Simple Linear Models

MAST90044

Thinking and Reasoning with Data

Dr Julia Polak

Chapter 6

School of Mathematics & Statistics
The University of Melbourne

Outline

One numerical explanatory variable

- Data display
- Estimation
- Diagnostics
- Interpretation
- R squared & Inference straight-line regression

One categorical explanatory variable

- Models for Categorical Predictors
- Hypothesis testing, ANOVA & F-test

Steps for statistical modeling

1. Estimate
2. Check
3. Interpret
4. Make prediction

Steps for statistical modeling

1. Estimate
2. Check
3. Interpret
4. Make prediction

Cancer mortality near Hanford Reactor, WA, USA

Source:
`http://www.statsci.org/data/general/hanford.html`

- Hanford: plutonium production plant for decades
- strontium 90 and cesium 137 leaked into the Columbia River
- Index of exposure and cancer mortality rate 1959–1964
- Index measured risk for 9 Oregon counties, using
  - county's stream distance from Hanford
  - average distance of population from any water frontage

Cancer mortality near Hanford Reactor*

Mortality = number of deaths per 100,000 person-years (sum of # yrs each person spent in study)

```
> hanford
```

	County	Exposure	Mortality
1	Umatilla	2.49	147.1
2	Morrow	2.57	130.1
3	Gilliam	3.41	129.9
4	Sherman	1.25	113.5
5	Wasco	1.62	137.5
6	HoodRiver	3.83	162.3
7	Portland	11.64	207.5
8	Columbia	6.41	177.9
9	Clatsop	8.34	210.3

