More Examples

Max Turgeon

SCI 2000–Introduction to Data Science

Explore data i

- We will use data on Forced Expiratory Volume (FEV) in children age 3 to 19 from East Boston recorded during the 1970s.
    - Can be downloaded from `https://hbiostat.org/data/`, but I also added a copy on UM Learn.
- The dataset contains information on age, height, sex, and smoking status.
- **Outcome**: FEV

Explore data ii

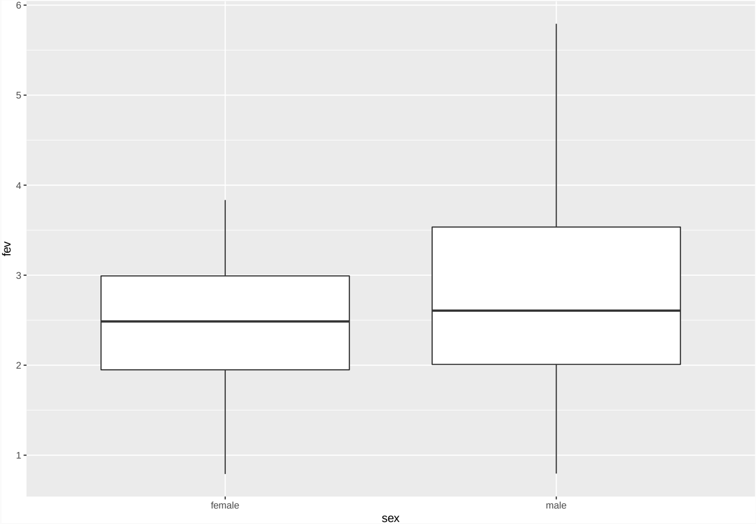
```
library(tidyverse)
# Import dataset into R
data_fev <- read_csv("FEV.csv")
glimpse(data_fev, width = 50) # So it fits the slide
```

```
## Rows: 654
## Columns: 6
## $ id <dbl> 301, 451, 501, 642, 901, 1701, 17~
## $ age <dbl> 9, 8, 7, 9, 9, 8, 6, 6, 8, 9, 6, ~
## $ fev <dbl> 1.708, 1.724, 1.720, 1.558, 1.895~
## $ height <dbl> 57.0, 67.5, 54.5, 53.0, 57.0,
```

```
61.~
## $ sex <chr> ”female”, ”female”, ”female”, ”ma~
## $ smoke <chr> ”non-current smoker”,
”non-curren~
```

```r
# Explore data
ggplot(data_fev, aes(x = sex, y = fev)) +
  geom_boxplot()
```

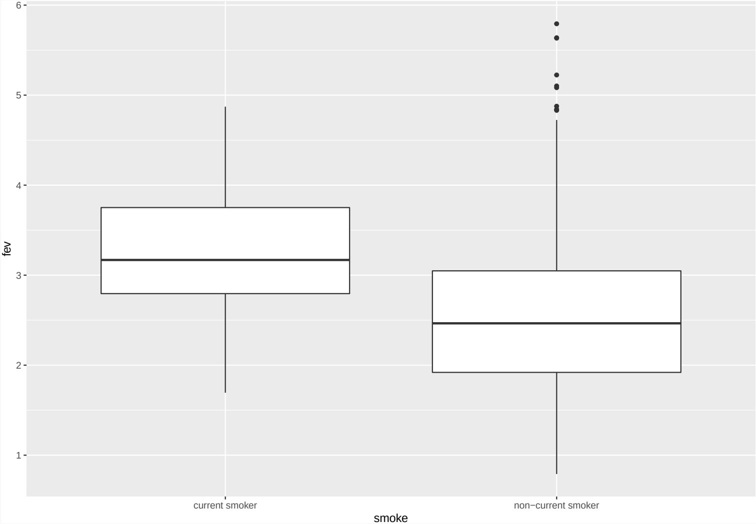
Explore data iv



Explore data v

```
ggplot(data_fev, aes(x = smoke, y = fev)) +
  geom_boxplot()
```

Explore data vi



Explore data vii

```r
# Smokers have higher FEV??
fit <- lm(fev ~ smoke, data = data_fev)
coef(fit)
```

```
## (Intercept) smokenon-current smoker
## 3.2768615 -0.7107189
```

```r
confint(fit)
```

