Lecture 2: Linear regression and ANOVA in R

Recap: regression basics (STAT0006/2002 or equivalent)

Data: $\{Y_i, x_{1i}, x_{2i}, \ldots, x_{pi} : i = 1, \ldots, n\}$

- Y is response variable; x's are covariates

Recap: regression basics (STAT0006/2002 or equivalent)

Data: $\{Y_i, x_{1i}, x_{2i}, \ldots, x_{pi} : i = 1, \ldots, n\}$

- Y is response variable; x's are covariates

Model: $Y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \varepsilon_i\ , \qquad \varepsilon_i \sim N(0, \sigma^2)$

Recap: regression basics (STAT0006/2002 or equivalent)

Data: $\{Y_i, x_{1i}, x_{2i}, \ldots, x_{pi} : i = 1, \ldots, n\}$

- Y is response variable; x's are covariates

Model: $Y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \varepsilon_i \,, \qquad \varepsilon_i \sim N(0, \sigma^2)$

- Linear relationship between response and covariates

UCL

Regression Models

Recap: regression basics (STAT0006/2002 or equivalent)

Data: $\{Y_i, x_{1i}, x_{2i}, \ldots, x_{pi} : i = 1, \ldots, n\}$

- Y is response variable; x's are covariates

Model: $Y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$

- Linear relationship between response and covariates
- 'Errors' $\{\varepsilon_i\}$ are independent and normally distributed, with zero mean and constant variance

Regression Models

Recap: regression basics (STAT0006/2002 or equivalent)

Data: $\{Y_i, x_{1i}, x_{2i}, \ldots, x_{pi} : i = 1, \ldots, n\}$

- Y is response variable; x's are covariates

Model: $Y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$

- Linear relationship between response and covariates
- 'Errors' $\{\varepsilon_i\}$ are independent and normally distributed, with zero mean and constant variance
- β_0 is intercept (expected response when all covariates are zero)

Recap: regression basics (STAT0006/2002 or equivalent)

Data: $\{Y_i, x_{1i}, x_{2i}, \ldots, x_{pi} : i = 1, \ldots, n\}$

- Y is response variable; x's are covariates

Model: $Y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \varepsilon_i \;, \qquad \varepsilon_i \sim N(0, \sigma^2)$

- Linear relationship between response and covariates
- 'Errors' $\{\varepsilon_i\}$ are independent and normally distributed, with zero mean and constant variance
- β_0 is intercept (expected response when all covariates are zero)
- β_j is regression coefficient (expected change in response when jth covariate increases by 1 unit)

Estimation in regression models

- Regression coefficients $\beta_0, \beta_1, \ldots, \beta_p$ via least squares:
  - Choose parameter estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ to minimise sum of squares

$$SS = \sum_{i=1}^{n} \left(Y_i - [\beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{ip}]\right)^2$$

Estimation in regression models

- Regression coefficients $\beta_0, \beta_1, \ldots, \beta_p$ via least squares:
  - Choose parameter estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ to minimise sum of squares

$$SS = \sum_{i=1}^{n} (Y_i - [\beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{ip}])^2$$

  - Minimised value is residual sum of squares (RSS)

Estimation in regression models

- Regression coefficients $\beta_0, \beta_1, \ldots, \beta_p$ via least squares:
  - Choose parameter estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ to minimise sum of squares

$$SS = \sum_{i=1}^{n} (Y_i - [\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}])^2$$

  - Minimised value is residual sum of squares (RSS)

- Error variance σ^2 as $\hat{\sigma}^2 = RSS/(n-k)$ where $k = p + 1$
  - $n - k$ is residual degrees of freedom (# of observations minus # of coefficients estimated)

Regression in R

- Command is `lm` ('linear model')
- Model structure specified using a formula:

R formula notation

Generic form:

```
Y ~ x1 + x2 + x3
```

- `Y` is the response variable
- `x1`, `x2`, `x3` are the covariates

- So, to regress `Y` on `x1`, `x2` and `x3`:

```
lm(Y ~ x1 + x2 + x3)
```

Example: US Minimum Temperature dataset

- Contains average daily minimum temperature in January at 56 US cities over the time period 1931-1960.
- Structure: city name, temperature, latitude (°W of Greenwich) and longitude (°N of the Equator):

```
> head(ustemp)
             city min.temp latitude longitude
1      Mobile, AL       44     31.2      88.5
2  Montgomery, AL       38     32.9      86.8
3     Phoenix, AZ       35     33.6     112.5
4 Little Rock, AR       31     35.4      92.8
5 Los Angeles, CA       47     34.3     118.7
6 San Francisco, CA     42     38.4     123.0
```

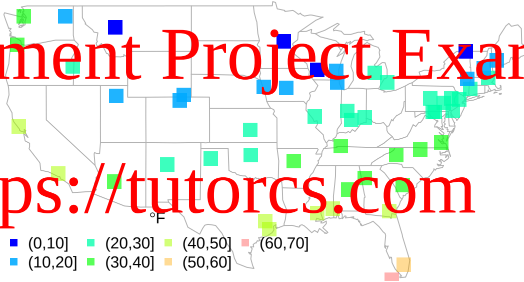
- Temperature is measured in degrees Fahrenheit.

