Introduction to hypothesis testing with R

Mark Johnson

Macquarie University
Sydney, Australia

October 15, 2014

Outline

Introduction

Hypothesis tests and confidence intervals

Tests for tabular data

One and two sample tests

Linear regression

Logistic regression

Mixed effects models

Conclusions

Hypothesis tests in R

- This material is based on Dalgard, P. 2008 *Introductory statistics with R*, which you should download (for free) from the MQ library
- The goal of this presentation is to make you aware of the kinds of statistical tests available
- If you use a statistical test in a publication *be sure to read a reliable textbook to make sure you are using it correctly*

Hypothesis testing: motivating examples

- I have a coin, which I'm not sure if is "fair". So I throw it 10 times, and it comes up tails 2 times. Is this evidence that the coin is biased?
- I measure the time it takes for a group of girls to push a button in an experiment, and then I do this for a group of boys. My data show that on average the girls are 10msec faster than the boys. Can I conclude that girls do this task faster than boys, and if so, by how much?
- I've modified my syntactic parser, but I'm not sure if my modifications have really made it more accurate. So I run both the old and the new parsers on the same set of "test sentences" and measure the accuracy of the parses they produce for each sentence. On average my new parser is 2% more accurate than the old one. Is it really better than the old parser, and by how much?

Statistics and Probability

- A statistic is a function of the data (usually chosen to summarise it)
  - example: the minumum, maximum and average are three different statistics
- Probability theory is the mathematics of random phenomena
- Hypothesis tests are statistics that indicate whether a hypothesis is consistent with the data (e.g., "Is this coin fair?")
- Confidence intervals are statistics that estimate a range of values that contains the true value of a parameter (e.g., "What are the lowest and highest values for the probability of heads?")
- There's a general move away from hypothesis tests to confidence intervals

Frequentists and Bayesian approaches

- Frequentist: the probability of an event is the frequency with which it appears in an infinite sequence of replications
- Bayesian: the probability of an event measures the degree of certainty or belief in that event
- Frequentist and Bayesian approaches have different notions of hypothesis testing and confidence intervals
- Frequentist approaches are more restrictive and unnatural, but computationally simple
- Bayesian approaches can easily integrate more diverse data, but computationally intensive
- Most "pre-packaged" software implements frequentist approaches, so that's what we'll cover here

Outline

Introduction

Hypothesis tests and confidence intervals

Tests for tabular data

One and two sample tests

Linear regression

Logistic regression

Mixed effects models

Conclusions

Is this coin fair?

- Hypothesis H_1: *this coin is not fair*, i.e., $p_{heads} \neq 0.5$
- Null hypothesis H_0: *this coin is fair*, i.e., $p_{heads} = 0.5$
- Data: out of 10 flips, 2 are tails
- Events as or more extreme than the data:
  - 0 tails, 1 tail, 2 tails, 0 heads, 1 head, 2 heads
- Probability of these extreme events under null hypothesis: $p = 0.109$
  - it's conventional to reject the null hypothesis H_0 when p is less than 0.05, 0.01 or 0.001

Hypothesis tests and the null hypothesis

- The Neymann/Pearson/Wald approach to hypothesis testing:
  - given a hypothesis to be tested H_1, formulate an alternative *null hypothesis* H_0
  - pick a *test statistic* T and a *significance level* α
  - calculate the value $T(D)$ of the test statistic on the data D
  - calculate the probability p of data sets with test statistics as or more extreme than $T(D)$
  - if $p < \alpha$ then accept H_1, otherwise reject H_1

Class discussion (1)

- Suppose our hypothesis H_1 were: *this coin is biased toward heads*, i.e., $p_{heads} > 0.5$
- What should the null hypothesis H_0 be for this H_1?
- Suppose the data is as before: out of 10 flips, 2 are tails
- What are the events as or more extreme than the data?
- What is the probability of these extreme events under the null hypothesis?
- (This is an example of a *one-sided test*)

Type 1 and type 2 errors

	H_0 is true coin really is fair	H_1 is true coin really is biased
Accept H_0 report coin is fair		Type 2 error false negative
Accept H_1 report coin is biased	Type 1 error false positive	

- In order to bound the probability of Type 2 errors below a small value α, we may have to accept a high probability of making a Type 1 error

Class discussion (2)

- What do type 1 and type 2 errors correspond to in the following contexts?
  - a medical test for whether the patient has a specific disease
  - a medical test for whether the patient is cured of a disease
  - a test for whether a computer program is functioning correctly
  - a test for whether a child knows a certain linguistic construction
- The standard approach to hypothesis testing bounds the probability of type 2 errors to a small value α, but doesn't bound the probability of type 1 errors. Is this reasonable? (Consider e.g., the cases above).

What could p_{heads} be?

- Data: out of 10 throws, 8 are heads
- The *maximum likelihood estimate* $\widehat{p}_{heads} = 0.8$, but 8/10 heads is not that unlikely if $p_{heads} = 0.7$
- A *95% confidence interval* is a statistic such were we to flip coins with various values of p_{heads} 10 times, 95% of the time p_{heads} would be within the confidence interval
  - A 95% confidence interval p_{heads} for this data is $[0.444, 0.975]$
- Confidence intervals can be derived from hypothesis tests
  - 0.5 is in the 95% confidence interval for p_{heads}
    $\iff$ $H_0 : p_{heads} = 0.5$ is not rejected at the 0.05 level

Warning about implicit stopping rules

- If the significance level $\alpha = 0.05$, then the null hypothesis will be rejected about one in every twenty experiments, *even if the null hypothesis is true*

⇒ If you just keep redoing your experiment, *eventually the results will be significant*

  - E.g., if we keep flipping a fair coin, eventually we'll see 10 heads in a row

- Doing this deliberately is scientific fraud, but it's easy to do this accidentally:
  - e.g., keep adjusting your program/experiment until the results are good
  - this is called a *stopping rule*, and significance levels are affected by the stopping rule
- This can be minimised by first selecting the experimental settings on development data, and then performing a single experiment on the test data

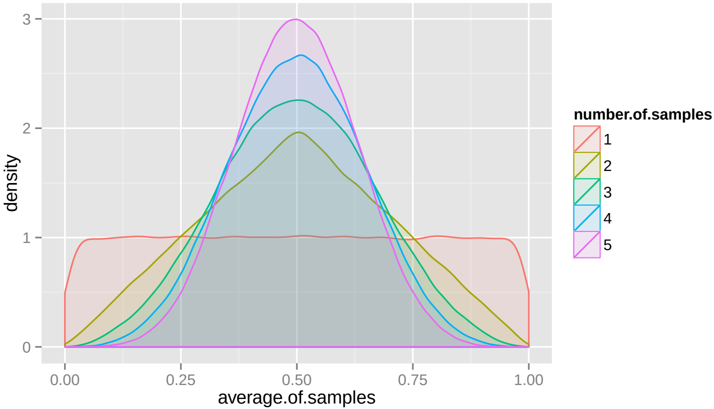
Class discussion (3)

- Can you come up with other examples where one might "accidentally" produce erroneous significance results?
- What could you do to guard against this?

