## Chapter 3  Section 3.1

Multiple Regression
- Model
- Prediction Equation
- Std. Deviation of Error
- Correlation Matrix

## Model Assumptions:

| Simple Linear Regression: | Multiple Regression: |
|---|---|
| 1.) Linearity | 1.) Linearity |
| 2.) Constant Variance | 2.) Constant Variance |
| 3.) Independent Errors | 3.) Independent Errors |
| 4.) Normality of the Errors | 4.) Normality of the Errors |

Notice that the assumptions are the same for both simple and multiple linear regression.

## Simple Linear Regression Model

$$Y = \boxed{\beta_0 + \beta_1 X} + \varepsilon$$

Data    Model    Error

where $\varepsilon \sim N(0, \sigma_\varepsilon)$ and independent

## Multiple Regression Model

$$Y = \boxed{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k} + \varepsilon$$

Data    Model    Error

Model: Consists of $k$ predictors for a total of $k+1$ parameters.

Error: Each error is Independent and distributed normally with constant variance, i.e. $\varepsilon \sim N(0, \sigma_\varepsilon)$

Data: For each of the $1, 2, \ldots, n$ cases we need a value for $Y$ and for all of $X_1, \ldots, X_k$

The 4 Step Process for Multiple Regression:

Collect data for $Y$ and all predictors.

CHOOSE a form of the model.
- Select predictors; possibly transform $Y$.
- Choose any functions of predictors.

FIT Estimate the coefficients $\hat{\beta}_0, \hat{\beta}_1, \ldots \hat{\beta}_k$
- Estimate the residual standard error: $\hat{\sigma}_\varepsilon$.

ASSESS the fit.
- Test individual predictors: t-tests.
- Test the overall fit: ANOVA, $R^2$.
- Examine residuals.

USE Predictions, CI's, and PI's.

## Multiple Regression Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

$k$ predictors

Recall in simple linear regression we fit the model using least squares, that is, we found the $\hat{\beta}$ that minimized $\Sigma(Y - \hat{Y})^2$.

We will do the same thing in multiple regression. The prediction model will be:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_k X_k$$

6

## Example: Multiple Predictors

Response Variable: $Y$ = Active pulse (in bpm)

after walking up and down 3 flights of stairs

Predictors: $X_1$ = Resting pulse (in bpm)
$X_2$ = Height (in inches)
$X_3$ = Gender (0 = M, 1 = F)

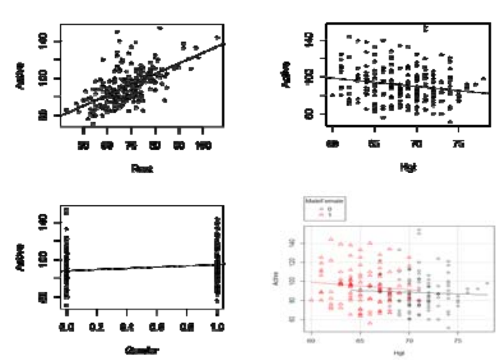
Sample size $n$ = 232, $k$ = 3

Data: **`Pulse.txt`** (has other variables too)

## Correlation "Matrix"

| | Active | Rest | Gender | Hgt |
|---|---|---|---|---|
| Active | 1.0000000 | 0.6041871 | 0.1780192 | -0.1808122 |
| Rest | 0.6041871 | 1.0000000 | 0.1665902 | -0.2426329 |
| Gender | 0.1780192 | 0.1665902 | 1.0000000 | -0.7520590 |
| Hgt | -0.1808122 | -0.2426329 | -0.7520590 | 1.0000000 |

Notice:
Correlations of $X$'s with $Y$ = Active
Correlations of $X$'s with each other
In particular, Gender & Hgt have high |r|




9

## Prediction Equation

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X + \cdots + \hat{\beta}_k X_k$$

where the coefficients
are chosen to minimize: $SSE = \sum \left(Y - \hat{Y}\right)^2$

Example: $Y$ = Active pulse rate

$$\hat{Y} = -6.37 + 1.13 Rest + 0.2685 Hgt + 4.46 Gender$$

10

## Multiple Regression in *R*

```
mymodel=lm(Active~Rest+Hgt+Gender)
```

*"Usual" commands still work.*

```
summary(mymodel)
anova(mymodel)
plot(mymodel)
...
```

## Regression Output

```
> mymodel=lm(Active~Rest+Hgt+Gender)
> summary(mymodel)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.3726    30.8934  -0.206    0.837
Rest          1.1300     0.1023  11.042   <2e-16 ***
Hgt           0.2685     0.4074   0.659    0.511
Gender        4.4610     2.9947   1.490    0.138
---
Residual standard error: 15.01 on 228 degrees of freedom
Multiple R-squared: 0.3724,    Adjusted R-squared: 0.3641
F-statistic:  45.1 on 3 and 228 DF,  p-value: < 2.2e-16
```

12

## Std. Deviation of Error Term = Residual standard error (in R)

$$\varepsilon \sim N(0, \sigma_\varepsilon)$$

$$S_\varepsilon = \sqrt{MSE} = \sqrt{\frac{SSE}{n-k-1}}$$

Given by R

## *R* Regression Output

```
> summary(mymodel)

 Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.3726    30.8934  -0.206    0.837
Rest          1.1300     0.1023  11.042   <2e-16 ***
Hgt           0.2685     0.4074   0.659    0.511
Gender        4.4610     2.9947   1.490    0.138
---
Residual standard error: 15.01 on 228 degrees of freedom
Multiple R-squared: 0.3724,       Adjusted R-squared: 0.3641
F-statistic:  45.1 on 3 and 228 DF,  p-value: < 2.2e-16

> anova(mymodel)
Response: Active
           Df Sum Sq Mean Sq  F value Pr(>F)
Rest        1  29868 29867.9 132.6144 <2e-16 ***
Hgt         1    102   101.8   0.4519 0.5021
Gender      1    500   499.8   2.2189 0.1377
Residuals 228  51351   225.2
```

$df = n - k - 1$

$= 232 - 4 = 228$

SSE MSE

14

## Correlation Matrix

```
> newpulse=pulse.df[,c(1,2,4,7)] #extract columns 1,2,4, and 7
> cor(newpulse)
```

| | Active | Rest | Gender | Hgt |
|---|---|---|---|---|
| Active | 1.0000000 | 0.6041871 | 0.1780192 | -0.1808122 |
| Rest | 0.6041871 | 1.0000000 | 0.1665902 | -0.2426329 |
| Gender | 0.1780192 | 0.1665902 | 1.0000000 | -0.7520590 |
| Hgt | -0.1808122 | -0.2426329 | -0.7520590 | 1.0000000 |

## Some *R* Linear Model Commands (some for later in the course)

Once you have fit, e.g., **`model=lm(Y~X1+X2+X3)`**

**`summary(model)`** → t-tests for coefficients, etc.

**`anova(model)`** → (sequential) sums of squares

**`plot(model)`** → modeling checking plots

**`rstandard(model)`** → standardized residuals

**`rstudent(model)`** → studentized residuals

**`hatvalues(model)`** → leverage ($h_i$)

## **Chapter 3 Section 3.2**

Multiple Regression
- Inference in Multiple Regression
- Partitioning Variability
- Adjusted $R^2$

CI, PI for Multiple Regression

## t-test for Correlation

$H_0: \rho = 0$

$H_1: \rho \neq 0$

$$t.s. = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

No change!

Compare to $t_{n-2}$

Use this to:
(1) Identify potential good predictors of $Y$.
(2) Look for relationships among predictors

## t-test for Slope

*Note:* We now have several "slopes" to test

$H_0$: $\beta_i = 0$

$H_1$: $\beta_i \neq 0$

$$t.s. = \frac{\hat{\beta}_i}{S_{\hat{\beta}_i}}$$

All given by *R* with a p-value

Compare to $t_{n-(k+1)}$

Lose 1 d.f. for each coefficient

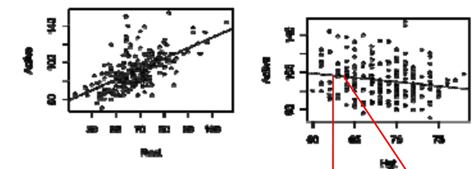
Reject $H_0 \Leftrightarrow$ The $i^{\text{th}}$ predictor is useful in this model, ***given*** others already in the model.

