Lecture 2

Simple Linear Regression

Reading: Forecasting, Time Series, and Regression (4th edition) by Bowerman, O'Connell, and Koehler: Chapter 3

MATH 4070: Regression and Time-Series Analysis

Whitney Huang
Clemson University

Notes

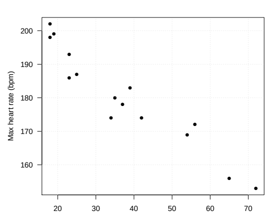
Agenda

1. **Simple Linear Regression**
2. **Parameter Estimation**
3. **Residual Analysis**
4. **Confidence/Prediction Intervals**
5. **Hypothesis Testing**
6. **Analysis of Variance (ANOVA) Approach to Regression**

Notes

What is Regression Analysis?

Regression analysis: A set of statistical procedures for estimating the relationship between a (numerical) response variable and predictor variable(s), at least one of which is numerical



Simple linear regression: The relationship between the response variable and the predictor variable is approximately linear

Notes

Simple Linear Regression (SLR)

Y: response variable; X: predictor variable

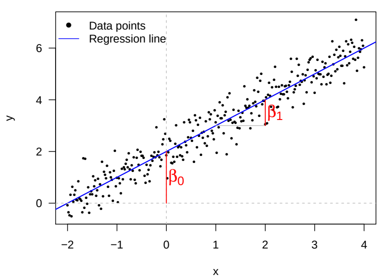
- In SLR we **assume** there is a **linear relationship** between X and Y:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- We need to estimate β_0 (intercept) and β_1 (slope) based on observed data $\{x_i, y_i\}_{i=1}^n$
- We can use the estimated regression equation to
  - make predictions
  - study the relationship between response and predictor
  - control the response
- Yet we need to quantify our estimation uncertainty regarding the linear relationship

Notes

Regression equation: $Y = \beta_0 + \beta_1 X$



- β_0: $\mathbb{E}[Y]$ when $X = 0$
- β_1: $\mathbb{E}[\Delta Y]$ when X increases by 1

Simple Linear Regression

Notes

Assumptions about the Random Error ε

In order to estimate β_0 and β_1, we make the following assumptions about ε

- $\mathbb{E}[\varepsilon_i] = 0$
- $\mathrm{Var}[\varepsilon_i] = \sigma^2$
- $\mathrm{Cov}[\varepsilon_i, \varepsilon_j] = 0, \quad i \neq j$

Therefore, we have

$$\mathbb{E}[Y_i] = \beta_0 + \beta_1 X_i, \text{ and}$$
$$\mathrm{Var}[Y_i] = \sigma^2$$

> The regression line $\beta_0 + \beta_1 X$ represents the **conditional mean curve** whereas σ^2 measures the magnitude of the **variation** around the regression curve

Simple Linear Regression

Notes

Parameter Estimation: Method of Least Squares

For given observations $\{x_i, y_i\}_{i=1}^n$, choose β_0 and β_1 to minimize the *sum of squared errors*:

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

Solving the above minimization problem requires some knowledge from Calculus (see notes `LS_SLR.pdf`)

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

We also need to **estimate** σ^2

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

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

Notes

Properties of Least Squares Estimators

- The estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased. That is

$$\mathbb{E}(\hat{\beta}_0) = \beta_0;$$
$$\mathbb{E}(\hat{\beta}_1 = \beta_1.$$

- The estimator $\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2}$ is unbiased. That is

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

We can write $\hat{\sigma}^2 = \frac{|\mathbf{y} - \hat{\mathbf{y}}|^2}{n-2}$, where $\mathbf{y} = (y_1, \cdots, y_n)^T$, $\hat{\mathbf{y}} = (\hat{\beta}_0 + \hat{\beta}_1 x_1, \cdots, \hat{\beta}_0 + \hat{\beta}_1 x_n)^T$.

