

Lecture 12: Polynomial Regression

- Polynomial regression
 - Chapter 5.2.1

Polynomial Regression

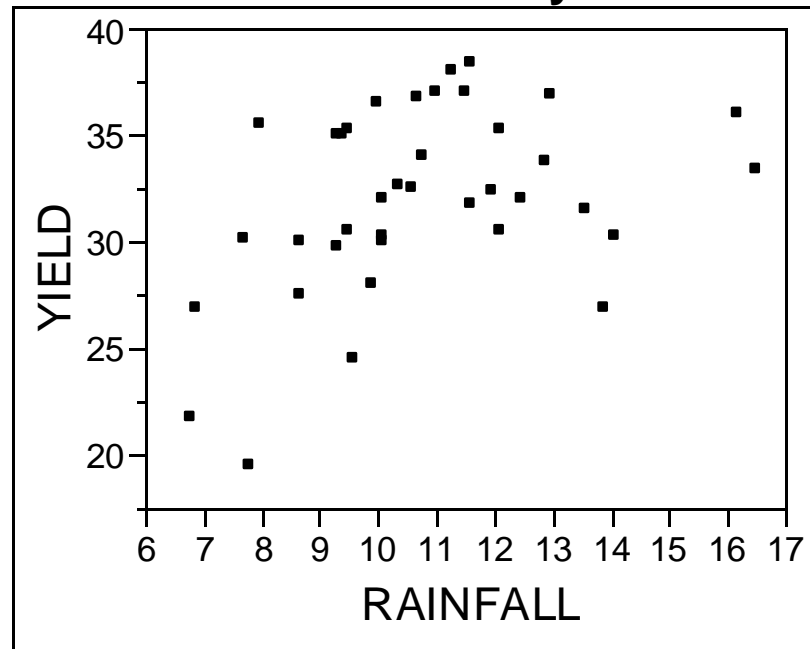
Another Method for Fitting Curvilinear Relationships

- Reconsider the **simple regression** problem of estimating $\mu_{Y|x}$ = the conditional mean of Y given x .
- For many problems, $\mu_{Y|x}$ is not linear in x .
- We have suggested transformations of x and (occasionally) transformations of Y to address this problem.
- In some situations these yield data in a form suitable for least squares analysis; but in others they do not work well.
- Polynomial regression is another convenient least-squares technique for fitting curvilinear data.
- We'll look at 3 examples, and then explain the theory.

Example 1: How does rainfall affect yield of corn?

- Data on annual corn yield and average rainfall in six US states (1890-1927). (See Cornyieldrainfall.jmp)

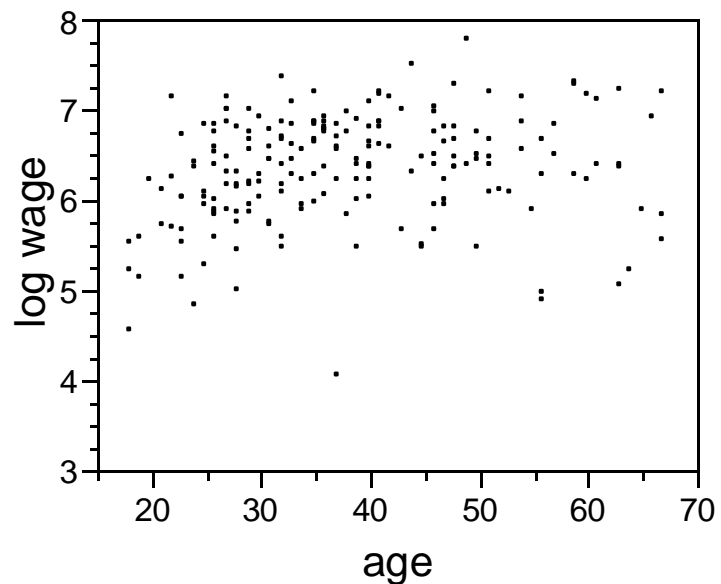
Bivariate Fit of YIELD By RAINFALL



- Note the generally curved pattern of points, with a max at about 11. Such a pattern cannot be well fit by transformations of x and/or Y .

