STAT 571A — Advanced Statistical Regression Analysis

Chapter 2 NOTES

Inferences in Regression and Correlation Analysis

Normal SLR Model

- Continuing with the normal SLR model, we have

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad (2.1)$$

  with $\varepsilon_i \sim$ i.i.d. $N(0,\sigma^2)$, $i = 1,\ldots,n$.

- This produces $Y_i \sim$ indep. $N(\, E[Y_i],\sigma^2\,)$, with mean response

$$E[Y_i] = \beta_0 + \beta_1 X_i + E[\varepsilon_i] = \beta_0 + \beta_1 X_i$$

$\beta_1 = 0$

- It is natural to focus on the slope parameter β_1. **Why?** Look at what happens to $E[Y_i]$ if, say, $\beta_1 = 0$:

$$E[Y_i] = \beta_0 + (0)X_i + E[\varepsilon_i]$$
$$= \beta_0 + 0 + 0 = \beta_0.$$

- That is, when $\beta_1 = 0$, $E[Y_i]$ is <u>independent</u> <u>of</u> X_i. There is no "regression" of Y on X.

Sampling Distribution of b_1

- We use the LS estimator b_1 to estimate β_1.
- Recall that b_1 can be written in the form

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i}{\sum_{m=1}^{n}(X_m - \bar{X})^2} = \sum_{i=1}^{n} k_i Y_i$$

$$\text{for} \quad k_i = \frac{(X_i - \bar{X})}{\sum_{m=1}^{n}(X_m - \bar{X})^2}$$

i.e., a linear combination of the Y_i's.

Distribution of b_1 (cont'd)

- So if $b_1 = \sum k_i Y_i$, then we know from Equ. (A.40) that
  $$\sum k_i Y_i \sim N(\ \sum k_i E[Y_i],\ \sum k_i^2 \sigma^2\ )$$
- But $\sum k_i E[Y_i] = \sum k_i(\beta_0 + \beta_1 X_i)$
  $$= \sum k_i \beta_0 + \sum k_i \beta_1 X_i = \beta_0 \sum k_i + \beta_1 \sum k_i X_i$$
- While $\sum k_i^2 \sigma^2 = \sigma^2 \sum k_i^2$
- So, what are $\sum k_i$, $\sum k_i X_i$, and $\sum k_i^2$?

Distribution of b_1 (cont'd)

- Since $k_i = \dfrac{(X_i - \bar{X})}{\sum_{m=1}^{n}(X_m - \bar{X})^2}$

  we need to find $\sum k_i$, $\sum k_i X_i$, and $\sum k_i^2$.
- (See handwritten PDF notes at http://math.arizona.edu/~piegorsch/571A/sumKnotes.pdf)
- We find:

  $\sum k_i = 0$ $\qquad$ $\sum k_i X_i = 1$ $\qquad$ and

$$\sum k_i^2 = \frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Distribution of b_1 (cont'd)

- **Thus we see:**
  - $E[b_1] = \beta_0 \sum k_i + \beta_1 \sum k_i X_i = \beta_0(0) + \beta_1(1) = \beta_1$ (unbiased!)
  - $\sigma^2[b_1] = \sigma^2 \sum k_i^2 = \dfrac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$

- **So, we can write**

$$b_1 \sim N\left(\beta_1 , \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)$$

Distribution of b_1 (cont'd)

- Now, σ^2 is unknown, so to estimate the variance of b_1, $\sigma^2\{b_1\}$, recall that

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

is unbiased for σ^2.

- Use this to estimate $\sigma^2\{b_1\}$ with

$$s^2\{b_1\} = MSE/\sum_{i=1}^{n}(X_i - \bar{X})^2$$

- The standard error of b_1 is then

$$s\{b_1\} = \sqrt{MSE/\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Distribution of b_1 (cont'd)

In addition, we can show that

$$U = \frac{(n-2)MSE}{\sigma^2} \sim \chi^2(n-2)$$

is independent of

$$b_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)$$

and therefore of

$$Z = \frac{b_1 - \beta_1}{\sigma / \sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}} \sim N(0,1)$$

Distribution of b_1 (cont'd)

Use these in the def'n of a t random variable from (A.44):

$$T = \frac{Z}{\sqrt{U/\nu}}$$

using Z and U from the b_1 construction. Need to 'do the math,' a <u>good</u> <u>exercise</u>: try to algebraically show this $T = (b_1 - \beta_1)/s\{b_1\}$, so that $T \sim t(n-2)$, where $s\{b_1\}$ is the std. error of b_1:

$$s\{b_1\} = \sqrt{MSE/\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Confidence Interval on β_1

- The t sampling distribution for b_1 allows for convenient inferences on β_1.
- For instance, a $1–\alpha$ conf. int. is based on

$$1 - \alpha = P\left[t(\tfrac{\alpha}{2};\, n-2) < T < t(1 - \tfrac{\alpha}{2};\, n-2)\right]$$

- In this, use $T = (b_1 - \beta_1)/s\{b_1\}$:

$$1 - \alpha = P\left[t(\tfrac{\alpha}{2};\, n-2) < (b_1 - \beta_1)/s\{b_1\} < t(1 - \tfrac{\alpha}{2};\, n-2)\right]$$

cont'd →

Confidence Interval on β_1 (cont'd)

The 1–α probability statement simplifies, as

$$1 - \alpha = P\Big[t(\tfrac{\alpha}{2};\, n-2)s\{b_1\} < (b_1 - \beta_1) < t(1-\tfrac{\alpha}{2};\, n-2)s\{b_1\}\Big]$$

$$= P\Big[-b_1 - t(1-\tfrac{\alpha}{2};\, n-2)s\{b_1\} < -\beta_1 < -b_1 + t(1-\tfrac{\alpha}{2};\, n-2)s\{b_1\}\Big]$$

$$= P\Big[b_1 + t(1-\tfrac{\alpha}{2};\, n-2)s\{b_1\} > \beta_1 > b_1 - t(1-\tfrac{\alpha}{2};\, n-2)s\{b_1\}\Big]$$

cont'd →

Confidence Interval on β_1 (cont'd)

