Statistics for Machine Learning and Big Data

An Introduction

Part 3: regression

Jeff Abrahamson

7 April 2015

Regression

Linear models

Problem: We have a set of points $\{(x_i, y_i)\}$. Given a new x value, we'd like to predict $\hat{y}$.

Linear model: We'll assume there exists a linear relationship $y = \beta_0 + \beta_1 x$ that offers a good approximation to the data.

Linear models

Problem: We have a set of points $\{(x_i, y_i)\}$. Given a new x value, we'd like to predict $\hat{y}$.

Linear model: We'll assume there exists a linear relationship $y = \beta_0 + \beta_1 x$ that offers a good approximation to the data.

We call x the **explanatory** or **predictor** variable.

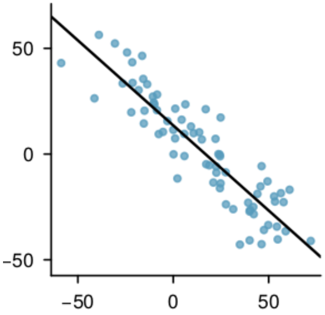
We call y the **response** variable.

Linear models

Problem: We have a set of points $\{(x_i, y_i)\}$. Given a new x value, we'd like to predict $\hat{y}$.

Linear model: We'll assume there exists a linear relationship $y = \beta_0 + \beta_1 x$ that offers a good approximation to the data.

Example:

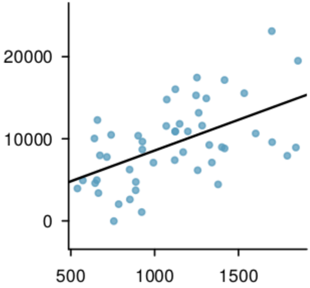


Linear models

Problem: We have a set of points $\{(x_i, y_i)\}$. Given a new x value, we'd like to predict $\hat{y}$.

Linear model: We'll assume there exists a linear relationship $y = \beta_0 + \beta_1 x$ that offers a good approximation to the data.

Example:

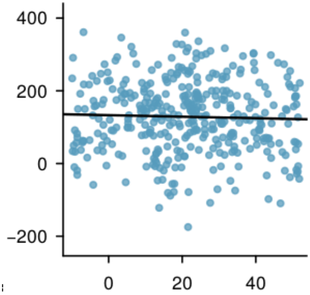


Linear models

Problem: We have a set of points $\{(x_i, y_i)\}$. Given a new x value, we'd like to predict $\hat{y}$.

Linear model: We'll assume there exists a linear relationship $y = \beta_0 + \beta_1 x$ that offers a good approximation to the data.

Example:

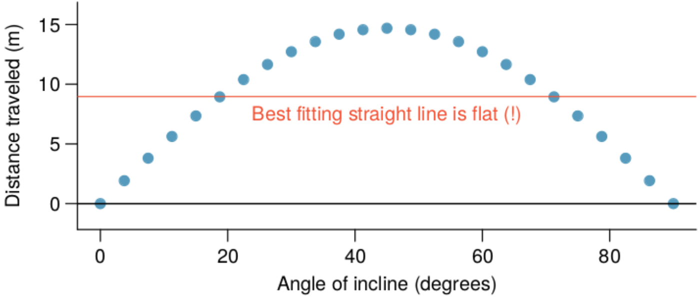


Linear models

Problem: We have a set of points $\{(x_i, y_i)\}$. Given a new x value, we'd like to predict $\hat{y}$.

Linear model: We'll assume there exists a linear relationship $y = \beta_0 + \beta_1 x$ that offers a good approximation to the data.

Example:



Residuals

What's left over.

$$\text{data} = \text{fit} + \text{residual}$$

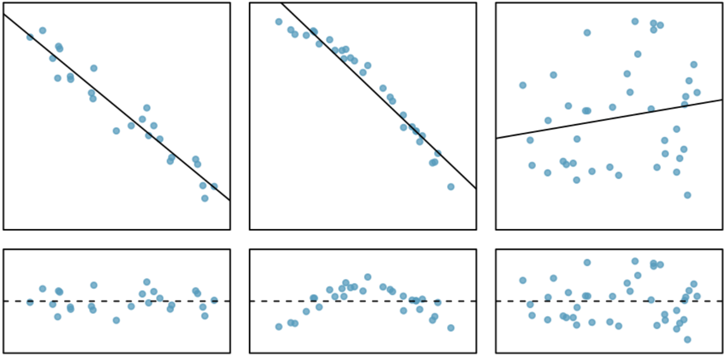
Residuals

What's left over.

$$y_i = \hat{y}_i + e_i$$

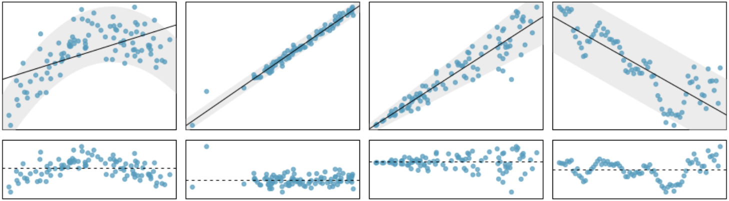
Residuals

What's left over.



