

Lecture 9: Multiple Regression (Introduction)

- The Multiple Regression Model – Chap. 4.1
 - Model description
 - Estimation of the coefficients
 - Interpretation of the coefficients in the multiple model vs. the simple regression model
- Inferences about the multiple regression coefficients - Chap 4.2
 - Confidence intervals for the coefficients
 - Hypothesis tests for the coefficients
- **Example:** Gasoline Mileage for New Cars (2004)

General Description

- In multiple regression analysis, we consider more than one independent variable x_1, \dots, x_k . We are interested in the conditional mean of y given x_1, \dots, x_k .
- Examples:
 - College admissions: The admissions officer wants to predict which students will be most successful. She wants to predict college GPA based on (x_1) GPA in high school, (x_2) total SAT score and (x_3) amount of time spent participating in extracurricular activities.
 - Business decision-making: La Quinta Inns wants to decide where to locate new inns. It wants to predict profit based on variables related to demand, demographics, competition and physical location.

Example: Gas Mileage

- A team charged with designing a new automobile is concerned about the gas mileage that they should aim to achieve. They want to aim for the average value that their competitors would achieve for such a car. We'll use **$Y = \text{MPG-City}$** .
- The new car is planned to have the following characteristics:
 - horsepower** – 225; **weight** – 4000 lbs;
 - seating** – 5 adults; **length** - 180"
- To estimate what the average competitor would achieve for such a vehicle they gathered data on all car models in 2004.
- The data includes information for each model about gas mileage and about the 4 variables: horsepower, weight, seating, & length.
- Here are some lines of the data:

Selected Automobile Data

Make/Model	MPG City	MPG Hwy	HP	WT (1000)	Seats	Length	Width	Displ	Cyl
Acura_RL	18	24	225	3.898	5	196.6	71.6	3.5	6
Acura_TL	20	28	270	3.575	5	189.3	72.2	3.2	6
Acura_TSX	23	32	200	3.318	5	183.3	69.4	2.4	4
Acura_RSX	25	34	160	2.771	4	172.2	67.9	2	4
Buick_Century	20	30	175	3.342	6	194.6	72.7	3.1	6
Buick_LeSabre	20	29	205	3.567	5	200	73.5	3.8	6
Chevrolet_Blazer	16	21	190	3.591	5	176.8	67.8	4.3	6
Chevrolet_Cavalier	24	34	140	2.676	5	180.9	67.9	2.2	4
Chevrolet_Corvette	18	25	350	3.214	2	179.7	73.6	5.7	8

There are 43 car Makes in the data, with a total of 242 Make/Model entries. We viewed 20 Make/Models as “not ordinary”, and will **exclude** them from future analyses.

Thus our analysis only holds for “ordinary” car types.

Answers via Simple Regression

- We could use simple regression to derive 4 prediction equations. Each could give a predicted value corresponding to their respective design parameter, x .

{See demonstration in Lecture.}

- The following table shows the coefficients of the 4 resulting linear prediction equations, and the corresponding **Prediction of Y** at each of the given values of our design parameters.

Table of Linear Predictions

x -Variable	b_0	b_1	x -value	Pred of Y
HP	28.34	- 0.0426	225	18.8
Wt(1000lb)	34.21	- 4.030	4	18.1
Seats	24.05	- 0.905	5	19.5
Length	46.47	- 0.145	180	20.4

- Note that the 4 predictions are different from each other.

Choosing the best of the simple linear regression predictors

- We could use _____ to choose the best prediction among these 4.
- Here is a table of correlations among the 5 variables: Y and our 4 x -variables. This can help choose the best prediction.

{See demonstration in lecture for JMP procedure to get this.}

	MPG_Cty	HP	Wt (1000lb)	Seats	Length
MPG_City	1.000	-0.737	-0.849	-0.322	-0.550
HP	-0.737	1.000	0.643	0.064	0.491
Wt (1000lb)	-0.849	0.643	1.000	0.586	0.702
Seats	-0.322	0.064	0.586	1.000	0.601
Length	-0.550	0.491	0.702	0.601	1.000

- Note that the maximum $|R|$ between Y and any of the x -variables is $|R|=0.849$ between **MPG_City** and **Wt**.
- It can be argued that among the 4 x -variables $x_2 = \mathbf{Wt}$ does the best job. The corresponding 95% CI is (17.8, 18.4).

Note that this CI does not contain any of the other predictions!

Summary Statement (for analysis up to here)

- Here, there are several possible x -variables that can be used as independent variables in a simple linear regression to predict Y .
- As is typical in such situations, they do not give the same answer.
- If one needs to choose a single one of the x -variables to be used, then the variable with the highest R^2 is the best choice. {Why?}
Highest $R^2 = (-.849)^2 = .721$.
- Here the “best” x is $x_2 = \text{Wt}$ with $\hat{Y} = 34.2 - 4.03 \times 4 = 18.1$.