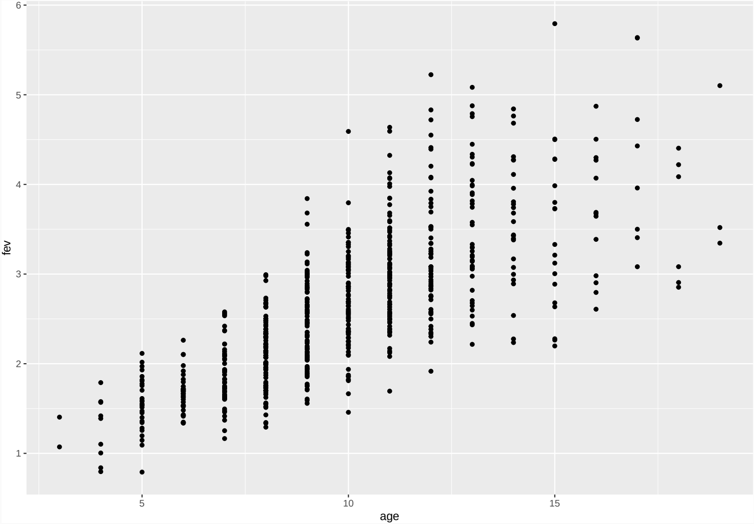
```
##                                  2.5 %     97.5 %
## (Intercept)                  3.0719861  3.4817370
## smokenon-current smoker -0.9266033 -0.4948346
```

Explore data viii

- Non-smokers have, on average, an FEV measure that is 0.7 lower than smokers... What can be going on?

```
# Look at FEV vs age and height
ggplot(data_fev, aes(x = age, y = fev)) +
  geom_point()
```

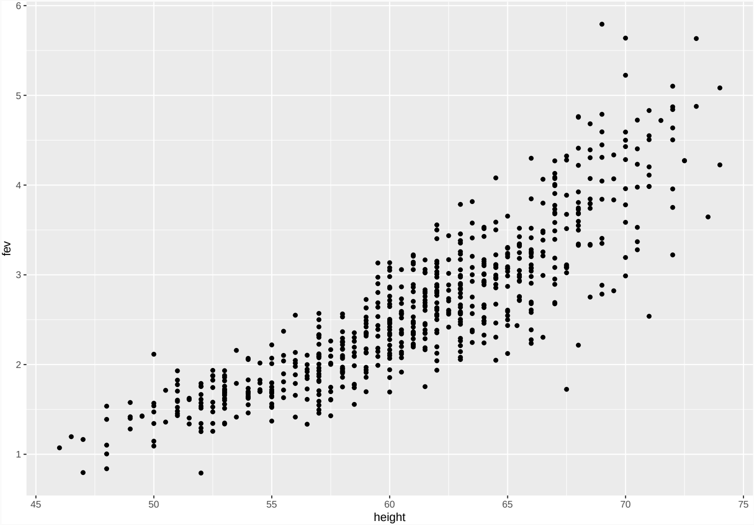
Explore data ix



Explore data x

```
ggplot(data_fev, aes(x = height, y = fev)) +
  geom_point()
```

Explore data xi



Explore data xii

- The association between FEV and smoking status is **spurious**: it looks like it is driven by the fact that:
  - Older children are taller, have larger lungs, and therefore higher FEV.
  - Older children are more likely to be smokers.
- We also say that age and height are **confounders** for the association between FEV and smoking status.

Fit a linear model i

- The idea is that we are comparing older and younger children together, thus creating this spurious association.
    - What if we only compared children of the same age?
- Linear regression actually allows us to **adjust** for the effect of age and height on FEV.

```
# Fit linear model
model <- lm(fev ~ smoke + sex + age + height,
            data = data_fev)
```

Fit a linear model ii

```
coef(model)
```

```
## (Intercept) smokenon-current smoker sexmale
## -4.54422029 0.08724639 0.15710293
## age height
## 0.06550932 0.10419943
```

