Multiple Regression

● Often we have data on several independent variables that can be used to predict / estimate the response.

Example: To predict Y = teacher salary, we may use:

Example: Y = sales at music store may be related to:

● A linear regression model with more than one independent variable is a multiple linear regression (MLR) model:

● In general, we have m independent variables and $m + 1$ unknown regression parameters.

Purposes of the MLR model

(1) Estimate the mean response E(Y | $\underline{X}$) for a given set of X_1, X_2, …, X_m values.

(2) Predict the response for a given set of X_1, X_2, …, X_m values.

(3) Evaluate the relationship between Y and the independent variables by interpreting the partial regression coefficients β_0, β_1, …, β_m (or their estimates).

Interpretations:

- **(Estimated intercept): the (estimated) mean response if all independent variables are zero (may not make sense)**
- **β_i (or $\hat{\beta}_i$): The (estimated) change in mean response for a one-unit increase in X_i, holding constant all other independent variables.**
- **May not be possible: What if X_1 = home runs and X_2 = runs scored?**

- **Note: The partial effects of each independent variable in a MLR model do not equal the effect of each variable in separate SLR models.**

- **Why? The independent variables tend to be correlated to some degree.**

• Partial effect: interpreted as the effect of an independent variable "<u>in the presence of</u> the other variables in the model."

• Finding least-squares estimates of $\beta_0, \beta_1, \ldots, \beta_m$ is typically done using matrices:

$$\underline{\hat{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T\underline{\mathbf{Y}}$$

where: $\underline{\mathbf{Y}}$ = vector of the n observed Y values in data set
X = matrix containing the observed values of the independent variables (see sec. 8.2)
$\underline{\hat{\beta}}$ = a vector of the least squares estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_m$

• We will use software to find the estimates of the regression coefficients in the MLR model.

<u>Example</u>: Data gathered for 30 California cities.
Y = annual precipitation (in inches)
X_1 = altitude (in feet)
X_2 = latitude (in degrees)
X_3 = distance from Pacific (in miles)

Estimated model is: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3$
From computer:

Interpretation of $\hat{\beta}_0$?
Interpretation of $\hat{\beta}_2$?

Interpretation of $\hat{\beta}_3$?

Inference with the MLR model

- **Again, we don't know σ^2 (the error variance), so we must estimate it.**
- **Again, we use as our estimate of σ^2:**

- **As in SLR, the total variation in the sample *Y* values can be separated: TSS = SSR + SSE.**

- **SS formulas given in book – for MLR, we will use software.**

Rain example: SSR = SSE =

Error df = MSE =

- **Most values in ANOVA table similar as for SLR.**

- ***m* d.f. associated with SSR**
- ***n* – *m* – 1 d.f. associated with SSE**

Overall F-test

- **Tests whether the model as a whole is useless.**

- **Null hypothesis: none of the independent variables are useful for predicting *Y*.**

H_0: $\beta_1 = \beta_2 = \ldots = \beta_m = 0$
H_a: At least one of these is not zero

- **Again, test statistic is $F^* = MSR / MSE$**

- **If $F^* > F_\alpha(m, n - m - 1)$, then reject H_0 and conclude at least one of the variables is useful.**

Rain data: F^* =

Testing about Individual Coefficients

- **Most easily done with t-tests.**
- **The *j*-th estimate, $\hat{\beta}_j$, is (approximately) normal with mean β_j and standard deviation $\sqrt{c_{jj}\sigma^2}$, where c_{jj} = *j*-th diagonal element of $(X^TX)^{-1}$ matrix.**

- **Replace σ^2 with its estimate, MSE:**

● **To test H_0: $\beta_j = 0$, note:**

● **For each coefficient, computer gives:** $\hat{\beta}_j$, $\sqrt{c_{jj} MSE}$, **and t statistic.**

H_a **<u>Reject H_0 if</u>:**

Software gives P-value for the (two-tailed) test about <u>each</u> β_j separately.

Rain data:

F-tests about sets of independent variables

● We can also test whether certain sets of independent variables are useless, in the presence of the other variables in the model.

Example: Suppose variables under consideration are X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8.

Question: Are X_2, X_4, X_7 needed, if the others are in the model?

● We want our model to have "large" SSR and "small" SSE. Why?

● If "full model" has much lower SSE than the "reduced model" (without X_2, X_4, X_7), then at least one of X_2, X_4, X_7 is needed.
→ conclude β_2, β_4, β_7 not all zero.

● To test: H_0: $\beta_2 = \beta_4 = \beta_7 = 0$
vs. H_a: β_2, β_4, β_7 not all zero

Use:

Reject H_0 if

Example above: numerator d.f. =

● Can test about more than one (but not all) coefficients within computer package (`TEST` statement in SAS or `anova` function in R)

Example:

Inferences for the Response Variable in MLR

As in SLR, we can find:

● CI for the mean response for a given set of values of $X_1, X_2, \ldots, X_m$.

● PI for the response of a new observation with a given set of values of $X_1, X_2, \ldots, X_m$.

Examples:

● Find a 90% CI for the mean precipitation for all cities with altitude 100 feet, latitude 40 degrees, and 70 miles from the coast.

● Find a 90% prediction interval for the precipitation of a new city having altitude 100 feet, latitude 40 degrees, and 70 miles from the coast.

Interpretations:

- **The coefficient of determination in MLR is denoted R^2.**

- **It is the proportion of variability in *Y* explained by the linear relationship between *Y* and <u>all</u> the independent variables (Note: $0 \leq R^2 \leq 1$).**

- **The higher R^2, the better the linear model explains the variation in *Y*.**
- **No exact rule about what a "good" R^2 is.**

Rain example:

Interpretation: