STAT2401 Analysis of Experiments

Lecture Week 8 **Dr Darfiana Nur**

Aims of this lecture

1. Fitting MLR in R: Interpretation

   Ch 5.2 Sheather (2009)

2. ANOVA for MLR — Ch 5.2 Sheather (2009)
   - Partial F-Test
   - Hypotheses concerning one term
3. Diagnostic plots for MLR — Ch 6.1 Sheather (2009)
   - Residual plots
   - Leverage, Cook's Distance
4. Categorical/Indicator variables in regression

   Ch 5.2 Sheather (2009)

Recap 1: Multiple linear regression in matrix notation

- If we have p predictors, we can write that $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{I}\sigma^2)$, and the least squares estimate is

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

- Hence, the fitted values can be written as $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$, or $\widehat{\boldsymbol{y}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$, and the matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is known as the 'hat' matrix*
- Residuals are $\hat{\boldsymbol{e}} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$, and RSS (residual sum of squares/SSE) can be written as

$$RSS = (\boldsymbol{y} - \widehat{\boldsymbol{y}})'(\boldsymbol{y} - \widehat{\boldsymbol{y}})$$

and as before, we estimate σ^2 from RSS, i.e.,

$$s^2 = \frac{RSS}{n-p-1} = \frac{1}{n-p-1}\hat{\mathbf{e}}'\hat{\mathbf{e}}$$

- Note that the number of degrees of freedom is $n - p - 1$

Recap 2 : Multiple linear regression in matrix notation

- The covariance matrix of the LS estimates is
$$\mathrm{var}\left(\widehat{\boldsymbol{\beta}}\right) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2$$
and as before, we estimate σ^2 using s^2
- Hence, for carrying out a t-test for testing $H_0\colon \beta_i = 0$, we use
$$\frac{\hat{\beta}_i - \beta_i}{\mathrm{se}(\hat{\beta}_i)} \sim t_{n-p-1}$$

Aim 1 Interpreting coefficients

- In simple linear regression, the slope $\hat{\beta}_1$ can be interpreted as the amount by which the mean of Y changes for a unit change in x
  - Wind speed example: For every $1\ \mathrm{m/s}$ increase in wind speed at the reference site, **mean** wind speed changes by $\hat{\beta}_1 = 0.7557\ \mathrm{m/s}$ at the candidate site
  - Mother/daughter example: For every 1 inch increase in the height of mothers, the mean height of daughters increases by $\hat{\beta}_1 = 0.5417$ inches
- In multiple linear regression, we can only interpret a coefficient in this way if we add the requirement that all other variables be held constant

Example 1: fuel consumption in US states

- **Objective:** to understand how fuel consumption varies across the 50 US states and DC, and particular, to understand the effect on fuel consumption of state gasoline tax
- Note transformations of explanatory variables

`Income`	Average personal income for 2000
`logMiles`	log2 of miles of Federal highways
`Tax`	State gasoline tax (cents per gallon)
`Fuel`	Fuel sold per thousand licensed drivers
`Dlic`	Licensed drivers per thousand people

Example 1 Interpreting coefficients – Fuel Data

- In a model with a single explanatory variable (`Fuel ~ logMiles`) the coefficient of `logMiles` is 25.25
  - Interpretation: a unit increase in `logMiles` is associated with a mean increase of 25.25 units of fuel sold per thousand drivers
- In a model with all explanatory variables, (`Fuel ~ Tax + Dlic + Income + logMiles`) the coefficient of `logMiles` is 18.54
  - Interpretation: a unit increase in logMiles is associated with a mean increase of 18.54 units of fuel sold per thousand drivers, **all other variables being held constant**
  - Assumes that we could in fact change one predictor without changing all the others; very unlikely with observational data!
  - Change in value of effect due to relationship between predictors

Model summary

```
> summary(Fuel.lm1)


Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 154.192845 194.906161   0.791 0.432938
Tax          -4.227983   2.030121  -2.083 0.042873 *
Dlic          0.471871   0.128513   3.672 0.000626 ***
Income       -0.006135   0.002194  -2.797 0.007508 **
logMiles     18.545275   6.472174   2.865 0.006259 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 64.89 on 46 degrees of freedom
Multiple R-squared:  0.5105,   Adjusted R-squared:  0.4679
F-statistic: 11.99 on 4 and 46 DF,  p-value: 9.331e-07
```

Aim 2 ANOVA in MLR

- $H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$
- H_A: at least one of the $\beta_i \neq 0$

Analysis of variance table

Source of variation	Degrees of freedom (df)	Sum of squares (SS)	Mean square (MS)	F
Regression	p	SSreg	SSreg/p	$F = \dfrac{\text{SSreg}/p}{\text{RSS}/(n-p-1)}$
Residual	$n-p-1$	RSS	S^2= RSS/ $(n-p-1)$	
Total	$n-1$	SST = SYY		

Example 2 Fuel data - ANOVA

```
Analysis of Variance Table

Response: Fuel
           Df Sum Sq Mean Sq F value    Pr(>F)
Regression  4 201994   50499  11.992  9.33e-07
Residuals  46 193700    4211
---
Total      50 395694
```

p=4, n=51

F-Test

1. Hypotheses: $H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$ $\quad H_A$: at least one of the $\beta_i \neq 0$
2. Test Statistic: F=11.992
3. Sampling distribution: F(df=p, n-p-1) ie F(4,46)
4. p-value= 9.33e-07

5 and 6. We reject Ho, one or more of the slopes are significantly different from 0.

F testing in general

- Here we are comparing the two models.

(Model 1) $H_0 : \mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$

(Model 2) $H_A : \mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$

- The first model already has some explanatory variables in it.
- The second model is suggesting that the variable(s) in $\mathbf{X}_2$ are also needed.
- $\mathbf{X}_2$ could include more than 1 explanatory variable

- Suppose Model 1 has *q* regression parameters and Model (2) adds *c* parameters
- The *F*-statistic compares the *MS12* (=*SS12*/c) to the *MSE* for Model 2.

$$F = \frac{(SSE_1 - SSE_2)/c}{MSE_2} = \frac{SS12/c}{\hat{\sigma}^2_{(2)}} \sim F_{c,n-q-c}$$

- So, if *SS12* is large then *F* will be large, if *SS12* is 'small' then *F* will be 'small'. Here this statistic has an F distribution with c and (n-q-c) degrees of freedom.

ANOVA for comparing models

- Adding more variables **always** decrease $SSE \ or \ RSS$ (Residual Sum of Squares), so we need to work out whether the trade-off – smaller SSE/RSS versus a more ‘complex’ (more explanatory variables) model – is favourable
- One way of assessing the trade-off is the **partial $\boldsymbol{F}$-test**
- Like all the F-tests we’ve seen so far, the partial F-test involves ratios of sums-of-squares

Partial F-test for comparing models

- 'small': model with only `logMiles` $(SSE = RSS = 325216)$
  - $y = \beta_0 + \beta_1 x_1 + \epsilon$ (q=2 parameters,n=51)
- 'big': model with all four explanatory variables $(SSE = RSS = 193700)$
  - $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$ (add c=3 parameters, $\beta_2, \beta_3, \beta_4$)

$H_0: \beta_2 = \beta_3 = \beta_4 = 0$ against $H_A: \beta_2, \beta_3, \beta_4$ are not all zero

$$F = \frac{(RSS_{\text{small}} - RSS_{\text{big}})/(c\ )}{(RSS_{\text{big}})/(\mathrm{df}_{\mathrm{n-q-c}}\ )} \sim F_{(c,n-q-c)}$$

$$F = \frac{(325216 - 193700)/(3)}{193700/(n-q-c)} = \frac{131516/3}{4210.87} = 10.411 \sim F_{3,46}$$

and $p\left(F_{3,46} > 10.411\right) = 2.4 \times 10^{-5}$

(Reject Ho, we need to add the at least 1 of the 3 variables)

In *R*: partial F-test for comparing models

```
> anova(Fuel.lm0, Fuel.lm1)
Analysis of Variance Table

Model 1: Fuel ~ logMiles
Model 2: Fuel ~ Tax + Dlic + Income + logMiles
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     49 325216
2     46 193700  3    131516 10.411 2.402e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Hypotheses concerning one term

- It may be that we require information about whether it’s worth keeping one of the terms in the (full) model
  - For example, what happens if we delete Tax from the full model?
  - RSS will increase, but is the increase significant?
- Hypotheses will be

$$H_0: \beta_{\text{Tax}} = 0, \text{given estimates of all other coefficients}$$
$$H_A: \beta_{\text{Tax}} \neq 0, \text{given estimates of all other coefficients}$$

- Strategy: same as before

$$F = \frac{(RSS_{\text{small}} - RSS_{\text{big}})/(\text{df}_{\text{small}} - \text{df}_{\text{big}})}{(RSS_{\text{big}})/(\text{df}_{\text{big}})} \sim F_{(\text{df}_{\text{small}} - \text{df}_{\text{big}},\, \text{df}_{\text{big}})}$$

In *R*: partial F-test for comparing models

```
> anova(Fuel.lm3, Fuel.lm1)

Analysis of Variance Table

Model 1: Fuel ~ Dlic + Income + logMiles
Model 2: Fuel ~ Tax + Dlic + Income + logMiles
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     47 211964
2     46 193700  1     18264 4.3373 0.04287 *
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Add 1 parameter c=1, q=4, n-c-q=46
- $F = \frac{18264/1}{193700/46} = 4.3373 \sim F_{1,46}$
- Compare with t–statistic of Tax in full model!
- Weak evidence to support keeping Tax in model **given other variables**