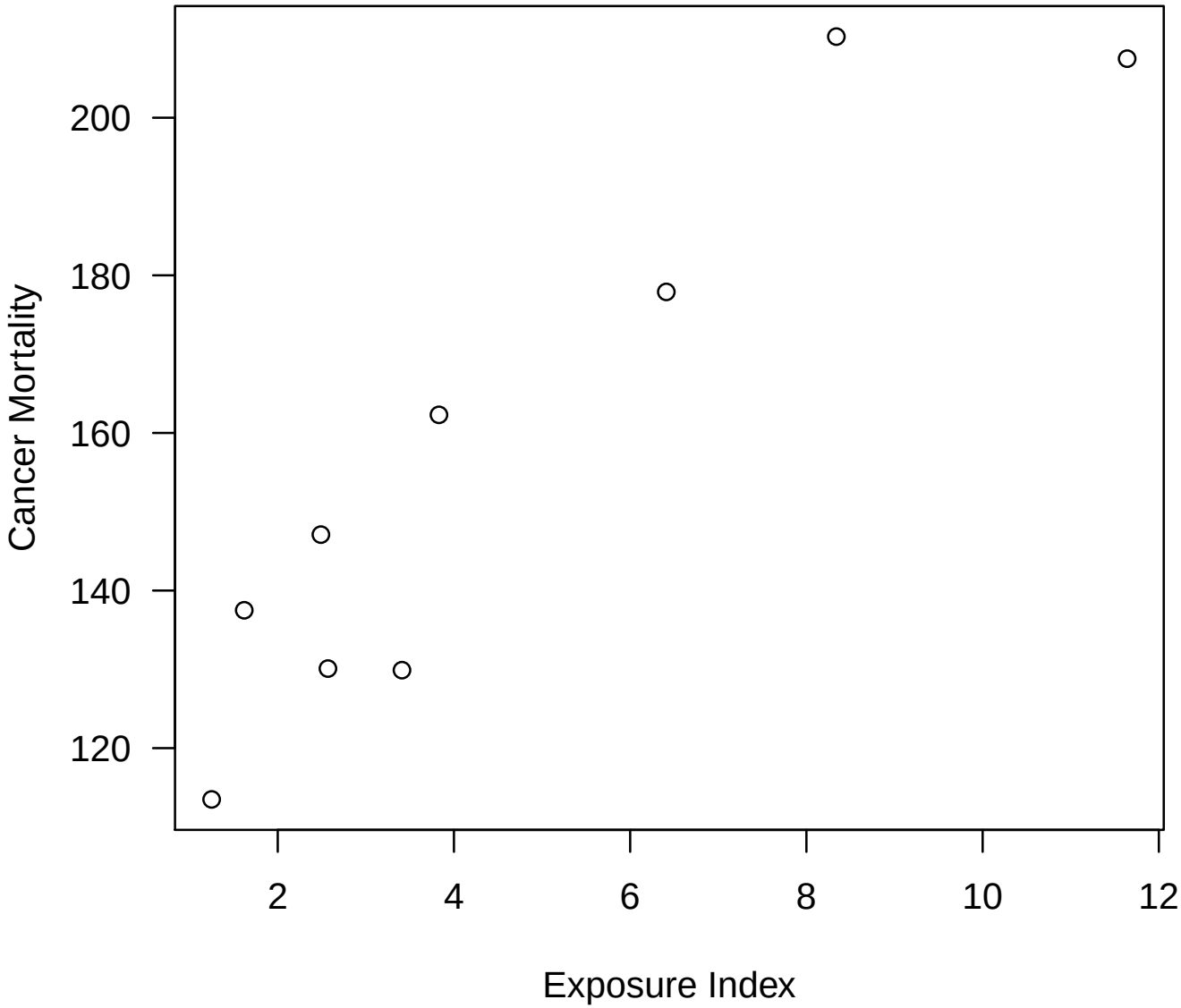


Figure: Plot of mortality per 100,0

One numerical explanatory variable

Simple linear regression model - one numeric response with one numeric explanatory.
Model:

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad e_i \stackrel{\mathrm{d}}{=} \mathrm{N}(0, \sigma)$$

Errors (e_i) **assumed** to be a random sample from $\mathrm{N}(0, \sigma)$

Predictions from the model: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

Parameters β_0 and β_1 estimated by method of least squares.