Thermometer image from

https://www.weather-station-products.co.uk

US temperatures: visualising the data

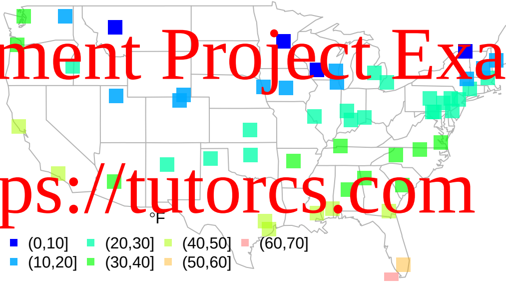
Mean US January daily minimum temperature, 1931−1960



- Colour scale divides temperature range into equal intervals (**NB** careful choice of colours — see comments in script during workshop).

US temperatures: visualising the data

Mean US January daily minimum temperature, 1931−1960



- Colour scale divides temperature range into equal intervals (**NB** careful choice of colours — see comments in script during workshop).
- Shows roughly linear (planar) variation of minimum temperature with latitude and longitude, of the form $\text{Temperature} \approx \beta_0 + (\beta_1 \times \text{Latitude}) + (\beta_2 \times \text{Longitude})$

US temperatures: fitting a linear model


```
> Temp.Model1 <- lm(min.temp ~ latitude + longitude, data=ustemp)
> print(Temp.Model1)

Call:
lm(formula = min.temp ~ latitude + longitude, data = ustemp)

Coefficients:
(Intercept)     latitude    longitude
     98.645       -2.164        0.134
```


- **Note** that `Temp.Model1` is an object of class `lm`, hence `print()` produces useful information about the fitted model (object classes were covered in Workshop 1).

Matrix representation of the linear regression model

Model for all observations can equivalently be written in matrix form as

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

or

$$\underbrace{\mathbf{Y}}_{n\times 1} = \underbrace{\mathbf{X}}_{n\times(p+1)} \underbrace{\boldsymbol{\beta}}_{(p+1)\times 1} + \underbrace{\boldsymbol{\varepsilon}}_{n\times 1}$$

X is the design matrix for the model

Inference about the coefficients

- Least-squares estimator of coefficient vector β can be written as $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.

Inference about the coefficients

- Least-squares estimator of coefficient vector β can be written as $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.

- Under model assumptions, $\hat{\beta}$ has multivariate normal distribution with mean vector β and covariance matrix $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.
  - Standard errors of coefficient estimators are square roots of diagonal elements of covariance matrix
  - Need to replace σ^2 with its estimate $\hat{\sigma}^2 = RSS/(n-k)$ where k is # of coefficients estimated

Inference about the coefficients

- Least-squares estimator of coefficient vector β can be written as $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.

- Under model assumptions, $\hat{\beta}$ has multivariate normal distribution with mean vector β and covariance matrix $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.
  - Standard errors of coefficient estimators are square roots of diagonal elements of covariance matrix

  - Need to replace σ^2 with its estimate $\hat{\sigma}^2 = RSS/(n-k)$ where k is # of coefficients estimated
- To test $H_0 : \beta_j = 0$, use test statistic $t_j = \hat{\beta}_j/\hat{\text{s.e.}}(\beta_j)$ and compare with t distribution on $n-k$ degrees of freedom.

  - **NB** beware collinearity: if covariates are highly correlated then test results can be sensitive to which other covariates are in the model.

Inference about the coefficients

- Least-squares estimator of coefficient vector β can be written as $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.
- Under model assumptions, $\hat{\beta}$ has multivariate normal distribution with mean vector β and covariance matrix $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.
  - Standard errors of coefficient estimators are square roots of diagonal elements of covariance matrix
  - Need to replace σ^2 with its estimate $\hat{\sigma}^2 = RSS/(n-k)$ where k is # of coefficients estimated
- To test $H_0 : \beta_j = 0$, use test statistic $t_j = \hat{\beta}_j / \hat{\text{s.e.}}(\beta_j)$ and compare with t-distribution on $n-k$ degrees of freedom.
  - **NB** beware collinearity: if covariates are highly correlated then test results can be sensitive to which other covariates are in the model.
- Can also use to derive confidence intervals for individual coefficients (command `confint()`), confidence intervals for the regression line and prediction intervals for future observations (both using command `predict()`) — details in workshop.

