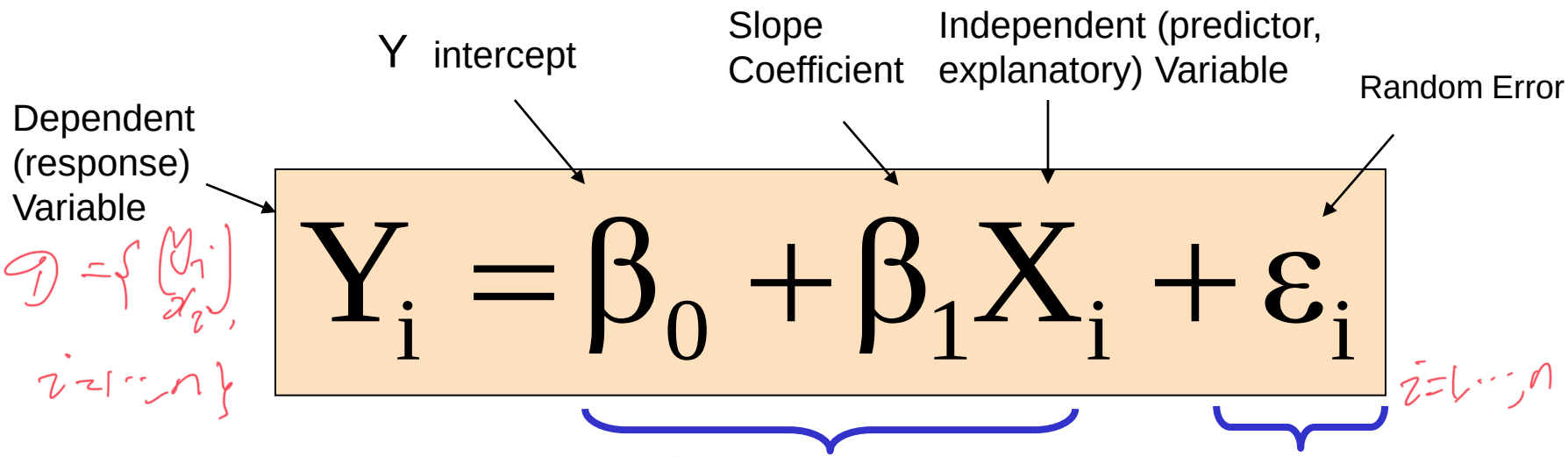
Ch2. Simple Linear Regression

- **Relationship between 2 variables**
- **The regression model**
- **Assumptions**
- **Estimation and method of least squares**
- **Inferences concerning $ß_1$ and $ß_0$**
- **Estimation of the mean of the response variable for a given level of X**
- **Prediction of new observation**
- **Analysis of variance approach to regression analysis**
- **Measures of linear association between X and Y**

Simple Linear Regression Model



Assumptions:

- $E(\varepsilon_i) = 0$
- Variance $(\varepsilon_i) = \sigma^2$
- Covariance $(\varepsilon_i, \varepsilon_j) = 0$

In matrix notation?

$$\underset{n\times 1}{Y} = \underset{n\times 2}{X}\ \underset{2\times 1}{\beta} + \underset{n\times 1}{\varepsilon}$$

Simple Linear Regression Equation

The simple linear regression equation provides an estimate of the population regression line

Estimated (or predicted) Y value for observation i

Estimate of the regression intercept

Estimate of the regression slope

Value of *X* for observation *i*

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

Interpretation of the Slope and the Intercept

- $\hat{\beta}_0$ is the estimated average value of Y when the value of X is zero
- $\hat{\beta}_1$ is the estimated change in the average value of Y as a result of a one-unit change in X

Example

- A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet)

- A random sample of 10 houses is selected
  - Dependent variable (*Y*) = house price in $1000s
  - Independent variable (*X*) = square feet

