Regression Modeling with Actuarial and Financial Applications

Chapters 2 to 6: Linear Models

Actuarial Short Course - Predictive Modeling 2024

Outline

1. Description of Data
2. Review of Linear Regression
3. Data Example

Medical Expenditure Panel Survey (MEPS)

- Goal: Predict y - measure of the health of an individual. We will focus on body mass index (BMI), an objective measure.
- We have knowledge of several characteristics of a person, include their age, sex, race, marital status. These will serve as predictor variables $x_1, \ldots, x_k$

Medical Expenditure Panel Survey (MEPS)

- Goal: Predict y - measure of the health of an individual. We will focus on body mass index (BMI), an objective measure.
- We have knowledge of several characteristics of a person, include their age, sex, race, marital status. These will serve as predictor variables $x_1, \ldots, x_k$
- Work with the 2009 data so there are $n = 750$ observations
- With this data, we seek to calibrate a model that can be used to understand an individual's health in terms of the predictor variables
- To see how this model fares in prediction, we utilize a new sample of 2010 data, also with 750 observations.
- We will then use the characteristics of the new sample to make predictions of an individual's health, and be able to assess the predictive ability of the model as we have 2010 values of y.

Linear Regression Sampling Assumptions I

Observables Representation Sampling Assumptions
F1. $\mathrm{E}\ y_i = \beta_0 + \beta_1 \cdot x_{i1} + \cdots + \beta_k \cdot x_{ik}$
F2. $\{x_1, \ldots, x_n\}$ are non-stochastic variables
F3. $\mathrm{Var}\ y_i = \sigma^2$
F4. $\{y_i\}$ are independent random variables

- The model parameters are $\beta_0, \ldots, \beta_k, \sigma^2$
- For F3, a common variance is known as *homoscedasticity*

Linear Regression Sampling Assumptions I

Observables Representation Sampling Assumptions
F1. $\mathrm{E}\ y_i = \beta_0 + \beta_1 \cdot x_{i1} + \cdots + \beta_k \cdot x_{ik}$
F2. $\{x_1, \ldots, x_n\}$ are non-stochastic variables
F3. $\mathrm{Var}\ y_i = \sigma^2$
F4. $\{y_i\}$ are independent random variables

- The model parameters are $\beta_0, \ldots, \beta_k, \sigma^2$
- For F3, a common variance is known as *homoscedasticity*
- We sometimes require

  F5. $\{y_i\}$ are normally distributed.

  Approximate normality is enough for central limit theorems needed for inference

Linear Regression Sampling Assumptions II

Error Representation Sampling Assumptions
E1. $y_i = \beta_0 + \beta_1 \cdot x_{i1} + \cdots + \beta_k \cdot x_{ik} + \varepsilon_i$
E2. $\{x_1, \ldots, x_n\}$ are non-stochastic variables
E3. $\mathrm{E}\ \varepsilon_i = 0$ and $\mathrm{Var}\ \varepsilon_i = \sigma^2$
E4. $\{\varepsilon_i\}$ are independent random variables

- These two sets of assumptions are equivalent
- The *error* representation is a useful springboard for residual analysis
- The *observable* representation is a useful springboard for extensions to nonlinear regression models

Overview of Linear Models Ingredients

Regression Function:

$$\mathrm{E}[y] = \beta_0 \cdot x_0 + \beta_1 \cdot x_1 + \cdots + \beta_k \cdot x_k$$

Overview of Linear Models Ingredients

Regression Function:

$$\mathrm{E}[y] = \beta_0 \cdot x_0 + \beta_1 \cdot x_1 + \cdots + \beta_k \cdot x_k \approx \widehat{y}$$

Overview of Linear Models Ingredients

Regression Function:

$$\mathrm{E}[y] = \beta_0 \cdot x_0 + \beta_1 \cdot x_1 + \cdots + \beta_k \cdot x_k \approx \widehat{y}$$

