




Live Classroom



Module 4 –Regression Diagnostics &
Multiple Linear Regression

Topics

- ▶ Scatterplot Matrices
- ▶ Multiple Linear Regression
- ▶ Regression Diagnostics
- ▶ Questions

Scatterplot Matrices

- ▶ When you have more than one explanatory variable, you should look at the relationships between each pair of variables
- ▶ Scatterplot matrix allows you to do this in a single figure
- ▶ Use `pairs(data)` function in R
 - ▶ Control contents of panels by specifying `upper.panel`, `lower.panel`, and `diag.panel` (defined ahead of time)

Panel specifications - correlation

```
panel.cor <- function(x, y, digits=2, prefix="", cex.cor, ...) {  
  usr <- par("usr")  
  on.exit(par(usr))  
  par(usr = c(0, 1, 0, 1))  
  r <- abs(cor(x, y, use="complete.obs"))  
  txt <- format(c(r, 0.123456789), digits=digits)[1]  
  txt <- paste(prefix, txt, sep="")  
  if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)  
  text(0.5, 0.5, txt, cex = cex.cor * (1 + r) / 2)  
}
```

Panel specifications - histogram

```
panel.hist <- function(x, ...) {  
  usr <- par("usr")  
  on.exit(par(usr))  
  par(usr = c(usr[1:2], 0, 1.5) )  
  h <- hist(x, plot = FALSE)  
  breaks <- h$breaks  
  nB <- length(breaks)  
  y <- h$counts  
  y <- y/max(y)  
  rect(breaks[-nB], 0, breaks[-1], y, col="white", ...)  
}
```

Panel specifications – linear regression

```
panel.lm <- function (x, y, col = par("col"), bg = NA, pch = par("pch"),  
  cex = 1, col.smooth = "black", ...) {  
  points(x, y, pch = pch, col = col, bg = bg, cex = cex)  
  abline(stats::lm(y ~ x), col = col.smooth, ...)  
}
```

