Chapter 8
Assessing Relationships with Correlation and Regression

Section 8.3
Model Fit and Forecast

David Gerbing

The School of Business
Portland State University

8.3a
Modeling Error

Fit of Model to the Data from which It is Estimated

How well does the model *describe* the sample data?

- Best OLS regression model can be found for any scatterplot
  - Best means "best" only relative to all other possible models
  - "Best" may be lousy, with much scatter about the line
- To forecast future values of y with the model, first validate with statistical criteria that the "best" line fits the data
- **Key Concept**: First adequately fit the past data values of y before attempting to predict the future values of y
- Two descriptive statistics describe the fit of the model to the data from which the model was estimated
  - **Absolute criterion**: standard deviation of the residuals, s_e, based on size of errors (residuals) for the regression analysis
  - **Relative criterion**: R squared, R^2, compares size of residuals to overall amount of variability of y

Regression Output: ANOVA

Fit of the model ultimately based on size of residuals

- The indicators of model fit, s_e and R^2, are based directly on the sum of the squared residuals, $\sum e^2$
- Although not usually interpreted per se, find $\sum e^2$ in the Regression output Analysis of Variance table, under the Sum Sq column

```
Analysis of Variance

          df    Sum Sq   Mean Sq   F-value   p-value
      Ht  1  1565.226  1565.226     8.363    0.0201
Residuals 8  1497.249   187.156
```

- For this model and data, $\sum e^2 = 1497.25$, which, in turn, enters into the expressions for both the standard deviation of the residuals, s_e, and R^2

Standard Deviation of the Residuals

Definition: Standard Deviation of the Residuals

Assess absolute fit

- First criterion of goodness of fit directly evaluates amount of scatter, the modeling error, about regression line
  - modeling error, the residuals, represent variation about the regression line
  - Assess fit with the standard deviation of residuals
- **Standard deviation of the residuals** (σ_e or s_e): Square root of average squared residual
- To understand the meaning of s_e, realize that the mean of the residuals is 0 for any one regression analysis
- So the sum of squared residuals is also the sum of squared deviations about the mean of the residuals

$$\sum e_i^2 = \sum (e_i - 0)^2 = \sum (e_i - \bar{e})^2$$

Formula: Standard Deviation of the Residuals

Describe the size of the residuals

- The sum of the squared residuals confounds their size with the number of the squared residuals
- Move from the sum to the mean, in this case, the variance
- Degrees of freedom is the size of the sample minus the number of constrained parameters already estimated from which to calculate the residuals, here b_0 and b_1, so $df = n - 2$
- **Mean Squared Error**: The average squared modeling error, with the average calculated with the degrees of freedom

$$MSE = \frac{\sum e_i^2}{df}$$

- **Standard deviation of e**: $s_e = \sqrt{MSE}$
- s_e illustrates the size of the typical residual, a summary of the size of modeling errors encountered across the data

Ex: $\hat{y} = -193.45 + 5.20X$

Ht	Wt	PredWt	Residual	Residual^2
71.75	182.25	179.76248	2.488	6.188
71.75	168.25	179.76248	-11.512	132.537
75.50	194.00	199.26847	-5.268	27.757
68.25	140.25	161.55688	-21.307	453.983
68.25	156.50	161.55688	-5.057	25.572
69.50	187.75	168.05888	19.691	387.740
70.50	193.50	173.26048	20.240	409.638
71.50	177.25	178.46208	-1.212	1.469
67.00	151.00	155.05488	-4.055	16.442
68.00	166.25	160.25648	5.994	35.922
Sum of Squared Errors (SSE)				1497.249
Mean Squared Error (MSE), SSE/df				187.156
Standard Error of Estimate, sqrt(MSE)				13.681

- $df = 10 - 2 = 8, \quad MSE = \frac{1497.25}{10 - 2} = 187.156$
- and $s_e = \sqrt{MSE} = 13.681$

Meaning of Standard Deviation of the Residuals

How big are the residuals?

