IEMS 304 Lecture 3: Multiple Linear Regression

Yiping Lu
yiping.lu@northwestern.edu

Industrial Engineering & Management Sciences
Northwestern University

Data Model

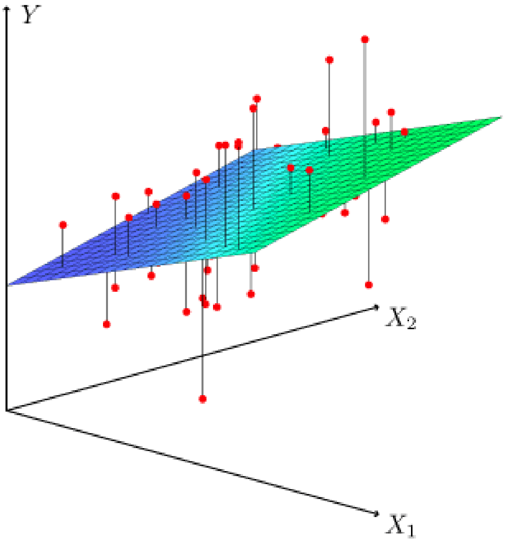
- There are p variables, $X_1, X_2, \ldots, X_p$. The variables can have arbitrary distributions, possibly deterministic. In particular, they may or may not be dependent. *Notation:* The single X refers to the collection of all these p variables.
- There is a scalar response variable $Y = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \varepsilon$, for some constants $\beta_0, \ldots, \beta_p$. Therefore there are $p+1$ coefficients.
- The noise variable ε has $\mathbb{E}[\varepsilon | X = x] = 0$ and $\mathrm{Var}(\varepsilon | X = x) = \sigma^2$, and is uncorrelated across observations.

In matrix form,

$$\mathrm{Y}_{n\times 1} = \mathrm{X}_{n\times(p+1)} \beta_{(p+1)\times 1} + \varepsilon_{n\times 1}.$$

Matrix View

Data Model



Least Square

Following the least-squares procedure, we solve for the estimator of β by minimizing the MSE:

$$\begin{aligned}
\widehat{\mathrm{MSE}} &= \frac{1}{n}(\mathsf{Y} - \mathsf{X}\beta)^{\mathsf{T}}(\mathsf{Y} - \mathsf{X}\beta) \\
&= \frac{1}{n}(\mathsf{Y}^{\mathsf{T}}\mathsf{Y} - 2\beta^{\mathsf{T}}\mathsf{X}^{\mathsf{T}}\mathsf{Y} + \beta^{\mathsf{T}}\mathsf{X}^{\mathsf{T}}\mathsf{X}\beta) \\
0 \overset{\text{set}}{=} \nabla_{\hat{\beta}}\widehat{\mathrm{MSE}} &= -2\mathsf{X}^{\mathsf{T}}\mathsf{Y} + 2\mathsf{X}^{\mathsf{T}}\mathsf{X}\hat{\beta} \\
\hat{\beta} &= (\mathsf{X}^{\mathsf{T}}\mathsf{X})^{-1}\mathsf{X}^{\mathsf{T}}\mathsf{Y}
\end{aligned}$$

In addition, the in-sample MSE is $\hat{\sigma}^2 = \frac{1}{n}\mathsf{e}^{\mathsf{T}}\mathsf{e}$, the mean squared *residuals*.

Matrix View

Bias and Variance

- Expectation: $\hat{\beta}$ is unbiased, i.e. $\mathbb{E}[\hat{\beta}|\mathsf{X}] = \beta$.
- Variance: $\mathrm{Var}(\hat{\beta}|\mathsf{X}) = \sigma^2(\mathsf{X}^\mathsf{T}\mathsf{X})^{-1}$

In-Sample MSE

$$\hat{\sigma}^2 = \frac{\sigma^2}{n}(n - (p + 1))$$

Degree of Freedom

Suppose that we observe

$$y_i = r(x_i) + \epsilon_i, \quad i = 1, \ldots, n,$$

where the errors ϵ_i, $i = 1, \ldots, n$ are uncorrelated with common variance $\sigma^2 > 0$. Now consider the fitted values $\hat{y}_i = \hat{r}(x_i)$, $i = 1, \ldots, n$ from a regression estimator $\hat{r}$. We define the **degrees of freedom** of $\hat{r}$ as

$$\mathrm{df}(\hat{y}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathrm{Cov}(\hat{y}_i, y_i).$$

Fact. $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(y_i' - \hat{y}_i)^2\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right] + \frac{2\sigma^2}{n}\,\mathrm{df}(\hat{y})$.

Example

- Simple average estimator: consider $\hat{y}^{\text{ave}} = (\bar{y}, \ldots, \bar{y})$, where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Then

$$\text{df}(\hat{y}^{\text{ave}}) = \frac{1}{\sigma^2}\sum_{i=1}^{n}\text{Cov}(\bar{y}, y_i) = \frac{1}{\sigma^2}\sum_{i=1}^{n}\frac{\sigma^2}{n} = 1,$$

  i.e., the effective number of parameters used by $\text{df}(\hat{y}^{\text{ave}})$ is just 1, which makes sense.

- Identity estimator: consider $\hat{y}^{\text{id}} = (y_1, \ldots, y_n)$. Then

$$\text{df}(\hat{y}^{\text{id}}) = \frac{1}{\sigma^2}\sum_{i=1}^{n}\text{Cov}(y_i, y_i) = n,$$

  i.e., $\hat{y}^{\text{id}}$ uses n effective parameters, which again makes sense.

Degree of Freedom for Linear Prediction (NOT REQUIRED)

$$\mathrm{df}(\hat{y}^{\mathrm{linreg}}) = p$$

$$\begin{aligned}
\mathrm{df}(\hat{y}^{\mathrm{linreg}}) &= \frac{1}{\sigma^2}\,\mathrm{tr}\big(\mathrm{Cov}\big(X(X^TX)^{-1}X^Ty, y\big)\big) \\
&= \frac{1}{\sigma^2}\,\mathrm{tr}\big((X^TX)^{-1}X^T\,\mathrm{Cov}(y, y)\big) \\
&= \mathrm{tr}\big(X(X^TX)^{-1}X^T\big) \\
&= \mathrm{tr}\big(X^TX(X^TX)^{-1}\big) = p
\end{aligned}$$

$\mathbf{tr}(AB) = \mathbf{tr}(BA)$

Accessing the Fit

Accessing the Fit

- As in simple regression, we calculate
  - fitted values: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik}$;
  - residuals: $e_i = y_i - \hat{y}_i$;
  - error sum of squares: $\mathrm{SSE} = \sum_{i=1}^n e_i^2$;
  - total sum of squares: $\mathrm{SST} = \sum_{i=1}^n (y_i - \bar{y})^2$;
  - regression sum of squares: $\mathrm{SSR} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$.
- We still have the decomposition

$$\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE}.$$

r^2 for Multiple Regression

- We can still look at $r^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}$;
- In multiple regression, r^2 is called **coefficient of multiple determination**. It still represents the proportion of variability in response that is accounted for by its linear dependence on the set of predictors;
- Mathematically, r^2 is equivalent to the square of the sample correlation coefficient between Y and $\hat{Y}$;
- **Beware**: r^2 is artificially high when $n \not\gg k$ because of **overfitting** — use something called "adjusted r^2" instead (coming up soon).

Fitting a Polynomial Using Linear Regression

Consider fitting a polynomial of degree p to data $\{(x_i, y_i)\}$:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p + \epsilon.$$

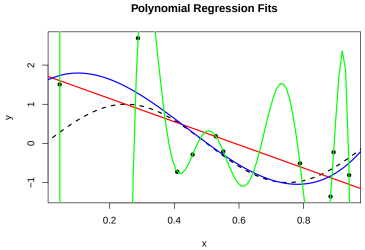
Define new variables: $z_1 = x, \quad z_2 = x^2, \quad \ldots, \quad z_p = x^p$. Then, the model can be written as:

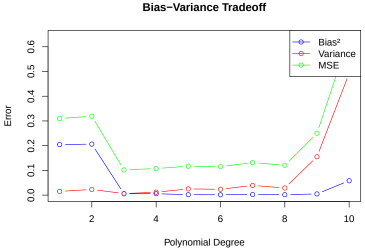
$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_p z_p + \epsilon,$$

which is linear in the parameters $\beta_0, \beta_1, \ldots, \beta_p$.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^p \\ 1 & x_2 & x_2^2 & \cdots & x_2^p \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^p \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

Is More Feature Better? (Homework)





Definition of r^2_{adj}

- Recall
$$r^2 = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\frac{\text{SSE}}{n-1}}{\frac{\text{SST}}{n-1}};$$
- Define the "mean squares" corresponding to the "sum of squares":
$$\text{MSE} = \frac{\text{SSE}}{n-(k+1)},$$
$$\text{MST} = \frac{\text{SST}}{n-1};$$
- For multiple regression, instead of r^2 you should look at "adjusted r^2":
$$r^2_{\text{adj}} = 1 - \frac{\text{MSE}}{\text{MST}} = 1 - \frac{\text{SSE}}{\text{SST}} \cdot \frac{n-1}{n-(k+1)}.$$

Statistical Inference

Statistical Inference on Coefficients

- A regression fit can seem **practically significant** (high r^2) without being **statistically significant**, and vice-versa.
- Three common tests of whether individual parameters or groups of parameters differ from zero are
  - A t-test of whether $\beta_j = 0$ is essentially testing whether including/excluding the individual predictor x_j in the model significantly changes the SSE.
  - For example, the t-test for β_1 compares the following two models:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon,$$

$$Y = \beta_0 + \qquad\qquad + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon.$$

t-test

$$H_0 : \beta_j = c \quad \text{v.s.} \quad H_1 : \beta_j \neq c$$

t-tests on Individual Coefficients

In order to develop a t-test on individual coefficients, we need the following statistical facts regarding the distribution of the estimated parameters $\hat{\beta}_j$ for $j = 0, 1, \ldots, k$:

- For $j = 0, 1, \ldots, k$, $\hat{\beta}_j \sim \mathsf{N}(\beta_j, \sigma^2 v_{jj})$, where v_{jj} denotes the $(j+1)$-th diagonal element of $\mathsf{V} = (\mathsf{X}^\top \mathsf{X})^{-1}$.
- That is, $\hat{\beta}_j$ is normally distributed with mean $\mathbb{E}[\hat{\beta}_j] = \beta_j$ and standard deviation $\mathrm{SD}(\hat{\beta}_j) = \sigma\sqrt{v_{jj}}$.
- Thus, a measure of precision in estimating β_j is

$$\mathrm{SE}(\hat{\beta}_j) = s\sqrt{v_{jj}},$$

where $s^2 = \mathrm{MSE} = \frac{\mathrm{SSE}}{n-(k+1)}$.

Fact. $\frac{\hat{\beta}_j - \beta_j}{\mathrm{SE}(\hat{\beta}_j)} \sim t_{n-(k+1)}$.

t-tests on Individual Coefficients Cont'd

To test

$$H_0 : \beta_j = c \quad \text{v.s.} \quad H_1 : \beta_j \neq c$$

for some specified constant c, e.g., $c = 0$.

- A two-sided $1 - \alpha$ confidence interval for β_j is

$$\hat{\beta}_j \pm t_{n-k-1,\alpha/2} \cdot \mathrm{SE}(\hat{\beta}_j).$$

- Use test statistic

$$t_j = \frac{\hat{\beta}_j - c}{\mathrm{SE}(\hat{\beta}_j)}.$$

- Reject H_0 if

$$|t_j| > t_{n-(k+1),\alpha/2} \quad \text{or} \quad c \text{ not in the confidence interval.}$$

Example: t-tests on All the Predictors

- We fit a multiple linear model on all 11 predictors.

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	17.339838	30.355375	0.571	0.5749
Displacement	-0.075588	0.056347	-1.341	0.1964
Hpower	-0.069163	0.087791	-0.788	0.4411
Torque	0.115117	0.088113	1.306	0.2078
Comp_ratio	1.494737	3.101464	0.482	0.6357
Rear_axle_ratio	5.843495	3.148438	1.856	0.0799
Carb_barrels	0.317583	1.288967	0.246	0.8082
No._speeds	-3.205390	3.109185	-1.031	0.3162
Length	0.180811	0.130301	1.388	0.1822
Width	-0.397945	0.323456	-1.230	0.2344
Weight	-0.005115	0.005896	-0.868	0.3971
Trans._type	0.638483	3.021680	0.211	0.8350

Almost all predictors are not statistically important?

Example: t-tests on Two Predictors

- We fit a multiple linear model on only two predictors: Rear_axle_ratio and Weight.

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    31.7594958  5.8348313   5.443 7.41e-06
Rear_axle_ratio  2.2141129  1.3146877   1.684    0.103
Weight         -0.0051025  0.0007106  -7.181 6.63e-08
```

- Why Weight becomes much more significant?

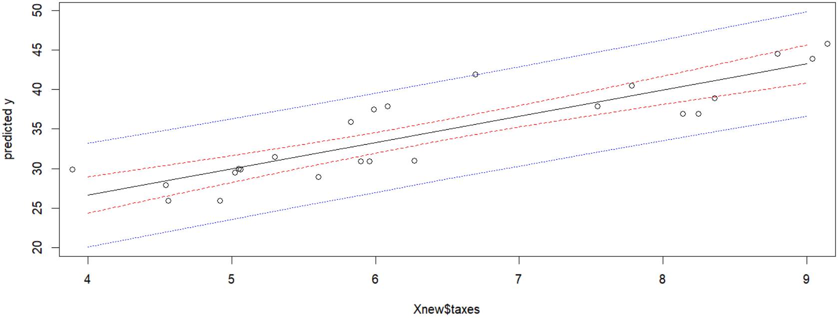
Prediction

Using the Model for Prediction

- For a fixed set of predictor values $(x_1^*, x_2^*, \ldots, x_k^*)$ for a new case, two "future" things on which we may want to make inferences are:
  - **actual response**: $Y^* = \beta_0 + \beta_1 x_1^* + \cdots + \beta_k x_k^* + \epsilon$;
  - **response mean**: $\mu^* = \mathbb{E}[Y^*] = \beta_0 + \beta_1 x_1^* + \cdots + \beta_k x_k^*$.
- The best **point prediction/estimate** is the same for both and is obvious (plug the predictors and the estimated coefficients into the model).
- If we want an interval that represents the uncertainty in the prediction/estimate, we use either:
  - A **confidence interval (CI) on** μ^* (considers uncertainty in the β's), or
  - A **prediction interval (PI) on** Y^* (considers uncertainty in the β's and in ϵ).

Example: Predicting Property Value

- `property_value.txt` contains home sales prices and nine other characteristics (taxes, lot size, living space, age, etc.) for a sample of 24 houses. The objective is to predict the sales price as a function of the other characteristics.
- We only use taxes to predict sales price.



Questions and Discussions

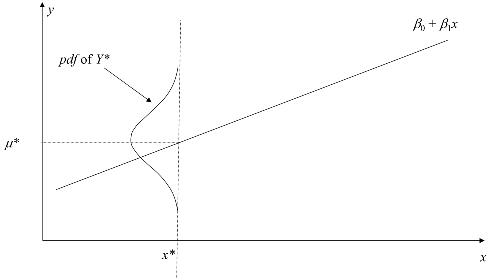
- Which is the PI and which is the CI in the previous figure?
- What is the interpretation of the PI?
- What is the interpretation of the CI?
- If someone is putting their house up for sale and wants to know the high end of the range for which it might sell, would the response PI or CI be more relevant?
- What is the relationship between the CI on μ^* versus a CI on one of the coefficients?
- How are the CI and PI calculated?

The Statistical View of Y^*

- For fixed x^*, we write

$$Y^* = \underbrace{\beta_0 + \beta_1 x_1^* + \cdots + \beta_k x_k^*}_{\mu^*} + \epsilon \sim \mathrm{N}(\mu^*, \sigma^2).$$

- In vector notation, we have $Y^* = (\mathrm{x}^*)^\top \boldsymbol{\beta} + \epsilon$.
- Point estimate of μ^* is $(\mathrm{x}^*)^\top \hat{\boldsymbol{\beta}}$.
- Point estimate of Y^* is $(\mathrm{x}^*)^\top \hat{\boldsymbol{\beta}}$. (The same as the previous one).



Calculating a CI on μ^* and PI on Y^*

- Two sources of uncertainty in future Y^*:
  (1) Don't know true $\beta_0, \ldots, \beta_k$;
  (2) Don't know future ϵ.
- To quantify (1), we use the fact $\mathrm{Var}(\underbrace{(\mathrm{x}^*)^\top \hat{\beta}}_{\hat{\mu}^*}) = \sigma^2 (\mathrm{x}^*)^\top (\mathrm{X}^\top \mathrm{X})^{-1} \mathrm{x}^*$.
- To quantify (2), we use $\mathrm{Var}(\epsilon) = \sigma^2$.
- Hence, we derive
  - two-sided $100(1-\alpha)\%$ PI for Y^*:
  $$\hat{\mu}^* \pm t_{n-(k+1),\alpha/2} \cdot s\sqrt{1 + (\mathrm{x}^*)^\top (\mathrm{X}^\top \mathrm{X})^{-1} \mathrm{x}^*}.$$
  - two-sided $100(1-\alpha)\%$ CI for μ^*:
  $$\hat{\mu}^* \pm t_{n-(k+1),\alpha/2} \cdot s\sqrt{(\mathrm{x}^*)^\top (\mathrm{X}^\top \mathrm{X})^{-1} \mathrm{x}^*}.$$

  Here, s^2 is sample variance of Y.

Example: PI on Property Value

- We consider two predictors taxes and baths.
- Let's predict a new home price with $x_1^* = 7$ and $x_2^* = 1.5$, i.e., taxes $= 7$ and baths $= 1.5$.
- After fitting the model, we find a point estimate

$$\hat{\mu}^* = 10.042 + 7 * 2.713 + 1.5 * 6.164 = 38.279.$$

- To find a 95% PI, we first find the data matrix

$$\mathsf{X} = \begin{bmatrix} 1 & 5.02 & 1 \\ 1 & 4.54 & 1 \\ \vdots & \vdots & \vdots \end{bmatrix}$$

Example: PI on Property Value Cont'd

- We calculate $\mathrm{X}^\top \mathrm{X}$ as

$$\mathrm{X}^\top \mathrm{X} = \begin{bmatrix} 1.12 & -0.04 & -0.69 \\ -0.04 & 0.03 & -0.13 \\ -0.69 & -0.13 & 1.30 \end{bmatrix}.$$

- Next, we calculate

$$(\mathrm{x}^*)^\top (\mathrm{X}^\top \mathrm{X})^{-1} \mathrm{x}^* = 0.146.$$

- Meanwhile, we find $s = 2.79$ and $t_{n-(k+1),\alpha/2} = 2.08$.
- Lastly, the PI is

$$\hat{\mu}^* \pm t_{n-(k+1),\alpha/2} \cdot s \sqrt{1 + (\mathrm{x}^*)^\top (\mathrm{X}^\top \mathrm{X})^{-1} \mathrm{x}^*} = [32.06, 44.50].$$

Example: R Computed PI

