The Effect of Vitamin C on Tooth Growth in Guinea Pigs

Available in *R* via *ToothGrowth*. From the description:
The response is the length of odontoblasts (teeth) in each of 10 guinea pigs at each of three dose levels of Vitamin C (0.5, 1, and 2 mg) with each of two delivery methods (orange juice or ascorbic acid). Each combination is tested 10 times.
Source: C. I. Bliss (1952) The Statistics of Bioassay. Academic Press.
References: McNeil, D. R. (1977) Interactive Data Analysis. New York: Wiley.
$n = 60$
Y_i = Tooth length
$S_i = \begin{cases} 0 & \text{orange juice} \\ 1 & \text{ascorbic acid} \end{cases}$
$d_{1i} = I(\text{dose} = 1)$
$d_{2i} = I(\text{dose} = 2)$

Model

$$Y_i = \beta_1 + S_i\beta_2 + d_{1i}\beta_3 + d_{2i}\beta_4 + \epsilon_i$$

Computing $\hat{\beta}$:

```
> Y = ToothGrowth$len
> X = cbind(1,ToothGrowth$supp=="VC",ToothGrowth$dose==1,ToothGrowth$dose==2)
> hbeta <- solve(t(X)%*%X)%*%t(X)%*%Y
> hbeta

        [,1]
[1,]  12.455
[2,]  -3.700  -
[3,]   9.130
[4,]  15.495
```

$\rightarrow (X^T X)^{-1} X^T Y$

We compute the fitted values $\hat{\mathbf{Y}}$ and residuals $\mathbf{e}$:

```
> P <- X%*%solve(t(X)%*%X)%*%t(X)
> Yhat <- P%*%Y
> e <- Y - Yhat
```

$\rightarrow X (X^T X)^{-1} X^T$

$\rightarrow \hat{Y} = PY$

$\rightarrow e = Y - \hat{Y}$

Is the delivery method important? Will test

$$H_0 : \beta_2 = 0 \text{ against } H_1 : \beta_2 \neq 0$$

```
> RSS <- t(e)%*%e
> c <- c(0,1,0,0)
> est <- c%*%hbeta
> est

     [,1]
[1,] -3.7
```

$\rightarrow e^T e$

$\rightarrow \hat{\beta}_2$

```
> sdhat <- sqrt(t(c) %*% solve(t(X)%*%X) %*% c*RSS/(60-4))
> sdhat

          [,1]
[1,] 0.9882795
```

$\rightarrow \sqrt{c^T (X^T X)^{-1} c \frac{RSS}{n-p}}$

A 95% confidence interval for $\mathbf{c}^T\beta$:

```
> L <- est+sdhat*qt(0.025,df=60-4)
> U <- est+sdhat*qt(0.975,df=60-4)
> cat("[",L,U,"]\n")

[ -5.679762 -1.720238 ]
```

$\hat{\beta}_2 \pm \widehat{SD(\beta_2)} \cdot c_{\alpha/2}$

WE REJECT H_0 BECAUSE $0 \notin$ CI

$\Theta_0 \cap A(y) = \emptyset$

0 is not in the CI -> reject $H_0 : \beta_2 = 0$.

Next, we want to compute the p-value of the test that rejects for large values of $|T|$, where

$$T = \frac{\mathbf{c}^T\hat{\beta}}{\sqrt{\mathbf{c}^T(X^TX)^{-1}\mathbf{c}\frac{\text{RSS}}{n-p}}}$$

Let t denote the observed value of T. Under H_0, $T \sim t_{60-4}$. Thus the p-value is

$$p = \mathrm{P}(|T| \geq |t|) = \mathrm{P}(T < -|t| \text{ or } T \geq |t|) = \underbrace{\mathrm{P}(T < -|t|)}_{=\mathrm{P}(T \geq |t|)} + \mathrm{P}(T \geq |t|) = 2(1 - \mathrm{P}(T \leq |t|))$$

```
> abst <- abs(est/sdhat)
> 2*(1-pt(abst,df=60-4))
```

→ $|T|$

```
           [,1]
[1,] 0.0004292793
```

FOR α ~~*~~ > 0.0004292793 WE REJECT H_0

One can get most of the above directly using the function *lm*

```
> summary(lm(len~supp+factor(dose), data=ToothGrowth))

Call:
lm(formula = len ~ supp + factor(dose), data = ToothGrowth)

Residuals:
   Min      1Q Median      3Q     Max
-7.085 -2.751 -0.800  2.446  9.650

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    12.4550     0.9883  12.603  < 2e-16 ***
suppVC         -3.7000     0.9883  -3.744 0.000429 ***
factor(dose)1   9.1300     1.2104   7.543 4.38e-10 ***
factor(dose)2  15.4950     1.2104  12.802  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.828 on 56 degrees of freedom
Multiple R-squared: 0.7623,     Adjusted R-squared: 0.7496
F-statistic: 59.88 on 3 and 56 DF,  p-value: < 2.2e-16
```

$$\bar{R}^2 = 1 - \frac{RSS}{\Sigma(Y_i-\bar{Y})^2} \cdot \frac{n-1}{n-p-1}$$

↳ OF THE F-STATISTIC

$$R^2 = 1 - \frac{RSS}{\Sigma(Y_i-\bar{Y})^2}$$

SUMMARY (LM (OPEN~ EMA50 + SMA50))

$C^T\beta \to$ ONE DIMENSIONAL

10.5 Confidence Regions

Suppose $\mathbf{E}\mathbf{Y} = X\beta$ is a linear model satisfying (FR), (NTA). We are interested in finding a confidence region for β, i.e. we want a random set D s.t. $\mathrm{P}(\beta \in D) \geq 1-\alpha$ for all β, σ^2.

Let

$$\underline{A} = \frac{(\hat\beta-\beta)^T X^T X(\hat\beta-\beta)}{\mathrm{RSS}} \cdot \frac{n-p}{p}$$

If we can work out the distribution of A, we can use A as a pivotal quantity for β.

The numerator of the first fraction can be rewritten as

$$(\mathbf{Y}-X\beta)^T P(\mathbf{Y}-X\beta)$$

where P is the projection onto the space spanned by the columns of X. Indeed,

$$(\mathbf{Y}-X\beta)^T P(\mathbf{Y}-X\beta) \overset{PP=P}{=} (\mathbf{Y}-X\beta)^T PP(\mathbf{Y}-X\beta) \underset{PP = P^TP}{=} (P(\mathbf{Y}-X\beta))^T P(\mathbf{Y}-X\beta)$$

Using $P = X(X^TX)^{-1}X^T$ this is equal to

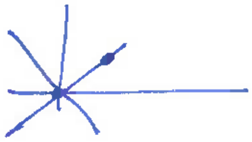
$$(X(\hat\beta-\beta))^T X(\hat\beta-\beta)$$

$P(Y-X\beta) = X(\hat\beta-\beta)$
BECAUSE $PY = \hat{Y} = X\hat\beta$
$PX\beta = X\beta$

Furthermore, RSS can be written as

$$\mathrm{RSS} = \underline{\mathbf{Y}^T Q\mathbf{Y}} \overset{QX\beta=0}{=} (\mathbf{Y}-X\beta)^T Q(\mathbf{Y}-X\beta)$$

where $Q = I-P$. Thus, letting $\mathbf{Z} = \frac{1}{\sigma}(\mathbf{Y}-X\beta)$,



$$\underline{A} = \frac{\mathbf{Z}^T P\mathbf{Z}}{\mathbf{Z}^T Q\mathbf{Z}} \cdot \frac{n-p}{p}$$

with $Z \sim N(\mathbf{0}, I)$, $P+Q=I$, $\operatorname{rank} P = p$, P and Q projection matrices.

Thus the Fisher-Cochran theorem shows that $A \sim F_{p,n-p}$.

Hence, a $1-\alpha$ confidence region R for β is defined by all $\gamma \in \mathbb{R}^p$ such that

$$\frac{(\hat\beta-\gamma)^T X^T X(\hat\beta-\gamma)}{\mathrm{RSS}} \cdot \frac{n-p}{p} \leq \underbrace{F_{p,n-p,\alpha}}_{\to\ \alpha\text{-CRITICAL VALUE}},$$

where $\mathrm{P}(Z \geq F_{p,n-p,\alpha}) = \alpha$ for $Z \sim F_{p,n-p}$. The region R is an ellipsoid centred at $\hat\beta$ (use diagonalisation).

Remark General definition of an ellipsoid: $\{\mathbf{z} \in \mathbb{R}^p : (\mathbf{z}-\mathbf{z}_0)^T A^{-1}(\mathbf{z}-\mathbf{z}_0) \leq 1\}$ where A is pos. def. and $\mathbf{z}_0 \in \mathbb{R}^p$.

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Example 62
$p = 2$, X has full rank.

