Lecture 21 - Model Selection

Sta102 / BME102

Colin Rundel

November 18, 2015

Modeling children's test scores

Predicting cognitive test scores of three- and four-year-old children using characteristics of their mothers. Data are a subsample from the National Longitudinal Survey of Youth.

	kid_score	mom_hs	mom_iq	mom_work	mom_age
1	65	yes	121.12	yes	27
⋮					
5	115	yes	92.75	yes	27
6	98	no	107.90	no	18
⋮					
434	70	yes	91.25	yes	25

Gelman, Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* (2007) Cambridge University Press.

Model output

```
summary(lm(kid_score ~ mom_hs + mom_iq + mom_work + mom_age, data = cognitive))

##
## Call:
## lm(formula = kid_score ~ mom_hs + mom_iq + mom_work + mom_age,
##     data = cognitive)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.134 -12.624   2.293  11.250  50.206
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20.82261    9.18765   2.266   0.0239 *
## mom_hs       5.56118    2.31345   2.404   0.0166 *
## mom_iq       0.56208    0.06077   9.249   <2e-16 ***
## mom_work     0.13373    0.76763   0.174   0.8618
## mom_age      0.21986    0.33231   0.662   0.5086
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.17 on 429 degrees of freedom
## Multiple R-squared:  0.215,Adjusted R-squared:  0.2077
## F-statistic: 29.38 on 4 and 429 DF,  p-value: < 2.2e-16
```


Conditions for MLR Inference

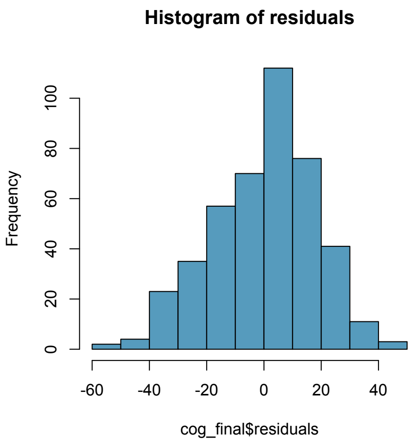
In order to conduct inference for multiple regression we require the following conditions:

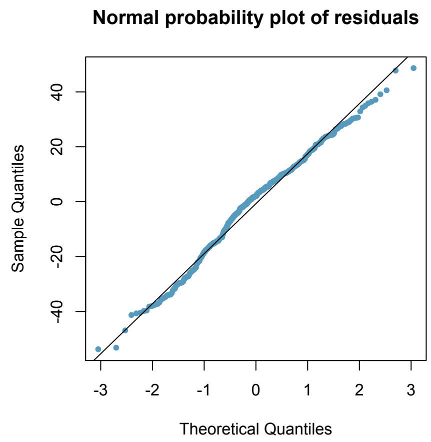
(1) Unstructured / nearly normal residuals

(2) Constant variability of residuals

(3) Independent residuals

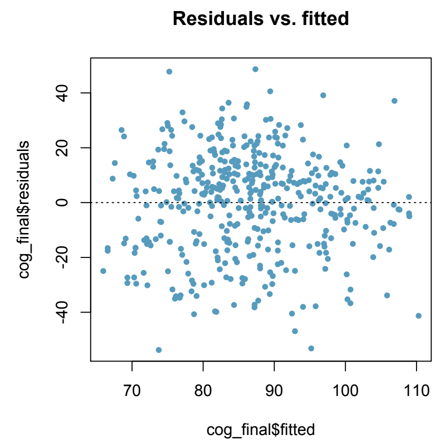
Nearly normal residuals

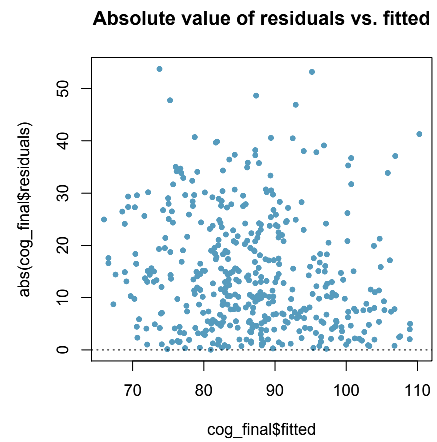




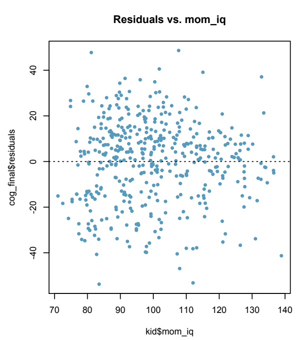
Unstructured / Constant variability of residuals

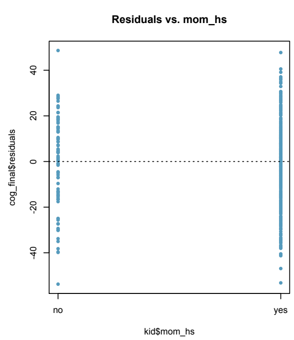
Why do we use the fitted (predicted) values in MLR?

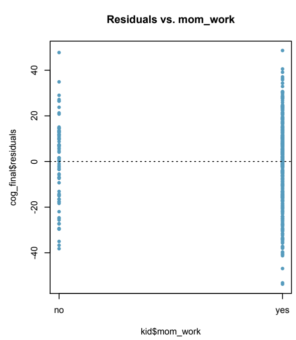




Constant variability of residuals (cont.)

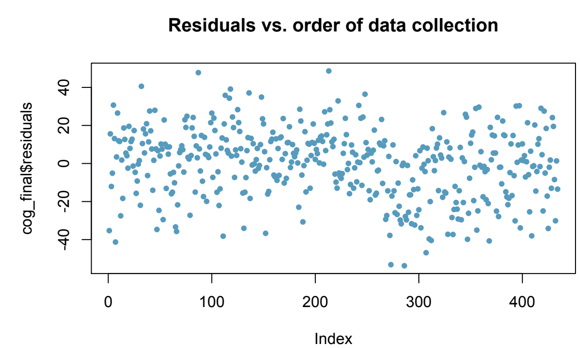






Independent residuals

- If we suspect that order of data collection may influence the outcome (mostly in time series data):



- If not, think about how data are sampled.

Model output

```
summary(lm(kid_score ~ mom_hs + mom_iq + mom_work + mom_age, data = cognitive))


##
## Call:
## lm(formula = kid_score ~ mom_hs + mom_iq + mom_work + mom_age,
##     data = cognitive)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.134 -12.624   2.293  11.250  50.206
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20.82261    9.18765   2.266   0.0239 *
## mom_hs       5.56118    2.31345   2.404   0.0166 *
## mom_iq       0.56208    0.06077   9.249   <2e-16 ***
## mom_work     0.13373    0.76763   0.174   0.8618
## mom_age      0.21986    0.33231   0.662   0.5086
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.17 on 429 degrees of freedom
## Multiple R-squared:  0.215,Adjusted R-squared:  0.2077
## F-statistic: 29.38 on 4 and 429 DF,  p-value: < 2.2e-16
```


Inference for the model as a whole

Is the model as a whole significant?

$$H_0 : \beta_0 = \beta_1 = \cdots = \beta_k = 0$$
$$H_A : \text{At least one of the } \beta_i \neq 0$$

```
F-statistic: 29.38 on 4 and 429 DF,  p-value: < 2.2e-16
```

Since p-value < 0.05, the model as a whole is significant.

- The F test yielding a significant result doesn't mean the model fits the data well, it just means at least one of the βs is non-zero.
- The F test not yielding a significant result doesn't mean individuals variables included in the model are not good predictors of y, it just means that the combination of these variables doesn't yield a good model.

