3.3 Other Considerations in the Regression Model

Prof. Lauren Perry

Qualitative Predictors

So far, we've focused on only quantitative predictors.

Often, datasets have one or more *qualitative* predictors.

We need to consider how to fit these into a numeric model fitting context.

Qualitative Predictors with Two Levels

Consider the variable `Own` from the `Credit` data.

```r
credit <- read.csv("~/Courses/STAT 196M/datasets/Credit.csv")
own <- as.factor(credit$Own)
summary(own)
```

```
##  No Yes
## 193 207
```

To put this into a regression model, we use a *dummy variable*:

$$x_i = I(\text{the } i\text{th person owns a house})$$

Qualitative Predictors with Two Levels

This results in the model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

which takes values

- $\beta_0 + \beta_1 + \epsilon_i$ if the ith person owns a house.

and

- $\beta_0 + \epsilon_i$ if the ith person does not own a house.

So β_1 is the average difference in credit card balance between owners and non-owners.

Qualitative Predictors with Two Levels

```r
summary(lm(Limit ~ Own, data = credit))
```

```
Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept)    4713.17   166.35    28.333   <2e-16 ***
OwnYes         43.35     231.24     0.187    0.851
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 2311 on 398 degrees of freedom
Multiple R-squared:  8.83e-05, Adjusted R-squared:  -0.002424
F-statistic: 0.03515 on 1 and 398 DF,  p-value: 0.8514
```

Qualitative Predictors with More than Two Levels

Consider the variable `region` from the `Credit` data.

```
##  East South  West
##    99   199   102
```

We can represent this by constructing *two* dummy variables.

$$x_{i,1} = I(i\text{th person is from the South})$$

$$x_{i,2} = I(i\text{th person is from the West})$$

Qualitative Predictors with More than Two Levels

Using `region` to predict `credit`,

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \epsilon_i$$

Why only two dummy variables? Consider:

- If the ith person is from the South, $y_i = \beta_0 + \beta_1 x_{i,1} + \epsilon_i$.
- If the ith person is from the West, $y_i = \beta_0 + \beta_2 x_{i,2} + \epsilon_i$
- If the ith person is from the East, $y_i = \beta_0 + \epsilon_i$

So each factor is represented in the model.

- Because `East` has no dummy variable, it is known as the *baseline*.

Qualitative Predictors with More than Two Levels

```
summary(lm(Limit ~ Region, data = credit))
```

```
Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept)      4881.6      232.4  21.009   <2e-16 ***
RegionSouth      -153.1      284.3  -0.539    0.590
RegionWest       -273.8      326.2  -0.839    0.402
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 2312 on 397 degrees of freedom
Multiple R-squared:  0.001781, Adjusted R-squared:  -0.003248
F-statistic: 0.3541 on 2 and 397 DF,  p-value: 0.702
```

Qualitative Predictors

We can also use this approach for a mix of qualitative and quantitative variables in a model.

```
mod2 <- lm(Limit ~ Income + Rating + Own + Region, data=credit)
summary(mod2)
```

```
Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) -539.62205   30.68155 -17.588   <2e-16 ***
Income         0.55281    0.42508   1.300    0.194
Rating        14.77373    0.09685 152.545   <2e-16 ***
OwnYes         2.78064   18.30426   0.152    0.879
RegionSouth    0.71509   22.49522   0.032    0.975
RegionWest    18.21038   25.82151   0.705    0.481
```

Accounting for Interactions

Sometimes, two predictor variables *interact* in their impact on the outcome.

Example:

- Suppose spending money on TV advertising increases the effectiveness of radio advertising.
- We want a way to let β_{radio} vary based on values of TV...

Accounting for Interactions

Consider

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

How does this let β_{radio} vary based on values of $X_2 = \text{TV}$?

$$\begin{aligned} Y &= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon \\ &= \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon \end{aligned}$$

We can interpret β_3 as the increase in effectiveness of TV advertising associated with a one-unit increase in radio advertising (or vice versa).

