Data Example on Hypothesis Tests and Confidence Intervals

1. Read Data

Let's look at an example concerning the number of species found on the various Galápagos Islands. There are 30 cases (Islands) and seven variables in the dataset. We start by reading the data into R and examining it:

```
library(faraway)
head(gala[,-2])
```

```
##              Species  Area Elevation Nearest Scruz Adjacent
## Baltra            58 25.09       346     0.6   0.6     1.84
## Bartolome         31  1.24       109     0.6  26.3   572.33
## Caldwell           3  0.21       114     2.8  58.7     0.78
## Champion          25  0.10        46     1.9  47.4     0.18
## Coamano            2  0.05        77     1.9   1.9   903.82
## Daphne.Major      18  0.34       119     8.0   8.0     1.84
```

The variables are

- Y Species: the number of species found on the island (response variable)
- X1 Area — the area of the island (km^2)
- X2 Elevation — the highest elevation of the island (m)
- X3 Nearest — the distance from the nearest island (km)
- X4 Scruz — the distance from Santa Cruz Island (km)
- X5 Adjacent — the area of the adjacent island (km^2)

We have omitted the second column (which has the number of endemic species) because we shall not use this alternative response variable in this analysis.

2. Fit OLS

```
lmod <- lm(Species ~ Area + Elevation + Nearest + Scruz + Adjacent, data =  gala)
summary(lmod)
```

```
##
## Call:
## lm(formula = Species ~ Area + Elevation + Nearest + Scruz + Adjacent,
##     data = gala)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -111.679  -34.898   -7.862   33.460  182.584
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.068221  19.154198   0.369 0.715351
## Area        -0.023938   0.022422  -1.068 0.296318
## Elevation    0.319465   0.053663   5.953 3.82e-06 ***
## Nearest      0.009144   1.054136   0.009 0.993151
## Scruz       -0.240524   0.215402  -1.117 0.275208
## Adjacent    -0.074805   0.017700  -4.226 0.000297 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.98 on 24 degrees of freedom
## Multiple R-squared:  0.7658, Adjusted R-squared:  0.7171
## F-statistic:  15.7 on 5 and 24 DF,  p-value: 6.838e-07
```

3. Hypothesis test examples of linear hypotheses

(1) Test all of coefficients inclusion or not.

H_0: Any of the predictors are significant or not.

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$

at the same time.

(2) Test one coefficients inclusion or not.

H_0: Can one particular predictor be dropped from the model?

$$H_0 : \beta_i = 0$$

for a given $i \in \{1, \ldots, 5\}$.

(3) Test a subvector includion or not.

$$H_0 : \beta_{\text{Area}} = \beta_{\text{Ajacent}} = 0.$$

(4) Test a particular value.

$$H_0 : \beta_{\text{Elevation}} = 0.5.$$

(5) Test a subspace.

Some tests cannot be expressed simply in terms of the inclusion or exclusion of subsets of predictors. Consider an example where we test whether the areas of the current and adjacent island can be added together and used in place of the two separate predictors.

$$H_0 : \beta_{\text{Area}} = \beta_{\text{Ajacent}}.$$

The model corresponding to this null hypothesis represents a linear subspace of the full model.

All the above examples fall into a class of hypotheses: linear hypothesis $H_0 : A\beta - c = 0$.

4. LRT test / F-test

For the general null hypothesis, we derived distribution

$$F_{\text{stat}} = \frac{(\text{RSS}_H - \text{RSS})/q}{\text{RSS}/(n-p)} = \frac{(\text{RSS}_H - \text{RSS})/(df_{.H} - df_{.F})}{\text{RSS}/df_{.F}} \sim F_{q,n-p}$$

(Clarification: F_{stat} represents the F test statistic, and $F_{q,n-p}$ represents a distribution.)

(1) Example 1: Test all of coefficients inclusion or not.

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$.

- The number of constraints is $q = 5$.
- The number of parameters is $p = 6$.
- The sample size is $n = 30$.

```
lmod <- lm(Species ~ Area + Elevation + Nearest + Scruz  + Adjacent, gala) #full model
nullmod <- lm(Species ~ 1, gala)  #null model
```

R codes of F-test

```
anova(nullmod, lmod)
```

```
## Analysis of Variance Table
##
## Model 1: Species ~ 1
## Model 2: Species ~ Area + Elevation + Nearest + Scruz + Adjacent
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     29 381081
## 2     24  89231  5    291850 15.699 6.838e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Reading numerical results:

- Results of Model 1 (null):
  - Resisual degrees of freedom (Res.Df) is $n-1 = n-p+q = 29$. (Minus 1 because only one parameter, i.e., intercept.)
  - Residual sum of squares (RSS) is $\|Y - \hat{Y}_H\|^2 = \|Y - X\hat{\beta}_H\|^2$
- Results of Model 2 (full):
  - Resisual degrees of freedom (Res.Df) is $n-p = 30-6 = 24$. (Minus p because p parameters in the model.)
  - Residual sum of squares (RSS) is $\|Y - \hat{Y}\|^2 = \|Y - X\hat{\beta}\|^2$.
- DF: $q = (n-1) - (n-p) = 5$.

- Sum of Sq. is $\text{RSS}_H - \text{RSS} = 381081 - 89231 = 291850$. By Geometric relationship, we have $\text{RSS}_H - \text{RSS} = \|Y - \hat{Y}_H\|^2 - \|Y - \hat{Y}\|^2 = \|\hat{Y}_H - \hat{Y}\|^2$.
- F-test
  - test statistic $F_{\text{stat}} = \frac{(\text{RSS}_H - \text{RSS})/q}{\text{RSS}/(n-p)} = 15.699$ above
  - p-value $P(F > F_{\text{stat}}) = P(F > 15.699) = 1 - P(F \leqslant 15.699)$ where F denotes a random variable following $F_{q,n-p}$ distribution.

Checking the F-statistic calculation with R codes below