ANOVA Table

```
anova(lm(kid_score~.,data=cognitive))

## Analysis of Variance Table
##
## Response: kid_score
##            Df Sum Sq Mean Sq F value    Pr(>F)
## mom_hs      1  10125 10125.0 30.6763 5.325e-08 ***
## mom_iq      1  28504 28504.1 86.3608 < 2.2e-16 ***
## mom_work    1     18    17.6  0.0533    0.8175
## mom_age     1    144   144.5  0.4377    0.5086
## Residuals 429 141595   330.1
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$MS_{Reg} = (18 + 144 + 10125 + 28504)/4 = 9697.75$$
$$F_{Reg} = 9697.75/330.1 = 29.38$$

```
F-statistic: 29.38 on 4 and 429 DF,  p-value: < 2.2e-16
```


Inference for the slope(s)

Is whether or not a mother graduated from high school a significant predictor of kid's cognitive test score, given all other variables in the model?

$$H_0 : \beta_1 = 0, \text{ when all other variables are included in the model}$$
$$H_A : \beta_1 \neq 0, \text{ when all other variables are included in the model}$$

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.59241    9.21906   2.125   0.0341
mom_hsyes    5.09482    2.31450   2.201   0.0282
mom_iq       0.56147    0.06064   9.259   <2e-16
mom_workyes  2.53718    2.35067   1.079   0.2810
mom_age      0.21802    0.33074   0.659   0.5101

Residual standard error: 18.14 on 429 degrees of freedom
```

$$T = 2.201, \; df = n - k - 1 = 434 - 4 - 1 = 429, \; \text{p-value} = 0.0282$$

Since p-value < 0.05, whether or not mom went to high school is a significant predictor of kid's test score, given all other variables in the model.

Interpreting the slope

What is the correct interpretation of the slope for `mom_work`?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.59	9.22	2.13	0.03
mom_hs:yes	5.09	2.31	2.20	0.03
mom_iq	0.56	0.06	9.26	0.00
mom_work:yes	2.54	2.35	1.08	0.28
mom_age	0.22	0.33	0.66	0.51

CI Recap from last time

Inference for the slope for a SLR model (only one explanatory variable):

- Hypothesis test:

$$T = \frac{b_1 - null\ value}{SE_{b_1}} \qquad df = n - 2$$

- Confidence interval:

$$b_1 \pm t^\star_{df} \times SE_{b_1}$$

The only difference for MLR is that we use b_i instead of b_1, and use $df = n - k - 1$

CI for the slope

Construct a 95% confidence interval for the slope of `mom_work`.

$$\begin{aligned} b_k &\pm t^\star SE_{b_k} \\ df &= n - k - 1 = 434 - 4 - 1 = 429 \to 400 \\ 2.54 &\pm 1.97 \times 2.35 \\ 2.54 &\pm 4.63 \\ (-2.0895 &, 7.1695) \end{aligned}$$

Interpretation?

Inference for the slope(s) (cont.)

Given all variables in the model, which variables are significant predictors of kid's cognitive test score?

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.59241    9.21906   2.125   0.0341
mom_hsyes    5.09482    2.31450   2.201   0.0282
mom_iq       0.56147    0.06064   9.259   <2e-16
mom_workyes  2.53718    2.35067   1.079   0.2810
mom_age      0.21802    0.33074   0.659   0.5101
```

Modeling kid's test scores (revisited)

Predicting cognitive test scores of three- and four-year-old children using characteristics of their mothers. Data are from a survey of adult American women and their children - a subsample from the National Longitudinal Survey of Youth.

	kid_score	mom_hs	mom_iq	mom_work	mom_age
1	65	yes	121.12	yes	27
⋮	⋮	⋮	⋮	⋮	⋮
5	115	yes	92.75	yes	27
6	98	no	107.90	no	18
⋮	⋮	⋮	⋮	⋮	⋮
434	70	yes	91.25	yes	25

Gelman, Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* (2007) Cambridge University Press.

Model output

```
cog_full = lm(kid_score ~ mom_hs + mom_iq + mom_work + mom_age,
              data = cognitive)
summary(cog_full)

