Ve406 Lecture 7

Jing Liu

UM-SJTU Joint Institute

June 4, 2018

A request from an insurance Company

- An insurance company wants to predict the damage (in $) to a home in a particular area if a fire occurs. The damage, and distance (in miles) from the fire station were recorded for 15 house fires in the area of interest.

```
> fire.df = read.table("~/Desktop/fire.txt",
+                      header = TRUE)
```

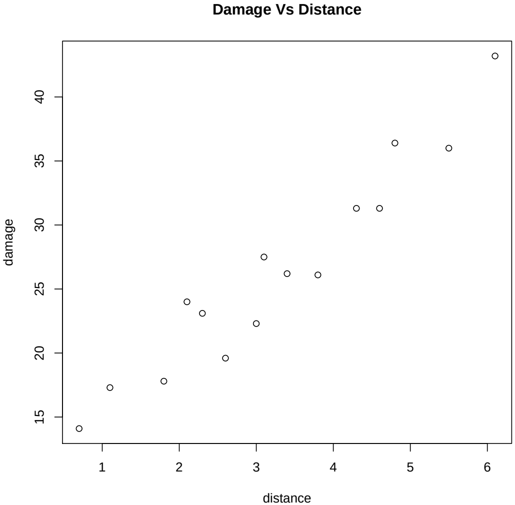
```
> str(fire.df)
```

```
'data.frame':   15 obs. of  2 variables:
 $ distance: num  3.4 2.6 1.8 4.3 4.6 ...
 $ damage  : num  26.2 19.6 17.8 31.3 31.3 ...
```

- We are required to predict the damage for house fires that are 1 and 4 miles from the fire station.

- Visualisation

```
> with(fire.df, plot(distance, damage,
+         main = "Damage Vs Distance"))
```

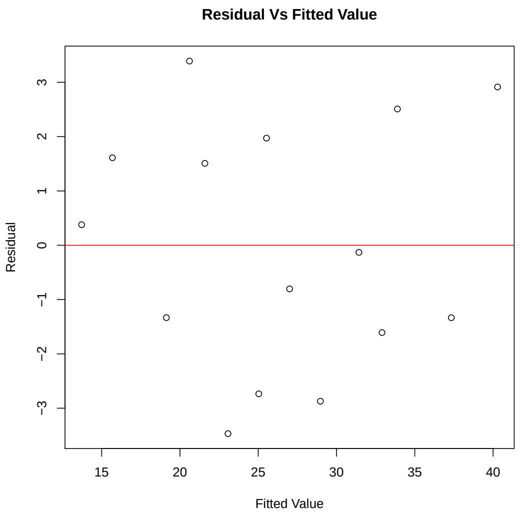


- Running a simple linear regression model

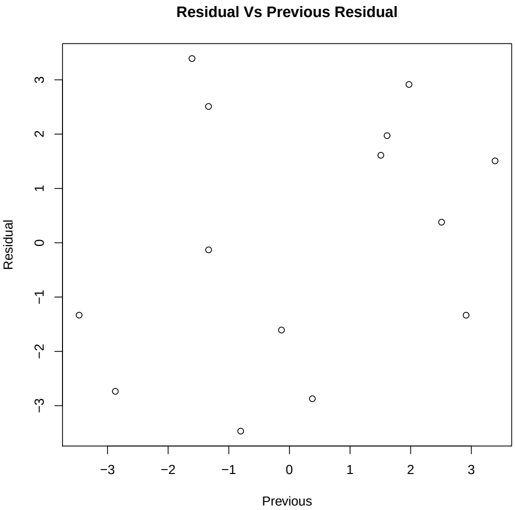
```
> fire.LM = lm(damage~distance, data = fire.df)
```