- For continuous x_j:
  Interpret β_j as expected change in y per unit change in x_j, holding other explanatory variables fixed

$$\beta_j = \frac{\partial \mathrm{E}[y]}{\partial x_j}$$

Overview of Linear Models Ingredients

Regression Function:

$$\mathrm{E}[y] = \beta_0 \cdot x_0 + \beta_1 \cdot x_1 + \cdots + \beta_k \cdot x_k \approx \widehat{y}$$

- For continuous x_j:
  Interpret β_j as expected change in y per unit change in x_j, holding other explanatory variables fixed

$$\beta_j = \frac{\partial \mathrm{E}[y]}{\partial x_j}$$

- For categorical x_j:
  Interpret β_j as expected change in y for observation in category x_j relative to the reference category, holding other explanatory variables fixed

Partitioning the Variability

Two estimators of y_i: $\overline{y}$ and $\widehat{y}_i$

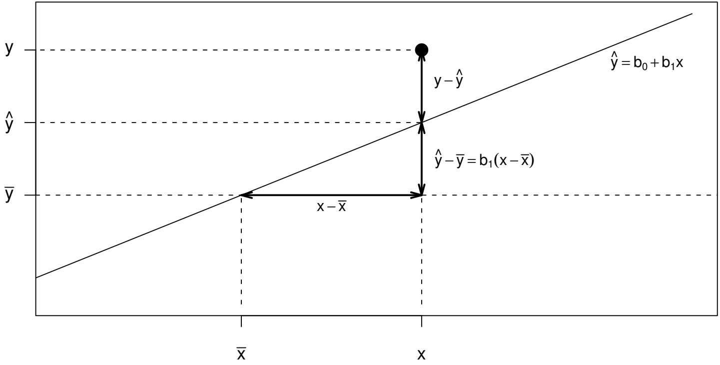
$$\underbrace{y_i - \overline{y}}_{\substack{\text{total}\\ \text{deviation}}} = \underbrace{y_i - \widehat{y}_i}_{\substack{\text{unexplained}\\ \text{deviation}}} + \underbrace{\widehat{y}_i - \overline{y}}_{\substack{\text{explained}\\ \text{deviation}}}$$

Partitioning the Variability

Two estimators of y_i: $\overline{y}$ and $\widehat{y}_i$

$$\underbrace{y_i - \overline{y}}_{\substack{\text{total}\\ \text{deviation}}} = \underbrace{y_i - \widehat{y}_i}_{\substack{\text{unexplained}\\ \text{deviation}}} + \underbrace{\widehat{y}_i - \overline{y}}_{\substack{\text{explained}\\ \text{deviation}}}$$

In one-variable regression:



Partitioning the Variability

After a little algebraic manipulation:

$$\underbrace{\sum_{i=1}^{n}(y_i-\overline{y})^2}_{Total\ \ SS}=\underbrace{\sum_{i=1}^{n}\left(y_i-\widehat{y}_i\right)^2}_{Error\ \ SS}+\underbrace{\sum_{i=1}^{n}\left(\widehat{y}_i-\overline{y}\right)^2}_{Regression\ \ SS}$$

Partitioning the Variability

After a little algebraic manipulation:

$$\underbrace{\sum_{i=1}^{n}(y_i-\overline{y})^2}_{Total\ \ SS}=\underbrace{\sum_{i=1}^{n}\left(y_i-\widehat{y}_i\right)^2}_{Error\ \ SS}+\underbrace{\sum_{i=1}^{n}\left(\widehat{y}_i-\overline{y}\right)^2}_{Regression\ \ SS}$$

Define R-square (*coefficient of determination*):

$$R^2=\frac{Regression\ SS}{Total\ SS}$$

Partitioning the Variability

After a little algebraic manipulation:

$$\underbrace{\sum_{i=1}^{n}(y_i-\overline{y})^2}_{Total\ \ SS}=\underbrace{\sum_{i=1}^{n}\left(y_i-\widehat{y}_i\right)^2}_{Error\ \ SS}+\underbrace{\sum_{i=1}^{n}\left(\widehat{y}_i-\overline{y}\right)^2}_{Regression\ \ SS}$$

