STAT 139: STATISTICAL SLEUTHING THROUGH LINEAR MODELS

Lecture 20
Nov 11, 2014

Victoria Liublinska

Previous lecture: Review

- Lack-of-fit *F*-test for checking the goodness of fit of the linear regression model.
  - Possible when a pure error is estimable (i.e., when some levels of X have more than one observed Y).
  - Three models for population means:

    1. $\mu(Y \mid X = x_i) = \mu_i$
    2. $\mu(Y \mid X = x_i) = \beta_0 + \beta_1 x_i$
    3. $\mu(Y \mid X = x_i) = \beta_0$
  - Models 1 and 2 may be compared using the lack-of-fit *F*-test,

$$R = \frac{(\text{SSRes}_{\text{LR}} - \text{SSRes}_{\text{SM}})/(\text{d.f.}_{\text{LR}} - \text{d.f.}_{\text{SM}})}{S_p^2}, \text{ where } S_p^2 = \frac{\text{SSRes}_{\text{SM}}}{\text{d.f.}_{\text{SM}}}$$

  - Models 2 and 3 are compared with an *F*-test in the default output of the `lm()` function.
  - Models 1 and 3 may be compared using a standard ANOVA *F*-test.

Previous lecture: Review

- Multiple Linear Regression
  - How Much Is Your Car Worth?

A representative sample of > 800 GM cars (2005) was selected, then retail price was calculated from the Kelly Blue Book.

Variables:

- Price: suggested retail price of the used 2005 GM car in excellent condition.
- Mileage: number of miles the car has been driven.
- Cruise: indicator variable representing whether the car has cruise control (1 = cruise).
- Type: body type such as Convertible, Coupe, Hatchback, Sedan, Wagon.

Today's overview

Multiple Linear Regression:

- Model
- Assumptions
- Graphical methods for data exploration & model checking
- Specially constructed explanatory variables
  - Interaction term;
  - Quadratic & polynomial term;
  - Sets of indicator variables for categorical variable with >2 categories.

Reading:

- **Required:** R&S Ch. 9; Ch. 9 R code
- **Supplementary Theory**: A. Sen and M. Srivastava. "Regression Analysis: Theory, Methods, and Applications", Ch. 2 (ignore Sec. 2.5, 2.9, 2.12), Ch. 4-6, and Ch. 9.

Modeling Price: Joint Model

$$\mu(\text{Price} \mid \text{Mileage, Cruise}) = \beta_0 + \beta_1 \text{Mileage} + \beta_2 \text{Cruise}$$

```
> regmodel_both <- lm(Price ~ Mileage + Cruise, data = CarData)
summary(regmodel_both)
...
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 17537.2052   966.4606  18.146  < 2e-16 ***
Mileage        -0.1857     0.0379  -4.898 1.17e-06 ***
Cruise       9950.5457   719.4055  13.832  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8801 on 801 degrees of freedom
Multiple R-squared: 0.2093,   Adjusted R-squared: 0.2073
F-statistic:   106 on 2 and 801 DF,  p-value: < 2.2e-16
```

Modeling Price: Separate Models vs. Joint Model

$$\hat{\mu}(\text{Price} \mid \text{Mileage, Cruise}) = 17{,}537 - 0.186 \cdot \text{Mileage} + 9{,}951 \cdot \text{Cruise}$$

Unadjusted model:

β_1=-0.173

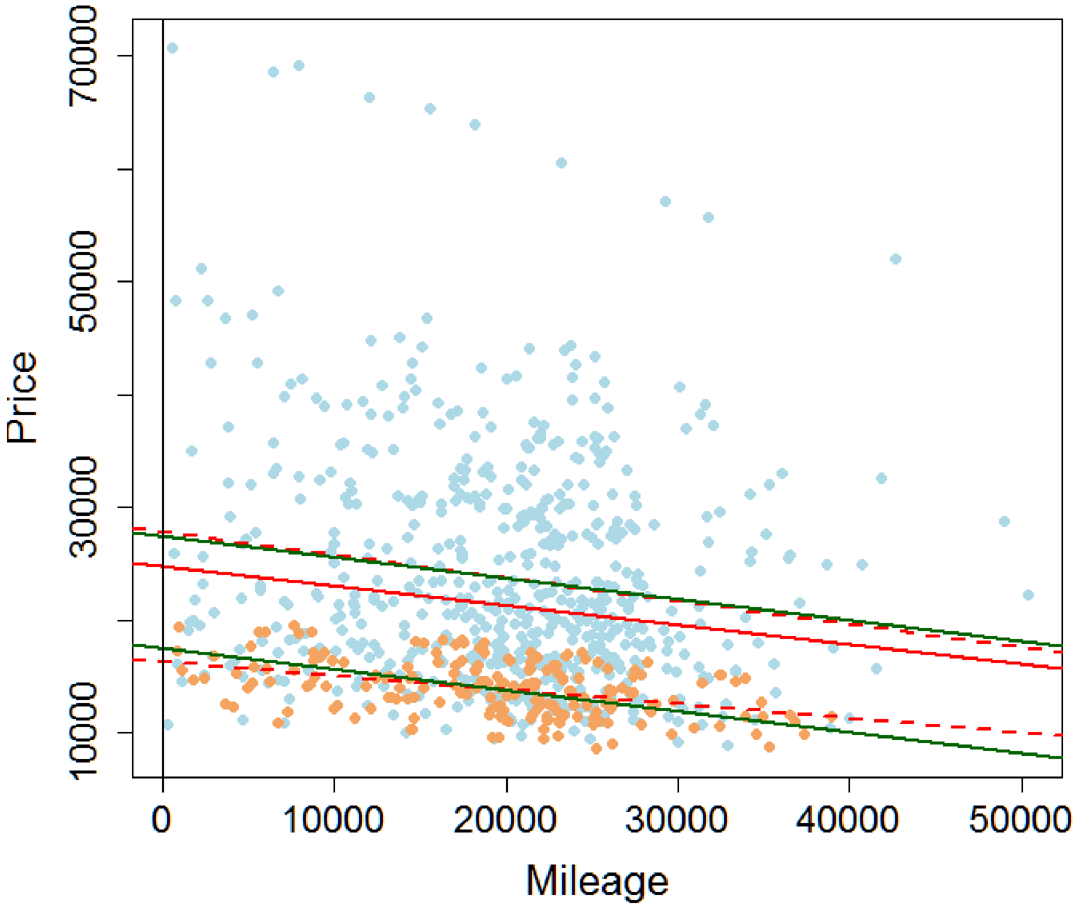
Separate models:

CC:

β_1=- 0.205

No CC:

β_1=-0.126



Joint model :

CC:

β_1=-0.186

No CC:

β_1=-0.186

Compare to Slide 39 in Lecture 19

Modeling Price: Joint Model with Interaction

$$\mu(\text{Price} \mid \text{Mileage, Cruise}) = \beta_0 + \beta_1 \text{Mileage} + \beta_2 \text{Cruise} + \beta_3 \text{Mileage} \cdot \text{Cruise}$$

- Two explanatory variables interact if the effect of one on mean response depends on the value of the other.
- An explanatory variable for interaction can be constructed by *multiplying* the two interacting variables.

Modeling Price: Joint Model with Interaction

$$\mu(\text{Price} \mid \text{Mileage}, \text{Cruise}) = \beta_0 + \beta_1 \text{Mileage} + \beta_2 \text{Cruise} + \beta_3 \text{Mileage} \cdot \text{Cruise}$$

```
> regmodel_both <- lm(Price ~ Mileage + Cruise +
  Mileage:Cruise, data = CarData)
summary(regmodel_both)
...
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       16,380  1.622e+03  10.098  < 2e-16 ***
Mileage          -0.126   7.689e-02  -1.642    0.101
Cruise            11,490  1.875e+03   6.128  1.4e-09 ***
Mileage:Cruise   -0.0785  8.837e-02  -0.888    0.375
---
```

Modeling Price: Interpretation of the Joint Model with Interaction

$$\mu(\text{Price} \mid \text{Mileage, Cruise}) = \beta_0 + \beta_1 \text{Mileage} + \beta_2 \text{Cruise} + \beta_3 \text{Mileage} \cdot \text{Cruise}$$

$$\mu(\text{Price} \mid \text{Mileage, Cruise} = 0) = \beta_0 + \beta_1 \text{Mileage}$$

$$\mu(\text{Price} \mid \text{Mileage, Cruise} = 1) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)\text{Mileage}$$

β_3 is the *difference between the slopes* (or *effects*) of Mileage on Price for subpopulations of cars with and without cruise control.

Modeling Price: Separate vs. Joint Model with Interaction Term

$$\hat{\mu}(\text{Price} \mid \text{Mileage, Cruise}) = 16{,}380 - 0.126 \cdot \text{Mileage} + 11{,}488 \cdot \text{Cruise} - 0.079 \cdot \text{Mileage} \cdot \text{Cruise}$$

Unadjusted model:

β_1=-0.173

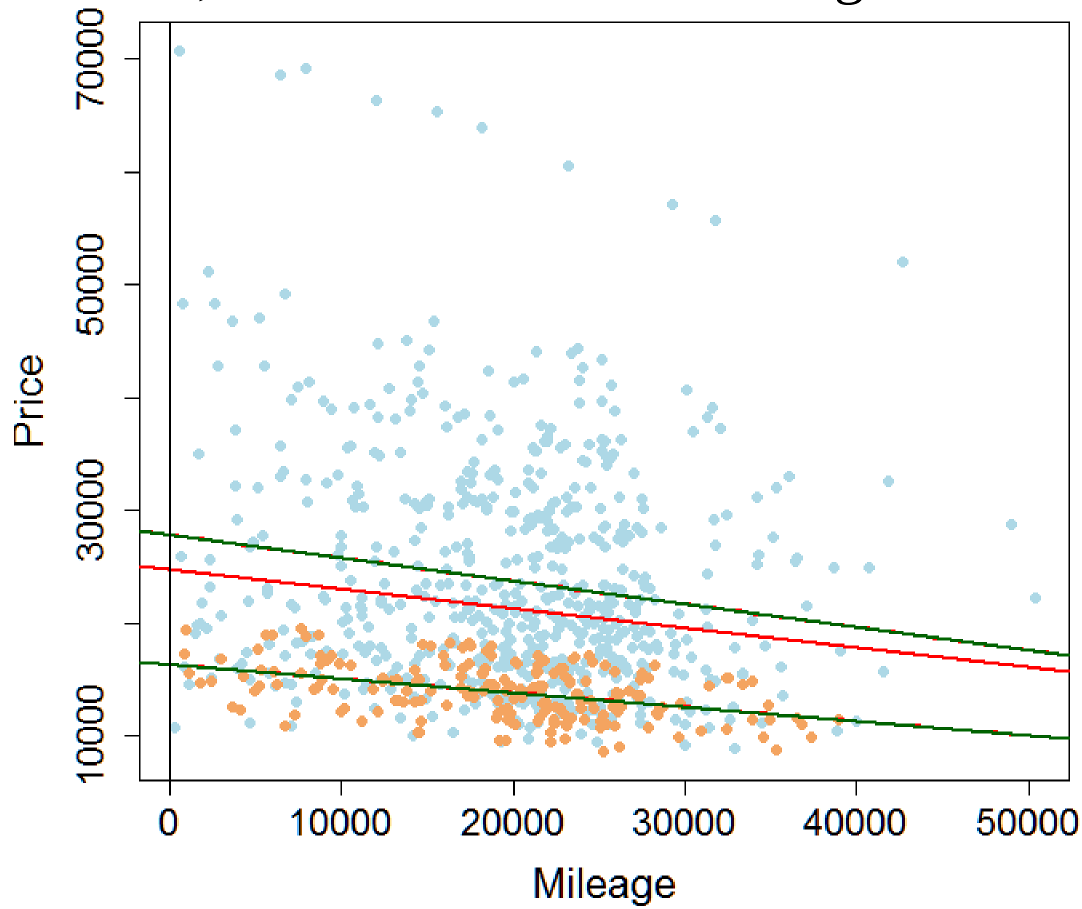
Separate models:

CC:

β_1=- 0.205

No CC:

β_1=-0.126



Joint model (with interaction term):

CC:

$\beta_1 + \beta_3$ = -0.126-0.079= -0.205

No CC:

β_1=-0.126

One Joint MR Model vs. Two SR Models

So what is the difference between fitting two simple regression models v.s. one multiple regression model with interaction term?

- Coefficients are the same in both cases;
- However, residual variance, $\hat{\sigma}^2$, is not the same!

Model	Separate SR models		Joint MR model	
Cruise Control	With CC	Without CC	With CC	Without CC
Slope (*SE*)	-0.205 (0.05)	-0.126 (0.019)	-0.205 (0.044)	-0.126 (0.077)
$\hat{\sigma}$ (d.f.)	10,060 (603)	2,160 (197)	8,802 (800)	8,802 (800)

S_p

Multiple Linear Regression: Model and Parameter Estimation

Multiple Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_K X_{iK} + \varepsilon_i, \text{ where } \varepsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$$

- $E(Y_i \mid X_{i1} X_{i2}, \ldots, X_{iK}) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_K X_{iK}$
- $Var(Y_i \mid X_{i1} X_{i2}, \ldots, X_{iK}) = \sigma^2$, where $i = 1, 2, .., n$.
- K+2 parameters: $\beta_0, \beta_1, \ldots, \beta_K$, and σ^2.

Multiple Regression: Least Squares Estimation of Regression Parameters

- Let

$$\boldsymbol{Y}_{n\times 1} = \begin{pmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{pmatrix}, \boldsymbol{X}_{(K+1)\times n} = \begin{pmatrix} 1 & X_{11} & X_{12} & \dots & X_{1K} \\ 1 & X_{21} & X_{22} & \dots & X_{2K} \\ 1 & \dots & \dots & \dots & \dots \\ 1 & X_{n1} & X_{n2} & \dots & X_{nK} \end{pmatrix},$$

$$\boldsymbol{\beta}_{(K+1)\times 1} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \dots \\ \beta_K \end{pmatrix}, \text{and } \boldsymbol{\varepsilon}_{n\times 1} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{pmatrix}$$

Multiple Regression: Least Squares Estimation of Regression Parameters

$$\left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_K\right) = \arg\min_{(\beta_0, \beta_1, \ldots, \beta_K)} \left[\sum_{i=1}^{n} \left(Y_i - \left(\beta_0 + \beta_1 X_{i1} + \ldots + \beta_K X_{iK} \right) \right)^2 \right]$$

- In matrix algebra notation:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

$$\hat{\boldsymbol{\beta}} = \operatorname{argmin}_{\boldsymbol{\beta}} \left[(\boldsymbol{Y} - \boldsymbol{X\beta})^{\mathrm{T}} (\boldsymbol{Y} - \boldsymbol{X\beta}) \right] = (\boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{X})^{-1} \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{Y}$$

See A. Sen and M. Srivastava. "Regression Analysis: Theory, Methods, and Applications", Sec. 2.2, 2.3 for more details.

- $\left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_K\right)$ are also maximum likelihood estimates (MLE).

