COLUMBIA
BUSINESS
SCHOOL

Linear Regression

Fall 2001
B6014: Managerial Statistics

Professor Paul Glasserman
403 Uris Hall

General Ideas of Linear Regression

1. **Regression analysis** is a technique for using data to identify relationships among variables and use these relationships to make predictions. We will be studying **linear** regression, in which we assume that the outcome we are predicting depends linearly on the information used to make the prediction. Linear dependence means constant rate of increase of one variable with respect to another (as opposed to, e.g., diminishing returns).

2. Some motivating examples.

   (a) Suppose we have data on sales of houses in some area. For each house, we have complete information about its size, the number of bedrooms, bathrooms, total rooms, the size of the lot, the corresponding property tax, etc., and also the price at which the house was eventually sold. Can we use this data to predict the selling price of a house currently on the market? The first step is to postulate a **model** of how the various features of a house determine its selling price. A **linear model** would have the following form:

$$\begin{aligned}\text{selling price} \;=\; & \beta_0 + \beta_1\ (\text{sq.ft.}) \;+ \beta_2\ (\text{no. bedrooms}) \\ & + \beta_3\ (\text{no. bath}) \;+ \beta_4\ (\text{no. acres}) \\ & + \beta_5\ (\text{taxes}) \;+\; \text{error}\end{aligned}$$

   In this expression, β_1 represents the increase in selling price for each additional square foot of area: it is the marginal cost of additional area. Similarly, β_2 and β_3 are the marginal costs of additional bedrooms and bathrooms, and so on. The **intercept** β_0 could in theory be thought of as the price of a house for which all the variables specified are zero; of course, no such house could exist, but including β_0 gives us more flexibility in picking a model.

   The last term in the equation above, the "error," reflects the fact that two houses with exactly the same characteristics need not sell for exactly the same price. There

is always some variability left over, even after we specify the value of a large number variables. This variability is captured by an error term, which we will treat as a random variable.

Regression gives us a method for computing estimates of the parameters β_0 and $\beta_1, \ldots, \beta_5$ from data about past sales. Once we have these estimates, we can plug in values of the variables for a new house to get an estimate of its selling price.

(b) Most economic forecasts are based on regression models. The methods used are more advanced than what we cover, but we can consider a simplified version. Consider the problem of predicting growth of the economy in the next quarter. Some of the relevant factors in such a prediction might be last quarter's growth, this quarter's growth, the index of leading economic indicators, total factory orders this quarter, aggregate wholesale inventory levels, etc. A linear model for predicting growth would then take the following form:

$$\begin{aligned} \text{next qtr growth} \;=\; & \beta_0 + \beta_1\ (\text{last qtr growth}) \ + \beta_2\ (\text{this qtr growth}) \\ & + \beta_3\ (\text{index value}) \ + \beta_4\ (\text{factory orders}) \\ & + \beta_5\ (\text{inventory levels}) \ +\ \text{error} \end{aligned}$$

We would then attempt to estimate β_0 and the coefficients $\beta_1, \ldots, \beta_5$ from historical data, in order to make predictions. This particular formulation is far too simplistic to have practical value, but it captures the essential idea behind the more sophisticated methods of economic experts.

(c) Consider, next, the problem of determining appropriate levels of advertising and promotion for a particular market segment. Specifically, consider the problem of managing sales of beer at large college campuses. Sales over, say, one semester might be influenced by ads in the college paper, ads on the campus radio station, sponsorship of sports-related events, sponsorship of contests, etc. Suppose we have data on advertising and promotional expenditures at many different campuses and we want to use this data to design a marketing strategy. We could set up a model of the following type:

$$\begin{aligned} \text{sales} \;=\; & \beta_0 + \beta_1\ (\text{print budget}) \ + \beta_2\ (\text{radio budget}) \\ & + \beta_3\ (\text{sports promo budget}) \ + \beta_4\ (\text{other promo}) \\ & + \text{error} \end{aligned}$$

We would then use our data to estimate the parameters. This would tell us the marginal value of dollars spent in each category.

3. We now put this in a slightly more general setting. A regression model specifies a relation between a **dependent variable** Y and certain **explanatory variables** $X_1, \ldots, X_K$. A linear model sets

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K + \epsilon.$$

Here, ϵ (the Greek letter epsilon) is the error term. To use such a model, we need to have data on values of Y corresponding to values of the X_i's. (E.g., selling prices for

Salary ($1,000s)	Budget ($100,000s)
59.0	3.5
67.4	5.0
50.4	2.5
83.2	6.0
105.6	7.5
86.0	4.5
74.4	6.0
52.2	4.0
82.6	4.5
59.0	5.0
44.8	2.5
111.4	12.5
122.4	9.0
82.6	7.5
57.0	6.0
70.8	5.0
54.6	3.0
111.0	8.5
86.2	7.5
79.0	6.5