Example 2: How do people's incomes change as they age?

- Weekly wages and age of 200 randomly chosen adult males
 - from the March 1998 Current Population Survey



- We'll see that there is also a curved pattern here, with max near 45
 - Note that we've already used $\log_e(\text{Wage})$ as the Y variable.

Arrival Pattern of Calls to a Financial Call Center

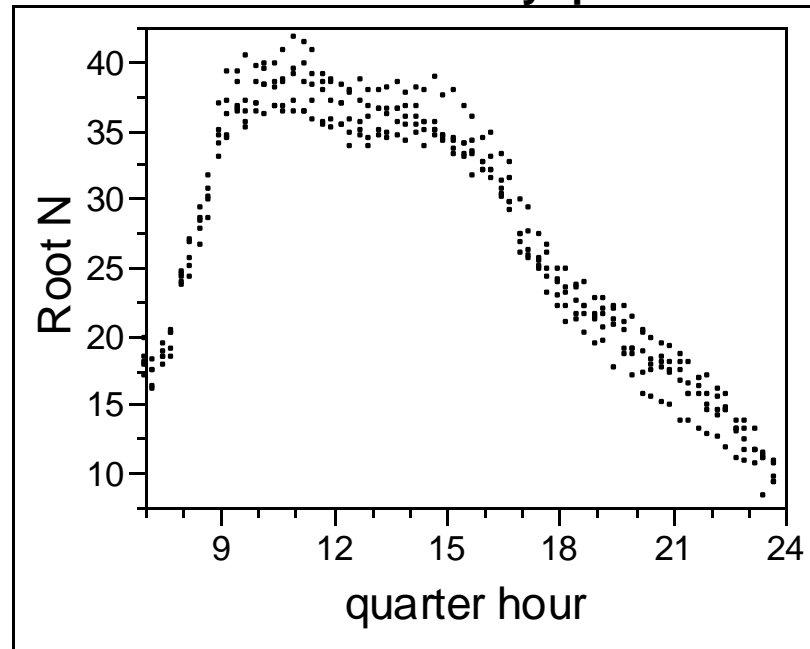
- Data is for week of 7/15/2002 – 7/19/2002.
- Data is from call center of a major US bank.
- Here's what a call center looks like:

A Call Center (picture is from England, ~1995)



- The number of calls made asking for service by an agent is automatically recorded (to the nearest second).
- The number of calls in each quarter hour of each weekday were totaled.
- For each data point $x = \text{time (to the quarter-hour)}$.
- For each data point $y = \sqrt{\# \text{ of calls in that quarter hour}}$.
 - The reason for transforming to the **sq rt** of the # of calls is explained in Brown, Gans, Mandelbaum, Sakov, Shen, Zeltyn and Zhao, “Statistical analysis of a telephone call center: a queueing science perspective”, *Jour. Amer. Statist. Assoc.*, **100**, 36-50.
- Information about arrival rates is needed in order to help plan staffing levels (# of agents on hand) at different times of day.
 - At peak times about 700 agents are at work in 3 physical locations.

Bivariate Fit of Root N By quarter hour



- Number of data points is **340** = **5**_[days] × **68**_[1/4 hr periods from 7AM-12PM]
- Total # of calls in 5 days = 277,680.

Polynomial Regression

- Add powers of x as additional explanatory variables in a multiple regression model.

$$\mu_{Y|x} = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_K x^K.$$

- Often $(x - \bar{x})$ is used in place of x .

$$\mu_{Y|x} = \beta_0 + \beta_1 x + \beta_2 (x - \bar{x})^2 + \dots + \beta_K (x - \bar{x})^K.$$

- These two forms of expressing a K -th degree polynomial are equivalent; they yield the same possible graphs -- only the values of the coefficients are different.
- A quadratic polynomial ($K = 2$) is often sufficient.
- Again, model fitting is via least squares; coefficients b_0, \dots, b_K are chosen to minimize the sum of squared error:

$$SSE = \sum (Y_i - \hat{Y}_i)^2 \quad \text{where } \hat{Y}_i = b_0 + b_1 x_i + \dots + b_K x_i^K.$$

Polynomial Regression in JMP

- Two ways to fit model:
 1. Create new variables with the formulas $(x - \bar{x})^2, \dots, (x - \bar{x})^K$. Use Fit Model to fit a multiple regression with the independent variables:
 $x, (x - \bar{x})^2, \dots, (x - \bar{x})^K$. OR
 2. Use Fit Y by X platform. Click on the red triangle... and click **Fit Polynomial** instead of the usual **Fit Line**. Click on a choice of K .
 - This gives the analysis tables *and* a nice plot of the curve.

Linear Fit: $YIELD = 23.55 + 0.776 RAINFALL$

Summary of Fit

RSquare 0.162
Root Mean Square Error 4.049
Observations 38

Polynomial Fit Degree=2:

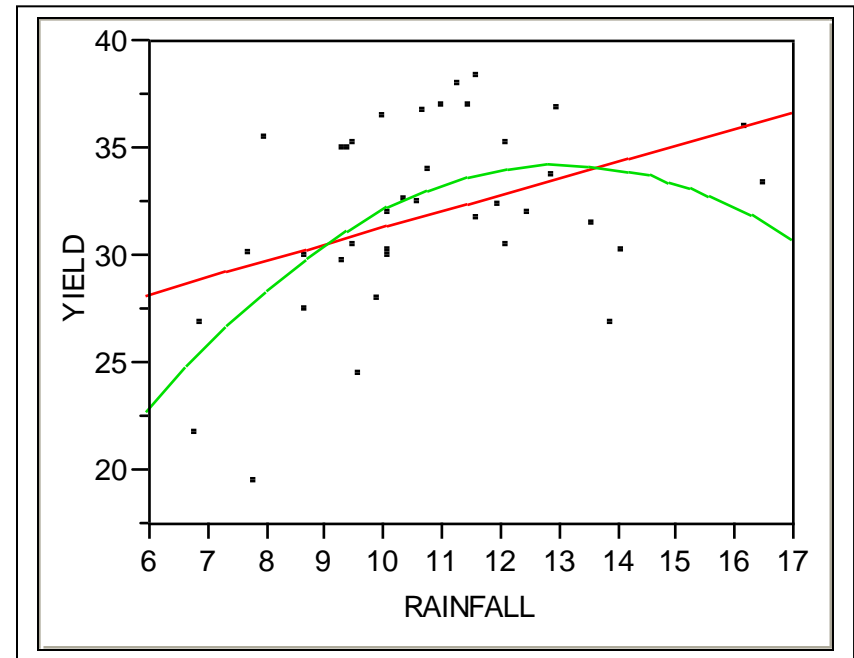
$YIELD = 21.66 + 1.057 RAINFALL - 0.229 (RAINFALL - 10.78)^2$

