Stat 608 Chapter 6

Regression Diagnostics for Multiple Regression

1. Draw scatterplots of the data:
   - Standardized residual plots
   - Marginal model plots
   - Inverse response plots
   - Plots for constant variance
2. Identify leverage points & outliers
3. Assess relationships between predictors
   - Added variable plots
   - Variance inflation factor
   - R^2 adjusted
   - Forward, backward, stepwise, AIC, SBC selection (Chapter 7)

Model Checking

- A multiple linear regression model is valid if:

$$E[Y|X=x] = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$
$$Var(Y|X=x) = \sigma^2$$

- When a valid model has been fit, plots of the residuals against ______ ________________ or ________________________________ will:
  - have a random scatter of points
  - have constant variability as the horizontal axis increases
- The residual plots should still have no patterns. Patterns indicate the model is not valid.

Model Checking

- If **both** of the following are true, then residual plots help determine the function g.

$$E[Y|X = x] = g(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)$$

$$E[X_i|X_j] \approx \alpha_0 + \alpha_1 X_j$$

- Otherwise, Cook & Weisberg: “Using residuals to guide model development will often result in misdirection, or at best more work than would otherwise be necessary.”

- Example:
  - True model: three predictors.
  - We fit a model with two.
  - Residual plots are potentially non-random.

- We can’t use residual plots to tell us what part of the model has been misspecified.

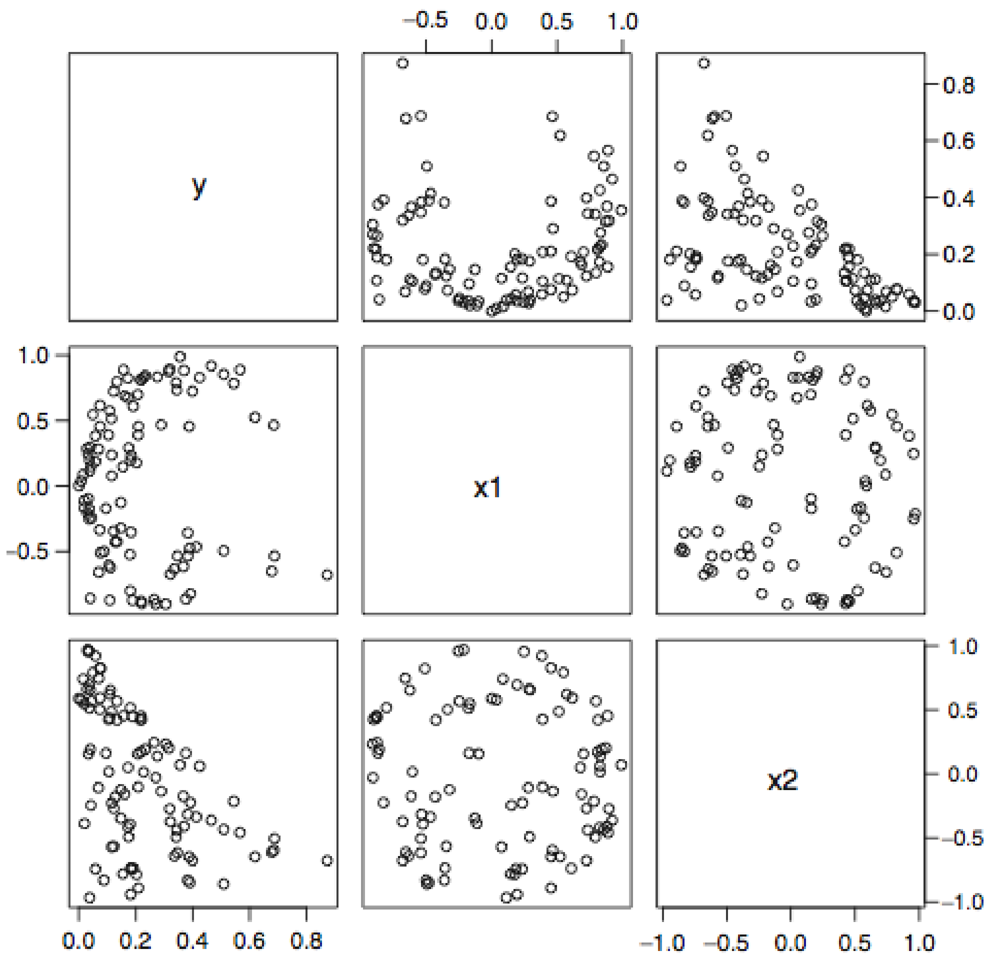
Example 1: Function g DNE

Mean function chosen as:

$$E[Y|X] = \frac{|x_1|}{2 + (1.5 + x_2)^2} = \frac{g_1(x_1)}{g_2(x_2)}$$

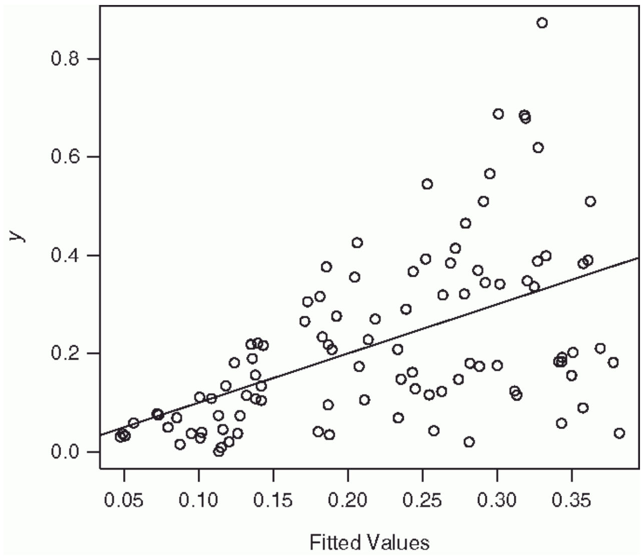
We need two functions to model the mean.

Example 1: Function g DNE



Example 1: Function g DNE

- The usual interpretation of the relationship between y and x2 (the fan shape) is that the variance is non-constant, but the data was generated with errors with constant variance!
- We can't use residual plots to tell us what part of the model has been misspecified.



Example 2: predictors are not linearly related

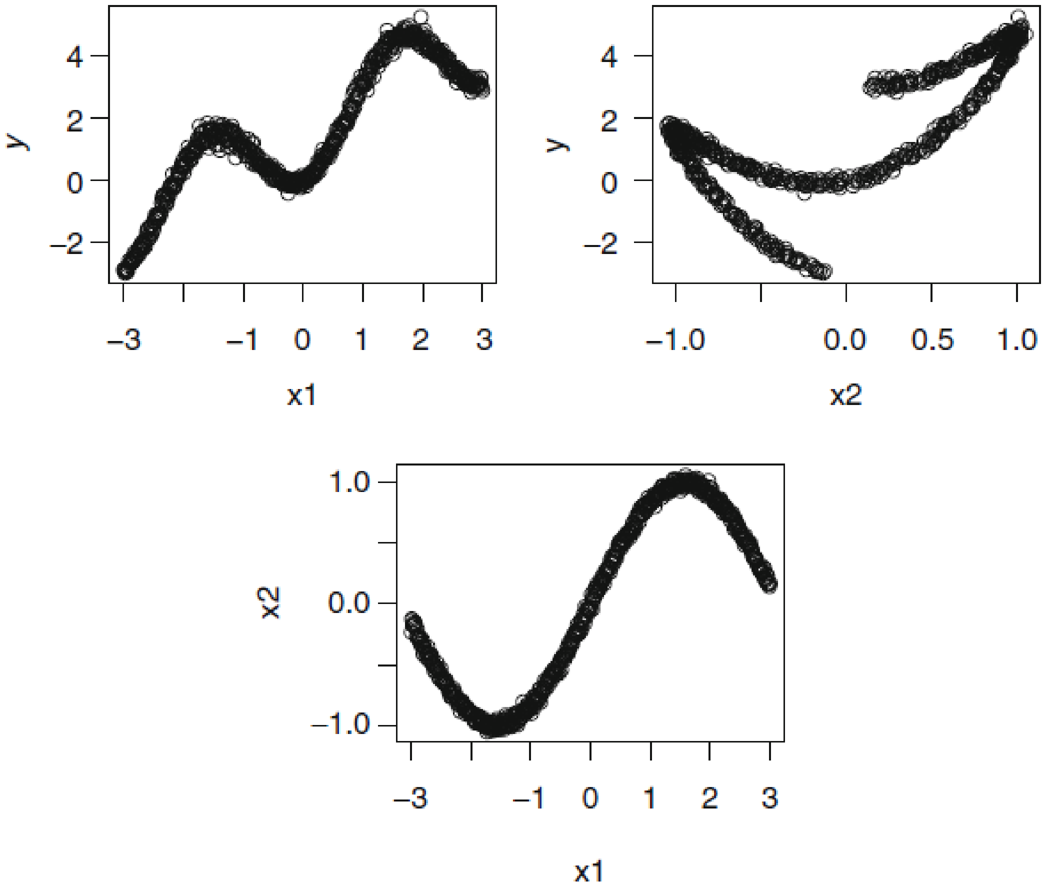
Mean function is not too crazy:

$$Y = x_1 + 3x_2^2 + e$$

But the predictors are related via sine:

$$E[X_2|X_1] = sin(X_1)$$

Example 2: predictors are not linearly related

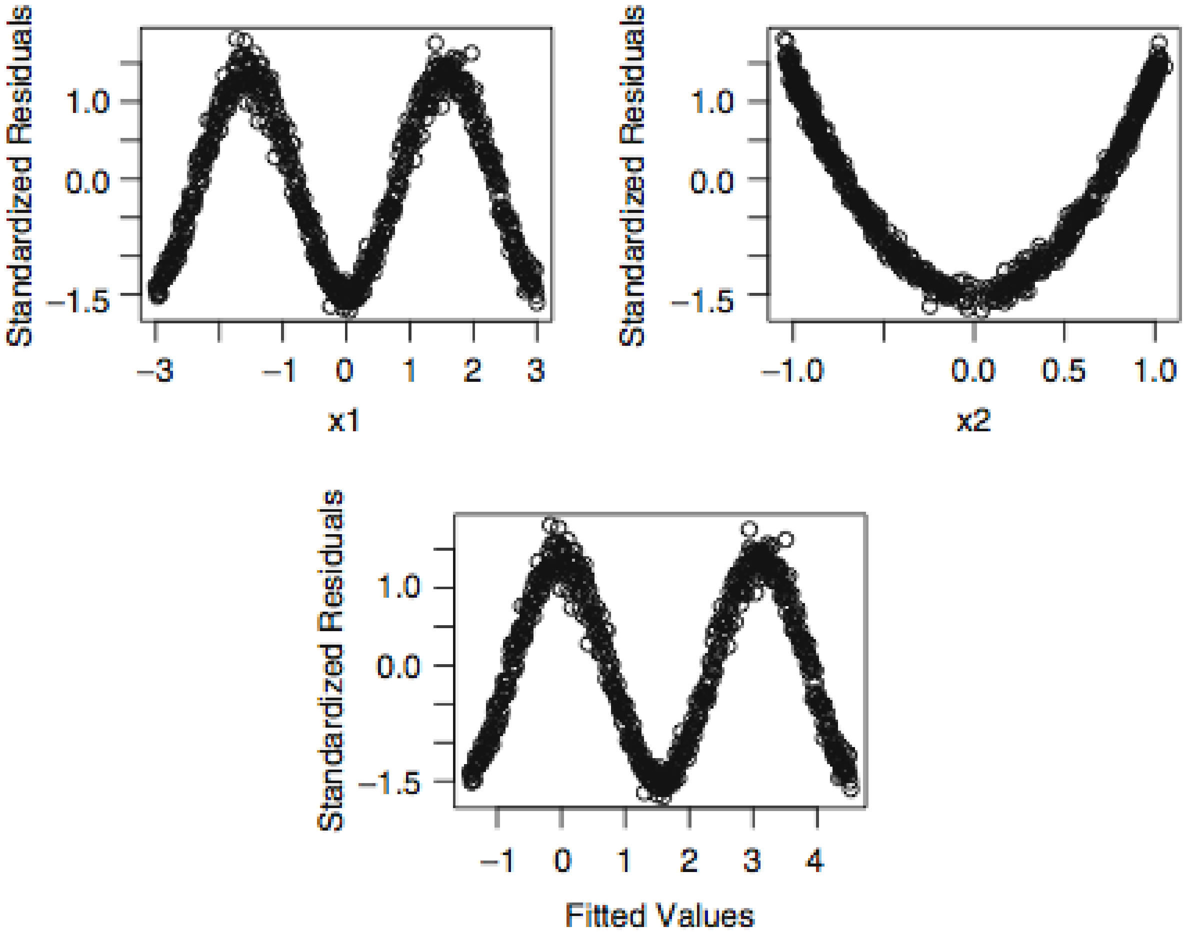


+ Example 2: predictors are not linearly related

- First we try the usual regression model to see what the residual plots look like:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$$

Example 2: predictors are not linearly related



+

Example 2: predictors are not linearly related

- The usual interpretation might be that we should use a periodic function in x_1 in the model, but that's not true in this case.

- The highly nonlinear relationship between ______________ has produced the nonrandom plot in the standardized residuals against x_1.

- Moral: We can't use residual plots to tell us what part of the model has been misspecified.

Leverage

- Recall that:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{H}\mathbf{Y}$$

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

$$\hat{Y}_i = h_{ii}Y_i + \sum_{j \neq i} h_{ij}Y_j$$

That is, each predicted value of Y is a linear combination of all the values of Y in the data set, generally with the other values being more lightly weighted than the one we are predicting for (Y_i).

- As with simple linear regression, if any of the h_{ii} values is much different from the others, it means that single observation *may* be changing the model much more than the others.

Leverage

- Rule of thumb:

$$h_{ii} > 2 \times \text{average}(h_{ii}) = 2 \times \frac{(p+1)}{n}$$

Classify a point as a point of high leverage if its hat value exceeds the above.

Marginal Model Plots: Assessing Mean

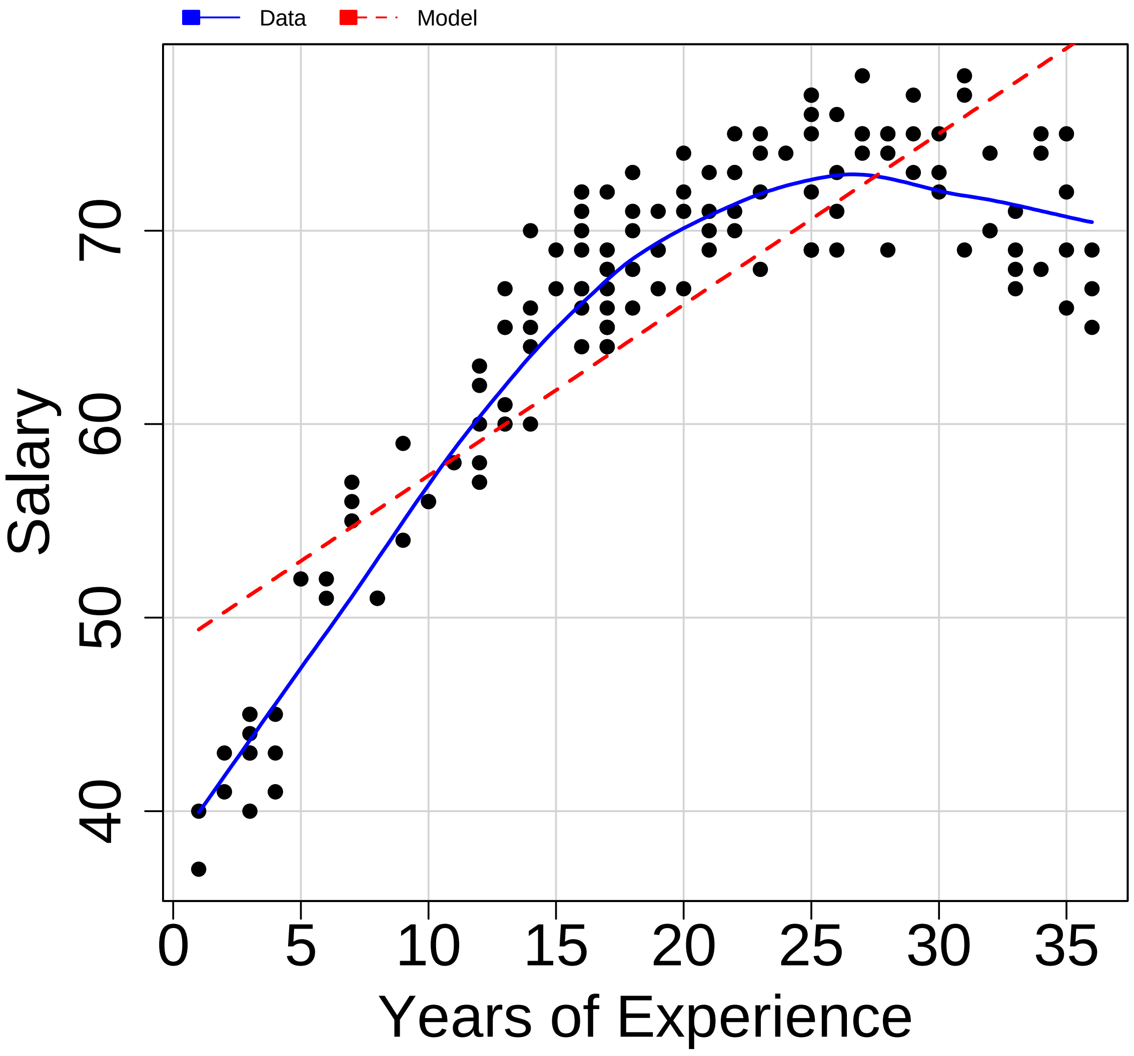
Does the simple linear regression model (1) model E[Y | X] adequately?

$$Y = \beta_0 + \beta_1 x + e \qquad (1)$$

One way to find out: Fit a nonparametric estimator like loess, and see whether it agrees with (1).

$$Y = f(x) + e \qquad (2)$$

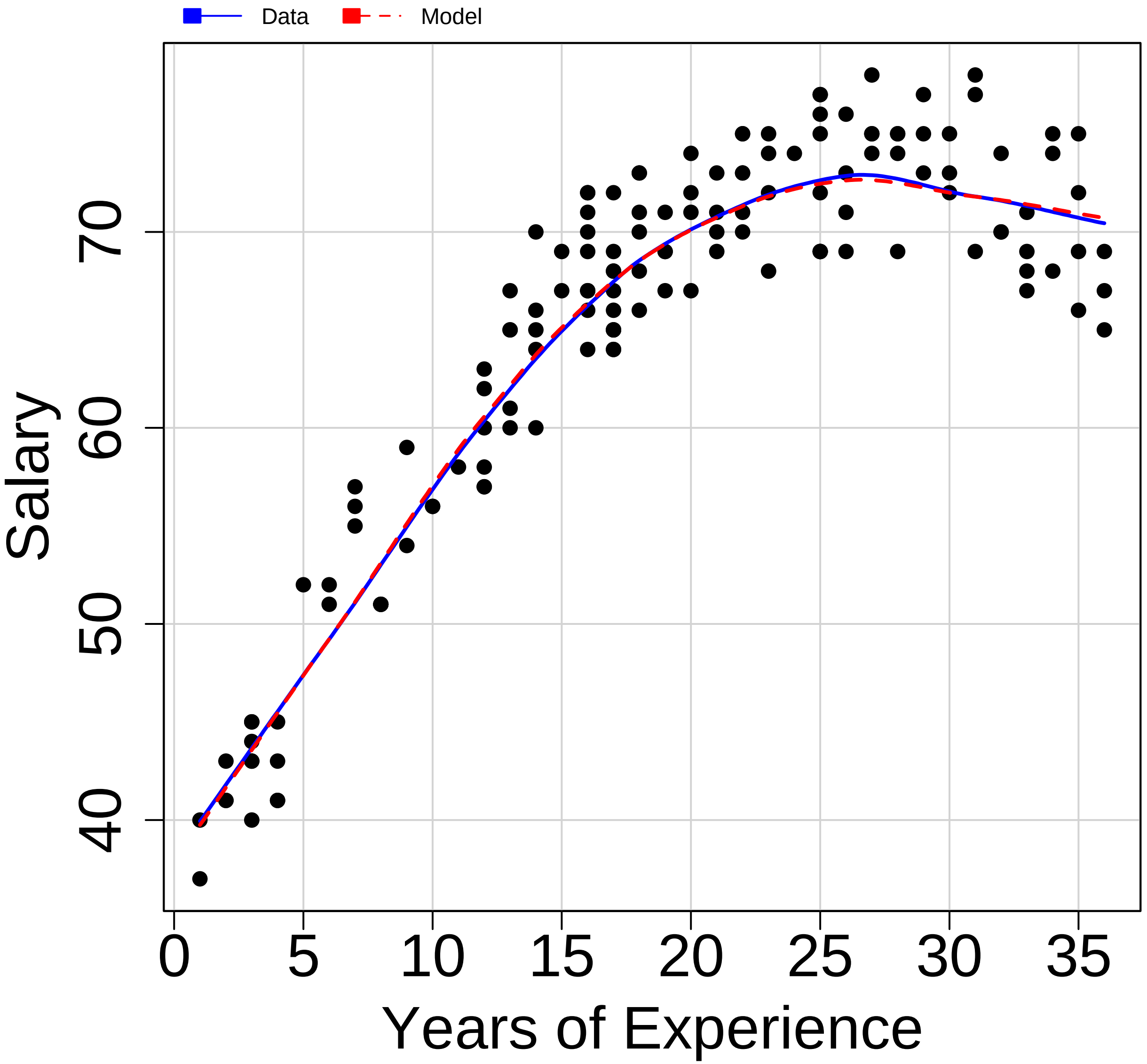
Marginal Model Plots: Assessing Mean



$$Y = \beta_0 + \beta_1 x + e$$

$$Y = f(x) + e$$

Marginal Model Plots: Assessing Mean



$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + e$$

$$Y = f(x) + e$$

Marginal Model Plots: Multiple Predictors

- If we have two predictor variables, we wish to compare models (1) and (2) below.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e \qquad (1)$$

$$Y = f(x_1, x_2) + e \qquad (2)$$

- It's less obvious what to do next. We don't want to make three-dimensional plots and we can't make k-dimensional plots in general.

Marginal Model Plots: Multiple Predictors

- Cook and Weisberg (1997) utilize the following result:

$$E[Y] = E\left[E[Y|X]\right]$$

- For our linear model context, we use:

$$E_1[Y|x_1] = E\left[E_1(Y|x)|x_1\right]$$

- To compare the left and right hand sides of the equation, we make two loess fits and compare to see that they match.

- Left hand side: Plot Y vs. x_1. Fit a loess smooth. Compare that fit to the right hand side (next slide), a plot of $\hat{y}$ from Model 1 against x_1.

Marginal Model Plots: Multiple Predictors

$$E_1[Y|x_1] = E\left[E_1(Y|x)|x_1\right]$$

- Right hand side (inside):

$$E_1[Y|x] = E_1(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + e|x)$$
$$= \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- This can be estimated by:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

- So we should plot the fitted values of model (1) against x_1, getting a loess smooth, and compare to the previous smooth of y vs. x_1.

Marginal Model Plots: Multiple Predictors

$$E_1[Y|x_1] = E\left[E_1(Y|x)|x_1\right]$$

- Proof of equality: Right hand side:

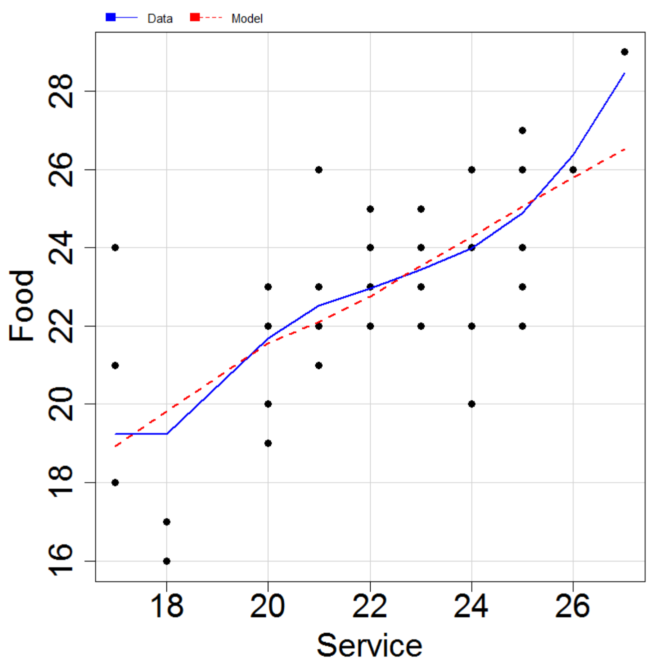
$$\begin{aligned} E\left[E_1[Y|x]|x_1\right] &= E(\beta_0 + \beta_1 x_1 + \beta_2 x_2 | x_1) \\ &= \beta_0 + \beta_1 x_1 + \beta_2 E[x_2|x_1] \end{aligned}$$