- The standard deviation of the residuals, s_e, conveys the size of the typical modeling error, or residual, $e_i = y_i - \hat{y}_i$
  - In this example, $s_e = 13.681$
  - There is no residual of exactly this size, but note that the residuals tend to be around this size in magnitude, and not, for example, .0004 or -115.82
- s_e is a standard deviation, so if the modeling errors are normally distributed, then the size of the standard error can also be gauged according to normal curve probabilities
- About 95% of the values of normally distributed modeling errors, e_i, lie within about ± 2 standard deviations of their mean of 0, for a range of 4 standard deviations

R: `Regression` Output: s_e

Goodness of fit indices

- Always report s_e for the analysis of any regression model

```
Standard deviation of residuals: 13.68
   for 8 degrees of freedom
```

- Assuming a normal distribution of the residuals, then about 95% of residuals will span the range of twice the 95% t-cutoff multiplied by the s_e

```
If normal, the approximate 95% range of residuals about
  each fitted value is 2*t-cutoff*13.6805, with a 95%
  interval t-cutoff of 2.306
95% range of variation: 63.09
```

- As expected from such a small sample size, this range of 63 lbs is much to large for the model to be of practical use

R^2 Fit Statistic

Basis of R^2

Assess relative fit

- R^2 compares amount of scatter for two different regression models, the model of interest with predictor variable X to the null model, a model without X
- **Null Model**: A model of response variable y with no contribution from any other variable X, which is either ...
  - included in model but unrelated to y, or
  - not included in the model
- Fit with null model is the worst case scenario
- In the absence of any information about a predictor variable X, the fitted value for y is the mean of *all* the values of y

  $\hat{y} = m$
- Null model specifies random variation about the mean

SSE and SSY as Error Summaries for Two Models

Compare the error terms

- modeling errors from the regression model
  - The fitted values of the model are the linear conditional means of y, one conditional mean, $\hat{y}$, for each value of X
  - So assess variability about the regression line with

    $SSE = \sum e_i^2 = \sum (y_i - \hat{y})^2$ and then s_e
- modeling errors from the null model
  - The null model is for the analysis of variability about the one unconditional mean of y, the same for all values of X
  - So assess variability about this null line with

    $SSY = \sum (y_i - m)^2$ and then s (i.e., s_y)

Definition: R^2

SSY vs SSE

- R^2 explicitly compares the variation about the conditional means of y that define the regression line for each value of X to the variation about the one unconditional mean of all of y
- Compare the ratio of SSE to SSY, the total sum of squares of y, subtracted from 1 so that a high value indicates good fit

$$R^2 = 1 - \frac{SSE}{SSY} = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - m)^2}$$

- Fundamental question: How much does X in the regression model reduce the error variability from the null model?

- For R these sum of squares are under Analysis of Variance
  - SSE is the Sum of Sq for Residuals
  - SSY is obtained from adding all the sums of squares

Range of R^2: From 0 to 1

Worst Case: X ignored or completely unrelated to y

- Null Model: $\hat{y}_i = m$, so SSE = SSY
- $R^2 = 1 - \frac{SSY}{SSY} = 1 - 1 = 0$

Best Case: X perfectly related to y

- Regression Model: $\hat{y}_i = y_i$, so SSE = 0
- $R^2 = 1 - \frac{SSE}{SSY} = 1 - 0 = 1$

Interpretation: R^2

Following are heuristics, informal "rules of thumb"

- An R^2 of ...
  - 1 indicates perfect fit
  - .6 is usually considered excellent fit
  - .3 is usually considered adequate fit
  - 0 indicates regression model provides no improvement over null model
- The higher the R^2 statistic, in general, the better the quality of the forecasts from the model
- Note, however, that R^2 does not directly address the quality of the forecasts such that a high value of R^2 does not imply quality forecasts
- Direct assessment of forecasting quality is provided in a following section

A More Realistic Version of R^2: R^2_{adj}

The adjusted R^2

- Adding predictor variables to a model reduces R^2 to the extent that if the number of predictors equals the sample size, $R^2 = 1.0$
- R^2_{adj} adjusts for this artificial increase in the fit of the model with too many predictor variables relative to the sample size
- The adjustment is to divide each of the two sums of squares in the definition of R^2 by their corresponding degrees of freedom
- In small samples the drop from R^2 to R^2_{adj} can be large
- In larger samples R^2_{adj} will still be somewhat smaller, but not usually much smaller

R: Regression Output: R^2

Goodness of fit indices

- Always report R^2 and R^2_{adj} for any regression analysis

```
R-squared: 0.511     Adjusted R-squared:  0.450
```

- Here, $R^2 = 0.51$, a value considerable larger than zero
- **Descriptive Result**: *For these data*, the use of Height to explain Weight better accounts for the value of Weight than if Height is not included in the model
- The drop of $R^2 = 0.51$ to $R^2_{adj} = 0.45$ is somewhat large because of the small sample size