Assessing the explanatory power of the model

- Explanatory power often measured using coefficient of determination or proportion of variance explained:

$$R^2 = 1 - \frac{RSS}{\text{Total Sum of Squares}}$$

where Total Sum of Squares $= \sum_{i=1}^{n}(Y_i - \overline{Y})^2$.

  - With a single covariate, R^2 is the square of the Pearson correlation coefficient.

UCL

Assessing the explanatory power of the model

- Explanatory power often measured using coefficient of determination or proportion of variance explained:

$$R^2 = 1 - \frac{RSS}{\text{Total Sum of Squares}}$$

where Total Sum of Squares $= \sum_{i=1}^{n} (Y_i - \bar{Y})^2$.

  - With a single covariate, R^2 is the square of the Pearson correlation coefficient.

- R^2 can be increased indefinitely by including more and more covariates: better to adjust for number of covariates in the model and use adjusted R^2:

$$R^2_{ADJ} = 1 - \frac{(RSS)\,/\,(n-p)}{(\text{Total Sum of Squares})\,/\,(n-1)}\ .$$

Inference and explanatory power in R

- Use `summary()` command on an object of class `lm`.

Example: linear model for US temperature


```
> summary(Temp.Model1)
 [snip snip]
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 98.64523    8.32708  11.846   <2e-16 ***
latitude    -2.16355    0.17570 -12.314   <2e-16 ***
longitude    0.13396    0.06314   2.122   0.0386 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 6.935 on 53 degrees of freedom
Multiple R-squared:  0.7411,  Adjusted R-squared:  0.7314
F-statistic: 75.88 on 2 and 53 DF,  p-value: 2.792e-16
```


- ‘Residual standard error’ is estimate of σ
- 53 degrees of freedom is $n - k$ (56 observations, 3 coefficients estimated)

Interactions in regression models

- With two covariates, suppose $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$.

Interactions in regression models

- With two covariates, suppose $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$.
- Suppose also that x_2 modulates effect of x_1: $\beta_1 = \gamma_0 + \gamma_1 x_{i2}$ (**NB** should therefore write β_{1i}). Then

Interactions in regression models

- With two covariates, suppose $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$.
- Suppose also that x_2 modulates effect of x_1: $\beta_1 = \gamma_0 + \gamma_1 x_{i2}$ (**NB** should therefore write β_{1i}). Then

$$
\begin{aligned}
Y_i &= \beta_0 + (\gamma_0 + \gamma_1 x_{i2}) x_{i1} + \beta_2 x_{i2} + \varepsilon_i \\
&= \beta_0 + \gamma_0 x_{i1} + \beta_2 x_{i2} + \gamma_1 x_{i1} x_{i2} + \varepsilon_i .
\end{aligned}
$$

- With two covariates, suppose $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$.
- Suppose also that x_2 modulates effect of x_1: $\beta_1 = \gamma_0 + \gamma_1 x_{i2}$ (**NB** should therefore write β_{1i}). Then

$$
\begin{aligned}
Y_i &= \beta_0 + (\gamma_0 + \gamma_1 x_{i2}) x_{i1} + \beta_2 x_{i2} + \varepsilon_i \\
&= \beta_0 + \gamma_0 x_{i1} + \beta_2 x_{i2} + \gamma_1 x_{i1} x_{i2} + \varepsilon_i .
\end{aligned}
$$

- Easily handled: just define extra covariate $x_1 x_2$.

Interactions in regression models

- With two covariates, suppose $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$.
- Suppose also that x_2 modulates effect of x_1: $\beta_1 = \gamma_0 + \gamma_1 x_{i2}$ (**NB** should therefore write β_{1i}). Then

$$
\begin{aligned}
Y_i &= \beta_0 + (\gamma_0 + \gamma_1 x_{i2}) x_{i1} + \beta_2 x_{i2} + \varepsilon_i \\
&= \beta_0 + \gamma_0 x_{i1} + \beta_2 x_{i2} + \gamma_1 x_{i1} x_{i2} + \varepsilon_i .
\end{aligned}
$$

- Easily handled: just define extra covariate $x_1 x_2$.