Multiple Regression: Estimation of Residual Variance

$$\hat{Y}_i = \hat{\mu}(Y_i \mid X_{i1} X_{i2}, \ldots, X_{iK}) = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \ldots + \hat{\beta}_K X_{iK}$$

Residual variance: $\hat{\sigma}^2 = \dfrac{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2}{n-(K+1)} = \dfrac{\text{SSRes}}{n-(K+1)}$

Exact sampling distribution of the variance:

$$\hat{\sigma}^2 \sim \frac{\sigma^2 \chi^2_{n-(K+1)}}{n-(K+1)}$$

Multiple Linear Regression: Model Assumptions and Diagnostics

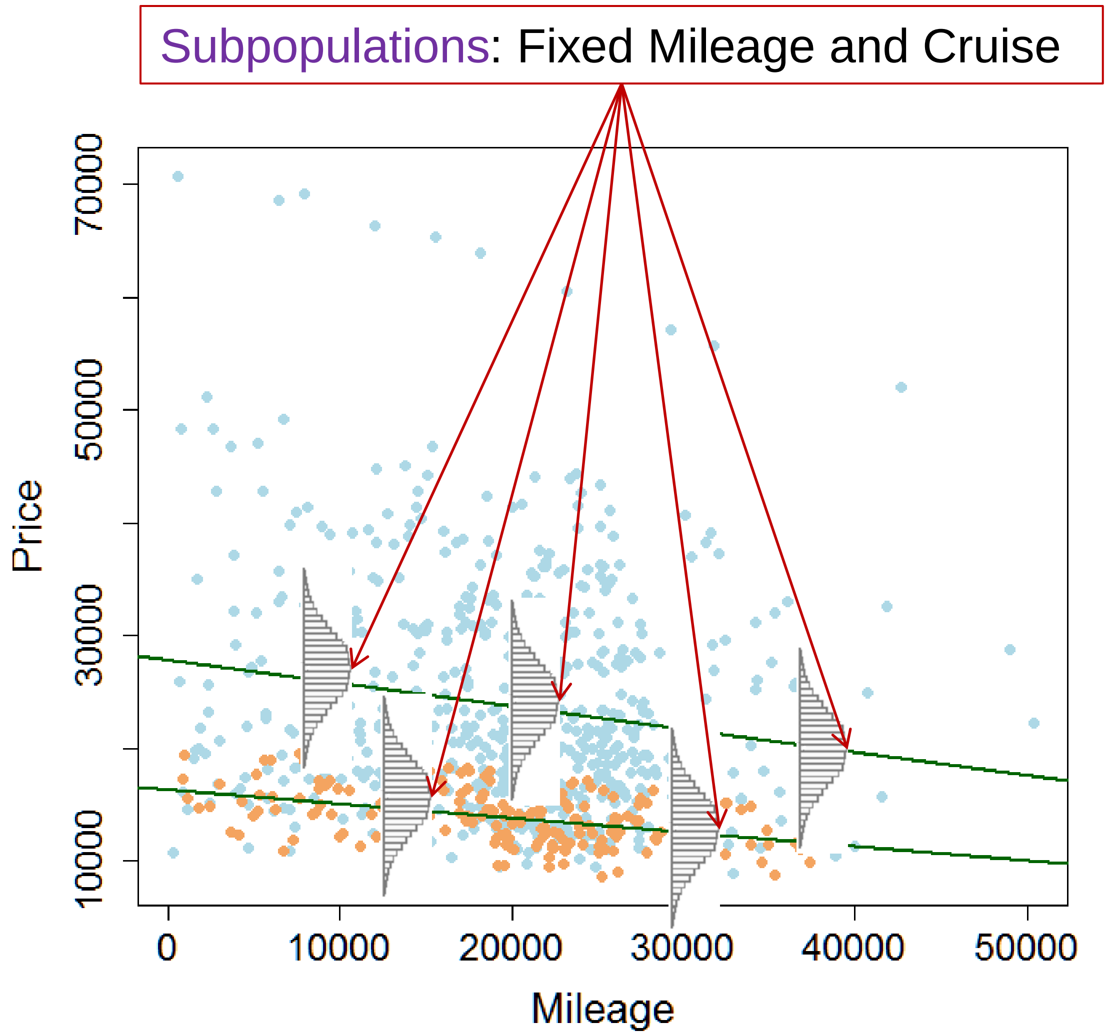
Multiple Linear Regression Assumptions

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_K X_{iK} + \varepsilon_i, \text{ where } \varepsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$$

