Statistics
Tutorial 03

Philipp Scherer & Jens Wiederspohn

13.05.2020

Disclaimer!

- **The content of the slides partly relies on material of Philipp Prinz, a former Statistics tutor. Like us, he's just a student. Therefore we provide no guarantee for the content of the slides or other data/information of the tutorial.**
- **Please note that the slides will not cover the entire lecture content. To pass the exam, it is still absolutely necessary to deal with the Wooldridge in detail!**

Logarithm in regressions

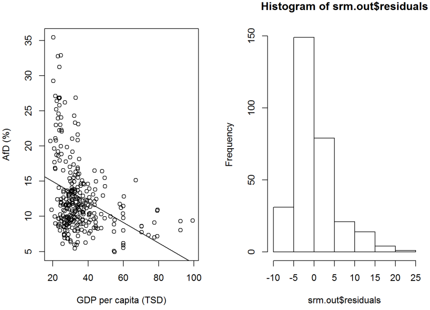

Figure 1: Original data

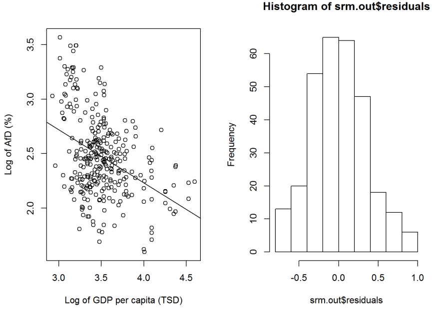

Figure 2: Log transformed data

1. If residuals are not normally distributed but right-skewed, taking the logarithm of a variable may improve fit
   - Distribution becomes more symmetric and normal
   - BUT: If we use log of left skewed distributions, it makes them even more left skewed!:(

Fun with logarithms

1. Model interpretation
   - $log(y)$ and $log(x)$: one % increase in X increases Y by $\beta\% \rightarrow \beta$ measures 'elasticity'
   - $log(y)$ and x: one unit increase in X increases Y by $\beta \cdot 100\% \rightarrow \beta$ measures 'semi-elasticity'
   - y and $log(x)$: one unit increase in X increases Y by $\beta \div 100$
2. Calculation
   - $log(x) = y$ is solution to $e^y = x$
     - $y = log(x) \Leftrightarrow exp(y) = x$
     - $log(1) = 0$ since $e^0 = 1$
   - Basic Rules
     - $log(a \cdot x) = log(a) + log(x)$
     - $log(a \div x) = log(a) - log(x)$
     - $log(x)^a = a \cdot log(x)$
     - $\frac{dlog(x)}{dx} = \frac{1}{x}$
     - Logarithmic function is inverse of exponential function
     - $log(exp(x)) = x = exp(log(x))$

Coefficient of determination

- Goodness-of-fit $\rightarrow$ How well does regression line fit the data?
  - R^2 = percentage of sample variation in y that is explained by x
    = ratio of explained variation compared to total variation
  - R^2 is bound between 0 and 1 (0% to 100%)
- $R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$
  - $SST = SSE + SSR \rightarrow \frac{SST}{SST} = \frac{SSE}{SST} + \frac{SSR}{SST} \rightarrow 1 = \frac{SSE}{SST} + \frac{SSR}{SST}$
  - *What assumption do we need to make so that this holds?*
- $R^2 = (Corr(y_i, \hat{y}_i))^2$ = squared correlation of y_i and $\hat{y}_i$
- Low/high R^2 does not always mean that model is bad/good
  - Quality of estimate does not depend directly on R^2
  - R^2 automatically grows with number of explanatory variables

Composition of OLS estimator

- β in population is unknown, we estimate $\hat{\beta}$ from our data

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2}, \text{ since } \sum -\bar{y}(x_i - \bar{x}) = 0 \\
&= \frac{\sum(x_i - \bar{x})(\beta_0 + \beta_1 x_i + u_i)}{SST_x} \\
&= \frac{\beta_0 \overbrace{\sum(x_i - \bar{x})}^{0}}{SST_x} + \frac{\beta_1 \overbrace{\sum(x_i - \bar{x})x_i}^{SST_x \text{ (see A.7)}}}{SST_x} + \frac{\sum(x_i - \bar{x})u_i}{SST_x} \\
&= \beta_1 + \frac{\sum(x_i - \bar{x})u_i}{SST_x} \\
&= \text{true } \beta_1 + \text{error}
\end{aligned}$$

$\rightarrow$ WS 2 Ex. 3c!