Consider: Why does estimating the coefficients not require any changes to our least squares approach?

```r
Advertising <- read.csv("~/Courses/STAT 196M/datasets/Advertising.csv")
mod3 <- lm(sales ~ TV + radio + TV*radio, data=Advertising)
summary(mod3)
```

```
Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.750e+00  2.479e-01  27.233   <2e-16 ***
TV            1.910e-02  1.504e-03  12.699   <2e-16 ***
radio         2.886e-02  8.905e-03   3.241   0.0014 **
TV:radio      1.086e-03  5.242e-05  20.727   <2e-16 ***
---
Residual standard error: 0.9435 on 196 degrees of freedom
Multiple R-squared:  0.9678,   Adjusted R-squared:  0.9673
F-statistic:  1963 on 3 and 196 DF,  p-value: < 2.2e-16
```

Consider R^2_{adj} for the main effects model:

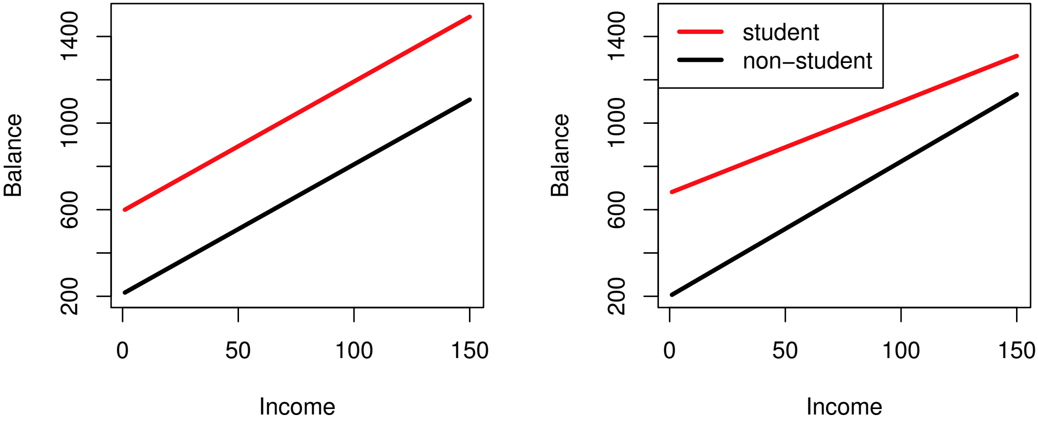
```r
Advertising <- read.csv("~/Courses/STAT 196M/datasets/Advertising.csv")
mod4 <- lm(sales ~ TV + radio, data=Advertising)
summary(mod4)
```

```
Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.92110    0.29449   9.919   <2e-16 ***
TV             0.04575    0.00139  32.909   <2e-16 ***
radio          0.18799    0.00804  23.382   <2e-16 ***
---
Residual standard error: 1.681 on 197 degrees of freedom
Multiple R-squared:  0.8972,   Adjusted R-squared:  0.8962
F-statistic: 859.6 on 2 and 197 DF,  p-value: < 2.2e-16
```