Sample Data for House Price Model

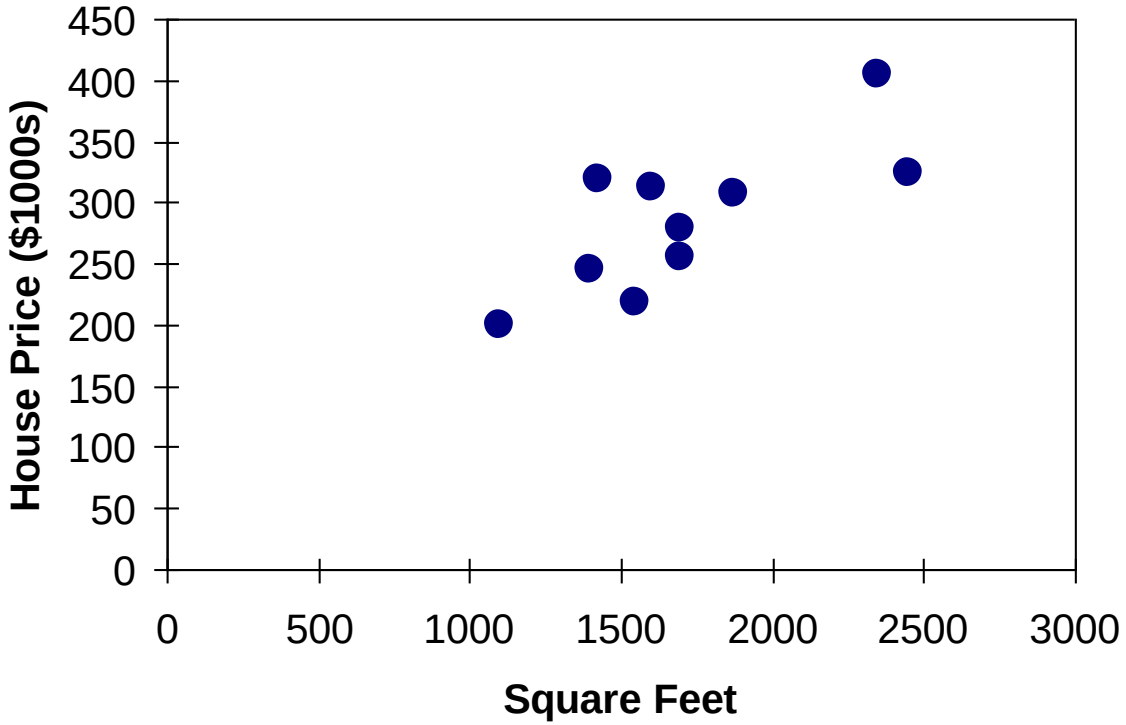
House Price in $1000s (*Y*)	Square Feet (*X*)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

y=c(245, 312, ..., 255)

x = c(140, 180, ... , 170)
plot(x, y)

