Chapter 2

STAT 3240

Michael McIsaac

UPEI

Learning Objectives for Sections 2.1-2.3

After Sections 2.1-2.3, you should be able to

- Compute and interpret confidence intervals for β_0 and β_1
- Conduct and interpret hypothesis tests concerning β_0 and β_1
- Define power and explain how it impacts inference

2: Inferences in Regression and Correlation Analysis

We assume that the *normal error regression model* is applicable:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- β_0 and β_1 are parameters
- X_i are known constants
- ε_i are independent $N(0, \sigma^2)$.

2.1. Inferences Concerning β_1

At times, tests concerning β_1 are of interest, particularly one of the form

$$H_0 : \beta_1 = 0 \qquad vs \qquad H_a : \beta_1 \neq 0$$

2.1. Inferences Concerning β_1

If we assume that

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \text{with } E[\varepsilon_i] = 0,$$

then $E[Y] = \beta_0 + \beta_1 X$. Thus, $H_0 : \beta_1 = 0$ is equivalent to

$$H_0 : E[Y] = \beta_0,$$

so the null hypothesis is equivalent to

$$H_0 : \text{There is no linear association between Y and X}$$

2.1. Inferences Concerning β_1

When we make the stronger assumption of the *normal error regression model*:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \text{ with } \varepsilon_i \sim N(0, \sigma^2),$$

then $H_0 : \beta_1 = 0$ is actually equivalent to the stronger statement that

$$H_0 : Y_i = \beta_0 + \varepsilon_i, \text{ where } \varepsilon_i \sim N(0, \sigma^2)$$

i.e.

$$H_0 : \text{There is no relation of any type between Y and X}$$

Sampling Distribution of b_1

β_1 can be estimated in a given sample by

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

Every sample would result in a different point estimate b_1 even if the *predictor variables*, X, were held constant across samples.

Under the *normal error regression model* ($Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ with $\varepsilon_i \sim N(0, \sigma^2)$), the sampling distribution of the *estimator* b_1 is

$$b_1 \sim N(\beta_1,\ \sigma\{b_1\}), \qquad \text{with } \sigma^2\{b_1\} = \frac{\sigma^2}{\sum(X_i - \bar{X})^2}$$

Therefore,

$$\frac{b_1 - \beta_1}{\sigma\{b_1\}} \sim N(0, 1).$$

Sampling Distribution of b_1 with $\sigma\{b_1\}$ unknown

When $\sigma\{b_1\}$, is estimated by $s\{b_1\}$, we rely on

$$\frac{b_1 - \beta_1}{s\{b_1\}} \sim t(n-2), \qquad \text{where } s^2\{b_1\} = \frac{MSE}{\sum(X_i - \bar{X})^2}$$

Note: There are n unique pieces of data, but 2 parameters are estimated in the regression model (β_0 and β_1), so there are $n-2$ degrees of freedom.

Confidence Interval for β_1

The $1-\alpha$ confidence limits for β_1 are $\boxed{b_1 \pm t(1-\alpha/2; n-2)s\{b_1\}}$.

Data Set for Warm Up Questions: SHS

The Canadian Survey of Household Spending is carried out annually across Canada.

(http://dli-idd-nesstar.statcan.gc.ca.proxy.library.upei.ca/webview/)

The main purpose of the survey is to obtain detailed information about household spending. Information is also collected about dwelling characteristics as well as household equipment.

The survey data are used by the following groups:

- Government departments use the data to help formulate policy;
- Community groups, social agencies and consumer groups use the data to support their positions and to lobby governments for social changes;
- Lawyers and their clients use the data to determine what is fair for child support and other compensation;
- Labour and contract negotiators rely on the data when discussing wage and cost-of-living clauses;
- Individuals and families can use the data to compare their spending habits with those of similar types of households.

```
###A subset of the latest Survey of Household Spending data are disp
spending_subset %>% datatable()
```

Show 20 entries Search:

	province	type_of_dwelling	income	marital_status	age_gr
1	NL	single_detached	68000	never_married	30-34
2	NL	single_detached	48000	never_married	25-29
3	NL	single_detached	30000	married	35-39
4	NL	row_house	30000	never_married	30-34
5	NL	single_detached	35000	married	25-29
6	NL	single_detached	26000	married	25-29
7	NL	single_detached	26000	other	55-59

Showing 1 to 20 of 30 entries Previous 1 2 Next

We are interested in the potential relationship between the income of working Canadians and the amount that they spend on clothing in a year.

Income and Clothing Expenditure for a small subset of the Survey of Household Spending are displayed below:

```
xyplot(clothing_expenditure~income, data=spending_subset,  type=c("p
```

A preliminary regression analysis follows:

```
clothing_model = lm(clothing_expenditure~income, data=spending_subse
msummary(clothing_model)
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.064e+03  7.832e+02   1.358   0.1853
## income      3.050e-02  1.651e-02   1.847   0.0753 .
##
## Residual standard error: 1663 on 28 degrees of freedom
## Multiple R-squared:  0.1086,    Adjusted R-squared:  0.07681
## F-statistic: 3.413 on 1 and 28 DF,  p-value: 0.07528
```

- Give a 95% CI for β_1
- Interpret your 95% CI for β_1 in the context of this problem
- We are 95% confident that for each $100 increase in income, people on average spend from 3 cents less to $6 more on clothing.

- Interpret your 95% CI for β_1 in the context of this problem
- We are 95% confident that the mean of clothing expenditures increases by an amount found between -0.00000 and 0.06099 for each additional unit in income. However, we cannot speak about spending a negative amount of money, so that would be extrapolating, so it is more reasonable to not interpret the negative part of the CI.
- There is a 95% chance that β_1 would fall in the interval (-0.0331, 0.06430)
- If we replicated the same study multiple times with different random samples and computed a confidence interval for each sample, we would expect 95% of the confidence intervals to say that there is no significant relationship between income and clothing expenditure.
- It is hard to interpret since 0 lays inside the interval.

CDI linear model: physicians vs hospital beds

This data set provides selected county demographic information (CDI) for 440 of the most populous counties in the United States.

Each line of the data set has an identification number with a county name and state abbreviation and provides information on 14 variables for a single county.

Counties with missing data were deleted from the data set.

```
cdi %>% datatable()
```

Show 20 entries Search:

	county	state	land_area	population	pop_18_to_34
1	Los_Angeles	CA	4060	8863164	32
2	Cook	IL	946	5105067	29.
3	Harris	TX	1729	2818199	31.
4	San_Diego	CA	4205	2498016	33
5	Orange	CA	790	2410556	32.
6	Kings	NY	71	2300664	28.
7	Maricopa	AZ	9204	2122101	29.

Showing 1 to 20 of 440 entries

Previous 1 2 3 4 5 ... 22 Next

```
mod_physician_beds = lm(number_physicians ~ number_hospital_beds, da
summary(mod_physician_beds)
```

```
##
## Call:
## lm(formula = number_physicians ~ number_hospital_beds, data = cdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3133.2  -216.8   -32.0    96.2  3611.1
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -95.93218   31.49396  -3.046  0.00246 **
## number_hospital_beds   0.74312    0.01161  63.995  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 556.9 on 438 degrees of freedom
## Multiple R-squared:  0.9034,    Adjusted R-squared:  0.9032
## F-statistic:  4095 on 1 and 438 DF,  p-value: < 2.2e-16
```

- What is a 95% Confidence Interval for β_1?
- How do we intepret this 95% Confidence Interval for β_1?

CDI linear model: physicians vs population

```
mod_physician_pop = lm(number_physicians ~ population, data=cdi)
summary(mod_physician_pop)
```

```
##
## Call:
## lm(formula = number_physicians ~ population, data = cdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1969.4  -209.2   -88.0    27.9  3928.7
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.106e+02  3.475e+01  -3.184  0.00156 **
## population   2.795e-03  4.837e-05  57.793  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 610.1 on 438 degrees of freedom
## Multiple R-squared:  0.8841,    Adjusted R-squared:  0.8838
## F-statistic:  3340 on 1 and 438 DF,  p-value: < 2.2e-16
```

- What is a 95% Confidence Interval for β_1?
- How do we intepret this 95% Confidence Interval for β_1?

CDI linear model: physicians vs total income

```
mod_physician_income = lm(number_physicians ~ total_personal_income,
summary(mod_physician_income)
```

```
##
## Call:
## lm(formula = number_physicians ~ total_personal_income, data = cdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1926.6  -194.5   -66.6    44.2  3819.0
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -48.39485   31.83333   -1.52    0.129
## total_personal_income   0.13170    0.00211   62.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 569.7 on 438 degrees of freedom
## Multiple R-squared:  0.8989,    Adjusted R-squared:  0.8987
## F-statistic:  3895 on 1 and 438 DF,  p-value: < 2.2e-16
```

- What is a 95% Confidence Interval for β_1?
- How do we intepret this 95% Confidence Interval for β_1?

CDI: confint

```
confint(mod_physician_beds, level=.95)
```

```
##                          2.5 %      97.5 %
## (Intercept)         -157.830247 -34.0341222
## number_hospital_beds    0.720294   0.7659389
```

```
confint(mod_physician_pop, level=.95)
```

```
##                     2.5 %       97.5 %
## (Intercept) -178.92443176 -42.34512271
## population     0.00270036   0.00289049
```

```
confint(mod_physician_income, level=.95)
```

```
##                           2.5 %     97.5 %
## (Intercept)          -110.9599089 14.1702110
## total_personal_income    0.1275537  0.1358487
```

Tests Concerning β_1

Since

$$\frac{b_1 - \beta_1}{s\{b_1\}} \sim t(n-2),$$

tests concering β_1 can be set up in ordinary fashion using the t distribution.

I.e., if $\beta_1 = 0$, then $\frac{b_1 - 0}{s\{b_1\}} \sim t(n-2)$.

So, we can test whether $\beta_1 = 0$ by looking at whether

$$t^* = \frac{b_1}{s\{b_1\}}$$

looks like it comes from a $t(n-2)$ distribution.

- p-value is the probability that we would get an estimate of β_1 that as far from 0 as the one that we saw, if $\beta_1 = 0$.
- p-value $= P\left(\left|\frac{b_1 - \beta_1}{s\{b_1\}}\right| > |t^*| | \beta_1 = 0\right) = P(|t(n-2)| > |t^*|)$

Remember that the p-value is the probability that we would see something as *extreme* as what we saw (t^*) if the null hypothesis were true.

"Extreme" is defined by the alternate hypothesis:

- $H_a : \beta_1 > 0 \quad \Longrightarrow \quad p^* = P(t_{n-2} > t^*)$
- $H_a : \beta_1 < 0 \quad \Longrightarrow \quad p^* = P(t_{n-2} < t^*)$
- $H_a : \beta_1 \neq 0 \quad \Longrightarrow \quad p^* = P(t_{n-2} > |t^*| \text{ or } t_{n-2} < -|t^*|) = P(|t_{n-2}| > |t^*|)$

SHS: Testing β_1

```
clothing_model = lm(clothing_expenditure~income, data=spending_subse
msummary(clothing_model)
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.064e+03  7.832e+02   1.358   0.1853
## income      3.050e-02  1.651e-02   1.847   0.0753 .
##
## Residual standard error: 1663 on 28 degrees of freedom
## Multiple R-squared:  0.1086,    Adjusted R-squared:  0.07681
## F-statistic: 3.413 on 1 and 28 DF,  p-value: 0.07528
```

```
confint(clothing_model)
```

```
##                     2.5 %       97.5 %
## (Intercept) -540.57211248 2.667947e+03
## income        -0.00331827 6.431123e-02
```

```r
confint(clothing_model) %>% round(5)
```

```
##                   2.5 %     97.5 %
## (Intercept) -540.57211 2667.94683
## income        -0.00332    0.06431
```

- Based on this CI, a two-sided test of $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$ at the 5% significance level would result in __

- Is clothing expenditure related to income? Justify your answer.

- At a 95% confidence level, it can not be concluded that clothing expenditure and income are related. This is because a slope of zero is in the confidence interval, thus it is possible that the slope is zero indicating no relationship. Therefore, the null hypothesis cannot be rejected.
- No, because we failed to reject the null hypothesis. Therefore the slope may not be effected by the relation of clothing expenditure and income.
- Since we fail to reject the null hypothesis and ß1 = 0 , the clothing expenditure is not related to income.
- It could be no relationship between clothing expenditure and income. Because the B1 could equal =0,

- Is clothing expenditure related to income? Justify your answer.
- According to the plot I would say that they are related but since 0 is in our CI they could be unrelated. According to my other calculation they seem to be positively related.
- Yes, clothing expenditure is related to income because according to the graph displayed above the higher your income the more the is spent on clothing.
- Yes it is, the R-squared can show the trend.

CDI: Testing β_1

```
tab_model(mod_physician_beds, show.ci=.95)
```

	number_physicians			
Predictors	*Estimates*	*std. Error*	*Statistic*	*p*
(Intercept)	-95.93 (-157.66 – -34.21)	31.49	-3.05	0.002
number_hospital_beds	0.74 (0.72 – 0.77)	0.01	63.99	<0.001
Observations	440			
R^2 / R^2 adjusted	0.903 / 0.903			

- Is $\beta_1 = 0$?
- Is $\beta_1 = .75$?
- If a county were to buy two additional hospital beds, what would you expect to happen to the number of physicians?

2.2 Inferences Concerning β_0

CDI: Inferences Concerning β_0

```
tab_model(mod_physician_beds, show.ci=.95)
```

	number_physicians			
Predictors	*Estimates*	*std. Error*	*Statistic*	*p*
(Intercept)	-95.93 (-157.66 – -34.21)	31.49	-3.05	0.002
number_hospital_beds	0.74 (0.72 – 0.77)	0.01	63.99	<0.001
Observations	440			
R^2 / R^2 adjusted	0.903 / 0.903			

- Is $\beta_0 = 0$?
- If a county were to have 0 hospital beds, what would you expect the number of physicians to be?

CDI: Inferences Concerning β_0 for centered predictor

```
###create a new column in the data set which is equal to the number
cdi = cdi %>% mutate(number_hospital_beds_c = number_hospital_beds-m
mod_physician_beds_centered = lm(number_physicians ~ number_hospital
tab_model(mod_physician_beds_centered, show.ci=.95)
```

	number_physicians			
Predictors	*Estimates*	*std. Error*	*Statistic*	*p*
(Intercept)	988.00 (935.96 – 1040.04)	26.55	37.21	<0.001
number_hospital_beds_c	0.74 (0.72 – 0.77)	0.01	63.99	<0.001
Observations	440			
R^2 / R^2 adjusted	0.903 / 0.903			

- How do we interpret the CI for β_0?

2.3 Some Considerations on Making Inferences.

- Thinking about Power

SHS: Thinking about Power

```
tab_model(clothing_model)
```

	clothing_expenditure			
Predictors	*Estimates*	*std. Error*	*Statistic*	*p*
(Intercept)	1063.69 (-471.31 – 2598.68)	783.17	1.36	0.185
income	0.03 (-0.00 – 0.06)	0.02	1.85	0.075
Observations	30			
R^2 / R^2 adjusted	0.109 / 0.077			

SHS: Thinking about Power

- Explain in your own words the concept of power. Do you believe that we have a lot of power in this setting?
- Power is the probability of avoiding a type 2 error when conducting a test. (Type 2 error is when the null hypothesis should be rejected, but we do not reject it). So, it is the probability that we reject the null hypothesis and reach to the conclusion of the alternate hypothesis, when the null hypothesis was indeed false and the alternate hypothesis truly holds.
- power refers to the ability to reject a null hypothesis when it is false. That is, we make the correct decision in rejecting the null hypothesis. This test uses a high sample size (using data from across Canada).

- Explain in your own words the concept of power. Do you believe that we have a lot of power in this setting?

- Power is the probability that a false null hypothesis will be rejected. Since alpha (the probability of rejecting a true null hypothesis) is low (0.05) power is also low. This is because when everything else is kept constant, as alpha increases so does power. This is because a larger alpha increases the chances of rejecting a true null hypothesis and thus a false null hypothesis.

- Power is a quantity that measures how well the test's ability in detecting a false null hypothesis is. I don't think we have a lot of power in this setting for the sample size is relatively small (n =30).

- Yes, I think that we have a lot of power because B1 is not equal to 0.

- Explain in your own words the concept of power. Do you believe that we have a lot of power in this setting?
- The power shows how high probability there is to make a correct conclusion, e.g. reject the null hypothesis when it is false. I would say that our test does not have a lot of power since we can not come to a conclusion whether to reject or fail to reject the null hypothesis.
- Power is the probability of rejecting the null hypothesis
- There is no power in this setting because we failed to reject the null hypothesis.
- power is the likelihood that the conclusion drawn from an analysis reflects the truth

Recap: Sections 2.1-2.3

After Sections 2.1-2.3, you should be able to

- Compute and interpret confidence intervals for β_0 and β_1
- Conduct and interpret hypothesis tests concerning β_0 and β_1
- Define power and explain how it impacts inference

Learning Objectives for Sections 2.4-2.6

After Sections 2.4-2.6, you should be able to

- 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

2.4 Interval Estimation of $E[Y_h]$

A common objective in regression analysis is to estimate the mean for one or more probability distributions of Y.

Consider a study of the relation between the level of piecework pay (X) and worker productivity (Y).

The mean productivity at high and medium levels of piecework pay may be of particular interest. Why?

Sampling Distribution of $\hat{Y}_h$

$E[Y_h]$, the mean response when $X = X_h$, can be estimated in a given sample by

$$\boxed{\hat{Y}_h = b_0 + b_1 X_h}$$

Every sample would result in a different point estimate $\hat{Y}_h$ even if the *predictor variables*, X, were held constant across samples.

Under the *normal error regression model* ($Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ with $\varepsilon_i \sim N(0, \sigma^2)$), the sampling distribution of the *estimator* $\hat{Y}_h$ is

$$\begin{aligned} \hat{Y}_h &\sim N(E[\hat{Y}_h],\ \sigma\{\hat{Y}_h\}), \\ \text{with } \ E[\hat{Y}_h] &= E[b_0 + b_1 X_h] = E[b_0] + E[b_1] X_h = \beta_0 + \beta_1 X_h \\ \text{and } \sigma\{\hat{Y}_h\} &= \sigma^2 \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right] \end{aligned}$$

Sampling Distribution of $\hat{Y}_h$ when σ^2 is unknown

Therefore, when σ^2 is unknown, the $1-\alpha$ confidence limits for $E[Y_h]$ are

$$\boxed{\hat{Y}_h \pm t(1-\alpha/2; n-2)s\{\hat{Y}_h\}}, \qquad \text{where } s^2\{\hat{Y}_h\} = MSE\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]$$

SHS: Interval Estimation of $\hat{Y}_h$

A 90% confidence interval for $E[Y_h]$ corresponding to $X_h = 60000$ can be found using the following code in R:

```
ci = predict(clothing_model, newdata=data.frame(income=60000), inter

ci
```

```
##        fit      lwr      upr
## 1 2893.476 2204.008 3582.944
```

```
ci
```

```
##        fit      lwr      upr
## 1 2893.476 2204.008 3582.944
```

- In your own words, interpret what the given 90% confidence interval for $E[Y_h]$ corresponding to $X_h = 60000$ actually means in the context of the problem.

- We are 90% confident that the mean clothing expenditure, when 60000 Canadian Dollars is the income, is somewhere between 2204.00784 and 3582.944245 Canadian Dollars for clothing expenditure.

- I have 90% of confidence to say that the mean expenditure on clothing of the Canadians at the income level of 60000 fall in the interval (2704.00784,3582.944245).

- In your own words, interpret what the given 90% confidence interval for $E[Y_h]$ corresponding to $X_h = 60000$ actually means in the context of the problem.
- For someone that earns 60000 dollars we can expect, with 90% certainty, them to spend somewhere between 2204 and 3583 on clothes.
- We are 90% confident that a person that makes $60000 in income spends between $2204 and $3583.
- That when the income of a home is at 60000 they spend at least 2204.00784 and at most 3582.944245.
- 90% of the time, we would expect to see the average person making $60 000/year spend between $2900 and $3600 annually on clothing.

SHS: Pointwise Confidence Intervals for $\hat{Y}_h$

```
plot(spending_subset$income, spending_subset$clothing_expenditure, x
abline(clothing_model)

newx = seq(20000, 85000)
confidence_intervals = predict(clothing_model, newdata=data.frame(in
lines(newx, confidence_intervals[,2], col="blue", lty=2); lines(newx
```

$$\boxed{\hat{Y}_h \pm t(1-\alpha/2; n-2)s\{\hat{Y}_h\}}, \qquad \text{where } s^2\{\hat{Y}_h\} = MSE\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]$$

CDI: Pointwise Confidence Intervals for $\hat{Y}_h$

This data set provides selected county demographic information (CDI) for 440 of the most populous counties in the United States.

Each line of the data set has an identification number with a county name and state abbreviation and provides information on 14 variables for a single county.

Counties with missing data were deleted from the data set.

```
cdi %>% datatable()
```

Show 20 entries Search:

	county	state	land_area	population	pop_18_to_34
1	Los_Angeles	CA	4060	8863164	32
2	Cook	IL	946	5105067	29.
3	Harris	TX	1729	2818199	31.
4	San_Diego	CA	4205	2498016	33
5	Orange	CA	790	2410556	32.
6	Kings	NY	71	2300664	28.
7	Maricopa	AZ	9204	2122101	29.

Showing 1 to 20 of 440 entries

Previous 1 2 3 4 5 ... 22 Next

```
mod_physician_beds = lm(number_physicians ~ number_hospital_beds, da
tab_model(mod_physician_beds)
```

	number_physicians			
Predictors	*Estimates*	*std. Error*	*Statistic*	*p*
(Intercept)	-95.93 (-157.66 – -34.21)	31.49	-3.05	0.002
number_hospital_beds	0.74 (0.72 – 0.77)	0.01	63.99	<0.001
Observations	440			
R^2 / R^2 adjusted	0.903 / 0.903			

```
plot(cdi$number_hospital_beds, cdi$number_physicians, xlab="Number o
abline(mod_physician_beds)

newx = seq(92, 27700)
confidence_intervals = predict(mod_physician_beds, newdata=data.fram
lines(newx, confidence_intervals[,2], col="blue", lty=2)
lines(newx, confidence_intervals[,3], col="blue", lty=2)
```

- How do we intepret this set of 90% Confidence Intervals?

2.6 Confidence Band for Regression Line

- If two confidence intervals each *independently* have a 90% chance of containing the true mean response at different levels of X_h, then what is the probability that they both *simultaneously* capture the true mean responses?

These *pointwise* 90% confidence intervals are *simultaneous* __% confidence intervals.

Therefore, if we want to obtain a confidence band (i.e., *simultaneous confidence intervals*) for the entire regression line $E[Y] = \beta_0 + \beta_1 X$, then we need to make each interval wider.

- If we wanted to capture two points simultaneously, we might use the Bonferroni method and consider $100 \cdot (1 - \alpha/2)$% pointwise confidence intervals $(.95 \cdot .95 \approx .90)$.
- Similarly, if we wanted to capture three points simultaneously, we might use the Bonferroni method and consider $100 \cdot (1 - \alpha/3)$% pointwise confidence intervals $(.966 \cdot .966 \cdot .966 \approx .90)$.
- Similarly, for four points, we could use $100 \cdot (1 - \alpha/4)$% pointwise confidence intervals $(.975 \cdot .975 \cdot .975 \cdot .975 \approx .90)$.
- However, this approach is conservative, and as the number of points grows, the Bonferroni method produces unreasonably wide intervals

Consider how wide the *confidence band* produced by the Bonferroni method would be.

One appropriate choice for *simultaneously capturing the regression line at all values* X_h is the Working-Hotelling confidence limits:

$$\boxed{\hat{Y}_h \pm \sqrt{2F(1-\alpha; 2; n-2)}\, s\{\hat{Y}_h\}}$$

where, again,

$$s^2\{\hat{Y}_h\} = MSE\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]$$

Notice that this is the same confidence interval that we are used to, but $t(1-\alpha/2; n-2)$ is replaced with the (often only slightly) larger $\sqrt{2F(1-\alpha; 2; n-2)}$.

$F(1-\alpha; 2; n-2)$ represents the $100 \cdot (1-\alpha)$ percentile of an F distribution with 2 and $n-2$ degrees of freedom. We will discuss the F distribution more as the class progresses.

SHS: Confidence Band for $\hat{Y}_h$

```
alpha_level = 0.1

tvalue = qt(p=1-alpha_level/2, df= clothing_model$df)
Fvalue = qf(p=1-alpha_level, df1=2, df2= clothing_model$df)
tvalue
```

```
## [1] 1.701131
```

```
sqrt(2*Fvalue)
```

```
## [1] 2.237302
```

```
# Bonferroni Method
qt(p=1-(alpha_level/2)/2, df= clothing_model$df)
## [1] 2.048407

qt(p=1-(alpha_level/3)/2, df= clothing_model$df)
## [1] 2.238312

qt(p=1-(alpha_level/4)/2, df= clothing_model$df)
## [1] 2.368452
```

```
plot(spending_subset$income, spending_subset$clothing_expenditure, x
abline(clothing_model)

newx = seq(20000, 85000)
clothing_model_Yh = predict(clothing_model, newdata=data.frame(incom

confidence_intervals_lb = clothing_model_Yh$fit - tvalue* clothing_m
confidence_intervals_ub = clothing_model_Yh$fit + tvalue* clothing_m

confidence_band_lb = clothing_model_Yh$fit - sqrt(2*Fvalue)* clothin
confidence_band_ub = clothing_model_Yh$fit + sqrt(2*Fvalue)* clothin


lines(newx, confidence_intervals_lb, col="blue", lty=2)
lines(newx, confidence_intervals_ub, col="blue", lty=2)

lines(newx, confidence_band_lb, col="green", lty=3)
lines(newx, confidence_band_ub, col="green", lty=3)
```

$$\hat{Y}_h \pm t(1-\alpha/2; n-2)s\{\hat{Y}_h\}$$

$$\hat{Y}_h \pm \sqrt{2F(1-\alpha; 2; n-2)}s\{\hat{Y}_h\}$$

- How do we intepret this set of 90% Confidence Intervals?
- How do we intepret this 90% Confidence Band?

CDI: Confidence Band for $\hat{Y}_h$

```
alpha_level = 0.1

tvalue = qt(p=1-alpha_level/2, df= mod_physician_beds$df)
Fvalue = qf(p=1-alpha_level, df1=2, df2= mod_physician_beds$df)
tvalue
```

```
## [1] 1.64834
```

```
sqrt(2*Fvalue)
```

```
## [1] 2.151619
```

```
plot(cdi$number_hospital_beds, cdi$number_physicians, xlab="Number o
abline(mod_physician_beds)

newx = seq(92, 27700)
mod_physician_beds_Yh = predict(mod_physician_beds, newdata=data.fra

confidence_intervals_lb = mod_physician_beds_Yh$fit - tvalue* mod_ph
confidence_intervals_ub = mod_physician_beds_Yh$fit + tvalue* mod_ph

confidence_band_lb = mod_physician_beds_Yh$fit -  sqrt(2*Fvalue)* mo
confidence_band_ub = mod_physician_beds_Yh$fit + sqrt(2*Fvalue)* mod


lines(newx, confidence_intervals_lb, col="blue", lty=2)
lines(newx, confidence_intervals_ub, col="blue", lty=2)

lines(newx, confidence_band_lb, col="green", lty=3)
lines(newx, confidence_band_ub, col="green", lty=3)
```

$$\hat{Y}_h \pm t(1-\alpha/2; n-2)s\{\hat{Y}_h\}$$

$$\hat{Y}_h \pm \sqrt{2F(1-\alpha; 2; n-2)}s\{\hat{Y}_h\}$$

- How do we intepret this set of 90% Confidence Intervals?
- How do we intepret this 90% Confidence Band?

2.5 Prediction of New Observation

So far, we have been concerned with capturing the *mean response* $E[Y_h]$ corresponding to a given level X_h of the predictor variable.

We now consider the prediction of a *new observation* Y corresponding to a given level X_h of the predictor variable.

Remember that $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, where ε_i are independent $N(0, \sigma^2)$.

Therefore, if we knew β_0 and β_1, we would expect new observations to satisfy

$$Y_h \sim N(E[Y_h],\ \sigma^2), \qquad \text{where } E[Y_h] = \beta_0 + \beta_1 X_h.$$

Thus a reasonable prediction interval for Y_h would be

$$E[Y_h]\ \pm\ z(1-\alpha/2)\sigma,$$

where $z(1-\alpha/2)$ is the $100 \cdot (1-\alpha/2)$th percentile of the standard normal distribution.

This prediction interval should capture about $100(1-\alpha)$% of new observations at X_h.

Prediction of New Observation When Parameters are Unknown

However, when β_0 and β_1 are not known, we don't know $E[Y_h]$, so we can't use

$$E[Y_h] \ \pm \ z(1-\alpha/2)\sigma.$$

Additionally, we don't know σ.

Instead, we need to centre our interval at $\hat{Y}_h$ and we need to account for the additional uncertainty that comes from using estimates:

$$\boxed{\hat{Y}_h \pm t(1-\alpha/2; n-2)s\{pred\}}, \qquad \text{where } s^2\{pred\} = MSE + s^2\{\hat{Y}_h\}$$

SHS: Prediction of New Observation

The pointwise confidence intervals are of the form

$$\hat{Y}_h \pm t(1-\alpha/2; n-2)s\{\hat{Y}_h\}, \qquad \text{where } s^2\{\hat{Y}_h\} = MSE\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]$$

The corresponding (pointwise) prediction intervals would be of the form

$$\hat{Y}_h \pm t(1-\alpha/2; n-2)s\{pred\}, \qquad \text{where } s^2\{pred\} = MSE + s^2\{\hat{Y}_h\}$$

- How would the corresponding prediction interval compare to the confidence interval discussed previously at the point $X_h = 60000$? In your own words, interpret this prediction interval in the context of this problem.

- The CI, previously mentioned, represents an inference on the mean clothing expenditure, when 6000 Canadian Dollars is the income... However, the corresponding prediction interval is a statement about what the value of the clothing expenditure would be in the next survey taken, if 6000 where the income.

- Prediction intervals must account for both the uncertainty in knowing the value of the population mean, plus data scatter, so it would always be wider. A prediction interval of Y of level X = 6000 says that the next time we pick somebody whose income is 6000, his/her clothing expenditure with most likely be within the prediction interval.

- The 90% prediction interval would be wider than the 90% confidence interval... In this case the prediction interval is (1605.15465,4181.79744). This indicates that there is a 90% chance that someone that makes $60000 a year is going to spend between $1605.15465 and $4181.79744 on clothes in a year.

- How would the corresponding prediction interval compare to the confidence interval discussed previously at the point $X_h = 60000$? In your own words, interpret this prediction interval in the context of this problem.

- prediction interval will have a wider range compare to confidence interval. prediction interval is to predict the interval of the mean of expenditure of Canadians at the income level of 6000.

- This prediction interval would create an upper and lower bound on the E[Y_h].

- Sorry, but for the model to work doesn't B1 has to be 0.30496 instead of 0.030496 or have I misunderstood something? ... I expected the prediction interval to be much more narrow than the CI.

```
plot(spending_subset$income, spending_subset$clothing_expenditure, x
abline(clothing_model)

newx = seq(20000, 85000)
confidence_intervals = predict(clothing_model, newdata=data.frame(in
lines(newx, confidence_intervals[,2], col="blue", lty=2); lines(newx


prediction_intervals = predict(clothing_model, newdata=data.frame(in
lines(newx, prediction_intervals[,2], col="red", lty=2); lines(newx,
```

CDI: Prediction of New Observation

```
mod_physician_beds = lm(number_physicians ~ number_hospital_beds, da
tab_model(mod_physician_beds)
```

	number_physicians			
Predictors	*Estimates*	*std. Error*	*Statistic*	*p*
(Intercept)	-95.93 (-157.66 – -34.21)	31.49	-3.05	0.002
number_hospital_beds	0.74 (0.72 – 0.77)	0.01	63.99	<0.001
Observations	440			
R^2 / R^2 adjusted	0.903 / 0.903			

```
plot(cdi$number_hospital_beds, cdi$number_physicians, xlab="Number o
abline(mod_physician_beds)

newx = seq(92, 27700)
confidence_intervals = predict(mod_physician_beds, newdata=data.fram
lines(newx, confidence_intervals[,2], col="blue", lty=2); lines(newx

prediction_intervals =  predict(mod_physician_beds, newdata=data.fra
lines(newx, prediction_intervals[,2], col="red", lty=2); lines(newx,
```

- How do we intepret this set of 90% Prediction Intervals?

Prediction of Mean of m New Observations for Given X_h

Occasionally, one would like to predict the mean of m new observations for Y for X_h, a given level of the predictor variable.

The appropriate $1-\alpha$ prediction limits are as follows, assuming the new Y observations are independent:

$$\boxed{\hat{Y}_h \pm t(1-\alpha/2; n-2)s\{predmean\}}$$

where $\boxed{s^2\{predmean\} = \frac{MSE}{m} + s^2\{\hat{Y}_h\}}$.

Contrast this with the prediction interval for a single observation:

$$\hat{Y}_h \pm t(1-\alpha/2; n-2)s\{pred\}, \qquad \text{where } s^2\{pred\} = MSE + s^2\{\hat{Y}_h\}$$

Note that we have just as much uncertainty surrounding our estimate of where the prediction interval should be centered ($s^2\{\hat{Y}_h\}$), but the variation in the distribution of $\bar{Y}_h$ for m individuals ($\frac{MSE}{m}$) is smaller than the variation in the distribution of a single point Y_h (MSE).

Pointwise confidence intervals
Pointwise prediction intervals

Recap: Sections 2.4-2.6

After Sections 2.4-2.6, you should be able to

- 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

Learning Objectives for Sections 2.7

After Section 2.7, you should be able to

- Construct and interpret an ANOVA table
- Conduct and interpret an ANOVA F test

2.7: Analysis of Variance Approach to Regression Analysis

We have developed the basic regression model and demonstrated is major uses.

We now consider the regression analysis from the perspective of analysis of variance.

This new perspective will not enable us to do anything new, but the analysis of variance approach will come into its own when we take up multiple regression models and other types of linear statistical models.

Partitioning of Total Sum of Squares

The analysis of variance (ANOVA) approach is based on the partitioning of sums of squares and degrees of freedom associated with the response variable Y

$$Y_i - \bar{Y} = \hat{Y}_i - \bar{Y} + Y_i - \hat{Y}_i$$

Total Deviation=Deviation of fitted regression value around mean

+ Deviation around fitted regression line

$$Y_i - \bar{Y} = \hat{Y}_i - \bar{Y} + Y_i - \hat{Y}_i$$

Total Deviation=Deviation of fitted regression value around mean

+ Deviation around fitted regression line

$$Y_i - \bar{Y} = \hat{Y}_i - \bar{Y} + Y_i - \hat{Y}_i$$

Total Deviation=Deviation of fitted regression value around mean

+ Deviation around fitted regression line

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

$$\text{total sum of squares} = \text{regression sum of squares} + \text{error sum of squares}$$

$$SSTO = SSR + SSE$$

- What happens to the regression sum of squares if the true regression line is horizontal (i.e., if $\beta_1 = 0$)?

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

$$\text{total sum of squares} = \text{regression sum of squares} + \text{error sum of squares}$$

$$SSTO = SSR + SSE$$

- What happens to the regression sum of squares if the regression line is far from horizontal (e.g., if $\beta_1 >> 0$)?

- Note that SSR is thus a measure of how far β_1 is from 0.

Analysis of Variance Table

Source of Variation	SS	df	MS	F
Regression	$SSR = \sum(\hat{Y}_i - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$	$F^* = \frac{MSR}{MSE}$
Error	$SSE = \sum(Y_i - \hat{Y}_i)^2$	$n-2$	$MSE = \frac{SSE}{n-2}$	
Total	$SSTO = \sum(Y_i - \bar{Y})^2$	$n-1$		

$$E[MSR] = \sigma^2 + \beta_1^2 \sum(X_i - \bar{X})^2$$
$$E[MSE] = \sigma^2$$

This shows that MSE is an unbiased estimator of σ^2;

- what else does it tell us?
- What happens to the F statistic if $\beta_1 = 0$?
- What happens to the F statistic if $\beta_1 >> 0$?
- What happens to the F statistic if $\beta_1 << 0$?

F Test of $\beta_1 = 0$ versus $\beta_1 \neq 0$

The analysis of variance approach provides us with a battery of highly useful tests for regression models.

For the simple linear regression case considered here, we can use the *test statistic* $F^* = \frac{MSR}{MSE}$ to test

$$H_0 : \beta_1 = 0 \qquad vs \qquad H_a : \beta_1 \neq 0.$$

- If H_0 is true, then F^* should be close to 1.
- If H_a is true, then F^* should be much bigger than 1.

Sampling Distribution of F^*

In order to develop a test, we need to be more precise than F^* *should be close to* 1.

By Cochran's theorem, $\textbf{when } H_0 \textbf{ holds}$

$$F^* = \frac{MSR}{MSE} = \frac{\frac{SSR}{\sigma^2}}{1} \div \frac{\frac{SSE}{\sigma^2}}{n-2} \sim \frac{\chi^2(1)}{1} \div \frac{\chi^2(n-2)}{n-2},$$

where the χ^2 variables are independent.

I.e.,

$$\boxed{F^* \sim F(1, n-2) \qquad \textbf{when } H_0 \textbf{ holds}} \qquad (\text{i.e., when } \beta_1 = 0)$$

When H_0 does not hold (i.e., when $\beta_1 \neq 0$), F^* follows the non-central F distribution.

- The important thing is that it will tend to be bigger than 1
  - how much bigger depends on what β_1 actually equals.

```
curve(df(x, df1=1, df2=100), from=0, to=15, xlab="", ylab="Density",
curve(df(x, df1=1, df2=100, ncp=5), from=0, to=15, col="green", add=
curve(df(x, df1=1, df2=100, ncp=10), from=0, to=15, col="green", add
abline(v=1, lty=4, lwd=.5, col="blue")
```

Density Functions for `$F$`

```
curve(pf(x, df1=1, df2=100), from=0, to=15, xlab="", ylab="Cumulativ
curve(pf(x, df1=1, df2=100, ncp=5), from=0, to=15, col="green", add=
curve(pf(x, df1=1, df2=100, ncp=10), from=0, to=15, col="green", add
abline(v=1, lty=4, lwd=.5, col="blue")
```

Distribution Functions for `$F$`

```
curve(pf(x, df1=1, df2=100), from=0, to=15, xlab="", ylab="Cumulativ
curve(pf(x, df1=1, df2=100, ncp=5), from=0, to=15, col="green", add=
curve(pf(x, df1=1, df2=100, ncp=10), from=0, to=15, col="green", add
abline(h=.95, lty=4, lwd=.5, col="blue")
```

Distribution Functions for `$F$`

```
curve(pf(x, df1=1, df2=100), from=0, to=15, xlab="", ylab="Cumulativ
curve(pf(x, df1=1, df2=100, ncp=5), from=0, to=15, col="green", add=
curve(pf(x, df1=1, df2=100, ncp=10), from=0, to=15, col="green", add
abline(h=.95, v=qf(.95, df1=1, df2=100), lty=4, lwd=.5, col="blue")
```

Distribution Functions for `$F$`

SHS: ANOVA

The Canadian Survey of Household Spending is carried out annually across Canada.

(http://dli-idd-nesstar.statcan.gc.ca.proxy.library.upei.ca/webview/)

The main purpose of the survey is to obtain detailed information about household spending. Information is also collected about dwelling characteristics as well as household equipment.

The survey data are used by the following groups:

- Government departments use the data to help formulate policy;
- Community groups, social agencies and consumer groups use the data to support their positions and to lobby governments for social changes;
- Lawyers and their clients use the data to determine what is fair for child support and other compensation;
- Labour and contract negotiators rely on the data when discussing wage and cost-of-living clauses;
- Individuals and families can use the data to compare their spending habits with those of similar types of households.

```
###A subset of the latest Survey of Household Spending data are disp
spending_subset %>% datatable()
```

Show 20 entries Search:

	province	type_of_dwelling	income	marital_status	age_gr
1	NL	single_detached	68000	never_married	30-34
2	NL	single_detached	48000	never_married	25-29
3	NL	single_detached	30000	married	35-39
4	NL	row_house	30000	never_married	30-34
5	NL	single_detached	35000	married	25-29
6	NL	single_detached	26000	married	25-29
7	NL	single_detached	26000	other	55-59

Showing 1 to 20 of 500 entries

Previous 1 2 3 4 5 ... 25 Next

We are interested in the potential relationship between the income of working Canadians and the amount that they spend on clothing in a year.

Income and Clothing Expenditure for a small subset of the Survey of Household Spending are displayed below:

```
xyplot(clothing_expenditure~income, data=spending_subset,  type=c("p
```

```
clothing_model = lm(clothing_expenditure~income, data=spending_subse
anova(clothing_model)
```

```
## Analysis of Variance Table
##
## Response: clothing_expenditure
##           Df     Sum Sq  Mean Sq F value    Pr(>F)
## income     1   71258930 71258930  26.309 4.174e-07 ***
## Residuals 498 1348849454  2708533
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Complete the *Total* line of the ANOVA table:

Source of Variation				SS	df
Total				__	__

- In your own words, interpret the p-value from the ANOVA table in the context of this problem.

- In your own words, interpret the p-value from the ANOVA table in the context of this problem.
- The p-value is the probability of obtaining a result at least this extreme, assuming that the null hypothesis is true.
- The small p-value allows us to reject the null hypothesis that there is no relation between clothing expenditure and income (b1 =0)
- The p-value is 4.174e-07 . Therefore, at a significance level of 0.01 the null hypothesis (income and amount spent on clothes are not related) can be rejected and it can be concluded that income and amount spent on clothing are related in a statistically significant way.
- we have to reject the null hypothesis and conclude that B1 is different than 0. Thus, there is a linear relationship between mean clothing expenditures and income.

- In your own words, interpret the p-value from the ANOVA table in the context of this problem.

- With such a small p-value it means we fail to reject the null hypothesis, meaning that the variances of clothes expenditure and income are not effected by each other.

- p value is the probability for the b1 of the sample to be true given that the beta1 = 0 is true. It is the probability for income and expenditure to have linear association. if the p value is too small, then there is no linear association between income and expenditure.

- P value shows that there is a very low probability to get the results we assumed in our null hypothesis, therefore showing there isn't a strong relationship between income and clothing expenditure

- The p-value shows the probability that our null hypothesis is true, which would be that there is no relationship between income and how much they spend on clothing. The p-value is a very small number in this case, hence there is a very low probability that there is no relationship between income and clothing expenditures.
- At 0.0000004174 we reject the null hypothesis as the probability of a Type 2 error is so low.

CDI: ANOVA - physicians vs hospital beds

This data set provides selected county demographic information (CDI) for 440 of the most populous counties in the United States.

Each line of the data set has an identification number with a county name and state abbreviation and provides information on 14 variables for a single county.

Counties with missing data were deleted from the data set.

```
cdi %>% datatable()
```

Show 20 entries Search:

	county	state	land_area	population	pop_18_to_34
1	Los_Angeles	CA	4060	8863164	32
2	Cook	IL	946	5105067	29.
3	Harris	TX	1729	2818199	31.
4	San_Diego	CA	4205	2498016	33
5	Orange	CA	790	2410556	32.
6	Kings	NY	71	2300664	28.
7	Maricopa	AZ	9204	2122101	29.

Showing 1 to 20 of 440 entries

Previous 1 2 3 4 5 ... 22 Next

```
mod_physician_beds = lm(number_physicians ~ number_hospital_beds, da
anova_table = anova(mod_physician_beds)
anova_table %>% round(2) %>% datatable(options=list(scrollY=150))
```

Show 20 entries Search:

	Df	Sum Sq	Mean Sq	F value	P
number_hospital_beds	1	1270342254.01	1270342254.01	4095.34	
Residuals	438	135864044.99	310191.88		

Showing 1 to 2 of 2 entries Previous 1 Next

- In your own words, interpret the p-value from the ANOVA table in the context of this problem.

CDI: ANOVA - physicians vs population

```
mod_physician_pop = lm(number_physicians ~ population, data=cdi)
anova_table = anova(mod_physician_pop)
anova_table %>% round(2) %>% datatable(options=list(scrollY=150))
```

Show 20 entries Search:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
population	1	1243181163.84	1243181163.84	3340.06	0
Residuals	438	163025135.15	372203.5		

Showing 1 to 2 of 2 entries Previous 1 Next

- In your own words, interpret the p-value from the ANOVA table in the context of this problem.

```
mod_physician_pop = lm(number_physicians ~ population, data=cdi)
summary(mod_physician_pop)
```

```
##
## Call:
## lm(formula = number_physicians ~ population, data = cdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1969.4  -209.2   -88.0    27.9  3928.7
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.106e+02  3.475e+01  -3.184  0.00156 **
## population   2.795e-03  4.837e-05  57.793  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 610.1 on 438 degrees of freedom
## Multiple R-squared:  0.8841,    Adjusted R-squared:  0.8838
## F-statistic:  3340 on 1 and 438 DF,  p-value: < 2.2e-16
```

- In your own words, interpret the p-value corresponding to the F-statistic in the context of this problem.

CDI: ANOVA - physicians vs total income

```
mod_physician_income = lm(number_physicians ~ total_personal_income,
anova_table = anova(mod_physician_income)
anova_table %>% round() %>% datatable(options=list(scrollY=150))
```

Show 20 entries Search:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
total_personal_income	1	1264058045	1264058045	3895	0
Residuals	438	142148254	324539		

Showing 1 to 2 of 2 entries Previous 1 Next

- In your own words, interpret the p-value from the ANOVA table in the context of this problem.

```r
mod_physician_income = lm(number_physicians ~ total_personal_income,
summary(mod_physician_income)
```

```
##
## Call:
## lm(formula = number_physicians ~ total_personal_income, data = cdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1926.6  -194.5   -66.6    44.2  3819.0
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -48.39485   31.83333   -1.52    0.129
## total_personal_income   0.13170    0.00211   62.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 569.7 on 438 degrees of freedom
## Multiple R-squared:  0.8989,    Adjusted R-squared:  0.8987
## F-statistic:  3895 on 1 and 438 DF,  p-value: < 2.2e-16
```

Equivalence of F-test and two-sided T-test

$$
\begin{aligned}
F^* &= \frac{MSR}{MSE} \\
&= \frac{b_1^2 \sum (X_i - \bar{X})^2}{MSE} \\
&= \frac{b_1^2}{s^2\{b_1\}} \\
&= \left(\frac{b_1}{s\{b_1\}}\right)^2 \\
&= (t^*)^2
\end{aligned}
$$

In addition,

$$F(1-\alpha; 1; n-2) = t(1-\alpha/2; n-2)^2$$

For simple linear regression, these are equivalent tests of $H_0 : \beta_1 = 0$.

Recap: Sections 2.7

After Section 2.7, you should be able to

- Construct and interpret an ANOVA table
- Conduct and interpret an ANOVA F test

Learning Objectives for Sections 2.8-2.10

After Sections 2.8-2.10, you should be able to

- Describe the general linear test approach
- Calculate and interpret R^2
- Understand the limitations of R^2
- Describe the limitations of linear regression analysis

2.8: General Linear Test Approach

The test of $\beta_1 = 0$ versus $\beta_1 \neq 0$ is a simple example of a general linear test.

The general linear test can be used in a wide variety of situations (some complex and some, like here, are simple) and has three parts

- Full Model
- Reduced Model
- Test Statistic

It tests whether the reduced model is "adequate".

That is, it tests whether the full model is significantly better than the reduced model at explaining the variability in the response.

Full Model

A full linear model is first fit to the data.

Then, the *error sum of squares* is obtained for this "full" model ($SSE(F)$).

In the context of simple linear regression, the full model is

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

and the error sum of squares for this model is

$$SSE(F) = \sum [Y_i - (b_0 + b_1 X_i)]^2 = \sum (Y_i - \hat{Y}_i)^2 = SSE.$$

Notice that for this full model, the error sum of squares is simply *SSE*, which measures the variability of Y_i observations around the fitted regression line.

Reduced Model

A reduced model is then fit to the data.

Here, we are considering a "reduced" simple linear regression model where the slope is zero (i.e., there is no relationship between input and output):

$$H_0 : \beta_1 = 0 \qquad \text{vs } H_a : \beta_1 \neq 0$$

The model when H_0 holds is called the reduced or restricted model. Here, it corresponds to the model

$$Y_i = \beta_0^* + \varepsilon_i$$

To develop the test statistic, we also need the *error sum of squares* for the reduced model:

$$SSE(R) = \sum [Y_i - b_0^*]^2 = \sum (Y_i - \bar{Y}_i)^2 = SSTO.$$

Test Statistic

The idea is to compare the two error sums of squares SSE(F) and SSE(R).

Note that we aren't just looking for which is bigger. Because the full model has more parameters than the reduced model, it is always true that

$$SSE(F) \leq SSE(R)$$

The question is whether $SSE(F)$ is *sufficiently* smaller than $SSE(R)$ to justify the additional parameters.

- I.e., the *full* model will always fit the data better than the *reduced* model, but is the fit *significantly* better?

In the general linear test, the test statistic is

$$\boxed{F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} \div \frac{SSE(F)}{df_F}}$$

which follows the F distribution when H_0 holds.

The degrees of freedom df_R and df_F are those associated with the reduced and full model error sums of squares respectively.

The decision rule is, therefore, based on whether $\boxed{F^* > F(1-\alpha; df_R - df_F, df_F)}$; this would be evidence against H_0.

For testing whether or not $\beta_1 = 0$, we therefore have

- $SSE(F) = SSE$
- $SSE(R) = SSTO$
- $df_F = n - 2$
- $df_R = n - 1$

So,

$$\boxed{F^* = \frac{SSTO - SSE}{1} \div \frac{SSE}{n-2} = \frac{SSR}{1} \div \frac{SSE}{n-2} = \frac{MSR}{MSE}},$$

which is identical to the ANOVA test statistic.

SHS: General Linear Test Approach

The Canadian Survey of Household Spending is carried out annually across Canada.

(http://dli-idd-nesstar.statcan.gc.ca.proxy.library.upei.ca/webview/)

The main purpose of the survey is to obtain detailed information about household spending. Information is also collected about dwelling characteristics as well as household equipment.

The survey data are used by the following groups:

- Government departments use the data to help formulate policy;
- Community groups, social agencies and consumer groups use the data to support their positions and to lobby governments for social changes;
- Lawyers and their clients use the data to determine what is fair for child support and other compensation;
- Labour and contract negotiators rely on the data when discussing wage and cost-of-living clauses;
- Individuals and families can use the data to compare their spending habits with those of similar types of households.

```
###A subset of the latest Survey of Household Spending data are disp
spending_subset %>% datatable()
```

Show 20 entries Search:

	province	type_of_dwelling	income	marital_status	age_gr
1	NL	single_detached	68000	never_married	30-34
2	NL	single_detached	48000	never_married	25-29
3	NL	single_detached	30000	married	35-39
4	NL	row_house	30000	never_married	30-34
5	NL	single_detached	35000	married	25-29
6	NL	single_detached	26000	married	25-29
7	NL	single_detached	26000	other	55-59

Showing 1 to 20 of 500 entries

Previous 1 2 3 4 5 ... 25 Next

We are interested in the potential relationship between the income of working Canadians and the amount that they spend on clothing in a year.

Income and Clothing Expenditure for a small subset of the Survey of Household Spending are displayed below:

```
xyplot(clothing_expenditure~income, data=spending_subset,  type=c("p
```

```
clothing_model = lm(clothing_expenditure~income, data=spending_subse
anova(clothing_model)
```

```
## Analysis of Variance Table
##
## Response: clothing_expenditure
##           Df     Sum Sq  Mean Sq F value    Pr(>F)
## income     1   71258930 71258930  26.309 4.174e-07 ***
## Residuals 498 1348849454  2708533
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- In your own words, briefly describe what the "General linear test approach" could accomplish in this setting.
- There is a large difference between SSE(R) and SSE(F) suggests that Ha holds. Therefore, B1 isn't zero, there is a relationship between the income of working Canadians and the amount that they spend on clothing in a year.

```
clothing_model = lm(clothing_expenditure~income, data=spending_subse
anova(clothing_model)
```

```
## Analysis of Variance Table
##
## Response: clothing_expenditure
##           Df     Sum Sq  Mean Sq F value    Pr(>F)
## income     1   71258930 71258930  26.309 4.174e-07 ***
## Residuals 498 1348849454  2708533
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- In your own words, briefly describe what the "General linear test approach" could accomplish in this setting.

- In the general linear test approach we know that SSE(R) is greater or equal to SSE(F). If they are equal or close to it, we know that the addition of the parameters in full models do not significantly reduce the variations around the fitted regression model. In this example and with the value of the CD at 0.05(approx), we can conclude that the predictor variable is not a significant factor in reducing the total variation.
- In the General linear test approach we compare the SSE and the SSTO, if those values are relatively close to each other it is more likely that the null hypothesis holds. In our test there is not so big difference between those numbers.which would indicate that our null hypothesis may hold.

```
clothing_model_reduced = lm(clothing_expenditure~1, data=spending_su
anova(clothing_model_reduced, clothing_model)
```

```
## Analysis of Variance Table
##
## Model 1: clothing_expenditure ~ 1
## Model 2: clothing_expenditure ~ income
##   Res.Df        RSS Df Sum of Sq      F    Pr(>F)
## 1    499 1420108384
## 2    498 1348849454  1  71258930 26.309 4.174e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

CDI: General Linear Test - physicians vs total income

```
mod_physician_income = lm(number_physicians ~ total_personal_income,
mod_physician_income_reduced = lm(number_physicians ~ 1, data=cdi)
anova(mod_physician_income_reduced, mod_physician_income)  %>%  roun
```

Show 20 entries Search:

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	439	1406206299				
2	438	142148254	1	1264058045	3895	0

Showing 1 to 2 of 2 entries Previous 1 Next

- In your own words, briefly describe what the "General linear test approach" could accomplish in this setting.

2.9: Descriptive Measures of Linear Association between X and Y.

- $SSTO$ measures the variation in the observations Y_i when X is not considered
- SSE measures the variation in the Y_i after a predictor variable X is employed
- A natural measure of the effect of X in reducing variation in Y is to express the reduction in variation $(SSTO - SSE = SSR)$ as a proportion of the total variation:

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

R^2 is called the Coefficient of Determination.

Note that since $0 \leq SSE \leq SSTO$ then $0 \leq R^2 \leq 1.$

1. When all observations fall on the fitted regression line, $SSE = 0$ and $R^2 = 1$

2. When the fitted regression line is horizontal so that $b_1 = 0$ and $\hat{Y}_i = \bar{Y}$, then $SSE = SSTO$ and $R^2 = 0$.

Misunderstandings about R^2

- high R^2 indicates that useful predictions can be made.
  - The prediction interval for a particular input of interest may still be wide even if R^2 is high.

- high R^2 means that there is a good linear fit between predictor and outcome.
  - It can be the case that an approximate (bad) linear fit to a truly curvilinear relationship might result in a high R^2.

- low R^2 means that there is no relationship between predictor and outcome.
  - Also not true since there can be clear and strong relationships between predictor and outcome that are not well explained by a linear functional relationship.

Coefficient of Correlation

$$\boxed{r = \pm\sqrt{R^2}}$$

- If $b_1 > 0$, then $r = \sqrt{R^2}$
- If $b_1 < 0$, then $r = -\sqrt{R^2}$

$$-1 \leq r \leq 1.$$

- We will discuss this more in Section 2.11

SHS: Coefficient of Determination

The Canadian Survey of Household Spending is carried out annually across Canada.

(http://dli-idd-nesstar.statcan.gc.ca.proxy.library.upei.ca/webview/)

The main purpose of the survey is to obtain detailed information about household spending. Information is also collected about dwelling characteristics as well as household equipment.

The survey data are used by the following groups:

- Government departments use the data to help formulate policy;
- Community groups, social agencies and consumer groups use the data to support their positions and to lobby governments for social changes;
- Lawyers and their clients use the data to determine what is fair for child support and other compensation;
- Labour and contract negotiators rely on the data when discussing wage and cost-of-living clauses;
- Individuals and families can use the data to compare their spending habits with those of similar types of households.

```
###A subset of the latest Survey of Household Spending data are disp
spending_subset %>% datatable()
```

Show 20 entries Search:

	province	type_of_dwelling	income	marital_status	age_gr
1	NL	single_detached	68000	never_married	30-34
2	NL	single_detached	48000	never_married	25-29
3	NL	single_detached	30000	married	35-39
4	NL	row_house	30000	never_married	30-34
5	NL	single_detached	35000	married	25-29
6	NL	single_detached	26000	married	25-29
7	NL	single_detached	26000	other	55-59

Showing 1 to 20 of 500 entries

Previous 1 2 3 4 5 ... 25 Next

We are interested in the potential relationship between the income of working Canadians and the amount that they spend on clothing in a year.

Income and Clothing Expenditure for a small subset of the Survey of Household Spending are displayed below:

```
xyplot(clothing_expenditure~income, data=spending_subset,  type=c("p
```

```
clothing_model = lm(clothing_expenditure~income, data=spending_subse
msummary(clothing_model)
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.578e+03  1.864e+02   8.465 2.89e-16 ***
## income      2.116e-02  4.126e-03   5.129 4.17e-07 ***
##
## Residual standard error: 1646 on 498 degrees of freedom
## Multiple R-squared:  0.05018,    Adjusted R-squared:  0.04827
## F-statistic: 26.31 on 1 and 498 DF,  p-value: 4.174e-07
```

- In your own words, interpret the coefficient of determination in this setting.
- The variation in clothing expenditure is reduced by 5.018% when income is considered.
- The variation in clothing expenditures is reduced by 5.018% when income is considered. Also, the coefficient of determination (0.05018) is not very close to 1, so the degree of linear association between income and clothing expenditures is not very large, meaning the points are not too close to the fitted regression line.
- The coefficient of determination is the ratio between the sum of squares regression and the total sum of squares. In this case, it was equal to 0.94982. This means that approximately 94.982% of the variation in Y can be explained by variation in X.

- In your own words, interpret the coefficient of determination in this setting.
- because the coefficient of determination is very close to zero, there is a lot of variation found in clothing-expenditure
- The coefficient of determination value in this setting of 0.05018 tells us that around 5% of clothing expenditures are predictable.
- Since our Coefficient of determination is very close to zero, we know that there is a lot of variation in the y-values.

CDI: R^2 - physicians vs hospital beds

This data set provides selected county demographic information (CDI) for 440 of the most populous counties in the United States.

Each line of the data set has an identification number with a county name and state abbreviation and provides information on 14 variables for a single county.

Counties with missing data were deleted from the data set.

```
cdi %>% datatable()
```

Show 20 entries Search:

	county	state	land_area	population	pop_18_to_34
1	Los_Angeles	CA	4060	8863164	32
2	Cook	IL	946	5105067	29.
3	Harris	TX	1729	2818199	31.
4	San_Diego	CA	4205	2498016	33
5	Orange	CA	790	2410556	32.
6	Kings	NY	71	2300664	28.
7	Maricopa	AZ	9204	2122101	29.

Showing 1 to 20 of 440 entries

Previous 1 2 3 4 5 ... 22 Next

```
mod_physician_beds = lm(number_physicians ~ number_hospital_beds, da
msummary(mod_physician_beds)
```

```
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -95.93218   31.49396  -3.046  0.00246 **
## number_hospital_beds   0.74312    0.01161  63.995  < 2e-16 ***
##
## Residual standard error: 556.9 on 438 degrees of freedom
## Multiple R-squared:  0.9034,    Adjusted R-squared:  0.9032
## F-statistic:  4095 on 1 and 438 DF,  p-value: < 2.2e-16
```

- In your own words, interpret the coefficient of determination in this setting.

CDI: R^2 - physicians vs population

```
mod_physician_pop = lm(number_physicians ~ population, data=cdi)
msummary(mod_physician_pop)
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.106e+02  3.475e+01  -3.184  0.00156 **
## population   2.795e-03  4.837e-05  57.793  < 2e-16 ***
##
## Residual standard error: 610.1 on 438 degrees of freedom
## Multiple R-squared:  0.8841,    Adjusted R-squared:  0.8838
## F-statistic:  3340 on 1 and 438 DF,  p-value: < 2.2e-16
```

- In your own words, interpret the coefficient of determination in this setting.

CDI: R^2 - physicians vs total income

```
mod_physician_income = lm(number_physicians ~ total_personal_income,
msummary(mod_physician_income)
```

```
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -48.39485   31.83333   -1.52    0.129
## total_personal_income   0.13170    0.00211   62.41   <2e-16 ***
##
## Residual standard error: 569.7 on 438 degrees of freedom
## Multiple R-squared:  0.8989,    Adjusted R-squared:  0.8987
## F-statistic:  3895 on 1 and 438 DF,  p-value: < 2.2e-16
```

- In your own words, interpret the coefficient of determination in this setting.

2.10: Considerations in Applying Regression Analysis

Be cautious whenever

1. making inferences about the future
   - e.g., predicting future school enrollments based on school demographics
2. predicting Y based on an X value that itself is predicted
   - e.g., predicting company sales based on demographic projections
3. extrapolating
   - e.g., predicting sales based on disposable income levels that are outside the range of what has been seen previously
4. attempting to establish cause-and-effect
   - e.g., just because cancer rates are predicted by smoking status $(\beta_1 \neq 0)$, that doesn't mean that smoking causes cancer
5. estimating several things from the same data
   - e.g., multiple testing or predicting the outcome for several new observations
6. X may be subject to measurement error
   - see Chapter 4.

SHS: Considerations in Regression Analyses

CDI: Considerations in Regression Analyses

This data set provides selected county demographic information (CDI) for 440 of the most populous counties in the United States.

Each line of the data set has an identification number with a county name and state abbreviation and provides information on 14 variables for a single county.

Counties with missing data were deleted from the data set.

Show 20 entries Search:

	county	state	land_area	population	pop_18_to_34
1	Los_Angeles	CA	4060	8863164	32
2	Cook	IL	946	5105067	29
3	Harris	TX	1729	2818199	31
4	San_Diego	CA	4205	2498016	33
5	Orange	CA	790	2410556	32

Showing 1 to 20 of 440 entries

```
xyplot(number_physicians ~ total_personal_income, data=cdi,  type=c(
```

```
xyplot(number_physicians ~ population, data=cdi,  type=c("p", "r", "
```

```
xyplot(number_physicians ~ number_hospital_beds, data=cdi,  type=c("
```

Recap: Sections 2.8-2.10

After Sections 2.8-2.10, you should be able to

- Describe the general linear test approach
- Calculate and interpret R^2
- Understand the limitations of R^2
- Describe the limitations of linear regression analysis

Learning Objectives for Sections 2.11

After Sections 2.11, you should be able to

- Contrast regression and correlation
- Conduct and interpret inference on correlation coefficients
- Estimate, interpret, test, and contrast Spearman rank correlation.

2.11: Normal Correlation Models

So far, we have been assuming that the X values are known constants.

- Confidence intervals, therefore, have been interpreted in terms of repeated sampling when the X values are kept the same.
- This might make sense when considering designed experiments.

In observational studies, the X values cannot be controlled, so it might make more sense to think about the joint distribution of the two variables Y_1 and Y_2 instead of trying to using one as the predictor (X) and the other as the outcome (Y)

- I.e., we might want to focus on correlation--understanding whether there is a relationship between two variables--rather than regression, where we try to predict one variable from the other

However, even in observational studies, the goal *is* often to predict one variable from others.

- To accomplish this, we can frame our regression as conditional inference.

Note that in observational studies, often the goal is actually to establish cause-and-effect relationships. This can be accomplished via regression *only if* a very precise set of additional assumptions are met.

Conditional Inferences

Your text describes how a bivariate normal distribution of Y_1 and Y_2 implies that the *conditional* distribution of $Y_1|Y_2$ is a normal distribution.

- Similarly, the conditional distribution of $Y_2|Y_1$ is normal.

This means that if we select (Y_1, Y_2) from a bivariate normal distribution and wish to make conditional inferences about one conditional on the other (say, $Y_1|Y_2$), then our usual normal-error regression model is entirely applicable because

1. The Y_1 observations are independent.
2. The Y_1 observations when Y_2 is considered fixed are normally distributed with mean of the form $E[Y_1|Y_2] = \alpha_{1|2} + \beta_{1|2}Y_2$ and constant variance $\sigma^2_{1|2}$.

However, the result is more general than just the situation where the variables are jointly-normally distributed.

If Y and X are random variables and

1. The conditional distributions of Y_i, given X_i, are normal and independent, with conditional means $\beta_0 + \beta_1 X_i$ and conditional variance σ^2, and
2. The X_i are independent random variables whose probability distribution $g(X_i)$ does not involve parameters $\beta_0, \beta_1, \sigma^2$,

then the regression model that we have been discussing is entirely valid;

- estimation, testing, and prediction work just as we have discussed despite X being a random variable.

The only caveat is that our concept of sampling variability now involves repeated sampling of pairs (X_i, Y_i) instead of repeated sampling of Y_i for fixed values of X_i.

- (Note also that power depends on the distribution of X, which is not under our control in an observational study).

Pearson Correlation Coefficient

In a bivariate normal model,

$$f(Y_1, Y_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho_{12}^2}} \exp\left\{ -\frac{1}{2(1-\rho_{12}^2)} \left[\left(\frac{Y_1-\mu_1}{\sigma_1} \right)^2 - 2\rho_{12} \left(\frac{Y_1-\mu_1}{\sigma_1} \right) \left(\frac{Y_2-\mu_2}{\sigma_2} \right) + \left(\frac{Y_2-\mu_2}{\sigma_2} \right.\right.\right.$$

the parameter ρ_{12} measures the association between the two variables Y_1 and Y_2.

The maximum likelihood estimator of ρ_{12} is the *Pearson product-moment correlation coefficient*:

$$\boxed{r_{12} = \frac{\sum(Y_{i1} - \bar{Y}_1)(Y_{i2} - \bar{Y}_2)}{\sqrt{\sum(Y_{i1} - \bar{Y}_1)^2 \sum(Y_{i2} - \bar{Y}_2)^2}}.}$$

- Even if Y_1 and Y_2 are not jointly-normally distributed, r_{12} provides information about the degree of linear relationship between the two variables.

- For simple linear regression of $Y_1|Y_2$ or $Y_2|Y_1$, we can see that $R^2 = r_{12}^2$
- More generally, R^2 describes the variance explained by a *model* (which will usually involve multiple predictors), while r_{12}^2 measures the linear relationship between two variables.

$$-1 \leq r_{12} \leq 1.$$

- Values of r_{12} near 1 indicate strong positive linear association
- Values of r_{12} near -1 indicate strong negative linear association
- Values of r_{12} near 0 indicate no linear association

Inferences on Correlation Coefficients

Consider testing

$$H_0 : \rho_{12} = 0 \text{ vs } H_a : \rho_{12} \neq 0.$$

This is testing whether there is a linear relationship between Y_1 and Y_2.

This is equivalent to testing for a linear relationship when regressing Y_1 on Y_2:

$$H_0 : \beta_{12} = 0 \text{ vs } H_a : \beta_{12} \neq 0.$$

It is also equivalent to testing for a linear relationship when regressing Y_2 on Y_1:

$$H_0 : \beta_{21} = 0 \text{ vs } H_a : \beta_{21} \neq 0.$$

All of these tests can be written in terms of ρ_{12}:

$$\boxed{t^* = \frac{r_{12}\sqrt{n-2}}{1-r_{12}^2}},$$

where t^* follows the $t(n-2)$ distribution *only if* H_0 is true.

Interval Estimation of ρ_{12}

When $\rho_{12} \neq 0$, the sampling distribution of r_{12} is complicated.

It is often much easier to think about the sampling distribution of the *Fisher z transformation* of r_{12}:

$$\boxed{z' = \frac{1}{2}\log\left(\frac{1+r_{12}}{1-r_{12}}\right).}$$

(Note that whenever I write $\log()$, I mean the natural logarithm $\log_e() = \ln()$. In R, for example, $log(10) = 2.3025851$, while $log(exp(1)) = 1$.)

When n is large (often $n \geq 25$ suffices), this transformed Pearson correlation is approximately normal with mean

$$E[z'] = \zeta = \frac{1}{2}\log\left(\frac{1+\rho_{12}}{1-\rho_{12}}\right)$$

and variance

$$\sigma^2\{z'\} = \frac{1}{n-3}.$$

An approximate $1-\alpha$ confidence interval for ζ is defined by

$$\boxed{z' \pm z(1-\alpha/2)\sigma\{z'\}},$$

where $z(1-\alpha/2)$ is the $(1-\alpha/2)\cdot 100$ percentile of the standard normal distribution.

An approximate $1-\alpha$ confidence interval for ρ_{12} is obtained by using the inverse of the Fisher z transformation:

$$\rho_{12} = \frac{\exp(2z') - 1}{\exp(2z') + 1}.$$

SHS: Pearson Correlation

The Canadian Survey of Household Spending is carried out annually across Canada.

(http://dli-idd-nesstar.statcan.gc.ca.proxy.library.upei.ca/webview/)

The main purpose of the survey is to obtain detailed information about household spending. Information is also collected about dwelling characteristics as well as household equipment.

The survey data are used by the following groups:

- Government departments use the data to help formulate policy;
- Community groups, social agencies and consumer groups use the data to support their positions and to lobby governments for social changes;
- Lawyers and their clients use the data to determine what is fair for child support and other compensation;
- Labour and contract negotiators rely on the data when discussing wage and cost-of-living clauses;
- Individuals and families can use the data to compare their spending habits with those of similar types of households.

```
###A subset of the latest Survey of Household Spending data are disp
spending_subset %>% datatable()
```

Show 20 entries Search:

	province	type_of_dwelling	income	marital_status	age_gr
1	NL	single_detached	68000	never_married	30-34
2	NL	single_detached	48000	never_married	25-29
3	NL	single_detached	30000	married	35-39
4	NL	row_house	30000	never_married	30-34
5	NL	single_detached	35000	married	25-29
6	NL	single_detached	26000	married	25-29
7	NL	single_detached	26000	other	55-59

Showing 1 to 20 of 500 entries

Previous 1 2 3 4 5 ... 25 Next

We are interested in the potential relationship between the income of working Canadians and the amount that they spend on clothing in a year.

Income and Clothing Expenditure for a small subset of the Survey of Household Spending are displayed below:

```
xyplot(clothing_expenditure~income, data=spending_subset,  type=c("p
```

```
clothing_model = lm(clothing_expenditure~income, data=spending_subse
msummary(clothing_model)
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.578e+03  1.864e+02   8.465 2.89e-16 ***
## income      2.116e-02  4.126e-03   5.129 4.17e-07 ***
##
## Residual standard error: 1646 on 498 degrees of freedom
## Multiple R-squared:  0.05018,    Adjusted R-squared:  0.04827
## F-statistic: 26.31 on 1 and 498 DF,  p-value: 4.174e-07
```

```
stats::cor(spending_subset$income, spending_subset$clothing_expendit
```

```
## [1] 0.2240056
```

```
mosaic::cor.test(clothing_expenditure~income, data=spending_subset, 
```

```
##
##     Pearson's product-moment correlation
##
## data:  clothing_expenditure and income
## t = 5.1292, df = 498, p-value = 4.174e-07
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1390463 0.3056914
## sample estimates:
##       cor
## 0.2240056
```

- In your own words, interpret this output.

CDI: Pearson Correlation - physicians vs hospital beds

This data set provides selected county demographic information (CDI) for 440 of the most populous counties in the United States.

Each line of the data set has an identification number with a county name and state abbreviation and provides information on 14 variables for a single county.

Counties with missing data were deleted from the data set.

```
cdi %>% datatable()
```

Show 20 entries Search:

	county	state	land_area	population	pop_18_to_34
1	Los_Angeles	CA	4060	8863164	32
2	Cook	IL	946	5105067	29.
3	Harris	TX	1729	2818199	31.
4	San_Diego	CA	4205	2498016	33
5	Orange	CA	790	2410556	32.
6	Kings	NY	71	2300664	28.
7	Maricopa	AZ	9204	2122101	29.

Showing 1 to 20 of 440 entries

Previous 1 2 3 4 5 ... 22 Next

```
xyplot(number_physicians ~ number_hospital_beds, data=cdi,  type=c("
```

```
mod_physician_beds = lm(number_physicians ~ number_hospital_beds, da
msummary(mod_physician_beds)
```

```
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -95.93218   31.49396  -3.046  0.00246 **
## number_hospital_beds   0.74312    0.01161  63.995  < 2e-16 ***
##
## Residual standard error: 556.9 on 438 degrees of freedom
## Multiple R-squared:  0.9034,    Adjusted R-squared:  0.9032
## F-statistic:  4095 on 1 and 438 DF,  p-value: < 2.2e-16
```

```
mosaic::cor.test(number_physicians ~ number_hospital_beds, data=cdi,
```

```
##
##     Pearson's product-moment correlation
##
## data:  number_physicians and number_hospital_beds
## t = 63.995, df = 438, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9405514 0.9587595
## sample estimates:
##       cor
## 0.9504644
```

- In your own words, interpret this output.

CDI: Pearson Correlation - physicians vs population

```
xyplot(number_physicians ~ population, data=cdi,  type=c("p", "r", "
```

```
mod_physician_pop = lm(number_physicians ~ population, data=cdi)
msummary(mod_physician_pop)
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.106e+02  3.475e+01  -3.184  0.00156 **
## population   2.795e-03  4.837e-05  57.793  < 2e-16 ***
##
## Residual standard error: 610.1 on 438 degrees of freedom
## Multiple R-squared:  0.8841,    Adjusted R-squared:  0.8838
## F-statistic:  3340 on 1 and 438 DF,  p-value: < 2.2e-16
```

```
mosaic::cor.test(number_physicians ~ population, data=cdi, method="p
```

```
##
##     Pearson's product-moment correlation
##
## data:  number_physicians and population
## t = 57.793, df = 438, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9283663 0.9502108
## sample estimates:
##       cor
## 0.9402486
```

- In your own words, interpret this output.

CDI: Pearson Correlation - physicians vs total income

```
xyplot(number_physicians ~ total_personal_income, data=cdi,  type=c(
```

```
mod_physician_income = lm(number_physicians ~ total_personal_income,
msummary(mod_physician_income)
```

```
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -48.39485   31.83333   -1.52    0.129
## total_personal_income   0.13170    0.00211   62.41   <2e-16 ***
##
## Residual standard error: 569.7 on 438 degrees of freedom
## Multiple R-squared:  0.8989,    Adjusted R-squared:  0.8987
## F-statistic:  3895 on 1 and 438 DF,  p-value: < 2.2e-16
```

```
mosaic::cor.test(number_physicians ~ total_personal_income, data=cdi
```

```
##
##     Pearson's product-moment correlation
##
## data:  number_physicians and total_personal_income
## t = 62.409, df = 438, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9377416 0.9567911
## sample estimates:
##       cor
## 0.9481106
```

- In your own words, interpret this output.

Spearman Rank Correlation Coefficient

When Y_1 and Y_2 differ considerably from the bivariate normal distribution, the Pearson correlation may be less useful:

- it will no longer be a complete description of the association, it will only be a measure of the linear relationship
- inference will no longer be exact, it will rely on the *central limit theorem* and, therefore, require large sample sizes

It might be possible to transform the variables to make them approximately bivariate normal, in which case we can proceed as we have discussed (as long as inference about the correlation of the transformed variables is really of interest).

Another approach that we can take is to focus on the Spearman Rank Correlation Coefficient:

$$\boxed{r_s = \frac{\sum(R_{i1} - \bar{R}_1)(R_{i2} - \bar{R}_2)}{\sqrt{\sum(R_{i1} - \bar{R}_1)^2 \sum(R_{i2} - \bar{R}_2)^2}},}$$

which is just the Pearson correlation coefficient, but with the actual values Y_{ij} replaced with their ranks R_{ij}

E..g, instead of using the actual values in the data set:

Show 20 entries Search:

	Y1	Y2
1	0.91	0.75
2	0.94	0.83
3	0.29	0.02
4	0.83	0.57

Showing 1 to 10 of 10 entries Previous 1 Next

we would use only the ranks (i.e., the ordering) of the data:

Show 20 entries Search:

	Y1	Y2
1	9	9
2	10	10
3	2	2

Showing 1 to 10 of 10 entries

This means that the Spearman Correlation measures something slightly different than the Pearson Correlation:

- Pearson Correlation measures the linear relationship
- Spearman Correlation measures a monotonic relationship
  - Spearman only looks at whether the *rank* in Y_1 increases at the same rate as the *rank* in Y_2;
  - i.e., it only looks for whether Y_1 and Y_2 vary together, not whether they vary at a constant rate.

Consider the following data:

The Pearson Correlation is

```
cor.test(Y1, Y2, method="pearson");
```

```
##
##     Pearson's product-moment correlation
##
## data:  x and y
## t = 7.0334, df = 8, p-value = 0.000109
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7170162 0.9831174
## sample estimates:
##       cor
## 0.9277897
```

The Pearson Correlation of the ranks is

```
cor.test(rank(Y1), rank(Y2), method="pearson");
```

```
##
##     Pearson's product-moment correlation
##
## data:  x and y
## t = 134217728, df = 8, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  1 1
## sample estimates:
## cor
##   1
```

Note that the Spearman Correlation is

```
cor.test(Y1, Y2, method="spearman");
```

```
##
##     Spearman's rank correlation rho
##
## data:  x and y
## S = 3.6637e-14, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
## rho
##   1
```

```
xyplot(Y2 ~ Y1,  type=c("p", "r"))
```

```
xyplot(rank(Y2) ~ rank(Y1),  type=c("p", "r"))
```

In addition,

- Pearson Correlation is *highly sensitive to outliers*, while
- Spearman Correlation is *robust to outliers*.

Consider the following data:

- What should the Correlation Coefficient equal here?

```r
cor.test(x,y, method="pearson")
```

```
##
##     Pearson's product-moment correlation
##
## data:  x and y
## t = 1.6747, df = 18, p-value = 0.1113
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.08998908  0.69650953
## sample estimates:
##       cor
## 0.3671525
```

```r
cor.test(x,y, method="spearman")
```

```
##
##     Spearman's rank correlation rho
##
## data:  x and y
## S = 1746, p-value = 0.1791
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## -0.312782
```

```
xyplot(y ~ x,  type=c("p", "r"))
```

```
xyplot(rank(y) ~ rank(x),  type=c("p", "r"))
```

Inference using Spearman Rank Correlation Coefficient

The Spearman Rank Correlation Coefficient can be used to test

H_0 : there is no association between Y_1 and Y_2 vs

H_a : there is an association between Y_1 and Y_2

or

H_a : there is a positive (negative) association between Y_1 and Y_2

This test is based on the same statistic we discussed earlier:

$$\boxed{t^* = \frac{r_s\sqrt{n-2}}{1-r_s^2}},$$

and the same t distribution with $n-2$ degrees of freedom.

SHS: Spearman Rank Correlation Coefficient

Income and Clothing Expenditure for a small subset of the Survey of Household Spending are displayed below:

```
xyplot(clothing_expenditure~income, data=spending_subset,  type=c("p
```

```
mosaic::cor.test(clothing_expenditure~income, data=spending_subset, r
```

```
##
##     Pearson's product-moment correlation
##
## data:  clothing_expenditure and income
## t = 5.1292, df = 498, p-value = 4.174e-07
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1390463 0.3056914
## sample estimates:
##       cor
## 0.2240056
```

```
mosaic::cor.test(clothing_expenditure~income, data=spending_subset, r
```

```
##
##     Spearman's rank correlation rho
##
## data:  clothing_expenditure and income
## S = 15169974, p-value = 6.409e-10
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.2718383
```

- Explain how and why the Spearman rank correlation coefficient differs from the Pearson correlation coefficient in this setting.

- The Spearman rank correlation is larger than the Pearson correlation coefficient because one variable increases when the other increases, but not at a consistent amount.

- what I know about Pearson correlation is it evaluates linear relationships between 2 continuous variables. while spearman rank correlation evaluates the monotonic relationship between two continuous or ordinal variables.

- The Spearman correlation is just the Pearson correlation using the order statistics instead of the actual numeric values, thus Pearson correlation coefficient implies linear trend, and Spearman's shows monotonic trend. In the plot above, the curve is a better fit for our data, hence the smaller Pearson correlation coefficient.

- Explain how and why the Spearman rank correlation coefficient differs from the Pearson correlation coefficient in this setting.

- the Spearman rank correlation coefficient [is] larger and more precise than the Pearson correlation coefficient in this setting...

```
mosaic::cor.test(clothing_expenditure~income, data=spending_subset, m
```

```
##
##     Spearman's rank correlation rho
##
## data:  clothing_expenditure and income
## S = 15169974, p-value = 6.409e-10
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.2718383
```

```
DescTools::SpearmanRho(spending_subset$income, spending_subset$cloth
```

```
##       rho    lwr.ci    upr.ci
## 0.2718383 0.1886446 0.3511581
```

CDI: Spearman Rank Correlation Coefficient - physicians vs hospital beds

```
xyplot(number_physicians ~ number_hospital_beds, data=cdi,  type=c("
```

```
mosaic::cor.test(number_physicians ~ number_hospital_beds, data=cdi,
```

```
##
##     Pearson's product-moment correlation
##
## data:  number_physicians and number_hospital_beds
## t = 63.995, df = 438, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9405514 0.9587595
## sample estimates:
##       cor
## 0.9504644
```

```
DescTools::SpearmanRho(cdi$number_hospital_beds, cdi$number_physicia
```

```
##       rho    lwr.ci    upr.ci
## 0.8873043 0.8656227 0.9056646
```

- In your own words, interpret this output.

CDI: Spearman Rank Correlation Coefficient - physicians vs population

```
xyplot(number_physicians ~ population, data=cdi,  type=c("p", "r", "
```

```
mosaic::cor.test(number_physicians ~ population, data=cdi, method="p
```

```
##
##     Pearson's product-moment correlation
##
## data:  number_physicians and population
## t = 57.793, df = 438, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9283663 0.9502108
## sample estimates:
##       cor
## 0.9402486
```

```
DescTools::SpearmanRho(cdi$population, cdi$number_physicians, conf.l
```

```
##       rho    lwr.ci    upr.ci
## 0.8786101 0.8553839 0.8983105
```

- In your own words, interpret this output.

CDI: Spearman Rank Correlation Coefficient - physicians vs total income

```
xyplot(number_physicians ~ total_personal_income, data=cdi,  type=c(
```

```
mosaic::cor.test(number_physicians ~ total_personal_income, data=cdi
```

```
##
##     Pearson's product-moment correlation
##
## data:  number_physicians and total_personal_income
## t = 62.409, df = 438, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9377416 0.9567911
## sample estimates:
##       cor
## 0.9481106
```

```
DescTools::SpearmanRho(cdi$total_personal_income, cdi$number_physici
```

```
##       rho    lwr.ci    upr.ci
## 0.8911783 0.8701906 0.9089378
```

- In your own words, interpret this output.

Recap: Sections 2.11

After Sections 2.11, you should be able to

- Contrast regression and correlation
- Conduct and interpret inference on correlation coefficients
- Estimate, interpret, test, and contrast Spearman rank correlation.