STAT 139: STATISTICAL SLEUTHING THROUGH LINEAR MODELS

Lecture 19
Nov 6, 2014

Victoria Liublinska

Previous lecture: Review

- Diagnostics of assumptions of the linear regression:
  - Linearity
  - Independence of errors
  - Equal variance of errors
  - Normality of errors
- Interpretation of results after log transformation.
- Sum of squares decomposition for the linear regression.
- R-squared (R^2) statistic - the proportion of variation in the response explained by the model for the means;
  - Useful when <u>all</u> assumptions of the linear regression are met.

Interpretation after Log Transformation

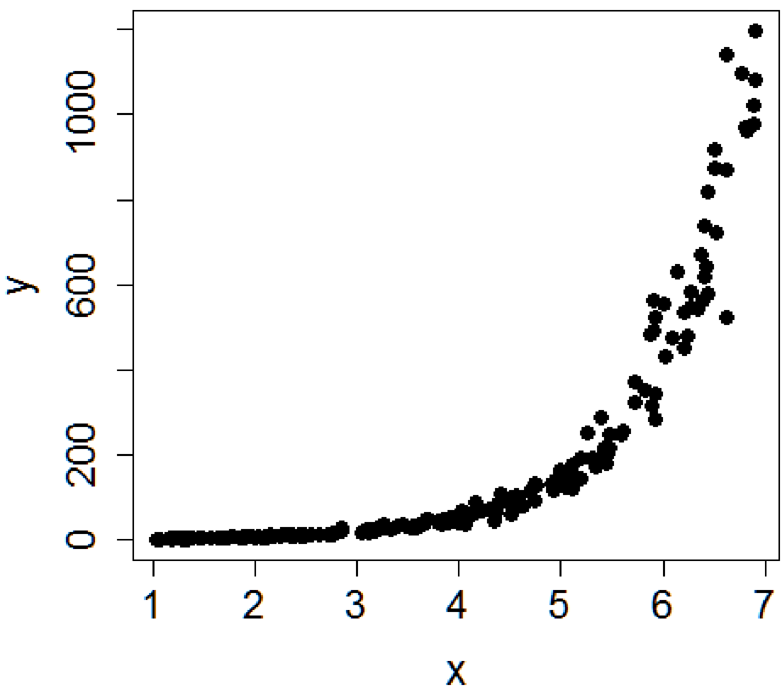
No transformation: increase in X by 1 is associated with a shift by β_1 in the mean of Y.

log(*Y*): increase in X by 1 is associated with a multiplicative change of e^{β_1} in the median of Y.

log(X): multiplying X by k is associated with a shift by $\beta_1 \log(k)$ in the mean of Y.

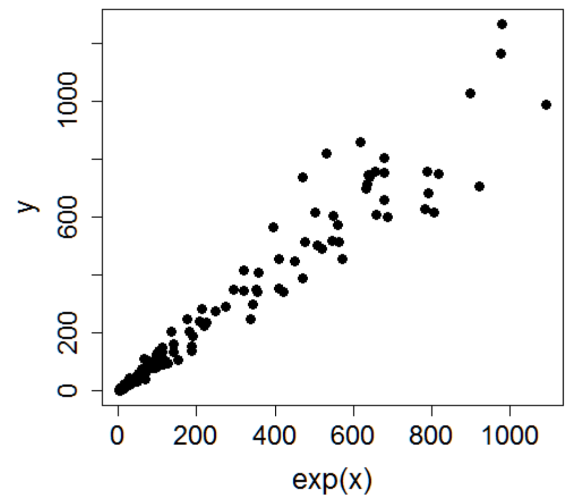
log(Y) and log(X): multiplying X by k is associated with a multiplicative change of k^{β_1} in the median of Y.

Transformations of the Response if it is Convex

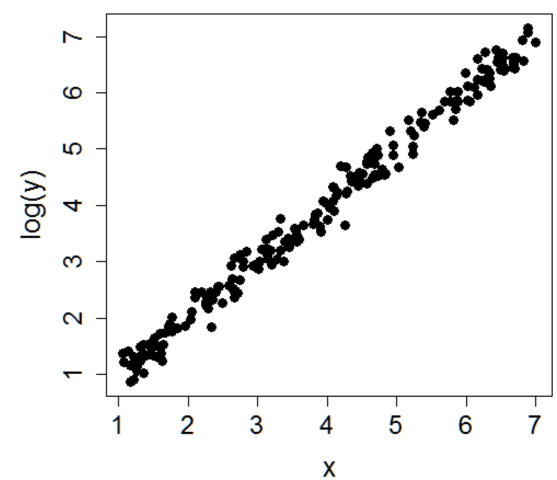


$$E(\log(Y)|X) = \beta_0 + \beta_1 X$$

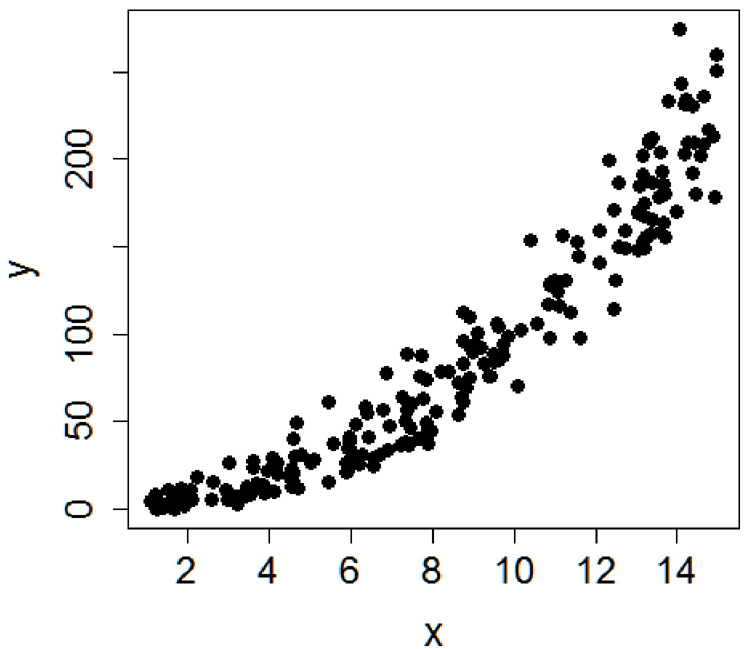
Y ~ exp(X)