```
(df <- df.residual(lmod)) #DF under full model
```

```
## [1] 24
```

```
(rss <- deviance(lmod)) #RSS under full model
```

```
## [1] 89231.37
```

```
(df0 <- df.residual(nullmod)) #DF under null
```

```
## [1] 29
```

```
(rss0 <- deviance(nullmod)) #RSS under null
```

```
## [1] 381081.4
```

```
(fstat <- ((rss0-rss)/(df0-df))/(rss/df)) #F-statistic
```

```
## [1] 15.69941
```

```
1-pf(fstat, df0-df, df) #p-value of F-test
```

```
## [1] 6.837893e-07
```

```
pf(fstat, df0-df, df, lower.tail = F) #p-value of F-test
```

```
## [1] 6.837893e-07
```

Checking the derived RSS calculation with R codes below

```
#directly calculate RSS under full model
x <- model.matrix( ~ Area + Elevation + Nearest + Scruz  + Adjacent, gala )
y <- gala$Species
```

```
xtx = crossprod(x,x)
beta_hat = solve(xtx,crossprod(x,y))
(RSS <- sum((y - x %*% beta_hat )^2))
```

```
## [1] 89231.37
```

```
#directly calculate RSS under the linear constraint Abeta - c = 0
q = 5
(A = cbind(rep(0, q), diag( q ) ))
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]    0    1    0    0    0    0
## [2,]    0    0    1    0    0    0
## [3,]    0    0    0    1    0    0
## [4,]    0    0    0    0    1    0
## [5,]    0    0    0    0    0    1
```

```
c = rep(0,q)
xtxinvA = solve(xtx, t(A))
beta_hat_H <- beta_hat - xtxinvA %*% solve( A %*% xtxinvA , A %*% beta_hat -  c)
(RSS0 <- sum( (y -  x%*% beta_hat_H)^2 ))
```

```
## [1] 381081.4
```

(2) Example 2: Test one coefficients inclusion or not.

$H_0 : \beta_{\text{Area}} = 0.$

```
lmods <- lm(Species ~  Elevation + Nearest + Scruz + Adjacent, gala) #null model
anova(lmods, lmod)
```

```
## Analysis of Variance Table
##
## Model 1: Species ~ Elevation + Nearest + Scruz + Adjacent
## Model 2: Species ~ Area + Elevation + Nearest + Scruz + Adjacent
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     25 93469
## 2     24 89231  1    4237.7 1.1398 0.2963
```

- Model 1 (null): Resisual degrees of freedom (Res.Df) is $n - 5 = n - p + q = 25$.
- DF: $q = 1$.

(3) Example 3: Test a pair of predictors

$H_0 : \beta_{\text{Area}} = \beta_{\text{Adjacent}} = 0$.

```
lmods <- lm(Species ~  Elevation + Nearest + Scruz, gala)
anova(lmods, lmod)
```

```
## Analysis of Variance Table
##
## Model 1: Species ~ Elevation + Nearest + Scruz
## Model 2: Species ~ Area + Elevation + Nearest + Scruz + Adjacent
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     26 158292
## 2     24  89231  2     69060 9.2874 0.00103 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Model 1 (null): Resisual degrees of freedom (Res.Df) is $n - 4 = n - p + q = 26$.
- DF: $q = 2$.

(4) Example 4: Test a particular value

$H_0 : \beta_{\text{Elevation}} = 0.5$.

```
lmods <- lm(Species ~  Area+ offset(0.5*Elevation) + Nearest + Scruz + Adjacent, gala)
anova(lmods, lmod)
```

```
## Analysis of Variance Table
##
## Model 1: Species ~ Area + offset(0.5 * Elevation) + Nearest + Scruz +
##     Adjacent
## Model 2: Species ~ Area + Elevation + Nearest + Scruz + Adjacent
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     25 131312
## 2     24  89231  1     42081 11.318 0.002574 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

"offset" can be used to specify an a priori known component to be included in the linear predictor during fitting.

- Model 1 (null): Resisual degrees of freedom (Res.Df) is $n - 5 = n - p + q = 25$.
- DF: $q = 1$.

(5) Example 5: Test a subspace

$H_0 : \beta_{\text{Area}} = \beta_{\text{Ajacent}}$.

```
lmods <- lm(Species ~  I(Area+Adjacent) + Elevation + Nearest + Scruz, gala)
anova(lmods, lmod)
```

```
## Analysis of Variance Table
##
## Model 1: Species ~ I(Area + Adjacent) + Elevation + Nearest + Scruz
## Model 2: Species ~ Area + Elevation + Nearest + Scruz + Adjacent
##   Res.Df    RSS Df Sum of Sq     F  Pr(>F)
## 1     25 109591
## 2     24  89231  1     20360 5.476 0.02793 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Model 1 (null): Resisual degrees of freedom (Res.Df) is $n - 5 = n - p + q = 25$.
- DF: $q = 1$.

5. Interpretation of results of F-test

Given significance level α, a hypothesis is rejected if

- p-value $< \alpha$
- or equivalently, F statistic $> F_{q,n-p}^{(\alpha)}$ (upper α-level quantile of F-distribution with degrees of freedom q and $n - p$)

Reject the null

When the null is rejected, this does not imply that the alternative model is the best model. We do not know whether all the predictors are required to predict the response or just some of them. Other predictors might also be added or existing predictors transformed or recombined.