Computing hypothesis tests and confidence intervals

- For simple cases (like the biased coin example) it is possible to directly compute the probability of the data under the null hypothesis
- Standard techniques use *normal approximations* which are reasonably accurate when there are more than (say) 5 data points
- It's often possible to compute significance levels by *sampling* from the null hypothesis

The central limit theorem



- The *central limit theorem* says that the average of a set of independent and identically-distributed samples approaches a normal distribution (a.k.a. a Gaussian distribution) as the number of samples grows
    - the normal distribution is a fairly good approximation when there are 5 or more samples

MACQUARIE UNIVERSITY

Outline

Introduction

Hypothesis tests and confidence intervals

Tests for tabular data

One and two sample tests

Linear regression

Logistic regression

Mixed effects models

Conclusions

Review: libraries and data in R

- We'll the `ggplot2` plotting package, and the `mpg` dataset that comes with it:

```
install.packages("ggplot2")
library(ggplot2)
```

```
xtable(summary(mpg))
```

	manufacturer	model	displ	year	cyl
1	dodge :37	caravan 2wd : 11	Min. :1.60	Min. :1999	Min. :4.00
2	toyota :34	ram 1500 pickup 4wd: 10	1st Qu.:2.40	1st Qu.:1999	1st Qu.:4.00
3	volkswagen:27	civic : 9	Median :3.30	Median :2004	Median :6.00
4	ford :25	dakota pickup 4wd : 9	Mean :3.47	Mean :2004	Mean :5.89
5	chevrolet :19	jetta : 9	3rd Qu.:4.60	3rd Qu.:2008	3rd Qu.:8.00
6	audi :18	mustang : 9	Max. :7.00	Max. :2008	Max. :8.00
7	(Other) :74	(Other) :177			

Review: Data frames in R

- Data in R is generally represented in a *data frame*, which is a list of equal-length columns.
- Numerical or boolean columns are *vectors*
- Categorical data is represented by a *factor*, which consists of:
  - a vector of *levels* or categories
  - an integer vector of values, which indexes the levels vector

```
levels(mpg$manufacturer)

##  [1] "audi"       "chevrolet"  "dodge"      "ford"
##  [5] "honda"      "hyundai"    "jeep"       "land rover"
##  [9] "lincoln"    "mercury"    "nissan"     "pontiac"
## [13] "subaru"     "toyota"     "volkswagen"

head(mpg$manufacturer)

## [1] audi audi audi audi audi audi
## 15 Levels: audi chevrolet dodge ford honda hyundai ... volkswagen

as.integer(head(mpg$manufacturer))

## [1] 1 1 1 1 1 1
```

Binomial test for a single proportion

- The binomial test tests whether binary samples were generated from a source with a given probability
- Example: *Did a fair coin produce 8 heads in 10 flips?*

```r
binom.test(8, 10, p=0.5, conf.level=0.95)

##
##  Exact binomial test
##
## data:  8 and 10
## number of successes = 8, number of trials = 10,
## p-value = 0.1094
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.4439 0.9748
## sample estimates:
## probability of success
##                     0.8
```

See Dalgaard (2008) section 8.1

Computational aside: R objects

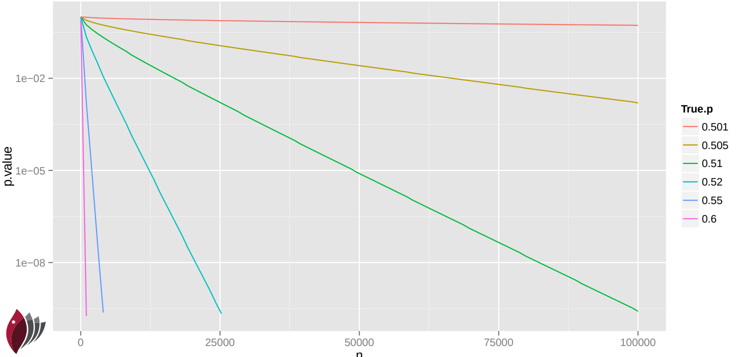
- R functions often produce class objects as output. Their print method produces the printed output, but they have other useful attributes.

```
t = binom.test(8, 10, p=0.5, conf.level=0.95)
t$p.value

## [1] 0.1094

t$estimate

## probability of success
##                    0.8

summary(t)

##             Length Class  Mode
## statistic   1      -none- numeric
## parameter   1      -none- numeric
## p.value     1      -none- numeric
## conf.int    2      -none- numeric
## estimate    1      -none- numeric
## null.value  1      -none- numeric
## alternative 1      -none- character
## method      1      -none- character
## data.name   1      -none- character
```

How does p-value vary with data size?

- Run `binom.test` on different data sizes and effect sizes

```
df = expand.grid(true.p = c(0.501, 0.505, 0.51, 0.52, 0.55, 0.6),
    n = floor(seq(10, 1e+05, length.out = 100)))
df$x = floor(df$true.p * df$n)
binom.test.p.value = Vectorize(function(x, n) binom.test(x, n)$p.value)
df$p.value = binom.test.p.value(df$x, df$n)
df$True.p = as.factor(df$true.p)
qplot(n, p.value, data = subset(df, p.value > 1e-10), log = "y", geom = c("line"),
    colour = True.p)
```



Fisher's exact test for comparing two proportions

- Fisher's exact test tests whether two sequences of binary samples come from the same distribution
    - Fisher's exact test is computationally very demanding, and when the numbers are large you should use `prop.test` or `chisq.test`
- Example: *Did the same biased coin generate a sequence of 8 heads and 2 tails, and another sequence of 4 heads and 7 tails?*

```
counts = matrix(c(8,2,4,7), 2)
fisher.test(counts)

##
##  Fisher's Exact Test for Count Data
##
## data:  counts
## p-value = 0.0805
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##   0.7307 91.0288
## sample estimates:
## odds ratio
##      6.303
```

See Dalgaard (2008) section 8.2

Tests for interaction in $r \times c$ tables

```r
d = table(mpg$class, mpg$drv)
xtable(d)
```

	4	f	r
2seater	0	0	5
compact	12	35	0
midsize	3	38	0
minivan	0	11	0
pickup	33	0	0
subcompact	4	22	9
suv	51	0	11

Chi-squared test for interaction

- The Chi-squared test can be used to determine if there's an interaction in $r \times c$ tables
  - Fisher's exact test can also be used on $r \times c$ tables, but it is computationally very demanding (use if counts are small)

```
chisq.test(table(mpg$class, mpg$drv))

## Warning:  Chi-squared approximation may be incorrect

##
##  Pearson's Chi-squared test
##
## data:  table(mpg$class, mpg$drv)
## X-squared = 221.6, df = 12, p-value < 2.2e-16
```

See Dalgaard (2008) section 8.4

Which cells are most “surprising”?

- The Chi-squared test predicts the values each cell should have if there is no interaction, and compares the actual cell values to those predictions
- We can display the relative difference between the predicted and actual values as follows

```r
results = chisq.test(table(mpg$class, mpg$drv))
observed = results$observed
expected = results$expected
squared.difference = (observed-expected)^2/expected
xtable(squared.difference)
```

	4	f	r
2seater	2.20	2.26	37.33
compact	3.65	8.83	5.02
midsize	12.55	20.32	4.38
minivan	4.84	7.27	1.18
pickup	23.50	14.95	3.53
subcompact	8.44	2.38	7.40
suv	20.60	28.09	2.89

Computational exercise (1)

- Read the Wikipedia entry about Fisher's "Lady tasting tea" experiment. Apparently the lady (Muriel Bristol) correctly identified which cups of tea had milk poured in them first. Run `fisher.test`, `prop.test` and `chi.sq` on the "lady tasting tea" data, and compare the results.

  Do your results agree with the significance level in the Wikipedia entry? If not, why not? If you run `help(fisher.test)` you'll discover additional arguments to `fisher.test` that may enable you to replicate the Wikipedia numbers.
- With the `mpg` data from the `ggplot2` library, determine if the proportion of car models with each kind of transmission has changed post-2000. Use the `fisher.test`, `prop.test` and `chisq.test` on your data. (Don't worry if you can't run one of the tests). You may find the following code helpful:

```
d = data.frame(mpg)  # copy mpg data frame into a new data frame d
d$post.2000 = (mpg$year > 2000)
```

- Use `qplot` from the `ggplot2` library to produce a bar graph showing the distribution of transmission types pre- and post-2000. Use colour to indicate transmission type. (Hint: see previous slides).
- Now repeat this experiment, but don't distinguish the different kinds of manual and automatic transmission. You may find the following code helpful:

```
d$manual = (d$trans %in% c("manual(m5)", "manual(m6)"))
```

MACQUARIE
UNIVERSITY

Outline

Introduction

Hypothesis tests and confidence intervals

Tests for tabular data

One and two sample tests

Linear regression

Logistic regression

Mixed effects models

Conclusions

Parametric and non-parametric tests

- Parametric tests assume that the data are generated from a certain family of probability distributions (usually normal distributions)
- Non-parametric tests don't assume that the data come from any specific family of probability distributions
- Parametric tests are typically easier to compute and more powerful when their assumptions are met
- When in doubt, it's probably safer to use non-parametric tests

