Chapter 3: Inference

Stats 500, Fall 2017
Brian Thelen, University of Michigan
443 West Hall, bjthelen@umich.edu

Inference

- Estimates: $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$
- Draw conclusions about $\beta_0, \beta_1, \ldots, \beta_p$
- Two main inference tools:
  - hypothesis tests
  - confidence intervals

Savings Example

- 50 different countries
- Data from 1960 – 1970
- Response: aggregate personal savings divided by disposable income (*sr*)
- Predictors:
  - per capital disposable income (*dpi*),
  - percentage rate of change in per capita disposable income (*ddpi*),
  - percentage of population under 15 (*pop15*),
  - percentage of population over 75 (*pop75*)

```
> data(savings)
> savings
                    sr pop15 pop75     dpi  ddpi
Australia        11.43 29.35  2.87 2329.68  2.87
Austria          12.07 23.32  4.41 1507.99  3.93
... ...
> result <- lm(sr ~ pop15 + pop75 + dpi + ddpi,
    savings)
```

```
> summary(result)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.5660865  7.3545161   3.884 0.000334
pop15       -0.4611931  0.1446422  -3.189 0.002603
pop75       -1.6914977  1.0835989  -1.561 0.125530
dpi         -0.0003369  0.0009311  -0.362 0.719173
ddpi         0.4096949  0.1961971   2.088 0.042471

Residual standard error: 3.803 on 45 degrees of freedom
Multiple R-Squared: 0.3385, Adjusted R-squared: 0.2797
F-statistic: 5.756 on 4 and 45 DF,  p-value: 0.0007904
```

Savings Example Ctd

- Is *pop75* significant in the full model?
- Estimation from the data:

$$\widehat{sr} = 28.6 - 0.46 \times pop15 - 1.69 \times pop75 - 0.0003 \times dpi + 0.41 \times ddpi$$

  Q: Is "-1.69" random fluctuation due to chance? Or does it indicate that the coefficient is truly different from 0?

- Each test only makes sense in the context of the fitted model

Hypothesis Tests

- Testing: use probability to decide whether data is consistent with hypothesis
- Null hypothesis H_0 (e.g. $\beta_{pop75} = 0$)
- Alternative hypothesis H_A (e.g. $\beta_{pop75} \neq 0$)
- Decide whether data is consistent with H_0:
  - If no, reject H_0 and accept H_A
  - Otherwise, fail to reject H_0

Errors in Hypothesis Testing

		True State	
		H_0 true	H_0 false
Our Decision	Not reject H_0	✔	Type II error
	Reject H_0	Type I error	✔

Usually type I error is more serious

The legal system analogy

- H_0: The accused is innocent
- H_A: The accused is guilty
- Type I error: convict an innocent person
- Type II error: acquit a guilty person

Presumption of innocence:

- H_0 assumed true unless there is convincing evidence for H_A
- H_A carries the "burden of proof"

Procedure

- Set $\alpha = Pr(\text{type I error})$. Typically $\alpha = 0.05$ or 0.01. α is called the significance level.
- Compute p-value: the probability of observed data or even more extreme departure from H_0 (in favor of H_A) when H_0 is true.
- If p-value $< \alpha$, reject H_0.

Savings Example

Full model:

$$sr = \beta_0 + \beta_{pop15} \times pop15 + \beta_{pop75} \times pop75 + \beta_{dpi} \times dpi + \beta_{ddpi} \times ddpi$$

- Null hypothesis: $\beta_{pop75} = 0$
- Alternative hypothesis: $\beta_{pop75} \neq 0$

We observe that

$$\widehat{sr} = 28.6 - 0.46 \times pop15 - 1.69 \times pop75 - 0.0003 \times dpi + 0.41 \times ddpi$$

Therefore, the p-value is

$$Pr(|\hat{\beta}_{pop75}| \geq 1.69 \mid \beta_{pop75} = 0)$$

Further Assumption on Errors

We have only assumed $E(\epsilon) = 0$, $var(\epsilon) = \sigma^2 I$, and ϵ_i are i.i.d. To compute the p-value, we also need to assume a distribution for the errors ϵ. Usually

$$\epsilon \sim Normal(0, \sigma^2 I)$$

Distribution of $\hat{\beta}$

Then

$$\begin{aligned}\hat{\beta} &\sim N(\beta, (X^T X)^{-1}\sigma^2) \\ \hat{\beta}_j &\sim N(\beta_j, (X^T X)^{-1}_{jj}\sigma^2)\end{aligned}$$

Distribution of $\hat{\beta}$ Ctd

Let

$$\begin{aligned} sd(\hat{\beta}_j) &= \sqrt{(X^TX)^{-1}_{jj}\sigma^2} \\ se(\hat{\beta}_j) &= \sqrt{(X^TX)^{-1}_{jj}\hat{\sigma}^2} \\ \text{Recall} \quad \hat{\sigma}^2 &= \frac{\sum_i (y_i - \hat{y}_i)^2}{n-(p+1)} \end{aligned}$$

