Prob

Joint, Cond.

Bayes Rule

Stats

Modeling

AB testing

Binom, Poiss

Compare means

Section 3: Linear Regression

Normal

Vadim Sokolov

Linear Regression model

Suggested Reading

OpenIntro Statistics, Chapters 4,5&6

Last Section

- Estimating Parameters and Fitting Distributions
- Confidence and Prediction Intervals
- Means, Proportions, Differences
- A/B Testing

This Section

- Linear Patterns in Data (Leavitt, House Price)
- Simple Linear Regression
- Predictions (Confidence and Prediction Intervals)
- Least Squares Principle
- Hypothesis Testing (Google vs SP500)
- Model Diagnostics (Cancer and Smoking Data)
- Data transformations (World's Smartest Mammal

Regression: Introduction

$$y = a + bx \; ; \; y = \beta_0 + \beta_1 x$$

intercept slope

Regression analysis is the most widely used statistical tool for understanding relationships among variables

- Regression provides a conceptuall approach for investigating relationships between one or more factors and an outcome of interest
- The relationship is expressed in the form of an equation or a model connecting the response or dependent variable and one or more explanatory or predictor variable

$$X \longrightarrow \boxed{\text{Rule/Model}} \longrightarrow y$$

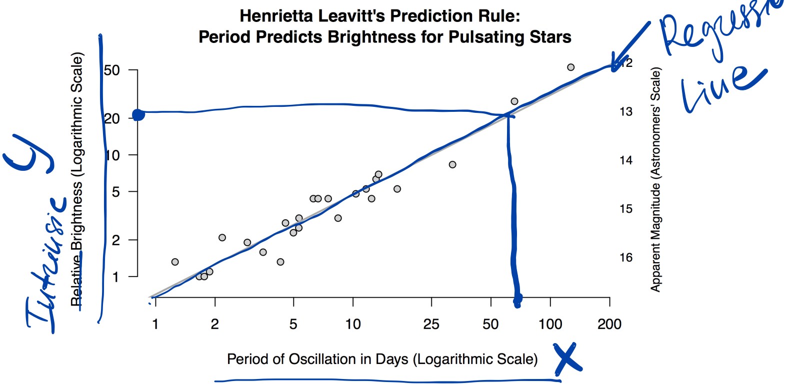
input
explanatory
predictor

Equation

output
Response
dependent

AIQ: Leavitt Stars Data

Bright Stars Have longer Period



Henrietta Leavitt's 1912 data on 25 pulsating stars. Pattern of period of oscillation with brightness allowed astronomers to measure cosmic distances over previously unimaginable scales.

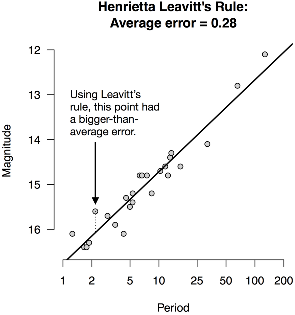
Fitting Prediction Rules to Data

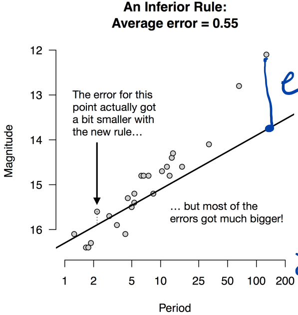
of obs

$$\frac{1}{n}\sum_{i=1}^{n} e_i^2 \rightarrow \underset{\beta_0,\beta_1}{\text{minimize}}$$

Easy to solve

In AI, the criterion for evaluating prediction rules is simple: How big are the errors the rule makes, on average?





e_i (residual error)

e^2

large → larger

small → smaller

Leavitt used "the principle of least squares" to fit a prediction rule to her data.

$$\frac{1}{n}\sum_{i=1}^{n} |e_i| \rightarrow \text{minimize?}$$

Hard to solve

Prediction

Straight prediction questions: X y

- For how much will my house sell? House Attr $\rightarrow$ Price
- Will the Chicago Cubs win the World Series? Team Attr $\rightarrow$ Prob of Win
- Will this person like that movie? (Netflix prize)

Explanation and understanding: x ? y

- What is the impact of an MBA on income? ✓ MBA $\rightarrow$ income
- How does the returns on Google relate to the market? SP500 $\xrightarrow{?}$ Goog ✓

Models, Parameters and Estimates

Values of β_0 & β_1

$$y = \underline{\beta_0} + \underline{\beta_1} x$$

params

calculated from sample

We'll use probability to talk about uncertainty ... and build models

- Define the random variable, Y, of interest
- Construct a regression model from historical data on characteristics, X This entails estimating parameters using their sample counterparts
- We are now ready to generate predictions, make decisions, evaluate risk, etc ...

Predicting House Prices

County Data

House Attr — Price

$X \longrightarrow y$

Problem: Predict market price based on observed characteristics (Zillow)

Solution:

- Look at property sales data where we know the price and some observed characteristics
- Build a decision model that predicts price as a function of the observed characteristics.

Zillow: Zestimate

R and Zestimate

R and AWS for analytics are helping Zillow real estate data.

Zillow and Big Data

Database behind the Zestimate is 20TB in size.

Zillow employs various decision tree, random forest, and regression algorithms

By averaging models, margin of error in pricing improved from 14% to 5%

Zillow Prize

Predicting whether you have waterfront property ...

