More about Regression*

Day 3, Afternoon

**Some of these power point slides are courtesy of Brooks-Cole, accompanying Mind On Statistics by Utts & Heckard.*

Making Inferences

1. Does the **observed relationship** in a sample **also occur in the population**?
2. For a linear relationship, what is the **slope** of the regression line in the population?
3. What is the **mean value of the response** variable (y) for cases with a specific value of the explanatory variable (x)?
4. What **interval of values predicts** an individual value of the response variable (y) for a case with a specific value of the explanatory variable (x)?

Sample and Population Regression Models

- If the sample represents a larger population, we need to distinguish between the **regression line for the sample** and the **regression line for the population.**
- The observed data can be used to determine the regression line for the sample, but the ***regression line for the population can only be imagined***.

Regression Line for the Sample

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

$\hat{y}$ is spoken as "***y*-hat**," and it is also referred to either as ***predicted*** y or ***estimated*** y.

b_0 is the **intercept** of the straight line. The intercept is the value of y when $x = 0$.

b_1 is the **slope** of the straight line. The slope tells us how much of an increase (or decrease) there is for the y variable when the x variable increases by one unit. The sign of the slope tells us whether y increases or decreases when x increases.

Deviations from the Regression Line in the Sample

For an observation y_i in the sample, the **residual** is:

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

y_i = value of response variable for i^{th} obs.

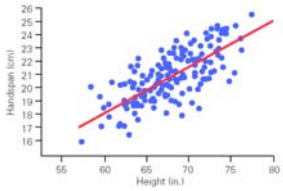
$\hat{y} = b_0 + b_1 x$, where x_i is the value of the explanatory variable for the observation.

Example: *Height and handspans of students*

Data: Heights (in inches) and Handspans (in centimeters) of 167 college students.

Regression equation: Handspan = -3 + 0.35 Height

Slope = 0.35 => Handspan increases by 0.35 cm, **on average**, for each increase of 1 inch in height.



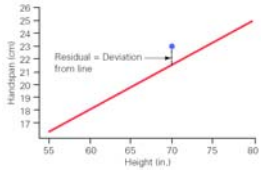
Example, continued

Consider a person 70 inches tall
whose handspan is 23 centimeters.

The sample regression line is $\hat{y} = -3 + 0.35x$

so $\hat{y} = -3 + 0.35(70) = 21.5$ cm for this person.

The *residual* =
observed y – predicted y
= 23 – 21.5 = 1.5 cm.



Regression Line for the Population

$$E(Y) = \beta_0 + \beta_1 x$$

$E(Y)$ represents the mean or *expected value* of y for cases in the population that all have the same x.

β_0 is the **intercept** of the straight line in the **population**.

β_1 is the **slope** of the straight line in the **population**. Note that if the population slope were 0, there is no linear relationship in the population.

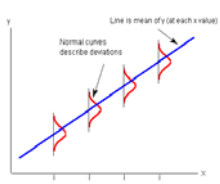
These population parameters are estimated using the corresponding statistics.

Assumptions about Deviations

1. Assume the general size of the deviations of y values from the line is the same for all values of the explanatory variable (x) – called the **constant variance** assumption.

2. For any x, the distribution of y values is normal

=> Deviations from the population regression line have a normal distribution. (This can be relaxed if n is large)



Simple Regression Model for a Population

y = Mean + Deviation

1. **Mean** in the population is the line $E(Y) = \beta_0 + \beta_1 x$ if the relationship is linear.
2. Individual case's **deviation = y − mean,** which is what is left unexplained after accounting for the mean y value at that case's x value.

Estimating the Standard Deviation around the Line

The **standard deviation for regression** measures …

- roughly, the average deviation of y values from the mean (the regression line).
- the general size of the residuals.

$$s = \sqrt{\frac{\text{Sum of Squared Residuals}}{n-2}}$$

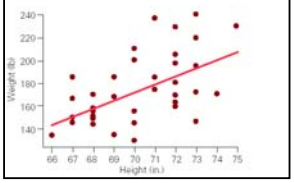
$$= \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$$

Example: *Height and Weight*

The regression equation is
Weight = − 318 + 7.00 Height

Predictor	Coef	SE Coef	T	P
Constant	−317.9	110.9	−2.87	0.007
Height	6.996	1.581	4.42	0.000

S = 24.00 R-Sq = 32.3% R-Sq(adj) = 30.7%



Data:
x = heights (in inches)
y = weight (pounds)
of n = 43 male students.

Standard deviation s = 24.00 (pounds):
Roughly measures, for any given height, the general size of the deviations of individual weights from the mean weight for the height.