Fit a linear model iii

```r
confint(model)
```

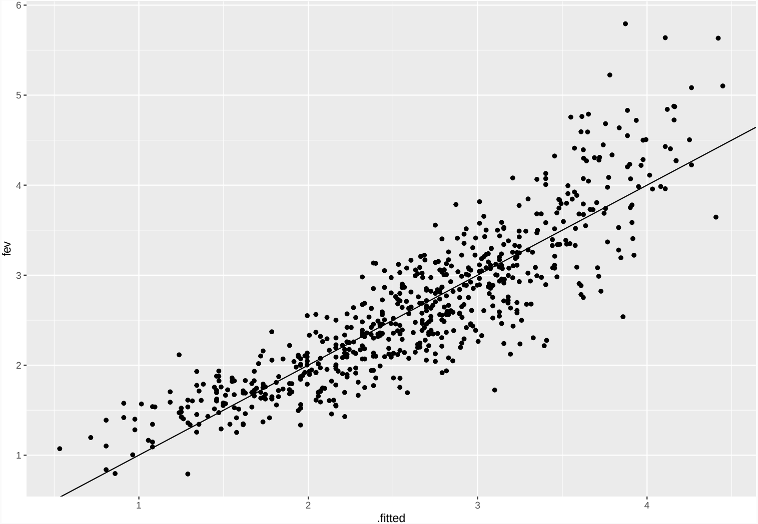
```
##                                2.5 %      97.5 %
## (Intercept)               -4.99987259 -4.08856799
## smokenon-current smoker -0.02910535  0.20359813
## sexmale                    0.09189669  0.22230917
## age                        0.04687736  0.08414129
## height                     0.09485705  0.11354180
```

- Non-smokers have, on average, an FEV measure that is 0.08 *higher* than smokers, when adjusting for age, height and sex.
    - And it's no longer significant (0 is in the confidence interval)

Residual analysis i

```r
library(broom)
# Plot outcome vs fitted values
augment(model) %>%
  ggplot(aes(x = .fitted, y = fev)) +
  geom_point() +
  geom_abline(intercept = 0,
              slope = 1)
```

Residual analysis ii



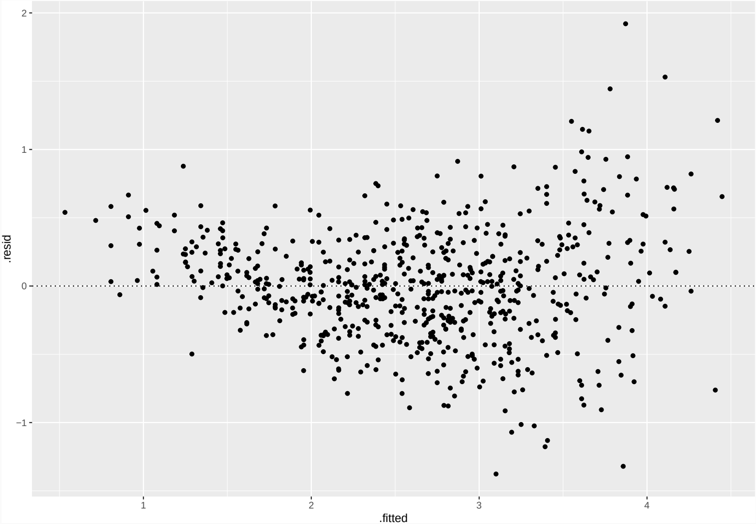
Residual analysis iii

```r
# Can also colour points according to smoking status
augment(model) %>%
  ggplot(aes(x = .fitted, y = fev)) +
  geom_point(aes(colour = smoke)) +
  geom_abline(intercept = 0,
              slope = 1)
```

Residual analysis iv

```r
# Plot residuals vs fitted values
augment(model) %>%
  ggplot(aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0,
             linetype = "dotted")
```

Residual analysis v



Residual analysis vi

- We found evidence that additivity/linearity is not met.
  - Outcome vs fitted plot.
  - Given our data visualizations, it is likely that relationship between FEV and height is nonlinear.
- We found evidence of unequal variance.
  - Residual vs fitted values: higher variance with larger fitted values.
- How can we use residual analysis to decide how we could improve the model?

Fit a second linear model i

```r
# We will add a quadratic term for height
# The function I() protects height^2
# Try removing it from the code below and
# see how it's different
model2 <- lm(fev ~ smoke + sex + age +
               height + I(height^2),
             data = data_fev)
```

Fit a second linear model ii

```r
coef(model2)
```

```
## (Intercept) smokenon-current smoker sexmale
## 6.761367559 0.133211169 0.094535151
## age height I(height^2)
## 0.069464619 -0.274234148 0.003125062
```

