STAT 139: STATISTICAL SLEUTHING THROUGH LINEAR MODELS

Lecture 21
Nov 13, 2014

Victoria Liublinska

Odds and Ends

- HW10 will be posted on Fri, 11/14 and due on Fri 11/21
- HW11 will be posted on Fri, 11/21, and due on Tuesday, 12/2, before class.

- Project update is due on Monday, 11/24.
- Poster Session: **Dec 9th**, 10am-4pm CGIS South building (1730 Cambridge Street, Cambridge), basement area.

Previous lecture: Review

- Multiple linear regression (MLR) model:
  - $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_K X_{iK} + \varepsilon_i,$

    where $\varepsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$ and $i = 1, 2, .., n.$
  - $\boldsymbol{Y}_{n\times 1} = \boldsymbol{X}_{n\times(K+1)} \boldsymbol{\beta}_{(K+1)\times 1} + \boldsymbol{\varepsilon}_{n\times 1},\ \boldsymbol{\varepsilon} \sim N_n\left(0, \sigma^2 I_{n\times n}\right).$
- Least squares estimators of regression parameters:
  - K+2 parameters: $\beta_0, \beta_1, \ldots, \beta_K,$ and $\sigma^2.$

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y}, \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2}{n - (K+1)} = \frac{\left(\boldsymbol{Y} - \hat{\boldsymbol{Y}}\right)^T \left(\boldsymbol{Y} - \hat{\boldsymbol{Y}}\right)}{n - (K+1)},$$

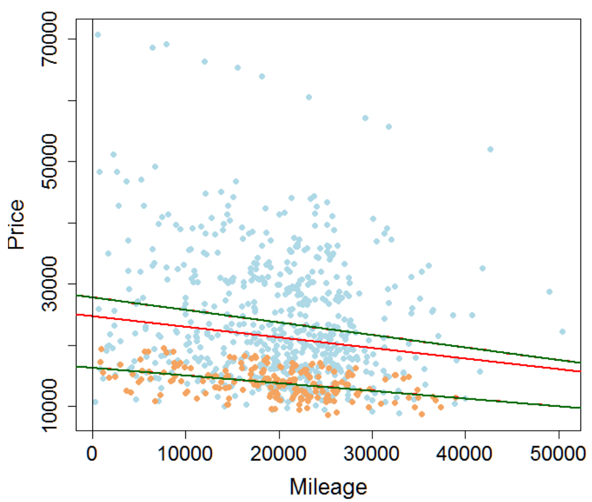
where $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \ldots + \hat{\beta}_K X_{iK}.$

Previous lecture: Review

- MLR model assumptions and diagnostics;
  - Pairwise scatterplot – graphical methods for data exploration;

- Specially constructed explanatory variables and their combinations:
  - Interaction term;
  - One continuous term and one indicator

(with or without interactions);
  - Categorical variable (factor);



Today’s overview

- Specially constructed explanatory variables and their combinations:
  - Categorical variable (factor), cont.;
  - One continuous and one categorical variable (with or without interactions);
  - Quadratic or polynomial terms;
  - Two or more continuous terms (with or without interaction).
- Parallels between linear regression and ANOVA;
- Inferential tools for multiple regression:
  - *t*-tests and CIs for coefficients and their linear combinations;

Today's overview

Reading:

- **Required:** R&S Ch. 9 (Ch. 9 R code), start Ch. 10 (Ch. 10 R code)
- **Optional Reading:** Gelman and Hill, Chapters 3, 4.
- **Supplementary Theory**: A. Sen and M. Srivastava. "Regression Analysis: Theory, Methods, and Applications", Ch. 3

Model Building: Specially Constructed Explanatory Variables, continued.

Multiple Regression: Constructing Explanatory Variables

- ✓ Indicator
- ✓ Continuous term
- 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

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

- Reference level: Convertible

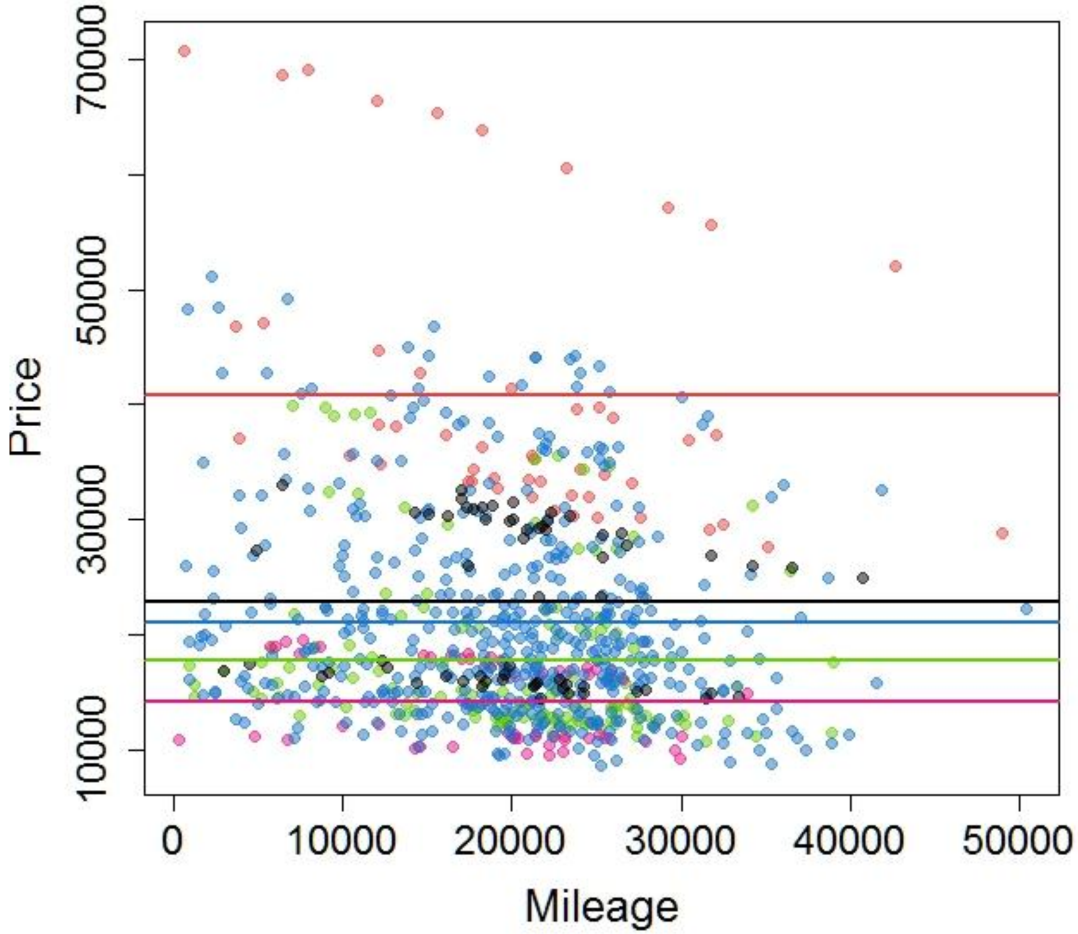
- Categorical variables are also called factors.
- Individual categories are called levels.
- Factor with *M* levels produce *M-1* slopes plus the intercept.

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 & Continuous Explanatory Variables

- Convertible, Coupe, Hatchback, Sedan, Wagon



Multiple Regression with Categorical Variable vs. ANOVA

```
> 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
--- --- --- --- --- --- --- --- --- --- --- --- --- ---
> 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
```

Correspondence Between Regression Model and ANOVA

- When X is binary, $\mu(\mathrm{Y} \mid \mathrm{X}) = \beta_0 + \beta_1 X$ corresponds to fitting a model with two means,

$$\mu(Y|X = 0) = \beta_0 \text{ and } \mu(Y|X = 1) = \beta_0 + \beta_1$$

  - A pooled two-sample *t*-test corresponds to testing

$$H_0 : \beta_1 = 0, H_a : \beta_1 \neq 0.$$

Correspondence Between Regression Model and ANOVA

- When $X \in \{\text{Cat}_0, \text{Cat}_1, \ldots, \text{Cat}_{\text{M-1}}\}$ is categorical,

$$\mu(Y|X) = \beta_0 + \beta_1 I(X = \text{Cat}_1) + \beta_2 I(X = \text{Cat}_2) + \ldots + \beta_{\text{M-1}} I(X = \text{Cat}_{\text{M-1}})$$

corresponds to fitting a model with *M* means:

$$\mu(Y|X = \text{Cat}_0) = \beta_0 \text{ and } \mu(Y|X = \text{Cat}_i) = \beta_0 + \beta_i,\ i = 1, \ldots, M-1$$

  - A one-way ANOVA *F*-test corresponds to testing

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_{\text{M-1}} = 0,\ H_a : \text{At least one} \neq 0$$

Correspondence Between Regression Model and ANOVA

- When X_1 and X_2 are categorical with M_1 and M_2 categories, then regressing Y on X_1, X_2 , and their interaction,

$$\mu(Y|X_1, X_2, X_1 \cdot X_2),$$

corresponds to fitting a model with $M_1 \cdot M_2$ means.

  - A two-way ANOVA *F*-test corresponds to testing for equality of all means in a *saturated* model (See R&S Ch. 13) (or testing whether all <u>slopes</u> are zero).

Multiple Regression: Constructing Explanatory Variables

- ✓ Indicators
- ✓ Continuous terms
- ✓ Categorical terms

- ✓ One continuous & one binary term (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 & Continuous Explanatory Variables

Model: $\mu(\text{Price}_i \mid \text{Mileage}_i, \text{Type}_i) = \beta_0 + \beta_1 \cdot \text{Mileage}_i$

$$+ \beta_2 \cdot I(\text{Type}_i = \text{Coupe})$$
$$+ \beta_3 \cdot I(\text{Type}_i = \text{Hatchback})$$
$$+ \beta_4 \cdot I(\text{Type}_i = \text{Sedan})$$
$$+ \beta_5 \cdot I(\text{Type}_i = \text{Wagon})$$

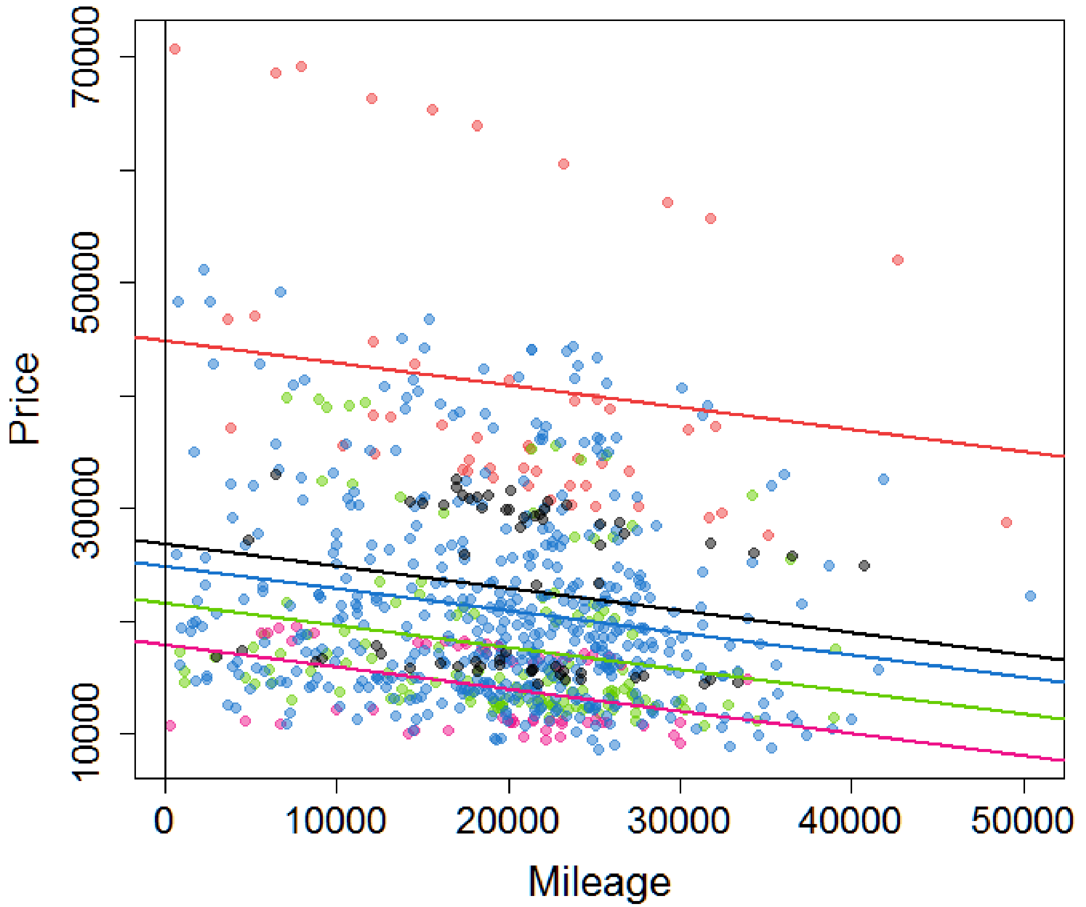
$$\mu(\text{Price}_i \mid \text{Mileage}_i, \text{Type}_i = \text{Convertible}) = \beta_0 + \beta_1 \cdot \text{Mileage}_i$$
$$\mu(\text{Price}_i \mid \text{Mileage}_i, \text{Type}_i = \text{Coupe}) = (\beta_0 + \beta_2) + \beta_1 \cdot \text{Mileage}_i$$
$$\mu(\text{Price}_i \mid \text{Mileage}, \text{Type}_i = \text{Hatchback}) = (\beta_0 + \beta_3) + \beta_1 \cdot \text{Mileage}_i$$
$$\mu(\text{Price}_i \mid \text{Mileage}_i, \text{Type}_i = \text{Sedan}) = (\beta_0 + \beta_4) + \beta_1 \cdot \text{Mileage}_i$$
$$\mu(\text{Price}_i \mid \text{Mileage}_i, \text{Type}_i = \text{Wagon}) = (\beta_0 + \beta_5) + \beta_1 \cdot \text{Mileage}_i$$

Categorical & Continuous Explanatory Variables

```
Call:
lm(formula = Price ~ Mileage + Type, data = CarData)

Residuals:
   Min      1Q Median      3Q     Max
-12573  -5797  -2292   3689   26652

Coefficients:
                Estimate   Std. Error t value Pr(>|t|)
(Intercept)     44,890     1,354       33.158  < 2e-16 ***
Mileage            -0.196     0.0349  -5.616  2.7e-08 ***
TypeCoupe      -23,270     1,334      -17.443  < 2e-16 ***
TypeHatchback  -26,980     1,551      -17.391  < 2e-16 ***
TypeSedan      -19,960     1,202      -16.596  < 2e-16 ***
TypeWagon      -18,000     1,528      -11.777  < 2e-16 ***
```

Categorical & Continuous Explanatory Variables

- Convertible, Coupe, Hatchback, Sedan, Wagon



How can we allow the slopes to vary as well?

