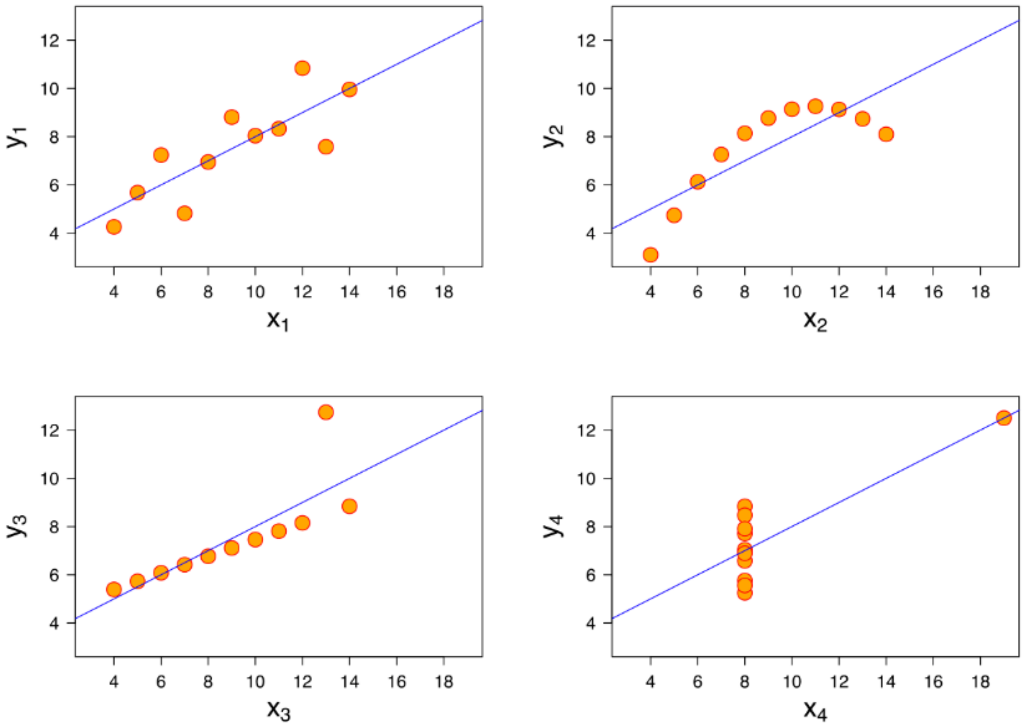
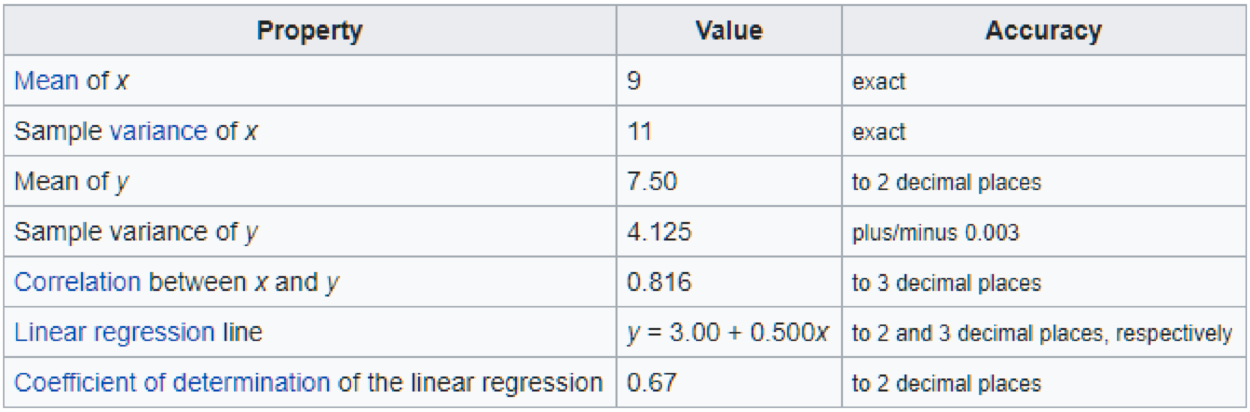
Session Agenda

- Anscombe’s Quartet
- Correlation
  - Pearson Product Moment
  - Testing hypotheses
- Simple Linear Regression
  - Basic Facts
  - Example
  - Coding in R
- Model Specification
- Confounding
  - ANOVA
  - Regression
- George E. P. Box Quote
- Review Problems
- Final Exam

Anscombe’s Quartet

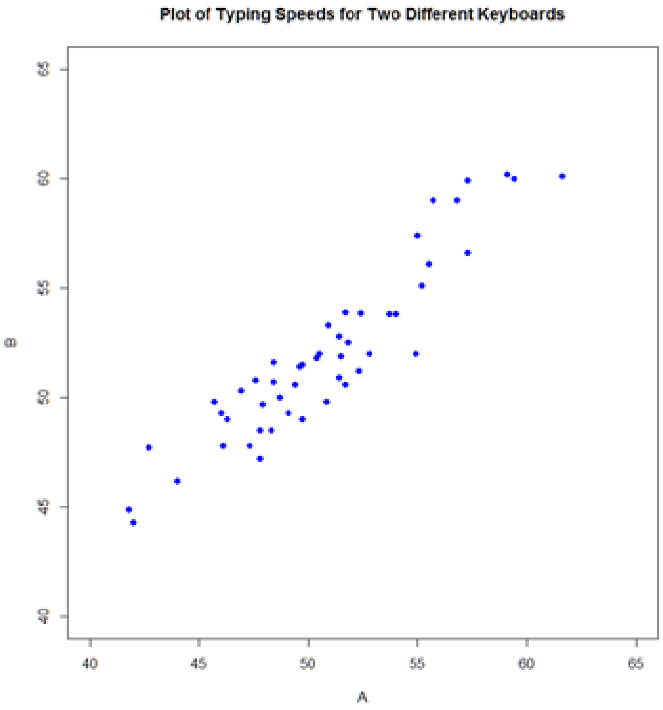
Beware of relying only on simple descriptive statistics!





Property	Value	Accuracy
Mean of *x*	9	exact
Sample variance of *x*	11	exact
Mean of *y*	7.50	to 2 decimal places
Sample variance of *y*	4.125	plus/minus 0.003
Correlation between *x* and *y*	0.816	to 3 decimal places
Linear regression line	*y* = 3.00 + 0.500*x*	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression	0.67	to 2 decimal places

Testing the Hypothesis of a Zero Correlation



```
> r <- cor(A,B)
> r
[1] 0.9325682
> n <- length(A)
> T <- r*sqrt((n-2)/(1-r^2))
> T
[1] 17.71045
> qt(0.95, n-2, lower.tail = TRUE)
[1] 1.677927
> pt(T, n-2, lower.tail = FALSE)
[1] 9.863848e-23
>
> cor.test(A, B, alternative = c("greater"), method = c("pearson"), conf.level = 0.95)

        Pearson's product-moment correlation

data:  A and B
t = 17.71, df = 47, p-value < 2.2e-16
alternative hypothesis: true correlation is greater than 0
95 percent confidence interval:
 0.8927321 1.0000000
sample estimates:
      cor
0.9325682
```

Wilcox, *Basic Statistics*, pages 173-174

Simple Linear Regression Model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Where-

y	**dependent variable,**
x	**independent variable,**
β_0 , β_1	**unknown constants**
ε	**random error term.**

The *method of least squares* is used to minimize the sum of squares:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

This provides a fitted equation:

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

The differences between observed values for y and predicted values are called residuals. Residuals play an important role in diagnosing model adequacy or "model fit".

A basic assumption of simple linear regression is that the random error term has a normal distribution with mean zero and constant variance for all observations. If investigation of the residuals reveals this is not true, the model must be changed.

Simple Linear Regression Estimators

Linear regression model estimators can be expressed in matrix terms. This is what the word “linear” denotes.

Differentiate the sum of squares for each parameter.

$$\frac{dS}{d\beta_0} = -2\left[\sum y_i - \beta_0 n - \beta_1 \sum x_i\right]$$

$$\frac{dS}{d\beta_1} = -2\left[\sum x_i y_i - \beta_0 \sum x_i - \beta_1 \sum x_i^2\right]$$

Set equal to zero and form the Normal Equations:

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

$$\sum x_i y_i / n = b_0 \bar{x} + b_1 \sum x_i^2 / n$$

Express in matrix algebra:

$$\begin{pmatrix} 1 & \bar{x} \\ \bar{x} & \sum x_i^2 / n \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{pmatrix} \bar{y} \\ \sum x_i y_i / n \end{pmatrix}$$

Inverting and solving gives the estimators:

$$\begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \left(\frac{1}{(\sum x_i^2 / n) - \bar{x}^2} \right) \begin{pmatrix} \sum x_i^2 / n & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \begin{pmatrix} \bar{y} \\ \sum x_i y_i / n \end{pmatrix}$$

Some Results

Fundamental Identity—

$$\sum (y-\bar{y})^2 = \sum (\hat{y}-\bar{y})^2 + \sum (y-\hat{y})^2 \,.$$

This states that the total variation about the mean of *y* equals the total variation of the predicted values of *y* about the mean of *y* plus the total variation of the residuals.