Distribution of $\hat{\beta}$ Ctd

It turns out that

$$\frac{\hat{\beta}_j - \beta_j}{sd(\hat{\beta}_j)} \sim N(0,1)$$

$$\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t_{n-(p+1)}$$

t_{df}-distribution

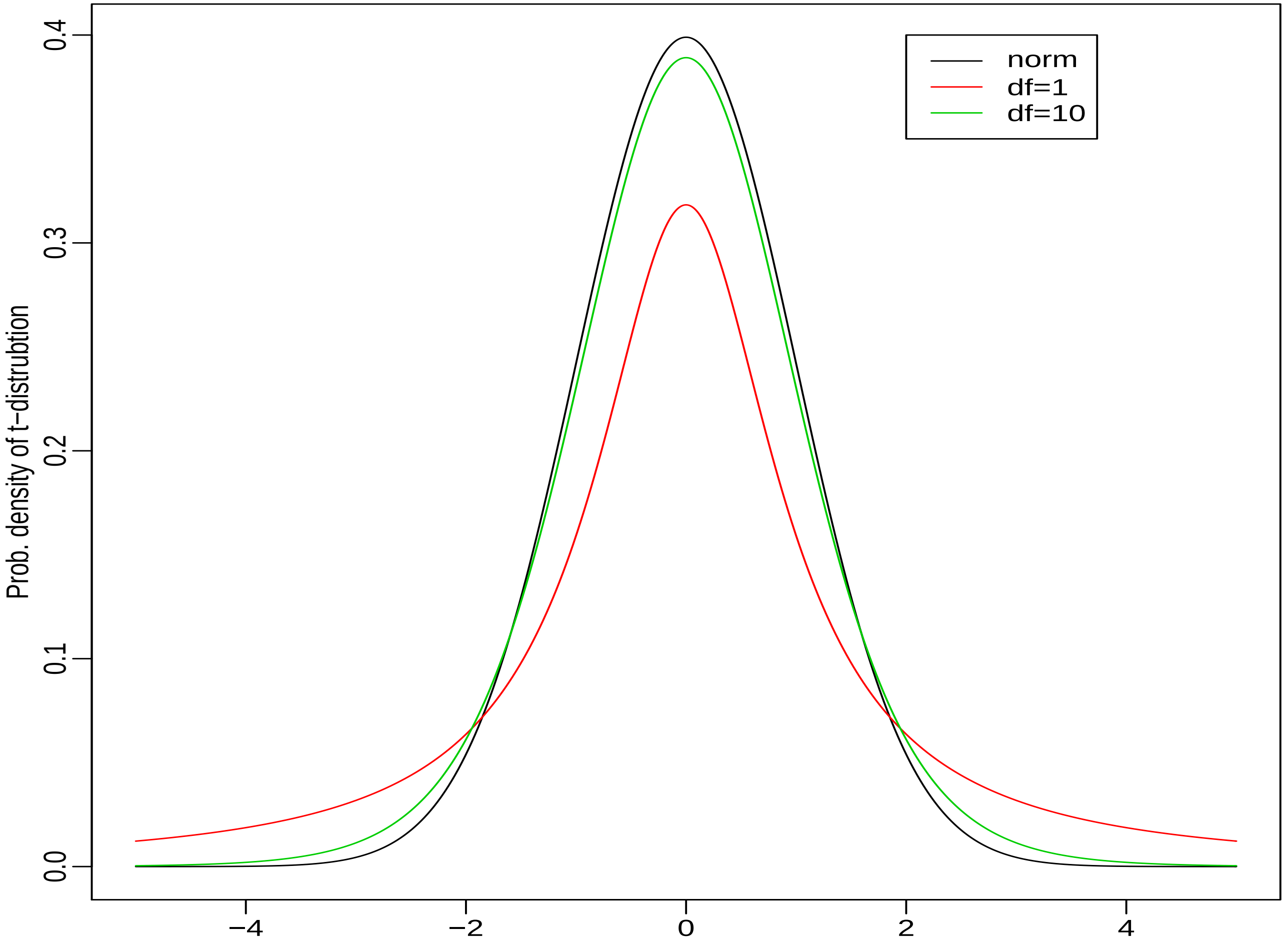
Probability density functions (pdf):

$$N(0,1) \sim \frac{1}{\sqrt{2\pi}} e^{-1/2 z^2}$$

$$t_{df} \sim \frac{\Gamma\left(\frac{df+1}{2}\right)}{\sqrt{df \cdot \pi} \cdot \Gamma(\frac{df}{2})} (1 + z^2/df)^{-(df+1)/2}$$

- Symmetric around 0, “bell-shaped”, but heavier tails than normal
- As $df \to \infty$, $t_{df} \to N(0,1)$

t_{df}-distribution Ctd



t_{df}: like normal distribution with wider tails

t-statistic (Savings Example)

If the Null is true, i.e. $\beta_{pop75} = 0$, then

$$\frac{\hat{\beta}_{pop75}}{se(\hat{\beta}_{pop75})} \sim t_{50-(4+1)}$$

From the R output, we have (t-statistic)

$$\frac{\hat{\beta}_{pop75}}{se(\hat{\beta}_{pop75})} = -1.56$$

t-statistic (Savings Example) Ctd

Is this value extreme for the t_{45} distribution? Or what is the probability of

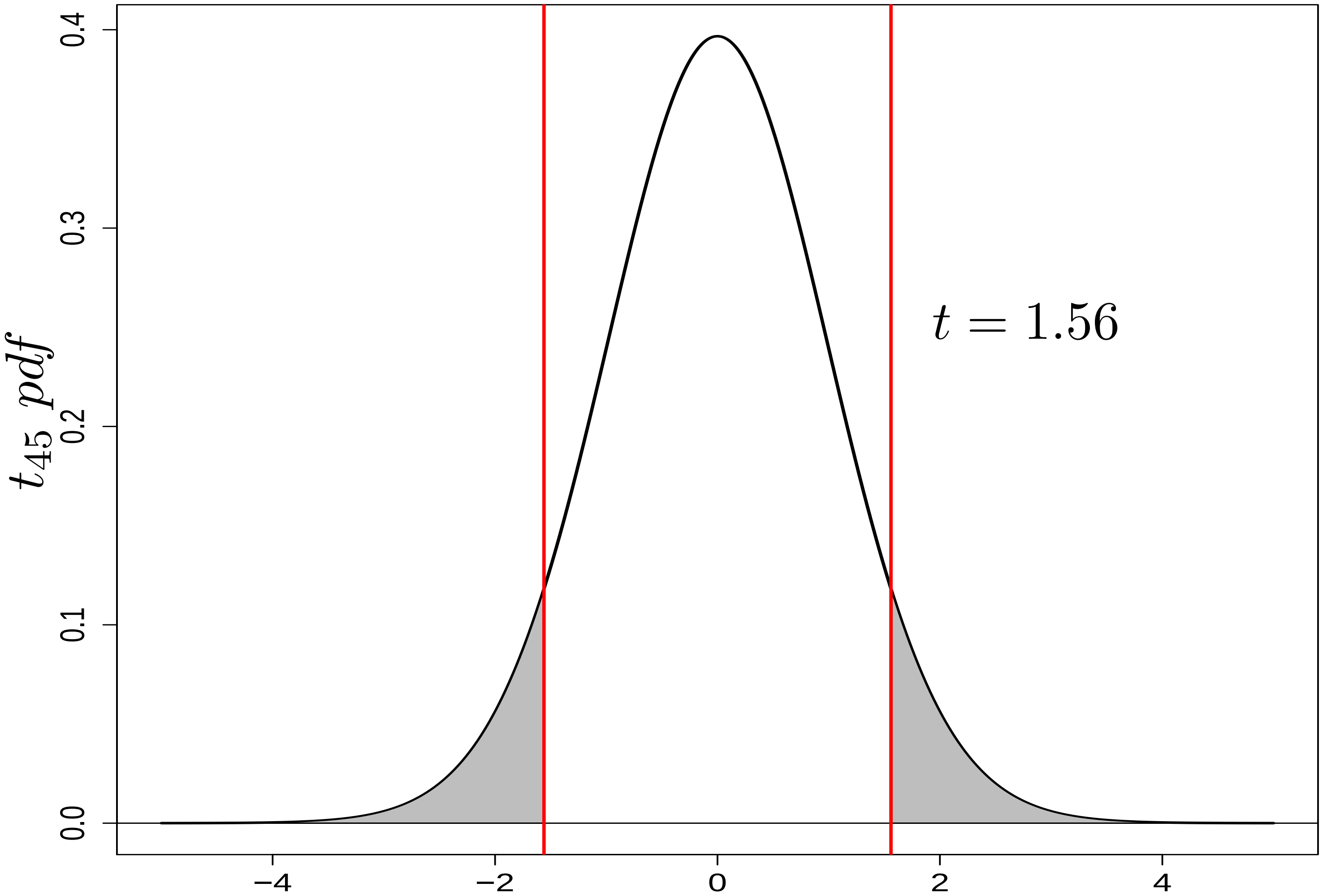
$$Pr(\text{observe “-1.56” or more extreme}|\beta_{pop75} = 0)$$

i.e.