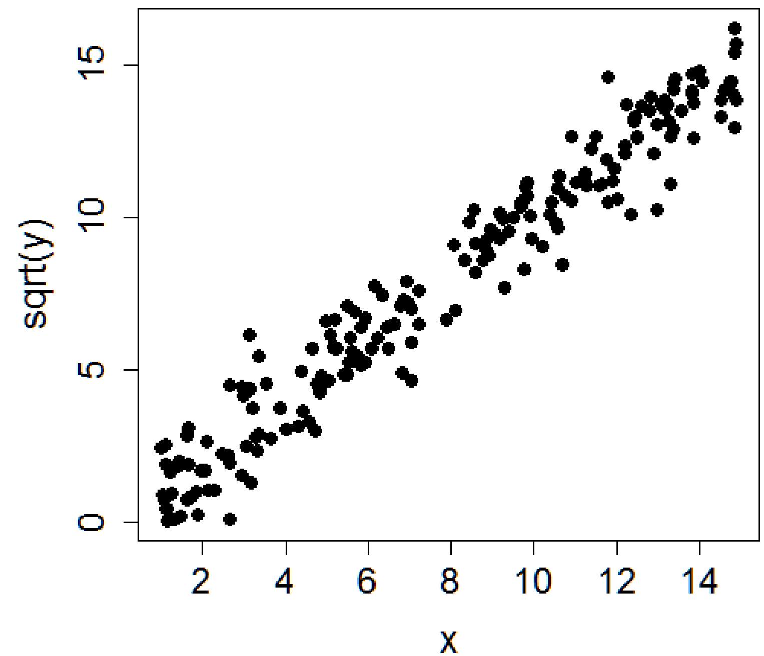


log(Y) ~ X



Transformations of the Response if it is Convex

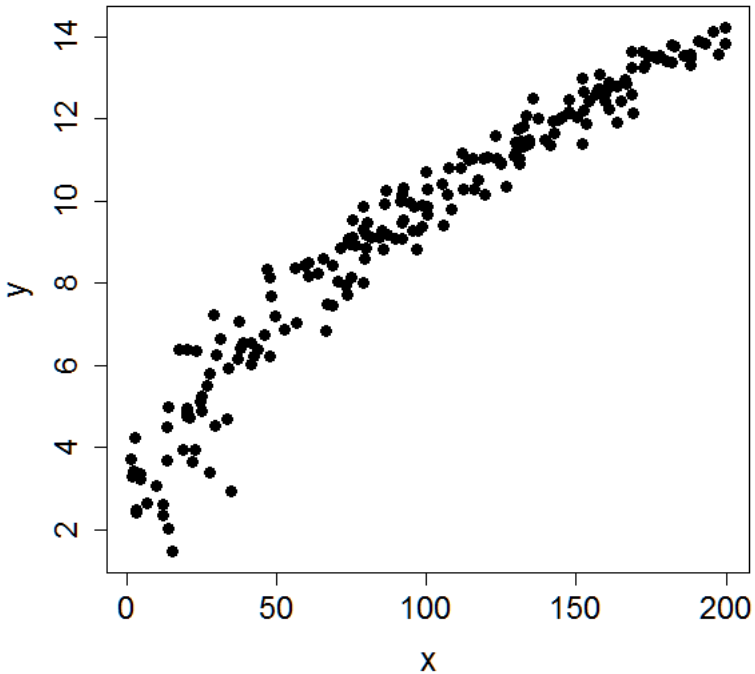


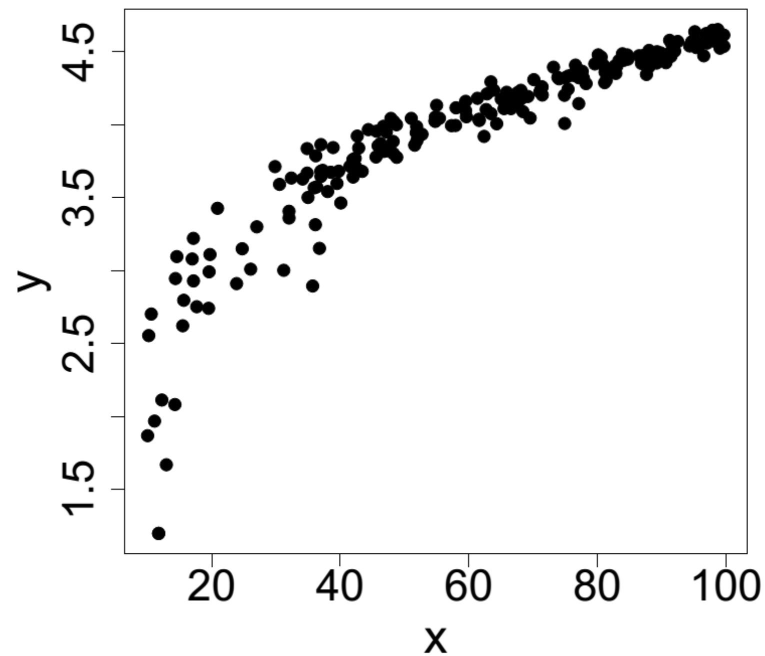


$$E(\sqrt{Y}|X) = \beta_0 + \beta_1 X$$

Transformations that compress larger values of Y more than smaller values: log(*Y*), sqrt(*Y*), $Y^{1/k}$, or 1/*Y* for *Y*>1.

Transformations of the Response if it is Concave



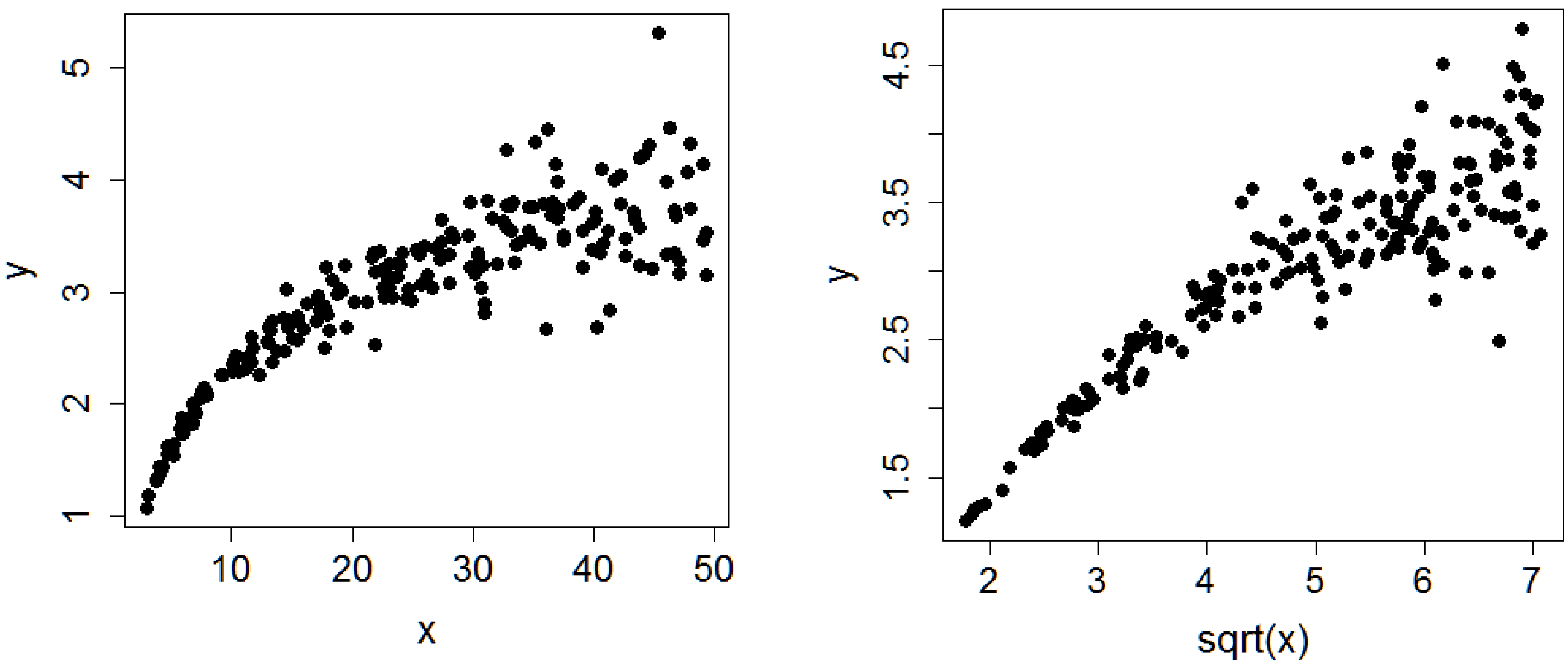


$$E(Y^2|X) = \beta_0 + \beta_1 X$$

$$E(e^Y|X) = \beta_0 + \beta_1 X$$

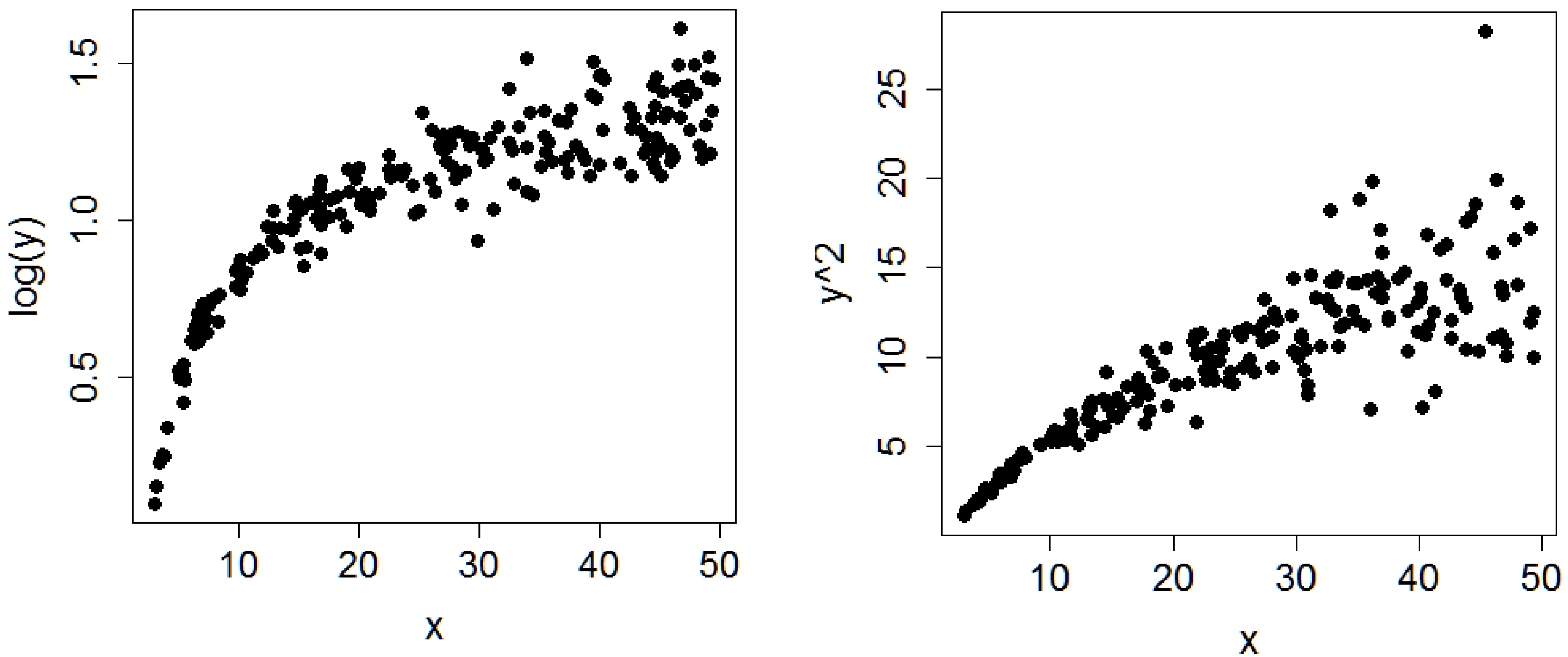
Transformations that expand larger values Y more than smaller values: $\exp(Y)$, Y^k, k^Y, or $1/Y$ for $0<Y<1$.

Which Transformation to Choose for Y?



Here, transformation of *X* does not help with unequal variances.

Which Transformation to Choose for Y?



- Transformations of Y will either help to correct unequal variances (e.g., $\log(Y)$) or non-linearity (e.g., Y^2).
- Possible solution: weighted regression (R&S Sec. 11.6.1)

Today's overview

- Lack-of-fit *F*-test for checking the goodness of fit of the linear regression model.
  - Pure error
  - Three models for population means
- Multiple Linear Regression
  - Motivation
  - Model
  - Interpretation of indicators, interactions

Reading:

- **Required:** Finish R&S Ch. 8, start Ch. 9, Ch. 9 R code
- **Supplementary Theory**: A. Sen and M. Srivastava. "Regression Analysis: Theory, Methods, and Applications", Ch. 2 (Multiple Regression, ignore Sec. 2.5, 2.9, 2.12) and Ch. 4 (Indicator Variables).

Lack-of-Fit F-test

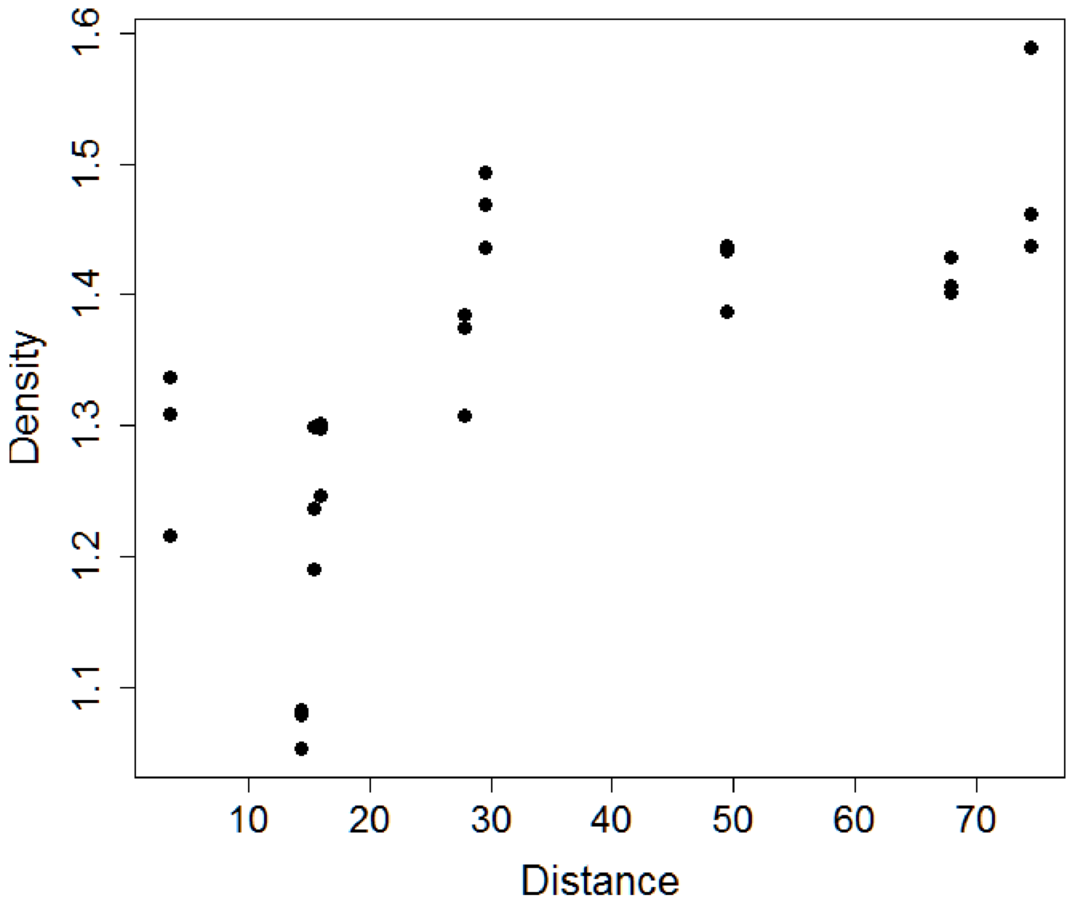
Example: Coral Reef

- We will examine data from 27 coral reef heads, *Porites lobata,* studied for nine different reefs that belong to the Great Barrier Reef, Australia.

- Risk and Sammarco (1991) found that the density of the coral skeletons *increases* with distance from the Australian shore due to differences in inshore and offshore environments.

Sample	Reef	Distance	Density
1	MiddleReef	3.5	1.337
2	MiddleReef	3.5	1.216
3	MiddleReef	3.5	1.309
4	AlmaBay	14.3	1.053
5	AlmaBay	14.3	1.082
…	...	...	...

Example: Coral Reef



Replications and Pure Error

- The estimate of residual variance in linear regression depends on the adequacy of the model.