Categorical & Continuous Explanatory Variables with Interaction

Model:

$$
\begin{aligned}
\mu(\text{Price}_i \mid \text{Mileage}_i, \text{Type}_i) = \beta_0 &+ \beta_1 \cdot \text{Mileage}_i \\
&+ \beta_2 \cdot I(\text{Type}_i = \text{Coupe}) \\
&+ \beta_3 \cdot I(\text{Type}_i = \text{Hatchback}) \\
&+ \beta_4 \cdot I(\text{Type}_i = \text{Sedan}) \\
&+ \beta_5 \cdot I(\text{Type}_i = \text{Wagon}) \\
&+ \beta_6 \cdot \text{Mileage}_i \cdot I(\text{Type}_i = \text{Coupe}) \\
&+ \beta_7 \cdot \text{Mileage}_i \cdot I(\text{Type}_i = \text{Hatchback}) \\
&+ \beta_8 \cdot \text{Mileage}_i \cdot I(\text{Type}_i = \text{Sedan}) \\
&+ \beta_9 \cdot \text{Mileage}_i \cdot I(\text{Type}_i = \text{Wagon})
\end{aligned}
$$

- One line for each car type, without restrictions on the slopes.


```
lm(Price ~ Mileage*Type, data = CarData)
```

Categorical & Continuous Explanatory Variables with Interaction

Corresponding models for each car type:

$$\mu(\text{Price}_i \mid \text{Mileage}_i, \text{Type}_i = \text{Convertible}) = \beta_0 + \beta_1 \cdot \text{Mileage}_i$$

$$\mu(\text{Price}_i \mid \text{Mileage}_i, \text{Type}_i = \text{Coupe}) = (\beta_0 + \beta_2) + (\beta_1 + \beta_6) \cdot \text{Mileage}_i$$

$$\mu(\text{Price}_i \mid \text{Mileage}_i, \text{Type}_i = \text{Hatchback}) = (\beta_0 + \beta_3) + (\beta_1 + \beta_7) \cdot \text{Mileage}_i$$

$$\mu(\text{Price}_i \mid \text{Mileage}_i, \text{Type}_i = \text{Sedan}) = (\beta_0 + \beta_4) + (\beta_1 + \beta_8) \cdot \text{Mileage}_i$$

$$\mu(\text{Price}_i \mid \text{Mileage}_i, \text{Type}_i = \text{Wagon}) = (\beta_0 + \beta_5) + (\beta_1 + \beta_9) \cdot \text{Mileage}_i$$

- Coefficient β_2 may be interpreted as the difference in average *Price* when *Mileage*=0 between convertible cars (i.e., the **reference** level!) and coupes.
- Coefficient β_6 may be interpreted as the difference in effects* of *Mileage* on average *Price* between convertible cars and coupes.

Categorical & Continuous Explanatory Variables with Interaction

Corresponding models for each car type:

$$\mu(\text{Price}_i \mid \text{Mileage}_i, \text{Type}_i = \text{Convertible}) = \beta_0 + \beta_1 \cdot \text{Mileage}_i$$

$$\mu(\text{Price}_i \mid \text{Mileage}_i, \text{Type}_i = \text{Coupe}) = (\beta_0 + \beta_2) + (\beta_1 + \beta_6) \cdot \text{Mileage}_i$$

$$\mu(\text{Price}_i \mid \text{Mileage}_i, \text{Type}_i = \text{Hatchback}) = (\beta_0 + \beta_3) + (\beta_1 + \beta_7) \cdot \text{Mileage}_i$$

$$\mu(\text{Price}_i \mid \text{Mileage}_i, \text{Type}_i = \text{Sedan}) = (\beta_0 + \beta_4) + (\beta_1 + \beta_8) \cdot \text{Mileage}_i$$

$$\mu(\text{Price}_i \mid \text{Mileage}_i, \text{Type}_i = \text{Wagon}) = (\beta_0 + \beta_5) + (\beta_1 + \beta_9) \cdot \text{Mileage}_i$$

- Coefficient $\boldsymbol{\beta_3}$ may be interpreted as the difference in average *Price* at *Mileage*=0 between convertible cars (i.e., the **reference** level!) and hatchbacks.
- Coefficient $\boldsymbol{\beta_7}$ may be interpreted as the difference in effects* of *Mileage* on average *Price* between convertible cars and hatchbacks.

Categorical & Continuous Explanatory Variables with Interaction

Corresponding models for each car type:

$$\mu(\text{Price}_i \mid \text{Mileage}_i, \text{Type}_i = \text{Convertible}) = \beta_0 + \beta_1 \cdot \text{Mileage}_i$$

$$\mu(\text{Price}_i \mid \text{Mileage}_i, \text{Type}_i = \text{Coupe}) = (\beta_0 + \beta_2) + (\beta_1 + \beta_6) \cdot \text{Mileage}_i$$

$$\mu(\text{Price}_i \mid \text{Mileage}_i, \text{Type}_i = \text{Hatchback}) = (\beta_0 + \beta_3) + (\beta_1 + \beta_7) \cdot \text{Mileage}_i$$

$$\mu(\text{Price}_i \mid \text{Mileage}_i, \text{Type}_i = \text{Sedan}) = (\beta_0 + \beta_4) + (\beta_1 + \beta_8) \cdot \text{Mileage}_i$$

$$\mu(\text{Price}_i \mid \text{Mileage}_i, \text{Type}_i = \text{Wagon}) = (\beta_0 + \beta_5) + (\beta_1 + \beta_9) \cdot \text{Mileage}_i$$

Analogously,

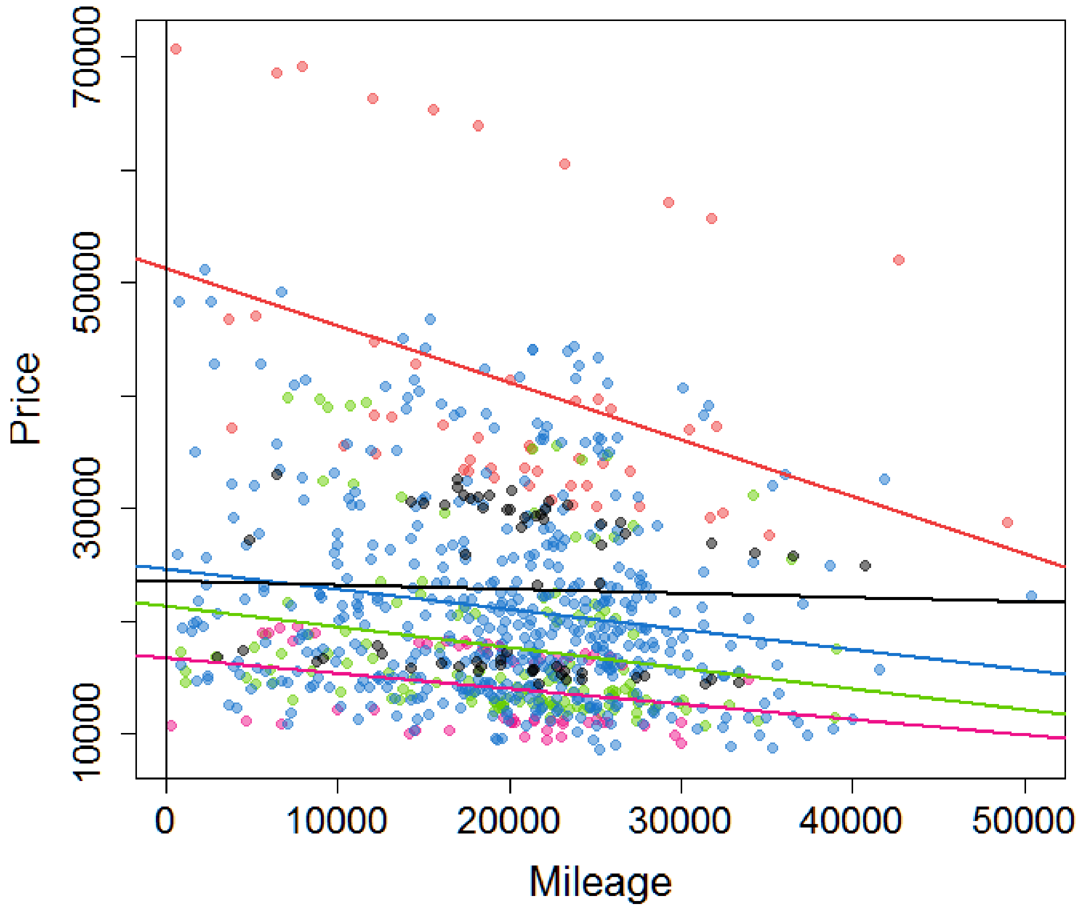
- Pair $\boldsymbol{\beta_4}$, $\boldsymbol{\beta_8}$ describes differences in intercepts and slopes of lines for convertible cars and for sedans.
- Pair $\boldsymbol{\beta_5}$, $\boldsymbol{\beta_9}$ describes differences in intercepts and slopes of lines for convertible cars and for wagons.

