

Modeling Nonlinear Relationships

**Piecewise Linear Regression Models:**

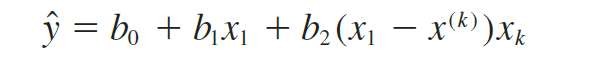
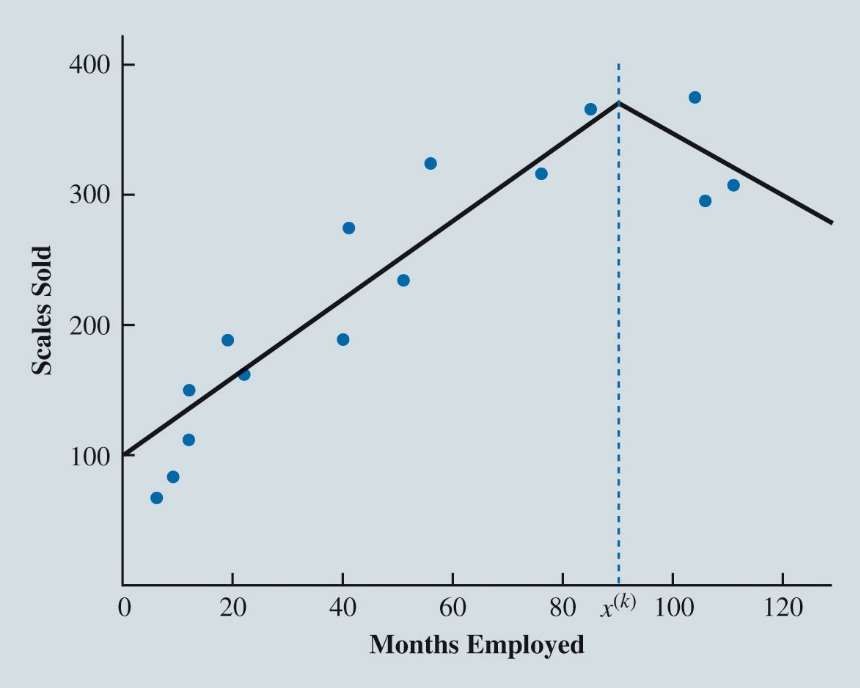
* For the Reynolds data, as an alternative to a quadratic regression model:
  + Recognize that up to a certain point of Months Employed
    - the relationship between Months Employed and Sales appears to be positive and linear.
  + After this point,
    - the relationship between Months Employed and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

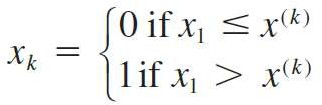
Figure 7.32: Possible Position of Knot *x*(*k*)



Modeling Nonlinear Relationships

**Piecewise Linear Regression Models (cont.):**

* Define a dummy variable:



*x*1 = Months.

*x*(*k*) = value of the knot (90 months for the Reynolds example).

*xk* = the knot dummy variable.

* Then fit the following estimated regression equation:





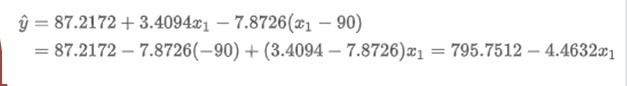
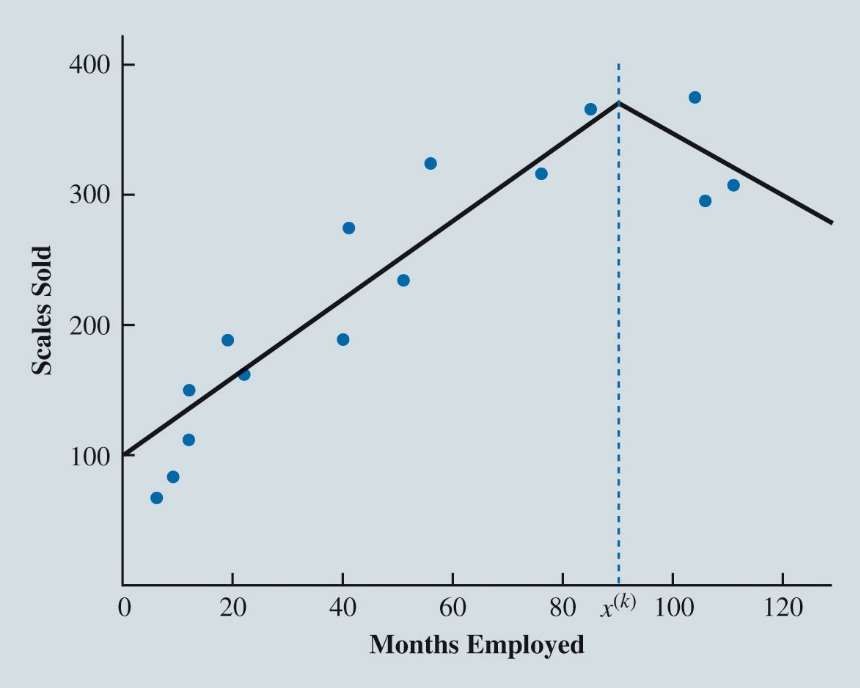
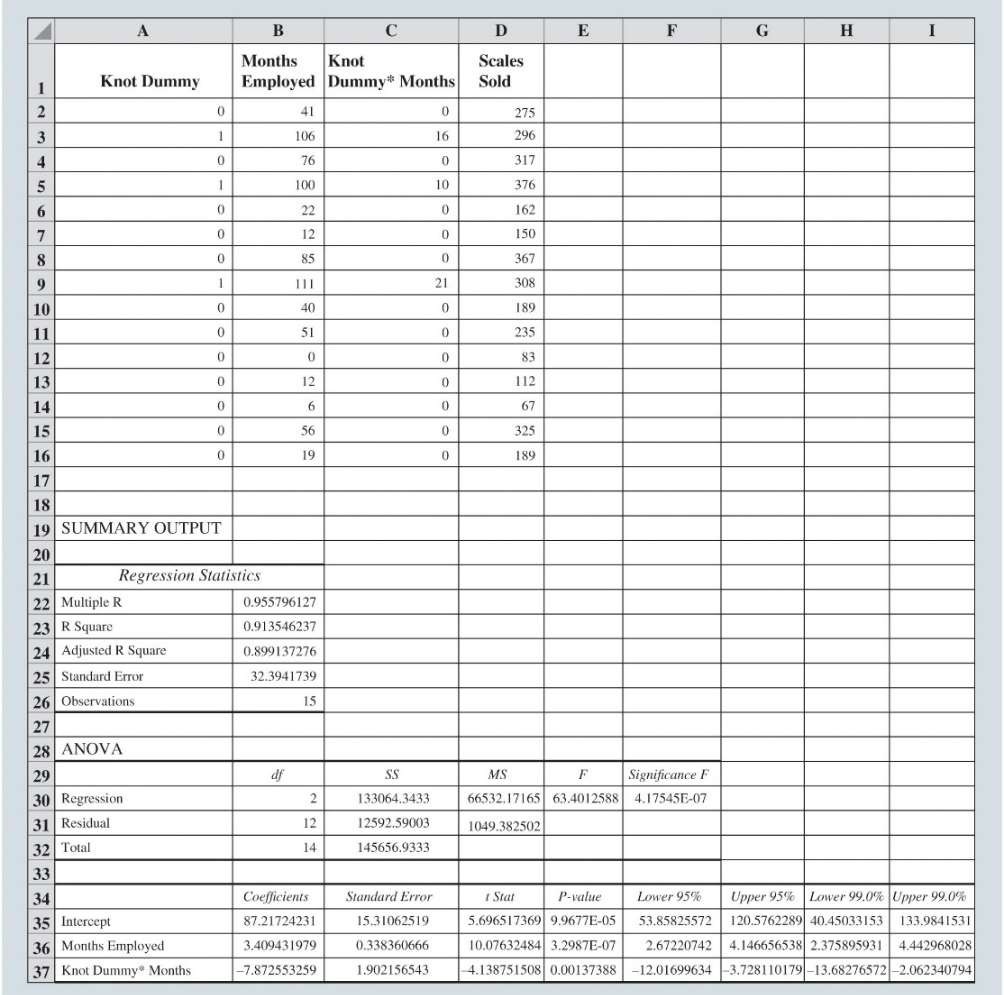
Modeling Nonlinear Relationships

