Chapter 4

Multiple Regression Analysis – Inference

Jim Bang

Sampling Distribution of $\hat{\beta}$

BLUE to BUE

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$$

Assumptions of Gauss-Markov

1. Linear in the parameters
2. Random Sampling
3. Zero conditional mean of errors
4. Homoskedasticity
5. No Perfect Multicollinearity
6. Large Outliers are Unlikely

Under these assumptions, OLS is BLUE - minimum variance among all *linear* unbiased estimators

Assumptions of the Classical Linear Model

1. Linear in the parameters
2. Random Sampling
3. Zero conditional mean of errors
4. Homoskedasticity
5. No Perfect Multicollinearity
6. Large Outliers are Unlikely
7. $E(y|X) \sim N(X\beta, \sigma^2)$

Under these assumptions, OLS is not only BLUE but MVUE - minimum variance among *all* unbiased estimators, not just linear ones.

Distribution of $\hat{\beta}$

Under the assumptions of the CLM,

$$\hat{\beta}_j \sim N(\beta_j, Var(\hat{\beta}_j))$$

When σ^2 is known,

$$\hat{\beta}_j / \sigma_{\hat{\beta}_j} \sim N(0, 1)$$

When σ^2 is unknown,

$$\hat{\beta}_j / s_{\hat{\beta}_j} \sim t_{n-k-1} = t_{df}$$

Hypothesis Testing

Steps

1. State the Null & Alternative Hypotheses
2. Determine the Significance Level (α)
3. Estimate the parameter(s) and the test statistic(s)
4. Calculate the critical value or p-value
5. Make a Rejection Decision

Types of Null/Alternative Hypotheses

1. One-Sided Alternatives

a. Left-Tailed Alternative

$$H_0 : \beta \geq 0$$
$$H_1 : \beta < 0$$

b. Right-Tailed Alternative

$$H_0 : \beta \leq 0$$
$$H_1 : \beta > 0$$

2. Two-Sided

$$H_0 : \beta = 0$$
$$H_1 : \beta \neq 0$$

Determining α

Type I and Type II Errors

		Null Hypothesis:	
		True	False
Data Conclusion:	Accept H_0	No Error	Type II Error
	Reject H_0	Type I Error	No Error

Significance (α) measures the probability of observing data different from H_0 when H_0 is true.

$$\alpha = P[\hat{T} > T_c | H_0] = P[\text{Type I Error}]$$

Significance and Power

Power (B) measures the probability of observing data different from H_0 when H_1 is true.

$$B = P[\hat{T} > T_c | H_1] = 1 - P[\text{Type II Error}] = 1 - F(T_c - \frac{\beta_{H_1}}{s_{\hat{\beta}}})$$

For a given magnitude of difference between β_{H_0} and β_{H_1} and a given sample size, there is a trade-off between achieving a lower significance level and achieving a higher power.

Survey designs typically determine their minimum sample size by calculating the number of observations required to detect a predetermined minimum effect size to be considered “important” at a predetermined highest-acceptable significance level (usually 0.05) at a predetermined lowest-acceptable power (often 0.8)

Does Experience *Increase* Wages?

Using `wage1`, regress `wage` on `exper`, controlling for `educ` and `tenure`. Call the output `wage.lm6` and print a `summary()` of the results.

```
wage.lm6 <- lm(wage ~ exper + educ + tenure, data = wage1)
summary(wage.lm6)
```

```
##
## Call:
## lm(formula = wage ~ exper + educ + tenure, data = wage1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6068 -1.7747 -0.6279  1.1969 14.6536
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.87273    0.72896  -3.941 9.22e-05 ***
## exper        0.02234    0.01206   1.853   0.0645 .
## educ         0.59897    0.05128  11.679  < 2e-16 ***
## tenure       0.16927    0.02164   7.820 2.93e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.084 on 522 degrees of freedom
## Multiple R-squared:  0.3064, Adjusted R-squared:  0.3024
## F-statistic: 76.87 on 3 and 522 DF,  p-value: < 2.2e-16
```