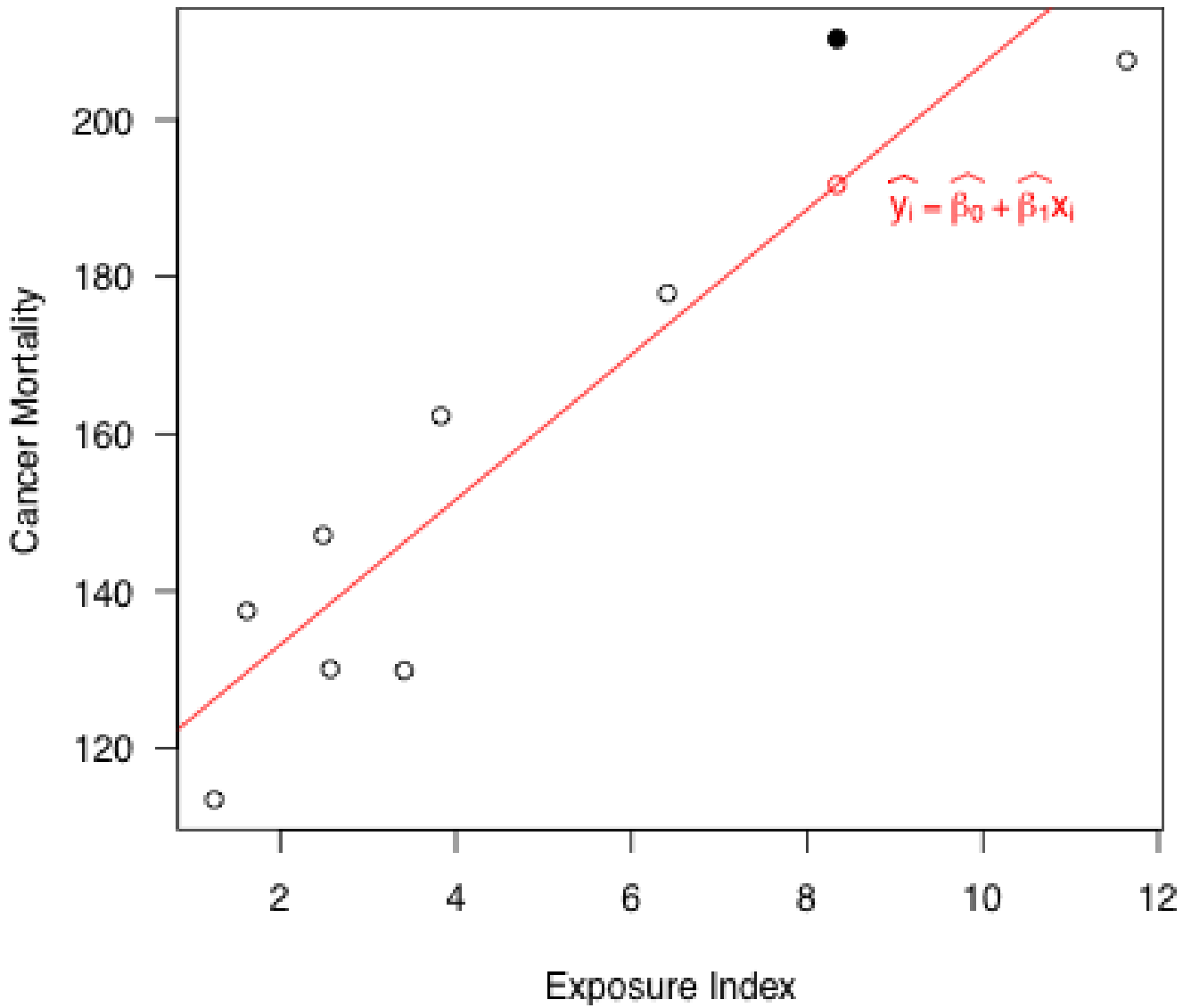


Figure: Plot of mortality per 100,000 person-years against exposure

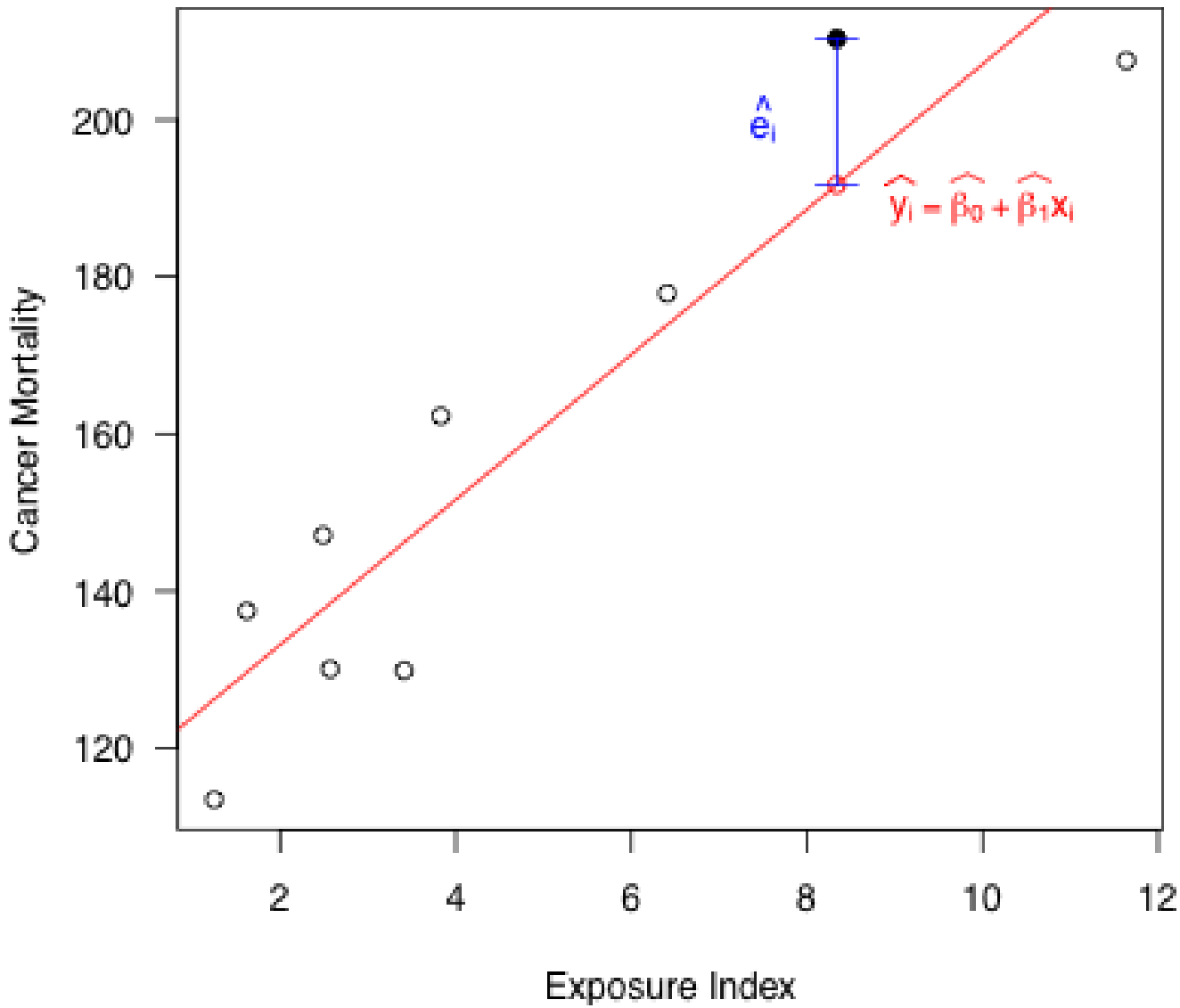


Figure: Plot of mortality per 100,000 person-years against exposure

Parameter estimation

- The parameters α, β and σ^2 can be estimated using the method of least squares. I.e. find the values of α and β that minimize the sum of squared residuals

$$SS_{res}(\alpha, \beta) = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2 \ .$$