Data and Excel Output for the Reynolds Piecewise Linear Regression Model

* Piecewise Regression



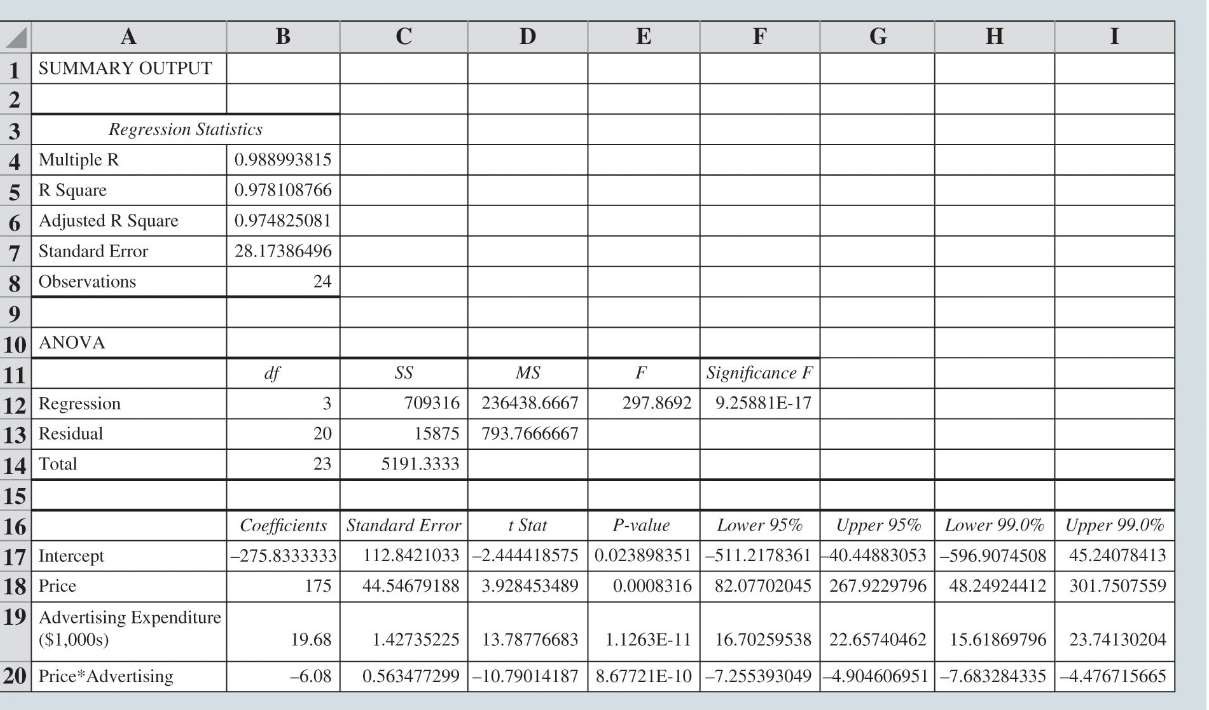
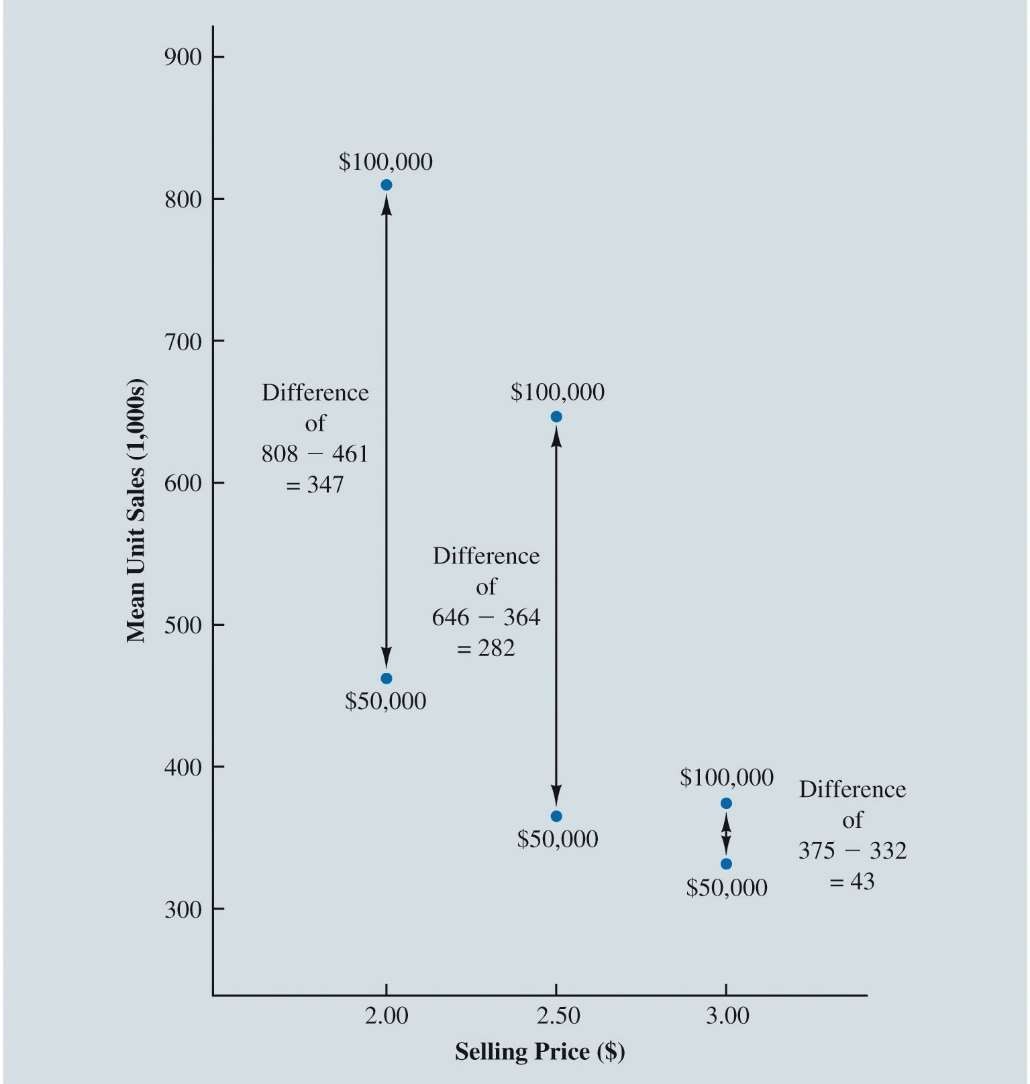
* When X1 < 90
* When X1 > 90



