ST517 Note Outline 10: Multiple Regression

Lecture 10.1: The Multiple Regression Equation

Recall: The simple linear regression model

- Sample line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
- Population line: $\hat{y} = \beta_0 + \beta_1 x$

Recall: The residuals $e = y - \hat{y}$

- Vertical distance from the point to the line
- It follows that $y = \hat{y} + e = \hat{\beta}_0 + \hat{\beta}_1 x + e$
    - On the population level: $y = \beta_0 + \beta_1 x + \varepsilon$

Recall BAC Example:

- When using the number of beers consumed to explain BAC:
    - BAC = −0.01270 + 0.01796∗beers
    - About 80% of variability is explained by number of beers
    - Meaning 20% is explained by other variables (e.g. body weight, amount of food consumed, sex [male or female] etc.)

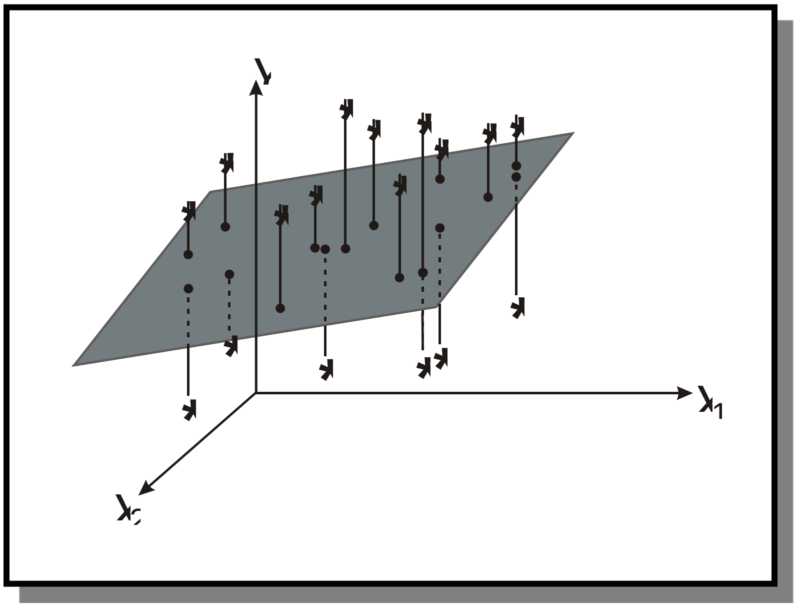
BAC	Weight	Sex	Beers
.10	132	female	5
.03	128	female	2
.19	110	female	9
.12	192	male	8
.04	172	male	3
.095	250	female	7
.07	125	female	3
0.06	175	male	5
.02	175	female	3
.05	275	male	5
.07	130	female	4
.10	168	male	6
.085	128	female	5
.09	246	male	7
.01	164	male	1
.05	175	male	4

Multiple Regression

- Using multiple variables to predict the response

- Types of predictors:
  - Other quantitative (numeric) variables that may help explain *y*
  - Other categorical variables that may help explain *y*
  - Higher order terms, such as:
    - Polynomial terms to model curved relationships
    - Interactions of two or more variables

Visually:

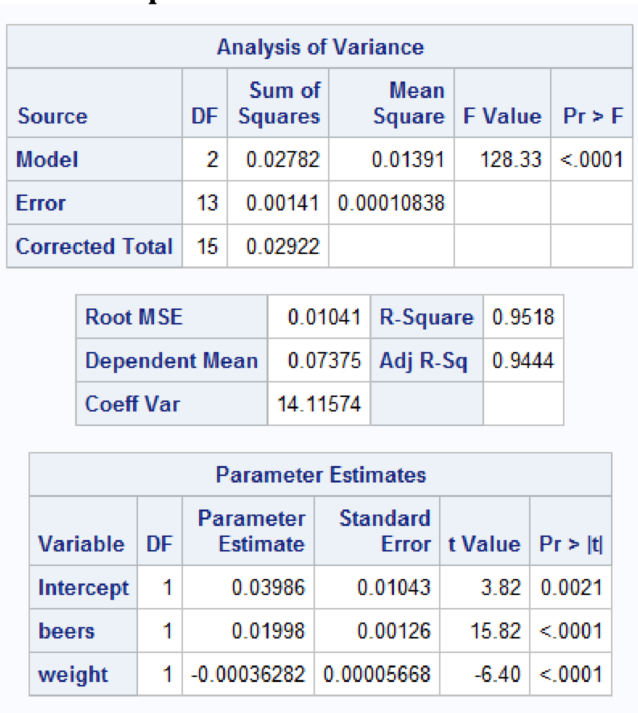


General form for the multiple regression model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_k x_k + \varepsilon$