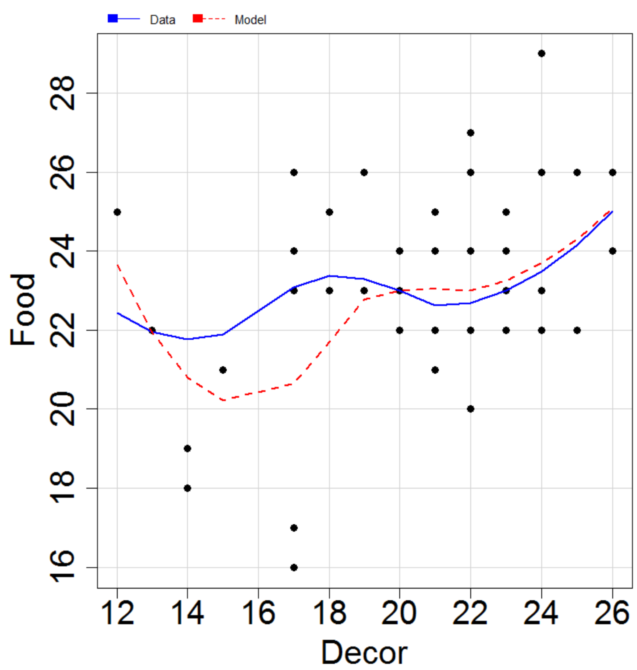
- Left hand side:

$$\begin{aligned} E_1\left[Y|x_1\right] &= E(\beta_0 + \beta_1 x_1 + \beta_2 x_2 | x_1) \\ &= \beta_0 + \beta_1 x_1 + \beta_2 E[x_2|x_1] \end{aligned}$$

Marginal Model Plots: Multiple Predictors

- It's easier to put both loess plots on the same graph.
- We repeat this marginal model plot for each predictor.

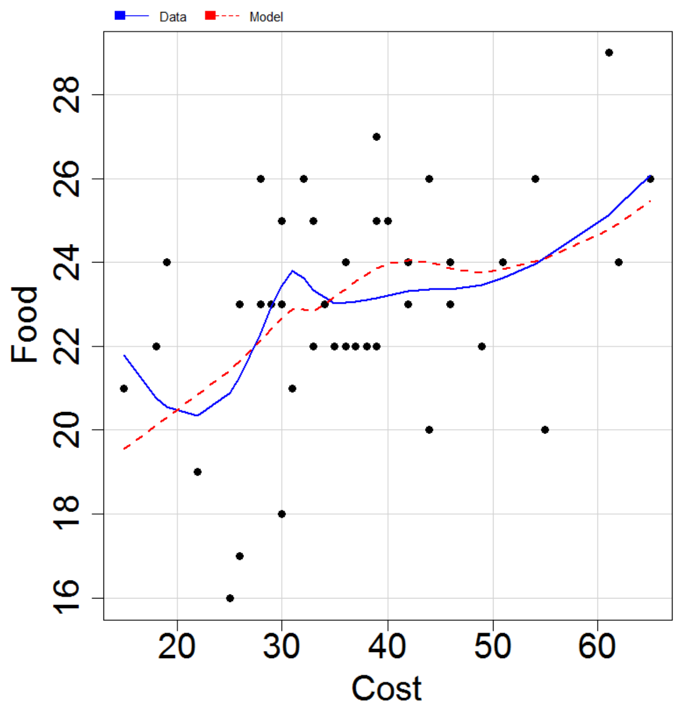




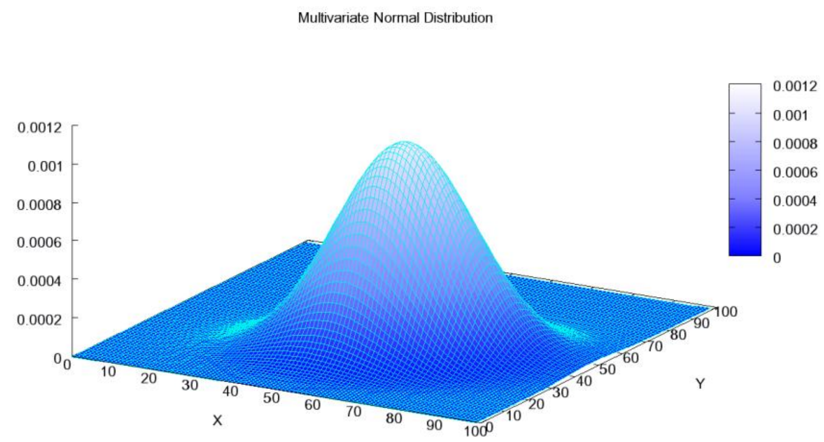
Marginal Model Plots: Multiple Predictors

- Since the two fits in the following plot differ markedly, we conclude that the model below is not a valid model for the data.

$$Food = \beta_0 + \beta_1 Service + \beta_2 Decor + \beta_3 Cost + e$$



Transformations (Box Cox: Approach 1)



- Step 1: Transform all of the predictors to multivariate normality

- Step 2: Transform Y given the predictors, so that the residuals are as normally distributed as possible. (I.e. Consider the model below.)

$$Y = g(\beta_0 + \beta_1 \psi_S(x_1, \lambda_{X_1}) + \ldots + \beta_p \psi_S(x_p, \lambda_{X_p}))$$

Transformations: State Spending

EX: Per capita state and local expenditures
ECAB: Economic ability index
MET: % of population in metropolitan areas
YOUNG: % of population aged 5 – 19 years
OLD: % of population aged 65 and older
WEST: 1 = western state, 0 = otherwise

