MSDS 660: Statistics and Experimental Design
Week 6

Correlation, Simple Linear Regression and Multiple Linear Regression

kpolson@regis.edu

Recap

Two-way ANOVA model

- What’s the alpha*beta term?

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

where μ is the grand mean
α*i* is the *ith* level of first factor
β*j* is the *j*th level of the second factor
k refers to the *kth* observation within the (*ij*)cell
E_{ijk} *is the error term*
Y_{123} = Refers to the 3rd observation in the first level of factor 1 and the second level of factor 2.

Two-way ANOVA model

- What’s the alpha*beta term?

– interaction (synergy)

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

- Their combined interaction have an effect on the outcome of y.

Agenda

- Understand the meaning of Correlation
- Understand and apply Simple Linear Regression
- Understand when a Multiple Linear Regression (MLR) model is appropriate
- Evaluate an MLR model based on scoring metrics

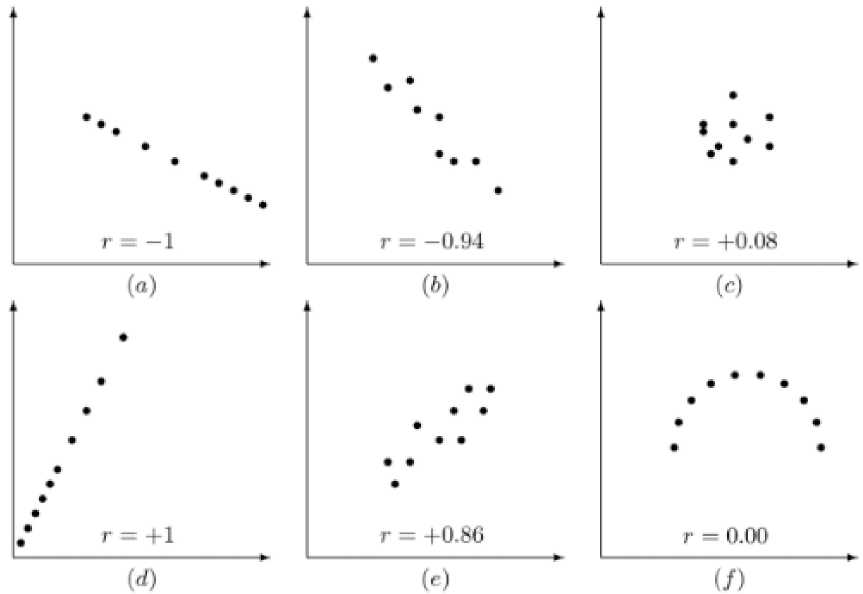
Correlation

A measure of the degree of relationship between two variables

- Linear correlation

What does it mean when two variables are correlated?

- The values of the two variables vary together in a systematic way.



Pearson Correlation Coefficient

$$\underline{r = \Sigma\, (x - \bar{x})(y - \bar{y})}$$

$$n\ s_x\ s_y$$

where $\bar{x}$ and $\bar{y}$ are the sample means and $s_x\ s_y$ are standard deviations for each variable.

It is a value that expresses quantitatively the magnitude and direction of the relationship.

Correlation and Causation

Just because two variables are highly correlated doesn’t mean that one caused the other.

There are five possible explanations for why there’s a correlation between *X* and *Y* :

- *X* causes *Y*
- *Y* causes *X*
- Two variables could relate systematically, meaning that they work together to cause a change.
- A third factor (*Z*) or group of factors (*a*, *b*, *c*, *d*) causes both *X* and *Y*.
- Correlation is spurious (accidents of sampling; chance)

Simple Linear Regression

Managerial decisions often are based on the relationship between two or more variables

Regression analysis can be used to develop an equation showing how the variables are related.

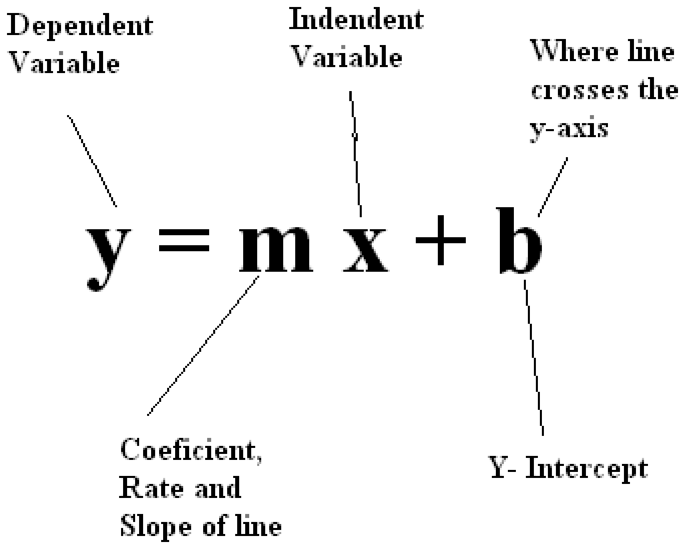
The variable being predicted is called the dependent variable and is denoted by y.

The variables being used to predict the value of the dependent variable are called the independent variables and are denoted by x.

Examples: Salary vs. Years Experience

Interest Rate vs. Loan Fee

Linear Regression math



$$y = m\,x + b$$

Dependent Variable

Indendent Variable

Where line crosses the y-axis

Coeficient, Rate and Slope of line

Y- Intercept

Linear regression math

$$Y \approx \beta_0 + \beta_1 X.$$

Approximately modeled as

$$\texttt{sales} \approx \beta_0 + \beta_1 \times \texttt{TV}.$$

Sales are approximately modeled as TV advertising spent

Linear regression notation

- Anything with a ‘hat’ is a parameter. Hat parameters have been fit or predicted from the data.

$$Y \approx \beta_0 + \beta_1 X.$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Y are approximately modeled as TV advertising spent

Error term

- Where did that e come from? What are the assumptions behind it?

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

Error term

- Where did that e come from? What are the assumptions behind it?
- Catch-all for what we miss with the model.
- Is it really linear? Missing variables? Measurement error? Uncertainty?
- It’s an unobserved random error term. It should be mean-zero and follow a normal distribution.

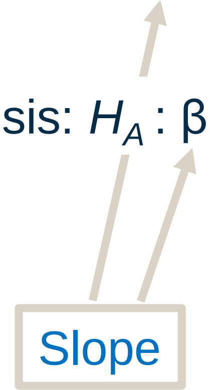
$$Y = \beta_0 + \beta_1 X + \epsilon$$

t-test with slope/intercept

- How do we know if there is a relationship between the x and y variables?

Null hypothesis: H_0 : $\beta_1 = 0$: There is no relationship between x and y.

Alternative hypothesis: H_A : $\beta_1 \neq 0$: There is some relationship between x and y.



Slope

t-test with slope/intercept

- How do we know if β_1 is measuring a relationship between x and y by chance?

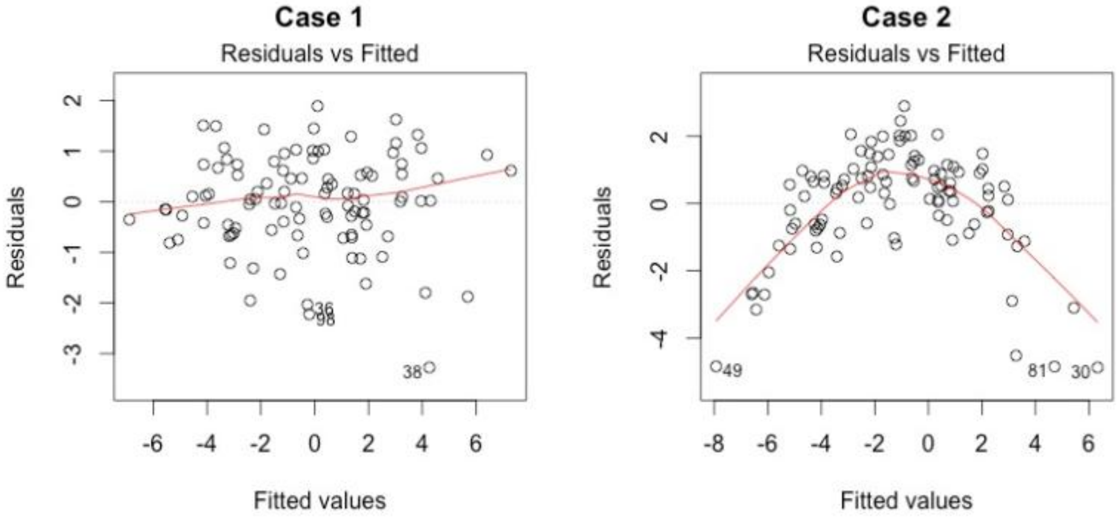
We calculate the t-value.
A high t-value corresponds to a small p-value and we reject the null hypothesis
Typically, we reject the null hypothesis if the p-value is < 0.05

