# Chapter 4

## STAT 3240

Michael McIsaac

UPEI

# 4: Simultaneous Inferences and Other Topics in Regression Analysis

## Learning Objectives for Sections 4.1-4.3

After Sections 4.1-4.3, you should be able to

- Compute and interpret Bonferroni and Working-Hotelling simultaneous CIs
- Compute and interpret simultaneous prediction intervals

# 4.1 Joint Estimation of $\beta_0$ and $\beta_1$

A procedure that provides a family confidence coefficient when estimating both $\beta_0$ and $\beta_1$ is often highly desirable since it permits the analyst to weave the two separate results together into an integrated set of conclusions, with an assurance that the entire set of estimates is correct.

# Bonferroni Joint Confidence Intervals

One procedure for constructing simultaneous confidence intervals for $\beta_0$ and $\beta_1$ with a specified family confidence coefficient is the Bonferroni procedure:

$$\boxed{b_0 \pm t(1-(\alpha/2)/2; n-2)s\{b_0\}}.$$

$$\boxed{b_1 \pm t(1-(\alpha/2)/2; n-2)s\{b_1\}}.$$

I.e., we use the usual CIs, but at a level of $1-\alpha/2$, so that together we have at least a simultaneous confidence of $1-\alpha$.

We reiterate that the Bonferroni $1-\alpha$ family confidence coefficient is actually a lower bound on the true (but often unknown) family confidence coefficient.

To the extent that incorrect interval estimates of $\beta_0$ and $\beta_1$ tend to pair up in the family, the families of statements will tend to be correct more than $(1-\alpha)100$ percent of the time.

# 4.2 Simultaneous Estimation of Mean Responses

Often the mean responses at a number of $X$ levels need to be estimated from the same sample data.

The combination of sampling errors in $b_0$ and $b_1$ may be such that the interval estimates of $E[Y_h]$ will be correct over some range of $X$ levels and incorrect elsewhere.

# Working-Hotelling Procedure

The Working-Hotelling procedure is based on the confidence band for the regression line discussed in Section 2.6.

The *confidence band* contains the entire regression line and therefore contains the mean responses at all $X$ levels. Hence, we can use the boundary values of the confidence band at selected $X$ levels as simultaneous estimates of the mean responses at these $X$ levels.

The family confidence coefficient for these simultaneous estimates will be at least $1-\alpha$ because the confidence coefficient that the entire confidence band for the regression line is correct is $1-\alpha$.

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

# Bonferroni Procedure

The Bonferroni procedure, discussed earlier for simultaneous estimation of $\beta_0$ and $\beta_1$ is a completely general procedure.

To construct a family of confidence intervals for mean responses at $g$ different $X$ levels with this procedure with family confidence coefficient $1-\alpha$, we use

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

For larger families, the Working-Hotelling confidence limits will always be the tighter, since $\sqrt{2F(1-\alpha;2;n-2)}$ stays the same for any number of statements in the family, whereas $t(1-(\alpha/2)/g;n-2)$ becomes larger as the number of statements increases.

In practice, once the family confidence coefficient has been decided upon, one can calculate both of these terms to determine which procedure leads to tighter confidence limits.

Note that both the Working-Hotelling and Bonferroni procedures provide lower bounds to the actual family confidence coefficient.

# SHS: Simultaneous Estimation

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:

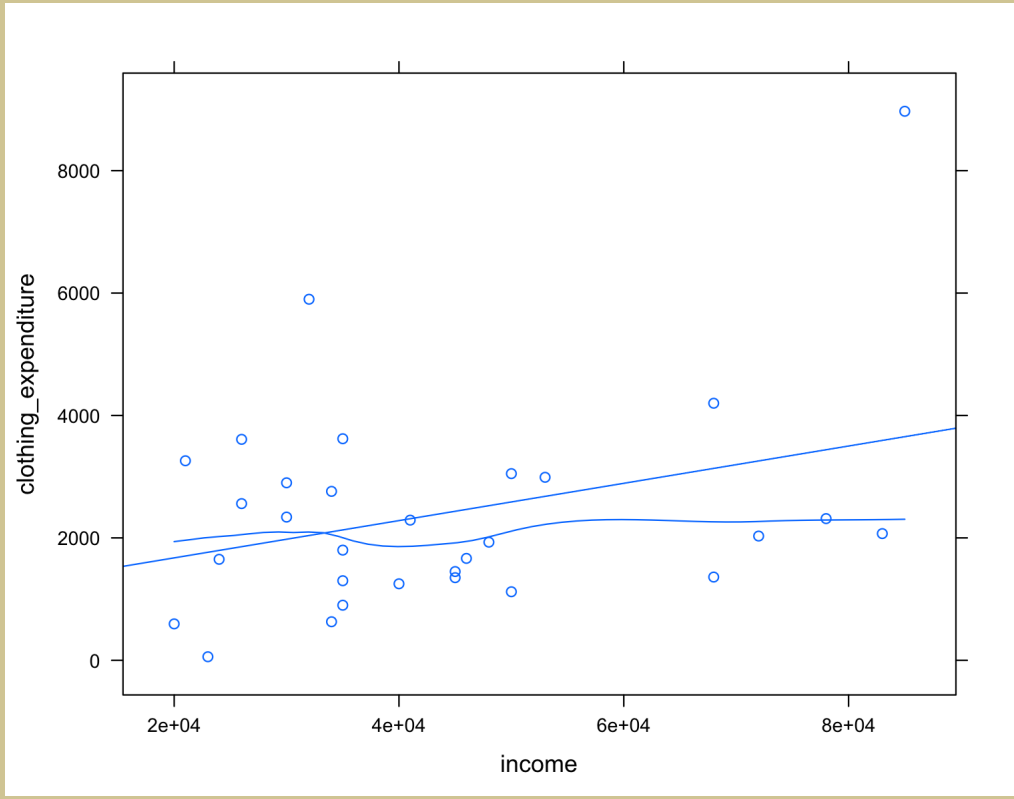
| | province | type_of_dwelling | income | marital_status | age_gr |
|---|---|---|---|---|---|
| 1 | NL | single_detached | 68000 | never_married | 30-34 |
| 2 | NL | single_detached | 48000 | never_married | 25-29 |
| 3 | NL | single_detached | 30000 | married | 35-39 |
| 4 | NL | row_house | 30000 | 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
```



```
clothing_model = lm(clothing_expenditure~income, data=spending_subse
confint(clothing_model, level=.95) %>% round(4)
```

```
##                  2.5 %    97.5 %
## (Intercept) -540.5721 2667.9468
## income        -0.0033    0.0643
```

```
confint(clothing_model, level=.975) %>% round(4)
```

```
##                 1.25 %   98.75 %
## (Intercept) -791.2228 2918.5975
## income        -0.0086    0.0696
```

- Interpret the confidence intervals in your own words.

Notice that the $[1-(\alpha/2)]\cdot 100$ percentile of the $t$-distribution is used in order to get a confidence level of $1-\alpha$.

Similarly the $[1-(\alpha/2)/2]\cdot 100$ percentile of the $t$-distribution is used in order to get a joint confidence level of $1-(\alpha/2)$.

A family of *confidence intervals* for simultaneous estimation of $E[Y_h]$ corresponding to $X_h = 40000, 50000$, and $60000$ can be found using the following code in **R**:

```
bonf_ci=predict(clothing_model, newdata=data.frame(income=c(40000, 5
```

```
##        fit      lwr      upr
## 1 2283.546 1590.178 2976.915
## 2 2588.511 1870.635 3306.388
## 3 2893.476 1986.288 3800.664
```

```
WH_ci=ALSM::ci.reg(clothing_model, newdata=data.frame(income=c(40000
```

```
##   income      Fit Lower.Band Upper.Band
## 1  40000 2283.546   1590.491   2976.602
## 2  50000 2588.511   1870.958   3306.064
## 3  60000 2893.476   1986.698   3800.254
```

```
B = qt(1-0.1/(2*3), clothing_model$df.residual); B
```

```
## [1] 2.238312
```

```
W = sqrt(2* qf(1-.1, 2, clothing_model$df.residual)); W
```

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

- Interpret the confidence intervals in your own words.
- For each individual confidence interval we can be 96.7 percent confident our Y value lies between the interval. When the confidence intervals for the three X values are run simultaneously we can say with 90% confidence that the family of statements will be correct.
- With family confidence coefficient 0.90, we conclude that the mean clothing expenditure is between (1590.177929, 2976.915035) for income = 40000$, between (1870.634534, 3306.387990) for income = 50000$, and between (1986.288259, 3800.663827) for income = 60000$, all at once.
- We are 90% confident that all three averages of clothing expenditures lays within respective interval at the same time.

- Interpret the confidence intervals in your own words.
- We can say that with 97% confidence that as X increase Y does as well meaning, that beta1 is positive.
- we are 90% confident at income equals 4000 the population mean is between 1590.177929 and 2976.915035. We are 90% confident at income equals 5000 the population mean is between 1870.634534 and 3306.387990. We are 90% confident at income equals 6000 the population mean is between 1986.288259 and 3800.663827.
- We are 90% confident that for an income of $40,000 the mean clothing expenditure is between $1590 and $2977. As the income values increase, the confidence intervals increase in value and therefore we can make the assumption that the more income people make, the more they spend on clothing.

# 4.3: Simultaneous Prediction Intervals for New Observations

Two procedures for making simultaneous predictions will be considered here: the Scheffe and Bonferroni procedures. Both utilize the same type of limits as those for predicting a single observation, and only the multiple of the estimated standard deviation is changed.

The Scheffe procedure uses the $F$ distribution, whereas the Bonferroni procedure uses the $t$ distribution.

The simultaneous prediction limits for $g$ predictions with the Scheffe procedure with family confidence coefficient $1-\alpha$ are:

$$\hat{Y}_h \pm S \cdot s\{pred\},$$

where $S^2 = g \cdot F(1-\alpha; g, n-2)$.

The Bonferroni procedure uses

$$\hat{Y}_h \pm B \cdot s\{pred\},$$

where $B = t(1-(\alpha/2)/g; n-2)$.

The $S$ and $B$ multiples can be evaluated in advance to see which procedure provides tighter prediction limits.

Note that both the $B$ and $S$ multiples for simultaneous predictions become larger as $g$ increases.

This contrasts with simultaneous estimation of mean responses where the $B$ multiple becomes larger but not the $W$ (Working-Hotelling) multiple.

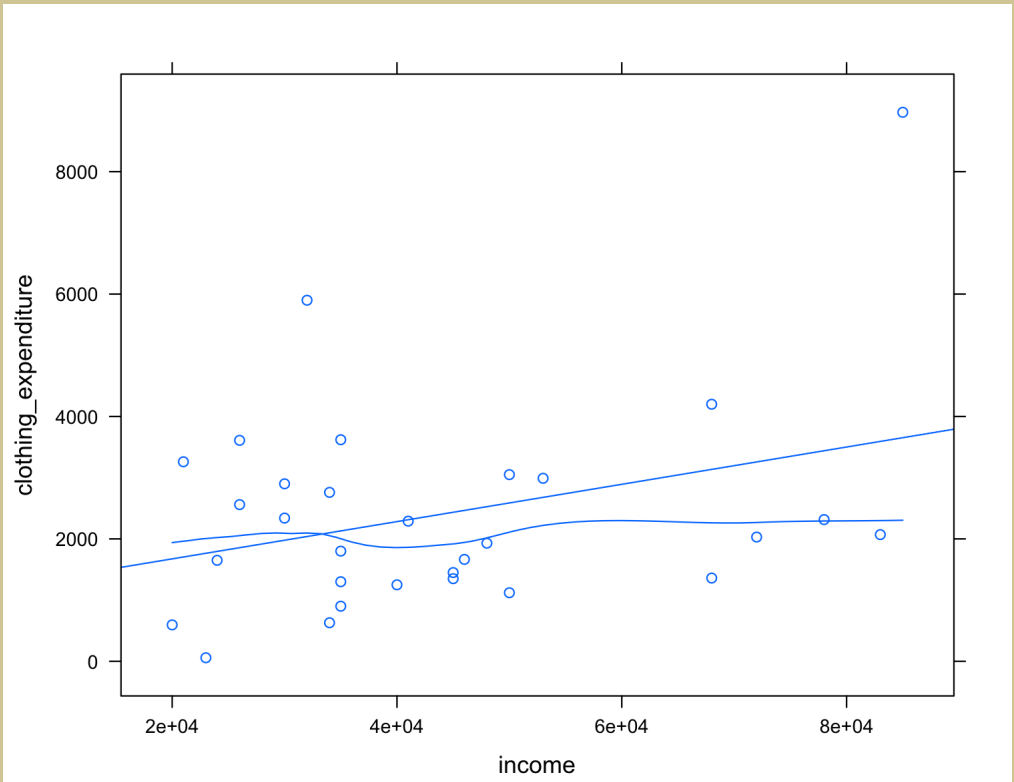
When $g$ is large, both the $B$ and $S$ multiples for simultaneous predictions may become so large that the prediction intervals will be too wide to be useful.

Other simultaneous estimation techniques might then be considered, as discussed in Reference 4.1.

# SHS: Simultaneous Prediction

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 family of *prediction intervals* for simultaneous prediction of $Y_h$:

```
bonf_ci = predict(clothing_model, newdata=data.frame(income=c(40000,
```

```
##        fit        lwr      upr
## 1 2283.546 -1502.3075 6069.400
## 2 2588.511 -1201.9079 6378.930
## 3 2893.476  -937.3098 6724.262
```

```
bonf_ci2 = ALSM::ci.reg(clothing_model, newdata=data.frame(income=c(
```

```
##   income      Fit Lower.Band Upper.Band
## 1  40000 2283.546 -1502.3075   6069.400
## 2  50000 2588.511 -1201.9079   6378.930
## 3  60000 2893.476  -937.3098   6724.262
```

```
  scheffe_ci = ALSM::ci.reg(clothing_model, newdata=data.frame(incom
```

```
##   income      Fit Lower.Band Upper.Band
## 1  40000 2283.546   -2858509    2863076
## 2  50000 2588.511   -2865107    2870284
## 3  60000 2893.476   -2926208    2931994
```

Those `type="s"` intervals do not seem plausible...

Let's look at what `ci.reg` is doing when `type="s"`:

```
ALSM::ci.reg
```

```
## function (model, newdata, type = c("b", "s", "w", "n", "m", "nm",
##     "gn"), alpha = 0.05, m = 1)
## {
##     type <- match.arg(type)
##     newdata <- as.data.frame(newdata)
##     if (dim(newdata)[2] == length(names(model$coeff))) {
##         colnames(newdata) <- names(model$coeff)
##     }
##     else {
##         colnames(newdata) <- names(model$coeff)[-1]
##     }
##     CI <- predict(model, newdata, se.fit = T)
##     g <- nrow(newdata)
##     p <- ncol(newdata) + 1
##     syh <- CI$se.fit
##     spred <- sqrt(CI$residual.scale^2 + (CI$se.fit)^2)
##     spredmean <- sqrt((CI$residual.scale^2)/m + (CI$se.fit)^2)
##     b <- qt(1 - alpha/(2 * g), model$df)
##     s <- sqrt(g * qf(1 - alpha, g, model$df))
##     w <- sqrt(p * qf(1 - alpha, p, model$df))
##     if (match.arg(type) == "b") {
##         s <- syh
##         z <- b
##     }
##     else if (match.arg(type) == "s") {
##         s <- spred
##         z <- s
##     }
##     else if (match.arg(type) == "w") {
##         s <- syh
```

```
ci.reg_fixed = function (model, newdata, type = c("b", "s", "w", "n"
{
    type <- match.arg(type)
    newdata <- as.data.frame(newdata)
    if (dim(newdata)[2] == length(names(model$coeff))) {
        colnames(newdata) <- names(model$coeff)
    }
    else {
        colnames(newdata) <- names(model$coeff)[-1]
    }
    CI <- predict(model, newdata, se.fit = T)
    g <- nrow(newdata)
    p <- ncol(newdata) + 1
    syh <- CI$se.fit
    spred <- sqrt(CI$residual.scale^2 + (CI$se.fit)^2)
    spredmean <- sqrt((CI$residual.scale^2)/m + (CI$se.fit)^2)
    b <- qt(1 - alpha/(2 * g), model$df)
    s <- sqrt(g * qf(1 - alpha, g, model$df))
    w <- sqrt(p * qf(1 - alpha, p, model$df))
    if (match.arg(type) == "b") {
        s <- syh
        z <- b
    }
    else if (match.arg(type) == "s") {
        z <- s
        s <- spred
    }
    else if (match.arg(type) == "w") {
        s <- syh
        z <- w
    }
    else if (match.arg(type) == "n") {
        s <- spred
        z <- qt(1 - alpha/2, model$df)
    }
```

21 / 61

A family of *prediction intervals* for simultaneous prediction of $Y_h$:

```
bonf_ci = predict(clothing_model
```

```
##        fit        lwr      upr
## 1 2283.546 -1502.3075 6069.400
## 2 2588.511 -1201.9079 6378.930
## 3 2893.476  -937.3098 6724.262
```

```
scheffe_ci = ALSM::ci.reg(clothi
```

```
##   Lower.Band Upper.Band
## 1   -2858509    2863076
## 2   -2865107    2870284
## 3   -2926208    2931994
```

```
S = sqrt(3*qf(1-.1, 3, clothing_
```

```
## [1] 2.621409
```

```
(6717.366- -2150.273) / (6069.40
```

```
## [1] 1.171154
```

```
bonf_ci2 = ALSM::ci.reg(clothing
```

```
##   Lower.Band Upper.Band
## 1 -1502.3075   6069.400
## 2 -1201.9079   6378.930
## 3  -937.3098   6724.262
```

```
scheffe_ci = ci.reg_fixed(clothi
```

```
##   Lower.Band Upper.Band
## 1  -2150.273   6717.366
## 2  -1850.655   7027.677
## 3  -1592.965   7379.917
```

```
B = qt(1-.1/(2*3), clothing_mode
```

```
## [1] 2.238312
```

```
S/B
```

```
## [1] 1.171154
```

- Interpret the prediction intervals in your own words.
- we are 90% confident that if we were to ask 3 people whose income fall in the range of the 3 levels 40000, 50000, and 60000), their corresponding clothing expenditure would ... be in the 3 obtained PIs, respectively.
- we predict that in the next sample taken, the observed Y(new) value of clothing expenditure will be between (-1502.3074691, 6069.400433) for income = 40000$, between (-1201.9078682, 6378.930393) for income = 50000$, and between (-937.3097597, 6724.261845) for income = 60000$, all at once.
- We are 90% confident that our next observation for an income of $40,000 will fall between -1502 and 6069, and so on for our other observations of Xh.
    - In this case, we disregard the negative values as it is not possible to spend negative dollars.

- Interpret the prediction intervals in your own words.
- With confidences of 97% you can say that when some ones income is 40000$ they are likely to spend between -1502.31$ and 6069.40$. You can say the same about each of the other points as well.
- we conclude that the predicted mean clothing expenditure is between -1502.3074691 and 6069.400433 when income is 40000, between -1201.9078682 and 6378.930393 when income is 50000 and between -937.3097597 and 6724.261845 when income is 60000.
- We are 90% sure that a person with income 40000, 50000 or 60000 will spend (-1502,6069), (-1201,6378) or (-937,6724) on clothes.
- For the first prediction interval, we can conclude with 90% confidence that the expected amount spent on clothes for an income of $40000 is between -$1502.3074691 and $6069.400433. Similarly, for the second prediction interval, we can conclude with 90% confidence that the expected amount spent on clothes for an income of $50000 is between -$1201.9078682 and $6378.930393. Finally, for the third prediction interval, we can conclude with 90% confidence that the expected amount spent on clothes for an income of $60000 is between -$937.3097597 and $6724.261845.

# CDI: Simultaneous CIs and PIs - 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()
```

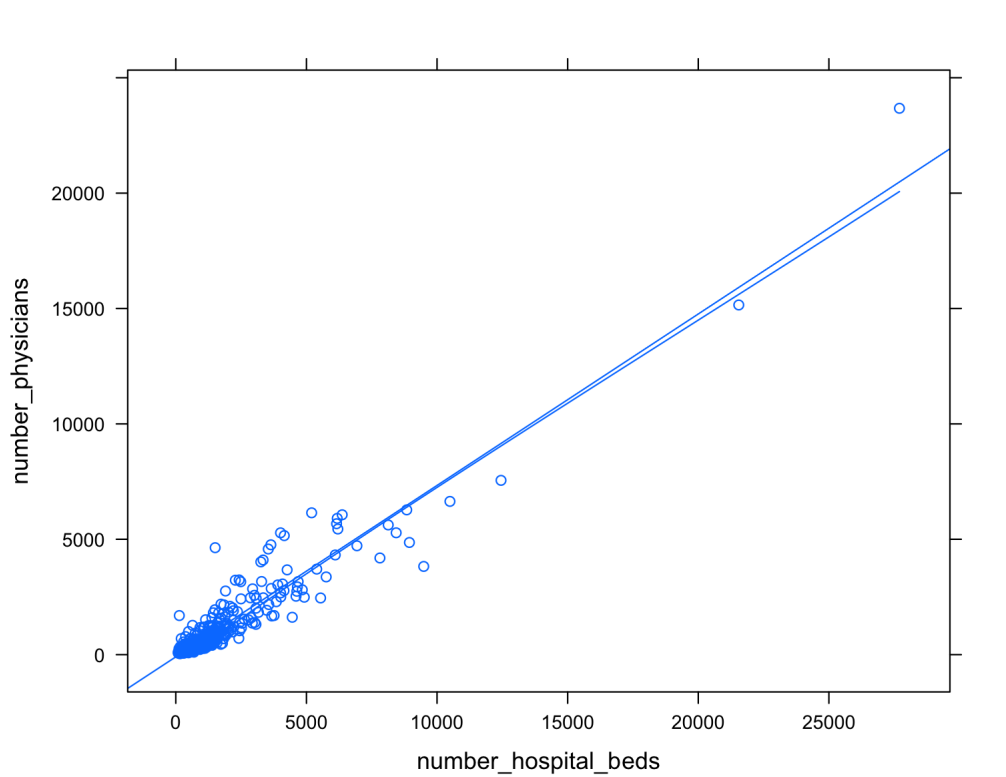
Show 20 entries Search:

| | 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
predict(mod_physician_beds, newdata=data.frame(number_hospital_beds=
```

```
##         fit        lwr       upr
## 1  1018.742   959.0148  1078.470
## 2  3619.650  3509.5556  3729.744
## 3 11050.814 10692.1448 11409.484
## 4 18481.979 17864.2500 19099.708
```

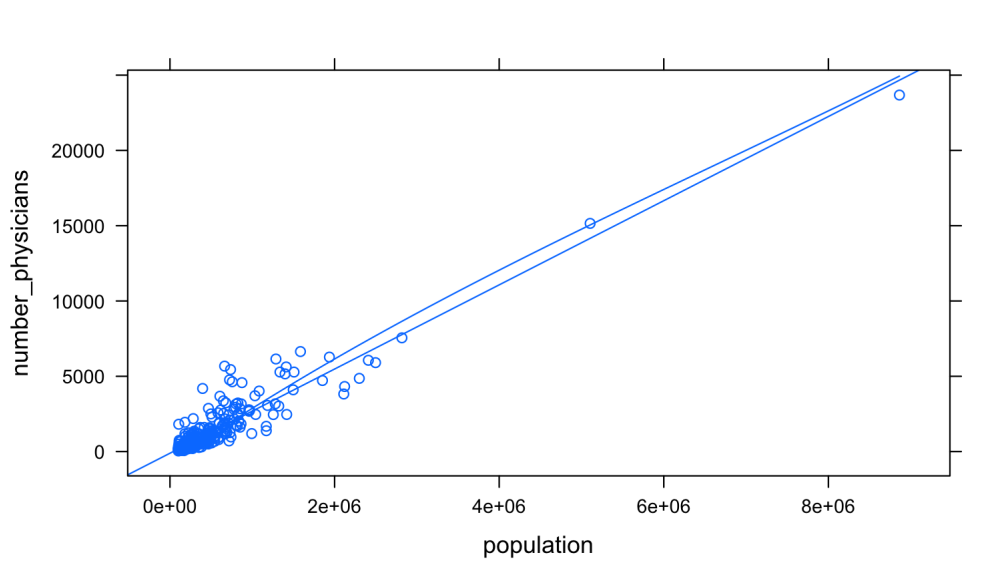
```
predict(mod_physician_beds, newdata=data.frame(number_hospital_beds=
```

```
##         fit        lwr       upr
## 1  1018.742  -235.3335  2272.818
## 2  3619.650  2362.1685  4877.132
## 3 11050.814  9747.8245 12353.804
## 4 18481.979 17085.2942 19878.664
```

- Interpret these intervals in your own words.

# CDI: physicians vs population

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



```
mod_physician_pop = lm(number_physicians ~ population, data=cdi)
predict(mod_physician_pop, newdata=data.frame(population=c(200000, 4
```

```
##          fit        lwr        upr
## 1   448.4502   369.4239   527.4765
## 2  1007.5352   932.2852  1082.7852
## 3 11071.0647 10613.4697 11528.6597
## 4 16661.9145 15956.2521 17367.5769
## 5 22252.7642 21297.8827 23207.6458
```

```
ci.reg(mod_physician_pop, newdata=data.frame(population=c(200000, 40
```

```
##   population        Fit Lower.Band Upper.Band
## 1      2e+05   448.4502   373.4242   523.4762
## 2      4e+05  1007.5352   936.0943  1078.9760
## 3      4e+06 11071.0647 10636.6332 11505.4963
## 4      6e+06 16661.9145 15991.9727 17331.8562
## 5      8e+06 22252.7642 21346.2188 23159.3097
```

```
B = qt(1-0.05/(2*5), mod_physician_pop$df.residual); B
```

```
## [1] 2.5871
```

```
W = sqrt(2* qf(1-.05, 2, mod_physician_pop$df.residual)); W
```

```
## [1] 2.456142
```

30 / 61

```
mod_physician_pop = lm(number_physicians ~ population, data=cdi)
predict(mod_physician_pop, newdata=data.frame(population=c(200000, 4
```

```
##          fit        lwr       upr
## 1   448.4502 -1131.8777  2028.778
## 2  1007.5352  -572.6084  2587.679
## 3 11071.0647  9427.7193 12714.410
## 4 16661.9145 14932.9985 18390.830
## 5 22252.7642 20408.0448 24097.484
```

```
ci.reg_fixed(mod_physician_pop, newdata=data.frame(population=c(2000
```

```
##   population        Fit Lower.Band Upper.Band
## 1      2e+05   448.4502  -1593.374   2490.274
## 2      4e+05  1007.5352  -1034.050   3049.121
## 3      4e+06 11071.0647   8947.821  13194.309
## 4      6e+06 16661.9145  14428.111  18895.718
## 5      8e+06 22252.7642  19869.340  24636.189
```

```
B = qt(1-.05/(2*5), mod_physician_pop$df.residual); B
```

```
## [1] 2.5871
```

```
S = sqrt(5*qf(1-.05, 5, mod_physician_pop$df.residual)); S
```

```
## [1] 3.3426
```

# Recap: Sections 4.1-4.3

After Sections 4.1-4.3, you should be able to

- Compute and interpret Bonferroni and Working-Hotelling simultaneous CIs
- Compute and interpret simultaneous prediction intervals

# Learning Objectives for Sections 4.5, 4.7

After Sections 4.5 and 4.7, you should be able to

- Understand the potential impact of measurement error
- Understand the challenges of choosing X levels when designing an experiment

# 4.5: Effects of Measurement Errors

# Measurement Errors in $Y$

When random measurement errors are present in the observations on the response variable $Y$, no new problems are created when these errors are uncorrelated and not biased (positive and negative measurement errors tend to cancel out).

Consider, for example, a study of the relation between the time required to complete a task $(Y)$ and the complexity of the task $(X)$. The time to complete the task may not be measured accurately because the person operating the stopwatch may not do so at the precise instants called for.

As long as such measurement errors are of a random nature, uncorrelated, and not biased, these measurement errors are simply absorbed in the model error term $\varepsilon$.

The model error term always reflects the composite effects of a large number of factors not considered in the model, one of which now would be the random variation due to inaccuracy in the process of measuring $Y$.

# SHS: Measurement Error in $Y$

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

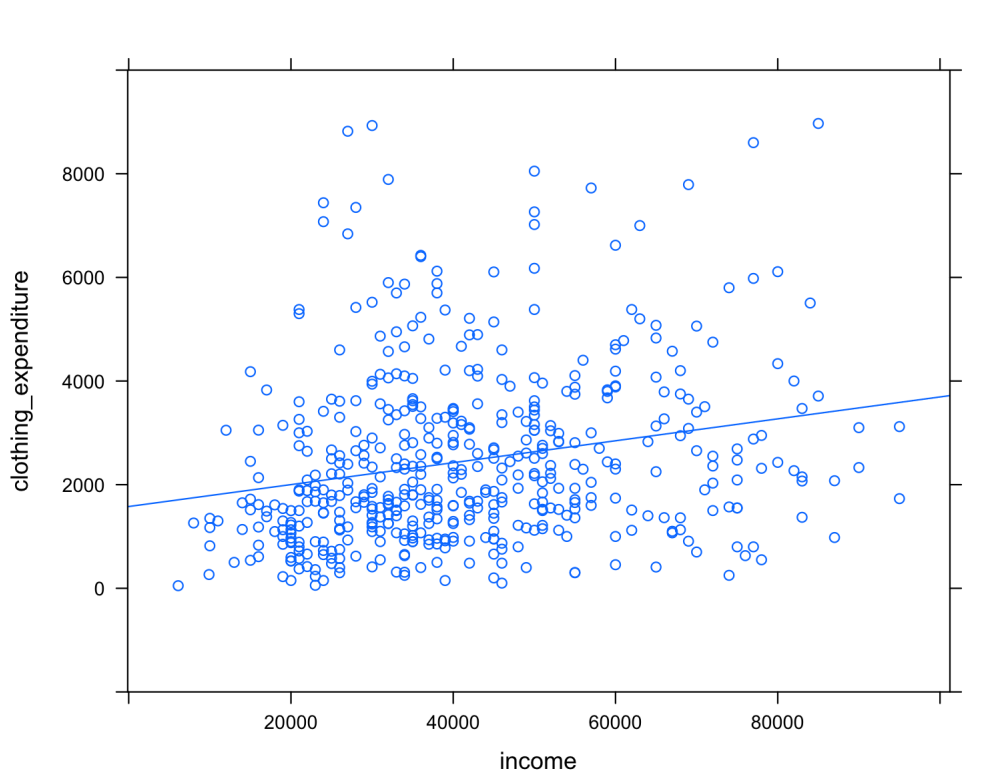
```
random_error = rnorm(n=length(spending_subset$income), mean=0, sd=sd
spending_subset$clothing_expenditure_plus_error = spending_subset$cl

spending_subset[, c("income", "clothing_expenditure", "clothing_expe
```

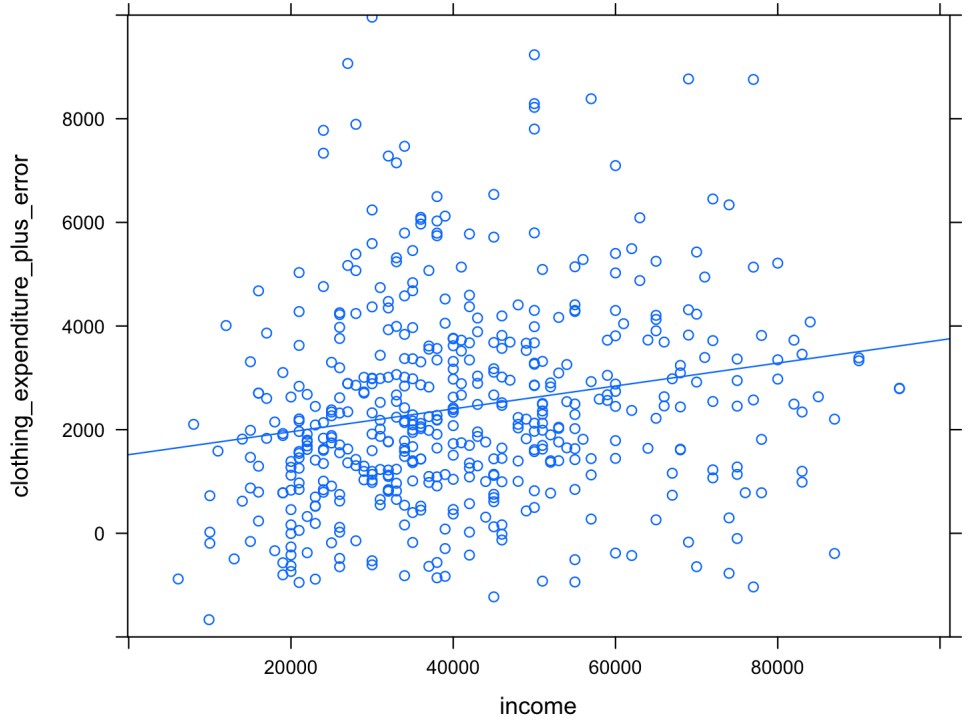
Show 20 entries Search:

| | income | clothing_expenditure | clothing_expenditure_plus_error |
|---|---|---|---|
| 1 | 68000 | 4200 | 3095.8 |
| 2 | 48000 | 1930 | 2033.3 |
| 3 | 30000 | 2340 | 2993.7 |
| 4 | 30000 | 2900 | 2886.5 |
| 5 | 35000 | 1300 | -177.2 |
| 6 | 26000 | 3610 | 4220.5 |
| 7 | 26000 | 2560 | 3975.0 |

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



```
xyplot(clothing_expenditure_plus_error~income, data=spending_subset,
```



```
msummary(lm(clothing_expenditure~income, data=spending_subset))
```

```
##               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
```

```
msummary(lm(clothing_expenditure_plus_error~income, data=spending_su
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.517e+03  2.200e+02   6.896 1.63e-11 ***
## income      2.211e-02  4.869e-03   4.541 7.02e-06 ***
##
## Residual standard error: 1942 on 498 degrees of freedom
## Multiple R-squared:  0.03977,    Adjusted R-squared:  0.03784
## F-statistic: 20.62 on 1 and 498 DF,  p-value: 7.019e-06
```

```
anova(lm(clothing_expenditure~income, data=spending_subset))
```

```
## 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
```

```
anova(lm(clothing_expenditure_plus_error~income, data=spending_subse
```

```
## Analysis of Variance Table
##
## Response: clothing_expenditure_plus_error
##             Df     Sum Sq  Mean Sq F value    Pr(>F)
## income       1   77787053 77787053  20.624 7.019e-06 ***
## Residuals 498 1878301288  3771689
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Measurement Errors in X

Unfortunately, a different situation holds when the observations on the predictor variable $X$ are subject to measurement errors.

At times measurement errors may enter the value observed for the predictor variable, for instance, when the predictor variable is pressure in a tank, temperature in an oven, speed of a production line, or reported age of a person.

# SHS: Measurement Error in $X$

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

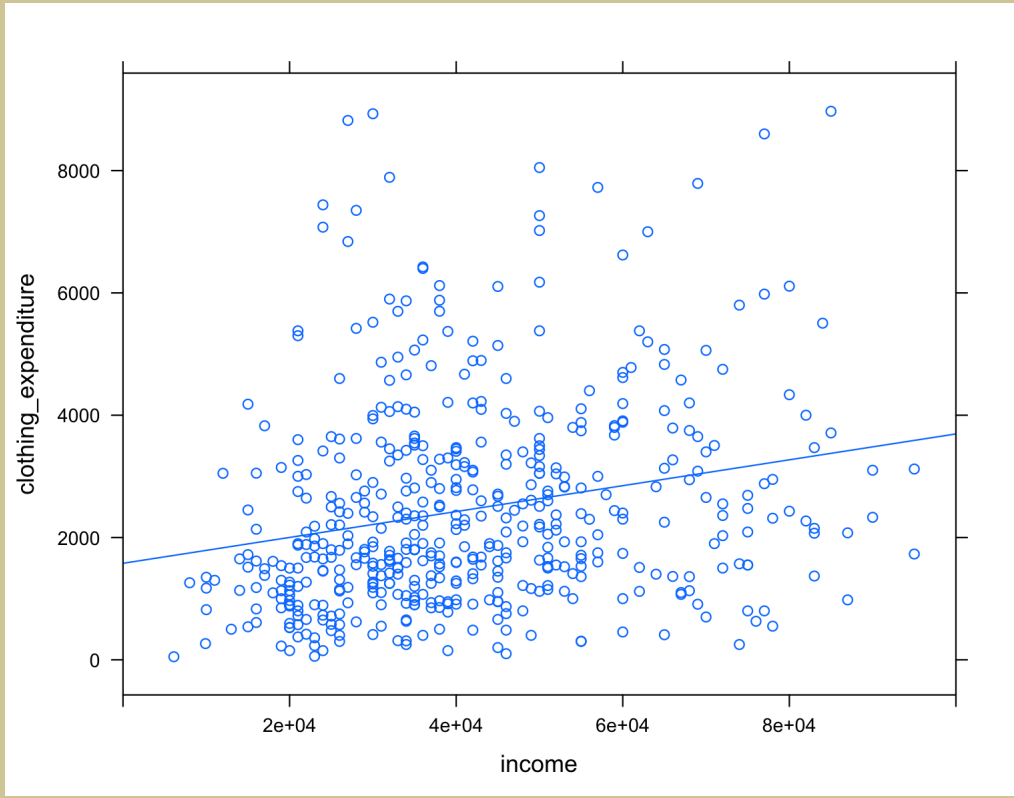
```
random_error = rnorm(n=length(spending_subset$income), mean=0, sd=sd
spending_subset$income_plus_error = spending_subset$income + random_

spending_subset[, c("income", "income_plus_error", "clothing_expendi
```

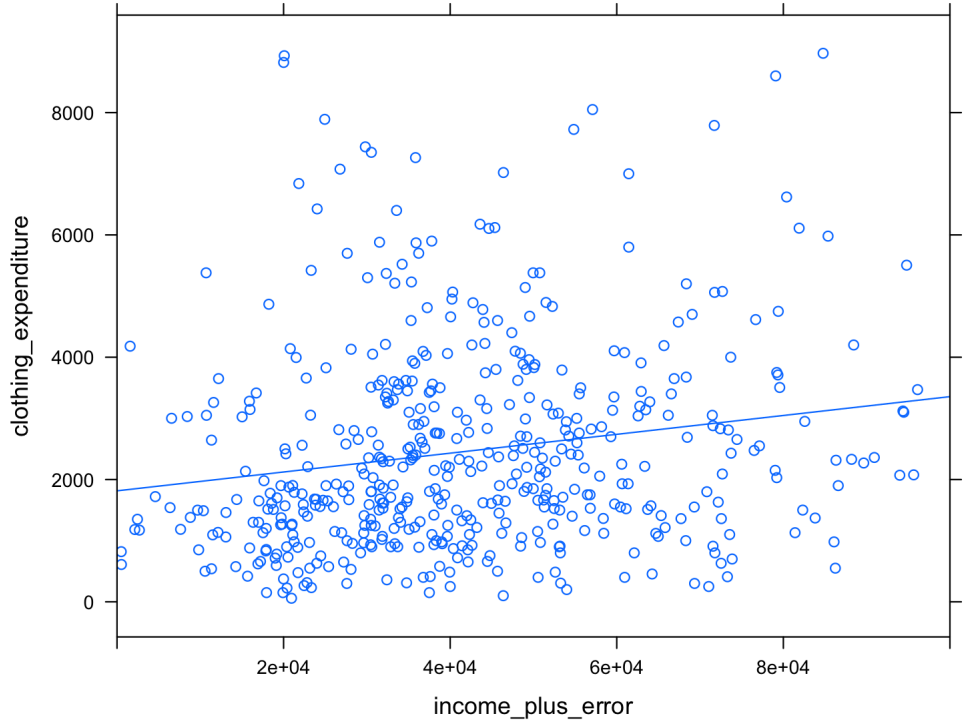
Show 20 entries Search:

| | income | income_plus_error | clothing_expenditure |
|---|---|---|---|
| 1 | 68000 | 88445.0366611871 | 4200 |
| 2 | 48000 | 47430.5200361185 | 1930 |
| 3 | 30000 | 31743.6961944071 | 2340 |
| 4 | 30000 | 36218.9115931983 | 2900 |
| 5 | 35000 | 34045.0566608039 | 1300 |
| 6 | 26000 | 35436.6200227398 | 3610 |
| 7 | 26000 | 22221.8548770456 | 2560 |

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



```
xyplot(clothing_expenditure~income_plus_error, data=spending_subset,
```



```
msummary(lm(clothing_expenditure~income, data=spending_subset))
```

```
##                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
```

```
msummary(lm(clothing_expenditure~income_plus_error, data=spending_su
```

```
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.815e+03  1.674e+02  10.838  < 2e-16 ***
## income_plus_error 1.541e-02  3.604e-03   4.277 2.27e-05 ***
##
## Residual standard error: 1658 on 498 degrees of freedom
## Multiple R-squared:  0.03543,    Adjusted R-squared:  0.03349
## F-statistic: 18.29 on 1 and 498 DF,  p-value: 2.271e-05
```

```
anova(lm(clothing_expenditure~income, data=spending_subset))
```

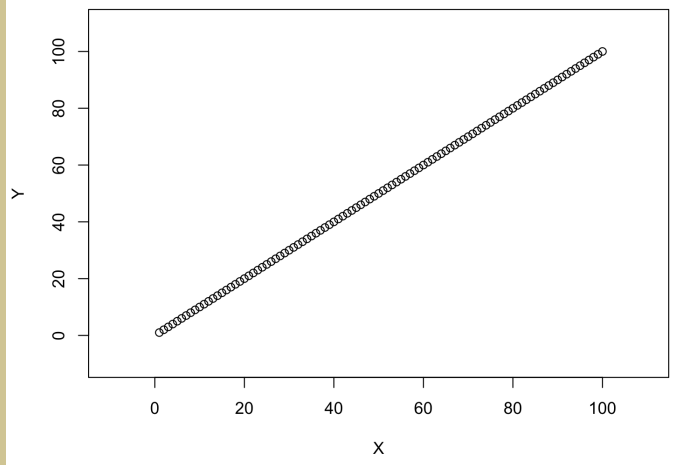
```
## Analysis of Variance Table
##
## Response: clothing_expenditure
##             Df     Sum Sq  Mean Sq F value    Pr(>F)
## income       1   71258930 71258930  26.309 4.174e-07 ***
## Residuals  498 1348849454  2708533
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(lm(clothing_expenditure~income_plus_error, data=spending_subse
```

```
## Analysis of Variance Table
##
## Response: clothing_expenditure
##                    Df     Sum Sq  Mean Sq F value    Pr(>F)
## income_plus_error   1   50316977 50316977  18.293 2.271e-05 ***
## Residuals         498 1369791407  2750585
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## Introducing measurement error into $Y$

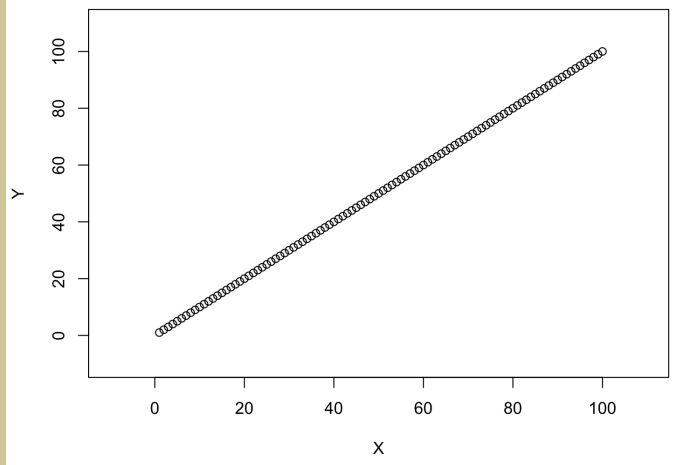
```
X = 1:100
Y = X
plot(X, Y, ylim=c(-10, 110), xli
```




```
##
## for (i in 1:1000) {
## Y = Y + rnorm(100)
## plot(X, Y, ylim=c(-10, 110),
## abline(lm(Y~X))
## lines(c(-10, 110), c(-10, 11(
```

## Introducing measurement error into $X$

```
X = 1:100
Y = X
plot(X, Y, ylim=c(-10, 110), xli
```

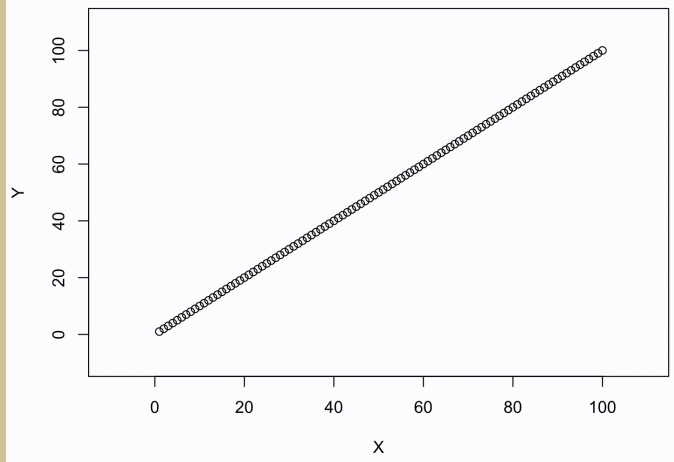



```
##
## for (i in 1:1000) {
## X = X + rnorm(100)
## plot(X, Y, ylim=c(-10, 110),
## abline(lm(Y~X))
## lines(c(-10, 110), c(-10, 11(
```

## Introducing measurement error into $Y$

```
set.seed(354)
X = 1:100
Y = X
plot(X, Y, ylim=c(-10, 110), xli
```

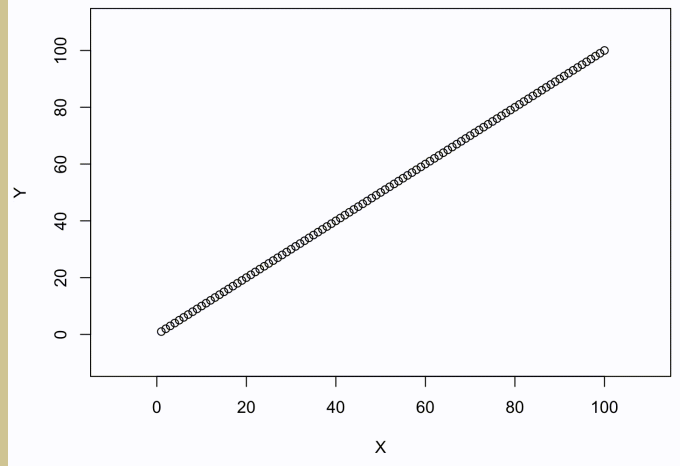
```
for (i in 1:5000) {
Y = Y + rnorm(100)
plot(X, Y, ylim=c(-10, 110), xli
abline(lm(Y~X), lwd=2)
lines(c(-10, 110), c(-10, 110),
```




## Introducing measurement error into $X$

```
X = 1:100
Y = X
plot(X, Y, ylim=c(-10, 110), xli
```

```
for (i in 1:5000) {
X = X + rnorm(100)
plot(X, Y, ylim=c(-10, 110), xli
abline(lm(Y~X), lwd=2)
lines(c(-10, 110), c(-10, 110),
```




48 / 61

# Classical Measurement Error

The regression model that we would like to study is

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

However, we observe only $X_i^* = X + \delta_i$, so we actually study
$Y_i = \beta_0^* + \beta_1^* X_i^* + \varepsilon_i^*$
Even if $E[X_i^*] = X_i$, we will not get the same estimates from these models.

Notice that

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \varepsilon_i \\ &= \beta_0 + \beta_1 (X_i^* - \delta) + \varepsilon_i \\ &= \beta_0 + \beta_1 X_i^* + (\varepsilon_i - \beta_1 \delta) \end{aligned}$$

isn't an ordinary regression model because $X_i^*$ and $\delta$ are correlated.

This leads to an *attenuation* of the parameter estimate. In fact, we can show that

$$\beta_1^* = \beta_1 \frac{\sigma_X^2}{\sigma_X^2 + \sigma_\delta^2}$$

```
msummary(lm(clothing_expenditure~income, data=spending_subset))
```

```
##                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
```

```
msummary(lm(clothing_expenditure~income_plus_error, data=spending_su
```

```
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.815e+03  1.674e+02  10.838  < 2e-16 ***
## income_plus_error 1.541e-02  3.604e-03   4.277 2.27e-05 ***
##
## Residual standard error: 1658 on 498 degrees of freedom
## Multiple R-squared:  0.03543,    Adjusted R-squared:  0.03349
## F-statistic: 18.29 on 1 and 498 DF,  p-value: 2.271e-05
```

$$\begin{aligned}\beta_1^* &= \beta_1 \frac{\sigma_X^2}{\sigma_X^2 + \sigma_\delta^2} \\ &= \beta_1 \frac{\sigma_X^2}{\sigma_X^2 + \sigma_X^2/4} \\ &= \beta_1(.8)\end{aligned}$$

50 / 61

- If it was discovered that income was measured with error, how would that impact your interpretation of the regression output given above?
- If income was measured with error, we would not be able to rely on the regression output given to us and therefore we would not be able to make an accurate analysis with this data.
- If it was discovered that income was measured with error, the $\beta 1$ estimate in the provided regression would be smaller than the $\beta 1$ of a regression conducted without error in measuring income.
- the output is misrepresenting reality, as observed clothing expenditure will be matched to an incorrect amount of income
- The regression coefficient would then be biased (always smaller than the true coefficient), with magnitude depending on the relative size of variance of X and Y (income and clothing expenditure).

- If it was discovered that income was measured with error, how would that impact your interpretation of the regression output given above?
- Then the assumption that the data at each point is normal would be wrong meaning that the way the data was analyzed was . wrong give us an inaccurate model.
- This depends on the circumstances under which this error was recorded. If the income levels we use in our study come from self-reports, this error should not be of any major concern. Some participants may accidentally report their income as being higher and others will report it as being lower. Overall, these errors should cancel out and the data should maintain its validity. In this case, I would not have to change my interpretation of the above data.
- if the income was measured with error then the regression output will not be helpful to provide the describe and predict function as the result is based on wrong data.

- If it was discovered that income was measured with error, how would that impact your interpretation of the regression output given above?

- If income was measured with error, our formula for regression would be different. This would mean that we need to use berkson model where we will set the predicted variable(50000,60000,70000). When the predictor variable(income) is a constant, and hence the error terms are not correlated with it. So it would provide us fixed value.

- Our interpretation of the regression output would not change. If we consider the Berkson Model, the reported income $X_i^*$ would be the person's salary mentioned on their contract. However, the true income $X_i$ would be variable, because income can vary based on sick days and overtime. So, $X_i$ would be a random variable and $X_i^*$ would be a fixed quantity. Thus, $X_i^*$ and the new error term are uncorrelated, so our regression model above can still be used to conduct tests, after applying the Least Squares procedures so that bo and b1 are considered unbiased.

# 4.7: Choice of $X$ Levels

Consider the impact of the choice of $X$ levels:

$$\sigma^2\{b_0\} = \sigma^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right] \quad (4.34)$$

$$\sigma^2\{b_1\} = \frac{\sigma^2}{\sum(X_i - \bar{X})^2} \quad (4.35)$$

$$\sigma^2\{\hat{Y}_h\} = \sigma^2\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right] \quad (4.36)$$

$$\sigma^2\{\text{pred}\} = \sigma^2\left[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right] \quad (4.37)$$

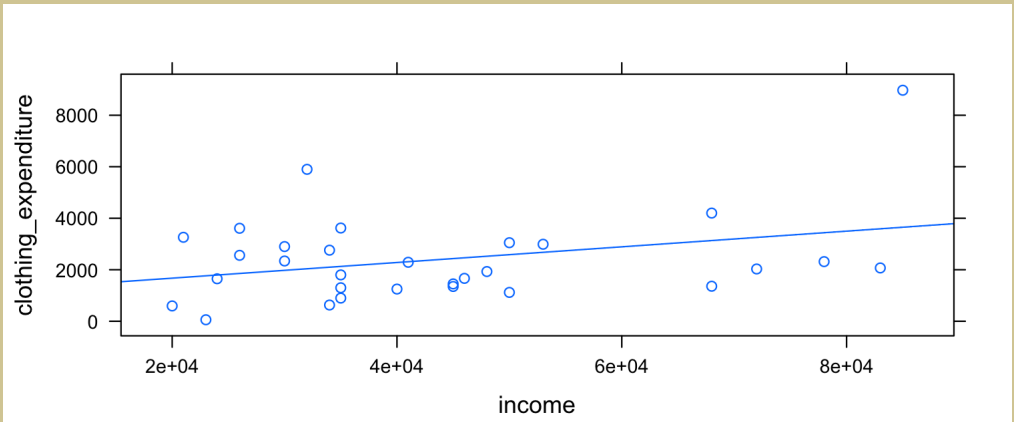
Although the number and spacing of $X$ levels depends very much on the major purpose of the regression analysis, the general advice given by D. R. Cox is still relevant:

> Use two levels when the object is primarily to examine whether or not ... (the predictor variable) . . . has an effect and in which direction that effect is.
>
> Use three levels whenever a description of the response curve by its slope and curvature is likely to be adequate; this should cover most cases.
>
> Use four levels if further examination of the shape of the response curve is important.
>
> Use more than four levels when it is required to estimate the detailed shape of the response curve, or when the curve is expected to rise to an asymptotic value, or in general to show features not adequately described by slope and curvature. Except in these last cases it is generally satisfactory to use equally spaced levels with equal numbers of observations per level.

# SHS: Choice of $X$ Levels

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
```



- If you could design an experiment where you controlled the values of income, how could that change your regression findings?
- Since we are interested in wether there is a relationship between income and clothing expenditures we want to estimate B1. Then I would choose two levels of X, the two extremes, and put one half of the population at one X and the other half at the other X. That could give us a better understanding if there is a linear relationship or not.
- For a more accurate $\beta 1$, two levels of X from two different extremes (really low income and really high income) should be chosen and half the observations should be at each of the two levels. For a more accurate $\beta 0$, observations should be made at X=0. This is not a goal of our model since no income means no money to spend on clothes. Therefore, the most practical goal for our model in finding $\beta 1$, therefore the first method is the most useful in this case.

- If you could design an experiment where you controlled the values of income, how could that change your regression findings?
- That would change the study from observational to experimental, making the regression findings more predictive of future data,
    - and a stronger correlation would be implied
    - This could mean external factors could have an effect on the data.
- Using regression analysis which will minimize the effect of confounding variables.
- If you were able to control the income values then the findings could be more narrow or broad depending on how you changed them.
- ... Berkson Model...

- If you could design an experiment where you controlled the values of income, how could that change your regression findings?
- By applying the Berkson model, ...
- I would focus on certain range of income and make the range of the X small. That means I will have larger sample size and small range of predictor variable. So that the prediction and description will be more accurate.
- the change of the values of income will change the slope of regression line.
- You can minimize variance by using income values that are equal to the mean income value of the regression analysis done above.
- If the experiment was in a controlled environment then we could control the X value getting rid of part of the random factor making the data more accurate.

# Recap: Sections 4.5, 4.7

After Sections 4.5 and 4.7, you should be able to

- Understand the potential impact of measurement error
- Understand the challenges of choosing $X$ levels when designing an experiment

# Examples of Measurement Error:

- Lester F, Arbuckle TE, Peng P, and McIsaac MA (2018). Impact of exposure to phenols during early pregnancy on birth weight in two Canadian Cohort studies subject to measurement errors. *Environment International*, 120, 231-237.

- Addressing measurement error in a cumulative exposure variable: the relationship between light at night and breast cancer risk