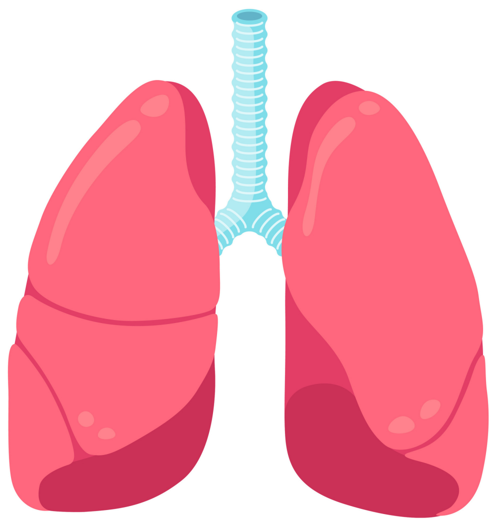
STAT 331 – Lecture 8 (Data analysis)

Here we study a data set that studies the lung function in children and teens. The data is taken from Kahn, Michael (2005). "An Exhalent Problem for Teaching Statistics", The Journal of Statistical Education, 13(2).



Consider a data set from $n = 655$ children between 3 and 19 years old. The variables in the data set include Forced Exhalation Volume (`FEV`) (the response variable), which is a measure of the amount of air an individual can forcibly exhale from their lungs, and `age` (the explanatory variable) in years. Other explanatory variables collected also include `ht` (height in inches), `sex` (1 = male, 0 = female) and `smoke` (1 = yes, 0 = no).

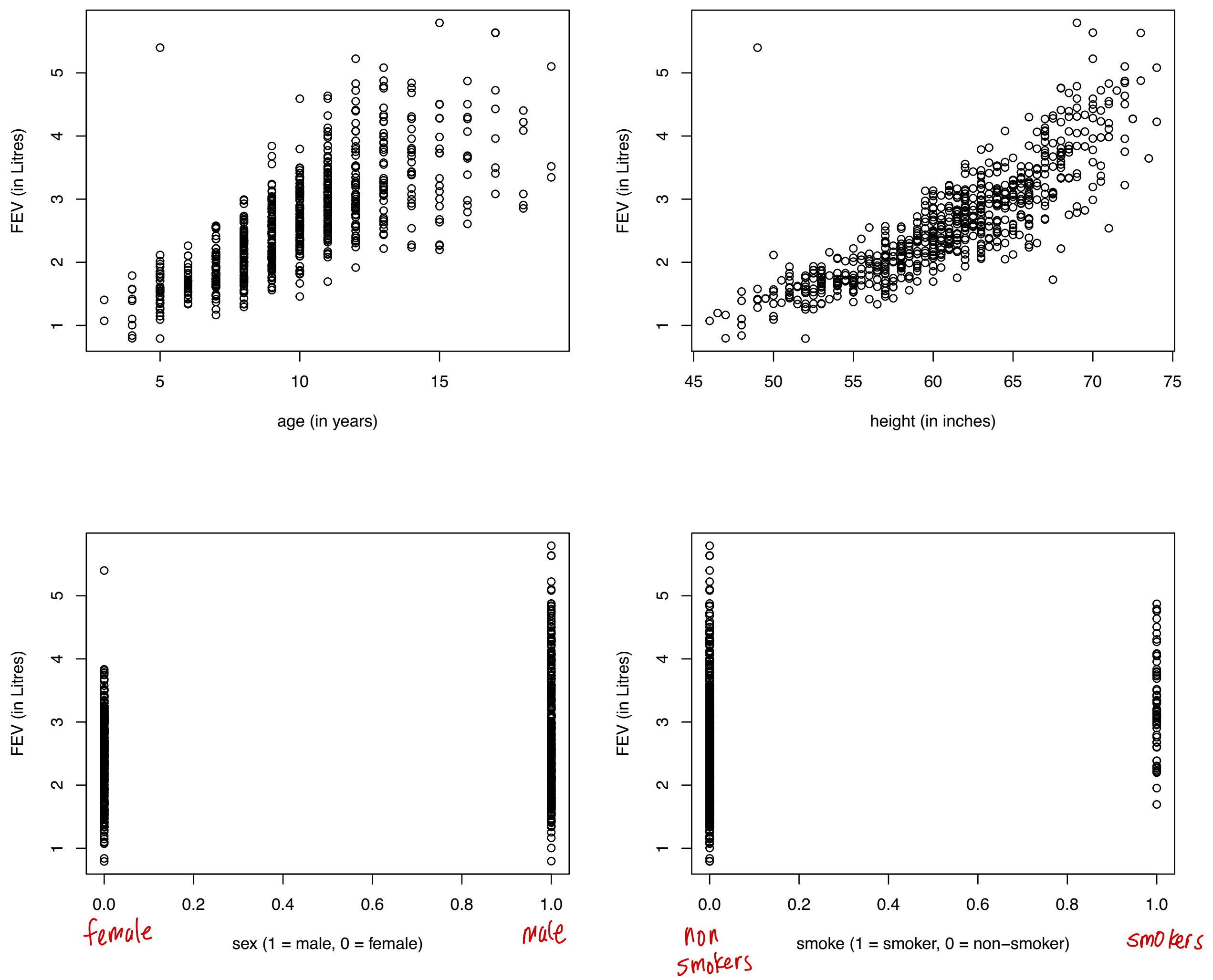
1 Read and view data from csv file

```
> lungdat = read.csv("lung_dat.csv", header=T)
> head(lungdat) ## View only the first 6 rows of data

  age   FEV   ht sex smoke
1   9 1.708 57.0   0     0
2   8 1.724 67.5   0     0
3   7 1.720 54.5   0     0
4   9 1.558 53.0   1     0
5   9 1.895 57.0   1     0
6   8 2.336 61.0   0     0
```