Regression Assumptions

- Relationship is linear between predictor and outcome variable (fitted values vs residuals plot, or just predictors vs outcome)
- Residuals are normally distributed (qqplot should be close to a diagonal line)
- Homoscedasticity: constant variance of errors over range of predicted variable (plot fitted values vs residuals)
- No outliers with large leverages (Cook’s distance plot)

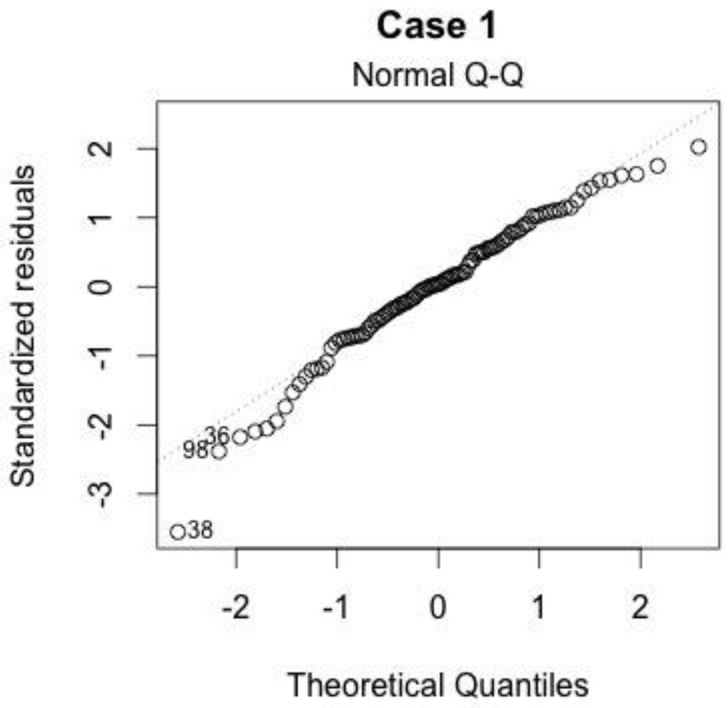
Residual vs. Fitted Values Plot

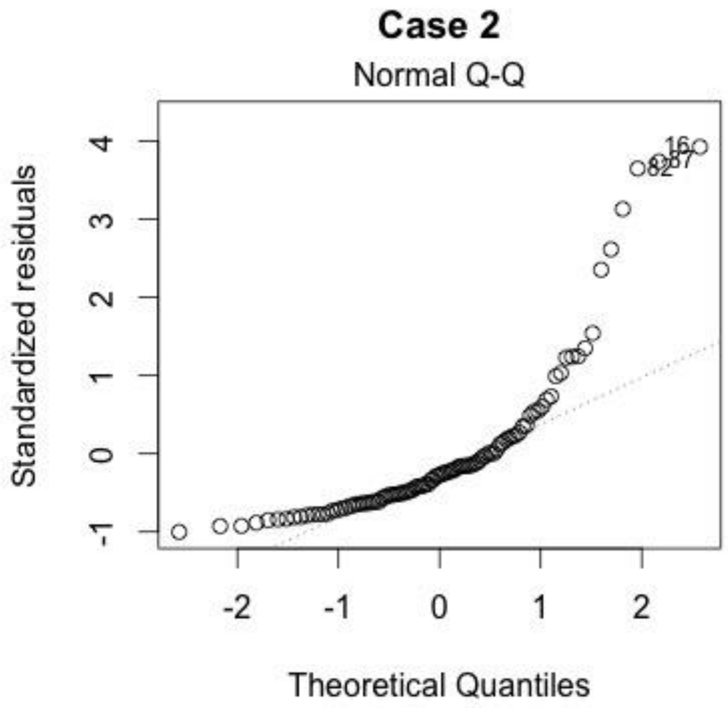
- If residuals are evenly distributed horizontally, it is a good indication your model is in fact linear.
- If residuals are parabolic there is a good chance your model is non-linear



