Lab 9 and PS 9 due Thursday
Poster session in lab on next Monday
Paper + project peer evaluation due next Wednesday at 5pm
Review materials will be posted over the weekend
Final review session - Thursday, May 2 (12:30 - 1:30pm or 5:30-6:30pm?)
Check grades on Sakai - no changes will be made after the final
Bring at least one laptop per team to class on Thursday

Design of scientific posters:
http://www.writing.engr.psu.edu/posters.html
Poster Perfect, The Scientist:
http://www.the-scientist.com/?articles.view/articleNo/31071/title/Poster-Perfect/
Poster session tips:
http://www.personal.psu.edu/drs18/postershow

Resizing plots: adjust fig.width and fig.height arguments

```r
fig.width=5,fig.height=5
```

Multiple plots at once:
1 row of 3 plots:
```r
par(mfrow=c(1,3))
```
2 rows of 3 plots each:
```r
par(mfrow=c(2,3))
```

Changing font size for code and output: Insert on top of your document, and adjust font size. Default is 12 point font. Do not use less than 9 px.

```r
# some code here to produce a plot
```

Hiding code: Especially useful for space savings due to code like load data, etc. The code will be ran, and any output will still be displayed.

```r
echo=FALSE
```

# some code that you don't want to show in your paper
Modeling kid’s test scores

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.

<table>
<thead>
<tr>
<th>kid_score</th>
<th>mom_hs</th>
<th>mom_iq</th>
<th>mom_work</th>
<th>mom_age</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>65</td>
<td>yes</td>
<td>121.12</td>
<td>27</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>5</td>
<td>115</td>
<td>yes</td>
<td>92.75</td>
<td>27</td>
</tr>
<tr>
<td>6</td>
<td>98</td>
<td>no</td>
<td>107.90</td>
<td>18</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>434</td>
<td>70</td>
<td>yes</td>
<td>91.25</td>
<td>25</td>
</tr>
</tbody>
</table>


Inference for MLR

Is the model as a whole significant?

H₀ : β₁ = β₂ = ... = βₖ = 0
Hₐ : At least one of the βᵢ ≠ 0

F-statistic: 29.74 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 βᵢs 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.

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 | 2.107 | 0.0281 |
| 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

Inference for the model as a whole

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

H₀ : β₁ = 0, when all other variables are included in the model
Hₐ : β₁ ≠ 0, when all other variables are included in the model

T = 2.201, df = n – k – 1 = 434 – 4 – 1 = 429, 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

Clicker question

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

|          | Estimate | Std. Error | t value | Pr(>|t|) |
|----------|----------|------------|---------|----------|
| (Intercept) | 19.592   | 9.219      | 2.125   | 0.0341   |
| mom_hs:yes | 5.094    | 2.314      | 2.201   | 0.0282   |
| mom_iq    | 0.561    | 0.061      | 9.259   | <2e-16   |
| mom_work:yes | 2.537 | 2.351      | 1.079   | 0.2810   |
| mom_age   | 0.218    | 0.331      | 0.659   | 0.5101   |

All else being equal, kids whose moms worked during the first three years of the kid’s life

(a) are estimated to score 2.54 points lower

(b) are estimated to score 2.54 points higher

than those whose moms did not work.

Model selection

Backward-elimination

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

|          | Estimate | Std. Error | t value | Pr(>|t|) |
|----------|----------|------------|---------|----------|
| (Intercept) | 19.592   | 9.219      | 2.125   | 0.0341   |
| mom_hs:yes | 5.094    | 2.314      | 2.201   | 0.0282   |
| mom_iq    | 0.561    | 0.061      | 9.259   | <2e-16   |
| mom_work:yes | 2.537 | 2.351      | 1.079   | 0.2810   |
| mom_age   | 0.218    | 0.331      | 0.659   | 0.5101   |

**R^2_adj** 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 highest increase in R^2_adj
- Repeat until none of the models yield an increase in R^2_adj

**p-value** approach:
- Start with the full model
- Drop the variable with the highest p-value and refit a smaller model
- Repeat until all variables left in the model are significant
### Backward-selection: $R^2_{adj}$ approach

#### Step 1
<table>
<thead>
<tr>
<th>Variables included</th>
<th>$R^2_{adj}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>kid_score ~ mom hs + mom iq + mom work + mom age</td>
<td>0.2098</td>
</tr>
</tbody>
</table>

#### Step 2
<table>
<thead>
<tr>
<th>Variables included</th>
<th>$R^2_{adj}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>kid_score ~ mom iq + mom work [-mom hs]</td>
<td>0.2024</td>
</tr>
<tr>
<td>kid_score ~ mom hs + mom iq [-mom hs]</td>
<td>0.0546</td>
</tr>
<tr>
<td>kid_score ~ mom hs + mom iq [-mom work]</td>
<td>0.2105</td>
</tr>
</tbody>
</table>

### Backward-selection: p-value approach

#### Full model:

| Estimate  | Std. Error | t value | Pr(>|t|) |
|-----------|------------|---------|----------|
| (Intercept) | 19.5924    | 9.2196  | 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 |

#### Step 2: Remove mom_age

| Estimate  | Std. Error | t value | Pr(>|t|) |
|-----------|------------|---------|----------|
| (Intercept) | 24.17944   | 6.04319 | 4.001    | 7.42e-05 ***|
| mom_hsyes  | 5.38225    | 2.27156 | 2.369    | 0.0183 * |
| mom_iq     | 0.56278    | 0.06057 | 9.291    | < 2e-16 ***|
| mom_workyes| 2.56640    | 2.34871 | 1.093    | 0.2751 |

#### Step 2: Remove mom_work

| 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 ***|

### Clicker question

The following model is used to predict income from hours worked, race, and birth quarter (using the ACS dataset from lab). Which of the following variables should be dropped from the model when doing backwards elimination using the p-value approach?

