Announcements

- Homework
  - Homework 7 and Lab due on Monday
- Test a week from today!
  - Over Chapters 2,3,4
  - I'm trying to get study materials posted today or tomorrow
  - I'm also trying hard to have grading done by that point
- Final Project Info sent out today!
  - I'm getting potential datasets uploaded over the weekend.
  - You are always free to find your own as well. Just make sure they have enough variables to give your options for multiple regression.
- Polling: `rembold-class.ddns.net`

ANOVA in R

Run the ANOVA test on a linear fit to the data:

```
anova(lm(aldrin$aldrin ~ aldrin$depth))
```

```
Response: aldrin$aldrin
             Df Sum Sq Mean Sq F value   Pr(>F)
aldrin$depth  2 16.961  8.4803  6.1338 0.006367 **
Residuals    27 37.329  1.3826
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

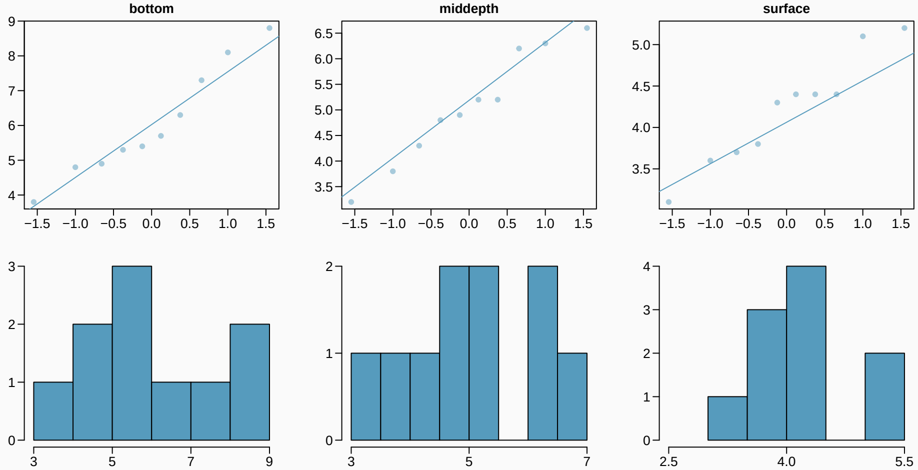
Checking Conditions: (1) independence

Does this condition appear to be satisfied?

In this study the we have no reason to believe that the aldrin concentration won't be independent of each other.

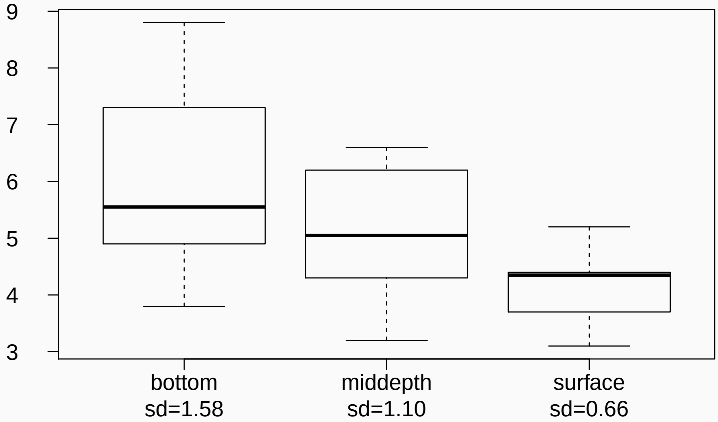
Checking Conditions: (2) approximately normal

Does this condition appear to be satisfied?



Checking Conditions: (3) constant variance

Does this condition appear to be satisfied?



Which means differ?

- Earlier we concluded that at least one pair of means differ. The natural question that follows is “which ones?”

Which means differ?

- Earlier we concluded that at least one pair of means differ. The natural question that follows is “which ones?”
- We can do two-sample t tests for differences in each possible pair of groups.

Which means differ?

- Earlier we concluded that at least one pair of means differ. The natural question that follows is “which ones?”
- We can do two-sample t tests for differences in each possible pair of groups.

Can you see any pitfalls with this approach?

Which means differ?

- Earlier we concluded that at least one pair of means differ. The natural question that follows is “which ones?”
- We can do two-sample t tests for differences in each possible pair of groups.

