BWH - Biostatistics

Intermediate Biostatistics for Medical Researchers

Robert Goldman
Professor of Statistics
Simmons College

Foundations of Correlation and Regression

Tuesday, March 7, 2017

March 7

Foundations of Correlation and Regression

March 14

Multiple Regression

March 21

Special Topics in Multiple Regression

March 28

Logistic Regression

Statistical Techniques in the Medical Literature

Switzer and Horton (2007)* counted how often various statistical techniques are used in articles in The New England Journal of Medicine.

Technique	1978-1979	1989	2004-2005
None/means/Stdevs	27	12	13
t-tests	44	39	26
Contingency Tables	27	36	53
Non-parametric tests	11	21	27
Odds ratios, Logistic regression	**9**	**22**	**35**
Pearson correlation	**12**	**19**	**3**
Simple linear regression	**8**	**9**	**6**
ANOVA	8	20	16
Multiple regression	**5**	**14**	**51**
Multiple comparisons	3	9	23
Power	3	3	39

*Switzer, S, and Horton, N, What your Doctor Should Know about Statistics, Chance, Vol 20, No 1, 2007.

Learning Objectives

Participants will be able to perform the process of constructing useful linear and logistic regression models

Participants will be able to interpret and draw conclusions from linear regression and logistic regression output.

Participants will understand the conditions for inference in linear and logistic regression and the role of diagnostics in checking these conditions.

Foundations of Correlation and Regression

- Descriptive Aspects of Correlation and Simple Linear Regression
- A Population Model for Regression
- Estimation and Hypothesis Tests in Regression
- Predictions with Regression
- Computer Intensive Methods in Regression

The *hd* Data Set

The *hd* data are from a study of 32 middle-aged patients with heart disease. The data is given below. The object of the exercise is to develop a model to predict systolic blood pressure (SBP) from one (and later more than one) of the other variables in the data set. The variable Smoke takes the value 1 for a smoker and the value 0 for a non-smoker.

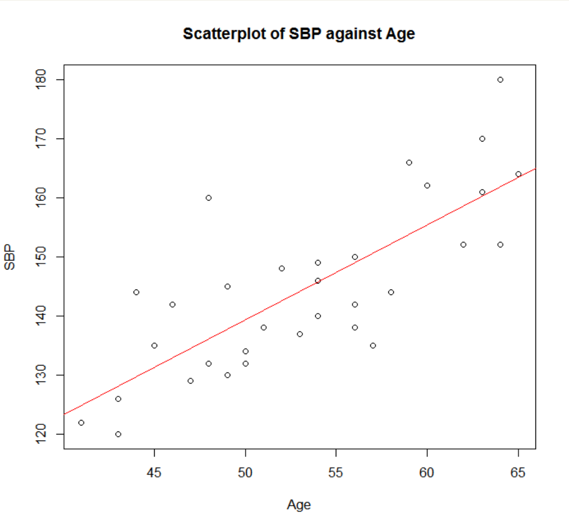
	Y	X				
Pat	SBP	Age	BMI	Height	Smoke	Race
1	135	45	22.8	70	0	Black
2	122	41	24.7	67	0	White
3	130	49	23.9	69	0	Black
4	148	52	27.4	70	0	White
5	146	54	23.3	71	1	White
6	129	47	22.3	76	1	Black
7	162	60	26.9	79	1	White
8	160	48	26.6	67	1	White
9	144	44	20.0	75	0	White
10	180	64	32.1	74	1	Hispanic
11	166	59	28.0	70	1	White
12	138	51	28.9	73	1	White
13	152	64	29.3	64	1	White
14	138	56	26.9	71	0	White
15	140	54	26.4	72	1	White
16	134	50	23.3	67	1	Hispanic
17	145	49	25.3	74	1	Hispanic
:	:	:	:	:	:	:
30	170	63	29.4	81	1	Black
31	152	62	28.5	69	0	White
32	164	65	28.7	66	1	Hispanic

$\bar{Y}=144.53$

Pat	SBP
1	135
2	122
3	130
4	148
5	146
6	129
7	162
8	160
9	144
10	180
11	166
12	?
13	152
14	138
15	140
16	134
17	145
:	:
30	170
31	152
32	164

Suppose one of the 32 Y-values is missing. In the absence of X, how should we estimate this missing value?

```
model <- lm(SBP ~ Age, hd)
plot(SBP ~ Age, hd,
   main = "Scatterplot of SBP against Age")
abline(model, col = "red")
```



```
model

Coefficients:
(Intercept)          Age
     59.092        1.605
```

$$\hat{Y} = 59.09 + 1.605X$$

$$\widehat{SBP} = 59.09 + 1.605Age$$

```
cor(hd$SBP, hd$Age)
[1] 0.7752041
```

$$r = 0.775$$

For each patient we can compute:

(i) the predicted or fitted value ($\hat{Y}$) and

(ii) residual value is $e = Y - \hat{Y}$.

```
Pat          SBP  Age
  :            :    :
  9          144   44
  :            :    :
```

$$Y = 144 \text{ mm}$$

$$\hat{Y} = 59.09 + 1.605\,(44) = 129.69 \text{ mm}$$

$$e = Y - \hat{Y} = 144 - 129.69 = 14.31 \text{ mm}$$

Note that: $\Sigma\, e = 0$ $\quad \bar{e} =$

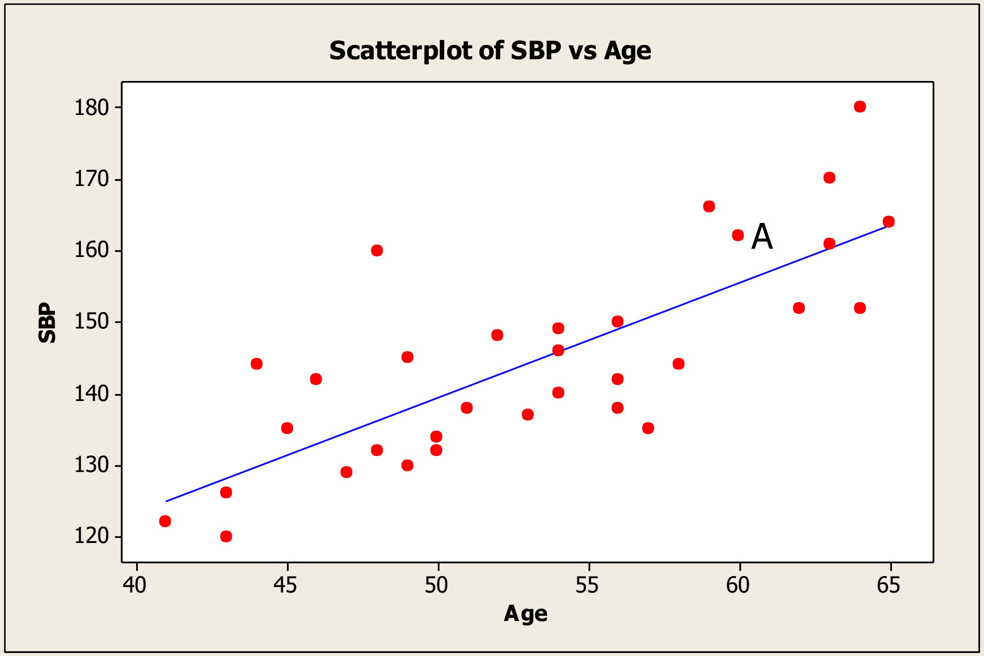
```
fit<- fitted(model)
res<- resid(model)
newdf<- data.frame(hd$SBP,hd$Age,fit,res)
newdf
```

