THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

Inference for simple regression 1

Lecture 3

STA 371G

Announcements and logistics

- Reminder: Pretest due at Thursday 11:59 PM

Announcements and logistics

- Reminder: Pretest due at Thursday 11:59 PM
- The course help site (“Lecture slides and R help” on the left side of Canvas) now has review advice and links to review videos

Announcements and logistics

- Reminder: Pretest due at Thursday 11:59 PM
- The course help site ("Lecture slides and R help" on the left side of Canvas) now has review advice and links to review videos
- Course help site also has lecture slides and scripts, and slides, scripts, and videos from Tuesday night R help sessions

Announcements and logistics

- Reminder: Pretest due at Thursday 11:59 PM
- The course help site ("Lecture slides and R help" on the left side of Canvas) now has review advice and links to review videos
- Course help site also has lecture slides and scripts, and slides, scripts, and videos from Tuesday night R help sessions
- The "Class kickoff survey" closes tomorrow night; please complete it before then in Learning Catalytics to let me know of your previous experience (and so I can send you a free MyStatLab access code if you previously bought the 3rd edition of the book)

Measuring goodness-of-fit

- R^2 measures the fraction of the variation in Y explained by X; in our analysis from last time, $R^2 = 0.03$.

Measuring goodness-of-fit

- R^2 measures the fraction of the variation in Y explained by X; in our analysis from last time, $R^2 = 0.03$.
- The standard error of the regression s_e can be roughly interpreted as the standard deviation of the residuals.

Interpreting the standard error of the regression $s_e = 2.96$

- The residual for the *i*th case is $Y_i - \hat{Y}_i$

Interpreting the standard error of the regression

$s_e = 2.96$

- The residual for the *i*th case is $Y_i - \hat{Y}_i$
- The residuals are approximately Normally distributed

Interpreting the standard error of the regression

$s_e = 2.96$

- The residual for the *i*th case is $Y_i - \hat{Y}_i$
- The residuals are approximately Normally distributed
- The mean of the residuals is 0 (why?)

Interpreting the standard error of the regression $s_e = 2.96$

- The residual for the *i*th case is $Y_i - \hat{Y}_i$
- The residuals are approximately Normally distributed
- The mean of the residuals is 0 (why?)
- Therefore: 95% of the residuals are roughly within $\pm 2s_e$

Interpreting the standard error of the regression $s_e = 2.96$

- The residual for the ith case is $Y_i - \hat{Y}_i$
- The residuals are approximately Normally distributed
- The mean of the residuals is 0 (why?)
- Therefore: 95% of the residuals are roughly within $\pm 2s_e$
- In other words, 95% of the time I expect my prediction to be off by at most 5.93

In our regression, $R^2 = 0.03$.

Is this “significant?”

In our regression, $R^2 = 0.03$.
Is this "significant?"

- **Statistical significance:** Can we reject the null hypothesis that the correlation between X and Y in the *population* is zero?

In our regression, $R^2 = 0.03$.
Is this "significant?"

- **Statistical significance:** Can we reject the null hypothesis that the correlation between X and Y in the *population* is zero?
- **Practical significance:** Is the relationship in our sample strong enough to be meaningful?

The overall null hypothesis for a regression model

The following are equivalent ways to express the overall null hypothesis:

- $R^2 = 0$ (in the population)

The overall null hypothesis for a regression model

The following are equivalent ways to express the overall null hypothesis:

- $R^2 = 0$ (in the population)
- $\mathrm{cor}(X, Y) = 0$ (in the population)

The overall null hypothesis for a regression model

The following are equivalent ways to express the overall null hypothesis:

- $R^2 = 0$ (in the population)
- $\text{cor}(X, Y) = 0$ (in the population)
- $\beta_1 = 0$

The overall null hypothesis for a regression model

The following are equivalent ways to express the overall null hypothesis:

- $R^2 = 0$ (in the population)
- $\text{cor}(X, Y) = 0$ (in the population)
- $\beta_1 = 0$
- The model has no predictive power

The overall null hypothesis for a regression model

The following are equivalent ways to express the overall null hypothesis:

- $R^2 = 0$ (in the population)
- $\text{cor}(X, Y) = 0$ (in the population)
- $\beta_1 = 0$
- The model has no predictive power
- Predictions from this model are no better than predicting $\overline{Y}$ for every case

Two ways to test the overall null hypothesis

- The *F*-test (tests $H_0 : R^2 = 0$ in the population vs $H_A : R^2 \neq 0$)
- The *t*-test for the *slope* (β_1) coefficient (tests $H_0 : \beta_1 = 0$ vs $H_A : \beta_1 \neq 0$)
- Note that both tests are two-tailed, since we would care about the null hypothesis being wrong in either direction (i.e. $\beta_1 > 0$ and $\beta_1 < 0$ are both of interest)