- (a) gender:female
- (b) race:other
- (c) race:black
- (d) race variable as a whole
- (e) none

### adjusted $R^2$ vs. p-value

- If you're interested in finding out which variables are significant predictors, use p-value approach.
- If you're interested in more reliable predictions, use adjusted $R^2$ method.
- Note that the p-value method depends on the (somewhat arbitrary) 5% significance level cutoff. Using a different significance level you could get a completely different model. But it's used commonly since it requires fitting fewer models (in the more commonly used backwards-selection approach).
### Forward-selection

**$R^2_{adj}$ approach:**
- Start with regressions 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 remaining variables does not result in a higher $R^2_{adj}$

**$p$ – value approach:**
- Start with regressions of response vs. each explanatory variable
- Pick the variable with the lowest significant p-value
- Add the remaining variables one at a time to the existing model, and pick the variable with the lowest significant p-value
- Repeat until any of the remaining variables does not have a significant p-value

*In forward-selection the p-value approach isn’t any simpler (you still need to fit a bunch of models), so there’s almost no incentive to use it.*

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

<table>
<thead>
<tr>
<th>Step</th>
<th>Variables included</th>
<th>$R^2_{adj}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 1</td>
<td>kid_score ~ mom_hs</td>
<td>0.0539</td>
</tr>
<tr>
<td></td>
<td>kid_score ~ mom_work</td>
<td>0.0097</td>
</tr>
<tr>
<td></td>
<td>kid_score ~ mom_age</td>
<td>0.0062</td>
</tr>
<tr>
<td></td>
<td>kid_score ~ mom_iq</td>
<td>0.1991</td>
</tr>
<tr>
<td>Step 2</td>
<td>kid_score ~ mom_iq + mom_work</td>
<td>0.2024</td>
</tr>
<tr>
<td></td>
<td>kid_score ~ mom_iq + mom_age</td>
<td>0.1999</td>
</tr>
<tr>
<td></td>
<td>kid_score ~ mom_iq + mom_hs</td>
<td>0.2105</td>
</tr>
<tr>
<td>Step 3</td>
<td>kid_score ~ mom_iq + mom_hs + mom_age</td>
<td>0.2095</td>
</tr>
<tr>
<td></td>
<td>kid_score ~ mom_iq + mom_hs + mom_work</td>
<td>0.2109</td>
</tr>
<tr>
<td>Step 4</td>
<td>kid_score ~ mom_iq + mom_hs + mom_age + mom_work</td>
<td>0.2098</td>
</tr>
</tbody>
</table>

### Clicker question

Which variable should be added to the model first?

(a) mom_hs  
(b) mom_iq  
(c) mom_work  
(d) mom_age

### 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 + mom_work, data = cognitive)
> summary(cog_final)

Call:
lm(formula = kid_score ~ mom_hs + mom_iq + mom_work, data = cognitive)

Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 24.17944 | 6.04319    | 4.001   | 7.42e-05 *** |
| mom_hsyes      | 5.38225  | 2.27156    | 2.369   | 0.0183 *   |
| mom_iq         | 0.56278  | 0.06057    | 9.291   | < 2e-16 *** |
| mom_workyes    | 2.56640  | 2.34871    | 1.093   | 0.2751     |

Residual standard error: 18.13 on 430 degrees of freedom
Multiple R-squared: 0.2163, Adjusted R-squared: 0.2109
F-statistic: 39.57 on 3 and 430 DF, p-value: < 2.2e-16

Conditions for MLR

Clicker question

Which of the following is not a condition for MLR?

(a) Nearly normal residuals with mean 0
(b) Constant variability of residuals
(c) Independent residuals
(d) Each (numerical) variable linearly related to outcome
(e) Nearly normal response variable with mean 0

Nearly normal residuals with mean 0

Histogram of residuals

Normal probability plot of residuals

Constant variability of residuals

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

Constant variability of residuals (cont.)

Each (numerical) variable linearly related to outcome

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.

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
Transformations

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:

\[
\hat{\text{price}} = b_0 + b_1 \text{ year}
\]

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

Truck prices - log transform of the response variable

Model:

\[
\log(\text{price}) = b_0 + b_1 \text{ year}
\]

We applied a log transformation to the response variable. The relationship now seems linear, and the residuals no longer have non-constant variance.

Interpreting models with log transformation

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

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.

\[
\log\left(\frac{\text{price at year } x + 1}{\text{price at year } x}\right) = 0.14
\]

\[
e^{0.14} \approx 1.15
\]

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 transformation on the response variable the interpretation of the slope changes:
  - For each unit increase in \( x \), \( y \) is expected on average to decrease/increase by a factor of \( e^b \).
- 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).