Statistical Inference and Multivariate Analysis (MA324)

LECTURE SLIDES
Lecture 27

Simple Linear Regression

Indian Institute of Technology Guwahati

Jan-May 2023

Hypothesis Testing: β_1

- The fourth assumption of linear regression is required for testing of hypothesis: Errors are independent and normally distributed.

- Want to test the hypothesis that the **slope** parameter (β_1) equals to a constant (a value, say β_{10}):

$$H_0 : \beta_1 = \beta_{10} \text{ ag. } H_1 : \beta_1 \neq \beta_{10}$$

- Note that, $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ and y_i's are independent.

- $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}}) \implies z = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/S_{xx}}} \sim N(0,1)$. But σ is unknown.

- $\frac{(n-2)MS_{Res}}{\sigma^2} \sim \chi^2_{n-2}$. Also MS_{Res} and $\hat{\beta}_1$ are independent.

- Therefore, the test statistic is

$$t = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{MS_{Res}/S_{xx}}} \sim t_{n-2}, \text{ under } H_0.$$

- Reject H_0 iff $|t| > t_{n-2,\alpha/2}$; (at level α).

Hypothesis Testing: β_0

- Want to test the hypothesis that the **intercept** parameter (β_0) equals to a constant (a value, say β_{00}):

$$H_0 : \beta_0 = \beta_{00} \text{ ag. } H_1 : \beta_0 \neq \beta_{00}$$

- $\hat{\beta}_0 \sim N(\beta_0, \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})) \implies z = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})}} \sim N(0, 1)$. But σ is unknown.
- $\frac{(n-2)MS_{Res}}{\sigma^2} \sim \chi^2_{n-2}$. Also MS_{Res} and $\hat{\beta}_1$ are independent.
- Therefore, the test statistic is

$$t = \frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{MS_{Res}(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})}} \sim t_{n-2}, \text{ under } H_0.$$

- Reject H_0 iff $|t| > t_{n-2,\alpha/2}$; (at level α).

Interval Estimation: β_0 and β_1

- To get the CI for β_0 and β_1, the pivots are

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{MS_{Res}(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})}}, \text{ and } \frac{\hat{\beta}_1 - \beta_1}{\sqrt{MS_{Res}/S_{xx}}}, \text{ respectively.}$$

- A $100(1-\alpha)\%$ CI for β_0 is

$$\left[\hat{\beta}_0 \pm t_{n-2,\alpha/2}\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}\right].$$

- A $100(1-\alpha)\%$ CI for β_1 is

$$\left[\hat{\beta}_1 \pm t_{n-2,\alpha/2}\sqrt{\frac{MS_{Res}}{S_{xx}}}\right].$$

Interval Estimation: CI for σ^2

- To get the CI for σ^2, the pivots is

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

- A $100(1-\alpha)\%$ CI for σ^2 is

$$\left[\frac{(n-2)MS_{Res}}{\chi^2_{n-2;\alpha/2}}, \frac{(n-2)MS_{Res}}{\chi^2_{n-2;1-\alpha/2}}\right].$$

Interval Estimation: CI for mean response

- A regression model can be used to **estimate the mean response** $E(y)$ for a particular value of the regressor variable x. Let x_0 be a value of the regressor variable (must be with the range of original data on x). Then $E(y|x_0) = \beta_0 + \beta_1 x_0$.
- Then, $\widehat{y_0} = \widehat{E(y|x_0)} = \hat{\beta}_0 + \hat{\beta}_1 x_0$
- And, $\widehat{y_0} \sim N\left(\beta_0 + \beta_1 x_0, \sigma^2\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)\right)$
- Pivot: $\frac{\widehat{y_0} - (\beta_0 + \beta_1 x_0)}{\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)}} \sim t_{n-2}$
- A $100(1-\alpha)\%$ CI for $\beta_0 + \beta_1 x_0$ is

$$\left[\widehat{y_0} \pm t_{n-2;\alpha/2}\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)}\right].$$

Prediction Interval for New Observation:

- Let x_0 be a value of the regressor variable (must be with the range of original data on x).
- The true value of the response is y_0 (corresponding to x_0).
- We want to provide an interval I such that $P(y_0 \in I) = 1 - \alpha$
- Note that the point estimate of y_0 is $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$.
- Consider $\psi = y_0 - \hat{y}_0$.
- Then, $E(\psi) = 0, Var(\psi) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)$
- $\psi \sim N\left(0, \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\right)$
- Pivot: $\frac{y_0 - \hat{y}_0}{\sqrt{MS_{Res}\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}} \sim t_{n-2}$

Prediction Interval for New Observation:

- A $100(1-\alpha)\%$ prediction interval is

$$\left[\hat{y}_0 \pm t_{n-2;\alpha/2}\sqrt{MS_{Res}\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}\right].$$