##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 19.59241    9.21906   2.125   0.0341
## mom_hsyes    5.09482    2.31450   2.201   0.0282
## mom_iq       0.56147    0.06064   9.259   <2e-16
## mom_workyes  2.53718    2.35067   1.079   0.2810
## mom_age      0.21802    0.33074   0.659   0.5101
##
## Residual standard error: 18.14 on 429 degrees of freedom
## Multiple R-squared: 0.2171, Adjusted R-squared: 0.2098
## F-statistic: 29.74 on 4 and 429 DF,  p-value: < 2.2e-16
```


Backward-elimination

- Adjusted R^2 approach:
  - Start with the full model
  - Drop one variable at a time and record R^2_{adj} of each smaller model
  - Pick the model with the largest increase in R^2_{adj}
  - Repeat until none of the reduced models yield an increase in R^2_{adj}
- When removing a categorical variable all levels should be included or removed *at the same time*

Backward-selection: R^2_{adj} approach

Step	**Variables included**	$\mathbf{R^2_{adj}}$
Full	kid_score ~ mom_hs + mom_iq + mom_work + mom_age	*0.2098*
Step 1	kid_score ~ mom_iq + mom_work + mom_age	0.2027
	kid_score ~ mom_hs + mom_work + mom_age	0.0541
	kid_score ~ mom_hs + mom_iq + mom_age	0.2095
	kid_score ~ mom_hs + mom_iq + mom_work	*0.2109*
Step 2	kid_score ~ mom_iq + mom_work	0.2024
	kid_score ~ mom_hs + mom_work	0.0546
	kid_score ~ mom_hs + mom_iq	*0.2105*
Step 3*	kid_score ~ mom_hs	0.2024
	kid_score ~ mom_iq	0.0546

Backward-selection: R^2_{adj} approach

Step	Variables included	R^2_{adj}
Full	kid_score ~ mom_hs + mom_iq + mom_work + mom_age	0.2098
Step 1	kid_score ~ mom_iq + mom_work + mom_age	0.2027
	kid_score ~ mom_hs + mom_work + mom_age	0.0541
	kid_score ~ mom_hs + mom_iq + mom_age	0.2095
	kid_score ~ mom_hs + mom_iq + mom_work	0.2109
Step 2	kid_score ~ mom_iq + mom_work	0.2024
	kid_score ~ mom_hs + mom_work	0.0546
	kid_score ~ mom_hs + mom_iq	0.2105
Step 3*	kid_score ~ mom_hs	0.2024
	kid_score ~ mom_iq	0.0546

Forward-selection

1. Adjusted R^2 approach:
   - Start with regression of response vs. each explanatory variable
   - Pick the model with the highest R^2_{adj}
   - Add the remaining variables one at a time to the existing model, and once again pick the model with the highest R^2_{adj}
   - Repeat until the addition of any of the remanning variables does not result in a higher R^2_{adj}

Forward-selection: R^2_{adj} approach

Step	Variables included	R^2_{adj}
Step 1	kid_score ~ mom_hs	0.0539
	kid_score ~ mom_work	0.0097
	kid_score ~ mom_age	0.0062
	kid_score ~ mom_iq	*0.1991*
Step 2	kid_score ~ mom_iq + mom_work	0.2024
	kid_score ~ mom_iq + mom_age	0.1999
	kid_score ~ mom_iq + mom_hs	*0.2105*
Step 3	kid_score ~ mom_iq + mom_hs + mom_age	0.2095
	kid_score ~ mom_iq + mom_hs + mom_work	*0.2109*
Step 4*	kid_score ~ mom_iq + mom_hs + mom_age + mom_work	0.2098

Forward-selection: R^2_{adj} approach

Step	Variables included	R^2_{adj}
Step 1	kid_score ~ mom_hs	0.0539
	kid_score ~ mom_work	0.0097
	kid_score ~ mom_age	0.0062
	kid_score ~ mom_iq	0.1991
Step 2	kid_score ~ mom_iq + mom_work	0.2024
	kid_score ~ mom_iq + mom_age	0.1999
	kid_score ~ mom_iq + mom_hs	0.2105
Step 3	kid_score ~ mom_iq + mom_hs + mom_age	0.2095
	kid_score ~ mom_iq + mom_hs + mom_work	0.2109
Step 4*	kid_score ~ mom_iq + mom_hs + mom_age + mom_work	0.2098

Expert opinion as criterion for model selection

In addition to the quantitative approaches we discussed, variables can be included in (or eliminated from) the model based on expert opinion.

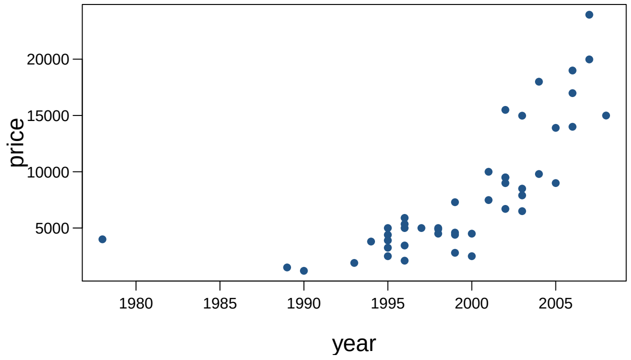
Final model choice

```
cog_final = lm(kid_score ~ mom_hs + mom_iq, data = kid)
summary(cog_final)

## Call:
## lm(formula = kid_score ~ mom_hs + mom_iq, data = kid)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.73154    5.87521   4.380 1.49e-05 ***
## mom_hsyes    5.95012    2.21181   2.690  0.00742 **
## mom_iq       0.56391    0.06057   9.309  < 2e-16 ***
##
## Residual standard error: 18.14 on 431 degrees of freedom
## Multiple R-squared: 0.2141, Adjusted R-squared: 0.2105
## F-statistic: 58.72 on 2 and 431 DF,  p-value: < 2.2e-16
```


Truck prices

The scatterplot below shows the relationship between year and price of a random sample of 43 pickup trucks. Describe the relationship between these two variables.

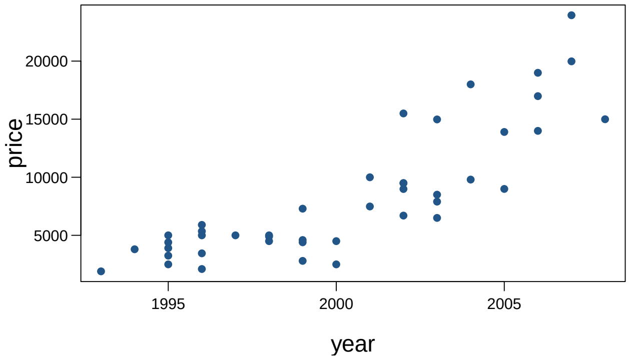


From: *http://faculty.chicagobooth.edu/robert.gramacy/teaching.html*

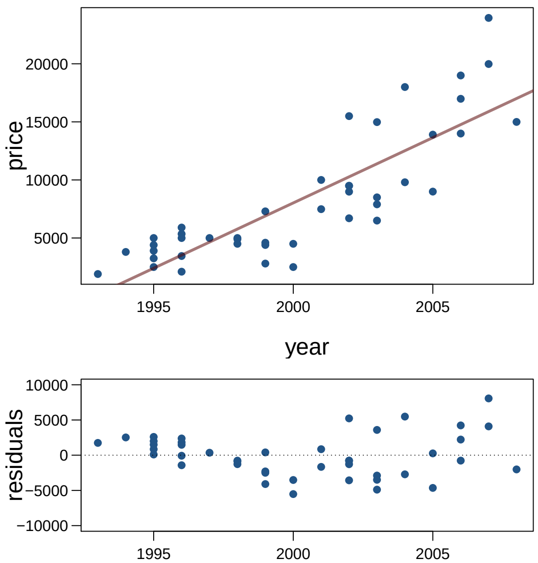
Remove unusual observations

Let's remove trucks older than 20 years, and only focus on trucks made in 1992 or later.

Now what can you say about the relationship?



Truck prices - linear model?



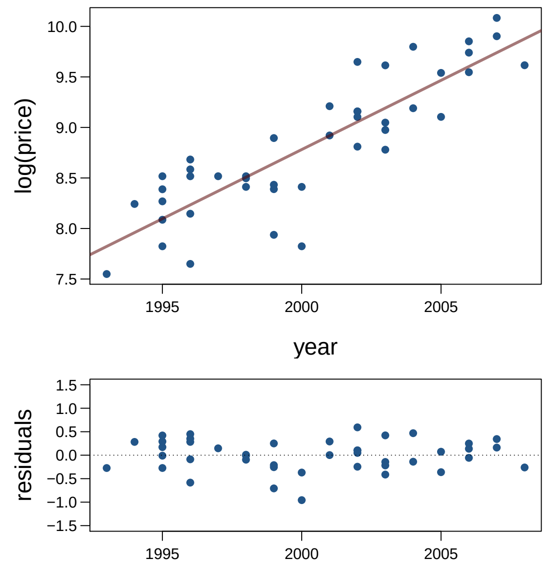
Model:

$$\widehat{price} = b_0 + b_1 \; year$$

The linear model doesn't appear to be a good fit since the residuals have non-constant variance.

In particular residuals for newer cars (to the right) have a larger variance than the residuals for older cars (to the left).

Truck prices - log transform of the response variable



Model:

$$\widehat{log(price)} = b_0 + b_1 \; year$$

We have applied a log transformation to the response variable. The relationship now seems linear, and the residuals have (more) constant variance.

Interpreting models with log transformation

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-265.07	25.04	-10.59	0.00
pu$year	0.14	0.01	10.94	0.00

Model: $\widehat{log(price)} = -265.07 + 0.14 \; year$

- For each additional year the car is newer (for each year decrease in car's age) we would expect the log price of the car to increase on average by 0.14 log dollars.
