Lab: Significance & p-values

1. Permutation test of independence

Replicate the permutation tests from slide 18 (see slide 10).

2. Parametric Inference

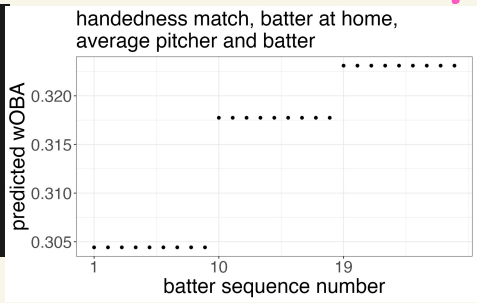
* Recall these models from the time through the order (TTO) analysis:

$$(*) \quad y_i = \beta_1 + \beta_2 \cdot \mathbb{1}\{t_i \geq 2\text{TTO}\} + \beta_3 \cdot \mathbb{1}\{t_i \geq 3\text{TTO}\} + \beta_{BQ} \cdot BQ_i + \beta_{PQ} \cdot PQ_i + \beta_{hand} \cdot hand_i + \beta_{home} \cdot home_i + \varepsilon_i$$

```
> m2 = lm(EVENT_WOBA_19 ~ 1 + as.numeric(ORDER_CT>=2) + as.numeric(ORDER_CT>=3) +
+                         HAND_MATCH + BAT_HOME_IND + WOBA_FINAL_BAT_19 + WOBA_FINAL_PIT_19,
+         data=df0)
> m2

Call:
lm(formula = EVENT_WOBA_19 ~ 1 + as.numeric(ORDER_CT >= 2) +
    as.numeric(ORDER_CT >= 3) + HAND_MATCH + BAT_HOME_IND + WOBA_FINAL_BAT_19 +
    WOBA_FINAL_PIT_19, data = df0)

Coefficients:
       (Intercept)  as.numeric(ORDER_CT >= 2)  as.numeric(ORDER_CT >= 3)        HAND_MATCH
         -0.299509                   0.013320                   0.005357        -0.016306
     BAT_HOME_IND          WOBA_FINAL_BAT_19          WOBA_FINAL_PIT_19
         0.009988                   0.969370                   0.962359
```

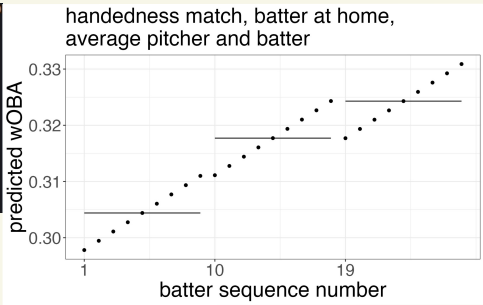


$$(**) \quad y_i = \beta_0 + \beta_1 \cdot t_i + \beta_2 \cdot \mathbb{1}\{t_i \geq 2\text{TTO}\} + \beta_3 \cdot \mathbb{1}\{t_i \geq 3\text{TTO}\} + \beta_{BQ} \cdot BQ_i + \beta_{PQ} \cdot PQ_i + \beta_{hand} \cdot hand_i + \beta_{home} \cdot home_i + \varepsilon_i$$

```
> m5 = lm(EVENT_WOBA_19 ~ 1 + as.numeric(ORDER_CT>=2) + as.numeric(ORDER_CT>=3) + BATTER_SEQ_NUM + HAND_MATCH + BAT_HO
ME_IND + WOBA_FINAL_BAT_19 + WOBA_FINAL_PIT_19, data=df0)
> m5

Call:
lm(formula = EVENT_WOBA_19 ~ 1 + as.numeric(ORDER_CT >= 2) +
    as.numeric(ORDER_CT >= 3) + BATTER_SEQ_NUM + HAND_MATCH +
    BAT_HOME_IND + WOBA_FINAL_BAT_19 + WOBA_FINAL_PIT_19, data = df0)

Coefficients:
       (Intercept)  as.numeric(ORDER_CT >= 2)  as.numeric(ORDER_CT >= 3)      BATTER_SEQ_NUM
         -0.316611                  -0.001528                  -0.008262         0.001649
        HAND_MATCH              BAT_HOME_IND          WOBA_FINAL_BAT_19  WOBA_FINAL_PIT_19
         -0.016837                   0.009994                   0.999090         0.962273
```



Model 1: $\hat{\beta}_2 = 0.013$, $\hat{\beta}_3 = 0.0054$

Model 2: $\hat{\beta}_2 = -0.0015$, $\hat{\beta}_3 = -0.0083$, $\hat{\beta}_1 = 0.0016$

These point estimates are a single "best guess" of the parameter value according to the model.

The thing is, you can fit a model to any dataset and get estimated coefficients.
You ought to wonder: what if the model is shit?
In our case, we've done enough validating of our model (via plotting) to be convinced it's not completely shit. And, "shit" is vague.

A more refined question is: are the values of these estimated coefficients due to a Real trend in the data or due to noise and random chance?
Since we're interested in whether these coefficients are zero or not, we ask: is there enough evidence in the data to include that the coefficient is highly likely nonzero (i.e., significantly nonzero)?

These questions underly parametric inference: we'd like to infer whether a parameter in a model is nonzero.