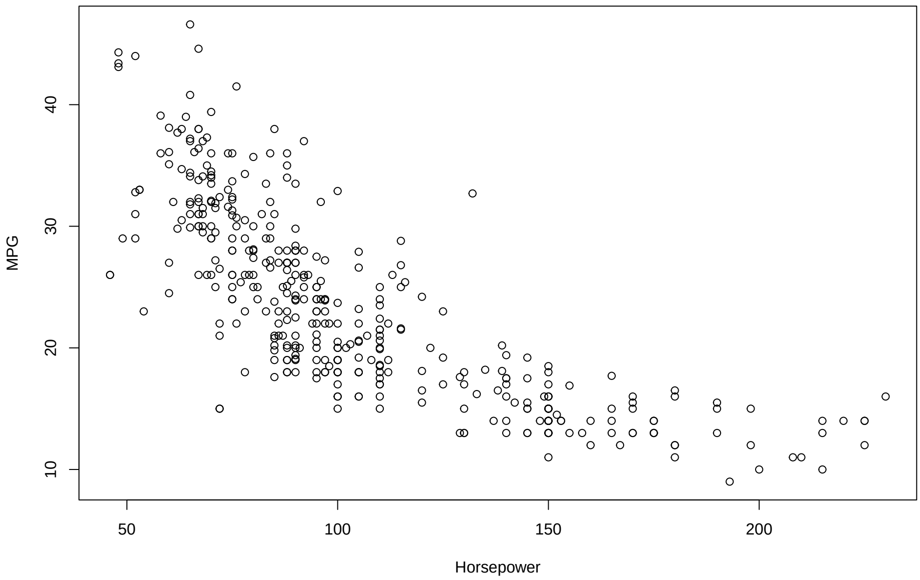
Hierarchical Principal

In general, if we include an interaction term in a model, we also include the main effects *even if the p-values associated with the main effects are not significant*.

Consider Credit `Balance` predicted by `Income` and `Student` status.



- The interaction allows the model for students to have a different slope than the model for non-students, while the main effects model only allows for different intercepts.



MPG
10
20
30
40
50
100
150
200
Horsepower

Nonlinear Relationships Between Predictors and Outcome

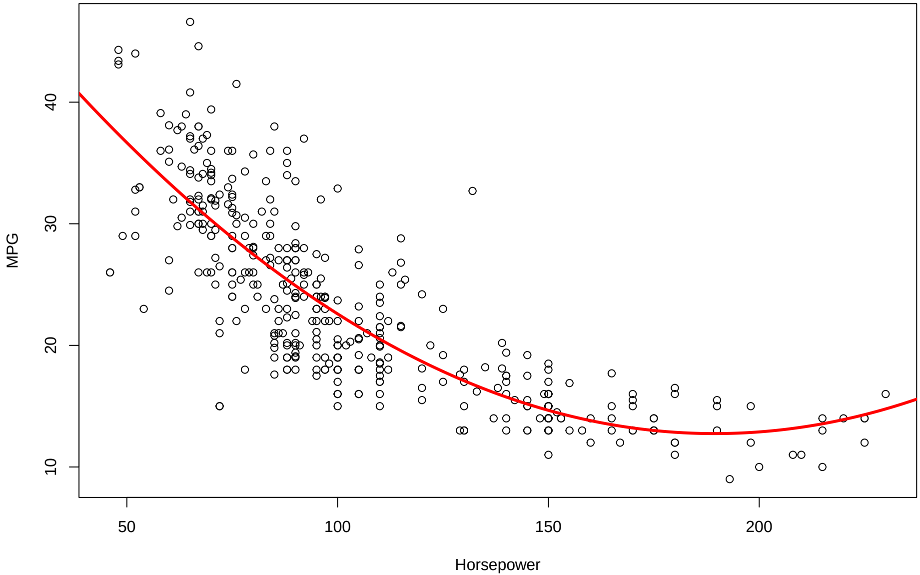
How can we deal with this using *linear* regression?

- The model fit requires the model to be linear *with respect to* β.
- This is much like including $X_1 X_2$ in the model by creating a "new variable" in the matrix X.
- Here, we just construct a "new variable", say, X_1^2 in X.

Nonlinear Relationships Between Predictors and Outcome

```r
mod5 <- lm(mpg ~ poly(horsepower, 2), data = Auto)
summary(mod5)
```

```
Coefficients:           Estimate Std. Error t value Pr(>|t|)
(Intercept)              23.4459     0.2209  106.13   <2e-16 ***
poly(horsepower, 2)1 -120.1377     4.3739  -27.47   <2e-16 ***
poly(horsepower, 2)2   44.0895     4.3739   10.08   <2e-16 ***
Residual standard error: 4.374 on 389 degrees of freedom
Multiple R-squared:  0.6876,   Adjusted R-squared:  0.686
F-statistic:   428 on 2 and 389 DF,  p-value: < 2.2e-16
```