Can you see any pitfalls with this approach?

- When we run too many tests, the Type 1 Error rate increases.
- This issue is resolved by using a modified significance level.

Multiple comparisons

- The scenario of testing many pairs of groups is called ***multiple comparisons***.

Multiple comparisons

- The scenario of testing many pairs of groups is called ***multiple comparisons***.
- The ***Bonferroni correction*** suggests that a more stringent significance level is more appropriate for these tests:

$$\alpha^\star = \alpha/K$$

where K is the number of comparisons being considered.

Multiple comparisons

- The scenario of testing many pairs of groups is called ***multiple comparisons***.
- The ***Bonferroni correction*** suggests that a more stringent significance level is more appropriate for these tests:

$$\alpha^{\star} = \alpha / K$$

  where K is the number of comparisons being considered.
- If there are k groups, then usually all possible pairs are compared and $K = \frac{k(k-1)}{2}$.

Understanding Check!

In the aldrin data set depth has 3 levels: bottom, mid-depth, and surface. If $\alpha = 0.05$, what should be the modified significance level for two sample t tests for determining which pairs of groups have significantly different means?

A) $\alpha^* = 0.05$

B) $\alpha^* = 0.05/2 = 0.025$

C) $\alpha^* = 0.05/3 = 0.0167$

D) $\alpha^* = 0.05/6 = 0.0083$

Understanding Check!

In the aldrin data set depth has 3 levels: bottom, mid-depth, and surface. If $\alpha = 0.05$, what should be the modified significance level for two sample t tests for determining which pairs of groups have significantly different means?

A) $\alpha^* = 0.05$

B) $\alpha^* = 0.05/2 = 0.025$

C) $\alpha^* = 0.05/3 = 0.0167$

D) $\alpha^* = 0.05/6 = 0.0083$

Which means differ?

If the ANOVA assumption of equal variability across groups is satisfied, we can use the data from all groups to estimate variability:

- Estimate any within-group standard deviation with $\sqrt{MSE}$, which is s_{pooled}
- Use the error degrees of freedom, $n - k$, for t-distributions

Difference in two means: after ANOVA

$$SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \approx \sqrt{\frac{MSE}{n_1} + \frac{MSE}{n_2}}$$

Bottom and Mid Check

Is there a difference between the average aldrin concentration at the bottom and at mid depth?

	n	mean	sd
bottom	10	6.04	1.58
middepth	10	5.05	1.10
surface	10	4.2	0.66
overall	30	5.1	1.37

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
depth	2	16.96	8.48	6.13	0.0063
Residuals	27	37.33	1.38		
Total	29	54.29			

Bottom and Mid Check

Is there a difference between the average aldrin concentration at the bottom and at mid depth?

	n	mean	sd
bottom	10	6.04	1.58
middepth	10	5.05	1.10
surface	10	4.2	0.66
overall	30	5.1	1.37

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
depth	2	16.96	8.48	6.13	0.0063
Residuals	27	37.33	1.38		
Total	29	54.29			

$$T_{df_E} = \frac{(\bar{x}_{bottom} - \bar{x}_{middepth})}{\sqrt{\frac{MSE}{n_{bottom}} + \frac{MSE}{n_{middepth}}}}$$

$$T_{27} = \frac{(6.04 - 5.05)}{\sqrt{\frac{1.38}{10} + \frac{1.38}{10}}} = \frac{0.99}{0.53} = 1.87$$

Bottom and Mid Check

Is there a difference between the average aldrin concentration at the bottom and at mid depth?

	n	mean	sd
bottom	10	6.04	1.58
middepth	10	5.05	1.10
surface	10	4.2	0.66
overall	30	5.1	1.37

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
depth	2	16.96	8.48	6.13	0.0063
Residuals	27	37.33	1.38		
Total	29	54.29			

$$T_{df_E} = \frac{(\bar{x}_{bottom} - \bar{x}_{middepth})}{\sqrt{\frac{MSE}{n_{bottom}} + \frac{MSE}{n_{middepth}}}}$$

$$T_{27} = \frac{(6.04 - 5.05)}{\sqrt{\frac{1.38}{10} + \frac{1.38}{10}}} = \frac{0.99}{0.53} = 1.87$$

$$0.05 < p - value < 0.10 \qquad \text{(two-sided)}$$

$$\alpha^{\star} = 0.05/3 = 0.0167$$

Bottom and Mid Check

Is there a difference between the average aldrin concentration at the bottom and at mid depth?

