Lab 9: Inference and Checking Assumptions in Simple Linear Regression

Objectives for this Lab

- Test $H_0 : \beta_0 = 0$.
- Estimate β_1 and test $H_0 : \beta_1 = 0$.
- Produce a normal Q-Q plot to check the normality assumption.

1. As usual, start up RStudio and open Lab8.R. Load the Sleuth3 and ggplot2 R packages.

```
> library(Sleuth3)
> library(ggplot2)
```

2. As in item 3 of Lab 8, save the lm object from a linear regression of Distance on Velocity.

```
> case0701_lm <- lm(Distance~Velocity, data=case0701)
```

3. Generic notation for the simple linear regression model is

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

where Y represents the response variable and X represents the predictor variable. β_0 is the intercept parameter, and β_1 is the slope parameter. Slope and intercept will have particular meanings in a given study.

There are four common inferences associated with simple linear regression:

- Estimating the mean of Y for a given X.
- Predicting a new Y for a given X.
- Estimating β_0, or testing $H_0 : \beta_0 = 0$.
- Estimating β_1, or testing $H_0 : \beta_1 = 0$.

We did the first of these in item 10 of Lab 8 when we estimated the population mean distance for velocities of -200 and 600 km/sec. We did the second inference in lecture (see page 17 of Outline 7) when we predicted the pH of a steer carcass 4 hours after slaughter. We estimated β_0 in lecture when we wrote a 99% confidence interval for β_0 on page 10 of Outline 7. In this lab, we'll test $H_0 : \beta_0 = 0$ and $H_0 : \beta_1 = 0$, and we'll write a confidence interval for β_1.

4. Testing $H_0 : \beta_0 = 0$. The intercept parameter β_0 is the mean Y when $X = 0$. In the context of the nebula study, β_0 is the population mean distance when the recession velocity is 0. According to the theory diagrammed in Display 7.2, β_0 should be equal to 0.

   (a) We will perform a t-test to test the null hypothesis $H_0 : \beta_0 = 0$. Recall that the general form of a t-statistic is

   $$\frac{\text{Point estimate} - \text{Value under} H_0}{\text{SE(Point estimate)}}$$

   (cf. formula at the bottom of page 35 of the *Sleuth*.)

   The point estimate of β_0 and its standard error are given in the coefficients table of the linear regression summary output.

```
> summary(case0701_lm)
```

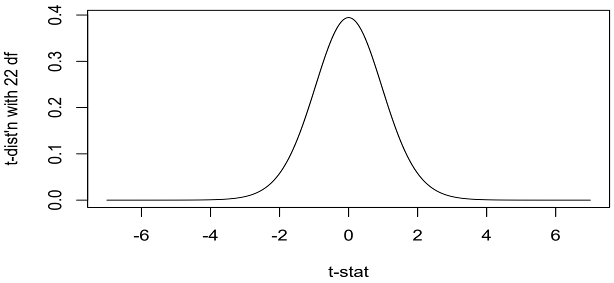
The line labeled `(Intercept)` corresponds to β_0. The output should tell you that the point estimate is $\hat{\beta}_0 = 0.3991704$ and $\text{SE}(\hat{\beta}_0) = 0.1186662$. The t-statistic for our test is therefore the ratio

```
> 0.3991704/0.1186662
```

You should get 3.363809. Notice that this is the value in the `(Intercept)` row and `t value` column of the coefficients table. The last item in the row is the two-sided p-value for our test. Confirm this by calculating the two-sided p-value. The function `pt()` gives the area to the left of its first argument, so subtract from 1 and multiply by 2.

```
> 2*(1-pt(3.363809,22))
```

A picture always helps:



Locate our calculated t-statistic and its negative on the horizontal axis. The p-value of the two-sided hypothesis test is the area to the right of the t-statistic plus the area to the left of its negative.

The degrees of freedom to use in the `pt()` function are always those associated with our estimate of σ. As with one-way ANOVA, it's the residual degrees of freedom: n minus the number of mean parameters. Here we have $n = 24$ and two parameters (β_0 and β_1, which, when combined with a value of X give the mean of the distribution of Y's associated with that X), so the residual degrees of freedom are 22. Find where this is stated in the `summary()` output.

The p-value is small. There is strong evidence that the population mean distance of nebulae with zero recession velocity is not zero.

(b) As we did on page 10 of Outline 7, we can use the point estimate of β_0 and its standard error to write a confidence interval, using the usual format:

$$\text{Point estimate} \pm t_{\text{df}}(1-\alpha/2)\text{SE(Point estimate)}$$

The t quantile for a 95% confidence interval is

```
> qt(0.975, 22)
```

and the limits of the confidence interval are

```
> 0.3991704 - qt(0.975, 22)*0.1186662
> 0.3991704 + qt(0.975, 22)*0.1186662
```

The statistical conclusion for a confidence interval for intercept parameter β_0 states that we have estimated the popultion mean response when the explanatory variable is 0. In particular, here we have estimated that the population mean distance of nebulae with a recession velocity of 0 km/sec is 0.153 to 0.645 megaparsecs (95% confidence interval, simple linear regression).

(c) As we saw in item 5 of Lab 8, we can get confidence intervals for both regression parameters from `confint()` applied to the `lm` object:

```
> confint(case0701_lm)
```

Confirm this gives the same interval for β_0 as calculated "by hand" above.

(d) The default confidence level is 95%. If you want 90% confidence intervals, specify `level=0.9`:

```
> confint(case0701_lm, level=0.9)
```

5. Estimating β_1, or testing $H_0 : \beta_1 = 0$.

The calculations are exactly analogous to those for estimating β_0 or testing $H_0 : \beta_0 = 0$.

As discussed in section 7.4.1 of the *Sleuth*, β_1 from the no-intercept model $\mu\{Y|X\} = \beta_1 X$ can be interpreted as the age of the universe. No-intercept models are extremely uncommon. If you're curious about the no-intercept model, see optional item 5(c) below. We will focus on the much more usual two-parameter simple linear regression model $\mu\{Y|X\} = \beta_0 + \beta_1 X$.

(a) The `summary()` output from item 4(a) gives the point estimate $\widehat{\beta}_1$ and the standard error of the point estimate. The 95% confidence interval for β_1 uses the same multiplier as the 95% confidence interval for β_0 calculated in 4(b).

```
> 0.0013724 - qt(0.975, 22)*0.0002278
> 0.0013724 + qt(0.975, 22)*0.0002278
```

Check that these calculations give you the same confidence interval as `confint()` in item 4(c).

The slope parameter β_1 in a simple linear regression model is the change in population mean response per one-unit increase in the explanatory variable. A statistical conclusion for the above confidence interval is "we have estimated the increase in population mean distance for every one km/sec increase in recession velocity is 0.0008999717 to 0.001844828 megaparsecs (95% confidence interval, simple linear regression)."

Because the confidence limits are so close to 0, you might want to multiply everything by, say, 10,000 and report, "we have estimated the increase in population mean distance for every 10,000 km/sec increase in recession velocity is 8.999717 to 18.44828 megaparsecs (95% confidence interval, simple linear regression)."

(b) The t-statistic to test $H_0 : \beta_1 = 0$ is the ratio of the point estimate to the standard error.

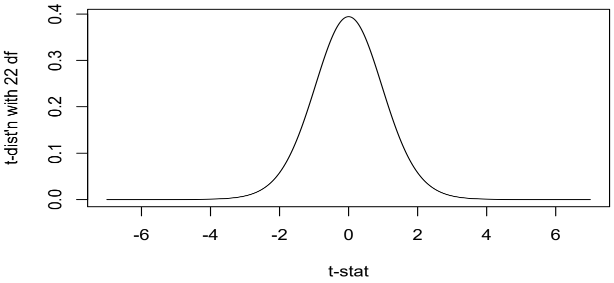
```
> 0.0013724/0.0002278
```

and, as in 4(a) the two-sided p-value is

```
> 2*(1-pt(6.024583,22))
```

Verify that this agrees with the `summary()` output.

The `summary()` shows two-sided p-values. If you are interested in a one-sided test, then you can get the p-value from the output if you're careful. Here, the p-value of the one-sided test with $H_A : \beta_1 > 0$ will be half the two-sided p-value, and the p-value of the one-sided test with $H_A : \beta_1 < 0$ will be one minus half the two-sided p-value. A picture always helps.



Locate the calculated t-statistic on the horizontal axis. The p-value for $H_A : \beta_1 > 0$ will be the area to the right of the t-statistic. The p-value for $H_A : \beta_1 < 0$ will be the area to the left of the t-statistic.

(c) (optional) Fit the no-intercept model and estimate β_1.

```
> case0701_noint <- lm(Distance~Velocity-1, data=case0701)
> summary(case0701_noint)
```

The -1 in the formula to `lm()` specifies "no intercept." Look at the one-line coefficients table in the summary output. You should see that $\hat{\beta}_1 = 0.0019214$ and $\text{SE}(\hat{\beta}_1) = 0.0001913$. Note that these values are different than $\hat{\beta}_1$ and $\text{SE}(\hat{\beta}_1)$ from the output we used in item 4(a). Fixing the intercept at $\beta_0 = 0$ changes the estimated slope.

The lower and upper bounds of a 95% confidence interval are

```
> 0.0019214 - qt(.975, 23)*0.0001913
> 0.0019214 + qt(.975, 23)*0.0001913
```

The degrees of freedom are 23 because the no-intercept model has only one parameter. You should get approximately $(0.0015256, 0.0023171)$ for the confidence interval, which you can verify with

```
> confint(case0701_noint)
```

6. We have used a plot of residuals vs. fitted values to assess the equal variance and normality assumptions. This plot is a good tool for checking equal variance, but not ideal for checking normality. A *Normal Probability Plot* of the residuals is specifically designed for checking the normality assumption. Another name for a normal probability plot is "normal quantile-quantile plot" or "normal Q-Q plot."

In item 11(b) of Lab 8, we produced a plot of the residuals vs. fitted values. Similarly, we can produce a normal Q-Q plot of these residuals.

```
> plot(case0701_lm, which=2)
```

The horizontal axis is labeled "Theoretical Quantiles," and the vertical axis is labeled "Sample Quantiles." The sample quantiles are just the (standardized) residuals, ordered from small to large. The theoretical quantiles are the expected ordered values from a standard normal sample. If the normality assumption is met, the residuals should be approximately normal, so the sample and theoretical quantiles should be similar, and so the normal Q-Q plot should be approximately linear.

This plot looks pretty good, even though the points at either end are not on the line. Refer to Display 8.13 for some possible patterns in normal Q-Q plots. Note, however, that the *Sleuth* plots the theoretical quantiles on the vertical axis and the sample quantiles on the horizontal axis, whereas R does the reverse.