By rearranging terms from left-to-right, the 1–α probability statement collapses to

$$1 - \alpha = P\Big[b_1 - t(1-\tfrac{\alpha}{2}; n-2)s\{b_1\} < \beta_1 < b_1 + t(1-\tfrac{\alpha}{2}; n-2)s\{b_1\}\Big]$$

or just $b_1 \pm t(1-\tfrac{\alpha}{2}; n-2)s\{b_1\}$

Example CH01TA01 (p. 19)

Recall from Ch. 1 (Table 1.1) the Toluca Co. example. To find LS fit for simple linear regression in R use:

```
> X = c(80, 30, ... , 70)
> Y = c(399, 121, ... , 323)
> CH01TA01.lm = lm( Y ~ X )
> summary( CH01TA01.lm )
```

summary() output for Toluca example

```
  Call:
lm(formula = Y ~ X)

Residuals:
    Min      1Q  Median      3Q     Max
-83.876 -34.088  -5.982  38.826 103.528

Coefficients:
           Estimate Std. Error t value Pr(>|t|)
(Intercept)  62.366     26.177   2.382   0.0259
 X            3.570      0.347  10.290 4.45e-10
```

(Std. errors of the regr. parameters highlighted in red here.)

Ex. CH01TA01 (cont'd): Conf. Int. on β_1

There are many ways to find a 95% Conf. Interval on the slope parameter, β_1, in R. Fastest is with `confint()`:

```
> confint( CH01TA01.lm )
                2.5 %      97.5 %
(Intercept) 8.213711  116.518006
 X          2.852435    4.287969
```

Ex. CH01TA01 (cont'd): Conf. Int. on β_1

Or, manipulate the various components of the `CH01TA01.lm` object:

The LS estimate is

```
> coef( CH01TA01.lm )[2]
       X
3.570202
```

The std. error $s\{b_1\}$ is

```
> summary( CH01TA01.lm )$coefficients[2,2]
[1] 0.3469722
```

Ex. CH01TA01 (cont'd): Conf. Int. on β_1

The 95% two-sided *t critical point is**

```
> qt( 0.975, df=CH01TA01.lm$df )
[1] 2.068658
```

So the 95% conf. int. is

```
> b1 = coef( CH01TA01.lm )[2]
> se1 = summary( CH01TA01.lm )$coefficients[2,2]
> tcrit = qt( 0.975, df=CH01TA01.lm$df )
> c( b1-tcrit*se1, b1+tcrit*se1 )
  2.852435  4.287969
```

Hypothesis tests on β_1

- Or, to test H_o:$\beta_1 = \beta_{1o}$ vs. H_a:$\beta_1 \neq \beta_{1o}$ (two-sided!), appeal to the t-reference distribution and build the test statistic

$$t^* = \frac{b_1 - \beta_{1o}}{s\{b_1\}}$$

- Under H_o, $t^* \sim t(n\text{–}2)$, so reject H_o when $|t^*| > t\left(1 - \frac{\alpha}{2}; n\text{–}2\right)$
- Special (why?) case: $\beta_{1o} = 0$.
- One-sided: reject H_o vs. (say) H_a:$\beta_1 > \beta_{1o}$ when $t^* > t(1 - \alpha; n\text{–}2)$, etc.

Ex. CH01TA01 (cont'd): Hypoth. tests on β_1

To test $H_o:\beta_1 = 0$ vs. $H_a:\beta_1 \neq 0$ just refer back to the `summary()` output:

```
  Call:
lm(formula = Y ~ X)
      ⋮
Coefficients:
           Estimate Std. Error t value  Pr(>|t|)
(Intercept)  62.366     26.177   2.382    0.0259
 X            3.570      0.347  10.290  4.45e-10
```

$t^* = 10.29$, with $P = 4.45 \times 10^{-10} < \alpha = 0.05$, so reject H_o and conclude
"x=lot size significantly affects Y=work hrs."

Distribution of b_0

- Since we saw that the LS estimator, b_0, for β_0 also has the form $b_0 = \sum k_i Y_i$ (not the same k_i's...), we can build similar sorts of t-based inferences for β_0.
- We find $b_0 \sim N(\ \beta_0, \sigma^2\{b_0\}\ )$, where the variance of b_0 is

$$\sigma^2\{b_0\} = \sigma^2 \left(\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \right)$$

Distribution of b_0

- Can show that

$$Z = (b_0 - \beta_0)/\sigma\{b_0\} \sim N(0,1)$$

is independent of

$$U = (n-2)MSE/\sigma^2 \sim \chi^2(n-2)$$

- From these, find the std. error of b_0:

$$s\{b_0\} = \sqrt{MSE\left(\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right)}$$

Inferences on β_0

- Use these various components to build the t-dist'n random variable

$$T = \frac{b_0 - \beta_0}{s\{b_0\}} \sim t(n-2)$$

- From this, t-test and conf. int's follow in similar form as with β_1.
- For instance, a $1 - \alpha$ conf. int. on β_0 is (no surprise):

$$b_0 \pm t(1 - \tfrac{\alpha}{2};\, n-2)s\{b_0\}$$

Extrapolation

- The textbook gives an example of a conf. int. for β_0 using the Toluca data; however, even they note that it's a silly exercise: who has a "lot size" of $X = 0$?!?
- The X values for these data are all well above $X = 0$, so the conf. int. is an **extrapolation** away from the core of the data.
- In general, extrapolation is tricky and can lead to trouble: try to avoid it!

Robustness

- Note that all these inferences are built under a normal assumption on ε_i. Deviations or departures from this will invalidate the inferences.
- But(!), slight departures from normality will not have a major effect: the conf. int's and hypoth. tests are fairly robust to (symmetric) departures from normality.
- They are much less robust to departures from the common variance assumption, however.

Power Analysis

- Recall that the **power** of a hypoth. test is

$$1 - \beta = 1 - P[\text{accept } H_o \mid H_o \text{ false}]$$
$$= P[\text{reject } H_o \mid H_o \text{ false}]$$

- For the t-test of $H_o\!:\beta_1 = \beta_{1o}$ vs. $H_a\!:\beta_1 \neq \beta_{1o}$, the power will depend on β_{1o} via the test's **noncentrality parameter**:

$$\delta = \frac{|\,\beta_1 - \beta_{1o}\,|}{\sigma\{b_1\}}$$

Power Analysis (cont'd)

- In particular,

$$\text{Power}(\delta) = P[\text{reject } H_o \mid H_o \text{ false}]$$
$$= P[|t^*| > t(1 - \tfrac{\alpha}{2}; n-2) \mid \delta]$$

  which depends upon an extension of the t-dist'n known as the **noncentral t-dist'n**.