MPG
10
20
30
40
Horsepower
50
100
150
200

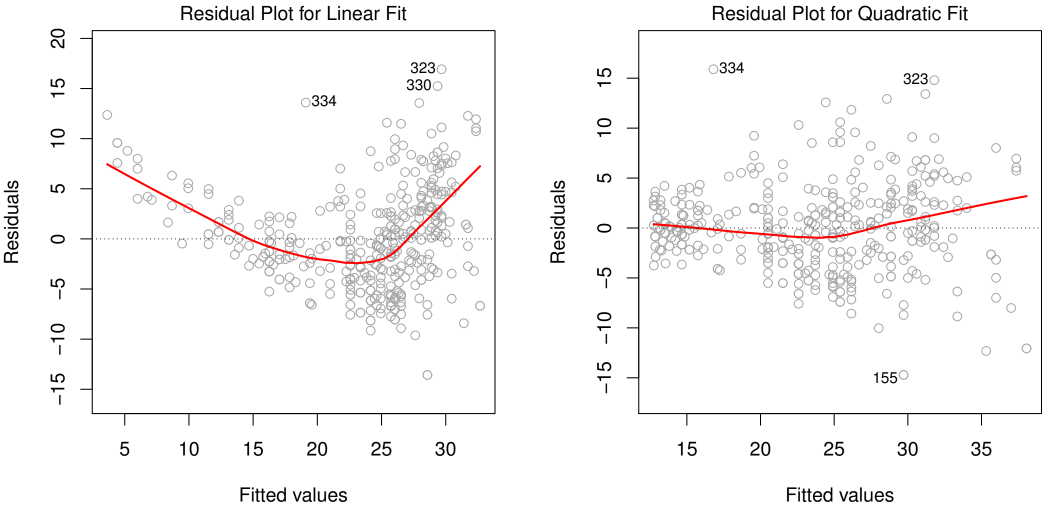
Potential Problems

1. Non-linearity of the response-predictor relationships.
2. Correlation of error terms.
3. Non-constant variance of error terms.
4. Outliers and high-leverage points.
5. Collinearity.

1. Non-linearity of the response-predictor relationships.

- We can examine non-linearity using *residual plots*.
- Ideally, these will show no discernible pattern (random scatter).
- We can work on fixing this problem by transforming the predictors:
  - Ex: $\log X$, $\sqrt(X)$, X^2, etc.

Example Residual Plots Showing Non-Linearity



2. Correlation of Error Terms

Assumption: error terms $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are uncorrelated.

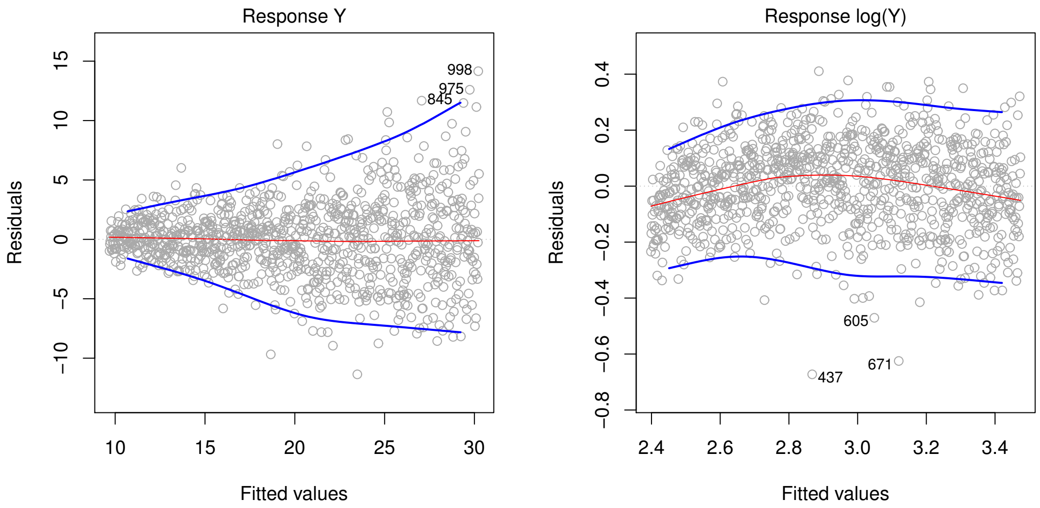
- That is, knowing something about ϵ_i, doesn't tell us anything about ϵ_{i+1}.
- Our standard error calculations rely on this.
  - Violations tend to result in std error being underestimated.
  - This causes erroneously narrow confidence/prediction intervals.
- These correlations can occur for data that is *time dependent*.
  - We should use different modeling techniques for this type of data.