Categorical & Continuous Explanatory Variables with Interaction

```
Call:
lm(formula = Price ~ Mileage * Type, data = CarData)

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)            5.127e+04  2.718e+03  18.863  < 2e-16 ***
Mileage               -5.043e-01  1.191e-01  -4.233 2.58e-05 ***
TypeCoupe             -2.989e+04  3.274e+03  -9.129  < 2e-16 ***
TypeHatchback         -3.447e+04  4.053e+03  -8.506  < 2e-16 ***
TypeSedan             -2.669e+04  2.876e+03  -9.282  < 2e-16 ***
TypeWagon             -2.762e+04  4.075e+03  -6.778 2.37e-11 ***
Mileage:TypeCoupe      3.203e-01  1.465e-01   2.186   0.0291 *  
Mileage:TypeHatchback  3.665e-01  1.897e-01   1.932   0.0538 .  
Mileage:TypeSedan      3.261e-01  1.269e-01   2.569   0.0104 *  
Mileage:TypeWagon      4.660e-01  1.832e-01   2.544   0.0112 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Categorical & Continuous Explanatory Variables with Interaction

- Convertible, Coupe, Hatchback, Sedan, Wagon



Type	N
Convertible	50
Coupe	140
Hatchback	60
Sedan	490
Wagon	64

- `lm(formula = Price ~ Mileage + Type + Mileage:Type, data = CarData)`

Multiple Regression: Constructing Explanatory Variables

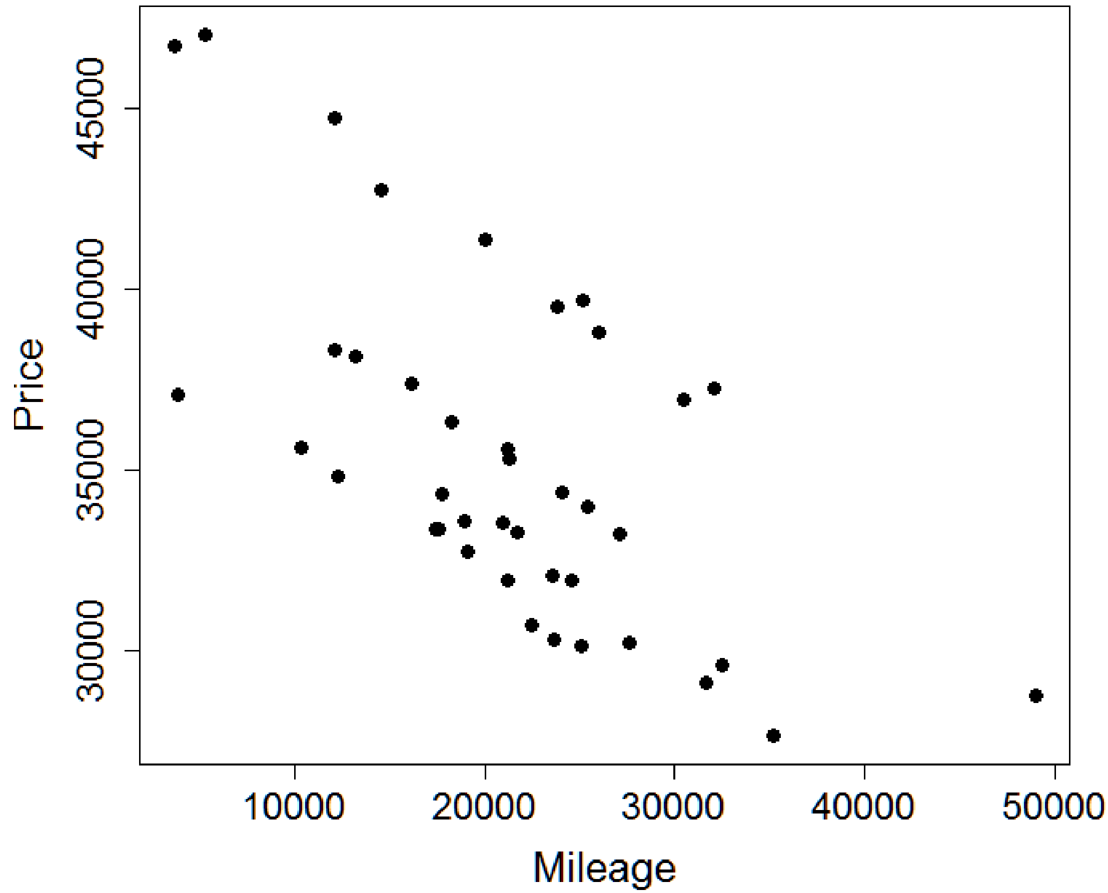
- ✓ Indicators
- ✓ Continuous terms
- ✓ Categorical terms

- ✓ One continuous & one binary term (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)

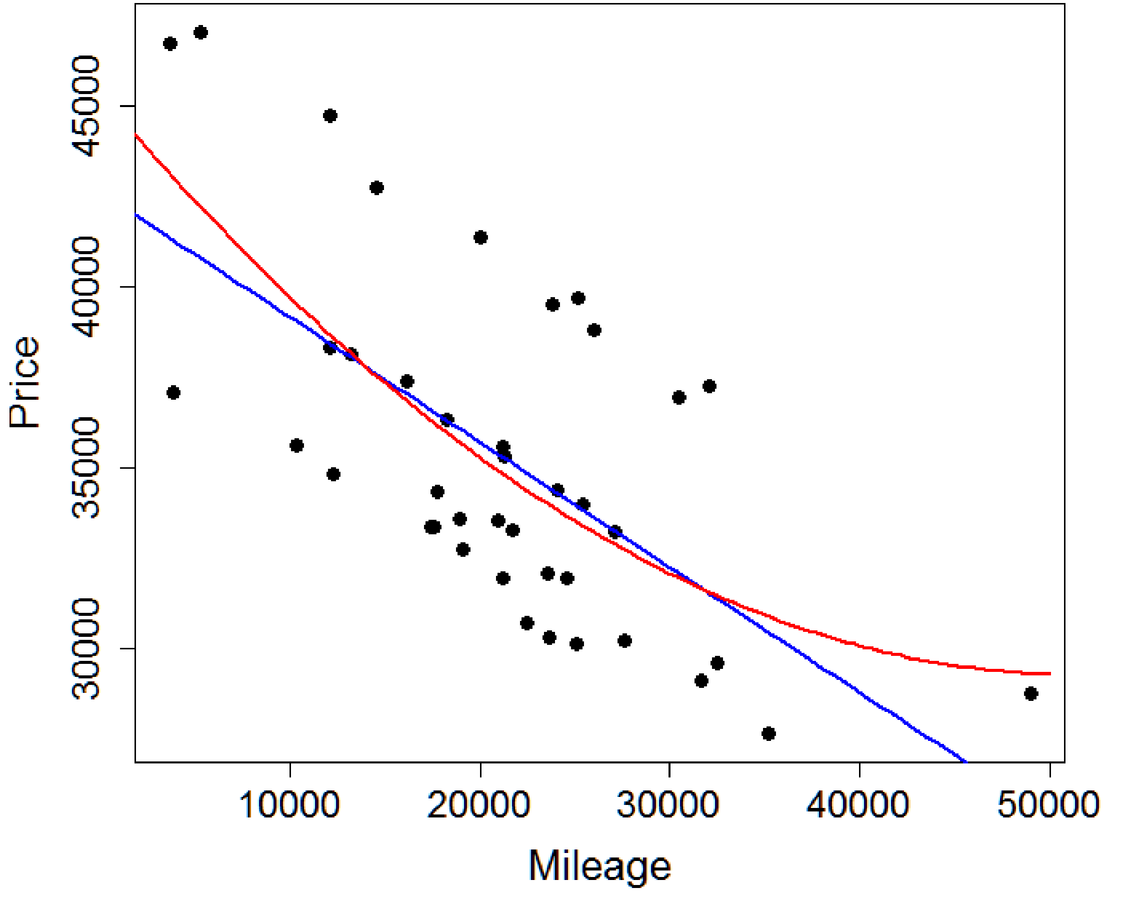
Polynomial Terms for Curvature

$$\mu(\text{Price} \mid \text{Mileage}, \text{Type}) = \beta_0 + \beta_1 \cdot \text{Mileage} + \beta_2 \cdot \text{Mileage}^2$$



Polynomial Terms for Curvature

$$\mu(\text{Price} \mid \text{Mileage}, \text{Type}) = \beta_0 + \beta_1 \cdot \text{Mileage} + \beta_2 \cdot \text{Mileage}^2$$



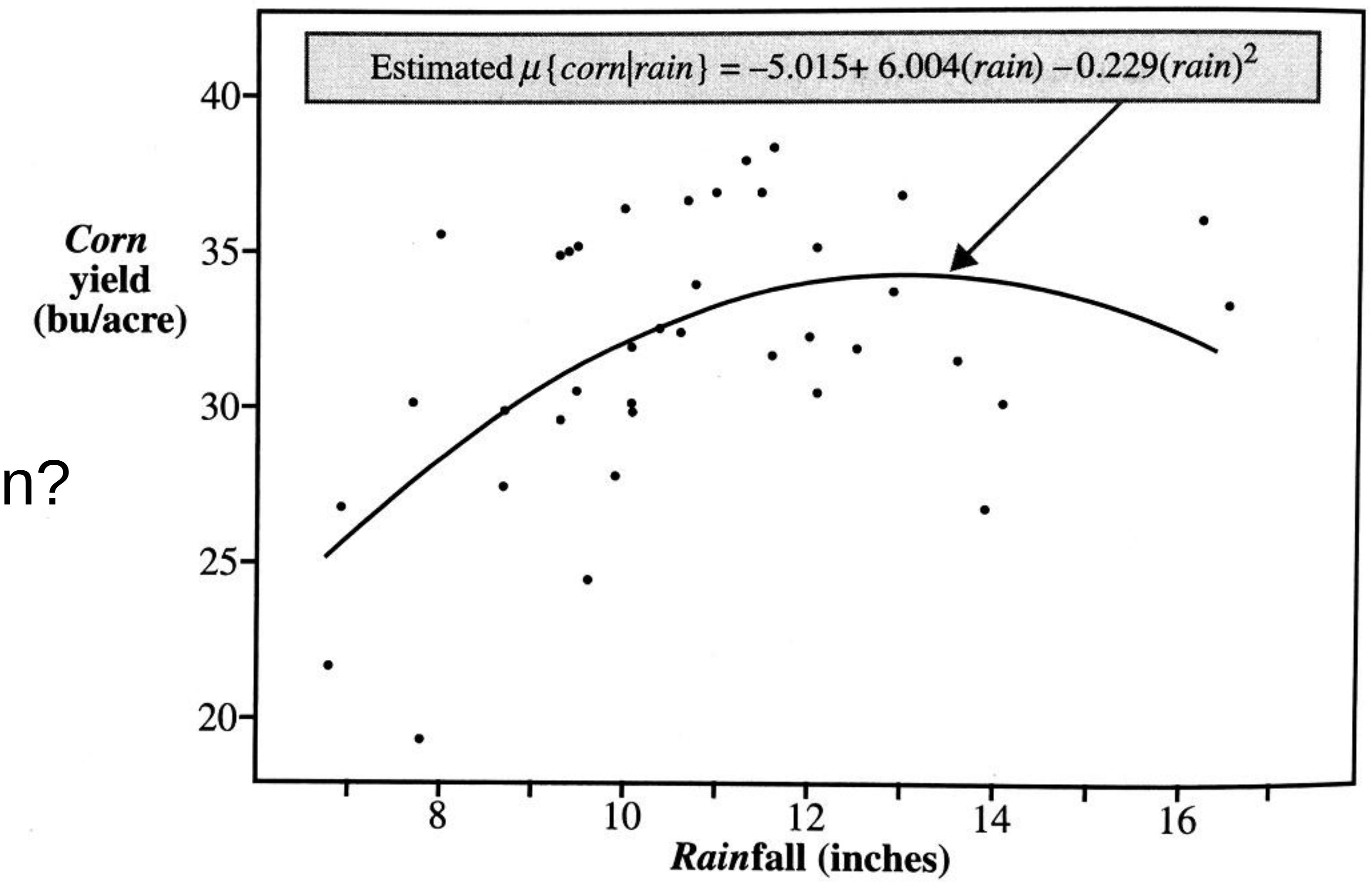