- For known δ, the power can be tabulated from Table B.5.
- (δ depends on β_1 and σ, so it can't be "known." But, it can be approximated.)

Ex. CH01TA01 (p. 51): Power analysis for β_1

- Consider again the Toluca data and focus on testing $H_o:\beta_1 = 0$ vs. $H_a:\beta_1 \neq 0$ (so $\beta_{1o} = 0$.) Set $\alpha = 0.05$.
- We found MSE = 2384 for these data, so a rough value for σ^2 here is $\sigma^2 \approx 2500$. Then $\sigma^2\{b_1\} \approx 2500/19800 = 0.1263$.
- Now, say we want to examine the power when $\beta_1 = 1.5$ ($\neq 0$). Then

$$\delta = \frac{|\ \beta_1 - \beta_{1o}\ |}{\sigma\{b_1\}} \approx \frac{|1.5 - 0|}{\sqrt{0.1263}} = 4.22$$

Toluca Power analysis (cont'd)

- Now, enter Table B.5 with:

  $\delta = 4.0$
  $\alpha = 0.05$ $\rightarrow$ Power = 0.97
  df = n–2 = 23

  $\delta = 5.0$
  $\alpha = 0.05$ $\rightarrow$ Power = 1.0
  df = n–2 = 23

- (Textbook uses linear interpolation at $\delta = 4.22$ to find Power $\approx$ 0.9766.)
- One-sided calculations are similar.

Toluca Power analysis (cont'd)

- In R, it's a little tricky (trust us...), but for

  $\delta = 4.22$, $\alpha = 0.05$, $df = n-2 = 23$

  can use

```
> delta=4.22
> a = 0.05
> nu = 23
> pt( qt(1-(a/2),df=nu), df=nu,
                   ncp=delta, low=F )
    + pt(-qt(1-(a/2),df=nu),
                  df=nu, ncp=delta, low=T )
```

This gives power = 0.98115, which is slightly larger than that found by interpolation.

Inference on the Mean Response

- Suppose we wish to estimate the mean response $E\{Y_h\}$ at some given predictor $X = X_h$ (doesn't have to be one of the orig. X_i's).
- The LS estimator is $\hat{Y}_h = b_0 + b_1X_h$
- This is (again!) of the form $\sum k_iY_i$, so the same sorts of operations we used for b_0 and b_1 can be applied here.
- (Details are left to the adventurous reader.)

The Mean Response $E\{Y_h\}$

We find:

$E\{\hat{Y}_h\} = \beta_0 + \beta_1 X_h$ (unbiased!)

$$\sigma^2\{\hat{Y}_h\} = \sigma^2\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)$$

$$s\{\hat{Y}_h\} = \sqrt{\text{MSE}\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)}$$

(so the variance and the std. error both ↑ as X_h departs from $\bar{X}$.)

The Mean Response $E\{Y_h\}$

Also: $\hat{Y}_h \sim N(\ \beta_0 + \beta_1 X_h,\ \sigma^2\{\hat{Y}_h\}\ )$

From this, we can construct the t random variable

$$T = \frac{\hat{Y}_h - (\beta_0 + \beta_1 X_h)}{s\{\hat{Y}_h\}} \sim t(n-2)$$

Hypoth. tests and conf. ints. can be built from this reference distribution. E.g., a $1-\alpha$ conf. int. for $E\{Y_h\}$ is

$$\hat{Y}_h \ \pm\ t(1-\tfrac{\alpha}{2};\ n-2)s\{\hat{Y}_h\}$$

(but, it's valid at only a single X_h !!)

Ex. CH01TA01 (cont'd): Conf. Interval on $E\{Y_h\}$

For the LS estimate of $E\{Y_h\}$ at any $X = X_h$, use `predict()`. E.g., at $X_h = 100$:

```
> predict( CH01TA01.lm,
   newdata=data.frame(X=100),
   interval="conf", level=.90 )
       fit       lwr      upr
1   419.3861  394.9251 443.847
```

First value (`'fit'`) is $\hat{Y}_h$ at $X_h = 100$; next two (`'lwr'`,`'upr'`) are 90% conf. limits.

Prediction of Y_h

We use $\hat{Y}_h$ to estimate the mean response $E\{Y_h\}$. But, what about predicting a future observed Y?

Call this $Y_{h(new)}$ at at $X = X_{h(new)}$.

The predictor itself isn't hard, just tricky:

$Y_{h(new)} = E\{Y_{h(new)}\} + \varepsilon_h$

so: (1) estimate $E\{Y_{h(new)}\}$ with $\hat{Y}_{h(new)}$

and (2) estimate ε_h with, well, $E\{\varepsilon_h\} = 0$.

Prediction (cont'd)

This gives the predicted value as

$$\hat{Y}_{h(new)} + 0$$

or simply

$$\hat{Y}_{h(new)} = b_0 + b_1 X_{h(new)}$$

(as might be expected).

But (!) the std. error is trickier $\longrightarrow$

Prediction Error

The std. error of prediction requires us to account for variation in ε_h:

Denote the prediction variance as $\sigma^2\{pred\}$.

This is $\sigma^2\{pred\} = \sigma^2\{\hat{Y}_{h(new)} + \varepsilon_h\}$

$$= \sigma^2\{\hat{Y}_{h(new)}\} + \sigma^2\{\varepsilon_h\}$$

$$= \sigma^2\{b_0 + b_1 X_{h(new)}\} + \sigma^2\{\varepsilon_h\}$$

(assuming the two terms are indep.)

Prediction Error (cont'd)

Now,

$$\sigma^2\{pred\} = \sigma^2\left(\frac{1}{n} + \frac{(X_{h(new)} - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right) + \sigma^2$$

$$= \sigma^2\left(1 + \frac{1}{n} + \frac{(X_{h(new)} - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)$$

The associated std. error of prediction is

$$s\{pred\} = \sqrt{MSE\left(1 + \frac{1}{n} + \frac{(X_{h(new)} - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)}$$

Prediction Interval

We can show that

$$T = \frac{Y_{h(new)} - \hat{Y}_{h(new)}}{s\{pred\}} \sim t(n-2)$$

so a 1–α **prediction interval** for $Y_{h(new)}$ is

$$\hat{Y}_{h(new)} \pm t\left(1 - \tfrac{\alpha}{2};\ n-2\right) s\{pred\}$$

(Notice that s{pred} > s{$\hat{Y}_h$}: prediction involves added variation/uncertainty.)

