left[(x_i - \bar{x})\, y_i \right]$$

  The highlighted y_i's are the only random variables and we have

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2),$$

  where β_0 and β_1 are the true parameters.
- Therefore, S_{xy} is a linear combination of normal random variables and is also normally distributed,

$$S_{xy} \sim N(\beta_1 S_{xx}, \sigma^2 S_{xx})$$

- Now we have

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \sim N(\beta_1, \sigma^2 S_{xx}^{-1})$$

Properties of OLS Estimators

- For the intercept estimator, we have

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \sim N(\beta_0, (n^{-1} + \bar{x}^2 S_{xx}^{-1})\sigma^2)$$

- For the variance estimator, we have

$$E(\hat{\sigma}^2) = \sigma^2.$$

Properties of OLS Estimators

Summary:

- All OLS estimators are **unbiased**:

$$E(\hat{\beta}_0) = \beta_0$$
$$E(\hat{\beta}_1) = \beta_1$$
$$E(\hat{\sigma}^2) = \sigma^2$$

- The estimated **standard errors (se)** of the estimators are:

$$s_{\hat{\beta}_0} = \sqrt{(n^{-1} + \bar{x}^2 S_{xx}^{-1})\hat{\sigma}^2}$$
$$s_{\hat{\beta}_1} = \sqrt{S_{xx}^{-1}\hat{\sigma}^2}$$
$$s_{\hat{\sigma}^2} = \sqrt{\frac{2\hat{\sigma}^4}{n-2}}$$

Confidence Interval

The $(1 - \alpha)$ confidence interval for β_1 is

$$\hat{\beta}_1 \pm t_{\alpha/2, n-2} s_{\hat{\beta}_1}.$$

- Confidence interval uses two-sided t-distribution with $n - 2$ degrees of freedom.
- It is t-distributed because we are estimating σ^2 from the data.

Hypothesis Testing

Consider the following hypothesis testing:

$$H_0 : \beta_1 = 0 \quad H_a : \beta_1 \neq 0.$$

Method 1: reject null if the CI does not cover 0:

$$\text{reject null if } 0 \notin (\hat{\beta}_1 - t_{\alpha/2,n-2} s_{\hat{\beta}_1}, \hat{\beta}_1 + t_{\alpha/2,n-2} s_{\hat{\beta}_1})$$

Method 2: reject null if the test statistic

$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$$

is greater than $t_{\alpha/2,n-2}$ in absolute value.

Hypothesis Testing

$$H_0 : \beta_1 = 0 \quad H_a : \beta_1 \neq 0.$$

Method 3: reject null if the p-value

$$p = 2\left(1 - F_{t,n-2}(|\hat{\beta}_1 / s_{\hat{\beta}_1}|)\right)$$

is less than α.

- To test $H_0 : \beta_1 > 0$, we should use one-sided t-test.
- Same process for testing $\beta_0 = 0$.

Goodness of Fit

The variation in the response variable y_i is

$$SST = \sum_i (y_i - \bar{y})^2$$

The variation explained by the regression model is

$$SSR = \sum_i (\hat{y}_i - \bar{y})^2$$

The variation not explained by the regression model is

$$SSE = \sum_i (y_i - \hat{y}_i)^2$$

We have

$$SST = SSR + SSE$$

Goodness of Fit

The **coefficient of determination** is defined as

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}.$$

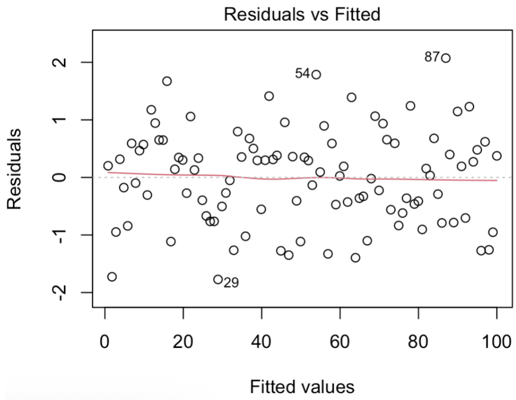
- R^2 is the proportion of the variation in the response variable that is explained by the regression model.
- R^2 is between 0 and 1.
- R^2 is a measure of the goodness of fit of the regression model.