$$Pr(|t_{45}| \geq 1.56) = ?$$

What if the test is one-sided?

t-statistic (Savings Example) Ctd



t-test

- Two-sided test: $Pr(|t_{45}| \geq 1.56) = 0.13 > \alpha = 0.05$, therefore we fail to reject H_0.
- Is *pop75* significant in the full model? Probably not.
- ```
  ## CDF of t-distribution
  > pt(1.56, df=45)
  [1] 0.937117
  > 2*(1 - pt(1.56, df=45))
  [1] 0.1257658
  ```
  ```

Another (General) Approach

- Recall RSS: residual sum of squares $\sum_i \hat{\epsilon}_i^2$
- Fit a model under H_0, compute RSS_{H_0} (e.g. with β_{pop75} set equal to 0)
- Fit another model under $H_0 \cup H_A$, compute $RSS_{H_0 \cup H_A}$ (e.g. no restriction on β_{pop75})
- Compute

$$F = \frac{(RSS_{H_0} - RSS_{H_0 \cup H_A})/(df_{H_0} - df_{H_0 \cup H_A})}{RSS_{H_0 \cup H_A}/df_{H_0 \cup H_A}}$$

General Approach Ctd

- If H_0 is true,

$$F \sim F_{df_1,df_2}; \quad df_1 = df_{H_0} - df_{H_0 \cup H_A}, df_2 = df_{H_0 \cup H_A}$$

- Compute p-value $= Pr(F_{df_1,df_2} > F)$

F-distribution

- $Z_1, \ldots, Z_n$ i.i.d. Normal(0,1). Then

$$U = Z_1^2 + \cdots + Z_n^2$$

has χ^2 (chi-square) distribution with n degrees of freedom.

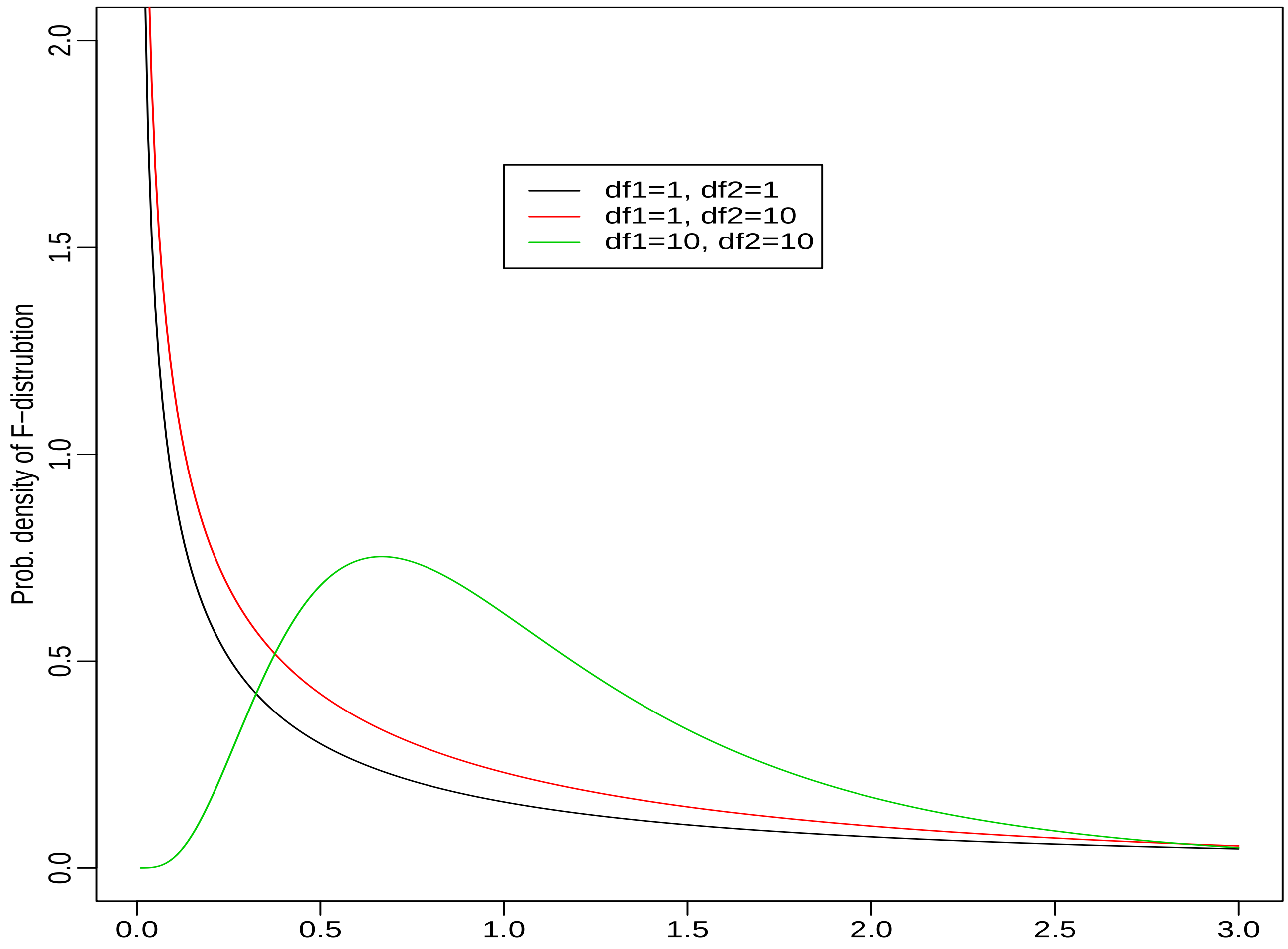
- χ_n^2 is the same as $Gamma(n/2, 2)$.

- Suppose $U \sim \chi_n^2$, $W \sim \chi_m^2$ are independent. Then

$$\frac{U/n}{W/m} \sim F_{n,m}$$

F-distribution with n and m degrees of freedom.

F-distribution



Important facts: (1) $F_{df_1,df_2} > 0$ (2) $t_{df}^2 \sim F_{1,df}$

F-test: Savings Example

```
## Model under H0
> h0 <- lm(sr ~ pop15 + dpi + ddpi, savings)
> summary(h0)

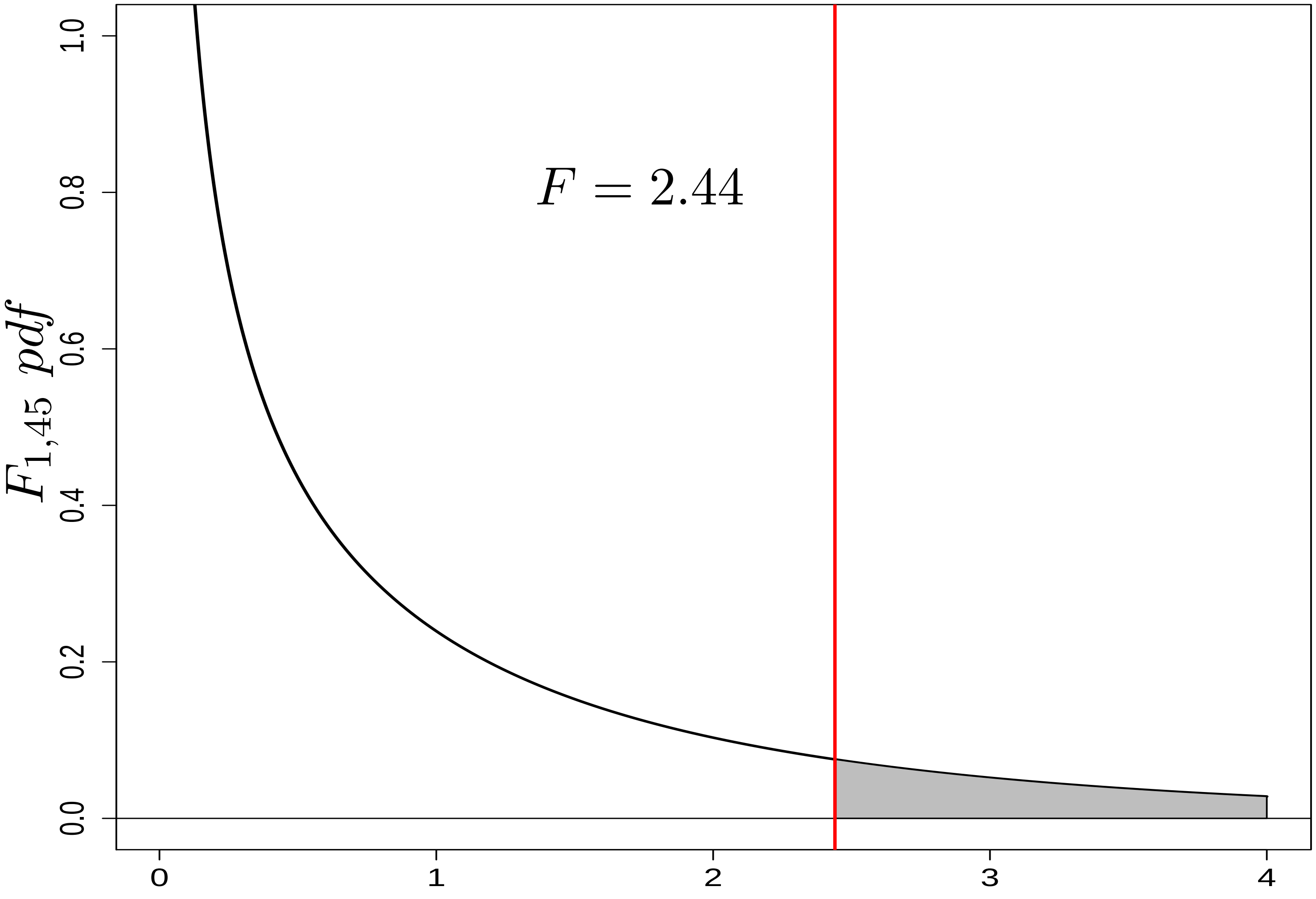
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.2771687  4.3888974   4.392 6.53e-05
pop15       -0.2883861  0.0945354  -3.051  0.00378
dpi         -0.0008704  0.0008795  -0.990  0.32755
ddpi         0.3929355  0.1989390   1.975  0.05427

## Model under (H0 U HA)
> h0a <- lm(sr ~ pop15 + pop75 + dpi + ddpi, savings)
```

```
> anova(h0, h0a)
Analysis of Variance Table

Model 1: sr ~ pop15 + dpi + ddpi
Model 2: sr ~ pop15 + pop75 + dpi + ddpi
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     46 685.95
2     45 650.71  1     35.24 2.4367 0.1255
```

F-test Ctd



F-test and t-test

- $Pr(F_{1,45} > 2.44) = 0.13 > \alpha = 0.05$, therefore we fail to reject H_0.
- Notice $t^2 = 1.56^2 = 2.44 = F$
- F-test and two-sided t-test are equivalent for testing a single predictor.

Test a Pair

- Whether both *pop75* and *dpi* can be excluded from the model.
- H_0: $\beta_{pop75} = \beta_{dpi} = 0$; H_A: not H_0.

```
> h0 <- lm(sr ~ pop15 + ddpi, savings)
> h0a <- lm(sr ~ pop15 + pop75 + dpi + ddpi, savings)
> summary(h0)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.59958    2.33439   6.682 2.48e-08
pop15       -0.21638    0.06033  -3.586 0.000796
ddpi         0.44283    0.19240   2.302 0.025837
```

```
> anova(h0, h0a)
Analysis of Variance Table

Model 1: sr ~ pop15 + ddpi
Model 2: sr ~ pop15 + pop75 + dpi + ddpi
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     47 700.55
2     45 650.71  2     49.84 1.7233 0.1900
```

- What if we want to test whether any of the predictors are useful in predicting the response?
- H_0: $\beta_{pop15} = \beta_{pop75} = \beta_{dpi} = \beta_{ddpi} = 0$

Test a Subspace

- Whether the effect of young people and the effect of old people on the savings rate are the same.
- H_0: $\beta_{pop15} = \beta_{pop75}$; H_A: $\beta_{pop15} \neq \beta_{pop75}$

```
> h0 <- lm(sr ~ I(pop15 + pop75) + dpi + ddpi, savings)
> h0a <- lm(sr ~ pop15 + pop75 + dpi + ddpi, savings)
> summary(h0)

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      21.6093051  4.8833633   4.425 5.87e-05
I(pop15 + pop75) -0.3336331  0.1038679  -3.212  0.00241
dpi              -0.0008451  0.0008444  -1.001  0.32212
```