Y X $\hat{Y}$ $e = Y - \hat{Y}$

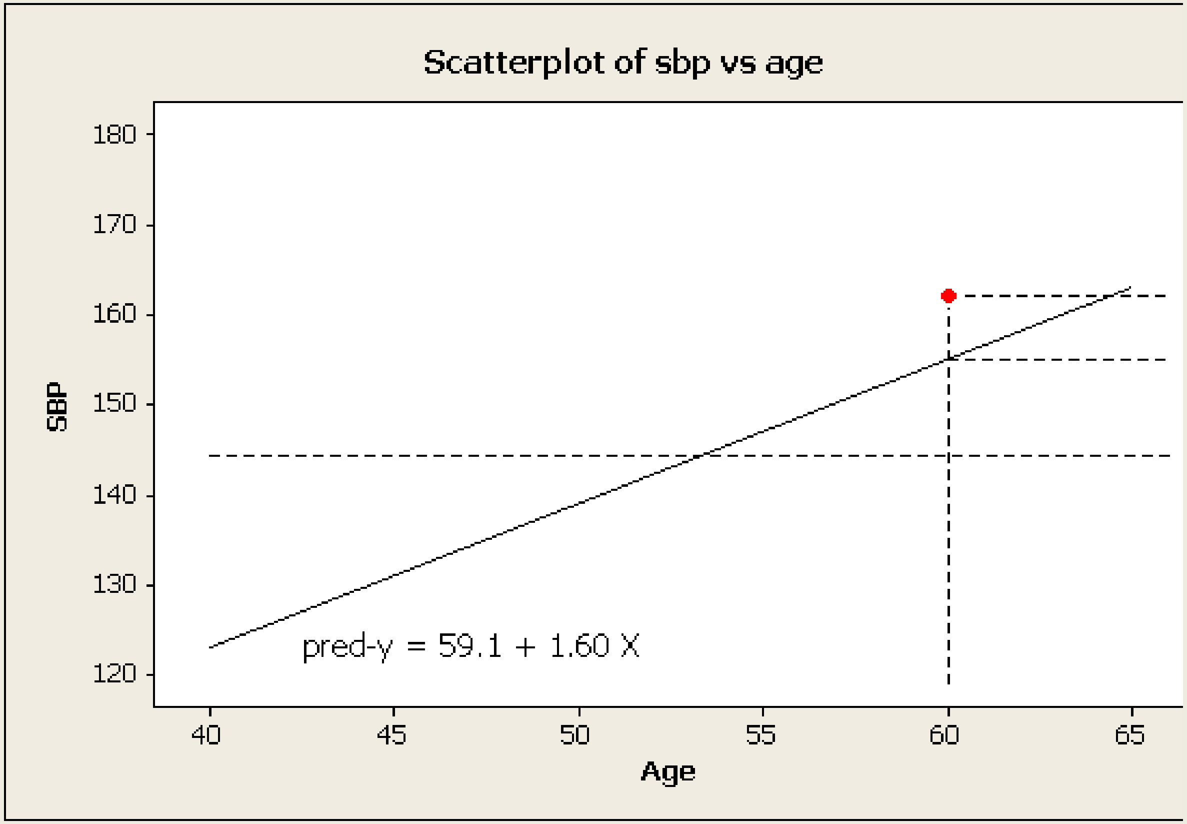
```
   hd.SBP hd.Age      fit        res
1     135     45 131.2941   3.705875
2     122     41 124.8761  -2.876125
3     130     49 137.7121  -7.712125
4     148     52 142.5256   5.474375
5     146     54 145.7346   0.265375
6     129     47 134.5031  -5.503125
7     162     60 155.3616   6.638375
8     160     48 136.1076  23.892375
9     144     44 129.6896  14.310375
10    180     64 161.7796  18.220375
11    166     59 153.7571  12.242875
12    138     51 140.9211  -2.921125
13    152     64 161.7796  -9.779625
14    138     56 148.9436 -10.943625
15    140     54 145.7346  -5.734625
16    134     50 139.3166  -5.316625
17    145     49 137.7121   7.287875
18    142     46 132.8986   9.101375
19    135     57 150.5481 -15.548125
20    142     56 148.9436  -6.943625
21    150     56 148.9436   1.056375
22    144     58 152.1526  -8.152625
23    137     53 144.1301  -7.130125
24    132     50 139.3166  -7.316625
25    149     54 145.7346   3.265375
26    132     48 136.1076  -4.107625
27    120     43 128.0851  -8.085125
28    126     43 128.0851  -2.085125
29    161     63 160.1751   0.824875
30    170     63 160.1751   9.824875
31    152     62 158.5706  -6.570625
32    164     65 163.3841   0.615875

                                --------
                                    0
```

Analysis of Variance in Regression

The basic idea in the ANOVA in regression is to break down a measure of the variability in SBP into (a) a component associated with its relationship to Age, and (b) a residual component associated with variables *other* than Age.



The patient, Albert, is 60 years old and has a SBP of Y = 162 mm. He has a predicted SBP of $\hat{Y}$ = 59.1 + 1.60(60) = 155.1 mm.



$Y = 162$

$\hat{Y} = 155.1$

$\bar{Y} = 144.5$

$$Y - \bar{Y} = (Y - \hat{Y}) + (\hat{Y} - \bar{Y})$$

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

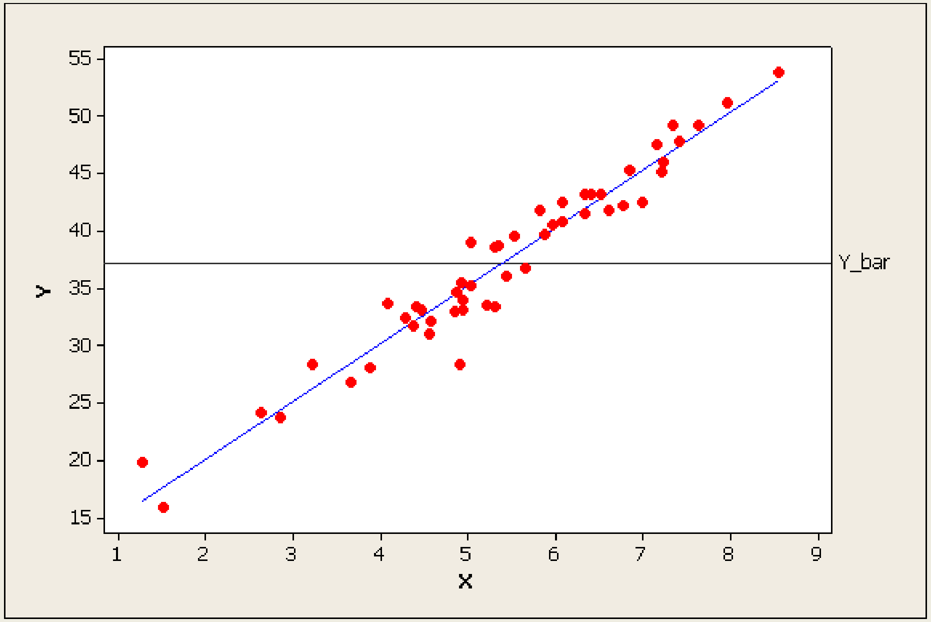
SSTOT SSRES SSREG

SSTOT (The ‘total’ sum of squares) is a measure of the variability in the Y’s. In fact, if you divide SSTOT by n – 1 you will have the variance of Y (the square of the standard deviation). So, SSTOT = $(n – 1)S_Y^2$. It is important to note that the SSTOT depends only on the values for Y; the values for X have no effect on this quantity.

SSRES is a measure of how spread out the points are around the regression line. The $Y - \hat{Y}$ values are simply the residuals. So, SSRES is the sum of the squared residuals. The SSRES captures the degree of spread of the points around the line—the lack of fit of the regression line.

SSREG captures how far the predicted values ($\hat{Y}$’s) are from $\bar{Y}$. If we are not given the X values, our best guess for a new person’s Y value would be $\bar{Y}$, the mean of the Y’s. So, think of SSREG as how much better off we are for knowing the X values.

Graph A



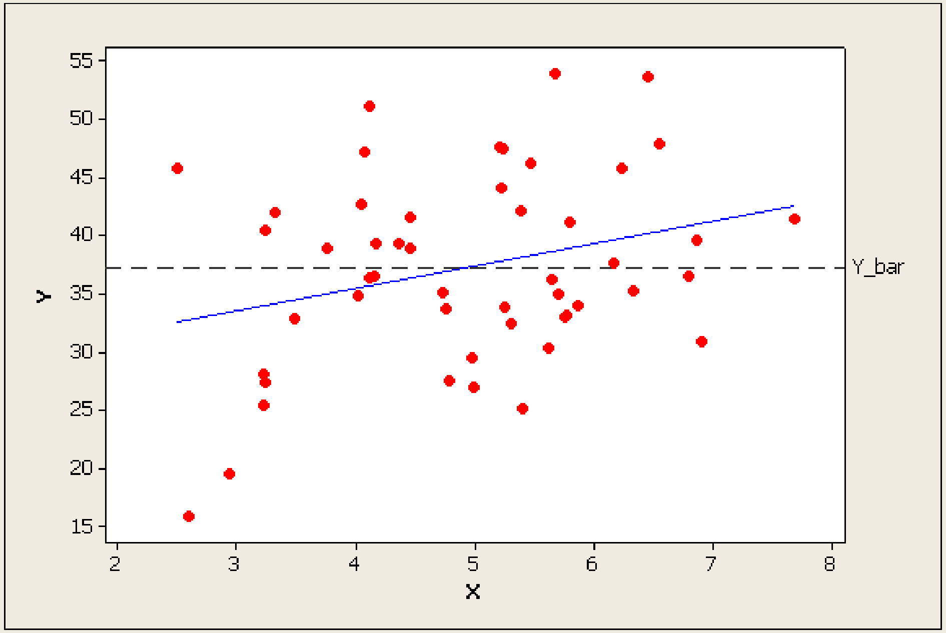
$$\sum (Y - \bar{Y})^2 = \sum (Y - \hat{Y})^2 + \sum (\hat{Y} - \bar{Y})^2$$

The Y's will be close to the $\hat{Y}$'s.

The $\hat{Y}$'s will be far from $\bar{Y}$

	SSTOT	SSRES	SSREG
A	1000	50	950

Graph B



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

The Y's will be far from the $\hat{Y}$'s.

The $\hat{Y}$'s will be close to $\bar{Y}$

	SSTOT	SSRES	SSREG
B	1000	950	50

For the Age, SBP Data:

One variable

Source of Variation	Sum of Squares	df	Mean Square	F	p
Regression	3861.630	1	3861.630	45.177	0.000
Residual	2564.338	30	85.478		
Total	6425.969	31			

n – 1

R

```
anova(model)
Analysis of Variance Table

Response: hd$SBP
          Df Sum Sq Mean Sq F value    Pr(>F)
hd$Age     1 3861.6  3861.6  45.177 1.894e-07
Residuals 30 2564.3    85.5
```

The Coefficient of Determination (r^2)

```
> cor(hd$SBP, hd$Age)
[1] 0.7752041
```

$$r^2 = (0.7752041)^2 = 0.601$$

Source of Variation	Sum of Squares	df	Mean Square	F	p
Regression	3861.630	1	3861.630	45.177	0.000
Residual	2564.338	30	85.478		
Total	6425.969	31			

$$\frac{\text{SSREG}}{\text{SSTOT}} = \frac{3861.630}{6425.969} = 0.601$$

$$r^2 = \frac{SSREG}{SSTOT}$$

The value for r^2 [100 r^2] is the proportion [percentage] of the variability in Y that can be ‘associated with’ the linear relationship between Y and X.

The value for r^2 [100 r^2] is the proportion [percentage] of the variability in Y that can be ‘associated with’ differences among the X values.

The correlation between Y (SBP) and X (Age) is

r = 0.775

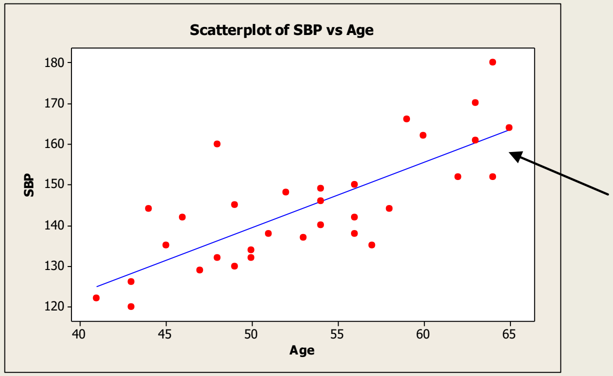
```
> cor(hd$SBP, hd$Age)
[1] 0.7752041
```

But notice also:

```
> cor(hd$SBP, fit)
[1] 0.7752041
```

That is $r(Y, \hat{Y}) = | r(Y, X) |$

Questions about the Population



$\widehat{\text{SBP}} = 59.09 + 1.605\text{Age}$

Our regression line is based on (what we assume is a random) sample of 32 patients with heart disease.

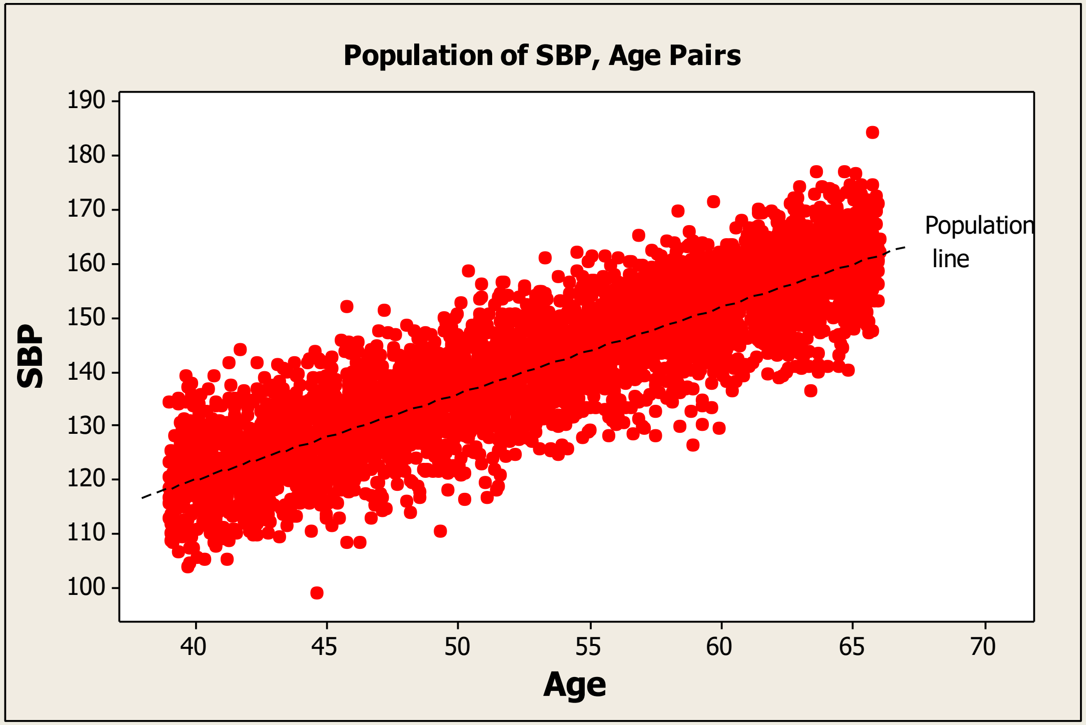
1. Our data seems to suggest a positive, relationship between SBP and Age. But is the relationship statistically significant? That is, might these two variables be independent over the entire population and our data due simply to chance/sampling variability?

