Linear Regression Review

(AST405) Lifetime data analysis

Lecture Outline

Section 1

0. Linear Regression Models

Subsection 1

0.1 Multiple Linear Regression Models

Multiple Linear Regression Models

- Let $\{(y_i, x_{i1}, \ldots, x_{ip}), i = 1, \ldots, n\}$ be the data obtained from the i^{th} subject
  - $y_i \rightarrow$ response
  - $x_{ij} \rightarrow jth$ independent variable

Multiple Linear Regression Models

- A multiple linear regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_{ij}$$

- In matrix notation

$$\mathbf{y} = \mathbf{x}'\beta + \epsilon$$

  - $\mathbf{y} = (y_1, \dots, y_n)'$
  - $\mathbf{x} \to n \times (p+1)$ matrix with first column is a vector of one's
  - $\beta = (\beta_0, \beta_1, \dots, \beta_p)'$
  - $\epsilon_{ij} \to$ error term

Multiple Linear Regression Models

- General assumptions
  - ϵ's are independent
  - $E(\epsilon_{ij}) = 0$
  - $V(\epsilon_{ij}) = \sigma^2$
  - $\epsilon_{ij} \sim N(0, \sigma^2)$

Multiple Linear Regression Models

- The fitted model

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$$

  - Ordinary least squares or maximum likelihood estimators

$$\hat{\beta} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y}$$

  - Asymptotically $\hat{\beta}$ follows a normal distribution with mean vector β and variance-covariance matrix $(\mathbf{x}'\mathbf{x})^{-1}\sigma^2$

Multiple Linear Regression Models

- Residuals

$$\hat{\epsilon} = \mathbf{y} - \hat{\mathbf{y}}$$

  - Estimate of error variance

$$\hat{\sigma}^2 = \frac{\epsilon' \epsilon}{n - p - 1}$$

  - Residuals are used for model diagnostics

Multiple Linear Regression Models

- Statistical inference regarding multiple linear regression models are based on t-test, F-test, and chi-square test

$$
\begin{aligned}
H_{01} &: \beta_j = 0 \ (j = 1, \ldots, p) \\
H_{02} &: \beta_1 = \cdots = \beta_p = 0 \\
H_{03} &: \beta_1 = \cdots = \beta_q = 0 \ (q < p) \\
H_{04} &: \beta_1 = \cdots = \beta_q \ (q < p)
\end{aligned}
$$

Subsection 2

0.2 An example with data on inheritance of height

Inheritance of height

- During 1893–1898 in the UK, K. Pearson (a famous statistician) organized the collection of heights of 1375 mothers aged 65 or less and one of their adult daughters aged 18 or more (Pearson and Lee 1903)
  - Mother height $(x) \rightarrow$ predictor
  - Daughter height $(y) \rightarrow$ response

Does taller mother tend to have taller daughter?

- Assumed model "Daughter height on mother height"

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Inheritance of height

- The data `heights`

```
library(alr4)
data("Heights")
```

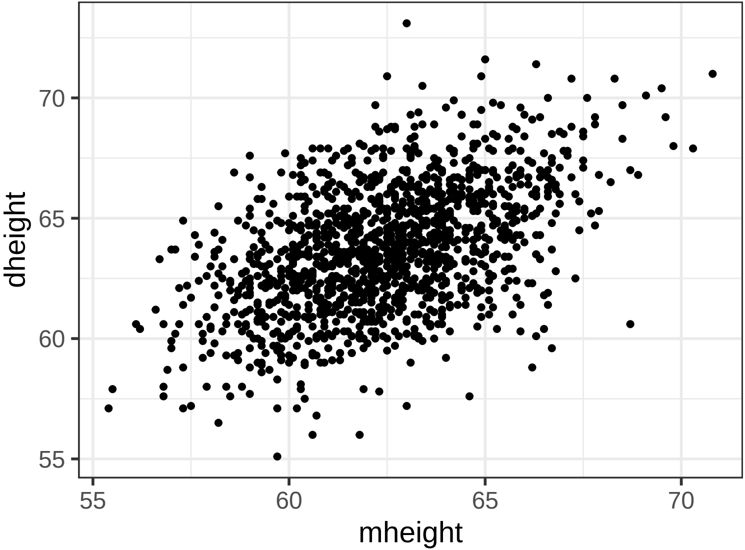
- Transform `Heights` from `data.frame()` to `tibble()`

```
heights <- as_tibble(Heights)
```

Scatterplot of mother height and daughter height

```
ggplot(heights) +
  geom_point(aes(mheight, dheight), size = .75) +
  theme_bw()
```

