Regression Modeling with Actuarial and Financial Applications

by Edward Frees

Ch. 2 notes

Regression - one independent variable

- Overall task: analyze the relationship between two variables.
- Identifying and summarizing the data.
- Basic linear regression model:
  - assumptions, estimation, interpretations.
- Is the Model useful at all?
- What the modeling procedure tells us.
- Improving the Model through residual analysis.

Y and X in a basic linear regression model

- Y is a quantitative response variable (a.k.a. dependent, outcome).
- X is a quantitative predictor variable (a.k.a. independent, explanatory, or covariates).
- Two variables play different roles, so important to identify which is which and define carefully. Consider, for example,

Scatter plot - a basic graphical tool

- The observation represents the information collected from a single individual although it consists of a pair of numbers.
- For cross-sectional observations, there is no natural ordering of the data.
- The scatter plot is the most common basic graphical tool to visually investigate the relationship between the two variables.
  - By graphing the data, we lose the exact values of the observations.
  - However, we gain a visual understanding of the relationship between purchase price and income.

Correlation coefficient - a basic summary measure

- The (Pearson) correlation coefficient provides a measure of the linear relationship between two variables:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \frac{(X_i - \overline{X})(Y_i - \overline{Y})}{s_X s_Y},$$

  where s_X and s_Y are the respective sample standard deviations.

- r has a value between -1 and 1, considered dimensionless.
- If r is positive (negative), then data is said to be positively (negatively) correlated.

More on correlation coefficients

- The correlation coefficient is said to be a “unitless” measure.
  - It is unaffected by scale and location changes of either, or both, variables. (Prove this!)
  - It can readily be compared across different data sets.
- Be careful in interpreting the correlation coefficient:
  - It gives the strength of the linearity between the variables.
  - It does not capture all the possible dependence between the two variables, e.g. quadratic relationships.
- Correlation coefficients take up less space to report than a scatter plot and are often the primary statistic of interest.
  - Scatter plots help us understand other aspects of the data, such as the range, and also provide indications of non-linear relationships in the data.

Fitting a line - least squares

- We would like to fit a regression line to the data so that we are able to guess the value of Y when $X = X_*$.
- A line may be defined as the vertical height (Y) equals the intercept (b_0) plus the slope (b_1) multiplied by the horizontal distance (X):

$$\mathrm{E}(Y|X) = b_0 + b_1 X$$

- The regression line is fit to the observations using the method of least squares where intercept and slope are obtained by minimizing the sum of squares defined by:

$$\mathrm{SS}(b_0, b_1) = \sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_i)]^2 ,$$

where the quantity $Y_i - (b_0 + b_1 X_i)$ represents the deviation (or mistake) of the actual observation from the height of the line and the sum of squares, SS, represents the sum of squared deviations for the line with the intercept b_0 and slope b_1.