Graphical Presentation

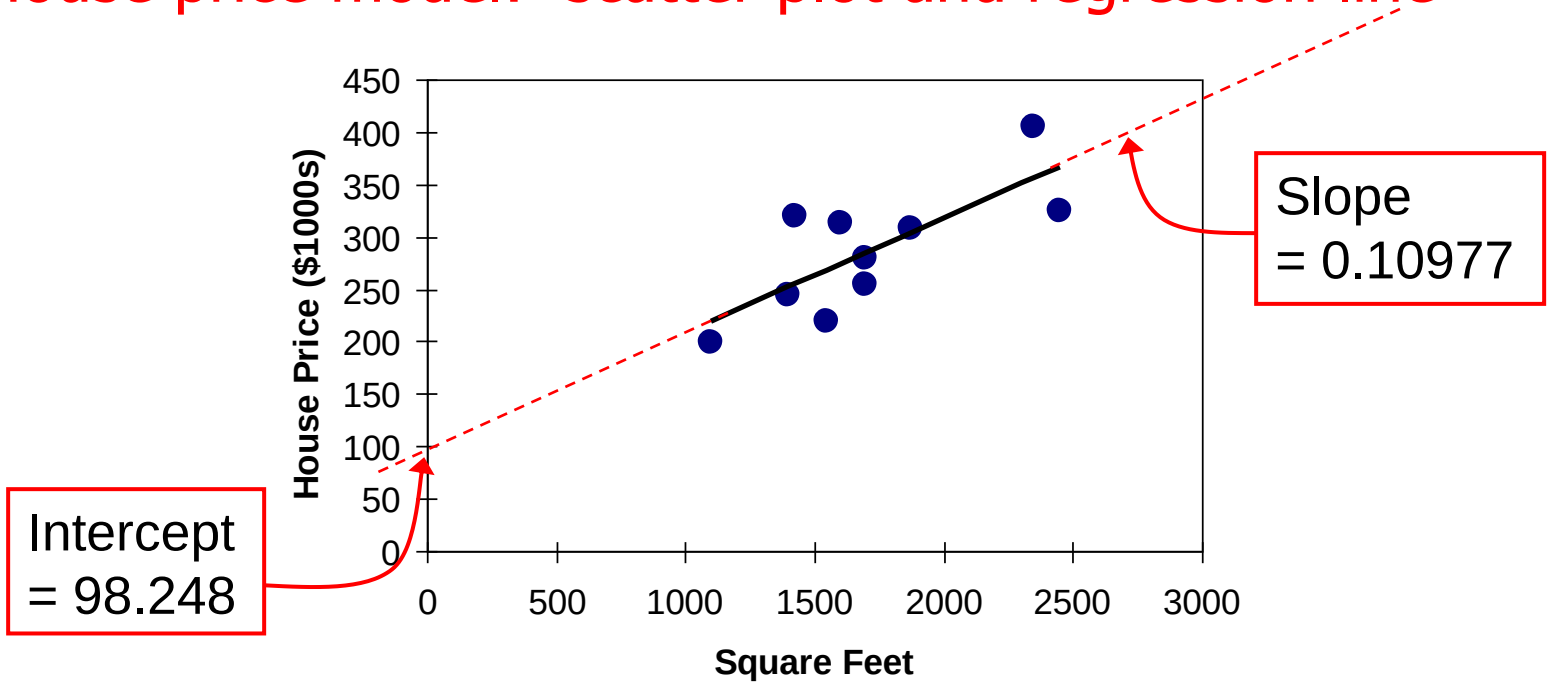
House price model: scatter plot



Graphical Presentation

abline(lm(Y ~ x))

House price model: scatter plot and regression line



$$\widehat{\text{house price}} = 98.24833 + 0.10977(\text{square feet})$$

Interpretation of the Intercept, $\hat{\beta}_0$

$$\widehat{\text{house price}} = 98.24833 + 0.10977(\text{square feet})$$

- $\hat{\beta}_0$ is the estimated average value of Y when the value of X is zero (if $X = 0$ is in the range of observed X values)
- Here, no houses had 0 square feet, so $\hat{\beta}_0 = 98.24833$ just indicates that, for houses within the range of sizes observed, \$98,248.33 is the portion of the house price not explained by square feet

Interpretation of the Slope Coefficient, $\hat{\beta}_1$

$$\widehat{\text{house price}} = 98.24833 + 0.10977(\text{square feet})$$

- $\hat{\beta}_1$ measures the estimated change in the average value of Y as a result of a one-unit change in X
  - Here, $\hat{\beta}_1$ = .10977 tells us that the average value of a house increases by .10977($1000) = $109.77, on average, for each additional one square foot of size