- Checking assumptions by looking at residuals

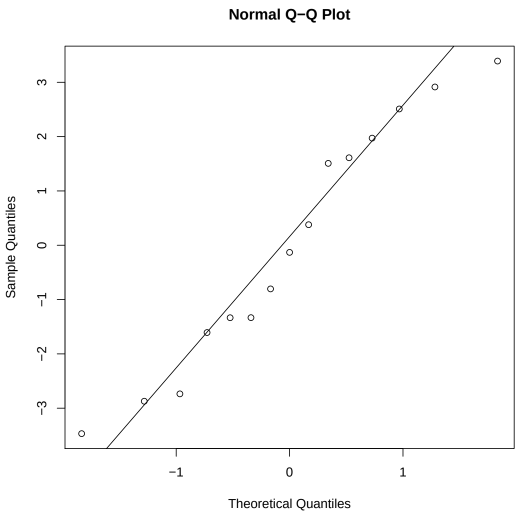
```
> fit = fire.LM$fitted.values
> res = fire.LM$residuals

> # Residual Plot
> plot(fit, res, main = "Residual Vs Fitted Value",
+       xlab = "Fitted Value", ylab = "Residual")
> abline(h = 0, col = "red")

> # Correlation Plot
> plot(res[-nrow(fire.df)], res[-1],
+       xlab = "Previous", ylab = "Residual",
+       main = "Residual Vs Previous Residual")

> # QQ plot
> qqnorm(res); qqline(res)
```



- There is no indication of nonlinearity, correlation or non-constant variance.



- There is no indication of autocorrelation.



- There is no indication of non-normality.

- Since the data size n is really small, we have to be careful with normality

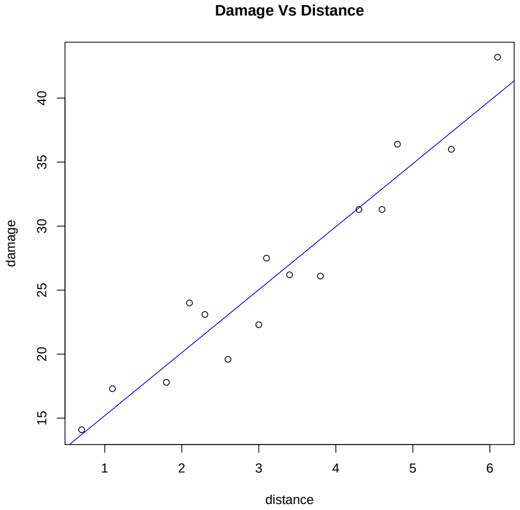
```
> # Shapiro-Wilk
> shapiro.test(res)
```

```

        Shapiro-Wilk normality test

data:  res
W = 0.94671, p-value = 0.4743
```

- Notice formal tests on normality are rather strict, use them if n is small.
- Since all the assumptions seem to be satisfied, we can now do inference.

```
> with(fire.df, plot(distance, damage,
+                    main = "Damage Vs Distance"))
> abline(fire.LM, col = "blue")
```

- So not only we can trust the line, but we do other meaningful things...



```
> (fire.sm = summary(fire.LM))
```

```
Call:
lm(formula = damage ~ distance, data = fire.df)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4682 -1.4705 -0.1311  1.7915  3.3915

Coefficients:
            Estimate Std. Error t value Pr(>t)
(Intercept)  10.2779     1.4203   7.237 6.59e-06 ***
distance      4.9193     0.3927  12.525 1.25e-08 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 2.316 on 13 degrees of freedom
Multiple R-squared:  0.9235,    Adjusted R-squared:  0.9176
F-statistic: 156.9 on 1 and 13 DF,  p-value: 1.248e-08
```

```
> names(fire.sm)
```

```
 [1] "call"          "terms"         "residuals"     "coefficients"  "aliased"
 [6] "sigma"         "df"            "r.squared"     "adj.r.squared" "fstatistic"
[11] "cov.unscaled"
```

```
> (coeff.m = fire.sm$coefficients)
```

```
             Estimate Std. Error   t value     Pr(>t)
(Intercept) 10.277929  1.4202778  7.236562 6.585564e-06
distance     4.919331  0.3927477 12.525421 1.247800e-08
```

```
> class(coeff.m)
```

```
[1] "matrix"
```

- Confidence interval for the parameters, e.g. for $\hat{\beta}_1$

```
> b1_hat = coeff.mat[2,1]
>
> b1_se = coeff.mat[2,2]
>
> (b1_ci = b1_hat + c(-1, 1) * b1_se *
+    qt(0.975, fire.LM$df.residual))
```

```
[1] 4.070851 5.767811
```

- In practice, we use the following to obtain the confidence interval

```
> confint(fire.LM, "(Intercept)", level = 0.95)
```

```
                2.5 %   97.5 %
(Intercept) 7.209605 13.34625
```

```
> confint(fire.LM, "distance", level = 0.95)
```

```
             2.5 %   97.5 %
distance 4.070851 5.767811
```

- The following gives the predictions according to our model.

```
> pred.df = data.frame(distance = c(1,4))
> predict(fire.LM, pred.df)
```

```
       1        2
15.19726 29.95525
```

Q: How can we construct a confidence interval for our prediction?

$$(\hat{y}_{n+1} - c, \hat{y}_{n+1} + c)$$

Q: How can we find the constant c for a given significant level?