Define R-square (*coefficient of determination*):

$$R^2=\frac{Regression\ SS}{Total\ SS}$$

R^2 = Proportion of variability explained by regression line

Partitioning the Variability

- If regression line fits data perfectly, $\textit{Error SS} = 0$ and $R^2 = 1$
- If regression line provides no information about response variable, $\textit{Regression SS} = 0$ and $R^2 = 0$
- Property: $0 \leq R^2 \leq 1$ (larger values implying better fit)

Model Adequacy and Goodness of Fit

- MSE (s^2) compared to s_y^2
- t-statistics of individual coefficients
- R^2 and R_a^2
- AIC, BIC, PRESS
  - Information criteria = measure of fit plus penalty for model complexity, e.g.
  - $AIC = -2 \times$ log-likelihood $+ 2 \times$ number of parameters
  - smaller is better
- Residual analysis

Prediction Validation

- Cross-validation is most common
- Sometimes, a natural ”in-sample” and ”out-sample” is available, such as by years
- In this case, summarize using, e.g., $SPSS = \sum_i (y_i - y_i^*)^2$

Medical Expenditure Panel Survey (MEPS)

- Conducted by Agency for Health Care Research and Quality (AHRQ) and National Center for Health Statistics (NCHS)
- Uses National Health Interview Survey (NHIS) as sampling frame
- 2 year panel
- Computer-assisted personal interviews (CAPI) to collect 2 full years of data
- Overlapping data collection

Data Used in Presentation

- Random sample of 750 observations from Panel 13 (cal yr 2009)
- Results not reflect weighting of observations

On your own:

- Download the Data
- Try out the R-code
- Review the presentation for an overview

Obesity Example

- More than 40% of U.S. adults are obese
- Obesity-related conditions include heart disease, stroke, type 2 diabetes and certain types of cancer
- In 2019, medical costs associated with obesity estimated at $173 billion
- Medical costs paid by third-party payors for people who are obese $1,861 higher than those of healthy weight.

(Source: `https://www.cdc.gov/bmi/adult-calculator/`)

Obesity Guidelines

BMI*	Weight Status
Less than 18.5	Underweight
18.5 to less than 25	Normal
25.0 to less than 30	Overweight
30.0 or greater	Obese

*Calculate BMI: 703 times weight in pounds divided by height in squared inches

(Source: `https://www.cdc.gov/bmi/adult-calculator/bmi-categories.html`)

Variables

Variable	Type	Descriptor
Log.BMI	num	Log (BMI)
Age	int	Age
AgeCat	Factor	3 Levels: Young ($<$25], Adult (25, 65], Senior ($>$65)
Sex	Factor	2 Levels: Male, Female
Race	Factor	3 Levels: Black, White, Other
Uninsured	Factor	2 Levels: Uninsured (1), Insured (0)

Variables

Variable	Type	Descriptor
Log.BMI	num	Log (BMI)
Age	int	Age
AgeCat	Factor	3 Levels: Young ($<$25], Adult (25, 65], Senior ($>$65)
Sex	Factor	2 Levels: Male, Female
Race	Factor	3 Levels: Black, White, Other
Uninsured	Factor	2 Levels: Uninsured (1), Insured (0)
Mental Health	Factor	3 Levels: Excellent, Good, Fair/Poor
Married	int	Indicator variable if person married
Smoker	Indicator	Indicator variable if person smokes
Income	Factor	5 Levels: Poor, Near Poor, Low, Middle, High

Variables

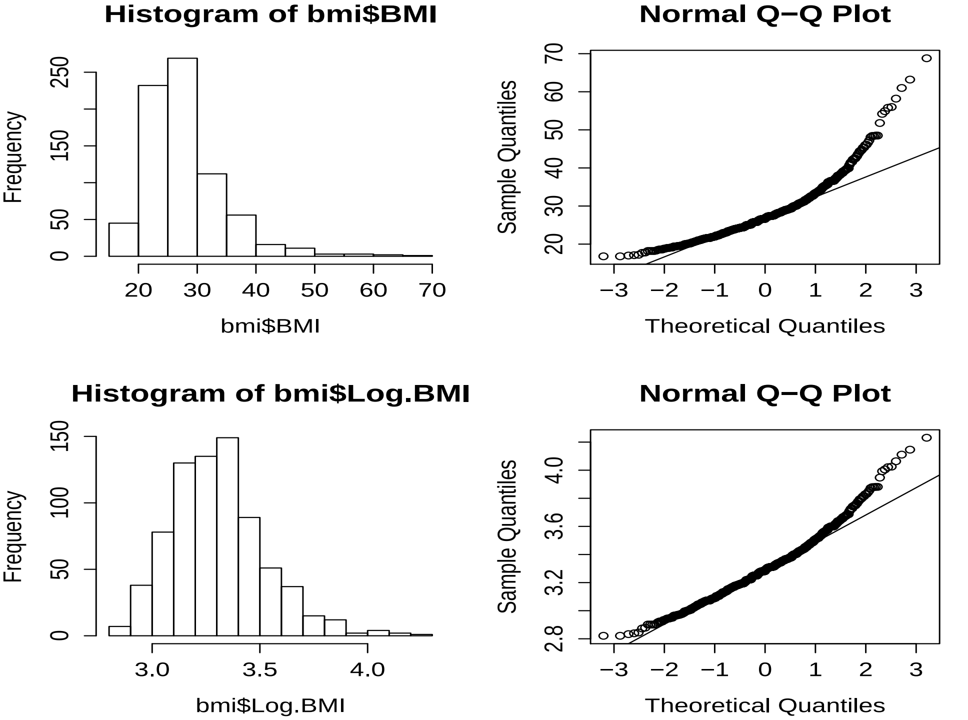
Variable	Type	Descriptor
Log.BMI	num	Log (BMI)
Age	int	Age
AgeCat	Factor	3 Levels: Young ($<$25], Adult (25, 65], Senior ($>$65)
Sex	Factor	2 Levels: Male, Female
Race	Factor	3 Levels: Black, White, Other
Uninsured	Factor	2 Levels: Uninsured (1), Insured (0)
Mental Health	Factor	3 Levels: Excellent, Good, Fair/Poor
Married	int	Indicator variable if person married
Smoker	Indicator	Indicator variable if person smokes
Income	Factor	5 Levels: Poor, Near Poor, Low, Middle, High
ASTHMA	int	Indicator variable if person has asthma
CANCER	int	Indicator variable if person has cancer
CHOLEST	int	Indicator variable if person has high cholesterol
CORONARY	int	Indicator variable if person has coronary heart disease
DIABETES	int	Indicator variable if person has diabetes
EMPHYSEMA	int	Indicator variable if person has emphysema
HIGHBP	int	Indicator variable if person has high blood pressure
STROKE	int	Indicator variable if person had stroke
Comord.Cnt	int	Sum of indicator variables of co-morbidities

Summary Statistics (n = 750)