- which is not very useful ...

Working with logs

- Subtraction and logs:

$$log(a) - log(b) = log\left(\frac{a}{b}\right)$$

- Natural logarithm:

$$e^{log(x)} = x$$

- We can use these identities to "undo" the log transformation

Interpreting models with log transformation (cont.)

The slope coefficient for the log transformed model is 0.14, meaning the log price difference between cars that are one year apart is predicted to be 0.14 log dollars.

$$\log(\text{price 2}) = -265.07 + 0.14\ (y + 1)$$
$$\log(\text{price 1}) = -265.07 + 0.14\ y$$

$$\begin{aligned}
\log(\text{price 2}) - \log(\text{price 1}) &= 0.14 \\
log\left(\frac{\text{price 2}}{\text{price 1}}\right) &= 0.14 \\
e^{log\left(\frac{\text{price 2}}{\text{price 1}}\right)} &= e^{0.14} \\
\frac{\text{price 2}}{\text{price 1}} &= 1.15
\end{aligned}$$

For each additional year the car is newer (for each year decrease in car's age) we would expect the price of the car to increase on average by *a factor of 1.15*.

Recap: dealing with non-constant variance

- Non-constant variance is one of the most common model violations, however it is usually fixable by transforming the response (y) variable
- The most common variance stabilizing transform is the log transformation: $log(y)$, especially useful when the response variable is (extremely) right skewed.
- When using a log (or any other) transformation on the response variable the interpretation of the slope changes:
  - For log - each unit increase in x, y is expected on average to decrease/increase by a factor of e^{b_1}.
- Another useful transformation is the square root: $\sqrt{y}$, especially useful when the response variable is counts.
- These transformations may also be useful when the relationship is non-linear, but in those cases a polynomial regression may also be needed (this is beyond the scope of this course, but you're welcomed to try it for your project, and I'd be happy to provide further guidance)