Predicting House Prices

What characteristics do we use?

There are many factors or variables that affect the price of a house (location, location, location, ...)

Some basics ones include

- size
- ZIP code
- location
- parking, ...

Let's run a simple linear regression on size

Predicting House Prices

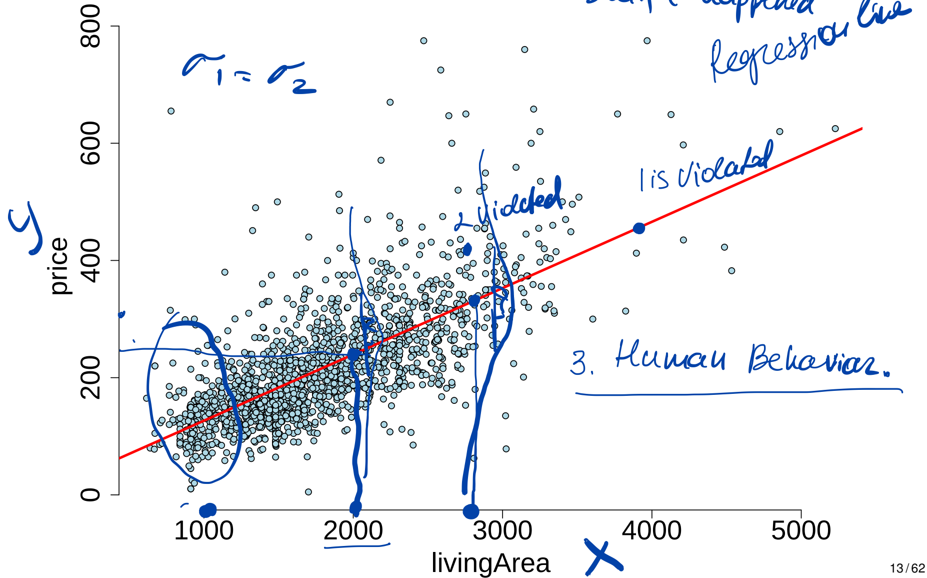
The value that we seek to predict is called the dependent (or output) variable, and we denote this by $Y =$ price of house (e.g. thousand of dollars)

The variable that we use to construct our prediction is the explanatory (or input) variable, and this is labeled $X =$ size of house (e.g. thousand of square feet)

Predicting House Prices

1. Unobserved Variables
2. Some distribution shift happened regression line

What's does the data look like?



Predicting House Prices

Price $\approx \beta_0 + \beta_1$ sqft

Simple Linear Regression (SLR) model

y \$ 1000

error

No house

$$\text{price} = \beta_0 + \beta_1 \text{ sqft} + \epsilon \text{ where } \epsilon \sim N(0, \sigma^2)$$

sqft = 0

price = β_0

where we add a random error term, ϵ.

The error term models the fact that not all prices will lie on our regression line

We find that $\boxed{\beta_1 = 0.11}$ $\beta_0 = 13.44$ \$1000

price of the land

Implication: every 1 sqft increase ups price by \$110

$\beta_1 = 0.11$ $\frac{\$}{\text{sqft}}$ Units of β_1 price of 1 sqft

Predicting House Prices

We can now predict the price of a house when we only know that size: take the value off the regression line.

For example, given a house size of $X = 2200$

Predicted Price: $\hat{Y} = 13.44 + 0.11(2200) = 262$

The intercept $\beta_0 = 13.44$ measures land value. In `R`: predict.lm(...)

Predicting House Prices

Now plot and run your regression ...

```
house = read.csv("data/SaratogaHouses.csv")
house$price =house$price/1000
plot(price~livingArea,data=house)
model=lm(price~livingArea,data=house)
coef(model)
abline(model,col="red",lwd=3)
coef(model)
```

The key command is lm(...) which stands for linear model.

R: will calculate everything for you!!

Simple Linear Regression (SLR)

What to do When violated?

1. More data?
2. Use non-linear model ✓
3. Transform data. ✓
4. Do nothing

The underlying assumptions about the linear regression model are:

1. For each value X, the Y values are *normally distributed*
2. The means of Y all lie on the regression line
3. The *standard deviations* of these normal distributions are *equal*
4. The Y values are statistically *independent*.

Violated for Time series

e.g daily sales = y

When violated: LM is not ideal

Cannot trust estimates ~~of~~ β_0 & β_1.

Cannot trust prediction intervals.

Simple Linear Regression (SLR)

fix X

$y = \underbrace{\beta_0 + \beta_1 X} + \varepsilon$

$\boxed{y \sim N(\underline{\beta_0 + \beta_1 x}, \sigma^2)}$

Stochastic part

The regression model looks like:

$$Y = \underbrace{\beta_0 + \beta_1 X}_{\text{determ. part}} + \boxed{\epsilon \text{ where } \epsilon \sim N(0, \sigma^2)}$$

does not depend on X

β_1 measures the effect on Y of increasing X by one

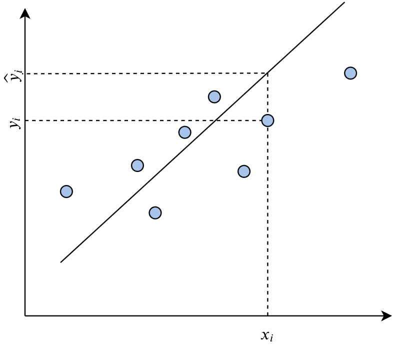
β_0 measures the effect on Y when $X = 0$.