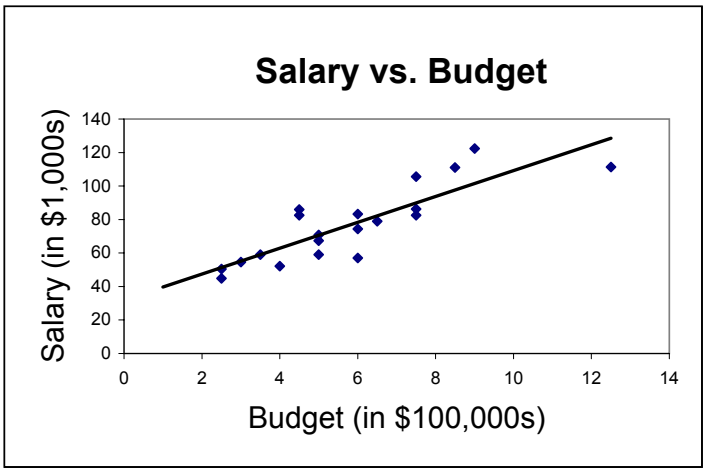

Figure 1: Salary and budget data

various house features, past growth values for various economic conditions, beer sales corresponding to various marketing strategies.) Regression software uses the data to find parameter estimates of $\beta_0, \beta_1, \ldots, \beta_K$, by implementing certain mathematical formulas. We will not discuss these formulas in detail. Instead, we will be primarily concerned with the proper interpretation of regression output.

Simple Linear Regression

1. A **simple** linear regression refers to a model with just one explanatory variable. Thus, a simple linear regression is based on the model

$$Y = \beta_0 + \beta_1 X + \epsilon$$

   In this equation, we say that X **explains** part of the variability in the dependent variable Y.

2. In practice, we rarely have just one explanatory variable, so we use **multiple** rather than simple regression. However, it is easier to introduce the essential ideas in the simple setting first.

3. We begin with a small example. A corporation is concerned about maintaining parity in salary levels of purchasing managers across different divisions. As a rough guide, it determines that purchasing managers responsible for similar budgets in different divisions should have similar compensation. Figure 1 displays salary levels for 20 purchasing managers and the sizes of the budgets they manage.

4. The scatter plot in Figure 1 includes a straight line fit to the data. The slope of this line gives the marginal increase in salary with respect to increase in budget responsibility.

5. Since this example is quite simple, we could fit a line to the data by drawing a line with a ruler. Regression analysis gives us a more systematic approach. Moreover, regression gives us the **best** line through the data. (In Excel, you can insert a regression line in a scatter plot by right-clicking on a data point and then selecting `Add Trendline...`.)

6. We need to define what we mean by the best line. Regression uses the **least squares criterion**, which we now explain. Any line we might come up with has a corresponding intercept β_0 and a slope β_1. This line may go through some of the data points, but it typically does not go through all of them. Let us label the data points by their coordinates $(X_1, Y_1), \ldots, (X_{20}, Y_{20})$. These are just the 20 pairs tabluated above. For the budget level X_i, our straight line predicts the salary level

$$\hat{Y}_i = \beta_0 + \beta_1 X_i.$$

Unless the line happens to go through the point (X_i, Y_i), the predicted value $\hat{Y}_i$ will generally differ from the observed value Y_i. The difference between the two is the **error** or **residual**

$$\begin{aligned} e_i &= Y_i - \hat{Y}_i \\ &= Y_i - (\beta_0 + \beta_1 X_i). \end{aligned}$$

(We think of ϵ_i as a random variable — a random error — and e_i as a particular outcome of this random variable.) The least squares criterion chooses β_0 and β_1 to **minimize the sum of squared errors**

$$\sum_{i=1}^{n} e_i^2,$$

where n is the number of data points.

7. A (non-obvious) consequence of this criterion is that the estimated regression line always goes through the point $(\overline{X}, \overline{Y})$ and the estimated slope is given by

$$\hat{\beta}_1 = \frac{Cov[X, Y]}{StdDev[X]}.$$

8. To summarize, of all possible lines through the data, regression picks the one that minimizes the sum of squared errors. This choice is reported through the estimated values of β_0 and β_1.

9. To give a preliminary indication of the use of regression, let's run a regression on the 20 points displayed in Figure 1. To get a complete analysis, we use Excel's Regression tool which can be found under `Tools/Data Analysis`. The results appear in Figure 2.

10. We begin by looking at the last two rows of the regression output, under the heading "Coeff." The two values displayed there are the estimated coefficients (intercept and slope) of the regression line. The estimated intercept is $\hat{\beta}_0 = 31.937$ and the estimated slope is $\hat{\beta}_1 = 7.733$. Thus, the estimated relation between salary and budget is

$$\text{Salary} = 31.937 + 7.733 \text{ Budget}$$

Regression Statistics	
Multiple R	0.850
R Square	0.722
Adjusted R Square	0.707
Standard Error	12.136
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*P-value*
Regression	1	6884.7	6884.7	46.74	0.000
Residual	18	2651.1	147.3		
Total	19	9535.8			

	Coeff	*Std Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	31.937	7.125	4.48	0.00	16.97	46.91
Budget	7.733	1.131	6.84	0.00	5.36	10.11

Figure 2: Results of regression of salary against budget

This says that each additional \$100,000 of budget responsibility translates to an expected additional salary of \$7,730. (Recall that Budget is in \$100,000s and Salary is in \$1,000s.) If we wanted to fit a salary corresponding to a budget of \$600,000, we could substitute 6.0 into this equation to get a salary of $31.937 + 7.733(6.0) = 78.335$.