Q: Why R has two types of confidence interval for prediction?

```r
> pred.df = data.frame(distance = c(1,4))
> predict(fire.LM, pred.df, interval = "predict")
```

```
       fit      lwr      upr
1 15.19726  9.67879 20.71573
2 29.95525 24.75100 35.15951
```

```r
> predict(fire.LM, pred.df, interval = "confidence")
```

```
       fit      lwr      upr
1 15.19726 12.87092 17.52360
2 29.95525 28.52604 31.38446
```

- Given a dataset, $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$,

  The variance of $\hat{y}$ can be derived based on the variance of $\hat{\beta}_1$

$$\begin{aligned}
\mathrm{Var}\left[\hat{Y} \mid X_1, X_2, \ldots, X_n\right] &= \mathrm{Var}\left[\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x \mid X_1, X_2, \ldots, X_n\right] \\
&= \frac{\sigma^2}{n-1}\left(\frac{n-1}{n} + \frac{(x-\bar{x})^2}{s_x^2}\right)
\end{aligned}$$

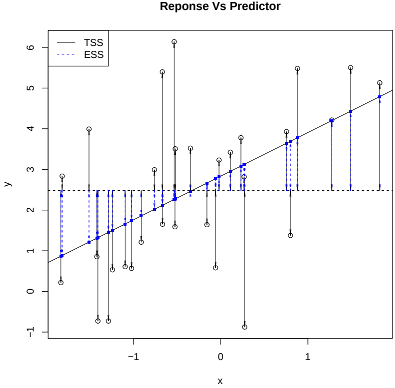
  and the sampling distribution of $\hat{y}$ is normal given σ^2.

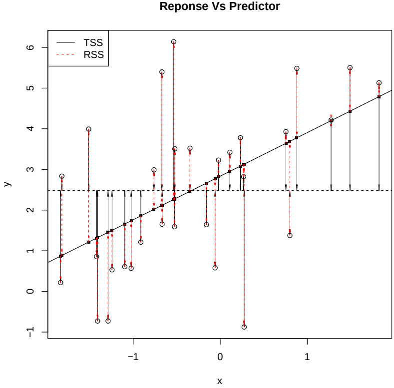
  The variance of Y_{n+1} based on our model consists two parts

$$\begin{aligned}
\mathrm{Var}\left[Y_{n+1} \mid X_1, X_2, \ldots, X_{n+1}\right] &= \mathrm{Var}\left[\hat{Y} \mid X_1, X_2, \ldots, X_{n+1}\right] \\
&\quad + \mathrm{Var}\left[\varepsilon_{n+1} \mid X_1, X_2, \ldots, X_{n+1}\right] \\
&= \frac{\sigma^2}{n-1}\left(n - \frac{1}{n} + \frac{(x-\bar{x})^2}{s_x^2}\right)
\end{aligned}$$

- This two classes of variability reflects the fact the total variability in the data

$$\text{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\text{ESS}} + \underbrace{\sum_{i=1}^{n}\hat{e}_i^2}_{\text{RSS}}$$

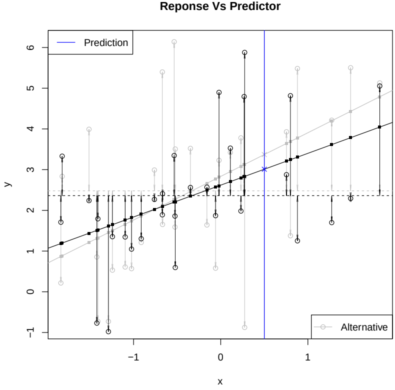


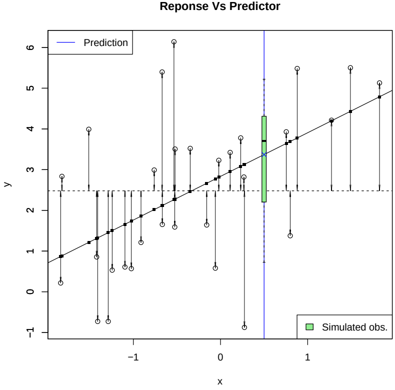


- The coefficient of determination

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

- The variability in the value of Y_{n+1} has two layers given $\{x_1, x_2, \cdots x_n\}$.