Defining interactions in R

Use "`:`" in model formula e.g.

```
lm(y ~ x1 + x2 + x1:x2)
```

Interactions in regression models

- With two covariates, suppose $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$.
- Suppose also that x_2 modulates effect of x_1: $\beta_1 = \gamma_0 + \gamma_1 x_{i2}$ (**NB** should therefore write β_{1i}). Then

$$
\begin{aligned}
Y_i &= \beta_0 + (\gamma_0 + \gamma_1 x_{i2}) x_{i1} + \beta_2 x_{i2} + \varepsilon_i \\
&= \beta_0 + \gamma_0 x_{i1} + \beta_2 x_{i2} + \gamma_1 x_{i1} x_{i2} + \varepsilon_i .
\end{aligned}
$$

- Easily handled: just define extra covariate $x_1 x_2$.

Defining interactions in R

Use ":" in model formula e.g.

```
lm(y ~ x1 + x2 + x1:x2)
```

- **Interpretation of interaction**: slope of (x_1, y) relationship depends on value of x_2 (or vice versa)
- **NB**: usually, if a model includes an interaction `x1:x2` then corresponding main effects `x1` and `x2` should also be included

UCL

Comparing models

- Often want to compare two nested models, $\mathcal{M}_0$ & $\mathcal{M}_1$ say, where $\mathcal{M}_0$ is a special case of $\mathcal{M}_1$ (e.g. obtained by dropping one or more terms)
  - Equivalently, where $\mathcal{M}_1$ is obtained by adding m extra terms to $\mathcal{M}_0$

Comparing models

- Often want to compare two nested models, $\mathcal{M}_0$ & $\mathcal{M}_1$ say, where $\mathcal{M}_0$ is a special case of $\mathcal{M}_1$ (e.g. obtained by dropping one or more terms)
  - Equivalently, where $\mathcal{M}_1$ is obtained by adding m extra terms to $\mathcal{M}_0$
- Can test hypothesis H_0: data were generated from $\mathcal{M}_0$ using F-statistic:

$$F = \frac{(RSS_0 - RSS_1)/m}{RSS_1/(n-k)}$$

where k is # of coefficients estimated in $\mathcal{M}_1$

  - Test statistic has F distribution under H_0
  - Use `anova()` command in R e.g. `anova(model1, model2, test="F")`

Comparing models

- Often want to compare two nested models, $\mathcal{M}_0$ & $\mathcal{M}_1$ say, where $\mathcal{M}_0$ is a special case of $\mathcal{M}_1$ (e.g. obtained by dropping one or more terms)
  - Equivalently, where $\mathcal{M}_1$ is obtained by adding m extra terms to $\mathcal{M}_0$
- Can test hypothesis H_0: data were generated from $\mathcal{M}_0$ using F-statistic:

$$F = \frac{(RSS_0 - RSS_1)/m}{RSS_1/(n-k)}$$

  where k is # of coefficients estimated in $\mathcal{M}_1$
  - Test statistic has F distribution under H_0
  - Use `anova()` command in R e.g. `anova(model1, model2, test="F")`

- **NB** if just one term is dropped ($m = 1$), gives same result as t-test for corresponding coefficient in $\mathcal{M}_1$
- **NB also**: models must be fitted to identical response data

Extending the US temperature model

- Exploratory analysis indicated that temperature variation isn't exactly linear — maybe better described by a quadratic function of latitude and longitude?

Extending the US temperature model

- Exploratory analysis indicated that temperature variation isn't exactly linear — maybe better described by a quadratic function of latitude and longitude?
- Quadratic function of two variables x_1 and x_2 has terms in x_1, x_2, x_1^2, x_2^2 and $x_1 x_2$
  - $x_1 x_2$ term is just an interaction!

Extending the US temperature model

- Exploratory analysis indicated that temperature variation isn't exactly linear — maybe better described by a quadratic function of latitude and longitude?
- Quadratic function of two variables x_1 and x_2 has terms in x_1, x_2, x_1^2, x_2^2 and $x_1 x_2$
  - $x_1 x_2$ term is just an interaction!