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:



Modeling Nonlinear Relationships

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

Y = Sales

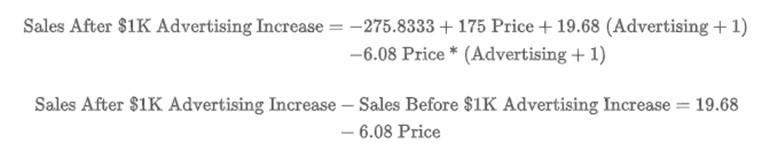
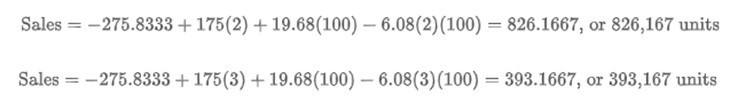
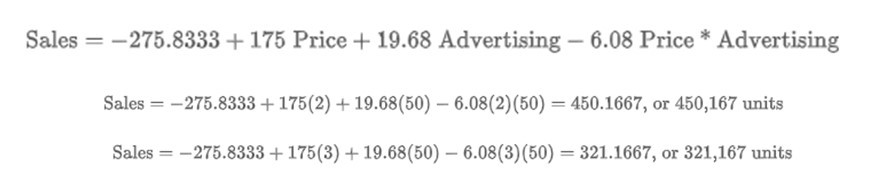
X1 = price of shampoo