One-sample *t*-test

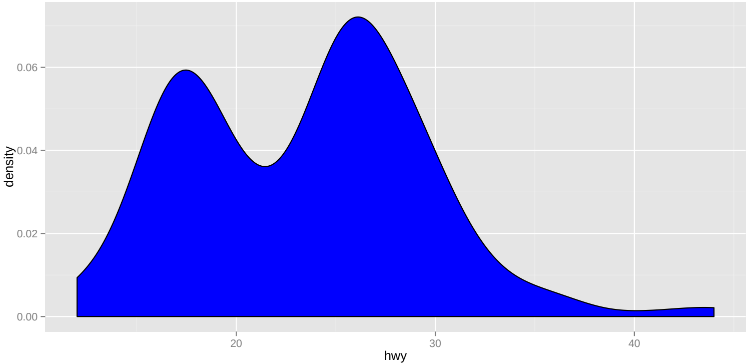
- A one-sample *t*-test tests whether a sequence of real-valued samples come from a distribution with a mean (population average) equal to some fixed value μ_0
  - the t-test is a parametric test: it assumes the data have a normal distribution
- Example: *Is the mean highway miles-per-gallon 25?*

```
t.test(mpg$hwy, mu=25)

##
##  One Sample t-test
##
## data:  mpg$hwy
## t = -4.007, df = 233, p-value = 8.274e-05
## alternative hypothesis: true mean is not equal to 25
## 95 percent confidence interval:
##  22.67 24.21
## sample estimates:
## mean of x
##     23.44
```

See Dalgaard (2008) section 5.1

Is hwy normally distributed?

```
qplot(hwy, data=mpg, geom="density", fill=I("blue"))
```



See Dalgaard (2008) section 4.2.3 for ways of determining non-normality

Wilcoxon signed rank test

- A Wilcoxon signed rank test tests whether a sequence of real-valued samples come from a distribution with a *median* equal to μ_0
  - it performs a binomial test on whether the observations are greater than μ_0
- It is more robust but less powerful than the t-test

```
wilcox.test(mpg$hwy, mu=25)

##
##  Wilcoxon signed rank test with continuity correction
##
## data:  mpg$hwy
## V = 8224, p-value = 4.535e-05
## alternative hypothesis: true location is not equal to 25
```

See Dalgaard (2008) section 5.2

Computational exercise (2)

- Run the `t.test` and the `wilcox.test` on the `mpg$hwy` data, but with the null hypothesis that $\mu = 24$. Is this result what you'd expect? Can you explain what's happened?
- The `wilcox.test` can compute a non-parametric confidence interval for the (pseudo)median. Use R's help facility to find out how to get this confidence interval, and then compare the confidence intervals from `t.test` and `wilcox.test` on the `mpg$hwy` data.
- Now run the `t.test` and the `wilcox.test` on the `mpg$cty` data, with the null hypothesis that $\mu = 16$. Is this result what you'd expect? You might need to plot `mpg$hwy` and `mpg$cty` to understand why they behave differently (see previous slides for how to make density plots)

Two-sample *t*-test

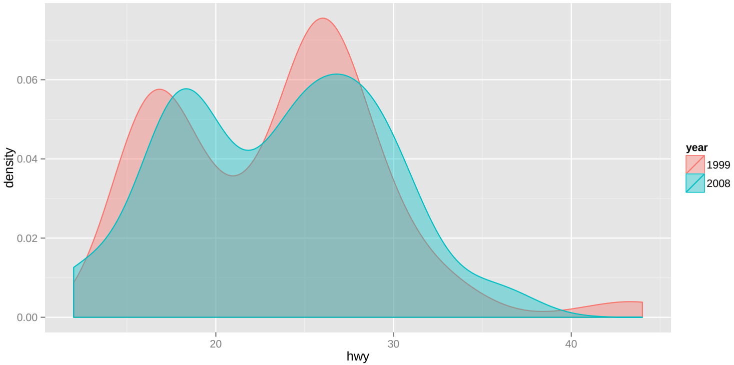
- A two-sample *t*-test tests whether two sequences of real-valued samples come from distributions with different means.
  - this is a parametric test, which assumes that both sequences are normally distributed
- Example: *Is the highway miles-per-gallon better in 2008 than in 1999?*

```
t.test(hwy~year, data=mpg)

##
##  Welch Two Sample t-test
##
## data:  hwy by year
## t = -0.0329, df = 231.6, p-value = 0.9738
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.563  1.512
## sample estimates:
## mean in group 1999 mean in group 2008
##              23.43              23.45
```

See Dalgaard (2008) section 5.3

Is the normal approximation justified?

```
d = data.frame(hwy=mpg$hwy, year=factor(mpg$year))
qplot(hwy, data=d, geom="density", fill=year, colour=year, alpha=I(0.4))
```



Two-sample Wilcoxon test

- A two-sample Wilcoxon test tests whether two sequences of real-valued samples come from distributions with different medians
  - ▸ it rank orders the values, and tests the distribution of ranks
  - ⇒ tied values can be problematic for this test
- It is more robust but less powerful than the two-sample t-test

```
wilcox.test(hwy~year, data=mpg)

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  hwy by year
## W = 6526, p-value = 0.5377
## alternative hypothesis: true location shift is not equal to 0
```

See Dalgaard (2008) section 5.5

Computational exercise (3)

- Do SUVs (what we call 4x4s) have better or worse fuel economy than minivans? From the `mpg` dataset, prepare a subset of the data that only contains `class==minivan` and `class==suv`. (Hint: use the `subset` function described in the previous slides). Then perform t-tests and two-sample Wilcoxon tests on both the city and highway miles-per-gallon values.

Paired t-test

- A paired t-test is used when there are two measurements on each item. The statistics are basically one-sample tests of the difference between the two measurements.
  - paired tests are more powerful than unpaired tests
  - this is a parametric test, which assumes that the differences are normally distributed
- Example: *Is the highway miles-per-gallon better than the city miles-per-gallon?*

```
t.test(mpg$hwy, mpg$cty, paired=TRUE)

##
##  Paired t-test
##
## data:  mpg$hwy and mpg$cty
## t = 44.49, df = 233, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  6.290 6.873
## sample estimates:
## mean of the differences
##                   6.581
```

See Dalgaard (2008) section 5.6

The matched-pairs Wilcoxon test

- The matched-pairs Wilcoxon test is a non-parametric version of the paired t-test

```
wilcox.test(mpg$hwy, mpg$cty, paired=TRUE)

##
##  Wilcoxon signed rank test with continuity correction
##
## data:  mpg$hwy and mpg$cty
## V = 27495, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

See Dalgaard (2008) section 5.7

Computational exercise (4)

- What happens if you run the unpaired two-sample versions of the t-test and the Wilcoxon test on the `mpg$hwy` and `mpg$cty` data? How do the results compare with the results when you analyse the data as paired (which, of course, it is)? Pay attention to the significance levels and the confidence intervals. Can you explain your results?
- The `diamonds` data set, distributed with the `ggplot2` library, contains measurements on 54,000 diamonds (run `help(diamonds)` to find out more about it). Each diamond has a `depth` and `table` measurement. What is the confidence interval for the difference in means and the difference in (pseudo)medians? Can you explain why these are different? (Hint: plot the density of the the differences in `depth` and `table`).

Class discussion (4)

- What would be the appropriate tests in the following situations:
  - We want to know if age affects a child's hearing acuity, so we measure the hearing acuity for a group of one year olds, wait a year and measure the same children again when they are two.
  - We want to do the same thing as above, but we can't wait a year, so we measure the hearing acuity of a group of one year olds, and another group of two year olds.
  - I'd like to know if modifications to a syntactic parser have improved its performance. How should I do this, and what tests should I perform?
  - We have developed two different experimental treatments for Ebola. How could we tell which works better, and what tests should we use?
  - Can you think of any situations where you'd use a binomial test for a single proportion? (Perhaps indirectly?)
- What are the advantages and disadvantages of experimental designs that use paired tests?

Outline

Introduction

Hypothesis tests and confidence intervals

Tests for tabular data

One and two sample tests

Linear regression

Logistic regression

Mixed effects models

Conclusions

What is linear regression?

- Regression estimates the relationship between two or more random variables
- In simple linear regression there is a *response or predicted variable* Y and a *explanatory or predictor variable* X, which we assume are related by:

  $$Y \quad \sim \quad \alpha + \beta X + N(0, \sigma^2)$$

  where $N(0, \sigma^2)$ is a normal distribution with zero mean and standard deviation σ.
- Given data $D = ((x_1, y_1), \ldots, (x_n, y_n))$ the goal of simple linear regression is to find the *regression coefficient* β and the *intercept* α
  - β is the slope of the line relating X and Y
  - α is the expected value of Y when $X = 0$

Regression on highway and city mpg