Proportion of Variation Explained

Squared correlation r^2 is between 0 and 1 and indicates the **proportion of variation in the response explained by x**.

SSTO = sum of squares total = sum of squared differences between observed y values and $\bar{y}$.

SSE = sum of squared errors (residuals) = sum of squared differences between observed y values and predicted values based on least squares line.

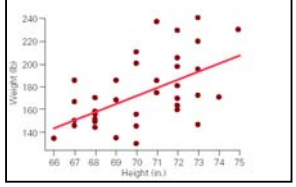
$$r^2 = \frac{SSTO - SSE}{SSTO}$$

Example: *Height and Weight, continued*

The regression equation is
Weight = − 318 + 7.00 Height

Predictor	Coef	SE Coef	T	P
Constant	−317.9	110.9	−2.87	0.007
Height	6.996	1.581	4.42	0.000

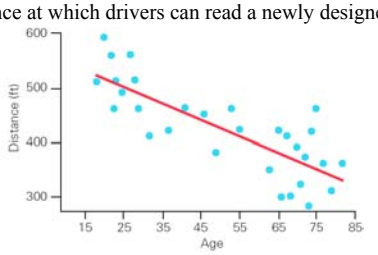
S = 24.00 R-Sq = 32.3% R-Sq(adj) = 30.7%



R-Sq = 32.3% => The variable height explains 32.3% of the variation in the weights of college men.

Example: *Driver Age and Maximum Legibility Distance of Highway Signs*

Study to examine relationship between **age** and maximum distance at which drivers can read a newly designed sign.



Average Distance = 577 – 3.01 × Age

Example: *Age and Distance, continued*

s = 49.76 and R-sq = 64.2% => Average distance from regression line is about 50 feet, and 64.2% of the variation in sign reading distances is explained by age.

$SSE = 69334$

$SSTO = 193667$

$$s = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{69334}{28}} = 49.76$$

$$r^2 = \frac{SSTO - SSE}{SSTO} = \frac{193667 - 69334}{193667} = .642$$

The regression equation is
Distance = 577 − 3.01 Age

Predictor	Coef	SE Coef	T	P
Constant	576.68	23.47	24.57	0.000
Age	**−3.0068**	**0.4243**	**−7.09**	**0.000**

S = 49.76 R-Sq = 64.2% R-Sq (adj) = 62.9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	124333	124333	50.21	0.000
Residual Error	28	69334	2476		
Total	29	193667			

Unusual Observations

Obs	Age	Distance	Fit	SE Fit	Residual	St Resid
27	75.0	460.00	351.17	13.65	108.83	2.27R

R denotes an observation with a large standardized residual

Hands-On Activity:
To be given in class

Applet to try to find least squares line (maximize R^2 and minimize MSE = SSE/n – 2)

http://onlinestatbook.com/stat_sim/reg_by_eye/index.html

Inference About Linear Regression Relationship

The **statistical significance** of a linear relationship can be evaluated by testing **whether or not the slope is 0**.

H_0: $\beta_1 = 0$ (the population slope is 0, so y and x are *not linearly related.*)

H_a: $\beta_1 \neq 0$ (the population slope is not 0, so y and x *are linearly related.*)

Alternative may be one-sided or two-sided.

Test for Zero Slope

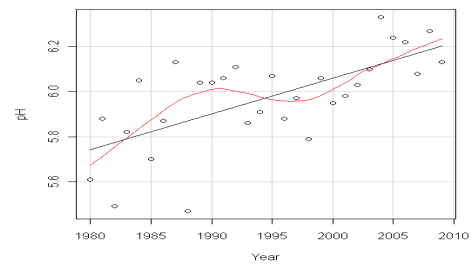
$$t = \frac{\text{Sample statistic} - \text{Null value}}{\text{Standard error}} = \frac{b_1 - 0}{s.e.(b_1)}$$

$$b_1 = r\frac{s_y}{s_x}$$

$$s.e.(b_1) = \frac{s}{\sqrt{\sum(x-\bar{x})^2}} \qquad \text{where } s = \sqrt{\frac{SSE}{n-2}}$$

Under the null hypothesis, this t statistic follows a t-distribution with **df = *n* – 2**.

Example: Is pH in Davis rainfall changing over time?



