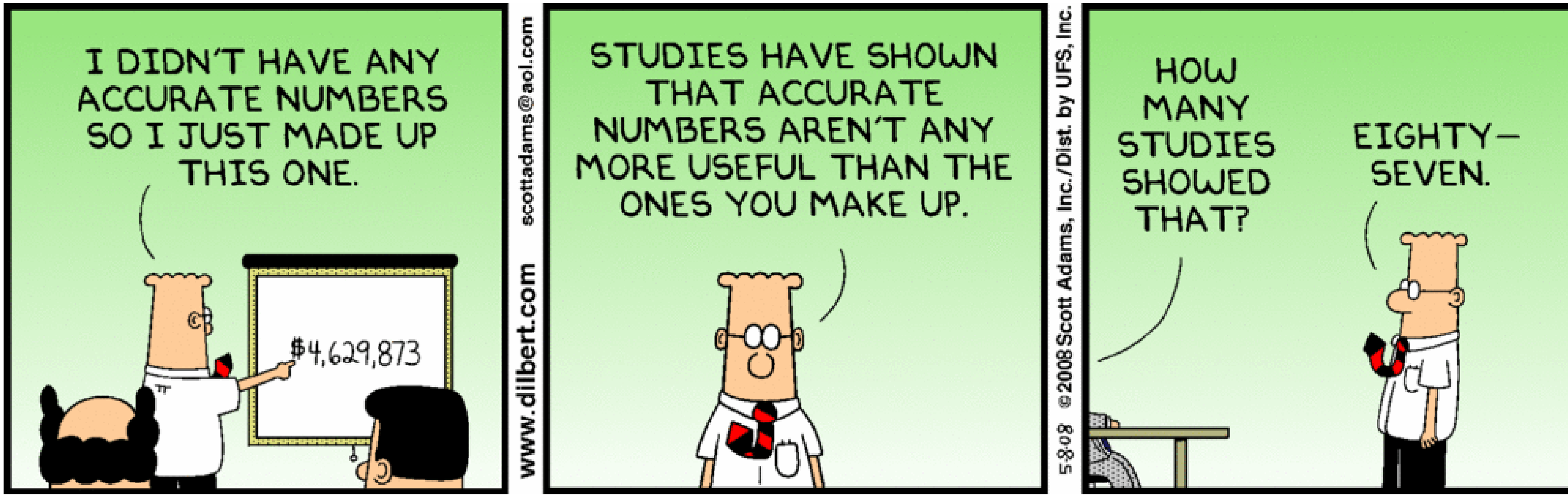
PART 4. STATISTICAL ANALYSIS


I DIDN'T HAVE ANY ACCURATE NUMBERS SO I JUST MADE UP THIS ONE.
$4,629,873
www.dilbert.com
scottadams@aol.com
STUDIES HAVE SHOWN THAT ACCURATE NUMBERS AREN'T ANY MORE USEFUL THAN THE ONES YOU MAKE UP.
5·8·08 © 2008 Scott Adams, Inc./Dist. by UFS, Inc.
HOW MANY STUDIES SHOWED THAT?
EIGHTY-SEVEN.

Part 4 Topics

- Basic statistical methods
  - descriptive statistics
  - frequency tables
  - correlation
  - t-tests
- ANOVA
- multiple linear regression

DESCRIPTIVE STATISTICS

summary(Salaries)

```
       rank     discipline yrs.since.phd    yrs.service
 AsstProf : 67   A:181      Min.   : 1.0   Min.   : 0.0
 AssocProf: 64   B:216      1st Qu.:12.0   1st Qu.: 7.0
 Prof     :266              Median :21.0   Median :16.0
                            Mean   :22.3   Mean   :17.6
                            3rd Qu.:32.0   3rd Qu.:27.0
                            Max.   :56.0   Max.   :60.0
    sex          salary
 Female: 39   Min.   : 57800
 Male  :358   1st Qu.: 91000
              Median :107300
              Mean   :113706
              3rd Qu.:134185
              Max.   :231545
```

library(psych)
describe (Salaries[c(3,4,6)])

```
                vars   n      mean       sd median
yrs.since.phd      1 397     22.31    12.89     21
yrs.service        2 397     17.61    13.01     16
salary             3 397 113706.46 30289.04 107300

                 trimmed      mad   min    max  range skew
yrs.since.phd      21.83    14.83     1     56     55 0.30
yrs.service        16.51    14.83     0     60     60 0.65
salary         111401.61 29355.48 57800 231545 173745 0.71

                kurtosis      se
yrs.since.phd      -0.81    0.65
yrs.service        -0.34    0.65
salary              0.18 1520.16
```

FREQUENCY TABLES

Frequency Tables

xtabs(~ rank, data=Salaries)

```
rank
 AsstProf AssocProf      Prof
       67        64       266
```

xtabs(~ rank + sex, data=Salaries)

```
 sex
rank        Female Male
  AsstProf      11   56
  AssocProf     10   54
  Prof          18  248
```