Coefficient of Determination—

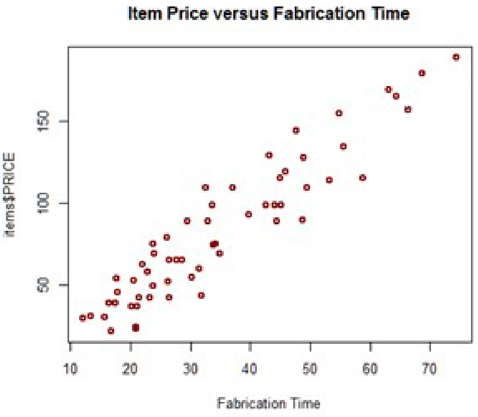
$$r^2 = \frac{\mathrm{exp}\,lained\ \mathrm{var}\ iation}{total\ \mathrm{var}\ iation} = \frac{\sum (\hat{y}-\bar{y})^2}{\sum (y-\bar{y})^2} \,.$$

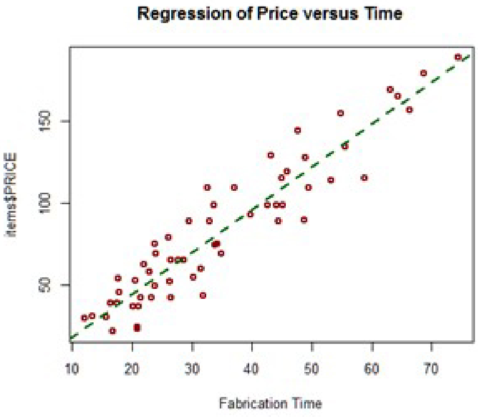
For a simple linear regression model, r^2 is the square of the Pearson Product Moment Correlation Coefficient. For a multiple linear regression model, the coefficient of multiple determination R^2 represents the proportion of variation of the dependent variable accounted for by the independent variables.

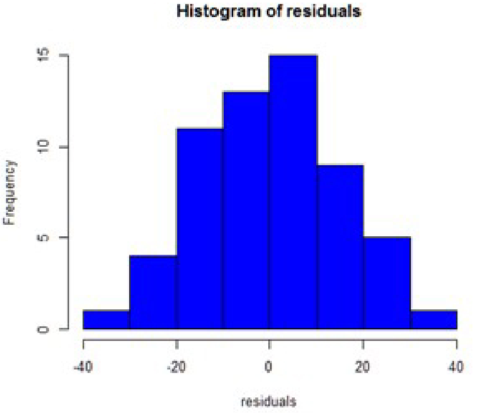
Variance of the Random Error Term—

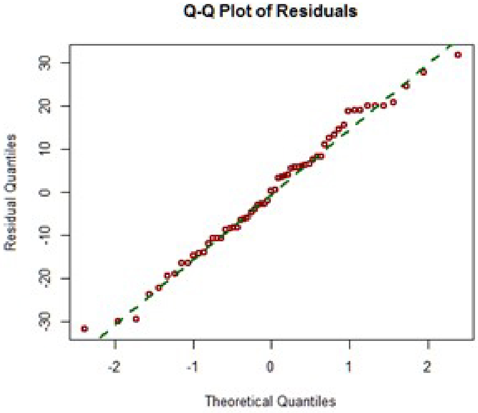
$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n} (y_i-\hat{y}_i)^2}{n-2} \,.$$

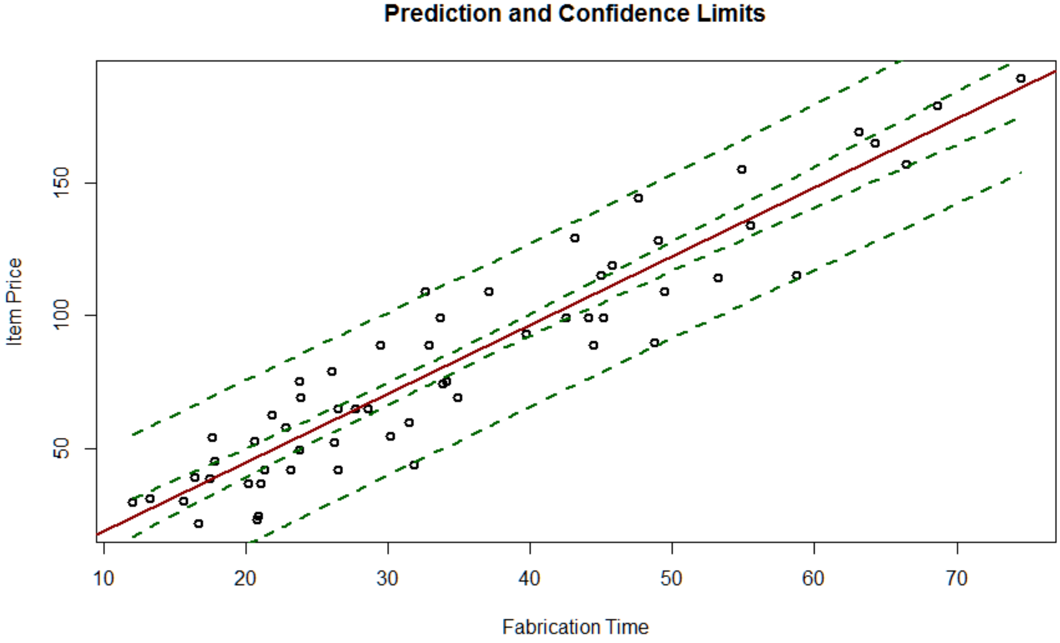
Example of Simple Linear Regression


Item Price versus Fabrication Time


Regression of Price versus Time


Histogram of residuals


Q-Q Plot of Residuals


Prediction and Confidence Limits

Linear Regression Example

```
> items <- read.csv(file.path("c:/R401/","pricing.csv"), sep=",")
> require(moments)
> object <- lm(PRICE~TIME, items)
> summary(object)
```

```
        lm(formula = PRICE ~ TIME,  data = items)

        Estimate   Std. Error   t value    Pr(>|t|)
        (Intercept)  -7.4467     4.8957    -1.521     0.134
        TIME          2.5942     0.1287    20.154    <2e-16 ***
        ---
        Residual standard error: 15.19 on 57 degrees of freedom
        Multiple R-squared:  0.8769,          Adjusted R-squared:  0.8748
        F-statistic: 406.2 on 1 and 57 DF,  p-value: < 2.2e-16
```

```
> cbind(object$coefficients, confint(object, parm= c(1,2), level= 0.95))
```

```
                                2.5 %      97.5 %
        (Intercept) -7.44667   -17.2502    2.3568
        TIME         2.5942      2.3365    2.8520
```

Bias-Corrected and Accelerated Bootstrap

```
> bs <- function(formula, data, indices){
+   d <- data[indices,]
+   fit <- lm(formula, data =d)
+   return(coef(fit))
+ }
> library(boot)
> set.seed(123)
> results <- boot(data=items, statistic=bs, R=1000, formula=PRICE~TIME)
>
> bca.intercept <- boot.ci(boot.out=results, type = "bca", index = 1)
> bca.slope <- boot.ci(boot.out=results, type= "bca", index = 2)
> bca.coef<- c(bca.intercept[2],bca.slope[2])
> bca.alpha <- c(bca.intercept$bca[4:5])
> bca.beta <- c(bca.slope$bca[4:5])
> comb.bca <- rbind(bca.intercept$bca[4:5], bca.slope$bca[4:5])
> colnames(comb.bca) <- c("2.5%", "97.5%")
> bca.estimates <- cbind(bca.coef, comb.bca)
> rownames(bca.estimates) <- c("(Intercept)", "TIME")
> bca.estimates
                      bca.coef       2.5%        97.5%
        (Intercept) -7.446681   -15.745730   0.2645653
        TIME          2.594250     2.346807   2.8087530
```

The bias-corrected and accelerated method is recommended for general use by Bradley Efron and Robert J. Tibshirani in "*An Introduction to the Bootstrap*" *(CRC Press)* Chapter 14 page 188.

Model Specification in Regression Analysis

Multiple linear regression relates a dependent variable to one or more independent variables.

Stages of regression analysis:

1. exploratory data analysis,
2. model specification,
3. estimation of the parameters of the model,
4. diagnostic checking and validation,
5. interpretation of the parameters.

EDA, theoretical considerations and prior experience contribute to model specification.

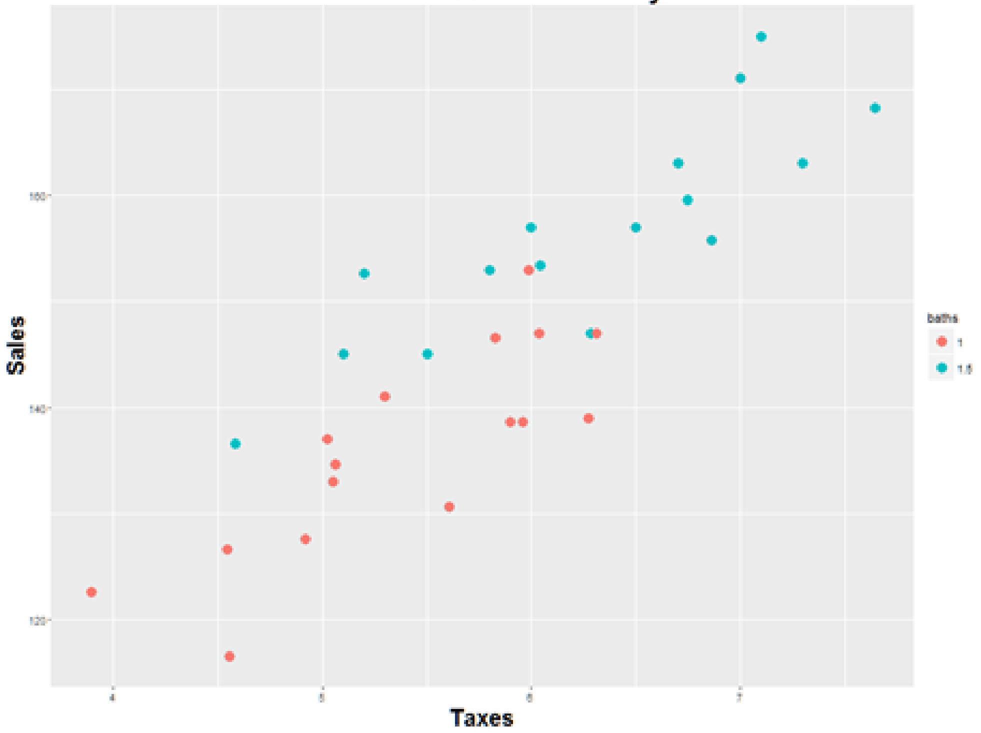
Model Specification Questions:

- Are the right independent variables included in the model?
- Are unnecessary variables excluded from the model?
- Are the variables expressed in proper functional form?

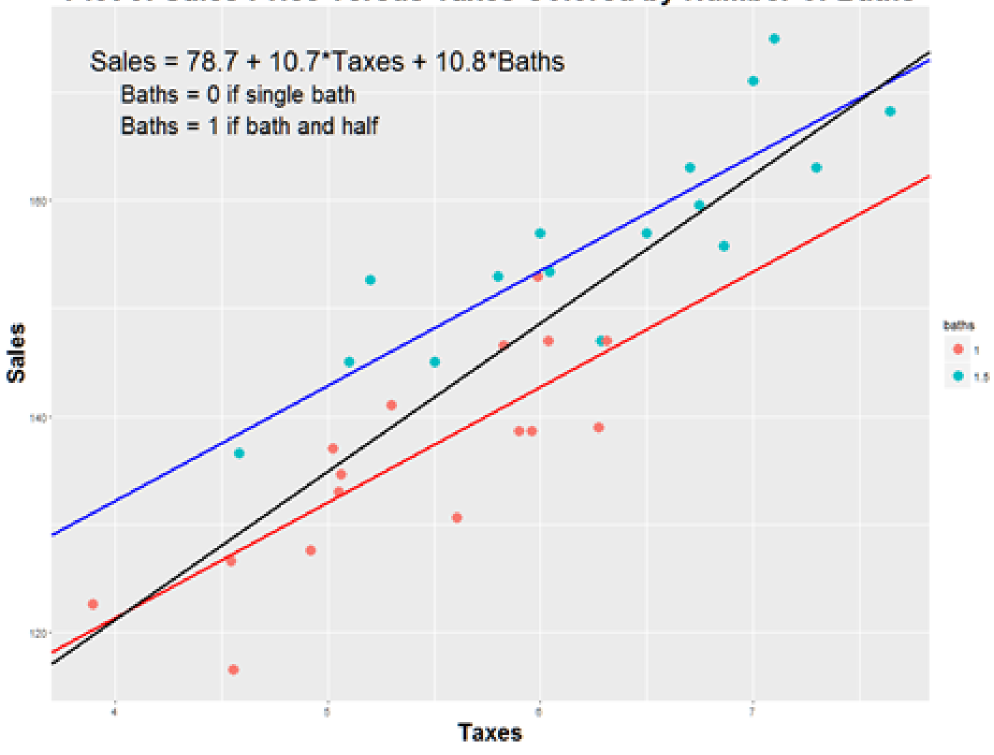
Specification errors can lead to problems of estimation, interpretation and erroneous prediction.

Model Specification and Dummy Variables

Plot of Sales Price versus Taxes Colored by Number of Baths



Plot of Sales Price versus Taxes Colored by Number of Baths

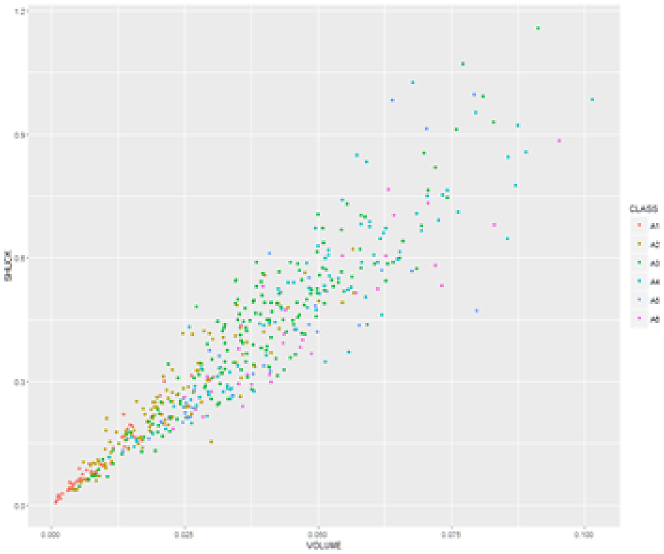


Sales = 78.7 + 10.7*Taxes + 10.8*Baths

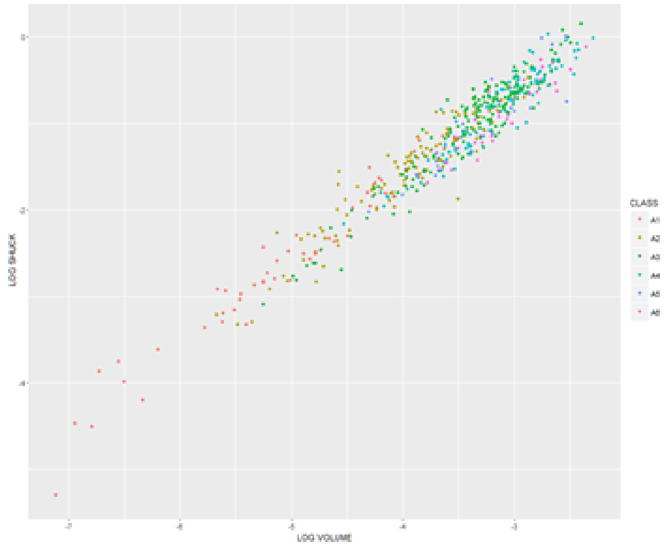
Baths = 0 if single bath

Baths = 1 if bath and half

Transformations



$$Shuck \cong k * Volume$$

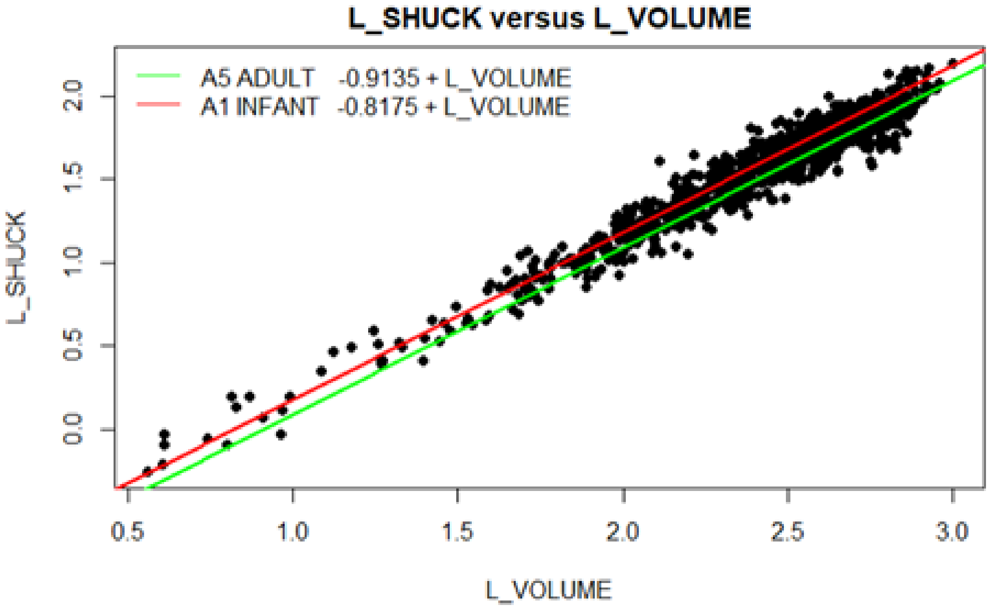


$$\log(Shuck) \cong c + m * \log(Volume)$$

http://kenbenoit.net/assets/courses/ME104/logmodels2.pdf

Abalone Regression Analysis

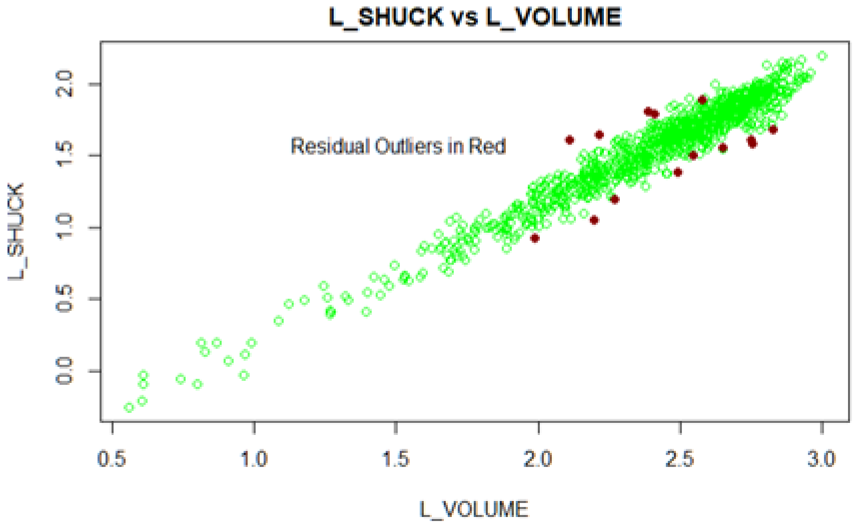
```
> model <- lm(L_SHUCK ~ L_VOLUME + CLASS + TYPE, data = mydata)
> summary(model)

Call:
lm(formula = L_SHUCK ~ L_VOLUME + CLASS + TYPE, data = mydata)

Residuals:
      Min         1Q     Median        3Q       Max
-0.270634 -0.054287  0.000159  0.055986  0.309718

Coefficients:
              Estimate  Std. Error  t value    Pr(>|t|)
(Intercept)  -0.817512  0.019040  -42.936   <2e-16 ***
L_VOLUME      0.999303  0.010262   97.377   <2e-16 ***
CLASSA2      -0.018005  0.011005   -1.636   0.102124
CLASSA3      -0.047310  0.012474   -3.793   0.000158 ***
CLASSA4      -0.075782  0.014056   -5.391   8.67e-08 ***
CLASSA5      -0.117119  0.014131   -8.288   3.56e-16 ***
TYPEADULT     0.021093  0.007688    2.744   0.006180 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08297 on 1029 degrees of freedom
Multiple R-squared:  0.9504,     Adjusted R-squared:  0.9501
F-statistic:  3287 on 6 and 1029 DF,  p-value: < 2.2e-16
```



L_SHUCK versus L_VOLUME

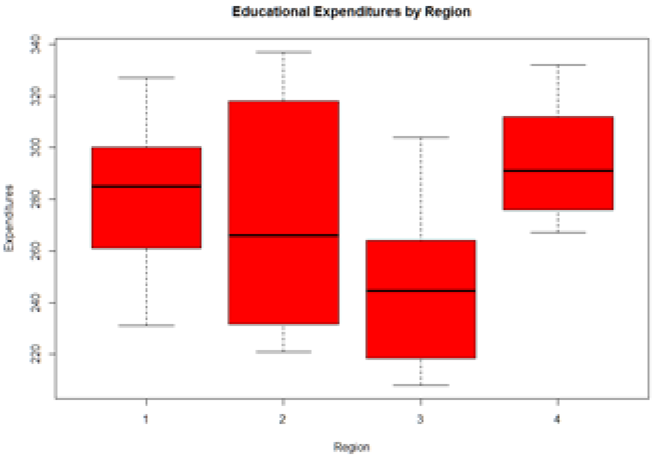
Drop the Residual Outliers

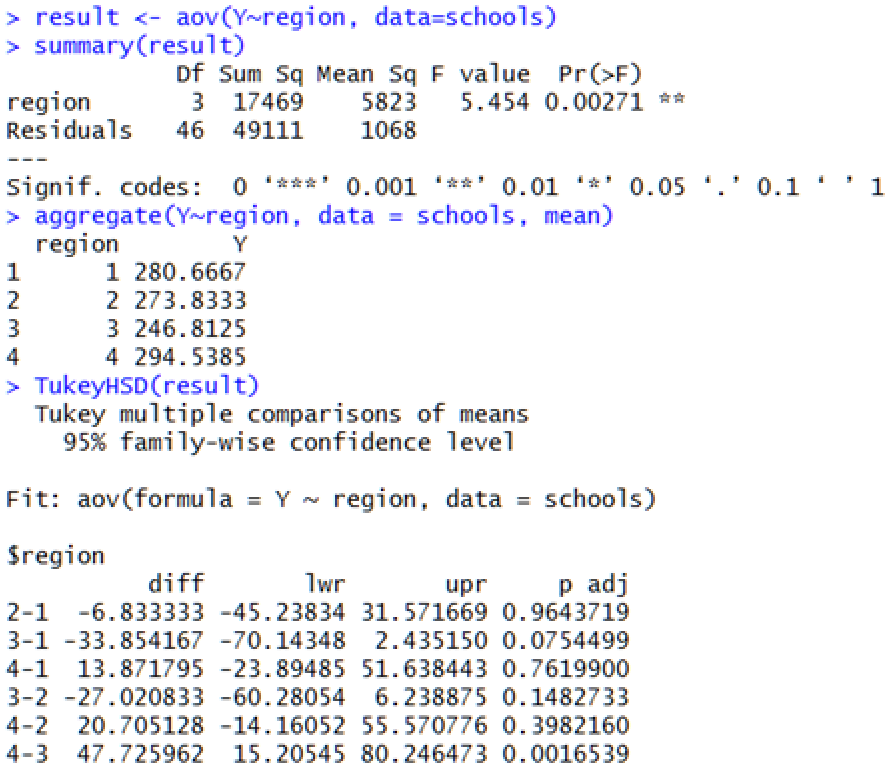
```
> summary(model$residuals)
     Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-0.2703072 -0.0540765  0.0002842  0.0000000  0.0558576  0.3096580

> results <- boxplot.stats(model$residuals, coef = 1.5)
> index <- abs(model$residuals) <= min(abs(results$out))
> 1036 - sum(index)
[1] 14
>
> model <- lm(L_SHUCK ~ L_VOLUME + CLASS + TYPE, data = mydata[index,])
> summary(model)

Call:
lm(formula = L_SHUCK ~ L_VOLUME + CLASS + TYPE, data = mydata[index, ])

Residuals:
      Min        1Q    Median        3Q       Max
-0.269362 -0.053679  0.000157  0.054796  0.260901

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.821168   0.018656 -44.017  < 2e-16 ***
L_VOLUME     1.001562   0.010104  99.123  < 2e-16 ***
CLASSA2     -0.020273   0.010874  -1.864 0.062559 .
CLASSA3     -0.046933   0.012315  -3.811 0.000147 ***
CLASSA4     -0.077358   0.013869  -5.578 3.12e-08 ***
CLASSA5     -0.117981   0.013960  -8.451  < 2e-16 ***
TYPEADULT    0.020088   0.007544   2.663 0.007872 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08108 on 1016 degrees of freedom
Multiple R-squared:  0.953,     Adjusted R-squared:  0.9527
F-statistic:  3434 on 6 and 1016 DF,  p-value: < 2.2e-16
```



Model Specification

Annual educational expenditure data are collected each of the fifty states. The states are grouped according to geographic region. It is of interest to find if regional differences can be detected. The initial analysis is a one-way ANOVA of Y = Per capita expenditure on education versus region (1, 2, 3, 4).





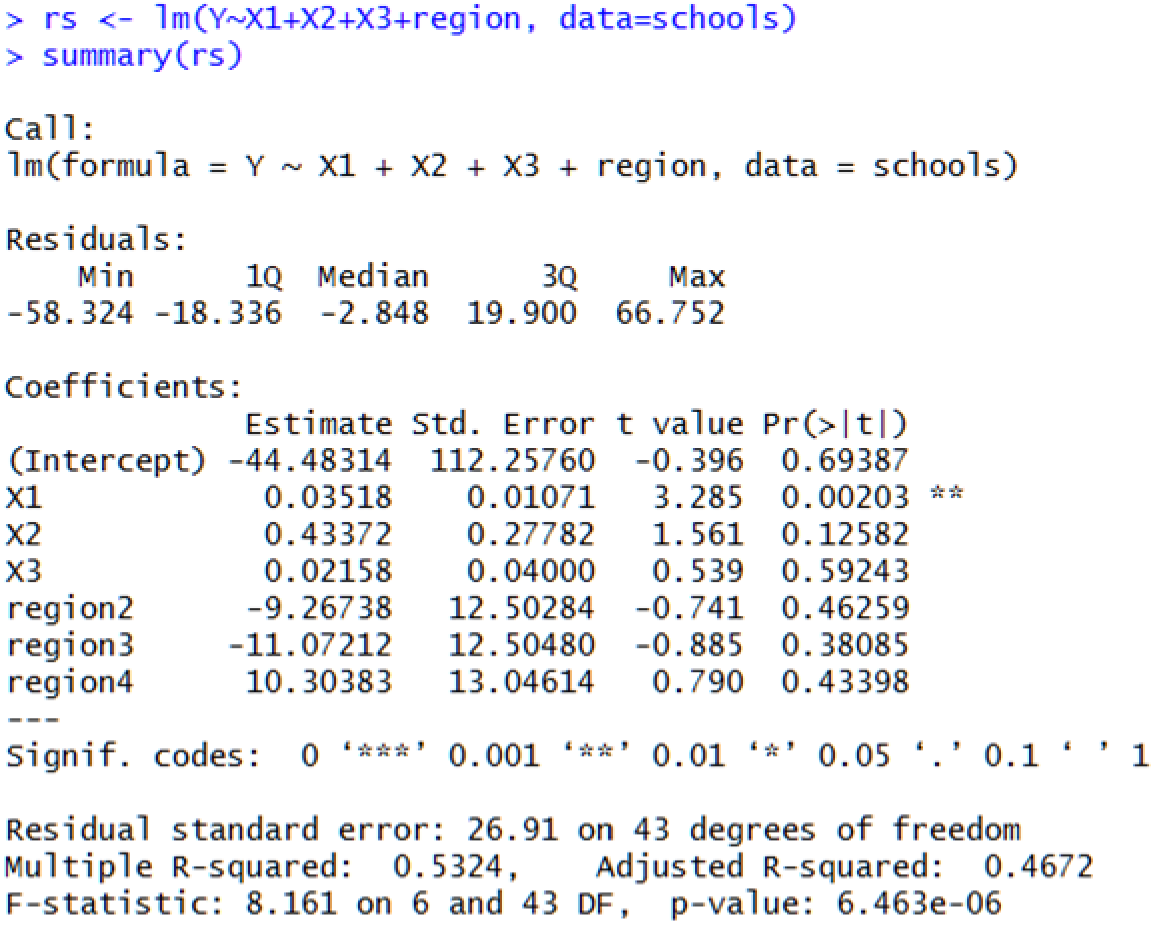
```
> result <- aov(Y~region, data=schools)
> summary(result)
            Df Sum Sq Mean Sq F value  Pr(>F)   
region       3  17469    5823   5.454 0.00271 **
Residuals   46  49111    1068                   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> aggregate(Y~region, data = schools, mean)
  region        Y
1      1 280.6667
2      2 273.8333
3      3 246.8125
4      4 294.5385
> TukeyHSD(result)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Y ~ region, data = schools)

$region
          diff       lwr       upr     p adj
2-1  -6.833333 -45.23834 31.571669 0.9643719
3-1 -33.854167 -70.14348  2.435150 0.0754499
4-1  13.871795 -23.89485 51.638443 0.7619900
3-2 -27.020833 -60.28054  6.238875 0.1482733
4-2  20.705128 -14.16052 55.570776 0.3982160
4-3  47.725962  15.20545 80.246473 0.0016539
```

Model Development

Now consider the analysis if covariates are included in a multiple linear regression analysis. They are:

X1 = Per capita income
X2 = Number of residents per thousand under 18 years of age
X3 = Number of residents per thousand living in urban areas

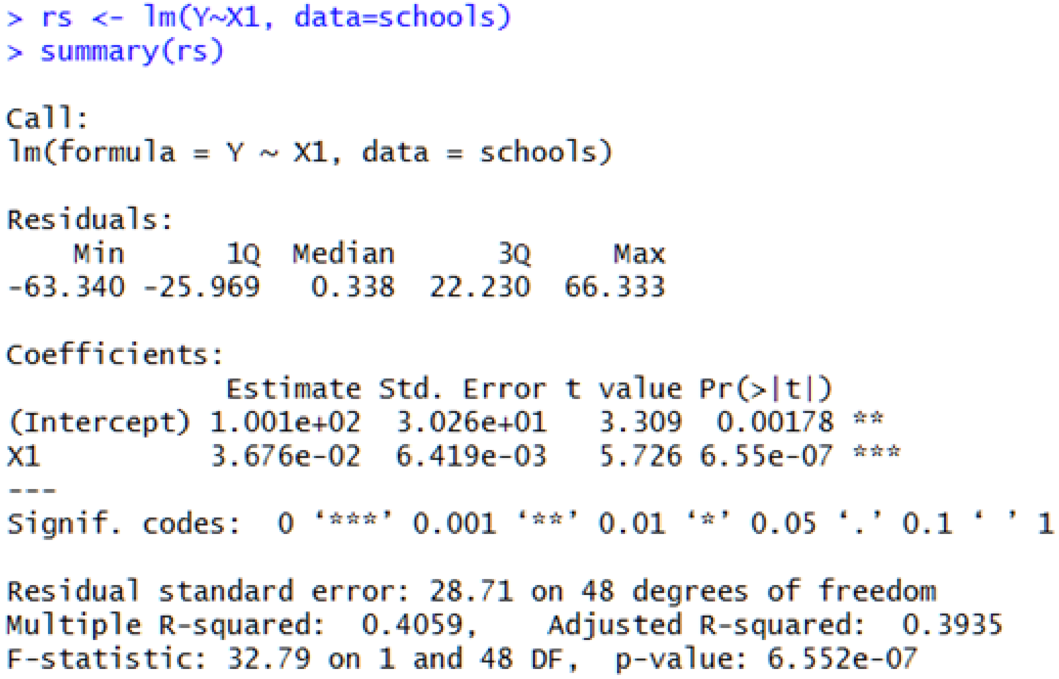
Revised analysis includes both continuous and categorical predictors.



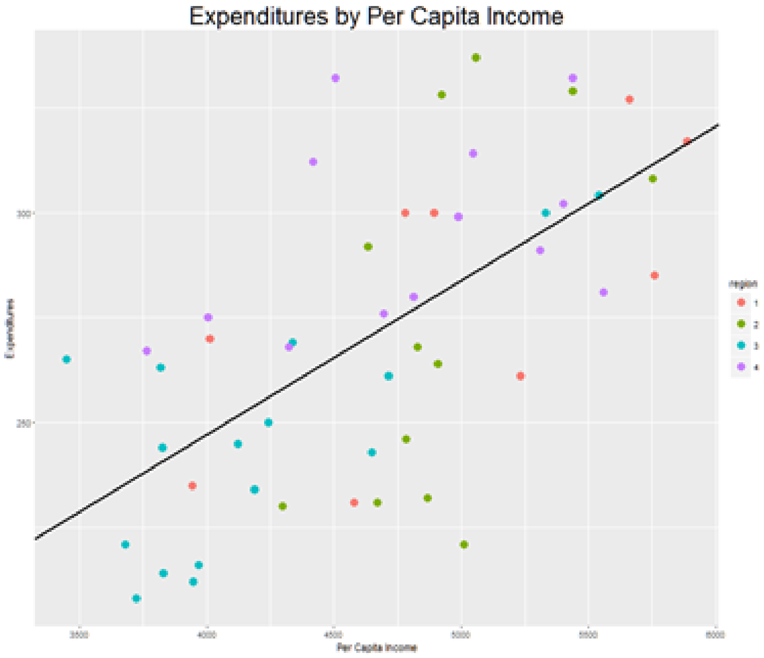
```
> rs <- lm(Y~X1+X2+X3+region, data=schools)
> summary(rs)

Call:
lm(formula = Y ~ X1 + X2 + X3 + region, data = schools)

Residuals:
    Min      1Q  Median      3Q     Max 
-58.324 -18.336  -2.848  19.900  66.752 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept) -44.48314  112.25760  -0.396  0.69387   
X1            0.03518    0.01071   3.285  0.00203 **
X2            0.43372    0.27782   1.561  0.12582   
X3            0.02158    0.04000   0.539  0.59243   
region2      -9.26738   12.50284  -0.741  0.46259   
region3     -11.07212   12.50480  -0.885  0.38085   
region4      10.30383   13.04614   0.790  0.43398   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26.91 on 43 degrees of freedom
Multiple R-squared:  0.5324,    Adjusted R-squared:  0.4672 
F-statistic: 8.161 on 6 and 43 DF,  p-value: 6.463e-06
```

Region is no longer a predictive factor. X1 emerges.

Subsequent Analysis

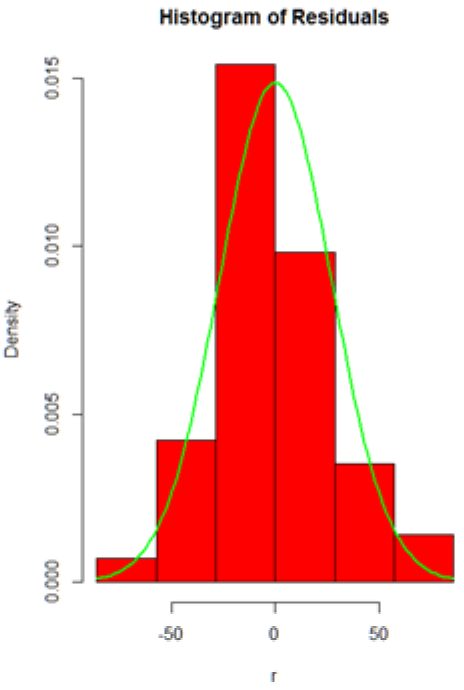


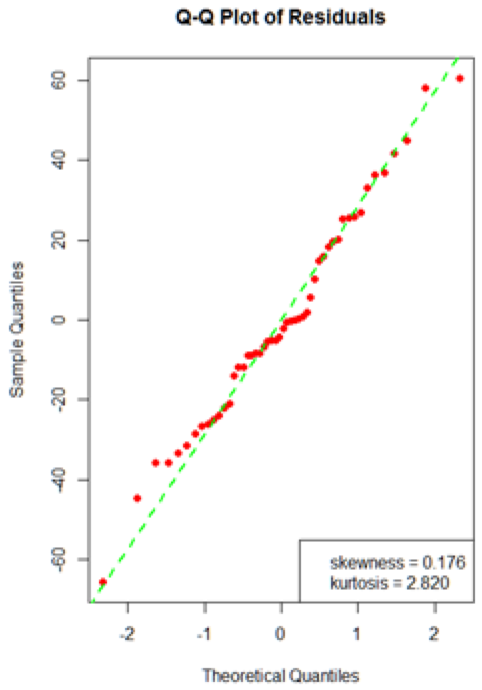
```
> rs <- lm(Y~X1, data=schools)
> summary(rs)

Call:
lm(formula = Y ~ X1, data = schools)

Residuals:
    Min      1Q  Median      3Q     Max 
-63.340 -25.969   0.338  22.230  66.333 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.001e+02  3.026e+01   3.309  0.00178 ** 
X1          3.676e-02  6.419e-03   5.726 6.55e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28.71 on 48 degrees of freedom
Multiple R-squared:  0.4059,    Adjusted R-squared:  0.3935 
F-statistic: 32.79 on 1 and 48 DF,  p-value: 6.552e-07
```

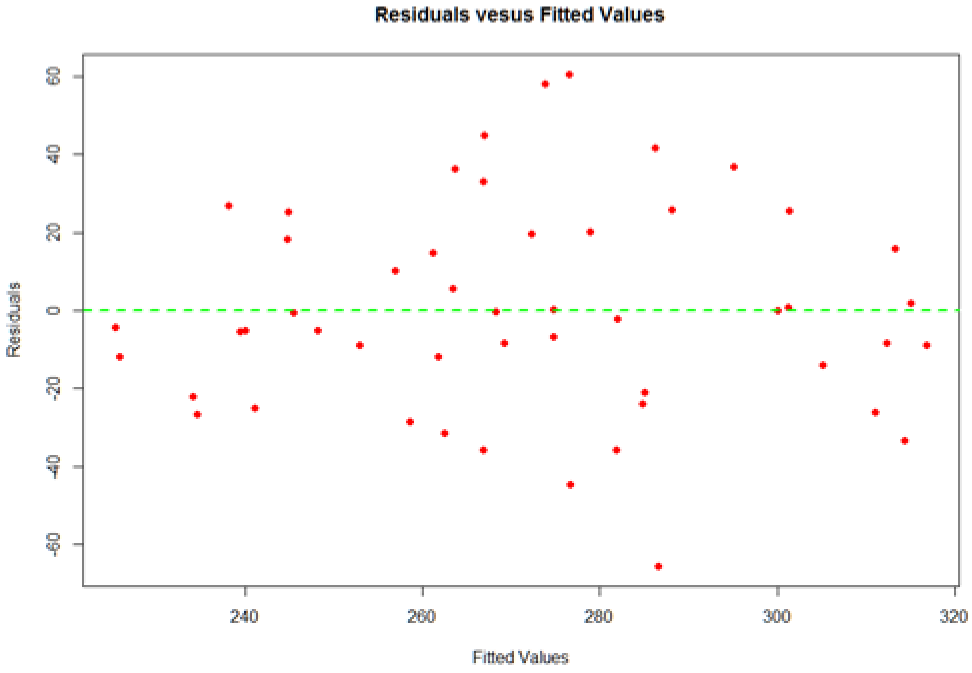


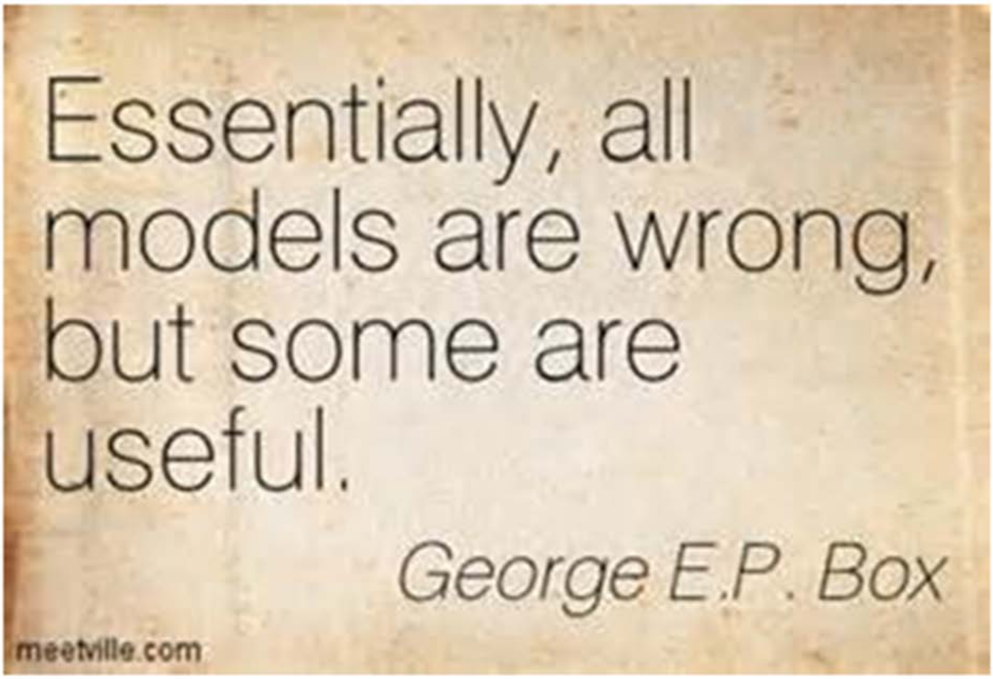
Subsequent regression analysis on X1, X2 and X3 indicates X1 and X2 should be retained as predictors in a multiple linear regression model.

Model Diagnostics (X1 and X2 predictors)









Essentially, all models are wrong, but some are useful.

George E.P. Box

meetville.com

George Edward Pelham Box FRS

(18 October 1919 – 28 March 2013) was a statistician, who worked in the areas of quality control, time-series analysis, design of experiments, and Bayesian inference. He has been called "one of the great statistical minds of the 20th century".

- **Statistics, as a science, is not algorithmic or deterministic. Data are rarely perfect. Judgment is necessary in the application of statistical methods to arrive at valid, useful conclusions.**
- **A model for data must be discarded or revised if it does not adequately fit the data and is misleading. This may lead to insights and progress.**
- **When faced with judgment calls, make the choice that best facilitates understanding the world as it is. It is fine to consider alternative models in the process of drawing conclusions.**

Selected Review Problems

	Approve of mayor	Do not approve of mayor
Republican	8	17
Democrat	18	13
Independent	7	37

One of the 100 test subjects is selected at random. Given that the person selected approves of the mayor, what is the probability they vote Democrat? Use Bayes Theorem.

$$\frac{P[\textit{Democrat who Approves of Mayor}]}{P[\textit{Person Approves of Mayor}]}$$

$$\frac{18/100}{(18/31)(31/100)+(8/25)(25/100)+(7/44)(44/100)}=18/33=0.545$$

Suppose there are three married couples: 1, 2 and 3: couple 1, both partners approve of the mayor, couple 2, both partners no not approve of the mayor, and couple 3, one partner approves of the mayor and the other partner does not approve of the mayor. Pick one couple at random and partner at random. If the selected partner does not approve of the mayor, what is the probability the other partner approves of the mayor?

Couple 1	**Couple 2**	**Couple 3**
A and A	**D and D**	**A and D**

Solution by enumeration: Couple 1 is out of consideration. Only couples 2 and 3 have partners who disapprove. Based on the stated sampling condition, there are three ways a partner who disapproves could be picked. Couple 2 partner 1, Couple 2 partner 2 or Couple 3 partner 2. Only one of these three possibilities has a partner who approves. Thus the conditional probability is 1 out of 3 possibilities or 1/3.

$$\frac{(1/2)(1/3)}{0(1/3)+1(1/3)+(1/2)(1/3)}=1/3$$

Fill in the missing entries in the following one-way analysis of variance table.

Source	df	SS	MS=SS/df	F-statistic
Treatment	3			11.16
Error		13.72	0.686	
Total				

Error degrees of freedom = 13.72/0.686 = 20. Total degrees of freedom = 3 + 20 = 23

Treatment MS = 11.16(.686) = 7.656. Treatment SS = 7.656(3) = 22.97

Total SS = 13.72 + 22.97 = 36.69

The systolic blood pressures of the patients at a hospital are normally distributed with a mean of 138 mm Hg and a standard deviation of 13.5 mm Hg. Find the two blood pressures having these properties: The mean is midway between them and 90% of all blood pressures are between them.

We are looking for an interval that is symmetric with the mean in the middle. To have 90% of the blood pressures between them, 95% of the readings must be to the left of the upper bound, and 5% to the left of the lower bound.

```
> qnorm(0.95, 138, 13.5, lower.tail = TRUE)
[1] 160.2055
> qnorm(0.05, 138, 13.5, lower.tail = TRUE)
[1] 115.7945
```

Assume a binomial experiment has been completed and two independent random samples were collected. For the first sample there were 18 successes out of 50 trials (p1 = 18/50). For the second sample there were 30 successes out of 60 trials (p2 = 30/60). Construct a 90% two-sided confidence interval for the difference $p_1 – p_2$. Determine if this is a statistically significant difference.

Read the problem carefully.

Correct solution	Wrong approach for this problem statement
> p1 <- 18/50 # 0.36	> p1 <- 18/50 # 0.36
> p2 <- 30/60 # 0.5	> p2 <- 30/60 # 0.5
> p1 - p2	> p2 - p1
[1] -0.14	[1] 0.14
See Section 10.4 Business Statistics	See Section 10.4 Business Statistics
> p_bar <- 48/110	> p_bar <- 48/110
> std <- sqrt(p_bar*(1-p_bar)*(1/50 + 1/60))	> std <- sqrt(p_bar*(1-p_bar)*(1/50 + 1/60))
> z <- (p1 - p2)/std	> z <- (p2 - p1)/std
> round(pnorm(z, 0, 1, lower.tail = TRUE), digits = 4)	> round(pnorm(z, 0, 1, lower.tail = FALSE), digits = 4)
[1] 0.0702	[1] 0.0702
> z.q <- qnorm(0.95, 0, 1, lower.tail = TRUE)	> z.q <- qnorm(0.95, 0, 1, lower.tail = TRUE)
> (p1 - p2) + round(c(-z.q, z.q)*std, digits = 4)	> (p2 - p1) + round(c(-z.q, z.q)*std, digits = 4)
[1] -0.2962 0.0162	[1] -0.0162 0.2962
Check page 278 Business Statistics.	Check page 278 Business Statistics.
Not statistically significant with alpha = 0.1.	Not statistically significant with alpha = 0.1.

What are 95% one-sided confidence intervals for the following alternatives?

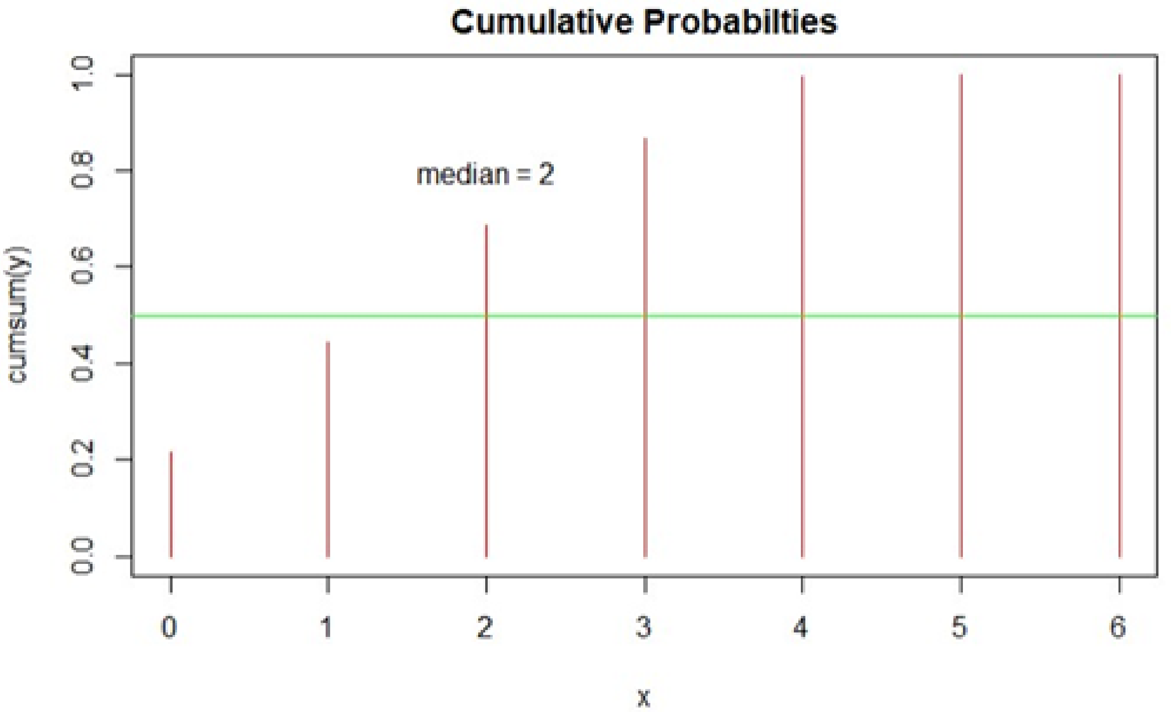
(p1 – p2) > 0   A lower confidence bound is needed. It is -0.2962 to plus infinity.

(p1 – p2) < 0   An upper confidence bound is needed. It is 0.0162 to minus infinity.

A discrete random variable has outcomes: 0, 1, 2, 3, 4, 5, 6. The corresponding probabilities in sequence are: 0.215, 0.230, 0.240, 0.182, 0.130, 0.003, 0.001. Determine the value of the median, mean, mode and variance for this variable.

0	1	2	3	4	5	6
0.215	0.230	0.240	0.181	0.130	0.003	0.001

Quantiles for a discrete random variable are selected from the possible values of the variable based upon the probability distribution.



Probability of x < 2 is 0.445.
Probability of x >= 2 is (1 – 0.445) or 0.555.

```
x <- seq(0,6)
y <- c(0.215, 0.230, 0.240, 0.181, 0.130, 0.003, 0.001)
avg <- round(sum(y*x),digits = 3)  # 1.794
var <- round(sum(y*(x-avg)^2), digits = 3)  #1.792
```

Largest probability is 0.240 so the mode is 2.

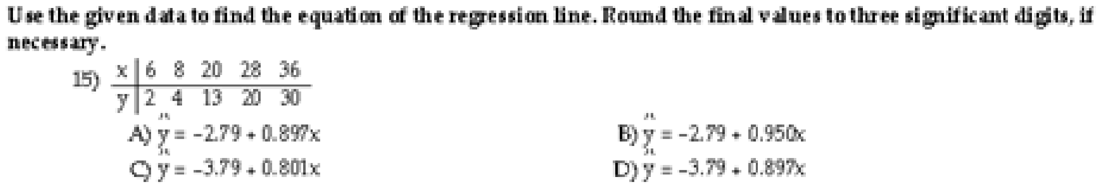
A researcher was interested in comparing the amount of time spent watching television by women and by men. Independent simple random samples of 14 women and 17 men were selected, and each person was asked how many hours he or she had watched television during the previous week. Assume the population variances are equal so that the sample variances can be pooled.

	Women	Men
Sample average	**11.4 hr**	**16.8 hr**
Standard deviation	**4.1 hr**	**4.7 hr**
Sample size	**14**	**17**

Use a 0.05 significance level to test the claim that the mean amount of time spent watching television by women is smaller than the mean amount of time spent watching television by men. Use the t radition method of hypothesis testing.

For this problem, a one-sided test is required. The alternative will be framed as a positive.

```
> s1.2 <- 4.1^2
> s2.2 <- 4.7^2
> n1 <- 14
> n2 <- 17
> pool <- sqrt((s1.2*(n1-1)+s2.2*(n2-1))/(n1+n2-2))
> den <- pool*sqrt(1/n1+1/n2)
> x1 <- 11.4
> x2 <- 16.8
> t <- (x2-x1)/den
> t
[1] 3.369099
> pt(t,29,lower.tail=FALSE)
[1] 0.001073162
> qt(0.95,29,lower.tail=TRUE)
[1] 1.699127
```



Use the given data to find the equation of the regression line. Round the final values to three significant digits, if necessary.

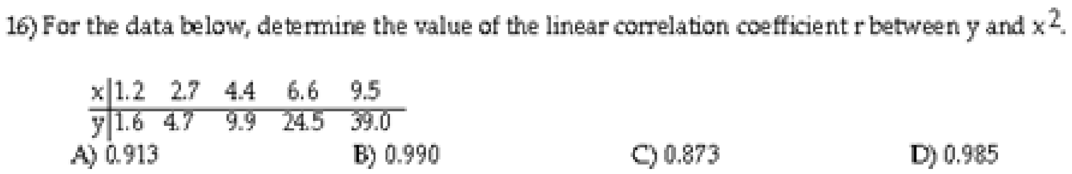
15)

x	6	8	20	28	36
y	2	4	13	20	30

A) $\hat{y} = -2.79 + 0.897x$ B) $\hat{y} = -2.79 + 0.950x$

C) $\hat{y} = -3.79 + 0.801x$ D) $\hat{y} = -3.79 + 0.897x$

```
> x <- c(6, 8, 20, 28, 36)
> y <- c(2, 4, 13, 20, 30)
> lm(y~x)

Call:  lm(formula = y ~ x)

Coefficients:
(Intercept)            x
    -3.7900       0.8975
```

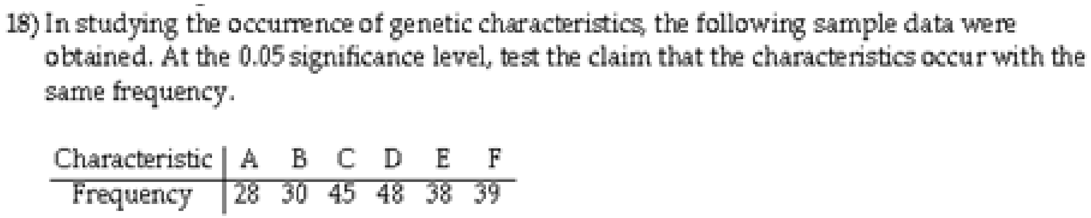


16) For the data below, determine the value of the linear correlation coefficient r between y and x^2.

x	1.2	2.7	4.4	6.6	9.5
y	1.6	4.7	9.9	24.5	39.0

A) 0.913 B) 0.990 C) 0.873 D) 0.985

```
> x <- c(1.2, 2.7, 4.4, 6.6, 9.5)
> y <- c(1.6, 4.7, 9.9, 24.5, 39.0)
> x <- x^2
> cor(x, y, method = c("pearson"))
[1] 0.9902759
```



18) In studying the occurrence of genetic characteristics, the following sample data were obtained. At the 0.05 significance level, test the claim that the characteristics occur with the same frequency.

Characteristic	A	B	C	D	E	F
Frequency	28	30	45	48	38	39

This is a Chi-square goodness-of-fit test. The counts are expected to be equal under the null hypothesis. This expectation is 38 which needed to be compared against the observed counts.

```
> obs <- c(28, 30, 45, 48, 38, 39)
> ec <- rep(sum(obs)/6, times = 6)
> diff <- sum((obs - ec)^2/ec)
> diff
[1] 8.263158
> pchisq(diff, df = 5, lower.tail = FALSE)
[1] 0.1423164
> qchisq(0.95, df = 5, lower.tail = TRUE)
[1] 11.0705
```

Customers in a store were selected at random to participate in a taste test. Two hundred people participated and expressed a preference for one of two beverages. Use the taste test preference data in the following table to test the hypothesis of equal preferences for the two beverages in the study. Test at the 5% level.
Use an uncorrected Chi-square test.

	children	men	women
Beverage A	35	23	42
Beverage B	25	30	45

```
> data
     [,1] [,2] [,3]
[1,]   35   23   42
[2,]   25   30   45
> chisq.test(data, correct = FALSE)

          Pearson's Chi-squared test
X-squared = 2.6946, df = 2, p-value = 0.2599
```

Some Sync Session Learning Points

- Essentially all models are wrong, but some are useful.

- It is perfectly proper to use both classical and robust methods routinely and only worry when they differ enough to matter.

- The Pearson Correlation Coefficient is intended to measure the association between two normally distributed random variables.

- Simple linear regression involves estimating two parameters in the equation and the variance of the error term.

- In simple linear regression, r^2 equals the square of the Pearson Correlation Coefficient.

- r^2 is the ratio of explained variation to total variation.

- Linear regression requires the normal equations to be amenable to linear algebra. It must be possible to isolate the coefficients.

- Multiple linear regression is not limited to straight line relationships. Polynomials may qualify.

- Model specification involves answering three questions:
    - Are the right variables included?
    - Are unnecessary variables excluded?
    - Are the variables in proper functional form?