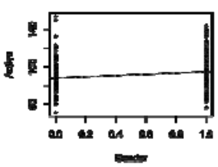
## Example: Hgt and Active

Test #1: Compute and test the *correlation* between Hgt and Active pulse rates.

Test #2: Compute and test the coefficient of Hgt in a multiple regression model (along with Rest and Gender) to predict Active pulse rates.

We will see that we get *different* results. What's going on?





```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 153.4130    22.3120   6.876 5.75e-11 ***
Hgt          -0.9102     0.3264  -2.788  0.00575 **
---
Residual standard error: 18.55 on 230 degrees of freedom
Multiple R-squared: 0.03269, Adjusted R-squared: 0.02849
F-statistic: 7.774 on 1 and 230 DF,  p-value: 0.005745
```

Negative Coefficient/Correlation when *only* Hgt is in model

## Correlation Matrix

| | Active |
|---|---|
| Active | 1.0000000 |
| Rest | 0.6041871 |
| Gender | 0.1780192 |
| Hgt | -0.1808122 |

Rest Gender Hgt

$H_0$: $\rho = 0$

$H_1$: $\rho \neq 0$

$$t.s. = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

$$\frac{-0.181\sqrt{232-2}}{\sqrt{1-0.181^2}} = -2.79$$

DF = 230, p-value = 0.0057

```
> cor.test(Active,Hgt)
data:  Active and Hgt
t = -2.7881, df = 230, p-value = 0.005745
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.30256468 -0.05325377
```

## Regression Output

```
> mymodel=lm(Active~Rest+Hgt+Gender)
> summary(mymodel)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.3726    30.8934  -0.206    0.837
Rest          1.1300     0.1023  11.042   <2e-16 ***
Hgt           0.2685     0.4074   0.659    0.511
Gender        4.4610     2.9947   1.490    0.138
---
Residual standard error: 15.01 on 228 degrees of freedom
Multiple R-squared: 0.3724,    Adjusted R-squared: 0.3641
F-statistic:  45.1 on 3 and 228 DF,  p-value: < 2.2e-16
```

## t-test for Correlation versus t-test for Slope

t-test for correlation: Assesses the linear association between two variables by themselves.

t-test for slope: Assesses the linear association *after accounting for the other predictors in the model.*

In this example, height and gender are correlated. So t-test is for slope of height, *once gender (and rest) already in model.*

## Partitioning Variability

$$\boxed{Y} = \boxed{\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k} + \varepsilon$$

$$\boxed{SSTotal} = \boxed{SSModel} + SSE$$

$$SSModel = \sum(\hat{Y}_i - \bar{Y})^2$$

$$+ \quad SSE = \sum(Y_i - \hat{Y}_i)^2$$

$$SSTotal = \sum(Y_i - \bar{Y})^2$$

25

## ANOVA F-test for Overall Fit

$H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0$ ← "Null" model (no X's used)

$H_1$: Some $\beta_i \neq 0$ ← Effective model

| Source | d.f. | Sum of Squares | Mean Square | t.s. | p-value |
|---|---|---|---|---|---|
| Model | $k$ | *SSModel* | *SSModel/k* | *MSModel/MSE* | $F_{k,n-k-1}$ |
| Error | $n-k-1$ | *SSE* | *SSE/(n-k-1)* | | |
| Total | $n-1$ | *SSTotal* | | | |

## Multiple Regression Model

Population model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

Fitted model (from sample):

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_k X_k$$

We can test:

Individual terms (t-test)

and overall fit (F-test from ANOVA table)