Unbiasedness of OLS estimator

- Recall: bias $= E(\hat{\theta}) - \theta \rightarrow$ unbiased if $E(\theta) - \theta = 0$

$$
\begin{aligned}
E(\hat{\beta}_1) &= E(\beta_1) + E\left(\frac{\sum(x_i - \bar{x})u_i}{SST_x}\right) \\
&= \beta_1 + \frac{1}{SST_x}\sum E(x_i - \bar{x})u_i \\
&= \beta_1 + \frac{1}{SST_x}\sum(x_i - \bar{x})\underbrace{E(u_i)}_{0} \\
&= \beta_1
\end{aligned}
$$

- $E(\hat{\beta}_1) - \beta_1 = \beta_1 - \beta_1 = 0 \rightarrow$ OLS estimator is unbiased!
  - ... as long as our assumptions 1-4 hold
  - If all assumptions hold, OLS estimator is BLUE
    $\rightarrow$ best linear unbiased estimator

Variance of the OLS estimator

- We need homoskedasticity assumption: $Var(u_i|x_i) = \sigma^2$

$$Var(\hat{\beta}_1) = \underbrace{Var(\beta_1)}_{0} + Var\left(\frac{\sum(x_i - \bar{x})u_i}{SST_x}\right)$$
$$= \left(\frac{1}{SST_x}\right)^2 \underbrace{\sum(x_i - \bar{x})^2}_{SST_x} \underbrace{Var(u_i)}_{\sigma^2}$$
$$= \frac{1}{SST}\sigma^2 = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$$

- When is $Var(\hat{\beta}_1)$ large?
  - Large error variance and little variability in x
    $\rightarrow$ increase n and try to account for unobservables $\rightarrow$ Why?

Error variance

- Remember difference between u_i (error) and $\hat{u}_i$ (residual)!
- u_i cannot be observed, but we can use $\hat{u}_i$ as an estimator

$$\sigma^2 = \frac{1}{n}\sum u_i^2$$
$$\rightarrow \hat{\sigma}^2 = \frac{1}{n}\sum \hat{u}_i^2 = \frac{SSR}{n}$$

- Biased! Does not account for $\sum \hat{u}_i = 0$ and $\sum x_i \hat{u}_i = 0$
  - "Give up" 2 df to guarantee that assumptions hold
    $\rightarrow df = n - 2$ gives unbiased estimator

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum u_i^2 = \frac{SSR}{n-2}$$

Standard error of $\hat{\beta}_1$

- Estimate σ with $\hat{\sigma}$ (standard error of regression, also RMSE)
- We can use $\hat{\sigma}$ to estimate SE of our regressors

$$sd(\hat{\beta}_1) = \sqrt{\frac{\sigma^2}{SST_x}} = \frac{\sigma}{\sqrt{SST_x}}$$

$$se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{SST_x}} = \frac{\hat{\sigma}}{\sqrt{\sum(x_i - \bar{x})^2}}$$

Multiple Linear Regression

- More than one explanatory variable
  - $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$
- x_2 is taken out of the error term u
  - Effect of x_2 was not accounted for before, but now it is
  - Explicitly accounting for x_2 allows to hold it constant
    $\rightarrow$ Effect of x_1 on y holding x_2 constant (& vice versa)
  - Relaxes assumption $Cov(x, u) = 0$
- Useful for generalizing functional relationships
  - Suppose too much learning can harm your grades
  - Include quadratic term to account for u-shaped relationship
    - $grade = \beta_0 + \beta_1 hours + \beta_2 hours^2 + u$
    - *What type of effects do you expect?*

OLS estimates for $k > 1$ (optional)

- OLS estimation works analogous to the case where k=1
  $\rightarrow$ Minimize sum of squared residuals

$$\begin{aligned} &\arg\min_{u^2} \sum_{i=1}^{n} u_i^2 \\ = &\arg\min_{u^2} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2})^2 \end{aligned}$$

- FOC w.r.t. $\hat{\beta}_1$:

$$\frac{\partial \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2})^2}{\partial \hat{\beta}_1} =$$

$$-2 \sum_{i=1}^{n} x_{i1} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}) = 0$$