```
lm(Price ~ Mileage + I(Mileage^2), data = CarData)
```

Polynomial Terms for Curvature

```
all:
lm(formula = Price ~ Mileage + I(Mileage^2), data = CarData)

...

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   4.532e+04  2.423e+03  18.706  < 2e-16 ***
Mileage      -6.218e-01  2.057e-01  -3.023  0.00453 ** 
I(Mileage^2)  6.027e-06  4.253e-06   1.417  0.16486    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3604 on 37 degrees of freedom
Multiple R-squared:  0.4456, Adjusted R-squared:  0.4156 
F-statistic: 14.87 on 2 and 37 DF,  p-value: 1.826e-05
```

Polynomial Terms for Curvature

Display 9.6 Yearly corn yield versus rainfall (1890–1927) in six U.S. states



Interpretation?

Further Notes About Polynomial Terms

- Squared term should be included when:
  - Analyst is suspecting nonlinear relationship between Y and X;
  - The response is maximized (or minimized) at a certain point X^*,

$$\mu(Y \mid X) = \beta_0 + \beta_1 \cdot X + \beta_2 \cdot X^2 \quad \Rightarrow \quad X^* = -\beta_1 / (2\beta_2)$$

  - A better model fit is needed for the purpose of prediction (especially if there aren't many explanatory variables).

Multiple Regression: Constructing Explanatory Variables

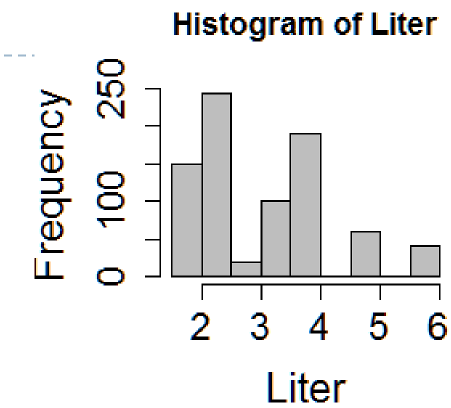
- ✓ Indicators
- ✓ Continuous terms
- ✓ Categorical terms

- ✓ One continuous & one binary term (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)

Two Continuous Predictors

- Consider Mileage and engine size (in liters):



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

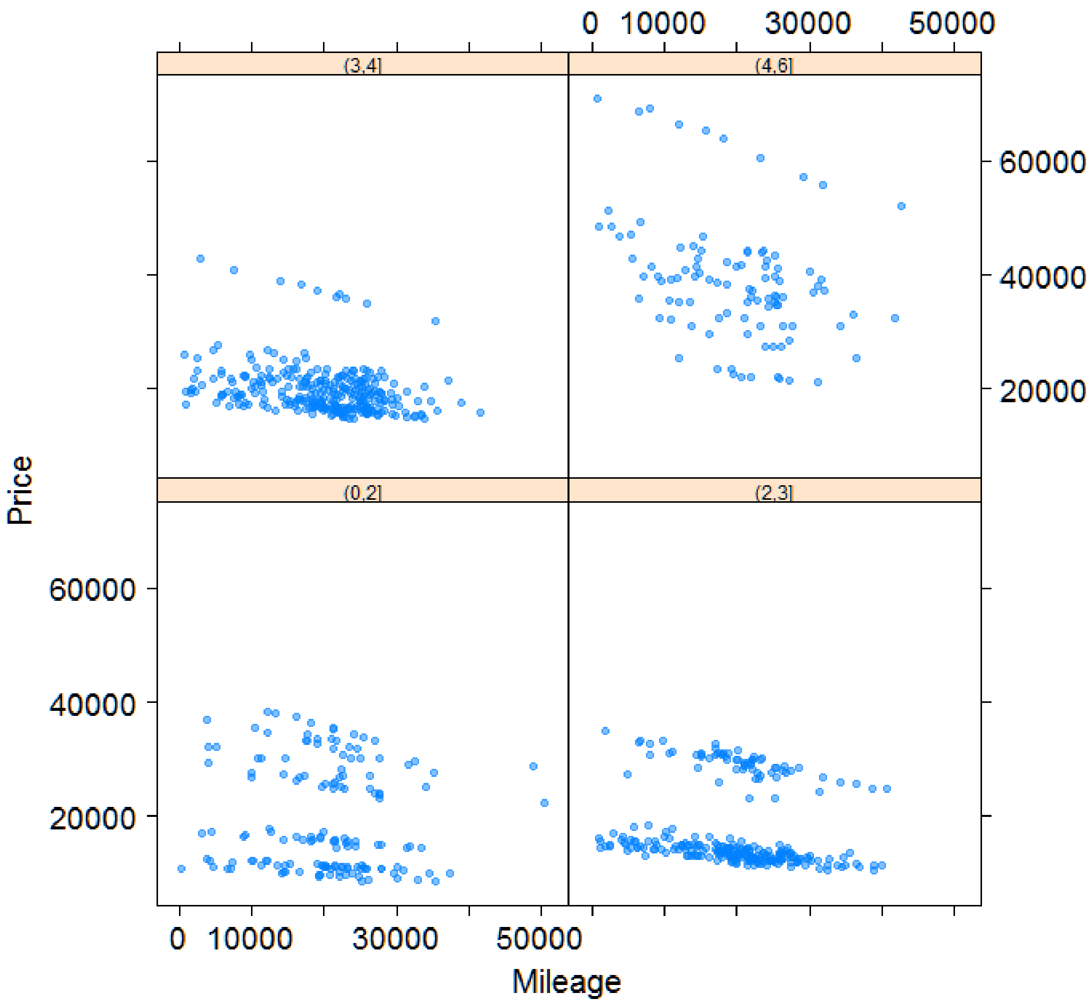
$$\hat{\mu}(\text{Price}_i \mid \text{Mileage}_i, \text{Liter}_i) = 9{,}426.6 - 0.16 \cdot \text{Mileage}_i + 4{,}968.3 \cdot \text{Liter}_i$$

Interpretation:

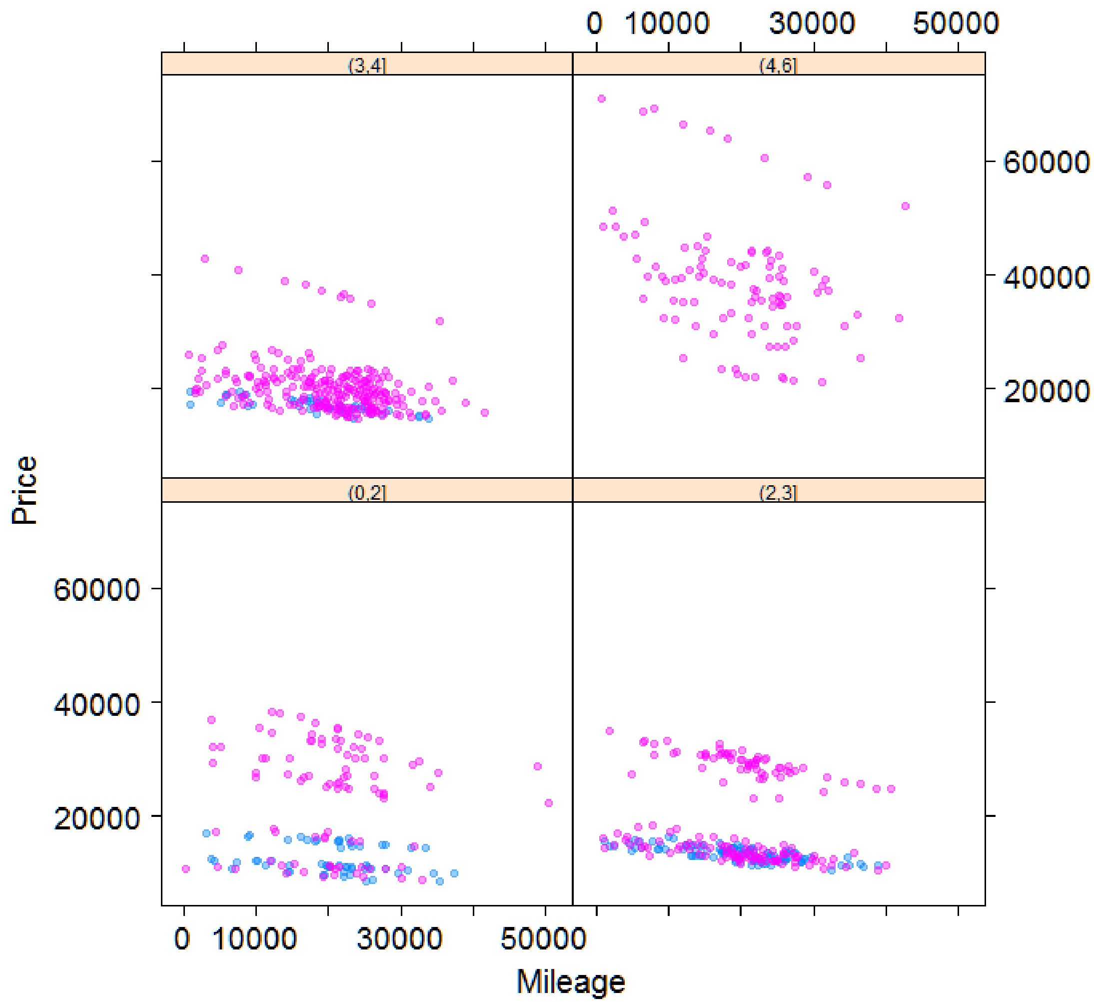
- One-mile increase in *Mileage* is associated with $\hat{\beta}_1$= -$**0.16** difference in average car *Price* after adjusting for engine size OR while engine size is held constant.
- One-unit increase in *Liter* is associated with $\hat{\beta}_2$= $**4,968.3** difference in average car *Price* after adjusting for mileage OR while mileage is held constant.

Always disclose what variables are **held constant!**

Plotting Two Continuous Predictors: Trellis Graph



Plotting Two Continuous Predictors: Trellis Graph



With or without cruise control.

Plotting Two Continuous Predictors: Trellis Graph

