IDS 702: Module 1.4

Hypothesis tests, confidence intervals, and predictions

Dr. Olanrewaju Michael Akande

Recap: model for baseline salary

```
regwage <- lm(bsal~ sex + senior + age + educ + exper, data= wages)
summary(regwage)
```

```
##
## Call:
## lm(formula = bsal ~ sex + senior + age + educ + exper, data = wages)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1217.36  -342.83   -55.61   297.10  1575.53
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6277.8934   652.2713   9.625 2.36e-15
## sexFemale   -767.9127   128.9700  -5.954 5.39e-08
## senior       -22.5823     5.2957  -4.264 5.08e-05
## age            0.6310     0.7207   0.876 0.383692
## educ          92.3060    24.8635   3.713 0.000361
## exper          0.5006     1.0553   0.474 0.636388
##
## Residual standard error: 508.1 on 87 degrees of freedom
## Multiple R-squared:  0.5152,    Adjusted R-squared:  0.4873
## F-statistic: 18.49 on 5 and 87 DF,  p-value: 1.811e-12
```

Recap: model for baseline salary with centered predictors

```
regwagec <- lm(bsal~ sex + seniorc + agec + educc + experc, data= wages)
summary(regwagec)
```

```
##
## Call:
## lm(formula = bsal ~ sex + seniorc + agec + educc + experc, data = wages)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1217.36  -342.83   -55.61   297.10  1575.53
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5924.0072    99.6588  59.443  < 2e-16
## sexFemale   -767.9127   128.9700  -5.954 5.39e-08
## seniorc      -22.5823     5.2957  -4.264 5.08e-05
## agec           0.6310     0.7207   0.876 0.383692
## educc         92.3060    24.8635   3.713 0.000361
## experc         0.5006     1.0553   0.474 0.636388
##
## Residual standard error: 508.1 on 87 degrees of freedom
## Multiple R-squared:  0.5152,    Adjusted R-squared:  0.4873
## F-statistic: 18.49 on 5 and 87 DF,  p-value: 1.811e-12
```

Hypothesis tests for coefficients

- The reported t-values and p-values in R are used to test whether a particular coefficient equals 0, GIVEN that all other variables are in the model.
- Specifically, for coefficient β_j,

$$\mathcal{H}_0 : \beta_j = 0; \quad \text{vs.} \quad \mathcal{H}_1 : \beta_j \neq 0$$

- Examples:
    - The test for whether the coefficient of education equals zero has p-value $\approx .0004$. Hence, reject the null hypothesis; it appears that education is a useful predictor of baseline salary when all the other predictors are in the model.
    - The test for whether the coefficient of experience equals zero has p-value $\approx .6364$. Hence, we cannot reject the null hypothesis; it appears that experience is not a particularly useful predictor of baseline salary when all other predictors are in the model.

IDS 702

Hypothesis tests for coefficients

- Fortunately, R (and pretty much all statistical software) computes both the t-values and p-values for us automatically.

How do we calculate the t-values and p-values manually?

- The t-values (test statistics) have the usual form:

$$T = \frac{\text{Point Estimate} - \text{Null Value}}{SE} = \frac{\hat{\beta}_j - 0}{\sqrt{\left[s_e^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1}\right]_{jj}}}$$

- For p-value, use area under a t-distribution with $n - (p+1)$ degrees of freedom, where p is the number of predictors (minus the intercept) in the model.
- In this problem, the degrees of freedom equal $93 - 6 = 87$.

You should know how to compute the p-values directly using the pt function in R (from the summer review materials).

CIs for regression coefficients

- A 95% CI for the coefficients is obtained in the usual way. Recall the general form for two-sided CIs from the online review material:

  $$CI = pe \pm SE \times C_\alpha$$

  where pe is the point estimate, and C_α is a multiplier (critical value) that depends on the confidence level.

- For MLR, we have

  $$CI = \hat{\beta}_j \pm SE \times C_\alpha = \hat{\beta}_j \pm C_\alpha \times \sqrt{\left[s_e^2 \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}\right]_{jj}},$$

  and the multiplier is obtained from the t-distribution with $n - (p+1)$ degrees of freedom.

- Example: A 95% "two-sided" CI for the population regression coefficient of age equals:
  $(0.63 - 1.988 \times 0.72, 0.63 + 1.988 \times 0.72) = (-0.80, 2.06)$.

  Find the multiplier (1.988) in R by using the command `qt(0.975,df=87)`.

CIs for regression coefficients

- We can compute "two-sided" confidence intervals very easily in R.

```
confint(regwage,level = 0.95)
```

```
##                     2.5 %      97.5 %
## (Intercept)  4981.4335106 7574.353262
## sexFemale   -1024.2545333 -511.570844
## senior        -33.1081429  -12.056463
## age            -0.8014178    2.063338
## educ           42.8870441  141.725002
## exper          -1.5968086    2.598088
```

- For employees with the same age, seniority, education, and experience, we expect the average starting salary for female employees to be between 511 and 1024 dollars less than the average starting salary for male employees.

- More succinctly, for employees with the same age, seniority, education, and experience, we expect female employees’ average starting salary to be around $767 less than male employees' average salary, with 95% CI in dollars = (-1024, -512).

Notes about tests and CIs

- When sample size is large enough, you will probably reject the null hypothesis $H_0 : \beta_j = 0$.
  - This is because as n increases, the SE will decrease, most likely blowing up the test statistic T.
  - Thus, you should consider practical significance, not just statistical significance.
- When sample size is small, there simply may not be enough evidence to reject null hypothesis $H_0 : \beta_j = 0$.
  - When you fail to reject the null hypothesis, don't be too hasty to say that predictor has no linear association with the outcome.
  - There may be an association, just not strong enough to detect with this sample (or perhaps a nonlinear one).
  - It may also be that the association is not significant because you are already controlling for other characteristics.

Predictions

- Making predictions using the fitted model is straightforward.
- For example, suppose we want to prediction baseline salary for a 25 year old woman with 12 years of education, 10 months of seniority, and two years of experience. We can simply plugin these values into the estimated model (without centering):

$$\hat{y}_i = 6277.9 - 767.9(1) - 22.6(10) + 0.63(300) + 92.3(12) + 0.50(24) = 6592.6$$

- Easier to do in R using the predict command. We can also get confidence and prediction intervals using the same command.

```
newdata <- data.frame(sex="Female",senior=10,age=25*12,
                      educ=12,exper=2*12)
pred1 <- predict(regwage,newdata,interval="confidence"); pred1
```

```
##          fit      lwr      upr
## 1 6593.133 5775.546 7410.721
```

```
pred2 <- predict(regwage,newdata,interval="prediction"); pred2
```

```
##          fit      lwr      upr
## 1 6593.133 5293.781 7892.486
```

IDS 702

PREDICTIONS

- Or using the model with centered predictors,

```
newdatac <- data.frame(sex="Female",
                       seniorc=(10 - mean(wages$senior)),
                       agec=(25*12 - mean(wages$age)),
                       educc=(12 - mean(wages$educ)),
                       experc=(2*12 - mean(wages$exper)))
predc1 <- predict(regwagec,newdatac,interval="confidence"); predc1
```

```
##        fit      lwr      upr
## 1 6593.133 5775.546 7410.721
```

```
predc2 <- predict(regwagec,newdatac,interval="prediction"); predc2
```

```
##        fit      lwr      upr
## 1 6593.133 5293.781 7892.486
```

- Notice that this is the same as what we had on the previous slide. Why is this so?

What's next?

Move on to the readings for the next module!