Null hypothesis $H_0: \beta = 0$

Alt hypothesis $H_1: \beta \neq 0$

Test statistic $t = \hat{\beta} / SE(\hat{\beta})$

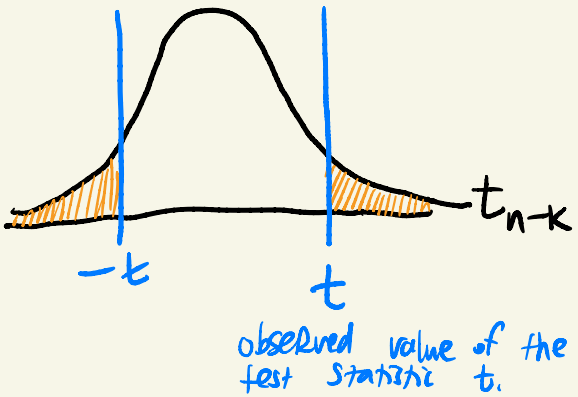
$SE(\hat{\beta})$ is the standard error of $\hat{\beta}$, which is an estimate of the standard deviation of $\hat{\beta}$. Since $\hat{\beta}$ is computed from the data (X, y) and y is generated from model $y = X^T\beta + \varepsilon$, y is a random variable and hence $\hat{\beta}$ is a random variable and hence $\hat{\beta}$ has a standard deviation which we estimate, $SE(\hat{\beta}) = \sqrt{Var(\hat{\beta})}$.

optional math HW: derive $SE(\hat{\beta})$ in multivariable linear regression.

Theorem: Suppose $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.
Suppose H_0 is true, so $\beta = 0$.
Then the test statistic t has distribution $t \sim t_{n-k}$, the t distribution with $n-k$ degrees of freedom (n = # datapoints, k = # predictors).

optional math HW: prove this

Under our assumptions, the test statistic t has a t_{n-k} distribution, which is nearly Normal when n is large,



The p-value is the orange area to the right of t and left of $-t$ under the t_{n-k} density curve. It is, assuming the null H_0 is true, the probability of observing a test statistic as extreme as the one we actually observed, where the randomness here is over the sampling distribution (generating y from x).

Theorem: the p-value $P = \mathbb{P}_{H_0}(|t_{n-k}| > t)$ has distribution $p \sim \text{Unif}[0,1]$.

optional math HW: prove this

If the null H_0 is true, we are unlikely to observe a $|t|$ in the far tails of the distribution, and so we are unlikely to observe a p-value that is small.

So, a large $|t|$ value and small p-value is associated with less evidence of β being zero.

The convention is to set a significance level, typically $\alpha = 0.05$, and then

Reject H_0 if $p < \alpha$ else do not Reject.

But this α is arbitrary and 0.05 is still a pretty big number.

p-values should be treated as a spectrum and not a strict cutoff.

R automatically runs this t-test for every coefficient. You can do this by `summary(m)` where `m ← lm(•)` is the object that stores the linear model:

```
> summary(m2)

Call:
lm(formula = EVENT_WOBA_19 ~ 1 + as.numeric(ORDER_CT >= 2) +
    as.numeric(ORDER_CT >= 3) + HAND_MATCH + BAT_HOME_IND + WOBA_FINAL_BAT_19 +
    WOBA_FINAL_PIT_19, data = df0)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0089 -0.3351 -0.2762  0.3995  1.8396

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -0.299509   0.010110 -29.626   < 2e-16 ***
as.numeric(ORDER_CT >= 2)  0.013320   0.002481   5.370 7.89e-08 ***
as.numeric(ORDER_CT >= 3)  0.005357   0.002922   1.833   0.0668 .
HAND_MATCH              -0.016306   0.002206  -7.392 1.45e-13 ***
BAT_HOME_IND             0.009988   0.002183   4.575 4.77e-06 ***
WOBA_FINAL_BAT_19        0.969370   0.017123  56.613  < 2e-16 ***
WOBA_FINAL_PIT_19        0.962359   0.026573  36.215  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5053 on 214379 degrees of freedom
Multiple R-squared:  0.0222,    Adjusted R-squared:  0.02218
F-statistic: 811.3 on 6 and 214379 DF,  p-value: < 2.2e-16
```

t statistic for each regression coefficient

```
> summary(m5)

Call:
lm(formula = EVENT_WOBA_19 ~ 1 + as.numeric(ORDER_CT >= 2) +
    as.numeric(ORDER_CT >= 3) + BATTER_SEQ_NUM + HAND_MATCH +
    BAT_HOME_IND + WOBA_FINAL_BAT_19 + WOBA_FINAL_PIT_19, data = df0)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0133 -0.3348 -0.2761  0.3995  1.8391

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -0.3166112  0.0113189 -27.972  < 2e-16 ***
as.numeric(ORDER_CT >= 2) -0.0015278  0.0050680  -0.301 0.763064
as.numeric(ORDER_CT >= 3) -0.0082619  0.0049974  -1.653 0.098285 .
BATTER_SEQ_NUM           0.0016491  0.0004909   3.360 0.000781 ***
HAND_MATCH              -0.0168367  0.0022114  -7.613 2.68e-14 ***
BAT_HOME_IND             0.0099937  0.0021831   4.578 4.70e-06 ***
WOBA_FINAL_BAT_19        0.9990902  0.0192728  51.839  < 2e-16 ***
WOBA_FINAL_PIT_19        0.9622731  0.0265726  36.213  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5053 on 214378 degrees of freedom
Multiple R-squared:  0.02225,   Adjusted R-squared:  0.02222
F-statistic:   697 on 7 and 214378 DF,  p-value: < 2.2e-16
```

p-value for each regression coefficient

- Fit the two models (*) and (**).
  In the first model, is pitcher decline from one time through the order to the next significant? Explain.
  Think carefully about how you'll explain it.

* I don't really like parametric inference.
  The theorems are based on strict mathematical assumptions like $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$
  that are often unreasonable.
  There are other better ways to do inference,
  which we will discuss:
  Bootstrapping & Bayesian statistics.
  But it is still good that y'all got an overview.