Specimen R code (more in workshop)

```
> Temp.Model2 <-
+       update(Temp.Model1, . ~ .
+                  + I(latitude^2) + I(longitude^2)
+                  + latitude:longitude)
> anova(Temp.Model1,Temp.Model2,test="F")
 [snip snip]
Model 1: min.temp ~ latitude + longitude
Model 2: min.temp ~ latitude + longitude + I(latitude^2) + ...
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     53 2548.65
2     50  830.72  3    1717.9 34.467 3.197e-12 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Model checking

- Recall model assumptions: errors $\{\varepsilon_i\}$ have zero mean, constant variance, normal distribution and are independent

UCL

Model checking

- Recall model assumptions: errors $\{\varepsilon_i\}$ have zero mean, constant variance, normal distribution and are independent
- Estimate errors using residuals $\{e_i : i = 1, \ldots, n\}$, where
  $e_i = y_i - \left[\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_p x_{ip}\right]$

Model checking

- Recall model assumptions: errors $\{\varepsilon_i\}$ have zero mean, constant variance, normal distribution and are independent
- Estimate errors using residuals $\{e_i : i = 1, \ldots, n\}$, where $e_i = Y_i - \left[\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_p x_{ip}\right]$
- When $n \gg k$, residuals should behave like errors $\Rightarrow$ use plots of residuals to assess validity of assumptions:
  1. Plot residuals against fitted values: should show random scatter about zero with no systematic structure in either mean or variability
  2. Scale-location plot: shows absolute values of residuals against fitted values (easier to see non-constant variance)
  3. Normal Q-Q plot to assess normality assumption
  4. May add smooth curve to plots to help interpretation (particularly useful for large datasets — and bear in mind that the curve will be subject to sampling variability so you shouldn't overinterpret it)

UCL

Model checking

- Recall model assumptions: errors $\{\varepsilon_i\}$ have zero mean, constant variance, normal distribution and are independent

- Estimate errors using residuals $\{e_i : i = 1, \dots, n\}$, where
$$e_i = Y_i - \left[\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_p x_{ip}\right]$$

- When $n \gg k$, residuals should behave like errors $\Rightarrow$ use plots of residuals to assess validity of assumptions:

  1. Plot residuals against fitted values: should show random scatter about zero with no systematic structure in either mean or variability
  2. Scale-location plot: shows absolute values of residuals against fitted values (easier to see non-constant variance)
  3. Normal Q-Q plot to assess normality assumption
  4. May add smooth curve to plots to help interpretation (particularly useful for large datasets — and bear in mind that the curve will be subject to sampling variability so you shouldn't overinterpret it)

- Independence: context-dependent e.g. unlikely to hold for data collected at successive time points

Consequences of assumption failure

- **Systematic structure in mean residuals**: structural component of model is inadequate (does it matter? Depends on application and magnitude of problem)

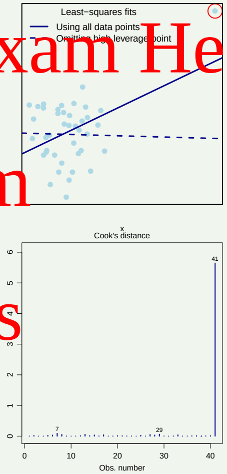
- **Non-constant variance**: inefficient estimation (doesn't make best use of data), incorrect standard errors etc.
- **Non-normal errors**: not critical in large samples except when computing prediction intervals (see later)

- **Lack of independence**: incorrect standard errors etc.

Another pitfall: influential observations

- Outlying observations can potentially have big impact on least-squares estimates:
  - Outliers in y-direction (large residuals) 'pull' regression line towards them
  - Outliers in x-direction ('leverage points') encourage regression line to pass close to them

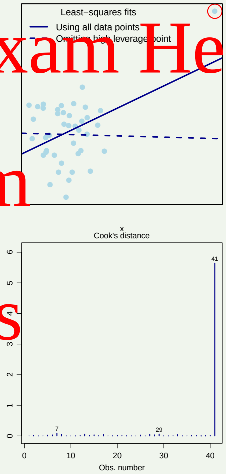
Illustration



Another pitfall: influential observations

- Outlying observations can potentially have big impact on least-squares estimates:
  - Outliers in y-direction (large residuals) 'pull' regression line towards them
  - Outliers in x-direction ('leverage points') encourage regression line to pass close to them
- Influence of each observation measured by Cook's distance
  - Measure of change in least-squares estimates if observation were omitted
  - Rule-of-thumb: observation influential if Cook's distance exceeds $8/(n-2k)$ where k is # of coefficients estimated

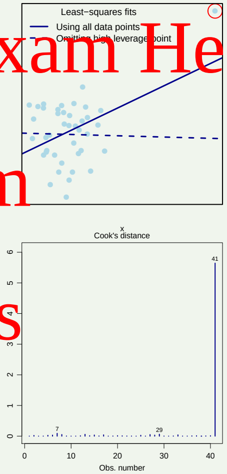
Illustration



Another pitfall: influential observations

- Outlying observations can potentially have big impact on least-squares estimates:
  - Outliers in y-direction (large residuals) 'pull' regression line towards them
  - Outliers in x-direction ('leverage points') encourage regression line to pass close to them
- Influence of each observation measured by Cook's distance
  - Measure of change in least-squares estimates if observation were omitted
  - Rule-of-thumb: observation influential if Cook's distance exceeds $8/(n-2k)$ where k is # of coefficients estimated
- Options for resolving problems: see workshop

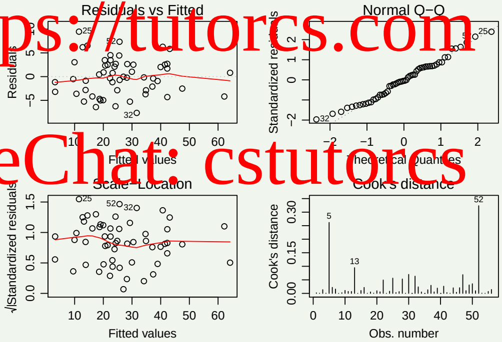
Illustration



- `plot()` for an object of class `lm` generates variety of residual plots

Diagnostics for quadratic US temperature model