X2 = Advertising expenditure



Modeling Nonlinear Relationships

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



Sales after a $1 increase in Price

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



Sales after a $1000 increase in Ads

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



Model Fitting

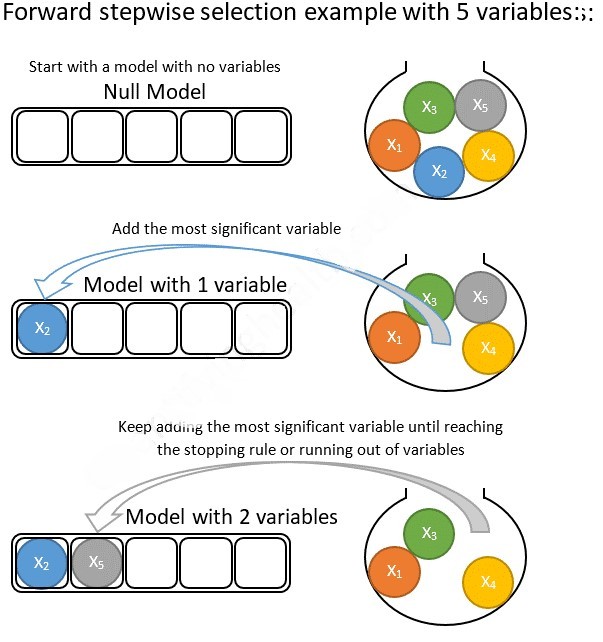
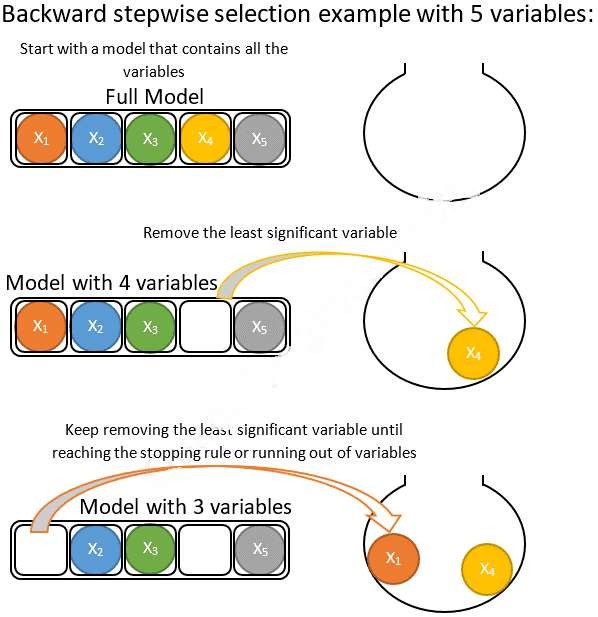
Variable Selection Procedures Overfitting



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.):**

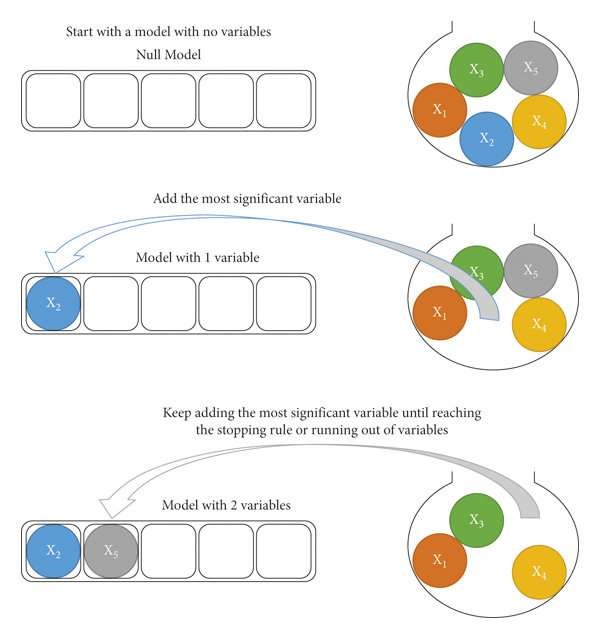
* **Backward elimination** procedure:
  + Begins with the regression model that includes all of the independent variables under consideration.
  + At each step, backward elimination considers the removal of an independent variable according to some criterion (Significance)
  + Stops when all independent 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.):**

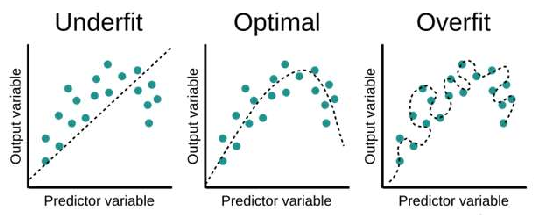
* **Stepwise selection** procedure:
  + Begins with none of the independent variables under consideration included in the regression model.
  + The analyst establishes both a criterion for allowing independent variables to enter the model and a criterion for allowing independent variables to remain in 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