```
lm(hwy~cty, data=mpg)

##
## Call:
## lm(formula = hwy ~ cty, data = mpg)
##
## Coefficients:
## (Intercept)          cty
##       0.892        1.337
```

- This says:

$$\text{Hwy} \quad \sim \quad 1.337\,\text{Cty} + 0.892 + N(0, \sigma^2)$$

See Dalgaard (2008) section 6.1

Understanding a model formula

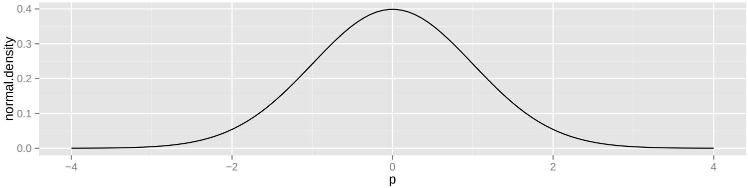
- "$\sim$" means "distributed as" or "distributed according to"
- So a formula like

$$\text{Hwy} \quad \sim \quad 1.337\,\text{Cty} + 0.892 + N(0, \sigma^2)$$

can be read as: *to generate a sample value for* Hwy, *sum the following values:*

  - 1.337 × Cty
  - 0.892
  - a sample from $N(0, \sigma^2)$ (a normal distribution with variance σ^2)

```r
df = expand.grid(p=seq(-4, 4, length.out=100))
df$normal.density = dnorm(df$p, mean=0, sd=1)
qplot(p, normal.density, data=df, geom="line")
```



Information about a regression

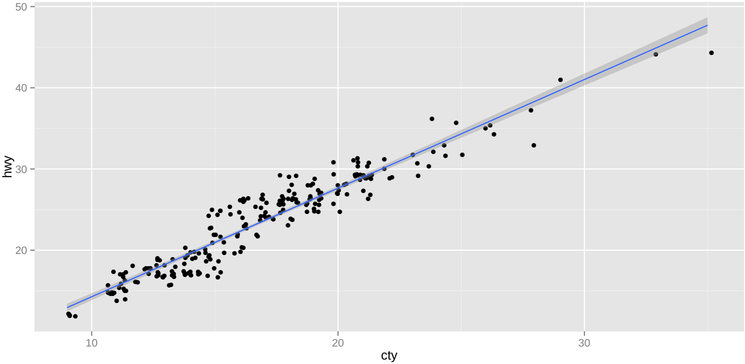
```
m = lm(hwy~cty, data=mpg)
summary(m)

##
## Call:
## lm(formula = hwy ~ cty, data = mpg)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.341 -1.279  0.021  1.034  4.046
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.892      0.469     1.9    0.058 .
## cty            1.337      0.027    49.6   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.75 on 232 degrees of freedom
## Multiple R-squared:  0.914,Adjusted R-squared:  0.913
## F-statistic: 2.46e+03 on 1 and 232 DF,  p-value: <2e-16
```

See Dalgaard (2008) section 6.1

Plotting a simple linear regression

- By default, ggplot plots a *95% confidence interval of the mean value of the prediction* (which is not the standard deviation of the data)
- See stat_smooth documentation for more informationm

```
ggplot(mpg, aes(x = cty, y = hwy)) + geom_jitter() +
    stat_smooth(method = "lm")
```



Using predict to predict values from a model

- What highway mpg would a car with 20 city mpg get?

```
m = lm(hwy~cty, data=mpg)
predict(m, newdata=data.frame(cty=c(20)))

##     1
## 27.64

predict(m, newdata=data.frame(cty=c(20)), se.fit=TRUE)

## $fit
##     1
## 27.64
##
## $se.fit
## [1] 0.1425
##
## $df
## [1] 232
##
## $residual.scale
## [1] 1.752
```

Computational exercise (5)

- Run this regression the "other way around", i.e., where `hwy` is the predictor variable and `cty` is the predicted variable, and compare the results with the regression in these notes. Can you explain the relationship between the regression coefficients from this regression, and the regression where `cty` is the predictor variable and `hwy` is the predicted variable?
- Use `predict` to predict the city mpg for a car with 30 highway mpg. What is the standard error of this prediction?

Multiple linear regression

- Example: *What's the relationship between engine displacement, number of cylinders and highway mpg?*

```
lm(hwy~displ+cyl, data=mpg)

##
## Call:
## lm(formula = hwy ~ displ + cyl, data = mpg)
##
## Coefficients:
## (Intercept)        displ          cyl
##       38.22        -1.96        -1.35
```

$$\text{Hwy} \quad \sim \quad 38.22 - 1.96\text{Displ} - 1.35\text{Cyl} + N(0, \sigma^2)$$

Model formulae in R

- General form of a model formula:

$$\text{Response variable} \quad \sim \quad \text{combination of explanatory variables}$$

- The "combination of explanatory variables" includes the following:

symbol	**example**	**meaning**
$+$	$+x$	include variable x
$-$	$-x$	exclude variable x
$:$	$x : y$	interaction between x and y
$*$	$x * y$	same as $x + y + x : y$
1	-1	exclude the intercept
.	.	include all variables in data
$f()$	$log(x)$	include $log(x)$
$I()$	$I(x * x)$	include x^2 as variable

Class discussion (5)

- In a model formula, `-1` means “remove the intercept term”. When and why might you want to do this?

```
lm(hwy~displ, data=mpg)

##
## Call:
## lm(formula = hwy ~ displ, data = mpg)
##
## Coefficients:
## (Intercept)        displ
##       35.70        -3.53

lm(hwy~displ-1, data=mpg)

##
## Call:
## lm(formula = hwy ~ displ - 1, data = mpg)
##
## Coefficients:
## displ
##  5.51
```

Interaction terms in multiple regression

- Example: *What's the relationship between engine displacement, number of cylinders and highway mpg?*

```
lm(hwy~displ*cyl, data=mpg)

##
## Call:
## lm(formula = hwy ~ displ * cyl, data = mpg)
##
## Coefficients:
## (Intercept)        displ          cyl    displ:cyl
##      52.193       -7.487       -3.340        0.756
```

$$\text{Hwy} \sim 52.193 - 7.487\,\text{Displ} - 3.340\,\text{Cyl} + 0.756\,\text{Cyl}\,\text{Displ} + N(0,\sigma^2)$$

Significance of interaction terms

- Example: *What's the relationship between engine displacement, number of cylinders and highway mpg?*

```
summary(lm(hwy~displ*cyl, data=mpg))