11. Two questions remain: Why is the least squares criterion the correct principle to follow? How do we evaluate and use the regression line? We touch on the first issue only briefly, then address the second one in detail.

12. **Assumptions Underlying Least Squares**

    - The errors $\epsilon_1, \ldots, \epsilon_n$ are independent of the values of $X_1, \ldots, X_n$.
    - The errors have expected value zero; i.e., $E[\epsilon_i] = 0$.
    - All the errors have the same variance: $Var[\epsilon_i] = \sigma^2$, for all $i = 1, \ldots, n$.
    - The errors are uncorrelated; i.e., $Corr[\epsilon_i, \epsilon_j] = 0$ if $i \neq j$.

    The first two assumptions imply that

    $$E[Y|X = x] = \beta_0 + \beta_1 x;$$

    i.e., they imply that the expected outcome of Y really does depend linearly on the value of x. When all four assumptions hold, the line selected by the least squares criterion is the optimal estimate.

13. The precise sense in which least squares is optimal is a theoretical issue that we do not address. It is, however, important to touch on the four assumptions made above. The first two are very reasonable: if the ϵ_i's are indeed random errors, then there is no reason to expect them to depend on the data or to have a nonzero mean. The second two assumptions are less automatic.

- Do we necessarily believe that the variability in salary levels among managers with large budgets is the same as the variability among managers with small budgets? Is the variability in price really the same among large houses and small houses? These considerations suggest that the third assumption may not be valid if we look at too broad a range of data values.
- Correlation of errors becomes an issue when we use regression to do forecasting. If we use data from several past periods to forecast future results, we may introduce correlation by overlapping several periods and this would violate the fourth assumption.

More advanced techniques address these considerations. For our purposes, we will always assume that the assumptions are in force. You should, however, be aware of possible limitations in these assumptions.

Evaluating the Estimated Regression Line

1. We feed data into the computer and we get back estimates of the model parameters β_0 and β_1. Is this estimated line any good? More precisely, does it accurately reflect the relation between the X and Y variables? Is it a reliable guide in predicting new Y values corresponding to new X values? (E.g., predicting the selling price of a house that just came on the market, or setting the salary for a newly defined position.)

2. Intuitively, the estimated regression line is useful if the points $(X_1, Y_1), \ldots, (X_n, Y_n)$, when plotted as in Figure 1, are pretty well lined up. The more they look like a cloud of dots, the less informative the regression will be.

3. The output of a regression gives us a lot of information to make this intuition precise in evaluating the explanatory power of a model. There is quite a bit of notation that goes with this information. As we go through it, keep the following principles in mind: Our goal is to determine how much of the variability in Y values is **explained** by the X values. We measure variability using sums of squared quantities.

4. To understand explained variability, consider the salary example. The Y_i's (the salary levels) exhibit considerable variability — not all managers have the same salary. We conduct the regression analysis to determine to what extent salary is tied to responsibility as measured by budget: the 20 managers have different budgets as well as different salaries. Thus, we ask to what extent the differences in salaries are explained by differences in budgets.

5. Continuing with this example, let's focus on the following portion of the regression output from Figure 2:

Regression Statistics	
Multiple R	0.850
R Square	0.722
Adjusted R Square	0.707
Standard Error	12.136
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*P-value*
Regression	1	6884.7	6884.7	46.74	0.000
Residual	18	2651.1	147.3		
Total	19	9535.8			

The lower table is called the **ANOVA table**. ANOVA is short for **analysis of variance**. This table breaks down the total variability into the explained and unexplained parts.

6. DF stands for degrees of freedom, SS for sum of squares, and MS for mean square. The mean squares are just the sum of squares divided by the degrees of freedom: MS = SS/DF.

7. We begin by looking at the SS, first giving an intuitive explanation before giving any formulas. A sum of squares measures variability. The Total SS (9535.8) measures the total variability in the salary levels. The Regression SS (6884.7) is the **explained variation**. It measures how much variability is explained by differences in budgets. What's left over, the Error SS (2651.1) is the **unexplained variation**. This reflects differences in salary levels that cannot be attributed to differences in budget responsibilities. The explained and unexplained variation sum to the Total SS.

8. How much of the original variability has been explained? The answer is given by the ratio of the explained variation to the total variation, which is

$$R^2 = \frac{\text{Explained variability}}{\text{Total variability}} = \frac{\text{SSR}}{\text{SST}} = \frac{6884.7}{9538.8} = 72.2\%$$

This quantity is the **coefficient of determination**, though everybody calls it **R-square**.

9. Other things being equal, a high R^2 indicates high explanatory power and a low R^2 indicates the opposite.

10. Fact: In simple linear regression, R^2 is also equal to the square of the sample correlation between the X_i's and Y_i's. Recall that correlation measures the strength of a linear relationship between two variables. Thus, high R^2 corresponds to a strong linear relationship (either positive or negative) between two variables.