2 Scatter Plots

```r
par(mfrow = c(2,2))
plot(lungdat$age, lungdat$FEV, ylab="FEV (in Litres)",
     xlab = "age (in years)")
plot(lungdat$ht, lungdat$FEV, ylab="FEV (in Litres)",
     xlab = "height (in inches)")
plot(lungdat$sex, lungdat$FEV, ylab="FEV (in Litres)",
     xlab = "sex (1 = male, 0 = female)")
plot(lungdat$smoke, lungdat$FEV, ylab="FEV (in Litres)",
     xlab = "smoke (1 = smoker, 0 = non-smoker)")
```



3 Fit a multiple linear regression model

First define our model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_3 x_{i4} + \epsilon_i, \quad i = 1, \ldots, 655, \quad \epsilon_i \sim N(0, \sigma^2) \text{ iid}$$

where x_1 denotes age, x_2 denotes height, x_3 denotes sex and x_4 denotes smoking status.

```
> myfit = lm(FEV ~ age + ht + sex + smoke, data = lungdat)
> summary(myfit) ## Shows a summary of our fitted model

Call:
lm(formula = FEV ~ age + ht + sex + smoke, data = lungdat)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3746 -0.2560 -0.0085  0.2408  4.3829

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
β0 (Intercept) -4.30470    0.24096 -17.865  < 2e-16 ***
β1 age          0.06492    0.01028   6.315 5.01e-10 ***
β2 ht           0.10198    0.00515  19.802  < 2e-16 ***
β3 sex          0.14769    0.03597   4.106 4.54e-05 ***
β4 smoke       -0.08171    0.06420  -1.273    0.204
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                                   σ̂             n-p-1 = 655-4-1
Residual standard error: 0.4466 on 650 degrees of freedom
Multiple R-squared:  0.7399,    Adjusted R-squared:  0.7383
F-statistic: 462.3 on 4 and 650 DF,  p-value: < 2.2e-16
```

Alternatively, we can estimate parameters manually using expression derived in Lecture 6: $\hat{\beta} = (X^T X)^{-1} X^T \vec{y}$

```
## First define X matrix
> X = cbind(rep(1, nrow(lungdat)), lungdat$age, lungdat$ht,
      lungdat$sex, lungdat$smoke)
> X[1:3,]
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    9 57.0    0    0
[2,]    1    8 67.5    0    0
[3,]    1    7 54.5    0    0

## Define y column vector
> y = matrix(lungdat$FEV, ncol=1)

## beta estimates
> beta_hat = solve(t(X) %*% X) %*% t(X) %*% y
                   (X^T X)^-1          X^T y
> beta_hat
            [,1]
[1,] -4.30470340
[2,]  0.06492341
[3,]  0.10198392
[4,]  0.14768938
[5,] -0.08170660
```