Normal Q-Q Plot

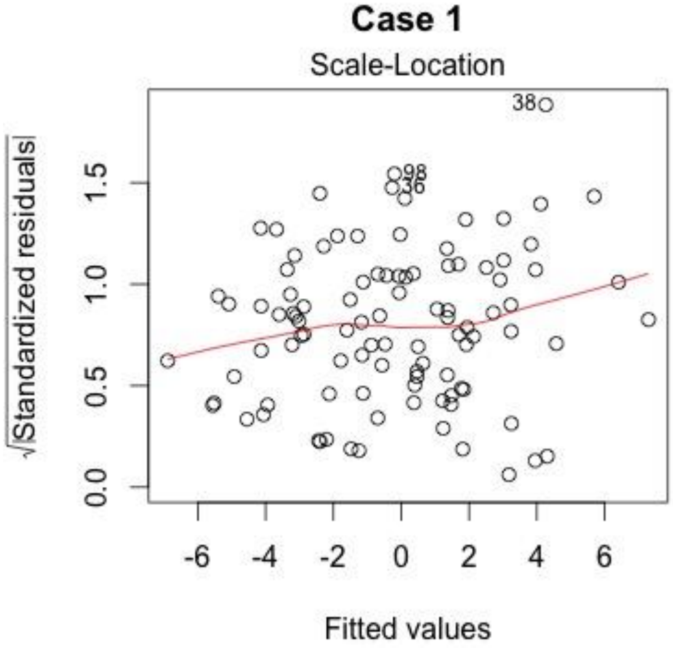
- Shows if residuals are normally distributed.
- Want residuals fitting on a straight line in QQ plots

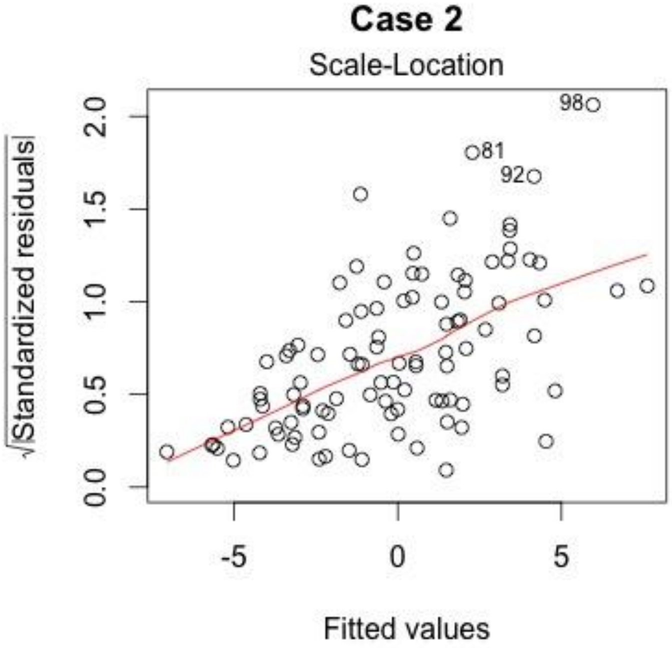




Scale - Location

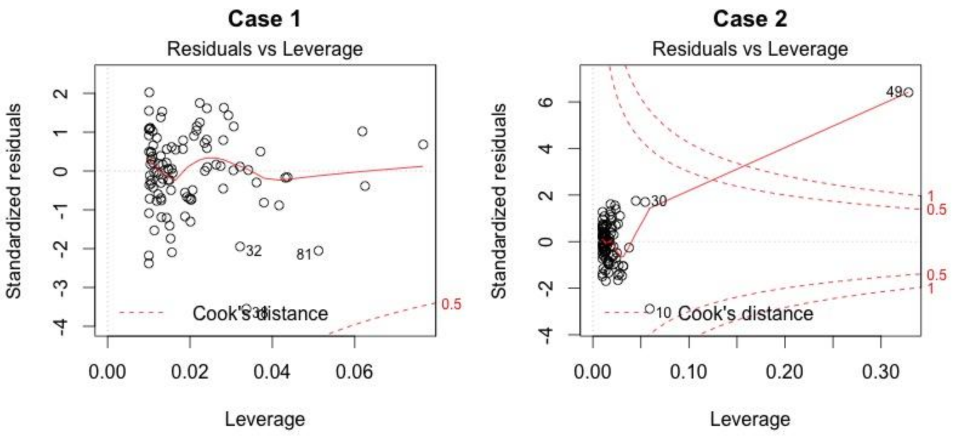
- Check for equal variance (homoscedasticity)
- Very similar to residual vs fitted. Want no discernable outliers or pattern to the plot.





