## Announcements

- Office hours today 1:30-2:30 in Weniger 273

- Midterm Exam Replacement Policy
    - If you perform better on the final exam than you did on the midterm I will replace your midterm grade with your final exam grade.
    - I will not replace your final exam with your midterm grade.

- No class on Thursday

- Extra credit activity will be available on Thursday, only. Please see the Canvas announcement I posted yesterday for more details.

*Week 9*

# Simple Linear Regression

ST 314
Introduction to Statistics for Engineers

Oregon State University

# Bivariate Relationships

Bivariate data looks at the relationship between two quantitative variables. Each unit has a measurement for both variables so the data are paired.

# Explanatory and Response Variables

| Variable Types for Bivariate Relationships | |
|---|---|
| Explanatory | Response |
| ❖ May help explain or predict changes in the response<br>❖ Quantitative<br>❖ Sometimes referred to as<br>❖ independent<br>❖ predictor<br>❖ x | ❖ Variable to be estimated/predicted<br>❖ Quantitative<br>❖ Sometimes referred to as<br>❖ dependent<br>❖ y |

# Interpreting Scatterplots

| | |
|---|---|
| Strength | weak, strong, or somewhere in between |
| Direction | positive or negative |
| Form | linear or nonlinear |
| Outliers | observations that fall outside the pattern of the data |

There is a moderately strong, positive, linear relationship between high school GPA and freshman year college GPA.

It appears there is an outlier at 3.3 and 3.6, where a student performed better in college than what we might expect.

# Strength

Perfect
Response
-10
-14
-18
10
12
14
16
18
20
Explanatory

Moderately Strong
Response
22
18
14
10
10
12
14
16
18
20
Explanatory

Moderately Weak
Response
25
20
15
10
10
12
14
16
18
20
Explanatory

No Relationship
Response
22
18
14
10
10
12
14
16
18
20
Explanatory

# Direction

Positive
Response
22
18
14
12
14
16
18
20
Explanatory
Negative
Response
28
24
20
32
34
36
38
Explanatory

# Form

Linear
-10
-14
-18
10
12
14
16
18
20

Non-Linear
50
30
10
0
1.0
1.5
2.0
2.5
3.0
3.5
4.0

# Outliers

# Measuring Strength with Correlation

The correlation coefficient, $R$ , measures the strength of a *linear* association between *two quantitative* variables.

$$R = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

$R = 0.82$

# Measuring Strength with Correlation

# Least Squares Regression Line

- A common goal in statistics is to model the relationship between two quantitative variables.
- The model that best fits the linear relationship is the **least squares regression line**:

$$\hat{y} = b_0 + b_1 x$$

predicted response

y-intercept

slope

explanatory variable

# Least Squares Regression Line

- What does "least squares" mean?
  - The LSRL is calculated by finding the smallest value for the **sum of squared error,** SSE, the sum of squared differences between the observed and predicted value of *y*.

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- The following are the outcomes for minimizing SSE:
  - $b_0 = \bar{y} - b_1 \bar{x}$
  - $b_1 = R\left(\frac{S_y}{S_x}\right)$

standard dev. of response
standard dev. of explanatory
correlation coefficient

# Interpreting the LSRL

$$\hat{y} = 0.69 + 0.78x$$

- Interpreting the intercept:
  - Expected value of y when $x=0$
  - Not always necessary to interpret
  - Only need to interpret this value if it makes sense for $x=0$
- Interpreting the slope:
  - For a 1 unit increase in $x$, we expect y to change by a value $b_1$
  - $b_1>0$, expect y to increase with $x$
  - $b_1<0$, expect y to decrease as $x$ increases
  - $b_1=0$, expect no change in y with $x$