Summary of Fit

RSquare 0.297
Root Mean Square Error 3.763
Observations 38

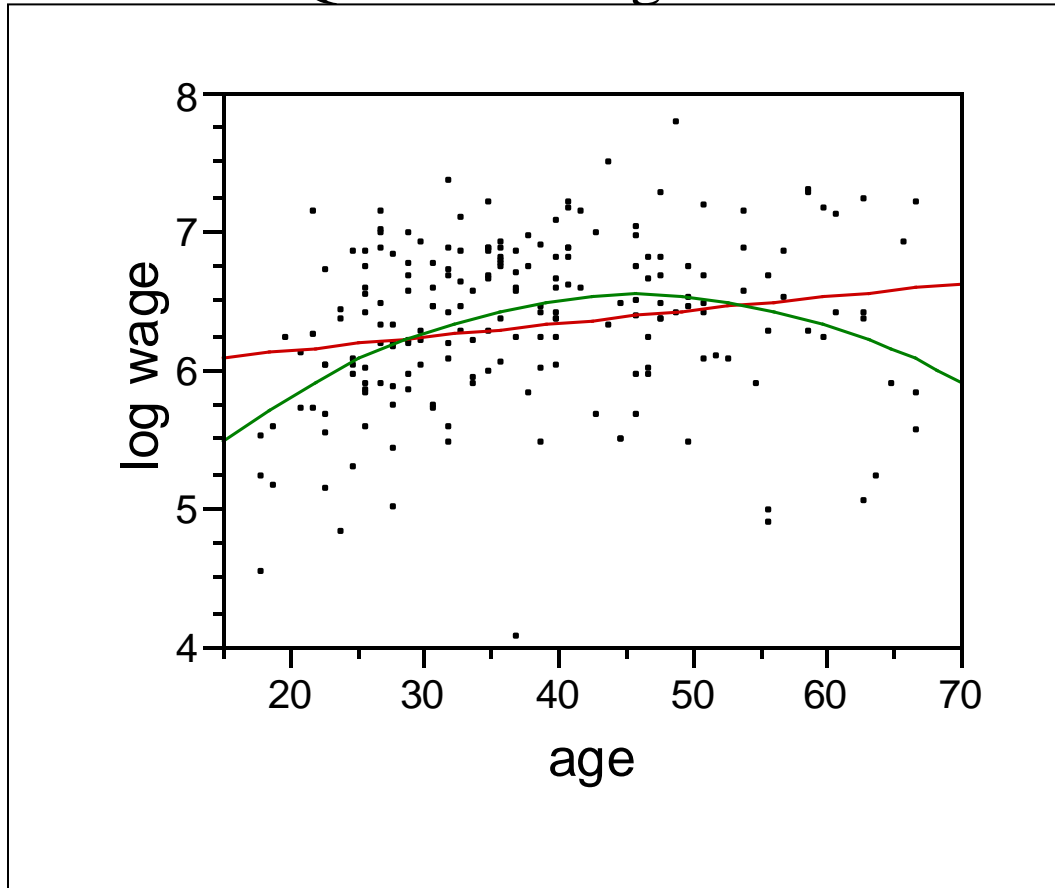
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	21.66	3.094	7.00	<.0001
RAINFALL	1.057	0.2940	3.60	0.0010
$(RAINFALL - 10.78)^2$	-0.229	0.0886	-2.59	0.0140



- Note that the quadratic regression fits better than the linear one.
(R^2 increases from 0.162 to 0.297)
- P-value is 0.014 that the quadratic coefficient is 0.

Plot of Linear and Quadratic Regression for the Weekly Wages Data



Output from JMP Analysis of Weekly Wages Data

Linear Fit: $\log \text{ wage} = 5.960 + 0.00974 \text{ age}$

Summary of Fit

RSquare	0.0389
Root Mean Square Error	0.585
Observations	200

Polynomial Fit Degree=2: $\log \text{ wage} = 5.843 + 0.0169 \text{ age} - 0.00109 (\text{age}-38.22)^2$

Summary of Fit

RSquare	0.123
Root Mean Square Error	0.561

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	8.65	4.32	13.76
Error	197	61.90	0.314	Prob > F
C. Total	199	70.55		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	5.843	0.135	43.37	<.0001
age	0.0169	0.00369	4.59	<.0001
(age-38.22)^2	-0.00109	0.000251	-4.33	<.0001

Interpretation of Coefficients in Polynomial Regression

- The usual interpretation of multiple regression coefficients doesn't make sense in polynomial regression: *eg* for

$$\mu_{Y|x} = \beta_0 + \beta_1 + \beta_2 (x - \bar{x})^2$$

- We can't hold x fixed and change $(x - \bar{x})^2$, nor *vice-versa*.
- Nevertheless we can use the Parameter estimates table to see whether adding the K -th coefficient (the last one) makes a statistically significant improvement.
- See the preceding tables to note that in each case the quadratic coefficient is statistically different from 0. (P=.014 and P<.0001, resp.)
- An ANOVA table, like that in the wages printout can show that the linear and quadratic terms taken together are statistically non-zero. (F=13.76 with 2&197 DF; P<.0001).