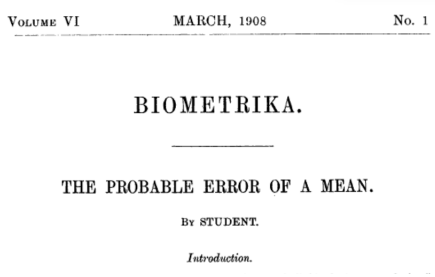
Failure to reject the null

A failure to reject the null hypothesis is not the end of the game — you must still investigate the possibility of nonlinear transformations of the variables and of outliers which may obscure the relationship. Even then, you may just have insufficient data to demonstrate a real effect, which is why we must be careful to say "fail to reject" the null rather than "accept" the null. It would be a mistake to conclude that no real relationship exists. This issue arises when a pharmaceutical company wishes to show that a proposed generic replacement for a brand-named drug is equivalent. It would not be enough in this instance just to fail to reject the null. A higher standard would be required.

The overall F-test (Example 1) is just the beginning of an analysis and not the end.

6. Single coefficient: T-test

- T-Distribution, also known as Student's t-distribution, gets its name from William Sealy Gosset who first published it in English in 1908 in the scientific journal Biometrika using the pseudonym "Student" because his employer, Guinness Breweries, preferred staff to use pen names when publishing scientific papers. ("`The probable error of a mean`") Originally motivated from conducting one-sample and two-sample mean tests, but is generalized to test regression coefficients.



VOLUME VI MARCH, 1908 No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

Introduction.

- In ususal R output of `summary()` fucntion, T-test is presented for each coefficient. It can actually be shown that T^2 is equal to the appropriate F-statistic computed using the method in Part 4 above. In particular, the full/alternative model needs to be consistent.

$$T_{\text{stat}} = \frac{\hat{\beta} - \beta}{s.e.(\hat{\beta})} \sim t_{n-p}$$

  where $\hat{\beta}$ denotes an 1-dimensional coefficient from a p-dimensional regression model.

- Given significance level α, the hypothesis is rejected if
  - p-value $= P(|t_{n-p}| > |T_{\text{stat}}|) = P(t_{n-p} > |T_{\text{stat}}|) + P(t_{n-p} < -|T_{\text{stat}}|) < \alpha$. (two-sided)
    - t_{n-p} represents a t-distribution with d.o.f. $n - p$.
    - Since t-distribution is symmetric around 0, $P(t_{n-p} > |T_{\text{stat}}|) = P(t_{n-p} < -|T_{\text{stat}}|)$.
    - $t^2_{n-p} \sim F_{1,n-p}$
  - or equivalently, $T_{\text{stat}} > t^{(\alpha/2)}_{n-p}$ or $T_{\text{stat}} < t^{(1-\alpha/2)}_{n-p}$.
    - $t^{(\omega)}_{n-p}$ denotes upper ω-level quantile of t-distribution with degrees of freedom $n - p$.

(6.1) Example 2: Test one coefficients inclusion or not.

$H_0 : \beta_{\text{Area}} = 0$. Under H_0,

$$T_{\text{stat}} = \frac{\hat{\beta}_{\text{Area}}}{s.e.(\hat{\beta}_{\text{Area}})} \sim t_{n-p}$$

```
(scoef <- summary(lmod)$coefficients)
```

```
##                 Estimate  Std. Error      t value     Pr(>|t|)
## (Intercept)  7.068220709 19.15419782  0.369016796 7.153508e-01
## Area        -0.023938338  0.02242235 -1.067610554 2.963180e-01
## Elevation    0.319464761  0.05366280  5.953187968 3.823409e-06
## Nearest      0.009143961  1.05413595  0.008674366 9.931506e-01
## Scruz       -0.240524230  0.21540225 -1.116628222 2.752082e-01
## Adjacent    -0.074804832  0.01770019 -4.226216850 2.970655e-04
```

```
(scoef['Area', 't value'])^2 #same as the F-statistic in Part 4 F-test Example 2 above.
```

```
## [1] 1.139792
```

```
scoef['Area', 'Pr(>|t|)'] #same as the F-test p-value in Part 4 F-test Example 2 above.
```

```
## [1] 0.296318
```

We can recalculate the p-value by the definition.

```
#p-value: two-sided p-value of t-statistic.
#P(|T| > |t_stat|) = P( T > |t_stat|) + P( T < -|t_stat| )
#This is the same as the F-test p-value above.
beta_hat_Area <- scoef['Area', 'Estimate']
s.e_Area <- scoef['Area', 'Std. Error']


#t-statistic
(tstat_Area <- (beta_hat_Area)/s.e_Area)
```

```
## [1] -1.067611
```

```
#two-sided p-value of t-test, same as R output
pt(abs(tstat_Area), 24, lower.tail = F) + pt( - abs(tstat_Area), 24, lower.tail = T)
```

```
## [1] 0.296318
```

(6.2) Example 4: Test a particular value

$H_0 : \beta_{\text{Elevation}} = 0.5$. (output of summary() does not give t-test results for the non-zero tested value.)
Under H_0,

$$T = \frac{\hat{\beta}_{\text{Elevation}} - 0.5}{s.e.(\hat{\beta}_{\text{Elevation}})} \sim t_{n-p}$$

```
beta_hat_Elevation <- scoef['Elevation', 'Estimate']
s.e_Elevation <- scoef['Elevation', 'Std. Error']
(tstat_Elevation <- (beta_hat_Elevation-0.5)/s.e_Elevation)  #t-statistic
```

```
## [1] -3.364253
```

```
#p-value: two-sided p-value of t-statistic.
#P(|T| > |t_stat|) = P( T > |t_stat|) + P( T < -|t_stat| )
#This is the same as the F-test p-value above.
pt(abs(tstat_Elevation), 24, lower.tail = F) +
  pt( - abs(tstat_Elevation), 24, lower.tail = T)
```

```
## [1] 0.002573836
```

```
#t-statistic squared. This is the same as the F-statistic in Part 4 above.
tstat_Elevation^2
```

```
## [1] 11.3182
```