Critical Values and p-Values

$$H_0 : \beta = 0$$

$$H_1 : \beta > 0$$

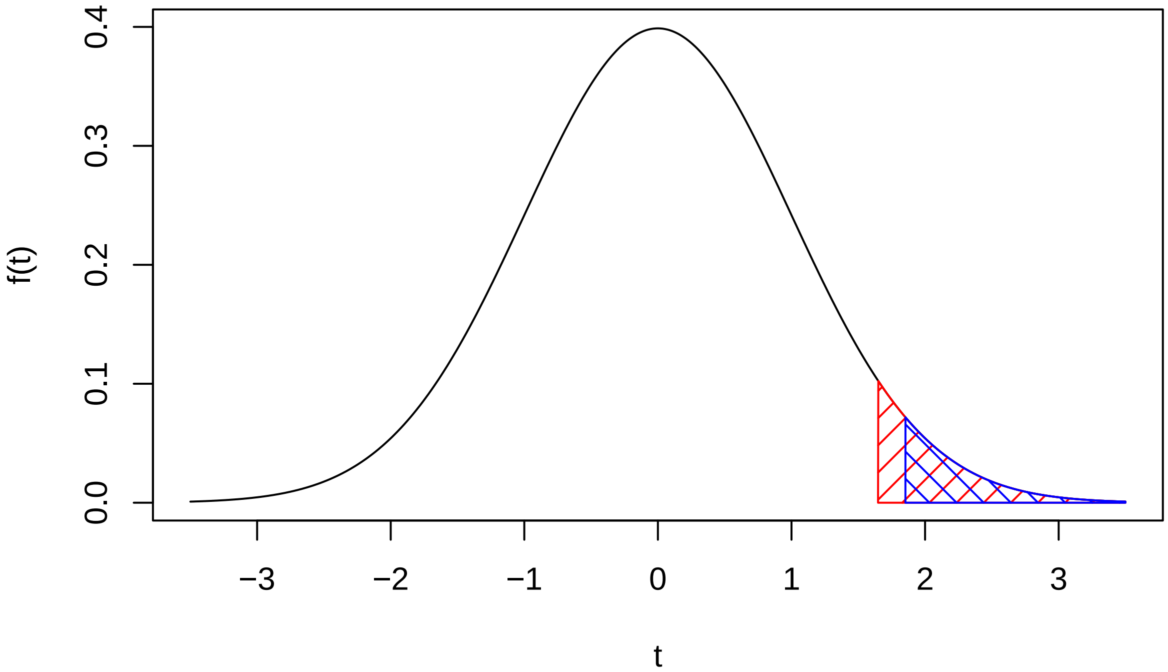
- What is the t-critical value for a test for whether the effect of experience is *greater than* zero and a significance of 0.05?
  a. 1.645
  b. -1.645
  c. 1.965
  d. -1.965
- What is the p-value?
  a. 0.0645
  b. 0.9355

c. 0.9678
d. 0.0322

Critical Values and p-Values in `R`

The following code would calculate and graph the critical value with its corresponding rejection region, along with a horizontal line for the test value.

```
# Define values for x and y axes, and the critical and test values.
x <- seq(-3.5, 3.5, length = 1000)
y <- dt(x, wage.lm6$df.residual)
t.critval <- qt(0.95, wage.lm6$df.residual)
t.testval <- summary(wage.lm6)$coefficients['exper', 't value']
# Plot the t distribution with n-k-1 degrees of freedom and sensibly-labeled axes.
plot(x,
     y,
     type = "l",
     ylab = "f(t)",
     xlab = "t")
# Add the polygon for the right-tailed, alpha = 0.05 t-critical value.
polygon(c(x[x >= t.critval], max(x), t.critval),
        c(y[x >= t.critval], 0, 0),
        col = "red",
        density = 10)
# Add the polygon for the p-value.
polygon(
  c(x[x >= t.testval], max(x), t.testval),
  c(y[x >= t.testval], 0, 0),
  col = "blue",
  density = 10,
  angle = -45
)
```



f(t)
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3
t

Do Sales Revenues *Affect* CEO Salaries?