X_f will denote a new/future value we wish to predict at

Fitted Values

The Fitted Values and Residuals have some special properties ...

Let's look at the fitted values



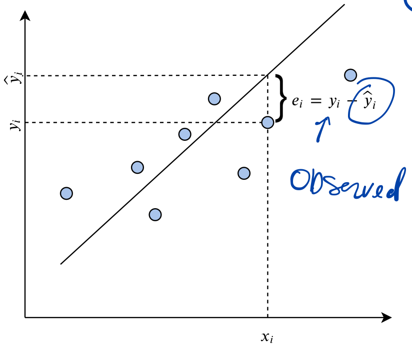
Our predictions $\hat{Y}_i = \beta_0 + \beta_1 X_i$ are given by the line!!

Residuals

$\hat{y}$ – predicted value

What is the "residual", e_i, for the ith observation?

$\hat{y} = \beta_0 + \beta_1 x$



We can write $Y_i = \hat{Y}_i + (Y_i - \hat{Y}_i) = \hat{Y}_i + e_i$

Standardized Residuals

M_1 M_2

$y = \beta_0 + \beta_1 x$ $\log(y) = \beta_0 + \beta_1 x$

- Compare models when y is on different scales

The residuals are $e_i = Y_i - \hat{Y}_i$. They estimate the errors from the line.

We re-scale the residuals by their standard errors. This lets us define

standardized residuals

$$r_i = \frac{e_i}{s_{e_i}} = \frac{Y_i - \hat{Y}_i}{s_{e_i}}$$

Outliers are points that are extreme relative to our model predictions.

They simply have large residuals!

- Don't need to think about scale/units of y

Residual Standard Error

How closely does the training dataset lie to our model?

- s is the residual standard error
- s is our estimate of σ
- $s = \sqrt{s^2}$ where

$$s^2 = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$\frac{r}{s}$ - std. residual

Lower s values means tighter predictions!!

Prediction

Suppose you have a regression of sales on price

$$\text{sales} = \beta_0 + \beta_1 \text{price}$$

You have to predict for a *given* level of price

Then the two intervals correspond to

1. A sales forecast for the next store (or next week's sales)
2. The average weekly sales (over *many* weeks)

Prediction

$y = \underline{\beta_0 + \beta_1 X} + \underline{\varepsilon}$; $\varepsilon \sim N(0, \sigma^2)$

95% CI $\boxed{\beta_0 + \beta_1 X \pm 1.96\sigma}$

Prediction is the most important application of your model Construct a new *X* variable

$\hat{\sigma} = S$

```
new = data.frame(price=5)
predict.lm(model,new,interval="prediction")
predict.lm(model,new,interval="confidence")
```

Define a vector for prediction

```
new1 = data.frame(price=c(4,5,6))
predict.lm(model,new1,interval="prediction")
predict.lm(model,new1,interval="confidence")
```

Confidence and Prediction Intervals

`lwr` lower limit, `upr` upper limit

```
       fit      lwr      upr
1 431.6129 397.0925 466.1333   # Prediction
       fit      lwr     upr
1 431.6129 416.7968 446.429    # Confidence
       fit      lwr      upr
1 474.1935 432.2873 516.0998   # Multiple Prediction
2 431.6129 397.0925 466.1333
3 389.0323 355.4325 422.6320
       fit      lwr      upr   # Multiple Confidence
1 474.1935 446.1938 502.1933
2 431.6129 416.7968 446.4290
3 389.0323 376.5104 401.5541
```

Least Squares Principle

Ideally we want to minimize the size of all of the residuals:

- If they were all zero we would have a perfect line

We'll use the least squares objective function to assess what constitutes a good "fit" to our empirical data The line fitting process:

- Minimize the "total" sums of squares of the residuals to get the "best" fit

Least Squares chooses β_0 and β_1 to minimize $\sum_{i=1}^{n} e_i^2$

$$\sum_{i=1}^{n} e_i^2 = e_1^2 + \ldots + e_n^2 = (Y_1 - \hat{Y}_1)^2 + \ldots + (Y_n - \hat{Y}_n)^2$$

Least Squares Principle

The formulas for β_0 and β_1 that minimize the least squares are:

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$
$$\beta_1 = r_{xy} \times \frac{s_y}{s_x}$$

where

- $\bar{x}$ and $\bar{y}$ are the sample means
- s_x and s_y are the sample standard deviations
- $r_{xy} = corr(x, y)$ is the sample correlation

Least Squares Principle

1. Intercept

$$\beta_0 = \bar{y} - \beta_1 \bar{x} \ \text{ or } \ \bar{y} = \beta_0 + \beta_1 \bar{x}$$

   The point $(\bar{x}, \bar{y})$ is *always* on the regression line.