	n	mean	sd
bottom	10	6.04	1.58
middepth	10	5.05	1.10
surface	10	4.2	0.66
overall	30	5.1	1.37

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
depth	2	16.96	8.48	6.13	0.0063
Residuals	27	37.33	1.38		
Total	29	54.29			

$$T_{df_E} = \frac{(\bar{x}_{bottom} - \bar{x}_{middepth})}{\sqrt{\frac{MSE}{n_{bottom}} + \frac{MSE}{n_{middepth}}}}$$

$$T_{27} = \frac{(6.04 - 5.05)}{\sqrt{\frac{1.38}{10} + \frac{1.38}{10}}} = \frac{0.99}{0.53} = 1.87$$

$$0.05 < p - value < 0.10 \qquad \text{(two-sided)}$$

$$\alpha^{\star} = 0.05/3 = 0.0167$$

Fail to reject H_0, the data do not provide convincing evidence of a difference between the average aldrin concentrations at bottom and mid depth.

Bottom and Surface

Is there a difference between the average aldrin concentration at the bottom and at surface?

Bottom and Surface

Is there a difference between the average aldrin concentration at the bottom and at surface?

$$T_{df_E} = \frac{(\bar{x}_{bottom} - \bar{x}_{surface})}{\sqrt{\frac{MSE}{n_{bottom}} + \frac{MSE}{n_{surface}}}}$$

Bottom and Surface

Is there a difference between the average aldrin concentration at the bottom and at surface?

$$
\begin{aligned}
T_{df_E} &= \frac{(\bar{x}_{bottom} - \bar{x}_{surface})}{\sqrt{\frac{MSE}{n_{bottom}} + \frac{MSE}{n_{surface}}}} \\
T_{27} &= \frac{(6.04 - 4.02)}{\sqrt{\frac{1.38}{10} + \frac{1.38}{10}}} = \frac{2.02}{0.53} = 3.81
\end{aligned}
$$

Bottom and Surface

Is there a difference between the average aldrin concentration at the bottom and at surface?

$$
\begin{aligned}
T_{df_E} &= \frac{(\bar{x}_{bottom} - \bar{x}_{surface})}{\sqrt{\frac{MSE}{n_{bottom}} + \frac{MSE}{n_{surface}}}} \\
T_{27} &= \frac{(6.04 - 4.02)}{\sqrt{\frac{1.38}{10} + \frac{1.38}{10}}} = \frac{2.02}{0.53} = 3.81 \\
p - value &< 0.01 \qquad \text{(two-sided)}
\end{aligned}
$$

Bottom and Surface

Is there a difference between the average aldrin concentration at the bottom and at surface?

$$
\begin{aligned}
T_{df_E} &= \frac{(\bar{x}_{bottom} - \bar{x}_{surface})}{\sqrt{\frac{MSE}{n_{bottom}} + \frac{MSE}{n_{surface}}}} \\
T_{27} &= \frac{(6.04 - 4.02)}{\sqrt{\frac{1.38}{10} + \frac{1.38}{10}}} = \frac{2.02}{0.53} = 3.81 \\
p - value &< 0.01 \qquad \text{(two-sided)} \\
\alpha^{\star} &= 0.05/3 = 0.0167
\end{aligned}
$$

Bottom and Surface

Is there a difference between the average aldrin concentration at the bottom and at surface?