7. Confidence interval

(7.1) For single coefficient

By

$$\frac{\hat{\beta}_i - \beta_i}{\text{s.e.}(\hat{\beta}_i)} \sim t_{n-p},$$

two end points of the confidence interval are given by

$$\hat{\beta}_i \pm t_{n-p}^{(\alpha/2)} \times \text{s.e.}(\hat{\beta}_i).$$

- $t_{n-p}^{(\alpha/2)}$ represents upper $\alpha/2$ quantile of t-distribution with d.o.f. $n-p$.
- s.e.$(\hat{\beta}_i)$ is the squared root of the i-th diagonal of $\hat{\sigma}^2(X^\top X)^{-1}$.

In the interval form:

$$\left(\hat{\beta}_i - t_{n-p}^{(\alpha/2)} \times \text{s.e.}(\hat{\beta}_i), \quad \hat{\beta}_i + t_{n-p}^{(\alpha/2)} \times \text{s.e.}(\hat{\beta}_i) \right).$$

$$\left(\hat{\beta}_i + t_{n-p}^{(1-\alpha/2)} \times \text{s.e.}(\hat{\beta}_i), \quad \hat{\beta}_i + t_{n-p}^{(\alpha/2)} \times \text{s.e.}(\hat{\beta}_i) \right).$$

where we note $t_{n-p}^{(1-\alpha/2)} = -t_{n-p}^{(\alpha/2)}$ by symmetricity of t-distribution.

```
#upper alpha/2= 0.05/2 = 0.025 level quantile of t distribution with d.o.f. n-p=30-6
alpha = 0.05
(t_upper_quantile_1 <- qt(alpha/2, 30-6, lower.tail = F ) )
```

```
## [1] 2.063899
```

```
(t_upper_quantile_2 <- qt(1-alpha/2, 30-6 , lower.tail = F ) )
```

```
## [1] -2.063899
```

Method using R function

```
confint(lmod)
```

```
##                   2.5 %      97.5 %
## (Intercept) -32.4641006 46.60054205
## Area         -0.0702158  0.02233912
## Elevation     0.2087102  0.43021935
## Nearest      -2.1664857  2.18477363
## Scruz        -0.6850926  0.20404416
## Adjacent     -0.1113362 -0.03827344
```

Example 1 for $H_0 : \beta_{\text{Area}} = 0$:

```
scoef
```

```
##                 Estimate  Std. Error      t value     Pr(>|t|)
## (Intercept)  7.068220709 19.15419782  0.369016796 7.153508e-01
## Area        -0.023938338  0.02242235 -1.067610554 2.963180e-01
## Elevation    0.319464761  0.05366280  5.953187968 3.823409e-06
## Nearest      0.009143961  1.05413595  0.008674366 9.931506e-01
## Scruz       -0.240524230  0.21540225 -1.116628222 2.752082e-01
## Adjacent    -0.074804832  0.01770019 -4.226216850 2.970655e-04
```

```
#interval same as R code output for Area
scoef['Area' , 'Estimate'] + c(-1,1) *  scoef['Area' , 'Std. Error'] * t_upper_quantile_1
```

```
## [1] -0.07021580  0.02233912
```

- CIs have a duality with two-sided hypothesis tests. If the interval contains zero, this indicates that the null hypothesis $H_0 : \beta_{\text{Area}} = 0$ would not be rejected at the $\alpha = 5\%$ level.
- We can see from the summary that the p-value is 29.6%, greater than 5%, confirming this point. Indeed, any point null hypothesis lying within the interval would not be rejected.

Example 2 for $H_0 : \beta_{\text{Adjacent}} = 0$:

```
#interval same as R code output for Adjacent
scoef['Adjacent' , 'Estimate'] +
  c(-1,1) *  scoef['Adjacent' , 'Std. Error'] * t_upper_quantile_1
```

```
## [1] -0.11133622 -0.03827344
```

- Because zero is not in this interval, the null is rejected at the significance level 5%.
- Nevertheless, this CI is relatively wide in the sense that the upper limit is about three times larger than the lower limit. This means that we are not really that confident about what the exact effect of the area of the adjacent island on the number of species really is, even though the statistical significance means we are confident it is negative.

(7.2) For multiple coefficients (confidence region)

If you are interested in jointly testing p-dimensional β vector, you can construct a $100(1-\alpha)\%$ confidence region for β using the F-test we derived.