  - Computation is usually performed with computer program

Dataframe & notation

y_i	x_{i1}	x_{i2}	x_{i3}	$\dots$	x_{ik}
y_1	x_{11}	x_{12}	x_{13}	$\dots$	x_{1k}
y_2	x_{21}	x_{22}	x_{23}	$\dots$	x_{2k}
y_3	x_{31}	x_{32}	x_{33}	$\dots$	x_{3k}
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
y_n	x_{n1}	x_{n2}	x_{n3}	$\dots$	x_{nk}

- Top-down: units from $i = 1$ to n
- Left-right: variables from $j = 1$ to k
    - x_{23} = value for second person on x_3

Interpretation of regression equation

- $\hat{\beta}_0$: intercept, predicted value of y when $x_k = 0$
- Coefficients $\hat{\beta}_k$ have partial effect (c.p.) interpretations

$$\Delta \hat{y} = \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2$$

- Multiple regression allows us to mimic a ceteris paribus style data collection without restricting values of any independent variables
- If x_2 is held fixed, we have $\Delta x_2 = 0$
  - $\Delta \hat{y} = \hat{\beta}_1 \Delta x_1$
  - $\hat{\beta}_1 =$ change in $\hat{y}$ due to a one-unit increase in x_1 if x_2 is fixed
  - Works similarly if we have more than two explanatory variables
- Likewise, if x_1 is held fixed, we have $\Delta x_1 = 0$
  - $\Delta \hat{y} = \hat{\beta}_2 \Delta x_2$

Regression coefficient

- Effect of x_1 = change in $\hat{y}$ per change in x_1

$$
\begin{aligned}
\Delta \hat{y} =& \hat{\beta}_1 \Delta x_1 \\
\rightarrow & \hat{\beta}_1 = \frac{\Delta \hat{y}}{\Delta x_1}
\end{aligned}
$$

Controlling for confounders

- Controlling = holding confounders fixed
  - Hold x_2,x_3,... fixed when we are interested in the effect of x_1
  - x_2 etc. should be confounders of x_1 and y
- Suppose we have $\hat{vote} = \hat{\beta}_0 + \hat{\beta}_1 wage + \hat{\beta}_2 foreign + \hat{\beta}_3 age$
  - If foreign and age are fixed, $\hat{\beta}_1$ = effect of wage
  - If only wage varies, it is responsible for changes in vote
- Allows us to keep other factors fixed, similar to laboratory
  - Requires correctly specified model $\rightarrow$ Lab is still ideal case!

Fitted values

- Fitted value = predicted value
- Value that $\hat{y}$ takes if certain values of x_k are inserted
  - Prediction of $\hat{y}$ for individual with certain combination of x_k
  - $E(grade)$ for x_1 = male, x_2=11 lectures, x_3=13 tutorials?

Coefficient of determination for MLR

- We still have $\underbrace{SST}_{\text{total}} = \underbrace{SSE}_{\text{explained}} + \underbrace{SSR}_{\text{residual}}$
- Divide by SST to get:

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

  - Sample variation in y_i that is explained by our model
- $R^2 = corr(y_i, \hat{y}_i)^2$
- In MLR, R^2 never decreases with additional variables!
  - More variables $\rightarrow$ larger SSE $\rightarrow$ larger R^2
  - Also applies if the connection is purely random and close to 0
  - Number of variables not taken into account $\rightarrow$ adjusted R^2

Gauss-Markov for $k > 1$

1. Model is linear in its parameters β_k
2. We have a random sample of n observations
3. No perfect collinearity
   - x_k varies ($= x_k$ is not constant)
   - No perfect correlation between the x_k
   - Fails if we have too few observations
     - For $k+1$ parameters, we need $n \geq k+1$
4. Zero conditional mean $\rightarrow E(u|x_1, x_2, ..., x_k) = 0$
   - Given all independent variables, u is 0 on average
5. Homoskedasticity $\rightarrow Var(u|x_1, x2, ..., x_k) = \sigma^2$
   - Variance in error term is equal for all combinations of x_k
   - $Var(u)$ does not change with explanatory variables

Perfect collinearity

- Perfect collinearity: $x_k^* =$ exact linear function of other x_k
  - Results in **perfect** correlation between independent variables
  - E.g. date of birth and age; p_{dem},p_{rep} and vote margin
- Stat software cannot compute $\hat{\beta}_k$ with perfect correlation
  - Suppose $corr(x_1, x_2) = 1 \rightarrow$ perfect correlation
  - *Can we keep x_2 constant and change only x_1 to get $\hat{\beta}_1$?*
  - Standard errors of perfectly correlated variables are infinite
- Does not apply to x and x^2! $\rightarrow x^2$ no linear function of x