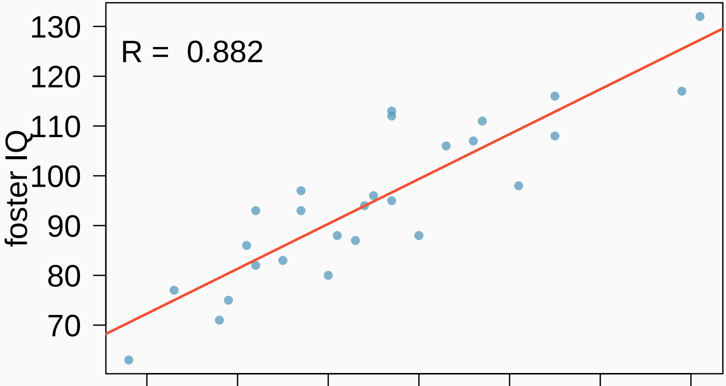
$$
\begin{aligned}
T_{df_E} &= \frac{(\bar{x}_{bottom} - \bar{x}_{surface})}{\sqrt{\frac{MSE}{n_{bottom}} + \frac{MSE}{n_{surface}}}} \\
T_{27} &= \frac{(6.04 - 4.02)}{\sqrt{\frac{1.38}{10} + \frac{1.38}{10}}} = \frac{2.02}{0.53} = 3.81 \\
p-value &< 0.01 \qquad \text{(two-sided)} \\
\alpha^{\star} &= 0.05/3 = 0.0167
\end{aligned}
$$

Reject H_0, the data provide convincing evidence of a difference between the average aldrin concentrations at bottom and surface.

Back to Linear Regression!

Nature or nurture?

In 1966 Cyril Burt published a paper called “The genetic determination of differences in intelligence: A study of monozygotic twins reared apart?” The data consist of IQ scores for [an assumed random sample of] 27 identical twins, one raised by foster parents, the other by the biological parents.



Testing for the slope

Assuming that these 27 twins comprise a representative sample of all twins separated at birth, we would like to test if these data provide convincing evidence that the IQ of the biological twin is a significant predictor of IQ of the foster twin. What are the appropriate hypotheses?

(a) $H_0 : b_0 = 0$; $H_A : b_0 \neq 0$

(b) $H_0 : \beta_0 = 0$; $H_A : \beta_0 \neq 0$

(c) $H_0 : b_1 = 0$; $H_A : b_1 \neq 0$

(d) $H_0 : \beta_1 = 0$; $H_A : \beta_1 \neq 0$

Remember: We use β_0 and β_1 for the population intercept and slope and b_0 and b_1 for our samples intercept and slope.

Testing for the slope

Assuming that these 27 twins comprise a representative sample of all twins separated at birth, we would like to test if these data provide convincing evidence that the IQ of the biological twin is a significant predictor of IQ of the foster twin. What are the appropriate hypotheses?

(a) $H_0 : b_0 = 0$; $H_A : b_0 \neq 0$

(b) $H_0 : \beta_0 = 0$; $H_A : \beta_0 \neq 0$

(c) $H_0 : b_1 = 0$; $H_A : b_1 \neq 0$

(d) $H_0 : \beta_1 = 0$; $H_A : \beta_1 \neq 0$

Remember: We use β_0 and β_1 for the population intercept and slope and b_0 and b_1 for our samples intercept and slope.

Testing for the slope (cont.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

Testing for the slope (cont.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

- We always use a t-test in inference for regression.

  Remember: *Test statistic,* $T = \frac{\textit{point estimate} - \textit{null value}}{SE}$

Testing for the slope (cont.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

- We always use a *t*-test in inference for regression.

  Remember: *Test statistic,* $T = \frac{\textit{point estimate} - \textit{null value}}{\textit{SE}}$

- Point estimate = b_1 is the observed slope.

Testing for the slope (cont.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

- We always use a *t*-test in inference for regression.

  Remember: Test statistic, $T = \frac{point\ estimate - null\ value}{SE}$
- Point estimate = b_1 is the observed slope.
- SE_{b_1} is the standard error associated with the slope.

Testing for the slope (cont.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

- We always use a t-test in inference for regression.

  Remember: *Test statistic,* $T = \frac{point\ estimate - null\ value}{SE}$
- Point estimate = b_1 is the observed slope.
- SE_{b_1} is the standard error associated with the slope.
- Degrees of freedom associated with the slope is $df = n - 2$, where n is the sample size.

Testing for the slope (cont.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

- We always use a *t*-test in inference for regression.

  Remember: *Test statistic,* $T = \frac{point\ estimate - null\ value}{SE}$

- Point estimate = b_1 is the observed slope.
- SE_{b_1} is the standard error associated with the slope.
- Degrees of freedom associated with the slope is $df = n - 2$, where n is the sample size.

  Remember: We lose 1 degree of freedom for each parameter we estimate, and in simple linear regression we estimate 2 parameters, β_0 *and* β_1.