- If we do the calculations, the values that minimize SS_{res} are

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2}$$

and

$$\hat{\alpha} = \bar{y}_n - \hat{\beta}\bar{x}_n \ .$$

Estimating σ

Model:

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad e_i \overset{\mathrm{d}}{=} \mathrm{N}(0, \sigma)$$

$$\text{residuals, } \hat{e}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i.$$

$$\hat{\sigma}^2 = \frac{\text{sum of (residuals)}^2}{\#\text{study units} - \#\text{coeff. estimated}}$$

$$= \frac{\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n-2}$$

.

Residual stdev = estimate of unexplained variability

The model and it's assumptions

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, \ldots, n$$

The "line of best fit" goes through the point $(\bar{x}, \bar{y})$, always!

Assumptions about the model

- the relationship between the predictor (x) and the outcome (y) is assumed to be linear.
- residuals have come from a normally distributed population of residuals (normality)
- there is the same variability in level of exposure index (constant variance)
- residuals represent random draws from a population (independence of observations)

Estimation*

```
> cancer.lm <- lm(Mortality ~ Exposure, data = hanford)
> summary(cancer.lm)$coefficients

Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 114.715631   8.045663 14.258070 1.984236e-06
Exposure      9.231456   1.418787  6.506584 3.320717e-04

> summary(cancer.lm)$sigma  #...gives the residual sd, s
[1] 14.00993



> options(scipen=999)
> options(digits=4)
> summary(ray.lm)$coefficients
Estimate Std. Error t value    Pr(>|t|)
(Intercept)  114.716      8.046  14.258 0.000001984
x              9.231      1.419   6.507 0.000332072
```

Steps for statistical modeling

1. Estimate
2. Check
3. Interpret
4. Make prediction

Diagnostics

Define residuals as
$\hat{e}_i = y_i - \hat{\beta}_0 - \hat{\beta} x_i$

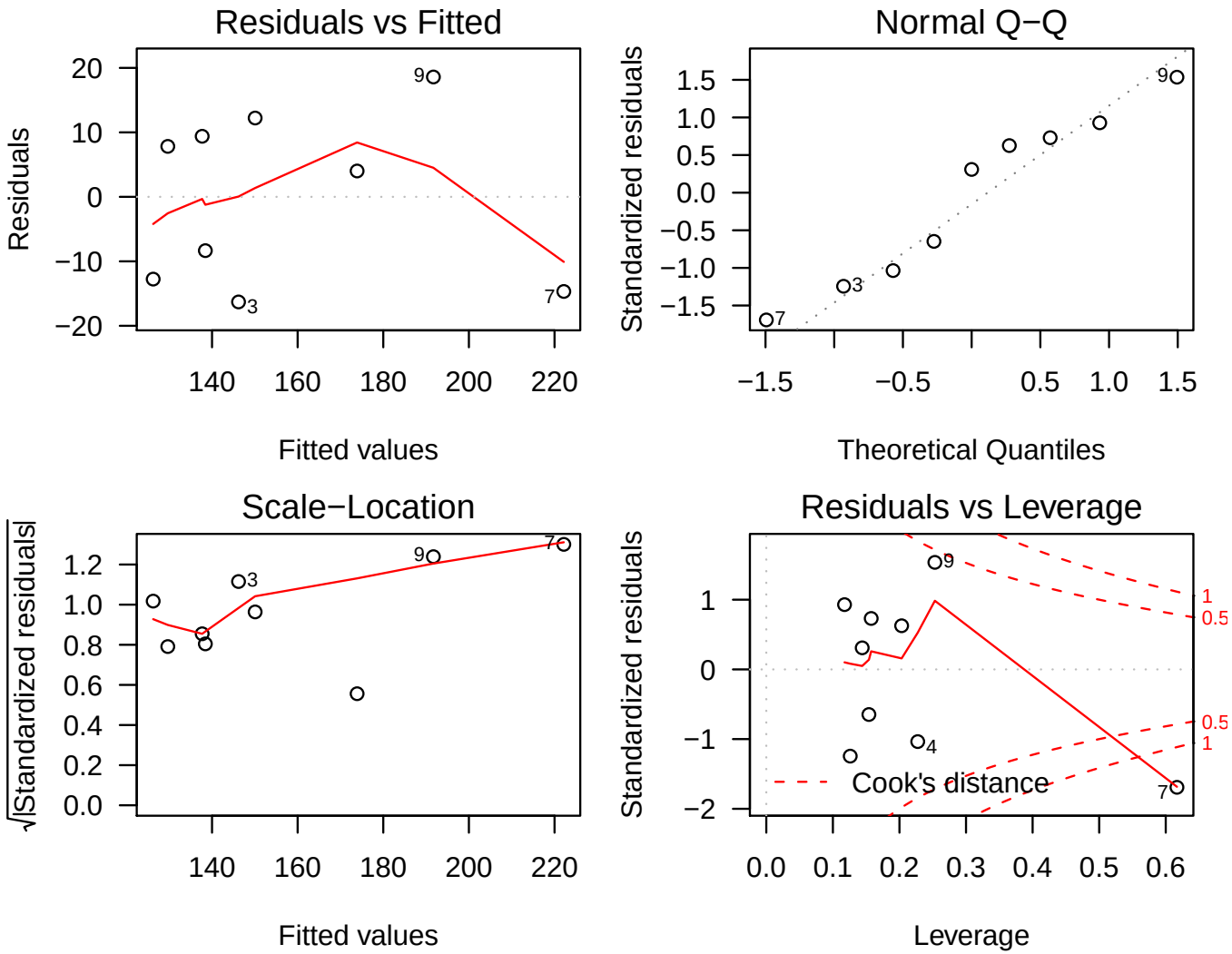
- residual is deviation of observation from fitted value in y-direction.
- These can be used to check the fit of the model:
- they should behave like the errors, e_i i.e. independent observations from $\mathrm{N}(0, \sigma)$.
- If they do not the model should be questioned.