Residuals

What's left over.



Residuals

What's left over.

Goal: small residuals.

$$\sum | e_i |$$

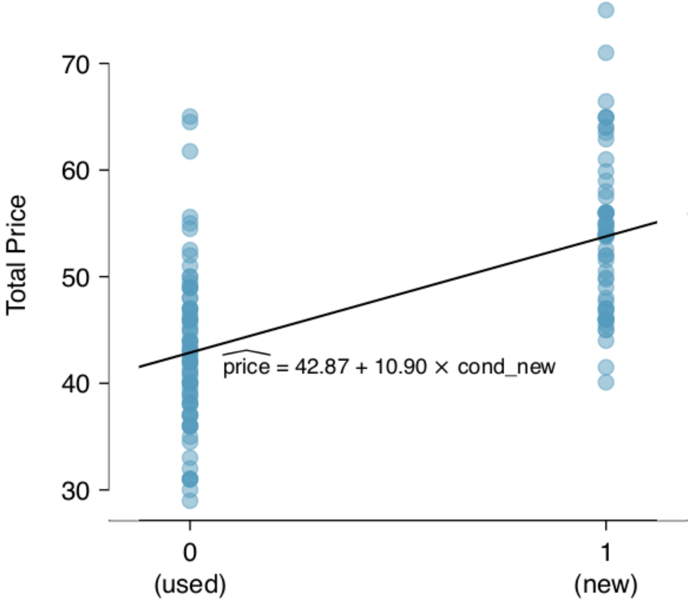
Residuals

What's left over.

Goal: small residuals.

$$\sum e_i^2$$

Categorical regression

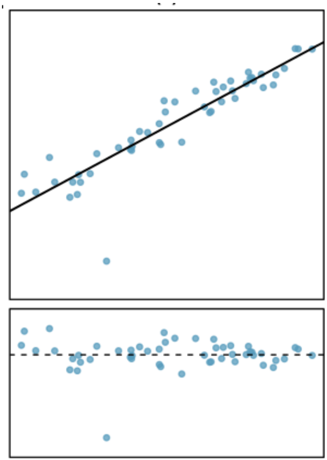


Outliers

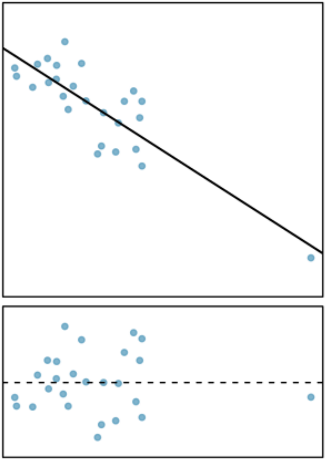
High leverage points fall far from the regression line.

Influential points make their leverage known.