##
## Call:
## lm(formula = hwy ~ displ * cyl, data = mpg)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.780 -2.254 -0.541  2.144 13.649
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   52.193      3.134   16.66  < 2e-16 ***
## displ         -7.487      1.275   -5.87  1.5e-08 ***
## cyl           -3.340      0.580   -5.75  2.8e-08 ***
## displ:cyl      0.756      0.161    4.71  4.3e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.6 on 230 degrees of freedom
## Multiple R-squared:  0.64,Adjusted R-squared:  0.635
## F-statistic:  136 on 3 and 230 DF,  p-value: <2e-16
```

Computational exercise (6)

- What is the difference between the models `hwy ~ displ+cyl+cty` and `hwy ~ displ*cyl*cty`. Fit both models and check that your understanding is correct.
- Use `summary` to determine which of the predictors in each of the models has a parameter significantly different to zero. Is `cty` a "significant predictor"?

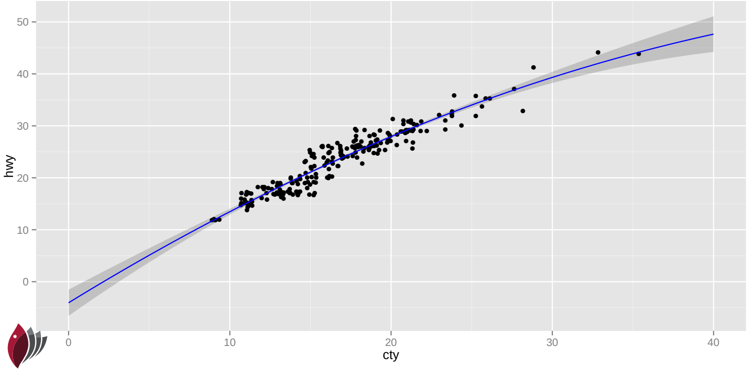
Using linear regression to find non-linear relationships

- The predictor variables can be non-linear functions of the data

```
m = lm(hwy ~ cty + I(cty*cty), data=mpg)
summary(m)

##
## Call:
## lm(formula = hwy ~ cty + I(cty * cty), data = mpg)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.289 -1.117 -0.053  0.918  3.918
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -4.05610    1.25893   -3.22   0.0015 **
## cty           1.90574    0.13740   13.87  < 2e-16 ***
## I(cty * cty) -0.01533    0.00364   -4.21  3.6e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.69 on 231 degrees of freedom
## Multiple R-squared:  0.92,Adjusted R-squared:  0.919
## F-statistic: 1.33e+03 on 2 and 231 DF,  p-value: <2e-16
```

Plotting a non-linear relationship

```
m = lm(hwy ~ cty + I(cty * cty), data = mpg)
d = expand.grid(cty = seq(0, 40, length.out = 100))
preds = predict(m, newdata = d, type = "response", se = TRUE)
d$y = preds$fit
d$ymin = preds$fit - 2 * preds$se.fit
d$ymax = preds$fit + 2 * preds$se.fit
ggplot(mpg, aes(x = cty, y = hwy)) + geom_jitter() + geom_ribbon(data = d,
    aes(y = y, ymin = ymin, ymax = ymax), alpha = 0.2) + geom_line(data = d,
    aes(y = y), colour = "blue")
```



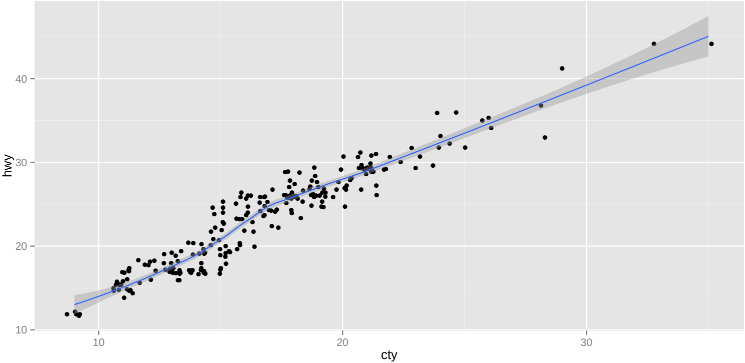
MACQUARIE UNIVERSITY

Computational exercise (7)

- This model is unbelievable, since it predicts that a vehicle could have negative highway mpg while having positive city miles per gallon
    - how can you force the city and highway mpg to be zero at the same time? (Hint: it involves a simple change to the model formula).
    - rerun the calculation above, but with your new model
    - what other changes does this make to the model's preductions?

Nonlinear curve fitting and “smoothing”

```
qplot(cty, hwy, data=mpg, geom=c("jitter","smooth"))

## geom_smooth:  method="auto" and size of largest group is <1000,
so using loess.  Use 'method = x' to change the smoothing method.
```



Discrete predictors

- Example: *What's the relationship between drive type and highway mpg?*

```
summary(mpg$drv)

##   4   f   r
## 103 106  25

lm(hwy~drv, data=mpg)

##
## Call:
## lm(formula = hwy ~ drv, data = mpg)
##
## Coefficients:
## (Intercept)         drvf         drvr
##       19.17         8.99         1.83
```

$$
\begin{aligned}
\text{Hwy} \quad &\sim \quad 19.17 + 8.99\ \llbracket \text{Drv=f} \rrbracket + 1.83\ \llbracket \text{Drv=r} \rrbracket + N(0, \sigma^2), \ \text{ where:} \\
& \quad\ \llbracket \text{Drv=r} \rrbracket = 1 \text{ if Drv=r and 0 if Drv} \neq \text{r}
\end{aligned}
$$

Significance of discrete predictors

- Example: *Do the different drive types affect highway mpg?*

```
summary(lm(hwy~drv, data=mpg))

##
## Call:
## lm(formula = hwy ~ drv, data = mpg)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -11.16  -2.17  -1.00   1.96  15.84
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   19.175      0.404    47.5   <2e-16 ***
## drvf           8.986      0.567    15.8   <2e-16 ***
## drvr           1.825      0.913     2.0    0.047 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.1 on 231 degrees of freedom
## Multiple R-squared:  0.531,Adjusted R-squared:  0.527
## F-statistic:  131 on 2 and 231 DF,  p-value: <2e-16
```

MACQUARIE
UNIVERSITY

Reordering levels to obtain desired predictor comparisons

- Example: *Which drive types are different to* `r` *in terms of highway mpg?*

```
mpg$new.drv = relevel(mpg$drv, 3)
summary(mpg$new.drv)

##   r   4   f
##  25 103 106

summary(lm(hwy~new.drv, data=mpg))

##
## Call:
## lm(formula = hwy ~ new.drv, data = mpg)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -11.16  -2.17  -1.00   1.96  15.84
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   21.000      0.819   25.63  < 2e-16 ***
## new.drv4      -1.825      0.913   -2.00    0.047 *
## new.drvf       7.160      0.911    7.86  1.4e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Computational exercise (8)

- How does highway mpg vary depending on manufacturer?
- The manufacturers in the `mpg` data are ordered alphabetically, so `audi` comes first. Arguably `volkswagen` is a more mainstream manufacturer. Reorder the levels so `volkswagen` is the baseline manufacturer. Do any manufacturers have a significantly better highway mpg than `volkswagen`?

Combining categorical predictors (1)

- Example: *What is the relationship between transmission, drive type and highway mpg?*

```
summary(lm(hwy~trans+drv, data=mpg))

##
## Call:
## lm(formula = hwy ~ trans + drv, data = mpg)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.493 -2.493 -0.586  2.240 14.507
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         19.038      1.826   10.43   <2e-16 ***
## transauto(l3)       -0.800      3.247   -0.25    0.806
## transauto(l4)       -1.307      1.812   -0.72    0.471
## transauto(l5)       -0.233      1.895   -0.12    0.902
## transauto(l6)       -2.712      2.383   -1.14    0.256
## transauto(s4)        3.708      2.859    1.30    0.196
## transauto(s5)        0.454      2.840    0.16    0.873
## transauto(s6)        1.627      2.007    0.81    0.418
## transmanual(m5)      2.114      1.823    1.16    0.248
## transmanual(m6)      1.007      1.977    0.51    0.611
```

Combining categorical predictors (2)

- Example: *What is the relationship between transmission, drive type and highway mpg?*

```
summary(lm(hwy~trans*drv, data=mpg))

##
## Call:
## lm(formula = hwy ~ trans * drv, data = mpg)
##
## Residuals:
##    Min      1Q Median      3Q     Max
## -9.676 -2.219 -0.676  1.644 14.324
##
## Coefficients: (6 not defined because of singularities)
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)              18.193      2.626    6.93  5.2e-11
## transauto(l3)            -0.800      3.212   -0.25    0.804
## transauto(l4)            -0.222      2.708   -0.08    0.935
## transauto(l5)            -0.227      2.721   -0.08    0.934
## transauto(l6)             0.807      3.777    0.21    0.831
## transauto(s4)             7.807      3.777    2.07    0.040
## transauto(s5)            -0.193      4.652   -0.04    0.967
## transauto(s6)             3.807      3.001    1.27    0.206
## transmanual(m5)           3.093      2.757    1.12    0.263
## transmanual(m6)           0.950      2.180    0.43    0.665
```