Ex. CH01TA01 (cont'd): Prediction Interval on Y_h

For a prediction of a future Y_h at any $X = X_h$, again use `predict()`. E.g., at $X_h = 100$:

```
> predict( CH01TA01.lm,
  newdata=data.frame(X=100),
  interval="pred", level=.90 )
          fit         lwr        upr
1    419.3861  332.2072 506.5649
```

First value ('`fit`') is $\hat{Y}_{h(new)}$ at $X_h = 100$; next two ('`lwr`','`upr`') are 90% prediction limits.

Prediction Caveats

Some caveats about prediction intervals:

- **They only apply for a single $X_{h(new)}$ ("pointwise")**
- **Normality matters: robustness here is poor!**

(Also see p. 60)

Confidence Bands

To build confidence statements at more than just a single X, we turn to **simultaneous inferences**.

A **simultaneous confidence band** is a confidence statement on the mean response

$$E\{Y\} = \beta_0 + \beta_1 X$$

at all possible values of X. (That is, it is valid for every X.)

WHS Band

A confidence band for E{Y} was given by Working & Hotelling (1929) and Scheffé (1953):

$$\hat{Y}_h \pm W_\alpha s\{\hat{Y}_h\}$$

where

$$W_\alpha = \sqrt{2\,F(1-\alpha;\,2,\,n-2)}$$

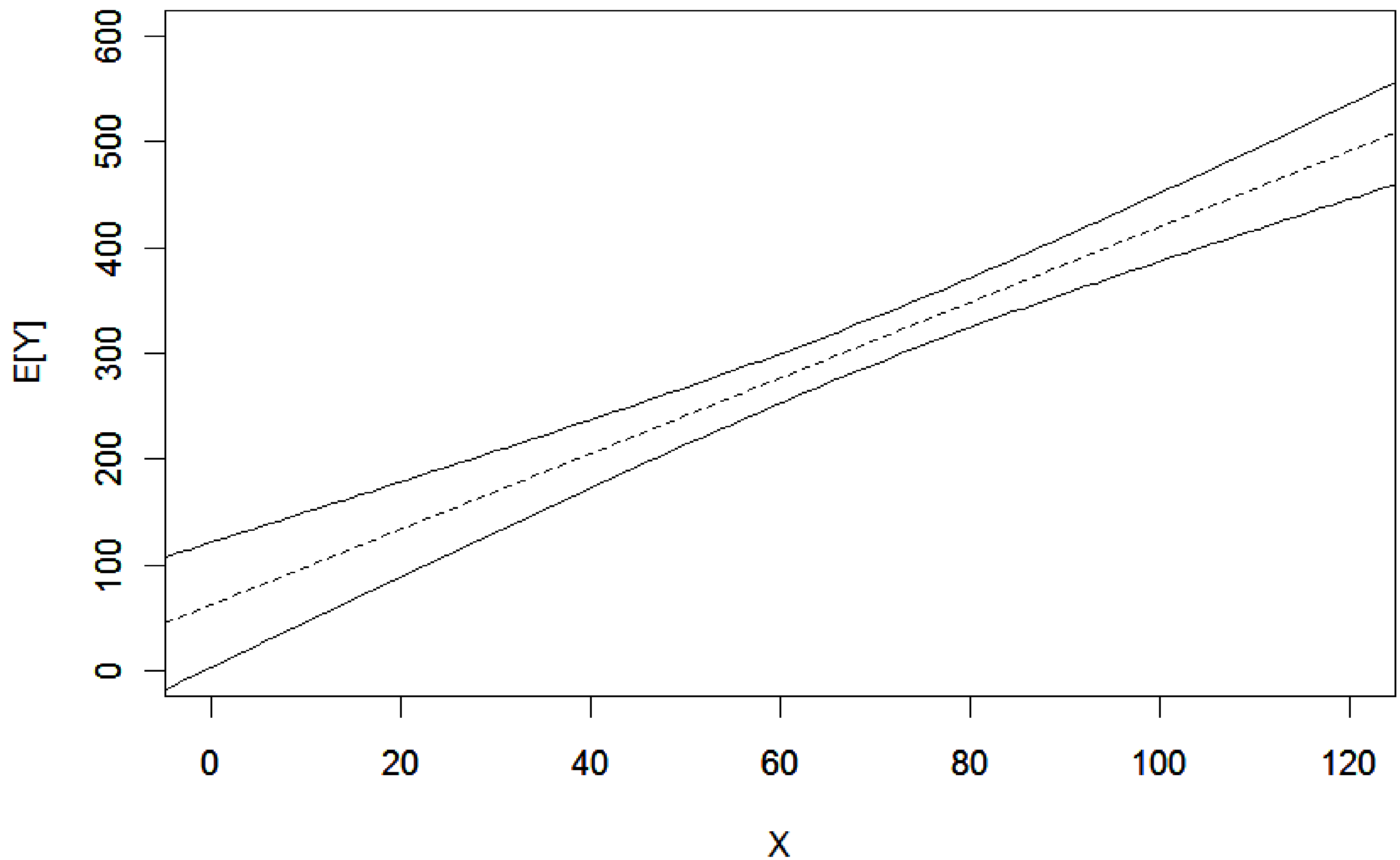
is the WHS upper-α critical point.

(Pretty simple!)