R: Regression Output: Hypothesis Test of R^2

Overall assessment of the model

- The fit indices s_e and R^2 are descriptive statistics
- Here consider the inferential analysis of R^2

```
Null hypothesis that population R-squared=0
  F-statistic: 8.363     df: 1 and 8     p-value:  0.020
```

- The p-value is for the test of the null hypothesis that the population $R^2 = 0$, that the inclusion of the predictor variable(s) in the model contributes to a significant reduction in the sum of squared residuals
- Here p-value $= 0.020 < \alpha = 0.05$, so reject the null
- **Interpretation**: Although the residuals can be large, the use of Height to account for Weight reduces the modeling errors for the value of Weight at a given value of Height compared to a model that does not include Height

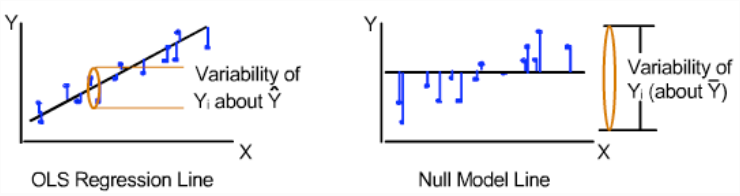
Comparing the Fit Indices

R^2 vs s_e

R^2 and s_e do not always agree

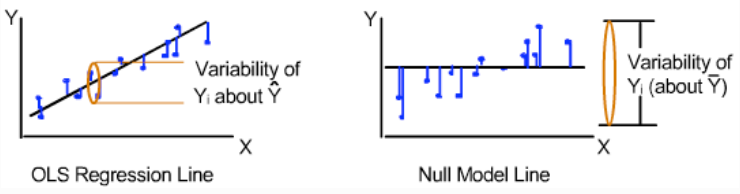
- R^2 and s_e provide complementary information, but not always the same information regarding fit
  - For the first two examples both fit indices agree
  - The third example demonstrates that R^2 can also indicate poor fit when s_e by itself would indicate good fit
- A low R^2 indicates that the regression model is not much better than the null model
- That is, scatter about the regression line is about as large as scatter about the unconditional mean, such as when
  - scatter about either line is large, as in previous example
  - scatter about either line is small, illustrated next

R^2 vs s_e: Good Fit is Large R^2 and Small s_e



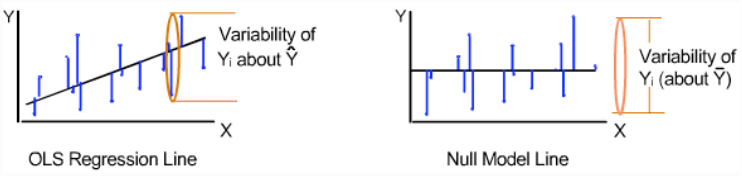
- Regression Model: Points in the scatterplot (data) fall relatively close to the regression line
- So sum of squared deviations about the conditional means, SSE, tends to be small and s_e tends to be small
- Null Model: Sum of squared deviations about the unconditional mean, SSY, is relatively large compared to SSE

R^2 vs s_e: Good Fit is Large R^2 and Small s_e



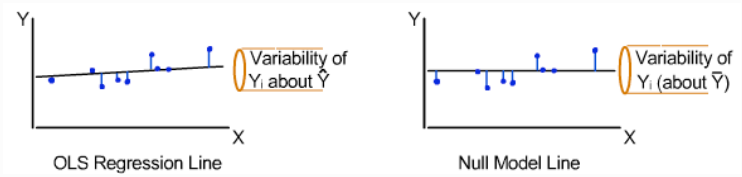
- A large decrease in variability about regression line relative to variability about the unconditional mean, leading to a large R^2
  - Regression model does much better than null model
  - Conclude that information provided by predictor variable X is useful

R^2 vs s_e: Poor Fit is Small R^2 and Large s_e



- A small decrease in variability about the regression line compared to variability about the mean, leads to a small R^2
  - The regression model does *not* do much better than null model
  - Conclude that information provided by predictor variable X is *not* particularly useful

R^2 vs s_e: Poor Fit for R^2 and Good Fit for s_e



- The amount of scatter about the regression line is low, so s_e is low, indicating good fit
- However, the variability of y is quite low, so there is not much variation of y for the model to explain
- With little variance to explain, the improvement using X in the regression model is also low, which means R^2 is low
- Small standard deviations of residuals do not necessarily mean that the model is worthwhile