Since $\hat{\mathbf{y}}$ has a dimension of 2 (regression slope and intercept), this leads to $n-2$ in the denominator

Simple Linear Regression
School of Mathematical and Statistical Sciences, Clemson University
Simple Linear Regression
Parameter Estimation
Residual Analysis
Confidence/Prediction Intervals
Hypothesis Testing
Analysis of Variance (ANOVA) Approach to Regression

Notes

Connection to Calculus: Derivation of β_1

Note that $\mathbb{E}[Y|X] = \beta_0 + \beta_1 X = \mu_Y + \beta_1(X - \mu_x)$. Now consider minimizing

$$g(b) = \mathbb{E}\left[(Y - \mu_Y - b(X - \mu_X))^2\right]$$

Note

$$\begin{aligned} g(b) &= \mathbb{E}\left[(Y-\mu_Y)^2\right] + b^2\mathbb{E}\left[(X-\mu_X)^2\right] - 2b\mathbb{E}\left[(Y-\mu_Y)(X-\mu_X)\right] \\ &= \sigma_Y^2 + b^2\sigma_X^2 - 2b\mathrm{Cov}(X,Y) \end{aligned}$$

Taking the derivative with respect to b:

$$g'(b) = 2b\sigma_X^2 - 2\mathrm{Cov}(X,Y)$$

Let β_1 solve $g'(b) = 0 \Rightarrow \beta_1 = \frac{\mathrm{Cov}(X,Y)}{\sigma_X^2}$

$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^n (x_i-\bar{x})^2} = \frac{\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})/(n-1)}{\sum_{i=1}^n (x_i-\bar{x})^2/(n-1)}$ is the sample counterpart

Simple Linear Regression
School of Mathematical and Statistical Sciences, Clemson University
Simple Linear Regression
Parameter Estimation
Residual Analysis
Confidence/Prediction Intervals
Hypothesis Testing
Analysis of Variance (ANOVA) Approach to Regression

Notes

Best Linear Predictor and Its Mean Square Error

Consider the mean square error (MSE) of the least square predictor

$$
\begin{aligned}
\mathbb{E}\left[(Y-\beta_0-\beta_1 X)^2\right] &= \operatorname{Var}(Y-\beta_0-\beta_1 X)\\
&= \operatorname{Cov}\left[(Y-\beta_1 X)(Y-\beta_1 X)\right]\\
&= \sigma_Y^2 - 2\beta_1 \operatorname{Cov}(X,Y) + \beta_1^2\sigma_X^2
\end{aligned}
$$

Now plug in $\beta_1 = \frac{\operatorname{Cov}(X,Y)}{\sigma_X^2}$, we have

$$
\begin{aligned}
\mathrm{MSE} &= \sigma_Y^2 - 2\frac{\operatorname{Cov}(X,Y)}{\sigma_X^2}\operatorname{Cov}(X,Y) + \left(\frac{\operatorname{Cov}(X,Y)}{\sigma_X^2}\right)^2\sigma_X^2\\
&= \sigma_Y^2 - 2\frac{\operatorname{Cov}(X,Y)^2}{\sigma_X^2} + \frac{\operatorname{Cov}(X,Y)^2}{\sigma_X^2}\\
&= \sigma_Y^2 - \frac{\operatorname{Cov}(X,Y)^2}{\sigma_X^2}\\
&= \sigma_Y^2(1-\rho^2)
\end{aligned}
$$

Notes

Geometric View of Least Squares Model Fit

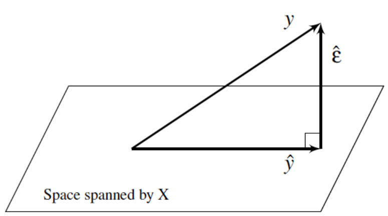


Figure courtesy of Faraway's *Linear Models with R* (2015, p. 15)

- $\mathbf{y} = (y_1, \cdots, y_n)^T$: The data vector
- $\hat{\mathbf{y}} = (\hat{y}_1 = \hat{\beta}_0 + \hat{\beta}_1 x_1, \cdots, \hat{y}_n = \hat{\beta}_0 + \hat{\beta}_1 x_n)^T$: The least squares fitted vector
- $\hat{\varepsilon} = (y_1 - \hat{y}_1, \cdots, y_n - \hat{y}_n)^T$: The residual vector

Notes

Example: Maximum Heart Rate vs. Age

The maximum heart rate `MaxHeartRate` of a person is often said to be related to age `Age` by the equation:

$$\text{MaxHeartRate} = 220 - \text{Age}.$$

Suppose we have 15 people of varying ages are tested for their maximum heart rate (bpm) (link to the "dataset": http://whitneyhuang83.github.io/maxHeartRate.csv)

1. Compute the estimates for the regression coefficients
2. Compute the fitted values
3. Compute the estimate for σ

Notes

Estimate the Parameters β_1, β_0, and σ^2

y_i and x_i are the Maximum Heart Rate and Age of the i[th] individual

- To obtain $\hat{\beta}_1$
  1. Compute $\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$, $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$
  2. Compute $y_i - \bar{y}$, $x_i - \bar{x}$, and $(x_i - \bar{x})^2$ for each observation
  3. Compute $\sum_i^n (x_i - \bar{x})(y_i - \bar{y})$ divided by $\sum_i^n (x_i - \bar{x})^2$
- $\hat{\beta}_0$: Compute $\bar{y} - \hat{\beta}_1 \bar{x}$
- $\hat{\sigma}^2$
  1. Compute the fitted values: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, \cdots, n$
  2. Compute the **residuals** $e_i = y_i - \hat{y}_i, \quad i = 1, \cdots, n$
  3. Compute the **residual sum of squares (RSS)** $= \sum_{i=1}^n (y_i - \hat{y}_i)^2$ and divided by $n-2$ (why?)

Notes

Let's Do the Calculations

Simple Linear Regression
School of MATHEMATICAL AND STATISTICAL SCIENCES Clemson University
Simple Linear Regression
Parameter Estimation
Residual Analysis
Confidence/Prediction Intervals
Hypothesis Testing
Analysis of Variance (ANOVA) Approach to Regression

$$\bar{x} = \sum_{i=1}^{15} \frac{18 + 23 + \cdots + 39 + 37}{15} = 37.33$$

$$\bar{y} = \sum_{i=1}^{15} \frac{202 + 186 + \cdots + 183 + 178}{15} = 180.27$$

X	18	23	25	35	65	54	34	56	72	19	23	42	18	39	37
Y	202	186	187	180	156	169	174	172	153	199	193	174	198	183	178
	-19.33	-14.33	-12.33	-2.33	27.67	16.67	-3.33	18.67	34.67	-18.33	-14.33	4.67	-19.33	1.67	-0.33
	21.73	5.73	6.73	-0.27	-24.27	-11.27	-6.27	-8.27	-27.27	18.73	12.73	-6.27	17.73	2.73	-2.27
	-420.18	-82.18	-83.04	0.62	-671.38	-187.78	20.89	-154.31	-945.24	-343.44	-182.51	-29.24	-342.84	4.56	0.76
	373.78	205.44	152.11	5.44	765.44	277.78	11.11	348.44	1201.78	336.11	205.44	21.78	373.78	2.78	0.11
	195.69	191.70	190.11	182.13	158.20	166.97	182.93	165.38	152.61	194.89	191.70	176.54	195.69	178.94	180.53

- $\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = -0.7977$
- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 210.0485$
- $\hat{\sigma}^2 = \frac{\sum_{i=1}^{15} (y_i - \hat{y}_i)^2}{13} = 20.9563 \Rightarrow \hat{\sigma} = 4.5778$

Notes

Let's Double Check

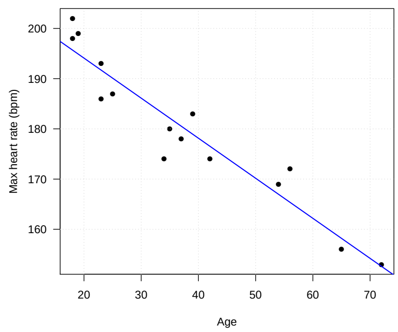
Simple Linear Regression
School of MATHEMATICAL AND STATISTICAL SCIENCES Clemson University
Simple Linear Regression
Parameter Estimation
Residual Analysis
Confidence/Prediction Intervals
Hypothesis Testing
Analysis of Variance (ANOVA) Approach to Regression

Output from R (RStudio)

```
> fit <- lm(MaxHeartRate ~ Age)
> summary(fit)

Call:
lm(formula = MaxHeartRate ~ Age)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.9258 -2.5383  0.3879  3.1867  6.6242 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 210.04846    2.86694   73.27  < 2e-16 ***
Age          -0.79773    0.06996  -11.40 3.85e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.578 on 13 degrees of freedom
Multiple R-squared:  0.9091,	Adjusted R-squared:  0.9021 
F-statistic:   130 on 1 and 13 DF,  p-value: 3.848e-08
```


Notes

Assessing Linear Regression Fit



Question: Is linear relationship between max heart rate and age reasonable? ⇒ Residual Analysis

Notes

Residuals

- The residuals are the differences between the observed and fitted values:

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

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

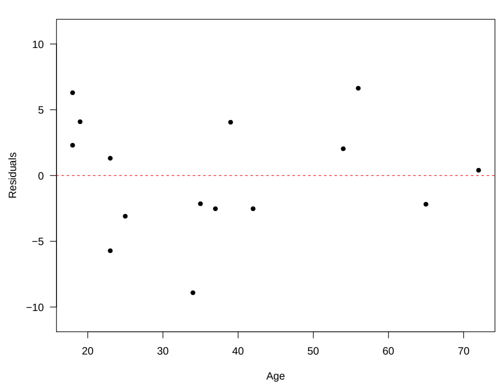
- Note that estimates aren't parameters, and residuals aren't random errors

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$
$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$$

- Nonetheless, residuals are very useful in assessing the appropriateness of the assumptions on ε_i. Recall
  - $\mathrm{E}[\varepsilon_i] = 0$
  - $\mathrm{Var}[\varepsilon_i] = \sigma^2$
  - $\mathrm{Cov}[\varepsilon_i, \varepsilon_j] = 0, \quad i \neq j$

Notes

Residuals Against Predictor Plot



Notes

Interpreting Residual Plots

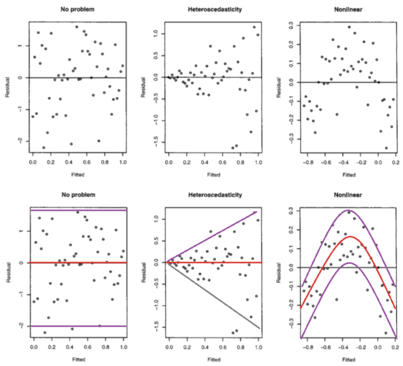
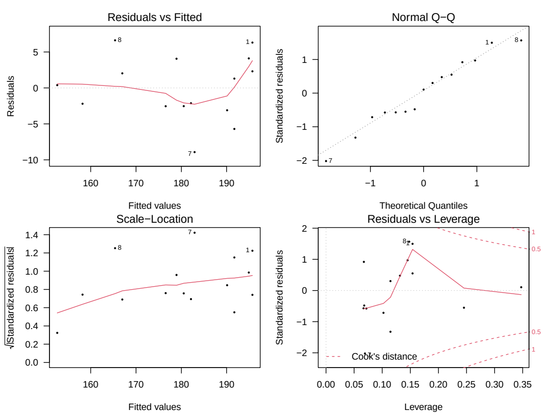


Figure courtesy of Faraway's Linear Models with R (2005, p. 59).

Notes

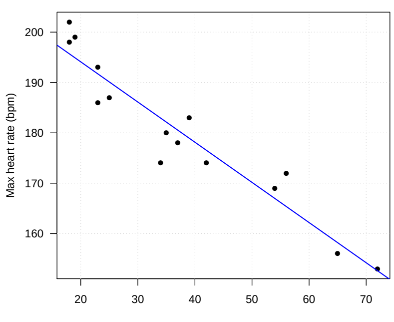
Diagnostic Plots in R



Notes

How (Un)certain We Are?

Remember: estimates are nor parameters



Can we formally quantify our estimation uncertainty?