- Take A matrix such that $A\beta = \beta_{\text{sub}}$ of interest.
- F-test statistic for $H_0 : A\beta = c$ is derived to be

$$\frac{(\text{RSS}_H - \text{RSS})/q}{\text{RSS}/(n-p)} = \frac{(A\hat{\beta}-c)^\top \{A(X^\top X)^{-1}A^\top\}^{-1}(A\hat{\beta}-c)/q}{\hat{\sigma}^2}$$

- Plugging in $c = A\beta$ gives

$$\begin{aligned}\frac{(\text{RSS}_H - \text{RSS})/q}{\text{RSS}/(n-p)} &= \frac{(\hat{\beta}-\beta)^\top A^\top \{A(X^\top X)^{-1}A^\top\}^{-1}A(\hat{\beta}-\beta)/q}{\hat{\sigma}^2} \\ &= \frac{(\hat{\beta}_{\text{sub}}-\beta_{\text{sub}})^\top \{(X^\top X)^{-1}_{\text{sub}}\}^{-1}(\hat{\beta}_{\text{sub}}-\beta_{\text{sub}})/q}{\hat{\sigma}^2} \sim F_{q,n-p}\end{aligned}$$

  Then a confidence region for jointly testing β_{sub} vector is

$$\frac{(\hat{\beta}_{\text{sub}}-\beta_{\text{sub}})^\top \{(X^\top X)^{-1}_{\text{sub}}\}^{-1}(\hat{\beta}_{\text{sub}}-\beta_{\text{sub}})}{q\hat{\sigma}^2} \leq F^{(\alpha)}_{q,n-p}$$

  These regions are ellipsoidally shaped. Because these ellipsoids lie in higher dimensions, they cannot easily be visualized except for the two-dimensional case.

- For example, if $\beta_{\text{sub}} = (\beta_{\text{Area}}, \beta_{\text{Adjacent}})^\top$, we can take

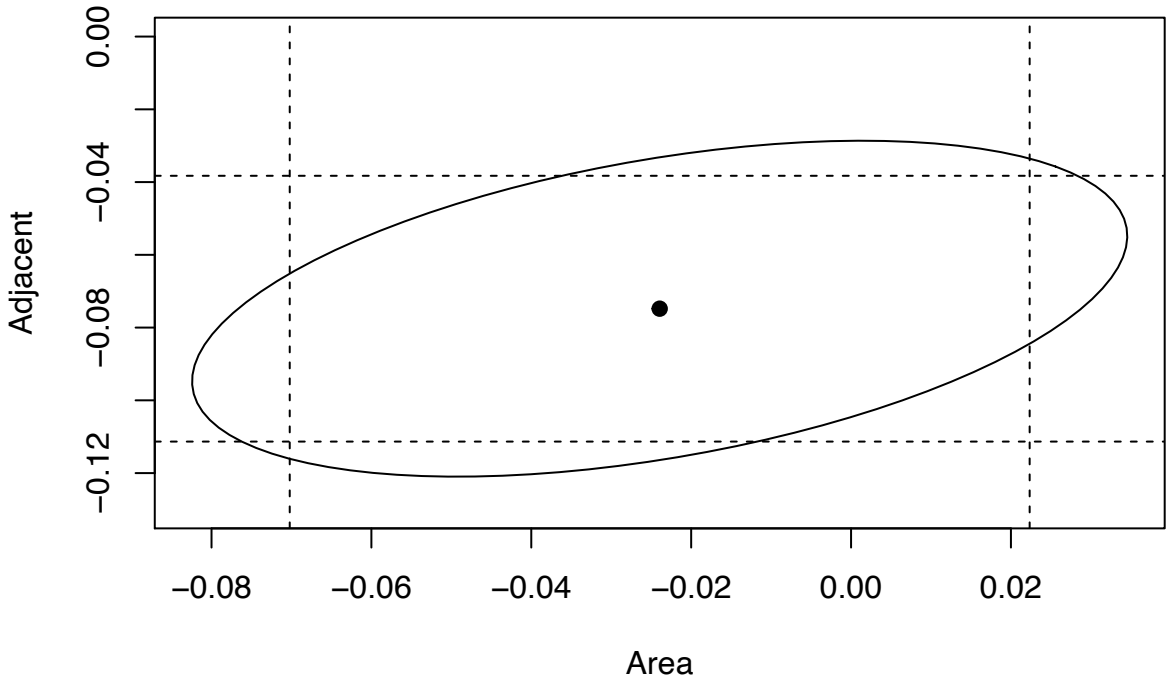
$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \Rightarrow A \begin{pmatrix} \beta_{\text{Intercept}} \\ \beta_{\text{Area}} \\ \beta_{\text{Adjacent}} \\ \beta_{\text{Elevation}} \\ \beta_{\text{Nearest}} \\ \beta_{\text{Scruz}} \end{pmatrix} = \begin{pmatrix} \beta_{\text{Area}} \\ \beta_{\text{Adjacent}} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

(7.3) Comparison

As an example, we compare ellipse given by $H_0 : \beta_{\text{Area}} = \beta_{\text{Adjacent}} = 0$ and two individual confidence intervals of β_{Area} and β_{Adjacent}, respectively.

```
require(ellipse)
plot(ellipse(lmod,c('Area','Adjacent')),type="l",ylim=c(-0.13,0))
points(coef(lmod)['Area'], coef(lmod)['Adjacent'], pch=19)
```

```
abline(v=confint(lmod)['Area',],lty=2)
abline(h=confint(lmod)['Adjacent',],lty=2)
```