11. Let us now define these quantities more generally and more precisely. Suppose our observed dependent variables are $Y_1, \ldots, Y_n$ and let their sample mean be $\overline{Y}$. The total sum of squares is

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

This is the same as the sample variance, except that we have not divided by $n - 1$. As before, let $\hat{Y}_i$ denote the predicted value of Y corresponding to X_i; that is,

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i,$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimates of the intercept and slope provided by the regression. The regression sum of squares (the explained variation) is

$$\text{SSR} = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2.$$

The difference between the observed value Y_i and the predicted value $\hat{Y}_i$ is the i-th residual

$$e_i = Y_i - \hat{Y}_i.$$

The error sum of squares (unexplained variation) is

$$\text{SSE} = \sum_{i=1}^{n} e_i^2.$$

12. It is a non-obvious mathematical fact that the explained and unexplained variation sum to equal the total variation:

$$\text{SSR} + \text{SSE} = \text{SST}$$

Just as before, we have

$$R^2 = \frac{\text{SSR}}{\text{SST}},$$

the fraction of the total variation explained by the regression. So, R^2 is a measure of the explanatory power of the model. (We will discuss adjusted R^2 later.)

Evaluating the Estimated Slope

1. Let's now go back to the regression output and look at some information about the estimated parameters β_0 and β_1. The relevant part of the output from Figure 2 is this:

	Coeff	*Std Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	31.937	7.125	4.48	0.00	16.97	46.91
Budget	7.733	1.131	6.84	0.00	5.36	10.11

This table gives more information about the estimates. The first row corresponds to β_0 (the intercept), the second to β_1 (the slope), which is the influence of budget on salary. The column `Coeff` displays the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$. The next column gives (estimated) standard errors associated with these estimates. These are valuable in assessing the uncertainty in the estimates.

2. The slope estimate $\hat{\beta}_1$ provided by the least squares method is unbiased; i.e., $E[\hat{\beta}_1] = \beta_1$. In this sense, it is accurate. Its precision (or efficiency) is measured by the estimated standard error — in our example, 1.131.

3. Does budget have a statistically significant impact on salary? The next two columns address this question. Notice that we could formulate it as a hypothesis test:

$$\begin{aligned} H_0: &\quad \beta_1 = 0 \ \text{(budget has no effect on salary)} \\ H_1: &\quad \beta_1 \neq 0 \ \text{(budget has some effect on salary)} \end{aligned}$$

The `t Stat` above is a **test statistic** for this hypothesis test. It is computed as follows:

$$t = \frac{\hat{\beta}_1 - 0}{s_{\hat{\beta}_1}} = \frac{7.733}{1.131} = 6.84;$$

$s_{\hat{\beta}_1}$ is the `Std Err` entry corresponding to the β_1 row (the estimated standard error of $\hat{\beta}_1$). This is a huge t-ratio, so we get a very small p-value, one that is zero to three decimal places. We conclude that there is very significant evidence in favor of $\beta \neq 0$; i.e., in favor of budget having some influence on salary.

4. It is typical to get very small p-values for this type of test. If you don't, it means you have somehow included a variable of absolutely no relevance to the dependent variable.

5. Under the null hypothesis ($\beta = 0$), the t statistic computed above, namely

$$t = \frac{\hat{\beta}_1 - 0}{s_{\hat{\beta}_1}}$$

has a t-distribution with $n-2$ degrees of freedom, not $n-1$. Intuitively, we have lost 2 df because we have estimated both β_0 and β_1.

6. More generally,

$$t = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}}$$

has a t distribution with $n-2$ df, where β_1 is the true, unknown slope. We can use this fact to test other hypotheses. To test

$$\begin{aligned} H_0: &\quad \beta_1 \leq 1 \\ H_1: &\quad \beta_1 > 1 \end{aligned}$$

compute the test statistic

$$t = \frac{\hat{\beta}_1 - 1}{s_{\hat{\beta}_1}};$$

reject H_0 if $t > t_{n-2,\alpha}$. In general, compute the t-ratio by subtracting off the value of β_1 in the null hypothesis.

7. The fact that

$$t = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}}$$

has a t distribution with $n-2$ df, where β_1 is the true slope, allows us to get a **confidence interval for the slope:**

$$\hat{\beta}_1 \pm t_{n-2,\alpha/2}\, s_{\hat{\beta}_1}$$

8. Example: Let's find a 95% confidence interval for the slope in the example above. Our point estimate is $\hat{\beta}_1 = 7.733$ and the estimated standard error is $s_{\hat{\beta}_1} = 1.131$. We have $20 - 2 = 18$ df. From the t-table we find that $t_{18,.025} = 2.101$. This gives the interval

$$7.733 \pm (2.101)(1.131);$$

so, we are 95% confident that the true slope lies in the interval

$$(5.357,\ 10.109)$$

9. Where does the estimated standard error $s_{\hat{\beta}_1}$ come from? The exact standard error of $\hat{\beta}_1$, denoted by $\sigma_{\hat{\beta}_1}$ satisfies

$$\sigma^2_{\hat{\beta}_1} = \frac{\sigma^2_\epsilon}{\sum_{i=1}^n X_i^2 - n\overline{X}^2},$$

where σ^2_ϵ is the (unknown) variance of the errors ϵ_i. The denominator in this expression is similar to the sample variance of the X_i's, except that is not divided by $n - 1$. Since we don't know σ_ϵ, we replace it with an estimate s_e to get $s_{\hat{\beta}_1}$. We discuss s_e in the next section.