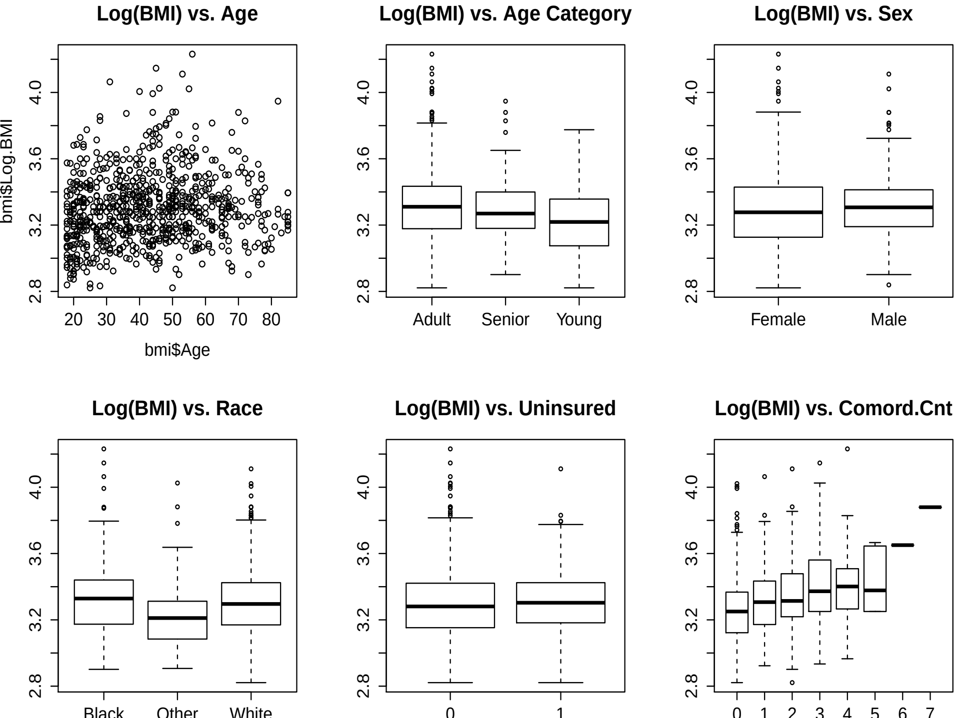
	Mean	Std.Dev.
BMI	27.94	6.73
Log.BMI	3.31	0.22
Age	42.87	17.01
Female	0.54	0.50
Married	0.51	0.50
educyr	12.55	2.90
FAMINC09	$59,595	$55,501
Smoker	0.17	0.38
Uninsured	0.21	0.41
CANCER	0.07	0.26
DIABETES	0.09	0.29
EMPHYSEMA	0.02	0.14
ASTHMA	0.08	0.28
STROKE	0.04	0.19
CORONARY	0.04	0.19
CHOLEST	0.28	0.45
HIGHBP	0.32	0.47
MI	0.03	0.18
Comord.Cnt	0.95	1.20

Categorical Variable Summary Statistics (n = 750)

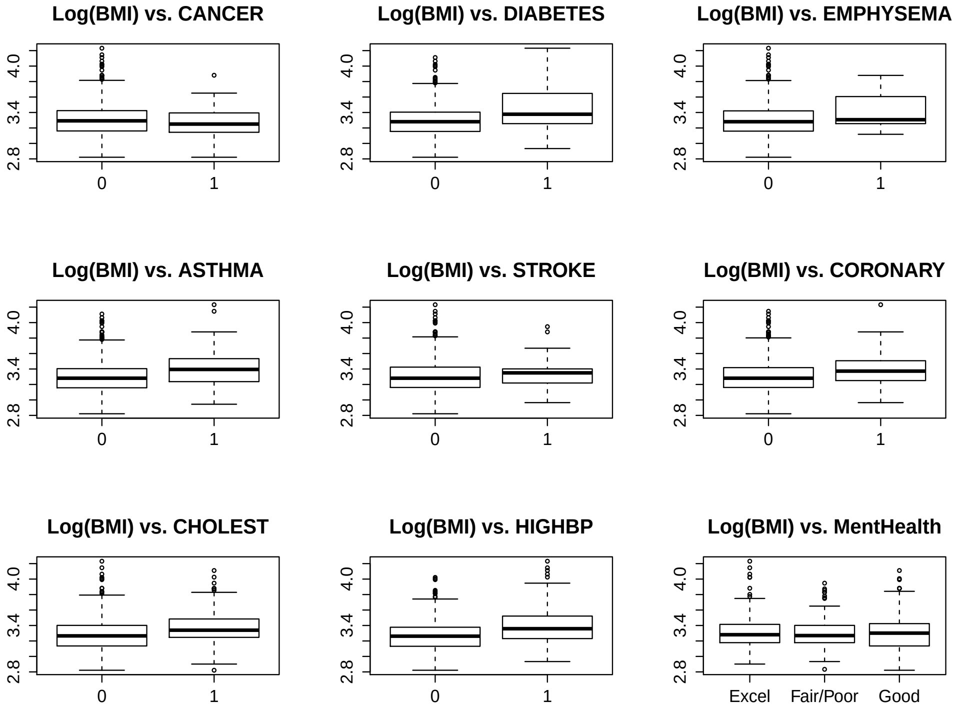
Variable	Category (Percentage)
Age Category	Adult (68%), Senior (12 %), Young (20 %)
Race	Black (20%), Other (12 %), White (68 %)
Mental Health	Excel (38 %), Fair/Poor (7 %), Good (55 %)
Income	High (31 %), Middle (29 %), Low (16 %), Near Poor (6 %), Poor (18 %)