Least squares estimates

- The method of least squares produces the following estimates for the slope

$$b_1 = r \cdot s_Y / s_X$$

  and intercept

$$b_0 = \overline{Y} - b_1 \overline{X}.$$

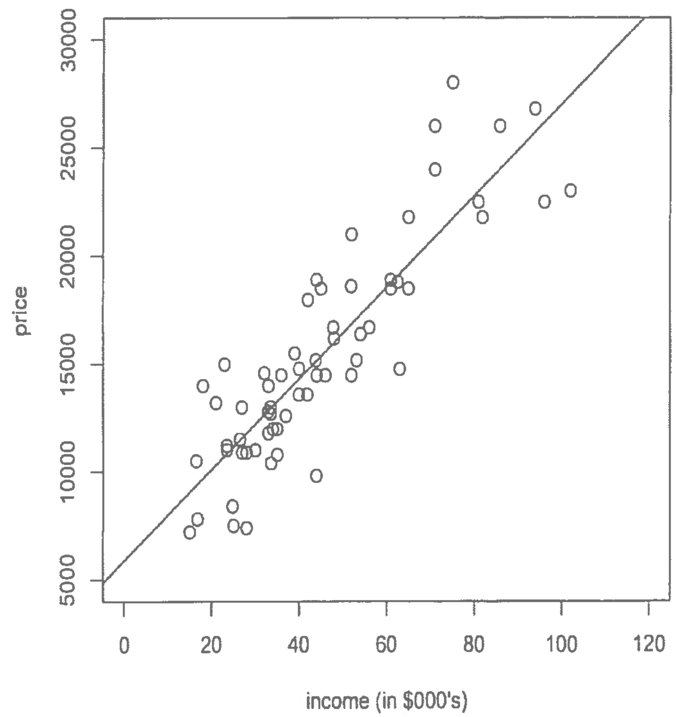
- The regression line is therefore

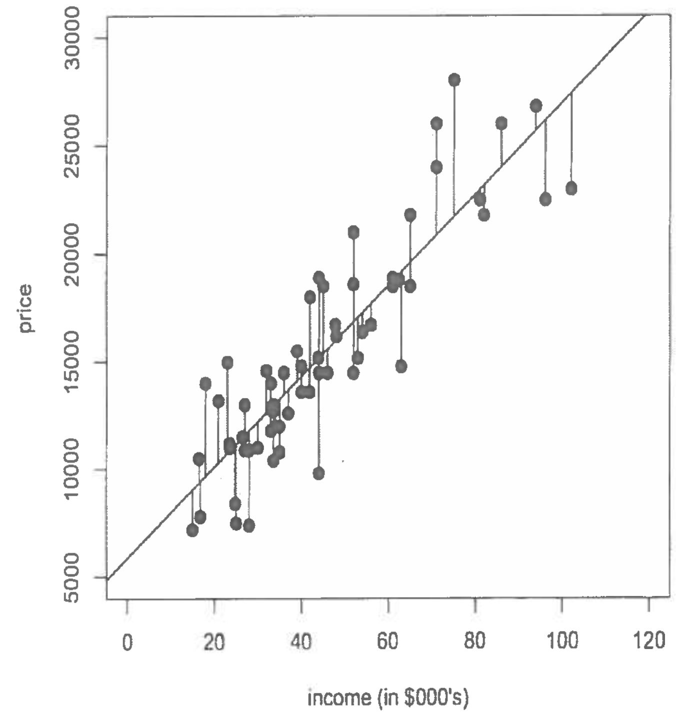
$$\widehat{Y} = b_0 + b_1 X$$

- These results can easily be proved and details will be provided in lecture.

Graphical representation of least squares estimates

Scatter plot of the data together with the super-imposed least squares regression line. The second plot shows the vertical distance of each observation to the regression line.





The basic linear regression model

- The basic linear regression model is given by:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \text{ for } i = 1, \ldots, n.$$

- The error terms $\{\varepsilon_i\}$ are assumed to be i.i.d. random variables with mean $\mathrm{E}(\varepsilon_i) = 0$ and variance $\mathrm{Var}(\varepsilon_i) = \sigma^2$.
- As a consequence, we find Y_i to have mean $\mathrm{E}(Y_i) = \beta_0 + \beta_1 X_i$ and variance $\mathrm{Var}(Y_i) = \sigma^2$.
- In addition, it is common to assume the errors have a Normal distribution, which then implies the Y_i's also have Normal distributions.
  - Normal distribution assumption allows us to derive sample properties of the least squares estimates.

Partitioning the variability

- Define the Total sum of squares as

$$\text{Total SS} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 .$$

- Interpret this as the total variation in the data set.
- Compute the fitted value $\widehat{Y}_i = b_0 + b_1 X_i$. We now have two "estimates" for Y_i - $\widehat{Y}_i$ and $\overline{Y}$.
- Decompose the total deviation as