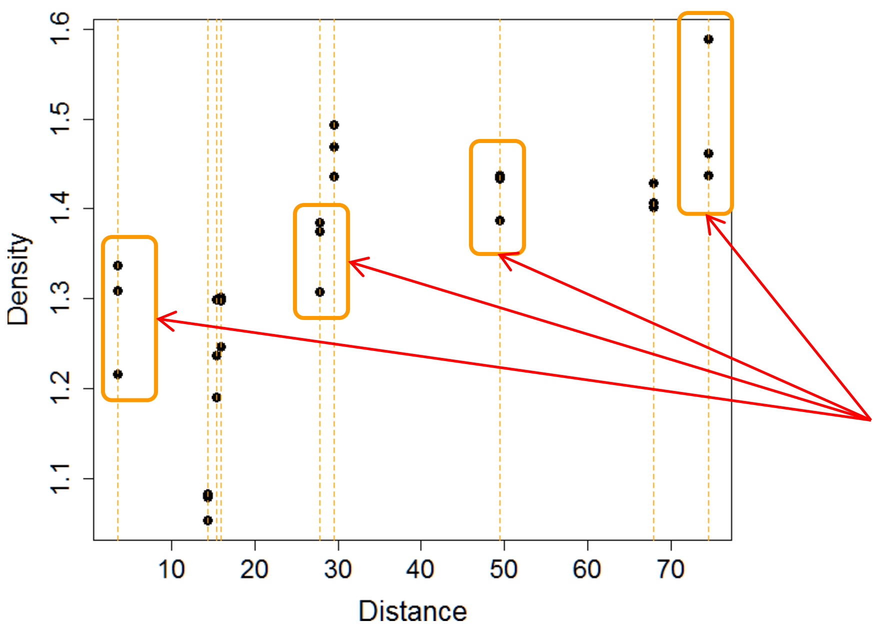
- Replication is repetition of an experiment or observation in the same or similar conditions.
  - I.e., when we have more than one unit at some levels of *X*.
  - It allows us to obtain

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(Y_{ij}-\overline{Y}_i\right)^2$$

and use if to get a *pure error* estimator of σ^2.

Replications and Pure Error

- When replicates are available at some or all values of *X*, a formal lack-of-fit *F*-test may be used to test the *adequacy of the straight-line regression model*,
  - i.e., compare simple LR model to separate-means (*one-way analysis of variance*) model.



Allow us to estimate the pure error.

Three Models for the Population Means

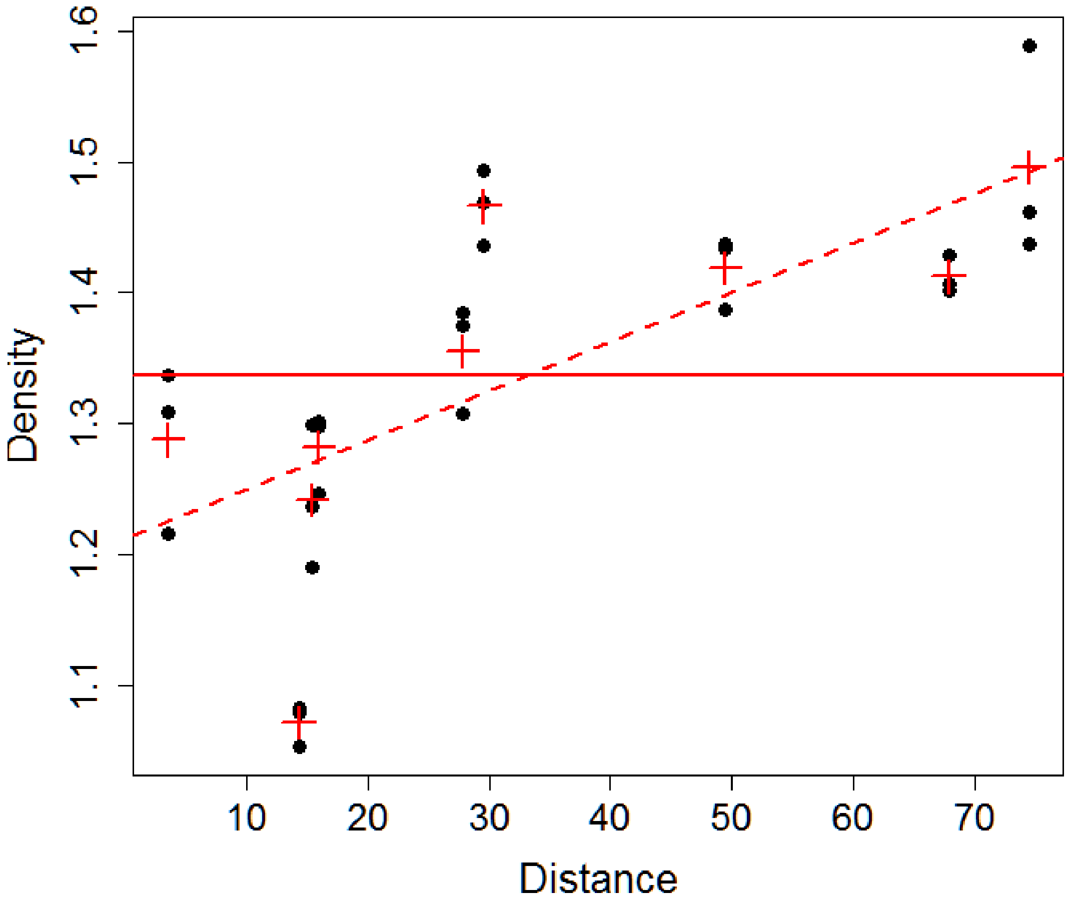
Separate Means: $E(Y \mid X = x_i) = \mu_i$

Simple Linear Regression: $E(Y \mid X = x_i) = \beta_0 + \beta_1 x_i$

Equal Means: $E(Y \mid X = x_i) = \mu = \beta_0$

These models are nested (i.e., form a *hierarchical set*).

Coral Reef: Three Models



+ $E(Y \mid X = x_i) = \mu_i$

- - $E(Y \mid X = x_i) = \beta_0 + \beta_1 x_i$

— $E(Y \mid X = x_i) = \mu = \beta_0$

Coral Reef: Equal-Means Model vs. Separate-Means Model

$$H_0 : E(Y \mid X = x_i) = \mu$$

$$H_a : E(Y \mid X = x_i) = \mu_i, \;\; i = 1,\ldots,9$$

$$n = 27, \;\; s_p^2 = 0.00205, \;\; s_Y^2 = 0.0175$$

Source	SSR	d.f.	Mean square	F-stat.	P-value
Between groups	0.4181	8	0.052	25.37	<0.0001
Within Groups	0.0369	18	0.00205		
Total	0.455	26	0.0175		

Conclusion?

In R: use `aov(Y ~ Groups)`

Sum of Squares Decomposition for the Simple Linear Regression

Sum of Squares	Calculation	D.f.	Distribu-tion	Mean Square
SSRes (full)	$\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$	n-2	$\sigma^2\chi^2_{n-2}$	$\hat{\sigma}^2$
SSReg (between)	$\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2$	1	$\sigma^2\chi^2_{1}$	
$\text{SSR}_{\text{reduced}}$ (reduced)	$\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2$	n-1	$\sigma^2\chi^2_{n-1}$	S_Y^2

$$R = \frac{\text{SSReg}/1}{\text{SSR}_{\text{full}}/(n-2)} \sim F_{1,n-2} \sim t_{n-2}$$

Coral Reef: Equal-Means Model vs. Simple Linear Regression Model

$$H_0 : E(Y \mid X = x_i) = \mu = \beta_0$$

$$H_a : E(Y \mid X = x_i) = \beta_0 + \beta_1 x_i$$

$$n = 27,\ \hat{\sigma}^2 = 0.0096,\ s_Y^2 = 0.0175$$

Source	SSR	d.f.	Mean square	F-stat.	P-value
Between	0.215	1	0.215	22.4	<0.0001
Within	0.240	25	0.0096		
Total	0.455	26	0.0175		

Conclusion?

In R: part of the `lm(Y~X)` output.