R Commander

Statistics → Fit model → Linear regression
Specify x (explanatory) and y (response = pH)

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.39811 -0.09337  0.00545  0.11777  0.27777

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -25.701060   6.719198  -3.825  0.00067 ***
Year          0.015880   0.003369   4.714 6.06e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 0.1597 on 28 degrees of freedom
Multiple R-squared: 0.4424, Adjusted R-squared: 0.4225
```


Example *Year and pH for Davis*

H_0: $\beta_1 = 0$ (y and x are *not linearly related.*)

H_a: $\beta_1 \neq 0$ (y and x *are linearly related.*)

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -25.701060   6.719198  -3.825  0.00067 ***
Year          0.015880   0.003369   4.714 6.06e-05 ***
```

Probability is close to 0 that observed slope could be as far from 0 or farther if there is no linear relationship in population (p-value shown in box) => Appears the relationship in the sample represents a real relationship in the population. So conclude that pH actually is increasing over time.

Confidence Interval for the Slope

A Confidence Interval for a Population Slope

$$b_1 \pm t^* \times s.e.(b_1) \Rightarrow b_1 \pm t^* \times \frac{s}{\sqrt{\sum(x-\bar{x})^2}}$$

where the **multiplier *t**** is the value in a t-distribution with degrees of freedom = df = n - 2 such that the area between -t* and t* equals the desired confidence level.

Testing Hypotheses about the Correlation Coefficient

The **statistical significance** of a linear relationship can be evaluated by testing **whether or not the correlation between *x* and *y* in the population is 0**.

H_0: $\rho = 0$ (x and y are *not correlated.*)

H_a: $\rho \neq 0$ (x and y are *correlated.*)

where ρ represents the population correlation

Results for this test will be the ***same*** as for the test of whether or not the population slope is 0.

Effect of Sample Size on Significance

With **very large sample sizes**, weak relationships
with **low correlation** values
can be **statistically significant**.

> ***Moral***: With a large sample size, saying two variables are significantly related may only mean the correlation is not precisely 0.
>
> We should carefully examine the observed strength of the relationship, the value of *r*.

Predicting for an Individual

A **95% prediction interval** estimates the value of *y* for an individual case with a particular value of *x*.

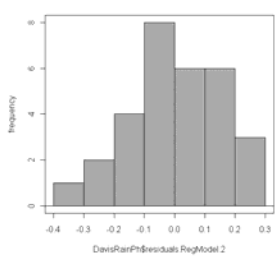
This interval can be ***interpreted*** in two equivalent ways:

1. It estimates the central 95% of the values of *y* for cases in a population with specified value of *x*.
2. Probability is .95 that a randomly selected case from population with a specified value of *x* falls into the 95% prediction interval.

R Commander: Storing residuals and predicted values

Models → Add observation statistics to data → Check "fitted values" and "residuals" to store these in the data set.

Histogram of residuals for pH example:



Prediction Interval

$$\hat{y} \pm t^* \sqrt{s^2 + [s.e.(fit)]^2}$$

where $s.e.(fit) = s\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum (x_i - \bar{x})^2}}$

Note:

- *t** found from t-distribution with df = *n* – 2.
- Width of interval depends upon how far the specified *x* value is from $\bar{x}$ (the further, the wider).
- When *n* is large, s.e.(fit) will be small, and prediction interval will be approximately … $\hat{y} \pm t^* s$

Estimating the Mean at given x

A **95% confidence interval for the mean** estimates the ***mean*** value of the response variable *y*, *E*(*Y*), for (all) cases with a particular value of *x*.

$$\hat{y} \pm t^* \times s.e.(fit)$$

where $s.e.(fit) = s\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum (x_i - \bar{x})^2}}$

*t** found from t-distribution with df = *n* – 2.

Checking Conditions for Regression Inference

Conditions:

1. **Form** of the equation that links the mean value of *y* to *x* **must be correct**.
2. **No extreme outliers** that influence the results unduly.
3. **Standard deviation** of values of *y* from the mean *y* is **same** regardless of value of *x*.
4. For cases in the population with same value of *x*, the distribution of *y* is a **normal distribution**. Equivalently, the distribution of deviations from the mean value of *y* is a normal distribution. This can be relaxed if the *n* is large.
5. **Observations** in the sample are **independent** of each other.

Hands-On Activity:
To be given in class

How outliers influence regression
http://illuminations.nctm.org/LessonDetail.aspx?ID=L455

Checking Conditions with Plots

Conditions 1, 2 and 3 checked using two plots:

Scatterplot of *y* versus *x* for the sample

Scatterplot of the residuals versus *x* for the sample

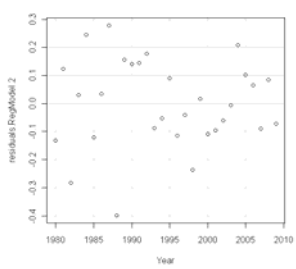
If Condition 1 holds for a linear relationship, then:
Plot of *y* versus *x* should show points randomly scattered around an imaginary straight line.
Plot of residuals versus *x* should show points randomly scattered around a horizontal line at residual 0.

If Condition 2 holds, extreme outliers should not be evident in either plot.

If Condition 3 holds, neither plot should show increasing or decreasing spread in the points as *x* increases.

Example: *Residuals vs Year for pH*

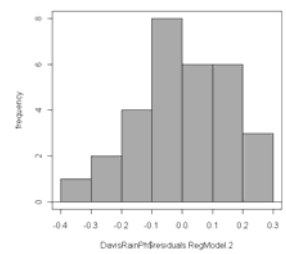
Residual plot:
Is a somewhat random-looking blob of points => linear model ok.
A few possible outliers?
Spread looks somewhat constant across years.



Conditions 4 and 5

Condition 4:
examine histogram or normal probability plot of the residuals

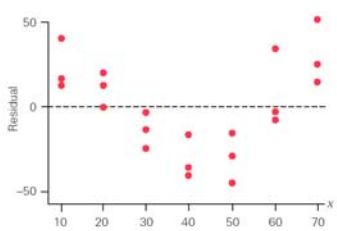
Histogram:
Residuals are approximately normally distributed



Condition 5: follows from the data collection process. Units must be measured independently. Is pH of rainfall across years independent?? Perhaps consider time series models.

When Conditions Are Not Met

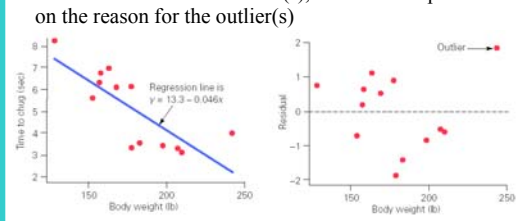
Condition 1 not met: use a more complicated model



Based on this residual plot, a curvilinear model, such as the quadratic model, may be more appropriate.

When Conditions Are Not Met

Condition 2 not met: if outlier(s), correction depends on the reason for the outlier(s)



Outlier is legitimate. Relationship appears to change for body weights over 210 pounds. Could remove outlier and use the linear regression relationship only for body weights under about 210 pounds.

When Conditions Are Not Met

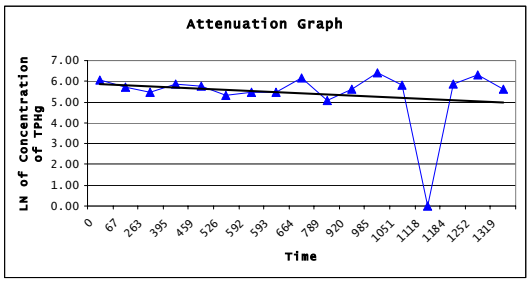
Either Condition 1 or 3 not met:
A **transformation** may be required. (Equivalent to using a different model.) Often the same transformation will help correct more than one condition.

Common transformation is the ***natural log*** of *y*.

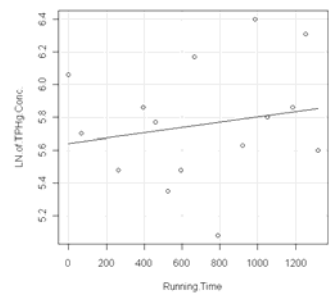
Example from Jim Tischler

- Trend analysis for monitoring cleanup of TPH (total petroleum hydrocarbons) gasoline
- Data (log of TPHg concentration) used to predict a 7.7 year time frame to achieve water quality objectives
- However, there is one non-detect that was replaced with a 0.
- See plots on next few slides.

Appears to be a decreasing trend across time; DL = log(50) = 3.91, non-detect replaced with 0.



Plot with the non-detect removed – increasing trend!



Regression results (not a significant trend in either case)

With 0 included:

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.9336117  0.7308107   8.119 7.17e-07
Running.Time -0.0007261  0.0008939  -0.812    0.429
```

Without 0 included:

```
Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.6417630  0.1834230   30.76 2.96e-14
Running.Time 0.0001593  0.0002307    0.69    0.501
```


Cautions about Regression

- Always look at plots of:
  - ✓*x* (horizontal axis) versus *y* (vertical axis)
  - ✓*x* versus residuals
  - ✓other possible explanatory variables versus residuals
- Methods that take dependence over time into account may be more appropriate when the explanatory variables is time.

Debriefing: Suggestions for future offerings of the course