Choosing the order in a polynomial regression

- Is it necessary to include a K -th order term in addition to a $(K-1)$ -th order term?
- Should you stop at

$$\mu_{Y|x} = \beta_0 + \beta_1 x + \beta_2 (x - \bar{x})^2 + \dots + \beta_{K-1} (x - \bar{x})^{K-1}$$

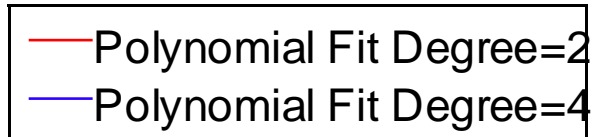
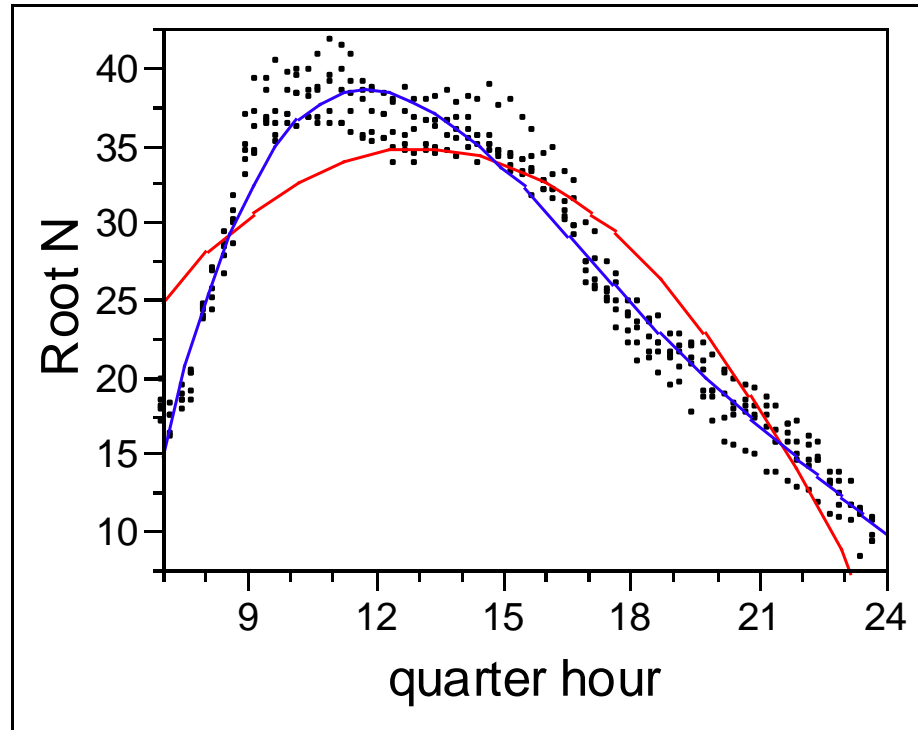
or go on to

$$\mu_{Y|x} = \beta_0 + \beta_1 x + \beta_2 (x - \bar{x})^2 + \dots + \beta_K (x - \bar{x})^K \text{ ?}$$

- Choose a K so that the **Parameter Estimates** test of the coefficient β_K still rejects H_0 (at, say level 0.05).
- And the improvement in R^2 is numerically noticeable from the $(K-1)$ -th order model to the K -th order one.

Telephone Data

Graphical Comparison of Quadratic and Quartic Polynomial Fits



The 4-th Degree Polynomial Fits Better!