Assumptions (in the order of importance):

1. Linearity of the relationship;
2. Independence of errors;
3. Equal variance of errors;
4. Normality of errors;
5. Random sampling from a larger population, hypothetical or real.

Multiple Linear Regression Diagnostics



Multiple Linear Regression Diagnostics

Methods of checking and consequences for violation are analogous to the ones for the simple linear regression:

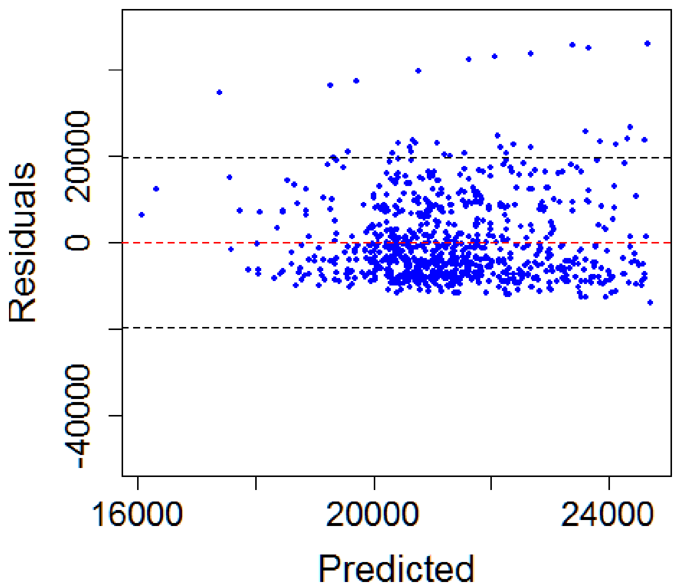
- See Lecture 17 & Lecture 18

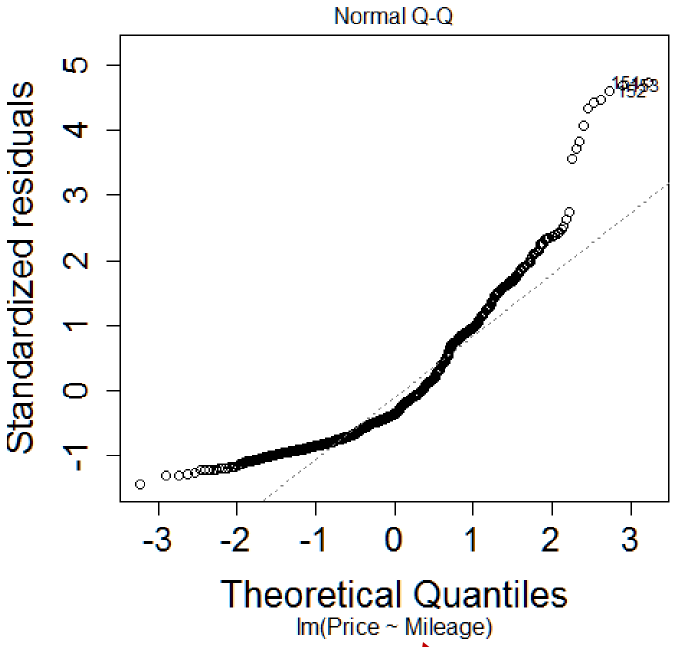
- Linearity of the relationship:
  - Difficult to check *conditional* linearity graphically;
  - Somewhat helpful plots: pairwise scatterplots, trellis graphs (R&S Sec. 9.5);

- Independence of errors & random sampling is assessed based on the study design;

Multiple Linear Regression Diagnostics

1. Equal variance of errors (& independence) may be checked using the plot of residuals vs. fitted values.
2. Normality of errors is checked using the QQ-plot.