# Calculating the LSRL in R

| Function | Function Values | What does it do? |
|---|---|---|
| `mod <- lm(y ~ x)` | `mod` saves the LSRL model in an object called mod<br>`x` = vector containing the explanatory variable values<br>`y` = vector containing the response variable values | Computes the linear model for the explanatory and response variable relationship |
| `summary(mod)` | `mod` call the saved model stored above | Gives the LSRL estimates, standard errors and performs a t test on the slope. Provides $R^2$ and adjusted $R^2$ value and results from a model utility F test. |

```
HSGPA =   c(3.4, 3.1, 3.7,3.3, 3.5, 4, 3.2, 3.8, 3.9, 3.6)
CollegeGPA = c(3.1,3,3.6,3.6,3.4,3.9,3.2,3.7,3.6, 3.3)

# Use the lm() command to get the linear model.
# Take the summary() of lm() to get more information.
mod = lm(CollegeGPA~HSGPA)
summary(mod)
```

# Reading Statistical Software Output

**R Output for Linear Regression**

```
lm(formula = CollegeGPA ~ HSGPA)

Residuals:
     Min       1Q   Median       3Q      Max
-0.22364 -0.10636  0.01515  0.06045  0.35394

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.6861     0.6897   0.995  0.34902
HSGPA         0.7758     0.1937   4.006  0.00392 **

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1759 on 8 degrees of freedom
Multiple R-squared:  0.6673,    Adjusted R-squared:  0.6257
F-statistic: 16.05 on 1 and 8 DF,  p-value: 0.003918
```

$\hat{y} = 0.6861 + 0.7758x$

predicted college GPA

HSGPA

# Basic Predictions from LSRL

Let's use the LSRL to predict the freshman year college GPA of a student who earned a 3.6 high school GPA.

$$\hat{y} = 0.69 + 0.78x$$

$$x = 3.6$$

$$\hat{y} = 0.69 + 0.78(3.6) = 3.498$$

# Residuals

- The difference between an observed value of the response and the value predicted by the regression equation for the same value of $x$ is called the **residual.**

Residual = observed − Predicted

$$e_i = y_i - \hat{y}_i$$

# Calculating Residuals Example

Calculate the residual for the student with the HS GPA and College GPA of 3.6 and 3.3, respectively, given the regression line $\hat{y} = 0.69 + 0.78x$.

$x_i = 3.6$

$y_i = 3.3$

$\hat{y}_i = 0.69 + 0.78(3.6) = 3.498$

$e_i = y_i - \hat{y}_i = 3.3 - 3.498 = -0.198$

# Residual Plot

| HS GPA | College GPA | Fitted Value | Residual |
|---|---|---|---|
| **3.4** | 3.1 | 3.342 | -0.242 |
| **3.1** | 3 | 3.108 | -0.108 |
| **3.7** | 3.6 | 3.576 | 0.024 |
| **3.3** | 3.6 | 3.264 | 0.336 |
| **3.5** | 3.4 | 3.42 | -0.02 |
| **4.0** | 3.9 | 3.81 | 0.09 |
| **3.2** | 3.2 | 3.186 | 0.014 |
| **3.8** | 3.7 | 3.654 | 0.046 |
| **3.9** | 3.6 | 3.732 | -0.132 |
| **3.6** | 3.3 | 3.498 | -0.198 |

- A residual plot is a scatterplot of all $n$ residuals.
- Used to assess:
  - fit of the regression line
  - conditions for inference using a LSRL
- It is desirable to see **no patterns** in the residual plot

# Conditions for Inference

If we want to make inference from sampled data, we need our data to be obtained from a random sample and the following conditions to be met. We can check the conditions from a **residual plot!**

1. Linear relationship

3. Response varies normally around the regression line

2. Observations are independent

4. constant variation in response

# Interpreting Residual Plots

| Condition | How to check |
|---|---|
| The relationship is linear in the population. | No "U" shape or curve in residual plot |
| The response varies normally about the population regression line. | Want random scatter above and below the 0 line in residual plot |
| Observations are independent. | Assess by looking sampling scheme<br>No pattern in residuals |
| The standard deviation of the responses is the same for all values of x. | Want no "funnel" shape in residuals |

# Interpreting Residual Plots

When we see these patterns in our residuals the least squares regression line is no longer reliable to make predictions about the data!

# Checking Conditions for Inference Example

Given the sample of 10 students was obtained randomly. Based on the residuals are the conditions for inference satisfied, to make inference from our sampled data to the population?