Coral Reef: Equal-Means Model vs. Simple Linear Regression Model

```
> regmodel <- lm(Density ~ Distance , data = coralData)
➢ summary(regmodel)
….
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.2119729  0.0324376  37.363  < 2e-16 ***
Distance    0.0037609  0.0007954   4.728 7.54e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09813 on 25 degrees of freedom
Multiple R-squared: 0.4721,      Adjusted R-squared: 0.4509
F-statistic: 22.35 on 1 and 25 DF,  p-value: 7.54e-05
-------------------------------------------------------
> AOV_RegModel <- aov(regmodel)
> summary(AOV_RegModel)
 Df Sum Sq Mean Sq F value   Pr(>F)
Distance     1 0.2153 0.21526   22.35 7.54e-05 ***
Residuals   25 0.2407 0.00963
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Coral Reef: Equal-Means Model vs. Simple Linear Regression Model

$$H_0 : E(Y \mid X = x_i) = \mu = \beta_0$$

$$H_a : E(Y \mid X = x_i) = \beta_0 + \beta_1 x_i$$

is equivalent to

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

Lack-of-Fit F-test

- A formal test of the adequacy of the straight-line regression model,

$$H_0 : E(Y \mid X = x_i) = \beta_0 + \beta_1 x_i \quad (\text{LR})$$

$$H_a : E(Y \mid X = x_i) = \mu_i \quad (\text{SM})$$

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

- Exact sampling distribution of R under H_0 is

$$F_{(\text{d.f.}_{\text{LR}} - \text{d.f.}_{\text{SM}},\, \text{d.f.}_{\text{SM}})}$$

Coral Reef: Simple Linear Regression Model vs. Separate-Means Model

- $SSRes_{LR}$ = 0.240, d.f. = 25 ("Within" variation)
- $SSRes_{SM}$ = 0.0369, d.f. = 18 ("Within" variation)

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

Verify that R = 14.2, and the d.f. for the F-distribution are 7, 18.

In R: `>anova(AOV_RegModel, AOV_modelSeparateMeans, test="F")`

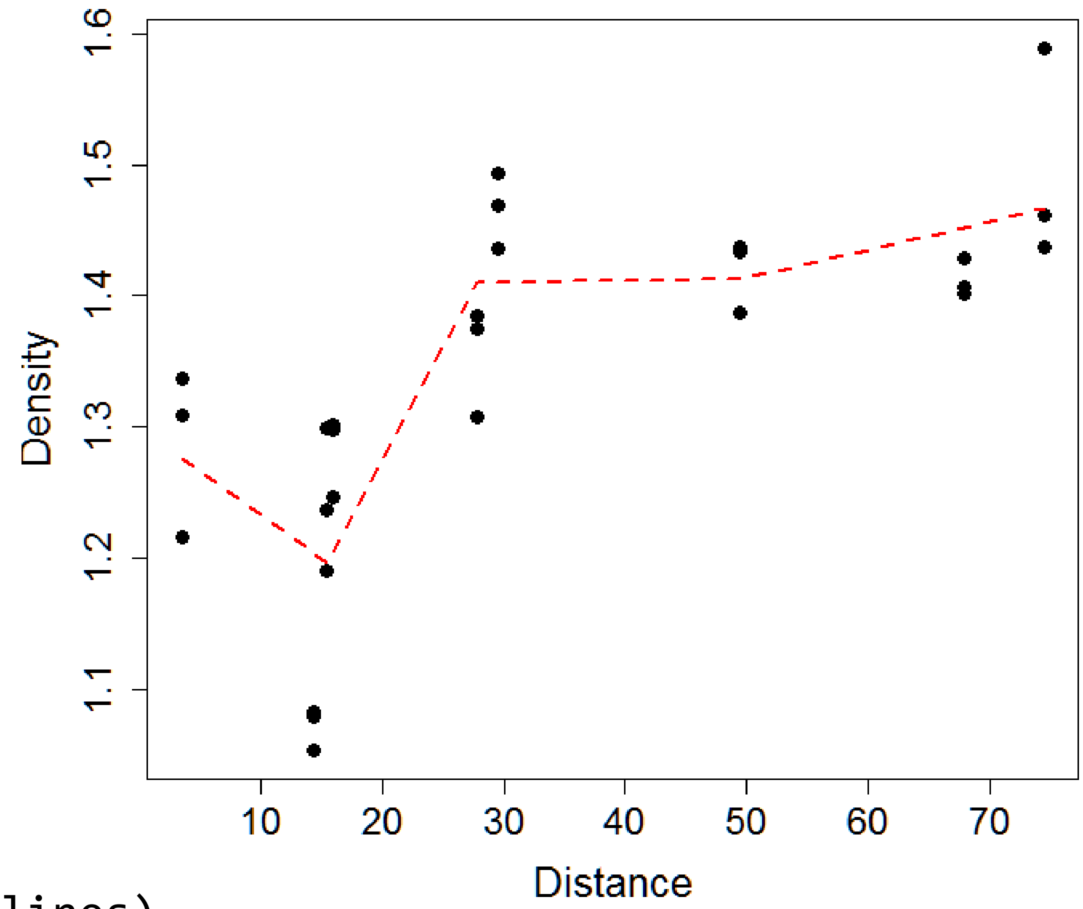
```
Analysis of Variance Table

Model 1: Density ~ Distance
Model 2: Density ~ as.factor(Distance)
  Res.Df      RSS Df Sum of Sq      F    Pr(>F)
1     25 0.240736
2     18 0.036908  7   0.20383 14.201 3.665e-06 ***
```


Conclusion?

R-code for the Lack-of-Fit *F*-test

```
AOV_modelSeparateMeans<- aov(Density ~ as.factor(Distance),
                                    data = coralData)


regmodel <- lm(Density ~ Distance , data = coralData)
AOV_RegModel <- aov(regmodel)


anova(AOV_RegModel, AOV_modelSeparateMeans, test="F")
```

Coral Reef: Spline Regression