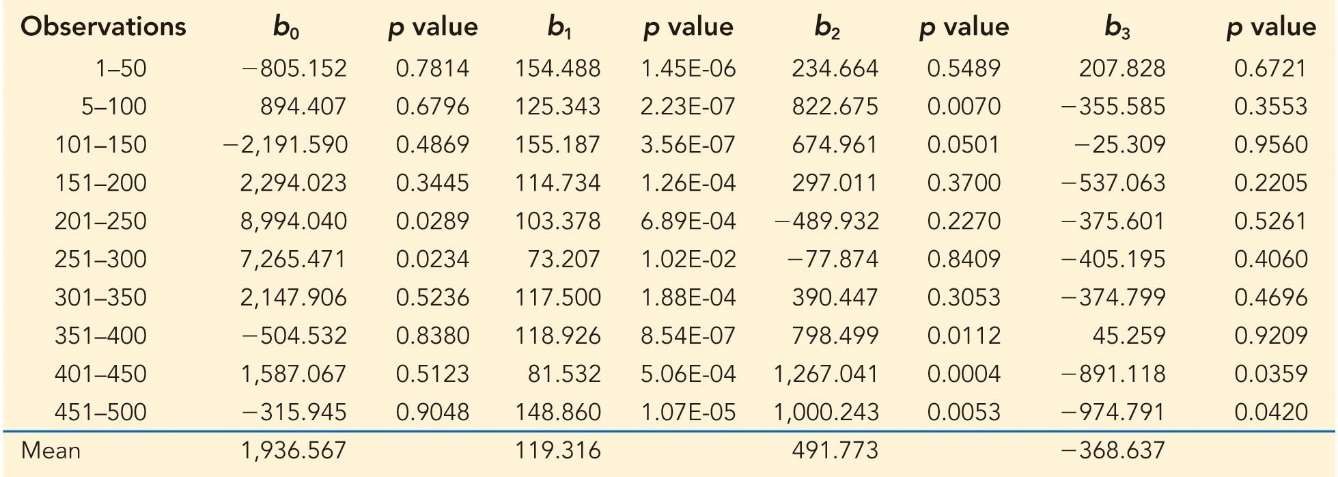
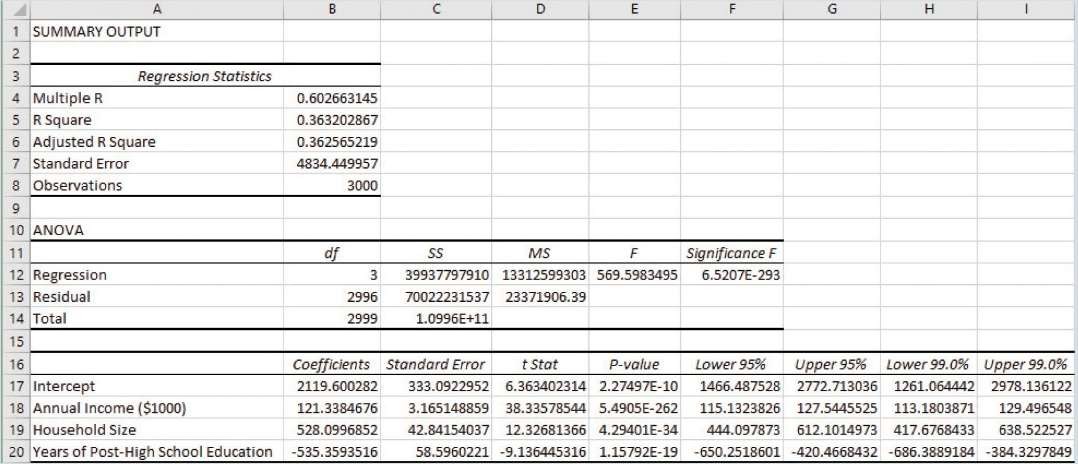


Big Data and Regression

**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

*bj* values generated by different random samples.