Frequency Tables (proportions)

tbl <- xtabs(~ rank + sex, data=Salaries)

```
              sex
rank          Female Male
  AsstProf        11   56
  AssocProf       10   54
  Prof            18  248
```

prop.table(tbl)

```
              sex
rank          Female  Male
  AsstProf     0.028 0.141
  AssocProf    0.025 0.136
  Prof         0.045 0.625
```

cells add up to 1

Frequency Tables (proportions)

prop.table(tbl, 1)

```
   sex
rank         Female  Male
  AsstProf    0.164 0.836
  AssocProf   0.156 0.844
  Prof        0.068 0.932
```

rows add up to 1

prop.table(tbl, 2)

```
   sex
rank         Female Male
  AsstProf     0.28 0.16
  AssocProf    0.26 0.15
  Prof         0.46 0.69
```

columns add up to 1

Chi-square test

chisq.test(tbl)

```
        Pearson's Chi-squared test

data:  tbl
X-squared = 8.5, df = 2, p-value = 0.01408
```

CORRELATION

Correlation

cor(*x*, use= , method=)

Option	Description
x	Matrix or data frame
use	Specifies the handling of missing data. everything (any correlation involving a case with missing values will be set to missing) complete.obs (listwise deletion) pairwise.complete.obs (pairwise deletion)
method	Specifies the type of correlation. The options are pearson, spearman, or kendall.

Correlation Matrix

cor(mtcars)

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
mpg	1.00	-0.85	-0.85	-0.78	0.681	-0.87	0.419	0.66	0.600	0.48	-0.551
cyl	-0.85	1.00	0.90	0.83	-0.700	0.78	-0.591	-0.81	-0.523	-0.49	0.527
disp	-0.85	0.90	1.00	0.79	-0.710	0.89	-0.434	-0.71	-0.591	-0.56	0.395
hp	-0.78	0.83	0.79	1.00	-0.449	0.66	-0.708	-0.72	-0.243	-0.13	0.750
drat	0.68	-0.70	-0.71	-0.45	1.000	-0.71	0.091	0.44	0.713	0.70	-0.091
wt	-0.87	0.78	0.89	0.66	-0.712	1.00	-0.175	-0.55	-0.692	-0.58	0.428
qsec	0.42	-0.59	-0.43	-0.71	0.091	-0.17	1.000	0.74	-0.230	-0.21	-0.656
vs	0.66	-0.81	-0.71	-0.72	0.440	-0.55	0.745	1.00	0.168	0.21	-0.570
am	0.60	-0.52	-0.59	-0.24	0.713	-0.69	-0.230	0.17	1.000	0.79	0.058
gear	0.48	-0.49	-0.56	-0.13	0.700	-0.58	-0.213	0.21	0.794	1.00	0.274
carb	-0.55	0.53	0.39	0.75	-0.091	0.43	-0.656	-0.57	0.058	0.27	1.000

be careful if categorical factors coded numerically have not been converted to factors

see corrplot package for methods of graphing correlation matricies

T TEST AND ANOVA

t-test

t.test(salary ~ sex, data=Salaries)

```
 Welch Two Sample t-test

data:  salary by sex
t = -3.2, df = 50, p-value = 0.002664
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -23038  -5138
sample estimates:
mean in group Female   mean in group Male
              101002               115090
```

ANOVA

fit <- aov(salary ~ rank, data=Salaries)
summary(fit)

```
             Df    Sum Sq  Mean Sq F value Pr(>F)
rank          2 1.43e+11 7.16e+10     128 <2e-16 ***
Residuals   394 2.20e+11 5.59e+08
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

TukeyHSD(fit) for post hoc comparisons

Factorial ANCOVA

```
fit <- aov(salary ~  yrs.since.phd +
     rank + sex + rank*sex, data=Salaries)
summary(fit)
```

```
               Df   Sum Sq  Mean Sq F value Pr(>F)
yrs.since.phd   1 6.39e+10 6.39e+10  113.76 <2e-16 ***
rank            2 7.96e+10 3.98e+10   70.91 <2e-16 ***
sex             1 9.07e+08 9.07e+08    1.62   0.20
rank:sex        2 5.06e+07 2.53e+07    0.05   0.96
Residuals     390 2.19e+11 5.61e+08
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Type I Sums of Squares