Residuals vs Leverage

- Used to see which points have the greatest leverage or influence on the model.
- Look for points in the upper right or lower right corners. Or look for points outside the ‘Cook’s” distance. These points have the greatest leverage on the model.



Multiple Linear Regression (MLR)

- We have more than one predictor variable we want to use, and we want the results to be easy to interpret/explain.
- Same basic equation as linear regression
- Remember the difference?
- The top equation is the model, the bottom equation is what we use to make predictions

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

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

Significant Coefficients

Null hypothesis is that all the coefficients are 0

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

Alternative is

$$H_a : \text{ at least one } \beta_j \text{ is non-zero}$$

Instead of a t-statistic, we use F-stat in MLR

- *p* is the number of predictors (betas)
- *n* is the sample size
- RSS refers to the Residual Sums of Squares
- TSS refers to Total Sums of Squares

$$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n - p - 1)}$$

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\text{TSS} = \sum (y_i - \bar{y})^2$$

What’s a way to describe the F-stat in common language?

- Is roughly the ratio of:

the difference between the variance of the data and size of the residuals divided by the size of the residuals

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}/(n - p - 1)}$$

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

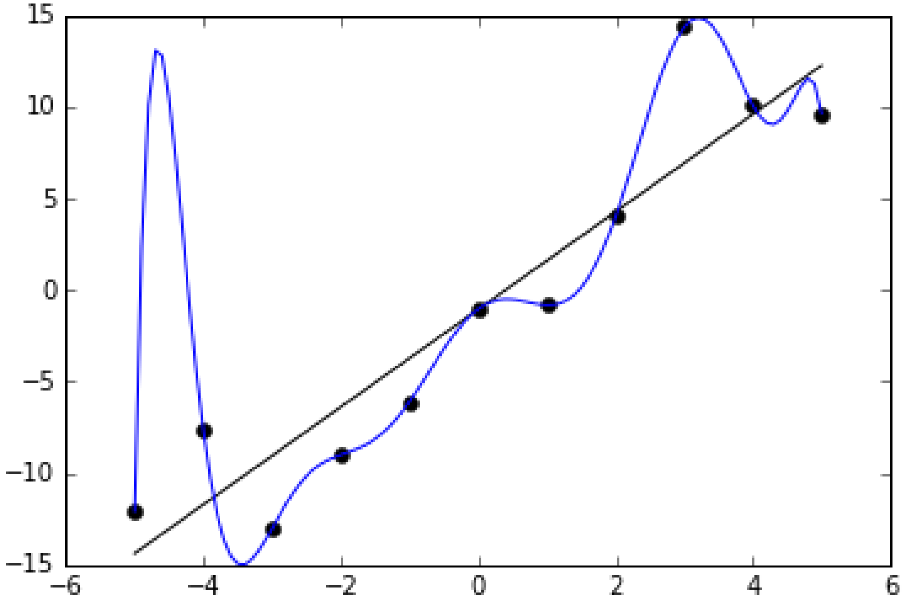
$$\mathrm{TSS} = \sum (y_i - \bar{y})^2$$

Evaluation Metrics in MLR

- Higher value is better:
  - F-statistic
  - Adjusted R^2
- Lower value is better:
  - Residual standard error (RSE)
  - Akaike Information Criterion (AIC)
  - Bayesian Information Criterion (BIC)
- Also CIC, DIC...many more

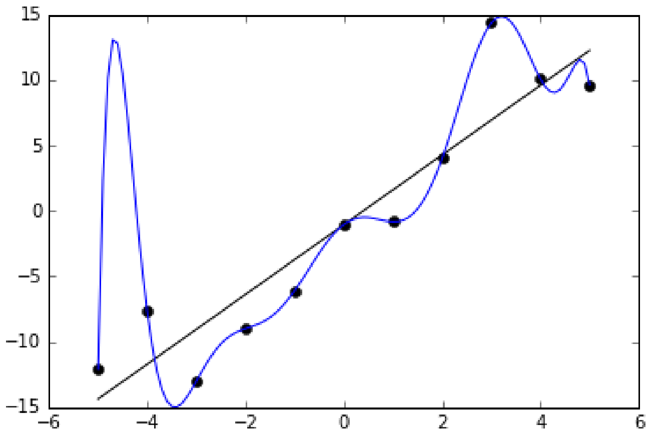
What’s wrong with R^2?