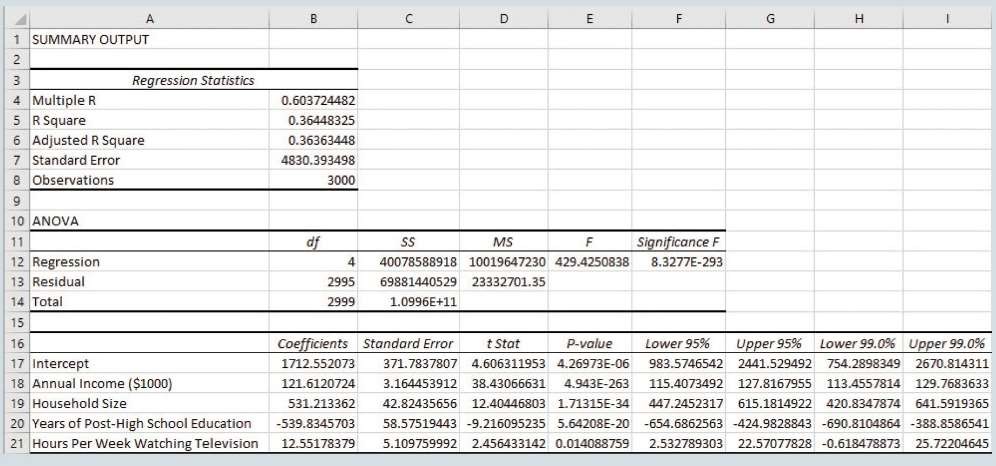
Big Data and Regression

Figure 7.36: Excel Regression Output for Credit Card Company Example



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



Big Data and Regression

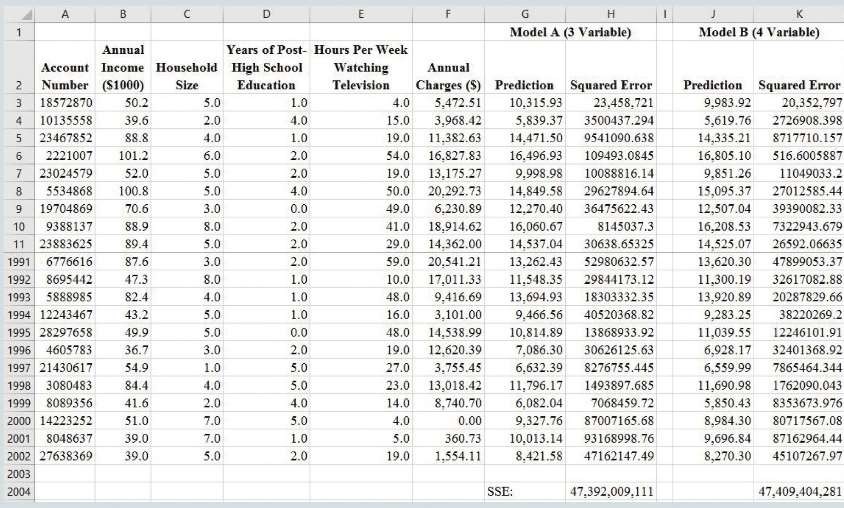
Figure 7.37: Excel Regression Output for Credit Card Company Example after Adding Number of Hours per Week Spent Watching Television



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.

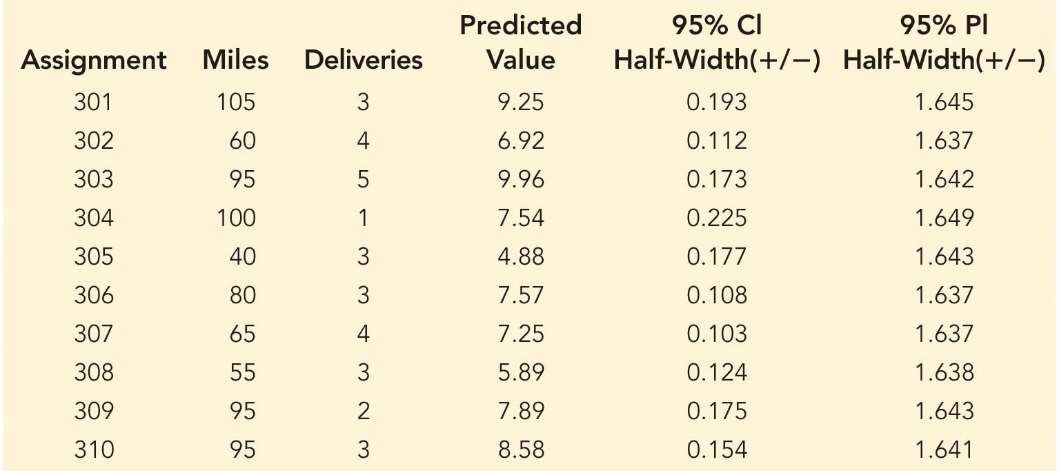


Big Data and Regression

Figure 7.38: Predictive Accuracy on *LargeCredit* Validation Set



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



Prediction with Regression

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