```
> predict(lm1, newdata=Xnew, se.fit = T, level=0.95, interval = "prediction")
$fit
       fit      lwr      upr
1 38.28195 32.06641 44.49749

$se.fit
[1] 1.066298

$df
[1] 21

$residual.scale
[1] 2.792114
```

Categorical Predictors

Handling Categorical Predictor Variables

- Represent the binary predictor by defining a single 0/1 indicator (dummy) variable.
- Use the resulting dummy variable in your regression model.

- We denote the response variable as Y, e.g., weight of a person. There are two predictor variables:
  - x_1 as the height, and
  - x_2 as the gender.
- We typically redefine $x_2 = 0/1$ indicator variable, where 1 represents male and 0 represents female.
- Then response is written as

$$Y = (\underbrace{\beta_0 + \beta_2 x_2}_{\text{Intercept for male}}) + \beta_1 x_1 + \epsilon.$$

- Note that we have $c = 2$ categories for x_2, which we have represented with $c - 1 = 1$ dummy variables.

Frame Title

- For a nominal predictor with c categories, create $c - 1$ 0/1 indicator (dummy) variables.

- We denote the response variable as Y, e.g., weight of a person. There are two predictor variables:
  - x_1 as the height, and x_2 as the country.
- Suppose there are 4 countries. We arbitrarily choose a base category (e.g., Canada) and create $c - 1 = 3$ binary indicator predictors:

$$x_2 = \begin{cases} 1 & \text{US} \\ 0 & \text{o.w.} \end{cases}, \quad x_3 = \begin{cases} 1 & \text{Mexico} \\ 0 & \text{o.w.} \end{cases}, \quad x_4 = \begin{cases} 1 & \text{China} \\ 0 & \text{o.w.} \end{cases}$$

y	x_1	x_2	x_3	x_4	
y_1	x_{11}	1	0	0	if from US
y_2	x_{21}	0	0	0	if from Canada
y_3	x_{31}	0	1	0	if from Mexico
y_4	x_{41}	0	0	1	if from China

Interpretation of Model

- From the previous slide, we have a model
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon.$$
- We examine the model for each country:
$$Y = \begin{cases} \beta_0 + \beta_1 x_1 & \text{Canada} \\ \beta_0 + \beta_2 + \beta_1 x_1 & \text{US} \\ \beta_0 + \beta_3 + \beta_1 x_1 & \text{Mexico} \\ \beta_0 + \beta_4 + \beta_1 x_1 & \text{China} \end{cases}.$$
- Net effect: A different intercept for each category.
- How would you depict the model graphically?

Why Not Use c Dummy Variables

- Suppose we had defined

$$x_5 = \begin{cases} 1 & \text{Canada} \\ 0 & \text{o.w.} \end{cases}.$$

- Then the model becomes

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \epsilon.$$

- For each country, we have

$$Y = \begin{cases} \beta_0 + \beta_5 + \beta_1 x_1 & \text{Canada} \\ \beta_0 + \beta_2 + \beta_1 x_1 & \text{US} \\ \beta_0 + \beta_3 + \beta_1 x_1 & \text{Mexico} \\ \beta_0 + \beta_4 + \beta_1 x_1 & \text{China} \end{cases}.$$

Why is this problematic?

Handling Categorical Predictors in R

- In R, any predictor of class "factor" is automatically treated as a categorical predictor, even if the factor levels are labeled as numbers. R internally converts the factor into a set of 0/1 dummy variables (i.e., you just enter the predictor as a single column of class factor).
- You may still want to manually convert the categorical predictor to $c - 1$ 0/1 dummy variables in the following R situations:
  - stepwise regression: R's step() command will add/drop entire categorical predictors. If you manually convert to 0/1 dummy variables, you can add/drop individual levels of a categorical predictor;
  - best subsets regression: R's leaps() command cannot handle categorical predictors. You must manually convert to 0/1 dummy variables.

Example: Converting Age into Categories

- We predict weight using age and gender in `pred_weight.txt`. Initially, age variable takes integer values. We convert it into a categorical variable accordingly to three ranges, namely, $(18, 20]$, $(20, 21]$ and $(21, 30]$.
- The fitted coefficients are as follows

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   44.068     65.063   0.677 0.504428
height         1.127      1.006   1.119 0.273650
gender        47.641     11.019   4.323 0.000215 ***
age(20,21]     7.354      7.999   0.919 0.366670
age(21,30]    19.255     10.551   1.825 0.079991 .
---
```

