Recap: Chapter 1-5

STAT 3240

Michael McIsaac

UPEI

Learning Objectives for Chapter 1

- Describe the uses of regression analysis
- Contrast regression vs causation
- Identify observational and experimental data and contrast these with respect to causation
- Label and interpret the components of a regression model
- Apply the method of least squares
- Define point estimates of mean response and residuals
- Define the normal error regression model
- Define and interpret SSE and MSE
- Apply the method of maximum likelihood

Learning Objectives for Chapter 2

- Compute and interpret confidence intervals for $E[Y]$
- Compute and interpret prediction intervals for a new observation
- Compute and interpret confidence bands for a regression line
- Construct and interpret an ANOVA table
- Conduct and interpret an ANOVA F test
- Describe the general linear test approach
- Calculate and interpret R^2
- Understand the limitations of R^2
- Describe the limitations of linear regression analysis
- Contrast regression and correlation
- Conduct and interpret inference on correlation coefficients
- Estimate, interpret, test, and contrast Spearman rank correlation.

Learning Objectives for Chapter 3

- Distinguish between residual, studentized residuals, and error term
- Identify outlying X values that could influence the regression function
- Use residual plots to conduct regression diagnostics
- Understand that their are formal tests for residual diagnostics
- Apply formal tests for normality and constant variance
- Carry out and interpret the F test for lack of fit.
- Understand the utility of transformations and when they could be applied.
- Assess the shape of the regression function using smoothed curves.

Learning Objectives for Chapter 4

- Compute and interpret Bonferroni and Working-Hotelling simultaneous CIs
- Compute and interpret simultaneous prediction intervals
- Understand the potential impact of measurement error
- Understand the challenges of choosing X levels when designing an experiment

Learning Objectives for Chapter 5

- Write simple linear regression in matrix terms
- Write simple least squares estimation in matrix terms
- Write fitted values and residuals in matrix terms
- Write ANOVA and regression inferences in matrix terms

Copier maintenance (CH01PR20.txt)

The Tri-City Office Equipment Corporation sells an imported copier on a franchise basis and performs preventive maintenance and repair service on this copier. The data below have been collected from 45 recent calls on users to perform routine preventive maintenance service; for each call, X is the number of copiers serviced and Y is the total number of minutes spent by the service person. Assume that first-order regression model (1.1) is appropriate.

Show 200 entries Search:

	minutes	copiers	machine_age	service_experience
1	20	2	20	4
2	60	4	19	5
3	46	3	27	4
4	41	2	32	1
5	12	1	24	4

Showing 1 to 45 of 45 entries Previous 1 Next

```r
copier_model = lm(minutes~copiers, data=copier_data)
msummary(copier_model)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.5802     2.8039  -0.207    0.837
## copiers      15.0352     0.4831  31.123   <2e-16 ***
##
## Residual standard error: 8.914 on 43 degrees of freedom
## Multiple R-squared:  0.9575,    Adjusted R-squared:  0.9565
## F-statistic: 968.7 on 1 and 43 DF,  p-value: < 2.2e-16
```