+ Transformations: State Spending

- Step 1: Transform Predictors

```
yjPower Transformations to Multinormality

      Est.Power Std.Err. Wald Lower Bound Wald Upper Bound
MET      1.0268   0.1715           0.6908           1.3629
ECAB     1.2373   0.6815          -0.0985           2.5731
YOUNG    0.4653   1.3559          -2.1923           3.1229
OLD      1.9089   0.8089           0.3235           3.4943

Likelihood ratio tests about transformation parameters
                                  LRT df         pval
LR test, lambda = (0 0 0 0) 63.741169  4 4.738432e-13
LR test, lambda = (1 1 1 1)  1.452826  4 8.349635e-01
```

Transformations: State Spending

- Step 2: Transform Y, given predictors

```
lm.1<-lm(EX ~ MET + ECAB+ YOUNG+ OLD + WEST)
tranmod <- powerTransform(lm.1, family="yjPower")
summary(tranmod)
```

```
Est.Power Std.Err. Wald Lower Bound Wald Upper Bound
Y1    0.1668   0.5829           -0.9757           1.3094

Likelihood ratio tests about transformation parameters
                              LRT df      pval
LR test, lambda = (0) 0.08138015  1 0.7754357
LR test, lambda = (1) 2.10124406  1 0.1471793
```

Transformations: Using Logs for % Effects

$$log(Y) = \beta_0 + \beta_1\, log(x) + \beta_2 x_2 + e$$

$$\begin{aligned}
\beta_2 &= \frac{\Delta\, log(Y)}{\Delta\, x_2} \\
&= \frac{log(Y_2) - log(Y_1)}{\Delta\, x_2} \\
&= \frac{log(Y_2/Y_1)}{\Delta\, x_2} \\
&\approx \frac{Y_2/Y_1 - 1}{\Delta\, x_2} \text{ (using } log(1+z) \approx z \text{ and assuming } \beta_2 \text{ is small)} \\
&= \frac{100(Y_2/Y_1 - 1)}{100\Delta\, x_2} \\
&= \frac{\%\Delta Y}{100\Delta\, x_2}
\end{aligned}$$

- For every 1 unit change in x_2, the model predicts a $100\times\beta_2$% change in Y.
- For every 1% change in x_1, the model predicts a β_1% change in Y.

Logarithms and % Effects

$$log(SundayCirculation) = \beta_0 + \beta_1\, log(WeekdayCirculation) \\ + \beta_2 Tabloidwithcompetitor + e$$