Coefficient of determination: R^2

- ***Coefficient of determination:***

$$R^2 = \frac{SS_{Reg}}{SS_T} = 1 - \frac{SS_{Res}}{SS_T} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2},$$

  where $SS_{Reg} = \sum_i (\hat{y}_i - \bar{y})^2$

- It is a bounded quantity: $0 \leq R^2 \leq 1$
- R^2 is interpreted as the **"percentage of variation explained by the model".**
- **Higher values of R^2 are desirable** (R^2 close to 1 indicates a good fit), but "how high is high" **depends on the context.**
- In **multiple linear regression**, **adding a variable to a model can increase the value of R^2**. To overcome this problem:

Adjusted R^2, $\quad R^2_{adj} = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2/(n-p)}{\sum_i (\hat{y}_i - \bar{y})^2/(n-1)}.$

F -Test for Regression

- Large value of the sum of squares for regression $SS_{Reg} = SS_T - SS_{Res}$ indicates the simple regression mean function $E(Y|X = x) = \beta_0 + \beta_1 x$ should be a **significant improvement over the mean function** $E(y|X = x) = \beta_0$.

- This is equivalent to saying that the **additional parameter** (β_1) in the simple regression mean function is **non-zero**. In other words, that $E(Y|X = x)$ is not constant with respect to X.

- To test

$$H_0 : E(Y|X = x) = \beta_0 \text{ ag. } H_1 : E(Y|X = x) = \beta_0 + \beta_1 x$$

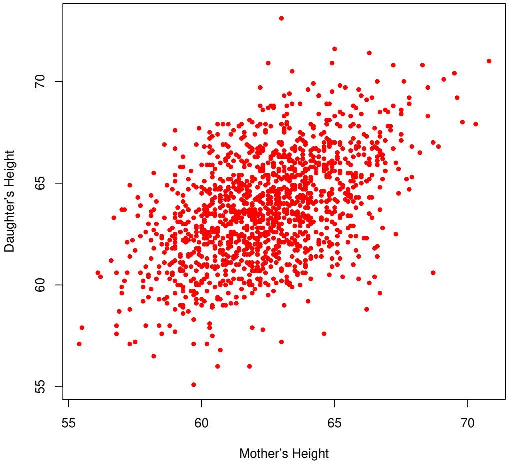
- Test statistics is a ratio, defined as F :

$$F = \frac{SS_{Reg}/1}{\hat{\sigma}^2} = \frac{SS_{Reg}/1}{SS_{Res}/(n-2)} \sim F_{1,n-2}$$

Simple Linear Regression in Heights Data

Heights Data Example

- Data on **heights of** $n = 1375$ **mothers** in the UK under the age of 65 and one of their **adult daughters over the age of 18** (collected and organized during the period 1893–1898 by the famous statistician Karl Pearson).
- A historical **use of regression to study inheritance of height from generation to generation.**

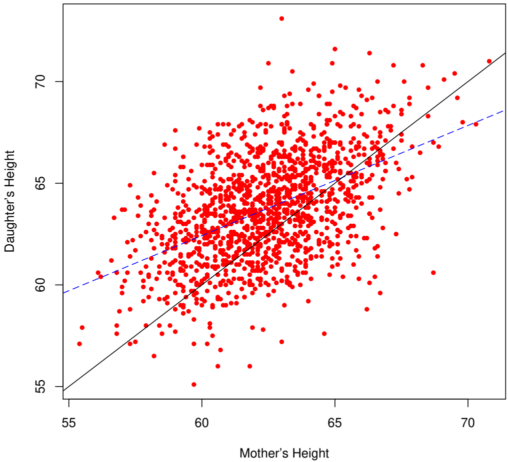

Daughter's Height
55
60
65
70
55
60
65
70
Mother's Height

Simple Linear Regression: Using R

```
> height.lm <- lm(Dheight~Mheight,data=heights)
> summary(height.lm)
...
Coefficients:
             Estimate  Std.Error  t-value Pr(>|t|)
(Intercept) 29.91744    1.62247   18.44   <2e-16 ***
Mheight      0.54175    0.02596   20.87   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

Residual standard error: 2.266 on 1373 degrees of
freedom Multiple R-squared: 0.2408,
adjusted R-squared: 0.2402
F-statistic: 435.5 on 1 and 1373 DF,
p-value: < 2.2e-16
```

Let's look at the Regression Line

```
> plot(Dheight ~ Mheight, heights,
xlab="Mother's Height", ylab="Daughter's Height",
pch=20, col=2)

> abline(coef(height.lm), lty=5, col=4)

> abline(0,1)
```


Daughter's Height
55
60
65
70
Mother's Height
55
60
65
70

Regression Effect

- It's an empirical phenomenon, also called "regression to the mean" or "regression to mediocrity", e.g.
  - Daughters of tall mothers tend to be tall, but not as tall as their mothers; daughters of short mothers tend to be short, but not as short as their mothers (same trend between sons and fathers)
  - Students doing well in Mid-term tend to do well in the Final, but not as well as the Mid-term; students doing poorly in Mid-term tend to do poorly in the Final, but not as poorly as the Mid-term!