2. Slope

$$\beta_1 = corr(x, y) \times \frac{s_Y}{s_X} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$
$$= \frac{cov(x, y)}{Var(x)}$$

   The estimate b is the correlation r times a scaling factor that ensures the proper units

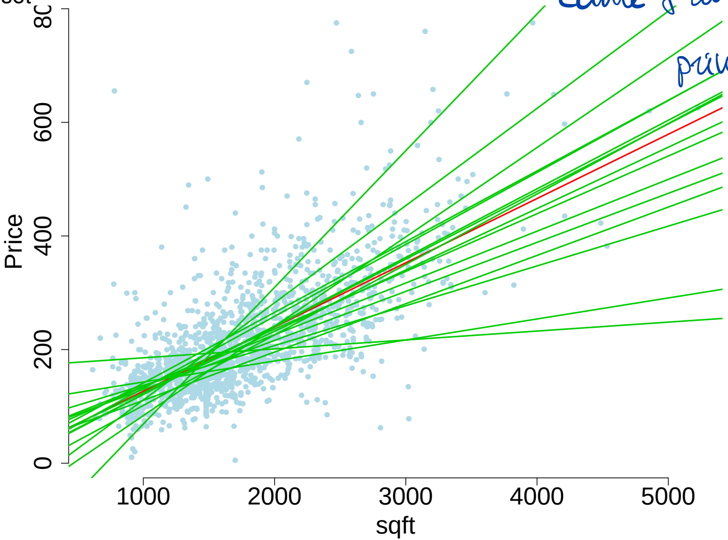
Sampling Distribution for β_1

$\beta_1 \sim N(\hat{\beta}_1 \quad)$

came from lm function

principle of least squares

Run linear regression several times using subsample of rows of the housing data set



Sampling Distribution for β_1

True $\Rightarrow$
X & y are lot
related

The sampling distribution of β_1 describes how it varies over different samples.

It allows us to calculate confidence and prediction intervals. Everything is uncertain!!

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

It turns out that β_1 is normally distributed: $\boxed{\beta_1 \sim N(\hat{\beta}_1, s_b^2)}$

- $\hat{\beta}_1$ is unbiased: $E(\beta_1) = \hat{\beta}_1$
- s_{β_1} is the standard error of β_1 The t-stat is $t_b = \beta_1 / s_{\beta_1}$
- The three factors: sample size (n), error variance (s^2), and x-spread, s_x

$$s_{\beta_1}^2 = \frac{s^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \boxed{\frac{s^2}{(n-1)s_x^2}}$$

$n \nearrow \quad S_{\beta_1}^2 \swarrow \checkmark$

$S_y \quad \nearrow \quad \checkmark$

$S_x^2 \downarrow \quad \nearrow \quad \bullet ?$

Prediction: revisited

CI for predicted y (ε)

Prediction Interval: ε + uncert in β_1 & β_0

How do we assess how much error that could be in our best prediction?

$$\hat{Y}_f = \beta_0 + \beta_1 X_f + e_f \text{ where } e_f \sim N(0, s^2)$$

There's error in everything, β_0, β_1, e_f, ...

After we account for all the uncertainty,

$$var(y_f) = s^2\left(1 + \frac{1}{n} + \frac{(x_f - \bar{x})^2}{(n-1)s_x^2}\right)$$

	Var(y)
$s^2 \uparrow$	$\uparrow$
$n \uparrow$	$\downarrow$
$S_x^2 \downarrow$	$\uparrow$
$(X_f - \bar{X})^2 \uparrow$	$\uparrow$

In R: predict.lm(...)

Housing

$\bar{X} = 2500$

Prediction errors

A large predictive error variance (high uncertainty) comes from four factors

1. Large s (i.e. large errors, ϵ's)
2. Small n (not enough data)
3. Small s_x (not enough spread in the covariates)
4. Large difference between x_f and $\bar{x}$ (predicting extremes)

As a practical matter, low s values are more important for prediction than high R^2-values.

Example: Google Stock Returns

Let's use the `quantmod` package to read in the data

```
library(quantmod)
Y = getSymbols("GOOG", from = "2005-01-01")
# Retrieve closing prices
y = GOOG$GOOG.Adj.Close
head(y)
[1] 101.25392  97.15301  96.65851  94.18098  96.82834  97.43274
tail(y)
[1] 796.42 794.56 791.26 789.91 791.55 785.05
```

Example: Google

Consider a CAPM regression for Google's stock

$$\underbrace{\text{Google}_t}_{y} = \alpha + \beta \underbrace{\text{sp500}_t}_{x} + \epsilon_t$$

time (daily)

In finance (α, β) are used instead of (β_0, β_1).

We'd like to know our estimates $(\hat{\alpha}, \hat{\beta})$.

Then formulate lots of hypothesis tests:

no relation

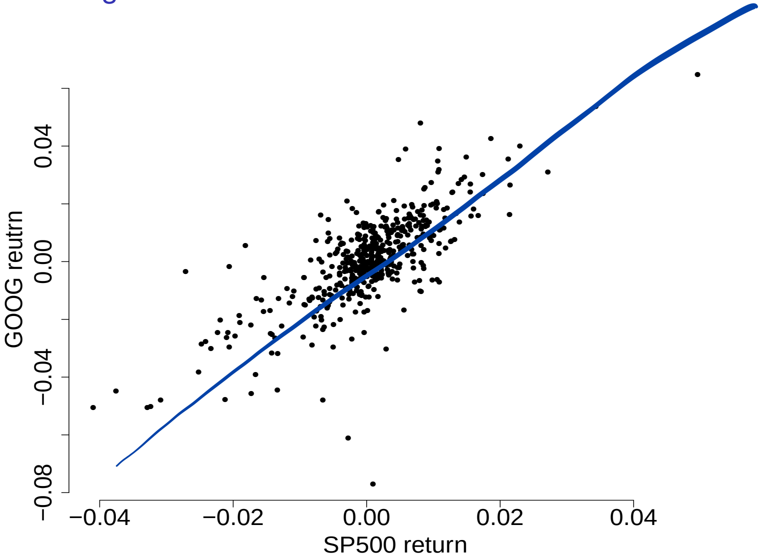
H_0 : is Google related to the market? $H_0: \beta = 0$; $H_1: \beta \neq 0$

H_0 : does Google out-perform the market in a consistent fashion?