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -0.44730    0.35138  -1.273    0.206
log(Weekday)    1.06133    0.02848  37.270  < 2e-16 ***
Tabloid        -0.53137    0.06800  -7.814 1.26e-11 ***
```

Because of the log transformation, the model above predicts:

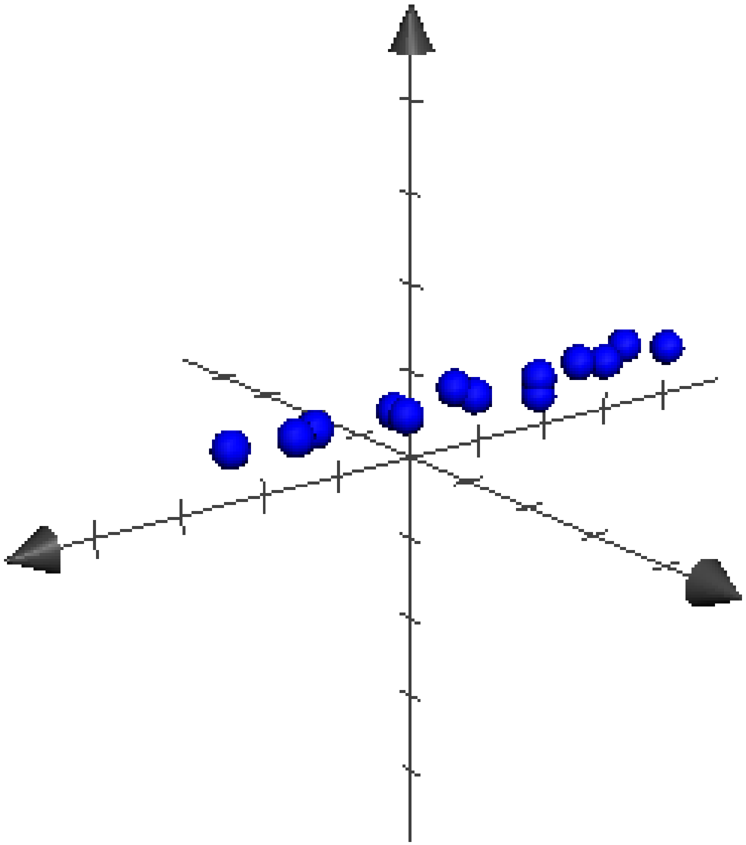
- A 1.06% increase in Sunday Circulation for every 1% increase in Weekday Circulation
- A 53.1% decrease in Sunday Circulation if the newspaper is a tabloid with a serious competitor

Multicollinearity

Multicollinearity: strong correlations between predictors

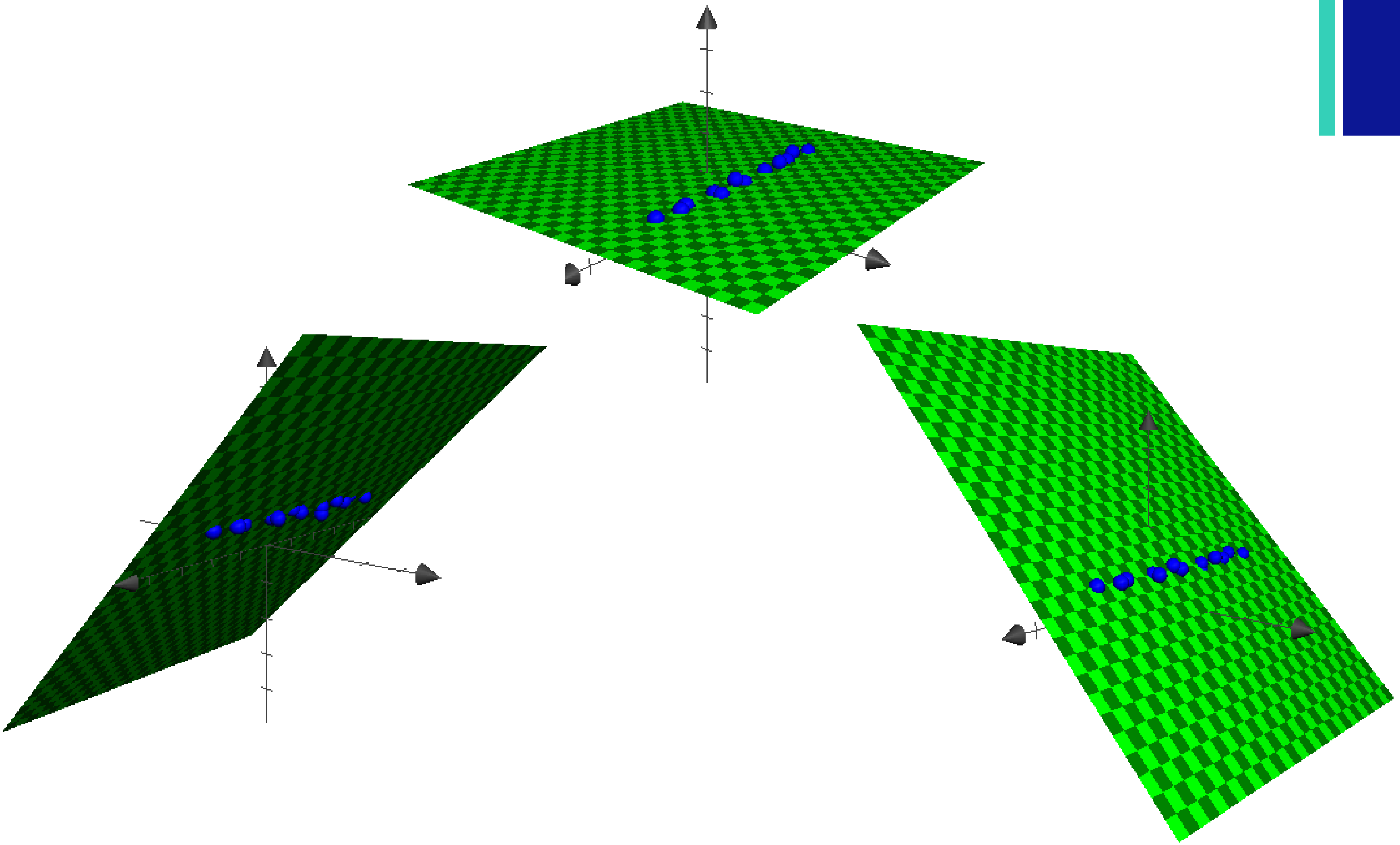
- Regression coefficients can have the wrong sign
- Many of the predictor variables may not be statistically significant when the overall F-test is highly significant.

+ Multicollinearity



- Problem: X1 and X2 are too strongly correlated with each other.
- When Rank(X) is not the number of columns of X, clearly we cannot estimate β.
- When the columns of X are pretty close to being linear combinations of one another:
  - The variables are effectively carrying very similar information about the response variable.
  - The parameter estimates become unstable, and variance estimates become large.

Multicollinearity



Added Variable Plots

- Goal: Find out whether X_2 adds anything to the model after X_1 has already been added.
- Idea: We're interested in the following Final Model:

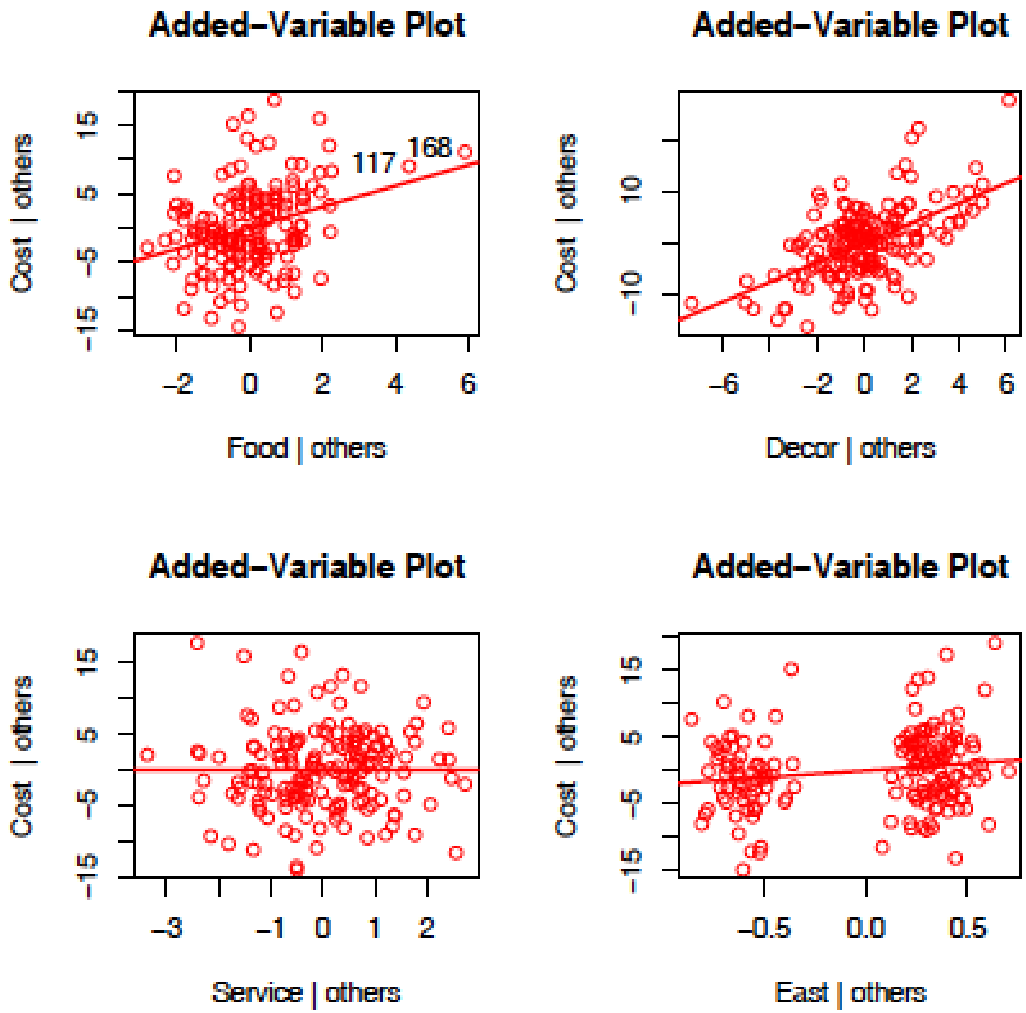
$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\alpha + \mathbf{e}$$

(Variable Z is a single variable, so α is a scalar.)

Added Variable Plots

Added Variable Plots

Added Variable Plots



Added Variable Plots

- Added variable plots enable us to visually assess the additional effect of each predictor, *after the others have been included in the model*.

- Added variable plots should display straight line relationships. If they don't, the model is misspecified.

- The slope from the added variable plot is the slope of the multiple linear regression model for that variable.

- The scatter of the points in the added variable plot visually indicates which points are most influential in determining the estimate of α.

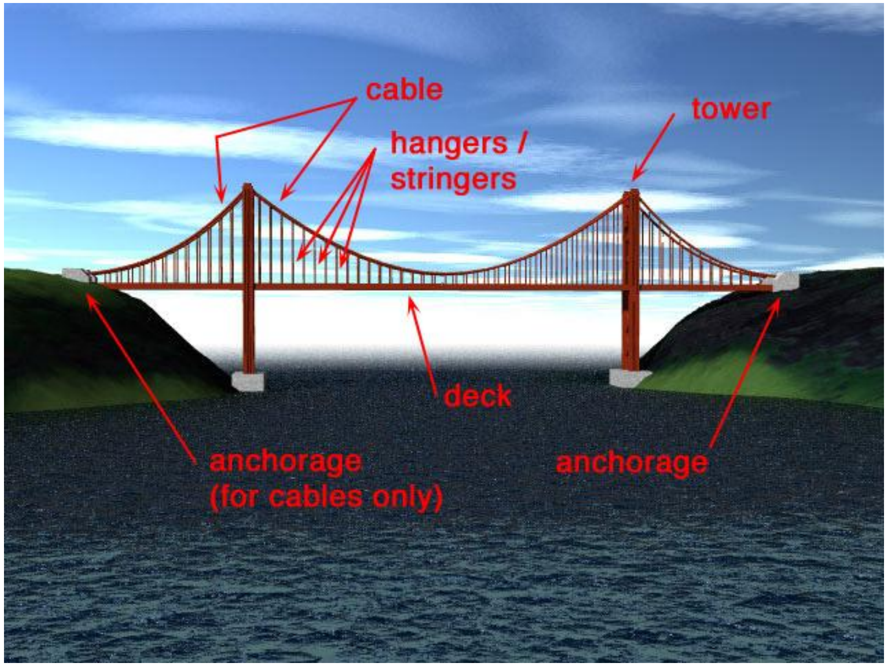
Review

- What do marginal model plots display vs. added variable plots?

a) Marginal model plots measure whether the mean is adequately modeled; added variable plots measure whether each variable contributes something the others don’t.

b) Marginal model plots measure whether each variable contributes something the others don’t; added variable plots measure additional contributions to the mean function.

Bridges



- Predicting design time of bridges is helpful for budgeting and scheduling purposes.

- The variables are as follows:
  - Y = Time = design time in person-days
  - Darea = Deck area of bridge
  - Ccost = Construction cost