```
ddpi              0.3909649  0.1968714   1.986  0.05302

Residual standard error: 3.827 on 46 degrees of freedom
Multiple R-Squared: 0.3152, Adjusted R-squared: 0.2705
F-statistic: 7.056 on 3 and 46 DF,  p-value: 0.0005328

> anova(h0, h0a)
Analysis of Variance Table

Model 1: sr ~ I(pop15 + pop75) + dpi + ddpi
Model 2: sr ~ pop15 + pop75 + dpi + ddpi
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     46 673.63
2     45 650.71  1     22.91 1.5847 0.2146
```

Test another Subspace

- Test whether β_{ddpi} is equal to 0.5
- H_0: $\beta_{ddpi} = 0.5$; H_A: $\beta_{ddpi} \neq 0.5$

```
> h0 <- lm(sr ~ pop15 + pop75 + dpi + offset(0.5*ddpi),
    savings)
> summary(h0)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 27.9287866  7.1608589   3.900 0.000311
pop15       -0.4543714  0.1426430  -3.185 0.002596
pop75       -1.7187908  1.0726662  -1.602 0.115923
dpi         -0.0002274  0.0008925  -0.255 0.800004
```

```
> h0a <- lm(sr ~ pop15 + pop75 + dpi + ddpi, savings)
> anova(h0, h0a)
Analysis of Variance Table

Model 1: sr ~ pop15 + pop75 + dpi + offset(0.5 * ddpi)
Model 2: sr ~ pop15 + pop75 + dpi + ddpi
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     46 653.78
2     45 650.71  1      3.06 0.2119 0.6475
```

- What about using t-test?

Confidence Intervals

Why do we care about CI?

- Hypothesis test: yes/no only
- Dependence on sample size
- Statistical significance vs. practical significance

Confidence Intervals for β_j

Consider each parameter individually.

$$\text{Recall } \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t_{n-(p+1)}$$

Hence

$$Pr\left(-t^{(\alpha/2)}_{n-(p+1)} \leq \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \leq t^{(\alpha/2)}_{n-(p+1)} \right) = 1 - \alpha$$

Or with probability $1 - \alpha$, i.e. confidence $100(1-\alpha)\%$

$$\hat{\beta}_j - t^{(\alpha/2)}_{n-(p+1)} \cdot se(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + t^{(\alpha/2)}_{n-(p+1)} \cdot se(\hat{\beta}_j)$$

$t^{(\alpha)}$ is the tail probability: $Pr(t > t^{(\alpha)}) = \alpha$.

Confidence Intervals for β_j Ctd

- General form:

$$\text{estimate} \pm \text{critical value} \times \text{s.e. of estimate}$$

- Two-sided t-test and CI

Savings Example

```
> result <- lm(sr ~ pop15 + pop75 + dpi + ddpi, savings)
> summary(result)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.5660865  7.3545161   3.884 0.000334
pop15       -0.4611931  0.1446422  -3.189 0.002603
pop75       -1.6914977  1.0835989  -1.561 0.125530
dpi         -0.0003369  0.0009311  -0.362 0.719173
ddpi         0.4096949  0.1961971   2.088 0.042471

