Financial Econometrics

Chapter 4. Prediciton, Goodness-of-fit, and Modeling Issues

Prof. Huei-Wen Teng

Spring, 2021

4.1 Least Square Prediction

confidence interval: E(y|x=x0)=beta1+beta2x
prediction interval: y0=beta1+beta2x0+e0

- To do prediction, assume y_0 and x_0 are related to one another by the same regression model,

$$y_0 = \beta_1 + \beta_2 x_0 + e_0,$$

where e_0 is a random error. The least squares predictor of y_0 comes from the fitted regression line

$$\hat{y}_0 = b_1 + b_2 x_0.$$

- To evaluate how well this predictor performs, we define the forecast error by

Least squares prediction

$$\sim N \quad f = \underset{\text{ideal prediction}}{y_0 - \hat{y}_0} = (\beta_1 + \beta_2 x_0 + \underset{\sim N}{e_0}) - (\underset{\sim N}{b_1} + \underset{\sim N}{b_2 x_0}).$$

- Because f is a linear combination of y_i and an additional random error e_0), $f \sim N(E(f) = 0, var(f))$ with

$$var(f) = \sigma^2 \left[1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}. \right] \tag{1}$$

$var(f)$ is proved in the appendix.

1. Because $E(f) = 0$, $\hat{y}_0$ is an unbiased predictor of y_0.
2. By the variance of this predictor in Eq. (1), the variance of the forecast is smaller when
   2.1 the overall uncertainty σ^2 is smaller;
   2.2 the sample N is larger;
   2.3 the variation in the explanatory variable is larger;
   2.4 the value of $(x_0 - \bar{x})^2$ is small.
3. $\hat{y}_0$ is the *best linear unbiased predictor* of y_0. [Proof is skipped.]

- For estimating variance, we use

$$\hat{var}(f) = \hat{\sigma}^2 \left[1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right].$$

Let the standard error of the forecast be $se(f) = \sqrt{\hat{var}(f)}$. Recall that $f = (y_0 - \hat{y}) \sim N(0, var(f))$, we have

$$\frac{f}{se(f)} = \frac{\hat{y}_0 - y_0}{se(f)} \sim t_{N-2}.$$

Similarly,

$$1 - \alpha = P\left(-t_c \le \frac{\hat{y}_0 - y_0}{se(f)} \le t_c \right),$$

where $t_c = qt(1 - \alpha/2, N - 2)$, the $100(1 - \alpha)\%$ *prediction interval* for y_0 is

$$\hat{y}_0 \pm t_c se(f).$$

Example: Find 95% prediction interval for y_0 when $x_0 = 20$. I

```
data("food")
alpha <- 0.05
x <- 20
xbar <- mean(food$income)
m1 <- lm(food_exp˜income, data=food)
b1 <- coef(m1)[[1]]
b2 <- coef(m1)[[2]]
yhatx <- b1+b2*x
sm1 <- summary(m1)
df <- df.residual(m1)
tcr <- qt(1-alpha/2, df)
N <- nobs(m1)   #number of observations, N
N <- NROW(food) #just another way of finding N
varb2 <- vcov(m1)[2, 2]
```

Example: Find 95% prediction interval for y_0 when $x_0 = 20$. II

```
sighat2 <- sm1$sigma^2 # estimated variance
varf <- sighat2+sighat2/N+(x-xbar)^2*varb2 #forecast variance
sef <- sqrt(varf) #standard error of forecast
lb <- yhatx-tcr*sef
ub <- yhatx+tcr*sef
```

The result is the prediction interval for the forecast (104.13,471.09). A different way of finding point and interval estimates for the predicted $E(y|x)$ and forecasted y

```
incomex=data.frame(income=20)
predict(m1, newdata=incomex, interval="confidence",level=0.95)
predict(m1, newdata=incomex, interval="prediction",level=0.95)
```

$$
\begin{aligned}
& var(\hat{y}_0) = var(b_1 + b_2 x_0) \\
= \quad & var(b_1) + x_0^2 var(b_2) + 2x_0 cov(b_1, b_2) \\
= \quad & \frac{\sigma^2 \sum x_i^2}{N \sum (x_i - \bar{x})^2} + x_0^2 \frac{\sigma^2}{\sum (x_i - \bar{x})^2} + 2x_0 \sigma^2 \frac{-\bar{x}}{\sum (x_i - \bar{x})^2} \\
= \quad & \left[\frac{\sigma^2 \sum x_i^2}{N \sum (x_i - \bar{x})^2} - \left\{ \frac{\sigma^2 N \bar{x}^2}{N \sum (x_i - \bar{x})^2} \right\} \right] + \\
& \left[\frac{x_0^2 \sigma^2}{\sum (x_i - \bar{x})^2} + \frac{\sigma^2 (-2x_0 \bar{x})}{\sum (x_i - \bar{x})^2} + \left\{ \frac{\sigma^2 N \bar{x}^2}{N \sum (x_i - \bar{x})^2} \right\} \right] \\
= \quad & \sigma^2 \left[\frac{\sum x_i^2 - N\bar{x}^2}{N \sum (x_i - \bar{x})^2} + \frac{x_0^2 - 2x_0 \bar{x} + \bar{x}^2}{\sum (x_i - \bar{x})^2} \right] \\
= \quad & \sigma^2 \left[\frac{\sum (x_i - \bar{x})^2}{N \sum (x_i - \bar{x})^2} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right] = \sigma^2 \left[\frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right]
\end{aligned}
$$

Therefore, we have

$$var(f) = var(y_0 - \hat{y}_0) = var(e_0 - \hat{y}_0) = var(\hat{y}_0) + \sigma^2,$$

or,

$$var(f) = \sigma^2 \left[1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}. \right]$$

4.2 Measuring Goodness-of-fit

- Recall that the regression model $y_i = \beta_1 + \beta_2 x_i + e_i$, and the fitted value $\hat{y}_i = b_1 + b_2 x_i$, and the residual $\hat{e}_i = y_i - \hat{y}_i$. We write $y_i = \hat{y}_i + \hat{e}_i$, and subtract $\bar{y}$,

  $$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + \hat{e}_i.$$

  Squaring and summing over i, with the fact $\sum(\hat{y}_i - \bar{y})\hat{e}_i = 0$ (proved in the appendix), we have

  $$\sum(y_i-\bar{y})^2 = \sum(\hat{y}_i-\bar{y})^2+\sum \hat{e}_i^2+2\sum(\hat{y}_i-\bar{y})\hat{e}_i = \sum(\hat{y}_i-\bar{y})^2+\sum \hat{e}_i^2.$$

- Define SST = total sum of squares = $\sum(y_i - \bar{y})^2$, SSR = sum of squares due to regression = $\sum(\hat{y}_i - \bar{y})^2$, and SSE = sum of squares due to error = $\sum \hat{e}_i^2$. That is,

  $$\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE}.$$

- Define the *coefficient of determination*, or, R^2, as the proportion of variation in y explained by x within the regression model:

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

When $0 < R^2 < 1$, then R^2 is interpreted as “the proportion of the variation in y about its mean that is explained by the regression model”.

- Example: Calculate R^2 and interpret in the food expenditure example.

  ```
  (rsq <- sm1$r.squared) #or sm1
  ```

  If you need SSR or SSE, use ANOVA

  ```
  anov <- anova(m1)
  dfr <- data.frame(anov)
  kable(dfr,
    caption="Output generated by the ‘anova‘ function")
  SSE = anov[2,2]; SSR = anov[1,2]; SST = SSE+SSR;
  ```

- The key ingredients in a report are:
  1. the coefficient estimates
  2. the standard errors (or *t*-values)
  3. an indication of statistical significance
  4. R^2

- The correlation coefficient ρ_{xy} between x and y is defined as

$$\rho_{xy} = \frac{cov(x,y)}{\sqrt{var(x)}\sqrt{var(y)}}.$$

Substituting sample values, we get the sample correlation coefficient:

$$r_{xy} = \frac{s_{xy}}{s_x s_y},$$

where

$$\begin{aligned} s_{xy} &= \sum(x_i - \bar{x})(y_i - \bar{y})/(N-1), \\ s_x &= \sqrt{\sum(x_i - \bar{x})^2/(N-1)}, \\ s_y &= \sqrt{\sum(y_i - \bar{y})^2/(N-1)}. \end{aligned}$$

The sample correlation coefficient r_{xy} has a value between -1 and 1, and it measures *the strength of linear association* between observed values of x and y.

- Two relationships between R^2 and r_{xy}:
  1. $R^2 = r_{xy}^2$. Note that r_{xy}^2 measures the strength of the linear association between x and y. This interpretation is not far from that of R^2: the proportion of variation in y about its mean explained by x in the linear regression model.
  2. $R^2 = r_{y\hat{y}}^2$, i.e., R^2 can be computed as the square of the sample correlation coefficient between y_i and $\hat{y}_i = b_1 + b_2 x_i$. R^2 measures the linear association, or goodness-of-fit, between the sample data and their predicted values. Consequently, R^2 is sometimes called a measure of “goodness-of-fit”.

- It is a general rule that the squared sample correlation between y and its fitted value $\hat{y}$ is a valid measure of good-of-fit. A general goodness-of-fit measure, or general R^2, is defined as

$$R_g^2 = [corr(y, \hat{y})]^2 = r_{y\hat{y}}^2.$$

Show that $\sum(\hat{y}_i - \bar{y})\hat{e}_i = 0$ under the linear regression model: $y_i = \beta_1 + \beta_2 x_i + e_i$, where $e_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$. To find the least squared estimates for β_1 and β_2, denoted by b_1 and b_2, respectively, we define

$$S(\beta_1, \beta_2) = \sum(y_i - \beta_1 - \beta_2 x_i)^2,$$

and b_1 and b_2 are solutions for the following first-order-conditions,

$$\begin{aligned} \sum(y_i - \beta_1 - \beta_2 x_i) &= 0, \\ \sum x_i(y_i - \beta_1 - \beta_2 x_i) &= 0. \end{aligned}$$

Plugging b_1 and b_2 into the above equations, we obtain

$$\begin{aligned} \sum(y_i - b_1 - b_2 x_i) &= 0, \\ \sum x_i(y_i - b_1 - b_2 x_i) &= 0. \end{aligned}$$

Using the definition of residuals $\hat{e}_i = y_i - b_1 - b_2 x_i$, we have equal expressions,

$$\begin{aligned} \sum \hat{e}_i &= 0, \\ \sum x_i e_i &= 0. \end{aligned}$$

Hence,

$$\sum(\hat{y}_i - \bar{y})\hat{e}_i = \sum(b_1 + b_2 x_i)\hat{e}_i + \bar{y}\sum \hat{e}_i = b_1 \sum \hat{e}_i + b_2 \sum x_i \hat{e}_i + 0 = 0.$$

4.3 Modelling issues

1. Transforming the variables:
   1.1 Power: quadratic (x^2), cubic (x^3)
   1.2 Nature logarithm: $ln(x)$.
2. Various models

Name	Model	slope = dy/dx	Elasticity
Linear	$y = \beta_1 + \beta_2 x + e$	β_2	$\beta_2 x/y$
Quadratic	$y = \beta_1 + \beta_2 x^2 + e$	$2\beta_2 x$	$(2\beta_2 x)x/y$
Cubic	$y = \beta_1 + \beta_2 x^3 + e$	$3\beta_2 x^2 + e$	$(3\beta_2 x^2)x/y$
Log-Log	$\ln(y) = \beta_1 + \beta_2 \ln(x) + e$	$\beta_2 y/x$	β_2
Log-Linear	$\ln(y) = \beta_1 + \beta_2 x + e$	$\beta_2 y$	$\beta_2 x$
	A 1-unit increase in x leads to approximately a 100 × β_2% change in y		
Linear-Log	$y = \beta_1 + \beta_2 \ln(x) + e$	β_2/x	β_2/y
	1 % change in x leads to approximately a $\beta_2/100$ unit change in y		

3. When specifying a regression model, we may inadvertently choose an inadequate or incorrect functional form. How to avoid this? We check *residuals*
    - 3.1 informal check using *visualization tools*: Subjective yet providing more information.
        - 3.1.1 residual plots (scatter plots of residuals to check if there are discernible patterns, homoskedacity)
        - 3.1.2 histogram, density plot, QQ plot (check normality)
    - 3.2 formal check using *statistical tests*.
        - 3.2.1 Check normality using *Jarque-Bera test*: adjective yet specific and narrow in conclusion.

$$JB = \frac{N}{6}\left(S^2 + \frac{(K-3)^2}{4}\right),$$

where N is ample size, S is skewness, and K is kurtosis.

4. Example: The residuals of the of the linear-log equation of the food expenditure example. One can notice that the spread of the residuals seems to be higher at higher incomes, which may indicate that the homoskedasticity assumption is violated.

```
mod2 <- lm(food_exp~log(income), data=food)
ehat <- mod2$residuals
plot(food$income, ehat, xlab="income", ylab="residuals")
```

Consider a simple moded: $y = 1 + x + e$, create artificial dataset, and check residuals.

```
#set.seed(12345)   #sets the seed for the random number gene
x <- runif(300, 0, 10)
e <- rnorm(300, 0, 1)
y <- 1+x+e
mod3 <- lm(y~x)
ehat <- resid(mod3)
plot(x,ehat, xlab="x", ylab="residuals")
```

Consider a different model: $y = 15 - 4x^2 + e$, create artificial dataset, fit a linear regression model, and check residuals.

```
# set.seed(12345)
x <- runif(1000, -2.5, 2.5)
e <- rnorm(1000, 0, 4)
y <- 15-4*x^2+e
mod3 <- lm(y~x)
ehat <- resid(mod3)
ymi <- min(ehat)
yma <- max(ehat)
plot(x, ehat, ylim=c(ymi, yma),
     xlab="x", ylab="residuals",col="grey")
```

1. For a linear-log model:

   - Slope or marginal effect. Note that for a linear-log model, the slope is

$$m = \frac{dy}{dx} = \frac{\beta_2}{x}.$$

   - Elasticity. The elasticity is

$$\varepsilon = \frac{dy/y}{dx/x} = m\frac{x}{y} = \frac{\beta_2}{y}.$$

   - Semi-elasticity. Look at the following identity:

$$\frac{dy}{100dx/x} = \frac{(dy/dx)x}{100} = \beta_2/100.$$

2. Example: Let us estimate a linear-log model, $y = \beta_1 + \beta_2 \ln(x) + e$, for the food dataset, draw the regression curve, and calculate the marginal effects for some given values of the dependent variable. We interpret the above by that a 1% change in x leads to (approximately) a $\beta_2/100$ unit change in y.

```
data(food)
mod2 <- lm(food_exp~log(income), data=food)
tbl <- data.frame(xtable(mod2))
kable(tbl, digits=5,
      caption="Linear-log model output for the *food* exampl
b1 <- coef(mod2)[[1]]
b2 <- coef(mod2)[[2]]
```

The results for an income of \$1000 are as follows: $dy/dx = 13.217$, which indicates that an increase in income of \$100 (i.e., one unit of x) increases expenditure approximately by \$13.217; for a 1% increase in income (that is, an increase of \$10), expenditure increases approximately by \$1.322 ; and, finally, for a 1% increase in income expenditure increases approximately by 0.638%.

```
> x <- 10 #for a household earning #1000 per week
> y <- b1+b2*log(x)
> DyDx <- b2/x    #marginal effect
> DyPDx <- b2/100 #1% change in x leads to (beta2/100)-unit
> PDyPDx <- b2/y  #elasticity
> DyDx
[1] 13.21658
> DyPDx
[1] 1.321658
> PDyPDx
[1] 0.638061
```

Test if residuals in the food expenditure example are normally distributed using the Jarque-Bera statistic. Note: need `tseries` package. While the histogram may not strongly support one conclusion or another about the normlity of ehat, the Jarque-Bera test is unambiguous. Conclusion: there is no evidence against the normality because the *p*-value is 0.9622.

```
> library(tseries)
> ebar <- mean(ehat)
> sde <- sd(ehat)
> qqnorm(ehat)
> hist(ehat, col="grey", freq=FALSE, main="",
+      ylab="density", xlab="ehat")
> curve(dnorm(x, ebar, sde), col=2, add=TRUE,
+       ylab="density", xlab="ehat")
> jarque.bera.test(ehat) #(in package 'tseries')
Jarque Bera Test
data:  ehat
X-squared = 0.076997, df = 2, p-value = 0.9622
```