Predictions using Regression Analysis

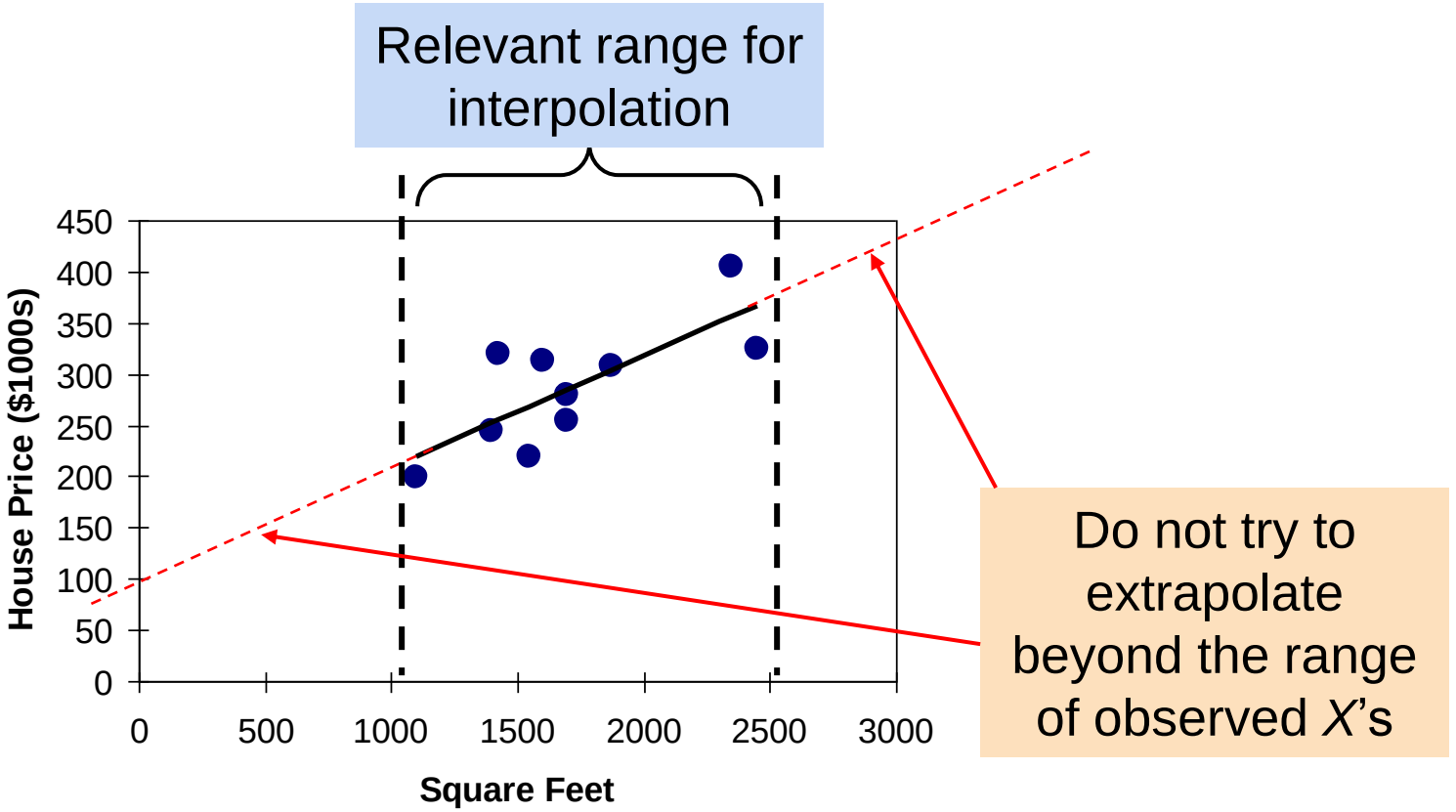
Predict the price for a house with 2000 square feet:

$$\widehat{\text{house price}} = 98.25 + 0.1098\,(\text{sq.ft.})$$
$$= 98.25 + 0.1098\,(2000)$$
$$= 317.85$$

The predicted price for a house with 2000 square feet is 317.85($1,000s) = $317,850

Interpolation vs. Extrapolation

- When using a regression model for prediction, only predict within the relevant range of data



Estimation (Method of Least Squares)

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are obtained by finding the values of β_0 and β_1 that minimize the sum of the squared differences between Y and $\hat{Y}$

Remark 2.1 (1) how to estimate β_0, β_1?
(2) how good is the estimation?