- *k* is number of predictors; *ε* is random error
- β_0 still represents y-intercept

- β_i , for *i*=1,…,*k*, represents one slope term

BAC Example:



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	0.02782	0.01391	128.33	<.0001
Error	13	0.00141	0.00010838		
Corrected Total	15	0.02922			

Root MSE	0.01041	R-Square	0.9518
Dependent Mean	0.07375	Adj R-Sq	0.9444
Coeff Var	14.11574		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.03986	0.01043	3.82	0.0021
beers	1	0.01998	0.00126	15.82	<.0001
weight	1	-0.00036282	0.00005668	-6.40	<.0001

“Simple” regression equation:

Multiple regression equation:

⇒ Slope terms are different than they would be if other variables were not in the model

At 180 pounds:

At 150 pounds:

⇒ Different values for one predictor (e.g. weight) lead to different lines

Lecture 10.2: Judging the fit of the model

Recall: Conditions necessary for model to be valid

1. A straight line is the correct model for the data

2. The spread of the points around the line have the same standard deviation for all x.
3. The random errors are independent of each other.
4. The points are normally distributed around the line.

Can re-write these conditions in terms of the random errors

- If the specified model is appropriate for the data,

- The standard deviation of the points around the line

- If the points are normally distributed around the line,

- If the points are centered at the line,

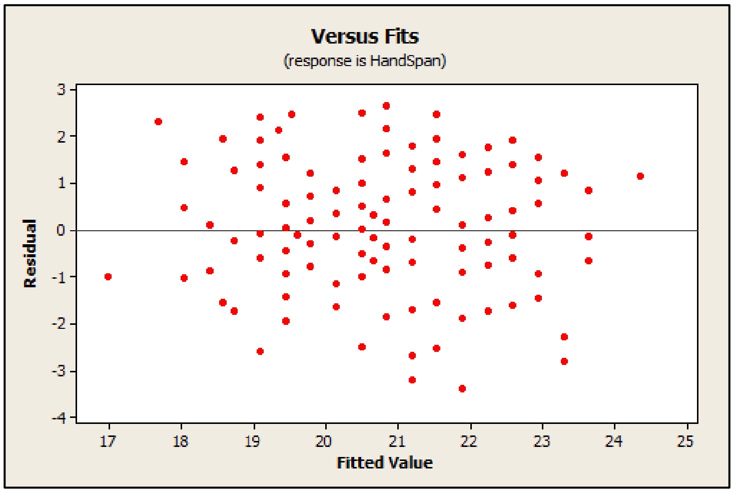
- Finally, we need the random errors to be independent.