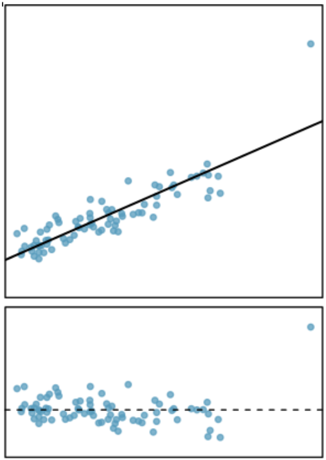
Outliers



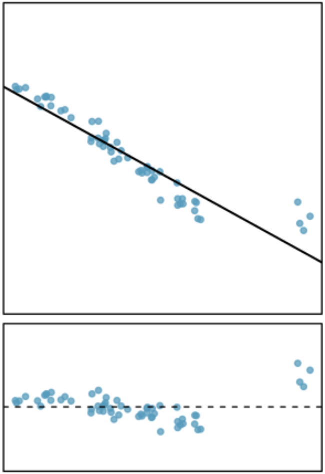
Outliers



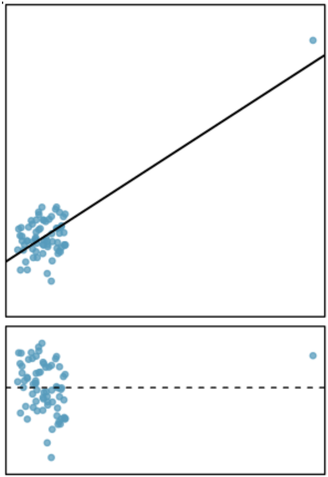
Outliers



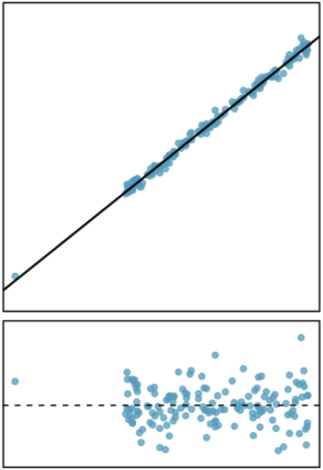
Outliers



Outliers

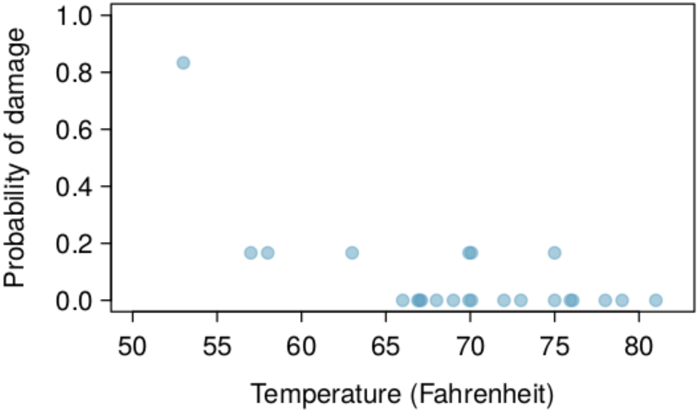


Outliers



Don't ignore outliers.

Outliers



Probability of damage
1.0
0.8
0.6
0.4
0.2
0.0
50
55
60
65
70
75
80
Temperature (Fahrenheit)

Correlation

Population correlation.

$$\rho_{X,Y} = \mathbf{Corr}(X, Y) = \frac{\mathbf{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\mathbf{E}[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

Correlation

Sample correlation.

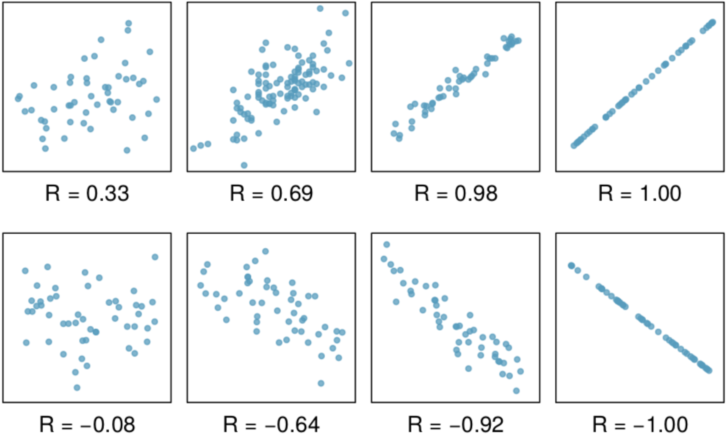
$$r_{x,y} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_i - \overline{y}}{s_y} \right)$$