```
> par(mfrow=c(2,2),mar=c(3,3,2,2),mgp=c(2,0.75,0))
> plot(Temp.Model2,which=1:4)
```

- Regression models represent variation of response with continuous covariates: what about factor covariates i.e. variation of response in different groups?

Example: petal lengths in the iris data (Workshop 1)

- **Response variable:** petal length

- **Covariate:** species (`Setosa`, `Versicolor` or `Virginica`)

ANOVA models

- Regression models represent variation of response with continuous covariates: what about factor covariates i.e. variation of response in different groups?

Example: petal lengths in the iris data (Workshop 1)

- **Response variable:** petal length
- **Covariate:** species (`Setosa`, `Versicolor` or `Virginica`)

- 'Classical' approach: Analysis of Variance (ANOVA)
  - Decompose variation into 'between-groups' and 'within-groups' sum of squares by computing overall mean and group means
  - *F*-test of null hypothesis of equal mean responses in each group
  - Presented in ANOVA table: sums of squares, mean squares, degrees of freedom etc.

- Command for classical ANOVA is `aov` — use `summary()` to get classical ANOVA table

Example: petal lengths for the iris species


```
> data(iris)
> summary(aov(Petal.Length ~ Species, data=iris))
             Df Sum Sq Mean Sq F value Pr(>F)
Species       2  437.1  218.55    1180 <2e-16 ***
Residuals   147   27.2    0.19
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

- Command for classical ANOVA is aov — use summary() to get classical ANOVA table

Example: petal lengths for the iris species


```
> data(iris)
> summary(aov(Petal.Length ~ Species, data=iris))
             Df Sum Sq Mean Sq F value Pr(>F)
Species       2  437.1  218.55    1180 <2e-16 ***
Residuals   147   27.2    0.19
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```


BUT · · ·

Example: petal lengths for the iris species

- Let:
  - Y_i be petal length for the ith specimen
  - x_{1i} = 1 if the ith specimen is `Setosa`, 0 otherwise
  - x_{2i} = 1 if the ith specimen is `Versicolor`, 0 otherwise
  - x_{3i} = 1 if the ith specimen is `Virginica`, 0 otherwise

Example: petal lengths for the iris species

- Let:
  - Y_i be petal length for the ith specimen
  - x_{1i} = 1 if the ith specimen is `Setosa`, 0 otherwise
  - x_{2i} = 1 if the ith specimen is `Versicolor`, 0 otherwise
  - x_{3i} = 1 if the ith specimen is `Virginica`, 0 otherwise
- Consider the linear model $Y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$ (**NB no intercept**)

ANOVA models are linear regressions!

Example: petal lengths for the iris species

- Let:
  - Y_i be petal length for the ith specimen
  - x_{1i} = 1 if the ith specimen is `Setosa`, 0 otherwise
  - x_{2i} = 1 if the ith specimen is `Versicolor`, 0 otherwise
  - x_{3i} = 1 if the ith specimen is `Virginica`, 0 otherwise
- Consider the linear model $Y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$ (**NB no intercept**)
  - For `Setosa` specimens we have $Y_i = \beta_1 + \varepsilon_i \Rightarrow \mathbb{E}(Y_i) = \beta_1$

ANOVA models are linear regressions!

Example: petal lengths for the iris species

- Let:
  - Y_i be petal length for the ith specimen
  - x_{1i} = 1 if the ith specimen is `Setosa`, 0 otherwise
  - x_{2i} = 1 if the ith specimen is `Versicolor`, 0 otherwise
  - x_{3i} = 1 if the ith specimen is `Virginica`, 0 otherwise
- Consider the linear model $Y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$ (**NB no intercept**)
  - For `Setosa` specimens we have $Y_i = \beta_1 + \varepsilon_i \Rightarrow \mathbb{E}(Y_i) = \beta_1$
  - For `Versicolor` and `Virginica` specimens we have $\mathbb{E}(Y_i) = \beta_2$ and $\mathbb{E}(Y_i) = \beta_3$ respectively.

ANOVA models are linear regressions!

Example: petal lengths for the iris species

- Let:
  - Y_i be petal length for the ith specimen
  - x_{1i} = 1 if the ith specimen is `Setosa`, 0 otherwise
  - x_{2i} = 1 if the ith specimen is `Versicolor`, 0 otherwise
  - x_{3i} = 1 if the ith specimen is `Virginica`, 0 otherwise
- Consider the linear model $Y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$ (**NB no intercept**)
  - For `Setosa` specimens we have $Y_i = \beta_1 + \varepsilon_i \Rightarrow \mathbb{E}(Y_i) = \beta_1$
  - For `Versicolor` and `Virginica` specimens we have $\mathbb{E}(Y_i) = \beta_2$ and $\mathbb{E}(Y_i) = \beta_3$ respectively.
- Least-squares coefficient estimates are sample means within each group

UCL

Example: petal lengths for the iris species

- Let:
  - Y_i be petal length for the ith specimen
  - x_{1i} = 1 if the ith specimen is `Setosa`, 0 otherwise
  - x_{2i} = 1 if the ith specimen is `Versicolor`, 0 otherwise
  - x_{3i} = 1 if the ith specimen is `Virginica`, 0 otherwise
- Consider the linear model $Y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$ (**NB no intercept**)
  - For `Setosa` specimens we have $Y_i = \beta_1 + \varepsilon_i \Rightarrow \mathbb{E}(Y_i) = \beta_1$
  - For `Versicolor` and `Virginica` specimens we have $\mathbb{E}(Y_i) = \beta_2$ and $\mathbb{E}(Y_i) = \beta_3$ respectively.
- Least-squares coefficient estimates are sample means within each group

- For a factor covariate with L levels, R generates L binary 'dummy variables' to represent effect in regression models
  - $L = 3$ for `Species` in iris data

Demonstration of equivalence for iris data