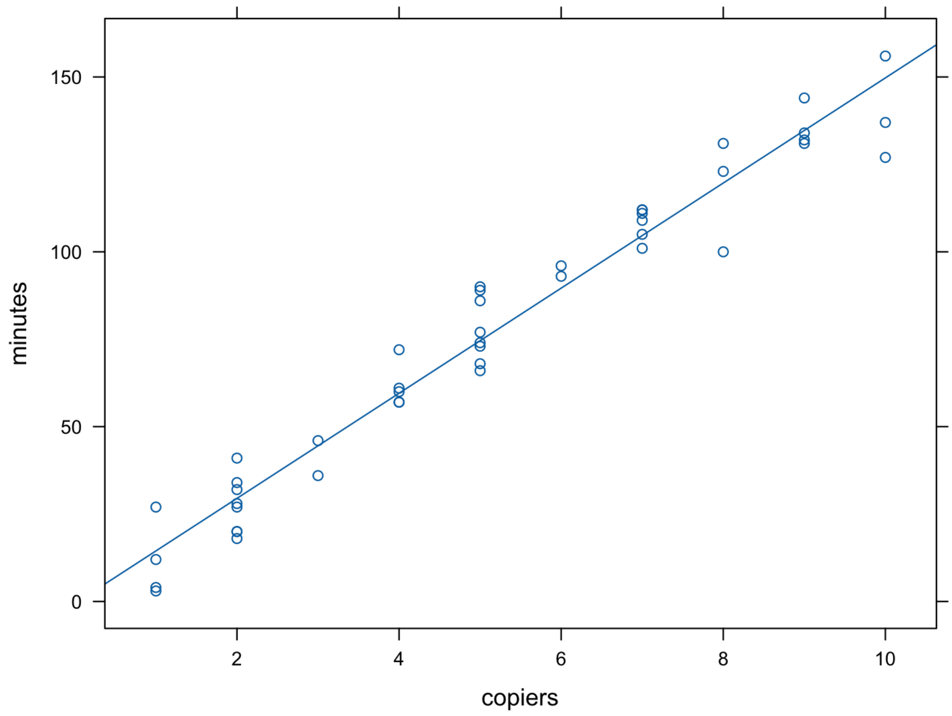
```r
anova(copier_model)
```

```
## Analysis of Variance Table
##
## Response: minutes
##           Df Sum Sq Mean Sq F value    Pr(>F)
## copiers    1  76960   76960  968.66 < 2.2e-16 ***
## Residuals 43   3416      79
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
confint(copier_model)
```

```
##                 2.5 %    97.5 %
## (Intercept) -6.234843  5.074529
```

```
xyplot(minutes~copiers, data=copier_data, type=c("p", "r"))
```



```r
predict(copier_model, newdata=data.frame(copiers=c(4,5,6,7,8)), interval
```

```
##           fit       lwr       upr
## 1  59.56084  55.69857  63.42310
## 2  74.59608  71.01205  78.18011
## 3  89.63133  85.86786  93.39480
## 4 104.66658 100.32234 109.01082
## 5 119.70183 114.50844 124.89522
```

```r
predict(copier_model, newdata=data.frame(copiers=c(4,5,6,7,8)), interval
```

```
##           fit      lwr       upr
## 1  59.56084 35.22952  83.89215
## 2  74.59608 50.30738  98.88478
## 3  89.63133 65.31551 113.94716
## 4 104.66658 80.25412 129.07904
## 5 119.70183 95.12405 144.27960
```

```r
mosaic::cor.test(minutes~copiers, data=copier_data, method="pearson")
```

```
##
##      Pearson's product-moment correlation
##
## data:  minutes and copiers
## t = 31.123, df = 43, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9610128 0.9882095
## sample estimates:
##       cor
## 0.978517
```

```r
mosaic::cor.test(minutes~copiers, data=copier_data, method="spearman", e
```

```
##
##      Spearman's rank correlation rho
##
## data:  minutes and copiers
## S = 310.64, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.9795363
```

```r
copier_model_reduced = lm(minutes~1, data=copier_data)
anova(copier_model_reduced, copier_model)
```

```
## Analysis of Variance Table
##
## Model 1: minutes ~ 1
## Model 2: minutes ~ copiers
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1     44 80377
## 2     43  3416  1     76960 968.66 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

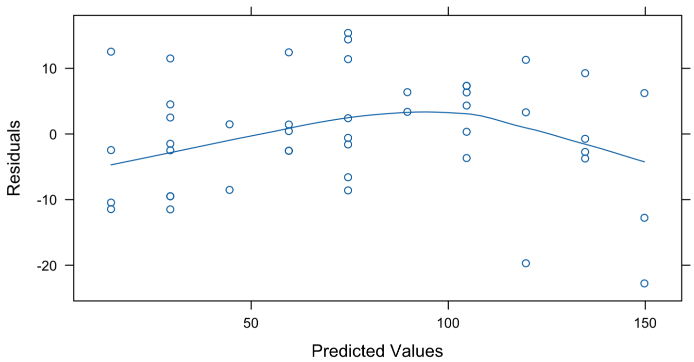
```r
copier_model_full = lm(minutes~factor(copiers), data=copier_data)
anova(copier_model, copier_model_full)
```