Let $\mathbf{a} = \hat\beta - \beta$, $B = X^T X$ and $c = p\frac{\text{RSS}}{n-p}F_{p,n-p,\alpha}$. Hence, in order to obtain the confidence region for β we want to find for which $\mathbf{a}$,

$$\boxed{\mathbf{a}^T B\mathbf{a} \le c.} \tag{2}$$

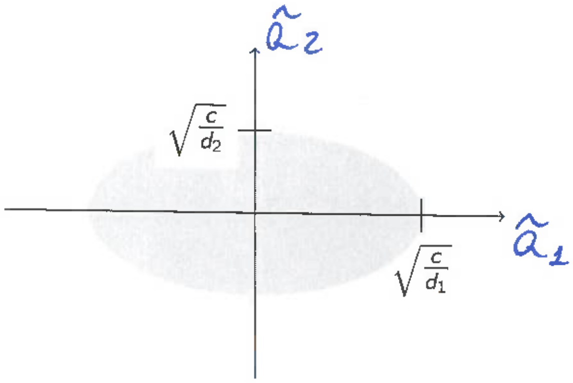
LEMMA 8

B is pos. def. Hence, $\exists$ an orthogonal matrix R and a diagonal matrix $D = \text{diag}(d_1, d_2)$ s.t. $\underline{B = R^T DR}$ and $\underline{d_1, d_2}$ are positive. [D consists of the eigenvalues of B, while R consists of the corresponding normalised eigenvectors.]

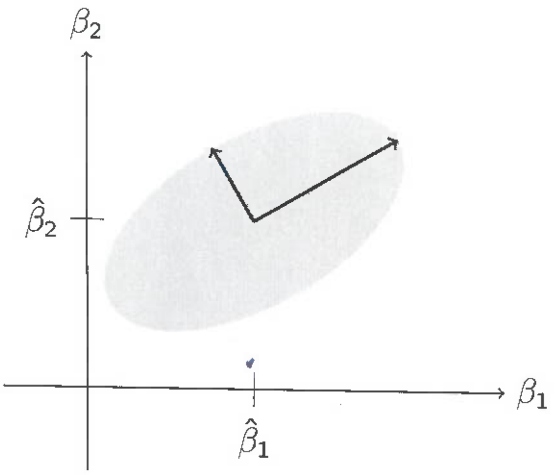
Let $\tilde{\mathbf{a}} = R\mathbf{a}$ (this rotates the coordinate axes). Then (2) is equivalent to

$$\boxed{\tilde a_1^2 \frac{d_1}{c} + \tilde a_2^2 \frac{d_2}{c} \le 1.}$$

This describes an ellipse with half-axes of lenght $\underline{\sqrt{\frac{c}{d_1}}}$ and $\underline{\sqrt{\frac{c}{d_2}}}$.



Transforming everything back via $\underline{\beta = \hat\beta - \mathbf{a}} = \underline{\hat\beta - R^T\tilde{\mathbf{a}}}$ gives a rotated and translated ellipse, centered at $\hat\beta$.



Remark We could construct individual CIs for β_1 and β_2 via Lemma 22 and combine them via the Bonferroni correction. The advantage of the above construction is that the resulting ellipsoid has a smaller area.

11 Diagnostics, Model Selection, Extensions

11.1 Outliers

An *outlier* is an observation that does not conform to the general pattern of the rest of the data. Potential causes:

- Error in the data recording mechanism (example - iron content of spinach).
- Data set may be "contaminated" - i.e. it may be the mixture of two or more populations.
- Indication that the model/underlying theory needs to be improved. Further investigations needed. Outliers may be the "most interesting" observations.

Practical method for spotting outliers: Look for residuals that are "too large". When is a residual too large?

$e = Y - \hat{Y}$

Recall: $\mathbf{e} = (I - P)\mathbf{Y}$, where P is the projection onto span(X). If X is full rank then $P = X(X^TX)^{-1}X^T$. Note that

$\mathrm{cov}(Y) = \sigma^2$ BECAUSE $I-P$ IS A PROJ. MATRIX

$$\mathrm{cov}\,\mathbf{e} = \underline{(I-P)\,\mathrm{cov}\,\mathbf{Y}(I-P)^T} = \underline{\sigma^2(I-P)}(I-P)^T \overset{P}{=} \sigma^2(I-P)$$

and $\mathrm{E}\,\mathbf{e} = \mathbf{0}$. Thus, unter NTA, $e_i \sim N(0, \underline{\sigma^2(1-P_{ii})})$, where P_{ii} is the ith diagonal element of P. Hence,

$$\frac{e_i}{\sqrt{(1-P_{ii})\sigma^2}} \sim N(0,1).$$

We do not know σ^2 → plug in the unbiased estimate

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}}{n-p}.$$

This gives the *standardised* residuals

$$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2\,(1-P_{ii})}}$$

This of course means that r_i are not (necessarily) N(0,1) distributed.

Nevertheless, the standardized residuals should be roughly independent, and their distribution should be close to a N(0,1)-distribution.