Aim 3 Diagnostics

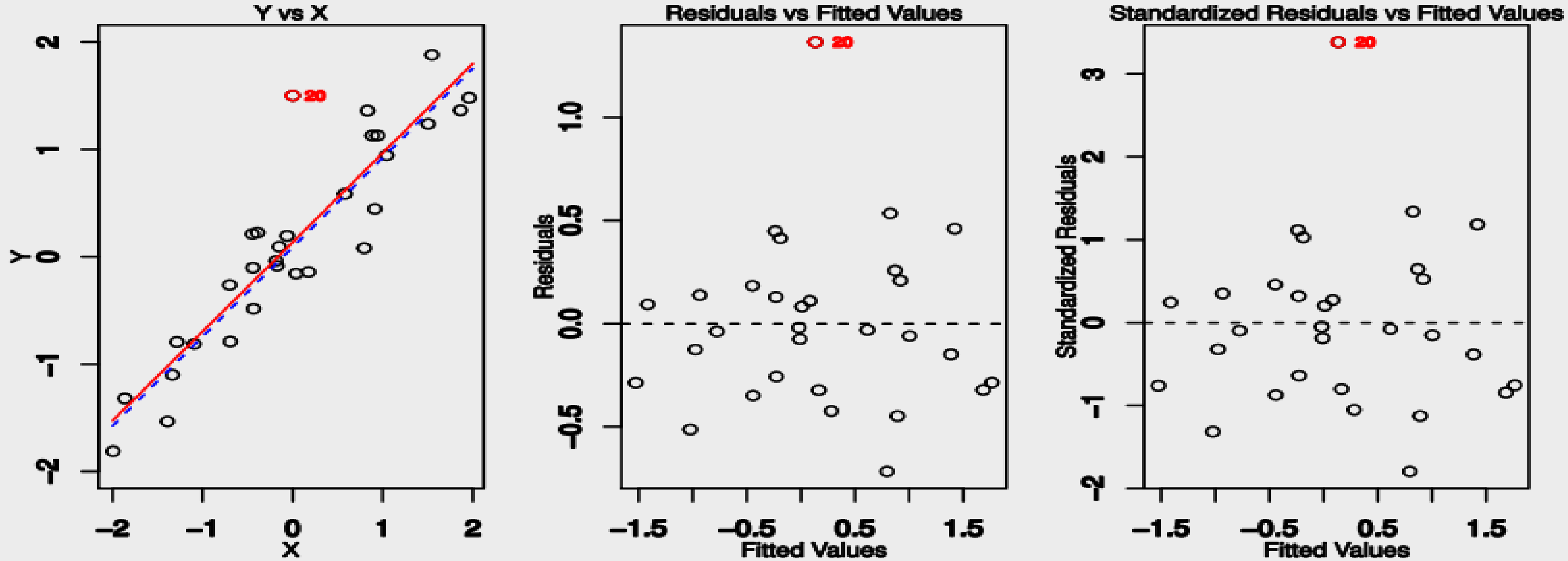
- Histogram and QQ plot of standardized residuals
- Plots of standardized residuals against each of the explanatory variables
- Plot of standardized residuals against fitted values
- Plot of fitted against actual values

Diagnostics for MLR

- As before, we examine as many residual plots as we can, but when we have many explanatory variables, it helps to have some numerical diagnostics
- Convention is to use standardized residuals – why?
- Recall that we can write
$$\hat{\boldsymbol{e}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$$
and we can show that
$$\mathrm{var}(\hat{\boldsymbol{e}}) = (\boldsymbol{I} - \boldsymbol{H})\sigma^2$$
which implies that $\mathrm{var}(\hat{e}_i) = (1 - h_{ii})\sigma^2$, where h_{ii} is the i^{th} diagonal element $\boldsymbol{H}$
- A better way of standardizing the residuals so they have constant variance is
$$r_i = \frac{\hat{e}_i}{s\sqrt{1 - h_{ii}}}$$

Outliers

- The red point (observation 20) in the left-hand panel of the following figures illustrates a typical *outlier*. The red solid line is the least squares regression fit, while the blue dashed line is the least squares fit after removal of the outlier.



Outliers

- But in practice, it can be difficult to decide how large a residual needs to be before we consider the point to be an outlier.
- To address this problem, instead of plotting the residuals, we can plot the *Standardized residuals*, computed by dividing each residual e_i by its estimated *standard error*.
- Observations whose *Standardized residuals* are greater than 2 in absolute value are possible *outliers*.
- In the right-hand panel, the *outlier*'s *Standardized residual* exceeds 3, while all other observations have *Standardized residuals* between -2 and 2.

Outliers

- It is typical for an *outlier* that does not have an unusual predictor value to have little effect on the least squares fit.
- However, even if an *outlier* does not have much effect on the least squares fit, it can cause other problems.

Outliers

- For instance, in this example, the RSE is 0.2927 when the outlier is included in the regression:

```
> summary(lm(Y~X,data=Data))

Residual standard error: 0.2927 on 28 degrees of freedom
Multiple R-squared:  0.9246,Adjusted R-squared:  0.9219
F-statistic: 343.2 on 1 and 28 DF,  p-value: < 2.2e-16
```

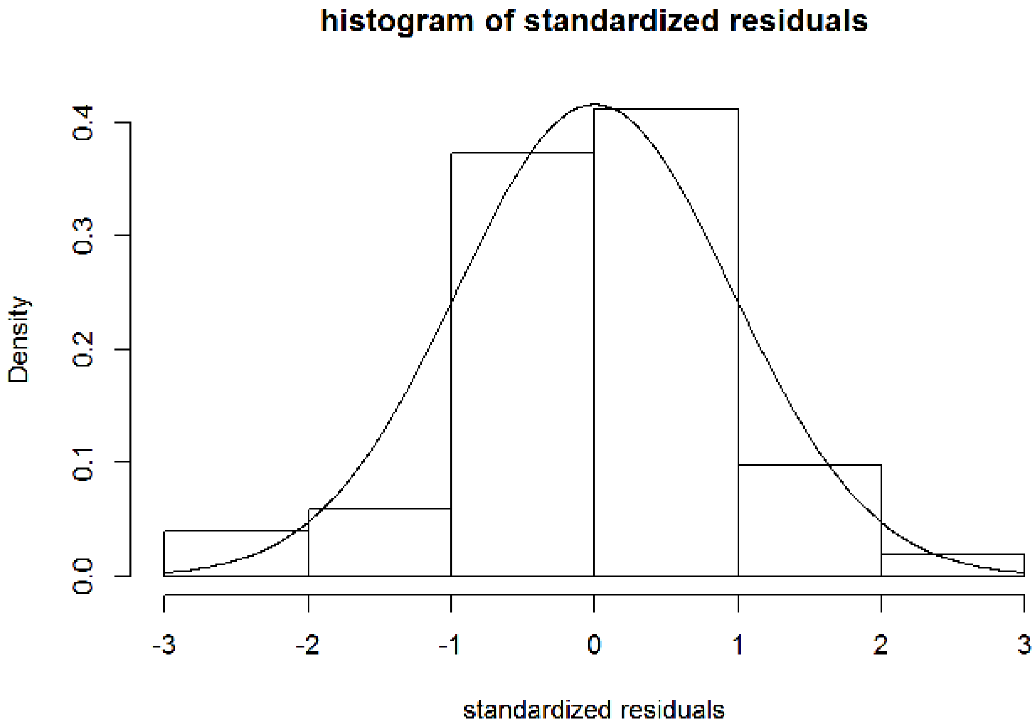
but it is only 0.107 when the outlier is removed:

```
> summary(lm(Y~X,data=Data[-20,]))

Residual standard error: 0.107 on 27 degrees of freedom
Multiple R-squared:  0.9896,Adjusted R-squared:  0.9892
F-statistic:  2573 on 1 and 27 DF,  p-value: < 2.2e-16
```

Outliers

- Since the RSE is used to compute all confidence intervals and p-values, such a dramatic increase caused by a single data point can have implications for the interpretation of the fit.
- Similarly, inclusion of the *outlier* causes the R^2 to decline from 0.9896 to 0.9246.
- If we believe that an *outlier* has occurred due to an error in data collection or recording, then one solution is to simply remove the observation.
- However, care should be taken, since an *outlier* may instead indicate a deficiency with the model, such as a missing predictor.

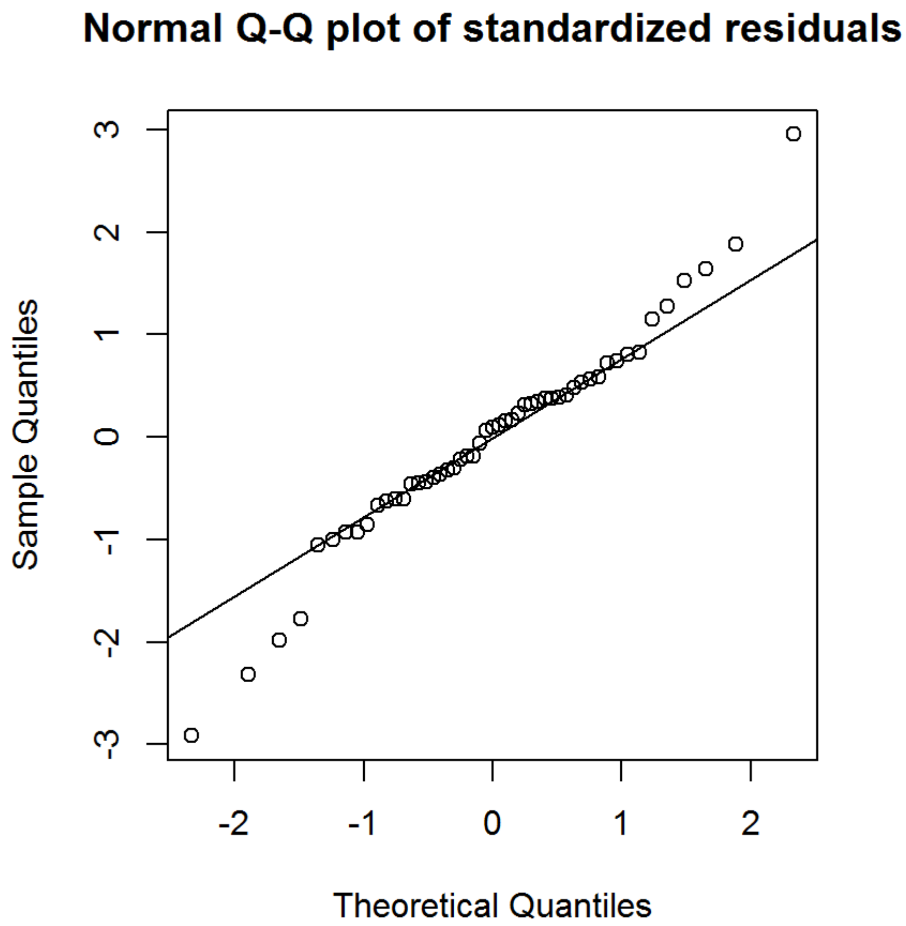
Example 3 Fuel data – histogram of standardized residuals