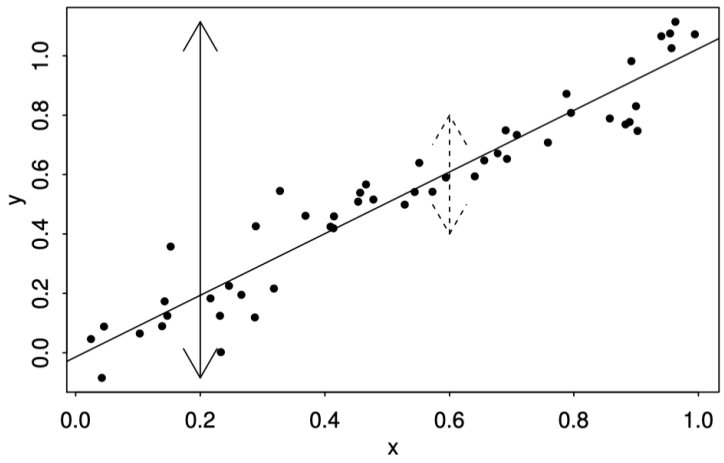
- We can determine the outcome of various hypotheses from the plot.
  - The joint hypothesis $H_0 : \beta_{\text{Area}} = \beta_{\text{Adjacent}} = 0$ is rejected because the origin does not lie inside the ellipse.
  - The hypothesis $H_0 : \beta_{\text{Area}} = 0$ is not rejected because zero does lie within the vertical dashed lines whereas the horizontal dashed lines do not encompass zero and so $H_0 : \beta_{\text{Adjacent}} = 0$ is rejected.
  - We must also specify all the other three predictors are part of the model used to make these tests and confidence statements.
- If you want to test multiple parameters, you need to use a joint testing procedure and not try to combine several univariate tests.
- In higher dimensions, confidence ellipses are not easily visualized so our example here is more of educational than practical value. Nevertheless, it should serve as a caution in interpreting a collection of univariate hypothesis tests or confidence intervals.

8. Coefficient of determination

$$R^2 = \widehat{\text{corr}}(Y, \hat{Y}) = 1 - \frac{\text{RSS}}{\text{TSS}} = \frac{\sum_{i=1}^{n}(\widehat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

Interpretation: percentage of variance explained.

- When x is not known, the best predictor of y is $\bar{y}$ and the variation is denoted by the dotted line.
- When x is known, we can predict y more accurately by the solid line.
- R^2 is related to the ratio of these two variances.



Evaluation:

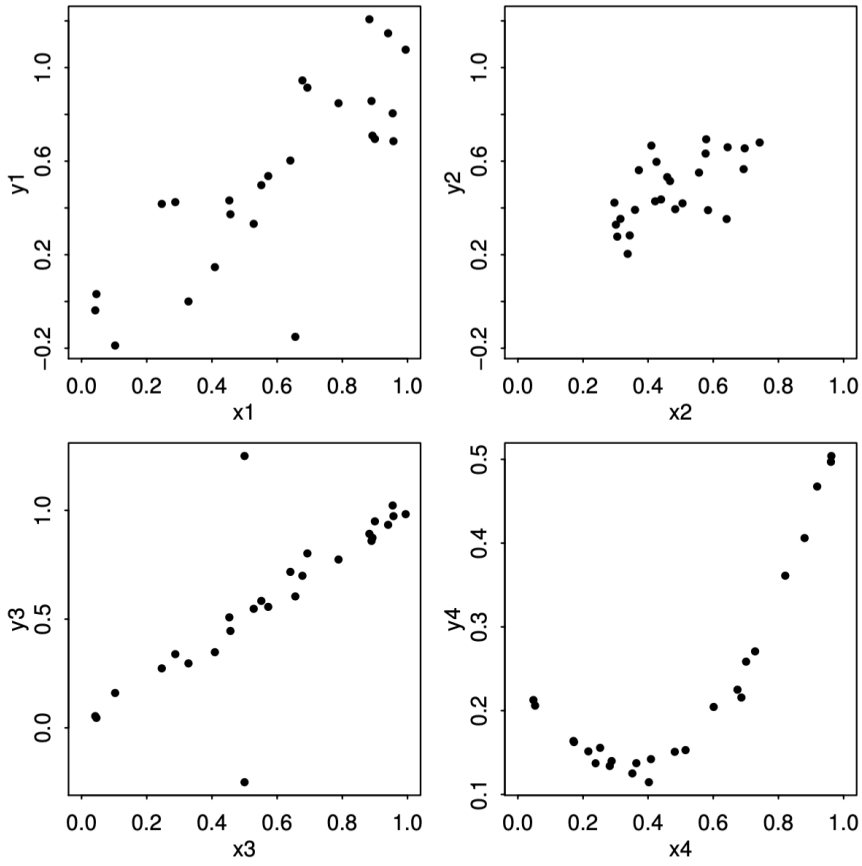
What is a good value of R^2 ? It depends on the area of application.

- In the biological and social sciences, variables tend to be more weakly correlated and there is a lot of noise. We would expect lower values for R^2 in these areas - a value of, say, 0.6 might be considered good.
- In physics and engineering, where most data come from closely controlled experiments, we typically expect to get much higher R^2 s and a value of 0.6 would be considered low.
- Some experience with the particular area is necessary for you to judge your R^2 s well.

Caution 1: It is a mistake to rely on R^2 as a sole measure of fit.

- In the next Figure we see some simulated datasets where the R^2 is around 0.65 for a linear fit in all four cases.
- Take the plot on the upper left as a baseline.

- In the plot on the upper right, the residual variation is smaller than the first plot but the variation in x is also smaller so R^2 is about the same. Predictions (within the range of x) would have less variation in the second case.
- In the plot on the lower left, the fit looks strong except for a couple of outliers.
- On the lower right, the relationship is quadratic, which shows us that R^2 doesn't tell us much about whether we have the right model.



Caution 2: Some care is necessary if there is no intercept in your model.