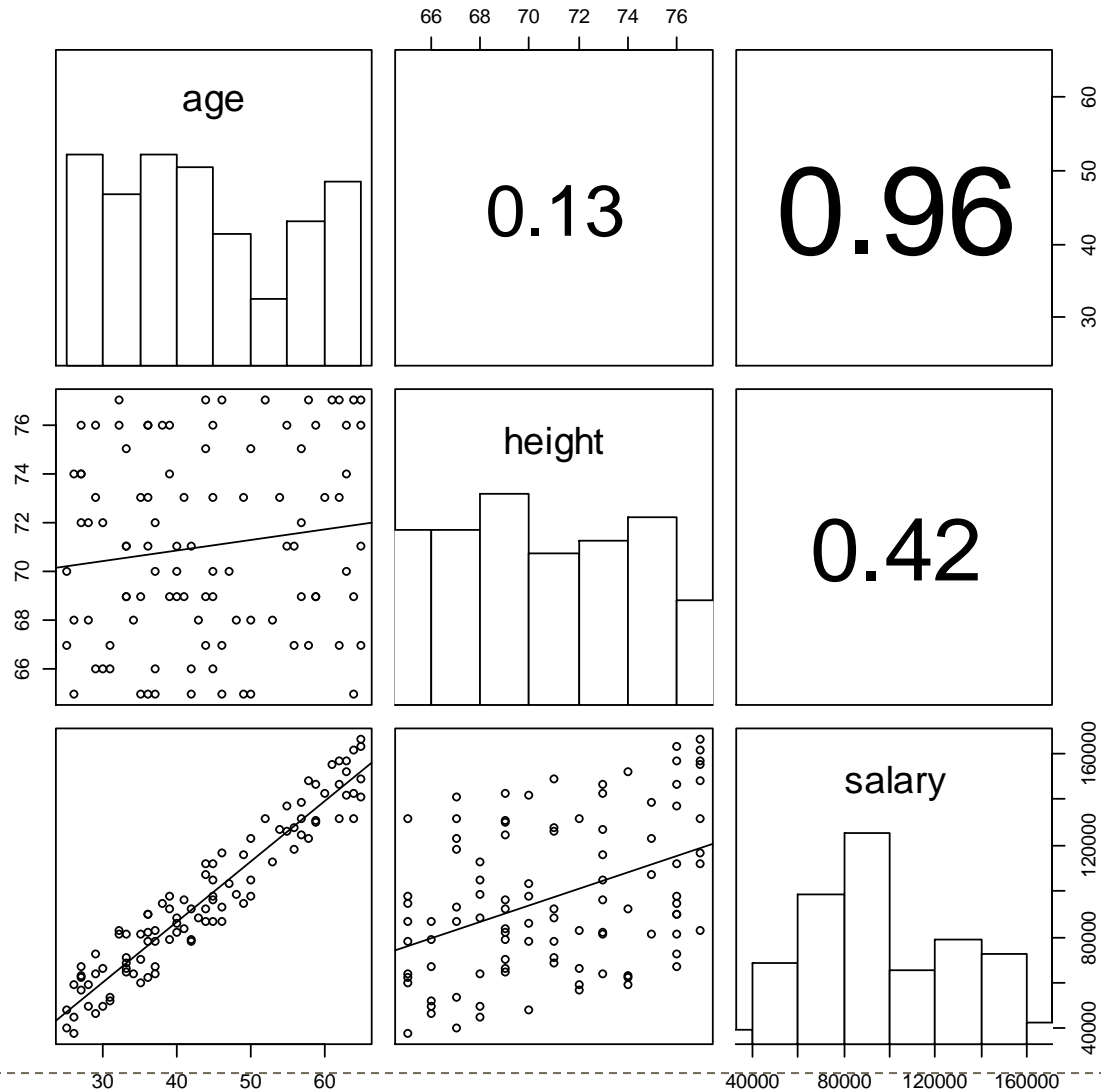
Example - Age, Height and Salary

In the book *Blink* by Malcolm Gladwell, Gladwell states that a study of CEOs of Fortune 500 companies found that these individuals tend to be taller than the average US population. In order to study this phenomenon in more detail and to see if height is associated with increased success in business (as measured by salary), 100 men between the ages of 25 and 65 were polled for their heights (in inches) and annual salaries.

Example - Age, Height and Salary – Scatterplot Matrix

- ▶ `pairs(data, upper.panel=panel.cor, diag.panel=panel.hist, lower.panel=panel.lm)`

Example - Age, Height and Salary – Scatterplot Matrix



Multiple Linear Regression

- ▶ Use `lm()` function to perform regression
 - ▶ `lm(data$response ~ data$explanatory1 + data$explanatory2 + ...)`
- ▶ Use `summary()` function on the `lm()` result to obtain the beta coefficients (t-test), R-squared value, F-statistic
- ▶ ANOVA table to be calculated by hand (if interested, but not necessary!)
- ▶ Use `confint()` on the `lm()` result to obtain confidence intervals for the regression coefficients

Example - Age, Height and Salary – F-test

- ▶ Are age and height when considered together predictors of annual salary?
Perform an F-test at the $\alpha = 0.01$ level to answer this question.
- ▶ (1) Set up the hypotheses and select the alpha level
 - $H_0: \beta_{age} = \beta_{height} = 0$ (age and height are not predictors of annual salary)
 - $H_1: \beta_{age} \neq 0$ and/or $\beta_{height} \neq 0$ (at least one of the slope coefficients is different than 0; age and/or height are predictors/is a predictor of annual salary)
 - $\alpha = 0.01$
- ▶ (2) Select the appropriate test-statistic
$$F = \frac{MS_{Reg}}{MS_{Res}} \quad df = k, n - k - 1$$

Example - Age, Height and Salary – F-test

(3) State the decision rule

Decision Rule: Reject H_0 if $p \leq \alpha$. Otherwise, do not reject H_0

Example - Age, Height and Salary – F-test

(4) Compute the test statistic

summary(m) gives

F – statistic: 1.299e + 05 on 2 and 97 DF, p – value: < 2.2e – 16

(5) Conclusion

Reject H_0 since $p \leq \alpha$. We have significant evidence at the $\alpha = 0.01$ level that age and height when taken together are predictive of annual salary. That is, there is evidence of a linear association between annual salary and age and height (here, $p < 0.001$).

Example - Age, Height and Salary – t-test

- Is height a predictor of annual salary after controlling for age? Perform a t -test at the $\alpha = 0.01$ level and calculate the 99% confidence interval for β_{height} .