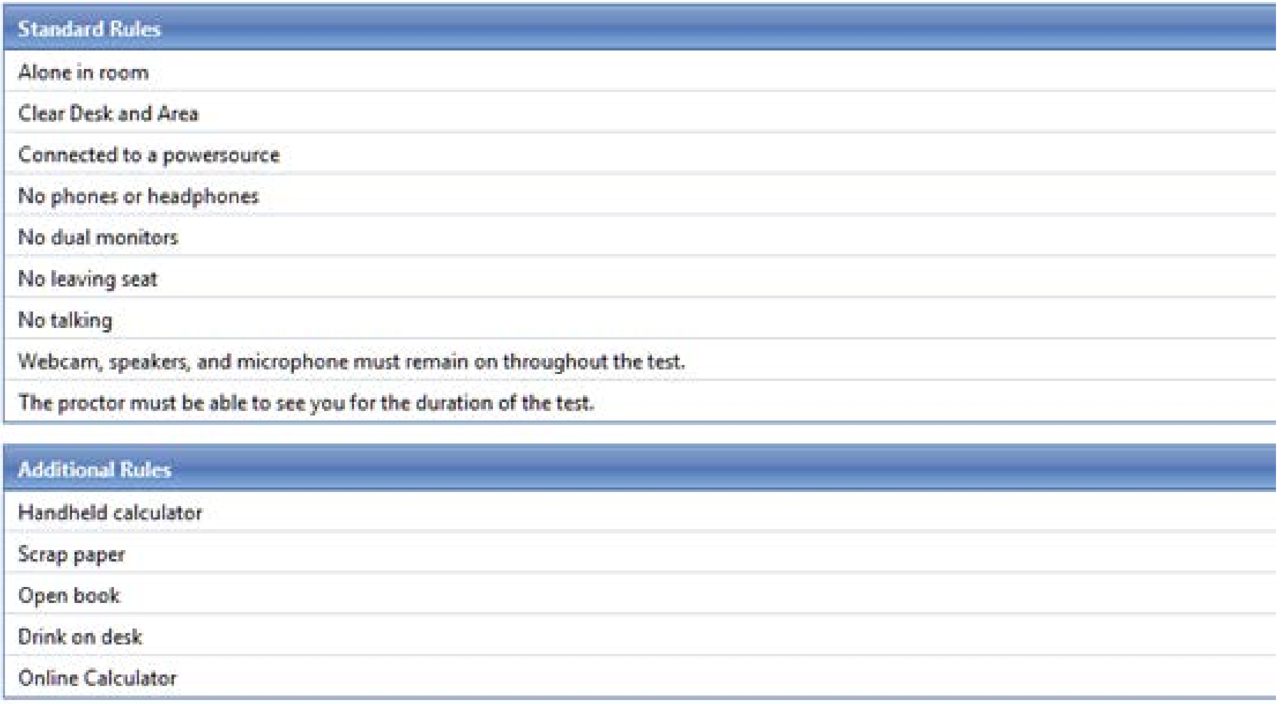
Topics on the Final Exam

- Probability
  - Calculations using R probability functions
  - Bayes' Theorem
  - Mean, median and variance for distributions
- Hypothesis Testing
  - One-sided and two-sided tests
  - Chi-square Test of Independence
  - Chi-square test of a variance
  - t tests
    - paired
    - two sample
- Confidence Interval Construction
- One-way AOV
  - F tests
  - p-values
  - critical values
- Linear Regression
  - Pearson Correlation Coefficient

Review questions are available in the Week Ten module.

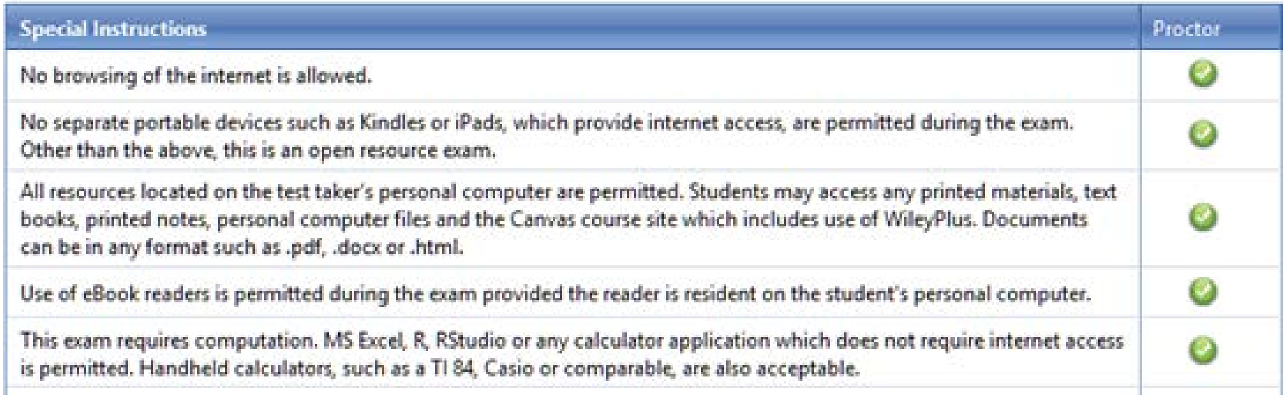
The test is two hours, proctored, with open book and open notes. The questions are multiple choice and true/false. Excel, R or any comparable calculator application may be used. The course site, WileyPlus, electronic files or hardcopy may be used. The use of portable devises such as kindles and iPads is not allowed unless special arrangements are made. No navigation from the testing site to the internet for browsing is allowed.

Examity Specifications



Standard Rules
Alone in room
Clear Desk and Area
Connected to a powersource
No phones or headphones
No dual monitors
No leaving seat
No talking
Webcam, speakers, and microphone must remain on throughout the test.
The proctor must be able to see you for the duration of the test.

Additional Rules
Handheld calculator
Scrap paper
Open book
Drink on desk
Online Calculator

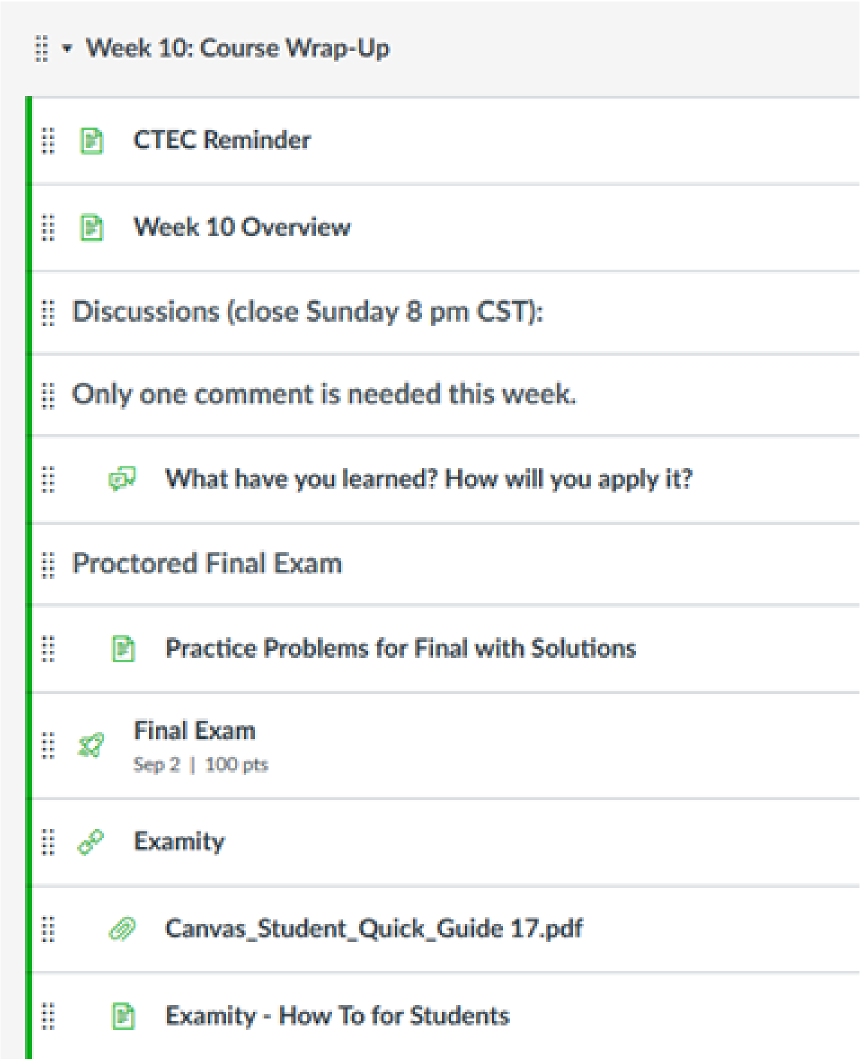


Special Instructions	Proctor
No browsing of the internet is allowed.	✓
No separate portable devices such as Kindles or iPads, which provide internet access, are permitted during the exam. Other than the above, this is an open resource exam.	✓
All resources located on the test taker's personal computer are permitted. Students may access any printed materials, text books, printed notes, personal computer files and the Canvas course site which includes use of WileyPlus. Documents can be in any format such as .pdf, .docx or .html.	✓
Use of eBook readers is permitted during the exam provided the reader is resident on the student's personal computer.	✓
This exam requires computation. MS Excel, R, RStudio or any calculator application which does not require internet access is permitted. Handheld calculators, such as a TI 84, Casio or comparable, are also acceptable.	✓

Examity has been instructed that this is an open book exam meaning "open resource". All resources located on the test taker's personal computer, external file storage as with the cloud, and printed materials are permitted.

No copying, saving or retaining entire exam questions. Students may copy and paste data from the exam into R for calculations.

Final Exam



Week 10: Course Wrap-Up

CTEC Reminder

Week 10 Overview

Discussions (close Sunday 8 pm CST):

Only one comment is needed this week.

What have you learned? How will you apply it?

Proctored Final Exam

Practice Problems for Final with Solutions

Final Exam
Sep 2 | 100 pts

Examity

Canvas_Student_Quick_Guide 17.pdf

Examity - How To for Students

- **You are responsible for scheduling and paying for your final exam.**
  - 2 hour exam: $23.00
  - Scheduling within 24 hours of exam: $5.00 per hour
  - Cancellations or schedule changes within 24 hours of exam: $5.00 per exam
  - No-shows: Full payment of all proctoring fees ($15.00 for the first hour plus $7.00 for each additional hour)
- **Arrange a "dry run" with Examity in advance to test your equipment and get any questions answered.**
- **This is an open-book exam, however only one screen is allowed.**
- **The two-hour exam consists of ten multiple choice questions. No questions about R, however R may be used for calculations.**