```
> tapply(iris$Petal.Length, INDEX=iris$Species, FUN=mean)
    setosa versicolor  virginica
     1.462      4.260      5.552
> lm(Petal.Length ~ Species - 1, data=iris)

Call:
lm(formula = Petal.Length ~ Species - 1, data = iris)

Coefficients:
    Speciessetosa  Speciesversicolor   Speciesvirginica
            1.462              4.260              5.552
```


- **NB** use of `-1` in formula to exclude the intercept
- Can use the `model.matrix()` command to see the design matrix (i.e. **X**) that R has used — see workshop.

ANOVA models with an intercept

Iris data: petal lengths again

```
> iris.Model1 <- lm(Petal.Length ~ Species, data=iris)
> summary(iris.Model1)
[snip snip]
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        1.46200    0.06086   24.02   <2e-16 ***
Speciesversicolor  2.79800    0.08607   32.51   <2e-16 ***
Speciesvirginica   4.09000    0.08607   47.52   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.4303 on 147 degrees of freedom
Multiple R-squared: 0.9414, Adjusted R-squared: 0.9406
F-statistic:  1180 on 2 and 147 DF,  p-value: < 2.2e-16
```


- **NB also** no coefficient for `Setosa` now — but intercept is equal to `Setosa` coefficient in previous model
- Other coefficients represent *differences* between other species and `Setosa`

ANOVA models with an intercept — what's going on?

- With intercept in model for iris data, design matrix $\mathbf{X}$ has
  - **First column**: all 1's
  - **Second column**: values of x_{i1} i.e. 1 for `Setosa`, 0 otherwise
  - **Third column**: values of x_{i2} i.e. 1 for `Versicolor`, 0 otherwise
  - **Fourth column**: values of x_{i3} i.e. 1 for `Virginica`, 0 otherwise

ANOVA models with an intercept — what's going on?

- With intercept in model for iris data, design matrix $\mathbf{X}$ has
  - **First column**: all 1's
  - **Second column**: values of x_{i1} i.e. 1 for `Setosa`, 0 otherwise
  - **Third column**: values of x_{i2} i.e. 1 for `Versicolor`, 0 otherwise
  - **Fourth column**: values of x_{i3} i.e. 1 for `Virginica`, 0 otherwise
- **NB** sum of columns 2–4 is always 1: so column 1 is linear combination of columns 2–4 & columns are linearly dependent.
  - **Consequence**: $(\mathbf{X}'\mathbf{X})$ isn't invertible $\Rightarrow$ least-squares estimate $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ isn't unique.

ANOVA models with an intercept — what's going on?

- With intercept in model for iris data, design matrix $\mathbf{X}$ has
  - **First column**: all 1's
  - **Second column**: values of x_{i1} i.e. 1 for `Setosa`, 0 otherwise
  - **Third column**: values of x_{i2} i.e. 1 for `Versicolor`, 0 otherwise
  - **Fourth column**: values of x_{i3} i.e. 1 for `Virginica`, 0 otherwise
- **NB** sum of columns 2–4 is always 1: so column 1 is linear combination of columns 2–4 & columns are linearly dependent.
  - **Consequence**: $(\mathbf{X}'\mathbf{X})$ isn't invertible $\Rightarrow$ least-squares estimate $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ isn't unique.
- **Solution to problem**: impose a constraint to remove the linear dependence
  - Default in R is to set $\beta_j = 0$ for one level of a factor — `Setosa` here

ANOVA models with an intercept — what's going on?

- With intercept in model for iris data, design matrix $\mathbf{X}$ has
  - **First column**: all 1's
  - **Second column**: values of x_{i1} i.e. 1 for `Setosa`, 0 otherwise
  - **Third column**: values of x_{i2} i.e. 1 for `Versicolor`, 0 otherwise
  - **Fourth column**: values of x_{i3} i.e. 1 for `Virginica`, 0 otherwise