## Convenient way to compute CIs
> conf <- confint(result)
```

```
> conf
                   2.5 %       97.5 %
(Intercept) 13.753330728 43.378842354
pop15       -0.752517542 -0.169868752
pop75       -3.873977955  0.490982602
dpi         -0.002212248  0.001538444
ddpi         0.014533628  0.804856227
```

Simultaneous Confidence Regions

Similarly,

$$\frac{(\hat{\beta}-\beta)^T X^T X(\hat{\beta}-\beta)}{(p+1)\hat{\sigma^2}} \sim F_{p+1,n-(p+1)}$$

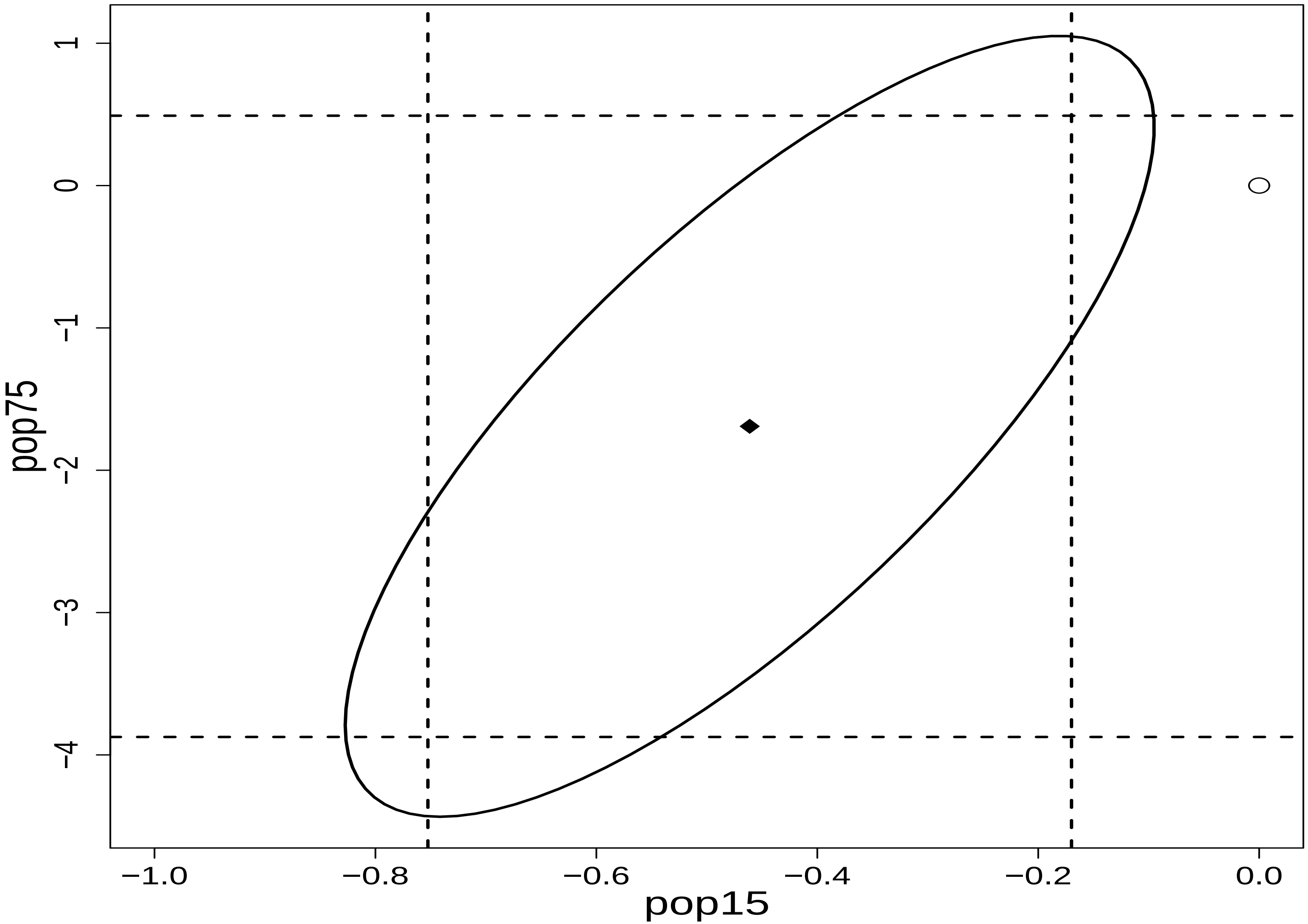
With probability $1-\alpha$, i.e. confidence $100(1-\alpha)\%$

$$(\hat{\beta}-\beta)^T X^T X(\hat{\beta}-\beta) \leq (p+1)\hat{\sigma^2} F^{(\alpha)}_{p+1,n-(p+1)}$$

Savings Example

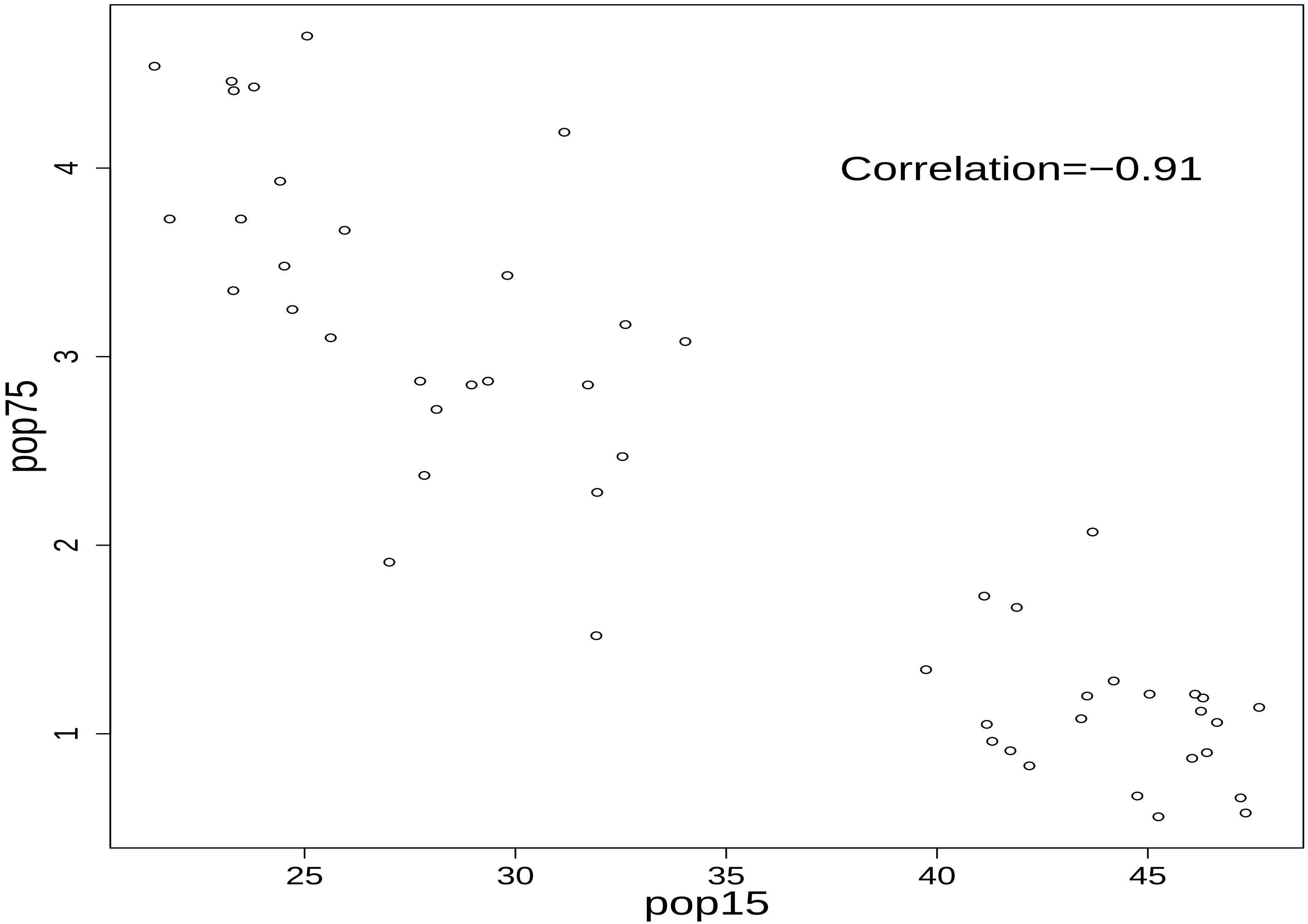
```
## Need to install the "ellipse" package
> library(ellipse)
## Plot the confidence region
> plot(ellipse(result, c('pop15', 'pop75')),
    type="l", xlim=c(-1,0))
## Add the estimates to the plot
> points(result$coef['pop15'], result$coef['pop75'],pch=18)
## Add the origin to the plot
> points(0, 0, pch=1)
## Add the confidence interval for pop15
> abline(v=conf['pop15',], lty=2)
## Add the confidence interval for pop75
> abline(h=conf['pop75',], lty=2)
```

Savings Example: Confidence region