2. How close is the sample slope (b_1 = 1.605 mm) to the slope that we would get if we had access to <u>all</u> patients with heart disease?

3. When we use the regression line to make predictions how accurate are these predictions?

What might the ‘population’ look like?

The population model specifies (a) the systematic (in this case, linear) relationship between Age (X) and mean SBP (Y), and (b) the nature of the scatter around that relationship.



$$\text{Mean SBP} = \beta_0 + \beta_1\text{Age}$$

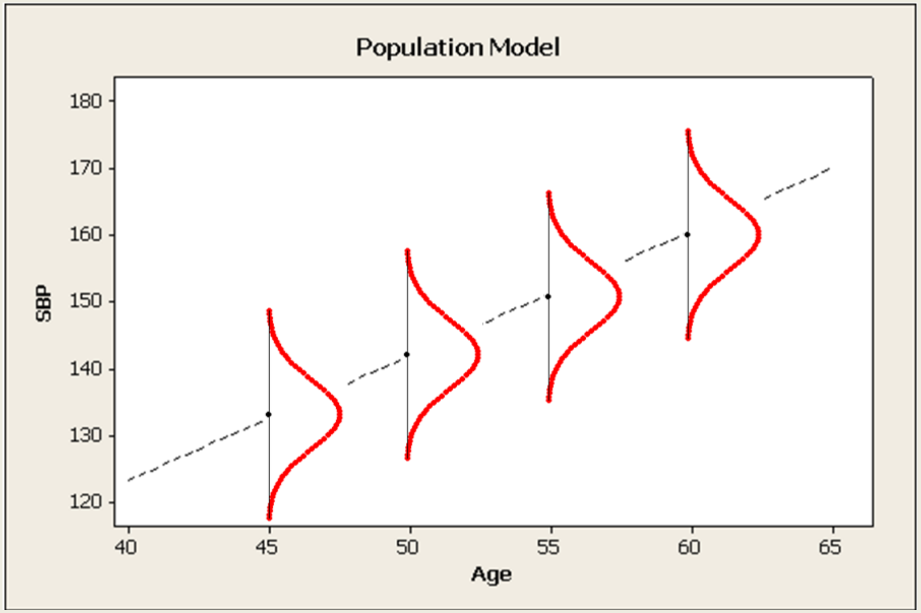
$$\mu_{\text{SBP}} = \beta_0 + \beta_1\text{Age}$$

The Population Model

1. **The linearity condition**: there is a straight line relationship of the form $\mu_{SBP} = \beta_0 + \beta_1 Age$ between age of patient (X) and mean SBP (μ_Y).

2. **The Normality condition**: for any particular age, the distribution of SBP is Normal.

3. **The equal standard deviation condition**: the standard deviation (σ) of SBP is the same for each age.



1. We estimate the (unknown) intercept of the population line, β_0 by b_0 = 59.09 mm.

2. We estimate the (unknown) slope of the population line, β_1 by b_1 = 1.605 mm.

3. We estimate σ^2 from the residuals. Specifically, our estimate of σ^2 is

$$S_e^2 = \frac{\Sigma(Y - \hat{Y})^2}{n - 2} = \frac{\Sigma(e - \bar{e})^2}{n - 2}$$

$$= \text{SSRES}/(n - 2) = \text{MSRES} = 85.478$$

Source of Variation	Sum of Squares	df	Mean Square	F	p
Regression	3861.630	1	3861.630	45.177	0.000
Residual	2564.338	30	85.478		
Total	6425.969	31			

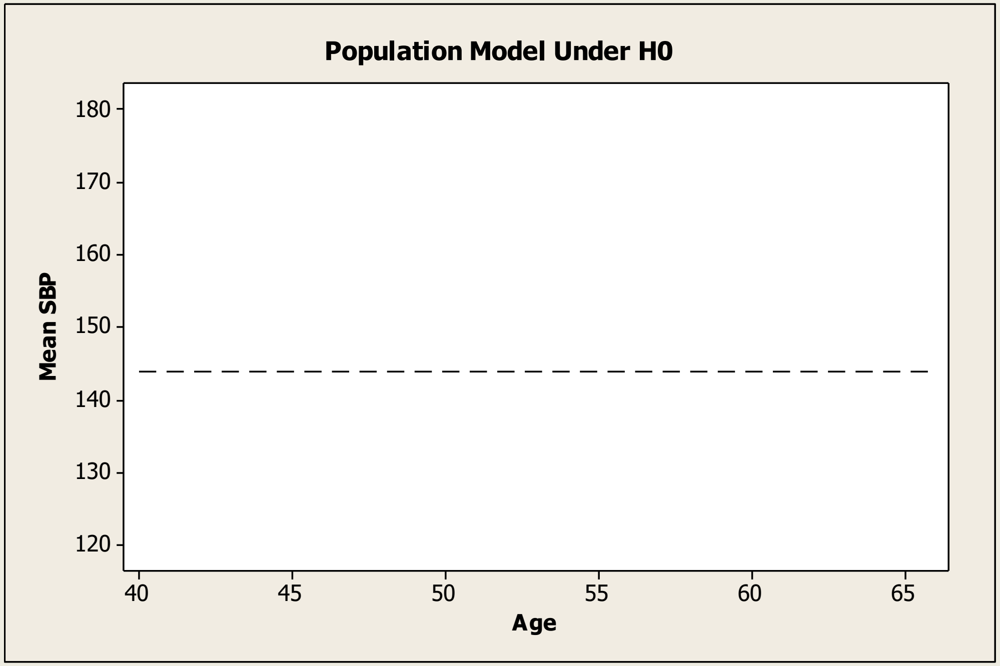
We estimate σ by $\hat{\sigma} = S_e = \sqrt{85.478} = 9.245$ mm.

S_e is called the residual standard error.

