Modeling Nonlinear Relationships

Piecewise Linear Regression Models:

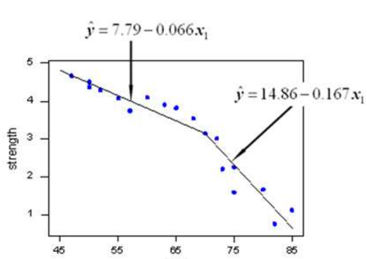
- For the Reynolds data, as an alternative to a quadratic regression model:
  - Recognize that up to a certain point of Months Employed
    - the relationship between Months Employed and Sales appears to be positive and linear.
  - After this point,
    - the relationship between Months Employed and Sales appears to be negative and linear

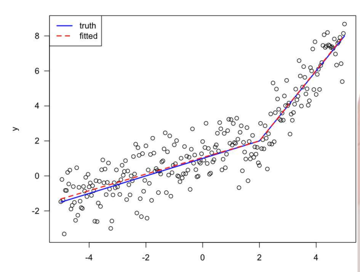
- **Piecewise linear regression model:**
  - This model will allow us to fit these relationships as two linear regressions
    - joined at the value of Months where the relationship between Months Employed and Sales changes.

Modeling Nonlinear Relationships

Piecewise Linear Regression Models (cont.):

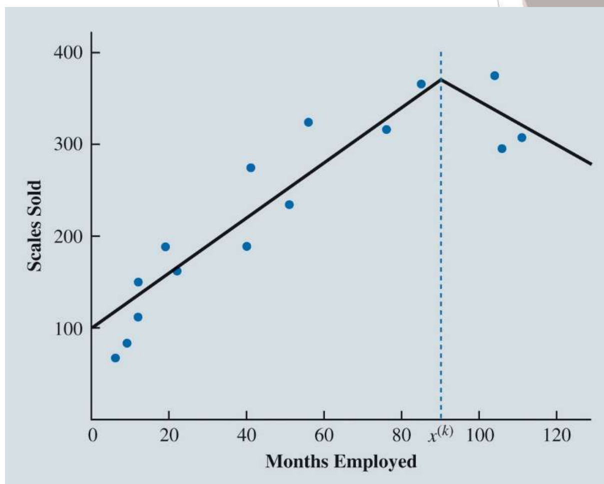
- Knot:
  - The value of the independent variable at which the relationship between dependent variable and independent variable changes;
  - also called *breakpoint*.





Modeling Nonlinear Relationships

Figure 7.32: Possible Position of Knot $x^{(k)}$



Modeling Nonlinear Relationships

Piecewise Linear Regression Models (cont.):

- Define a dummy variable:

$$x_k = \begin{cases} 0 \text{ if } x_1 \leq x^{(k)} \\ 1 \text{ if } x_1 > x^{(k)} \end{cases}$$

x_1 = Months.

$x^{(k)}$ = value of the knot (90 months for the Reynolds example).

x_k = the knot dummy variable.

- Then fit the following estimated regression equation:

$$\hat{y} = b_0 + b_1 x_1 + b_2(x_1 - x^{(k)})x_k$$

Modeling Nonlinear Relationships

Data and Excel Output for the Reynolds Piecewise Linear Regression Model

- Piecewise Regression

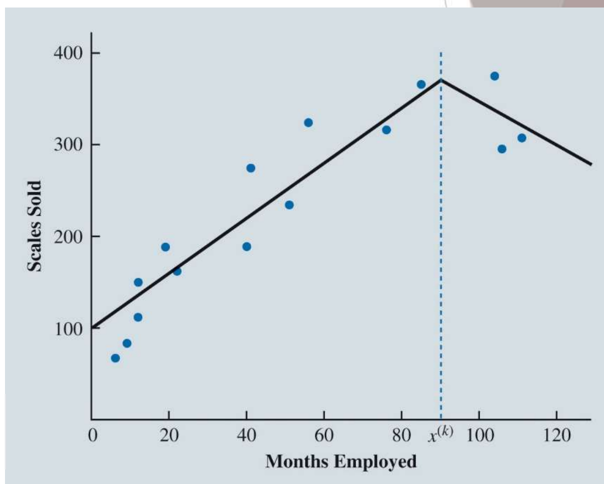
$$\hat{y} = 87.2172 + 3.4094x_1 - 7.8726(x_1 - 90)x_k$$