Scatterplot of mother height and daughter height



Subsection 3

0.3 Model fit

`lm()`

- `lm()` is the most popular R function to fit linear model (with continuous response)
  - A typical syntax of `lm()` function `lm(formula, data)`, which returns a `list`
- For example, codes for fitting the model "Daughter height on mother height"

```
mod_h <- lm(formula = dheight ~ mheight,
            data = heights)
```

lm()

- Elements of `lm()` output object contain useful objects related to the corresponding fit of linear model

```
names(mod_h)
```

```
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.resid
 [9] "xlevels"       "call"          "terms"         "model"
```

```
mod_h$coefficients
```

```
(Intercept)     mheight
  29.917437    0.541747
```

lm()

- Elements of `lm()` output object contain useful objects related to the corresponding fit of linear model

```
names(summary(mod_h))
```

```
 [1] "call"          "terms"         "residuals"     "coeffici
 [5] "aliased"       "sigma"         "df"            "r.square
 [9] "adj.r.squared" "fstatistic"    "cov.unscaled"
```

```
summary(mod_h)$coefficients
```

```
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 29.917437 1.62246940 18.43945 5.211879e-68
mheight      0.541747 0.02596069 20.86797 3.216915e-84
```

`lm()`

- Output of `lm()` object can also be used as an argument of some useful functions, such as
    - `coefficients()` returns the estimates of regression parameters
    - `residuals()` returns associated residuals (there are different types of residuals, use `type` argument to specify this)
    - `fitted()` returns fitted values corresponding to the predictor values of the data
    - `anova()` returns ANOVA table
    - `summary()` returns objects related to linear model fits, some of them are not included in the `lm()` object
    - `confint()` returns confidence intervals of the regression parameters

Subsection 4

0.4 `broom` package

broom

- Most of the built-in R objects related to model fits (e.g. `lm()`, `t.test()`, etc.) require tidy data as input, but its outputs are messy (not tidy), which cannot be used as input for the methods of `tidyverse`
    - e.g. `lm()` returns a list, not a data frame
- `broom` package has functions that transform messy data into tidy data, which are used as inputs of different `tidyverse` functions, such as `ggplot()`, `kable()`, etc.

`broom`

- `broom` has three functions that takes model fit object as an argument and returns a `tibble` (tidy data)
  - `glance()` returns a signle row summary of the model fit, which contains estimates of coefficient of determination, error variance, etc.
  - `tidy()` returns different values corresponding to each parameter, such as estimates, t-stat, p-value, etc.
  - `augment()` returns fitted information corresponding to each observations, e.g. residuals, fitted values, SE of fitted values, etc.

glance()

Single row summary of the model fit

```
glance(mod_h)
```

```
# A tibble: 1 x 12
  r.squared adj.r.squared sigma statistic  p.value    df logLik   AI
      <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>  <dbl> <dbl
1     0.241         0.240  2.27      435. 3.22e-84     1 -3075. 6156
# i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

tidy()

Summary of the parameter estimates

```r
tidy(mod_h)
```

```
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)   29.9      1.62        18.4 5.21e-68
2 mheight        0.542    0.0260      20.9 3.22e-84
```

augment()

Observation-wise values of model fit