  - Dwgs = Number of structural drawings
  - Length = Length of bridge
  - Spans = Number of spans (space between towers)

Summary

- Remember that scatterplot matrices in two dimensions and correlation matrices only measure whether each individual predictor is correlated with the others; it doesn't account for relationships like $x_1 = x_2 + x_5$.

- A consequence of multicollinearity is that the determinant of X'X is near 0; that means variances of the parameter estimates are going wild. One more point could completely change the parameter estimates from positive to negative.

- Multicollinearity can invalidate a model which is otherwise valid. Our bridges model is invalid.

Summary

- When two or more highly correlated predictor variables are included in a regression model, they are effectively carrying very similar information about the response variable. Thus, it is difficult for least squares to distinguish their separate effects on the response variable.

- In this situation the overall F-test will be highly statistically significant but very few of the regression coefficients may be statistically significant.

Multicollinearity and Variance Inflation Factors

Consider the multiple regression model:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + e$$

If R_j^2 denotes the value of R^2 from regressing x_j on the other predictors, then:

$$Var(\hat{\beta}_j) = \frac{1}{1 - R_j^2} \times \frac{\sigma^2}{(n-1)S_{x_j}^2}, j = 1 \ldots, p$$

The first fraction is called the j^{th} **variance inflation factor**. We say that our model has problems with multicollinearity if VIF > 5.

Omitted Variables

- **Spurious correlation** is found when two variables being studied are related because both are related to a third variable currently omitted from the regression model.
- Ex: Number of ice cream cones sold and number of shark attacks are positively correlated. Weather is called a **lurking variable**.
- Ex: Hormone replacement therapy and estrogen replacement therapy for women were associated with a lower risk of coronary heart disease. But in randomized controlled trials, the association wasn't found. Why not?

Omitted Variables

Model we should fit: $Y = \beta_0 + \beta_1 x + \beta_2 \nu + e_{Y \cdot x, \nu}$

Relationship between predictors: $\nu = \alpha_0 + \alpha_1 x + e_{\nu \cdot x}$

Model we actually fit if we don't use v:

$$Y = (\beta_0 + \beta_2 \alpha_0) + (\beta_1 + \beta_2 \alpha_1) x + (e_{Y \cdot x, \nu} + \beta_2 e_{\nu \cdot x})$$

- Two cases:
  - $\alpha_1 = 0$ and/or $\beta_2 = 0$: The omitted variable has no effect on the regression model that has only x.
  - $\alpha 1 \neq 0$ and $\beta_2 \neq 0$: The omitted variable does have an effect on the model that has only x.
    - Ex: Y and x could be highly correlated even when $\beta_1 = 0$.
    - Ex: Y and x could be strongly negatively associated even when $\beta_1 > 0$.