- When X1 < 90

$$\hat{y} = 87.2172 + 3.4094x_1$$

- When X1 > 90

$$\begin{aligned}\hat{y} &= 87.2172 + 3.4094x_1 - 7.8726(x_1 - 90)\\ &= 87.2172 - 7.8726(-90) + (3.4094 - 7.8726)x_1 = 795.7512 - 4.4632x_1\end{aligned}$$



Modeling Nonlinear Relationships

Interaction Between Independent Variables:

- **Interaction:**
  - This occurs when the relationship between the dependent variable and one independent variable is different at various values of a second independent variable.
  - Capture whether the relationship between y and x1 changes because of another x2
- The estimated multiple linear regression equation is given as:

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_1x_2$$

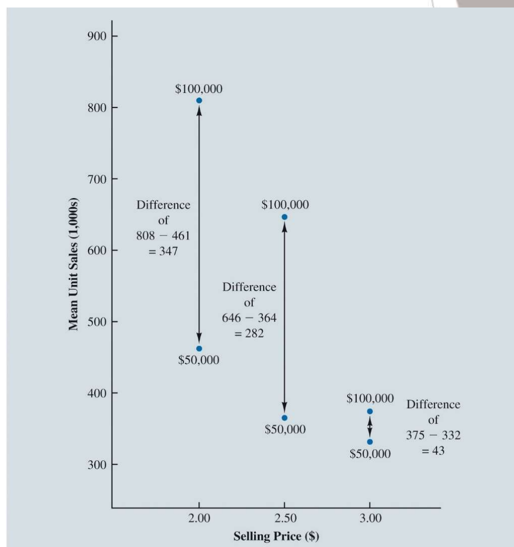
Modeling Nonlinear Relationships

Figure 7.34: Mean Unit Sales (1,000s) as a Function of Selling Price and Advertising Expenditures

Y = Sales

X1 = price of shampoo

X2 = Advertising expenditure



Modeling Nonlinear Relationships

Excel Output for the Tyler Personal Care Linear Regression Model with Interaction

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.988993815							
5	R Square	0.978108766							
6	Adjusted R Square	0.974825081							
7	Standard Error	28.17386496							
8	Observations	24							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	3	709316	236438.6667	297.8692	9.25881E-17			
13	Residual	20	15875	793.7666667					
14	Total	23	5191.3333						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
17	Intercept	–275.8333333	112.8421033	–2.444418575	0.023898351	–511.2178361	–40.44883053	–596.9074508	45.24078413
18	Price	175	44.54679188	3.928453489	0.0008316	82.07702045	267.9229796	48.24924412	301.7507559
19	Advertising Expenditure ($1,000s)	19.68	1.42735225	13.78776683	1.1263E-11	16.70259538	22.65740462	15.61869796	23.74130204
20	Price*Advertising	–6.08	0.563477299	–10.79014187	8.67721E-10	–7.255393049	–4.904606951	–7.683284335	–4.476715665

Sales after a $1 increase in Price

The relationship between Price and Sales is different at various values of Advertising

$$\text{Sales} = -275.8333 + 175\ \text{Price} + 19.68\ \text{Advertising} - 6.08\ \text{Price} * \text{Advertising}$$

$$\text{Sales} = -275.8333 + 175(2) + 19.68(50) - 6.08(2)(50) = 450.1667, \text{ or } 450{,}167 \text{ units}$$

$$\text{Sales} = -275.8333 + 175(3) + 19.68(50) - 6.08(3)(50) = 321.1667, \text{ or } 321{,}167 \text{ units}$$

$$\text{Sales} = -275.8333 + 175(2) + 19.68(100) - 6.08(2)(100) = 826.1667, \text{ or } 826{,}167 \text{ units}$$

$$\text{Sales} = -275.8333 + 175(3) + 19.68(100) - 6.08(3)(100) = 393.1667, \text{ or } 393{,}167 \text{ units}$$