$\Rightarrow$ We need additional (distributional) assumption on ε

Notes

Normal Error Regression Model

Recall the SLR model:

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

- Further assume $\varepsilon_i \overset{i.i.d}{\sim} \mathrm{N}(0, \sigma^2) \Rightarrow Y_i|X_i \sim \mathrm{N}(\beta_0 + \beta_1 X_i, \sigma^2)$
- With normality assumption, we can derive the **sampling distribution** of $\hat{\beta}_1$ and $\hat{\beta}_0 \Rightarrow$

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\mathrm{se}}(\hat{\beta}_1)} \sim t_{n-2}, \quad \hat{\mathrm{se}}(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$
$$\frac{\hat{\beta}_0 - \beta_0}{\hat{\mathrm{se}}(\hat{\beta}_0)} \sim t_{n-2}, \quad \hat{\mathrm{se}}(\hat{\beta}_0) = \hat{\sigma}\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)}$$

where t_{n-2} denotes the Student's t distribution with $n-2$ degrees of freedom

Notes

Deviation of $\mathrm{se}(\hat{\beta}_1)$

Recall $\hat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}_1) &= \mathrm{Var}\left(\frac{\sum_{i=1}^n (Y_i - \bar{Y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) \\
&= \mathrm{Var}\left(\frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) \\
&= \left(\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)^2 \left(\sum_{i=1}^n (x_i - \bar{x})^2\right) \mathrm{Var}(Y_i) \\
&= \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}
\end{aligned}$$

$\mathrm{se}(\hat{\beta}_1) = \sqrt{\mathrm{Var}(\hat{\beta})} = \frac{\sigma}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$. Replacing σ by $\hat{\sigma}$ to get $\hat{\mathrm{se}}(\hat{\beta}_1)$

Notes

Deviation of $\mathrm{se}(\hat{\beta}_0)$

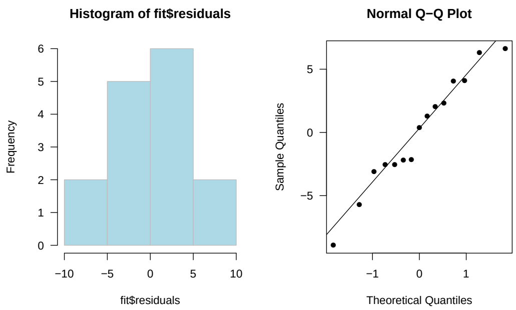
Recall $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}_0) &= \mathrm{Var}\left(\bar{Y} - \hat{\beta}_1 \bar{x}\right) \\
&= \mathrm{Var}(\bar{Y}) + \mathrm{Var}(-\hat{\beta}_1 \bar{x}) - 2\mathrm{Cov}(\bar{Y}, \bar{x}\hat{\beta}_1) \\
&= \frac{\sigma^2}{n} + \bar{x}^2 \left(\frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})}\right) - \cancel{2\mathrm{Cov}(\bar{Y}, \bar{x}\hat{\beta}_1)} \\
&= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)
\end{aligned}$$

Taking the square root and replacing σ with $\hat{\sigma}$ yields $\hat{\mathrm{se}}(\hat{\beta}_0)$

Notes

Assessing Normality Assumption on ε



The Q-Q plot is more effective in detecting subtle departures from normality, especially in the tails.

Notes

Confidence Intervals

- Recall $\frac{\hat{\beta}_1-\beta_1}{\hat{\text{se}}_{\hat{\beta}_1}} \sim t_{n-2}$, we use this fact to construct **confidence intervals (CIs)** for β_1:

$$\left[\hat{\beta}_1 - t_{1-\alpha/2,n-2}\hat{\sigma}_{\hat{\beta}_1}, \hat{\beta}_1 + t_{1-\alpha/2,n-2}\hat{\sigma}_{\hat{\beta}_1}\right],$$

  where α is the **confidence level** and $t_{1-\alpha/2,n-2}$ denotes the $1-\alpha/2$ percentile of a student's t-distribution with $n-2$ degrees of freedom

- Similarly, we can construct CIs for β_0:

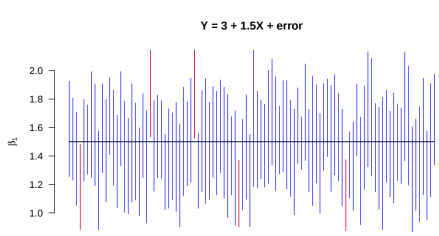
$$\left[\hat{\beta}_0 - t_{1-\alpha/2,n-2}\hat{\sigma}_{\hat{\beta}_0}, \hat{\beta}_0 + t_{1-\alpha/2,n-2}\hat{\sigma}_{\hat{\beta}_0}\right]$$