Diagnostics

- Regression diagnostics plots in R created using the function *plot()*

```
cancer.lm <- lm(Mortality ~ Exposure, data = hanford)
par(mfrow = c(2, 2))
plot(cancer.lm)
```

Diagnostics



Diagnostics - Plots

Residuals vs fitted values

- Used to check the linear relationship assumptions.
- Any patterns indicate lack of independence i.e. non-random.
- Evidence of curvature implies non-constant variance (*heteroscedasticity*).
- Look for outliers.
- Expect 95% to be within $\pm 2s$ of the line since residuals are distributed normally.

Diagnostics - Plots

Normal QQ plot

- Quantile plot of standardised residuals against normal distribution.
- Large departures from the straight line indicate non-normality (or non-constant variance).

Standardised residual is residual divided by stdev of residual i.e. resid. standardised to have stdev=1, so that $e_i \stackrel{\mathrm{d}}{=} N(0,1)$: $e_i = \frac{y_i - \hat{y}_i}{se(y_i - \hat{y}_i)}$. 95% of e_i should be within ± 2

Diagnostics - Plots

Standardised residuals vs fitted values

- Used to check the homogeneity of residuals' variance.
- We are looking for straight line with equally spread points.
- Departures from a straight line indicate heteroscedasticity.

Diagnostics - Plots

Residuals vs leverage and Cook's distance

Leverage -

- Used to identify influential observations, that might influence parameter estimated when included or excluded from the analysis.
- These points also may reduce the R^2 value.
- Point with high leverage is extreme in x-direction.
- Points outside the contour line of greater than 1 indicate a potential problem.

Diagnostics - Plots

Residuals vs leverage and Cook's distance

Outlier -

- A point that has an extreme y value.
- Might influence parameter estimated when included or excluded from the analysis.
- Observations whose standardized residuals are > 3 or < -3 are possible outliers.