```
augment(mod_h) %>%
  slice(1:6)
```

```
# A tibble: 6 x 8
  dheight mheight .fitted .resid      .hat .sigma .cooksd .std.resid
    <dbl>   <dbl>   <dbl>  <dbl>     <dbl>  <dbl>   <dbl>      <dbl>
1    55.1    59.7    62.3  -7.16 0.00172     2.26 0.00862     -3.16
2    56.5    58.2    61.4  -4.95 0.00310     2.26 0.00743     -2.19
3    56      60.6    62.7  -6.75 0.00118     2.26 0.00523     -2.98
4    56.8    60.7    62.8  -6.00 0.00113     2.26 0.00397     -2.65
5    56      61.8    63.4  -7.40 0.000783    2.26 0.00418     -3.27
6    57.9    55.5    60.0  -2.08 0.00707     2.27 0.00303     -0.923
```

Subsection 5

0.5 Model diagnostics

Model diagnostics

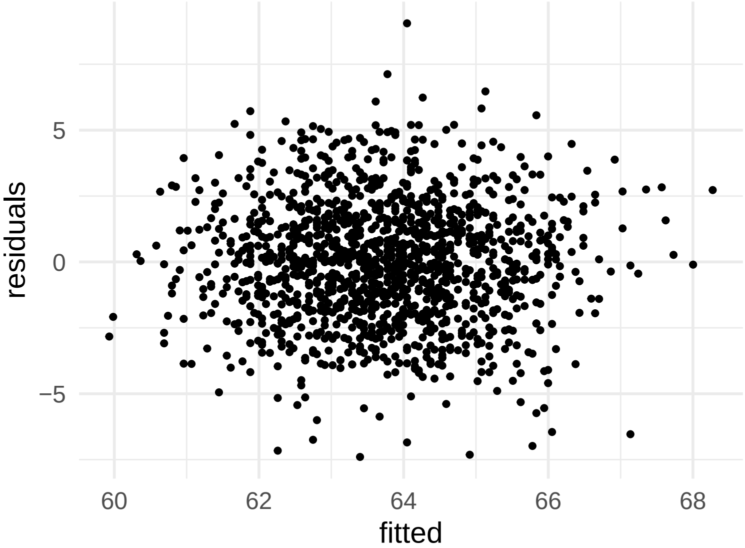
- Residual vs fitted (Independence of errors, Constant variance)
- Residual vs predictor (Linearity, Zero mean)
- Q-Q normal plot of residuals (Normality of error)

Residual plots

Scatterplot of fitted values and residuals

```
ggplot(augment(mod_h)) +
  geom_point(aes(.fitted, .resid), size = .75) +
  labs(x = "fitted", y = "residuals") +
  theme_minimal()
```

Residual plots

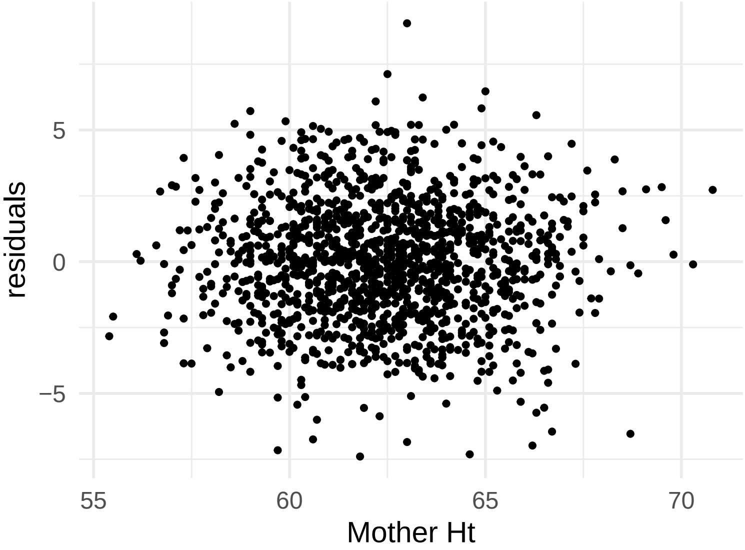


Residual plots

Scatterplot of predictor and residuals

```
ggplot(augment(mod_h)) +
  geom_point(aes(mheight, .resid), size = .75) +
  labs(x = "Mother Ht", y = "residuals") +
  theme_minimal()
```

Residual plots

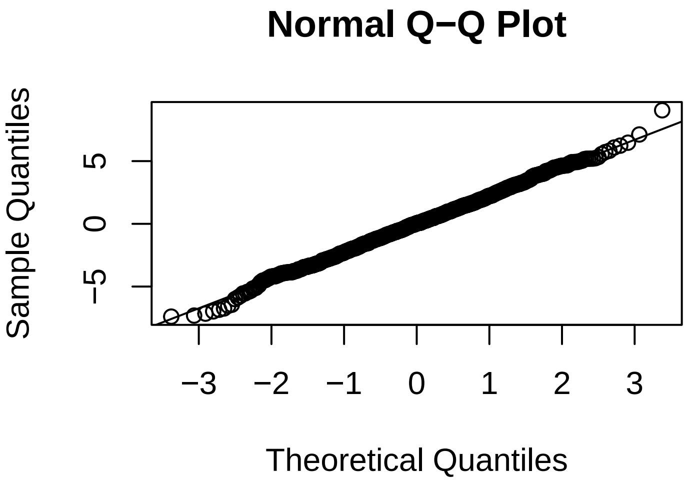


Q-Q normal plot

Using base R functions