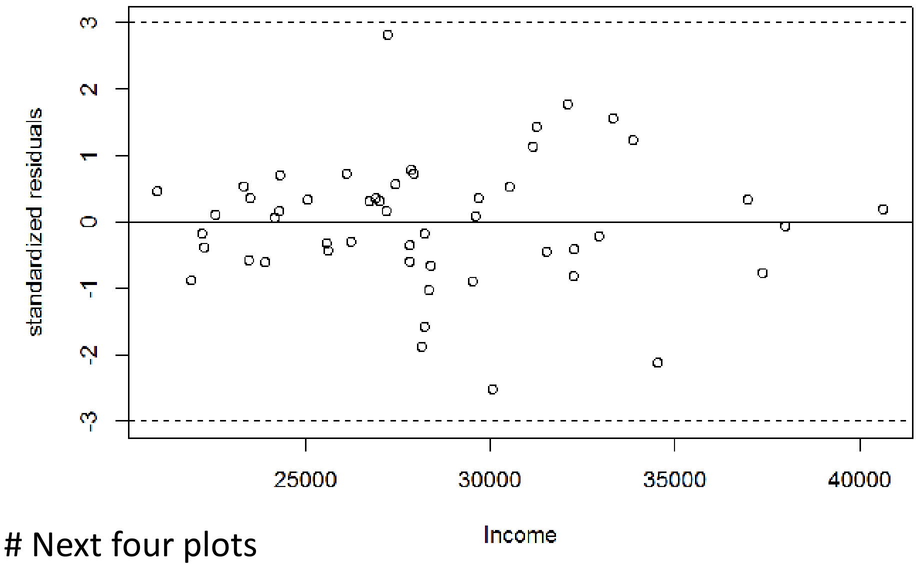
The histogram of standardised residuals depicts of Normality, visually.

Fuel data: QQ plot of standardized residuals

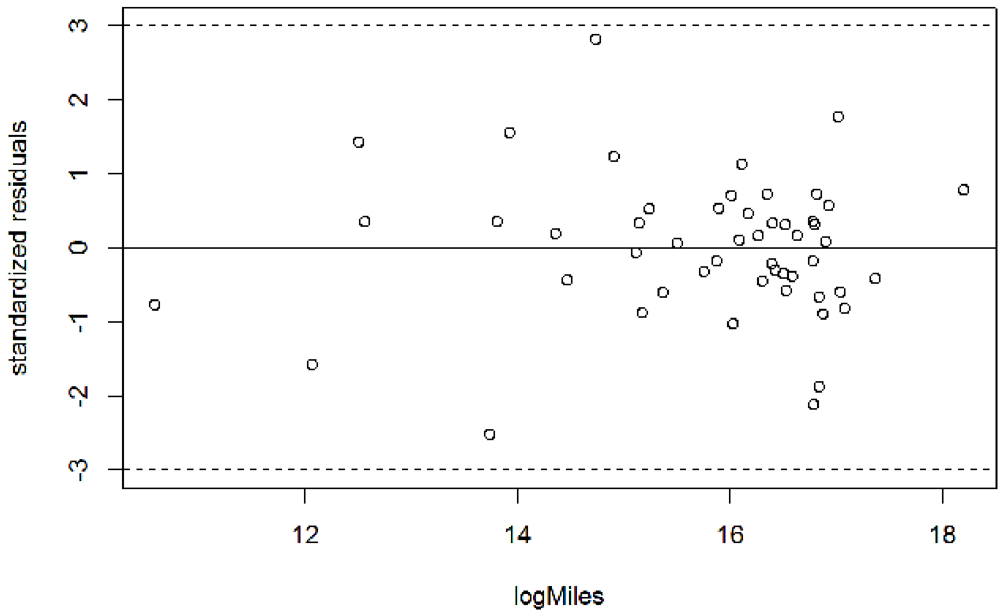
Most of the points lie on the straight line except a few on the upper and lower parts.



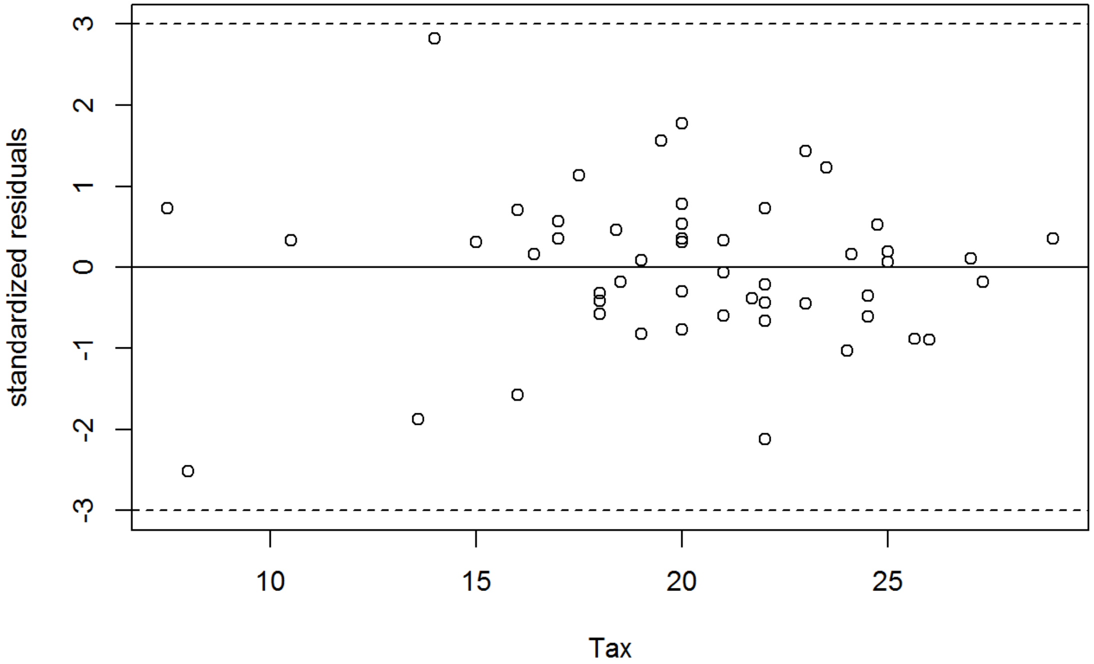
Fuel data - r_i against Income
Fuel data - r_i against logMiles

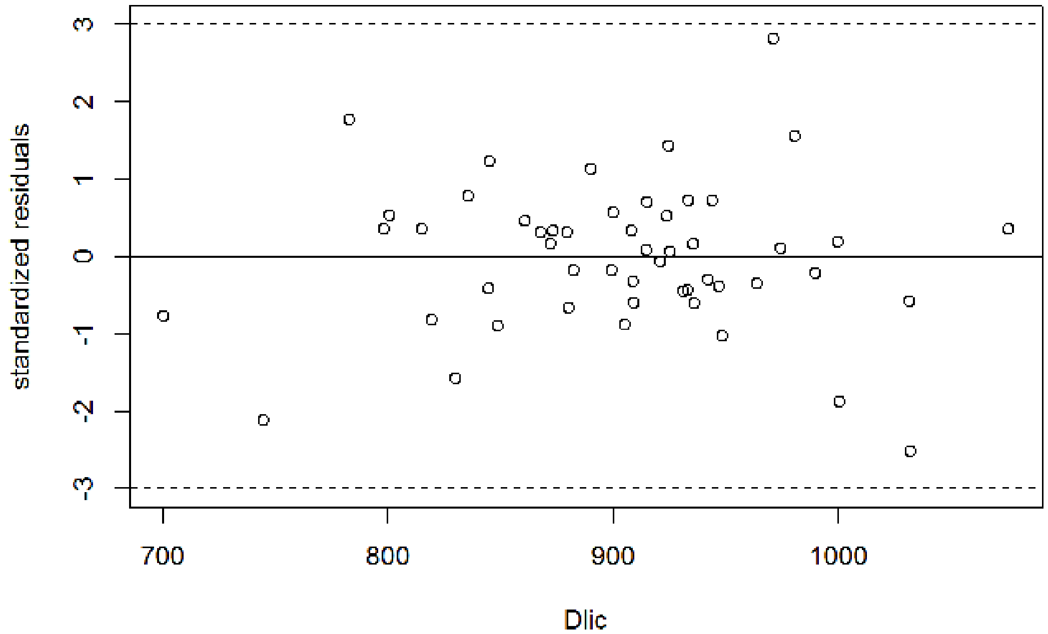



```
# Next four plots
for(i in c(1, 2, 3, 5)){
  plot(Fuel2001[, i], stdResid, xlab = names(Fuel2001)[i],
ylab = "standardized residuals",
     ylim = c(-3, 3))
  abline(h = 0)
  abline(h = c(-3, 3), lty = 2)
}
```