Simple Linear Regression
Simple Linear Regression
Parameter Estimation
Residual Analysis
Confidence/Prediction Intervals
Hypothesis Testing
Analysis of Variance (ANOVA) Approach to Regression

Notes

Understanding Confidence Intervals

- Suppose $Y = \beta_0 + \beta_1 X + \varepsilon$, where $\beta_0 = 3$, $\beta_1 = 1.5$ and $\sigma^2 \sim \mathrm{N}(0,1)$
- We take 100 random sample each with sample size 20
- We then construct the 95% CI for each random sample ($\Rightarrow$ 100 CIs)



Simple Linear Regression
Simple Linear Regression
Parameter Estimation
Residual Analysis
Confidence/Prediction Intervals
Hypothesis Testing
Analysis of Variance (ANOVA) Approach to Regression

Notes

Interval Estimation of $\mathrm{E}(Y_h)$

- We often interested in estimating the **mean** response for a particular value of predictor, say, X_h. Therefore we would like to construct CI for $\mathrm{E}[Y_h]$
- We need sampling distribution of $\hat{Y}_h$ to form CI:
  - $\frac{\hat{Y}_h - Y_h}{\hat{\sigma}_{\hat{Y}_h}} \sim t_{n-2}, \quad \hat{\sigma}_{\hat{Y}_h} = \hat{\sigma}\sqrt{\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)}$
  - CI:

$$\left[\hat{Y}_h - t_{1-\alpha/2,n-2}\hat{\sigma}_{\hat{Y}_h}, \hat{Y}_h + t_{1-\alpha/2,n-2}\hat{\sigma}_{\hat{Y}_h}\right]$$

- **Quiz:** Use this formula to construct CI for β_0

Notes

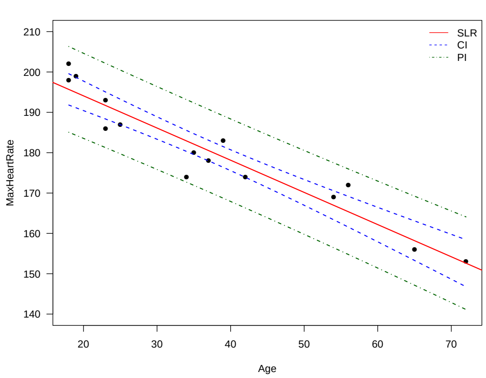
Prediction Intervals

- Suppose we want to predict the response of a future observation given $X = X_h$
- We need to account for added variability as a new observation does not fall directly on the regression line (i.e., $Y_{\text{h(new)}} = \mathrm{E}[Y_h] + \varepsilon_h$)
- Replace $\hat{\sigma}_{\hat{Y}_h}$ by $\hat{\sigma}_{\hat{Y}_{\text{h(new)}}} = \hat{\sigma}\sqrt{\left(1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)}$ to construct CIs for $Y_{\text{h(new)}}$

Simple Linear Regression
School of Mathematical and Statistical Sciences, Clemson University
Simple Linear Regression
Parameter Estimation
Residual Analysis
Confidence/Prediction Intervals
Hypothesis Testing
Analysis of Variance (ANOVA) Approach to Regression

Notes

Confidence Intervals vs. Prediction Intervals



Simple Linear Regression
School of Mathematical and Statistical Sciences, Clemson University
Simple Linear Regression
Parameter Estimation
Residual Analysis
Confidence/Prediction Intervals
Hypothesis Testing
Analysis of Variance (ANOVA) Approach to Regression

Notes

Maximum Heart Rate vs. Age Revisited

The maximum heart rate `MaxHeartRate` (HR_{max}) of a person is often said to be related to age `Age` by the equation:

$$\text{HR}_{max} = 220 - \text{Age}.$$

Suppose we have 15 people of varying ages are tested for their maximum heart rate (bpm)

Age	18	23	25	35	65	54	34	56	72	19	23	42	18	39	37
HR_{max}	202	186	187	180	156	169	174	172	153	199	193	174	198	183	178

- Construct the 95% CI for β_1
- Compute the estimate for mean `MaxHeartRate` given `Age` $= 40$ and construct the associated 90% CI
- Construct the prediction interval for a new observation given `Age` $= 40$

