Multiple Linear Regression

- Response Variable: Y
- Explanatory Variables: $X_1,\ldots,X_k$
- Model (Extension of Simple Regression):

  $E(Y) = \alpha + \beta_1 X_1 + \ldots + \beta_k X_k \qquad V(Y) = \sigma^2$
- Partial Regression Coefficients (β_i): Effect of increasing X_i by 1 unit, holding all other predictors constant.
- Computer packages fit models, hand calculations very tedious

Prediction Equation & Residuals

- Model Parameters: $\alpha, \beta_1, \ldots, \beta_k, \sigma$
- Estimators: $a, b_1, \ldots, b_k, \hat{\sigma}$
- Least squares prediction equation: $\hat{Y} = a + b_1 X_1 + \cdots + b_k X_k$
- Residuals: $e = Y - \hat{Y}$
- Error Sum of Squares: $SSE = \sum e^2 = \sum (Y - \hat{Y})^2$
- Estimated conditional standard deviation:

$$\hat{\sigma} = \sqrt{\frac{SSE}{n-k-1}}$$

Commonly Used Plots

- **Scatterplot**: Bivariate plot of pairs of variables. Do not adjust for other variables. Some software packages plot a matrix of plots
- **Conditional Plot (Coplot)**: Plot of Y versus a predictor variable, seperately for certain ranges of a second predictor variable. Can show whether a relationship between Y and X_1 is the same across levels of X_2
- **Partial Regression (Added-Variable) Plot**: Plots residuals from regression models to determine association between Y and X_2, after removing effect of X_1 (residuals from (Y , X_1) vs (X_2 , X_1))

Example - Airfares 2002Q4

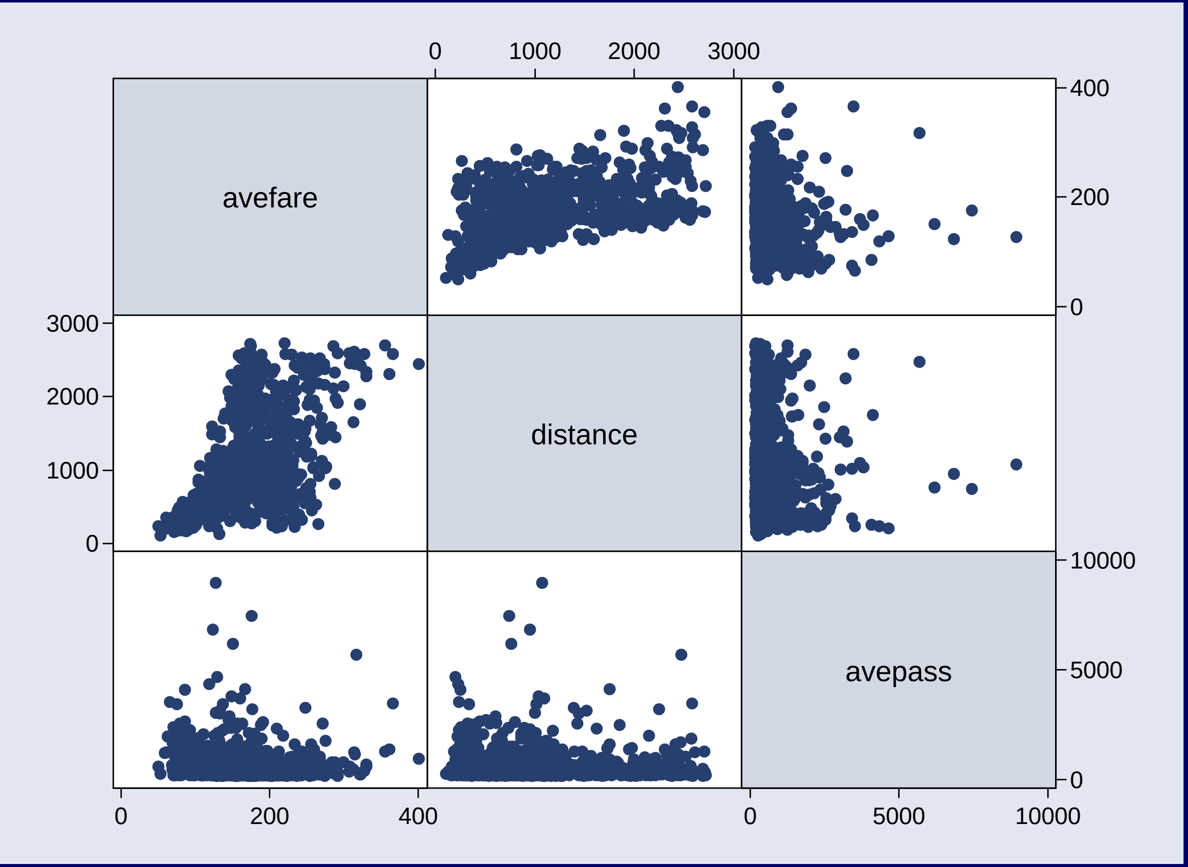
- Response Variable: Average Fare (Y, in $)
- Explanatory Variables:
  - Distance (X_1, in miles)
  - Average weekly passengers (X_2)
- Data: 1000 city pairs for 4th Quarter 2002
- Source: U.S. DOT

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
AVEFARE	1000	50.52	401.23	163.3754	55.36547
DISTANCE	1000	108.00	2724.00	1056.9730	643.20325
AVEPASS	1000	181.41	8950.76	672.2791	766.51925
Valid N (listwise)	1000				

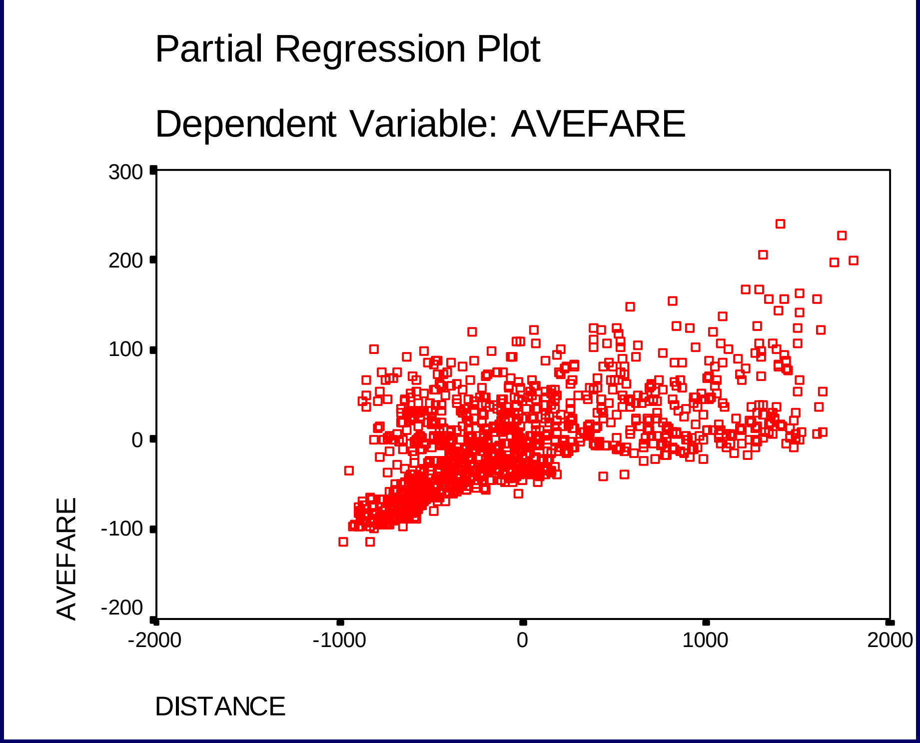
Example - Airfares 2002Q4

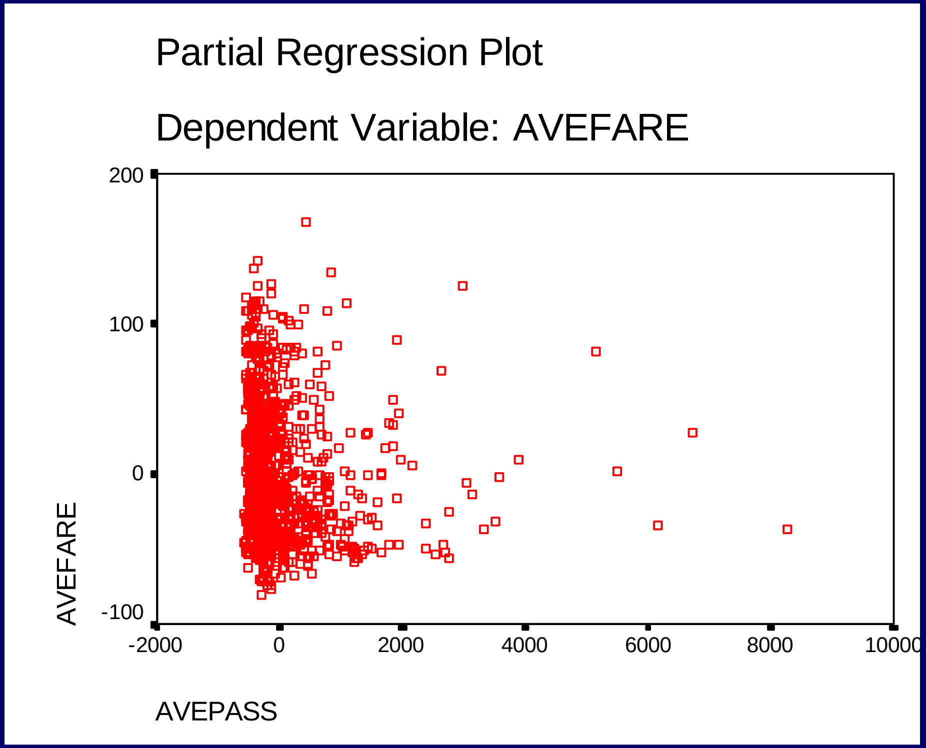
Scatterplot Matrix of Average Fare, Distance, and Average Passengers (produced by STATA):



Example - Airfares 2002Q4

Partial Regression Plots: Showing whether a new predictor is associated with *Y*, after removing effects of other predictor(s):





After controlling for AVEPASS, DISTANCE is linearly related to FARE

After controlling for DISTANCE, AVEPASS not related to FARE

Standard Regression Output

- Analysis of Variance:
  - Regression sum of Squares: $SSR = \sum(\hat{Y} - \overline{Y})^2 \quad df_R = k$
  - Error Sum of Squares: $SSE = \sum(Y - \hat{Y})^2 \quad df_E = n - k - 1$
  - Total Sum of Squares: $TSS = \sum(Y - \overline{Y})^2 \quad df_T = n - 1$
- Coefficient of Correlation/Determination: $R^2 = SSR/TSS$
- Least Squares Estimates
  - Regression Coefficients
  - Estimated Standard Errors
  - *t*-statistics
  - *P*-values (Significance levels for 2-sided tests)

Example - Airfares 2002Q4

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.592[a]	.350	.349	44.67574

a. Predictors: (Constant), AVEPASS, DISTANCE

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1072336	2	536168.162	268.632	.000[a]
	Residual	1989934	997	1995.921		
	Total	3062270	999			

a. Predictors: (Constant), AVEPASS, DISTANCE

b. Dependent Variable: AVEFARE

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	114.146	3.084		37.018	.000
	DISTANCE	.050	.002	.581	22.646	.000
	AVEPASS	-.005	.002	-.074	-2.881	.004

a. Dependent Variable: AVEFARE

Multicollinearity

- Many social research studies have large numbers of predictor variables
- Problems arise when the various predictors are highly related among themselves (collinear)
  - Estimated regression coefficients can change dramatically, depending on whether or not other predictor(s) are included in model.
  - Standard errors of regression coefficients can increase, causing non-significant t-tests and wide confidence intervals
  - Variables are explaining the same variation in Y