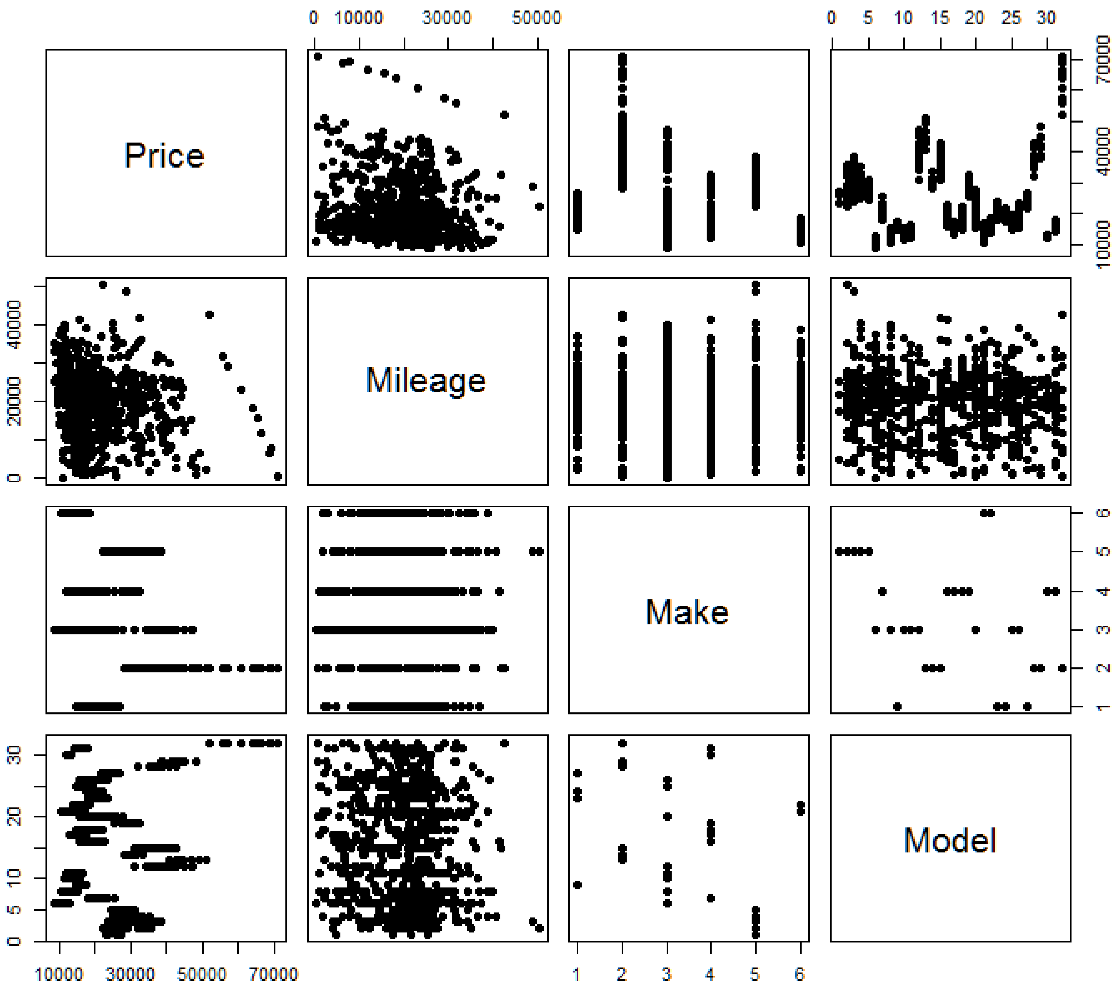


In R: `plot(regmodel, which = 1) # or “which = 2”`

Model Building: Specially Constructed Explanatory Variables

Multiple Regression: Model Building

Matrix of pairwise scatterplots:



Useful for

- observing marginal relationships;
- detecting outliers;
- suggesting the need for transformations.