Notes

Simple Linear Regression
School of MATHEMATICAL AND STATISTICAL SCIENCES Clemson University
Simple Linear Regression
Parameter Estimation
Residual Analysis
Confidence/Prediction Intervals
Hypothesis Testing
Analysis of Variance (ANOVA) Approach to Regression

Maximum Heart Rate vs. Age: Hypothesis Test for Slope

1. $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$
2. Compute the **test statistic**:
   $t^* = \frac{\hat{\beta}_1 - 0}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{-0.7977}{0.06996} = -11.40$
3. Compute **p-value**: $\mathbb{P}(|t^*| \geq |t_{obs}|) = 3.85 \times 10^{-8}$
4. Compare to α and draw conclusion:

> Reject H_0 at $\alpha = .05$ level, evidence suggests a negative linear relationship between `MaxHeartRate` and `Age`

Notes

Simple Linear Regression
School of MATHEMATICAL AND STATISTICAL SCIENCES Clemson University
Simple Linear Regression
Parameter Estimation
Residual Analysis
Confidence/Prediction Intervals
Hypothesis Testing
Analysis of Variance (ANOVA) Approach to Regression

Maximum Heart Rate vs. Age: Hypothesis Test for Intercept

1. $H_0 : \beta_0 = 0$ vs. $H_a : \beta_0 \neq 0$
2. Compute the **test statistic**:
   $t^* = \frac{\hat{\beta}_0 - 0}{\hat{\sigma}_{\hat{\beta}_0}} = \frac{210.0485}{2.86694} = 73.27$
3. Compute **p-value**: $\mathbb{P}(|t^*| \geq |t_{obs}|) \simeq 0$
4. Compare to α and draw conclusion:

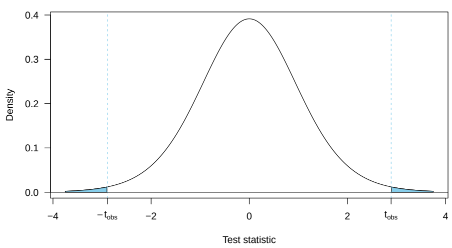
> Reject H_0 at $\alpha = .05$ level, evidence suggests evidence suggests the intercept (the expected `MaxHeartRate` at age 0) is different from 0

Notes

Hypothesis Tests for $\beta_{\text{age}} = -1$

$H_0 : \beta_{\text{age}} = -1$ vs. $H_a : \beta_{\text{age}} \neq -1$

Test Statistic: $\frac{\hat{\beta}_{\text{age}} - (-1)}{\hat{\sigma}_{\hat{\beta}_{\text{age}}}} = \frac{-0.79773 - (-1)}{0.06996} = 2.8912$



p-value: $2 \times \mathbb{P}(t^* > 2.8912) = 0.013$, where $t^* \sim t_{df=13}$

Notes

Analysis of Variance (ANOVA) Approach to Regression

Partitioning Sums of Squares

- Total sums of squares in response

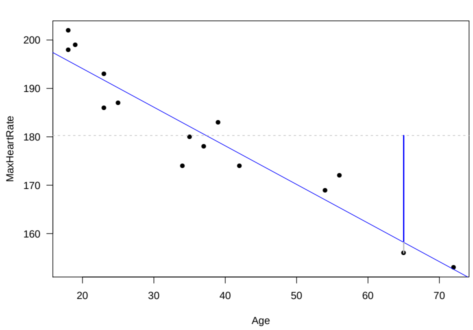
$$\text{SST} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

- We can rewrite SST as

$$\begin{aligned}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 \\ &= \underbrace{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}_{\text{Error}} + \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{\text{Model}}\end{aligned}$$

Notes

Partitioning Total Sums of Squares



Notes

Total Sum of Squares: SST

- If we ignored the predictor X, the $\bar{Y}$ would be the best (linear unbiased) predictor

$$Y_i = \beta_0 + \varepsilon_i \tag{1}$$

- SST is the sum of squared deviations for this predictor (i.e., $\bar{Y}$)
- The **total mean square** is $\mathrm{SST}/(n-1)$ and represents an unbiased estimate of σ^2 under the model (1)

