Applied Regression Modeling:
A Business Approach
Chapter 5: Regression Model Building II
Sections 5.3–5.4

by Iain Pardoe

Exploratory data analysis

1. Carefully frame your questions and identify the data that will help answer the questions.
2. Collect and organize data—usually the most time-consuming part of process (haven't said much about this step because our focus has been elsewhere).
3. Organize data for analysis—includes checking for mistakes and coding qualitative variables using indicator variables.
4. Graph data (using a scatterplot matrix, say) and calculate summary statistics to get a feel for dataset—this can highlight potential data mistakes and might suggest transformations (e.g., highly skewed variables are often best transformed to natural logarithms before further analysis).

Building model up

5. Fit initial model using quantitative predictors and qualitative indicator variables—use untransformed Y as response unless transformed Y suggested (e.g., replace highly skewed Y with $\log_e(Y)$).
6. Check four regression assumptions for initial model using residual plots—if one or more is clearly violated then proceed to step 7, otherwise if everything checks out proceed to step 8.
7. Add interactions/transformations so assumptions check out—start simple and try more complicated models as needed (once model is adequate and assumptions check out, proceed to step 8):
   - first try interactions between indicator variables and quantitative predictors (e.g., DX_1, DX_2);
   - if model still inadequate, next try adding/replacing transformations (e.g., add X_1^2, or replace X_1 with $\log_e(X_1)$ or $1/X_1$);
   - if model still inadequate, next try adding all-quantitative interactions (e.g., X_1X_2).

Simplifying model

8. Simplify model in stages (aim for parsimonious model that captures major, important population relationships between variables without getting distracted by minor, unimportant sample relationships):
   - evaluate each model to identify which predictors, interactions, and transformations to retain and which to remove—methods include R^2, adjusted R^2, regression standard error (s), and regression parameter hypothesis tests (global usefulness, nested model, individual);
   - remove all redundant predictors, interactions, and transformations (retain hierarchy and proceed a few predictor terms at a time).

Model interpretation

9. Evaluate the final model and confirm that the regression assumptions still hold.
10. During the whole process identify any outliers or other influential points, and look out for potential pitfalls (autocorrelation, multicollinearity, excluding important predictors, overfitting, missing data)—address any problems as they occur.
11. Interpret the final model, including understanding predictor effects on Y, and estimating expected values of Y and predicting individual values of Y at particular values of the predictors (taking care not to extrapolate outside the sample data region).
12. Interpret the final model using graphics—see next section.

- Examples—see case studies in Chapter 6.

5.4 Model interpretation using graphics

Predictor effect plots

- Transformations and interactions can enhance our ability to model multiple linear regression relationships.
- But they can make the resulting models more difficult to understand and interpret.
- One approach to interpreting such models is to use graphs to plot how Y changes as each predictor changes (and all other predictors are held constant)—called predictor effect plots.
- The use of the term "predictor effect" does *not* imply any kind of causal effect.

CREDITCARD example

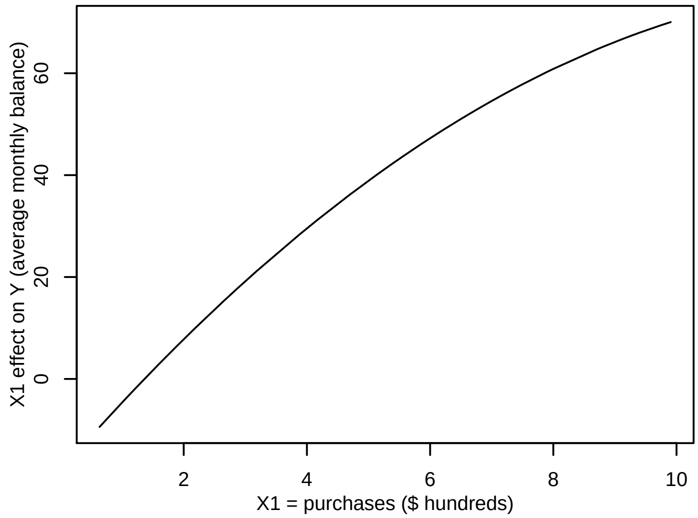
- Simulated **CREDITCARD** dataset contains average outstanding monthly balance (Y, in dollars) for 50 individual credit card accounts.
- Possible predictors include:
  - $X_1 =$ average monthly purchases (in hundreds of dollars);
  - $X_2 =$ average monthly housing expenses (in hundreds of dollars);
  - D_3 is an indicator variable that is 1 for renters and 0 for homeowners;
  - D_4 is an indicator variable that is 1 for males and 0 for females.
- Final model results: $\hat{Y} = 14.35 + 13.94X_1 - 4.92X_2 + 13.15D_3$
  $-5.37D_4 + 12.65D_3X_2 - 0.51X_1^2$.

