STATS 412

Twenty-second and Twenty-third Class Note

In Son Zeng

11 December, 2018

My Office Hour:

The office hours this week are on **16:30 - 18:00 Tuesday** and **13:30 - 15:00 Friday**, at **USB 2165**. You may check the campus map to get to my office. I am prepared for your questions, so please feel free to come to my office hours.

Exam 3 Updates:

I will be updating the revision for the final exam day by day, so check closely to the update for the final exam.

• The revision contents are various: concepts, test convention, sample questions and solutions. Hope you find the materials helpful for performing better in the final exam.

• If our class can improve the handwriting for the rest of the homeworks, I can have more time to write the revision day by day to improve the general quality. Also, when I have more time, I can be more efficiently respond questions on Piazza and provide better designed questions and solutions (with video).

Homework Grading Policy:

Please include the final answer for each homework question. If the final answer is not included, you will risk 0.5 points for each missing part.

Homework 12 Reminder:

Implementation of R codes: Remember for simple linear regression that we have discussed in class so far, the model is $y = \beta_0 + \beta_1 x + \epsilon$ and the predicted value for observation $i = 1, 2, ..., n$ is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. We now implement R if you are going to use R as an alternative to JMP.

```r
# Set up data
# Independent variable
x = c(26.6, 26.0, 27.4, 21.7, 14.9, 11.3, 15.0, 8.7, 8.2)
# Dependent variable
y = c(1.58, 1.45, 1.13, 0.96, 0.99, 1.05, 0.82, 0.68, 0.56)

# Step 1: Build linear model

linearmodel <- lm(y ~ x)  # build linear regression model on full data
summary_mod = summary(linearmodel); summary_mod
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.23691 -0.12513 -0.02289  0.13280  0.25479
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.393960   0.171536   2.297  0.05526 .
## x           0.035509   0.008927   3.978  0.00534 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1967 on 7 degrees of freedom
## Multiple R-squared:  0.6933, Adjusted R-squared:  0.6494
## F-statistic: 15.82 on 1 and 7 DF,  p-value: 0.00534
```

• The summary here gives the least-square line for prediction is $\hat{y} = 0.39396 + 0.035509 \cdot x$, one unit increase of x is associated with the increase of y by 0.035509. Then, we look at the **Std. Error** part, the summary gives that $SE(\hat{\beta}_0) = 0.171536$ and $SE(\hat{\beta}_1) = 0.008927$ and degree of freedom $n - 2 = 7$. In **Multiple R-squared** part, the summary table gives that the $R^2 = 0.6933$, indicating that approximately 69.33 % of the observed variation in the model can be explained by the simple linear regression relationship between x and y.

```r
# Step 2: Confidence interval for intercept and slope
conf = confint(linearmodel); round(conf, 4)
```

```
##               2.5 % 97.5 %
## (Intercept) -0.0117 0.7996
## x            0.0144 0.0566
```

```r
# For intercept, the first value is estimate, the second value is SE
c(summary_mod$coefficients[1,1],summary_mod$coefficients[1,2])
```

```
## [1] 0.3939595 0.1715359
```

```r
conf_intercept = c(summary_mod$coefficients[1,1] + qt(0.025, 7) * summary_mod$coefficients[1,2],
                   summary_mod$coefficients[1,1] + qt(0.975, 7) * summary_mod$coefficients[1,2])
round(conf_intercept, 4) # This will give the same result as the confint() function
```

```
## [1] -0.0117  0.7996
```

```r
# For slope, the first value is estimate, the second value is SE
c(summary_mod$coefficients[2,1],summary_mod$coefficients[2,2])
```

```
## [1] 0.035509163 0.008927328
```

```r
conf_intercept = c(summary_mod$coefficients[2,1] + qt(0.025, 7) * summary_mod$coefficients[2,2],
                   summary_mod$coefficients[2,1] + qt(0.975, 7) * summary_mod$coefficients[2,2])
round(conf_intercept, 4) # This will give the same result as the confint() function
```

```
## [1] 0.0144 0.0566
```

```r
# Step 3: Predicted interval for each observation
predval = predict(linearmodel); round(predval, 4)
```

```
##      1      2      3      4      5      6      7      8      9
## 1.3385 1.3172 1.3669 1.1645 0.9230 0.7952 0.9266 0.7029 0.6851
```

• From the confint(), we are approximately 95% confident that the intercept term is between -0.0117 and 0.7996. Also, we are approximately 95% confident that the slope term is between 0.0144 and 0.0566. Particularly, we can extract the values from the summary table, with the aids of qt() function for t-score, to compute the confidence interval by hand.