```
library(splines)
SplineReg <- lm(Density ~ bs(Distance, degree=1, knots = 3),
  data = coralData)
```

Separate-Means Model v.s. Simple Linear Regression

When both fit, prefer regression:

- Allows interpolation.
- Fewer parameters -> more degrees of freedom for error estimation, $\hat{\sigma}^2$,
- … and, therefore, smaller standard errors for parameters and predicted responses.

Multiple Linear Regression

How Much Is Your Car Worth?

A representative sample of over eight hundred 2005 GM cars was selected, then retail price was calculated from the tables provided in the 2005 Central Edition of the Kelly Blue Book.

Variables:

1. Price: suggested retail price of the used 2005 GM car in excellent condition. The condition of a car can greatly affect price. All cars in this data set were less than one year old when priced and considered to be in excellent condition.
2. Mileage: number of miles the car has been driven
3. Make: manufacturer of the car such as Saturn, Pontiac, Chevrolet, etc.
4. Model: specific models for each car manufacturer such as Ion, Vibe, Cavalier, etc.
5. Trim (of car): specific type of car model such as Sedan 4D, Quad Coupe 2D, etc…

How Much Is Your Car Worth?

6. Type: body type such as sedan, coupe, etc.
7. Cylinder: number of cylinders in the engine
8. Liter: a more specific measure of engine size
9. Doors: number of doors
10. Cruise: indicator variable representing whether the car has cruise control (1 = cruise)
11. Sound: indicator variable representing whether the car has upgraded speakers (1 = upgraded)
12. Leather: indicator variable representing whether the car has leather seats (1 = leather)

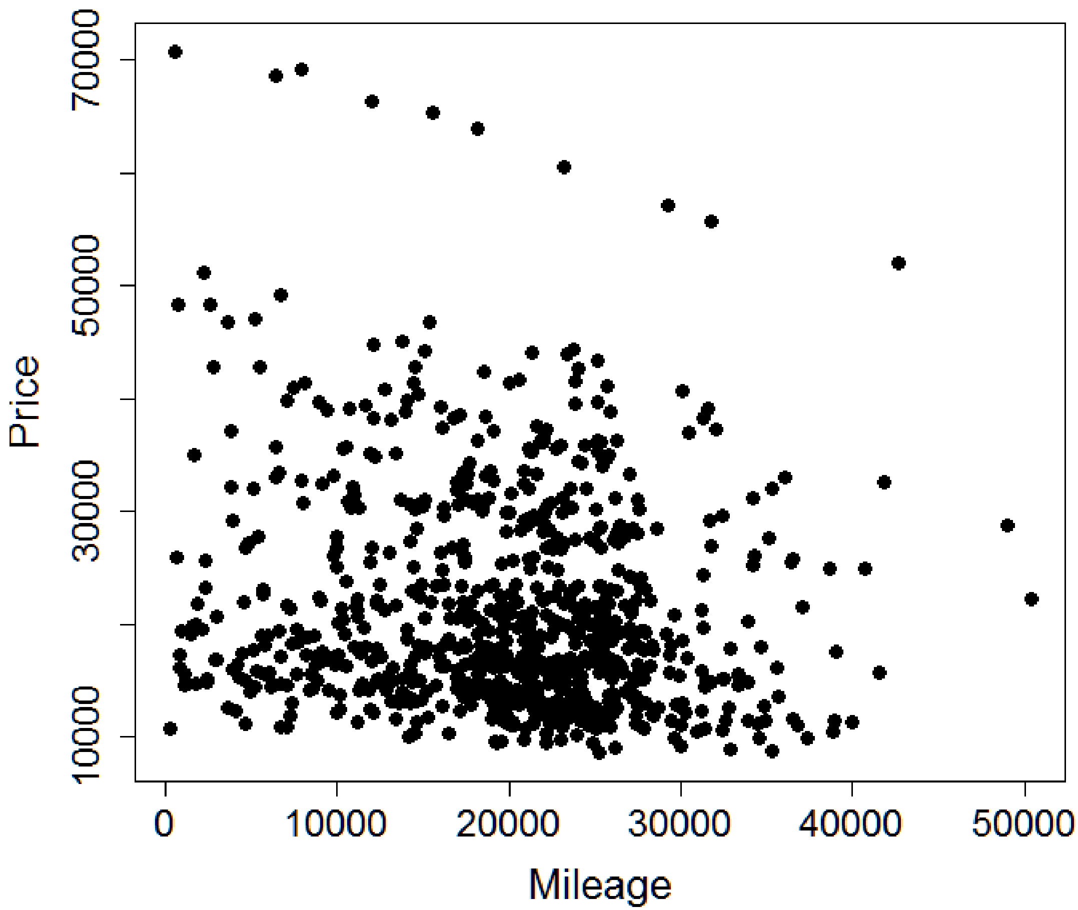
How Much Is Your Car Worth?

A representative sample of over eight hundred 2005 GM cars were selected, then retail price was calculated from the tables provided in the 2005 Central Edition of the Kelly Blue Book.

	Price	Mileage	Make	Model	Trim	Type	Cylinder	Liter	Doors	Cruise	Sound	Leather
1	17314.10	8221	Buick	Century	Sedan 4D	Sedan	6	3.1	4	1	1	1
2	17542.04	9135	Buick	Century	Sedan 4D	Sedan	6	3.1	4	1	1	0
3	16218.85	13196	Buick	Century	Sedan 4D	Sedan	6	3.1	4	1	1	0
4	16336.91	16342	Buick	Century	Sedan 4D	Sedan	6	3.1	4	1	0	0
5	16339.17	19832	Buick	Century	Sedan 4D	Sedan	6	3.1	4	1	0	1
6	15709.05	22236	Buick	Century	Sedan 4D	Sedan	6	3.1	4	1	1	0

Given that all cars were in excellent condition, what is one of the most important determining factors of the price of a car?

How Much Is Your Car Worth?



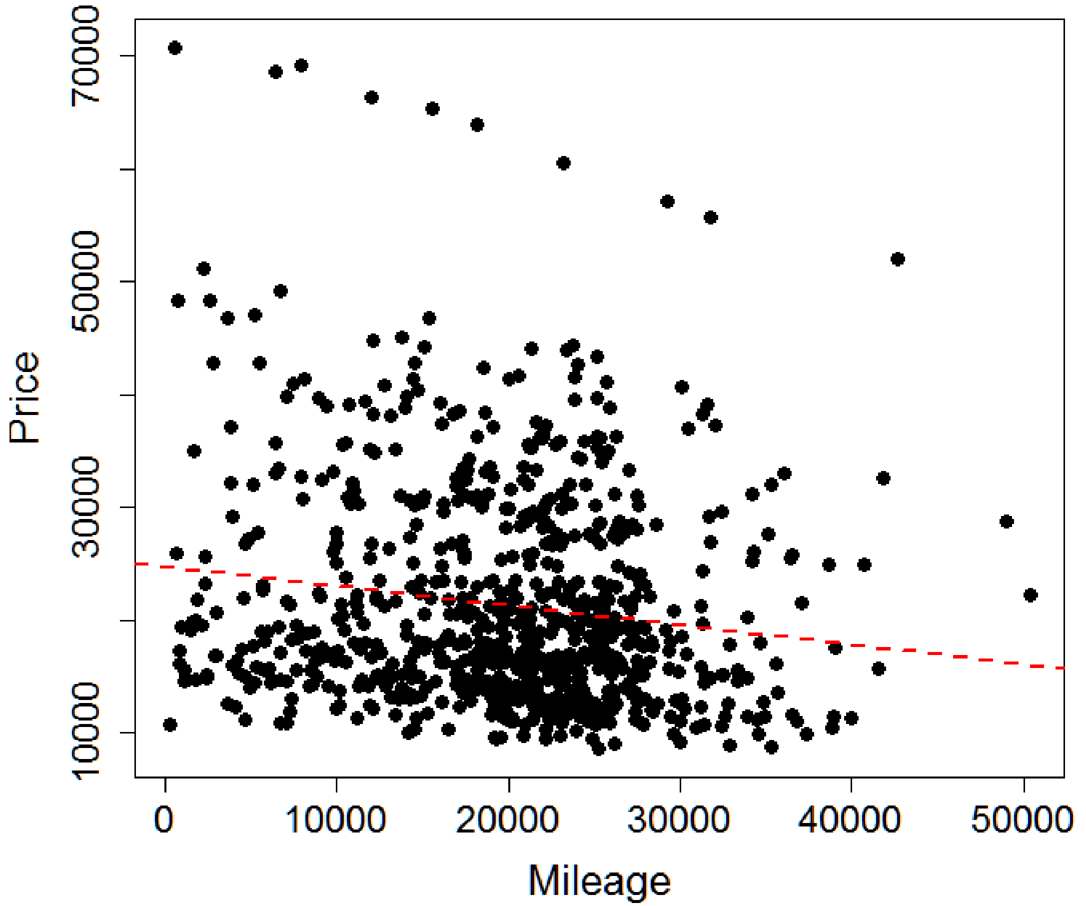
Modeling Price: Simple LR

```
> regmodel <- lm(Price ~ Mileage, data = CarData)
> summary(regmodel)