```
## Analysis of Variance Table
##
## Model 1: minutes ~ copiers
## Model 2: minutes ~ factor(copiers)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     43 3416.4
## 2     35 2797.7  8    618.72 0.9676 0.4766
```

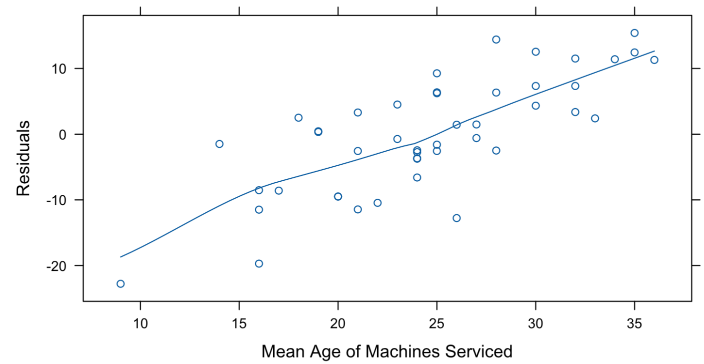
```r
alr3::pureErrorAnova(lm(minutes~copiers, data=copier_data))
```

```
## Analysis of Variance Table
##
## Response: minutes
##              Df Sum Sq Mean Sq  F value Pr(>F)
## copiers       1  76960   76960 962.8105 <2e-16 ***
## Residuals    43   3416      79
##  Lack of fit  8    619      77   0.9676 0.4766
##  Pure Error  35   2798      80
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

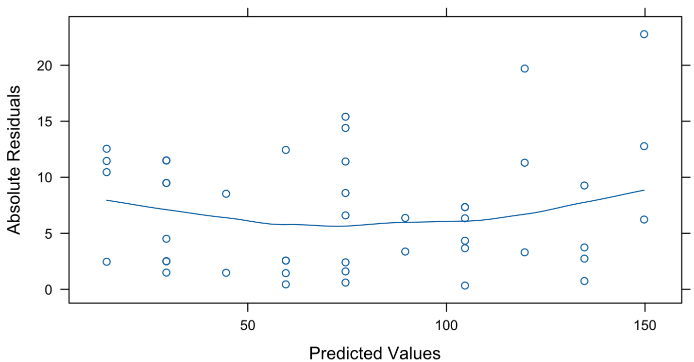
```
xyplot(resid(copier_model)~predic
```



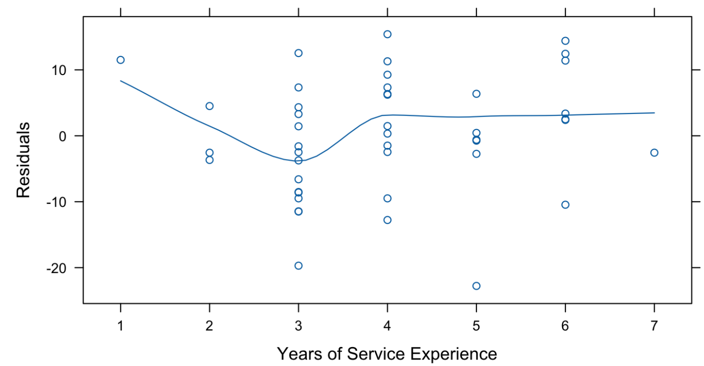

```
xyplot(resid(copier_model)~copier
```