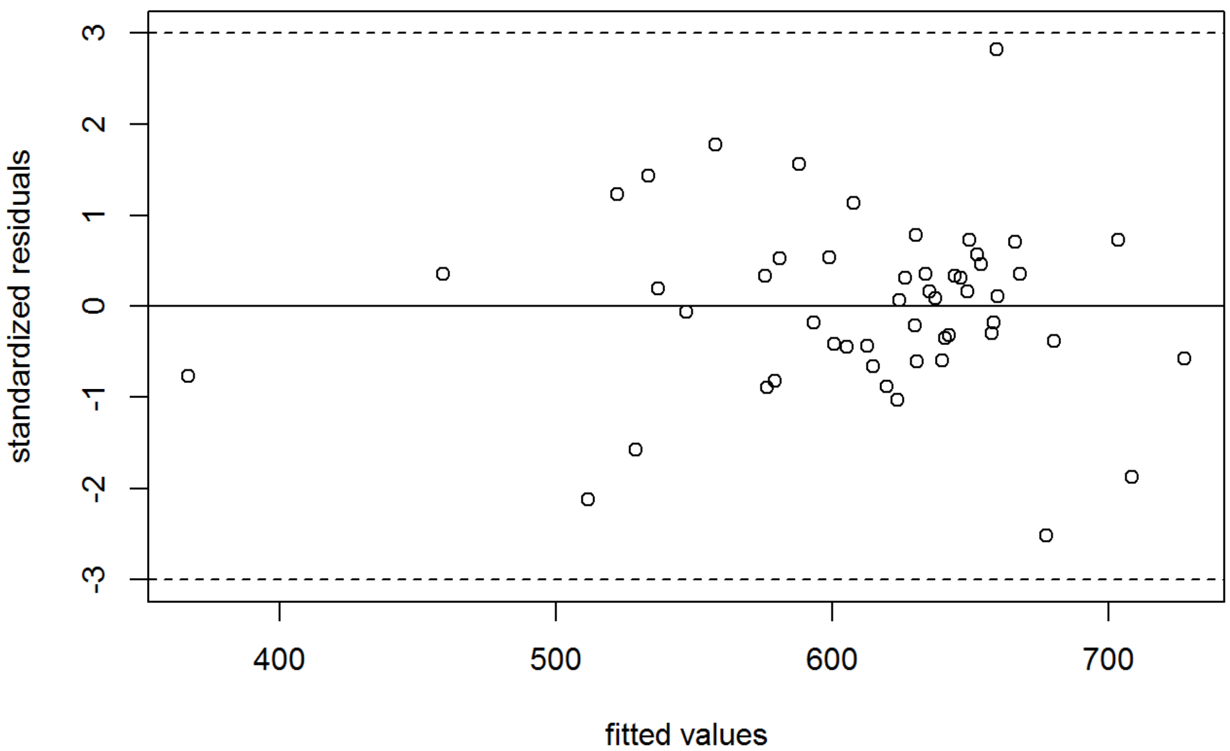


Fuel data - r_i against Tax
Fuel data - r_i against Dlic



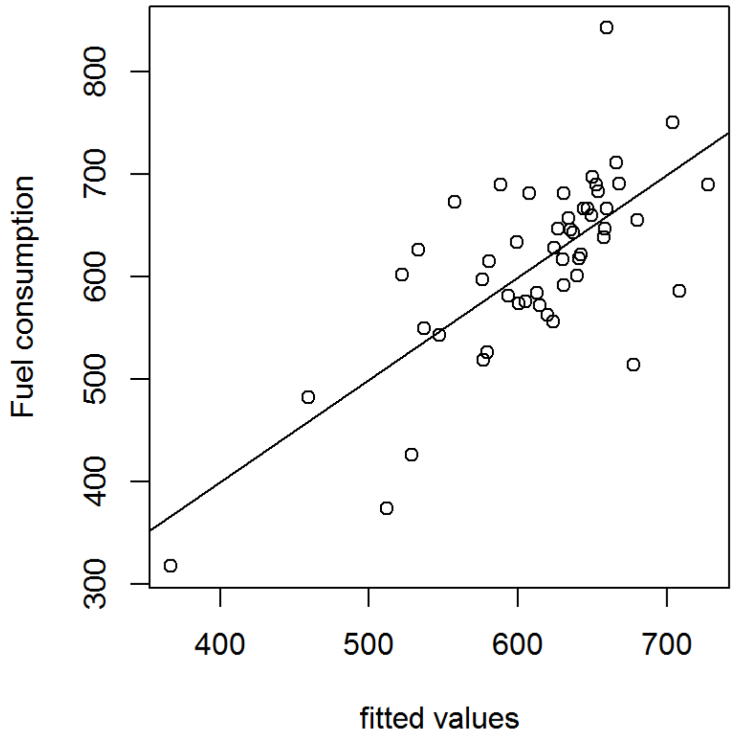


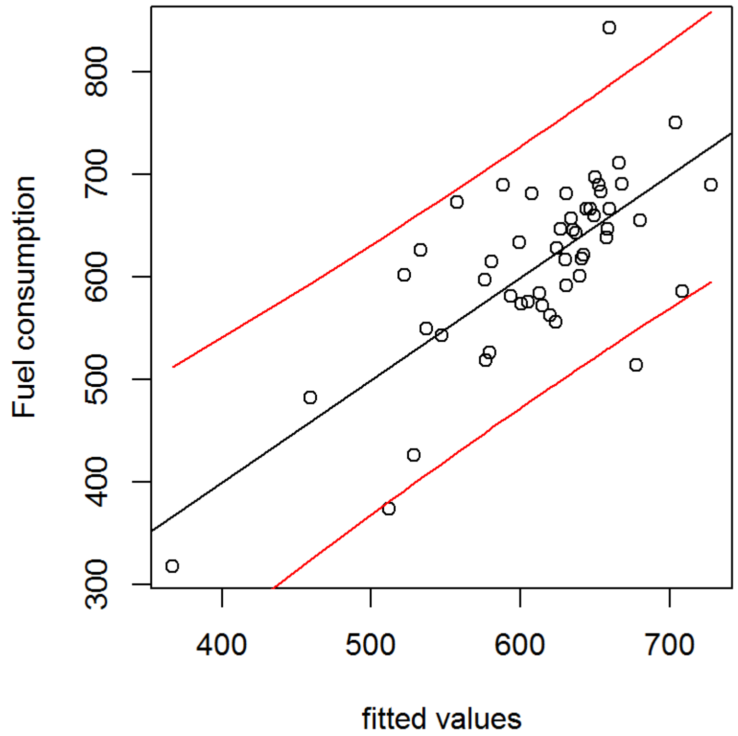
Fuel data - r_i against $\hat{y}_i$ (fitted)



plot(stdResid, Fuel.lm1$fitted, xlab = “fitted values”, ylab =“standardized residuals”)
abline(h = c(-3, 3), lty = 2)
abline(h = 0)

Fuel data: actual vs fitted



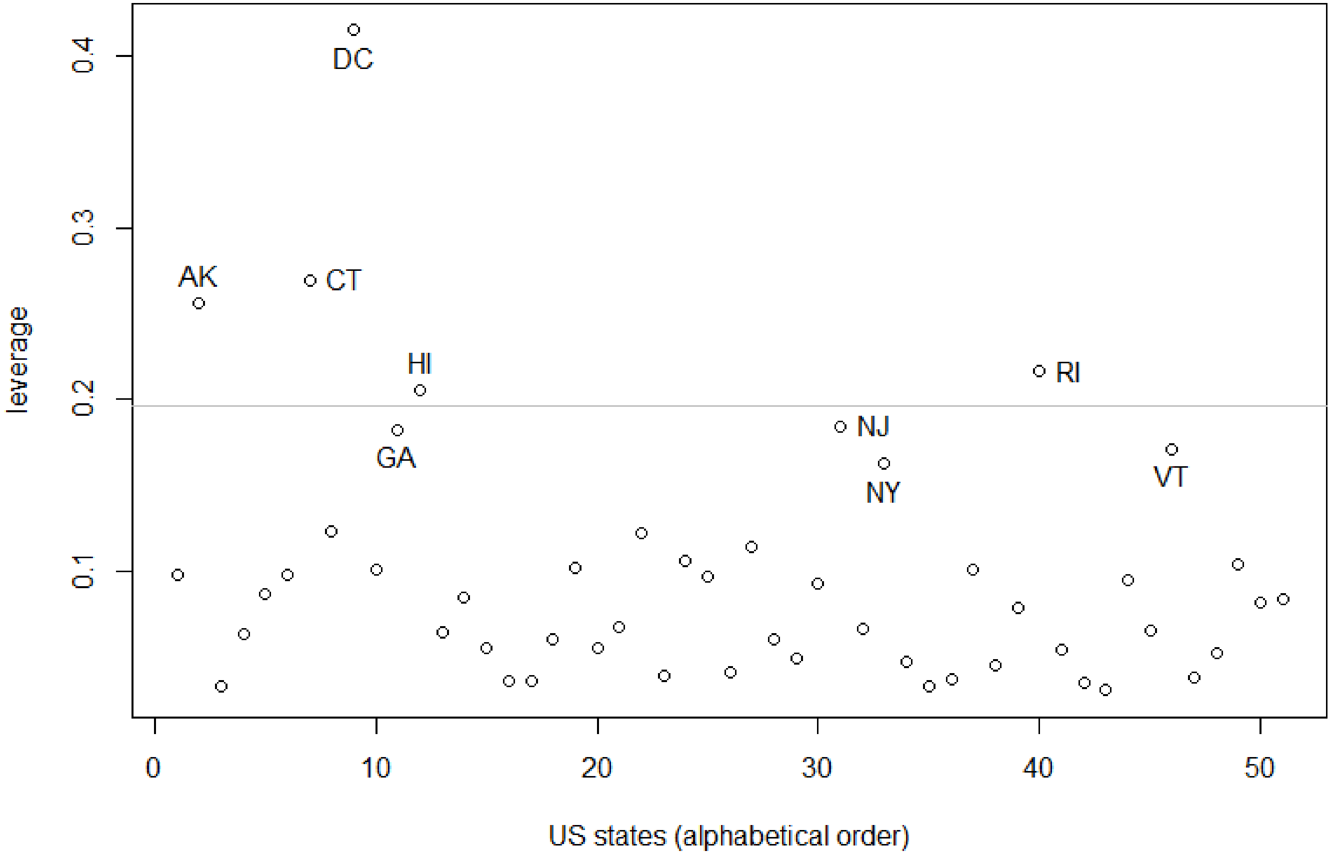


Leverage

- A high leverage point is easy to identify when we have only a single explanatory variable; when there are many xs, a numerical measure would be useful
- A point of high leverage $\boldsymbol{x}_{\boldsymbol{i}}' = (x_{1i}, x_{2i}, \dots, x_{pi})$ is one which:
  - is 'far away' from the bulk of the other $\boldsymbol{x}$s
  - 'attracts' the fitted regression line
- A useful measure of leverage is the i^{th} diagonal element h_{ii} of the hat matrix $\boldsymbol{H}$, and a useful rule of thumb is that a point has high leverage if