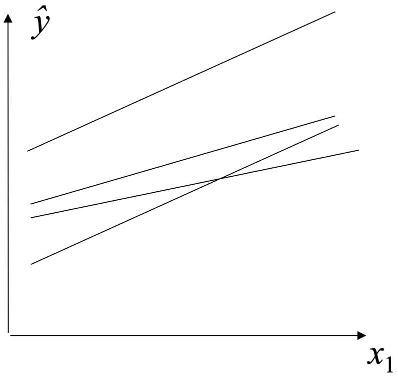
Questions and Discussions

- In the previous fitted linear model, what is the form of the regression model that was fit, and what age category did R use as the base category?
- How do you interpret the two age coefficients that were produced? Do they seem to make sense?

Interaction

What if Slopes Differ in Different Categories?

- If we suspected the Y v.s. x_1 relationship for the four categories may look like the following, what terms could we add to the model to represent this?



Interactions Between Categorical and Quantitative Predictors

- Adding interactions between x_1 and the dummy variables, the model for our earlier weight example becomes

$$\begin{aligned} Y = {} & \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \\ & + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{14} x_1 x_4 + \epsilon. \end{aligned}$$

- We evaluate according to different countries

$$Y = \begin{cases} \beta_0 + \beta_1 x_1 & \text{Canada} \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_{12}) x_1 & \text{US} \\ (\beta_0 + \beta_3) + (\beta_1 + \beta_{13}) x_1 & \text{Mexico} \\ (\beta_0 + \beta_4) + (\beta_1 + \beta_{14}) x_1 & \text{China} \end{cases}.$$

- This allows for different slopes and/or intercepts for each predictor category.

Example: Handling Interactions in R

- Using `pred_weight.txt` data again, we add an interaction term:

$$\text{weight} \sim \text{height} + \text{gender} + \text{height} \times \text{gender}.$$