Making Predictions

1. Our ultimate objective in building a regression model is to make predictions about the dependent variable Y for new values of the independent variable X. We want to predict selling price for a house with given characteristics, or growth in the next quarter given information about the current state of the economy, or sales of beer as a result of a new marketing strategy.

2. In the salary/budget example, making a "prediction" actually means recommending a salary level Y for a given budget responsibility X. The regression is useful in ensuring that the recommended salary is in line with existing levels of compensation.

3. It is necessary to distinguish two types of predictions: predictions of individual values and predictions of expected values. Examples:

   - Predicting the selling price of a particular house vs. predicting the average selling price among all houses with certain characteristics.
   - Predicting salary level for a particular position with a particular budget responsibility vs. predicting average salary level over all positions with that budget.
   - Predicting beer sales at a particular campus based on an advertising/promotion mix vs. predicting average sales over all campuses at which that mix is used.

   The average value of Y corresponding to an X value of x is $E[Y|X = x]$, the conditional expectation of Y given $X = x$.

4. Naturally, we expect to have more uncertainty in our estimate of a particular value than in our estimate of an average value.

5. This additional uncertainty is reflected in wider confidence intervals for the particular value. However, the point estimates for the two types of predictions are exactly the same. In either case, our prediction corresponding to an X value of x is

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

   Graphically, this corresponds to the height of the regression line at point x.

6. Example: Consider the budget level $x = 6.0$. The predicted corresponding salary level is

$$\hat{Y} = 31.937 + 7.733(6.0) = 78.335.$$

7. Of course, by itself a point estimate is not very informative. We need to supplement it with a confidence interval.

8. An important ingredient for this type of confidence interval is the common variance σ_ϵ^2 of the errors ϵ_i. Since we don't know σ_ϵ^2 in practice, we estimate it. The estimate is

$$s_e^2 = \frac{\text{SSE}}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2,$$

   where $e_i = Y_i - \hat{Y}_i$ is the i-th residual, as before. Notice that s_e^2 is similar to the sample variance of the residuals; once again, we have divided by $n-2$ because we lost 2 df in estimating β_0 and β_1. Taking the square root of s_e^2 yields s_e.

9. In the regression output of Figure 2, the s_e is labeled `Standard Error` and is the fourth values from the top of the output, 12.136. The value of s_e^2 also appears in the ANOVA table. Recall that the MS column is the ratio of the SS and DF columns. Thus, MSE = 147.3 is the same as s_e^2.

10. Back to confidence intervals. A confidence interval for the **average** Y value at level x, namely $E[Y|X = x]$ is

$$\hat{Y} \pm t_{n-2,\alpha/2}\sqrt{\left[\frac{1}{n} + \frac{(x-\overline{X})^2}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}\right] s_e^2}.$$

    For a particular Y value at level x, the confidence interval is

$$\hat{Y} \pm t_{n-2,\alpha/2}\sqrt{\left[1 + \frac{1}{n} + \frac{(x-\overline{X})^2}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}\right] s_e^2}.$$

    Notice that the only difference is that we have added one more s_e^2 inside the square root. This makes sense: the additional uncertainty in predicting a particular value rather than an average value is the uncertainty in the errors ϵ_i, which is σ_ϵ^2, which we approximate by s_e^2.

11. Now look at the complicated ratio appearing inside the square root. This expression becomes large if its numerator $(x-\overline{X})$ is large, which means that we have more uncertainty in predictions for values of x that are far from $\overline{X}$ than in predictions for values close to $\overline{X}$. This should not be surprising: our predictions should be most reliable when they are close to values for which we have data. If we go out to extreme values for which we have little or no data, it is harder to make accurate predictions.

12. Example: Let's get a confidence interval for the average salary level corresponding to a budget of \$600,000 to supplement our earlier point estimate. Directly from the regression output we get $s_e^2 = 147.3$. To get the other term inside the square root, we need more information. By taking the average of the data displayed in Figure 1, we find that $\overline{X} = 5.825$. We similarly find that the sample standard deviation of the X_i's is 2.462; i.e.,

$$\sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right)} = 2.462.$$

From this we can find the expression we need:

$$\sum_{i=1}^{20} X_i^2 - n\overline{X}^2 = (n-1)(2.462)^2 = 19(2.462)^2 = 115.17$$

Since the x value we want is 6.0, we get

$$\sqrt{\left[\frac{1}{n} + \frac{(x-\overline{X})^2}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}\right] s_e^2} = \sqrt{\left[\frac{1}{20} + \frac{(6.0-5.825)^2}{115.17}\right] 147.3} = 2.72.$$

From the t table, we get $t_{18,.025} = 2.101$. So, our confidence interval is

$$78.335 \pm (2.101)(2.72) \;=\; 78.335 \pm 5.71$$

Multiple Regression

1. We now turn to the more interesting case of building a model with several explanatory variables. In practice, we almost always need more than one variable to get a meaningful model. However, one of the principles of regression is that we should use as few variables as possible: only include the most important explanatory variables and avoid redundancies among these.