```
qqnorm(residuals(mod_h))
qqline(residuals(mod_h))
```

Q-Q normal plot

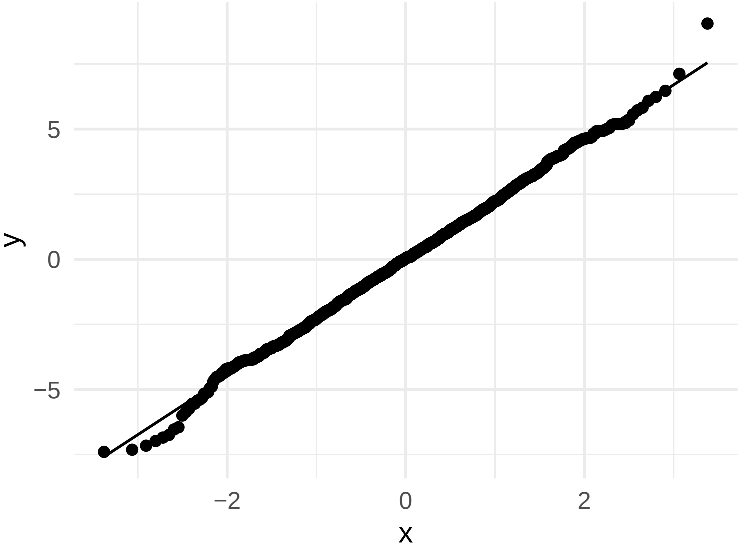


Q-Q normal plot

- For Q-Q plot, `ggplot2` functions `stat_qq()` and `stat_qq_line()` can be used

