Handout 7

STAT3530 Applied Linear Models

The omnibus F test compares the full model against an intercept-only model. It turns out this is one case of something much more general: a test comparing *any* two nested models.

Comparing nested models

1. A model $\mathcal{M}$ is a claim about where the mean of Y lives.

   a. Written as a subspace condition, $\mathcal{M}$ says

   $$\mathbb{E}_{\mathcal{M}}[Y] \in _____ \subseteq \mathbb{R}^n$$

   b. For the general linear model $\mathbb{E}[Y] = X\beta$, that subspace is $V = _____$.

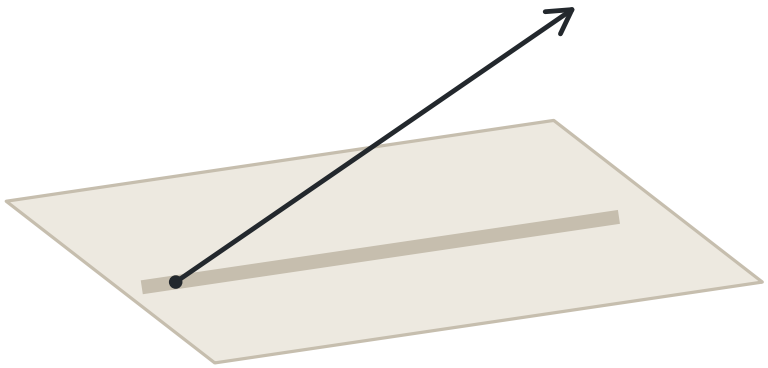
   c. So $\mathcal{M}_0$ is **nested** in $\mathcal{M}_1$ exactly when _____.

   d. Check the omnibus case: the intercept-only model is nested in the full model because

   $$_____ \subseteq _____$$

The geometry of the test

2. The figure shows the two subspaces for a nested comparison: the shaded plane and the heavy line inside it. Label V_0, V_1 and the vector Y. Then **draw in**: the two projections H_0Y and H_1Y, the residual vector for each model, the segment joining the two projections, and right-angle marks wherever two things are perpendicular.



3. Your figure contains a right triangle. Write down what Pythagoras gives you, in terms of SSE_0, SSE_1, and the segment joining the two projections.

$$SSE_0 =$$

   a. The piece you just added to SSE_1 is called the **extra sum of squares**. In words, it measures

   b. It is never negative. Which feature of the picture guarantees that?

4. Each subspace has a dimension, and each sum of squares has degrees of freedom.

   a. Fill in the table for a comparison of two nested models.

	subspace	its dimension	residual df
reduced model $\mathcal{M}_0$	V_0	_____	_____
full model $\mathcal{M}_1$	V_1	_____	_____

   b. The statistic compares the relative change in fit to the relative change in degrees of freedom. Fill in both pieces, and the null distribution.

$$F = \underbrace{\text{_____}}_{\text{change in fit}} \Big/ \underbrace{\text{_____}}_{\text{change in df}} \sim F_{\text{_____},\text{_____}}$$

   c. Check your answer against the omnibus test by setting $V_0 = \text{span}\{\mathbf{1}\}$ and $V_1 = \text{col}(X)$. Do you recover the statistic from Handout 6b?

The extra sum of squares test in practice

Back to the car seat data: 38 drivers, eight body measurements, and the question of where they position the seat.

```
round(summary(seat)$coefficients, 3)
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  436.432    166.572   2.620    0.014
Age            0.776      0.570   1.360    0.184
Weight         0.026      0.331   0.080    0.937
HtShoes       -2.692      9.753  -0.276    0.784
Ht             0.601     10.130   0.059    0.953
Seated         0.534      3.762   0.142    0.888
Arm           -1.328      3.900  -0.341    0.736
Thigh         -1.143      2.660  -0.430    0.671
Leg           -6.439      4.714  -1.366    0.182
```

5. Recall from Handout 6b that the omnibus test is significant here while not one individual t test is. The extra sum of squares test lets us ask about predictors in *groups*. For each question below, write the null hypothesis in terms of coefficients, give the reduced model formula, and give both degrees of freedom. Note $n = 38$ and $\dim V_1 = 9$.

   a. Do the body measurements matter at all, once age and weight are accounted for?