Two ways to test the overall null hypothesis

- The F-test (tests $H_0 : R^2 = 0$ in the population vs $H_A : R^2 \neq 0$)
- The t-test for the *slope* (β_1) coefficient (tests $H_0 : \beta_1 = 0$ vs $H_A : \beta_1 \neq 0$)
- Note that both tests are two-tailed, since we would care about the null hypothesis being wrong in either direction (i.e. $\beta_1 > 0$ and $\beta_1 < 0$ are both of interest)

Both of these methods are equivalent; the p-values will be exactly the same!

```
> model <- lm(num.drinks ~ age)
> summary(model)



Call:
lm(formula = num.drinks ~ age)

Residuals:
   Min     1Q Median     3Q    Max
-4.204 -1.853 -0.853  0.810 15.160

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.5542     0.2653    24.7   <2e-16 ***
age          -0.1688     0.0159   -10.6   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3 on 3600 degrees of freedom
  (2902 observations deleted due to missingness)
Multiple R-squared:  0.0304,Adjusted R-squared:  0.0302
F-statistic:  113 on 1 and 3600 DF,  p-value: <2e-16
```

What is our conclusion about β_1?

- There is a **statistically significant** relationship between the age someone starts drinking and how much they drink as an adult.

What is our conclusion about β_1?

- There is a **statistically significant** relationship between the age someone starts drinking and how much they drink as an adult.
- Or: People that start drinking earlier in life consume **significantly more** alcohol when they drink as adults.

What is our conclusion about β_1?

- There is a **statistically significant** relationship between the age someone starts drinking and how much they drink as an adult.
- Or: People that start drinking earlier in life consume **significantly more** alcohol when they drink as adults.
- Each additional year you wait to start drinking is associated with consuming 0.17 fewer drinks as an adult.

What is our conclusion about β_1?

- There is a **statistically significant** relationship between the age someone starts drinking and how much they drink as an adult.
- Or: People that start drinking earlier in life consume **significantly more** alcohol when they drink as adults.
- Each additional year you wait to start drinking is associated with consuming 0.17 fewer drinks as an adult.
- Is this relationship **practically significant**?

Practical significance

- To assess **statistical significance**, we look at the p-value
- To assess **practical significance**:
  - We only consider it if we already have statistical significance (why?)
  - Look at R^2, the standard error of the regression, and the magnitude of the coefficients
  - It's ultimately a judgement call!

Put a confidence interval on it

- Our best estimate for the *effect* of a year's postponement of drinking is 0.17 fewer drinks as an adult

Put a confidence interval on it

- Our best estimate for the *effect* of a year's postponement of drinking is 0.17 fewer drinks as an adult
- We can use a confidence interval to give a range of plausible values for what this effect size is in the population

Put a confidence interval on it

A confidence interval is always of the form

$$\text{estimate} \pm (\text{critical value})(\text{standard error}).$$

Put a confidence interval on it

A confidence interval is always of the form

$$\text{estimate} \pm (\text{critical value})(\text{standard error}).$$

Recall that the critical value for a 95% confidence interval is the cutoff value that cuts off 95% of the area in the middle of the distribution; the sampling distribution of $\hat{\beta}_1$ is a t-distribution.

```
> n <- nobs(model)
> qt(0.975, n-2)

[1] 1.960623
```

Put a confidence interval on it

R will also calculate confidence intervals for us:

```
> confint(model)

                 2.5 %     97.5 %
(Intercept)  6.0339847  7.0743549
age         -0.1999713 -0.1376959
```

Put a confidence interval on it

R will also calculate confidence intervals for us:

```
> confint(model)

                 2.5 %     97.5 %
(Intercept)  6.0339847  7.0743549
age         -0.1999713 -0.1376959
```

In other words, we are 95% confident that the effect of each additional year's delay in starting to drink is between 0.14 and 0.2.

Put a confidence interval on it, part 2

We can also put a confidence interval on a prediction!

Two kinds of intervals:

Confidence	Predicting the mean value of Y for a particular X.	Among all people that start drinking at age 21, how many drinks do have on average as adults?
Prediction	Predicting Y for a single new case.	If Bob started drinking at age 21, how many drinks do we think will have as an adult?