2. The general multiple linear regression model with K explanatory variables has the following form:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K + \epsilon,$$

Our data consists of observations

$$\begin{gathered}(Y_1, X_{11}, \ldots, X_{K1}) \\ (Y_2, X_{12}, \ldots, X_{K2}) \\ (Y_3, X_{13}, \ldots, X_{K3}) \\ \cdots \\ (Y_n, X_{1n}, \ldots, X_{Kn}).\end{gathered}$$

In other words, we have n observations of the outcome Y, and for each one we have the corresponding values of the explanatory variables $X_1, \ldots, X_K$. The symbol X_{ij} denotes the value of variable X_i corresponding to the j-th observation. We put this data into a regression package and get estimates of $\beta_0, \beta_1, \ldots, \beta_K$.

3. As before, these estimates are based on minimizing the sum of squared residuals $\sum_{i=1}^{n} e_i^2$, where

$$e_i = Y_i - \hat{Y}_i,$$

and

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_K X_{Ki}.$$

The underlying least squares assumptions are the same as in simple regression, with one new assumption added. The new assumption basically rules out the possibility of redundancy among the explanatory variables. For example, we cannot have X_1 measure area in square feet and X_2 measure area in square yards. Two such variables would contain exactly the same information but in different units.

4. The possibility of redundancy among the explanatory variables is known as **multicollinearity**. If it is present, the regression may give poor results or may fail to run altogether. Most regression software checks for multicollinearity. As a user, you should be careful not to introduce redundant variables.

5. Let's look at an example. Here are the first few lines of a data set consisting of 228 assessed property values along with some information about the houses in question:

ROW	VALUE	LOC	LOTSZ	BDRM	BATH	ROOMS	AGE	GARG	EMEADW	LVTTWN
1	190.00	3	6.90	4	2.0	8	38	1	0	1
2	215.00	1	6.00	2	2.0	7	30	1	1	0
3	160.00	3	6.00	3	2.0	6	35	0	0	1
4	195.00	1	6.00	5	2.0	8	35	1	1	0
5	163.00	3	7.00	3	1.0	6	39	1	0	1
6	159.90	3	6.00	4	1.0	7	38	1	0	1
7	160.00	1	6.00	2	1.0	7	35	1	1	0
8	195.00	3	6.00	3	2.0	7	38	1	0	1
9	165.00	3	9.00	4	1.0	6	32	1	0	1
10	180.00	3	11.20	4	1.0	9	32	1	0	1
11	181.00	3	6.00	5	2.0	10	35	0	0	1
. . .										

The second column gives the assessed value in thousands of dollars. The third encodes the location: 1 for East Meadow, 3 for Levittown, and 4 for Islip. The next gives lot size in thousands of square feet, then bedrooms, bathrooms, total rooms, age, and number of garage units. The last two columns encode location in **dummy variables**. We discusse these in more detail later. For now, just note that a 1 under EMEADW indicates a house

in East Meadow, a 1 under LVTTWN indicates a house in Levittown, and 0's in both columns indicate a house in Islip.

6. Our goal is to use this data to predict assessed values from characteristics of a house. **The best model need not include all variables**.

7. Appendix 2 of these notes gives the regression output for a model using all explanatory variables. (Again, that's not necessarily the best thing to do.) From the coefficients displayed there, we find the regression equation

```
VALUE = 78.737 + 0.679 LOTSZ - 3.687 BDRM + 19.003 BATH + 8.484 ROOMS
        - 0.348 AGE + 4.014 GARG + 57.082 EMEADW + 24.418 LVTTWN
```

   This says, for example, that the marginal value of an additional 1000 square feet of lot is .679 thousand dollars, given that everything else is held fixed.

8. It is important to understand that the estimated slopes $\hat{\beta}_i$ **depend on which variables are included**. Adding and deleting variables changes the other $\hat{\beta}_i$'s. Notice that BDRM has an estimated **negative** slope of -3.687. Does this mean that additional bedrooms detract from the value of a house? Not necessarily. This may simply reflect the fact that the relevant information in the number of bedrooms is already captured in other variables. Deleting some variables may eliminate this anomaly.