Notes

Regression Sum of Squares: SSR

- SSR: $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
- Degrees of freedom is 1 due to the inclusion of the slope, i.e.,

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{2}$$

- "Large" $\mathrm{MSR} = \mathrm{SSR}/1$ suggests a linear trend, because

$$\mathrm{E}[\mathrm{MSR}] = \sigma^2 + \beta_1^2 \sum_{i=1}^{n}(X_i - \bar{X})^2$$

Simple Linear Regression
School of Mathematical and Statistical Sciences Clemson University
Simple Linear Regression
Parameter Estimation
Residual Analysis
Confidence/Prediction Intervals
Hypothesis Testing
Analysis of Variance (ANOVA) Approach to Regression

Notes

Error Sum of Squares: SSE

- SSE is simply the sum of squared residuals

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

- Degrees of freedom is $n-2$ (Why?)
- SSE large when $|$residuals$|$ are "large" $\Rightarrow$ Y_i's vary substantially around fitted regression line
- $\mathrm{MSE} = \mathrm{SSE}/(n-2)$ and represents an unbiased estimate of σ^2 **when taking X into account**

Simple Linear Regression
School of Mathematical and Statistical Sciences Clemson University
Simple Linear Regression
Parameter Estimation
Residual Analysis
Confidence/Prediction Intervals
Hypothesis Testing
Analysis of Variance (ANOVA) Approach to Regression

Notes

ANOVA Table and F-Test

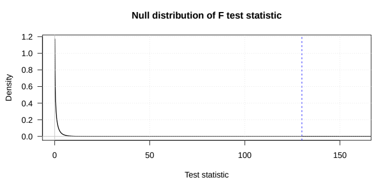
Source	df	SS	MS
Model	1	$\text{SSR} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$	$\text{MSR} = \text{SSR}/1$
Error	$n-2$	$\text{SSE} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	$\text{MSE} = \text{SSE}/(n-2)$
Total	$n-1$	$\text{SST} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$	

- **Goal:** To test $H_0 : \beta_1 = 0$
- Test statistics $F^* = \frac{\text{MSR}}{\text{MSE}}$
- If $\beta_1 = 0$ then F^* should be near one $\Rightarrow$ reject H_0 when F^* "large"
- We need sampling distribution of F^* under $H_0 \Rightarrow F_{1,n-2}$, where F_{d_1,d_2} denotes a F distribution with degrees of freedom $d_1 = 1$ and $d_2 = n-2$

Notes

F-Test: $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$

```
fit <- lm(MaxHeartRate ~ Age)
anova(fit)
```
```

Analysis of Variance Table

Response: MaxHeartRate
          Df  Sum Sq Mean Sq F value
Age        1 2724.50 2724.50  130.01
Residuals 13  272.43   20.96
             Pr(>F)
Age       3.848e-08 ***
```



Notes

SLR: F-Test vs. T-test

ANOVA Table and F-Test

```
Analysis of Variance Table

Response: MaxHeartRate
          Df  Sum Sq Mean Sq
Age        1 2724.50 2724.50
Residuals 13  272.43   20.96
          F value    Pr(>F)
Age        130.01 3.848e-08
```

Parameter Estimation and T-Test

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 210.04846    2.86694   73.27  < 2e-16
Age          -0.79773    0.06996  -11.40 3.85e-08
```


Notes

Summary

This week, we have learned

- Simple Linear Regression:
  $Y = \beta_0 + \beta_1 X + \varepsilon,\ \varepsilon \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$

- Method of Least Squares for parameter estimation

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}=(\beta_0,\beta_1)}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

- Residual analysis to check model assumptions

- Confidence/Prediction Intervals and Hypothesis Testing

School of Mathematical and Statistical Sciences, Clemson University

Notes

R Funcations

- Fitting linear models

  `object <- lm(formula, data)` where the formula is specified via `y ~ x` $\Rightarrow y$ is modeled as a linear function of x

- Diagnostic plots

```
plot(object)
```

- Summarizing fits

```
summary(object)
```

- Making predictions

```
predict(object, newdata)
```

- Confidence Intervals for Model Parameters

```
confint(object)
```


School of Mathematical and Statistical Sciences, Clemson University

Notes

Notes