Correlation

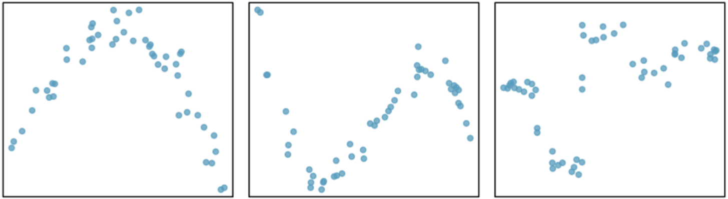
Coefficient of determination

$$r^2 = \frac{\mathbf{Var}\, e_i}{\mathbf{Var}\, y_i}$$

Correlation



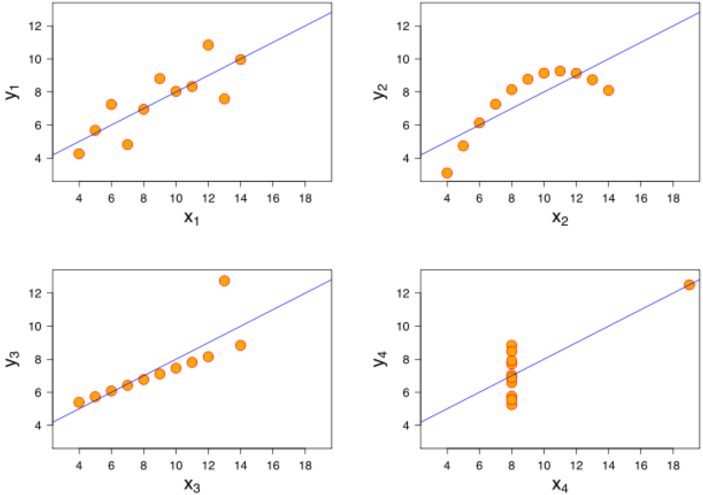
Correlation



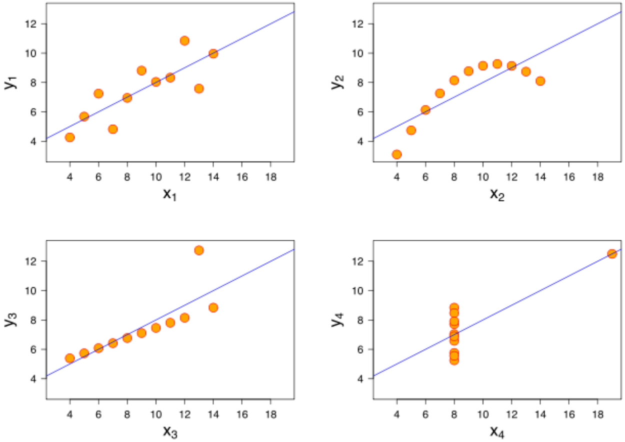
R = −0.23 R = 0.31 R = 0.50

Correlation

Anscombe's Quartet

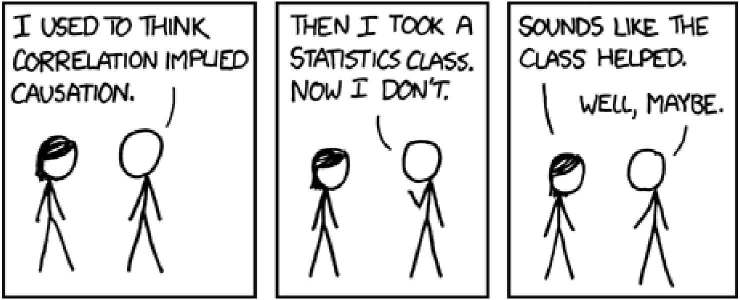


Anscombe's Quartet



http://en.wikipedia.org/wiki/File:Anscombe%27s_quartet_3.svg

Correlation does not imply causation



https://xkcd.com/552/

Correlation does not imply causation

Tufte:

"Empirically observed covariation is a necessary but not sufficient condition for causality."

http://en.wikipedia.org/wiki/Correlation_does_not_imply_causation

Correlation does not imply causation

Tufte:

"Correlation is not causation but it sure is a hint."

http://en.wikipedia.org/wiki/Correlation_does_not_imply_causation

Code lab: linear regression

For each data set:

Tasks:

- Visualise the data
- Compute mean, variance, standard deviation, correlation
- Compute linear regression and visualise
- Compute residuals and visualise
- Discuss what the data tells you

Linear regression and inference

Remember about inference?

- Standard error
- *p* values

Multiple regression

One response $\hat{y}$ but many predictors $x_1, \ldots, x_n$.

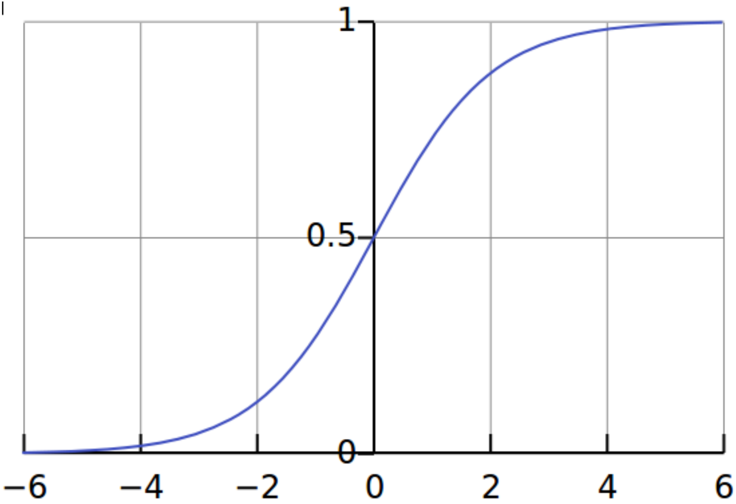
$$\hat{y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

Logistic regression

Also: *logit regression*, *logit model*

Dependent variable (y) is binary.

Logistic regression



Logistic regression

Note that $f(-x) = 1 - f(x)$.

And it satisfies this differential equation:

$$\frac{\mathrm{d}}{\mathrm{d}x} f(x) = f(x) \cdot (1 - f(x))$$

Integrating,

$$f(x) = \frac{e^x}{e^x + 1} = \frac{1}{1 + e^{-x}}$$

Fitting polynomials

- Under/over-fitting
- Test is not validation

Questions?

`purple.com/talk-feedback`

Break