

# Notes: MS 204 Chapter 1/5

## Overview

- Simple linear regression
- Linear model assumptions
- R-squared
- Categorical predictors

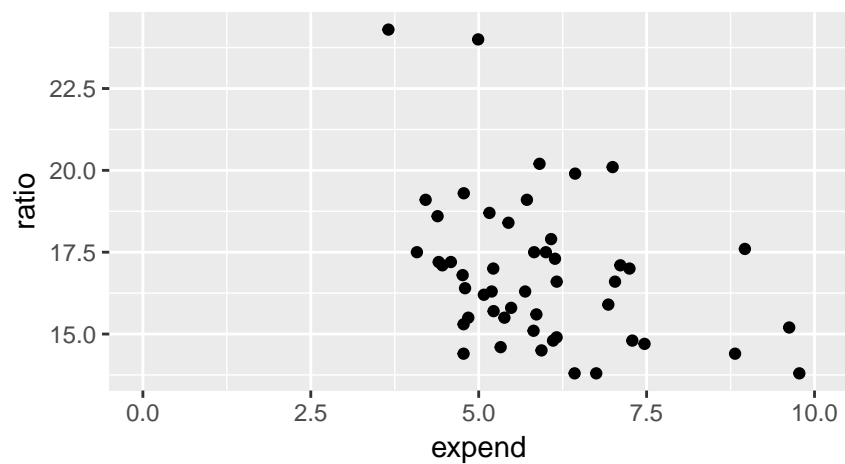
## Simple linear regression

Ex:  $X$  = salary,  $Y$  = SAT

```
library(tidyverse); library(mosaic)
SAT %>% summarise(cor.SAT = cor(expend, ratio))
```

```
##      cor.SAT
## 1 -0.3710254
```

```
qplot(x = expend, y = ratio, data = SAT) + xlim(c(0, 10))
```



- Reminders: slope, intercept, estimated line

```
fit <- lm(ratio ~ expend, data = SAT)
msummary(fit)

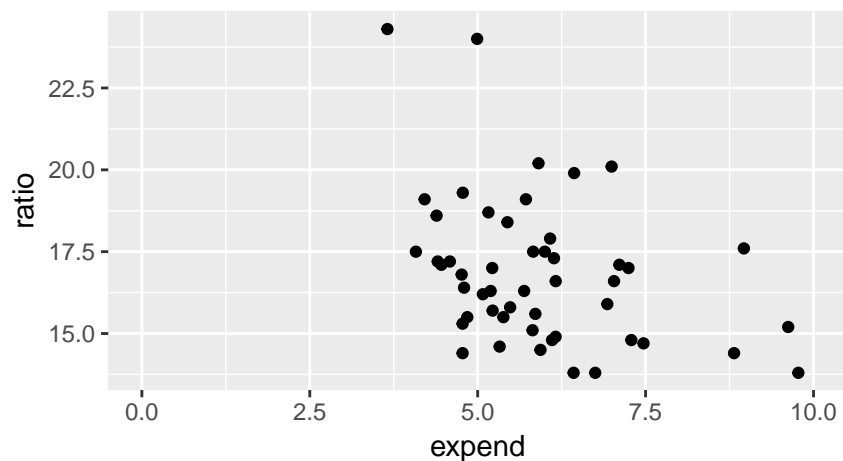
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  20.5016      1.3502  15.184  < 2e-16 ***
## expend      -0.6170      0.2229  -2.768  0.00799 **
##
## Residual standard error: 2.126 on 48 degrees of freedom
## Multiple R-squared:  0.1377, Adjusted R-squared:  0.1197
## F-statistic: 7.662 on 1 and 48 DF,  p-value: 0.007987
```

- Residuals and line fitting

```
head(SAT, 2)

##      state expend ratio salary frac verbal math  sat
## 1 Alabama  4.405  17.2 31.144    8    491  538 1029
## 2 Alaska   8.963  17.6 47.951   47    445  489  934

qplot(x = expend, y = ratio, data = SAT) + xlim(c(0, 10))
```



```
fit <- lm(ratio ~ expend, data = SAT)
resid.fit <- resid(fit)
```

```
resid.fit[1:2]
```

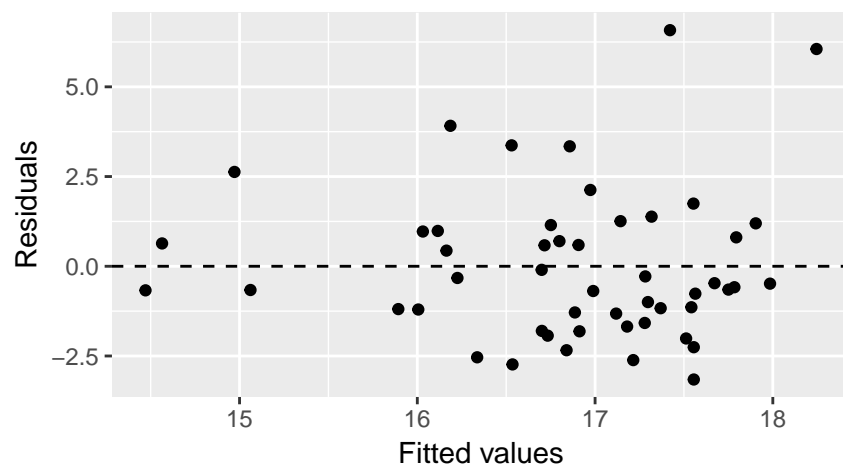
```
##          1          2  
## -0.5836862  2.6286782
```