Testing for the slope (cont.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

Testing for the slope (cont.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

$$T = \frac{0.9014 - 0}{0.0963} = 9.36$$

Testing for the slope (cont.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

$$
\begin{aligned}
T &= \frac{0.9014 - 0}{0.0963} = 9.36 \\
df &= 27 - 2 = 25
\end{aligned}
$$

Testing for the slope (cont.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

$$\begin{aligned} T &= \frac{0.9014 - 0}{0.0963} = 9.36 \\ df &= 27 - 2 = 25 \\ p - value &= P(|T| > 9.36) < 0.01 \end{aligned}$$

Confidence interval for the slope

Remember that a confidence interval is calculated as *point estimate* $\pm$ *ME* and the degrees of freedom associated with the slope in a simple linear regression is $n - 2$. Which of the below is the correct 95% confidence interval for the slope parameter? Note that the model is based on observations from 27 twins.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

(a) $9.2076 \pm 1.65 \times 9.2999$

(b) $0.9014 \pm 2.06 \times 0.0963$

(c) $0.9014 \pm 1.96 \times 0.0963$

(d) $9.2076 \pm 1.96 \times 0.0963$

Confidence interval for the slope

Remember that a confidence interval is calculated as *point estimate* $\pm$ *ME* and the degrees of freedom associated with the slope in a simple linear regression is $n - 2$. Which of the below is the correct 95% confidence interval for the slope parameter? Note that the model is based on observations from 27 twins.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

(a) $9.2076 \pm 1.65 \times 9.2999$

(b) $0.9014 \pm 2.06 \times 0.0963$

(c) $0.9014 \pm 1.96 \times 0.0963$

(d) $9.2076 \pm 1.96 \times 0.0963$

Recap

- Inference for the slope for a single-predictor linear regression model:

Recap

- Inference for the slope for a single-predictor linear regression model:
  - Hypothesis test:

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

Recap

- Inference for the slope for a single-predictor linear regression model:
  - Hypothesis test:

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

  - Confidence interval:

$$b_1 \pm t^{\star}_{df=n-2} SE_{b_1}$$

Recap

- Inference for the slope for a single-predictor linear regression model:
  - Hypothesis test:

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

  - Confidence interval:

$$b_1 \pm t^{\star}_{df=n-2} SE_{b_1}$$

- The null value is often 0 since we are usually checking for ***any*** relationship between the explanatory and the response variable.

Recap

- Inference for the slope for a single-predictor linear regression model:
  - Hypothesis test:
$$T = \frac{b_1 - \text{null value}}{SE_{b_1}} \qquad df = n - 2$$
  - Confidence interval:
$$b_1 \pm t^{\star}_{df=n-2} SE_{b_1}$$
- The null value is often 0 since we are usually checking for ***any*** relationship between the explanatory and the response variable.
- The regression output gives b_1, SE_{b_1}, and ***two-tailed*** p-value for the t-test for the slope where the null value is 0.

Recap

- Inference for the slope for a single-predictor linear regression model:
  - Hypothesis test:

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

  - Confidence interval:

$$b_1 \pm t^{\star}_{df=n-2} SE_{b_1}$$

- The null value is often 0 since we are usually checking for ***any*** relationship between the explanatory and the response variable.
- The regression output gives b_1, SE_{b_1}, and ***two-tailed*** p-value for the t-test for the slope where the null value is 0.
- We rarely do inference on the intercept, so we'll be focusing on the estimates and inference for the slope.

Caution

- Always be aware of the type of data you're working with: random sample, non-random sample, or population.

Caution

- Always be aware of the type of data you're working with: random sample, non-random sample, or population.
- Statistical inference, and the resulting p-values, are meaningless when you already have population data.

Caution

- Always be aware of the type of data you're working with: random sample, non-random sample, or population.
- Statistical inference, and the resulting p-values, are meaningless when you already have population data.
- If you have a sample that is non-random (biased), inference on the results will be unreliable.

Caution

- Always be aware of the type of data you're working with: random sample, non-random sample, or population.
- Statistical inference, and the resulting p-values, are meaningless when you already have population data.
- If you have a sample that is non-random (biased), inference on the results will be unreliable.
- The ultimate goal is to have independent observations.