Diagnostics - Plots

Cook's distance, d_i Residuals vs leverage and Cook's distance

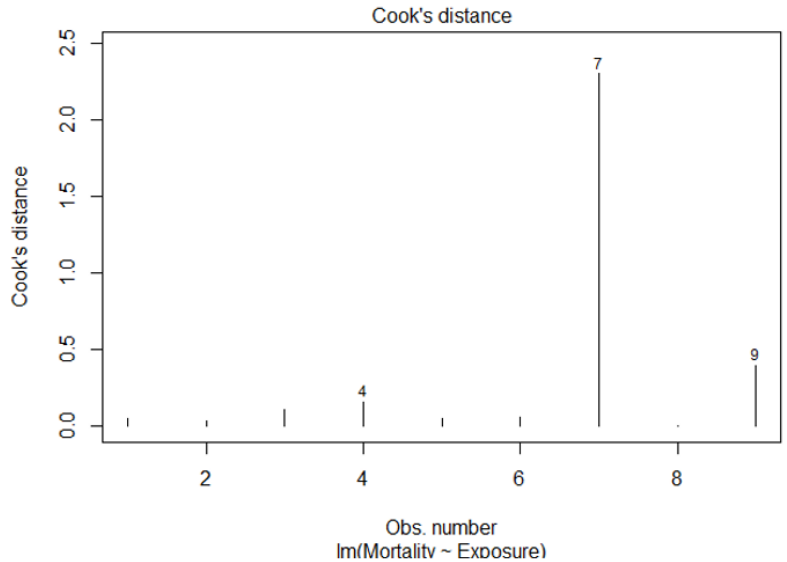
- A metric to determine the influence of an observation.
- Is a combination of leverage (deviation in x-direction) and residual (deviation in y-direction).
- Points with large Cook's distance should be examined.
- Rule of thumb: $d_i > 1$ indicates a problem.
-

Diagnostics - Plots

Cook's distance, d_i

```
> cancer.lm <- lm(Mortality~Exposure, data = hanford)
> plot(cancer.lm)
> plot(cancer.lm, 4) # Cook's dist. plot
> hanford
```

	County	Exposure	Mortality
1	Umatilla	2.49	147.1
2	Morrow	2.57	130.1
3	Gilliam	3.41	129.9
4	Sherman	1.25	113.5
5	Wasco	1.62	137.5
6	HoodRiver	3.83	162.3
7	Portland	11.64	207.5
8	Columbia	6.41	177.9
9	Clatsop	8.34	210.3



Steps for statistical modeling

1. Estimate
2. Check
3. Interpret
4. Make prediction

Estimation*

```
> cancer.lm <- lm(Mortality ~ Exposure, data = hanford)
> summary(cancer.lm)$coefficients

Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 114.715631   8.045663 14.258070 1.984236e-06
Exposure      9.231456   1.418787  6.506584 3.320717e-04

> summary(cancer.lm)$sigma  #...gives the residual sd, s
[1] 14.00993



> options(scipen=999)
> options(digits=4)
> summary(ray.lm)$coefficients
Estimate Std. Error t value    Pr(>|t|)
(Intercept)  114.716      8.046  14.258 0.000001984
x              9.231      1.419   6.507 0.000332072
```

Parameter estimate interpretation

$$\text{cancer mortality} = \beta_0 + \beta_1 \times \text{exposure index} + \text{error}$$

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad e_i \overset{\text{d}}{=} N(0, \sigma)$$

slope ($\hat{\beta}_1$): a 9.23 expected ("on average") increase in cancer mortality rate (# of deaths per 100 000 person years) as exposure index increases by 1 unit, for years between 1959-1964.

intercept ($\hat{\beta}_0$): expected cancer mortality rate ($= 114.7$) when exposure index $= 0$.

Note: Extrapolating from these data to cancer mortality outside the time range 1959 - 1964 years cannot be justified by the data; e.g. are these data relevant to (or indicative of) cancer mortality in 2016?

Confidence intervals*

95% confidence interval for the slope β_1:

$$9.23 \pm c_{0.975}(t_7) \times 1.42 \;=\; 9.23 \pm 2.365 \times 1.42 \;=\; (5.87, 12.59).$$

t distribution: 7 df = 9 observations − 2 parameters $(df = n-p)$

```
> confint(cancer.lm)
                2.5 %    97.5 %
(Intercept) 95.690661 133.74060
Exposure     5.876558  12.58635
```

R squared*

R^2: The proportion of variation in the response variable explained by the explanatory variables in the model.

Or another way: a measure of the variation explained by the model relative to the natural variation in the response values.

$$R^2 = \text{SS}_{\text{regression}}/\text{SS}_{\text{total}} \text{ (Multiple R-squared in R)}$$

```
> summary(cancer.lm)$r.squared
[1] 0.8581147
```

Alternative: $R^2_{\text{adj}} = 1 - \frac{s^2}{s^2_Y} \quad < \quad R^2,$

adjusts for number of parameters in the model (model df). Often preferred. s^2 is residual variance and s^2_y the total variance in observations. (Note that R^2 may be larger due to point of high leverage.)