• From the predict(), we can obtain the predicted value for the 9 observations after performing simple linear regression.

```r
# Step 4: Inference on the mean response, and Prediction interval for future observation

# a) Let us assume we want to know the inference when x = 30
newdata = data.frame(x=30)
# Confidence interval for mean response
conf30 = predict(linearmodel, newdata, interval="confidence"); round(conf30,4)
```

```
##      fit    lwr    upr
## 1 1.4592 1.1578 1.7606
```

```r
# Prediction interval for one future observation
pred30 = predict(linearmodel, newdata, interval="predict"); round(pred30, 4)
```

```
##      fit   lwr    upr
## 1 1.4592 0.905 2.0135
```

```r
newdata1 = data.frame(x=7)
conf7 = predict(linearmodel, newdata1, interval="confidence");round(conf7, 4)
```

```
##      fit    lwr    upr
## 1 0.6425 0.3676 0.9175
```

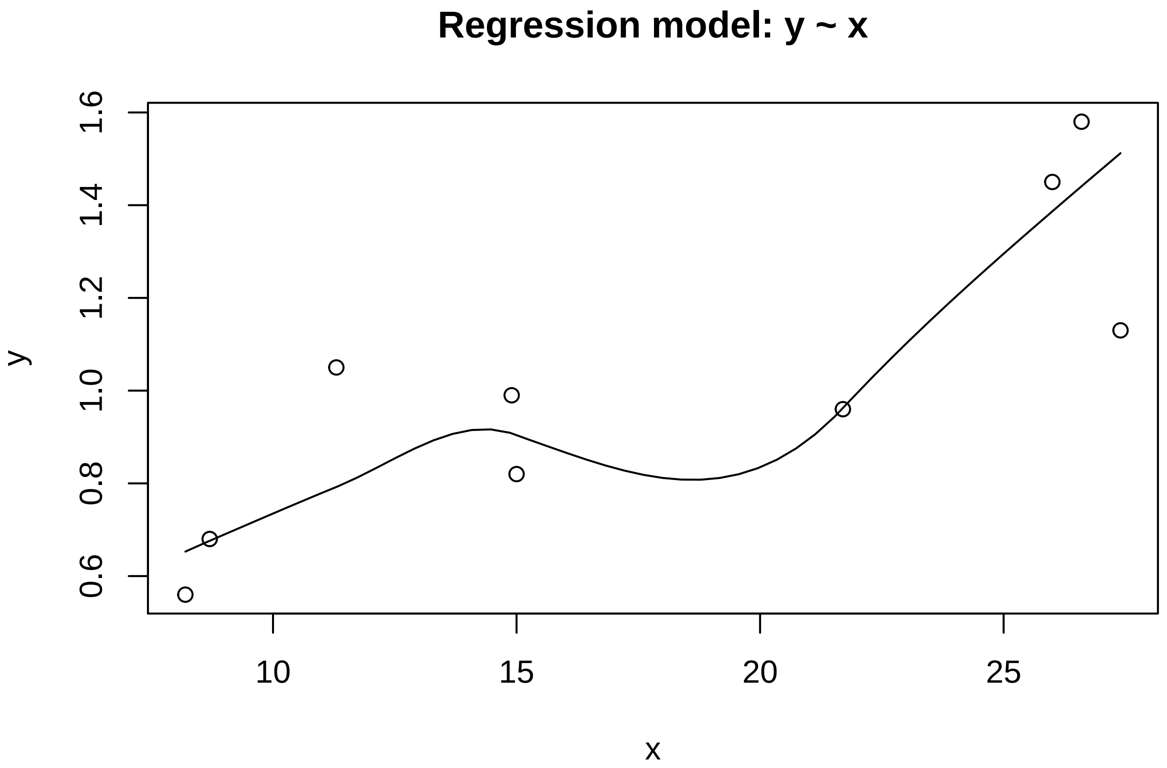
```r
pred7 = predict(linearmodel, newdata1, interval="predict"); round(pred7, 4)
```

```
##      fit    lwr    upr
## 1 0.6425 0.1022 1.1828
```

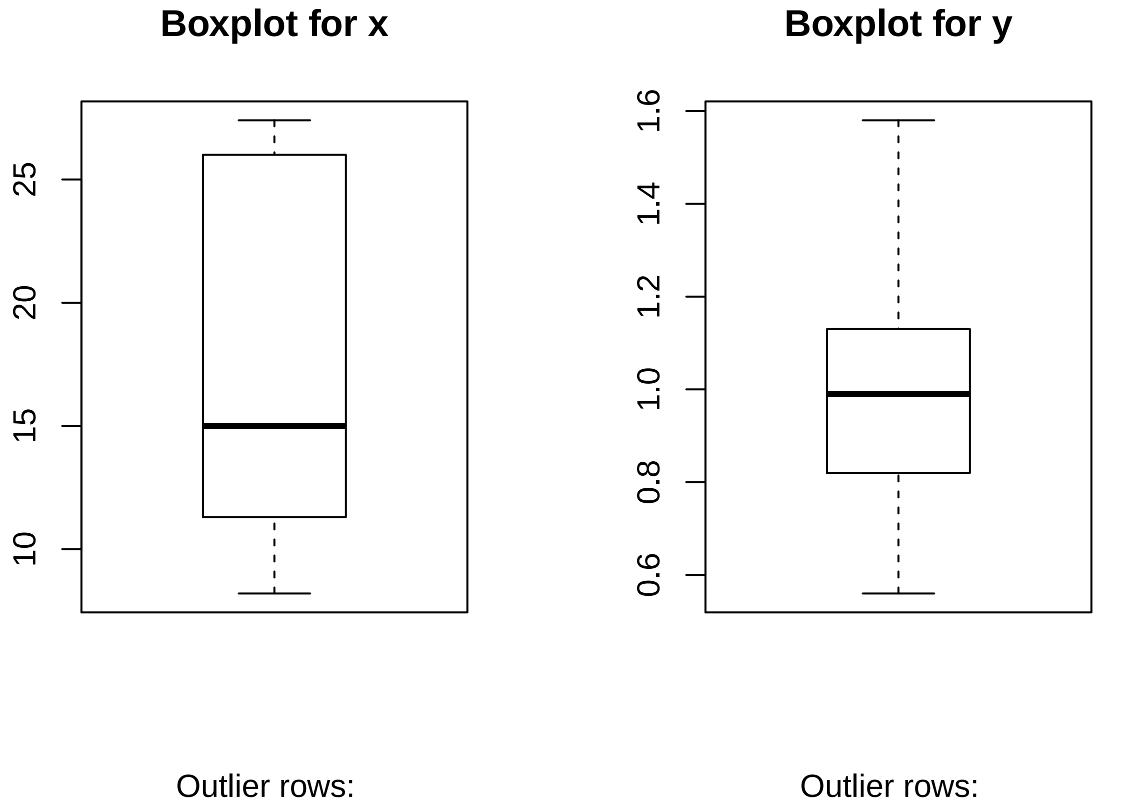
• By adding new data and performing the predict() function, we obtain that when $x = 30$, the confidence interval for mean response is (1.1578, 1.7606), and the prediction interval for a future observation is (0.905, 2.0135). By similar method, we obtain that when $x = 7$, the confidence interval for mean response is (0.3676, 0.9175), and the prediction interval for a future observation is (0.1022, 1.1828).

• We see that the prediction interval for future observation is wider than the confidence interval for mean response for both cases. Do you remember the formula to derive these two intervals? See the note below before attempting the homework. Good Luck!

```r
# Step 5: Plot Diagnostics
# Scatterplot
scatter.smooth(x, y, main="Regression model: y ~ x")  # scatterplot
```

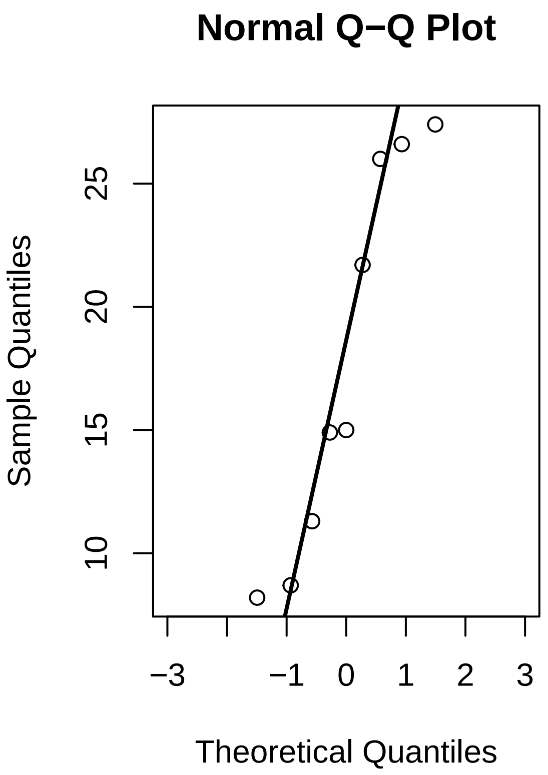


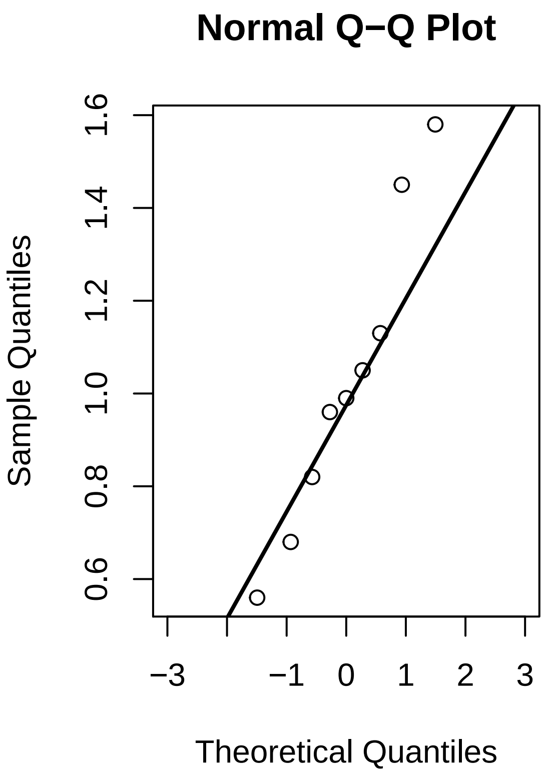
```
# Boxplot (No outlier in this case)
par(mfrow=c(1, 2))  # divide graph area in 2 columns
boxplot(x, main="Boxplot for x", sub=paste("Outlier rows: ", boxplot.stats(x)$out))  # Boxplot for x
boxplot(y, main="Boxplot for y", sub=paste("Outlier rows: ", boxplot.stats(y)$out))  # Boxplot for y
```



```
# Q-Q plot
par(mfrow=c(1, 2))
qqnorm(x, xlim = c(-3,3))
qqline(x, lwd=2)

qqnorm(y, xlim = c(-3,3))
qqline(y, lwd=2)
```


Normal Q−Q Plot
Sample Quantiles
10
15
20
25
−3
−1
0
1
2
3
Theoretical Quantiles


Normal Q−Q Plot
Sample Quantiles
0.6
0.8
1.0
1.2
1.4
1.6
−3
−1
0
1
2
3
Theoretical Quantiles

Key Concepts During Lecture 22 and 23:

Correlation, Coefficient of Determination: See class note 21.

JMP outputs in Lecture note and ANOVA: See class note 21.

Simple Linear Model: The simple linear model (true model for least-squares regression) is given by:

$$\boxed{y_i = \beta_0 + \beta_1 x_i + \epsilon_i \ \ or \ \ Y = \beta_0 + \beta_1 X + \epsilon} \tag{1}$$

, where y_i is the observed value of the response variable, β_0 is the least-squares y-intercept, β_1 is the least-squares slope for the model. The ϵ_i is the error term with the ith measurement.

Normal Assumption for Linear Model:

• 1) We need the observations are independent.

• 2) We need the errors $\epsilon_1, \epsilon_2, ..., \epsilon_n$ are random and independent. In particular, the magnitude of any error term does not influence the value of the other error terms. Again, if the error terms appear to be monotone increasing or decreasing, U-shape or inverted-U-shape, the assumption for linear models is violated.

• 3) We need the error terms to be normally distributed, with mean 0 and with same variance (homoscedasticity). Notationally speaking, we need $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$

If the normal assumption is fulfilled, we can further have the response variable y_i follows normal distribution. Notationally speaking, $y_i \overset{\text{i.i.d.}}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$, for $i = 1, 2, ..., n$.

Intercept and Slope for regression: I will provide only the formulas of $\hat{\beta}_0$, $\hat{\beta}_1$, $SE(\hat{\beta}_0)$ and $SE(\hat{\beta}_1)$ here, while skipping the proofs for the derivation for simplicity. To see the proofs, I will upload side notes for further study after the final examination. If you are interested in regression, take STATS 500 or other related courses.

$$\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}} \tag{2}$$

$$\boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}} \tag{3}$$

$$\boxed{SE(\hat{\beta}_1) = \sqrt{Var(\frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2})} = \frac{\sigma}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}} \tag{4}$$

$$\boxed{SE(\hat{\beta}_0) = \sqrt{Var(\bar{y} - \hat{\beta}_1 \bar{x})} = \sigma \cdot \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \tag{5}$$

$$\boxed{\hat{\sigma}^2 = s^2 = MSE = (RMSE)^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2}} \tag{6}$$

Hypothesis Tests and Confidence Intervals for the Intercept:

• For Intercept, our hypothesis is $H_0 : \beta_0 = 0 \ \ vs \ \ H_1 : \beta_0 \neq 0$.

• We perform standardization for $\beta_0 \to \frac{Observed - Mean}{SD} = \frac{\hat{\beta}_0 - \beta_0}{SE(\hat{\beta}_0)}$. This value should follow the t-distribution with $n-2$ degree of freedom, because the degree of freedom is computed by (sample size - number of coefficients estimated β_0, β_1), which is $n-2$.

• If the p-value is smaller than the significance level α, we reject the null hypothesis and we have sufficient evidence to conclude that the intercept is significantly different from 0.

• The $100(1-\alpha)\%$ confidence interval (you know how to derive the CUB and CLB already) of intercept, we have: $\hat{\beta}_0 \pm t_{n-2, \frac{\alpha}{2}} \cdot SE(\hat{\beta}_0)$. Then, we are **approximately** $100(1-\alpha)\%$ confident that the intercept value falls (between/below/above) . . . .

Hypothesis Tests and Confidence Intervals for the Slope:

• For Slope, our hypothesis is $H_0 : \beta_1 = 0 \ \ vs \ \ H_1 : \beta_1 \neq 0$.

• We perform standardization for $\beta_1 \to \frac{Observed - Mean}{SD} = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)}$. This value should follow the t-distribution with $n-2$ degree of freedom, because the degree of freedom is computed by (sample size - number of coefficients estimated β_0, β_1), which is $n-2$.

• If the p-value is smaller than the significance level α, we reject the null hypothesis and we have sufficient evidence to conclude that the slope (regression effect) is significantly different from 0. In this case, the predictor is significant in our linear model.

• On the other hand, if the p-value is greater than the significance level α, we fail to reject the null hypothesis and we do have sufficient evidence to conclude that the slope (regression effect) is significantly different from 0. In this case, the predictor can be dropped in our linear model.

• The $100(1-\alpha)\%$ confidence interval (you know how to derive the CUB and CLB already) of slope, we have: $\hat{\beta}_1 \pm t_{n-2,\frac{\alpha}{2}} \cdot SE(\hat{\beta}_1)$. Then, we are **approximately** $100(1-\alpha)\%$ confident that the intercept value falls (between/below/above) . . . .

Inference for Mean Response: Recap: $Y|x_i \sim N(\beta_0 + \beta_1 x_1, \sigma^2)$, $Y|x^* \sim N(\beta_0 + \beta_1 x^*, \sigma^2)$.

• For $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^* \to E[\hat{Y}] = \beta_0 + \beta_1 x^*$, $Var(\hat{Y}) = \sigma^2 \Big[\frac{1}{n} + \frac{(x^*-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\Big]$

• Now for the **estimated standard deviation** of $\hat{Y}$, we have

$$\boxed{s_{\hat{Y}} = s \cdot \sqrt{\frac{1}{n} + \frac{(x^*-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}} \quad , \quad s = RMSE = \sqrt{MSE} = \sqrt{\frac{\sum_{i=1}^n (y_i-\hat{y}_i)^2}{n-2}} = \sqrt{\frac{(1-r^2)\sum_{i=1}^n (y_i-\bar{y})^2}{n-2}}} \tag{7}$$

• Therefore, for hypothesis testing, we use the test statistics $\frac{\hat{Y}-(\beta_0+\beta_1 x^*)}{s_{\hat{Y}}} \sim t_{n-2}$, remember we have degree of freedom $n-2$ because we are estimating intercept β_0 and slope β_1. In JMP output, we can see the degree of freedom $n-2$ by looking at the **Df for Error** in Analysis of Variance.

• To construct $100(1-\alpha)\%$ confidence interval for mean response, we use

$$\boxed{(\hat{\beta}_0 + \hat{\beta}_1 x^*) \pm t_{\frac{\alpha}{2},n-2} \cdot s_{\hat{Y}} = (\hat{\beta}_0 + \hat{\beta}_1 x^*) \pm t_{\frac{\alpha}{2},n-2} \cdot \sqrt{MSE \cdot [\frac{1}{n} + \frac{(x^*-\bar{x})^2}{Sum\ of\ Square}]}} \tag{8}$$

• The MSE in JMP can be found in Analysis of Variance → Mean Square → Error row, while $\sum_{i=1}^n (x_i-\bar{x})^2$ can be cound in **Corrected SS** in Summary Statistics.

Inference for Future Observations (Prediction Interval): Since the error in prediction is calculated by $y - \hat{y} = (\beta_0 + \beta_1 x^* + \epsilon) - (\hat{\beta}_0 + \hat{\beta}_1 x^*)$, the prediction interval for a future observation has more variability which makes it wider than a confidence interval for mean response.

• Since the future observation $Y_{new} = \beta_0 + \beta_1 x^* + \epsilon$ is assumed to be independent of the observed values, we have $cov(Y_{new}, \hat{Y}) = 0$. Therefore, the error of prediction has variance:

$$\boxed{Var(Y_{new} - \hat{Y}) = Var(Y_{new}) - 2cov(Y_{new}, \hat{Y}) + Var(\hat{Y}) = Var(\beta_0 + \beta_1 x^* + \epsilon) + Var(\hat{Y})} \tag{9}$$

$$\boxed{= Var(\epsilon) + \sigma^2 \cdot [\frac{1}{n} + \frac{(x^*-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}] = \sigma^2 + \sigma^2 \cdot [\frac{1}{n} + \frac{(x^*-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}] = \sigma^2 \cdot [1 + \frac{1}{n} + \frac{(x^*-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}]} \tag{10}$$

• Therefore, with the similar logic of mean response, the **estimated standard deviation** of Y_{new}, we have

$$\boxed{s_{Y_{new}} = \sqrt{s_{\hat{Y}}^2 + s^2} = s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x^*-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}} = (RMSE) \cdot \sqrt{1 + \frac{1}{n} + \frac{(x^*-\bar{x})^2}{Sum\ of\ Square}}} \tag{11}$$

• Therefore, for hypothesis testing, we use the test statistics $\frac{Y_{new}-(\hat{\beta}_0+\hat{\beta}_1 x^*)}{s_{Y_{new}}} \sim t_{n-2}$, remember we have degree of freedom $n-2$ because we are estimating intercept β_0 and slope β_1. In JMP output, we can see the degree of freedom $n-2$ by looking at the **Df for Error** in Analysis of Variance.

• To construct $100(1-\alpha)\%$ confidence interval for mean response, we use

$$\boxed{(\hat{\beta}_0 + \hat{\beta}_1 x^*) \pm t_{\frac{\alpha}{2},n-2} \cdot s_{Y_{new}} = (\hat{\beta}_0 + \hat{\beta}_1 x^*) \pm t_{\frac{\alpha}{2},n-2} \cdot \sqrt{MSE \cdot [1 + \frac{1}{n} + \frac{(x^*-\bar{x})^2}{Sum\ of\ Square}]}} \tag{12}$$

Last Comment:

Please inform me to fix the typos and grammatical mistakes if they exist. It is a great practice of writing and I appreciate your help!