3. Non-constant variance of error terms.

Assumption: error terms have constant variance, $\text{Var}(\epsilon_i) = \sigma^2$.

- We can check for homoscedasticity using residual plots.
- There should be no discernible pattern in the variability.
- Standard errors rely on this assumption.
- This assumption is often violated, but we can usually fix (or at least improve) it!
- We work on fixing this problem by transforming the outcome variable:
  - Ex: $\log Y$, $\sqrt{(}Y)$, Y^2, etc.

Example Residual Plot - Before and After log(Y) Transformation



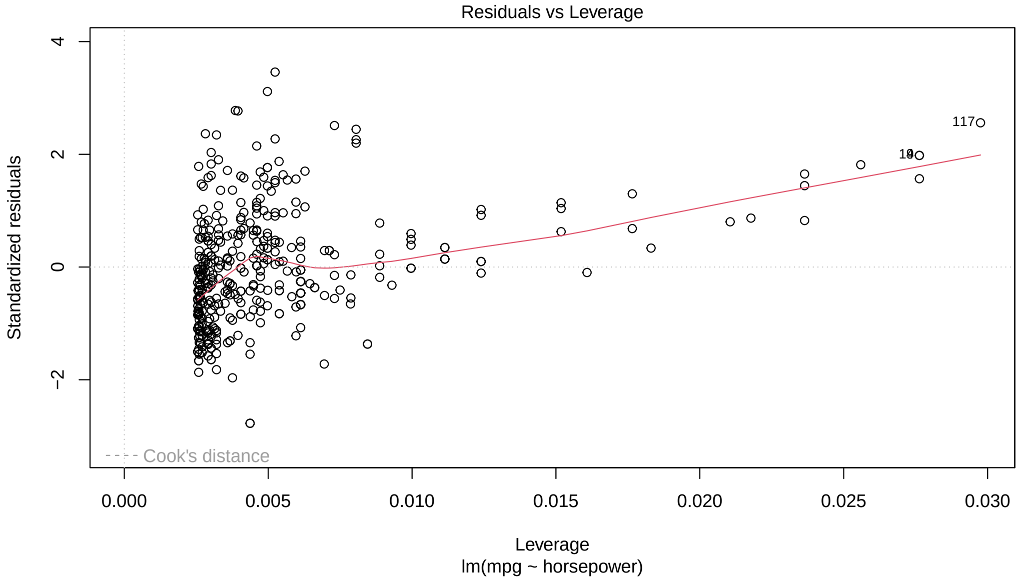
4. Outliers and High-Leverage Points

An *outlier* is a point for which y_i is far from the value predicted by the model.

- If we think the outlier resulted from an error in data collection, we can remove it.
- ... but there is nothing inherently wrong with outliers.

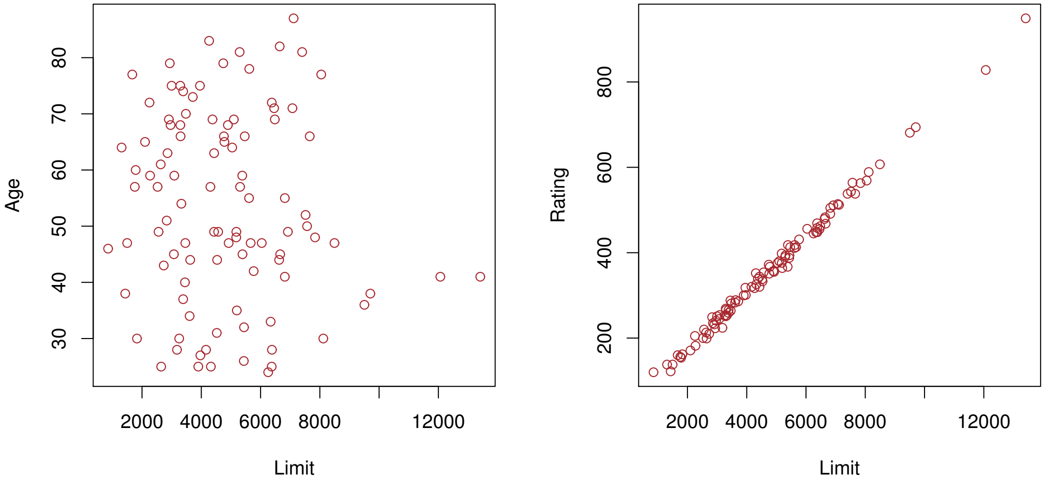
From a model fittng perspective, we are much more interested in *high-leverage points*.