- The fitted coefficients are as follows

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -102.407    141.569  -0.723    0.476
height           3.466      2.191   1.582    0.126
gender         198.204    160.401   1.236    0.228
height:gender   -2.309      2.437  -0.947    0.352
```

- How can we explain the result of individual coefficient t-test? How we interpret the coefficient of the interaction term?

General Comments and Discussions

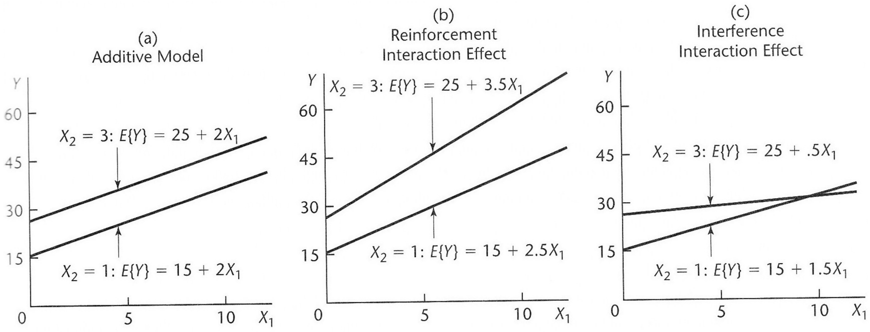
- In a model relating y to the **quantitative** predictors, using indicator variables for the categorical predictors can account for different intercepts (and different slopes if interaction terms are included) in each category.
- Using indicator variables is more efficient than fitting separate models in each category if we suspect that some of the parameters are common across categories.
  - We can pick and choose which parameters are common and use all the data to estimate the common values.
  - This is especially important if we have multiple categorical variables with many categories each.
- Multiple categorical predictors, each with many categories, are extremely common in "analytics" problems. Regression and classification trees (coming up later) are very good at handling this.

Interactions between Quantitative Predictors

- An **interaction between two quantitative predictors** is interpreted analogously to an interaction between a qualitative and quantitative predictor: The slope of y w.r.t. one predictor depends on the level of the other predictor.
- Example: Study of the effect of point-of-sales and TV add expenditures on locality sales. The variables are
  - y: locality sales;
  - x_1: point-of-sales add expenditure;
  - x_2: TV add expenditure.
- The model is

$$\begin{aligned} Y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon \\ &= \underbrace{(\beta_0 + \beta_2 x_2)}_{\text{Intercept for fixed } x_2} + \underbrace{(\beta_1 + \beta_{12} x_2)}_{\text{Slope for fixed } x_2} x_1. \end{aligned}$$

Reinforcement and Interference Interactions



Leverage and influence

Sensitive to Outlier

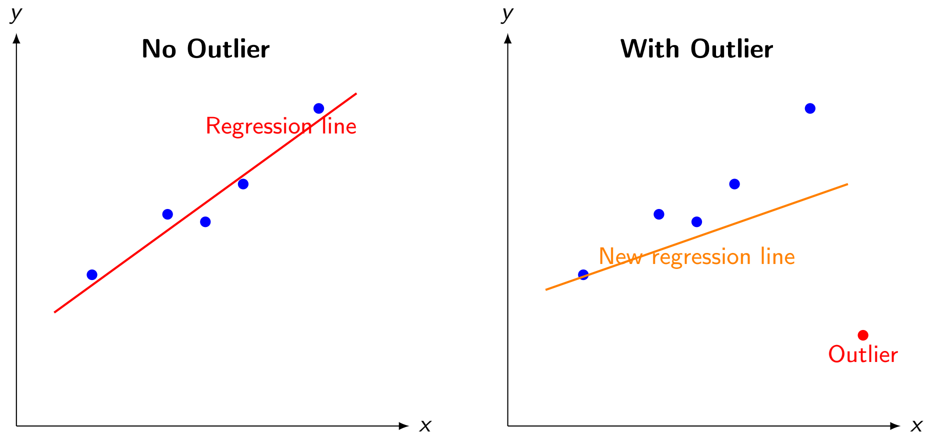
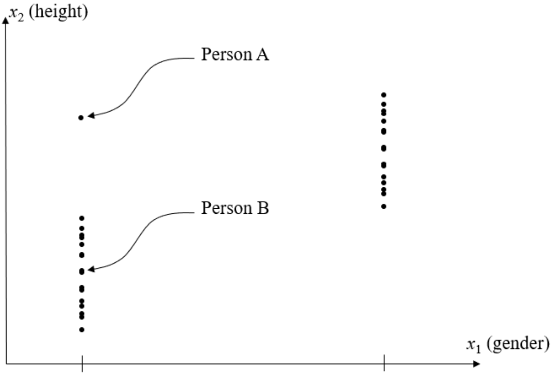


Illustration of How One Observation Can Be Influential

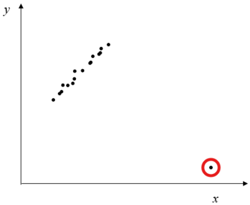
- Suppose the objective is to predict the weight (y) of a person, based on their gender (x_1) and height (x_2). Imagine a third axis coming out of the page to represent y.

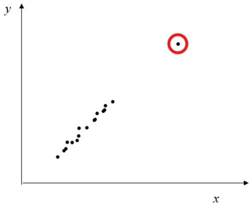


Leverage and Influence

- Suppose we have n multivariate observations $\{(y_i, \mathrm{x}_i) : i = 1, 2, \ldots, n\}$, and denote the predictors for the i-th observation by $\mathrm{x}_i = [1, x_{i1}, x_{i2}, \ldots, x_{ik}]^\top$;
- Any "unusual" x_i is called a **high-leverage observation**;
- Any $\{y_i, \mathrm{x}_i\}$ that significantly changes the estimated coefficients is called an **influential observation**, i.e.,