Ex. CH01TA01 (cont'd): 1 – α confidence band on E{Y}

```
> alpha = .10; n = length(Y)
> W = sqrt( 2*qf(1-alpha,2,CH01TA01.lm$df) )
> Xh = seq( from=0, to=max(X), length=100 )
> Yhat = coef( CH01TA01.lm )[1] +
                              coef( CH01TA01.lm )[2]*Xh
> se = sqrt( summary(CH01TA01.lm)$sigma^2 *( (1/n) +
                      ((Xh-mean(X))^2)/((n-1)*var(X)) ) )
> WHSlwr = Yhat - W*se
> WHSupr = Yhat + W*se
> plot( WHSlwr ~ Xh, type='l', xlim=c(0,max(X)),
                      ylim=c(0,600), xlab='', ylab='' )
> par(new = T)
> plot( WHSupr ~ Xh, type='l', xlim=c(0,max(X)),
                 ylim=c(0,600), xlab='X', ylab='E[Y]' )
```

Ex. CH01TA01 (cont'd): 1 – α confidence band on E{Y}



cf. Figure 2.6

Total Sum of Squares

The secret of statistics: to understand the mean (response), analyze the variability...

Consider the following decomposition of how Y_i varies: at the core, Y_i varies from its mean $\bar{Y}$: $Y_i - \bar{Y}$

Squaring and summing these deviations gives the Total Sum of Squares:

$$SSTO = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

Error Sum of Squares

Next, posit some model (say, the SLR) and find the predicted value $\hat{Y}_i$. This is another form of variation: $Y_i - \hat{Y}_i$

with its own sum of squares

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

(we already saw this as the error sum of squares, a.k.a. residual sum of squares)

SSTO vs. SSE

Now, if the model estimates in $\hat{Y}_i$ are no better (in terms of squared deviations) than $\bar{Y}$, we expect SSTO ≈ SSE.

But <u>if</u> the model improves upon the fit, SSTO > SSE. (Fig. 2.7 gives a nice visual.)

What makes up this difference??

SS Decomposition

$$\text{SSTO} = \sum\{Y_i - \bar{Y}\}^2 = \sum\{(Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})\}^2$$

$$= \sum\{(Y_i - \hat{Y}_i)^2 + 2(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) + (\hat{Y}_i - \bar{Y})^2\}$$

$$= \sum(Y_i - \hat{Y}_i)^2 + 2\sum(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) + \sum(\hat{Y}_i - \bar{Y})^2$$

cont'd →

SS Decomposition (cont'd)

But now,

$$\sum(Y_i - \hat{Y}_i)(\hat{Y}_i - \overline{Y})$$

$$= \sum e_i (\hat{Y}_i - \overline{Y})$$

$$= \sum e_i \hat{Y}_i - \sum e_i \overline{Y}$$

$$= \sum e_i \hat{Y}_i - \overline{Y} \sum e_i$$

$$= (0) - \overline{Y}(0) = 0$$

(from relationships seen in Ch. 1)

Regression Sum of Squares

So, we find

$$SSTO = \sum(Y_i - \hat{Y}_i)^2 + (2)(0) + \sum(\hat{Y}_i - \bar{Y})^2$$

$$= SSE + \sum(\hat{Y}_i - \bar{Y})^2$$

The latter term is what separates SSE from SSTO.

We call this the Model Sum of Squares, or for an SLR model, the Regression Sum of Squares:

$$SSR = \sum(\hat{Y}_i - \bar{Y})^2 \quad \Rightarrow \quad SSTO = SSR + SSE.$$

Degrees of Freedom

As with the sample variance, each of these SS terms is associated with a set of d.f.:

- We saw $df_E = n - 2$
- From S^2, we know $df_{TO} = n - 1$
- For SSR, it turns out that $df_R = 2 - 1 = 1$

Conveniently, $df_{TO} = df_R + df_E$

Mean Squares

With these, divide the SS terms by their d.f.'s to produce Mean Squares:

$$\mathrm{MSTO} = \frac{\mathrm{SSTO}}{df_{TO}} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}$$

$$\mathrm{MSR} = \frac{\mathrm{SSR}}{df_{R}} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{1}$$

$$\mathrm{MSE} = \frac{\mathrm{SSE}}{df_{E}} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}$$

Expected Mean Squares

We can show (p. 69) that

$$E[MSR] = \sigma^2 + \beta_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2$$

and we know

$$E[MSE] = \sigma^2 \qquad \text{(unbiased for } \sigma^2\text{)}$$

Notice that <u>if</u> $\beta_1 = 0$, MSR is another unbiased estimator of σ^2; but if not, its expectation always exceeds σ^2.

ANOVA Table

We collect all these terms together into an **Analysis of Variance (ANOVA) Table**:

Source	d.f.	SS	MS	E{MS}
Regr.	1	SSR	MSR	$\sigma^2 + \beta_1^2 \sum (X_i - \bar{X})^2$
Error	$n-2$	SSE	MSE	σ^2
Total	$n-1$	SSTO		

F-Statistic

What makes the ANOVA Table so handy is its layout of the pertinent statistics for inferences on β_1.

In partic., to test $H_o: \beta_1 = 0$ vs. $H_a: \beta_1 \neq 0$, construct the F-statistic $F^* = MSR/MSE$.

Notice that if H_o is true, $F^* \approx 1$, but if H_a is true, $F^* > 1$. This suggests a use for F^* in testing H_o.

Cochran's Theorem

We employ F^* based on a famous result:

<u>Cochran's Thm.</u>: Given $Y_i \sim$ indep.$N(\mu_i,\sigma^2)$, $i = 1,\ldots,n$, where $\mu_i = E[Y_i]$. Let

$$SSTO = SS_1 + SS_2 + \cdots + SS_{k-1}$$

where each SS_r has d.f.$=df_r$. Then if $\mu_i = \mu =$ const., the terms $SS_r/\sigma^2 \sim$ indep. $\chi^2(df_r)$ are <u>indep</u>. of $SSE/\sigma^2 \sim \chi^2(n-2)$ when

$$\sum df_r + df_E = n-1.$$

F-Reference Dist'n

From Cochran's Thm., we find for the LSR model that

$$F^* = \frac{\frac{SSR}{\sigma^2}/1}{\frac{SSE}{\sigma^2}/(n-2)} = \frac{MSR}{MSE} \sim F(1,\ n-2)$$

whenever $E\{Y_i\}$ is constant. But, a constant mean equates to $\beta_1 = 0$, i.e., H_o is true. This gives the reference dist'n for F^*.

F-Test

So, when H_o is true, the null reference dist'n for F^* is $F^* \sim F(1, n-2)$.

(When H_o is false, F^* has a <u>noncentral</u> F-dist'n.)

We reject H_o at signif. level α when

$$F^* > F(1-\alpha;\ 1,\ n-2).$$

This is called the 'full' F-test from the ANOVA table.