Fit a second linear model iii

```r
confint(model2)
```

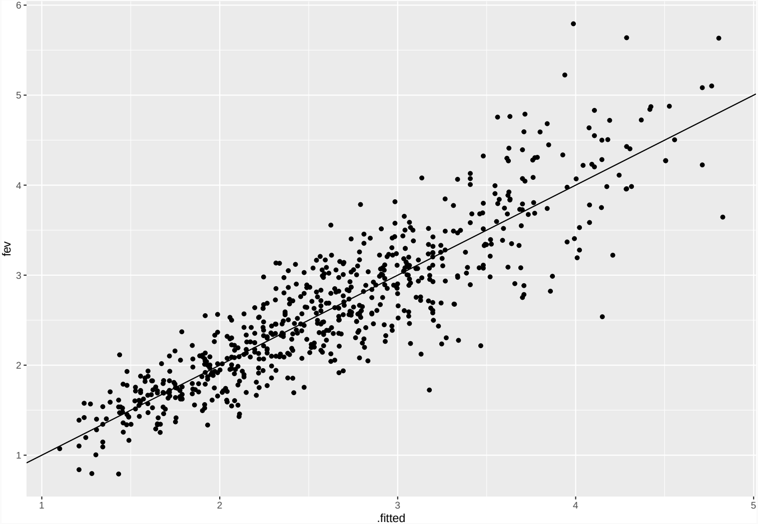
```
##                                  2.5 %       97.5 %
## (Intercept)                 3.826331931  9.696403187
## smokenon-current smoker     0.021072270  0.245350068
## sexmale                     0.030007679  0.159062623
## age                         0.051578126  0.087351111
## height                     -0.371797141 -0.176671155
## I(height^2)                 0.002322798  0.003927326
```

Fit a second linear model iv

- Non-smokers have, on average, an FEV measure that is 0.13 *higher* than smokers, when adjusting for age, height and sex.
    - And now it's back to being significant

```r
augment(model2) %>%
  ggplot(aes(x = .fitted, y = fev)) +
  geom_point() +
  geom_abline(intercept = 0,
              slope = 1)
```

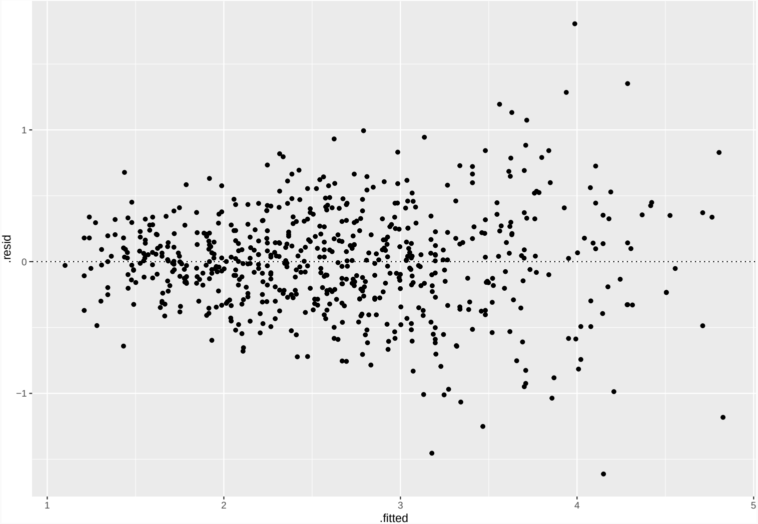
Residual analysis redux ii



Residual analysis redux iii

```r
augment(model2) %>%
  ggplot(aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0,
             linetype = "dotted")
```

Residual analysis redux iv



Residual analysis redux v

We still have evidence of unequal variance, but at least additivity/linearity now seem to hold!

```
# Let's use robust standard errors
library(lmtest)
library(sandwich)
coefci(model2, vcov. = vcovHC(model2))
```

Residual analysis redux vi

```
##                                  2.5 %       97.5 %
## (Intercept)                  3.659722511  9.863012607
## smokenon-current smoker -0.018764668  0.285187006
## sexmale                      0.031332609  0.157737694
## age                          0.049814635  0.089114603
## height                      -0.381834166 -0.166634130
## I(height^2)                  0.002210479  0.004039645
```

Summary

- We still have the same interpretation for our regression coefficient:
  - Non-smokers have, on average, an FEV measure that is 0.13 *higher* than smokers, when adjusting for age, height and sex.
- With the robust standard errors, the confidence interval is wider, and so the association between FEV and smoking status (accounting for age, height and sex) is no longer significant.
- Because we are now confident our assumptions hold, the right conclusion from our analysis is the one based on the last model.
  - Quadratic term for height
  - Robust standard errors