Histogram of bmi$BMI
Frequency
0
50
150
250
20
30
40
50
60
70
bmi$BMI
Normal Q−Q Plot
Sample Quantiles
20
30
40
50
60
70
−3
−2
−1
0
1
2
3
Theoretical Quantiles
Histogram of bmi$Log.BMI
Frequency
0
50
100
150
3.0
3.5
4.0
bmi$Log.BMI
Normal Q−Q Plot
Sample Quantiles
2.8
3.2
3.6
4.0
−3
−2
−1
0
1
2
3
Theoretical Quantiles



Log(BMI) vs. Age
bmi$Log.BMI
bmi$Age
Log(BMI) vs. Age Category
Adult
Senior
Young
Log(BMI) vs. Sex
Female
Male
Log(BMI) vs. Race
Black
Other
White
Log(BMI) vs. Uninsured
Log(BMI) vs. Comord.Cnt


Log(BMI) vs. CANCER
Log(BMI) vs. DIABETES
Log(BMI) vs. EMPHYSEMA
Log(BMI) vs. ASTHMA
Log(BMI) vs. STROKE
Log(BMI) vs. CORONARY
Log(BMI) vs. CHOLEST
Log(BMI) vs. HIGHBP
Log(BMI) vs. MentHealth
2.8
3.4
4.0
0
1
Excel
Fair/Poor
Good

Linear Model Example

Run a linear regression model:
Log.BMI $\sim$ Age + I(Age * Age) + Female + Race.f + Comord.Cnt

Compared to:

Log.BMI $\sim$ AgeCat + Female + Race.f + CANCER + DIABETES + EMPHYSEMA + ASTHMA + STROKE + CORONARY + CHOLEST + HIGHBP