(1) Set up the hypotheses and select the alpha level

$H_0: \beta_{height} = 0$ (height is not associated with salary, after controlling for age)

$H_1: \beta_{height} \neq 0$ (height is associated with salary, after controlling for age)

$\alpha = 0.01$

(2) Select the appropriate test-statistic

$$t = \frac{\hat{\beta}_{height}}{SE_{\hat{\beta}_{height}}} \text{ with df} = n - k - 1$$

Example - Age, Height and Salary – t-test

(3) State the decision rule

Decision Rule: Reject H_0 if $p \leq \alpha$. Otherwise, do not reject H_0

Example - Age, Height and Salary – t-test

(4) Compute the test statistic

summary(m)

$$t = 147.5, \quad df = 97, \quad p\text{-value} < 2e - 16$$

(5) Conclusion

Reject H_0 since $p \leq \alpha$. We have significant evidence at the $\alpha = 0.01$ level that $\beta_{height} \neq 0$ after controlling for age. That is, height is predictive of annual salary after adjusting for age ($p < 0.0001$). We are 99% confident that the true value of β_{height} is between \$2462 and \$2552, after controlling for age. That is, for every additional inch of height, we are 99% confident that annual salary is generally between \$2462 and \$2552 higher.

Regression Diagnostics

- ▶ Check the assumptions
 - ▶ Linearity
 - ▶ Independence
 - ▶ Constant variance
 - ▶ Normally distributed residuals
- ▶ Residual Plots (to assess linearity and constant variance)
 - ▶ `plot([variable for x-axis],resid(m))`
 - ▶ Check each explanatory variable, and the fitted values (`fitted(m)`)
- ▶ Histograms (to check the distribution of the residuals)
 - ▶ `hist(resid(m))`
 - ▶ Regression assumptions least sensitive to departures from this assumption

Example - Salary versus Age and Height – Regression Diagnostics

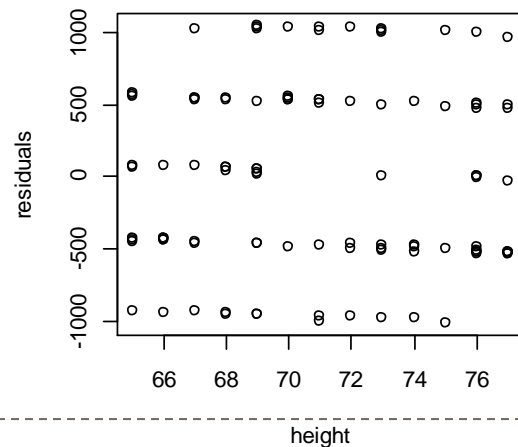
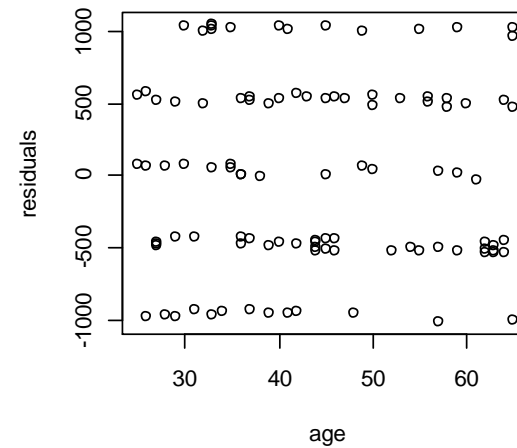
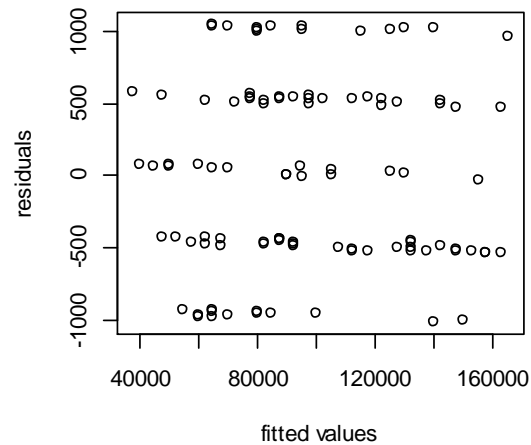
```
plot(fitted(m),resid(m), axes=TRUE, frame.plot=TRUE, xlab = "fitted values",  
ylab="residuals")
```

```
plot(data$age,resid(m), axes=TRUE, frame.plot=TRUE, xlab = "age",  
ylab="residuals")
```

```
plot(data$height,resid(m), axes=TRUE, frame.plot=TRUE, xlab = "height",  
ylab="residuals")
```

```
hist(resid(m))
```

Example – Salary versus Age and Height – regression diagnostics



Histogram of resid(m)