Notation

$$y = M(\ ) + e$$

observations = model + error

LS $\Rightarrow$ parameter estimates to minimise residual SS
(explain as much as possible with the model)

Analysis of Variance (ANOVA)

$\downarrow$	due to	SS	MS=SS/df
$M(\ )$	model	$\text{model.SS} = \sum(\hat{y} - \bar{y})^2$	s^2_M
$+$			
e	error	$\text{res.SS} = \sum(y - \hat{y})^2$	s^2
$=$			
y	data	$\text{total.SS} = \sum(y - \bar{y})^2$	s^2_Y

R squared*

Simple linear regression (one explanatory variable)

Correlation coefficient $r = \sqrt{R^2} = \sqrt{0.8581} = 0.926$

$$r = \frac{\hat{\beta} S_x}{S_y}$$

```
> cor(cancer)
```

	County	Exposure	Mortality
County	1	NA	NA
Exposure	NA	1.0000000	0.9263448
Mortality	NA	0.9263448	1.0000000

Making predictions

What is the predicted cancer mortality rate for another county in Oregon with an exposure index of 5.3?

$114.7 + 9.231 \times 5.3 = 163.6$ (deaths per 100 000 person-years).

But what is the uncertainty around this prediction?

There are two sorts of intervals, depending on what we are trying to predict:

(1) A **confidence interval** for the mean of y for specified x

$$\text{est} \pm \text{“2”}\,\text{se} \qquad \text{est} = \hat{\beta}_0 + \hat{\beta}_1 x,\ \text{“2”} = \text{t}_{n-2}\text{-quantile}.$$

(2) A **prediction interval** for an observation y for specified x

$$\text{est} \pm \text{“2”}\,\sqrt{\text{se}^2 + s^2}$$

se(fit) = error *in* line; how line changes from sample-to-sample,

s = error *about* line; how individs vary about the line.

Making predictions*

In R:

```
> predict(cancer.lm,
+         newdata = data.frame(Exposure=5.3),
+         interval = "confidence")
       fit      lwr      upr
1 163.6423 152.3649 174.9198
> predict(cancer.lm,
+         newdata = data.frame(Exposure=5.3),
+         interval = "prediction")
       fit      lwr      upr
1 163.6423 128.6472 198.6375
```

The prediction interval is always much wider.
The CI goes to zero as $n \to \infty$; the PI does not.
Both get wider as x moves away from $\bar{x}$.

One categorical explanatory variable

1. Models for Categorical Predictors
2. Hypothesis testing, F-test
3. Point & Interval estimation

One categorical explanatory variable

COLOURS ATTRACTING BUGS:

An experiment to examine how effective various colours were in attracting cereal leaf beetles to coloured boards in an oat field.

Colour	Beetles trapped					
Yellow	45	59	48	46	38	47
White	21	12	14	17	13	17
Green	37	32	15	25	39	41
Blue	16	11	20	21	14	7

One categorical explanatory variable

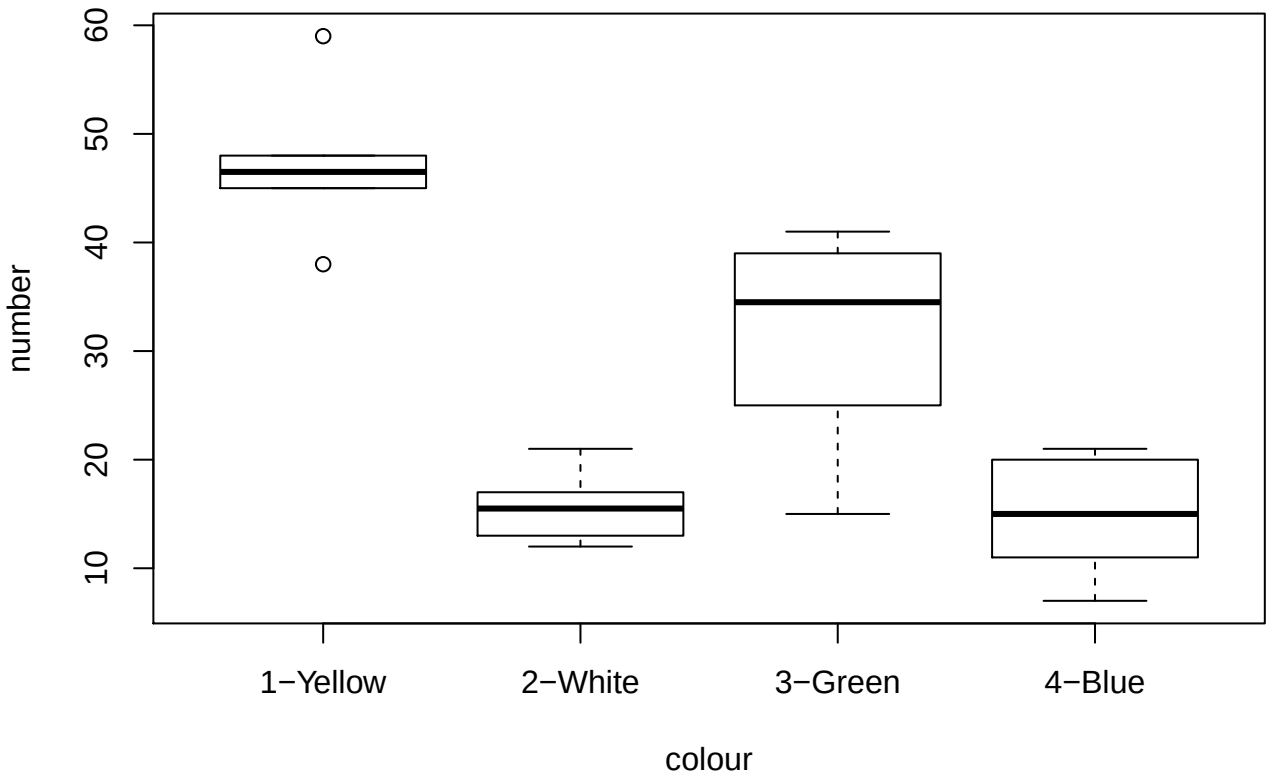
COLOURS ATTRACTING BUGS:

```
> bugs <- data.frame(colour = c(rep("1-Yellow",6),
+ rep("2-White",6),rep("3-Green",6),rep("4-Blue",6)),
+      number = c(45,59,48,46,38,47,21,12,14,17,13,17,
+                 37,32,15,25,39,41,16,11,20,21,14,7)

plot(bugs)
```