Ex. CH01TA01 (cont'd): ANOVA table

Recall the Toluca data. For the ANOVA table, use `anova()`:

```
> anova( CH01TA01.lm )

Analysis of Variance Table
Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X          1 252378  252378  105.88 4.45e-10
Residuals 23  54825    2384
```

Ex. CH01TA01 (cont'd): F-test

For the Toluca data, the ANOVA shows

$$F^* = 252378/2384 = 105.9.$$

Reject $H_o\!:\beta_1 = 0$ vs. $H_a\!:\beta_1 \neq 0$ when $F^* > F(1-\alpha;\, 1,\, n-2)$. At $\alpha=0.05$ this is $F^* > F(.95;\, 1,\, 23)$. Find the critical point in R:

```
> qf( 0.95,df1=1,df2=CH01TA01.lm$df )
[1] 4.279344
```

Clearly, $F^* = 105.9 > F(.95;\, 1,\, 23) = 4.28$, so we <u>reject</u> H_o.

Ex. CH01TA01 (cont'd): F vs. *t*

Note the equivalence between the F-test and the t-test for $H_o\!:\beta_1 = 0$ vs. $H_a\!:\beta_1 \neq 0$.

P-values are the same (*P* = `4.45e-10`).
And, can show $F^* = (t^*)^2$:

```
> anova( CH01TA01.lm )[1,4]
[1] 105.8757

> summary( CH01TA01.lm )$coef[2,3]^2
[1] 105.8757
```

Reduction Sum of Squares (1)

We can extend the ANOVA F-test to any form of statistical model, via 3 basic steps:

(1) Define a **FULL MODEL** (FM) with all desired components. For the SLR this is $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$. From the FM, find the SSE: $SSE(F) = \sum(Y_i - \hat{Y}_i)^2$, with $\hat{Y}_i$ found under the FM via LS.

Reduction Sum of Squares (2)

(2) For a given H_o, determine how the constraint *reduces* the model. (The **REDUCED MODEL** (RM) holds under H_o.) Then find the SSE under the RM, say $SSE(R) = \sum\{Y_i - \hat{Y}_i(R)\}^2$.

For instance, with SLR, under $H_o: \beta_1 = 0$ the RM is $Y_i = \beta_0 + \varepsilon_i$ and $SSE(R) = \sum(Y_i - \bar{Y})^2$ (which happens to = SSTO.)

Reduction Sum of Squares (3)

(3) If SSE(F) << SSE(R), the reduction in SS is "significant." An F-statistic to quantify the discrepancy is

$$F^* = \frac{SSE(R) - SSE(F)}{df_{ER} - df_{EF}} \Big/ \frac{SSE(F)}{df_{EF}}$$

Under appropriate conditions,

$F^* \sim F(df_{ER} - df_{EF},\ df_{EF})$ so reject H_o when

$$F^* > F(1-\alpha;\ df_{ER} - df_{EF},\ df_{EF})$$

as in the ANOVA Table.

Linear Association

Besides the slope parameter β_1, we can measure the linear association between Y and X using the SS terms from the ANOVA.

The reduction SS for the SLR model is SSE(R) – SSE(F) = SSTO – SSE = SSR. So, consider the ratio

$$\frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

Linear Association (cont'd)

Since SSE(R) is always ≥ SSE(F), that says SSTO ≥ SSE. But then 1 ≥ SSE/SSTO, i.e.

$$0 \leq 1 - \frac{SSE}{SSTO}$$

And, since SSE/SSTO ≥ 0, we have

$$1 - \frac{SSE}{SSTO} \leq 1$$

$$\Rightarrow \qquad 0 \leq 1 - \frac{SSE}{SSTO} \leq 1$$

R^2

We denote this as

$$R^2 = 1 - \frac{SSE}{SSTO} = \frac{SSR}{SSTO}$$

and call it the **Coefficient of Determination**.

Interpretation: $R^2 = SSR/SSTO$ is the % of total variation in the Y_is explained by the regression model.

R^2 (cont'd)

R^2 is easy to understand, but also <u>easy to overuse</u>!! (So, employ with care.)

Some features:

(a) $R^2 = 1$ when every point sits <u>on</u> the (straight) line.

(b) $R^2 = 0$ when the data are an amorphous cloud (i.e., $\beta_1 = 0$)

(c) $R^2 \to 1$ is good, but "how big is big" depends on the subject matter.

Ex. CH01TA01 (cont'd): R^2

The coeff. of determination (R^2) is in the `summary( )` output

(near bottom; previously suppressed):

```
> summary( CH01TA01.lm )
  Call:
lm(formula = Y ~ X)
        ⋮
Residual std. error: 48.82 on 23 degr. of freedom
Multiple R-squared: 0.8215,
Adjusted R-squared: 0.8138
F-stat.: 105.9 on 1 and 23 DF, p-value: 4.449e-10
> summary( CH01TA01.lm )$r.squared
[1] 0.8215335
```

R^2 Limitations

Some limitations:

(a) $R^2 \rightarrow 1$ indicates strong <u>linear</u> association, but it may be a poor fit.

See Fig. 2.9(a).

(b) $R^2 \rightarrow 0$ indicates weak <u>linear</u> association, but it may be a good nonlinear fit.

See Fig. 2.9(b).

Comments on the SLR Model

(1) **If using $\hat{Y}_h$ for future estimation or prediction at $X = X_h$, the model assumptions must continue to hold.**

(2) **If using $\hat{Y}_h$ for future estimation or prediction at $X = X_h$, *and* if X_h is also predicted, the inferences are conditional on that X_h value.**

(3) **If X_h falls outside the range of the orig. X_is, watch for extrapolation errors.**

Comments (cont'd)

(4) If we reject H_o:$\beta_1 = 0$, we <u>don't</u> necess. establish a causal relationship between X and Y. (Don't do lazy statistics!)

(5) Except for the WHS conf. band, every inference we've described is <u>pointwise</u> and valid only once. (Adjust this with "multiplicity corrections" as in Ch. 4.)

(6) If X is itself random, the inferences are approximate (or, can be "conditional").

Correlation Analysis

- **Analysis of data pairs can also be performed via measures of correlation.**
- **Similar to the SLR model on the surface, and sharing many calculations, correlation is actually a totally different model built using two random variables, Y_1 and Y_2.**
- **If the paired components are both random and prediction is not an issue, the correlation model is more appropriate.**

Correlation Model

Assume Y_1 and Y_2 have a joint probability function of the form

$$f(y_1,y_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho_{12}^2}}\exp\left\{-\frac{1}{2(1-\rho_{12}^2)}\left[\left(\frac{y_1-\mu_1}{\sigma_1}\right)^2 - 2\rho_{12}\left(\frac{y_1-\mu_1}{\sigma_1}\right)\left(\frac{y_2-\mu_2}{\sigma_2}\right) + \left(\frac{y_2-\mu_2}{\sigma_2}\right)^2\right]\right\}$$

This is the **Bivariate Normal** model, denoted as

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \sim N_2\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}\right).$$

vectors

matrix

Correlation Model (cont'd)

Marginally, we have $E\{Y_j\} = \mu_j$ and $\sigma^2\{Y_j\} = \sigma_j^2$, with $Y_j \sim N(\mu_j,\sigma_j^2)$, $j = 1,2$.

The correlation coefficient between Y_1 and Y_2 is $\rho_{12} = \sigma\{Y_1,Y_2\}/\sigma\{Y_1\}\sigma\{Y_2\}$.

If Y_1 and Y_2 are indep., then $\rho_{12} = 0$. The reverse isn't always true; however for the bivariate normal it *is*:

Y_1 and Y_2 are indep. $\Leftrightarrow$ $\rho_{12} = 0$

Conditional Distribution (1|2)

Under the bivariate normal model, the **conditional distributions** are intriguing:

Use $f(y_1|y_2) = \dfrac{f(y_1,y_2)}{f(y_2)}$ to find

$$Y_1|Y_2=y_2 \sim N(\alpha_{1|2} + \beta_{12}y_2,\ \sigma_{1|2}^2),$$

where $\alpha_{1|2} = \mu_1 - \mu_2\rho_{12}\sigma_1/\sigma_2$

$\beta_{12} = \rho_{12}\sigma_1/\sigma_2$

$\sigma_{1|2}^2 = \sigma_1^2(1-\rho_{12}^2).$

Conditional Distribution (2|1)

Similarly, $Y_2|Y_1=y_1 \sim N(\alpha_{2|1} + \beta_{21}y_1, \sigma_{2|1}^2)$,

where $\alpha_{2|1} = \mu_2 - \mu_1\rho_{12}\sigma_2/\sigma_1$

$\beta_{21} = \rho_{12}\sigma_2/\sigma_1$

$\sigma_{2|1}^2 = \sigma_2^2(1-\rho_{12}^2)$.

Notice that $E\{Y_2|Y_1=y_1\} = \alpha_{2|1} + \beta_{21}y_1$ is a linear relationship. This is often described as a "regression" of Y_2 on y_1. (Same holds for $E\{Y_1|Y_2=y_2\}$.)

Lots of Confusion...

- **The linear relation apparent in the conditional models means that <u>given</u> $Y_1=y_1$, $\alpha_{2|1}$ and β_{21} can be computed using the SLR normal equs.**
- **But *that doesn't mean the models are the same*! It's just a convenient computational coincidence.**
- **This leads to lots of confusion between correlation and regression. Bottom line: they are two different models.**

PPMCC

The goal in correlation analysis is determination of the (strength of) association between Y_1 and Y_2, using the ρ_{12} measure.

Estimate ρ_{12} with the (sample) Pearson Product-Moment Correlation Coefficient:

$$r_{12} = \frac{\sum_{i=1}^{n}(Y_{i1} - \overline{Y}_1)(Y_{i2} - \overline{Y}_2)}{\sqrt{\sum_{i=1}^{n}(Y_{i1} - \overline{Y}_1)^2 \sum_{i=1}^{n}(Y_{i2} - \overline{Y}_2)^2}}$$

(a slightly biased, ML estimator).

r_{12}

The sample correlation coeff. r_{12} satisfies

$$-1 \le r_{12} \le 1,$$

where

$r_{12} \rightarrow -1$ if Y_1, Y_2 are negatively associated

$r_{12} \rightarrow +1$ if Y_1, Y_2 are positively associated

$r_{12} \rightarrow 0$ if Y_1, Y_2 are not associated.

(Oh, by the way: $r_{12}^2 = R^2$.)

Hypothesis Test of ρ_{12}

The natural null hypoth. here is H_o: $\rho_{12} = 0$, vs. H_a: $\rho_{12} \neq 0$. Under the bivariate normal model,

$$t^* = \frac{r_{12}\sqrt{n-2}}{\sqrt{1 - r_{12}^2}} \sim t(n-2)$$

so reject H_o when $|t^*| > t\left(1 - \frac{\alpha}{2}; n-2\right)$.

The P-value is $2P[\, t(n-2) > |t^*| \,]$.

t^* is numerically identical to the t^* in (2.20) for testing $\beta_1 = 0 \Rightarrow$ tends to create confusion.

Example p. 84: Correlation

Oil Co. sales example:
study $n = 23$ gas stations and record
Y_1 = {gasoline sales}
and
Y_2 = {auxiliary product sales}.

We are given $r_{12} = 0.52$.

Wish to test if ρ_{12} is positive. Set α = 0.05.

Can do this in R →

Example p.84: Correlation

For the Oil Co. sales example, with $r_{12} = 0.52$ we can find $t^* = 2.79$ on 21 df.

To test $H_o: \rho_{12} \leq 0$ vs. $H_a: \rho_{12} > 0$, the one-sided P-value is $P[t(21) > 2.79]$. Find this in R via:

```
> pt( 2.79, df=21, lower.tail=F )
[1] 0.005486405
```

At $\alpha = 0.05$ we see $P < \alpha$, so <u>reject</u> H_o.

Confidence Limits on ρ_{12}

Conf. limits on ρ_{12} are trickier (since, e.g., ρ_{12} doesn't appear in t*).

We use the **Fisher z-Transform**:

$$z' = \frac{1}{2}\,ln\left(\frac{1 + r_{12}}{1 - r_{12}}\right)$$

For $n \geq 8$, $z' \stackrel{\cdot}{\sim} N(\,\zeta, \sigma^2\{z'\}\,)$ where

$\zeta = \frac{1}{2}\,ln\left(\frac{1 + \rho_{12}}{1 - \rho_{12}}\right)$ and $\sigma^2\{z'\} = 1/(n-3)$.