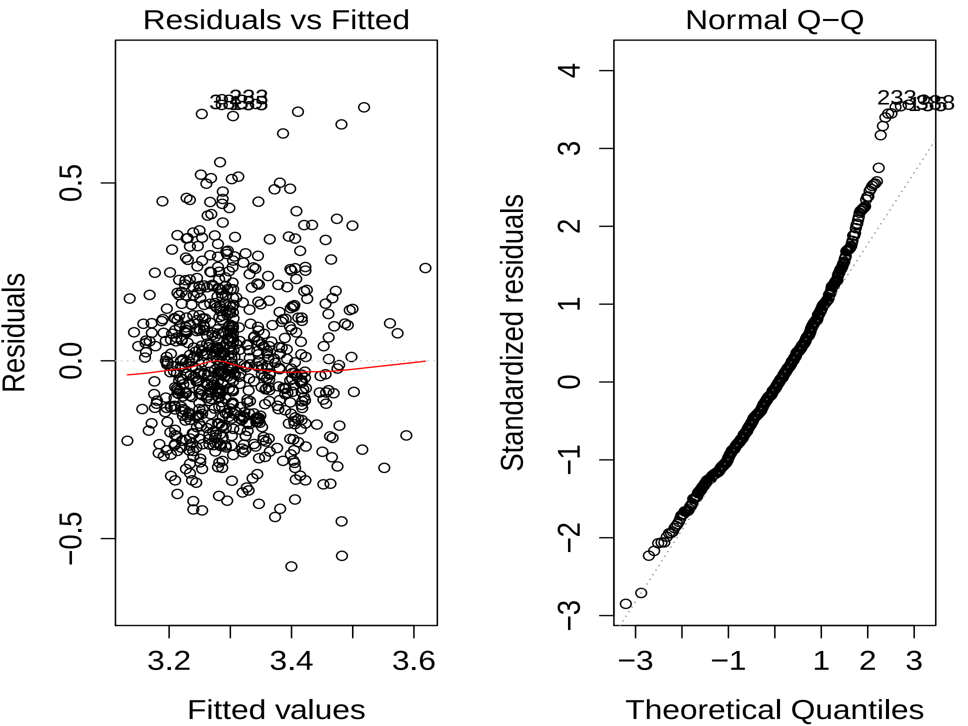
Model A

```
lm(formula = Log.BMI ~ Age + I(Age * Age) + Female + Race.f +
    Comord.Cnt, data = bmi)

               Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.037e+00  4.910e-02  61.861  < 2e-16 ***
Age           1.234e-02  2.237e-03   5.517 4.77e-08 ***
I(Age * Age) -1.422e-04  2.348e-05  -6.055 2.23e-09 ***
Female       -1.765e-02  1.495e-02  -1.180  0.23821
Race.fBlack   1.683e-02  1.885e-02   0.892  0.37243
Race.fOther  -7.798e-02  2.367e-02  -3.295  0.00103 **
Comord.Cnt    5.912e-02  7.845e-03   7.537 1.40e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2036 on 743 degrees of freedom
Multiple R-squared: 0.1325,  Adjusted R-squared: 0.1255
F-statistic: 18.92 on 6 and 743 DF,  p-value: < 2.2e-16

AIC   =  -250.17
PRESS =    31.38
```



Residuals vs Fitted
Residuals
Fitted values
0.5
0.0
−0.5
3.2
3.4
3.6
233
Normal Q−Q
Standardized residuals
Theoretical Quantiles
4
3
2
1
0
−1
−2
−3
−3
−1
1
2
3
233

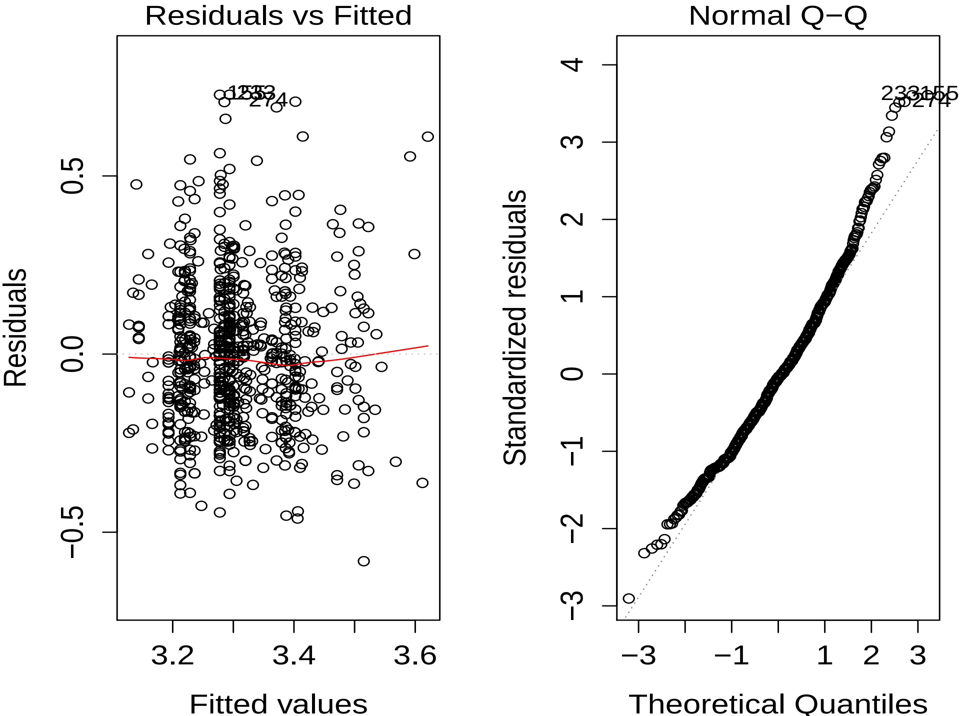
Model B

```
lm(formula = Log.BMI ~ AgeCat + +Female + Race.f + CANCER + DIABETES +
    EMPHYSEMA + ASTHMA + STROKE + CORONARY + CHOLEST + HIGHBP,
    data = bmi)

               Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.293755   0.014860 221.655  < 2e-16 ***
AgeCatSenior -0.081097   0.026495  -3.061 0.002288 **
AgeCatYoung  -0.065155   0.020062  -3.248 0.001217 **
Female       -0.016020   0.015155  -1.057 0.290845
Race.fBlack   0.007559   0.019064   0.397 0.691830
Race.fOther  -0.084599   0.023699  -3.570 0.000381 ***
CANCER       -0.053052   0.030070  -1.764 0.078102 .
DIABETES      0.112799   0.028255   3.992 7.20e-05 ***
EMPHYSEMA     0.046063   0.054626   0.843 0.399359
ASTHMA        0.107432   0.027137   3.959 8.26e-05 ***
STROKE       -0.018196   0.040705  -0.447 0.654994
CORONARY      0.029317   0.041951   0.699 0.484873
CHOLEST       0.022618   0.019439   1.164 0.244994
HIGHBP        0.086012   0.018985   4.531 6.86e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2023 on 736 degrees of freedom
Multiple R-squared: 0.1512,  Adjusted R-squared: 0.1362
F-statistic: 10.09 on 13 and 736 DF,  p-value: < 2.2e-16

AIC   =   -252.53
PRESS =     31.42
```