Call:
lm(formula = Price ~ Mileage, data = CarData)
…
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     24,760  9.044e+02   27.383  < 2e-16 ***
Mileage        -0.1725  4.215e-02  -4.093   4.68e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9789 on 802 degrees of freedom
Multiple R-squared: 0.02046,     Adjusted R-squared: 0.01924
F-statistic: 16.75 on 1 and 802 DF,  p-value: 4.685e-05
```

Modeling Price: Simple LR

$$\hat{\mu}(\text{Price}) = 24{,}760 - 0.173 \cdot \text{Mileage}$$

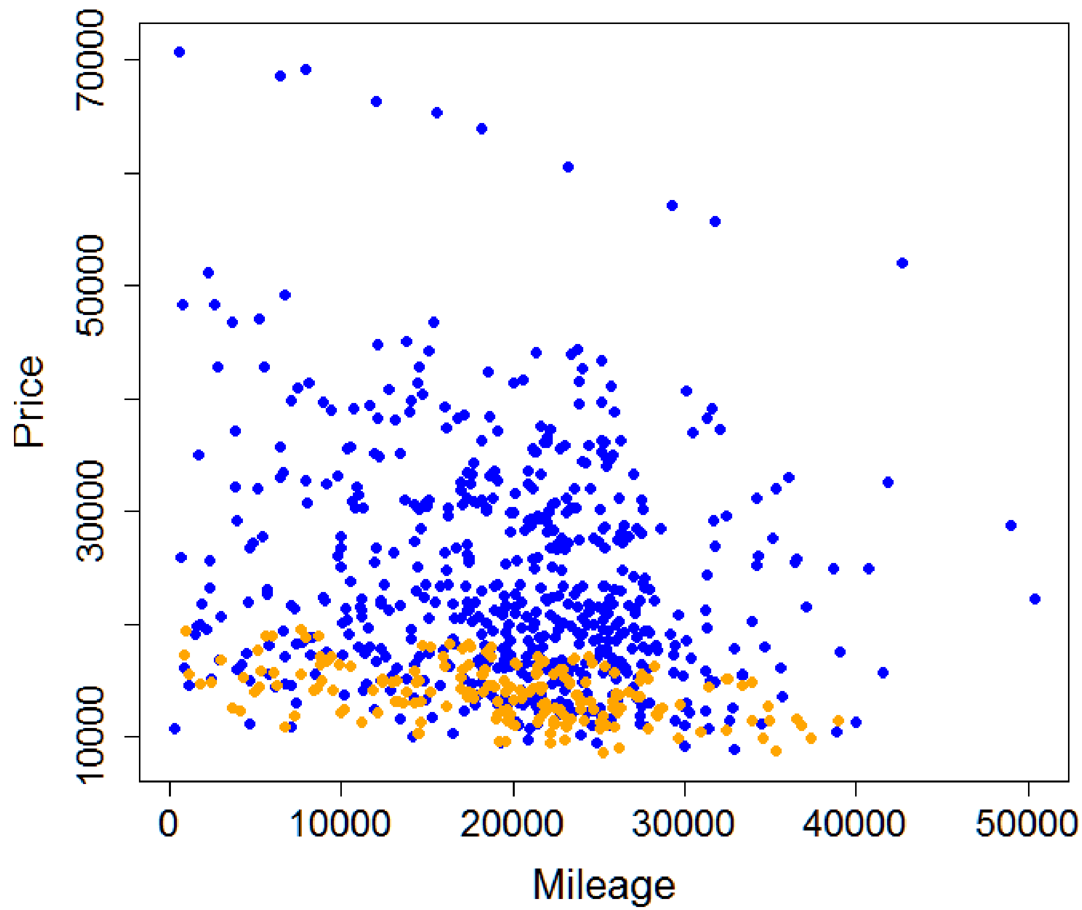
$$\text{SE } (p\text{ - value}) \quad 904 (< 0.001) \qquad 0.04 (< 0.001)$$



R-squared: **0.02046**

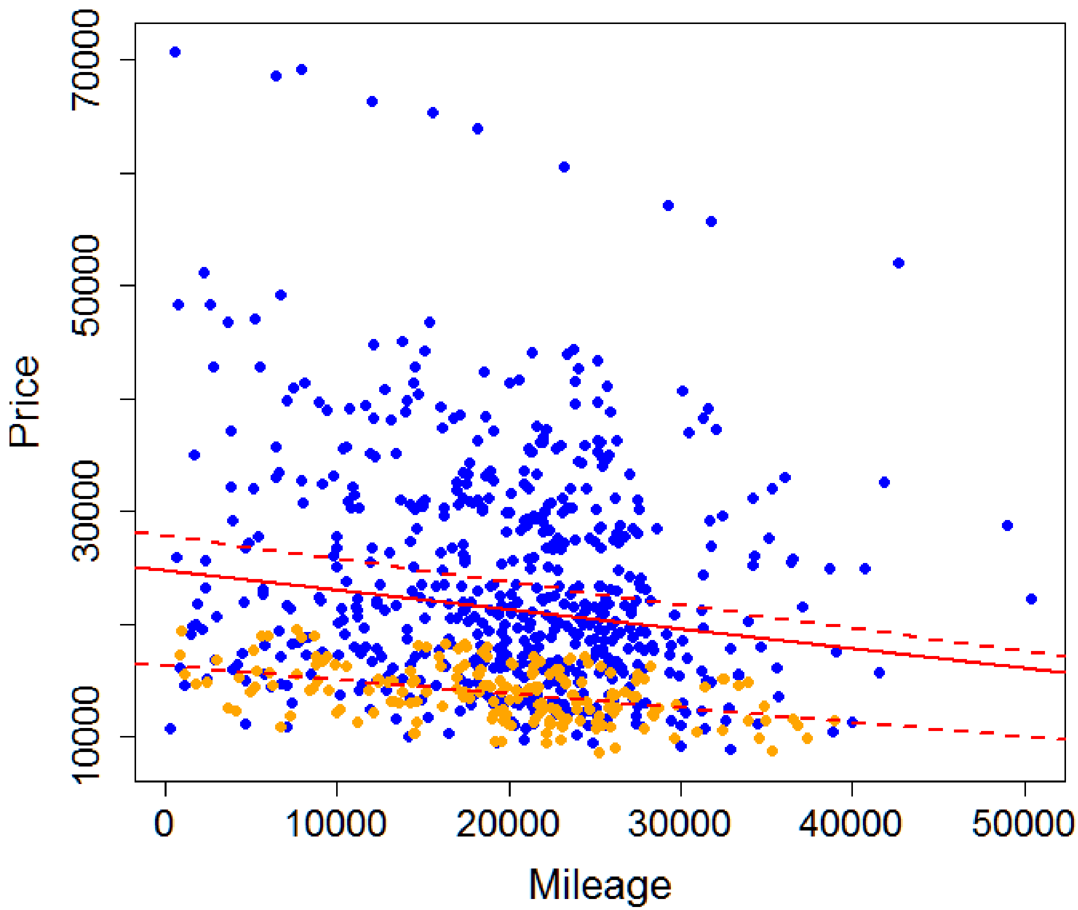
Modeling Price: Split the Data in Two Groups

with cruise control and without cruise control



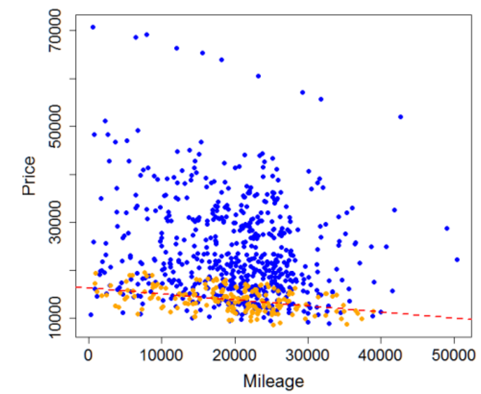
Modeling Price: Split the Data in Two Groups

with cruise control and without cruise control

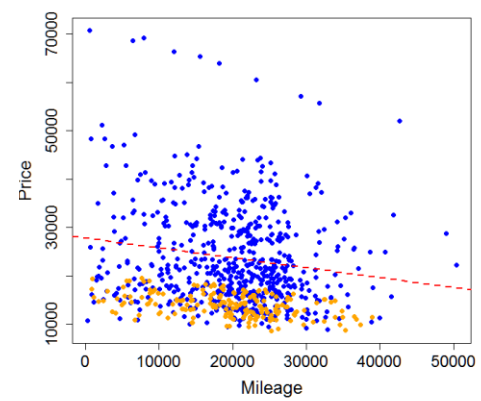


Modeling Price: Simple LR

```
lm(formula = Price ~ Mileage, data = CarData, subset = (Cruise == 0))
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   16,380     398.1     41.15   < 2e-16 ***
Mileage      -0.1262     0.0189   -6.69    2.25e-10 ***

Residual standard error: 2160 on 197 degrees of freedom
Multiple R-squared: 0.1851, Adjusted R-squared: 0.181
```



```
lm(formula = Price ~ Mileage, data = CarData, subset = (Cruise == 1))
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    27,870     1.075    25.93  < 2e-16 ***
Mileage       -0.2047     0.0498  -4.11   4.51e-05 ***