1. Using `ceosal1`, regress the log of `salary` on log of `sales`, controlling for `roe` and firm industry group (industry, finance, consumer product, or utility)
2. Look out for perfect multicollinearity and leave on group - industry - out!).
3. Call the output `salary.lm1` and print a `summary()` of the results.

```
salary.lm1 <- lm(salary ~  log(sales) + roe + finance + consprod + utility, data = ceosal1)
summary(salary.lm1)
```

```
##
## Call:
## lm(formula = salary ~ log(sales) + roe + finance + consprod +
##     utility, data = ceosal1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1335.2  -403.6  -138.9    65.5 13278.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1196.373    846.488  -1.413  0.15909
## log(sales)    262.180     91.915   2.852  0.00479 **
## roe             7.976     12.337   0.647  0.51865
## finance       233.465    255.367   0.914  0.36168
## consprod      571.984    243.221   2.352  0.01965 *
## utility      -285.817    284.725  -1.004  0.31665
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1319 on 203 degrees of freedom
## Multiple R-squared:  0.09819,    Adjusted R-squared:  0.07598
## F-statistic: 4.421 on 5 and 203 DF,  p-value: 0.000764
```

Critical Values and p-Values

- What is the t-critical value of a test for whether the salaries of CEOs in finance firms differ from those of the baseline group (indistrial firms) at the 0.05 level?
  a. -1.652
  b.