$\rightarrow$ More on that in section on multicollinearity

Expected value & unbiasedness

- Like in bivariate case, $E(\hat{\beta}_k) = \beta_k$ for $k = 0, 1, ..., k$
  - If Gauss-Markov assumptions 1-4 hold
  - Conditional on the explanatory variables (see App. 3A)
- Estimator $\hat{\beta}_k$ is unbiased estimator of population β_k
- *Can estimates of regression coefficients be unbiased? Why?*

Overspecification

- Inclusion of an irrelevant variable (let's call it x_{irr})
  - Independent variable has no partial effect on y ($\rightarrow \beta_{irr} = 0$)
  - If all other variables are controlled for, x_{irr} has no effect on y
- True population model vs. overspecified regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_{irr} x_{irr}$$

- What if $\beta_{irr} = 0$, but we include $\hat{\beta}_{irr}$ in the regression?
  - **Estimators are still unbiased**
    - $E(\hat{\beta}_0) = \beta_0, E(\hat{\beta}_1) = \beta_1, ..., E(\hat{\beta}_{irr}) = 0$
      $\rightarrow \hat{\beta}_{irr}$ is 0 on average
  - **Negative effect on variances** of the OLS estimators
    - See multicollinearity

Underspecification

- Omitting a variable that belongs in the true model
- True population model vs. underspecified regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$
$$\hat{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$$

- Are estimators still unbiased? $\rightarrow$ We know $E(\tilde{\beta}_1) = E(\hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}_1)$
  - $\tilde{\beta}_1$ coefficient in underspecified model
  - $\hat{\beta}_1, \hat{\beta}_2$ partial effects of x_1, x_2 on $\hat{y}$
  - $\tilde{\delta}_1$ = slope parameter of the regression of x_2 on x_1.

$$\begin{aligned} \rightarrow Bias(\tilde{\beta}_1) &= E(\tilde{\beta}_1) - \beta_1 \\ &= \beta_1 + \beta_2 \tilde{\delta}_1 - \beta_1 \qquad \text{(MLR.1 - MLR.4)} \\ &= \beta_2 \tilde{\delta}_1 \end{aligned}$$

  - *When is $\tilde{\beta}_1$ unbiased?*

Omitted variable bias (OVB)

- No need to worry about OVB in the model above if:
  1. β_2 that is omitted is 0 in the population model
     - Model is correctly specified
  2. $Corr(x_1, x_2) = 0$, i.e. x_1 and x_2 are uncorrelated
     - If $Corr(x_1, x_2) = 0$, x_2 cannot be a confounder of y and x_1
     - Controlling for non-confounders does not change estimates

- If $\beta_2 \neq 0$ and $Corr(x_1, x_2) \neq 0$, we have $OVB = \beta_2 \tilde{\delta}_1$
  - Even if β_2 is unknown, direction of bias can be assessed!

TABLE 3.2 Summary of Bias in $\tilde{\beta}_1$ when x_2 Is Omitted in Estimating Eqution (3.40)

	$Corr(x_1, x_2) > 0$	$Corr(x_1, x_2) < 0$
$\beta_2 > 0$	Positive bias	Negative bias
$\beta_2 < 0$	Negative bias	Positive bias

Figure 3: Table 3.2: Bias in $\tilde{\beta}_1$ if x_2 is omitted

- *What could lead to OVB in $p_{cdu} = \tilde{\beta}_0 + \tilde{\beta}_{christ}$? Is $\beta_2 \tilde{\delta}_1 \gtrless 0$?*

OVB for $k > 2$

- Suppose we omit x_3 from the model
  - x_3 in error term u might be correlated with x_1 and x_2
    $\rightarrow$ violates MLR.4 $\rightarrow$ biased OLS estimators!
- Bias, if there is pairwise correlation between x_1, x_2 and x_3
  - e.g. $\tilde{\beta}_2$ is only unbiased if x_2 is not correlated with x_1 and x_3
- With assumptions, we can (easily) obtain direction of bias:
  - If $corr(x_1, x_2) = 0 \rightarrow OVB(\tilde{\beta}_1) = E(\tilde{\beta}_1) - \beta_1 = \beta_3 \tilde{\delta}_{1,3}$
    - *Why can we treat x_2 as absent if $corr(x_1, x_2) \approx 0$?*
  - Interpretation like in the case with two explanatory variables

Variance of β_k

- $Var(\beta_k)$ determines test statistics and precision of estimators
  - Confidence intervals and accuracy of hypothesis testing
- $Var(\beta_k) = \frac{\sigma^2}{SST_k(1-R_k^2)}$
  - Requires that all Gauss-Markov assumptions hold
    1. $E(y|x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$ GM1-GM4
    2. $Var(y|x) = \sigma^2$ GM5

Variance of β_k

- $Var(\beta_k) = \frac{\sigma^2}{SST_k(1-R_k^2)}$ depends on
  1. Error variance σ^2
     - More noise in regression equation $\rightarrow$ larger $Var(\beta_k)$
     - Population property, therefore independent of n
     - Can only be reduced if some factors are taken out of error term
  2. Total sample variation in x_k
     - Larger total variation in x_k $\rightarrow$ smaller $Var(\beta_k)$
     - Total variation can be increased by increasing n
     - If $SST_k = 0$, GM3 is violated
  3. Goodness-of-fit R_k^2
     - Regression of all independent variables on x_k instead of y
     - R_k^2 increases if explanatory variables are strongly correlated
     - $x_1, x_2, \ldots$ explain x_k very well $\rightarrow$ large R_k^2 $\rightarrow$ large $Var(\beta_k)$
     - As R_k^2 approximates 1, $Var(\beta_k)$ approximates ∞
       $\rightarrow$ $R_k^2 = 1$ violates GM3 (no perfect linear combination)
       $\rightarrow$ For R_k^2 close to 1, we have "*multicollinearity*"

Multicollinearity

- Extent to which independent variables are correlated
  - Explanatory variables can predict another one
    - Height & weight, education & income, product age & price
- Large $R_k^2 \rightarrow$ large $Var(\beta_k)$
- Results in unstable and unreliable regression estimates
  - Difficulties to measure effect of an independent variable on y
  - Small test statistic and large confidence intervals
  - Imprecise regression coefficients ($+$ large se)
- Perfect correlation $\rightarrow$ perfect multicollinearity
- Solutions:
  - Increase sample size (smaller sampling error)
  - Remove highly-correlated independent variables
  - Replace highly-correlated variables by new variable (e.g. index)

Trade-off unbiasedness vs. variance

- In an underspecified model we have:
  - $bias(\tilde{\beta}) > bias(\hat{\beta})$
  - $Var(\tilde{\beta}) < Var(\hat{\beta})$, if $corr(x_1, x_2) \neq 0$
- Bias does not depend on n $\rightarrow$ choose unbiased estimator
  - Variance shrinks to 0 as n gets larger
  - Multicollinearity issues can be countered with larger n

Standard error

- Recall that $\hat{\sigma}^2 = \frac{SSR}{n-k-1}$ is an unbiased estimator for σ^2

$$
\begin{aligned}
sd(\hat{\beta}_k) &= \sqrt{Var(\hat{\beta}_k)} \\
&= \frac{\sigma}{\sqrt{SST_k(1 - R_k^2)}} \\
se(\hat{\beta}_k) &= \frac{\hat{\sigma}}{\sqrt{SST_k(1 - R_k^2)}} \\
&= \frac{\hat{\sigma}}{\sqrt{n}sd(x_k)\sqrt{1 - R_k^2}}
\end{aligned}
$$

- *What sample size is needed to cut the standard error in half?*

Gauss-Markov Theorem

- There are many unbiased estimators, why use OLS estimator?
- If all GM assumptions hold, OLS is BLUE
  - Best Linear Unbiased Estimator
  - Linear estimator with the smallest variance
  - If GM assumptions are violated, theorem does not hold. E.g.:
    - GM4 $\rightarrow$ no longer unbiased
    - GM5 $\rightarrow$ no longer smallest variance

Exercise 1a and b

What is SSR?

- Sum of squared residuals
- $SSR = \sum(\hat{u}_i)^2$

How can you obtain the left hand side of Equation 2 from the right hand side of Equation 1? Describe it briefly.

$$SSR = \sum_i [\hat{\beta}_1^2 x_{i1}^2 - 2\hat{\beta}_1 x_{i1}(y_i - \hat{\beta}_0 - \hat{\beta}_2 x_{i2}) + (y_i - \hat{\beta}_0 - \hat{\beta}_2 x_{i2})^2] \quad (1)$$

From the right hand side of this Equation, we can obtain

$$\sum_i [2\hat{\beta}_1 x_{i1}^2 - 2x_{i1}(y_i - \hat{\beta}_0 - \hat{\beta}_2 x_{i2})] \quad (2)$$

for $\hat{\beta}_1$ by doing a partial derivation with respect to $\hat{\beta}_1$ and NOT as usual to x. Simple example:

- If $F(x_1, x_2) = \hat{\beta}_1^3 x_1^2 + x_2^2$, its derivation with respect to $\hat{\beta}_1$ would be:
- $f(x_1, x_2) = 3\hat{\beta}_1^2 x_1^2$

Exercise 1c and d

Above, we could obtain the unique solution for the estimates. But this is not always the case. Which assumption assures such an unique solution?

- No Perfect Collinearity
- $SSR = \sum(\hat{u}_i)^2$

In which scenario is the assumption of the last task violated? Mark all the scenarios in which the assumption is violated.

A another binary variable for a revised ordinance of the State Government $\rightarrow$ perhaps similar, but certainly not identical to x_1

B **another variable** $x_3 = x_1 \times 2 \rightarrow x_3$ increases simultaneous to x_1

C another variable $x_3 = (x_1)^2 \rightarrow x^2$ no linear function of x!

D **another variable** $x_3 = x_1 + x_2 \rightarrow x_3$ increases simultaneous to x_1 & x_2

Exercise 1e

Before enforcement of the ordinance of the State Government, how much increase of the infected persons is predicted for a day?

- Remember:
  - $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u = -7376.70 + 669.89 x_1 - 2796.09 x_2$
  - $\Delta y = \beta_1 * \Delta x_1 + \beta_2 * \Delta x_2$
- We are interested in the marginal change (change from one day to the next) for $\Delta x_1 = 1$ and $\Delta x_2 = 0$
- Therefore we note $\Delta y = 669.89 * 1 - 2796.09 * 0 = 669.89$

Exercise 1f

After enforcement of the ordinance of the State Government, how much increase of the infected persons is predicted for a day?

- We are still interested in the marginal change Δy, but this time for $x_1 = 1$ and $x_2 = 1$
- If x_2 would constantly increase as x_1 does, we would note $\Delta y = 669.89 * 1 - 2796.09 * 1 = -2126.2$
- Seems like the ordinance of the State Government is having the desired effect. Or perhaps not?
- Because x_2 stays constant (since day 23), the increase of the infected persons per day still is
- $\Delta y = 669.89 * 1 - 2796.09 * 0 = 669.89$

Exercise 1g

Predict the number of infected persons at the 21st day and 24th day.

- Now, we are not longer interested in the marginal change but in the total number of infected persons!
- Remember:
  - $y = -7376.70 + 669.89x_1 - 2796.09x_2$
- Insert the appropriate values for 21st ($x_1 = 21$ and $x_2 = 0$) and for 24th ($x_1 = 24$ and $x_2 = 1$)
- $y_1 = -7376.70 + 669.89 * 21 - 2796.09 * 0 = 6690.99$
- $y_2 = -7376.70 + 669.89 * 24 - 2796.09 * 1 = 5904.57$

Exercise 1h

Calculate the residual at the 21st day and 24th day.

- Formula for calculating residuals:
  - $\hat{u}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = y_i - \hat{y}_i$
  - where y_i present the true values (see the table on WS, p.1) and $\hat{y}_i$ our estimates (see 1g)
- $\hat{u}_1 = y_i - \hat{y}_i = 1105 - 6690.99 = -5585.99$
- $\hat{u}_2 = y_i - \hat{y}_i = 2748 - 5904.57 = -3156.57$

Exercise 1i

If the variance of all residuals is 7768923**, calculate SSR.**

- Starting point:
  - $\frac{1}{n}\sum \hat{u}_i^2 = \frac{SSR}{n}$
- Arranging formula yields:
  - $SSR = \frac{1}{n}\sum \hat{u}_i^2 * n = 7768923 * 66 = 512748918$

Exercise 1j

If the variance of y is 144894582, calculate SST.