Remark r_i is also *not* student t; the usual proof (that we used for t-tests) does not work because $\hat{\sigma}^2$ and e_i are not independent.

Remark Let X ~N(0,1). Then the probabilities for larges values of X are very rapidly decreasing as the following table shows. [The normal distribution has *light tails*.]

x	3	4	5	6	7	8
$P(X > x)$	1.3e-03	3.2e-05	2.9e-07	9.9e-10	1.3e-12	6.2e-16

Thus if (NTA) holds then the standardized residuals should be relatively small.

Remark An entire branch of statistics, called "robust statistics", is concerned with the development of methods/statistics that give meaningful results even in the presence of outliers.

Suppose we are interested in the "centre" of a distribution and have observations $x_1, \dots x_n$. The sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is heavily affected by outliers - in fact, just one outlier can make $\bar{x}$ take any value. Other statistics are far more *robust* to outliers. For example the median cannot be changed arbitrarily by one outlier [you would have to move half of the observations.]

11.2 Leverage

What is the potential impact of individual observations on the model fit?

$$\underline{\operatorname{cov}(\mathbf{e}) = \sigma^2 (I_n - P)}$$

and $\underline{\operatorname{Var} e_i = \sigma^2(1 - P_{ii})}$.

IF $P_{ii} \approx 1$ THEN VAR $e_i \approx 0$

Definition 27
The *leverage* of the ith observation in a linear model is $\underline{P_{ii}}$, the ith diagonal matrix of the hat matrix P.

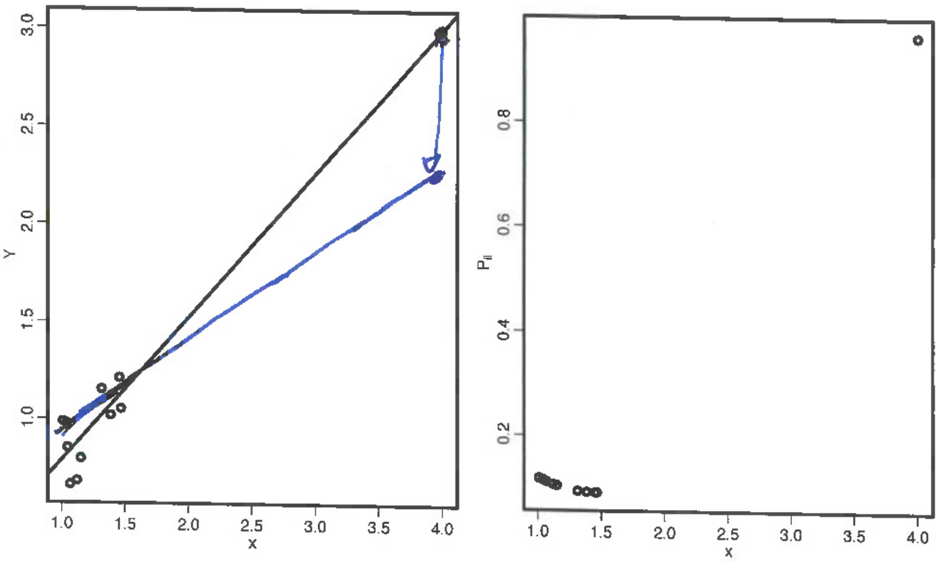
(Recall: P is the projection matrix onto span(X), where X is the design matrix).

If $P_{ii} \approx 1$ then the variance of the ith residual is very low. This is totally determined by X, i.e. the design matrix is forcing the model fit to be good at the covariates of the ith observation. In this case the ith observation is said to have **high leverage**.

$\sum_{i=1}^{n} P_{ii} = \text{trace}(P) = \text{rank}(X) =: r$ (see Lemma 12), so the "average" is r/n and a rule of thumb is to take notice when

$$P_{ii} > \frac{2r}{n}.$$

Example 63 (Linear regression)
$\text{E}\, Y_i = \beta_1 + \beta_2 x_i$



11.3 Cook's Distance

To measure how much the ith observation changes the estimator $\hat{\beta}$ one can consider the following measure, called Cook's distance:

$$D_i = \frac{(\hat{\beta}_{(\text{i})} - \hat{\beta})^T X^T X (\hat{\beta}_{(\text{i})} - \hat{\beta})}{p\, \text{RSS}\,/(n-p)},$$

where $\hat{\beta}_{(\text{i})}$ is the least squares estimator with the ith observation removed. Alternatively,

$$D_i = \frac{(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)})^T (\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)})}{p\, \text{RSS}\,/(n-p)},$$

where $\hat{\mathbf{Y}}_{(i)} = X\hat{\beta}_{(i)}$. Rule of thumb: take notice if D_i gets close to 1.