Testing for the Overall Model - *F*-test

- Tests whether **any** of the explanatory variables are associated with the response
- H_0: $\beta_1 = \cdots = \beta_k = 0$ (None of X^s associated with Y)
- H_A: Not all $\beta_i = 0$

$$T.S.: F_{obs} = \frac{MSR}{MSE} = \frac{R^2 / k}{(1 - R^2)/(n - (k+1))}$$

$$P - val : P(F \geq F_{obs})$$

The *P*-value is based on the *F*-distribution with *k* numerator and (n-(k+1)) denominator degrees of freedom

Testing Individual Partial Coefficients - *t*-tests

- Wish to determine whether the response is associated with a single explanatory variable, after controlling for the others

- H_0: $\beta_i = 0$ H_A: $\beta_i \neq 0$ (2-sided alternative)

$$T.S.: t_{obs} = \frac{b_i}{\hat{\sigma}_{b_i}}$$

$$R.R.: |t_{obs}| \geq t_{\alpha/2,\, n-(k+1)}$$

$$P-val: 2P(t \geq |t_{obs}|)$$

Modeling Interactions

- Statistical Interaction: When the effect of one predictor (on the response) depends on the level of other predictors.
- Can be modeled (and thus tested) with cross-product terms (case of 2 predictors):
  - $E(Y) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$
  - $X_2=0 \Rightarrow E(Y) = \alpha + \beta_1 X_1$
  - $X_2=10 \Rightarrow E(Y) = \alpha + \beta_1 X_1 + 10\beta_2 + 10\beta_3 X_1$

    $= (\alpha + 10\beta_2) + (\beta_1 + 10\beta_3) X_1$
- The effect of increasing X_1 by 1 on $E(Y)$ depends on level of X_2, unless $\beta_3=0$ (t-test)

Comparing Regression Models

- Conflicting Goals: Explaining variation in *Y* while keeping model as simple as possible (parsimony)
- We can test whether a subset of *k*-*g* predictors (including possibly cross-product terms) can be dropped from a model that contains the remaining *g* predictors. H_0: $\beta_{g+1} = \ldots = \beta_k = 0$
  - Complete Model: Contains all *k* predictors
  - Reduced Model: Eliminates the predictors from H_0
  - Fit both models, obtaining the Error sum of squares for each (or R^2 from each)

Comparing Regression Models

- H_0: $\beta_{g+1} = \ldots = \beta_k = 0$ (After removing the effects of $X_1, \ldots, X_g$, none of other predictors are associated with Y)
- H_a: H_0 is false

$$\text{Test Statistic}: F_{obs} = \frac{(SSE_r - SSE_c)/(k-g)}{SSE_c/[n-(k+1)]}$$

$$P = P(F \geq F_{obs})$$

P-value based on F-distribution with k-g and n-(k+1) d.f.

Partial Correlation

- Measures the strength of association between Y and a predictor, controlling for other predictor(s).
- Squared partial correlation represents the fraction of variation in Y that is not explained by other predictor(s) that is explained by this predictor.

$$r_{YX_2 \bullet X_1} = \frac{r_{YX_2} - r_{YX_1} r_{X_1X_2}}{\sqrt{\left(1 - r_{YX_1}^2\right)\left(1 - r_{X_1X_2}^2\right)}} \qquad -1 \le r_{YX_2 \bullet X_1} \le 1$$

Coefficient of Partial Determination

- Measures proportion of the variation in Y that is explained by X_2, out of the variation not explained by X_1
- Square of the partial correlation between Y and X_2, controlling for X_1.

$$r^2_{YX_2 \bullet X_1} = \frac{R^2 - r^2_{YX_1}}{1 - r^2_{YX_1}} \qquad 0 \le r^2_{YX_2 \bullet X_1} \le 1$$

- where R^2 is the coefficient of determination for model with both X_1 and X_2: $R^2 = SSR(X_1,X_2) / TSS$
- Extends to more than 2 predictors (pp.414-415)

Standardized Regression Coefficients

- Measures the change in $E(Y)$ in standard deviations, per standard deviation change in X_i, controlling for all other predictors (β_i^*)
- Allows comparison of variable effects that are independent of units
- Estimated standardized regression coefficients:

$$b_i^* = b_i \left(\frac{s_{X_i}}{s_Y} \right)$$

- where b_i , is the partial regression coefficient and s_{Xi} and s_Y are the sample standard deviations for the two variables