Using anova to compare nested models

- Model m0 is a submodel “nested” within model m1, so we can compare the fit of these two models with anova

```
m0 = lm(hwy~trans+drv, data=mpg)
m1 = lm(hwy~trans*drv, data=mpg)
anova(m0, m1)

## Analysis of Variance Table
##
## Model 1: hwy ~ trans + drv
## Model 2: hwy ~ trans * drv
##   Res.Df  RSS Df Sum of Sq   F Pr(>F)
## 1    222 3343
## 2    210 3096 12       247 1.4   0.17
```

See Dalgaard (2008) section 12.6

Computational exercise (9)

- We can use anova to compare nested models in general, but it's often not necessary with continuous predictors as lm gives a significance level for each parameter weight.
  - anova is useful with categorical predictors because each categorical predictor is usually associated with many parameters (one for each category, except for the baseline)
- Use anova to calculate the significance of the interaction term in the linear model hwy~displ*cyl, and compare that value with the significance of the parameter associated with the interaction term computed by lm.

Combining continuous and categorical predictors (1)

- Example: *What's the relationship between displacement, drive type and highway mpg?*

```
summary(lm(hwy~new.drv+displ, data=mpg))

##
## Call:
## lm(formula = hwy ~ new.drv + displ, data = mpg)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.000 -1.907 -0.394  1.578 13.921
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   36.083      1.287   28.04   <2e-16 ***
## new.drv4      -5.258      0.734   -7.17    1e-11 ***
## new.drvf      -0.467      0.892   -0.52      0.6
## displ         -2.914      0.218  -13.35   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.08 on 230 degrees of freedom
## Multiple R-squared:  0.736,Adjusted R-squared:  0.732
## F-statistic:  213 on 3 and 230 DF,  p-value: <2e-16
```

Combining continuous and categorical predictors (2)

- Example: *What's the relationship between displacement, drive type and highway mpg?*

```
lm(hwy ~ drv*displ, data=mpg)

##
## Call:
## lm(formula = hwy ~ drv * displ, data = mpg)
##
## Coefficients:
## (Intercept)         drvf         drvr        displ
##      30.683        6.695       -4.903       -2.878
##  drvf:displ   drvr:displ
##      -0.724        1.955
```

- The extra "crossed" terms allows the slope of the line to vary with Drv. It's equivalent to fitting 3 different linear models, one for each value of `drv`!

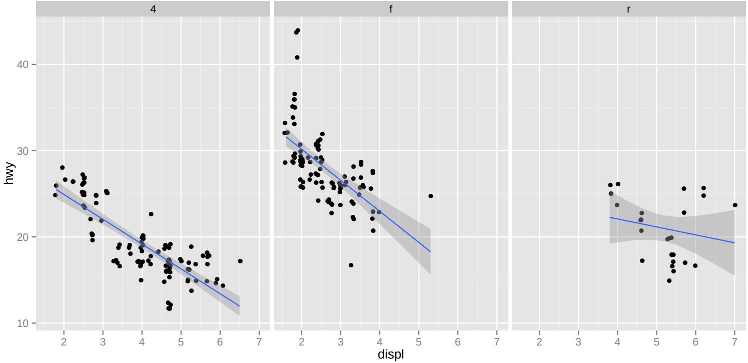
$$\begin{aligned}\text{Hwy} \quad \sim \quad & 30.7 + 6.7\ \llbracket\text{Drv}=\text{f}\rrbracket - 4.9\ \llbracket\text{Drv}=\text{r}\rrbracket \\ & - 2.88\,\text{Displ} - 0.72\,\text{Displ}\ \llbracket\text{Drv}=\text{f}\rrbracket + 1.96\,\text{Displ}\ \llbracket\text{Drv}=\text{r}\rrbracket\end{aligned}$$

MACQUARIE
UNIVERSITY

Plotting linear and categorical predictors

- Example: *What's the relationship between displacement, drive type and highway mpg?*

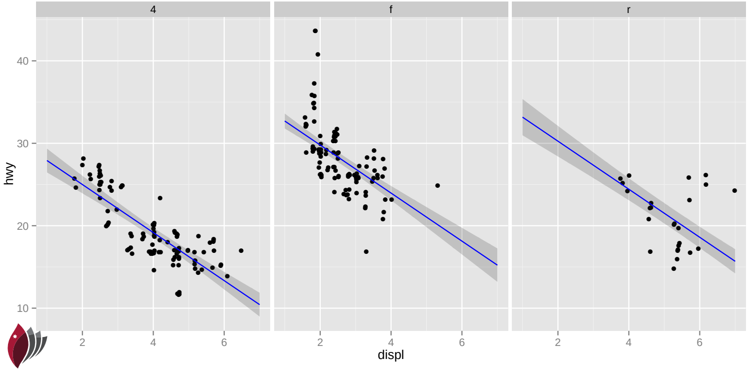
```
ggplot(mpg, aes(x = displ, y = hwy)) + geom_jitter() +
    facet_wrap(~drv) + stat_smooth(method = "lm")
```



- Warning: this corresponds to the model formula `hwy ~ drv*displ`. It's equivalent to fitting 3 different linear models, one for each value of `drv`!

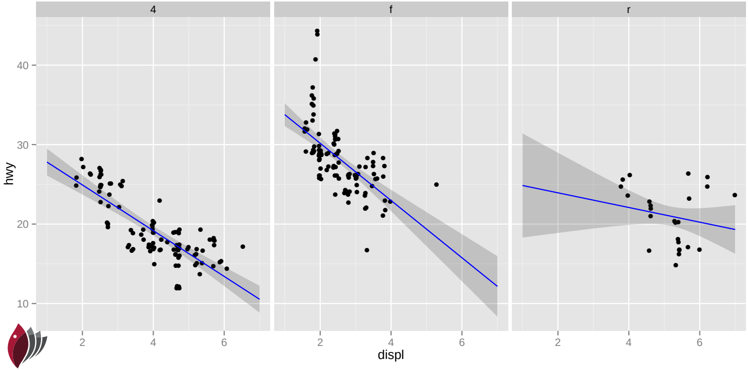
Explicitly plotting a model’s predictions

```
m = lm(hwy ~ drv + displ, data = mpg)
d = expand.grid(displ = seq(1, 7, length.out = 100), drv = levels(mpg$drv))
preds = predict(m, newdata = d, type = "response", se = TRUE)
d$y = preds$fit
d$ymin = preds$fit - 2 * preds$se.fit
d$ymax = preds$fit + 2 * preds$se.fit
ggplot(mpg, aes(x = displ, y = hwy)) + geom_jitter() + facet_wrap(~drv) +
    geom_ribbon(data = d, aes(y = y, ymin = ymin, ymax = ymax), alpha = 0.2) +
    geom_line(data = d, aes(y = y), colour = "blue")
```



Modifying the model to allow interaction

```
m = lm(hwy ~ drv * displ, data = mpg)
d = expand.grid(displ = seq(1, 7, length.out = 100), drv = levels(mpg$drv))
preds = predict(m, newdata = d, type = "response", se = TRUE)
d$y = preds$fit
d$ymin = preds$fit - 2 * preds$se.fit
d$ymax = preds$fit + 2 * preds$se.fit
ggplot(mpg, aes(x = displ, y = hwy)) + geom_jitter() + facet_wrap(~drv) +
    geom_ribbon(data = d, aes(y = y, ymin = ymin, ymax = ymax), alpha = 0.2) +
    geom_line(data = d, aes(y = y), colour = "blue")
```



Warning: whether a predictor is significant depends on the other predictors! (1)

```
summary(lm(hwy~displ, data=mpg))

##
## Call:
## lm(formula = hwy ~ displ, data = mpg)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.104 -2.165 -0.224  2.059 15.010
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   35.698      0.720    49.5   <2e-16 ***
## displ         -3.531      0.195   -18.1   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.84 on 232 degrees of freedom
## Multiple R-squared:  0.587,Adjusted R-squared:  0.585
## F-statistic:  329 on 1 and 232 DF,  p-value: <2e-16
```