- **NB** sum of columns 2–4 is always 1: so column 1 is linear combination of columns 2–4 & columns are linearly dependent.
  - **Consequence**: $(\mathbf{X}'\mathbf{X})$ isn't invertible $\Rightarrow$ least-squares estimate $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ isn't unique.
- **Solution to problem**: impose a constraint to remove the linear dependence
  - Default in R is to set $\beta_j = 0$ for one level of a factor — `Setosa` here
  - Iris model becomes $Y_{ij} = \beta_0 + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_{ij}$ so for `Setosa` we have $\mathbb{E}(Y_i) = \beta_0$, for `Versicolor` we have $\mathbb{E}(Y_i) = \beta_0 + \beta_2$ and for `Virginica` we have $\mathbb{E}(Y_i) = \beta_0 + \beta_3$.

Factor covariates: what you need to know

- To include a factor with L levels as a covariate in a linear regression model, $L-1$ coefficients will be estimated ('$L-1$ degrees of freedom')

Factor covariates: what you need to know

- To include a factor with L levels as a covariate in a linear regression model, $L-1$ coefficients will be estimated ('$L-1$ degrees of freedom')
  - By default, coefficient for 'reference level' is set to zero so that other coefficients are differences relative to this reference level (e.g. `Setosa` in iris example)

Factor covariates: what you need to know

- To include a factor with L levels as a covariate in a linear regression model, $L-1$ coefficients will be estimated ('$L-1$ degrees of freedom')
  - By default, coefficient for 'reference level' is set to zero so that other coefficients are differences relative to this reference level (e.g. `Setosa` in iris example)
  - Other constraints also possible (see workshop) — but choice of constraint does not affect the fitted values (i.e. estimates of $\mathbb{E}(Y_i)$)
  - **NB** Care required when interpreting p-values in `summary`: interpretation of coefficients depends on constraint that has been applied

Factor covariates: what you need to know

- To include a factor with L levels as a covariate in a linear regression model, $L-1$ coefficients will be estimated ('$L-1$ degrees of freedom')
  - By default, coefficient for 'reference level' is set to zero so that other coefficients are differences relative to this reference level (e.g. `Setosa` in iris example)
  - Other constraints also possible (see workshop) — but choice of constraint does not affect the fitted values (i.e. estimates of $\mathbb{E}(Y_i)$)
  - **NB** Care required when interpreting p-values in `summary`: interpretation of coefficients depends on constraint that has been applied
- Factor covariates can be included alongside continuous covariates in linear regression models

Factor covariates: what you need to know

- To include a factor with L levels as a covariate in a linear regression model, $L-1$ coefficients will be estimated ('$L-1$ degrees of freedom')
  - By default, coefficient for 'reference level' is set to zero so that other coefficients are differences relative to this reference level (e.g. `Setosa` in iris example)
  - Other constraints also possible (see workshop) — but choice of constraint does not affect the fitted values (i.e. estimates of $\mathbb{E}(Y_i)$)
  - **NB** Care required when interpreting p-values in `summary`: interpretation of coefficients depends on constraint that has been applied
- Factor covariates can be included alongside continuous covariates in linear regression models
- Interactions between factor and continuous covariates imply different regression slopes within each group (see workshop)