- These are observations which have a significant individual impact on the regression line.
  - We can examine this by removing a point from the data and refitting the model, and then examining how much the regression line changed.


Residuals vs Leverage
Standardized residuals
4
2
0
−2
0.000
0.005
0.010
0.015
0.020
0.025
0.030
Leverage
lm(mpg ~ horsepower)
117
14
Cook's distance

6. Collinearity

Collinearity is the situation in which one or more predictor variables are closely related to one another.



Collinearity

When two variables are *collinear*:

- It can be difficult to separate out their individual effects on the response.
- The accuracy of regression coefficient estimates is decreased.
- Standard error is increased, which shrinks the test statistics (toward 0).
  - This results in larger p-values and potentially a failure to reject H_0.

Dealing with Collinearity

- Examine the correlation matrix for the predictors.

.	mpg	cyl	disp	hp	wt	acc	yr	orgn
mpg	1.00	-0.78	-0.81	-0.78	-0.83	0.42	0.58	0.57
cyl	-0.78	1.00	0.95	0.84	0.90	-0.50	-0.35	-0.57
disp	-0.81	0.95	1.00	0.90	0.93	-0.54	-0.37	-0.61
hp	-0.78	0.84	0.90	1.00	0.86	-0.69	-0.42	-0.46
wt	-0.83	0.90	0.93	0.86	1.00	-0.42	-0.31	-0.59
acc	0.42	-0.50	-0.54	-0.69	-0.42	1.00	0.29	0.21
yr	0.58	-0.35	-0.37	-0.42	-0.31	0.29	1.00	0.18
orgn	0.57	-0.57	-0.61	-0.46	-0.59	0.21	0.18	1.00

Multicollinearity

Sometimes, we can run into collinearity between three or more variables that will not appear in the two-way correlations shown in the correlation matrix.

- To examine possible multicollinearity, we compute the *variance inflation factor* (VIF).
  - This is the ratio of (variance of $\hat{\beta}_j$ when fitting the full model) to (the variance of $\hat{\beta}_j$ if fit on its own).
  - The minimum value for VIF is 1.
  - There are different ideas for what constitutes a "high" VIF, but people often use 5 or 10.

Multicollinearity

```
cylinders   displacement   horsepower
10.737535   21.836792      9.943693
weight      acceleration   year       origin
10.831260   2.625806       1.244952   1.772386
```

- Now what?

Multicollinearity

Let's try removing the variable with the highest VIF:

```
cylinders     horsepower    weight
6.008253      9.088413      9.219674
acceleration  year          origin
2.598356      1.239409      1.594220
```

- Notice that removing `displacement` also slightly improved the VIF for all the other variables!
- At this point, we can stop (if we're using 10) or try removing another variable.

Multicollinearity

Let's try removing one more variable (`displacement`):

```
##    cylinders   horsepower acceleration          year        origin
##     4.155143     5.323311     1.996560      1.209909      1.495100
```

- That made a big difference!

The Final Model

```
Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept)    -7.87876    5.05154  -1.560     0.12
cylinders      -1.22202    0.22524  -5.425 1.02e-07 ***
horsepower     -0.08815    0.01130  -7.802 5.75e-14 ***
acceleration   -0.40305    0.09654  -4.175 3.69e-05 ***
year            0.66601    0.05628  11.833  < 2e-16 ***
origin          1.82772    0.28612   6.388 4.84e-10 ***
---
Residual standard error: 3.727 on 386 degrees of freedom
Multiple R-squared:  0.7749,   Adjusted R-squared:  0.772
F-statistic: 265.7 on 5 and 386 DF,  p-value: < 2.2e-16
```