Sales after a $1000 increase in Ads

The relationship between Advertising Expenditure and Sales is different at various values of Price

$$\text{Sales After \$1K Advertising Increase} = -275.8333 + 175\ \text{Price} + 19.68\ (\text{Advertising} + 1) - 6.08\ \text{Price} * (\text{Advertising} + 1)$$

$$\text{Sales After \$1K Advertising Increase} - \text{Sales Before \$1K Advertising Increase} = 19.68 - 6.08\ \text{Price}$$

$$\text{Sales} = -275.8333 + 175(2) + 19.68(50) - 6.08(2)(50) = 450.1667, \text{ or } 450{,}167 \text{ units}$$

$$\text{Sales} = -275.8333 + 175(2) + 19.68(100) - 6.08(2)(100) = 826.1667, \text{ or } 826{,}167 \text{ units}$$

$$\text{Sales} = -275.8333 + 175(3) + 19.68(50) - 6.08(3)(50) = 321.1667, \text{ or } 321{,}167 \text{ units}$$

$$\text{Sales} = -275.8333 + 175(3) + 19.68(100) - 6.08(3)(100) = 393.1667, \text{ or } 393{,}167 \text{ units}$$

Model Fitting

Variable Selection Procedures

Overfitting

Model Fitting

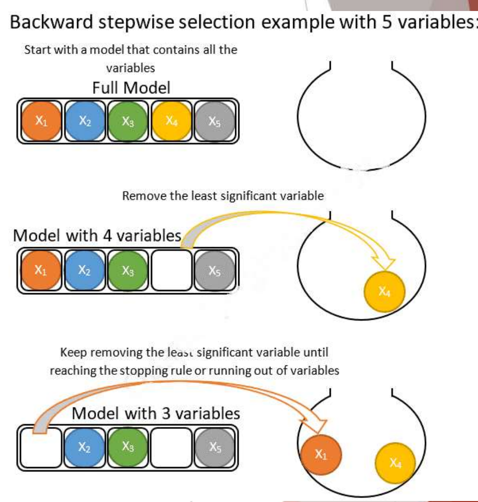
Variable Selection Procedures:

- Special procedures are sometimes employed to select the independent variables to include in the regression model.
  - Iterative procedures: At each step of the procedure, a single independent variable is added or removed and the new model is evaluated. Iterative procedures include:
    - **Backward elimination.**
    - **Forward selection.**
    - **Stepwise selection.**
  - **Best subsets** procedure: Evaluates regression models involving different subsets of the independent variables.

Model Fitting

Variable Selection Procedures (cont.):

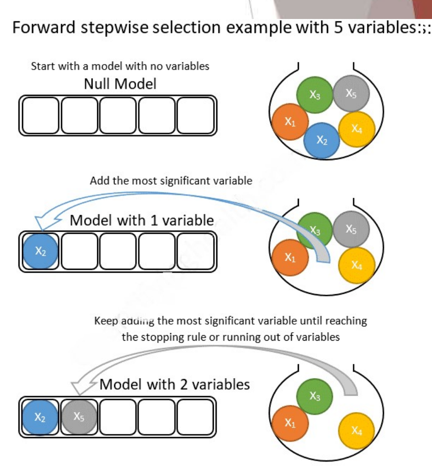
- **Backward elimination** procedure:
  - Begins with the regression model that includes all of the independent variables under consideration.
  - At each step, backward elimination considers the removal of an independent variable according to some criterion (Significance)
  - Stops when all independent variables in the model are significant at a specified level of significance.