9. The negative slope on the AGE variable seems appropriate: increasing the age may well decrease the value.

10. Let's proceed to information about the estimated parameters:

	Coeff	Std Err	t Stat	P-value
Intercept	78.74	10.52	7.49	0.000
LOTSZ	0.6792	0.3706	1.83	0.068
BDRM	-3.687	2.224	-1.66	0.099
BATH	19.003	2.802	6.78	0.000
ROOMS	8.484	1.491	5.69	0.000
AGE	-0.3475	0.1201	-2.89	0.004
GARG	4.014	2.336	1.72	0.087
EMEADW	57.082	3.972	14.37	0.000
LVTTWN	24.418	3.887	6.28	0.000

   This table gives the estimated parameters and the corresponding estimated standard errors for these estimates. Each t-stat is a test statistic to test which variables have a significant effect on assessed value; i.e., to test

$$H_0: \quad \beta_i = 0$$
$$H_1: \quad \beta_i \neq 0$$

Notice that BDRM has a large p-value, suggesting that the true slope may be zero **when the other variables are included**. This may prompt us to remove the BDRM variable in our next attempt to build a model. GARG and LOTSZ also have relatively large p-values, but these may change once we delete BDRM.

11. Just as in simple regression, we can use the information in this table to carry out any test on the slopes. For example, to test

$$H_0: \quad \beta_{\text{BATH}} \leq 10$$
$$H_1: \quad \beta_{\text{BATH}} > 10,$$

we would compute

$$t = \frac{19.003 - 10}{2.802}$$

and reject H_0 if $t > t_{n-K-1,\alpha}$, where n is the number of data points and K is the number of explanatory variables. In our case, $n = 228$ and $K = 8$. A t-distribution with 219 df is very well approximated by the standard normal.

12. We now proceed to the **ANOVA Table**:

Regression Statistics	
Multiple R	0.841
R Square	0.708
Adjusted R Sq	0.697
Standard Erro	20.569
Observations	228

ANOVA

	df	*SS*	*MS*	*F*	*P-value*
Regression	8	224290.81	28036.4	66.3	0.000
Residual	219	92657.67	423.1		
Total	227	316948.48			

13. The Total DF is one less than the sample size, $n - 1$. The Regression DF is the number of explanatory variables K. The Error DF is $n - K - 1$.

14. The interpretation of SS is the same as before. The total sum of squares, SST = 316948, measures total variation; the regression sum of squares, SSR = 224291, is the explained variation and the error sum of squares, SSE = 92658, is the unexplained variation. The proportion of variation that is explained by the model is

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 70.8\%.$$

So, in this case, we have explained a rather large fraction of the variation. However, adding extraneous explanatory variables will artificially inflate the R^2, so we must be careful in interpreting this number; more on this later.

15. The formulas for SST, SSR and SSE are exactly the same as in simple regression.

16. Continuing with the ANOVA table, notice the F value of 66.3. This is the ratio of the MSR to the MSE. (As in simple regression, these mean squares are obtained from the sum of squares by dividing by the DF.) The F value is a test statistic for the hypotheses

$$\begin{aligned} H_0: &\quad \beta_1 = \beta_2 = \cdots = \beta_K = 0 \\ H_1: &\quad \text{some } \beta_i \neq 0. \end{aligned}$$

Thus, the null hypothesis is that none of the variables has any effect; this is a **simultaneous** test on all the slopes. We have not discussed the F-test, but its interpretation is the same as for other tests. The very small p-value of 0.000 indicates that we may safely reject H_0. (It is very unusual not to reject this null hypothesis.)

17. As already mentioned, introducing extra variables can lead to spurious results and can interfere with the proper estimation of slopes for the important variables. On the other hand, introducing more variables will virtually always increase R^2. In order to penalize an excess of variables, we also consider the **adjusted** R^2, which is

$$\text{adjusted } R^2 \ = 1 - \frac{\text{SSE}/(n-K-1)}{\text{SST}/(n-1)}.$$

This should be contrasted with

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}},$$

an alternative expression for the ordinary R^2. The adjusted thus divides numerator and denominator by their DF. Since we divide by $n-K-1$, increasing the number of variables will not necessarily increase the adjusted R^2.

18. In the example above, the adjusted R^2 is 69.7%. When we compare different models, we should compare the adjusted R^2 as well as the ordinary R^2.

Dummy Variables

1. Often, some of the variables in a regression are **categorical** rather then **numeric**. A typical example is the location variable in the example above. The possible locations considered are East Meadow, Levittown and Islip. These were originally encoded as locations 1, 3, and 4, respectively. However, the numbers 1, 3, and 4 are arbitrary; their values carry no information but simply provide a means of distinguishing categories.

2. Nothing would stop us from carrying out a regression using the values 1, 3, and 4 for the three towns, but the results of such a model would be meaningless. How would we interpret the slope for such a variable?

3. The correct way to incorporate categorical data is through **dummy variables**. A dummy variable takes the value 0 or 1 to distinguish between two categories. To distinguish m different categories, we need $m-1$ dummy variables.

4. Example: To distinguish three towns, we need two dummy variables:

$$\begin{aligned} \text{EMEADW} &= 1 \text{ if East Meadow, 0 otherwise} \\ \text{LVTTWN} &= 1 \text{ if Levittown, 0 otherwise} \end{aligned}$$

   Using these variables, we get the following encoding:

$$\begin{aligned} 1 \;\; 0 &= \text{ a house in East Meadow} \\ 0 \;\; 1 &= \text{ a house in Levittown} \\ 0 \;\; 0 &= \text{ a house in Islip} \end{aligned}$$

   Notice that we don't need to introduce a separate variable for Islip.