$\mathbf{X_1}$ effect

- Consider how Y changes as X_1 changes.
- Isolate this change in Y when we hold X_2, renter status, and gender constant:
  - hold X_2 constant at sample mean of X_2, $m_{X_2} = 6.5882$;
  - hold renter status constant at reference level, $D_3 = 0$ (homeowners);
  - hold gender constant at reference level,
    $D_4 = 0$ (female).
- $$\begin{aligned} X_1 \text{ effect} &= 14.348 + 13.94X_1 - 4.919(6.5882) + 13.15(0) \\ &\quad -5.37(0) + 12.65(0)(6.5882) - 0.51X_1^2 \\ &= 14.348 + 13.94X_1 - 32.407 - 0.51X_1^2 \\ &= -18.06 + 13.94X_1 - 0.51X_1^2. \end{aligned}$$

$\mathbf{X_1}$ effect plot

X_1 effect $= -18.06 + 13.94X_1 - 0.51X_1^2$.



Interpreting $\mathbf{X_1}$ effect

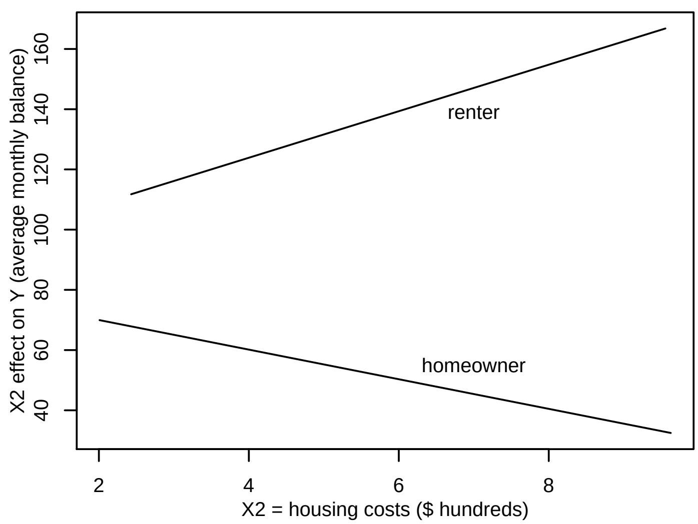
- High spending female homeowners with average monthly housing expenses tend to carry a higher credit card balance than low spenders, but balance increases become smaller the more that is spent each month.
- When X_1 increases from \$200 to \$300, expect Y to increase $(-18.06 + 13.94(3) - 0.51(3^2)) - (-18.06 + 13.94(2) - 0.51(2^2)) = \$11.39$.
- When X_1 increases from \$800 to \$900, expect Y to increase $(-18.06 + 13.94(9) - 0.51(9^2)) - (-18.06 + 13.94(8) - 0.51(8^2)) = \$5.27$.
- These increases also hold for other values of X_2, D_3, and D_4 (e.g., try male renters with $X_2 = 4$).
- So, vertical differences between points along line represent changes in Y as X_1 changes for *all* individuals in the population.

$\mathbf{X_2}$ effect

- Consider how Y changes as X_2 changes.
- Since X_2 is included in an interaction term, D_3X_2, we need to take into account renter status.
- Isolate this change in Y when we hold X_1 and gender constant:
  - hold X_1 constant at sample mean of X_1, $m_{X_1} = 6.0242$;
  - hold gender constant at reference level, $D_4 = 0$ (female).
- $$\begin{aligned} X_2 \text{ effect} &= 14.348 + 13.937(6.0242) - 4.92X_2 + 13.15D_3 \\ &\quad -5.37(0) + 12.65D_3X_2 - 0.509(6.0242^2) \\ &= 14.348 + 83.959 - 4.92X_2 + 13.15D_3 \\ &\quad +12.65D_3X_2 - 18.472 \\ &= 79.84 - 4.92X_2 + 13.15D_3 + 12.65D_3X_2. \end{aligned}$$

$\mathbf{X_2}$ effect plot

X_2 effect $= 79.84 - 4.92X_2 + 13.15D_3 + 12.65D_3X_2$.



Interpreting $\mathbf{X_2}$ effect

- Renters tend to carry a higher credit card balance than homeowners, and renters' balances increase as housing expenses increase, whereas homeowners' balances decrease as housing expenses increase.
- An additional $\$100$ in housing expenses tends to increase the monthly balance for a renter by $\$7.73$, but it tends to decrease the monthly balance for a homeowner by $\$4.92$.
- Expected difference in Y between renters ($D_3=1$) and homeowners ($D_3=0$):

$$\begin{aligned}&(14.35 + 13.94X_1 - 4.92X_2 + 13.15(1) - 5.37D_4 + 12.65(1)X_2 - 0.51X_1^2)\\ &- (14.35 + 13.94X_1 - 4.92X_2 + 13.15(0) - 5.37D_4 + 12.65(0)X_2 - 0.51X_1^2)\\ &= 13.15 + 12.65X_2.\end{aligned}$$

- In other words, renters tend to carry higher credit card balances and the difference becomes greater as average monthly housing expenses increases.

Gender effect

- Expected difference in Y between males ($D_4=1$) and females ($D_4=0$):

$$\begin{aligned}&(14.35 + 13.94X_1 - 4.92X_2 + 13.15D_3 - 5.37(1) + \ldots)\\ &- (14.35 + 13.94X_1 - 4.92X_2 + 13.15D_3 - 5.37(0) + \ldots).\\ &= -5.37.\end{aligned}$$

- Thus, males tend to carry lower balances on their credit cards than women ($\$5.37$ lower on average).