```
## Correlation between pop15 and pop75
> plot(x=savings$pop15, y=savings$pop75)
> cor(savings$pop15, savings$pop75)
[1] -0.9084787
```

Correlation between predictors



Confidence Intervals for Predictions

- Given new predictors, x_0, what is the predicted response?

$$\hat{y}_0 = x_0^T \hat{\beta}$$

- Two types of predictions:
  - Prediction of a future observation
  - Prediction of the future mean response
- Prediction intervals vs. confidence intervals

Confidence Intervals for Predictions Ctd

For a future observation:

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

For the future mean response:

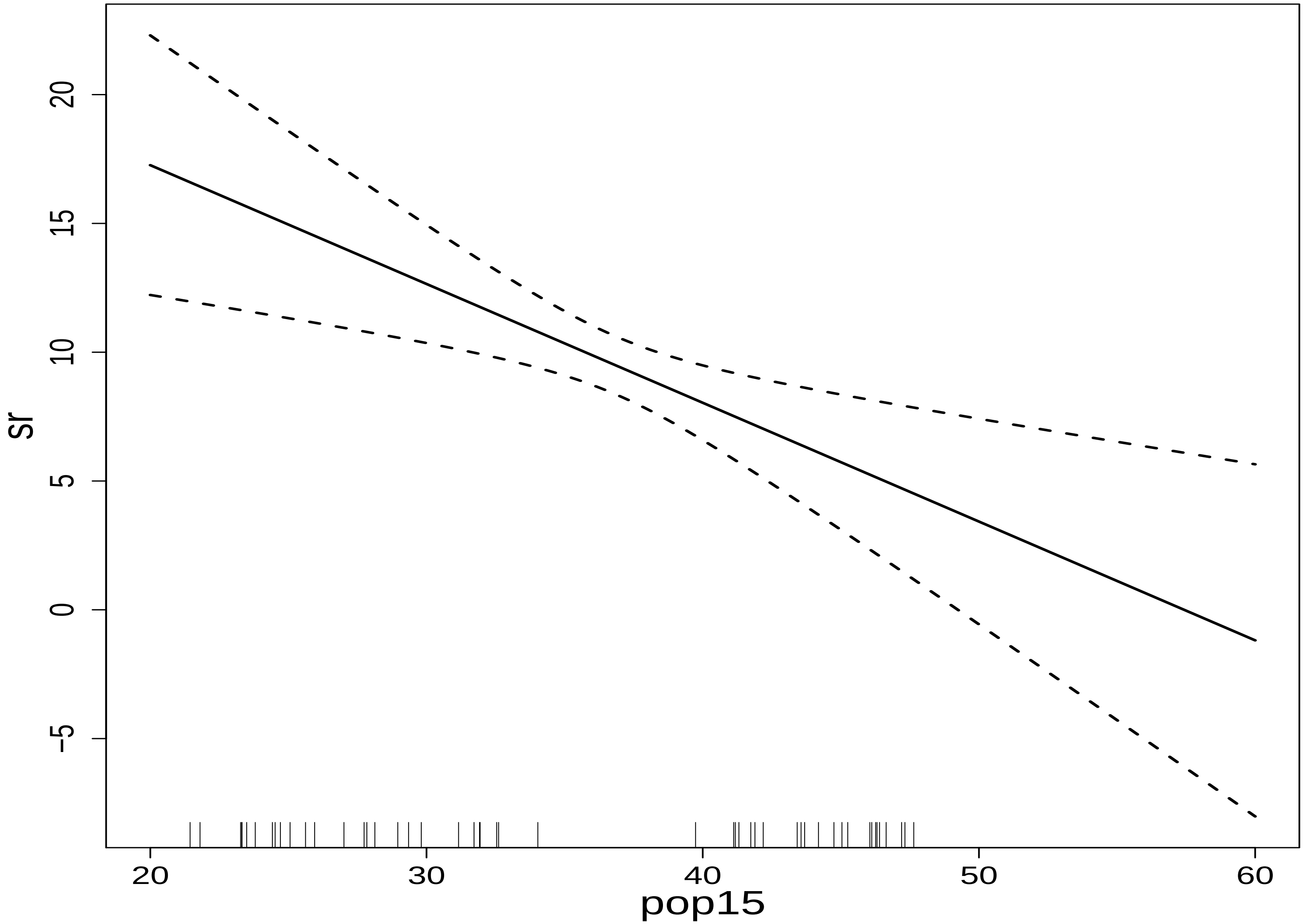
$$\hat{y}_0 \pm t_{n-(p+1)}^{(\alpha/2)} \hat{\sigma} \sqrt{x_0^T (X^T X)^{-1} x_0}$$

Savings Example

```
> result <- lm(sr ~ pop15 + pop75 + dpi + ddpi, savings)
## Convenient way to compute CI and PI
> x0 <- data.frame(pop15=35, pop75=2, dpi=1000, ddpi=4)
> predict(result, x0, interval="confidence")
          fit      lwr      upr
[1,] 10.34321 9.093452 11.59297
> predict(result, x0, interval="prediction")
          fit      lwr      upr
[1,] 10.34321 2.582946 18.10347
```