One categorical explanatory variable

COLOURS ATTRACTING BUGS:



Models for Categorical Predictors

$$y_{ij} = \mu_i + e_{ij}; \quad e_{ij} \overset{\mathrm{d}}{=} \mathrm{N}(0, \sigma) \quad i = Y, W, G, B; \quad j = 1, \ldots, 6$$

parameter list $\theta = (\mu_Y, \mu_W, \mu_G, \mu_B)$
R alternative parameter list: $\theta_R = (\beta_0, \beta_1, \beta_2, \beta_3)$

where $\beta_0 = \mu_Y$, $\beta_1 = \mu_W - \mu_Y$, $\beta_2 = \mu_G - \mu_Y$, $\beta_3 = \mu_B - \mu_Y$.
OR

$$y_{ij} = \beta_1 + \beta_i + e_{ij}; \quad e_{ij} \sim \mathrm{N}(0, \sigma) \quad i = 2, 3, 4; \quad j = 1, \ldots, 6$$

(A different parameterisation)

β_1 = mean for Yellow (the "baseline" level)

β_2 = mean for White − mean for Yellow

β_3 = mean for Green − mean for Yellow

β_4 = mean for Blue − mean for Yellow

Hypothesis Testing

Null Model:

H_0: $y_{ij} = \mu + e_{ij}$, $e_{ij} \overset{\mathrm{d}}{=} \mathrm{N}(0, \sigma)$; i.e. H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ $(= \mu)$

Alternative Model:

H_1: $y_{ij} = \mu_i + e_{ij}$; $e_{ij} \overset{\mathrm{d}}{=} \mathrm{N}(0, \sigma)$

H_0 is a particular case of H_1; the null model is *nested* within the (more general) alternative model.

The F-Test

Under H_0, the test statistic F follows an F-distribution with $(t-1, n-t)$ df.

Colours attracting bugs:

$t = 4$ colours; $n = 24$ observations

Under H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$, F statistic $\overset{d}{=} F_{3,20}$.

```
> bugs.lm <- lm(number ~ colour, data = bugs)
> summary(bugs.lm)
```

The F-Test

```
Call:
lm(formula = number ~ colour, data = bugs)

Residuals:
Min       1Q   Median       3Q      Max
-16.5000  -2.9167   0.1667   5.2083  11.8333

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)     47.167      2.770  17.030 2.27e-13 ***
colour2-White  -31.500      3.917  -8.042 1.07e-07 ***
colour3-Green  -15.667      3.917  -4.000 0.000704 ***
colour4-Blue   -32.333      3.917  -8.255 7.16e-08 ***
---
Signif. codes:  0 â***â 0.001 â**â 0.01 â*â 0.05 â.â 0.1 â â 1

Residual standard error: 6.784 on 20 degrees of freedom
Multiple R-squared:  0.8209,Adjusted R-squared:  0.794
F-statistic: 30.55 on 3 and 20 DF,  p-value: 1.151e-07
```