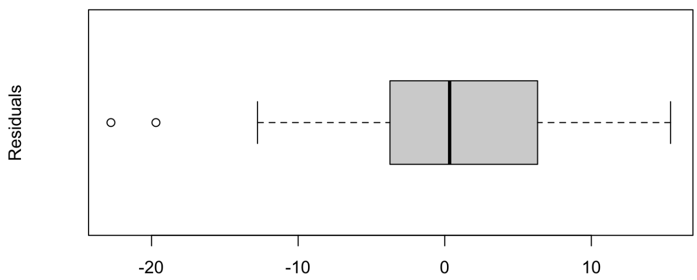

```
xyplot(abs(resid(copier_model))~p
```



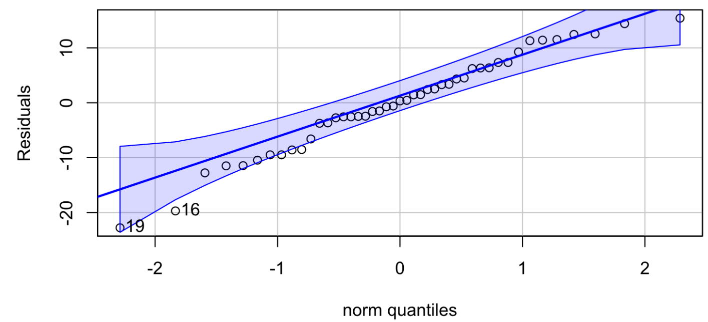

```
xyplot(resid(copier_model)~copier
```



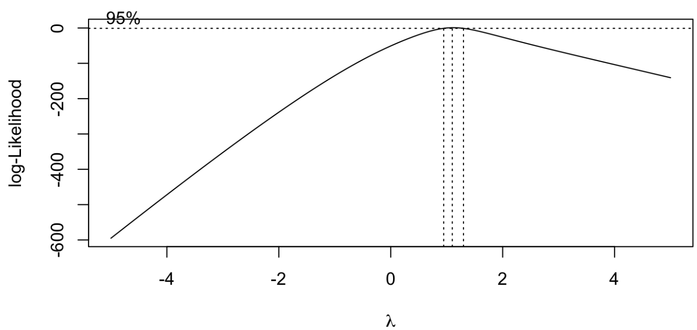
```
boxplot(resid(copier_model), ylab
```



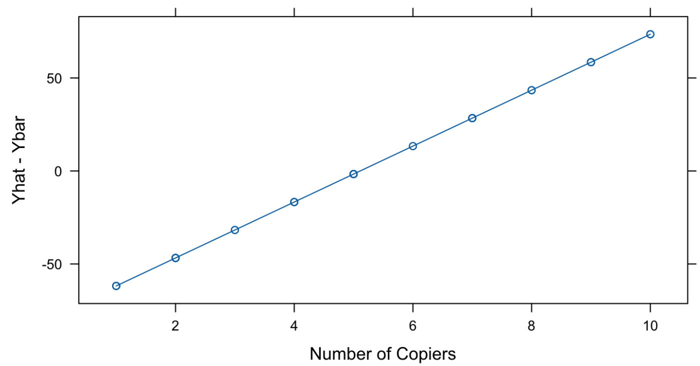

```
qqPlot(resid(copier_model), ylab=
```