The denominator in the first definition of R^2 has a null model with an intercept in mind when the sum of squares is calculated. Unfortunately, R uses this definition and will give a misleadingly high R^2. If you must have an R^2 use the $\text{cor}^2(\hat{y}, y)$ definition when there is no intercept.

```
#full model R squared
lmods <- lm(Species ~  Area + Elevation + Nearest + Scruz + Adjacent , gala)
summary(lmods)$r.squared
```

```
## [1] 0.7658469
```

```r
# + 0 implies no intercept in linear model
lmods_no <- lm(Species ~  Area + Elevation + Nearest + Scruz + Adjacent + 0 , gala)
summary(lmods_no)$r.squared
```

```
## [1] 0.8501933
```

9. Summary of R outputs calculations

We now have explained all of the outpus from R `summary()`.

```
lmod <- lm(Species ~ Area + Elevation + Nearest + Scruz + Adjacent , data =  gala)
summary(lmod)
```

```
##
## Call:
## lm(formula = Species ~ Area + Elevation + Nearest + Scruz + Adjacent,
##     data = gala)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -111.679  -34.898   -7.862   33.460  182.584
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.068221  19.154198   0.369 0.715351
## Area        -0.023938   0.022422  -1.068 0.296318
## Elevation    0.319465   0.053663   5.953 3.82e-06 ***
## Nearest      0.009144   1.054136   0.009 0.993151
## Scruz       -0.240524   0.215402  -1.117 0.275208
## Adjacent    -0.074805   0.017700  -4.226 0.000297 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.98 on 24 degrees of freedom
## Multiple R-squared:  0.7658, Adjusted R-squared:  0.7171
## F-statistic:  15.7 on 5 and 24 DF,  p-value: 6.838e-07
```

- Coefficients estimates: $\hat{\beta} = (X^\top X)^{-1} X^\top Y$

```
X <- model.matrix( ~ Area + Elevation + Nearest + Scruz  + Adjacent, gala )
Y <- gala$Species
xtx = crossprod(X,X)
```

```
beta_hat = solve(xtx,crossprod(X,Y))
t(beta_hat)
```

```
##      (Intercept)        Area Elevation     Nearest      Scruz    Adjacent
## [1,]    7.068221 -0.02393834 0.3194648 0.009143961 -0.2405242 -0.07480483
```

- Residuals: $Y - \hat{Y}$

```
Y_hat <- X %*% beta_hat
summary( Y - Y_hat)
```

```
##        V1
##  Min.   :-111.679
##  1st Qu.: -34.898
##  Median :  -7.862
##  Mean   :   0.000
##  3rd Qu.:  33.460
##  Max.   : 182.584
```

- Residual standard error: $\hat{\sigma} = \sqrt{\frac{\text{RSS}}{n-p}}$. ($\text{E}(\hat{\sigma}^2) = \sigma^2$ unbiased.)

```
n = 30
p = 6
sigma_hat <- sqrt( sum((Y - Y_hat)^2)/(n-p) ) #n-p=30-6=24
```

- Standard errors: squared roots of diagonals of $\hat{\sigma}^2(X^\top X)^{-1}$. (By $\text{var}(\hat{\beta}) = \sigma^2(X^\top X)^{-1}$)

```
(sd_errors <- sigma_hat * sqrt(diag( solve( xtx ))))
```

```
## (Intercept)        Area   Elevation     Nearest       Scruz    Adjacent
## 19.15419782  0.02242235  0.05366280  1.05413595  0.21540225  0.01770019
```

- t-values: Estimate/Std. Error

```
(t_stats <- t(beta_hat)/sd_errors)
```

```
##      (Intercept)      Area Elevation     Nearest     Scruz  Adjacent
## [1,]   0.3690168 -1.067611  5.953188 0.008674366 -1.116628 -4.226217
```

- p-values of t-tests

```r
pt(abs(t_stats), 24, lower.tail = F) + pt( - abs(t_stats), 24, lower.tail = T)
```

```
##      (Intercept)     Area    Elevation   Nearest     Scruz     Adjacent
## [1,]   0.7153508 0.296318 3.823409e-06 0.9931506 0.2752082 0.0002970655
```

- Multiple R-squared

```r
(cor( Y, Y_hat ))^2
```

```
##           [,1]
## [1,] 0.7658469
```

```r
TSS <- sum((Y - mean(Y))^2)
RSS <- sum((Y - Y_hat)^2)
1 - RSS/TSS
```

```
## [1] 0.7658469
```

- Adjusted R-squared

```r
1 - (RSS/(n-p)) / (TSS / (n-1))
```

```
## [1] 0.7170651
```

- F-test for all coefficients excluding intercept $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

```r
lmod_null <- lm( Species ~ 1, data = gala )
RSS0 = sum((Y - lmod_null$fitted.values)^2)
q = 5
(F_stat = ((RSS0 - RSS)/q)/(RSS/(n-p))) #F-statistic
```

```
## [1] 15.69941
```

```r
( pf(F_stat, q, n-p, lower.tail = F) )
```

```
## [1] 6.837893e-07
```

Remark:

F or T-test can suffer from loss of power as p approaches n (high-dimensional issue). See, e.g., `Effect of high dimension: by an example of a two sample problem` (1996) by Bai, Zhidong and Saranadasa, Hewa.

Prediction: point estimate and prediction interval

Motivation

- Besides model inference, another main use of regression analysis is prediction.
- Suppose we build a model $Y = X\beta + \epsilon$ and obtain estimates $\hat{\beta}$.
- Given a new set of predictors x_0, the predicted response is: $\hat{y}_0 = x_0^\top \hat{\beta}$.
- Besides this point estimate, one may want to further assess the uncertainty in this prediction.
  - Decision makers need more than just a point estimate to make rational choices.
  - If the prediction has a wide "confidence" interval, we need to allow for outcomes far from the point estimate.
  - For example, suppose we need to predict the high water mark of a river. We may need to construct barriers high enough to withstand floods much higher than the predicted maximum when "confidence" interval is wide.