5. The two dummy variables just defined are X_7 and X_8 in the example above. These variables take only the values 0 and 1. How do we interpret the "slopes" β_7 and β_8? If $X_7 = 1$, then the expected increase in assessed value is β_7, compared with $X_7 = 0$; if $X_8 = 1$, then the expected increase in assessed value is β_8. Thus, β_7 is the premium for a house in East Meadow over a comparable house in Islip and β_8 is the premium for a house in Levittown over a comparable house in Islip.

6. In our example, these premiums are estimated at \$57,082 and \$24,418. The p-values for both are 0.000, indicating that a non-zero premium does in fact exist.

7. Using the corresponding standard errors, we can get confidence intervals for these premiums. For the first one, we get

$$\$57,082 \pm t_{219,\alpha/2}(3.972).$$

   The DF of 219 is large enough to replace the t-value with the corresponding z-value of $z_{\alpha/2}$.

Prediction

1. As in simple linear regression, we use estimates of $\beta_0, \beta_1, \ldots, \beta_k$ to make predictions by substituting specific values for the explanatory variables. Regardless of whether we predict a particular outcome or an expected outcome, our point estimate of Y for values $x_1, \ldots, x_K$ of the explanatory variables is

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_K x_K.$$

2. Here is an example of a prediction in the housing value example. The predicted value of a house with lot size 6.8, 3 bedrooms, 2 baths, 7 rooms, 32 years old, 1 garage unit in Levittown is given by

```
    Fit  Stdev.Fit          95% C.I.          95% P.I.
 187.00       3.27   ( 180.55, 193.45)  ( 145.94, 228.06)
```

The fitted value of 187 is found by plugging the specified values into the regression equation.

3. The formula for the standard deviation in the multiple case is too complicated to be discussed here. Statistical software (such as MINITAB) gives the value 3.27 automatically.

4. A quick-and-dirty approximation to the confidence interval can be based on s_e, given just before the ANOVA table. In our example, $s_e = 20.57$. An approximate confidence interval for a predicted individual value is

$$\hat{Y} \pm t_{n-K-1,\alpha/2}(s_e)$$

and for a predicted average value it is

$$\hat{Y} \pm t_{n-K-1,\alpha/2}\frac{s_e}{\sqrt{n}}$$

5. These approximate confidence intervals are adequate for ball-park information but are not very precise. They are the best available option with the current state of spreadsheet software.

Appendix 1

Regression Statistics	
Multiple R	0.252
R Square	0.064
Adjusted R Squ	0.059
Standard Error	36.238
Observations	228

ANOVA

	df	*SS*	*MS*	*F*	*P-value*
Regression	1	20160.5	20160.5	15.4	0.000
Residual	226	296788.0	1313.2		
Total	227	316948.5			

	Coefficients	*Std Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	205.292	7.389	27.785	0.000	190.733	219.851
AGE	-0.796	0.203	-3.918	0.000	-1.196	-0.396

Appendix 2

Regression Statistics	
Multiple R	0.841
R Square	0.708
Adjusted R Sq	0.697
Standard Erro	20.569
Observations	228

ANOVA

	df	*SS*	*MS*	*F*	*P-value*
Regression	8	224290.81	28036.4	66.3	0.000
Residual	219	92657.67	423.1		
Total	227	316948.48			

	Coeff	*Std Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	78.737	10.516	7.487	0.000	58.011	99.463
LOTSZ	0.679	0.371	1.833	0.068	-0.051	1.410
BDRM	-3.687	2.224	-1.658	0.099	-8.069	0.696
BATH	19.003	2.802	6.783	0.000	13.482	24.525
ROOMS	8.484	1.491	5.690	0.000	5.545	11.422
AGE	-0.348	0.120	-2.894	0.004	-0.584	-0.111
GARG	4.014	2.336	1.718	0.087	-0.590	8.617
EMEADW	57.082	3.972	14.373	0.000	49.254	64.909
LVTTWN	24.418	3.887	6.282	0.000	16.757	32.079

Appendix 3

Regression Statistics	
Multiple R	0.834
R Square	0.696
Adjusted R Squ	0.689
Standard Error	20.828
Observations	228

ANOVA

	df	*SS*	*MS*	*F*	*P-value*
Regression	5	220644.2	44128.8	101.7	0.000
Residual	222	96304.2	433.8		
Total	227	316948.5			

	Coefficients	*Std Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	80.573	9.794	8.227	0.000	61.272	99.874
BATH	19.592	2.795	7.009	0.000	14.084	25.101
ROOMS	8.092	1.315	6.152	0.000	5.500	10.684
AGE	-0.368	0.119	-3.093	0.002	-0.603	-0.134
EMEADW	53.484	3.550	15.067	0.000	46.489	60.480
LVTTWN	18.711	3.312	5.649	0.000	12.183	25.238