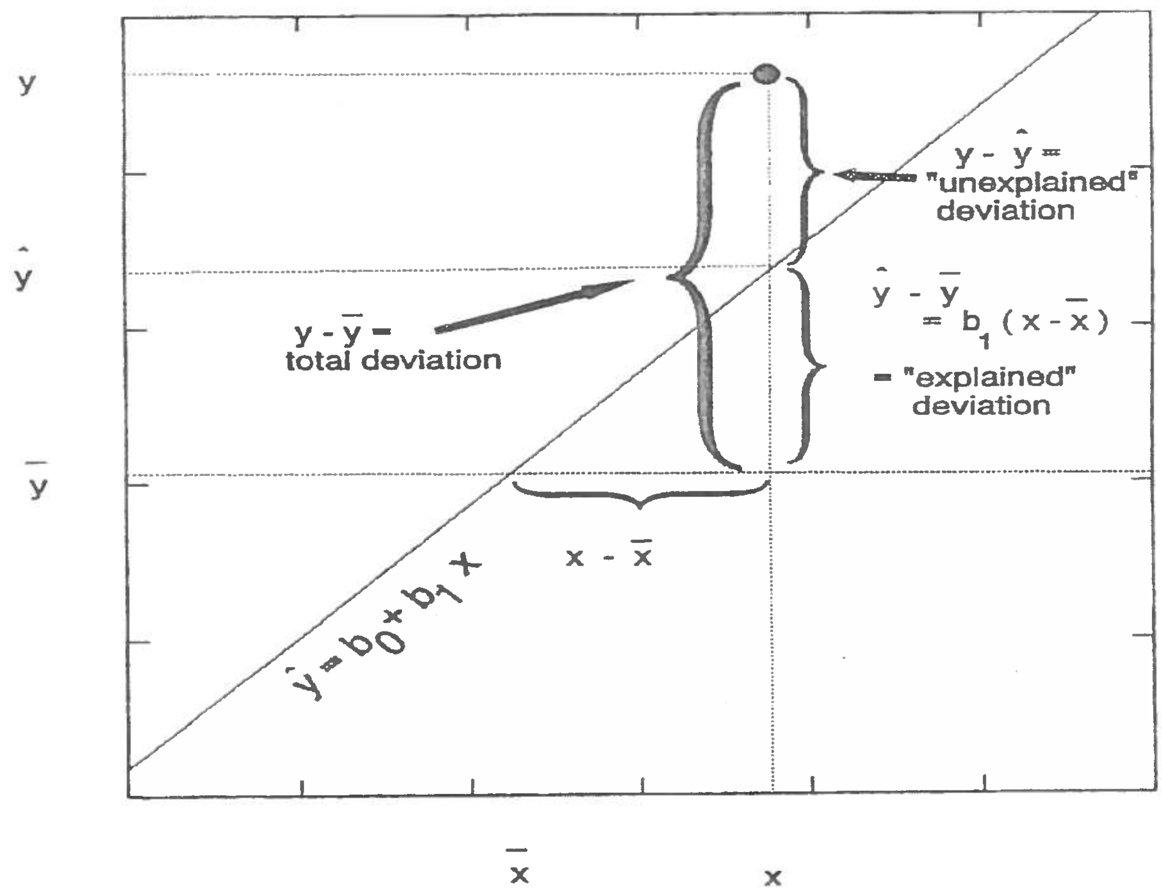
$$\underbrace{Y_i - \overline{Y}}_{\text{total deviation}} = \underbrace{Y_i - \widehat{Y}_i}_{\text{unexplained deviation}} + \underbrace{\widehat{Y}_i - \overline{Y}}_{\text{explained deviation}}$$

- Square both sides and then sum over all the observations. With some algebraic manipulation, we get

$$\underbrace{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}_{\text{Total SS}} = \underbrace{\sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2}_{\text{Error SS}} + \underbrace{\sum_{i=1}^{n} (\widehat{Y}_i - \overline{Y})^2}_{\text{Regression SS}},$$

where 'SS' stands for sum of squares.

Geometric display of the deviation decomposition



Coefficient of determination

- Interpretation of the decomposition:
  - Total SS is total variation without knowledge of X,
  - Error SS is total variation with knowledge of X, and
  - Regression SS is the difference, or the total variation "explained" by the regression line (or through knowing X).
- Define the coefficient of determination as

$$R^2 = \frac{\text{Regression SS}}{\text{Total SS}},$$

  and interpret it as the "proportion of the variability that is explained by the regression line". We have $0 \le R^2 \le 100\%$.
- Note that $s_Y^2 = \dfrac{\text{Total SS}}{n-1}$.
- Also note that in the case of one independent variable, $R^2 = r^2$, where r is the correlation coefficient between Y and X.

Residuals and mean square error

- The random error $\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$ is estimated by the "estimated error"

$$\widehat{\varepsilon}_i = Y_i - \widehat{Y}_i = Y_i - (b_0 + b_1 X_i),$$

  also called the residual.

- The "mean square error" defined by

$$\text{Error MS} = s^2 = \frac{\text{Error SS}}{n-2} = \frac{\sum_{i=1}^{n} \widehat{\varepsilon}_i^{\,2}}{n-2},$$

  is used as an estimator for σ^2.