Residual Plot

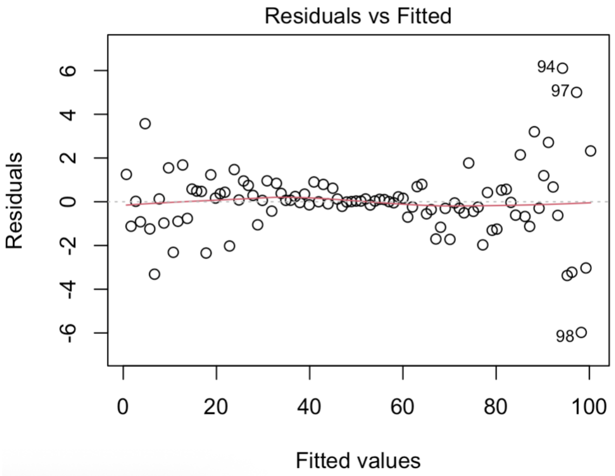
The **residual** is defined as the difference between the observed value and the fitted value:

$$\hat{\epsilon}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i.$$

The **residual plot** is a scatter plot of the residuals against the fitted values.

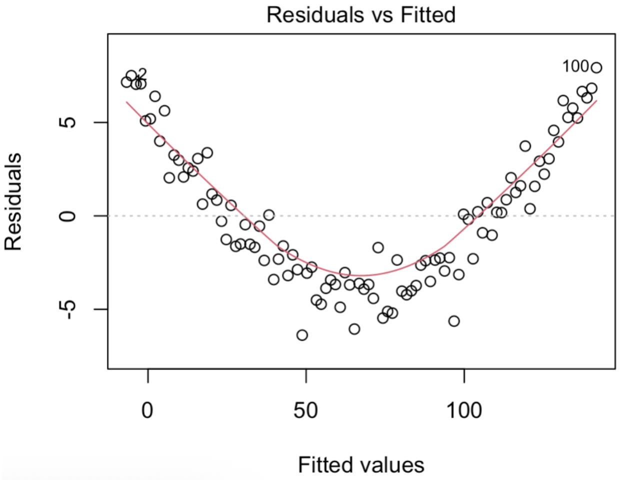


Residual Plot



The variance is not equal for all ϵ_i's. **Solution**: data need to be transformed.