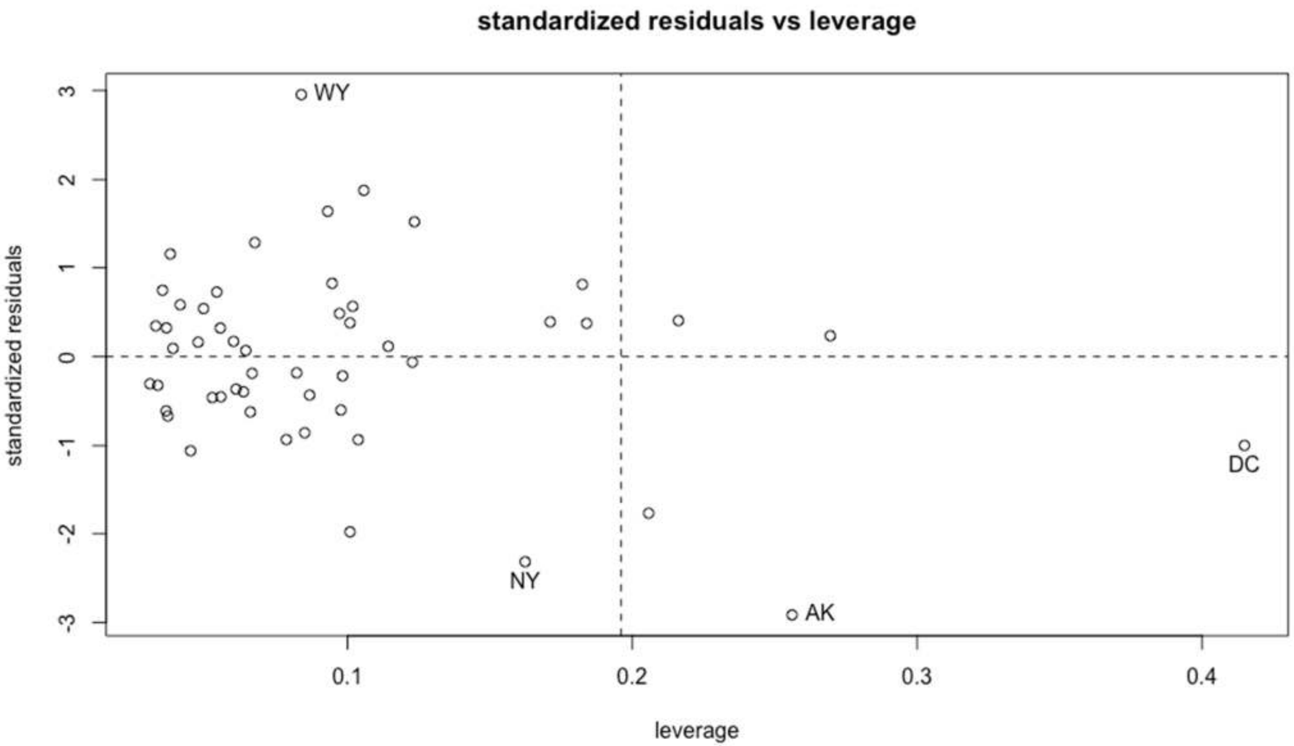
$$h_{ii} > \frac{2(p+1)}{n}$$

Example 4: Leverage plot- fuel data




```
plot(1:51, hatvalues(Fuel.lm1), xlab = "US states (alphabetical order)", ylab = "leverage")
abline(h = 2 * 5/51, col = "grey")

# Can only do the next step interactively in the console
identify(1:51, hatvalues(Fuel.lm1), labels = rownames(Fuel2001))
```

Fuel data: full model




```
plot(hatvalues(Fuel.lm1),
stdres(Fuel.lm1), xlab =
"leverage", ylab =
"standardized residuals",
    main = "standardized
residuals vs leverage")
abline(v = 2 * 5 / 51, lty = 2)
abline(h = 0, lty = 2)

# Can only do the next step
interactively in the # console
identify(hatvalues(Fuel.lm1),
rstandard(Fuel.lm1), labels =
rownames(Fuel2001))
```

Influential observations

- Single or small groups of observations can strongly influence the fit of a regression model
- ***Influence analysis*** studies changes in a specific part of an analysis under the assumption that the model is correct
  - 'Easy' way would be to delete observations from the data one at a time and then study its effects, for example, changes in coefficients $\widehat{\boldsymbol{\beta}}$
  - Observations whose removal causes major changes are called *influential*
- A useful measure of influence is *Cook's Distance*, D_i, which reflects two aspects: a large residual *and* a large leverage:

$$D_i = \frac{r_i^2}{2} \frac{h_{ii}}{1 - h_{ii}},$$

- A useful rule of thumb is that a point is an influential observation if

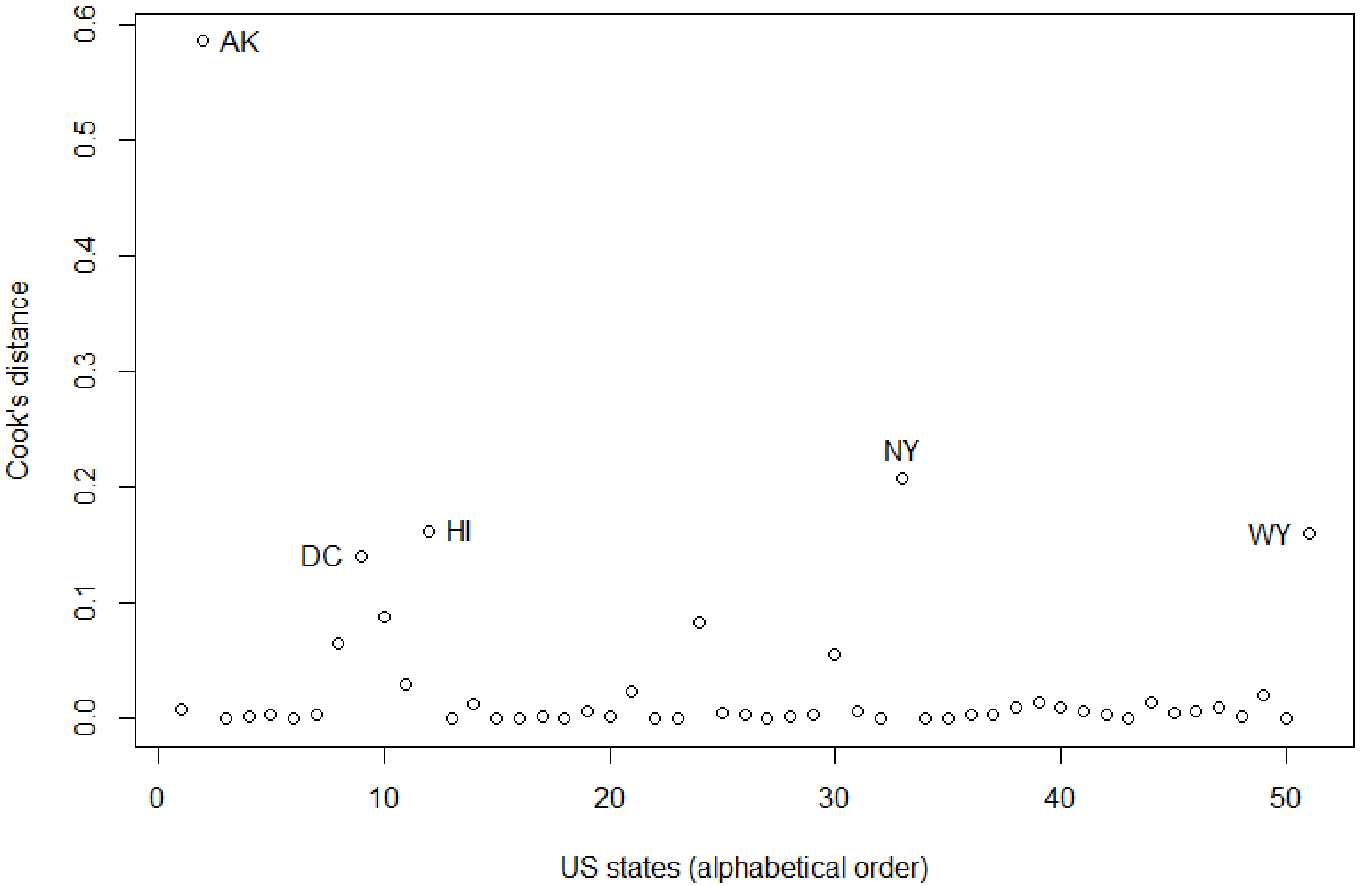
$$D_i > \frac{2(p+1)}{n-(p+1)}$$

Leverage – bad or good?

According to Sheather (2009):