8.3b
Forecasting Error

The Process of Forecasting the Unknown

Obtain the forecasts from new data

- Gather two columns of data, for variables y and X
- To calculate the estimates b_0 and b_1 for the "best-fitting" line drawn through the scatterplot, enter the data values into a regression application such as R or Excel
- There are no forecasts of the original y values because the value of y for these observations are already known
- Indeed both X and y values are required to estimate the regression model, b_0 and b_1
- **Forecasted value**: A value of $\hat{y} = b_0 + b_1 X$ in which the value of X is from *new* data, beyond the original data from which the model was estimated, and for which the value of y is not yet known
- The forecasted value is the fitted value, the same value, but a different interpretation when applying a model to *new* data

Ex: The Process of Forecasting the Unknown

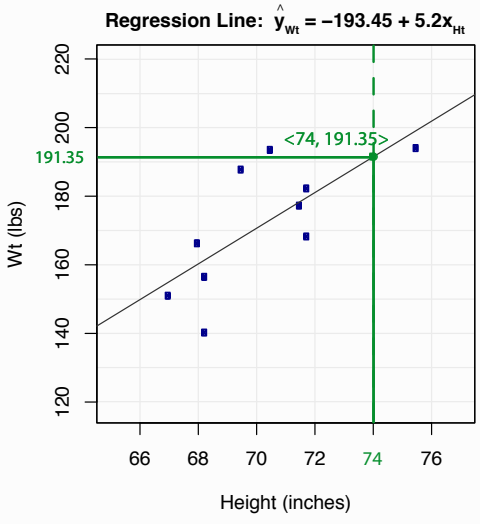
Forecasting an unknown weight from height

- A man of unknown weight is 74 inches tall, for which there is no corresponding data value in the scatter plot
- To forecast his weight, enter 74 into the regression model

$$\hat{y} = b_0 + b_1 X = -193.45 + 5.2(X) = -193.45 + 5.2(74) = 191.35$$

- **Interpretation**: For a man with a Height of 74 inches, from the same population as the original sample of 10 men from which the regression model was estimated, his forecasted Weight is 191.35 lbs

Ex: The Process of Forecasting the Unknown



A Forecast is Based on New Data

Need to consider an X value with an unknown y value

- The variability of the fitted values in the *same* sample from which the model is estimated is assessed by modeling error
- A forecast is calculated from a value of X in a *new* sample
- In a new sample, the original regression model is no longer optimal, that is, does not minimize $\sum e_i^2$
- **Key Concept**: To explain the variability of $\hat{y}$ for a specific value of X from a new sample implies that sampling error, the variability of the sample regression line from sample to sample, must be considered
- As previously discussed, these fitted values, $\hat{y}$, are conditional means, the means of all the values of y for a given value of X
- **95% Confidence Interval of a Conditional Mean**: For a specific value of X, the range of values that contains 95% of all fitted values from all possible sample regression lines

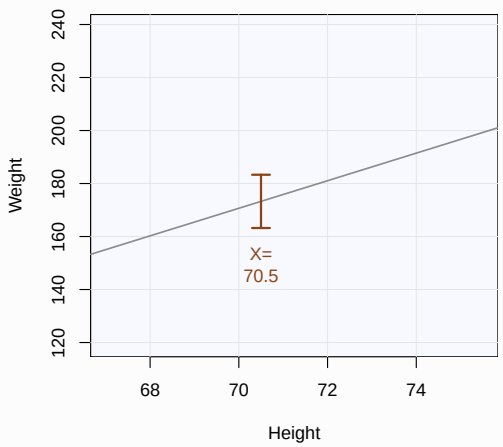
Confidence Interval of a Conditional Mean

Consider a specific value of predictor X

- There is a different confidence interval of the conditional mean of X for each value of X
- The fluctuation of the sample regression line from sample to sample tends to resemble a teeter-totter
- For the value of X equal to its mean, there is the least amount of fluctuation from sample to sample of the point on the regression line, the fitted value
- The further the value of X is from its mean, the larger the fluctuations of the fitted value across the hypothetical samples
- **Key Concept**: The confidence interval of the conditional mean, the point on the regression line for the corresponding value of X, becomes larger as the value of X becomes farther from its mean