Estimation (Method of Least Squares)

- $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \dfrac{S_{xy}}{S_{xx}}$

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$$

$$= \sum x_i y_i - n\bar{x}\bar{y}$$

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

$$S_{xx} = \sum (x_i - \bar{x})^2$$

$$= \sum x_i^2 - n\bar{x}^2$$

Relationship between slope (b_1) and sample correlation (r)

- $\hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}} = r \cdot \sqrt{\dfrac{S_{yy}}{S_{xx}}}$

- $r = \dfrac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$

Similarity:

* $H_0: r = 0$ and $H_0: \beta_1 = 0$ } same conclusion

* signs of $\hat{\beta}_1$ and r are the same.

Differences

* meanings of r and $\hat{\beta}_1$ are different.

* interpretation of $\hat{\beta}_1$

$S_{yy} = \sum (Y_i - \bar{Y})^2 = \sum Y_i^2 - n\bar{Y}^2$

Estimation of error terms variance σ^2

- The estimator of σ^2 is

$$S^2 = MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

- S^2 is an unbiased estimator of σ^2

Remark 2.2

Estimation (Method of Maximum Likelihood)

Normal Error Model

- $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$

- $\hat{\beta}_0 = \overline{Y} - b_1\overline{X}$

Remark 2.3

- The estimator of σ^2 is $\dfrac{SSE}{n} = \dfrac{n-2}{n}S^2$

- MLE of β_0 = LSE of β_0 (unbiased)
- MLE of β_1 = LSE of β_1 (unbiased)
- MLE of σ^2 (< unbiased estimator of σ^2) asymptotically unbiased

More on inference

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

Assumptions:

- X_i are known constants, $i = 1,\ldots n$
- $\varepsilon_i \overset{ind}{\sim} N(0, \sigma^2)$

Therefore, $Y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2)$

Distribution of $\hat{\beta}_1$

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

$$= \sum_{i=1}^{n} k_i Y_i$$

where $$k_i = \frac{(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Distribution of $\hat{\beta}_1$

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

(1) $\hat{\beta}_1 \sim \mathrm{N}(\beta_1, \sigma^2/S_{xx})$ where $S_{xx} = \sum_{i=1}^{n} (X_i - \overline{X})^2$

(2) $\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{(n-2)}$

(3) $\hat{\beta}_1$ and S^2 are independent

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma / S_{xx}^{1/2}} \Big/ \left(\frac{(n-2)S^2}{\sigma^2} / n-2 \right)^{1/2} \sim t_{n-2}$$

$$t = \frac{\hat{\beta}_1 - \beta_1}{S / S_{xx}^{1/2}} \sim t_{n-2}$$

Testing (Two-sided test of β_1)

H_0: $\beta_1 = 0$ (no linear relationship)
H_1: $\beta_1 \neq 0$ (linear relationship does exist between X and Y)

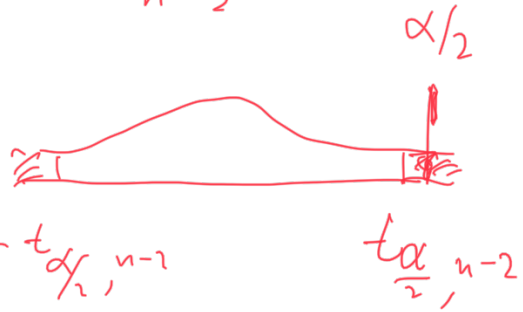
Test statistics:

$$t = \frac{\hat{\beta}_1}{S/S_{xx}^{1/2}} \overset{H_0}{\sim} t_{n-2}$$

Decision rule:

Reject H_0 if $|t| > t_{\alpha/2, n-2}$

OR p-value $= 2 \Pr(t_{n-2} > |t|) < \alpha$



Testing (Two-sided test of β_1)

$H_0: \beta_1 = k$
$H_1: \beta_1 \neq k$

Test statistics:

Decision rule:

How to construct C.I. of β_1 ?

Distribution of $\hat{\beta}_0$

$$\hat{\beta}_0 = \overline{Y} - b_1 \overline{X}$$

Hence,

$$\hat{\beta}_0 \sim \mathrm{N}(\beta_0, \sigma^2 \left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \right])$$

C.I. of β_0 = ?