```
library(lattice)

CarData$LiterCategories <- cut(CarData$Liter, c(0,2,3,4,6))

xyplot(Price ~ Mileage | LiterCategories, data = CarData,
  pch=19, alpha=0.5)

xyplot(Price ~ Mileage | LiterCategories,  groups =
  Cruise, data = CarData, pch=19, alpha=0.4)
```

Two Continuous Predictors with Interaction

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

```
Call:
lm(formula = Price ~ Mileage * Liter, data = CarData)
…
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    3669.79882 2137.39353   1.717  0.08638 .
Mileage           0.13038    0.09906   1.316  0.18849
Liter          6844.85522  652.42025  10.491  < 2e-16 ***
Mileage:Liter    -0.09493    0.03033  -3.130  0.00181 **
```

$$\hat{\mu}(\text{Price}_i \mid \text{Mileage}_i, \text{Liter}_i) = 3{,}669.8 + 0.13 \cdot \text{Mileage}_i + 6{,}845 \cdot \text{Liter}_i - - 0.095 \cdot \text{Liter}_i \cdot \text{Mileage}_i$$

Two Continuous Predictors with Interaction

$$\hat{\mu}(\text{Price}_i \mid \text{Mileage}_i, \text{Liter}_i) = 3{,}669.8 + 0.13 \cdot \text{Mileage}_i + 6{,}845 \cdot \text{Liter}_i - 0.095 \cdot \text{Liter}_i \cdot \text{Mileage}_i$$

Interpretation:

- For *Liter*=0, one-mile increase in *Mileage* is associated with $\hat{\beta}_1$ = $**0.13** difference in average car *Price*.
- For *Mileage*=0, one-unit increase in *Liter* is associated with $\hat{\beta}_2$ = $**6,845** difference in average car *Price*.
- For a fixed *Liter*, one-mile increase in *Mileage* is associated with $\hat{\beta}_1 + \hat{\beta}_3 \text{Liter}$ = $\mathbf{0.13 - 0.095 \cdot Liter}$ difference in average car *Price*.
- For a fixed *Mileage*, one-unit increase in *Liter* is associated with $\hat{\beta}_2 + \hat{\beta}_3 \text{Mileage}$ = $\mathbf{6{,}845 - 0.095 \cdot Mileage}$ difference in average car *Price*.

Transforming Predictors for a More Convenient Interpretation

- Center continuous predictors around their sample means:

$$\text{c.Mileage}_i = \text{Mileage}_i - \overline{\text{M}}, \text{ where } \overline{\text{M}} = \sum_{j=1}^{n} \text{Mileage}_j / n,$$

$$\text{c.Liter}_i = \text{Liter}_i - \overline{\text{L}}, \text{ where } \overline{\text{L}} = \sum_{j=1}^{n} \text{Liter}_j / n$$

$$\mu(\text{Price}_i \mid \text{c.Mileage}_i, \text{c.Liter}_i) = \beta_0 + \beta_1 \cdot \text{c.Mileage}_i + \beta_2 \cdot \text{c.Liter}_i + \beta_3 \cdot \text{c.Liter}_i \cdot \text{c.Mileage}_i$$

$$\hat{\mu}(\text{Price}_i \mid \text{c.Mileage}_i, \text{c.Liter}_i) = 21{,}327 - 0.16 \cdot \text{c.Mileage}_i + 4{,}962 \cdot \text{c.Liter}_i - \boxed{-0.095} \cdot \text{c.Liter}_i \cdot \text{c.Mileage}_i$$

- Here, intercept and slopes are interpreted with reference to average values of covariates,
  - e.g., for an average engine size, one-mile increase in *Mileage* is associated with β_1= -$**0.16** difference in average car *Price*.

Transforming Predictors for a More Convenient Interpretation