  - define $\hat{\beta}_{(i)}$ = estimated coefficients if delete the i-th observation;
  - $\{y_i, \mathrm{x}_i\}$ is high influential if $\hat{\beta}_{(i)} - \hat{\beta}$ is "large".

Which is high leverage? Which is high influence?





Measuring Leverage

- Define $\mathrm{H} = \mathrm{X}(\mathrm{X}^\top \mathrm{X})^{-1}\mathrm{X}^\top$, where, as usual $\mathrm{X} = [\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_n]^\top$.
- Denote the i-th diagonal element of H as

$$h_{ii} = [\mathrm{H}]_{ii} = \mathrm{x}_i^\top (\mathrm{X}^\top \mathrm{X})^{-1} \mathrm{x}_i,$$

  which is the **measure of leverage** for x_i.
- Average leverage should be $\frac{1}{n}\sum_{i=1}^n h_{ii} = \frac{k+1}{n}$.
- Common rule-of-thumb: x_i flagged when $h_{ii} > \frac{2(k+1)}{n}$.

Understanding Leverage

The fitted (predicted) values are:$\hat{\mathrm{y}} = \mathrm{X}\hat{\beta} = \underbrace{\mathrm{X}(\mathrm{X}^\top\mathrm{X})^{-1}\mathrm{X}^\top}_{H}\mathrm{y}$. Thus, the fitted value $\hat{y}_i$ (the i-th element of $\hat{\mathrm{y}}$) is given by:

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j,$$

Measuring Influence

- Recall that $\hat{\boldsymbol{\beta}}_{(i)}$ is the estimated coefficients if we delete i-th row $\{y_i, \mathrm{x}_i\}$ of data.
- A common measure of influence is
$$D_i = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})^\top (\mathrm{X}^\top \mathrm{X})^{-1} (\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{(k+1)s^2} = \textbf{Cook's distance}$$
- If D_i is large, the i-th observation changes $\hat{\boldsymbol{\beta}}$ significantly.
- Rules of thumb for flagging an observation as influential:
  (1) $D_i > 1$, which is standard, but perhaps too conservative;
  (2) $D_i > 4/n$, which translates some well-known criterion, but perhaps too liberal.

Relation Between Leverage and Influence

- Fact: $\mathrm{Var}(e_i) = \sigma^2(1 - h_{ii})$.
- We define the "standardized residuals" as
$$e_i^* = \frac{e_i}{\mathrm{SE}(e_i)} = \frac{e_i}{s\sqrt{1 - h_{ii}}}.$$
- A surprising and useful result $D_i = \frac{1}{k+1}\frac{h_{ii}}{1-h_{ii}}(e_i^*)^2$.
- This tells us that the influence of the i-th observation depends on two things:
  (1) leverage of the i-th observation, and
  (2) residual of the i-th observation.

Example: Use Handy R Built-in Functions

- We can use methods function to show various post-fit regression analyses.

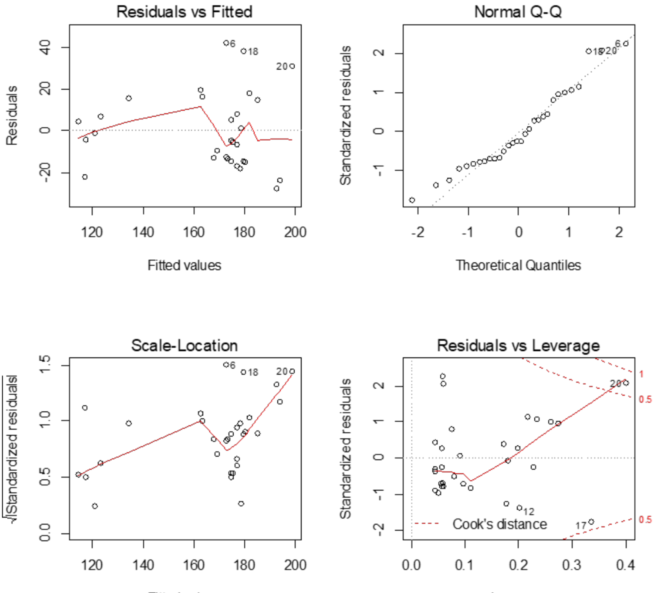
```
lm1<-lm(weight~.,data=WT[,c(1,2,3,6)])
summary(lm1)
methods(class="lm")  #shows various post-fit regression analyses
e<-residuals(lm1)  #regular residuals
estar<-rstandard(lm1)  #standardized residuals
Inf1<-influence(lm1)  #calculates a number of influence-related quantities
round(Inf1$coefficients, 2)  #change in estimated coefficients after deleti
cook1<-cooks.distance(lm1)  #Cook's distance
round(data.frame(estar,Inf1$hat,cook1,WT[,c(1,2,3,6)]), 3)
```

Example: High Influence and Leverage Data

- Can we identify and explain some high influence (or leverage) data points?

