GEOG 5680
Introduction to R
10: Statistical modeling in R

Simon Brewer

Geography Department
University of Utah
Salt Lake City, Utah 84112

simon.brewer@geog.utah.edu

May 05, 2020

Statistical modeling

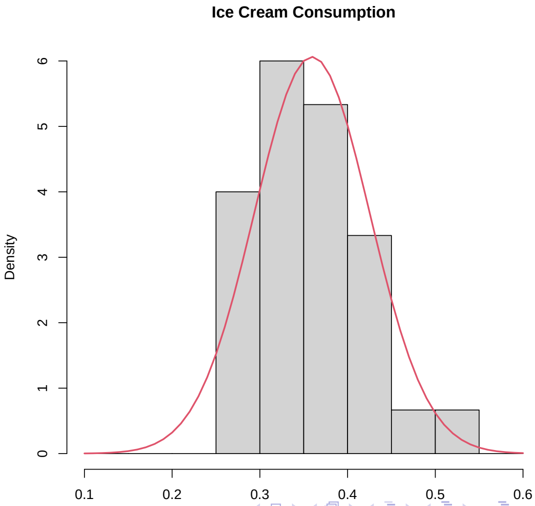
Data = predictable component + unpredictable component

$$y = f + \epsilon \tag{1}$$

- Interest in understanding the function f which explains observations (y):
    - should explain as much variation as possible
- The unpredictable part is also important:
    - should be random noise (i.e. nothing left that we can explain)
- Used for *understanding process* and *prediction*

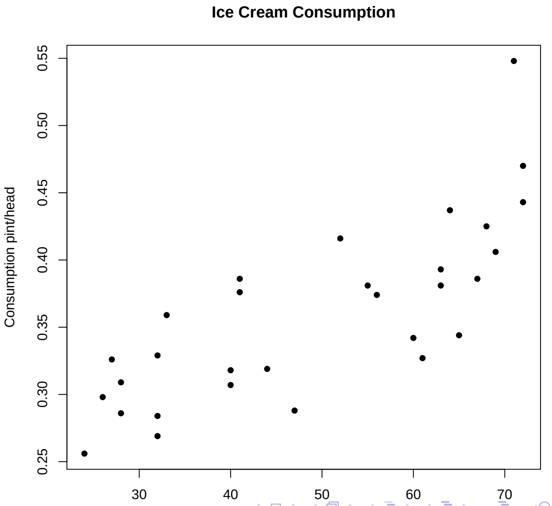
A Simple Model

- Consumption of ice cream
- Simplest model is just $E(y) = \mu$
- With unexplained variance described by a normal distribution $N(0, \sigma^2)$



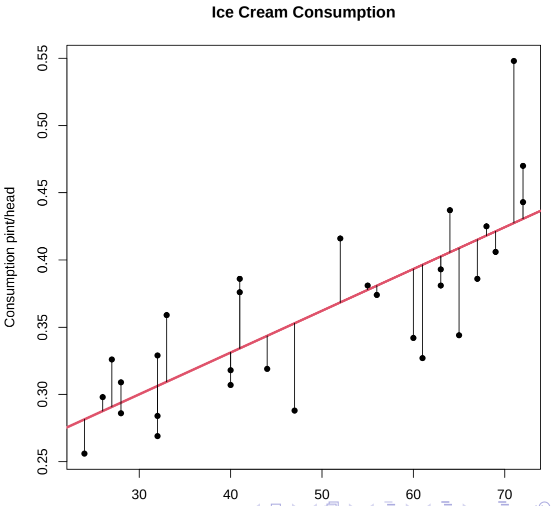
A Simple Model

- Most statistical modeling introduces independent variables
- Can we improve on simple model by introducing x?
- E.g. daily temperature
- Expected value ($E(y|x) = \beta_0 + \beta_1 x + \epsilon$)
- Where $\epsilon = N(0, \sigma^2)$



A Simple Model

- Simple linear regression fit by minimizing the sum of squares (distance between observed and modeled y)
- The slope gives the strength of the relationship (the *rate* of change)
- The intercept is expected value of y at $x = 0$



Linear models in R

R syntax — 'formula' method uses the *tilde* ($\sim$)

- Dependent variable on left, explanatory variable(s) on right: $\texttt{lm}(y \sim x_1 + x_2 \ldots)$
- Model fitting produces a model *object* as output, so create a variable to store this:

```
fit = lm(cons ~ temp, Icecream)
fit

##
## Call:
## lm(formula = cons ~ temp, data = Icecream)
##
## Coefficients:
## (Intercept)         temp
##    0.206862     0.003107
```

Centering data

- If the value of $x = 0$ is not meaningful, we can center the data by subtracting the mean from covariates
- $x_{i,cen} = x_i - \bar{x}$

```
Icecream$temp.c = Icecream$temp - mean(Icecream$temp)
fit = lm(cons ~ temp.c, Icecream)
fit

##
## Call:
## lm(formula = cons ~ temp.c, data = Icecream)
##
## Coefficients:
## (Intercept)       temp.c
##    0.359433     0.003107
```

Diagnostics

- Coefficients
- Goodness of fit
  - ANOVA
  - F-statistic
  - r-squared: variance explained
- Residuals and diagnostic plots
- These ideas can be applied to most models