H_0 : is Google better than Nvidia? $H_0: \alpha = 0$ $H_1: \alpha \neq 0$

Goog does not outperform

Example: Google



Example: Google

Goog does not out perform

summary(model) command provides all of our estimates ...

```
> summary(model)
lm(formula = ret ~ SP500)
                Estimate  Std. Error  t value   P(>|t|)
(Intercept)  0.0004086   0.0002936    1.392    0.164
SP500        0.9232752   0.0232625   39.689   <2e-16 ***
Residual standard error: 0.01546
Multiple R-squared:  0.3622
```

α ✓

β

Cannot Reject $H_0: \alpha = 0$

$H_0: \beta = 0$

Reject H_0

Accept $H_1: \beta \neq 0$

Our best estimates are: $\hat{\alpha} = 0.0004$, $\hat{\beta} = 0.9232$

How much will Google move if the market goes up 10%?

What do the t-ratios show?

Outliers

Residuals allow us to define outliers:

95% of the time we expect the standardized residuals to satisfy $-2 < r_i < 2$

Any observation with $\boxed{| r_i | \geq 3}$ is an extreme outlier

Residuals will also help in assessing the validity of our model ...

Influential Points

infl. of point i is change in slope β_1 when point i is removed.

Influential points are observations that affect the magnitude of our estimates $\hat{\beta}_1$.

They are important to find as they typically have economic consequences.

We will use Cook's D distance to assess the significance of an influential point

They are typically extreme in the characteristics, X-space

We will delete observations with Cook's D greater than one and assess the sensitivity of our conclusions

Comes out of lm function
by-product of the estimation algorithm.

Influential Points: Cook's D

Cook's D depends on the standardized residual, r_i, and leverage, $0 < h_i < 1$

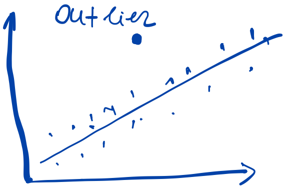
$$\text{CookD}_i = \frac{1}{p} r_i^2 \frac{h_i}{1 - h_i}$$

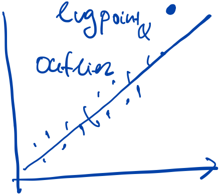
where p is the number of variables

```
plot(model)

plot(cooks.distance(model))

datanew = data[-i,]  # Deletes ith row
```

Is $\beta_{1(-i)}$ is different from β_1?

Influential Points: Cook's D





They are three ranges: $0 < D_i < 0.5$, $0.1 < D_i < 1$ and $D_i > 1$

We will delete all observations with Cook's D > 1

To see how stable our β_1's are to these data points

Quite often, I also delete the point with the largest Cook's D just to check it doesn't affect my conclusions

All this is done, before I use summary(model) and interpret my model.

each poin

lev	Out
not	not
y	n
y	y
n	n

Regression: Strategy

Five point basic strategy

1. Input and Plot Data: Use `plot` and `boxplot` commands
2. Build Regression Model: Use the `model = lm (y ~ x)` command
3. Diagnostics: `plot(model)` Fitted vs standardized residuals.

   QQplot Residuals for Outliers and

   Cook's D for Influential
4. Interpretation: `summary(model)`. Regression β's
5. Prediction: `predict.lm`.

   A model is only as good as its predictions. Do some out-of-sample

   forecasting

R Regression Commands

Given input-output vectors *x* and *y*

`cor( ... )` computes correlation table

`model = lm(y ~ x)` for linear model (a.k.a regression)

`model = glm(y ~ x)` for logistic regression

`model = lm(y ~ x1+ ... + xp)` for linear multiple regression model

R provides diagnostics in

`plot(model)` 4-in-1 diagnostics plot

`plot(cooks.distance(model))` influential points

`rstudent(model)` outliers

Output and Prediction

`R` provides model output in

`summary(model)` provides a summary analysis of our model

`R` provides predictions in

`newdata = data.frame( ...  )` constructs a new input variable

`predict.lm(model,newdata)` provides a prediction at a new input

Diagnostics: plot(model) 4-in-1 plot

Everything in plot(model) our 4-in-1 residual plot

1. Residuals vs Fitted: Straight line. Random looking pattern
2. Scale-location: Ought to be a straight line. Otherwise changing variance
3. Normal Q-Q Plot: Standardized residuals. This should be a straight line.

   You're plotting quantiles of the standardized residuals vs what you'd expect if the assumptions are true, a standard normal
4. Residuals vs Leverage: Contours of Cook's D. If $D \geq 1$ then influential.

   Remove and see what happens!!

In `R`: simply use plot(model)

Example: Lung Cancer Data

Famous dataset linking lung cancer and cigarette consumption.