$SE(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 (X^T X)^{-1}_{jj}}$ $\qquad SS(res) = \sum_{i=1}^{n} e_i^2 = \vec{e}^T \vec{e}$ $\qquad \vec{e} = \vec{y} - \underbrace{X\hat{\vec{\beta}}}_{\text{fitted values}}$

$\hat{\sigma}^2 = SS(res)/n-p-1$

We can also manually estimate the standard errors of each $\hat{\beta}_j$ $(j = 0, \dots, p)$ in $\hat{\vec{\beta}}$ using expressions derived in Lecture 6:

```
## sigma estimate
> fitted_values = X %*% beta_hat ## fitted values
> res = y - fitted_values ## residuals
> ssres = t(res) %*% res ## SS(Res)
> ssres = sum(res^2) ##same as above
> ssres
[1] 129.6679
> n = nrow(lungdat); p = length(beta_hat)-1
> sigma_hat = sqrt(ssres/(n-p-1))
> sigma_hat
[1] 0.446642
```

$\vec{e}$

$\hat{\sigma}^2 (X^T X)^{-1}$ (estimate of $Var(\hat{\vec{\beta}}) = \sigma^2 (X^T X)^{-1}$)

```
## standard error of beta estimates
> var_beta_hat = sigma_hat^2*(solve(t(X) %*% X)); var_beta_hat
              [,1]       [,2]       [,3]       [,4]       [,5]
[1,]  0.0580592  1.461e-03 -1.194e-03  1.325e-03  3.888e-04
[2,]  0.0014613  1.057e-04 -4.115e-05  4.748e-05 -1.958e-04
[3,] -0.0011939 -4.115e-05  2.652e-05 -4.055e-05  1.714e-05
[4,]  0.0013252  4.748e-05 -4.055e-05  1.294e-03  1.899e-04
[5,]  0.0003888 -1.958e-04  1.714e-05  1.899e-04  4.122e-03

> se_beta_hat = sqrt(diag(var_beta_hat)); se_beta_hat
[1] 0.240954805 0.010280803 0.005150149 0.035967626 0.064199532
```

$SE(\hat{\beta}_0) = \sqrt{[\hat{\sigma}^2 (X^T X)^{-1}]_{00}}$ $\quad SE(\hat{\beta}_1)$ $\quad SE(\hat{\beta}_2)$ $\quad SE(\hat{\beta}_3)$ $\quad SE(\hat{\beta}_4)$

$\left(= \sqrt{\hat{\sigma}^2 [(X^T X)^{-1}]_{00}}\right)$

Alternatively, we can extract the standard error from the variance-covariance matrix in R:

$\hat{\sigma}^2 (X^T X)^{-1}$

```
## Covariance matrix of beta_hat
> vcov(myfit)
             (Intercept)        age         ht        sex      smoke
(Intercept)    0.0580592  1.461e-03 -1.194e-03  1.325e-03  3.888e-04
age            0.0014613  1.057e-04 -4.115e-05  4.748e-05 -1.958e-04
ht            -0.0011939 -4.115e-05  2.652e-05 -4.055e-05  1.714e-05
sex            0.0013252  4.748e-05 -4.055e-05  1.294e-03  1.899e-04
smoke          0.0003888 -1.958e-04  1.714e-05  1.899e-04  4.122e-03

## Standard errors of individual betas
> sqrt(diag(vcov(myfit)))
(Intercept)         age          ht         sex       smoke
0.240954805 0.010280803 0.005150149 0.035967626 0.064199532
```