R Model Diagnostics

```
summary(fit)

##
## Call:
## lm(formula = cons ~ temp.c, data = Icecream)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.069411 -0.024478 -0.007371  0.029126  0.120516
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.3594333  0.0077159  46.584  < 2e-16 ***
## temp.c      0.0031074  0.0004779   6.502 4.79e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04226 on 28 degrees of freedom
## Multiple R-squared:  0.6016,Adjusted R-squared:  0.5874
## F-statistic: 42.28 on 1 and 28 DF,  p-value: 4.789e-07
```

ANOVA

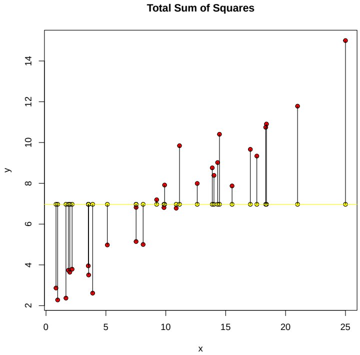
ANOVA can be used to test model goodness-of-fit

$$F = \frac{MSS/(df1)}{RSS/(df2)} \tag{2}$$

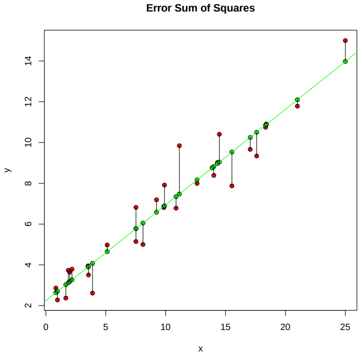
- Ratio of how much of the variance is explained by the model (MSS) to the variance in the residuals (RSS)
- Compare to an F-distribution, using degrees of freedom based on the number of parameters and the number of observations

ANOVA with a linear model

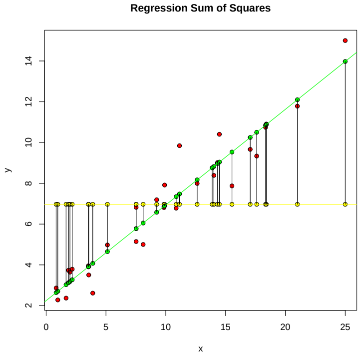
Total SS



Residual SS



Model SS



ANOVA with a linear model

```
anova(ex1.lm)

## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## x          1 283.947 283.947   368.4 < 2.2e-16 ***
## Residuals 28  21.581   0.771
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

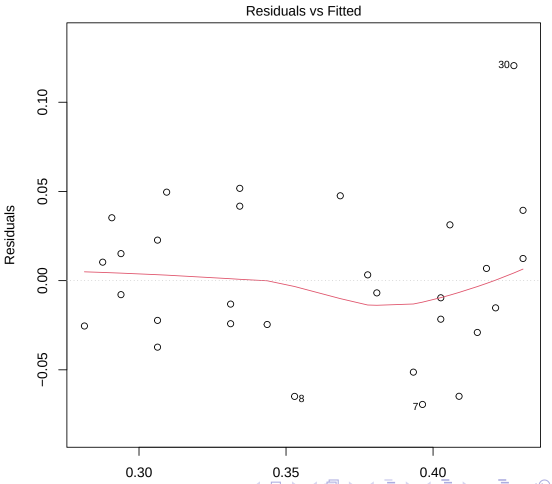
Residuals and Diagnostic Plots

- Regression function can be wrong (quadratic or other effects)
- Model for the errors may be incorrect:
  - may not be normally distributed.
  - may not be independent.
  - may not have the same variance.
- If model is correct then residuals should resemble random variables with mean $= 0$ and a normal distribution
- Detecting problems is more art then science, i.e. we cannot test for all possible problems in a regression model.

Residuals and Diagnostic Plots

- Plot of residuals vs. fitted values
- Look for *bias* in residuals to indicate a poor model fit
- Also use histograms, Q-Q plots, etc

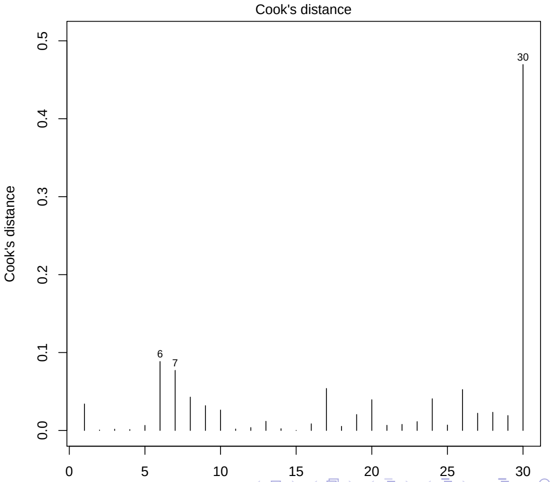
```
plot(fit, which=1)
```



Residuals and Diagnostic Plots

- Plot of Cook's distance
- Low if x_i is close to other x's
- High if x_i is distant: indicates high leverage and influence in regression

```
plot(fit, which=4)
```



Predictions

- Predicting for new values of x
- Requires new data frame containing variable(s) with the same name(s) as the independent x's used in model
- `interval` parameter estimates 95% prediction CIs

```
newtemp = data.frame(temp.c = 70 - mean(Icecream$temp))
predict(fit, newdata = newtemp, interval = "pred")

##          fit     lwr       upr
## 1 0.4243771 0.33403 0.5147241
```

Extensions to basic model

- Multiple linear regression
- Dummy variables
- Interactions
- Generalized linear modeling
  - Logistic regression
  - Poisson regression
- Many other non/semi-parametric, Bayesian and machine learning methods available