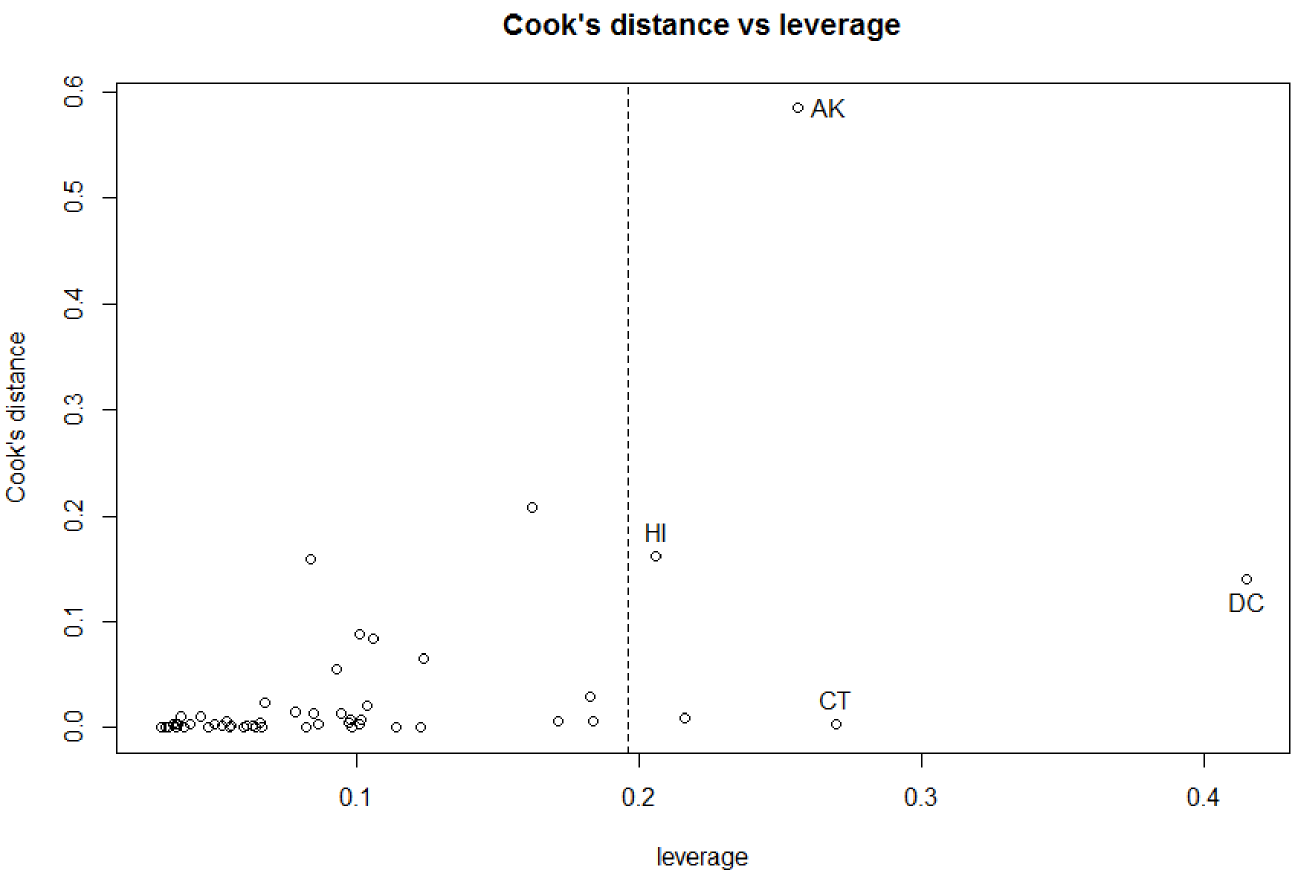
- for small to moderate sample sizes, points are considered as **outliers** if the standardized residual for the point falls outside the interval from –2 to 2 .
- for very large data sets, this rule may change to –4 to 4 based on standardised residuals.
- a **bad leverage point** is a **leverage point** (if leverage $h_{ii} > 2(p+1)/n$) which is also an **outlier.** Thus, a bad leverage point is a leverage point whose standardized residual falls outside the interval from –2 to 2.
- On the other hand, a **good leverage point** is a leverage point whose standardized residual falls inside the interval from –2 to 2 .

Example 5: Cook's distance – fuel data



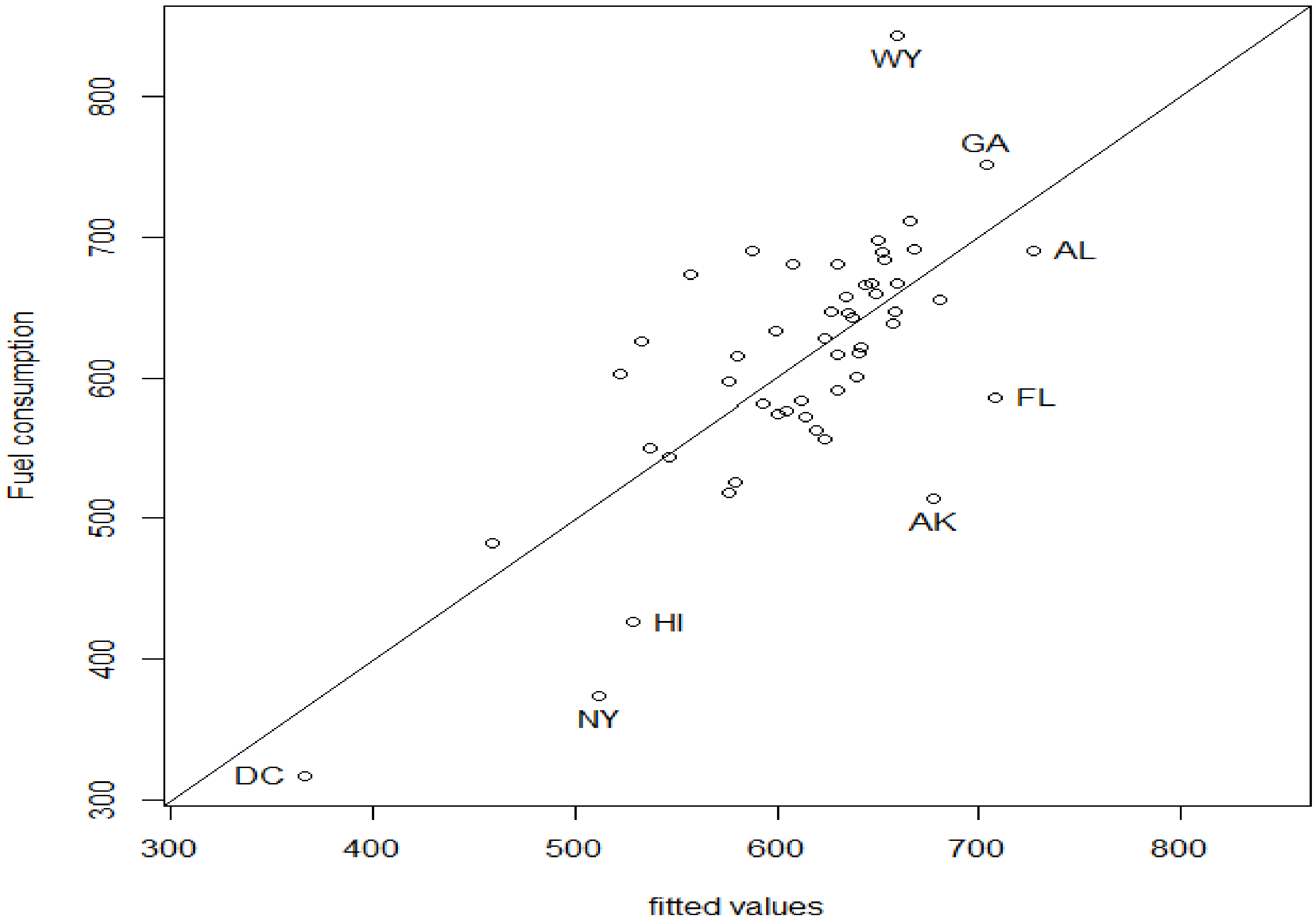

```
plot(cooks.distance(Fuel.lm1),
xlab = "US states (alphabetical
order)", ylab = "Cook's
distance")

# Can only do the next step
interactively in the console
identify(1:51,
cooks.distance(Fuel.lm1), label
= rownames(Fuel2001))
```

Fuel data: full model




```
plot(hatvalues(Fuel.lm1),
cooks.distance(Fuel.lm1),
xlab = "leverage", ylab =
"Cook's distance",
    main = "Cook's
distance vs leverage")
abline(v = 2 * 5 / 51, lty =
2)
```

Fuel data - fitted vs actual ('big' model)