Testing for Zero Slope

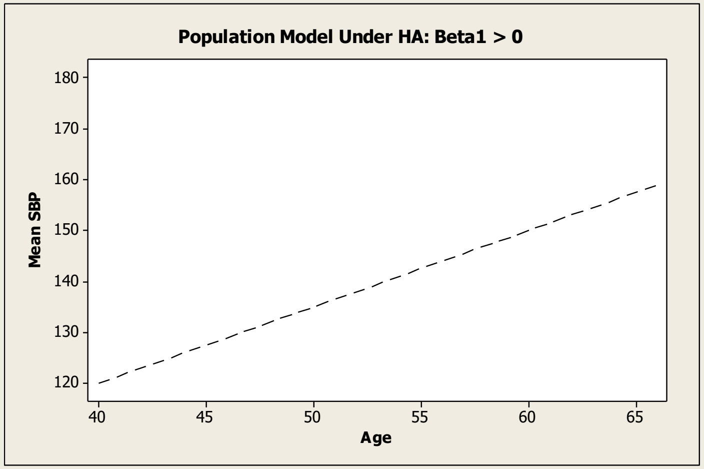
H_0: $\beta_1 = 0$

If H_0 is true: $\mu_{Y|X} = \beta_0 + \beta_1 X = \beta_0 + (0)X = \beta_0$



$\mu_{Y|X} = \beta_0$

H_A: $\beta_1 \neq 0$



$\mu_{Y|X} = \beta_0 + \beta_1 X$

We test the null hypothesis with the t-test for zero slope.

The test statistic is $t = \dfrac{b_1 - 0}{SE(b_1)}$

If the null hypothesis is true and our population model is correct, t has the t_{n-2} distribution.

```
summary(model)


Call:
lm(formula = SBP ~ Age, data = hd)


Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  59.0916    12.8163   4.611 6.98e-05 ***
Age           1.6045     0.2387   6.721 1.89e-07 ***
```

We can reject the null hypothesis at the 1% level of significance. The data suggest that β_1 is significantly greater than 0. This suggests a significant, positive linear relationship between Age and mean SBP

A Confidence Interval for β_1

$$b_1 \quad \pm \quad t_{n-2}\,\mathrm{SE}(b_1)$$

```
confint(model, level = 0.9)

                   5 %      95 %
(Intercept) 37.339086 80.844164
Age          1.199337  2.009663
```

We can be 90% confident that the slope of the population line (the change in mean SBP for each additional year of age) lies between 1.12 and 2.09 mm.

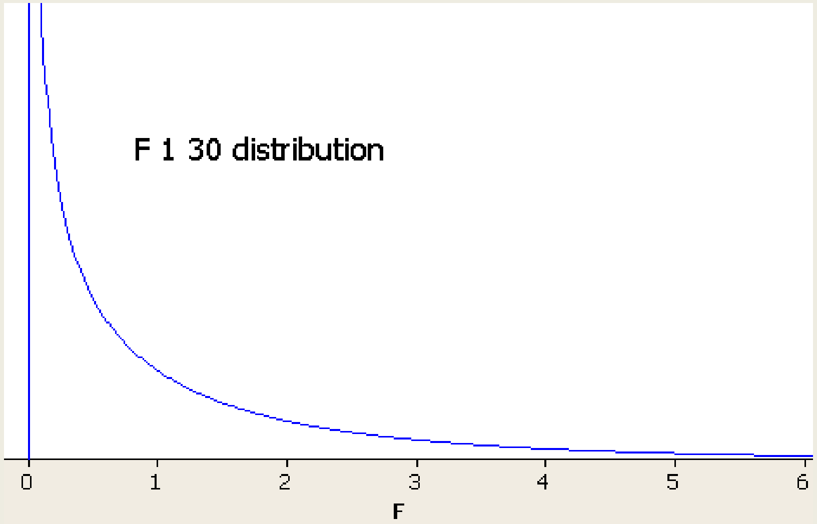
The F test in Regression is based on the ANOVA table.

Source of Variation	Sum of Squares	df	Mean Square	F	p
Regression	3861.630	1	3861.630	45.177	0.000
Residual	2564.338	30	85.478		
Total	6425.969	31			

H_0: $\beta_1 = 0$ H_A: $\beta_1 \neq 0$

If H_0 is true both MSREG and MSRES are estimates for σ^2 and so the ratio F = MSREG/MSRES should be around 1. The larger this ratio, the greater the support for H_A.

The p-value is the area under the F $_{1, 30}$ distribution to the right of the 45.18.



Same conclusion as that of the t test!

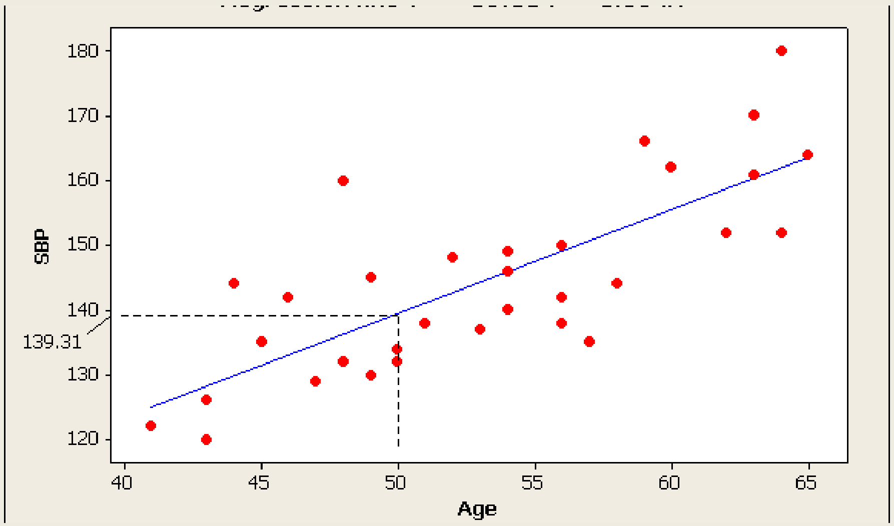
Interestingly, the F test in simple regression is essentially equivalent to the two-sided t test in regression. In fact, you can easily verify that the F value is simply the square of the t value.

$F = 45.18 = 6.72^2 = t^2$

The Accuracy of Predictions

When X = 50:

$\hat{Y}$ = 59. 09 + 1.605 (50) = 139.31 mm.



The value 139.31 is an estimate for Y|50, the (unknown) blood pressure for 50-year old Florence.

The value 139.31 is also an estimate for $\mu_{Y|50}$, the *mean* SBP over all 50-year old patients.

Can we estimate Y|50 and $\mu_{Y|50}$ equally accurately?