```
> round(data.frame(estar,Inf1$hat,cook1,WT[,c(1,2,3,6)]), 3)
    estar Inf1.hat cook1 weight height gender age
1   1.143    0.217 0.090    182     62      1  20
2   0.389    0.172 0.008    130     66      0  20
3  -0.251    0.226 0.005    113     60      0  21
4  -0.967    0.049 0.012    160     72      1  21
5  -0.777    0.059 0.010    165     73      1  21
6   2.260    0.058 0.078    215     68      1  21
7  -0.683    0.058 0.007    160     68      1  21
8   0.435    0.044 0.002    185     71      1  21
9  -0.893    0.044 0.009    160     71      1  21
10  0.070    0.090 0.000    180     74      1  20
11 -0.829    0.111 0.021    165     75      1  20
12 -1.378    0.200 0.119    170     74      1  27
13  0.285    0.056 0.001    180     71      1  20
14 -0.502    0.079 0.005    160     67      1  20
15  0.272    0.197 0.005    119     61      0  19
16 -0.784    0.056 0.009    160     71      1  20
17 -1.759    0.334 0.388    165     70      1  29
18  2.061    0.059 0.067    218     73      1  21
19  1.070    0.233 0.087    200     78      1  19
20  2.090    0.400 0.728    230     73      1  30
```

Example: Residual and Influence/leverage Plots



Residual Diagnostics

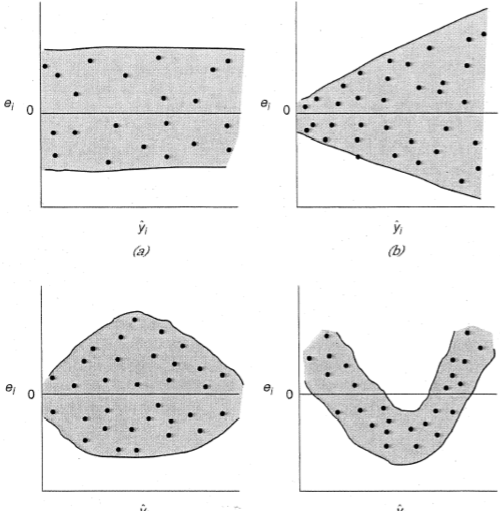
What is Residual Diagnostics?

Regression diagnostics refers to checking for pitfalls, problems, and violations of the underlying assumptions that are corrupting the model and/or that should be accounted for to improve the model. These include (but not limited):

- ❒ unusual observations that are influencing the fit
- ❒ nonlinearities that should be accounted for additional important predictors that should be included in the model
- ❒ strong departures from normality and the constant variance assumption (mild departures are OK)

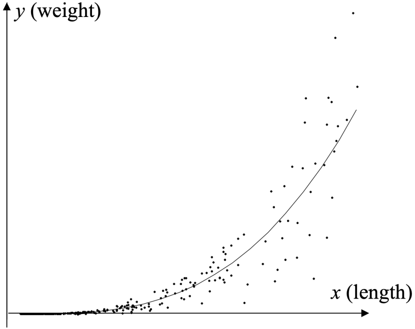
Plot of Residuals Versus Fitted Values

Perhaps the most useful residual plot. Good for checking for nonlinearity and non-constant variance.



Simple Example of Nonconstant Variance

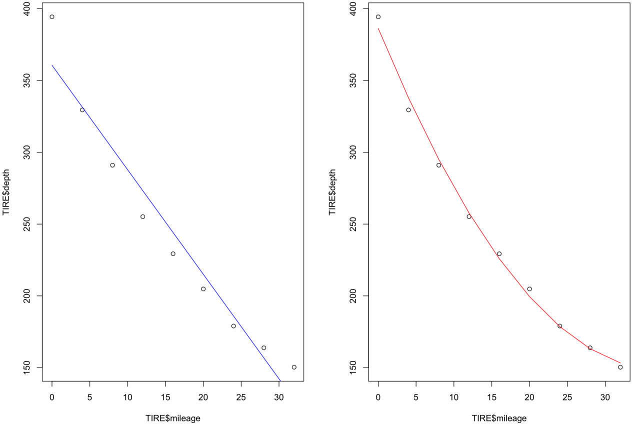
- We denote x =length of animal and y =weight of animal.
- Typical data for a sample of 500 animals look like



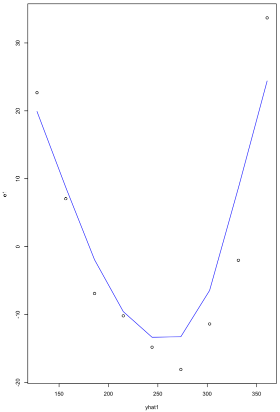
- We fit a cubic polynomial to these data, what would a plot of the residuals versus fitted values look like?

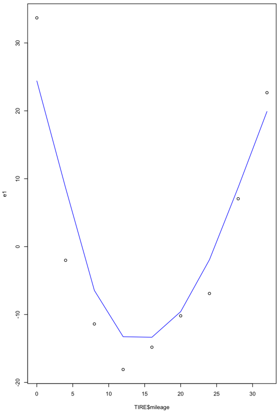
Example: Tire Wear Fitted Plots

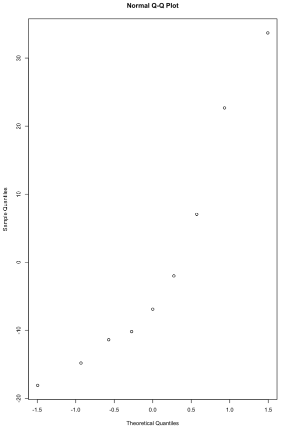
- We use mileage to predict tire wear.
- We fit two models: 1) simple linear model and 2) quadratic model.



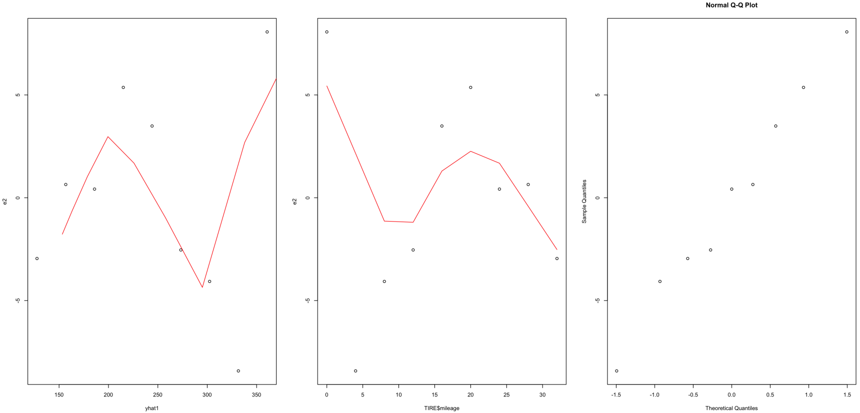
Example: Residual Plots — Linear Model







Example: Residual Plots — Quadratic Model



Discussion

- Nonlinearities are usually much more visible in the residuals than in the raw data.
- What is the relationship between the plots of e_i versus x_i and of e_i versus $\hat{y}_i$ for the linear model? What is the relationship between the plots of e_i versus x_i and of e_i versus $\hat{y}_i$ for the quadratic model or, more generally, if there is more than one predictor variable?
- Has the quadratic model captured the nonlinearity, or is some other nonlinear model perhaps necessary? Where is the quadratic fit the poorest?
- Are the errors normally distributed?