Warning: whether a predictor is significant depends on the other predictors! (2)

```
summary(lm(hwy~displ+cty, data=mpg))

##
## Call:
## lm(formula = hwy ~ displ + cty, data = mpg)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.312 -1.242  0.005  1.030  4.124
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.1514     1.2127    0.95     0.34
## displ        -0.0343     0.1479   -0.23     0.82
## cty           1.3291     0.0449   29.60   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.76 on 231 degrees of freedom
## Multiple R-squared:  0.914,Adjusted R-squared:  0.913
## F-statistic: 1.22e+03 on 2 and 231 DF,  p-value: <2e-16
```

Interaction between predictors

- In general, adding a new predictor reduces the significance of the other predictors
  - the significance of a predictor depends on the *amount of the variance in the predicted variable it explains*
  - the new predictor may be a better "explanation" of the variance than the old predictors
- Highly correlated predictors have very large confidence intervals in predictor weights
  - if two predictor variables are exact copies (e.g., by mistake) the confidence intervals on their predictor weights is infinite

$$Y \quad \sim \quad \alpha + \beta_1 X_1 + \beta_2 X_2 + N(0, \sigma^2)$$

$\Rightarrow$ neither will be significant, even if both would be significant on their own

Class discussion (6)

- Suppose I'm interested in the relationship between hearing impairment, phonetic awareness and reading ability in children. What model formula should be used here? How should I interpret the significance levels?
- Is there a danger of "just so" stories justifying combinations of predictors that yield significant results? If so, what might be done to reduce this?
  - would *experiment pre-registration* help?

Searching for significance and exploratory data analysis

- It's easy to run many regressions with different combinations of predictors
- But it's not clear what significance levels mean if you do this
  - you shouldn't trust the significance levels the software returns if you do this
- *Exploratory data analysis* is searching for relationships between variables (instead of testing whether a relationship exists)

Automatically searching for significant predictors

- step searches for significant combinations of predictors
    - verb|direction="forward"| means add variables, direction="backward" means remove variables, and direction="both" means both add and remove variables
    - it attempts to minimise the *Akaike Information Criterion* (AIC)
    - recall that . abbreviates “all possible predictors”
- This is very much exploratory data analysis; minor changes to the data can produce very different results!

```
m = lm(hwy ~ cty+manufacturer+model+displ+cyl+trans+drv+class+year, data=mpg)
step(m, direction="both")

## Start:  AIC=62.51
## hwy ~ cty + manufacturer + model + displ + cyl + trans + drv +
##     class + year
##
##
## Step:  AIC=62.51
## hwy ~ cty + manufacturer + model + displ + cyl + trans + class +
##     year
##
##
## Step:  AIC=62.51
## hwy ~ cty + model + displ + cyl + trans + class + year
##
```

Computational exercise (10)

- Run the step procedure from the previous slide to see how it adds and removes predictors in models. What set of predictors does it converge on?
- You can use the step procedure on models with complex interaction terms as well. Try it on the model formula `displ ~ class*trans*drv`. What do the results mean?

Outline

Introduction

Hypothesis tests and confidence intervals

Tests for tabular data

One and two sample tests

Linear regression

Logistic regression

Mixed effects models

Conclusions

Why logistic regression?

- Logistic regression is used to predict *binary responses*, such as:
  - an experimental subject completes a task correctly,
  - a patient dies after taking a drug,
  - a program crashes under particular circumstances, etc.
- It can also be used to avoid the normality assumptions in linear regression, as in:
  - identifying the circumstances in which program 1 performs better than program 2

See Dalgaard (2008) chapter 13

Preparing some binary data

```r
levels(mpg$manufacturer)
```

```
[1] "audi" "chevrolet" "dodge" "ford" [5] "honda" "hyundai" "jeep" "land rover"
[9] "lincoln" "mercury" "nissan" "pontiac" [13] "subaru" "toyota" "volkswagen"
```

```r
american = c("chevrolet", "dodge", "ford", "jeep", "lincoln",
    "mercury", "pontiac")
mpg$american = (mpg$manufacturer %in% american)
xtable(mpg[17:20, c("manufacturer", "model", "american")])
```

	manufacturer	model	american
17	audi	a6 quattro	FALSE
18	audi	a6 quattro	FALSE
19	chevrolet	c1500 suburban 2wd	TRUE
20	chevrolet	c1500 suburban 2wd	TRUE

MACQUARIE
UNIVERSITY

Does fuel efficiency predict American?

```
summary(glm(american~hwy+cty, data=mpg, family=binomial))

##
## Call:
## glm(formula = american ~ hwy + cty, family = binomial, data = mpg)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.358  -0.604  -0.151   0.651   2.755
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   9.2022     1.2299    7.48  7.3e-14 ***
## hwy           0.1272     0.0958    1.33     0.18
## cty          -0.7632     0.1687   -4.52  6.1e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 320.00  on 233  degrees of freedom
## Residual deviance: 195.32  on 231  degrees of freedom
## AIC: 201.3
##
```

What does logistic regression compute?

- Logistic regression predicts the probability p that the response variable is true:

$$logit(p) \quad = \quad \alpha + \beta_1 X_1 + \ldots + \beta_n X_n,$$

where $logit(p) = \log(p/1 - p)$ is known as the *log odds* of p.

  - the log odds of a probability vary from $-\infty$ to ∞, while a probability is always between 0 and 1

⇒ Logistic regression is just like linear regression, except that it *predicts the log odds of the response variable*

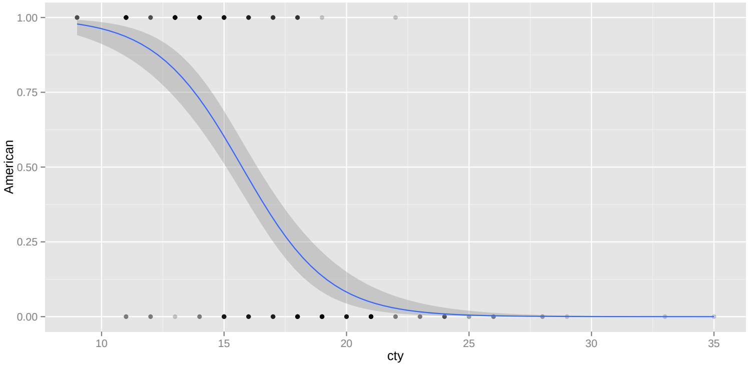
- So the previous logistic regression showed:

$$logit(p_{American}) \quad = \quad 9.2 + 0.13\,\mathrm{Hwy} - 0.76\,\mathrm{Cty}$$

I.e., an increase in Hwy increases the probability of American, while an increase in Cty decreases the probability of American.

Plotting a logistic regression

```
mpg$American = 1 * mpg$american  # convert Boolean to 0/1 values
ggplot(mpg, aes(x = cty, y = American)) + geom_point(alpha = I(0.2))
    stat_smooth(method = "glm", family = "binomial")
```



Categorical predictors in logistic regression

```
summary(glm(american~drv+displ, data=mpg, family=binomial))

##
## Call:
## glm(formula = american ~ drv + displ, family = binomial, data = mpg)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.926  -0.412  -0.203   0.303   2.239
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -9.501      1.436   -6.61  3.7e-11 ***
## drvf           1.282      0.555    2.31    0.021 *
## drvr          16.693   1061.712    0.02    0.987
## displ          2.416      0.347    6.95  3.6e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 320.00  on 233  degrees of freedom
## Residual deviance: 143.24  on 230  degrees of freedom
## AIC: 151.2
```

Warning: the significance of a predictor depends on the other predictors! (1)

```
summary(glm(american~hwy, data=mpg, family=binomial))

##
## Call:
## glm(formula = american ~ hwy, family = binomial, data = mpg)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.100  -0.715  -0.353   0.639   2.247
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   6.6189     0.8737    7.58  3.6e-14 ***
## hwy          -0.3020     0.0381   -7.94  2.1e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

Warning: the significance of a predictor depends on the other predictors! (2)

```
summary(glm(american~hwy+cty, data=mpg, family=binomial))

##
## Call:
## glm(formula = american ~ hwy + cty, family = binomial, data = mpg)
##
## Deviance Residuals:
##    Min       1Q   Median       3Q      Max
## -2.358  -0.604  -0.151   0.651   2.755
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   9.2022     1.2299    7.48  7.3e-14 ***
## hwy           0.1272     0.0958    1.33     0.18
## cty          -0.7632     0.1687   -4.52  6.1e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Class discussion (7)

- Suppose I wanted to identify the circumstances in which my new syntactic parser performs better than a baseline parser. How could I use `glm` to do this?
  - Hint: think of the Wilcoxon paired sample test (i.e., the non-parametric test for difference in paired samples)
- What are the advantages and disadvantages of using logistic regression on the binary "better/worse" data, as compared to linear regression on the differences in (say) accuracy scores.
- How would you handle tied scores (i.e., cases where the two parsers behaved identically)?

