Lecture 16

Review

STAT 8020 Statistical Methods II
September 25, 2019

Whitney Huang
Clemson University

Simple Linear Regression

Y: response variable; X: predictor variable

- Simple Linear Regression Model:

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon, \quad \varepsilon \sim \mathrm{N}(0, \sigma^2)$$

  where β_0 is the intercept and β_1 is the slope

- Use method of least squares to estimate the parameters
  - $\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$
  - $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$
  - $\hat{\sigma}^2 = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 / (n-2)$

Residual Analysis

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

$$e_i = Y_i - \hat{Y}_i,$$

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

We use residuals to assess the assumptions on ε:

- $\mathrm{E}[\varepsilon_i] = 0$
- $\mathrm{Var}[\varepsilon_i] = \sigma^2$
- $\mathrm{Cov}[\varepsilon_i, \varepsilon_j] = 0, \quad i \neq j$
- ε follows a normal distribution

Sampling distribution of $\hat{\beta}_1$ and $\hat{\beta}_1$

CLEMSON UNIVERSITY

- $\hat{\beta}_1 \sim N(\beta_1, \sigma^2_{\hat{\beta}_1})$ where $\sigma_{\hat{\beta}_1} = \sigma / \sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}$

$$\Rightarrow \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim t_{n-2}$$

- $\hat{\beta}_0 \sim N(\beta_0, \sigma^2_{\hat{\beta}_0})$ where $\sigma_{\beta_0} = \sigma \sqrt{\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)}$

$$\Rightarrow \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\hat{\beta}_0}} \sim t_{n-2}$$

Hypothesis Test for Slope

1. $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$
2. Compute the **test statistic**: $t^* = \frac{\hat{\beta}_1 - 0}{\hat{\sigma}_{\hat{\beta}_1}}$
3. Compute the P-value
4. Compare to α, the prespecified significant level, and draw conclusion

Confidence Intervals (CIs) for β_1 and β_0

- $100 \times (1-\alpha)$% CI for β_1:

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

- $100 \times (1-\alpha)$% CI for β_0:

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

Confidence/Prediction Intervals for Response

Let Y_h be the response given that $X = X_h$

- CI for $\mathbb{E}(Y_h)$:

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

where $\hat{\sigma}_{\hat{Y}_h} = \hat{\sigma}\sqrt{\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)}$

- PI for Y_h: Replace $\hat{\sigma}_{\hat{Y}_h}$ by

$$\hat{\sigma}_{\hat{Y}_{\text{h(new)}}} = \hat{\sigma}\sqrt{\left(1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)}$$

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

- **F-test:** To test $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$
- Test statistics $F^* = \frac{\text{MSR}}{\text{MSE}}$
- 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

Coefficient of Determination R^2

Defined as the proportion of total variation explained by a simple regression model:

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

Multiple Linear Regression

Goal: To model the relationship between two or more explanatory variables (X's) and a response variable (Y) by fitting a **linear equation** to observed data:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1} + \varepsilon_i, \quad \varepsilon_i \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma^2)$$

Matrix form: $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

- All Quantitative Predictors
- Both Quantitative and Qualitative Predictors
- Polynomial Regression

ANOVA Table

Source	df	SS	MS	F Value
Model	$p-1$	SSR	MSR = SSR/$(p-1)$	MSR/MSE
Error	$n-p$	SSE	MSE = SSE/$(n-p)$	
Total	$n-1$	SST		

- F-test: Tests if the predictors $\{X_1, \cdots, X_{p-1}\}$ collectively help explain the variation in Y
  - $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$
  - $H_a :$ at least one $\beta_k \neq 0, \quad 1 \leq k \leq p-1$
  - $F^* = \frac{\text{MSR}}{\text{MSE}} = \frac{\text{SSR}/(p-1)}{\text{SSE}/(n-p)} \overset{H_0}{\sim} F(p-1, n-p)$
  - Reject H_0 if $F^* > F(1-\alpha, p-1, n-p)$

CLEMSON UNIVERSITY

Testing Individual Predictor

- $\hat{\boldsymbol{\beta}} \sim \mathrm{N}_p\left(\boldsymbol{\beta}, \sigma^2\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\right) \Rightarrow \hat{\beta}_k \sim \mathrm{N}(\beta_k, \sigma^2_{\hat{\beta}_k})$
- Perform t-test:
  - $H_0 : \beta_k = 0$ vs. $H_a : \beta_k \neq 0$
  - $\frac{\hat{\beta}_k - \beta_k}{\hat{\sigma}_{\hat{\beta}_k}} \sim t_{n-p} \Rightarrow t^* = \frac{\hat{\beta}_k}{\hat{\sigma}_{\hat{\beta}_k}} \overset{H_0}{\sim} t_{n-p}$
  - Reject H_0 if $|t^*| > t_{1-\alpha/2,n-p}$
- Confidence interval for β_k: $\hat{\beta}_k \pm t_{1-\alpha/2,n-p}\hat{\sigma}_{\hat{\beta}_k}$

General Linear Test

- Comparison of a “full model” and “reduced model” that involves a subset of full model predictors
- Consider a full model with k predictors and reduced model with ℓ predictors ($\ell < k$)
- Test statistic: $F^* = \frac{\text{SSE(R)} - \text{SSE}(F)/(k-\ell)}{\text{SSE}(F)/(n-k-1)}$ $\Rightarrow$ Testing H_0 that the regression coefficients for the extra variables are all zero

Multicollinearity

Multicollinearity is a phenomenon of high inter-correlations among the predictor variables

- β's are not well estimated
- Spurious regression coefficient estimates
- R^2 and predicted values are usually OK

Model Selection

- Model Selection Criteria
  - Mallows' Cp
  - Adjusted R^2
  - Predicted Residual Sum of Squares (PRESS)
  - AIC
  - BIC
- Automatic Search Procedures
  - Stepwise Search
  - All Subset Selection

Model Diagnostics

- Leverage
- Studentized & Studentized Deleted Residuals
- Influential Observations: DFFITS
- Variance Inflation Factor (VIF)