Model Fitting

Variable Selection Procedures (cont.):

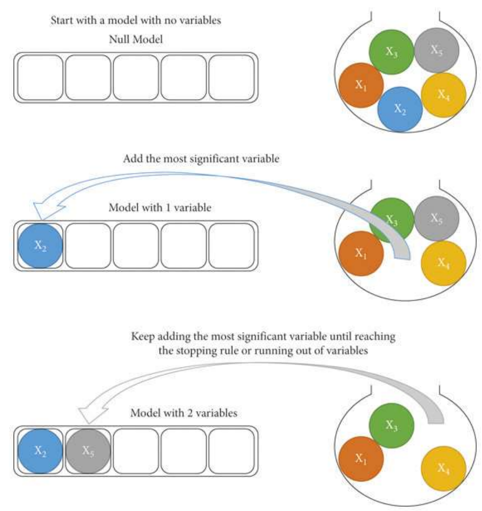
- **Forward selection** procedure:
  - Begins with none of the independent variables under consideration included in the regression model.
  - At each step, forward selection considers the addition of an independent variable according to some criterion (Significance).
  - Stops when there are no independent variables not currently in the model that meet the criterion for being added to the regression model.



Model Fitting

Variable Selection Procedures (cont.):

- **Stepwise selection** procedure:
  - Begins with none of the independent variables under consideration included in the regression model.
  - The analyst establishes both a criterion for allowing independent variables to enter the model and a criterion for allowing independent variables to remain in the model.
  - To initiate the procedure, the most significant independent variable is added to the empty model if its level of significance satisfies the entering threshold.



Model Fitting

Variable Selection Procedures (cont.):

- Stepwise selection procedure (cont.):
  - Each subsequent step involves two intermediate steps:
    - First, the remaining independent variables not in the current model are evaluated, and the most significant one is added to the model.
    - Then the independent variables in the current model are evaluated, and the least significant one is removed.
  - Stops when no independent variables not currently in the model have a level of significance for remaining in the regression model.

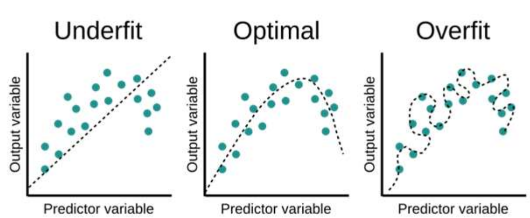
Model Fitting

Variable Selection Procedures (cont.):

- **Best subsets** procedure:
  - Estimate a regression for every combination of independent variables
  - Compare and evaluate the entire collection of regression models

Model Fitting

- **Overfitting**
  - Generally results from creating an overly complex model to explain idiosyncrasies in the sample data.
  - Typically includes independent variables that do not have meaningful relationships with the dependent variable.
- If a model is overfit to the sample data
  - it will perform better on the sample data used to fit the model than it will on other data from the population.
- An overfit model
  - can be misleading about its predictive capability and its interpretation.



Model Fitting

- **How does one avoid overfitting a model?**
  - Use only independent variables that you expect to have real and meaningful relationships with the dependent variable.
  - Use complex models, such as quadratic models and piecewise linear regression models, only when reasonable
  - Do not let software dictate your model;
  - Use iterative modeling procedures, such as the stepwise and best-subsets procedures, only for guidance

Model Fitting

- **How does one avoid overfitting a model? (cont.):**
  - **Cross-Validate**
    - Assess your model on data other than the sample data (if you have it)
    - One possible ways to execute cross-validation is the holdout method.
  - **Holdout method:** The sample data are randomly divided into mutually exclusive and collectively exhaustive training and validation sets.
    - **Training set:**
      - The data set used to build the candidate models that appear to make practical sense.
    - **Validation set:**
      - The set of data used to compare model performances and ultimately select a model for predicting values of the dependent variable.

Big Data and Regression

Inference and Very Large Samples

Model Selection

Big Data and Regression

Inference and Very Large Samples:

- Virtually all relationships between independent variables and the dependent variable will be statistically significant if the sample is sufficiently large.

- That is, if the sample size is very large, there will be little difference in the b_j values generated by different random samples.