But,

is there a way to use all the x -variables together?

Yes!

The way to use all the x -variables together is ...

Multiple Regression Model

- For each Y_i -value there are K independent predictors, labeled as x_{1i}, \dots, x_{Ki} . And, $i = 1, \dots, n$. {For our example, $K=4, n=223$.}

- The mean of each Y_i is “linear” in its x -values. Symbolically,

$$\mu_{Y_i|x_{1i}, \dots, x_{Ki}} = \beta_0 + \beta_1 x_{1i} + \dots + \beta_K x_{Ki} .$$

- As with simple linear regression, there is also random error. *ie*

$$Y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_K x_{Ki} + e_i .$$

- β_0 is the *Y-intercept*.
- The β_1, \dots, β_K are the *regression coefficients* corresponding to the respective independent variables.
- Each β_k describes how much $\mu_{Y_i|x_{1i}, \dots, x_{Ki}}$ changes when x_{ki} changes by one unit ***and all other x-values are held fixed.***

Assumptions about the random errors, e_i :

- Their expected values are all 0, *ie*

$$E(e_i) = 0.$$

– This makes $\mu_{Y_i|x_{1i}, \dots, x_{Ki}} = \beta_0 + \beta_1 x_{1i} + \dots + \beta_K x_{Ki}$.

- Homoscedasticity: The variances of every e_i are equal, *ie*

$$\text{Var}(e_i) = \sigma_e^2 \text{ for every } e_i.$$

- The e_i are normally distributed.
- The e_i are independent.

Least Squares Estimators

- Statistical software (such as JMP) produces the least-squares estimators, denoted by b_0, b_1, \dots, b_K .
- Corresponding to these is a least-squares estimator of $\mu_{Y_i|x_{1i}, \dots, x_{Ki}}$, for every $i = 1, \dots, n$; namely

$$\hat{Y}_i = b_0 + b_{1i}x_{1i} + \dots + b_{Ki}x_{Ki} .$$

- Least Squares in JMP: Click Analyze→Fit Model; put the dependent variable into Y; select the independent variables and click them to “Add” them into the “Construct Model Effects” box. Then click on “Standard Least Squares”.

Why are these called “Least-Squares” estimators?

- The sum of squared deviations between the observed Y_i and their estimates is called the Sum of Squares for Error, (= SSE).
- Symbolically,

$$(*) \quad SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \square \sum_{i=1}^n (Y_i - (b_0 + b_{1i}x_{1i} + \dots + b_{Ki}x_{Ki}))^2 .$$

- “*Least-Squares*” means that b_0, b_1, \dots, b_K are chosen to minimize the sum of squared deviations on the right of (*) among all possible choices of b^s .
- Symbolically, for any choice of constants $\tilde{b}_0, \tilde{b}_1, \dots, \tilde{b}_K$

$$\sum_{i=1}^n (Y_i - (b_0 + b_{1i}x_{1i} + \dots + b_{Ki}x_{Ki}))^2 \leq \sum_{i=1}^n (Y_i - (\tilde{b}_0 + \tilde{b}_{1i}x_{1i} + \dots + \tilde{b}_{Ki}x_{Ki}))^2 .$$

- JMP (and all other software) does this by **MATH-MAGIC!**

SSE, DF and MSE

- We have already described the **SSE** (Sum of Squares for Error) as

$$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - (b_0 + b_{1i}x_{1i} + \dots + b_{Ki}x_{Ki}))^2,$$

where b_0, b_1, \dots, b_K are the least squares coefficients.

- There are n observations and we're estimating $K + 1$ coefficients. Hence this SSE has **Degrees of Freedom**
 $\mathbf{DF} = n - K - 1$.
- We correspondingly define the **Mean Squared Error**, denoted by s_e^2 , as

$$s_e^2 = \text{MSE} = \frac{SSE}{DF} = \frac{SSE}{n - 1 - K}.$$

JMP Output

Here is some of the JMP output for our example:

Summary of Fit

RSquare	0.795
Root Mean Square Error	$s_e = 1.649$
Observations	$n = 222$

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	$b_0 = 31.50$	1.77	17.84	<.0001
HP	$b_1 = -0.0154$	0.00280	-5.50	<.0001
Wt (1000)	$b_2 = -3.77$	0.278	-13.56	<.0001
Seats	$b_3 = 0.337$	0.136	2.47	0.0141
Length	$b_4 = 0.0175$	0.0126	1.39	0.1660