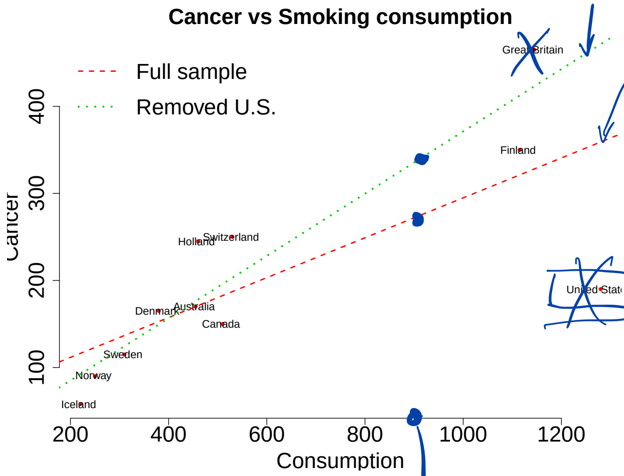
$Y =$ lung cancer deaths/million in 1950

$X =$ cigarette consumption/capita in 1930

Country	Y	X
1. Iceland	58	220
2. Norway	90	250
3. Sweden	115	310
4. Canada	150	510
5. Denmark	165	380
6. Australia	170	455
7. United States	190	1280
8. Holland	245	460
9. Switzerland	250	530
10. Finland	350	1115
11. Great Britain	465	1145

Cancer and Smoking Data

Without U.S.



Full Model

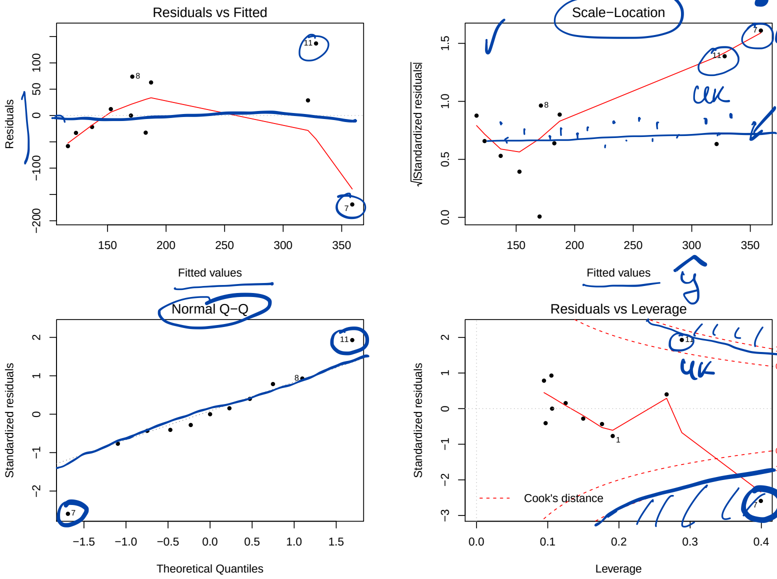
log & outl ?

Brazil

The US and UK are well off the regression line

4 in 1 Residual Plots for Model 1

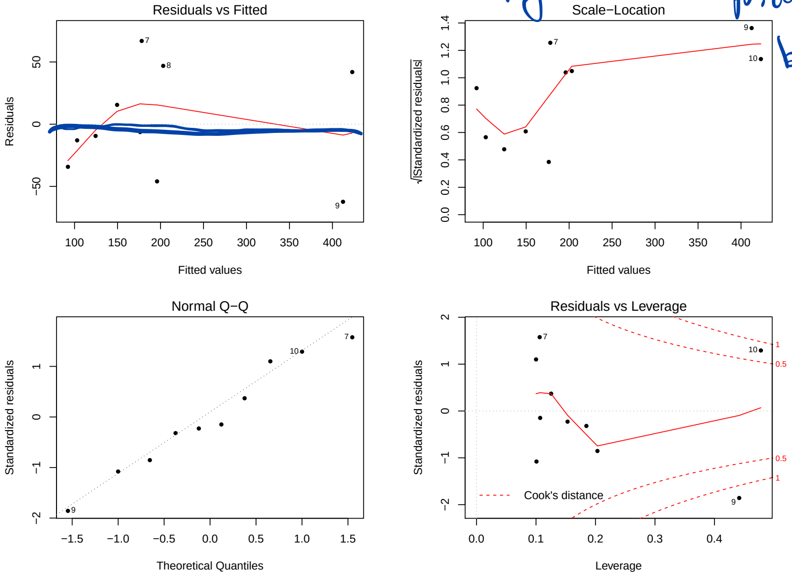
plot(model) UK out
US out & lvg



US
ideal assump. of lm satisfied

4 in 1 Residual Plots for Model 2

With o.s Removed satisfy the assumptions better.



Model Coefficients

There are two ways to get the model coefficients in `R`

1. coef(model)

```
(Intercept) Consumption
 66.8434535   0.2286585
```

2. lm(formula = Cancer ~ Consumption)

```
Coefficients:
(Intercept)  Consumption
    66.8435       0.2287
```

Transformations

Basic assumption is linearity

What if this doesn't hold?

1. A simple solution is to transform the variables.
2. Re-run the regression on the transformed
3. If all is fine then the model holds on the transformed scale.

   Then transform back to the original nonlinear scale.