Residuals vs Fitted
Residuals
Fitted values
0.5
0.0
−0.5
3.2
3.4
3.6
Normal Q−Q
Standardized residuals
Theoretical Quantiles
4
3
2
1
0
−1
−2
−3
−3
−1
1
2
3
274

Model A vs. Model B Goodness of Fit Statistics

Statistic	Model A	Model B
R_a^2	0.1255	0.1362
s	0.2036	0.2023
AIC	-250.17	-252.53
PRESS	31.38	31.42
SSPE	29.08	29.18

Other Possible Variables in Sample Data

Married	int	0 1 1 0 1 0 1 1 0 0 ...
educyr	int	10 17 8 15 12 13 17 16 16 9 ...
FAMINC09	int	25000 320126 0 45000 97432 4800 185936 45000 18000 ...
Pov.f	Factor	w/ 5 levels "High","Low","Middle",..: 3 1 5 1 1 5 1 3 2 3 ...
Smoker	int	0 0 0 0 0 0 0 0 0 0 ...
Uninsured	int	1 0 0 0 1 1 0 0 1 0 ...
MentHealth	Factor	w/ 3 levels "Excel","Fair/Poor",..: 3 3 2 3 3 1 1 3 3 1 ...
MI	int	0 0 0 0 0 0 0 0 0 0 ...
DIABETES.AGEDX	int	NA NA NA 35 NA NA NA NA NA 80 ...
EMPHYSEMA.AGEDX	int	NA NA NA NA NA NA NA NA NA NA ...
ASTHMA.AGEDX	int	NA NA NA NA NA NA NA 25 NA NA ...
CORONARY.AGEDX	int	NA NA NA NA NA NA NA NA NA NA ...
CHOLEST.AGEDX	int	NA NA NA NA NA NA NA NA NA NA ...
HIGHBP.AGEDX	int	NA NA 58 NA NA NA 59 NA NA 80 ...
MI.AGEDX	int	NA NA NA NA NA NA NA NA NA NA ...

Issues To Explore

1. Given the output from Models A and B, what you could you include to improve the model?
2. Does collinearity impact the models as shown?
3. What does an added variable plot show?
4. What additional variables could you add to the model?
5. What other considerations could impact the usefulness of the model?