   H_0 :

   reduced model: `lm(hipcenter ~` _____ `, data = seatpos)`

   $\dim V_0 =$ _____, so the test has _____ and _____ degrees of freedom.

   b. `HtShoes` and `Ht` are two ways of measuring the same person. Do we need both, given everything else in the model?

   H_0 :

   reduced model: `lm(hipcenter ~` _____ `, data = seatpos)`

   $\dim V_0 =$ _____, so the test has _____ and _____ degrees of freedom.

   c. Does `Age` matter, given everything else?

   H_0 :

   reduced model: `lm(hipcenter ~` _____ `, data = seatpos)`

   $\dim V_0 =$ _____, so the test has _____ and _____ degrees of freedom.

Here are the three tests, once you have built the right reduced model for each.

```
     (a)    (b)    (c)
F 3.2402 0.3433 1.8499
p 0.0146 0.7123 0.1843
```

6. Using the output above:

   a. Reconcile the result of test (a) with the fact that no individual t test was significant.

   b. Compare the statistic from (c) to the t statistic for `Age` in the summary on the previous page. What is the relationship, and why?

Contrasts

A **contrast** is a linear function of the parameters, written $c^T\beta$ for a fixed vector c.

7. A contrast hypothesis $H_0 : c^T\beta = 0$ is a linear constraint, so it defines a reduced subspace

$$V_0 = \{X\beta : _____\} \qquad \text{with} \qquad \dim V_0 = _____$$

   and the test has _____ and _____ degrees of freedom.

Activity

The seat position model has coefficients in this order:

$$\beta_0 \;\; \beta_{\texttt{Age}} \;\; \beta_{\texttt{Weight}} \;\; \beta_{\texttt{HtShoes}} \;\; \beta_{\texttt{Ht}} \;\; \beta_{\texttt{Seated}} \;\; \beta_{\texttt{Arm}} \;\; \beta_{\texttt{Thigh}} \;\; \beta_{\texttt{Leg}}$$

8. For each question, write the hypothesis in terms of coefficients and then as a vector c.

	question	H_0 in coefficients	c^T
a.	Do shoes change the height effect?	_____	_____
b.	Is the leg effect the same as the thigh effect?	_____	_____
c.	Does `Age` have no effect?	_____	_____
d.	Do arm and thigh length together offset leg length?	_____	_____

9. Two of the four above are tests you already know how to run without contrasts. Which, and why?

Here is the test for (a).

```
car::linearHypothesis(seat, 'HtShoes = Ht')[['Pr(>F)']][2]
```

```
[1] 0.8684083
```

10. State the conclusion in context.

Beyond contrasts

Not every question about coefficients is a linear constraint. Suppose we want to know which predictor has the largest effect. On standardized predictors the estimand is

$$\theta(\beta) = \operatorname{argmax}_{j>0} |\beta_j|$$

11. The hypothesis $H_0 : \theta(\beta) = 2$ says that β_2 beats every other coefficient in magnitude. Write that as a set of β, then check whether the set is closed under scaling and under addition — try $\beta = (0, 5, 1)$ and $\beta' = (0, -5, 1)$ for the second. What kind of object is it, and why does that rule out an F test?

$$\{\beta : \qquad\qquad\qquad\qquad\qquad\qquad \}$$

Inference by simulation

When the theory runs out, build the sampling distribution by computation instead.

12. The recipe:
    a. simulate many datasets _____
    b. compute _____ for each one
    c. build up _____

13. Both schemes keep the fitted part of each observation and replace the error. They differ in where the new errors come from.

	new errors drawn from	what it assumes
randomization	_____	_____
parametric bootstrap	_____	_____

14. Write the simulated response for observation i on replicate b under each scheme.

    randomization: $y_i^{(b)} =$ bootstrap: $y_i^{(b)} =$

Activity: simulating by hand

We will be the computer. Each of you holds a few rows of the cheddar data; your card lists the fitted value $x_i^T \hat{\beta}$ for each row. One deck of cards carries the residuals from the fitted model, and a second carries fresh errors drawn from $N(0, \hat{\sigma}^2)$.

Each round, you will be dealt an error for each of your rows, add it to that row's fitted value to get a simulated taste score $y_i^{(b)}$, and whoever holds the largest score in the room stands up. Record both the winning score $T = \max_i y_i^{(b)}$ and who held it.

Observed value in the real data: $T =$ _____

Randomization rounds. We shuffle the residual deck and deal it out, so every residual is used exactly once.

round	1	2	3	4	5
your largest $y^{(b)}$					
class max T					
who held it					

Parametric bootstrap rounds. Now set the residuals aside and deal from the second deck of fresh errors. Cards go back after each draw.

round	1	2	3	4	5
your largest $y^{(b)}$					
class max T					
who held it					

Sketch the two distributions of T on the same axis and mark the observed value. Then count how many different people won across the ten rounds. The fitted values never moved — so what decided the winner?

Inverse prediction

Average SAT total, educational expenditure, and the share of test takers in 50 states.

$$\text{SAT}_i = \beta_0 + \beta_1 \,\text{spending}_i + \beta_2 \log(\text{takers}_i) + \epsilon_i$$

```
sat.fit <- lm(total ~ expend + log(takers), data = sat)

round(coef(sat.fit), 2)
```

```
(Intercept)      expend log(takers)
    1147.11       11.13      -78.20
```

15. The fit answers "given spending, what score?". Run it backwards: what level of spending reaches a target of $y_0 = 975$ at 30% participation?

    a. Solve the fitted equation for spending.

    $$\hat{x} =$$

    b. Evaluate it using the coefficients above.

    c. Why can you not read a standard error for $\hat{x}$ off the `summary()` output?

$\hat{x}$ is a *nonlinear* function of $\hat{\beta}$, so simulate instead.

```
x.star <- function(f, y0, tk) unname((y0 - coef(f)[1] - coef(f)[3] * log(tk)) / coef(f)[2])
refit  <- function(y) lm(y ~ expend + log(takers), data = sat)

set.seed(5)
x.null <- replicate(2000, {
  y <- fitted(sat.fit) + sample(resid(sat.fit))
  refit(y) |> x.star(y0 = 975, tk = 30)
})

quantile(x.null, c(0.025, 0.975))     # bootstrap percentile interval
```

```
    2.5%    97.5%
7.571457 11.308079
```

16. Using the interval above:

    a. Interpret it in context, in one sentence, on the dollar scale.

    b. How far does the interval reach below the point estimate, and how far above? What does that asymmetry tell you?

Two ways to build an interval

The symmetry of a usual parametric confidence interval comes from the normal sampling distribution. Simulation gives two ways to get a symmetric interval without assuming one.

Substitute a bootstrap standard error. Replace $\widehat{\text{SE}}$ in $\hat{\theta} \pm 1.96\,\widehat{\text{SE}}$ by the standard deviation of the simulated values.

```
x.hat <- x.star(sat.fit, y0 = 975, tk = 30)

x.hat + c(-1, 1) * 1.96 * sd(x.null)
```

```
[1]  1.768779 15.101143
```

17. Using the output above:

    a. Compare this to the percentile interval in 16. What has gone wrong?

    b. This method is fine when the simulated distribution is symmetric, which sometimes only happens after a transformation. What would you check before trusting it?

Take the half-width from the bootstrap. Keep the symmetry, but let the data set the width: find the distance c that captures 95% of the simulated values around $\hat{\theta}$.

$$\hat{\theta} \pm c \qquad \text{where} \qquad P(|\hat{\theta}^{(b)} - \hat{\theta}| \le c) = _____$$

```
cc <- unname(quantile(abs(x.null - x.hat), 0.95))

c(x.hat - cc, x.hat + cc)
```

```
[1]  6.226979 10.642943
```

18. Using the output above:

    a. Why does folding the simulated values at $\hat{\theta}$ and taking a quantile of $|\hat{\theta}^{(b)} - \hat{\theta}|$ give a symmetric interval?

    b. Compare all three intervals. The symmetric one pushes its lower limit well below anything the simulation produced. Explain why, and say which you would report here.