- The positive square root $s = \sqrt{s^2}$ is called the "residual standard deviation" and gives the typical size of an error.
- The d.f. $n-2$ is also often referred to as the "error degrees of freedom".

Tracking sources of variability

Summary Measures of the Population and Sample

Data	Summary measures	Regression line		Variance
		Intercept	Slope	
Population	parameters	β_0	β_1	σ^2
Sample	statistics	b_0	b_1	s^2

The ANOVA (analysis of variance) Table

source	sum of squares (SS)	d.f.	mean sqaure (MS)
regression	regression SS	1	regression MS
error	error SS	$n-2$	error MS
total	total SS	$n-1$	

Weighted sum of the responses

- The least squares estimates b_0 and b_1 can be expressed as weighted sum of the responses as follows:

$$b_1 = \sum_{i=1}^{n} w_i Y_i$$

and

$$b_0 = \sum_{i=1}^{n} \left(\frac{1}{n} - w_i \overline{X} \right) Y_i,$$

where the weights $w_i = \dfrac{X_i - \overline{X}}{\sum_{i=1}^{n} (X_i - \overline{X})^2} = \dfrac{X_i - \overline{X}}{(n-1)s_X^2}$.

- Simple algebra leads to $\sum_{i=1}^{n} w_i = 0$.

Properties of least squares estimates

- MEAN (unbiased):

$$\mathrm{E}(b_1) = \beta_1 \text{ and } \mathrm{E}(b_0) = \beta_0.$$

- VARIANCES:

$$\begin{aligned} \mathrm{Var}(b_1) &= \frac{\sigma^2}{(n-1)s_X^2}, \quad \text{and} \\ \mathrm{Var}(b_0) &= \sigma^2 \left(\frac{1}{n} - \frac{\overline{X}^2}{(n-1)s_X^2} \right). \end{aligned}$$

- COVARIANCE:

$$\mathrm{Cov}(b_0, b_1) = \frac{-\overline{X}\sigma^2}{(n-1)s_X^2}.$$

- <u>Confidence Intervals</u> – provides us with a range of what the slope (b_1) can reasonably be.
    - $CI = b_1 \pm t * se(b_1)$
- This uses the same t score that we looked up in the table, with df=n-2 and ($\alpha/2$) in each tail.
- For the WiscLottery example,
    - $.64709 \pm (2.0106)(0.04881) = (0.549, 0.745)$
- This means that we are 95% confident that the slope is between 0.549 and 0.745. Interpreted within the regression equation, this means that for every 1 person in the population, sales of lotto tickets will go up between 0.549 and 0.745 of a ticket.

- <u>Prediction Intervals</u> – provides us with a range of what the y variable will reasonably be at a specific x.
    - $PI = y \pm t * se(pred)$
- This uses the same t score that we looked up in the table, with df=n-2 and ($\alpha/2$) in each tail.

- For the WiscLottery example, suppose we have a population of 10,000, this means POP=10,000

- This means that we are 95% confident that when the population is 10,000, the lottery sales will be between (-\$759.76, \$14,640.96). Since negative numbers are not possible for sales, we can just truncate (or cut off) this number at 0. This gives us an expected range of lotto sales between (0, \$14,640.96).

- I’m not going to discuss <u>residuals</u> right now because I don’t want to talk about them until we do them. The main idea is that <u>any</u> pattern in a residual plot is bad.

- <u>Outlier</u> – unusual point in the vertical direction
- <u>Leverage point</u> – unusual point in the horizontal direction
- Both of these are undesirable and any observation can be one or the other or both.
- We will often delete outliers, this is considered pretty common practice, as long as you provide some justification.

- Often, we will find the z scores for all observations and any point with a z score larger than ± 3 is taken out of the dataset.
- The fit of a regression line is almost always improved by the removal of these points. Which you can see in the vast improvement in the R^2 and t-score values in table 2.6 when we delete A or C. B improves the dataset with none of A,B, or C included because it would fall directly in the center of the regression that you would draw through the other points. A and C would not, they would bend down or up the regression line drastically, making many of the other points then fall farther from the regression line than they would now.

- See the code and output for these examples in the notes.
- I am not covering section 2.7.