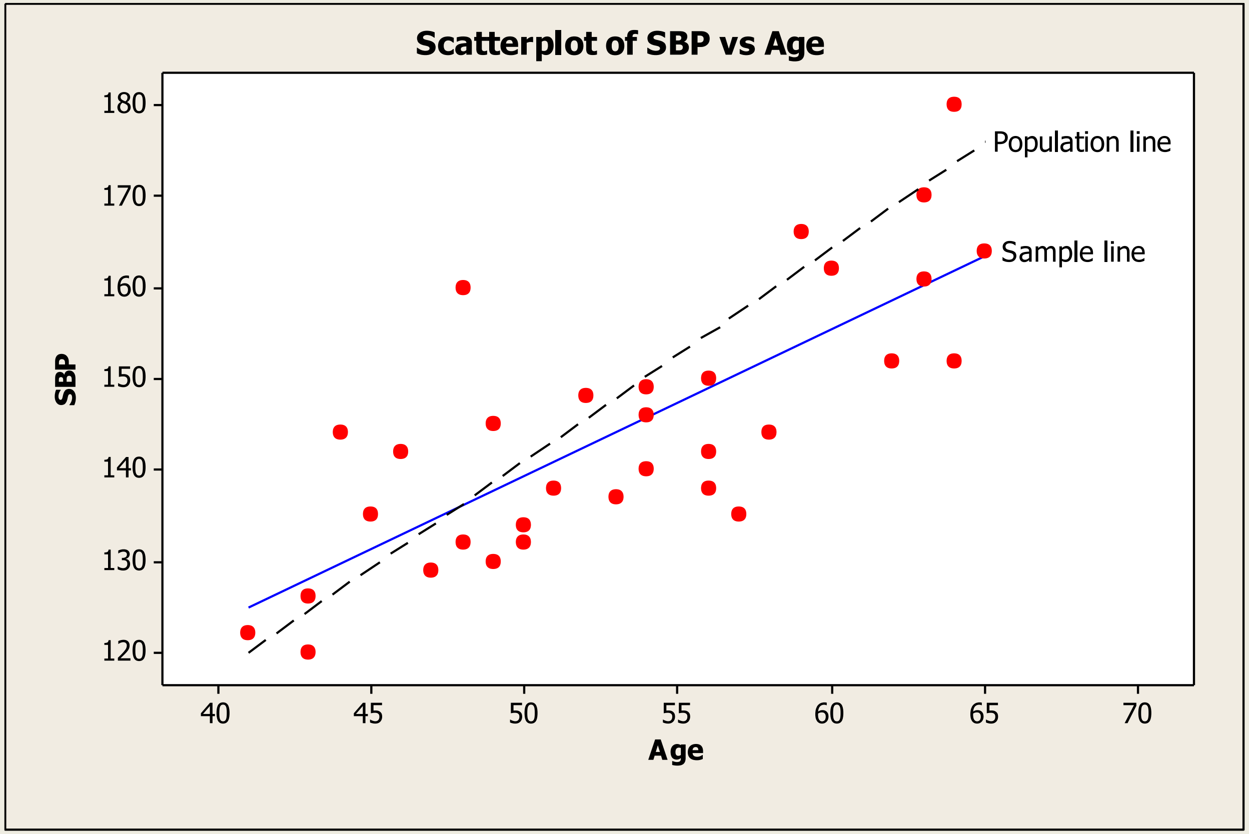
If the population model is valid, a **confidence interval for $\mu_{Y|x}$** is

$$\hat{Y} \quad \pm \quad t_{n-2} \sqrt{S_e^2 \left(\frac{1}{n} + \frac{(x - \bar{X})^2}{\Sigma(X - \bar{X})^2} \right)}$$

X (Age)	$\hat{Y}$	CI for $\mu_{Y\|x}$	Width
40	123.27	116.00 - 130.54	14.54 mm
50	**139.32**	**135.62 - 143.01**	**7.39 mm**
60	155.36	150.67 - 160.05	9.38 mm
70	171.41	162.58 - 180.23	17.65 mm

$\bar{X}$ = 53.25 years

Predictions using the regression line are more accurate the closer the value for x (for which we wish to make the prediction) is to $\bar{X}$.



If the population model is valid, a confidence interval for $\mu_{Y|x}$ is:

$$\hat{Y} \quad \pm \quad M$$

$$\hat{Y} \quad \pm \quad t_{n-2}\, SE(\hat{Y})$$

$$\hat{Y} \quad \pm \quad t_{n-2} \sqrt{S_e^2\left(\frac{1}{n} + \frac{(x - \bar{X})^2}{\Sigma(X - \bar{X})^2}\right)}$$

The form of $SE(\hat{Y})$ reflects the uncertainty due to estimating $\mu_{Y|x}$ from the sample (least-squares) line.

(If we knew the exact form of the population line

$\mu_{SBP} = \quad \beta_0 \; + \; \beta_1 X$

We could compute $\mu_{Y|x}$ for any age (X).)

If the population model is valid, a prediction interval for Y|x is:

$$\hat{Y} \quad \pm \quad M$$

$$\hat{Y} \quad \pm \quad t_{n-2}\, SE(\hat{Y})$$

$$\hat{Y} \;\pm\; t_{n-2}\sqrt{\underbrace{S_e^{2}} + \underbrace{S_e^{2}\left(\frac{1}{n} + \frac{(x - \bar{X})^{2}}{\Sigma(X - \bar{X})^{2}}\right)}}$$

Here, the form of $SE(\hat{Y})$ reflects two sources of variability in estimating Y|x.

One is, as before, due to estimating $\mu_{Y|x}$ from the sample (least-squares) line.

The second source of variability is the fact that the individual values for Y vary around their mean, $\mu_{Y|x}$. We estimate this second source of variability by S_e^2.

Even if we knew the population line (and hence $\mu_{Y|x}$) we would not know Y|x.

Comparing Confidence and Prediction Intervals

	x = 40 yrs	x = 50 yrs
$\hat{Y}$	123.3 mm	139.3 mm
95% CI for $\mu_{Y\|x}$	116.0 - 130.5	135.6 – 143.0
Width	14.5 mm	7.4 mm
95% PI for Y|x	103.0 – 143.5	120.1 – 158.6
Width	40.5 mm	38.5 mm

Obtaining Confidence Intervals for $\mu_{Y|x}$

Estimating the mean SBP for ages 40, 50, 60, and 70, with a 90% confidence interval

```
model <- lm(SBP ~ Age,  data = hd)
a <- c(40, 50, 60, 70)  # new values for Age
k <- data.frame(Age = a)
p <- predict(model, newdata = k, interval =
     "confidence", level = 0.9)
p
       fit      lwr      upr
1 123.2716 117.2289 129.3144
2 139.3166 136.2460 142.3873
3 155.3616 151.4662 159.2570
4 171.4066 164.0751 178.7381



width <- p[,3] - p[,2]
width <- round(width,2)
d <- data.frame(a, round(p,2), width)
d
   a    fit    lwr    upr width
1 40 123.27 117.23 129.31 12.09
2 50 139.32 136.25 142.39  6.14
3 60 155.36 151.47 159.26  7.79
4 70 171.41 164.08 178.74 14.66
```

```
mean(hd$Age)
[1] 53.25
```