## Conditions for least squares regression

1. Linearity
2. Nearly normal residuals
3. Constant variability

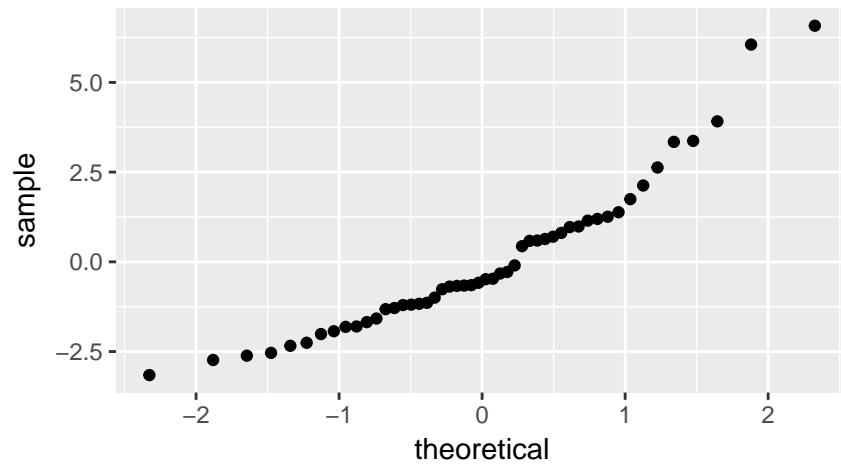
### Linearity

```
qplot(x = .fitted, y = .resid, data = fit) +  
  geom_hline(yintercept = 0, linetype = "dashed") +  
  xlab("Fitted values") +  
  ylab("Residuals")
```



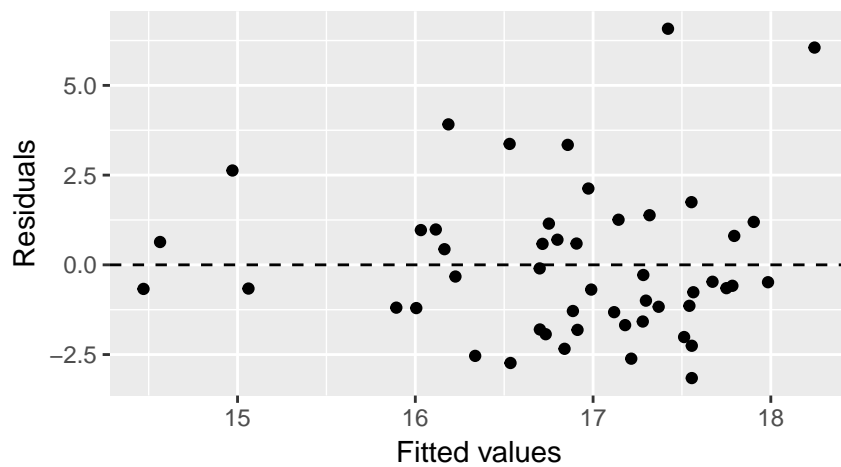
### Nearly normal residuals

```
qplot(sample = .resid, data = fit, geom = "qq")
```



### Constant variability

```
qplot(x = .fitted, y = .resid, data = fit) +
  geom_hline(yintercept = 0, linetype = "dashed") +
  xlab("Fitted values") +
  ylab("Residuals")
```



**R-squared**

**Coefficient of determination:**

Example interpretation: expenditure versus ratio (correlation -0.37)

Incorrect interpretations:

## Categorical predictors

Flights data set – departure delay as a function of origin

```
library(nycflights13)
tally(~origin, data = flights)
```

```
## origin
##      EWR      JFK      LGA
## 120835 111279 104662
```

```
fit1 <- lm(dep_delay ~ origin, data = flights)
msummary(fit1)
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  15.1080      0.1171  129.00  <2e-16 ***
## originJFK    -2.9958      0.1687  -17.76  <2e-16 ***
## originLGA    -4.7611      0.1721  -27.67  <2e-16 ***
##
## Residual standard error: 40.16 on 328518 degrees of freedom
## (8255 observations deleted due to missingness)
## Multiple R-squared:  0.002411, Adjusted R-squared:  0.002405
## F-statistic: 396.9 on 2 and 328518 DF, p-value: < 2.2e-16
```

Interpretations: