

Lecture 11: Multiple Regression III

- Full and Reduced Model Comparisons – Chapter 4.4 (cont)
 - Testing a group of variables
- Prediction in multiple regressions – Chapter 4.5
 - Confidence intervals
 - For the conditional mean, $\mu_{Y|x_1, \dots, x_K}$.
 - ~ Called “Mean Confidence Interval” in JMP.
 - For the value of an individual Y given x_1, \dots, x_K .
 - ~ Called “Individual Confidence Interval” in JMP.

Testing a Group of Variables

After controlling for other variables in a model

- In Lecture 10 we discussed the test that one coefficient is zero, after controlling for all others in the model.
 - This test is given directly in the JMP's Effect Test table.
- It's also possible to test for a group of coefficients, after controlling for all others in the model.
- The procedure is similar (with a few differences) but it is not given directly in JMP.
 - Instead, you need to construct the test statistics from two separate JMP tables. Here's the general theory. Then we'll use this procedure in our Cars 2004 example.

General Theory

- You want to test that the (linear) effect of a subset of the coefficients is 0; after controlling for the (linear) effect of all the others in the model.
- For notational convenience, we'll label the tested coefficients as $\beta_{L+1}, \dots, \beta_K$. But the procedure is similar if applied to any other subset of coefficients.
- In formal terms, the null and alternative hypotheses are
$$H_0 : \beta_{L+1} = \dots = \beta_K = 0 \text{ vs } H_a : H_0 \text{ is false.}$$
- To test the null hypothesis begin by examining the SSE terms in both
 - a. The ANOVA table for the full model having all the coefficients $\beta_1, \dots, \beta_L, \dots, \beta_K$ (Call this **SSE_F**); and
 - b. The ANOVA table for the “reduced” model that has only the effects corresponding to β_1, \dots, β_L . (Call this **SSE_R**.)

- Note that $\mathbf{SSE}_R \geq \mathbf{SSE}_F$. { Why? }
- Then define the statistic

$$F = \frac{(\mathbf{SSE}_R - \mathbf{SSE}_F) / (K - L)}{\mathbf{SSE}_F / (n - K - 1)} .$$

- Note that the term $\mathbf{SSE}_F / (n - K - 1)$ in the denominator of this F-statistic. This is the MSE from the full model.
- And note the term $(K - L)$ that appears inside its numerator. This is the DF of the difference $\mathbf{SSE}_R - \mathbf{SSE}_F$, since there are $K-L$ extra parameters, $\beta_{K+1}, \dots, \beta_L$.
- This F-statistic should be compared with critical values (say, $\mathbf{F}(\alpha; K - L, n - K - 1)$). And then **Reject H_0** if $F > \mathbf{F}$.
- For the critical values use Tables B.3-5 with numerator DF = $K-L$ and denominator DF = $n-K-1$. Or use JMP as follows: Create new table; add rows; right-click on column heading, & create “formula” for F-quantile and type the desired values of $p = 1 - \alpha$ and the DF entries into formula boxes.

Example from Cars 2004

- We have looked at an analysis of MPG_City on the **4** predictor variables: **Wt(1000lb)**, **HP**, **Seating**, and **Length**.
- In the complete data set there are **3** more possible numerical predictor variables: **Displacement**, **Cylinders**, and **Width**.
- Let's test whether these **3** new variables, as a group, have a statistically significant effect, after controlling for the (linear) effect of the **4** original variables.
- The full model has $4 + 3 = 7$ possible effects. (So, here $L = 4$ and $K = 7$.)
- Here are the two ANOVA tables:

- Full Model (with $K = 7$ factors)

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	7	2297.8	328.2	120.4
Error	214	SSE_F = 583.3	2.73	Prob > F
C. Total	221	2881.1		<.0001

- Reduced Model (with $L = 4$ factors) [We've seen this table on p. 4 of Lecture 10.]

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	2291.2	572.8	210.7
Error	217	SSE_R = 589.9	2.72	Prob > F
C. Total	221	2881.1		<.0001

- The F-statistic is

$$F = \frac{(589.9 - 583.3) / (7 - 4)}{583.3 / 214} = \frac{6.6 / 3}{2.73} = 0.81.$$

- From Table B.5, $F(0.05; 3, 214) \approx 2.65$.

Conclusion of the Test

- The observed $F = \mathbf{0.81}$ is smaller than the critical value of $\mathbf{F(0.05; 3, 214) \approx 2.65}$.
- So we **Fail to Reject** H_0 *at level* $\alpha = 0.05$.
- You can find the exact P-value of the data {details in lecture} as $P = 1 - 0.51 = 0.49$
- We **conclude** that there is **no statistically significant** evidence *at level* $\alpha = 0.05$ that **one or more of the coefficients** of **Displacement, Cylinders, or Width**, is **not 0**, after controlling for the other four variables.

Confidence Intervals for Predictions

- In Lecture 9 we expressed interest in estimating the mean value of MPG_City for a car model with
 - **horsepower** – 225; **weight** – 4000 lbs;
 - **seating** – 5 adults; **length** – 180''

- From the linear regression having those four x-variables in the model and for our car of interest with

$x_1 = 225, x_2 = 4, x_3 = 5, x_4 = 180$ we found the estimate

$$\hat{Y} = 31.5 - .0154 \times 225 - 3.77 \times 4 + .337 \times 5 + .0175 \times 180 = 17.79.$$

The values in pink are the estimated coefficients from the Parameter Estimates table.

- Its possible to state Cis for this mean value.
- The general idea is just the same as in ordinary regression
 - The algebraic formula is much more complicated, and you need to rely on JMP to find the CI.*

Mean Confidence Intervals (from JMP)

- Use the ▼ → Save Columns option inside the Fit Model output.
 - Do this after creating a row containing the desired x-values.
 - This will directly give the “Mean Confidence Interval” with confidence 95%. (Look for this inside the data table.)
 - Do this You can also get the corresponding value for the SE, as “Std Error of Predicted”.
 - Use this SE along with tabled values of $t_{\alpha/2; DFE}$ to get 100(1- α)% CIs at other confidence levels.
(NOTE: Use DF for Error here.)

- Here is the JMP output

	HP	Wt	Seats	Lngh	Lower 95%	Upper 95%	SE of Pred
Design	225	4	5	180	17.46	18.13	0.169

- So, the 95% CI for the mean at the values 225,4,5,180 is (17.46, 18.13)
- Note that $t_{.025;217} = 1.97$ (from JMP).
- So, this same CI can also be found as

$$17.79 \pm 1.97 \times .169 = 17.79 \pm .33$$

Individual Confidence Intervals

- The mean confidence interval from the preceding pages was sought as a target for the production design of a new model of car.
- Now we want to predict the MPG of this particular car model. (*ie*, of a model having $x_1 = 225$, $x_2 = 4$, $x_3 = 5$, $x_4 = 180$.)
- For such a situation one would want an “individual prediction interval”
- This is centered on the same prediction as before, but the individual prediction SE is larger than the value used in the previous CI for the mean.

Individual Prediction CI in JMP

- The procedure is as before with the mean prediction CI, but now we need “Indiv Confidence Intervals” and their “Std Error”
- Here is the JMP output:

	HP	Wt	Seats	Lngh	Lower 95%	Upper 95%	SE of Pred
Design	225	4	5	180	14.53	21.06	1.657

- Note that the SE is much larger than for the mean CI previously, and the prediction CI is correspondingly much wider.

Sample Questions on Multiple Regression (Sections 4.1-4.5, only)

Questions Based on Dataset: CS Grades

Background: This data has core-course grades (~first 1.5 years) for CS students at Purdue U. (from about 1990), along with various scores from the entrance applications. The goal of the analysis is to see which of the indicators are most useful in predicting core-course performance, and to develop an equation for predicting such performance. The possible predictor variables are

hsm = overall average of high school math grades,
hss = overall average of high school science grades,
hse = overall average of high school English grades,
SATV = SAT Verbal score, SATM = SAT Math Score

1. Multiple regressions of GPA on the five predictors.

Here is the output from JMPIN:

Summary of Fit

RSquare	0.211
Root Mean Square Error	0.700
Mean of Response	4.635
Observations	224

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	?	28.64	5.73	?
Error	?	106.82	0.49	Prob > F
C. Total	?	135.46		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	2.327	0.399996	5.82	<.0001
hsm	0.1460	0.039261	3.72	0.0003
hss	0.03591	0.037798	0.95	0.3432
hse	0.05529	0.039569	1.40	0.1637
SATV	-0.000408	0.000592	-0.69	0.4915
SATM	0.000944	0.000686	1.38	0.1702

Effect Tests

Source	DF	Sum of Squares	F Ratio	Prob > F
hsm	1	6.7724	13.82	0.0003
hss	1	0.4421	0.90	0.3432
hse	1	0.9568	1.95	0.1637
SATV	1	0.2327	0.47	0.4915
SATM	1	0.9280	?	?

- a) Fill in the 4 missing entries in the ANOVA table.
- b) What is the least squares equation?
- c) Predict the GPA of a student with $hsm = 9$, $hss = 7$, $hse = 8$, $SATV = 460$, $SATM = 450$.
- d) Is the full model useful to predict the response variable? Set up the hypotheses and carry out the test.
- e) Report R^2 and interpret the value.
- f) Fill in the two missing entries in the Effect Test table.

2. t-tests and partial F-tests

a) Is hse a useful predictor, after controlling for the effect of all the other variables in the model?

State the hypotheses for this test, and give the P-value as well as the conclusion of the test at level $\alpha = 0.05$.

[You should understand how to use the output provided to produce the test using both the t-method and the F-method.]

b) Are the SATM and SATV scores useful if you do not have available the high school grades?

c) Are they useful in addition to the high school grades?

{For questions b} and c) you will need some of the preceding output, as well as some of the following.}

Output for model with $Y = \text{GPA}$ and Model variables: hsm, hss, and hse.

Summary of Fit

RSquare	0.2046
Root Mean Square Error	0.6998
Mean of Response	4.6352
Observations	224

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	27.71	9.24	18.86
Error	220	107.75	0.49	Prob > F
C. Total	223	135.46		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	2.590	0.2942	8.80	<.0001
hsm	0.169	0.0355	4.75	<.0001
hss	0.0343	0.0376	0.91	0.3619
hse	0.0451	0.0387	1.17	0.2451

Output for model with $Y = \text{GPA}$ and Model variables: SATM and SATV.

Summary of Fit

RSquare	0.0634
Root Mean Square Error	0.758
Mean of Response	4.64
Observations	224

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	8.58	4.29	7.48
Error	221	126.88	0.57	Prob > F
C. Total	223	135.46		0.0007

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	3.289	0.3767	8.75	<.0001
SATV	-0.000025	0.000618	-0.04	0.9684
SATM	0.002283	0.000663	3.44	0.0007