R: `pairs(DataSetName)`

Why adjust for covariates?

- Improve precision of the actual effect(s) of interest.

- Allow for effect modification: variation in magnitude of effect across levels of a third variable,
  - Subgroups with different responses to therapy.
  - Interactions between risk factors.

- Eliminate confounding: systematic differences between exposure groups;
  - May lead to distortion of effect due to a third measure which results in biased estimates.

Multiple Regression: Constructing Explanatory Variables

- ✓ Indicator
- ✓ Continuous term
- $\boxed{\text{Categorical term}}$

- ✓ One continuous & one indicator (with or without interactions)
- One continuous & one categorical term (with or without interactions)
- Quadratic or polynomial terms

- Two or more continuous terms (with or without interaction)

Categorical Explanatory Variable

- There are five types of cars in the dataset: Convertible, Coupe, Hatchback, Sedan, Wagon.

```
> summary(CarData$Type)
Convertible       Coupe   Hatchback       Sedan       Wagon
         50         140          60         490          64
```

- How would we specify the following model,

$$\mu(\text{Price} \mid \text{Type})?$$

Categorical Explanatory Variable in R

```
> summary(lm(Price ~ Type, data = CarData))
…
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     40,832       1167   35.00   <2e-16 ***
TypeCoupe      -23,105       1359  -17.00   <2e-16 ***
TypeHatchback  -26,661       1580  -16.88   <2e-16 ***
TypeSedan      -19,764       1225  -16.14   <2e-16 ***
TypeWagon      -17,973       1557  -11.54   <2e-16 ***
---
Residual standard error: 8249 on 799 degrees of freedom
Multiple R-squared: 0.3071,  Adjusted R-squared: 0.3036
F-statistic: 88.51 on 4 and 799 DF,  p-value: < 2.2e-16
--- --- --- --- --- --- --- --- --- --- --- --- --- ---
```

Categorical Explanatory Variable

$$\begin{aligned}\mu(\text{Price} \mid \text{Type}) = \; & \beta_0 + \beta_1 \cdot I(\text{Type} = \text{Coupe}) \\ & + \beta_2 \cdot I(\text{Type} = \text{Hatchback}) \\ & + \beta_3 \cdot I(\text{Type} = \text{Sedan}) \\ & + \beta_4 \cdot I(\text{Type} = \text{Wagon})\end{aligned}$$

- Reference level: Convertible

- Categorical variables are also called factors.
- Individual categories are called levels.
- Factor with *M* levels produce *M-1* slopes plus the intercept.

Multiple Regression with Categorical Variable vs. ANOVA

```
> summary(lm(Price ~ Type, data = CarData))
…
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)      40832       1167   35.00   <2e-16 ***
TypeCoupe       -23105       1359  -17.00   <2e-16 ***
TypeHatchback   -26661       1580  -16.88   <2e-16 ***
TypeSedan       -19764       1225  -16.14   <2e-16 ***
TypeWagon       -17973       1557  -11.54   <2e-16 ***
---
Residual standard error: 8249 on 799 degrees of freedom
Multiple R-squared: 0.3071,  Adjusted R-squared: 0.3036
F-statistic: 88.51 on 4 and 799 DF,  p-value: < 2.2e-16
--- --- --- --- --- --- --- --- --- --- --- --- --- ---
> summary(aov(Price ~ Type, data = CarData))
             Df    Sum Sq   Mean Sq F value Pr(>F)    
Type          4 2.409e+10 6.023e+09   88.51 <2e-16 ***
Residuals   799 5.437e+10 6.805e+07
```