4.4 Polynomials models

1. The general form of a quadratic equation is

$$y = a_0 + a_1 x + a_2 x^2.$$

The general form of a cubic equation is

$$y = a_0 + a_1 + a_2 x^2 + a_3 x^3.$$

2. Example: the wa_wheat dataset gives annual wheat yield in tonnes per hectare in Greenough Shire in Western Australia over a period of 48 years. Since there is a pattern in the residuals. Consider a cubic model.

```
data("wa_wheat")
mod1 <- lm(greenough~time, data=wa_wheat)
ehat <- resid(mod1)
plot(wa_wheat$time, ehat, xlab="time", ylab="residuals")
mod2 <- lm(wa_wheat$greenough~I(time^3), data=wa_wheat)
ehat <- resid(mod2)
plot(wa_wheat$time, ehat, xlab="time", ylab="residuals")
```

4.5 Log-linear models

1. The log-linear model:

$$\ln(y) = \beta_1 + \beta_2 x.$$

A natural choice for prediction is

$$\hat{y}_n = \exp(b_1 + b_2 x),$$

where the subscript *n* is for *natural*. A better alternative is

$$\hat{y}_c = E(y) = \exp(b_1 + b_2 x + \hat{\sigma}^2/2) = \hat{y}_n \exp(\hat{\sigma}^2/2),$$

because of the property of log-normal distribution (see appendix).

2. In a log-linear model, $\log(y) = \beta_1 + \beta_2 x + e$, focus on $y = \exp(\beta_1 + \beta_2 x)$.
   - The slope is
   $$m = \frac{dy}{dx} = \beta_2 y.$$
   - The elasticity is
   $$\varepsilon = m\frac{x}{y} = \beta_2 x.$$
   - Semi-elasticity: While x_0 increases to x_1, y_0 changes to y_1. Therefore, we have

$$\begin{aligned}(\beta_1 + \beta_2 x_1) - (\beta_1 + \beta_2 x_0) &= \ln(y_1) - \ln(y_0) \\ &= \ln(\frac{y_1}{y_0}) = \ln(1 + \frac{y_1 - y_0}{y_0}) \\ &= \ln(1 + \frac{\Delta y}{y_0}) \approx \frac{\Delta y}{y_0}.\end{aligned}$$

Multiplying by 100 in the above approximation, we have

$$100\beta_2(x_1 - x_0) \approx 100\frac{\Delta y}{y_0}.$$

If x_0 changes to x_1 by just one unit, we have

$$100\beta_2 \approx 100\frac{\Delta y}{y_0}.$$

We interpret this as a 1-unit change in x leads to approximately a $100\beta_2\%$ change in y.
The percentage change in y is β means that that

$$\frac{dy}{y}100 = \beta.$$

While x changes 1 unit, the percentage change in y is

$$\frac{100\frac{dy}{y}}{dx} = 100\frac{dy}{dx}\frac{1}{y} = 100\beta_2.$$

3. Example: A growth model. Suppose that the yield in year t is

$$YIELD_t = (1+g)YIELD_{t-1},$$

with g being a fixed growth rate in 1 year. By substituting repeatedly, we obtain

$$YIELD_t = YIELD_0(1+t)^t.$$

Taking logarithm, we obtain

$$\ln(YILED_t) = \ln(YIELD_0) + \ln(1+g)t = \beta_1 + \beta_2 t.$$

This is simply a log-linear model with dependent variable $\ln(YIELD_t)$ and explanatory variable t, or time. Since $\ln(1+g) \approx g$ while g is small, β_2 is interpreted as the growth rate.

Consider the log-linear model: $log(y) = \beta_1 + \beta_2 x + e$. Let us calculate the prediction: $\hat{y}_n$ and $\hat{y}_c$, marginal effect, and semi-elasticity.
Because b2 = 0.017844, we estimate that the rate of growth in wheat production has increased at an average rate of approximately 1.78 percent per year.

```
> data(wa_wheat)
> mod4 <- lm(log(greenough)˜time, data=wa_wheat)
> smod4 <- summary(mod4)
> tbl <- data.frame(xtable(smod4))
> kable(tbl, caption="Log-linear model for the *yield* equat
```

```
|            |   Estimate| Std..Error|   t.value| Pr...t..|
|:-----------|----------:|----------:|---------:|--------:|
|(Intercept) | -0.3433665|  0.0584042| -5.879140|    4e-07|
|time        |  0.0178439|  0.0020751|  8.599107|    0e+00|
```