Big Data and Regression

Figure 7.36: Excel Regression Output for Credit Card Company Example

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.602663145							
5	R Square	0.363202867							
6	Adjusted R Square	0.362565219							
7	Standard Error	4834.449957							
8	Observations	3000							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	3	39937797910	13312599303	569.5983495	6.5207E-293			
13	Residual	2996	70022231537	23371906.39					
14	Total	2999	1.0996E+11						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
17	Intercept	2119.600282	333.0922952	6.363402314	2.27497E-10	1466.487528	2772.713036	1261.064442	2978.136122
18	Annual Income ($1000)	121.3384676	3.165148859	38.33578544	5.4905E-262	115.1323826	127.5445525	113.1803871	129.496548
19	Household Size	528.0996852	42.84154037	12.32681366	4.29401E-34	444.097873	612.1014973	417.6768433	638.522527
20	Years of Post-High School Education	-535.3593516	58.5960221	-9.136445316	1.15792E-19	-650.2518601	-420.4668432	-686.3889184	-384.3297849

Big Data and Regression

Table 7.4: Regression Parameter Estimates and the Corresponding *p* values for 10 Multiple Regression Models, Each Estimated on 50 Observations from the *LargeCredit* Data

Observations	b_0	*p* value	b_1	*p* value	b_2	*p* value	b_3	*p* value
1–50	−805.152	0.7814	154.488	1.45E-06	234.664	0.5489	207.828	0.6721
5–100	894.407	0.6796	125.343	2.23E-07	822.675	0.0070	−355.585	0.3553
101–150	−2,191.590	0.4869	155.187	3.56E-07	674.961	0.0501	−25.309	0.9560
151–200	2,294.023	0.3445	114.734	1.26E-04	297.011	0.3700	−537.063	0.2205
201–250	8,994.040	0.0289	103.378	6.89E-04	−489.932	0.2270	−375.601	0.5261
251–300	7,265.471	0.0234	73.207	1.02E-02	−77.874	0.8409	−405.195	0.4060
301–350	2,147.906	0.5236	117.500	1.88E-04	390.447	0.3053	−374.799	0.4696
351–400	−504.532	0.8380	118.926	8.54E-07	798.499	0.0112	45.259	0.9209
401–450	1,587.067	0.5123	81.532	5.06E-04	1,267.041	0.0004	−891.118	0.0359
451–500	−315.945	0.9048	148.860	1.07E-05	1,000.243	0.0053	−974.791	0.0420
Mean	1,936.567		119.316		491.773		−368.637	

Big Data and Regression

Figure 7.37: Excel Regression Output for Credit Card Company Example after Adding Number of Hours per Week Spent Watching Television

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.603724482							
5	R Square	0.36448325							
6	Adjusted R Square	0.36363448							
7	Standard Error	4830.393498							
8	Observations	3000							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	4	40078588918	10019647230	429.4250838	8.3277E-293			
13	Residual	2995	69881440529	23332701.35					
14	Total	2999	1.0996E+11						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
17	Intercept	1712.552073	371.7837807	4.606311953	4.26973E-06	983.5746542	2441.529492	754.2898349	2670.814311
18	Annual Income ($1000)	121.6120724	3.164453912	38.43066631	4.943E-263	115.4073492	127.8167955	113.4557814	129.7683633
19	Household Size	531.213362	42.82435656	12.40446803	1.71315E-34	447.2452317	615.1814922	420.8347874	641.5919365
20	Years of Post-High School Education	-539.8345703	58.57519443	-9.216095235	5.64208E-20	-654.6862563	-424.9828843	-690.8104864	-388.8586541
21	Hours Per Week Watching Television	12.55178379	5.109759992	2.456433142	0.014088759	2.532789303	22.57077828	-0.618478873	25.72204645

Big Data and Regression

Model Selection:

- When dealing with large samples, it is often more difficult to discern the most appropriate model.
- For explanatory purposes, the practical significance of the estimated regression coefficients should be considered when interpreting the model and considering which variables to keep in the model.
- For future predictions, the independent variables included in the regression model should be based on the predictive accuracy on observations that have not been used to train the model.