```
## Generate a sequence of points
> grid <- seq(20, 60, 1)
> pred <- predict(result, data.frame(pop15=grid, pop75=2,
    dpi=1000, ddpi=4), interval="confidence")
## Plot a matrix
> matplot(grid, pred, lty=c(1,2,2), col=1, type="l",
    xlab="pop15", ylab="sr")
> rug(savings$pop15)
```

Prediction Band Plot



Interpreting Parameter Estimates

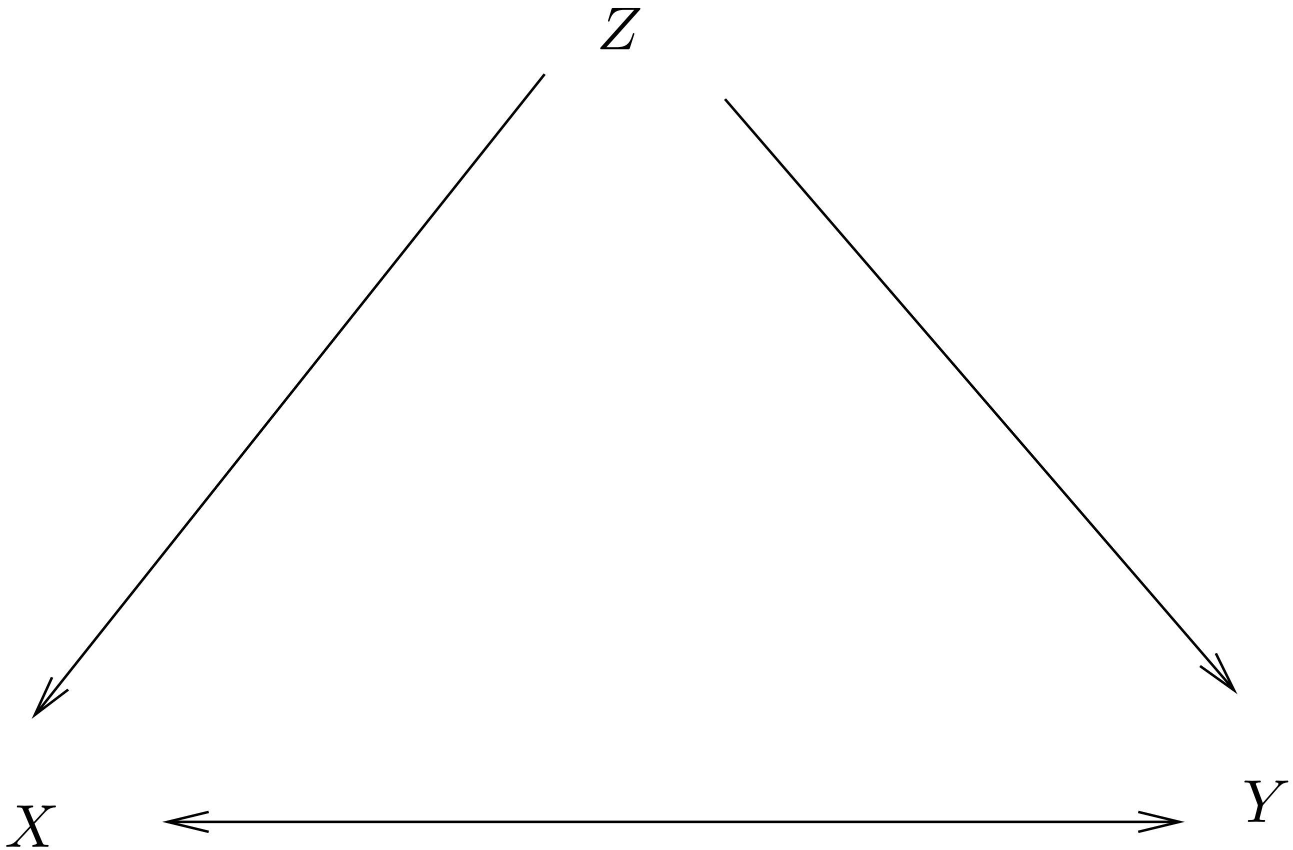
- What does $\hat{\beta}_1$ mean? A unit change in x_1 will produce a change of $\hat{\beta}_1$ in the response?
- Causal conclusion?
- Easier for orthogonal predictors (designed experiments): if $X = [X_1, X_2]$ and $X_1^T X_2 = 0$, then the coefficients of predictors in X_1 are the same regardless of whether X_2 is in the model or not.
- Randomization can reduce the effect of unknown predictors not included in the model

Interpreting Parameter Estimates Ctd

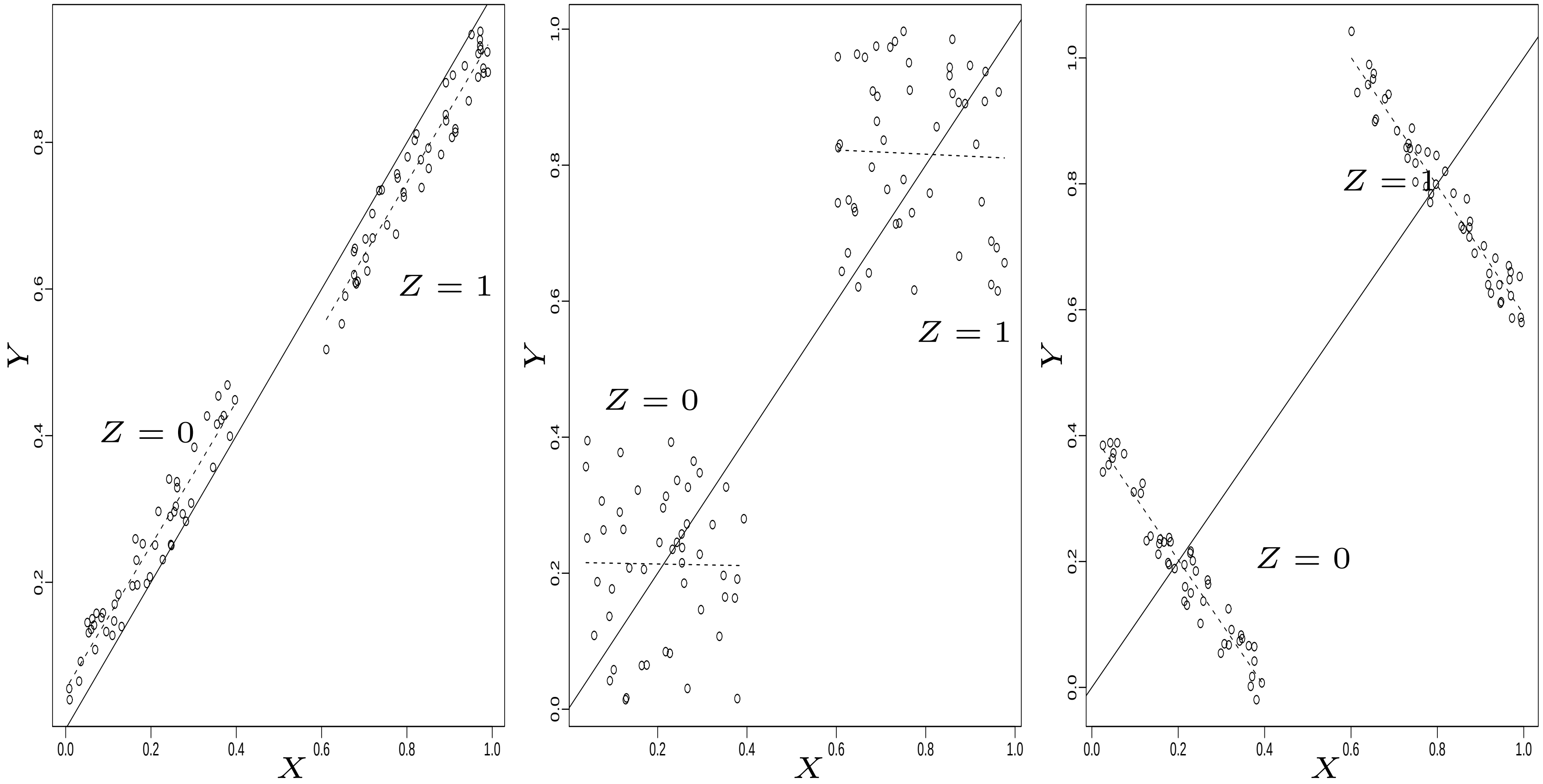
- For observational data, causal conclusion problematic
  - An unmeasured lurking variable Z may be the real cause or be a confounding variable.
  - Still not OK even if all relevant variables have been measured

$$
\begin{aligned}
y &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \\
y &= \hat{\beta}_0 + (\hat{\beta}_1 - \hat{\beta}_2) x_1 + \hat{\beta}_2 (x_1 + x_2)
\end{aligned}
$$

Lurking Variables



Confounding Variables



Interpreting Parameter Estimates Ctd

- Interpretation cannot be separately done for each variable.
- New interpretation: $\hat{\beta}_1$ is the effect of x_1 when all other predictors are held constant.
- Technically correct, but problematic in practice
- Conclusion: no simple solution

Interpreting Predictions

- True parameter values may never be known
- Concentrate on predicting future responses
- Conceptually simpler, but need to worry about extrapolation