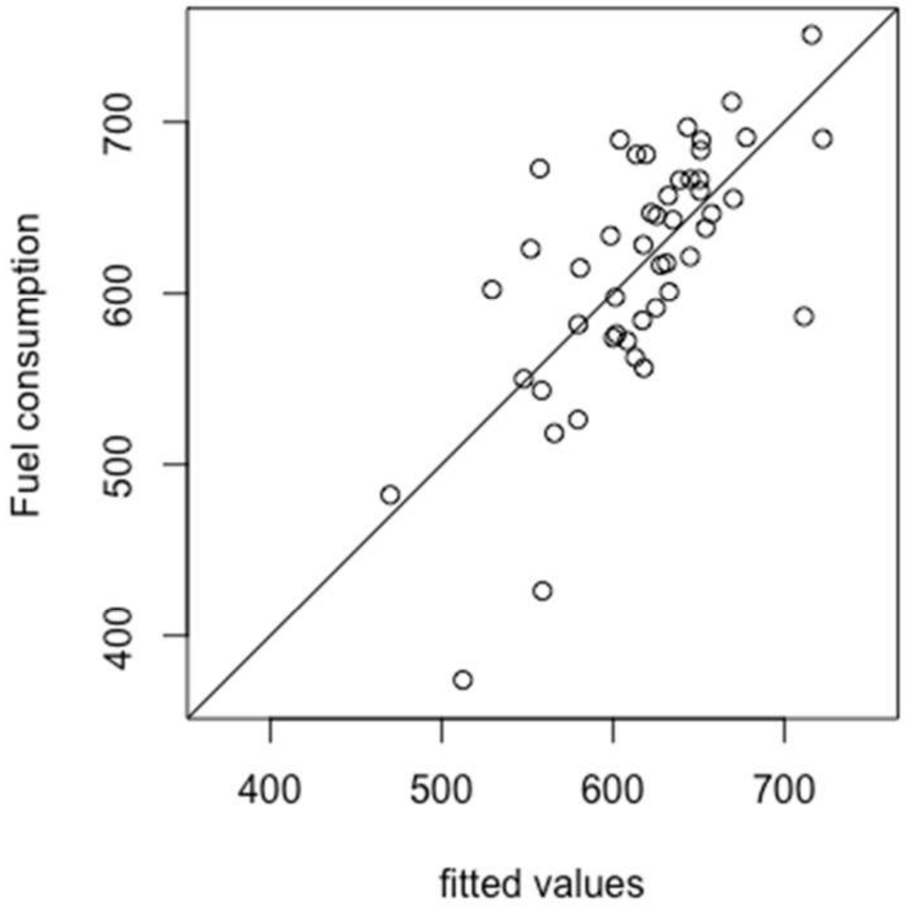


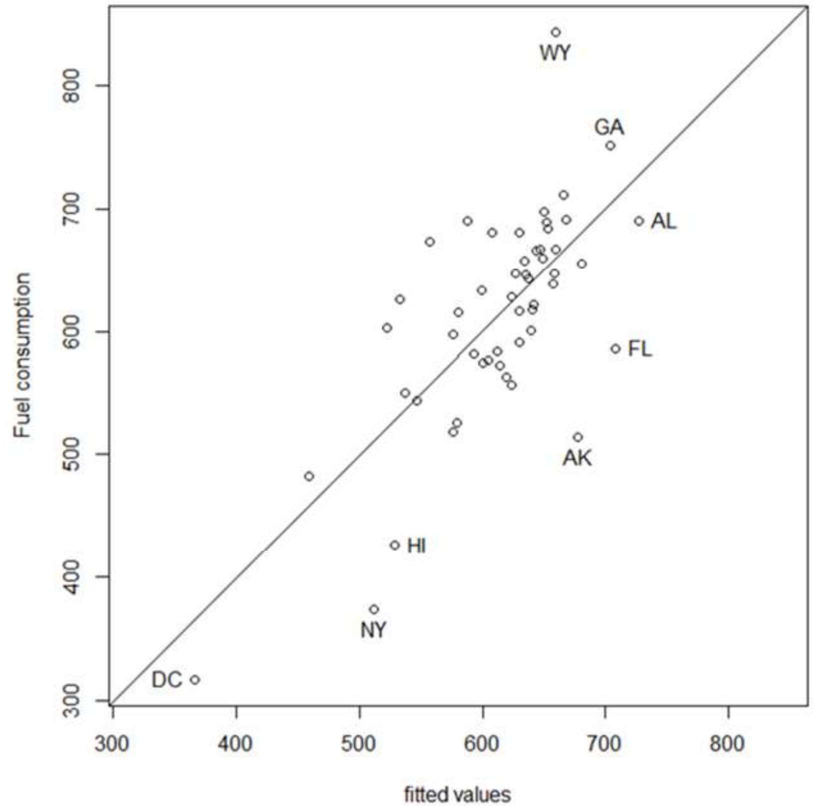
Example 6 Fuel data: refit without AK, DC, WY

```
> summary(Fuel.lm1)

Call:
lm(formula = Fuel ~ ., data = Fuel2001)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 154.192845 194.906161   0.791 0.432938
Tax          -4.227983   2.030121  -2.083 0.042873
Dlic          0.471871   0.128513   3.672 0.000626
Income       -0.006135   0.002194  -2.797 0.007508
logMiles     18.545275   6.472174   2.865 0.006259
---
Residual standard error: 64.89 on 46 df
Multiple R-squared:  0.5105,
Adjusted R-squared:  0.4679
F-statistic: 11.99 on 4 and 46 DF,
p-value: 9.331e-07
```

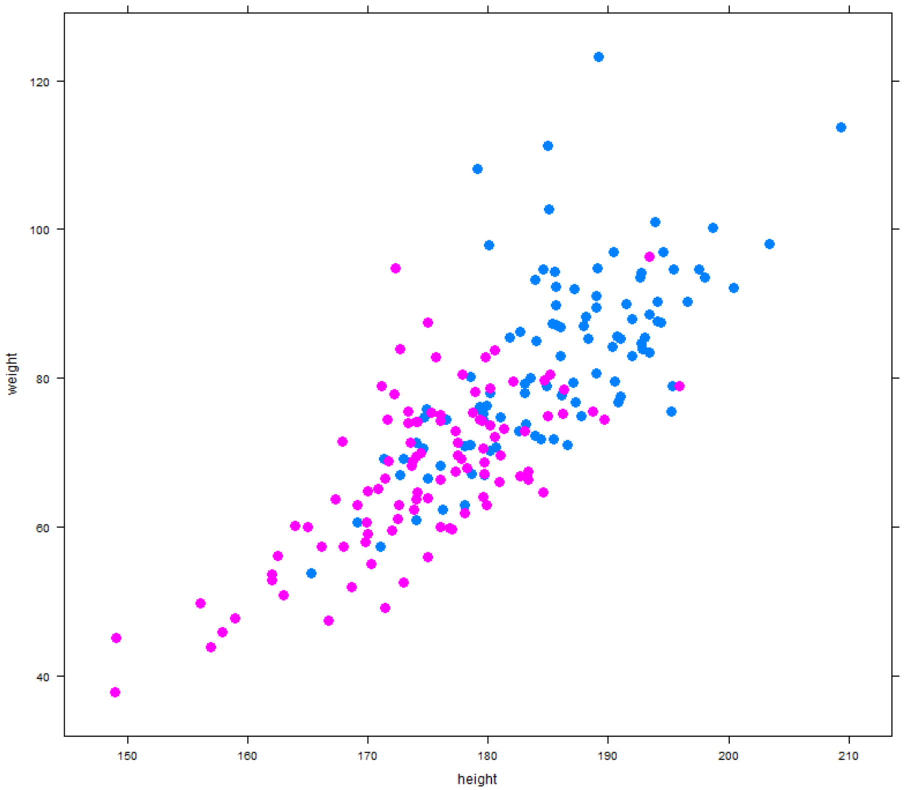
```
> summary(Fuel.lm2)

Call: lm(formula = Fuel ~ ., data = Fuel2001,
subset = -c(2, 9, 51))

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 268.368526 189.074455   1.419 0.162997
Tax          -5.601341   1.910106  -2.932 0.005371
Dlic          0.456181   0.122482   3.724 0.000565
Income       -0.005413   0.001848  -2.929 0.005423
logMiles     12.793390   6.475399   1.976 0.054630
---
Residual standard error: 54.29 on 43 degrees of
freedom
Multiple R-squared:  0.4832,
Adjusted R-squared:  0.4351
F-statistic: 10.05 on 4 and 43 DF,
p-value: 7.824e-06
```

Fuel data - fitted vs actual (without/with AK, DC, WY)





Aim 4 Categorical/indicator variables




```
require(lattice)
xyplot(Wt ~ Ht, groups = Sex, data = ais, pch = 16, xlab = "height", ylab =
"weight")
```

Example 7

- Consider the athlete height/weight data
- We might be interested in determining whether linear regressions are different for males and females
- How might we express such a model?

Categorical/indicator variables

- We could fit two separate regression models, but it makes more sense to combine them into a single regression model as follows:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i + \epsilon_i$$

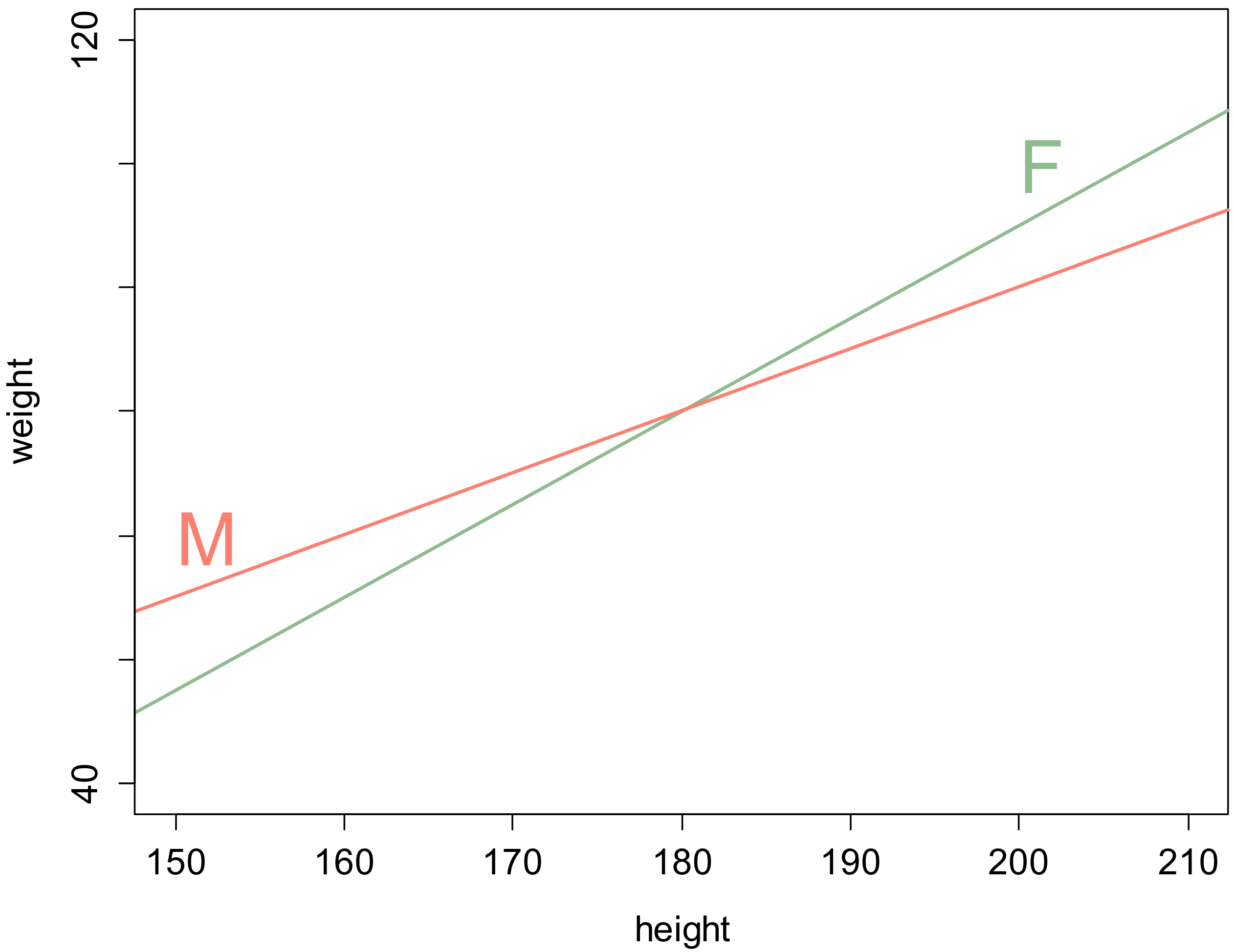
- For males, $z_i = 0$, so

$$E(Y|X) = \beta_0 + \beta_1 x$$

- For females, $z_i = 1$, so

$$E(Y|X) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x$$

Categorical/indicator variables



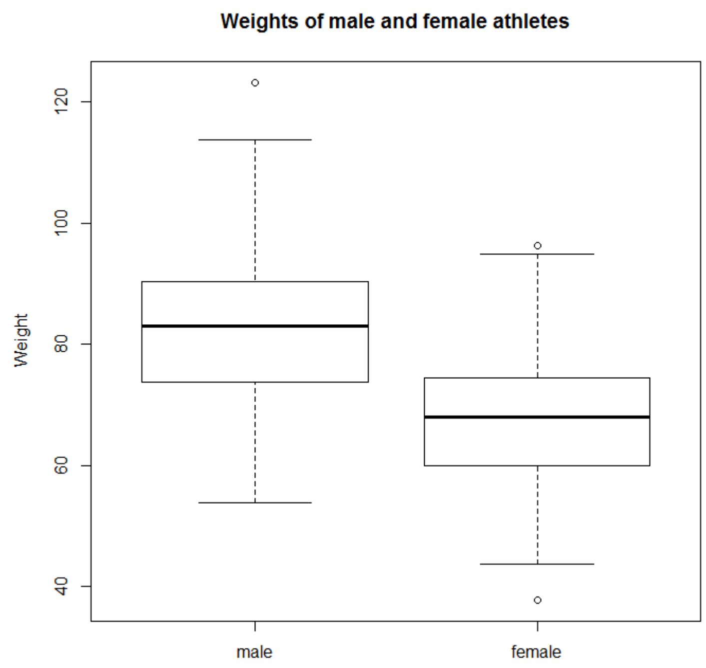
Categorical/indicator variables

```
> WtHtS.lm <- lm(Wt ~ Ht + Sex + Ht*Sex, data = ais)
> summary(WtHtS.lm)

Call:
lm(formula = Wt ~ Ht + Sex + Ht * Sex, data = ais)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -111.5476    20.0116  -5.574 8.07e-08 ***
Ht             1.0462     0.1078   9.707  < 2e-16 ***
Sex           15.0144    27.0809   0.554    0.580
Ht:Sex        -0.1076     0.1500  -0.717    0.474
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.561 on 198 degrees of freedom
Multiple R-squared:  0.6277,     Adjusted R-squared:  0.6221
F-statistic: 111.3 on 3 and 198 DF,  p-value: < 2.2e-16
```