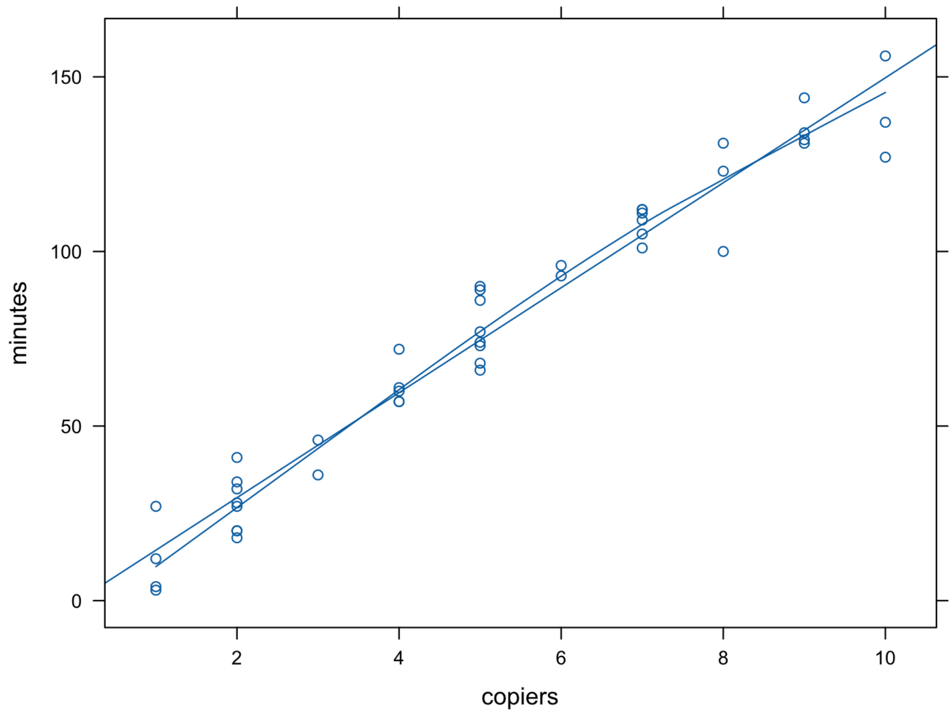

```
MASS::boxcox(copier_model, seq(-5
```




```
xyplot(I(predict(copier_model) -
```



```r
xyplot(minutes~copiers, data=copier_data, type=c("p", "r", "smooth"))
```



- Test $H_0 : \gamma_1 = 0$ in $\ln \sigma_i^2 = \gamma_0 + \gamma_1 X_i$

```
lmtest::bptest(copier_model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  copier_model
## BP = 1.4187, df = 1, p-value = 0.2336
```

Consider the following data

X:	1	2	3	4
Y:	1	2	3	5

Use appropriate matrix algebra to conduct simple linear regression by hand by completing the following tasks:

- Write down the design matrix $\boldsymbol{X}$
- Calculate $\boldsymbol{X}'\boldsymbol{X}$
- Calculate $(\boldsymbol{X}'\boldsymbol{X})^{-1}$
- Calculate $\boldsymbol{X}'\boldsymbol{Y}$
- Calculate $\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$
- Find $\hat{\boldsymbol{Y}}$
- Find $\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$
- Find $\boldsymbol{SSE} = \boldsymbol{e}'\boldsymbol{e}$
- Find $\boldsymbol{SSTO}$ and $\boldsymbol{SSR}$
- Find $\boldsymbol{MSE}$ and $\boldsymbol{MSR}$
- Complete the corresponding ANOVA table
- Find $\boldsymbol{R}^2$

Interpret each of the above quantities.

Now, suppose that we want to conduct regression through the origin. That is, suppose that instead of estimating β_0 and β_1 in the model $Y = \beta_0 + \beta_1 X + \varepsilon$, we assume that we know that $\beta_0 = \mathbf{0}$ and we fit the model $Y = \beta_1 \boldsymbol{X} + \varepsilon$. Use appropriate matrix algebra to conduct this linear regression by hand by completing the following tasks:

- Write down the design matrix $\boldsymbol{X}$
- Calculate $\boldsymbol{X}'\boldsymbol{X}$
- Calculate $(\boldsymbol{X}'\boldsymbol{X})^{-1}$
- Calculate $\boldsymbol{X}'\boldsymbol{Y}$
- Calculate $\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$
- Find $\hat{\boldsymbol{Y}}$
- Find $\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$
- Find $SSE = \boldsymbol{e}'\boldsymbol{e}$
- Find $SSTO$ and SSR
- Find R^2

Interpret each of the above quantities and contrast the two regression models.