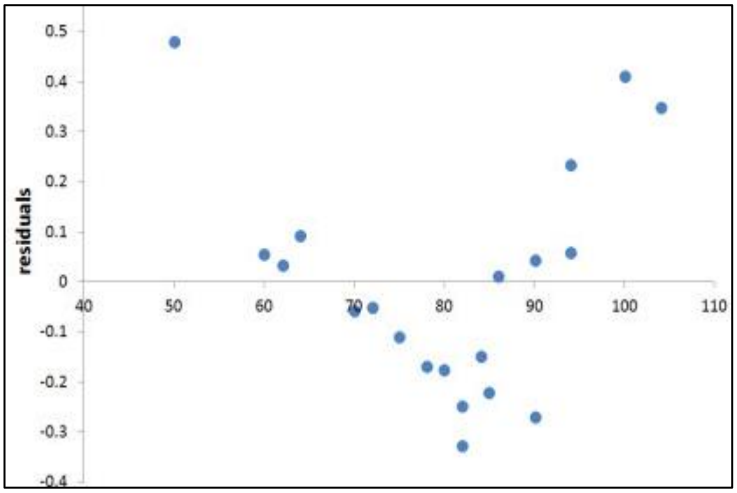
Visually: Check <u>residual plot</u>

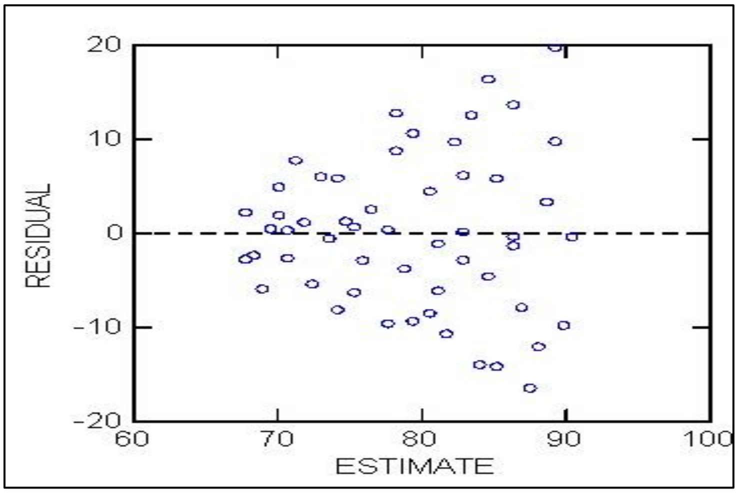
- Plot of residuals vs.

- Plot of residuals vs.

Examples of residual plots (from a variety of software)







Numeric summary: Coefficient of Determination

- Recall: $R^2 =$ the square of the correlation coefficient
- Can also be found directly using model sums of squares:

$$R^2 = \frac{SSR}{SS_{yy}} = 1 - \frac{SSE}{SS_{yy}}$$

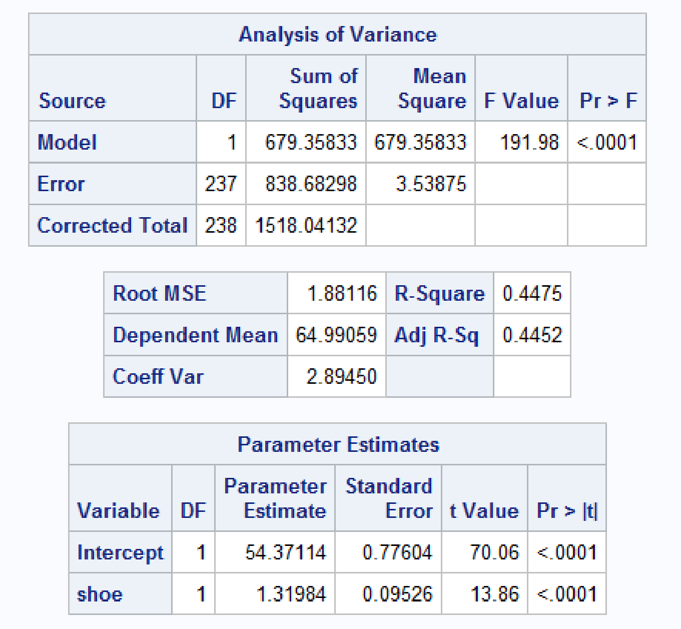
- Interpretation: In the multiple regression model, R^2 is the proportion of variability in y explained by the model overall

BAC example:

Number of beers consumed explained 80% of the variability in BAC.

Number of beers consumed <u>and</u> weight together explain _______ of the variability in BAC.

Example: Using shoe size to predict height



Analysis of Variance

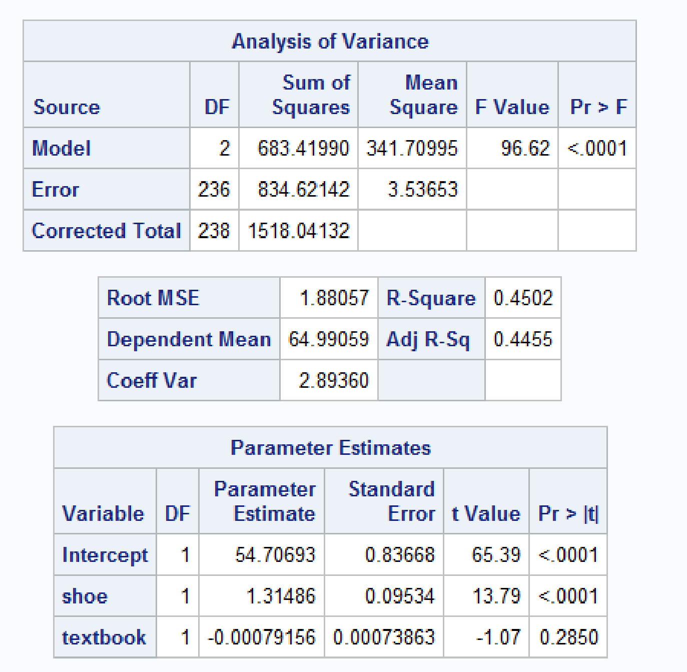
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	679.35833	679.35833	191.98	<.0001
Error	237	838.68298	3.53875		
Corrected Total	238	1518.04132			