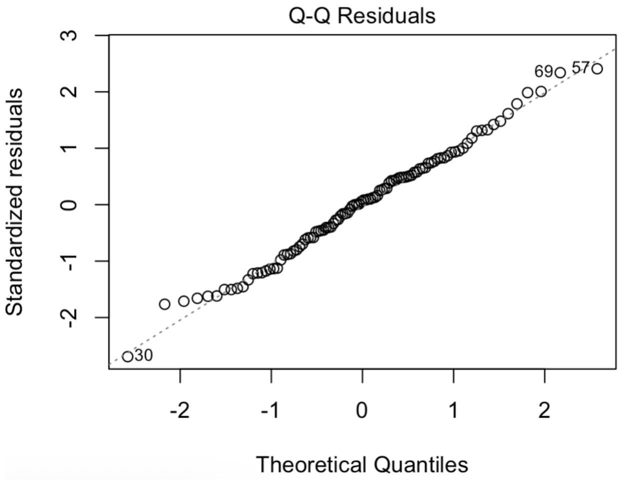
Residual Plot



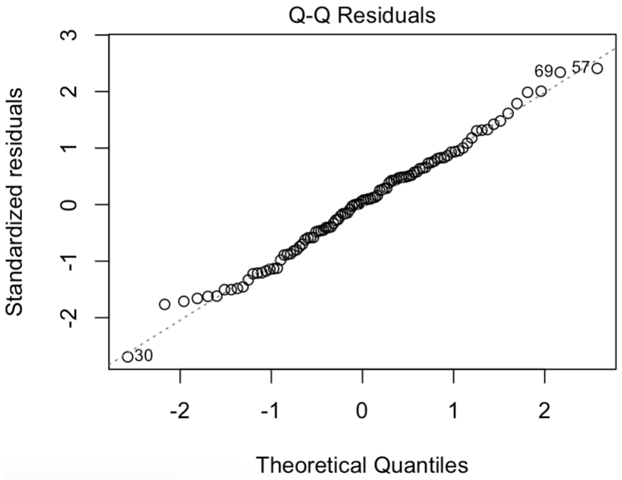
The residual is not independent with the fitted value. **Solution**: add more predictors.

QQ Plot

The **QQ plot** is a scatter plot of the quantiles of the residuals against the quantiles of the normal distribution.

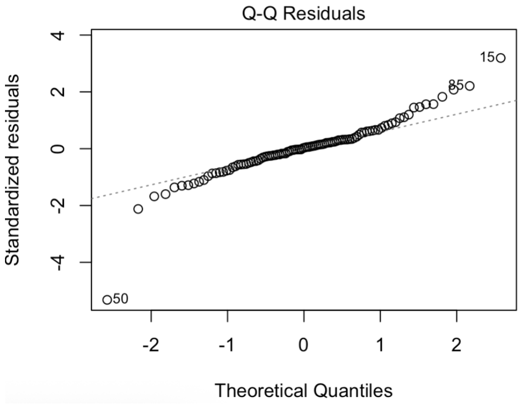


QQ Plot



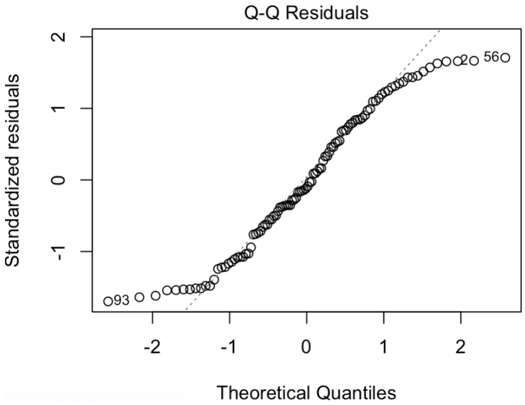
If all the points are on the line, then the residuals are normally distributed.

QQ Plot



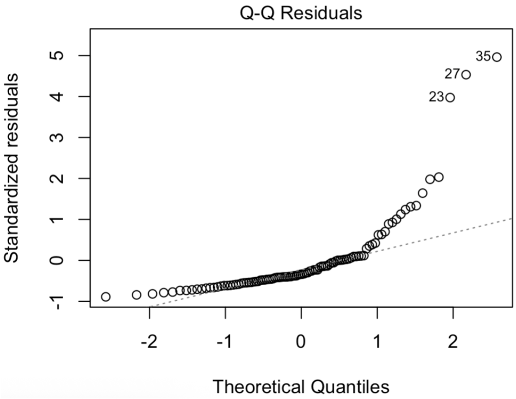
If the left tail is bended down and the right tail is bended up, then the residuals are **heavy-tailed**.

QQ Plot



If the left tail is bended up and the right tail is bended down, then the residuals are **light-tailed**.

QQ Plot



If the two tails are bended to the same direction, then the residuals are **skewed**.

QQ Plot

- If the points are on the line, then the residuals are normally distributed.
- If the points are not on the line, then the residuals are not normally distributed.
- Light tails is usually not a problem.
- But heavy tails is a problem.