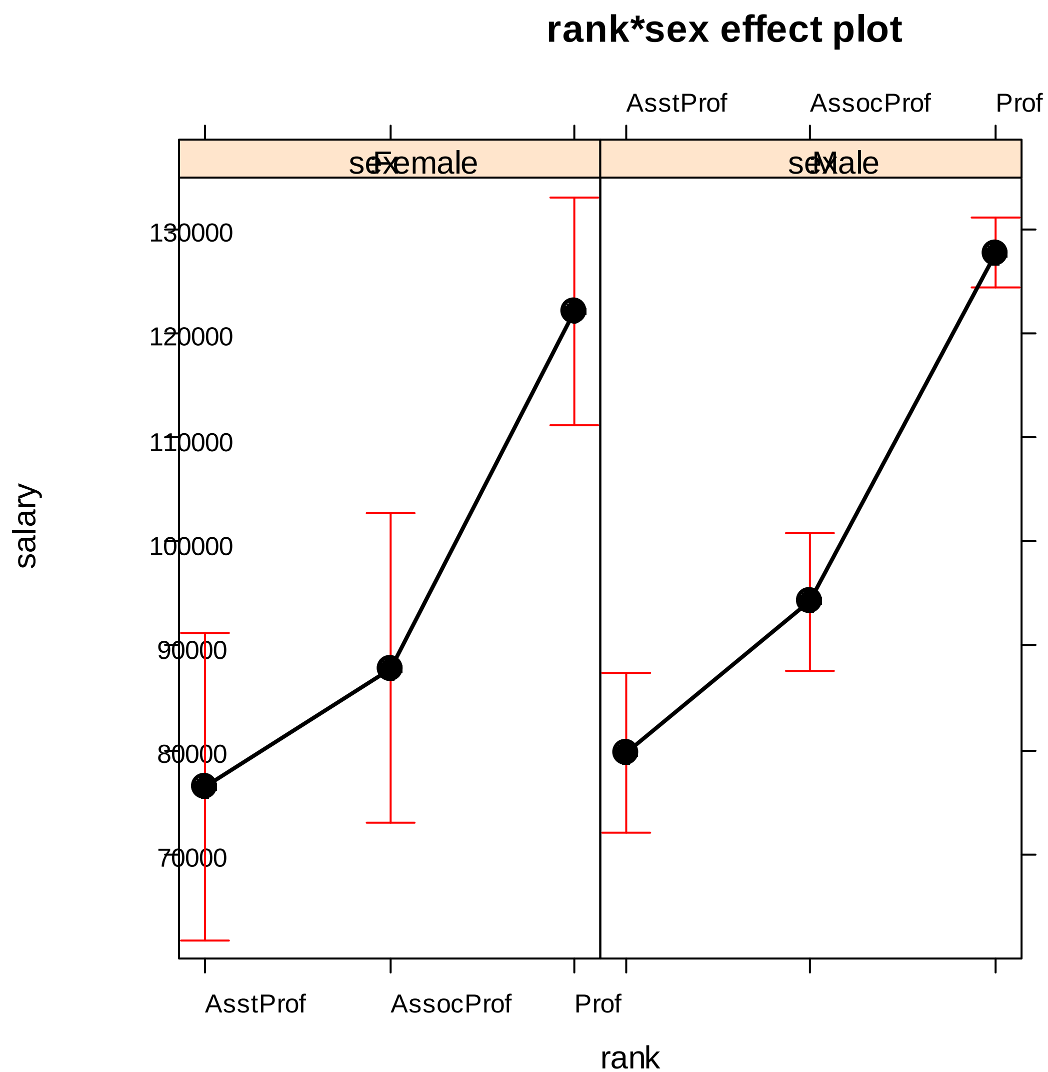
Factorial ANCOVA

library(car)
Anova(fit, type="III")

```
Response: salary
                 Sum Sq  Df F value  Pr(>F)
(Intercept)    6.72e+10   1  119.74 < 2e-16 ***
yrs.since.phd 2.89e+08   1    0.51    0.47
rank          1.54e+10   2   13.70 1.8e-06 ***
sex           9.43e+07   1    0.17    0.68
rank:sex      5.06e+07   2    0.05    0.96
Residuals     2.19e+11 390
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Type III Sums of Squares

library(effects)
plot(effect("rank*sex"), fit)



adjusted means

REGRESSION

Regression

Fit a model:

fit <- lm(y ~ x1 + x2 + x3, data=mydata)

Evaluate the model:

plot(fit) or use the many functions available in the **car** package

Use the model:

predict(fit, newdata)

```r
data(Prestige, package="car")
fit <- lm(prestige ~ education + income + women,
          data=Prestige)
summary(fit)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.794334   3.239089   -2.10    0.039 *
education    4.186637   0.388701   10.77  < 2e-16 ***
income       0.001314   0.000278    4.73  7.6e-06 ***
women       -0.008905   0.030407   -0.29    0.770
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.8 on 98 degrees of freedom
Multiple R-squared:  0.798,     Adjusted R-squared:  0.792
F-statistic:  129 on 3 and 98 DF,  p-value: <2e-16
```


Don't ask
But yes, I could use some help