Thus, for our car of interest with $x_1 = 225$, $x_2 = 4$, $x_3 = 5$, $x_4 = 180$

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

Meaning of the LS coefficients

- Recall that each β_k describes how much $\mu_{Y_i|x_{1i}, \dots, x_{Ki}}$ changes when x_{ki} changes by one unit *and all other x-values are held fixed*.
- Each b_k has the same interpretation relative to $\hat{Y}_i = \hat{\mu}_{Y_i|x_{1i}, \dots, x_{Ki}}$, *ie*, each b_k describes how much \hat{Y}_i changes when x_{ki} changes by one unit *and all other x-values are held fixed*.
- An alternate description is: b_k is the estimate of “the per-unit effect of changing x_{ki} *after controlling for* all the other variables in the model”.
- This interpretation may need to be kept in mind when explaining the values of the b_k . Otherwise some results may seem quite strange. FOR EXAMPLE

A Paradoxical Result

- Note that the coefficient of seats is

$$b_3 = +.337.$$

- This *seems* to say that “Cars with **more** seats get **more** gas mileage”.
- That’s counter-intuitive.
- Also, our ordinary regression analysis of Y on only x_3 yielded a slope of $-.905$. This is more intuitively sensible; it says that cars with more seats get **less** mileage. (For this slope see slide 5.)
- Can the positive value of b_3 make any sense?

Can the positive value of b_3 make any sense?

- Possibly so. Remember that b_3 measures what is the relation between number of seats and gas-mileage for the car models **so long as** their Wt and HP and length are **fixed**.
- **Hence maybe:** For a fixed basic car structure (especially Wt and engine HP) the car designers think, “Is this particular model for ‘families’ or for ‘singles’”.
 - If it’s for families then we’ll put in more seats and tweak the performance, etc, so as to save fuel – since more seats and better gas mileage are important for family buyers.
 - If it’s for ‘singles’ we’ll do the reverse – take out seating capacity in favor of more space, fancier interiors, etc, and tweak the performance so as to make the car sportier even if that reduces gas mileage.

- If the **brown** explanation is the correct one, then it does make sense that we found $b_3 > 0$, since **for fixed Wt and HP** the cars with more seating capacity would on average provide better gas mileage.
 - On the other hand, it would also be reasonable that the simple regression of Y on Seating had a negative slope, because
 - In spite of the **brown** explanation it can still be true that cars with more seating capacity are generally heavier and have more powerful engines.
 - **Actually**, what the correlations on slide 6 show is that there's very little (linear) relation between seating and HP, but there's a **large, positive correlation** (of 0.586) between seats and Wt. This means our **brown story** is possible. It doesn't prove it's true
- For further analysis see the optional note at the end of the lecture.*

Tests and CIs for the Coefficients

- Statistical Software (JMP) produces “standard errors” for each of the coefficients.
 - These come from an extension of the same MATH-MAGIC that produces the values of the coefficients b_i .
- They can be used in the usual fashion to produce
 - CIs for the true values of the β ’s.
 - Tests of hypotheses about the β ’s.
- CI^s:
 - These involve the t-distribution with $n - 1 - K$ **D**egrees of **F**reedom. {Recall, K denotes the number of dependent variables.}
 - A typical CI for β_i is of the form

$$b_i \pm t_{\alpha/2;DF} \times SE(b_i).$$

Example CI

- CI for β_3 , the model's coefficient of **Seats**:

Recall from the Parameter Estimates table (*on p. 12*) that

$$b_3 = 0.337 \text{ and } SE(b_3) = 0.136.$$

- For our data $n = 222$ entries and $K = 4$ independent variables.
Hence

$$DF = n - 1 - K = 222 - 5 = 217 .$$

- For a 95% CI the critical value is

$$t_{\alpha/2;217} = 1.96$$

- Hence the 95% CI for is

$$0.337 \pm 1.96 \times 0.136 = 0.337 \pm .267 = (0.070, 0.604).$$

- This CI does not contain 0; hence the **test** of $H_0 : \beta_3 = 0$ vs $H_a : \beta_3 \neq 0$ **Rejects** the null hypothesis.
- We can also directly construct the test and its P-value.

Hypothesis Test and P-Value

- To test $H_0 : \beta_3 = 0$ vs $H_a : \beta_3 \neq 0$ calculate the corresponding t-statistic

$$t = \frac{b_3 - 0}{SE(b_3)}.$$

- This alternative is two-sided. So,

$$\text{Reject if } |t| > t_{\alpha/2; DF},$$

(with $DF = n - 1 - K$, as before.)