4 Study the association between age and FEV

```
> myfit = lm(FEV ~ age + ht + sex + smoke, data = lungdat)
> summary(myfit) ## Shows a summary of our fitted model

Call:
lm(formula = FEV ~ age + ht + sex + smoke, data = lungdat)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3746 -0.2560 -0.0085  0.2408  4.3829

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.30470    0.24096 -17.865  < 2e-16 ***
age          0.06492    0.01028   6.315 5.01e-10 ***
ht           0.10198    0.00515  19.802  < 2e-16 ***
sex          0.14769    0.03597   4.106 4.54e-05 ***
smoke       -0.08171    0.06420  -1.273    0.204
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4466 on 650 degrees of freedom
Multiple R-squared:  0.7399,    Adjusted R-squared:  0.7383
F-statistic: 462.3 on 4 and 650 DF,  p-value: < 2.2e-16

> crit_val = qt(0.975, n-p-1); crit_val
[1] 1.96362
```

Q: How do we interpret the estimate $\hat{\beta}_1$?

A: $\hat{\beta}_1$ indicates that the average change in FEV for every year increase in age is 0.065 liters, holding height, sex and smoking status constant.

Carry out a t-test:

$$H_0 : \beta_1 = 0 \text{ vs. } H_a : \beta_1 \neq 0$$

$$\mid t \mid = \mid 0.06492/SE(\hat{\beta}_1) \mid = 6.315, \quad \text{where SE}(\hat{\beta}_1) = 0.01028$$

The degree of freedom is $n - p - 1 = 650$, so the critical value $c = t_{0.975,650} = 1.964$. Since $\mid t \mid > c$, we reject the null hypothesis and conclude that there is a strong association between age and FEV, given other covariates such as height, sex and smoking status.

5 Estimate the mean response for a 7 year old boy who is 47 inches and does not smoke. Give a 95% confidence interval.

age=7 sex=1 ht=47 smoke=0

First, we calculate manually:

$\vec{x}_0 = (1, 7, 47, 1, 0)^T$ (order matters!)

$\hat{\mu}_0 = \vec{x}_0^T \hat{\beta}$

```
> x0 = matrix(c(1, 7, 47, 1 , 0), ncol=1);
> mu0_hat = t(x0) %*% beta_hat; mu0_hat
         [,1]
[1,] 1.090694
```

$SE(\hat{\mu}_0) = \hat{\sigma}\sqrt{\vec{x}_0^T (X^T X)^{-1} \vec{x}_0}$

```
> se_mu0 = sigma_hat * sqrt(t(x0) %*% solve(t(X) %*% X) %*% x0)
> crit_val = qt(0.975, n-p-1); crit_val
[1] 1.96362
```

$\hat{\mu}_0 \pm c\,SE(\hat{\mu}_0)$

```
> CI_low = mu0_hat - crit_val*se_mu0
> CI_high = mu0_hat + crit_val*se_mu0
> c(CI_low, CI_high)
[1] 0.9698424 1.2115454
```

Next, we use predict() in R:

```
> predict(object = myfit, newdata = data.frame(age = 7, ht = 47,
          sex = 1, smoke = 0), interval = "confidence", level = 0.95)
       fit       lwr      upr
1 1.090694 0.9698424 1.211545
```

6 Predict the response for a 7 year old boy who is 47 inches and does not smoke. Give a 95% prediction interval.

First, we calculate manually:

$\hat{\sigma}\sqrt{1 + \vec{x}_0^T (X^T X)^{-1} \vec{x}_0}$

```
> x0 = matrix(c(1, 7, 47, 1 , 0), ncol=1);
> y0_hat = t(x0) %*% beta_hat; y0_hat
         [,1]
[1,] 1.090694

> se_y0 = sigma_hat * sqrt(1 + t(x0) %*% solve(t(X) %*% X) %*% x0)
> crit_val = qt(0.975, n-p-1); crit_val
[1] 1.96362

> CI_low = y0_hat - crit_val*se_y0
> CI_high = y0_hat + crit_val*se_y0
> c(CI_low, CI_high)
[1] 0.2053714 1.9760164
```

Next, we use `predict()` **in R:**

```
> predict(object = myfit, newdata = data.frame(age = 7, ht = 47,
          sex = 1, smoke = 0), interval = "prediction", level = 0.95)
       fit       lwr      upr
1 1.090694 (0.2053714 1.976016)
```