Two types of predictions

- There are two kinds of predictions made from regression models.
  - One is a predicted mean response: $\mathrm{E}(y_0 \mid x_0) = x_0^\top \beta$.
  - Another is a prediction of a future observation: Y_{future}. Intuitively, $y_0 = \mathrm{E}(y_0 \mid x_0) + \epsilon = x_0^\top \beta + \epsilon$ (an additional mean zero random error term.)
- Example: Suppose we have built a regression model that predicts the rental price of houses in a given area based on predictors such as the number of bedrooms and closeness to a major highway. There are two kinds of predictions that can be made for a given x_0:
  - 1. Suppose we ask the question — "What would a house with characteristics x_0 rent for on average?" This selling price is $x_0^\top \beta$ and is again predicted by $x_0^T \hat{\beta}$ but now only the variance in $\hat{\beta}$ needs to be taken into account.
  - 2. Suppose a specific house comes on the market with characteristics x_0. Its rental price will be $x_0^\top \beta + \varepsilon$. Since $E\varepsilon = 0$, the predicted price is $x_0^\top \hat{\beta}$, but in assessing the variance of this prediction, we must include the variance of ε.

Most times, we consider the second case, which is called "prediction of a future value," while the first case, called "prediction of the mean response" is less commonly required.

I. First type of prediction

- The confidence interval (CI) for the mean response for given x_0, i.e., $\mathrm{E}(y_0 \mid x_0) = x_0^\top \beta$ is:

$$\hat{y}_0 \pm t_{n-p}^{(\alpha/2)} \hat{\sigma} \sqrt{x_0^T \left(X^T X\right)^{-1} x_0}$$

  - This is by

$$\frac{(\hat{y}_0 - x_0^\top \beta)/\sqrt{\mathrm{var}(\hat{y}_0 - x_0^\top \beta)}}{\hat{\sigma}/\sigma} \sim t_{n-p},$$

  - where $\hat{y}_0 - x_0^\top \beta = x_0^\top \hat{\beta} - x_0^\top \beta \sim N(0, \mathrm{var}(x_0^\top \hat{\beta}))$ and $\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}/(n-p)$
  - Variance:

$$\mathrm{var}(\hat{y}_0 - x_0^\top \beta) = \mathrm{var}\left(x_0^\top \hat{\beta}\right) = x_0^T \left(X^T X\right)^{-1} x_0 \sigma^2.$$

II. Second type of prediction

- A future observation for $y_0 = x_0^\top \beta + \epsilon$ should be predicted to be $x_0^\top \hat{\beta} + \epsilon$.
- But we do not know the future ϵ but we expect it has mean zero so the point prediction is $\hat{y}_0 = x_0^\top \hat{\beta}$.
- Uncertainty: (similar to the derivation of t-test)

$$\frac{(\hat{y}_0 - y_0)/\sqrt{\mathrm{var}(\hat{y}_0 - y_0)}}{\hat{\sigma}/\sigma} \sim t_{n-p},$$

  - where $\hat{y}_0 - y_0 = x_0^\top \hat{\beta} - (x_0^\top \beta + \epsilon) \sim N(0, \mathrm{var}(x_0^\top \hat{\beta} - \epsilon))$ and $\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}/(n-p)$
  - Variance: It is usually reasonable to assume future ϵ is independent of $\hat{\beta}$ and has variance σ^2. Then

$$\mathrm{var}(\hat{y}_0 - y_0) = \mathrm{var}\left(x_0^\top \hat{\beta} - \epsilon\right) = \mathrm{var}\left(x_0^\top \hat{\beta}\right) + \mathrm{var}(\epsilon) = \sigma^2 \left\{x_0^\top (X^\top X)^{-1} x_0 + 1\right\}.$$

    * Parameter uncertainty + Model uncertainty
- ***Prediction interval***: $100(1-\alpha)\%$ prediction interval for a single future response is:

$$\hat{y}_0 \pm t_{n-p}^{(\alpha/2)} \hat{\sigma} \sqrt{1 + x_0^T \left(X^T X\right)^{-1} x_0}.$$

  - Previous confidence intervals have been for model parameters β. Parameters are considered to be fixed but unknown — they are not random under the Frequentist approach we are using here.

  - However, a future observation is a random variable. For this reason, it is better to call this a "prediction interval" not confidence interval.
- This prediction interval is typically much wider than the CI above. Although we would like to have a narrower interval generally, we should not make the mistake of using CI when forming prediction intervals for predicted values.

Data Example: Predicting body fat

Problem background

- Measuring body fat is not simple. Muscle and bone are denser than fat so an estimate of body density can be used to estimate the proportion of fat in the body. Measuring someone's weight is easy but volume is more difficult.
- One method requires submerging the body underwater in a tank and measuring the increase in the water level. Most people would prefer not to be submerged underwater to get a measure of body fat so we would like to have an easier method.
- In order to develop such a method, researchers recorded age, weight, height, and 10 body circumference measurements for 252 men. Each man's percentage of body fat was accurately estimated by an underwater weighing technique.
- Can we predict body fat using just the easy-to-record measurements?

```
library(faraway)
data(fat,package="faraway")
head(fat, 2)
```

```
##   brozek siri density age weight height adipos  free neck chest abdom  hip
## 1   12.6 12.3  1.0708  23 154.25  67.75   23.7 134.9 36.2  93.1  85.2 94.5
## 2    6.9  6.1  1.0853  22 173.25  72.25   23.4 161.3 38.5  93.6  83.0 98.7
##   thigh knee ankle biceps forearm wrist
## 1  59.0 37.3  21.9   32.0    27.4  17.1
## 2  58.7 37.3  23.4   30.5    28.9  18.2
```

- Use `brozek` as the response (Brozek's equation estimates percent body fat from density).
- Fit a model using all thirteen predictors.