- Same procedure: $Var(y) = \frac{SST}{n}$
- Arranging formula yields:
  - $SST = Var(y) * n = 144894582 * 66 = 9563042412$
  - Including df correction:
    $SST = Var(y) * (n - 1) = 144894582 * 65 = 9418147830$

If the population variance was estimated, we would had to take account for uncertainty, by including a Degrees of Freedom correction $n - 1$ into our formula.

Exercise 1k and l

Calculate the R-squared

- $R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$
- $R^2 = 1 - \frac{SSR}{SST}$
- $R^2 = 1 - \frac{512748918}{9563042412}$
- $R^2 = 0.9463822$

Which (in)equation is true?

- A is true: $\delta > 0$
    - δ is bigger than zero because x_2 and x_1 are positive correlated. With a higher value on x_1 it is more likely that x_2 has also an higher value, respectively is more likely to be 1 and not 0.

Exercise 1m

Assume that $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ are BLUE. Is $\tilde{\beta}_1$ biased?

- B is true: Yes, and the bias is negative.
- $\delta > 0$ and $\hat{\beta}_2 < 0$, because of that is the bias negative
- $E(\tilde{\beta}_1) = E(\hat{\beta}_1 + (-\hat{\beta}_2)\tilde{\delta}_1)$
- $E(\tilde{\beta}_1) = \beta_1 + (-\beta_2)\tilde{\delta}_1$
- $Bias(\tilde{\beta}_1) = E(\tilde{\beta}_1) - \beta_1$
- $= \beta_1 + (-\beta_2)\tilde{\delta}_1 - \beta_1$
- $= (-\beta_2)\tilde{\delta}_1$

Exercise 1n

What is the name of the assumptions assuring the OLS estimates being BLUE?

- Gauss-Markov-Assumption
  - If all GM assumptions hold, OLS is BLUE
    - Best Linear Unbiased Estimator
    - Linear estimator with the smallest variance
    - If GM assumptions are violated, theorem does not hold. E.g.:
    - GM4 $\rightarrow$ no longer unbiased
    - GM5 $\rightarrow$ no longer smallest variance

Exercise 1o and p

It is known that the variance of $\hat{\beta}$ is as follows: $Var(\hat{\beta}) = \frac{\sigma^2}{SST_j(1-R_j^2)}$**. What is** σ^2

- A is true: Variance of the errors

We can estimate σ^2 **as follows:** $\hat{\sigma}^2 = \frac{1}{???}\sum_i \hat{u}_i$**. What value do we have for ??? for Model 1?**

- $\hat{\sigma}^2 = \frac{1}{(n-k-1)}\sum_{i=1}^n \hat{u}_i^2$
- $n = 66$
- $k = 2$ (number of slope parameters)
- and 1 for the intercept in the model
- $df = n - k - 1 = 63$
- $\hat{\sigma}^2 = \frac{1}{63}\sum_{i=1}^n \hat{u}_i^2$

Exercise 1q and r

How do we call the value of the last task?

- degrees of freedom
  - The term $n - k - 1$ in the last task is the degrees of freedom (df) for the general OLS problem with n observations and k independent variables. Since there are $k + 1$ parameters in a regression model with k independent variables and an intercept, we can write:
  - $df = n - (k + 1)$
  - $df =$ (number of observations) $-$ (number of estimated parameters)

Concerning the variance of estimates, what is true?

- A is true: $Var(\tilde{\beta}) < Var(\hat{\beta})$
- $Var(\tilde{\beta})$ is always smaller than $Var(\hat{\beta})$, unless x_1 and x_2 are uncorrelated in the sample, in which case the two estimators $\tilde{\beta}$ and $\hat{\beta}$ are the same.

Exercise 1s and t

If the sample size grows infinitely, which of the three (in)equations in the last task is approximately true?

- C is true: $Var(\tilde{\beta}) = Var(\hat{\beta})$
  - $Var(\tilde{\beta})$ and $Var(\hat{\beta})$ both shrink to zero as n gets large, which means that the multicollinearity induced by adding x_2 becomes less important as the sample size grows.

Is the R-squared of Model 2 is in comparison with that of Model 1:

- A is true: smaller
  - An important fact about R-squared is that it never decreases, and it usually increases when another independent variable is added to a regression. This algebraic fact follows because, by definition, the sum of squared residuals never increases when additional regressors are added to the model.
  - Because Model 1 has an additional independent variable with an non-zero effect in the sample R-squared of model 1 is bigger that R-squared of model 2.