Computational exercise (11)

- Add a Boolean variable `pre.2000` in the `mpg` data frame, which is `TRUE` when `year > 2000`. Use `head` and `summary` to check that you have correctly assigned this variable (there should be 117 pre-2000 rows and 117 post-2000 rows).
- Now develop different models for predicting whether a car is pre-2000 or post-2000. Can you find predictors that seem highly predictive?
- You can use `anova` to compare nested `glm` models and `step` to search for `glm` models in the same way as we did for `lm` models

Outline

Introduction

Hypothesis tests and confidence intervals

Tests for tabular data

One and two sample tests

Linear regression

Logistic regression

Mixed effects models

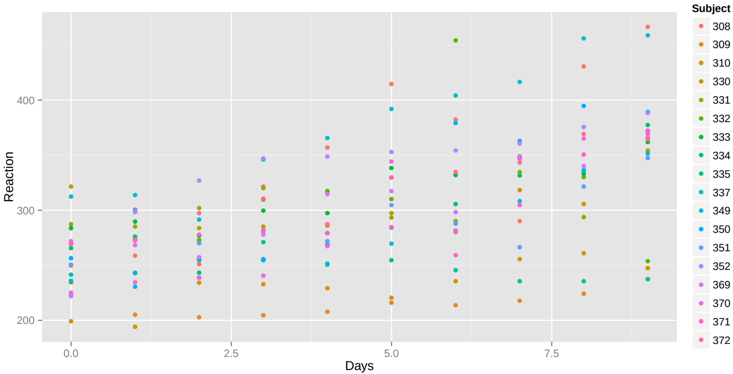
Conclusions

Why mixed effects models? (1)

- `sleepstudy` data: reaction times after varying days of sleep deprivation
- Reactions get slower as the days go on

```
install.packages("lme4")
library(lme4)
```

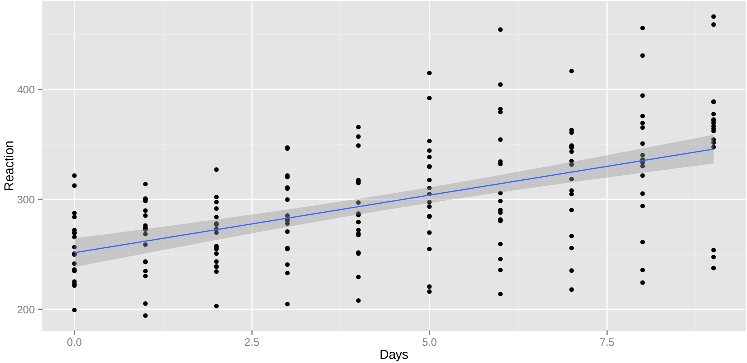
```
qplot(Days, Reaction, data = sleepstudy, geom = c("point"), colour = Subject)
```



Why mixed effects models? (2)

- We can fit a line to all the data, but it doesn't fit that well

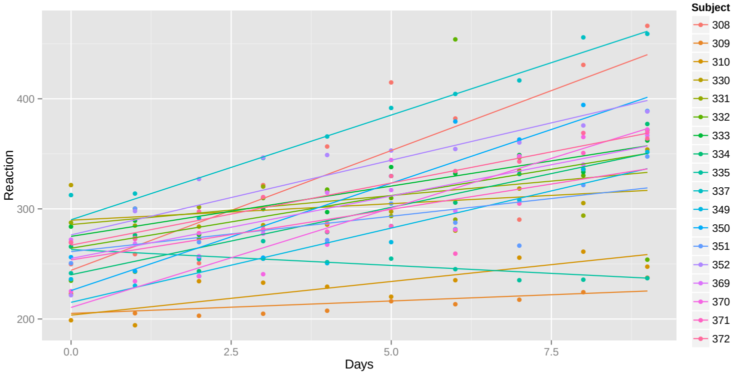
```
qplot(Days, Reaction, data = sleepstudy, geom = c("point")) +
    geom_smooth(method = "lm")
```



Why mixed effects models? (3)

- We can fit a line for each subject, i.e., make `Subject` a predictor
- But this tells us nothing about subjects in general, i.e., do reaction times get slower with more sleep deprivation?

```
qplot(Days, Reaction, data = sleepstudy, geom = c("point"), colour = Subject) +
    geom_smooth(method = "lm", se = FALSE)
```



Why mixed effects models? (4)

- A *mixed effects model* estimates a group mean slope from which the individual slopes are generated

```
m <- lmer(Reaction ~ Days + (Days | Subject), sleepstudy)
m

## Linear mixed model fit by REML ['lmerMod']
## Formula: Reaction ~ Days + (Days | Subject)
##    Data: sleepstudy
## REML criterion at convergence: 1744
## Random effects:
##  Groups   Name        Std.Dev. Corr
##  Subject  (Intercept) 24.74
##           Days         5.92    0.07
##  Residual             25.59
## Number of obs: 180, groups:  Subject, 18
## Fixed Effects:
## (Intercept)         Days
##       251.4         10.5
```

MACQUARIE
UNIVERSITY

Fixed and random effects

- The predictors we used with linear and logistic regression are *fixed effects*
- The parameters associated with each subject are *random effects*
- A model with both fixed and random effects is called a *mixed effects model*

Outline

Introduction

Hypothesis tests and confidence intervals

Tests for tabular data

One and two sample tests

Linear regression

Logistic regression

Mixed effects models

Conclusions

Hypothesis testing and confidence intervals

- Formulate an alternative *null hypothesis* H_0, which would be true if your hypothesis H_1 is true
- Pick a test statistic T and a significance level α
- Run your experiment and collect data D
- Calculate the probability p of all possible experimental outcomes with values for T more extreme than $T(D)$
- If $p \leq \alpha$, then reject the null hypothesis H_0, i.e., accept your hypothesis H_1

- A $1 - \alpha$ *confidence interval* for a parameter is a range such that if the experiment were repeated many times, the parameter's true value would be outside this range at most α of the time

Hypothesis tests for tabular data

- `binom.test` tests whether binary samples are generated with a given probability
- `fisher.test`, `prop.test` and `chisq.test` can test whether two or more categorical samples came from the same source

One and two sample tests

- These tests are for *real-valued data*

null hypothesis	parametric test	non-parametric test
mean equal to fixed value	one-sample t-test	Wilcoxon signed rank test
equal means	two-sample t-test	two-sample Wilcoxon test
zero mean difference	paired t-test	matched pairs Wilcoxon test

Regression

- Linear and logistic regression describe the behaviour of a *predicted or response variable* in terms of one or more *predictor or explanatory variables*
- Confidence intervals on the predictor variable weights let us identify whether the contribution of a variable is significant
- Beware: whether a variable is a significant predictor depends on the other predictors in the model!

Statistics is not magic!

- It is easy to produce erroneous results by:
  - repeating (e.g., tweaking) an experiment until the results are significant
  - searching for (combinations of) predictors that are significant
- "Throwing in the kitchen sink" is probably not a good way to do statistical analysis
  - (close to) colinear predictors will appear insignificant