## *R* Regression Output

```
> summary(mymodel)

 Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.3726    30.8934  -0.206    0.837
Rest          1.1300     0.1023  11.042   <2e-16 ***
Hgt           0.2685     0.4074   0.659    0.511
Gender        4.4610     2.9947   1.490    0.138
---
Residual standard error: 15.01 on 228 degrees of freedom
Multiple R-squared: 0.3724,       Adjusted R-squared: 0.3641
F-statistic:  45.1 on 3 and 228 DF,  p-value: < 2.2e-16

> anova(mymodel)
Response: Active
           Df Sum Sq Mean Sq  F value Pr(>F)
Rest        1  29868 29867.9 132.6144 <2e-16 ***
Hgt         1    102   101.8   0.4519 0.5021
Gender      1    500   499.8   2.2189 0.1377
Residuals 228  51351   225.2
```

Test individual terms (*given* other terms)

Test for Overall model

Will learn next what these test.

Note that R does *not* provide SSModel, *and* overall F test is not given by anova command anymore.

28

## *R* Multiple Regression Output, so far we have covered these:

```
> summary(mymodel)

Residuals:
    Min      1Q  Median      3Q     Max
-35.287  -9.637  -2.219   7.221  64.993

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.3726    30.8934  -0.206    0.837
Rest          1.1300     0.1023  11.042   <2e-16 ***
Hgt           0.2685     0.4074   0.659    0.511
Gender        4.4610     2.9947   1.490    0.138
---
Residual standard error: 15.01 on 228 degrees of freedom
Multiple R-squared: 0.3724,    Adjusted R-squared: 0.3641
F-statistic:  45.1 on 3 and 228 DF,   p-value: < 2.2e-16
```

## *R* Multiple Regression Output

```
> anova(mymodel)
Analysis of Variance Table

Response: Active
           Df Sum Sq Mean Sq  F value Pr(>F)
Rest        1  29868 29867.9 132.6144 <2e-16 ***
Hgt         1    102   101.8   0.4519 0.5021
Gender      1    500   499.8   2.2189 0.1377
Residuals 228  51351   225.2
```

"Usual" F test and p-value is in `summary( )` (last slide), not in "anova" output.

"Sequential" sum of squares: New variability "explained" as each predictor is added.

*SSModel* = 29868 + 102 + 500 = 30470 with 3 d.f.

*SSTotal* = 30470 + 51351 = 81821

Coefficient of Multiple Determination

$$R^2 = \frac{SSModel}{SSTotal}$$

Now interpreted as the % of variability in the response variable ($Y$) that is "explained" by a linear combination of these predictors.

$$R^2 = \frac{SSModel}{SSTotal} = \frac{30470}{81821} = 0.3724$$

The % of variability in the response variable (*active pulse*) that is "explained" by a linear combination of the predictors (*resting pulse, height, gender*).

Why Do We Call It $R^2$?

$$R^2 = \frac{SSModel}{SSTotal}$$

For a simple linear model:

If $r$ is the correlation between $X$ and $Y$, then $r^2 = R^2$.

Does this make sense for multiple regression?

Each predictor has a different correlation with $Y$.

Why Do We Call It $R^2$?

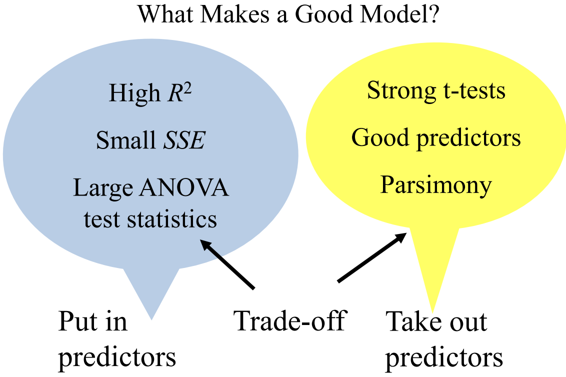
Another way to get $R^2$:

Compute the correlation $r$ between the $Y$ values and the predicted $Y$ values: $r^2 = R^2$.

For a simple model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

$$\Rightarrow \quad |Corr(X,Y)| = Corr(\hat{Y}, Y)$$

What Makes a Good Model?