ANOVA for Regression

Since we have computed SSR, SSE and SST. We can print the ANOVA table for the simple lienar regression:

Source	SS	d.f.	MS	F stat
Regression	SSR	1	MSR = SSR	F=MSR/MSE
Error	SSE	n-2	MSE = SSE/(n-2)	
Total	SST	n-1		

The hypothesis testing of $H_0 : \beta_1 = 0$ can be done by the F-test:

$$\text{reject null when } F > F_{\alpha,n-2}$$

T-test vs. F-test

$$H_0 : \beta_1 = 0 \quad \text{v.s.} \quad H_a : \beta_1 \neq 0.$$

T-test:

$$\text{reject null when } |t| = \left| \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} \right| > t_{\alpha/2, n-2}$$

F-test:

$$\text{reject null when } F = \frac{MSR}{MSE} > F_{\alpha, 1, n-2}$$

T-test vs. F-test

For t-test, we have

$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \frac{S_{xy}/S_{xx}}{\sqrt{S_{xx}^{-1}\hat{\sigma}^2}} = \frac{S_{xy}}{\sqrt{S_{xx} \cdot \mathrm{MSE}}}$$

For F-test, we have

$$F = \frac{MSR}{MSE} = \frac{SSR}{MSE} = \frac{\hat{\beta}_1^2 S_{xx}}{MSE} = \frac{S_{xy}^2}{S_{xx} \cdot \mathrm{MSE}}$$

Therefore, we have

$$F = t^2$$

Then

$$|t| > t_{\alpha/2,n-2} \iff t^2 > t^2_{\alpha_2,n-2} \iff F > F_{\alpha,n-2},$$

using the fact that $t^2_{\alpha_2,n-2} = F_{\alpha,1,n-2}$.

Therefore, the t-test and F-test for β_1 are equivalent.

Prediction

- Suppose we have fitted a simple linear regression model with $\hat{\beta}_0$ and $\hat{\beta}_1$.
- Let x_* be a new value of x.
- The **point prediction** for y_* is

$$\hat{y}_* = \hat{\beta}_0 + \hat{\beta}_1 x_*$$

- $\hat{y}_*$ is a random variable
  because $\hat{\beta}_0$ and $\hat{\beta}_1$ are random variables depending on the data.

Prediction

- The expectation of $\hat{y}_*$ is

$$E(\hat{y}_*) = E(\hat{\beta}_0) + E(\hat{\beta}_1)x_* = \beta_0 + \beta_1 x_* = \bar{y}_*$$

  $\bar{y}_*$ is the **mean response** for x_*.(it does not have the error term ϵ_*)

- The variance of $\hat{y}_*$ is

$$Var(\hat{y}_*) = Var(\hat{\beta}_0) + Var(\hat{\beta}_1)x_*^2 + 2Cov(\hat{\beta}_0, \hat{\beta}_1)x_* = \sigma^2 \left(\frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{xx}} \right)$$

  - The variance scales as $1/n$ (because $S_{xx} \propto n$).
  - The variance negatively depends on the distance from x_* to $\bar{x}$.

- An estimate of the variance is

$$\hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{xx}} \right).$$

Prediction

The $(1-\alpha)$ **confidence interval for the mean response** is

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

The interpreation is:
The probability that this CI covers the **mean response** $\bar{y}_*$ is $1-\alpha$.

- The response $y_* = \bar{y}_* + \epsilon_*$ is the mean response plus the error term.
- y_* is more noisy than $\bar{y}_*$.
- Above CI has a less coverage for y_* than $\bar{y}_*$.
- We need a wider CI for y_*.

Prediction

The $(1-\alpha)$ **prediction interval for the response** is

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

The interpretation is:
The probability that this PI covers the **response** y_* is $1-\alpha$.

- The constant 1 in the above formula accounts for the variance of the error term ϵ_*.
- The prediction interval is wider than the confidence interval for the mean response.