Polynomial Fit Degree=2

Summary of Fit

RSquare	0.822
Root Mean Square Error	3.889

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	52.32	0.7327	71.40	<.0001
quarter hour	-1.223	0.0430	-28.47	<.0001
(quarter hour-15.375)^2	-0.267	0.00980	-27.30	<.0001

Polynomial Fit Degree=2

Summary of Fit

RSquare	0.957
Root Mean Square Error	1.924

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	74.90	0.841	89.07	<.0001
quarter hour	-2.752	0.0532	-51.73	<.0001
(quarter hour-15.375)^2	-0.1413	0.0170	-8.32	<.0001
(quarter hour-15.375)^3	0.0353	0.00113	31.34	<.0001
(quarter hour-15.375)^4	-0.00204	0.000263	-7.75	<.0001

Confidence Intervals for the Mean and for Individual Predictions

- Statistical software (JMP) can find CIs for the conditional mean $\mu_{Y|x}$, and for the value of an individual having a given value of x .

This is because software can find such CIs in multiple regression problems, and polynomial regressions are just special kinds of multiple regressions.

Sample Questions:

1. Based on our data find a 95% CI for the mean wage of a 40 year-old male.
2. Fred is a 40 year old male. Based on our data find a 95% CI for his wage.

1. Based on our data find a 95% CI for the mean wage of a 40 year-old male. ANS: Do this in two parts. ---

Our polynomial regression predicts log wage. So, the **Fit Model** platform in JMP will provide a 95% CI for the mean log wage.

To use **Fit Model** for this you need to create a variable $(\text{age})^2$ with the appropriate formula. Then perform a multiple regression using age and $(\text{age})^2$ as the x -variables.

(Or you could create $(\text{age} - \overline{\text{age}})^2$, and etc.; you'll get the same least squares curve and the same CIs either way.)

log wage vs. age and age²

log wage vs. age and (age-38.22)²

<p>Summary of Fit RSquare 0.122561 Analysis of Variance</p> <table border="1"> <thead> <tr> <th>Source</th> <th>DF</th> <th>SSquares</th> <th>MSquare</th> <th>F Ratio</th> <th></th> </tr> </thead> <tbody> <tr> <td>Model</td> <td>2</td> <td>8.646713</td> <td>4.32336</td> <td>13.7585</td> <td></td> </tr> <tr> <td>Error</td> <td>197</td> <td>61.903775</td> <td>0.31423</td> <td>Prob > F</td> <td></td> </tr> <tr> <td>C. Total</td> <td>199</td> <td>70.550488</td> <td></td> <td></td> <td><.0001</td> </tr> </tbody> </table> <p>Parameter Estimates</p> <table border="1"> <thead> <tr> <th>Term</th> <th>Estimate</th> <th>Std Error</th> <th>t Ratio</th> <th>Prob> t </th> </tr> </thead> <tbody> <tr> <td>Intercept</td> <td>4.2530088</td> <td>0.41556</td> <td>10.23</td> <td><.0001</td> </tr> <tr> <td>age</td> <td>0.1001411</td> <td>0.021123</td> <td>4.74</td> <td><.0001</td> </tr> <tr> <td>age²</td> <td>-0.001089</td> <td>0.000251</td> <td>-4.33</td> <td><.0001</td> </tr> </tbody> </table>						Source	DF	SSquares	MSquare	F Ratio		Model	2	8.646713	4.32336	13.7585		Error	197	61.903775	0.31423	Prob > F		C. Total	199	70.550488			<.0001	Term	Estimate	Std Error	t Ratio	Prob> t	Intercept	4.2530088	0.41556	10.23	<.0001	age	0.1001411	0.021123	4.74	<.0001	age ²	-0.001089	0.000251	-4.33	<.0001	<p>Summary of Fit RSquare 0.122561 Analysis of Variance</p> <table border="1"> <thead> <tr> <th>Source</th> <th>DF</th> <th>SSquares</th> <th>MSquare</th> <th>F Ratio</th> <th></th> </tr> </thead> <tbody> <tr> <td>Model</td> <td>2</td> <td>8.646713</td> <td>4.32336</td> <td>13.7585</td> <td></td> </tr> <tr> <td>Error</td> <td>197</td> <td>61.903775</td> <td>0.31423</td> <td>Prob > F</td> <td></td> </tr> <tr> <td>C. Total</td> <td>199</td> <td>70.550488</td> <td></td> <td></td> <td><.0001</td> </tr> </tbody> </table> <p>Parameter Estimates</p> <table border="1"> <thead> <tr> <th>Term</th> <th>Estimate</th> <th>Std Error</th> <th>t Ratio</th> <th>Prob> t </th> </tr> </thead> <tbody> <tr> <td>Intercept</td> <td>5.8431841</td> <td>0.134721</td> <td>43.37</td> <td><.0001</td> </tr> <tr> <td>age</td> <td>0.0169294</td> <td>0.003688</td> <td>4.59</td> <td><.0001</td> </tr> <tr> <td>(age-38.22)²</td> <td>-0.001089</td> <td>0.000251</td> <td>-4.33</td> <td><.0001</td> </tr> </tbody> </table>						Source	DF	SSquares	MSquare	F Ratio		Model	2	8.646713	4.32336	13.7585		Error	197	61.903775	0.31423	Prob > F		C. Total	199	70.550488			<.0001	Term	Estimate	Std Error	t Ratio	Prob> t	Intercept	5.8431841	0.134721	43.37	<.0001	age	0.0169294	0.003688	4.59	<.0001	(age-38.22) ²	-0.001089	0.000251	-4.33	<.0001
Source	DF	SSquares	MSquare	F Ratio																																																																																															
Model	2	8.646713	4.32336	13.7585																																																																																															
Error	197	61.903775	0.31423	Prob > F																																																																																															
C. Total	199	70.550488			<.0001																																																																																														
Term	Estimate	Std Error	t Ratio	Prob> t																																																																																															
Intercept	4.2530088	0.41556	10.23	<.0001																																																																																															
age	0.1001411	0.021123	4.74	<.0001																																																																																															
age ²	-0.001089	0.000251	-4.33	<.0001																																																																																															
Source	DF	SSquares	MSquare	F Ratio																																																																																															
Model	2	8.646713	4.32336	13.7585																																																																																															
Error	197	61.903775	0.31423	Prob > F																																																																																															
C. Total	199	70.550488			<.0001																																																																																														
Term	Estimate	Std Error	t Ratio	Prob> t																																																																																															
Intercept	5.8431841	0.134721	43.37	<.0001																																																																																															
age	0.0169294	0.003688	4.59	<.0001																																																																																															
(age-38.22) ²	-0.001089	0.000251	-4.33	<.0001																																																																																															