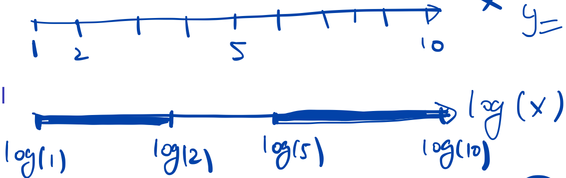
The two most common models are

Power relationship

Exponential relationship

The log-log Model

Power/Multiplicative Model



$y = \beta_0 + \beta_1 x$

Multiplicative Model: $Y = AX^b$ where $A = e^a$

$y = A x^{b}$

Log-Log Transformation: $\log(Y) = \beta_0 + \beta_1 \log(X)$

x - price
y - demand

Why? Taking logs of both sides gives

$$\log Y = \log A + \log X^b = \beta_0 + \beta_1 \log X$$

$x \to kx$, $k = 1.1$

$Ax^b \to Ax^b k^b$

The slope, β_1, is an elasticity. % change in Y versus % change in X

When $x \uparrow$ by 10%

Variables are related on a multiplicative, or percentage, scale.

$y \uparrow$ by 1.1^b %

In R: model = lm(log(y) ~log(x))

Recall: log is the natural $\log_e$ with base $e = 2.718\ldots$ and that $\log(ab) = \log a + \log b$ and $\log(a^b) = b \log a$.

Both X & y change on % scale

$\log(y) = \log(AX^b) = \log(A) + \log(X^b) = \log(A) + b\log(X)$

The Exponential Model

Compounding Growth

Suppose that we have an equation: $Y = Ae^{bX}$ where $A = e^a$.

This is equivalent to $\log(Y) = \beta_0 + \beta_1 X$

Taking logs of the original equation gives

$$\log Y = \log A + \beta_1 X$$

$$\boxed{\log Y = \beta_0 + \beta_1 X}$$

Therefore, we can run a regression of $\log Y$ on X!!

Caveat: not all variables can be logged!

$Y > 0$ needs to be positive.

Dummy variables $X = 0$ or 1 can't be logged.

Counting variables are usually left alone as well.

X

Every 10 years

SP500 doubles y

X: changes on linear scale

y: changes on % scale

$$y = Ae^{bx}$$

$$\log(y) = \log(A) + \log(e^{bx})$$

$$= \log(A) + bx \, \log(e) \quad 1$$

$$= \log(A) + bx$$

Example: World's Smartest Mammal

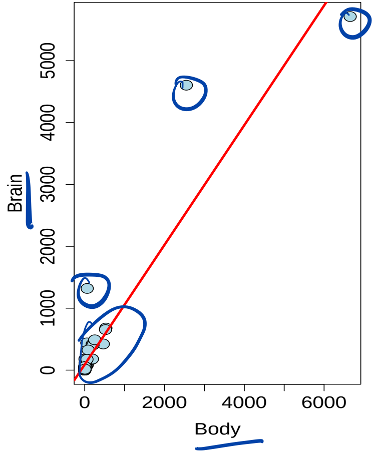
First of all, read in and attach our data ...

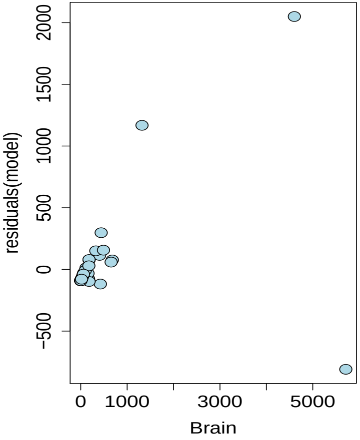
name

```
mammals = read.csv("data/mammals.csv")
attach(mammals)
head(mammals)
                      Mammal    Brain   Body
1           African_elephant 6654.000 5712.0
2 African_giant_pouched_rat     1.000    6.6
3                 Arctic_Fox    3.385   44.5
4     Arctic_ground_squirrel    0.920    5.7
5             Asian_elephant 2547.000 4603.0
6                     Baboon   10.550  179.5
> tail(mammals)
                 Mammal         Brain Body
57               Tenrec         0.900  2.6
58           Tree_hyrax         2.000 12.3
59           Tree_shrew         0.104  2.5
60               Vervet         4.190 58.0
61        Water_opossum         3.500  3.9
62 Yellow-bellied_marmot        4.050 17.0
```

Residual Diagnostics

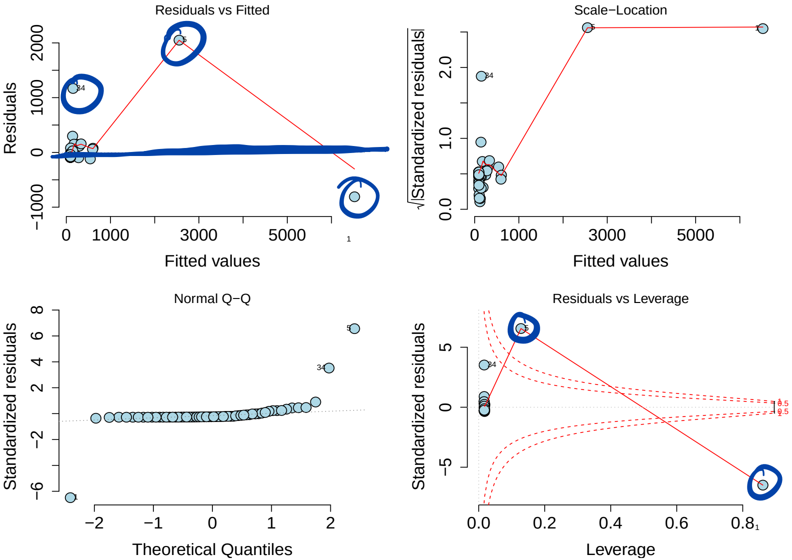
The residuals show that you need a transformation ....