Residual standard error: 10060 on 603 degrees of freedom
Multiple R-squared: 0.02725, Adjusted R-squared: 0.02563
```



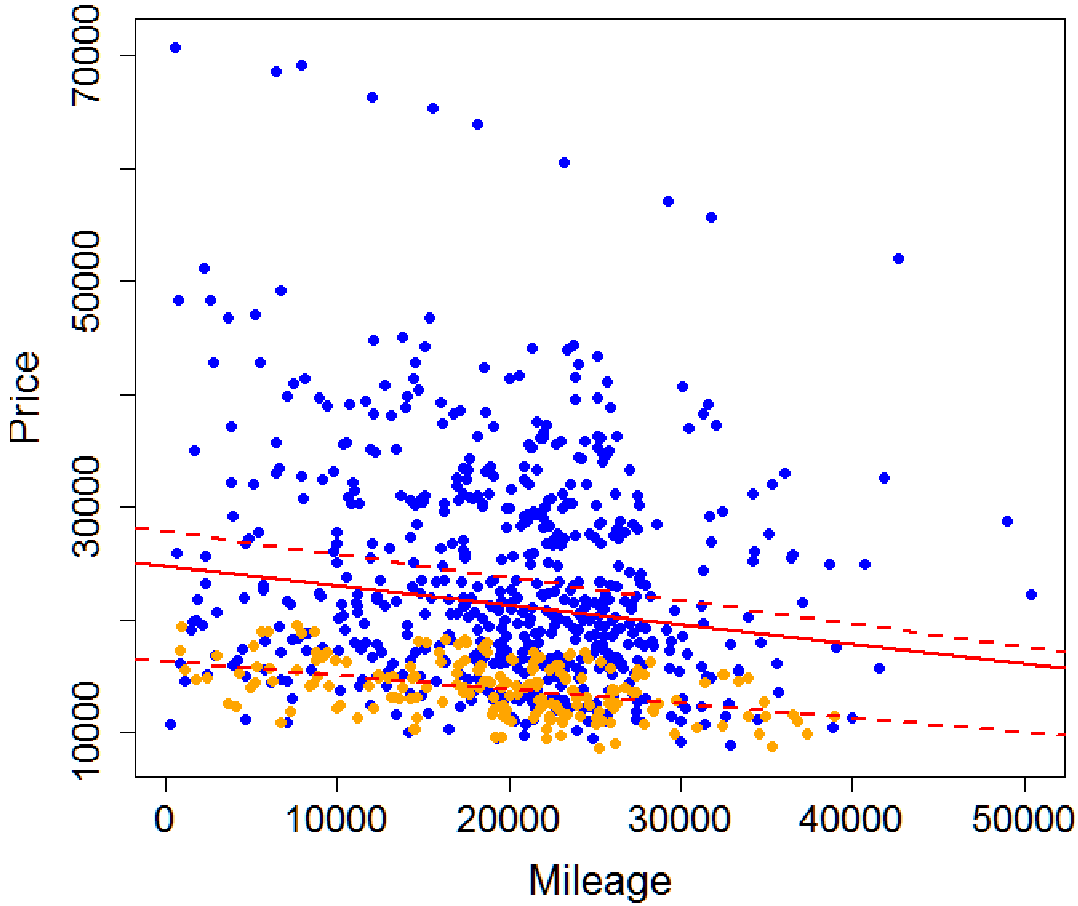
Modeling Price:
Split the Data in Two Groups

with cruise control and without cruise control

β_1=-0.2047
SE=0.05

β_1=-0.173
SE=0.04

β_1=-0.126
SE=0.019



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 vs. Joint Models

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

Unadjusted model:

β_1=-0.173

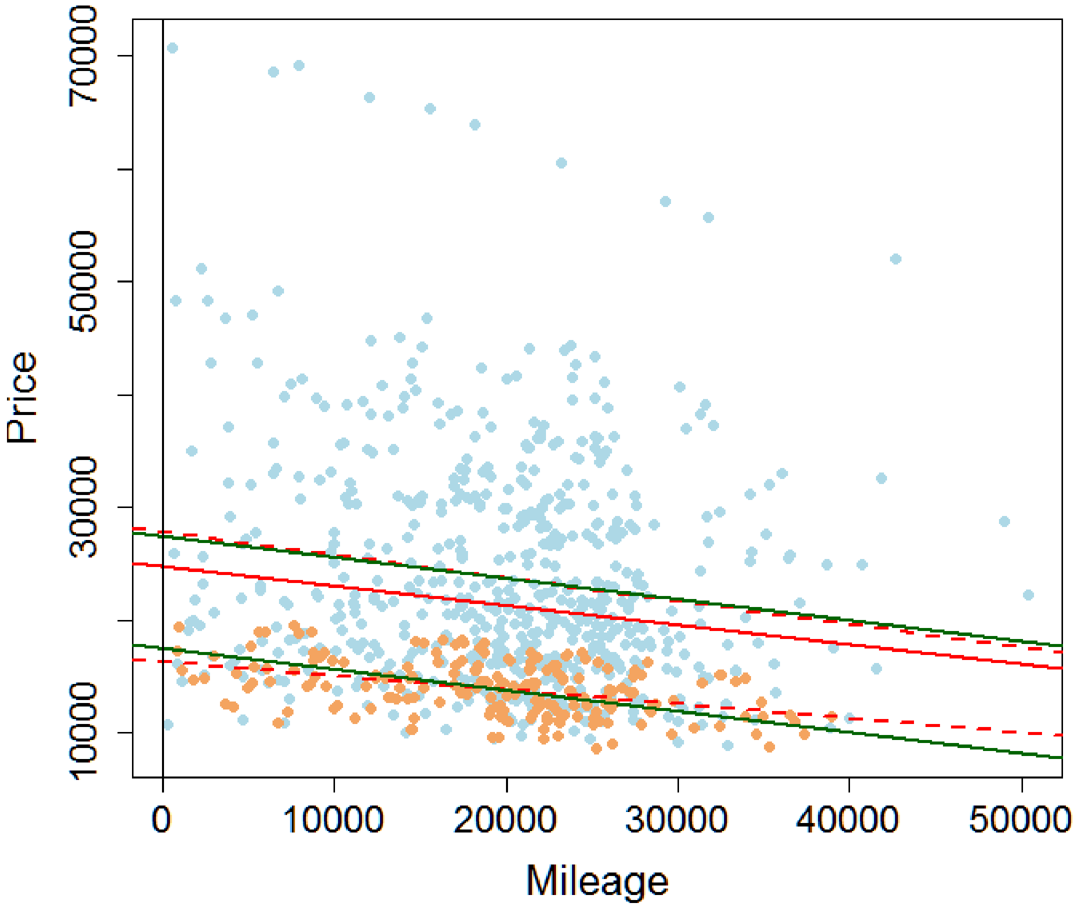
Separate models:

CC:

β_1=- 0.205

No CC:

β_1=-0.126



Joint model :

CC:

β_1=-0.186

No CC:

β_1=-0.186

Modeling Price: Interpretation of Parameters of the Joint Model

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

Interpretation of the first slope:

- $\beta_1 = -0.186$ is the change in *average Price* when *Mileage* increases by one **after adjusting for cruise control**.
  - The textbook uses "*effect*" – not the best choice.
  - For a subpopulation of GM cars *with cruise control*, the estimated reduction in *average Price* is \$0.186 per one additional mile.
  - For a subpopulation of GM cars *without cruise control*, the estimated reduction in *average Price* will be \$0.186 per one additional mile.
  - Note it is NOT the same as *unadjusted estimate* $\beta_1 = -0.173$.

Modeling Price: Interpretation of Parameters of the Joint Model

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

Interpretation of the intercept:

- $\beta_0 = 17{,}537$ is the *average Price* of a car if *Mileage*=0 and *Cruise*=0 (no cruise control).
  - Interpretation depends on what level of each indicator variable was chosen to be the reference level.
  - Sometimes, it is advised to center continuous *X*'s to make the intercept *more interpretable*,

$$\hat{\mu}(\text{Price} \mid \text{Mileage*, Cruise}) = 13{,}855 - 0.186 \cdot \text{Mileage}* + 9{,}951 \cdot \text{Cruise},$$

$$\text{where } \text{Mileage*} = \text{Mileage} - \overline{\text{Mileage}}$$

Modeling Price: Interpretation of Parameters of the Joint Model

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

Interpretation of the second slope:

- $\beta_2 = 9{,}951$ is the change in *average Prices* of a car with and without cruise control **after adjusting for Mileage**.
  - For a subpopulation of cars with *Mileage=Mileage*$_0$, the difference in average prices of cars with and without cruise control is estimated to be $9,951.

$$\hat{\mu}(\text{Price} \mid \text{Mileage}_0, \text{Cruise} = 0) = \beta_0 + \beta_1 \text{Mileage}_0 = 17{,}537 - 0.186 \cdot \text{Mileage}_0$$

$$\hat{\mu}(\text{Price} \mid \text{Mileage}_0, \text{Cruise} = 1) = (\beta_0 + \beta_2) + \beta_1 \text{Mileage} = 27{,}488 - 0.186 \cdot \text{Mileage}_0$$