Put a confidence interval on it, part 2

```
> predict(model, list(age=21),
+   interval='confidence')

       fit      lwr      upr
1 3.008664 2.83616 3.181167

> predict(model, list(age=21),
+   interval='prediction')

       fit       lwr      upr
1 3.008664 -2.802894 8.820221
```

Put a confidence interval on it, part 2

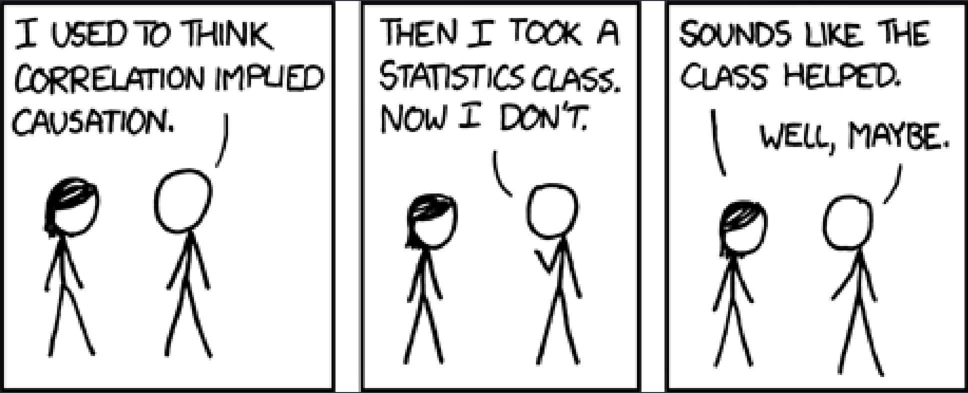
```
> predict(model, list(age=21),
+   interval='confidence')

       fit      lwr      upr
1 3.008664 2.83616 3.181167

> predict(model, list(age=21),
+   interval='prediction')

       fit       lwr      upr
1 3.008664 -2.802894 8.820221
```

Why is the prediction interval wider?


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Correlation $\neq$ Causation

Because the *p*-value is small, we can be highly confident that there is a relationship in the population between age of first drink and number of drinks consumed as an adult.

This could be because:

Correlation $\neq$ Causation

Because the *p*-value is small, we can be highly confident that there is a relationship in the population between age of first drink and number of drinks consumed as an adult.

This could be because:

- Starting to drink earlier causes you to drink more as an adult.

Correlation $\neq$ Causation

Because the p-value is small, we can be highly confident that there is a relationship in the population between age of first drink and number of drinks consumed as an adult.

This could be because:

- Starting to drink earlier causes you to drink more as an adult.
- Being predisposed to drink more will cause you to start drinking sooner.

Correlation $\neq$ Causation

Because the *p*-value is small, we can be highly confident that there is a relationship in the population between age of first drink and number of drinks consumed as an adult.

This could be because:

- Starting to drink earlier causes you to drink more as an adult.
- Being predisposed to drink more will cause you to start drinking sooner.
- There is a third ("lurking") variable that causes both early drinking and drinking more as an adult.

Correlation $\neq$ Causation

Because the *p*-value is small, we can be highly confident that there is a relationship in the population between age of first drink and number of drinks consumed as an adult.

This could be because:

- Starting to drink earlier causes you to drink more as an adult.
- Being predisposed to drink more will cause you to start drinking sooner.
- There is a third (“lurking”) variable that causes both early drinking and drinking more as an adult.

Correlation $\neq$ Causation

Because the *p*-value is small, we can be highly confident that there is a relationship in the population between age of first drink and number of drinks consumed as an adult.

This could be because:

- Starting to drink earlier causes you to drink more as an adult.
- Being predisposed to drink more will cause you to start drinking sooner.
- There is a third ("lurking") variable that causes both early drinking and drinking more as an adult.

We can't tell just by looking at this data set!


EXCLUSIVE | THE ATHLETICISM OF LEBRON + THE DRIVE OF KOBE = RUSSELL WESTBROOK
BY LEE JENKINS P. 38
Sports Illustrated
APRIL 6, 2015
SI.COM
@SINOW
FROM THE BRINK. TO THE BRINK.
KENTUCKY
CLOSES IN ON...
40-0
Karl-Anthony Towns leads Kentucky's quest for college hoops' first undefeated season since 1976.

Regression to the mean

The value of Y will tend to be closer to the mean than X, on average (even if you switch X and Y!).

- Students who take the SAT again after getting a very low score tend to improve even if they don't receive any coaching
- Children of tall parents tend to be tall, but not as tall as their parents
- Olympic champions tend to have poorer performances following their Olympic victories
- The "*Sports Illustrated* curse" of athletes that appear on the cover

Regression to the mean

- In this graph, variables have been standardized, $r = 0.8$, and $\hat{Y} = 0.8X$
- Parent at mean $\longrightarrow$ predict child height is also at the mean
- Parent 1 SD above mean $\longrightarrow$ predict child height 0.8 SD above the mean
- Parent 2 SD above mean $\longrightarrow$ predict child height 1.6 SD above the mean