4. Example. The wage model:

$$\log(WAGE) = \beta_1 + \beta_2 EDUC + e.$$

The predictions and the slope are calculated for educ=12 years. Here are the results of these calculations: “natural” prediction $y_n = 14.796$; corrected prediction, $y_c = 16.996$; growth rate g=9.041; and marginal effect $dy/dx = 1.34$. The growth rate indicates that an increase in education by one unit increases hourly wage approximately by 9.041 %. (Or, an additional year of education leads to approximately 9.941% increases in wages.)

```
> data("cps4_small")
> xeduc <- 12
> mod5 <- lm(log(wage)~educ, data=cps4_small)
> data("cps4_small")
> xeduc <- 12
> mod5 <- lm(log(wage)~educ, data=cps4_small)
> smod5 <- summary(mod5)
> tabl <- data.frame(xtable(smod5))
> kable(tabl, caption="Log-linear 'wage' regression output")


|            |  Estimate| Std..Error|  t.value| Pr...t..|
|:-----------|---------:|----------:|--------:|--------:|
|(Intercept) | 1.6094445|  0.0864229| 18.62288|        0|
|educ        | 0.0904082|  0.0061456| 14.71102|        0|
>
> b1 <- coef(smod5)[[1]]
> b2 <- coef(smod5)[[2]]
> sighat2 <- smod5$sigma^2
> yhatn <- exp(b1+b2*xeduc) #"natural" predictiction
> yhatc <- exp(b1+b2*xeduc+sighat2/2) #corrected prediction
> DyDx <- b2*yhatn           #marginal effect
> yhatn
[1] 14.7958
> yhatc
[1] 16.99643
> b2
[1] 0.09040825
> DyDx
[1] 1.337662
```

The regular R^2 cannot be used to compare two regression models having different dependent variables such as a linear-log and a log-linear models; when such a comparison is needed, one can use the generalized R^2, which is R2g=$corr(y, \hat{y})^2$. Let us calculate the generalized R^2 for the quadratic and the log-linear wage models.
The quadratic model yields R2g=0.188, and the log-linear model yields R2g=0.186; since the former is higher, we conclude that the quadratic model is a better fit to the data than the log-linear one. (Note: using generalized R2 is just one criterion for model selection. Different criteria may give different suggestions.)

```
> mod4 <- lm(wage~I(educ^2), data=cps4_small)
> yhat4 <- predict(mod4)
> mod5 <- lm(log(wage)~educ, data=cps4_small)
> smod5 <- summary(mod5)
> b1 <- coef(smod5)[[1]]
> b2 <- coef(smod5)[[2]]
> sighat2 <- smod5$sigma^2
> yhat5 <- exp(b1+b2*cps4_small$educ+sighat2/2)
> rg4 <- cor(cps4_small$wage, yhat4)^2
> rg5 <- cor(cps4_small$wage,yhat5)^2
> rg4
[1] 0.1881762
> rg5
[1] 0.1859307
```

Now, we find the prediction interval estimate for the forecast about the wage in the log-linear model for educ = 12. The logic goes as follows: a log-linear model is $\log(y) = \beta_1 + \beta_2 x + e$ equals to $w = \beta_1 + \beta_2 + e$, where $w = \log(y)$ is the log-transformation of y. To forecast y, we first forecast w and obtain its prediction interval

$$(b_1 + b_2 x) + t_{cr} se(f),$$

where $se(f)$ can be calculated as the usual simple linear regression. Then we took exponential for every part to obtain the prediction interval for a forecast in y,

$$exp((b_1 + b_2 x) + t_{cr} se(f)).$$

The prediction interval for a forecast of y while educ = 12 is (5.26,41.62). We can also draw a 95% prediction band for the log-linear wage model.

```
> alpha <- 0.05
> xeduc <- 12
> xedbar <- mean(cps4_small$educ)
> mod5 <- lm(log(wage)~educ, data=cps4_small)
> b1 <- coef(mod5)[[1]]
> b2 <- coef(mod5)[[2]]
> df5 <- mod5$df.residual
> N <- nobs(mod5)
> tcr <- qt(1-alpha/2, df=df5)
> smod5 <- summary(mod5)
> varb2 <- vcov(mod5)[2,2]
> sighat2 <- smod5$sigma^2
> varf <- sighat2+sighat2/N+(xeduc-xedbar)^2*varb2
> sef <- sqrt(varf)
> lnyhat <- b1+b2*xeduc
> lowb <- exp(lnyhat-tcr*sef)
> upb <- exp(lnyhat+tcr*sef)
> lowb
```

```
[1] 5.260398
> upb
[1] 41.61581
```

Suppose $Y \sim N(\mu, \sigma^2)$. Consider $W = \exp(Y)$. Then, W is said to have a log-normal distribution with mean μ and variance σ^2. We can show that

$$E(W) = \mathrm{e}^{\mu+\sigma^2/2},$$

and

$$var(W) = \mathrm{e}^{2\mu+\sigma^2}(e^{\sigma^2} - 1).$$

Then, for a log-linear model, $\ln(y) = \beta_1 + \beta_2 x + e$ with $e \sim N(0, \sigma^2)$, then

$$E(y_i) = E(\mathrm{e}^{\beta_1+\beta_2 x_i+e_i}) = E(\mathrm{e}^{\beta_1+\beta_2 x_i}\mathrm{e}^{e_i}) = \mathrm{e}^{\beta_1+\beta_2 x_i}\mathrm{e}^{\sigma^2/2} = \mathrm{e}^{\beta_1+\beta_2+\sigma^2/2}.$$

Therefore, we predict $E(y)$ with the corrected formula:

$$E(y_i) = \mathrm{e}^{b_1+b_2 x_i+\hat{\sigma}^2/2}.$$

4.6 Log-log models

1. The log-log-function, $\ln(y) = \beta_1 + \beta_2 \ln(x)$, is widely used to describe demand equations and production functions. For a log-log model, we focus on $y = exp(\beta_1 + \beta_2 \ln(x))$. Then, the first-order differentiation (slope) is

$$\frac{dy}{dx} = m = \frac{\beta_2 \exp(\beta_1 + \beta_2 \ln(x))}{x}.$$

The second-order differentiation is

$$\begin{aligned} \frac{d^2y}{dx^2} &= \frac{\beta_2^2 \exp(\beta_1 + \beta_2 \ln(x)) - \beta_2 \exp(\beta_1 + \beta_2 \ln(x))}{x^2} \\ &= \frac{\beta_2 \exp(\beta_1 + \beta_2 \ln(x))(\beta_2 - 1)}{x^2}. \end{aligned}$$

1.1 If $\beta_2 > 0$, then y is an increasing function of x at an an increasing rate for $\beta_2 > 1$ but at a decreasing rate for $0 < \beta_2 < 1$.

1.2 If $\beta_2 < 0$, there is an inverse relationship between y and x.

2. For a log-log model, focusing on $y = exp(\beta_1 + \beta_2 \ln(x))$:
    - The slope is
    $$m = \frac{dy}{dx} = \frac{\beta_2 y}{x}.$$
    - The elasticity is β_2,
    $$\varepsilon = m\frac{x}{y} = \frac{\beta_2 y}{x}\frac{x}{y} = \beta_2.$$

3. Similar to the log-linear model, a natural choice for prediction (or the fitted value) is

$$\hat{y}_n = \exp(b_1 + b_2 \log(x)).$$

A better alternative is

$$\hat{y}_c = \exp(b_1 + b_2 \log(x) + \hat{\sigma}^2/2).$$

4. Example: calculating log-log demand for chicken. The coefficient on *p* indicates that an increase in price by 1% changes the quantity demanded by -1.121%. The generalized R-squared is 0.88. We can also draw the fitted values of the log-log equation by using the corrected formula of the fitted value.

```
> data("newbroiler", package="PoEdata")
> mod6 <- lm(log(q)~log(p), data=newbroiler)
> b1 <- coef(mod6)[[1]]
> b2 <- coef(mod6)[[2]]
> smod6 <- summary(mod6)
> tbl <- data.frame(xtable(smod6))
> kable(tbl, caption="The log-log poultry regression equatio
```

```
|            |  Estimate| Std..Error|   t.value| Pr...t..|
|:-----------|---------:|----------:|---------:|--------:|
|(Intercept) |  3.716944|  0.0223594| 166.23619|        0|
|log(p)      | -1.121358|  0.0487564| -22.99918|         0|

> yhatc <- exp(b1+b2*log(newbroiler$p)+sighat2/2)
> rgsq <- cor(newbroiler$q, yhatc)^2
> rqsq
> rgsq
[1] 0.8817758
```