```
> lm(formula = Price ~ Mileage * Liter, data = CarData)
...
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)    3669.79882 2137.39353   1.717  0.08638 .
Mileage           0.13038    0.09906   1.316  0.18849
Liter          6844.85522  652.42025  10.491  < 2e-16 ***
Mileage:Liter    -0.09493    0.03033  -3.130  0.00181 **
---
Multiple R-squared: 0.3372,    Adjusted R-squared: 0.3348
F-statistic: 135.7 on 3 and 800 DF,  p-value: < 2.2e-16
------------------------------------------------------------------
> RegModel <- lm(Price ~ c.Mileage * c.Liter, data = CarData)
...
Coefficients:
                     Estimate   Std. Error t value Pr(>|t|)
(Intercept)          21327.13   2.844e+02  74.995  < 2e-16 ***
c.Mileage            -0.15795   3.472e-02  -4.549 6.23e-06 ***
c.Liter              4962.154   2.574e+02  19.278  < 2e-16 ***
c.Mileage:c.Liter    -0.09493   3.033e-02  -3.130  0.00181 **
---
Multiple R-squared: 0.3372,    Adjusted R-squared: 0.3348
F-statistic: 135.7 on 3 and 800 DF,  p-value: < 2.2e-16
```

Transforming Predictors for a More Convenient Interpretation

- Center continuous predictors around sample means and scale by sample SD:

$$\text{z.Mileage}_i = (\text{Mileage}_i - \overline{\text{M}})/s_M \text{, where } s_M \text{ is the sample s.d. of Mileage,}$$

$$\text{z.Liter}_i = (\text{Liter}_i - \overline{\text{L}})/s_L \text{, where } s_M \text{ is the sample s.d. of Liter.}$$

$$\hat{\mu}(\text{Price}_i \mid \text{z.Mileage}_i, \text{z.Liter}_i) = \boxed{21{,}327} - 1{,}295 \cdot \text{z.Mileage}_i + 5{,}486 \cdot \text{z.Liter}_i - \\ -860 \cdot \text{z.Liter}_i \cdot \text{z.Mileage}_i$$

- Slopes are interpreted with reference to units of standard deviations of the corresponding explanatory variables,
  - e.g., for an average engine size, one-SD increase in *Mileage* is associated with β_1= **-$1,295** difference in average car *Price*.
- Fit quality is not affected!

When to Include Interaction Terms

1. It is part of the question of interest.
2. Scientific reasons suggest different slopes ("effects") of one predictor at different levels of another one.
3. Creating general model for comparison with main effects-only model.

Predictors with large main effects tend to have large interactions (Gelman and Hill, p. 36).

In general, if including interaction term, also include main effects.

Multiple Regression: Constructing Explanatory Variables

- ✓ Indicators
- ✓ Continuous terms
- ✓ Categorical terms

- ✓ One continuous & one binary term (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)

Cool animations: http://www.math.yorku.ca/SCS/spida/lm/visreg.html

Further Notes on Log Transformation

- Interpretation of slopes after log-transformation of covariates or outcomes is the same for MLR as it is for SLR (see Lecture 18).
  - With an added disclaimer about other covariates held fixed.

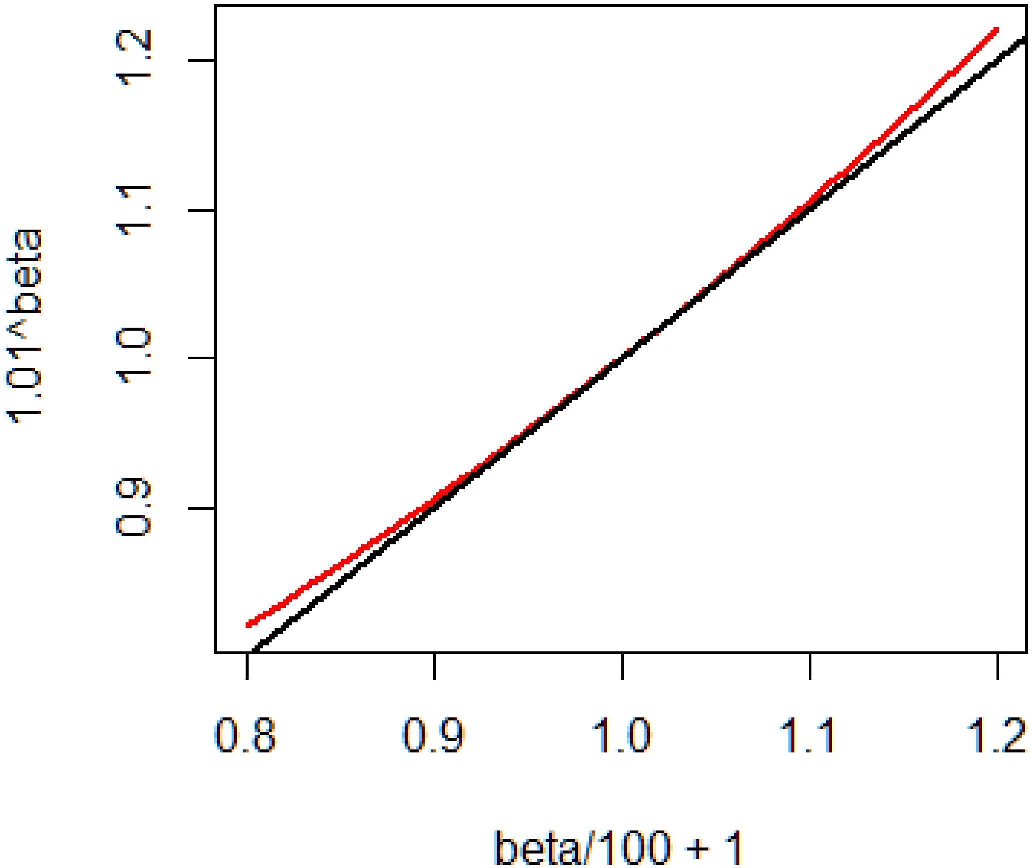
- If the outcome is log-transformed, then the intercept is interpreted as follows:

$$\text{Median}(Y \mid X_1 = 0, X_2 = 0, \ldots, X_K = 0) = \exp(\beta_0).$$

- Alternative interpretation of the slope after a log-log transformation:

One percent *increase in X yields a change in the median of Y by* $\boldsymbol{\beta_1}$ ***percent.***

One percent** increase in X yields an change in the median of Y by* $\boldsymbol{\beta_1}$ ***percent.



This approximation works well for -20< $\boldsymbol{\beta_1}$ <20.