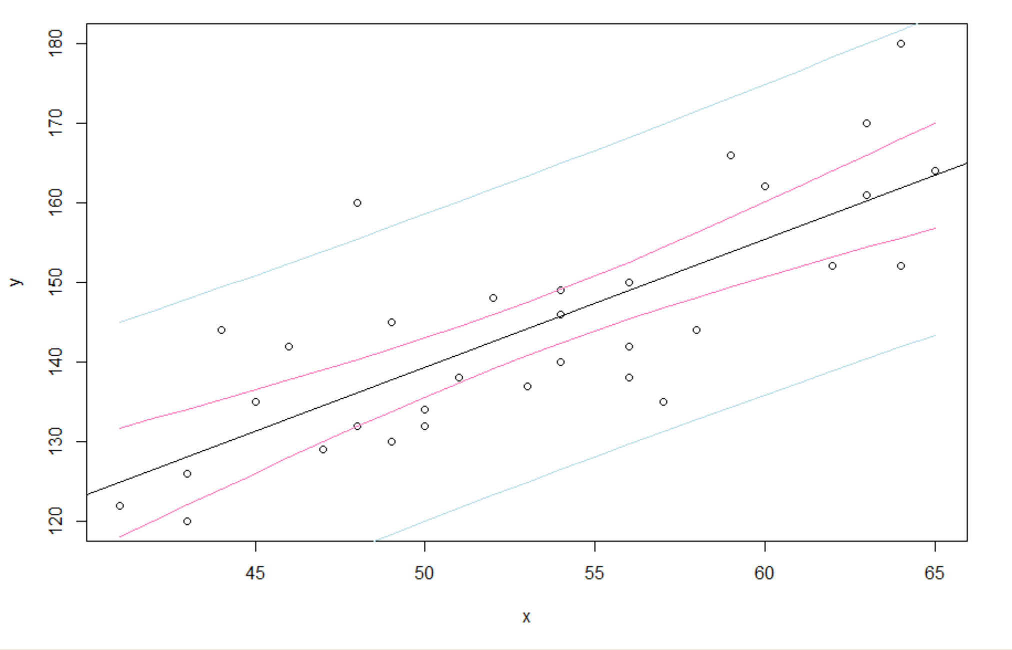
Obtaining Prediction Intervals for Y|x

Estimating the SBP for four individual patients aged 40, 50, 60, and 70 respectively, with a 90% prediction interval.

```
model <- lm(SBP ~ Age,  data = hd)
a <- c(40, 50, 60, 70)  # new values for Age
k <- data.frame(Age = a)
p <- predict(model, newdata = k, interval =
     "predict", level = 0.9)
p
       fit      lwr      upr
1 123.2716 106.4564 140.0868
2 139.3166 123.3271 155.3061
3 155.3616 139.1934 171.5298
4 171.4066 154.0865 188.7268



width <- p[,3] - p[,2]
width <- round(width,2)
d <- data.frame(a, round(p,2), width)
d
   a    fit    lwr    upr width
1 40 123.27 106.46 140.09 33.63
2 50 139.32 123.33 155.31 31.98
3 60 155.36 139.19 171.53 32.34
4 70 171.41 154.09 188.73 34.64
```

Obtaining a Confidence/Prediction Band

```
df = data.frame(x = hd$Age, y = hd$SBP)
mod = lm(y ~ x, data = df)
allx = seq(min(df$x), max(df$x))
k = data.frame(x=allx)
preds = predict(mod, k, interval =
"confidence")
preds2 = predict(mod, k, interval = "predict")
# plot
plot(y ~ x, data = df)
# model
abline(mod)
# intervals
lines(allx, preds[ ,3], col = "hotpink")
lines(allx, preds[ ,2], col = "hotpink")
lines(allx, preds2[ ,3], col = "lightblue")
lines(allx, preds2[ ,2], col = "lightblue")
```



Options Available with the lm Function

model <- lm(y ~ x, df)

summary(model) Displays detailed results for the fitted model

coef(model) Lists the intercept and the slope(s) for the fitted model

confint(model) Provides the CI’s for the population model slopes (95%)

fitted(model) Lists the predicted/fitted values in a fitted model

resid(model Lists the residual values in a fitted model

anova(model) Generates an ANOVA table for a fitted model, or an ANOVA table comparing two or more fitted models

Computer Intensive Inference Methods in Regression

The ‘traditional’ inference methods in regression rely on the validity of the linearity, Normality, and the equal standard deviation conditions.

If these conditions are valid, the statistical theory tells us that the sampling distribution of the statistic:

$$\frac{b_1 - \beta_1}{SE(b_1)}$$

is a t distribution with n – 2 degrees of freedom.

This structure is the basis for the confidence intervals and tests we have looked at. This structure is suspect on a number of grounds:

What if one or more of the conditions are invalid?

The entire structure is opaque to non-statisticians

A more transparent approach is to use computer-intensive methods

A Bootstrap Confidence Interval for β_1

(a) Select many, many samples of size n = 32 with replacement from the 32 SBP, Age pairs in *hd.csv*.

(b) For each sample, compute and store the slope (b_1) of the regression line relation SBP to Age. This will form a pseudo-sampling distribution for $b_{1.}$

(c) Compute the mean of the b_1's [b_1*] and the standard deviation of the b_1's [S*]

(d) If the sampling distribution of the b_1's looks approximately bell-shaped, an attractive CI for β_1 can be obtained from the appropriate percentiles of the distribution of b_1. For example a 95% bootstrap CI for β_1 is formed by the 2.5th percentile and the 97.5th percentile of the b_1's

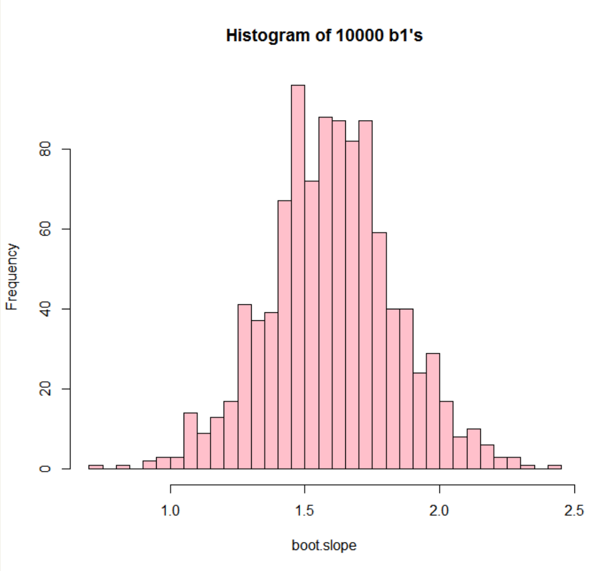
$b_{1,\ 0.025}$ to $b_{1,\ 0.975}$

Here is the annotated R script that will compute and store 10000 bootstrap values for b_1.

```
boot.slope <- numeric(1000)
for (i in 1:1000)
{
# take a sample of 32 rows with replacement
s <- hd[sample(1:32, 32, replace = T),]
# now regress SBP on Age
l <- lm(SBP ~ Age, data = s)
# c is a vector containing the intercept and the slope
c <- coef(l)
boot.slope[i] <- c[2]
}
```

Here is a histogram of the 1000 values for b_1.

```
hist(boot.slope, main =
 "Histogram of 10000 b1's", breaks = 30,
  col = "pink")
```



Here is the 95% bootstrap percentile interval for β_1

```
quantile(boot.slope, c(0.025, 0.975))

    2.5%    97.5% 
1.110327 2.087187 
```

As a reminder, here is our theory-based 95% confidence interval for β_1.

```
confint(model)

                2.5 %    97.5 %
(Intercept) 32.917327 85.265923
Age          1.116977  2.092023
```

A Permutation Test of H_0: $\beta_1 = 0$

```
model <- lm(SBP ~ Age, data = hd)
model


Coefficients:
(Intercept)          Age
     59.092        1.605
```

If the null hypothesis is true how unusual is a sample slope as large as $b_1 = 1.605$?

The key idea here is that, if the null hypothesis is true, Age and SBP are linearly independent, and so the assignment of the Ages to the SBP's can be viewed as occurring at random.

We take advantage of this fact to produce a permutation test of H_0: $\beta_1 = 0$ against the (conservative) alternative hypothesis H_A: $\beta_1 \neq 0$.