Reporting

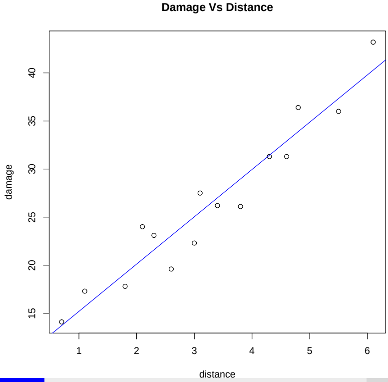
- When using regression in a project, you report should consists of two sections

1. Technical Notes
    - Exploratory Analysis
    - Model Specification
    - Checking Assumptions
    - Statistical Inference
2. Executive Summary
    - Overall quality of the model
    - Explaining the relationship between the response and the predictors
    - Making predictions
    - Addressing specific questions the project is about

Technical Notes

- Exploratory Analysis:

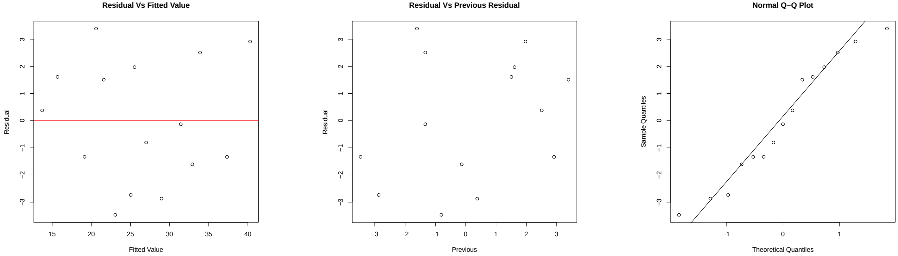
  The scatter plot of fire damage versus distance shows a strong, increasing, linear relationship. The greater the distance from the fire station, the greater the mean amount of damage that is caused by the house fire.



Technical Notes

- Checking Assumptions:

  The observations appear to be independent. The plot of residuals versus fitted values shows random scatter about 0. The Normal Q-Q plot shows the points lying close to the straight line indicating that the errors could have a normal distribution. The Shapiro-Wilk test provides no evidence against the hypothesis that the errors have a normal distribution (P-value $= 0.4743$).



Technical Notes

- Statistical Inference:

The F-test for regression provides extremely strong evidence against the hypothesis that distance from the fire station is not related to the amount of damage (P-value $= 1.248 \times 10^{-8}$). The Multiple R^2 is 0.9235 indicating that 92% of the variation in damage is explained by the variation in distance from the fire station, so the model should be very accurate for prediction. We have extremely strong evidence against the hypothesis that the intercept is equal to 0 (P-value $= 6.59 \times 10^{-6}$). We have extremely strong evidence against the hypothesis that the slope coefficient associated with distance is equal to 0 (P-value $= 1.25 \times 10^{-8}$).

```
Coefficients:
            Estimate Std. Error t value Pr(>t)
(Intercept)  10.2779     1.4203   7.237 6.59e-06 ***
distance      4.9193     0.3927  12.525 1.25e-08 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 2.316 on 13 degrees of freedom
Multiple R-squared:  0.9235,    Adjusted R-squared:  0.9176
F-statistic: 156.9 on 1 and 13 DF,  p-value: 1.248e-08
```

Executive Summary

Our model explains 92% of the variation in house fire damage and should therefore be a very accurate model for prediction. We have extremely strong evidence that as the distance from the fire station increases, the average amount of damage increases. We estimate that if the house next to the fire station catches fire, the mean fire damage will be between $7,200 and $13,300. We estimate that for each additional mile from the fire station, the mean fire damage increases by between $4,100 and $5,800. Using our model, we predict that if a new fire occurs in a house that is 1 mile from the fire station, the damage will be between $9,700 and $20,700. The mean damage for house fires that are 1 mile from the fire station will be between $12,900 and $17,500. For a house that is 4 miles from the fire station, we predict the damage will be between $24,800 and $35,200. The mean damage for house fires that are 4 miles from the fire station will be between $28,500 and $31,400.

SLR Summary

- The question that simple linear regression model tries to address.

$$Y \qquad \text{and} \qquad X$$

- Assumptions:

1. The conditional mean of the response is linear in terms of β_0, β_1, x_i

$$\mathbb{E}\left[Y_i \mid X_i = x_i\right] = \beta_0 + \beta_1 x_i$$

2. The errors have zero mean and constant variance

$$\mathbb{E}\left[\varepsilon_i \mid X_i\right] = 0 \quad \text{and} \quad \mathrm{Var}\left[\varepsilon_i \mid X_i\right] = \sigma^2 \qquad \text{where} \quad \varepsilon_i = Y_i - \beta_0 - \beta_1 X_i$$

3. The errors are independent of X_i, and of each other.
4. The errors follow the normal distribution of $\mathrm{N}\left(0, \sigma^2\right)$.

- Model description:

$$y_i = \beta_0 + \beta_1 x_i + e_i \implies y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{e}_i \implies y_i = \hat{y}_i + \hat{e}_i$$

- The LSE/MLE for β_0 and β_1 are unbiased and consistent.

$$\hat{\beta}_1 = \frac{c_{xy}}{s_x^2} \qquad \text{and} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- Given a dataset, $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$,

  The variances

$$\begin{aligned} \mathrm{Var}\left[\hat{\beta}_1 \mid X_1, X_2, \ldots, X_n\right] &= \frac{\sigma^2}{(n-1)s_x^2} \\ \mathrm{Var}\left[\hat{\beta}_0 \mid X_1, X_2, \ldots, X_n\right] &= \frac{\sigma^2}{n}\left(1 + \frac{n\bar{x}^2}{(n-1)s_x^2}\right) \end{aligned}$$

  can be derived in terms of σ^2 with assumption

2. The errors have zero mean and constant variance σ^2.

- Given a dataset, $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$,

  The sampling distribution of $\hat{\beta}_1$ or $\hat{\beta}_0$ is normal given σ^2 can be found with

4. The errors follow the normal distribution of $\mathrm{N}\left(0, \sigma^2\right)$.

- Estimator

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \hat{e}_i^2$$

  which is based on adjusting the LSE/MLE, is unbiased and consistent.

- Given a dataset, $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$,

  The sampling distribution of $\hat{\beta}_1$ or $\hat{\beta}_0$ can be found in terms of $\hat{\sigma}^2$.

- Estimator

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x \qquad \text{which is for} \quad \mathbb{E}\left[Y \mid X\right]$$

  is unbiased and consistent.

- Given a dataset, $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$,

  The variance of $\hat{y}$ can be derived based on the variance of $\hat{\beta}_1$

$$\begin{aligned}\mathrm{Var}\left[\hat{Y} \mid X_1, X_2, \ldots, X_n\right] &= \mathrm{Var}\left[\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x \mid X_1, X_2, \ldots, X_n\right] \\ &= \frac{\sigma^2}{n-1}\left(\frac{n-1}{n} + \frac{(x-\bar{x})^2}{s_x^2}\right)\end{aligned}$$

  and the sampling distribution of $\hat{y}$ is normal given σ^2.

  The variance of Y_{n+1} based on our model consists two parts

$$\begin{aligned}\mathrm{Var}\left[Y_{n+1} \mid X_1, X_2, \ldots, X_{n+1}\right] = &\ \mathrm{Var}\left[\hat{Y} \mid X_1, X_2, \ldots, X_{n+1}\right] \\ &+ \mathrm{Var}\left[\varepsilon_{n+1} \mid X_1, X_2, \ldots, X_{n+1}\right] \\ = &\ \frac{\sigma^2}{n-1}\left(n - \frac{1}{n} + \frac{(x-\bar{x})^2}{s_x^2}\right)\end{aligned}$$