Estimation of the mean of the response variable for a given level of X

Example

Suppose you want to develop a model to predict selling price of homes based on assessed value. A sample of 30 recently sold single-family houses in a small city is selected to study the relationship between selling price (Y, in $000) and assessed value (X, in $000). The houses in the city had been reassessed at full value 1 year prior to the study.

Remark 2.4

Estimate of the **average** selling price for houses with an assessed value of $70,000.

Estimation of the mean of the response variable for a given level of X (Estimate $E(Y_h)$)

Let X_h denote the level of X for which we wish to estimate the mean response [to be estimated by $\hat{Y}_h$]. (Note: Given X_h, the mean response is $E(Y_h) = \beta_0 + \beta_1 X_h$ according to the model)

For estimation, Given X_h

$$\hat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 X_h$$

Distribution of $E(Y_h) - \hat{Y}_h$:

$$E(Y_h) - \hat{Y}_h \sim N\left(0, \sigma^2 \left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]\right)$$

Confidence Interval for $E(Y_h)$

Two-sided $100(1-\alpha)\%$ C.I. for $E(Y_h)$:

$$\left(\hat{Y}_h - t_{\alpha/2,n-2} s \sqrt{\frac{1}{n} + \frac{\left(X_h - \overline{X}\right)^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2}}, \quad \hat{Y}_h + t_{\alpha/2,n-2} s \sqrt{\frac{1}{n} + \frac{\left(X_h - \overline{X}\right)^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2}} \right)$$

Prediction of a new observation $Y_{h(new)}$

Example

Suppose you want to develop a model to predict selling price of homes based on assessed value. A sample of 30 recently sold single-family houses in a small city is selected to study the relationship between selling price (Y, in $000) and assessed value (X, in $000). The houses in the city had been reassessed at full value 1 year prior to the study.

Estimate the selling price of **an individual** house with an assessed value of $70,000.

Prediction of a new observation $Y_{h(new)}$

Prediction of $Y_{h(new)}$ corresponding to a given level X of the predictor variable by $\hat{Y}_h$.

$$\hat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 X_h$$

Distribution of $Y_{h(new)} - \hat{Y}_h$:

$$Y_{h(new)} - \hat{Y}_h \sim N\left(0, \sigma^2\left[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]\right)$$

Confidence Interval for $Y_{h(new)}$

Two-sided 100(1-α)% C.I. for $Y_{h(new)}$:

$$\left(\hat{Y}_h - t_{\alpha/2,n-2} s \sqrt{\left[1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \right]}, \quad \hat{Y}_h + t_{\alpha/2,n-2} s \sqrt{\left[1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \right]} \right)$$

Analysis of variance approach to regression analysis

- Partitioning of Total Sum of Squares
- Mean Squares
- Analysis of Variance (ANOVA) Table

Remark 2.5

Analysis of variance approach to regression analysis

Total Variation = Unexplained Variation + Explained Variation

$$1 = \frac{\text{Unexplained Variation}}{\text{Total Variation}} + \frac{\text{Explained Variation}}{\text{Total Variation}}$$

$$\frac{\text{Explained Variation}}{\text{Total Variation}} = 1 - \frac{\text{Unexplained Variation}}{\text{Total Variation}}$$

Analysis of variance approach to regression analysis

ANOVA Table

	Sum of Squares (SS)	Degrees of freedom (df)	Mean squares (MS)	*F*
Regression	SSR	1	MSR = SSR/1	MSR/MSE
Error	SSE	n-2	MSE = SSE/(n-2)	
Total	SST	n-1		

H_0: $\beta_1 = 0$ (no linear relationship)
H_1: $\beta_1 \neq 0$ (linear relationship does exist between X and Y)

Test Statistics: F = MSR/MSE
Rejection Rule: reject the null hypothesis if $F > F_{(\alpha, 1, n-2)}$

Measures of linear association between X and Y

- Coefficient of Determination R^2

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

  b) $0 \leq R^2 \leq 1$

- Coefficient of Correlation

  a) $r = \pm\sqrt{R^2}$

  b) A plus or minus sign is attached to this measure according to whether the slope of the fitted regression line is positive or negative.