Two purposes for regression: (1) to model and understand; (2) to predict.

(1) → parsimony, construct a simple model

(2) → increase $R^2$, construct a complex model

But can we believe that a model will yield good predictions for points that weren't used to fit the model in the first place?

Adding additional predictors will:

Increase *SSModel*

Decrease *SSE*

Increase $R^2$

But is the increase in $R^2$ worth it?

## Adjusted $R^2$

Recall:

$$R^2 = \frac{SSModel}{SSTotal} = 1 - \frac{\text{SSE}}{\text{SSTotal}}$$

$$R^2_{adj} = 1 - \frac{\text{SSE}/(n-k-1)}{\text{SSTotal}/(n-1)} = 1 - \frac{\hat{\sigma}^2_\varepsilon}{S^2_Y}$$

(Adjusts for the number of predictors in the model)

## *R* Multiple Regression Output

```
>summary(mymodel)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.3726    30.8934  -0.206    0.837
Gender        4.4610     2.9947   1.490    0.138
Hgt           0.2685     0.4074   0.659    0.511
Rest          1.1300     0.1023  11.042   <2e-16 ***
---
Residual standard error: 15.01 on 228 degrees of freedom
Multiple R-squared:  0.3724,    Adjusted R-squared:  0.3641
F-statistic:  45.1 on 3 and 228 DF,  p-value: < 2.2e-16
```

## *Compare Models using Adjusted R-Squared*

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   26.8130    21.4598   1.249    0.213
Hgt           -0.1830     0.2730  -0.670    0.503
Rest           1.1262     0.1026  10.979   <2e-16 ***
Residual standard error: 15.05 on 229 degrees of freedom
Multiple R-squared:  0.3663,     Adjusted R-squared:  0.3608
F-statistic: 66.18 on 2 and 229 DF,  p-value: < 2.2e-16
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.18257    6.86443    1.92    0.056 .
Rest         1.14288    0.09939   11.50   <2e-16 ***
Residual standard error: 15.03 on 230 degrees of freedom
Multiple R-squared:  0.365,      Adjusted R-squared:  0.3623
F-statistic: 132.2 on 1 and 230 DF,  p-value: < 2.2e-16
```

40

## CI's and PI's for *Y*

Recall: For a simple linear model, when we predict *Y* for a particular value of $X = x_p$

(1) CI for $\mu_Y$ $\quad \hat{Y} \pm t_{\alpha/2} S_\varepsilon \sqrt{\frac{1}{n} + \frac{(x_p - \overline{X})^2}{SSX}}$

Where is the average *Y* for all with $X = x_p$?

(2) PI for individual *Y* $\quad \hat{Y} \pm t_{\alpha/2} S_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(x_p - \overline{X})^2}{SSX}}$

Where are most *Y*'s when $X = x_p$?

What about predicting *Y* with multiple $X_i$'s?

## CI's and PI's for Multiple Regression

For a particular set of predictor values: $(x_1, x_2, \ldots, x_k)$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$$

CI for $\mu_Y$

SE Fit

$$\hat{y} \pm t^* \hat{\sigma}_\varepsilon \sqrt{Stuff}$$

PI for Individual *Y*

$$\hat{y} \pm t^* \hat{\sigma}_\varepsilon \sqrt{1 + Stuff}$$

d.f.$=n-k-1$

## *R*: CI and PI for Multiple Regression

```
 Read the file Pulse
> model<-lm(Active~Rest+Hgt+Gender, data=Pulse)

> newx=data.frame(Rest=63,Hgt=65,Gender=1)

> predict(model,newx,interval="confidence")
      fit      lwr      upr
1 86.7275 83.53862 89.91638

> predict(model,newx,interval="prediction")
      fit      lwr    upr
1 86.7275 56.98501 116.47

> predict(model,Pulse,interval="prediction")
          fit        lwr       upr
1   103.14026  73.35331 132.92721
2    89.25875  59.55785 118.95965
3    83.01580  53.30042 112.73119
Etc…
```

All cases in the “Pulse” dataset