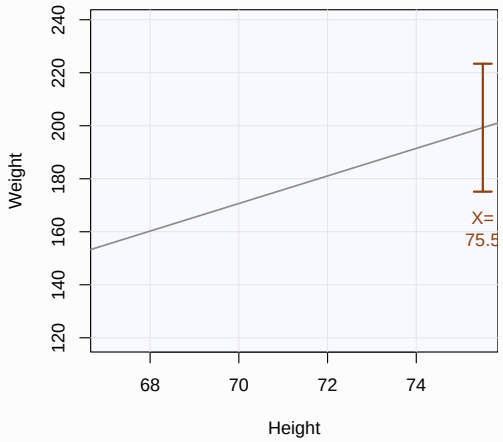
Confidence Interval of Conditional Mean for X=70.5

- This confidence interval of $\hat{y}$ is for the value of X=70.5 in, close to its mean of $\bar{x} = 70.2$



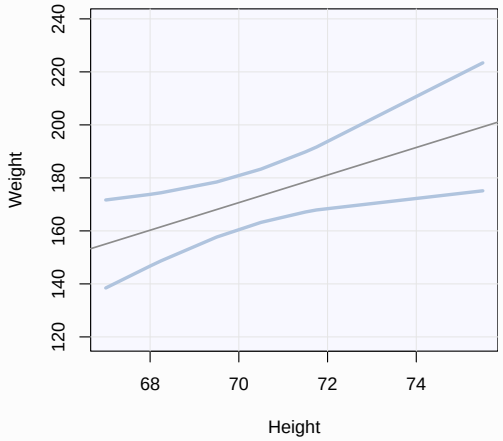
Confidence Interval of Conditional Mean for X=75.5

- This confidence interval of $\hat{y}$ is for the value of X=75.5 in, the largest value of X in the data set



Confidence Intervals of Conditional Mean: All Values of X

- Plotting all confidence intervals of the conditional mean, $\hat{y}$, for all values of X, illustrates the teeter-totter effect



Intervals for the Fitted Values

How much forecasting error is expected?

- Unfortunately the usual notation does not distinguish between
  - $\hat{y}$ as a conditional mean, a point on the regression line
  - $\hat{y}$ as a forecasted data value, a point in the scatterplot
- The future true value of y for a given value of X will not likely equal the forecasted value, $\hat{y}$
- **Key Concept**: A meaningful forecast of a specific data value includes the range of values that likely contain the actual future value of response variable y
- **Forecasting Error**: The difference between the actual value and the forecasted value of response variable y for a given value of predictor variable X
- **95% Prediction Interval**: Range of values that contains 95% of all future values of response variable y for a given future value of the predictor variable, X

The Prediction Intervals of Likely Forecasting Error

How much forecasting error is expected?

- The size of the prediction interval depends on the standard error of forecast, the standard deviation of the residuals from true forecasting to new data
- The bad news is that forecasting error is larger than the modeling error described by s_e and R^2, which are descriptive statistics that apply only to the original sample of data
- **Key Concept**: The size of a prediction interval for a fitted value depends on two sources of random variability
  - The extent of modeling error, assessed by s_e
  - The extent of sampling variability from sample to sample as the regression line changes due to sampling variability, assessed by the confidence interval
- In the analysis of prediction intervals, the smaller confidence intervals for the regression line are usually included

`Regression` Output: Prediction Intervals

Assess forecasting error

- The `lessR` function `Regression` lists for each row of data its 95% prediction interval, sorted by the value of predictor variable X

```
Data, Fitted Values, Confidence and Prediction Intervals
---------------------------------------------------------
      Ht     Wt fitted ci:lwr ci:upr pi:lwr pi:upr  width
9  67.00 151.00 155.05 138.45 171.66 119.40 190.70  71.30
10 68.00 166.25 160.26 146.74 173.78 125.93 194.58  68.64
4  68.25 140.25 161.56 148.71 174.40 127.50 195.62  68.12
5  68.25 156.50 161.56 148.71 174.40 127.50 195.62  68.12
6  69.50 187.75 168.06 157.67 178.45 134.84 201.27  66.43
7  70.50 193.50 173.26 163.21 183.31 140.15 206.37  66.22
8  71.50 177.25 178.46 167.12 189.80 144.94 211.99  67.05
1  71.75 182.25 179.76 167.89 191.63 146.06 213.47  67.41
2  71.75 168.25 179.76 167.89 191.63 146.06 213.47  67.41
3  75.50 194.00 199.27 175.13 223.41 159.54 238.99  79.45
```


Interpretation of Forecasting Error

How much forecasting error is expected?