Root MSE	1.88116	R-Square	0.4475
Dependent Mean	64.99059	Adj R-Sq	0.4452
Coeff Var	2.89450		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	54.37114	0.77604	70.06	<.0001
shoe	1	1.31984	0.09526	13.86	<.0001

Example: Using shoe size and *amount of money spent on textbooks* to predict height



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	683.41990	341.70995	96.62	<.0001
Error	236	834.62142	3.53653		
Corrected Total	238	1518.04132			

Root MSE	1.88057	R-Square	0.4502
Dependent Mean	64.99059	Adj R-Sq	0.4455
Coeff Var	2.89360		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	54.70693	0.83668	65.39	<.0001
shoe	1	1.31486	0.09534	13.79	<.0001
textbook	1	-0.00079156	0.00073863	-1.07	0.2850

Adjusted R^2

- R^2 can be driven artificially close to 100% by adding (sometimes unnecessary) more predictors to the model
- Adjusted R^2 corrects for this by "penalizing" the number of predictors used:

BAC example: Accounting for the number of variables in the model, the number of beers consumed <u>and</u> weight together explain _______ of the variability in BAC.

Lecture 10.3: Inference for multiple regression

Two types of tests:

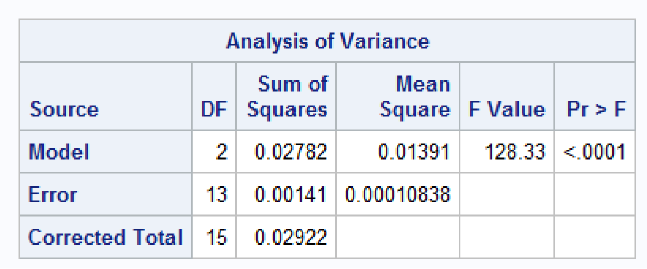
1. F-test for overall significance of the model
2. t-tests for individual predictors

- In simple linear regression, these are basically the same test (since there is only one predictor), but in multiple regression they have slightly different goals

F-test for overall effect:

- Steps 1, 6, and 7 same as always
- Hypotheses:
- Conditions:
- Test statistic:
- Null distribution:
- p-value represents proportion of null distribution that is greater than or equal to the test statistic

BAC Example (F-test):



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	0.02782	0.01391	128.33	<.0001
Error	13	0.00141	0.00010838		
Corrected Total	15	0.02922			

t-test for individual predictors

- Steps 1, 6, and 7 same as always
- Hypotheses:

- Conditions:

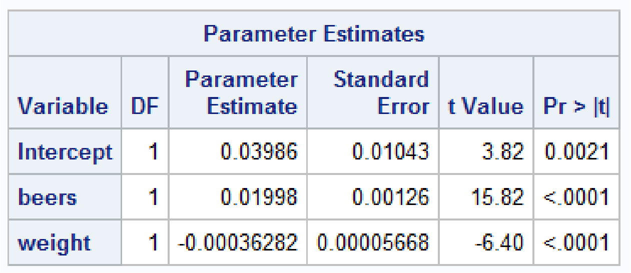
- Test statistic:

- Null distribution:

- P-value found under null distribution in direction of alternative

- New interpretation:

BAC Example (t-tests):



Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.03986	0.01043	3.82	0.0021
beers	1	0.01998	0.00126	15.82	<.0001
weight	1	-0.00036282	0.00005668	-6.40	<.0001

Lecture 10.4: Categorical Predictors

Indicator variables

- To use categorical variables in regression we need to convert them to indicator variables:

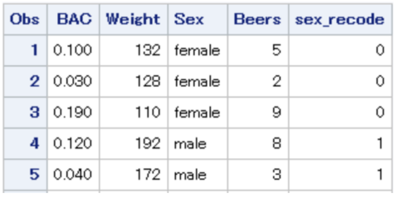
$$x = \begin{cases} 1 \text{ if yes} \\ 0 \text{ if no} \end{cases}$$

- Also called

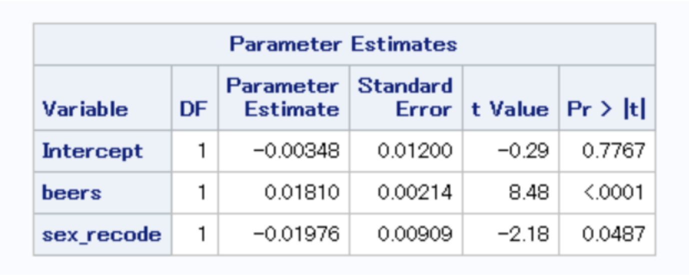
- General form of model is the same as before

BAC Example: Sex

$$X = \begin{cases} 1 \text{ if male} \\ 0 \text{ if female} \end{cases}$$



Obs	BAC	Weight	Sex	Beers	sex_recode
1	0.100	132	female	5	0
2	0.030	128	female	2	0
3	0.190	110	female	9	0
4	0.120	192	male	8	1
5	0.040	172	male	3	1



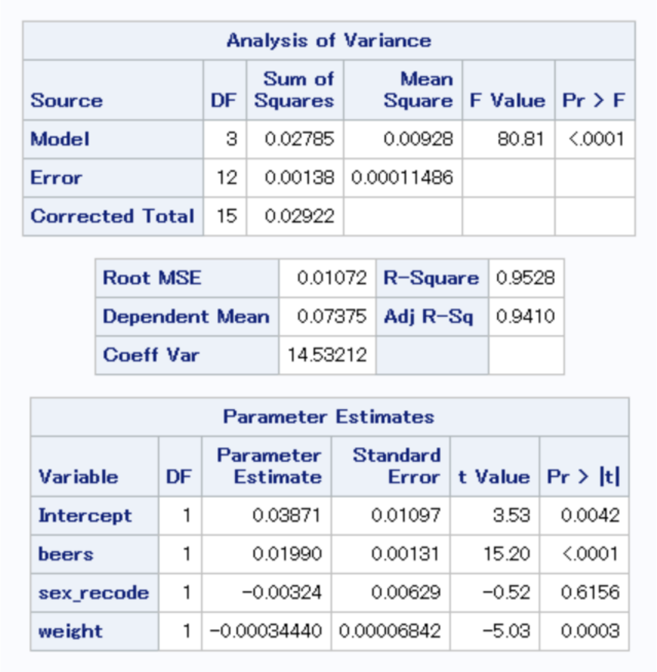
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	−0.00348	0.01200	−0.29	0.7767
beers	1	0.01810	0.00214	8.48	<.0001
sex_recode	1	−0.01976	0.00909	−2.18	0.0487

If male:

If female:

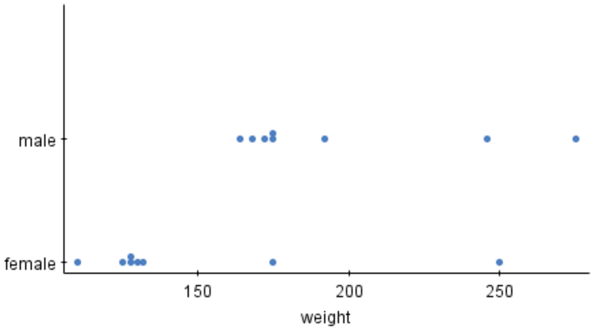
BAC Example:



Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	0.02785	0.00928	80.81	<.0001
Error	12	0.00138	0.00011486		
Corrected Total	15	0.02922			

Root MSE	0.01072	R-Square	0.9528
Dependent Mean	0.07375	Adj R-Sq	0.9410
Coeff Var	14.53212		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.03871	0.01097	3.53	0.0042
beers	1	0.01990	0.00131	15.20	<.0001
sex_recode	1	-0.00324	0.00629	-0.52	0.6156
weight	1	-0.00034440	0.00006842	-5.03	0.0003