- Linear fit
  - (y = B0 + B1*x) → $R^2 \approx 0.7$
- Polynomial fit
  - (y = B0 + B1*x + … + B9*x^9) → $R^2 \approx 1$



What’s wrong with R^2?

- The highest R^2 value is the blue line, so use that as our model, right?
- Wrong. We’re overfitting – our model is too complex and it won’t predict new data well.
- Bias = Underfitting. High bias in algorithm misses relevant relationships
- Variance = Overfitting. High variance causes algorithm to model random noise.



RSE is a little better than R^2

- Lower is better
- Penalty for adding more predictor variables (*p*)

$$\mathrm{RSE} = \sqrt{\frac{1}{n-p-1}\mathrm{RSS}}.$$

Adjusted R^2

- Closer to 1 is better
- Penalty for adding predictors
- “Adjusted R^2 not as well motivated in statistical theory as Cp, AIC, and BIC” - ISLR

$$\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)}$$

Akaike Information Criterion (AIC)/Bayesian Information Criterion (BIC)

- Penalty for adding more predictors (*d*)
- Lower is better

$$\mathrm{AIC} = \frac{1}{n\hat{\sigma}^2}\left(\mathrm{RSS} + 2d\hat{\sigma}^2\right)$$

$$\mathrm{BIC} = \frac{1}{n}\left(\mathrm{RSS} + \log(n)d\hat{\sigma}^2\right)$$

Ways to find the right number of variables

- Best subset selection = 2^p different combination of models. $2^{30} > 1x10^9$ models!
- We use the stepAIC() function
- Backward search (default)
  - Start with all variables, remove one by one and evaluate metrics.
- Forward search
  - Start with one variable, add others and evaluate metrics.
- Both
  - A combination of forward and backward selection

An Introduction to Statistical Learning with applications in R: ISLR: pg. 227

Qualitative predictors? → Add dummy variables

- For qualitative predictors we need to make dummy variables.
- Cannot use factors because factors imply ordinality to data.
- Typically, there is one less dummy variable than number of levels.
    - The last dummy variable is incorporated into the intercept.
    - fastDummies library will make all of the dummies for you

Diagnostics: Collinearity

- Two or more predictor variables are closely related.
- It can be difficult to separate the individual effects of collinear variables and can mask the variable importance.

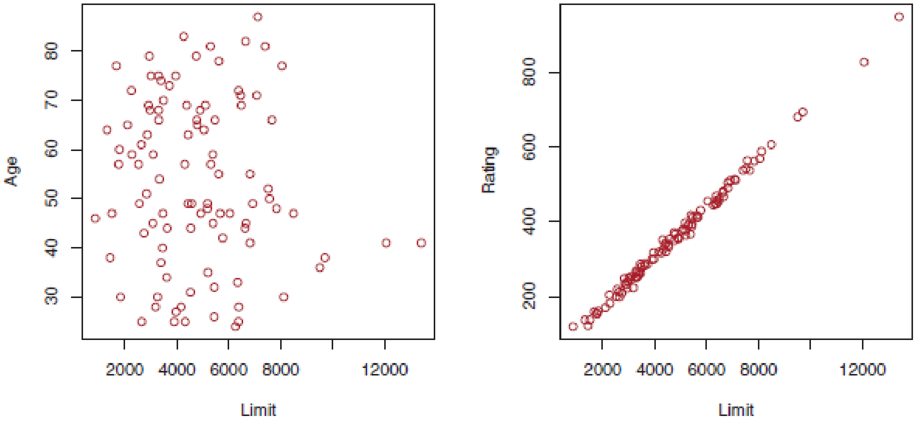
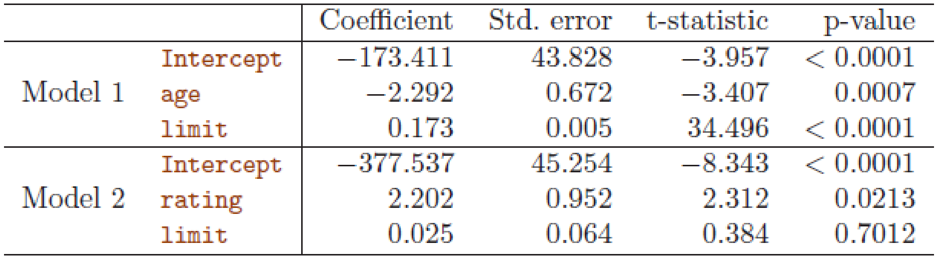


