Lecture 4

Simple Linear Regression III

Reading: Chapter 11

STAT 8020 Statistical Methods II
August 28, 2019

Whitney Huang
Clemson University

Agenda

1. **Review of Last Class**
2. **Confidence/Prediction Intervals**
3. **Analysis of Variance (ANOVA) Approach to Regression**

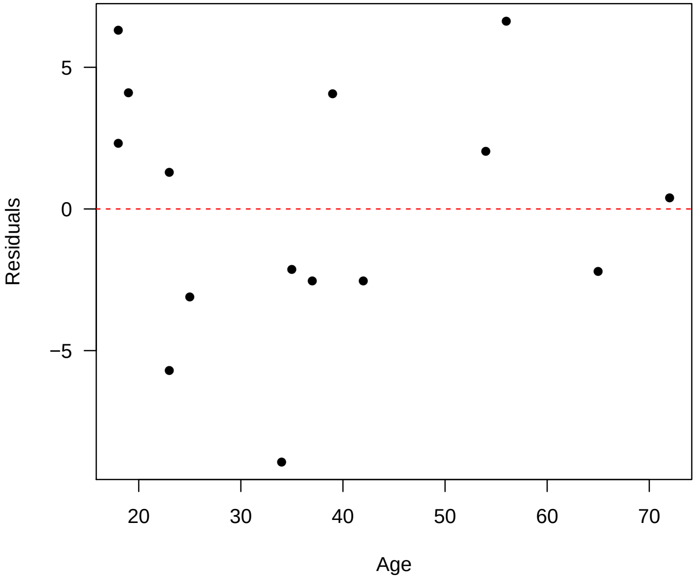
Last Class

- **Residual Analysis:** To check the appropriateness of SLR model
  - Is the regression function linear?
  - Do ε_i's have constant variance σ^2?
  - Are ε_i's indepdent to each other?

  We could plot residuals (e_i's) against predictor variable to assess these

- **Hypothesis Tests for β_1 and β_0**
  - With additional normality assumption on ε, we obtained the **sampling distribution** for $\hat{\beta}_{1,\text{LS}}$ and $\hat{\beta}_{0,\text{LS}}$
  - Test statistic $\left(\hat{\beta}_{1,\text{LS}} - \beta_1\right) / \hat{\sigma}_{\hat{\beta}_{1,\text{LS}}} \sim t_{n-2}$. With hypothesized value β_1^* (i.e., $H_0 : \beta_1 = \beta_1^*$), H_a and significant level α, we can compute the **P-value** to perform a test

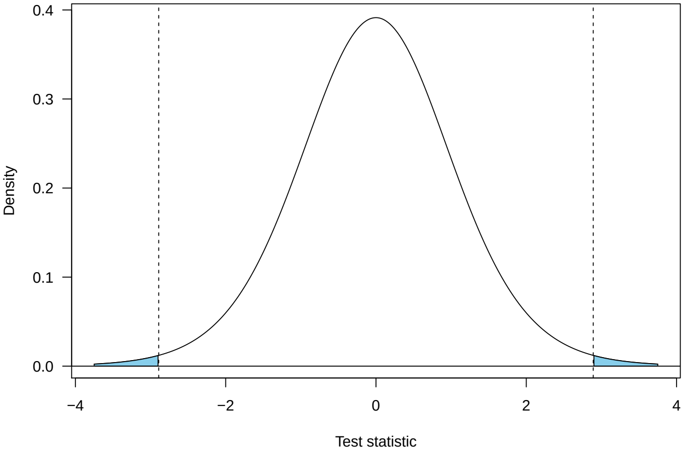
Residual Plot: e_i's vs. X's



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

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

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



Confidence Intervals

- Recall $\frac{\hat{\beta}_{1,\text{LS}}-\beta_1}{\hat{\sigma}_{\hat{\beta}_{1,\text{LS}}}} \sim t_{n-2}$, we use this fact to construct **confidence intervals (CIs)** for β_1:

$$\left[\hat{\beta}_{1,\text{LS}} - t(1-\alpha/2, n-2)\hat{\sigma}_{\hat{\beta}_{1,\text{LS}}}, \hat{\beta}_{1,\text{LS}} + t(1-\alpha/2, n-2)\hat{\sigma}_{\hat{\beta}_{1,\text{LS}}}\right],$$

  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:

$$\left[\hat{\beta}_{0,\text{LS}} - t(1-\alpha/2, n-2)\hat{\sigma}_{\hat{\beta}_{0,\text{LS}}}, \hat{\beta}_{0,\text{LS}} + t(1-\alpha/2, n-2)\hat{\sigma}_{\hat{\beta}_{0,\text{LS}}}\right]$$

- Interpretation?

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

- We often interested in estimating the **mean** response for 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:

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

Prediction Intervals

- Suppose we want to “predcit” 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 $Y_{\text{h(new)}}$

Maximum Heart Rate vs. Age Revisited

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)

Age	18	23	25	35	65	54	34	56	72	19	23	42	18	39	37
MaxHeartRate	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 $= 40$

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

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

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 SST$/(n-1)$ and represents an unbiased estimate of σ^2 under the model (1).

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" MSR $=$ SSR$/1$ suggests a linear trend, because

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

Error Sum of Squares: SSE

- SSE is simply the sum of squared residuals

$$\text{SSE} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

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

ANOVA Table and F test

Source	df	SS	MS
Model	1	SSR $= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$	MSR = SSR/1
Error	$n-2$	SSE $= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	MSE = SSE/(n-2)
Total	$n-1$	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 and d_2