Example (textbook example 12.13)

x = carbonation depth (mm) and y = strength (MPa).

x	8.0	15.0	16.5	20.0	20.0	27.5	30.0	30.0	35.0
y	22.8	27.2	23.7	17.1	21.5	18.6	16.1	23.4	13.4
x	38.0	40.0	45.0	50.0	50.0	55.0	55.0	59.0	65.0
y	19.5	12.4	13.2	11.4	10.3	14.1	9.7	12.0	6.8

Summary statistics:

$$n = 18 \qquad \sum_i x_i = 659.0 \qquad \sum_i x_i^2 = 28967.50$$

$$\sum_i y_i = 293.2 \qquad \sum_i y_i^2 = 5335.76 \qquad \sum_i x_i y_i = 9293.95$$

Example (textbook example 12.13)

We first compute:

$$S_{xx} = 28967.50 - \frac{659^2}{18} = 4840.778$$

$$S_{xy} = 9293.95 - \frac{659 \times 293.2}{18} = -1440.428$$

$$S_{yy} = 5335.76 - \frac{293.2^2}{18} = 559.858$$

The estimators are:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = -0.2976$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 27.183$$

$$\hat{\sigma}^2 = \frac{S_{yy} - \hat{\beta}_1 S_{xy}}{n-2} = 8.203$$

Example (textbook example 12.13)

Suppose we have a new observation $x_* = 45.0$ mm. The prediction is

$$\hat{y}_* = \hat{\beta}_0 + \hat{\beta}_1 x_* = 27.183 - 0.2976 \times 45 = 13.79$$

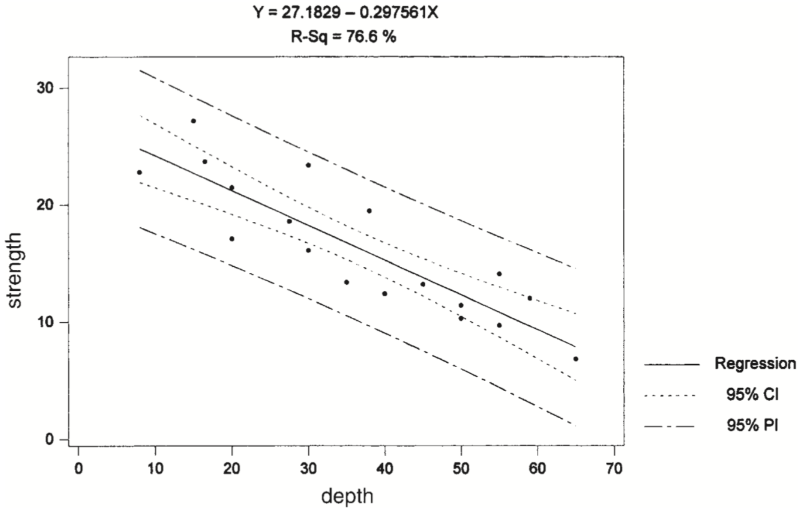
The 95% confidence interval for the mean response is

$$13.79 \pm t_{0.025,16}\sqrt{8.203\left(\frac{1}{18} + \frac{(45 - 36.611)^2}{4840.778}\right)} = (12.18, 15.40)$$

The 95% prediction interval for the response is

$$13.79 \pm t_{0.025,16}\sqrt{8.203\left(1 + \frac{1}{18} + \frac{(45 - 36.611)^2}{4840.778}\right)} = (7.50, 20.08)$$

Example (textbook example 12.13)



Confidence Band

The confidence intervals can be constructed for **any** value of x_*.
The confidence intervals for all values of x_* can be plotted to form a **confidence band**.

- The **pointwise confidence band for the mean response** is the region that

$$|y - (\hat{\beta}_0 + \hat{\beta}_1 x)| < t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{xx}}\right)}$$

- Interpreation: for any given x, the probability that the mean response at x is in the band is $1 - \alpha$.

Confidence Band

The **Working-Hotelling simultaneous confidence band** is the region that

$$|y - (\hat{\beta}_0 + \hat{\beta}_1 x)| < \sqrt{2F_{\alpha,2,n-2}}\sqrt{\hat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{xx}}\right)}$$

- Interpreation: the probability that the confidence band covers the whole mean response curve is $1 - \alpha$.
- The simultaneous confidence band is wider than the pointwise confidence band.

$$2F_{\alpha,2,n-2} > F_{\alpha,1,n-2} = t^2_{\alpha/2,n-2}$$