```
ggplot(augment(mod_h), aes(sample = .resid)) +
  stat_qq() +
  stat_qq_line() +
  theme_minimal()
```

Q-Q normal plot



Summary

- `lm()` function is used to fit the models
- Functions of `broom` package can be used to obtain tidy data from the output of `lm()` function, which is usually messy
- Output of functions of `broom` package can be used as arguments of `ggplot` and other `tidyverse` functions

Exercise

- Using the FEV data (Download the FEV data), fit the following models and perform the model diagnostics
  1. `fev` on `Age`
  2. `fev` on `Age` and `Hgt`

Subsection 6

0.6 Regression models with categorical predictors

Regression models with categorical predictors

- In general, for a predictor with two levels, define a dummy or binary variable that takes the values 1 and 0 corresponding to the two levels of the original variable
  - For example, if the variable X has the levels "male" and "female", we can define a dummy variable D such that

$$D = \begin{cases} 1 & \text{if } X = \text{male} \\ 0 & \text{if } X = \text{female} \end{cases}$$

  - The level "female" is considered as the "reference" category in this case
  - Dummy variable can also be defined with "male" as the reference category

Regression models with categorical predictors

- The model "Y on X" can be expressed in terms of "D" as

$$E(Y \mid D) = \beta_0 + \beta_1 D$$

  - $\beta_0 = E(Y \mid D = 0)$
  - $\beta_1 = E(Y \mid D = 1) - E(Y \mid D = 0)$

- Interpretations of regression parameters depend on the reference category considered

Regression models with categorical predictors

- For a categorical variable with more that two categories, more than one dummy variables needed to be defined
- Let X be a categorical variable with three categories, say "poor", "middle", and "rich"
  - To consider X as a predictor, two dummy variables need to be defined

$$D_1 = \begin{cases} 1 & \text{if } X = \text{poor} \\ 0 & \text{otherwise} \end{cases} \quad D_2 = \begin{cases} 1 & \text{if } X = \text{middle} \\ 0 & \text{otherwise} \end{cases}$$

  - In this case, the category "rich" is considered as the reference category and dummy variables can be defined with other category as the reference

Regression models with categorical predictors

- The regression model "Y on X", where X has three categories can be defined as

$$E(Y \mid D_1, D_2) = \beta_0 + \beta_1 D_1 + \beta_2 D_2$$

  - $\beta_0 = E(Y \mid D_1 = 0, D_2 = 0)$
  - $\beta_1 = E(Y \mid D_1 = 1, D_2 = 0) - E(Y \mid D_1 = 0, D_2 = 0)$
  - $\beta_2 = E(Y \mid D_1 = 0, D_2 = 1) - E(Y \mid D_1 = 0, D_2 = 0)$

Subsection 7

0.7 Interaction

Regression models with one categorical and one continuous predictors

- Consider a regression model "Y on X_1 and X_2", where X_1 has two levels ("male" and "female") and X_2 is continuous (say age in years)
- Define a dummay variable for X_1
  - $D_{1M} = I(X_1 = \text{Male})$

Regression models with one categorical and one continuous predictors

Consider the models

$$
\begin{aligned}
&(1) \;\; E(Y \mid D_{1M}, X_2) = \beta_0 + \beta_1 D_{1M} + \beta_2 X_2 \\
&(2) \;\; E(Y \mid D_{1M}, X_2) = \beta_0 + \beta_1 D_{1M} + \beta_2 X_2 + \theta D_{1M} X_2
\end{aligned}
$$

- How would you interpret the parameters in model (1)?
- **How would you interpret θ in model (2)?**

Regression models with one categorical and one continuous predictors

$$(2) \quad E(Y \mid D_{1M}, X_2) = \beta_0 + \beta_1 D_{1M} + \beta_2 X_2 + \theta D_{1M} X_2$$
$$= \begin{cases} \beta_0 + \beta_2 X_2 & \text{for female} \\ (\beta_0 + \beta_1 + (\beta_2 + \theta) X_2 & \text{for male} \end{cases}$$

- β_0 $\rightarrow$ mean response of female of age 0
- β_1 $\rightarrow$ difference of mean response between male and female when both of them at age of 0
- β_2 $\rightarrow$ change of mean response for 1 year change in age *for female*
- θ $\rightarrow$ difference in the change of the mean response between males and females when their age changes by 1 year
  - it represents how the effect of age differs between males and females

Regression model with two categorical variables

- Consider a regression model "Y on X_1 and X_2", where X_1 has two levels ("male" and "female") and X_2 has three levels ("poor", "middle", and "rich")
- Define the dummy variables for the categorical variables X_1 and X_2
  - For X_1, $D_{1M} = I(X_1 = \text{male})$
  - For X_2, $D_{21} = I(X_2 = \text{poor})$ and $D_{22} = I(X_2 = \text{middle})$

Regression model with two categorical variables

Model 1

$$E(Y \mid D_{1M}, D_{21}, D_{22}) = \beta_0 + \beta_1 D_{1M} + \beta_{21} D_{21} + \beta_{22} D_{22}$$

- β_0 $\rightarrow$ mean response of rich female individuals
- β_1 $\rightarrow$ mean difference between male and female when X_2 is fixed
- β_{21} $\rightarrow$ mean difference between poor and rich when X_1 is fixed
- β_{22} $\rightarrow$ mean difference between middle and rich when X_1 is fixed

Regression model with two categorical variables

Model 2

- The "Model 2" contains both main effects and interactions

$$E(Y \mid D_{1M}, D_{21}, D_{22}) = \beta_0 + \beta_1 D_{1M} + \beta_{21} D_{21} + \beta_{22} D_{22} + \theta_1 D_{1M} D_{21} + \theta_2 D_{1M} D_{22}$$

  - $\beta_0 \rightarrow$ mean response of rich female individuals
  - Interpretations of other parameters are complicated!!

Regression model with two categorical variables

- The following table of expected response would help us to define parameters of "Model 2"

gender	Poor	Middle	Rich
Male	$\beta_0 + \beta_1 + \beta_{21} + \theta_1$	$\beta_0 + \beta_1 + \beta_{22} + \theta_2$	$\beta_0 + \beta_1$
Female	$\beta_0 + \beta_{21}$	$\beta_0 + \beta_{22}$	β_0

- Interpret θ's

Regression model with two categorical variables

- Difference of mean response between male and female among the "rich"

$$E(Y \mid \text{Male, Rich}) - E(Y \mid \text{Female, Rich}) = \beta_1$$

- Difference of mean response between male and female among the "middle"

$$E(Y \mid \text{Male, Middle}) - E(Y \mid \text{Female, Middle}) = \beta_1 + \theta_1$$

Regression model with two categorical variables

- Difference in differences (DID)

$$\Big\{E(Y \,|\, \text{Male, Middlw}) - E(Y \,|\, \text{Female, Middle})\Big\} \\ -\Big\{E(Y \,|\, \text{Male, Rich}) - E(Y \,|\, \text{Female, Rich})\Big\} = \theta_1$$

- Interaction term θ_1 measures whether the effect of "gender" is the same at the levels "Rich" and "Middle"
- Similarly, the interaction term θ_2 measures whether the effect of "gender" is the same at the levels "Rich" and "Poor"

Regression model with two categorical variables

- In the presence of significant interactions, the main effects have no interesting interpretations
- Interaction terms should not be in the model if both the corresponding main effects are not significant
- Insignificant interaction terms should not be in the model

Acknowledgements

This lecture is adapted from materials created by Mahbub Latif

References

Pearson, Karl, and Alice Lee. 1903. "On the Laws of Inheritance in Man: I. Inheritance of Physical Characters." *Biometrika* 2 (4): 357–462.