- In our example we reject since

$$t = \frac{.337}{.136} = 2.48 \ \& \ t_{\alpha/2; 217} \approx 1.96$$

- The P-value can be found from a detailed t-table or by accessing the t-distribution in JMP.
- All the above answers can also be found directly in JMP, but for CI^s that are not 95% you need to calculate as above.

Optional notes about the “Paradoxical result concerning Seats”

- The explanation in **brown** isn't the only *plausible* possibility. Two others are:
 - (1) Statistically, the coefficient b_3 isn't significantly different from 0, so we shouldn't be bothered with trying to explain that. Or,
 - (2) The assumptions for a multiple regression aren't valid and so the entire analysis is misleading.
- *Answers to these possibilities:*
 - (1) *We've now seen that the coefficient b_3 IS significantly different from 0 within the model in question, so this possibility doesn't apply.*
 - (2) *We'll later find that the assumptions don't seem quite valid – although they're not too far afield – but a more valid model and corresponding analysis still yield essentially the same statistically significant paradox with respect to length (and it still holds with respect to Seats, but isn't statistically significant there) .*

cont.

Here's a more detailed look at the situation:

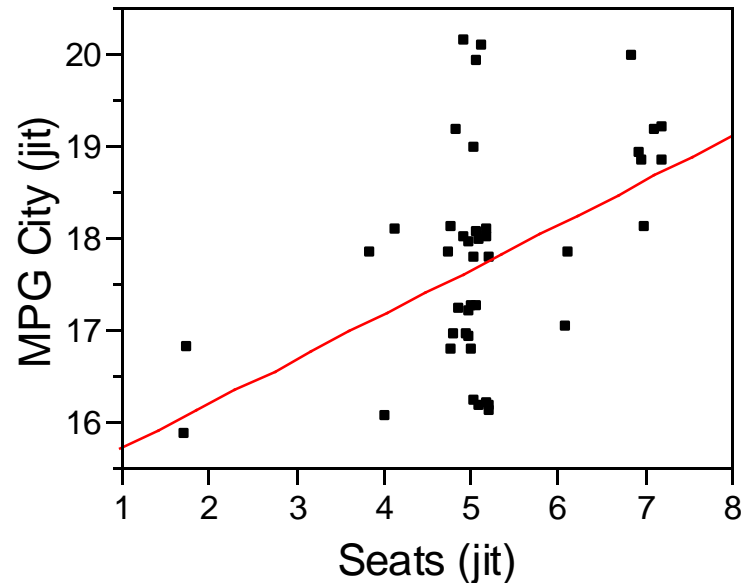
- The analysis suggests that *for given* Wt, HP and Length the mileage *increases* with increasing Seats.
- To examine this, we'd like to consider only those cars with $Wt \approx 4$, $HP \approx 225$, and $Length \approx 180$. However, there aren't many car models satisfying **all** these conditions.
- So, let's concentrate on the story involving only **Wt** and **Seats** to predict MPG, and ignore the other two variables.
- This smaller model has the same paradox. This is revealed in its Parameter Estimates table. Here is the table for the multiple regression of MPG on Wt(1000lb) and Seating.

Parameter Estimates				
Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	33.09	0.606	54.59	<.0001
Wt (1000lb)	-4.77	0.191	-24.94	<.0001
Seating	+0.747	0.113	6.61	<.0001

Note that the coefficient of Seating is **positive**. So, the analysis says that *on average* for given values of **Wt** the car models with *more* **Seats** give *better* **MPG**.

- Let's see whether this holds for cars having $Wt(1000lb) \approx 4$, like our special model. Specifically, let's look at ALL car models in the data that have $3.75 \leq Wt(1000lb) \leq 4.25$.
- There are 42 models with Wt in this range. Here is the scatterplot of **MPG** on **Seats** for these 42 models, along with the least squares line. [All values of Wt and $Seats$ are integers, so we've slightly *jittered* all the values so that they will show up more clearly on the scatterplot.]
- This scatterplot and least squares analysis shows that **among these 42 models**, on average the (linear) relationship between Seating and MPG has a clearly positive slope.
- This happens in spite of the fact that in a simple regression **for all cars** of **MPG** on **Seats** the regression coefficient is significantly negative.
 - From p. 6 the corresponding value of b_1 is $b_1 = -.905$.

Least Squares Analysis of **MPG** on **Seats** for the Models with $3.75 \leq \text{Wt}(1000\text{lb}) \leq 4.25$



	Parameter Estimates			
Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	15.24	0.752	20.27	<.0001
Seats (jit)	0.486	0.141	3.44	0.0014

Note that the coefficient for Seats is very significantly positive. (This is also true for the non-jittered analysis.) This is what was claimed to be shown.