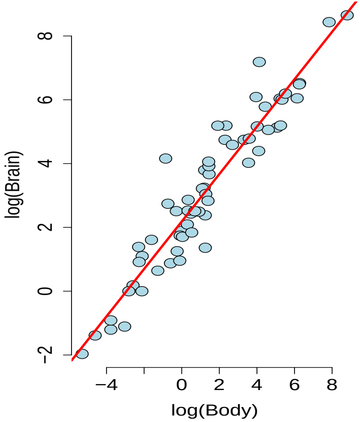


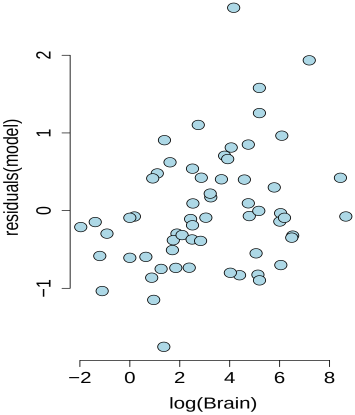
Residual Plots

lin assum are violated



log-log model

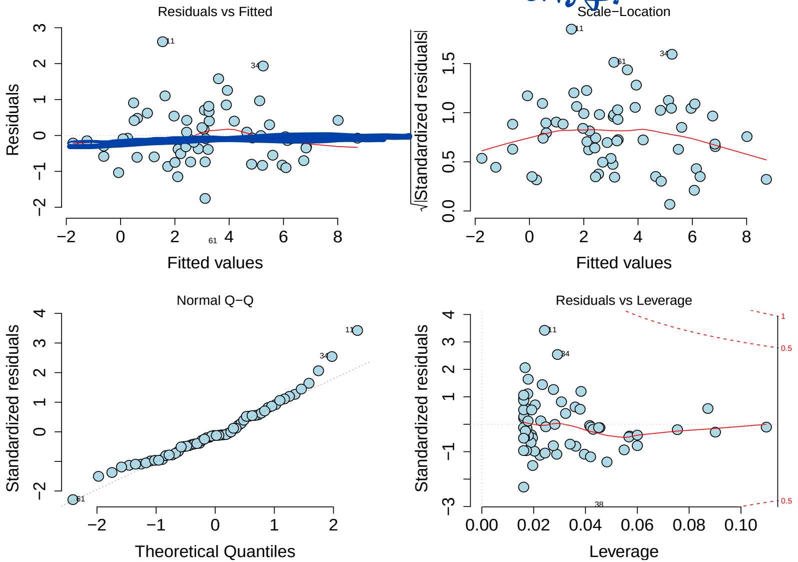




That’s better!

4 in 1 Residuals: log-log model

Ass. of lm are somewhat satisf.



log-log Model

$\log(\text{Brain}) = \beta_0 + \beta_1 \log(\text{Body})$

lm(formula = log(Brain) ~log(Body))

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.18328    0.10682   20.44   <2e-16 ***
log(Body)    0.74320    0.03166   23.48   <2e-16 ***
```

$$\log(\text{Body}) = 2.18 + 0.74 \log(\text{Brain})$$

The coefficients are highly significant $R^2 = 90\%$

$\text{Body} = 1g$

$\exp(\log(\cancel{\text{Body}}\ \text{Brain})) = \exp(2.18)$ $\quad \log(\cancel{\text{Brain}}\ \text{Body}) = 0$

$\cancel{\text{Body}}\ \text{Brain} = \exp(2.18)\,\text{kg}$ $\quad \log(\cancel{\text{Body}}\ \text{Brain}) = 2.18$

Outliers

Galton

6.4

```
rstudent(model)
```

Predicted

Brain size

	Mammal	Brain	Body	Residual	Fit
11	Chinchilla	64.0	0.425	3.7848652	4.699002
34	Man	1320.0	62.000	2.6697886	190.672827
50	Rhesus_monkey	179.0	6.800	2.1221002	36.889735
6	Baboon	179.5	10.550	1.6651361	51.128826
42	Owl_monkey	15.5	0.480	1.4589815	5.143815
10	Chimpanzee	440.0	52.160	1.2734358	167.690600

There is a residual value of 3.78 extreme outlier.

It corresponds to the Chinchilla.

This suggests that the Chinchilla is a ~~master race of supreme intelligence~~!

Inference

NO!!! I checked and there was a data entry error.

- The brain weight is given as 64 grams and should only be 6.4 grams.
- The next largest residual corresponds to mankind

In this example the log-log transformation used seems to achieve two important goals, namely linearity and constant variance.

Glossary of Symbols

Intercept, β_0

Slope, β_1

Error, e

Residual standard error, s

Standardised residual, r_i

Leverage, h_i

Summary

- Linear Patterns in Data (Leavitt, House Price)
- Simple Linear Regression
- Predictions (Confidence and Prediction Intervals)
- Least Squares Principle
- Hypothesis Testing (Google vs SP500)
- Model Diagnostics (Cancer and Smoking Data)
- Data transformations (World's Smartest Mammal