Conf. Limits on ρ_{12} (cont'd)

Notice that $(z' - \zeta)/\sigma\{z'\} \stackrel{\cdot}{\sim} N(0,1)$. So, an approx. $1-\alpha$ conf. int. for ζ is clearly

$$z' \pm z\left(1 - \tfrac{\alpha}{2}\right)\frac{1}{\sqrt{n-3}}$$

[Use the ∞ row of Table B.2 to find $z(1 - \frac{\alpha}{2})$.]

Now, reverse-transform to the ρ-scale:

$$r_{12} = \frac{e^{2z'} - 1}{e^{2z'} + 1}$$

(Table B.8 gives selected values of both transforms.)

Conf. Limits on ρ_{12} (cont'd)

So, if the z-transform produces 1–α limits on ζ of, say,

$$z'_L < \zeta < z'_U,$$

the corresp. 1–α limits on ρ_{12} are

$$\frac{e^{2z'_L} - 1}{e^{2z'_L} + 1} < \rho_{12} < \frac{e^{2z'_U} - 1}{e^{2z'_U} + 1}$$

Example p. 86: Correlation

<u>Grocery purchase example</u>:
study n = 200 households and record
Y_1 = {beef purchases}
and
Y_2 = {poultry purchases}.

We are given $r_{12} = -0.61$.

Wish to find a 95% conf. int. on the true correlation coeff. ρ_{12}.

Can do this in R →

Ex. p. 86: 1–α conf. limits on ρ_{12}

Direct R code for Fisher z′-transform:

```
> r12 = -0.61
> alpha = .05
> n = 200
> zprime = 0.5*( log(1+r12) - log(1-r12) )
> se = 1/sqrt( n-3 )
> zlwr = zprime - qnorm( 1-alpha/2 )*se
> zupr = zprime + qnorm( 1-alpha/2 )*se
> rholwr = (exp(2*zlwr)-1)/(exp(2*zlwr)+1)
> rhoupr = (exp(2*zupr)-1)/(exp(2*zupr)+1)
> c(rholwr, rhoupr)
[1] -0.6903180 -0.5148301
```

Ex. p. 86: 1–α conf. limits on ρ_{12}

Even faster, for Fisher z′-transform, are the hyperbolic tangent functions:

```
> r12 = -0.61
> alpha = .05
> n = 200

> zprime = atanh( r12 )
> se = 1/sqrt( n-3 )
> zlwr = zprime - qnorm( 1-alpha/2 )*se
> zupr = zprime + qnorm( 1-alpha/2 )*se
> c( tanh( zlwr ), tanh( zupr ) )
[1] -0.6903180 -0.5148301
```

$1-\alpha$ conf. limits on ρ_{12}

In R, can also use

- `CIr()` from *psychometric* package
- `fisherz()` suite in *psych* package
- `cor.test()` (in base *stats*) if original data pairs are available; see `help(cor.test)`

Testing $H_o: \rho_{12} = \rho_0$

The t-test for H_o: $\rho_{12} = 0$ doesn't naturally extend to testing any H_o: $\rho_{12} = \rho_o$.

Fastest solution is to build a Fisher z-transform conf. int. for ρ_{12} (as above) and reject H_o if the interval fails to contain ρ_o.

(Appeal here is to the tautology between hypoth. tests and conf. int's)

Spearman's Rank Correlation

- If the bivariate normal model doesn't hold (and a transformation of the Y_j's can't help), there is a rank-based form available, known as **Spearman's rank correlation**.
- Basic idea: replace the observations with their ranks, and then perform the corrl'n calculations on the ranks.

Rank Correlation

Step 1: Find all the Y_{i1}'s and rank them from min. to max. Call these R_{i1}.

Step 2: Repeat Step 1 for Y_{i2} to find R_{i2}. (If ties exist, give each tied value the average of the tied ranks.)

Step 3: Calculate

$$r_s = \frac{\sum_{i=1}^{n}(R_{i1} - \overline{R}_1)(R_{i2} - \overline{R}_2)}{\sqrt{\sum_{i=1}^{n}(R_{i1} - \overline{R}_1)^2 \sum_{i=1}^{n}(R_{i2} - \overline{R}_2)^2}}$$

Notice that $-1 \le r_s \le 1$.

Rank Correlation (cont'd)

Step 4: For $n \geq 10$, calculate appox. t-statistic $t^* = \dfrac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}} \mathrel{\dot{\sim}} t(n-2)$.

Step 5: Set
H_o: {no assoc. between Y_1 & Y_2}
vs.
H_a: {some assoc. between Y_1 & Y_2}

Step 6: Reject H_o when $|t^*| > t\left(1-\frac{\alpha}{2};\ n-2\right)$.

Example p. 88: Rank Correlation

<u>New Food Marketing example</u>:

study n = 12 test markets and record

Y_1 = {popl'n of market} and

Y_2 = {per cap. spending on new food product}.

Data are in Table 2.4.

Wish to test for association between Y_1 and Y_2 but can't appeal to normality

$\Rightarrow$ use Spearman's rank corrl'n.

Can do this in R $\rightarrow$

Example CH02TA04: Spearman Rank Correlation

The New Food Marketing data from Table 2.4 are

```
> Y1 = c(29, 435, ... , 89)
> Y2 = c(127, 214, ... , 103)
```

We can find r_s in R:

```
> cor( Y1, Y2, method="spearman" )
[1] 0.8951049
```

Ex. CH02TA04 (cont'd): Spearman Corrl'n Testing

To test H_o:No Y_1-vs.-Y_2 association against H_a:Some Y_1-vs.-Y_2 association via t* statistic in R, use:

```
> cor.test( Y1, Y2, method="spearman", exact=F )

        Spearman's rank correlation rho
data:  Y1 and Y2
S = 30, p-value = 8.367e-05
alternative hypothesis: true rho is not equal to 0
```

At $\alpha = 0.01$ we see $P = 8.37\times10^{-5} < \alpha$, so <u>reject</u> H_o. (For an 'exact' test, use `exact=T` option.)