The least squared equation from both outputs:

$$\log(wage) = 4.25 + .1 \times age - .001 \times age^2$$

When age=40, the estimated $\log(wage) = 4.25 + .1 \times 40 - .001 \times 40^2 = 6.52$

a) Use JMP output we have:

The CI bounds appear in new columns of the data table. They are

$$95\% \text{ CI for } \mu_{Y|x=40} : (6.407, 6.627)$$

Note also that the estimate for $\mu_{Y|x=40}$ is $\hat{Y} = 6.517 =$ the center of this interval.

b) Convert this interval to an interval for wage, rather than an interval for log wage; So, we get

$$95\% \text{ CI for Wage: } (e^{6.407}, e^{6.627}) = (606, 755).$$

PS. The above procedure is standard practice, but a 2003 Wharton PhD thesis by H. Shen shows that there's a better way to do this. The probability that the mean is in the above CI won't quite be as much as the advertised 95%. However, the Individual CI, below, does have truth in its advertising, and doesn't need modification.

2. Based on our data find a 95% CI for the Fred's wage, where Fred is an individual 40 year-old male.

ANS: Similar to the previous. --

a) The individual CI from JMP for the value of Fred's log wage is
Individual CI for log wage at age = 40: (5.41, 7.63).

b) This yields a CI for Fred's wage as
95% CI for Fred: $(e^{5.41}, e^{7.63}) = (224, 2059)$.

Not surprisingly, this is a very wide interval. We can't say much about Fred's wage if all we know about him is that he's a 40 year-old male. (Gathering more data than the 200 observations we have in our data set might help a very small amount, but it wouldn't help very much