- For a Height of 67 inches, the fitted value is 155.05

$$\hat{y} = -193.45 + 5.20X = -193.45 + 5.20(67) = 155.05$$

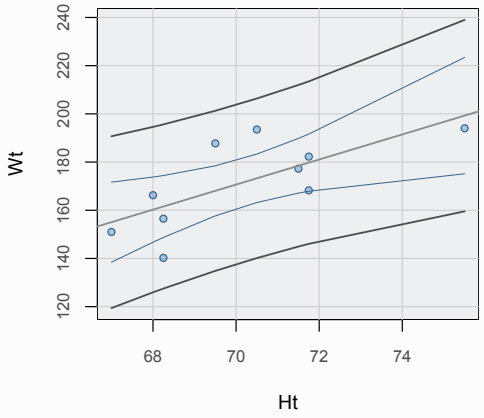
- However, the actual value of y, the person's weight, when known, will almost certainly not be 155.05 lbs
- **Interpretation**: For a man who is 67 inches tall, his actual weight is forecasted, at the 95% prediction level, to be within the range of 119.40 lbs to 190.71 lbs
- This specific prediction interval is $190.71 - 119.4 = 71.31$ lbs wide, much too wide to be of practical use
- To decrease the width of the prediction intervals
  - Decrease sampling error by increasing sample size
  - Decrease modeling error by improving the fit of the model, such as by adding additional predictor variables

Regression Output: Prediction Intervals

Prediction intervals illustrated in a scatterplot

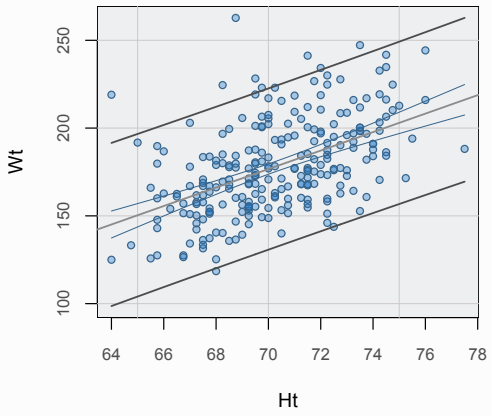
- The lessR function Regression for a model with one predictor variable, X, generates a graph that displays the
  - a scatterplot of the data
  - fitted regression line
  - confidence intervals of fitted values
  - prediction intervals of likely forecasting error
- The confidence intervals are the set of curved lines closest to the regression line, displayed in the color blue
- The prediction intervals are the set of curved lines farthest from the regression line, displayed in the orange/rust color

Regression Output: Prediction Intervals, n=10



- A small sample size, with large confidence intervals and even larger prediction intervals

Regression Output: Prediction Intervals, n=248



- A larger sample size, with small confidence intervals and still fairly large prediction intervals

Forecasts from New Data

Have R calculate a forecast for any values of the predictor

- By default, the prediction interval is provided for each set of values for the predictor variable in the data
- One of these prediction intervals is applicable to a forecast from new data if the value of the predictor variable in the new data matches the given value of the predictor variable in the original data
- To have `Regression` provide prediction intervals for new data, specify the new data values for the first predictor variable using the `X1.new` option
  - Can specify the values individually with the combine function, such as `X1.new=c(2.3,4.1)`
  - Can specify the values systematically with the sequence function, such as `X1.new=seq(0,4,0.25)`, which specifies a range of values from 0 to 4 in intervals of .25

`Regression` Output: Forecasts from New Data

- Specify three new values of X, Ht: 64, 64.5 and 65 inches

```
> Regression(Wt ~ Ht, X1.new=c(64,64.5,65))
```

```
Data, Fitted Values, Confidence and Prediction Intervals
   [sorted by lower bound of prediction interval]
-------------------------------------------------------------
     Ht Wt fitted ci:lwr ci:upr pi:lwr pi:upr  width
1 64.00    139.45 111.87 167.03  97.54 181.36  83.81
2 64.50    142.05 116.39 167.71 101.39 182.72  81.33
3 65.00    144.65 120.89 168.42 105.16 184.15  78.99
```

- Ht is for new data with unknown values of Wt, so the values for Wt are empty in the output

- The End