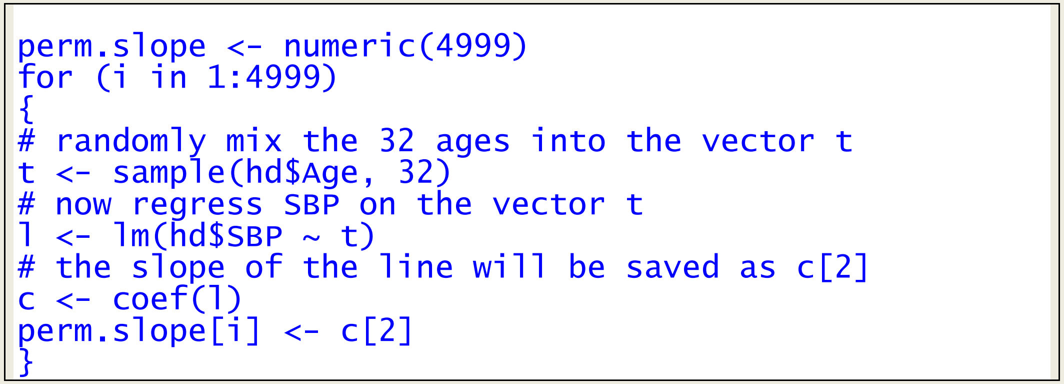
(a) Randomly assign the 32 ages to the 32 SBP's. For this assignment compute and save the slope (b_1) of the regression line relating SBP to Age.

(b) Repeat this process a large number (999, 4999, etc) of times.

(c) Compute the p-value as the proportion of occasions on which the slope of the sample regression line is more extreme than the slope of the line for the observed data.

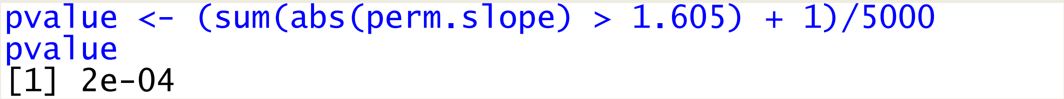
You need to include the observed occasion in both the numerator and the denominator of this proportion.

Here is the annotated R script that will compute and store 4999 values for b_1 under the assumption that H_0: $\beta_1 = 0$ is true.



```
perm.slope <- numeric(4999)
for (i in 1:4999)
{
# randomly mix the 32 ages into the vector t
t <- sample(hd$Age, 32)
# now regress SBP on the vector t
l <- lm(hd$SBP ~ t)
# the slope of the line will be saved as c[2]
c <- coef(l)
perm.slope[i] <- c[2]
}
```

For a two-sided test, compute the p-value with the command:

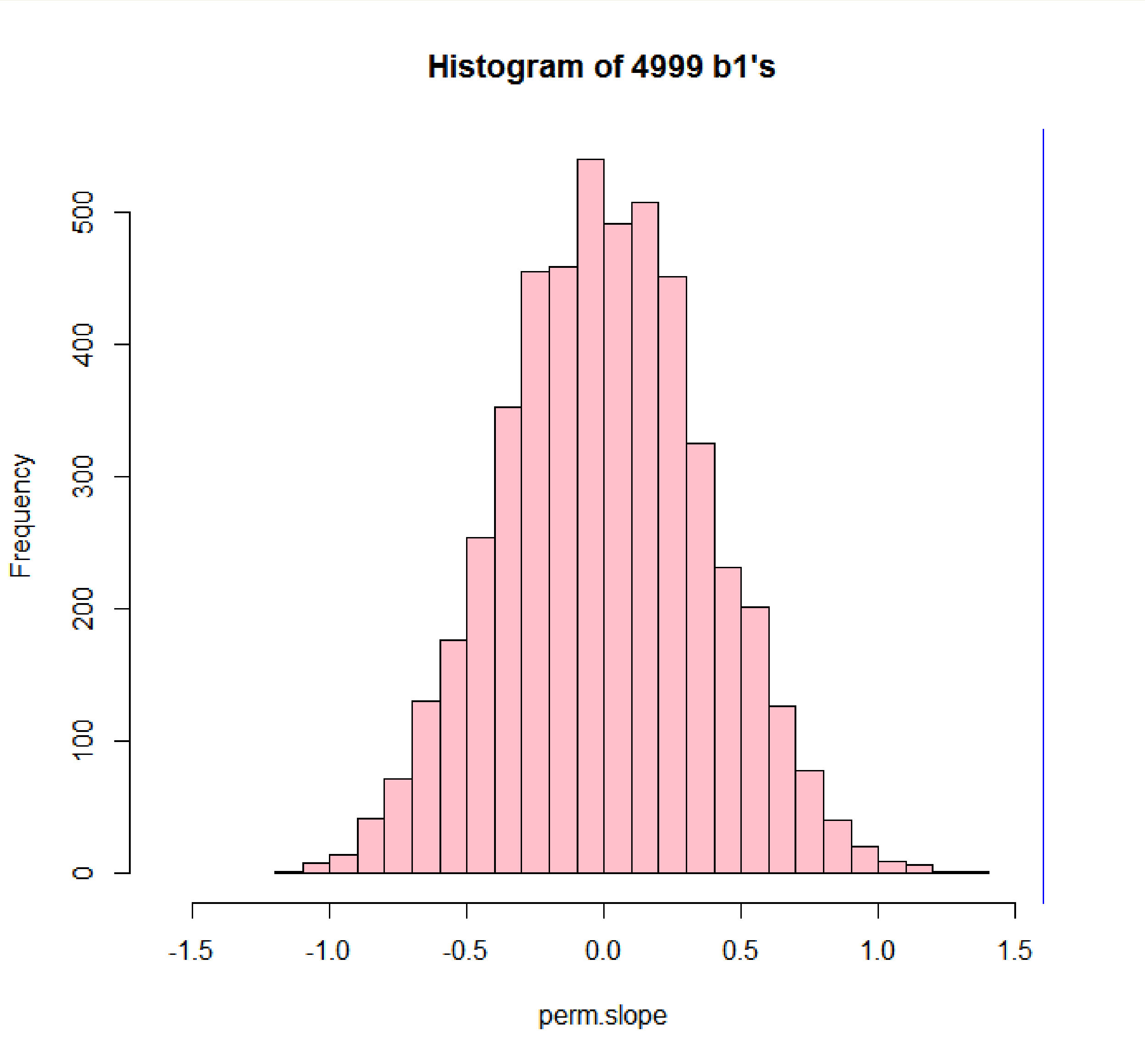


```
pvalue <- (sum(abs(perm.slope) > 1.605) + 1)/5000
pvalue
[1] 2e-04
```

The p-value is 1/5000. Assuming Age and SBP are linearly independent, not one of the 4999 random assignments of Ages to SBP created a slope as large as the one we observed (1.605).

We can reject the null hypothesis at the 1% level of significance.

```
hist(perm.slope, main =
 "Histogram of 4999 b1's", breaks = 30,
  col = "pink")
abline(v = 1.6, col = "blue")
```



Helpful Guides to R

http://polisci.msu.edu/jacoby/apsa07/graphics/refers/Maindonald,%20Using%20R.pdf

http://cran.r-project.org/doc/contrib/Owen-TheRGuide.pdf

http://cran.r-project.org/doc/manuals/r-release/R-intro.pdf

http://cyclismo.org/tutorial/R/

http://ww2.coastal.edu/kingw/statistics/R-tutorials/

http://cran.r-project.org/doc/contrib/Torfs+Brauer-Short-R-Intro.pdf

http://www.tfrec.wsu.edu/TFREConly/r4beginners_v3.pdf