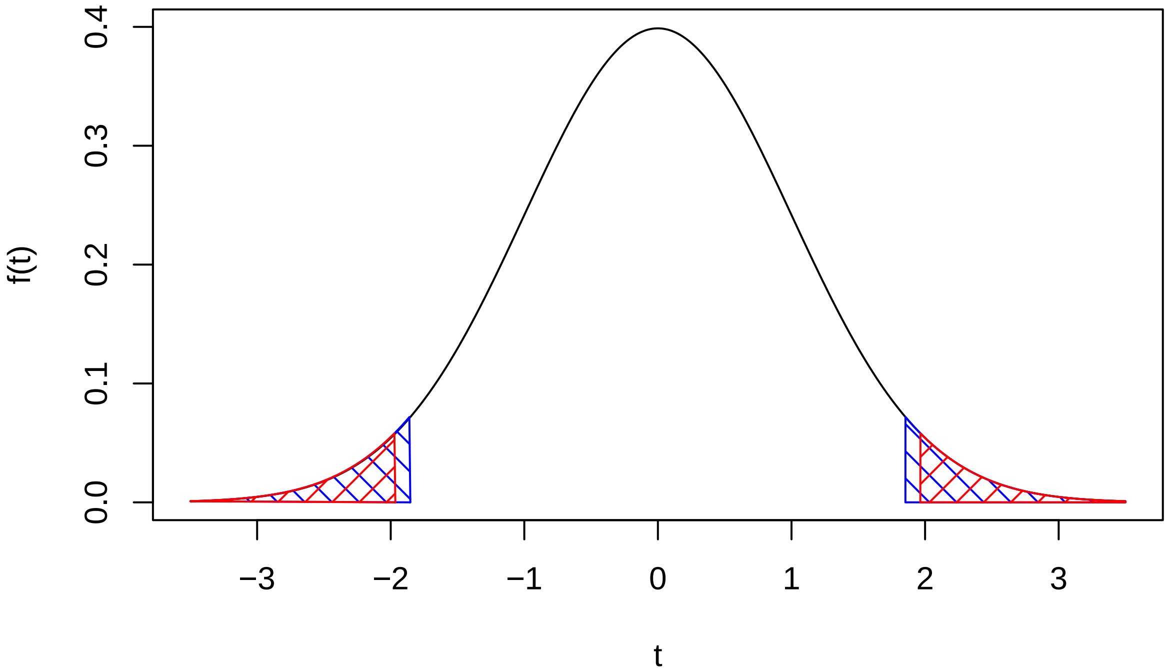
$pm1.652$

c. $pm1.972$
d. -1.972

- What is the p-value?
    a. 0.914
    b. 0.362
    c. 0.05
    d. 0.025

Critical Values and P-Values in R

```
# Define values for x and y axes, and the critical and test values.
x <- seq(-3.5,3.5,length=1000)
y <- dt(x,wage.lm6$df.residual)
t.critval <- qt(0.025, wage.lm6$df.residual)
t.testval <- summary(wage.lm6)$coefficients['exper', 't value']
# Plot the t distribution with n-k-1 degrees of freedom and sensibly-labeled axes.
plot(x, y, type="l", ylab = "f(t)", xlab = "t")
# Add the polygons for the p-value.
polygon(c(x[x>=abs(t.testval)], max(x), abs(t.testval)), c(y[x>=abs(t.testval)], 0, 0), col="blue", density = 10, angle = -45)
polygon(c(min(x), x[x<=-abs(t.testval)], -abs(t.testval)), c(y[x<=-abs(t.testval)], 0, 0), col="blue", density = 10, angle = -45)
# Add the polygons for BOTH alpha = 0.05 t-critical values.
polygon(c(x[x>=-t.critval], max(x), -t.critval), c(y[x>=-t.critval], 0, 0), col="red", density = 10)
polygon(c(min(x), x[x<=t.critval], t.critval), c(y[x<=t.critval], 0, 0), col="red", density = 10)
```


0.0
0.1
0.2
0.3
0.4
f(t)
−3
−2
−1
0
1
2
3
t

Confidence Intervals

The 95% confidence interval for $\hat{\beta}_j$ solves

$$P(\hat{\beta}_j - t^c_{0.025} \cdot s_{\hat{\beta}_j} \leq \mu \leq \hat{\beta}_j + t^c_{0.025} \cdot s_{\hat{\beta}_j}) = 0.95$$

- The bounds of the confidence interval for
  hat
  $beta_j$ are:
  a. Random.
  b. Centered around the true value of β
  c. Centered around zero.
  d. Centered around the null-hypothesized value of
     $beta_j$, $\beta_0.

Confidence Intervals

1. Using the `wage1` data, regress wage on education, experience, experience squared, tenure, and occupation (profesional services, professional occupations, clerical occupations, and service occupations).
2. Name the result `wage.lm7` and summarize the results using `summary()`.
3. Calculate 95-percent confidence intervals for the coefficients using the `confint()` function.
4. Calculate 99-percent confidence intervals for the coefficients.

```r
wage.lm7 <- lm(wage ~ educ + exper + I(exper^2) + tenure  + clerocc + servocc, data = wage1)
summary(wage.lm7)
```

```
##
## Call:
## lm(formula = wage ~ educ + exper + I(exper^2) + tenure + clerocc +
##     servocc, data = wage1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9721 -1.6092 -0.5683  1.1482 13.9203
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.2746819  0.7394668  -3.076  0.00221 **
## educ         0.5187697  0.0501466  10.345  < 2e-16 ***
## exper        0.1861238  0.0355888   5.230 2.46e-07 ***
```

```
## I(exper^2)  -0.0038585  0.0007715  -5.002 7.79e-07 ***
## tenure       0.1511597  0.0207354   7.290 1.16e-12 ***
## clerocc     -1.4282768  0.3506907  -4.073 5.37e-05 ***
## servocc     -1.5606360  0.3882824  -4.019 6.70e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.935 on 519 degrees of freedom
## Multiple R-squared:  0.3758, Adjusted R-squared:  0.3686
## F-statistic: 52.08 on 6 and 519 DF,  p-value: < 2.2e-16
```

```
confint(wage.lm7)
```

```
##                    2.5 %      97.5 %
## (Intercept) -3.727398020 -0.82196580
## educ         0.420254327  0.61728505
## exper        0.116208015  0.25603966
## I(exper^2)  -0.005374097 -0.00234298
## tenure       0.110424031  0.19189540
## clerocc     -2.117224682 -0.73932896
## servocc     -2.323434369 -0.79783762
```

```
confint(wage.lm7, level = 0.99)
```

```
##                    0.5 %       99.5 %
## (Intercept) -4.186451616 -0.362912200
## educ         0.389123792  0.648415584
## exper        0.094114841  0.278132839
## I(exper^2)  -0.005853009 -0.001864068
## tenure       0.097551686  0.204767746
## clerocc     -2.334929979 -0.521623666
## servocc     -2.564476190 -0.556795794
```