Conclude that a single linear model is a good description of male and female athletes

t-test as a linear model

```
    Two Sample t-test

data:  ais$Wt[ais$Sex == 0] and
ais$Wt[ais$Sex == 1]
t = 9.226, df = 200, p-value <2e-16
alternative hypothesis: true
difference in means is not equal to
0
95 percent confidence interval:
 11.9365 18.4256
sample estimates:
mean of x mean of y
  82.5235   67.3425
```



boxplot(Wt ~ Sex, data = ais, names = c("male", "female"), ylab = "Weight", main = "Weights of male and female athletes")
t.test(Wt ~ Sex, var.equal=TRUE, data = ais)

t-test as a linear regression

```
> lm6 <- lm(Wt ~ Sex, data = ais)
> summary(lm6)

Call:
lm(formula = Wt ~ Sex, data = ais)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    82.52       1.16   71.28   <2e-16 ***
Sex           -15.18       1.65   -9.23   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.7 on 200 degrees of freedom
Multiple R-squared:  0.299,   Adjusted R-squared:  0.295
F-statistic: 85.1 on 1 and 200 DF,  p-value: <2e-16
```

t-test with equality of variances is equivalent to fitting $y_i = \beta_0 + \beta_1 z_i + \epsilon_i$, where z is an indicator variable that takes a value of 0 for males, and 1 for females

Two categorical variables

- We analyse data where the response represents the amount of suspended solids in a coal cleaning system
- The continuous explanatory variable is the pH of the system, and there are **three** polymer flocculants whose effect we wish to assess
- If there are l categories, there will be $l-1$ indicator variables
- A simple model with two categorical variables is

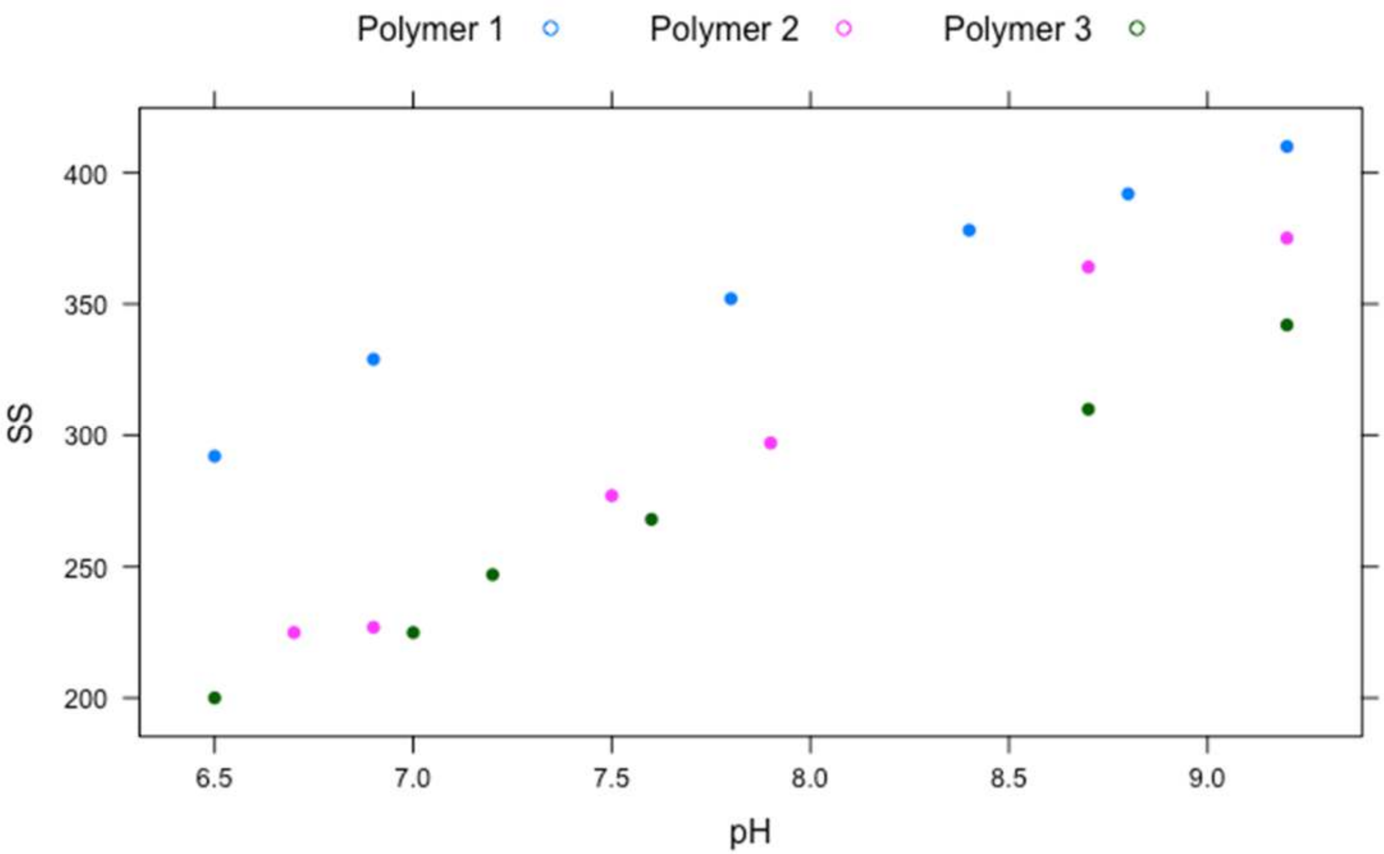
$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_{1i} + \beta_3 z_{2i} + \epsilon_i$$

$z_1 = 1$ for polymer **1**, $z_1 = 0$ otherwise

$z_2 = 1$ for polymer **2**, $z_2 = 0$ otherwise

(ie. polymer 3 when $z_1 = 0$ and $z_2 = 0$)

Example 8 Polymer flocculants



pH	Solids	z1	z2
6.5	292	1	0
6.9	329	1	0
7.8	352	1	0
8.4	378	1	0
8.8	392	1	0
9.2	410	1	0
6.7	225	0	1
6.9	227	0	1
7.5	277	0	1
7.9	297	0	1
8.7	364	0	1
9.2	375	0	1
6.5	200	0	0
7	225	0	0
7.2	247	0	0
7.6	268	0	0
8.7	310	0	0
9.2	342	0	0

Example 8 Polymer flocculants

- A more complex model might be

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_{1i} + \beta_3 z_{2i} + \beta_4 x_i z_{1i} + \beta_5 x_i z_{2i} + \epsilon_i$$

- In *R*, we don't have to create the indicator variables explicitly: instead, we only need a 'factor' variable that has as many levels as categories, and *R* will carry out the expansion internally
- For fitting different intercepts only,

```
lm(SS ~ pH + Type, data = Polymer)
```

- For fitting different intercepts and slopes,

```
lm(SS ~ pH + Type + pH:Type, data = Polymer)
```

Example 8 Polymer flocculants

```
Polymer.lm1 <- lm(SS ~ pH + Type, data = Polymer)
summary(Polymer.lm1)


Call:
lm(formula = SS ~ pH + Type, data = Polymer)

Residuals:
    Min      1Q  Median      3Q     Max 
-20.126  -6.183  -1.180   7.174  24.503 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    -48.285     26.278  -1.837   0.0875 .  
pH              51.317      3.244  15.818 2.52e-10 ***
TypePolymer 2  -58.680      7.511  -7.812 1.80e-06 ***
TypePolymer 3  -81.526      7.540 -10.813 3.52e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.99 on 14 degrees of freedom
Multiple R-squared:  0.9672,	Adjusted R-squared:  0.9602 
F-statistic: 137.7 on 3 and 14 DF,  p-value: 1.258e-10
```

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_{1i} + \beta_3 z_{2i} + \epsilon_i$$

```
Polymer.lm2 <- lm(SS ~ pH + Type + Type:pH, data = Polymer)
summary(Polymer.lm2)


Call:
lm(formula = SS ~ pH + Type + Type:pH, data = Polymer)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.1517 -4.8480 -0.6488  3.4620 12.3730 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)        39.417     26.273   1.500 0.159390    
pH                 40.263      3.287  12.249 3.85e-08 ***
TypePolymer 2    -253.720     38.369  -6.613 2.49e-05 ***
TypePolymer 3    -163.613     37.059  -4.415 0.000843 ***
pH:TypePolymer 2   24.787      4.841   5.121 0.000253 ***
pH:TypePolymer 3   10.326      4.707   2.194 0.048667 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.857 on 12 degrees of freedom
Multiple R-squared:  0.9897,	Adjusted R-squared:  0.9854 
F-statistic: 231.3 on 5 and 12 DF,  p-value: 1.699e-11
```

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_{1i} + \beta_3 z_{2i} + \beta_4 x_i z_{1i} + \beta_5 x_i z_{2i} + \epsilon_i$$

Example 8 Polymer flocculants: ‘full’ model