If you have multiple categories, create multiple indicators

- A series of yes/no variables
- Each takes on 1 or 0
- Need one less variable than there are categories

Example: Political parties

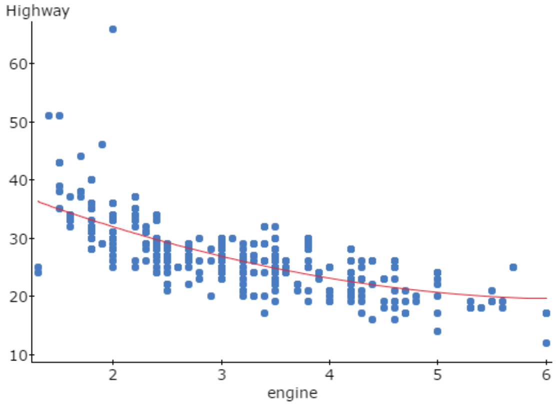
party	X1	X2
Republican	1	0
Democrat	0	1
Republican	1	0
Democrat	0	1
Democrat	0	1
Republican	1	0
Other	0	0

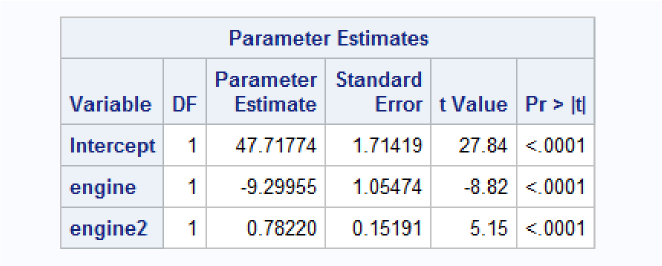
Lecture 10.5: Higher Order Terms

Polynomial Terms

- Used to model curved (rather than straight line) relationships
- Most common are quadratic models:

Example: Using engine size to predict gas mileage on the highway





Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	47.71774	1.71419	27.84	<.0001
engine	1	-9.29955	1.05474	-8.82	<.0001
engine2	1	0.78220	0.15191	5.15	<.0001

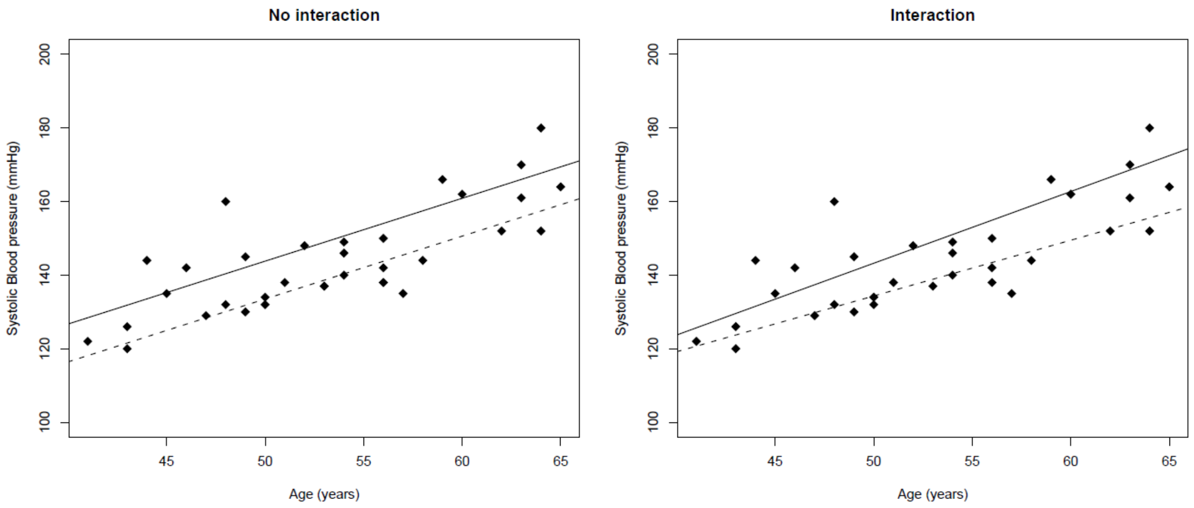
Estimated model:

Predicted value for a car that has a 1.8 liter engine:

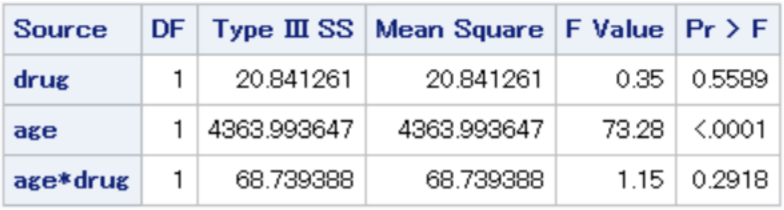
Interaction Terms

- Used to model differential effects
- Model with an interaction:

- Interactions change slope and/or intercept of model
  - Let x_1 be an indicator variable
  - Basic model:
  - Interaction model:



Example (Hypertension study): A study of 32 men randomly assigned to receive either a placebo or an anti-hypertension drug. Also recorded was the age and systolic blood pressure of each subject.



Source	DF	Type III SS	Mean Square	F Value	Pr > F
drug	1	20.841261	20.841261	0.35	0.5589
age	1	4363.993647	4363.993647	73.28	<.0001
age*drug	1	68.739388	68.739388	1.15	0.2918

A final note about higher order terms:

- What if the higher order term is significant, but one of its supporting lower order terms is not?

Ex: $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2$$

Lecture 10.6: Additional Examples

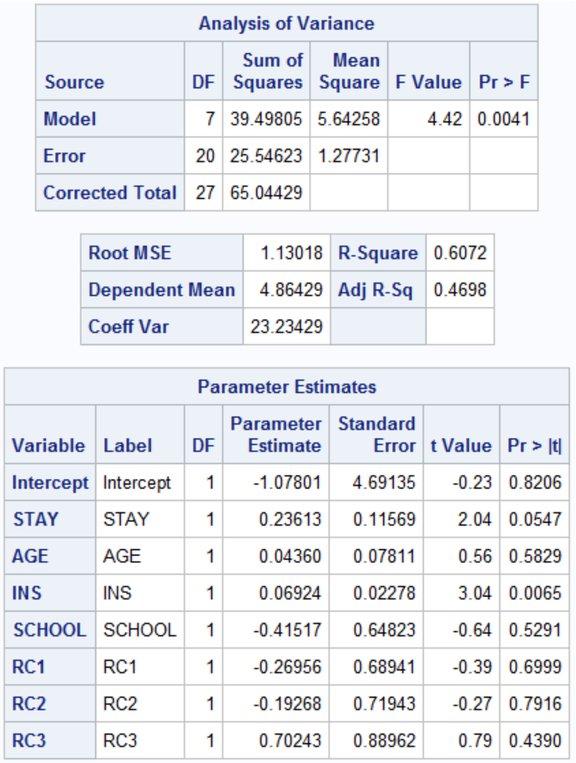
Use the following to answer questions 1 to 14: A study was conducted to determine whether infection surveillance and control program have reduced the rates of hospital-acquired infections. We have data on a random sample of 28 US hospitals with the following variables:

- RISK = average estimated probability of acquiring an infection while in the hospital (in percent)
- STAY = average length of stay of all patients in the hospital (in days)
- AGE = average age of all patients in the hospital (in years)
- INS = ratio of number of cultures performed to number of patients without signs or symptoms of hospital acquired infection (times 100)
- SCHOOL = indicator that the hospital is affiliated with a medical school (1 = yes, 0 = no)
- RC = region of country where the hospital is located (Northeast, North central, South, West)

1. How many indicator variables do we need to create so that RC could be included as a predictor in a regression model?
2. Define the appropriate indicator variables for RC using WEST as the baseline (reference) category.
3. SAS output (OUTPUT 1) on the next page shows results of a regression model predicting RISK based on all of the other variables. Predict percent RISK for a hospital that is located in the south, not affiliated with a medical school, and has an average length of stay of 7.13 days, average patient age of 55.7 years, and INS ratio of 9. Round the estimated coefficients to 2 decimal places. Round your final answer to 3 decimal places.
4. What percent of the variability in RISK is explained by the model?
5. What percent of the variability in RISK is explained by the model, after penalizing for the number of predictors?
6. On average, how far are the actual values of RISK from what we would predict based on the model?
7. Note that RC1 is insignificant. Could we remove just this variable from the model? Explain.
8. Suppose we wanted to see if there was an interaction between STAY and AGE; write the model equation (using generic $\hat{\beta}_i$ notation) that could be estimated to test this. Include all other predictor variables as well.
9. For the model in the previous question, imagine that the interaction term is significant. Which main effects could be removed from the model if they were found to insignificant? Explain.
10. SAS output (OUTPUT 2) on the next page shows the estimated model with the interaction. It would not be reasonable to interpret the estimated slope coefficient for STAY. Explain why not.

11. What test statistic and p-value would be used to determine if the model is significant overall? Draw a well-labeled picture of this p-value.
12. What test statistic and p-value would be used to determine if the interaction is significant? Draw a well-labeled picture of this p-value.
13. Suppose we wanted to see if there was a non-linear relationship between RISK and INS; write the model equation (using generic $\hat{\beta}_i$ notation) that could be estimated to test this. Include all other predictor variables as well. (Note: do not include the interaction from question 8.)
14. For the model in the previous question, imagine that the quadratic term is significant. Which main effects could be removed from the model if they were found to insignificant? Explain.

OUTPUT 1

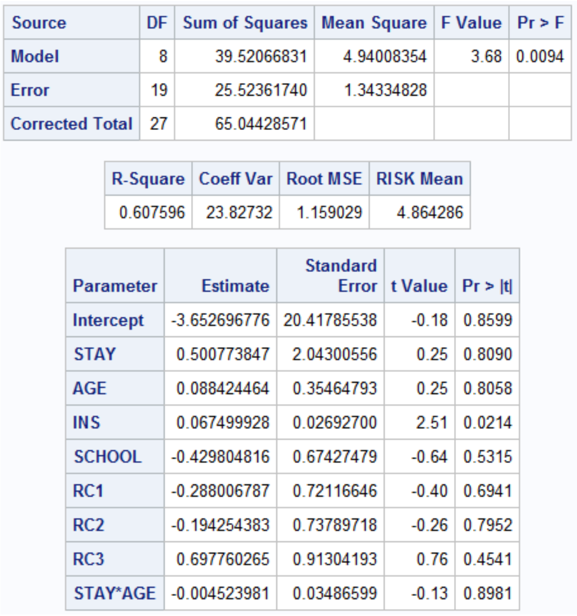


Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	39.49805	5.64258	4.42	0.0041
Error	20	25.54623	1.27731		
Corrected Total	27	65.04429			

Root MSE	1.13018	R-Square	0.6072
Dependent Mean	4.86429	Adj R-Sq	0.4698
Coeff Var	23.23429		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-1.07801	4.69135	-0.23	0.8206
STAY	STAY	1	0.23613	0.11569	2.04	0.0547
AGE	AGE	1	0.04360	0.07811	0.56	0.5829
INS	INS	1	0.06924	0.02278	3.04	0.0065
SCHOOL	SCHOOL	1	-0.41517	0.64823	-0.64	0.5291
RC1	RC1	1	-0.26956	0.68941	-0.39	0.6999
RC2	RC2	1	-0.19268	0.71943	-0.27	0.7916
RC3	RC3	1	0.70243	0.88962	0.79	0.4390

OUTPUT 2



Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	39.52066831	4.94008354	3.68	0.0094
Error	19	25.52361740	1.34334828		
Corrected Total	27	65.04428571			

R-Square	Coeff Var	Root MSE	RISK Mean
0.607596	23.82732	1.159029	4.864286

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-3.652696776	20.41785538	-0.18	0.8599
STAY	0.500773847	2.04300556	0.25	0.8090
AGE	0.088424464	0.35464793	0.25	0.8058
INS	0.067499928	0.02692700	2.51	0.0214
SCHOOL	-0.429804816	0.67427479	-0.64	0.5315
RC1	-0.288006787	0.72116646	-0.40	0.6941
RC2	-0.194254383	0.73789718	-0.26	0.7952
RC3	0.697760265	0.91304193	0.76	0.4541
STAY*AGE	-0.004523981	0.03486599	-0.13	0.8981

Note:

- RC1 is an indicator that the hospital is located in the Northeast region
- RC2 is an indicator that the hospital is located in the North central region
- RC1 is an indicator that the hospital is located in the South