FIGURE 3.14. *Scatterplots of the observations from the* Credit *data set.* Left: *A plot of* age *versus* limit. *These two variables are not collinear.* Right: *A plot of* rating *versus* limit. *There is high collinearity.*



		Coefficient	Std. error	t-statistic	p-value
Model 1	Intercept	−173.411	43.828	−3.957	< 0.0001
	age	−2.292	0.672	−3.407	0.0007
	limit	0.173	0.005	34.496	< 0.0001
Model 2	Intercept	−377.537	45.254	−8.343	< 0.0001
	rating	2.202	0.952	2.312	0.0213
	limit	0.025	0.064	0.384	0.7012

TABLE 3.11. *The results for two multiple regression models involving the* Credit *data set are shown. Model 1 is a regression of* balance *on* age *and* limit*, and Model 2 a regression of* balance *on* rating *and* limit*. The standard error of* $\hat{\beta}_{\text{limit}}$ *increases 12-fold in the second regression, due to collinearity.*

Diagnostics: Variance Inflation Factor (VIF)

- Plot a correlation plot
- A better approach is to compute VIF (variance inflation factor)
    - The ratio of variance in a model with multiple terms, divided by the variance of a model with one term alone.
    - Smallest value VIF can be is 1.
    - VIF over 5 or 10 indicates a problematic amount of collinearity.
        - Page 101-102 in ILSR 4th edition.

Action to remedy collinearity:

- Remove a variable based on VIF.
- Or combine the VIF collinear variables into a new variable.

Summary on MLR development

Predicting an outcome from multiple independent variables.

1. Check for multicollinearity:
   - plot everything vs. everything and eyeball trends
   - check correlations (Pearson is used for numerical values, the default for cor() in R)
   - check VIF (variance inflation factor): value above 5 or 10 is bad
2. Minimize residuals:
   - Adjusted $R^2 \rightarrow 1$
   - RSE → 0
3. stepAIC() to remove unnecessary variables:
   - then check p-values on variables, any greater than 0.05 try dropping and see how metrics change
   - repeat all steps until model does not improve.
4. Check p-value < 0.05 to reject null hypothesis
   - less than 5% chance X and Y are correlated by chance

Let’s do an MLR in R

- Get Lab_wk6.Rmd from WorldClass and file=loans_full_schema.csv

Discussion Activity

1. Pick another DV (not total_credit_limit) and make a hypothesis of variables that maybe related. You need to include at least 3 IVs in the analysis.

2. Run several MLR models. Be sure to consider if you need to add/remove or transform variables.

3. Perform tests of diagnostics, i.e. with plot(), correlation, and vif.

4. Are there variables currently not in the data set that may be beneficial to your analysis? Does your initial hypothesis hold?

5. Post your Rmd and knitted file and responses to the Week 6 discussion.

Assignment

Objective: Find the best multilinear regression model on the marketing data set or on the dataset of your choosing. (Data sources on next slide)

You must include:

1. Box plot and histogram of the dependent variable
2. A correlation plot of all the numerical variables
3. A MLR model that has summary, residual plots, and a VIF analysis
4. Comment in the R code a justification on why you removed, combined, or left all variables/observations.
5. Kudos bonus if you create dummy variables from character values.
6. a stepAIC analysis and AIC scores of all models built
7. Summary, residual plots, and a VIF analysis of the final best model you created.
8. A brief summary that includes:

- An interpretation of the model and which variable is most positively correlated to your dependent variable and which variable is most negatively correlated with the dependent variable. Are there variables currently not in the data set that may be beneficial to your analysis? Does your initial hypothesis hold?
- Submit an Rmd file and a knitted pdf file to the Assignment dropbox.

Data Sources

UCI data repository

Kaggle datasets

data sets included in base R