Big Data and Regression

Figure 7.38: Predictive Accuracy on *LargeCredit* Validation Set

	A	B	C	D	E	F	G	H	I	J	K
1							Model A (3 Variable)			Model B (4 Variable)	
2	Account Number	Annual Income ($1000)	Household Size	Years of Post-High School Education	Hours Per Week Watching Television	Annual Charges ($)	Prediction	Squared Error		Prediction	Squared Error
3	18572870	50.2	5.0	1.0	4.0	5,472.51	10,315.93	23,458,721		9,983.92	20,352,797
4	10135558	39.6	2.0	4.0	15.0	3,968.42	5,839.37	3500437.294		5,619.76	2726908.398
5	23467852	88.8	4.0	1.0	19.0	11,382.63	14,471.50	9541090.638		14,335.21	8717710.157
6	2221007	101.2	6.0	2.0	54.0	16,827.83	16,496.93	109493.0845		16,805.10	516.6005887
7	23024579	52.0	5.0	2.0	19.0	13,175.27	9,998.98	10088816.14		9,851.26	11049033.2
8	5534868	100.8	5.0	4.0	50.0	20,292.73	14,849.58	29627894.64		15,095.37	27012585.44
9	19704869	70.6	3.0	0.0	49.0	6,230.89	12,270.40	36475622.43		12,507.04	39390082.33
10	9388137	88.9	8.0	2.0	41.0	18,914.62	16,060.67	8145037.3		16,208.53	7322943.679
11	23883625	89.4	5.0	2.0	29.0	14,362.00	14,537.04	30638.65325		14,525.07	26592.06635
1991	6776616	87.6	3.0	2.0	59.0	20,541.21	13,262.43	52980632.57		13,620.30	47899053.37
1992	8695442	47.3	8.0	1.0	10.0	17,011.33	11,548.35	29844173.12		11,300.19	32617082.88
1993	5888985	82.4	4.0	1.0	48.0	9,416.69	13,694.93	18303332.35		13,920.89	20287829.66
1994	12243467	43.2	5.0	1.0	16.0	3,101.00	9,466.56	40520368.82		9,283.25	38220269.2
1995	28297658	49.9	5.0	0.0	48.0	14,538.99	10,814.89	13868933.92		11,039.55	12246101.91
1996	4605783	36.7	3.0	2.0	19.0	12,620.39	7,086.30	30626125.63		6,928.17	32401368.92
1997	21430617	54.9	1.0	5.0	27.0	3,755.45	6,632.39	8276755.445		6,559.99	7865464.344
1998	3080483	84.4	4.0	5.0	23.0	13,018.42	11,796.17	1493897.685		11,690.98	1762090.043
1999	8089356	41.6	2.0	4.0	14.0	8,740.70	6,082.04	7068459.72		5,850.43	8353673.976
2000	14223252	51.0	7.0	5.0	4.0	0.00	9,327.76	87007165.68		8,984.30	80717567.08
2001	8048637	39.0	7.0	1.0	5.0	360.73	10,013.14	93168998.76		9,696.84	87162964.44
2002	27638369	39.0	5.0	2.0	19.0	1,554.11	8,421.58	47162147.49		8,270.30	45107267.97
2003											
2004							SSE:	47,392,009,111			47,409,404,281

Prediction with Regression

Prediction with Regression

- In addition to the point estimate, there are two types of interval estimates associated with the regression equation:
  - A confidence interval is an interval estimate of the mean *y* value given values of the independent variables.

$$\hat{y} \pm t_{\alpha/2} s_{\hat{y}} \qquad \textbf{(7.23)}$$

  - A **prediction interval** is an interval estimate of an individual y value given values of the independent variables.

$$\hat{y} \pm t_{\alpha/2} \sqrt{s_{\hat{y}}^2 + \frac{SSE}{n - q - 1}} \qquad \textbf{(7.24)}$$

Prediction with Regression

Table 7.5: Predicted Values and 95% Confidence Intervals and Prediction Intervals for 10 New Butler Trucking Routes

Assignment	Miles	Deliveries	Predicted Value	95% CI Half-Width(+/−)	95% PI Half-Width(+/−)
301	105	3	9.25	0.193	1.645
302	60	4	6.92	0.112	1.637
303	95	5	9.96	0.173	1.642
304	100	1	7.54	0.225	1.649
305	40	3	4.88	0.177	1.643
306	80	3	7.57	0.108	1.637
307	65	4	7.25	0.103	1.637
308	55	3	5.89	0.124	1.638
309	95	2	7.89	0.175	1.643
310	95	3	8.58	0.154	1.641